TECHNICAL REPORT
R-116
Aug, 2024

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).

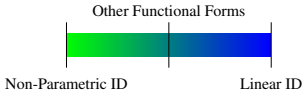
The question of non-parametric identification of causal queries from one layer of the PCH using data from another layer has received a lot of attention in the literature. Various versions of this problem have been studied extensively, from Pearl's celebrated result known as do-calculus to other more 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). 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).

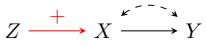
Interestingly, the space between the two extremes of the spectrum in Fig. 1a has received relatively less attention, and yet many interesting possibilities exist for considering effect identification under other functional form assumptions. These include examples such as additive noise models (Peters, 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, we begin with the following example.

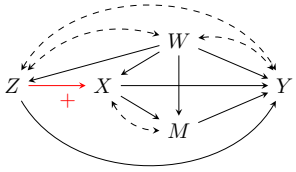
Example 1 (Local Average Treatment Effect or LATE (Imbens and Angrist 1994))**.** *Consider the instrumental variable setting in Fig. 1b with variables X (binary), Y, and an instrument Z (binary). The Local Average Treatment Effect*



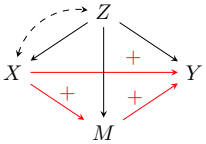
(a) Identification spectrum.



(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

(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 401k dataset studied in (Abadie 2003), represented here by the causal diagram in Fig. 1c, with the following variables: income (W), 401k eligibility (Z), 401k participation (X), net financial assets (M), and total wealth (Y). We wish to compute the LATE of 401k 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. 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 $W = w$,

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

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

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

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

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

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:

(i) 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.

(ii) 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).

(iii) 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 $\mathrm{pa}(V_i) \subseteq \mathbf{V}$ 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.

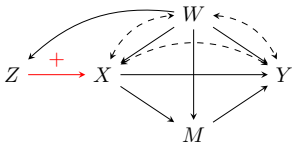
The lemma can be used to identify PoCs in CMGA shown Fig. 1d. Note that the condition is necessary to guarantee that all SCMs consistent with the CGMA satisfy Eq. 10. Hence, we remark that assuming global monotonicity, as done in previous works including (Pearl 2022), is equivalent to assuming these 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 $t \leq t'$.

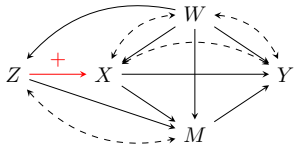
Lemma 2 (Monotonicity Reduction Lemma (MRL))**.** *Let $\mathbf{T}, \mathbf{S}$ be a partition of the parents of W. such that $\mathbf{T}$ and $\mathbf{S}$ are the set of monotonic and non-monotonic parents of W respectively. Let $P(Y_*, W_{\mathbf{t},\mathbf{s}} = w, W_{\mathbf{t}',\mathbf{s}} = w')$ 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*

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



(a) LNDE setting.



(b) LNDE+ setting.

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

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

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

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:

$$P(X_{z_0} \leq x < X_{z_1}) = P(X_{z_0} \leq x) - P(X_{z_1} \leq x) \tag{14}$$

$$= P(X_{z_1} > x) - P(X_{z_0} > x) \tag{15}$$

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:*

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

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

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

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

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:*

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

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

$$\sum_{m,z} P(m \mid w, z, x_0)P(z \mid w) \Big[P(y \mid w, m, x_i, z_i)P(x_i \mid w, z_i) - P(y \mid w, m, x_i, z_{1-i})P(x_i \mid w, z_{1-i})\Big]. \tag{19}$$

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).

$$\text{PN} := P(Y_{x_0} = 0 \mid x_1, y_1) \tag{20}$$

$$\text{PS} := P(Y_{x_1} = 1 \mid x_0, y_0) \tag{21}$$

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 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

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

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 $x \neq x'$.

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

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

---

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

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-factor that needs 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 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 ideal 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 $w \neq w'$*
2. *There exists $W_{\mathbf{t_1}} = w_1, W_{\mathbf{t_2}} = w_2 \in \mathbf{W}_* = \mathbf{w}_*$, such that $\mathbf{t_1} < \mathbf{t_2}$ and $w_1 > w_2$.*

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 Inconsistency* 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.

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: $\mathbf{Q} = \{+1, \mathbf{W}_* = \mathbf{w}_*\}, \mathbf{z}_* = \mathbf{y}_*, \mathbf{V} = \mathbf{V}(\mathbf{W}_*)$

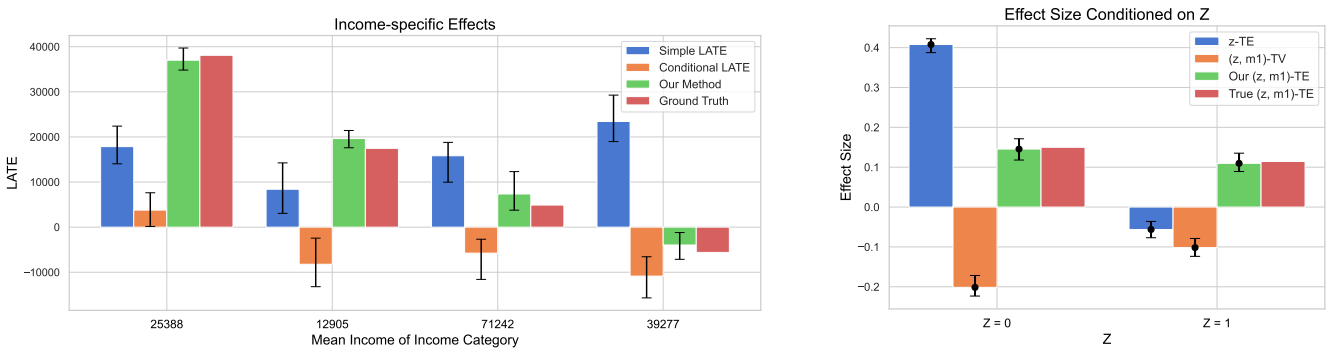
1: $\mathbf{C}_{i*} = \mathbf{c}_{i*}$ $(i \in [k]) \leftarrow$ FACTORIZE$(\mathbf{T}_* = \mathbf{t}_*, G)$
2: **for** $V \in \mathbf{V}$ and $s, \mathbf{T}_* = \mathbf{t}_* \in \mathbf{Q}$ **do**
3: **if** $V \in C_i \setminus Y$ and causes inconsistency in $\mathbf{C}_{j*} = \mathbf{c}_{j*}$ for any j **then**
4: 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: **end if**
6: **end for**
7: **for** $s, \mathbf{T}_* = \mathbf{t}_* \in \mathbf{Q}$ **do**
8: **if** $\mathbf{T}_* = \mathbf{t}_*$ is impossible, remove item from $\mathbf{Q}$
9: Apply Simplification Rule with M for all variables
10: **if** any condition of Lem. 6 holds, **return** FAIL.
11: **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: **for** $s, \mathbf{T}_* = \mathbf{t}_* \in \mathbf{Q}$ **do**
15: **if** the conditions of Difference Rule are not applicable, **continue**
16: **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: **if** Differene Rule cannot be applied, **return** FAIL
18: Replace $(s, \mathbf{T}_* = \mathbf{t}_*)$ with $s, \mathbf{T}'_* = \mathbf{t}'_*$ and $(-s, \mathbf{T}''_* = \mathbf{t}''_*)$
19: **end for**
20: **end for**
21: **for** $s, \mathbf{T}_* = \mathbf{t}_* \in \mathbf{Q}$ **do**
22: **if** any factors in the ctf-factorization of $\mathbf{T}_* = \mathbf{t}_*$ is inconsistent, **return** FAIL.
23: **end for**

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 401k 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 addresses effectively.



(a) (b)

Figure 3: Experiments comparing the newly proposed method versus different baselines. (a) LATE computed on 401k 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.

To evaluate the performance of these methods, we first discretize income, net financial assets, and total wealth into groups corresponding roughly to quartiles, with the group average taken as the representative value. Using this 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, naive-LATE, and conditional-LATE, the expressions for which are shown in Ex. 1. We also show a 95% confidence interval with bootstrapping. The ground truth is also depicted in Fig. 3a.

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 with values from $\{a, c, n\}$.

$$Z = U_Z; X = U_X, \quad P(U_X = 1) = P(U_Z = 1) = 0.5$$

$$M = \begin{cases} \mathbb{1}\{U_m^1 = a\} + X \cdot \mathbb{1}\{U_m^1 = c\} & \text{if } Z = 1 \\ \mathbb{1}\{U_m^0 = a\} + X \cdot \mathbb{1}\{U_m^0 = c\} & \text{if } Z = 0 \end{cases}$$

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

	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

.

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 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)-TE$_{x_0,x_1}(y)$. However, the (z, m)-TV quantity does not equal the (z, m)-TE$_{x_0,x_1}(y)$ effect, and using 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 the experiments are in Appendix E.

5 Conclusions

In conclusion, this work represents a significant step toward encoding and utilizing local monotonicity constraints into the graphical approach to causality. Our proposed lemmas and algorithms broaden the scope of identifiable counterfactual queries, extending existing methods to more complex settings, as also witnessed by real-world examples. Future work could explore additional shape constraints (e.g., convexity/concavity) to further extend the graphical approach for query identification under shape constraints.

Acknowledgements

This research is supported in part by the NSF, 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.; 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*.

Supplementary Material *Counterfactual Identification Under Monotonicity Constraints*

A Background and Previous Results

A.1 Local Average Treatment Effect

Let X be a binary treatment variable and Y be its effect. Define an instrument variable Z, which is independent of Y_{x_0}, Y_{x_1} and X is dependent on Z. The local average treatment effect is the effect of X on Y for the group of units whose treatments comply with the instrument Z and is given quantitatively as follows:

$$\text{LATE} := \mathbb{E}[Y_{x_1} - Y_{x_0} \mid X_{z_0} = 0, X_{z_1} = 1] \tag{26}$$

(Imbens and Angrist 1994) shows that under the following two assumptions, LATE is identifiable.

Assumption 1 (Existence of Instruments)**.** *Let Z be a random variable such that*

1. *For all $z \in D(Z)$, Y_{x_0}, Y_{x_1}, X_z are jointly independent of Z, and,*
2. *$P(X \mid Z)$ is a non-trivial function of Z.*

Assumption 2 (Monotonicity)**.** *For all units, $X_{z_1} \geq X_{z_0}$.*

Under these two assumptions, we have

$$\begin{aligned}
&\mathbb{E}[Y \mid z_1] - \mathbb{E}[Y \mid z_0] && (27)\\
&= \mathbb{E}[Y_{x_1} \cdot X_{z_1} + Y_{x_0} \cdot (1 - X_{z_1}) \mid z_1] - \mathbb{E}[Y_{x_1} \cdot X_{z_0} + Y_{x_0} \cdot (1 - X_{z_0}) \mid z_0] && (28)\\
&\overset{(i)}{=} \mathbb{E}[(X_{z_1} - X_{z_0}) \cdot (Y_{x_1} - Y_{x_0})] && (29)\\
&\overset{(ii)}{=} P(X_{z_1} - X_{z_0} = 1) \cdot \mathbb{E}[Y_{x_1} - Y_{x_0} \mid X_{z_1} - X_{z_0} = 1] && (30)
\end{aligned}$$

(i) follows from Assumption 1 and (ii) follows from Assumption 2. Hence,

$$\text{LATE} = \frac{\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0]}{\mathbb{E}[X \mid Z = 1] - \mathbb{E}[X \mid Z = 0]}. \tag{31}$$

For more details, refer to (Imbens and Angrist 1994).

A.2 Assumptions for Identification of LNDE and LNIE

(Yamamoto 2013) proposes that LNDE and LNIE can be identified under the following assumptions, mapped to the notation used in this paper.

1. **Assumption 1: Exclusion Restriction**

$$M_{z,x}(\mathbf{u}) = M_{z',x}(\mathbf{u}) \qquad Y_{z,x,m}(\mathbf{u}) = Y_{z',x,m}(\mathbf{u}) \tag{32}$$

for all $z, z' \in \{0,1\}, t \in \{0,1\}, m \in \mathcal{M}, \mathbf{u} \in \mathbf{U}$

2. **Assumption 2: Monotone Treatment Reception**

$$X_{z_0}(\mathbf{u}) \leq X_{z_1}(\mathbf{u}) \tag{33}$$

for all $\mathbf{u} \in \mathbf{U}$.

3. **Assumption 3: Conditionally Ignorable Treatment Assignment**

$$\{Y_{x,m}, M_{x'}, X_z : x, x', z \in \{0,1\}, m \in \mathcal{M}\} \perp\!\!\!\perp Z \mid W \tag{34}$$

4. **Assumption 4: Conditionally Ignorable Observed Mediator among Compliers**

$$Y_{x',m} \perp\!\!\!\perp M_x \mid X = x, W, X_{z_0} = 0, X_{z_1} = 1 \tag{35}$$

for all $x, x' \in \{0,1\}, m \in \mathcal{M}$.

A.3 Counterfactual Identification

In this section, we provide the necessary background needed for understanding counterfactual evaluation and identification. Most of the results presented here are from (Correa, Lee, and Bareinboim 2021). We first begin by defining counterfactual distributions:

Definition 4 (Counterfactual Distribution). *An SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$ induces a family of joint distributions over counterfactual events $\mathbf{y}_{\mathbf{x}}, \ldots, \mathbf{Z}_{\mathbf{w}}$ for any $\mathbf{Y}, \mathbf{Z}, \ldots, \mathbf{X}, \mathbf{W} \subseteq \mathbf{V}$*

$$P^{\mathcal{M}}(\mathbf{y}_{\mathbf{x}}, \ldots, \mathbf{z}_{\mathbf{w}}) = \sum_{\substack{\mathbf{u} | \mathbf{Y}_{\mathbf{x}} = y, \\ \ldots, \mathbf{Z}_{\mathbf{w}} = \mathbf{z}}} P(\mathbf{u}) \tag{36}$$

The above definition describes how counterfactual distributions are defined through the elements of the SCM, namely the functional mechanisms $\mathcal{F}$ and the distribution over the exogenous variables $P(\mathbf{u})$. Since the latent variables $\mathbf{U}$ are unobserved, a key challenge in practice is on how to evaluate the right-hand side of Equation 4 based on observational/interventional data and the causal diagram (this task is usually called *identification*). (Correa, Lee, and Bareinboim 2021) proposes ways to identify counterfactual distributions from observational distributions. Some definitions and background needed to understand their identification algorithm are provided below.

Definition 5 (Ancestors of a counterfactual). *Let $Y_{\mathbf{X}}$ be a counterfactual term where $Y \in \mathbf{V}, \mathbf{X} \subseteq \mathbf{V}$. Then the set of ancestors of $Y_{\mathbf{X}}$, denoted by $An(Y_{\mathbf{X}})$ consists of each $W_{\mathbf{Z}}$, such that $W \in An(Y)_{G_{\underline{\mathbf{X}}}}$ and $\mathbf{z} = \mathbf{x} \cap An(W)_{G_{\overline{\mathbf{X}}}}$.*

Similarly, we can define the *Parents of a counterfactual* to be $W_{\mathbf{z}}$ such that $W \in Pa(Y)_{G_{\underline{X}}}$ and $\mathbf{z} = \mathbf{x} \cap An(W)_{G_{\overline{X}}}$.

Definition 6 (Counterfactual Factor (ctf-factor)). *A counterfactual factor is a distribution of the form*

$$P(W_{1[\mathbf{pa}_1]} = w_1, \ldots, W_{l[\mathbf{pa}_l]} = w_l) \tag{37}$$

where each $W_i \in \mathbf{V}$ and W_i can be equal to W_j for some $i, j \in \{1, \ldots, l\}$.

Theorem 3 (Counterfactual factorization). *Let $P(\mathbf{W}_* = \mathbf{w}_*)$ is a ctf-factor and let $W_1 < W_2 < \ldots$ be a topological order over the variables in $G[V(\mathbf{W})]$. If $\mathbf{C}_1, \mathbf{C}_2, \ldots$ are c-components in the same graph, define $\mathbf{C}_{j*} = \{W_{\mathbf{pa}_w} \in \mathbf{W}_* \mid W \in \mathbf{C}_j\}$ and $\mathbf{c}_{j*}$ are the values in $\mathbf{w}_*$ corresponding to the values in $\mathbf{C}_{j*}$, then*

$$P(\mathbf{W}_* = \mathbf{w}_*) = \prod_j P(\mathbf{C}_{j*} = \mathbf{c}_{j*}) \tag{38}$$

(Correa, Lee, and Bareinboim 2021) introduced the concept of inconsistency in a ctf-factor and shows that if a ctf-factor is inconsistent, then the ctf-factor is non-ID.

Definition 7 (Inconsistent ctf-factor). *$P(\mathbf{W}_* = \mathbf{w}_*)$ is an inconsistent ctf-factor if it is a ctf-factor, $G[V(\mathbf{W}_*)]$ has a single c-component, and one of the following conditions hold*

1. ***Parent-Child Inconsistency:*** *there 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 Inconsistency:*** *there 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 $t \neq t'$.*

The counterfactual query may need to be simplified before it can be represented as a ctf-factor. Unnesting and minimization are two ways to do that.

Theorem 4 (Counterfactual Unnesting Theorem (CUT)). *Let $\hat{\mathbf{X}}, \hat{\mathbf{Z}}$ be any natural interventions on disjoint sets $\mathbf{X}, \mathbf{Z} \subseteq \mathbf{V}$, and $\mathbf{Y} \subseteq \mathbf{V}$ be disjoint from $\mathbf{X}, \mathbf{Z}$ such that $\mathbf{X} \in An(\mathbf{Y})$. Then $P(\mathbf{Y}_{\hat{\mathbf{X}}, \hat{\mathbf{Z}}} = \mathbf{y})$ is identifiable iff $P(\mathbf{Y}_{\hat{\mathbf{Z}}, \mathbf{x}} = \mathbf{y}, \hat{\mathbf{X}} = \mathbf{x})$ and is given by*

$$P(\mathbf{Y}_{\hat{\mathbf{X}}, \hat{\mathbf{Z}}} = \mathbf{y}) = \sum_{\mathbf{x}} P(\mathbf{Y}_{\hat{\mathbf{Z}}, \mathbf{x}} = \mathbf{y}, \hat{\mathbf{X}} = \mathbf{x}) \tag{39}$$

Definition 8 (Interventional Minimization). *Let $\|\mathbf{Y}_*\| := \cup_{Y_{\hat{\mathbf{X}}} \in \mathbf{Y}_*}$ such that $\left\|Y_{\hat{\mathbf{X}}}\right\| := Y_{\hat{\mathbf{Z}}}$ where $\mathbf{Z} = \mathbf{X} \cap An(Y)_{G_{\bar{X}}}$ and $\hat{\mathbf{Z}}$ is consistent with $\left\|\hat{\mathbf{X}}\right\|$, $\|\mathbf{x}\| = \mathbf{x}$ and $\|\emptyset\| = \emptyset$. Then $\mathbf{Y}_* = \|\mathbf{Y}_*\|$.*

If the counterfactual query to be evaluated is in the conditional form, then (Correa, Lee, and Bareinboim 2021) proposes the following to get the final counterfactual query we need to evaluate.

Definition 9 (Ancestral components). *Let $\mathbf{X}_*, \mathbf{Y}_*$ be two sets of counterfactual variables and G be the causal diagram. Then the ancestral components induced by $\mathbf{W}_* = \mathbf{X}_* \cup \mathbf{Y}_*$ given $\mathbf{X}_*$ are sets $\mathbf{A}_{1*}, \mathbf{A}_{2*}, \ldots$ that form a partition over $An(\mathbf{W}_*)$, made of the union of ancestral sets $An(W_{\mathbf{t}})_{G_{\underline{\mathbf{X}_*(W_{\mathbf{t}})}}}$ for $W_{\mathbf{t}} \in \mathbf{W}_*$. Sets $An(W_{1[\mathbf{t}_1]})_{G_{\underline{\mathbf{X}_*(W_{1[\mathbf{t}_1]})}}}$ and $An(W_{2[\mathbf{t}_2]})_{G_{\underline{\mathbf{X}_*(W_{2[\mathbf{t}_2]})}}}$ are put together if they are not disjoint or there exists a bidirected arrow in G connecting variables in those sets. Here $\overline{\mathbf{X}_*(W_{\mathbf{t}})} = V(\|X\|_* \cap An(W_{\mathbf{t}})$.*

This is used for the following reduction.

Lemma 7 (Conditional Query Reduction)**.** *Let $\mathbf{X}_*, \mathbf{Y}_*$ be two sets of counterfactual variables and $\mathbf{D}_*$ be the set of variables in the same ancestral component given $\mathbf{X}_*$ as any variable in $\mathbf{Y}_*$, then*

$$P(\mathbf{Y}_* = \mathbf{y}_* \mid \mathbf{X}_* = \mathbf{x}_*) = \frac{\sum_{\mathbf{d}_* \setminus (\mathbf{y}_* \cup \mathbf{x}_*)} P(\wedge_{D_t \in \mathbf{D}_*} D_{\mathbf{pa}_d} = d)}{\sum_{\mathbf{d}_* \setminus \mathbf{x}_*} P(\wedge_{D_t \in \mathbf{D}_*} D_{\mathbf{pa}_d} = d)} \tag{40}$$

where $\mathbf{pa}_d$ is consistent with $\mathbf{t}$ and $\mathbf{d}_*$ for each $D_\mathbf{t} \in \mathbf{D}_*$. Moreover, $P(\mathbf{Y}_* = \mathbf{y}_* \mid \mathbf{X}_* = \mathbf{x}_*)$ is ID iff $P(\wedge_{D_t \in \mathbf{D}_*} D_{\mathbf{pa}_d} = d)$ is ID.

For more details, refer to (Correa, Lee, and Bareinboim 2021).

B Proofs

B.1 Proof of Lem. 1

Proof. **Proof of first part:** First, we use Theorem 1 from (VanderWeele and Robins 2010) to convert all edges on all paths from X to Y to be positive monotonic. Consider a particular assignment of exogenous variables $\mathbf{u}$. Suppose $X(\mathbf{u}) = x_0$ and we are changing it to x_1. Now, for all children Z of X on a path from X to Y, we have, $Z_{x_1}(\mathbf{u}) \geq Z_{x_0}(\mathbf{u})$.

Consider any node M on the path from X to Y. None of the non-monotonic parents of M gets affected since that would violate our assumption that all paths from X to Y consist of only monotonic edges. If the monotonic parents $Pa(M)_{x_1}(\mathbf{u}) \geq Pa(M)_{x_0}(\mathbf{u})$, then $M_{x_1}(\mathbf{u}) \geq M_{x_0}(\mathbf{u})$.

We know the base case is true for children of X. Hence, by induction on the children, we get $Y_{x_1}(\mathbf{u}) \geq Y_{x_0}(\mathbf{u})$. Similarly, it can be shown when $X(\mathbf{u}) = x_1$.

Proof of second part: In this case, we will find an assignment such that when there is a non-monotonic edge, the condition of monotonicity is violated. Let on the path from X to Y, the edge $W \to Z$ is non-monotonic. Let's call this path P. If a node is not in P it is just a random coin toss. For all nodes in P except Z, they can be equal to $0, 1$ or parent on this path. For Z, it can be either $0, 1, W, 1 - W$.

Consider the assignment of exogenous variables $\mathbf{u}$ as follows: all nodes in P are equal to their parents, except for Z, which is $1 - W$. Now, $W_{x_1}(\mathbf{u}) = 1, W_{x_0}(\mathbf{u}) = 0$, which implies $Z_{x_1}(\mathbf{u}) = 0, Z_{x_0}(\mathbf{u}) = 1$ and $Y_{x_1}(\mathbf{u}) = 0, Y_{x_0}(\mathbf{u}) = 1$, violating the condition of monotonicity. □

Note that the condition of Lem. 1 is not necessary for every SCM to have the global monotonicity, that is, there exists SCMs which does not satisfy the condition, but can still have global monotonicity. The situation is illustrated through two SCMs M_1, M_2 below:

$$Z = U_Z, \quad \text{(both } M_1, M_2\text{)} \tag{41}$$

$$W = \begin{cases} U_W \cdot (1 - Z) + (1 - U_W) \cdot Z & (M_1) \\ U_W & (M_2) \end{cases} \tag{42}$$

$$X = \mathbb{1}(U_X = 1) + \mathbb{1}(U_X = 2) \cdot (W \oplus Z) \tag{43}$$

Here U_Z, U_W are binary with $P(U_Z = 1) = P(U_W = 1) = 0.5$. U_X is ternary and $P(U_X = 1) = P(U_X = 2) = P(U_X = 3) = 1/3$. Note that for model M_1, $X_{z_0}(\mathbf{u}) = X_{z_1}(\mathbf{u})$ implying that $Z \to X$ is monotonic. However, in M_2

$$X_{z_1}(U_Z, U_W = 1, U_X = 2) < X_{z_0}(U_Z, U_W = 1, U_X = 2) \tag{44}$$

implying that $Z \to Y$ is not monotonic. Observe that M_1, M_2 have the same observational and interventional distribution. So, the two models cannot be distinguished by observations or experiments.

B.2 Proof of Lem. 2

Proof. Let the ctf-factor be of the form $P(Y_*, W_{\mathbf{t},\mathbf{s}} = w, W_{\mathbf{t}',\mathbf{s}} = w')$ where $\mathbf{t}' \geq \mathbf{t}$, and W is binary. By definition of monotonicity, we have

$$W_{\mathbf{t},\mathbf{s}}(\mathbf{u}) \leq W_{\mathbf{t}',\mathbf{s}}(\mathbf{u}) \tag{45}$$

The definition has the following two implications:

1. No $\mathbf{u}$ satisfies $W_{\mathbf{t},\mathbf{s}}(\mathbf{u}) = 1, W_{\mathbf{t}',\mathbf{s}}(\mathbf{u}) = 0$
2. $W_{\mathbf{t},\mathbf{s}}(\mathbf{u}) = 0, W_{\mathbf{t}',\mathbf{s}}(\mathbf{u}) = 0 \iff W_{\mathbf{t}',\mathbf{s}}(\mathbf{u}) = 0$
3. $W_{\mathbf{t},\mathbf{s}}(\mathbf{u}) = 1, W_{\mathbf{t}',\mathbf{s}}(\mathbf{u}) = 1 \iff W_{\mathbf{t},\mathbf{s}}(\mathbf{u}) = 1$

We will use the above implications to first prove the Simplification Rule and then the Difference Rule.

Proof of Simplification Rule:

$$P(Y_*, W_{\mathbf{t},\mathbf{s}} = 0, W_{\mathbf{t}',\mathbf{s}} = 0) = \sum_{\mathbf{u}} P(\mathbf{u}) \mathbb{1}\{Y_*(\mathbf{u}) = y_*, W_{\mathbf{t},\mathbf{s}}(\mathbf{u}) = 0, W_{\mathbf{t}',\mathbf{s}}(\mathbf{u}) = 0) \tag{46}$$

$$= \sum_{\mathbf{u}} P(\mathbf{u}) \mathbb{1}\{Y_*(\mathbf{u}) = y_*, W_{\mathbf{t}',\mathbf{s}}(\mathbf{u}) = 0) \quad \text{(Definition of Monotonicity)} \tag{47}$$

$$= P(Y_*, W_{\mathbf{t}',\mathbf{s}} = 0) \tag{48}$$

$$P(Y_*, W_{\mathbf{t},\mathbf{s}} = 1, W_{\mathbf{t}',\mathbf{s}} = 1) = \sum_{\mathbf{u}} P(\mathbf{u}) \mathbb{1}\{Y_*(\mathbf{u}) = y_*, W_{\mathbf{t},\mathbf{s}}(\mathbf{u}) = 0, W_{\mathbf{t}',\mathbf{s}}(\mathbf{u}) = 0) \tag{49}$$

$$= \sum_{\mathbf{u}} P(\mathbf{u}) \mathbb{1}\{Y_*(\mathbf{u}) = y_*, W_{\mathbf{t},\mathbf{s}}(\mathbf{u}) = 1) \quad \text{(Definition of Monotonicity)} \tag{50}$$

$$= P(Y_*, W_{\mathbf{t},\mathbf{s}} = 1) \tag{51}$$

This proves Rule 1 of MRL. Now, we use this to prove the Difference Rule.

Proof of Difference Rule:

$$P(Y_*, W_{\mathbf{t}',\mathbf{s}} = 1) = P(Y_*, W_{\mathbf{t},\mathbf{s}} = 1, W_{\mathbf{t}',\mathbf{s}} = 1) + P(Y_*, W_{\mathbf{t},\mathbf{s}} = 0, W_{\mathbf{t}',\mathbf{s}} = 1) \tag{52}$$

$$= P(Y_*, W_{\mathbf{t},\mathbf{s}} = 1) + P(Y_*, W_{\mathbf{t},\mathbf{s}} = 0, W_{\mathbf{t}',\mathbf{s}} = 1) \quad \text{(Simplification Rule)} \tag{53}$$

This can be rewritten as

$$P(Y_*, W_{\mathbf{t},\mathbf{s}} = 0, W_{\mathbf{t}',\mathbf{s}} = 1) = P(Y_*, W_{\mathbf{t}',\mathbf{s}} = 1) - P(Y_*, W_{\mathbf{t},\mathbf{s}} = 1) \tag{54}$$

This proves Eq. 11. Eq. 12 can be proved similarly.

□

B.3 Proof of Prop. 3

Causal Graph in Fig. 1c

Proof. First, we will show how to compute $P(Y_{x_1}, X_{z_0} = 0, X_{z_1} = 1)$ for the causal graph in Fig. 1c. Once we can compute this term, the claim follows. Note that, $An(Y_x) = \{Y_x, M_x, W, Z\}$, $An(X_z) = \{X_z, W\}$.

$$P(Y_{x_1} = y, X_{z_0} = 0, X_{z_1} = 1) = \sum_{w,z,m} P(Y_{x_1 mzw} = y, M_{x_1 w} = m, X_{z_0 w} = 0, X_{z_1 w} = 1, Z_w = z, W = w) \tag{55}$$

$$\overset{(i)}{=} \sum_{w,z,m} P(Y_{x_1 mzw} = y, Z_w = z, W) P(M_{x_1 zw} = m, X_{z_0 w} = 0, X_{z_1 w} = 1) \tag{56}$$

$$\overset{(ii)}{=} \sum_{w,z,m} P(Y_{x_1 mzw} = y, Z_w = z, W) \big[P(M_{x_1 w} = m, X_{z_1 w} = 1) - \tag{57}$$

$$P(M_{x_1 w} = m, X_{z_0 w} = 1) \big] \tag{58}$$

$$= \sum_{w,z,m} P(y, z, w \mid do(x, m)) \big[P(m, x_1 \mid do(w, z_1, y)) - P(m, x_1 \mid do(w, z_0, y)) \big] \tag{59}$$

$$= \sum_{w,z,m} P(y \mid w, z, x_1, m) P(w, z) \big[P(m, x_1 \mid w, z_1) - P(m, x_1 \mid w, z_0) \big] \tag{60}$$

(i) follows from counterfactual factorization and (ii) follows from Difference Rule Eq. 11. For clarity, let's denote

$$Q_1(y) = \sum_{m,z} P(y \mid w, z, x_1, m) P(z \mid w) \big[P(m, x_1 \mid w, z_1) - P(m, x_1 \mid w, z_0) \big] \tag{61}$$

Using a similar calculation as above and replacing x_1 with x_0, and applying Difference Rule Eq. 12 instead of Eq. 11. Then,

$$Q_0(y) = \sum_{m,z} P(y \mid w, z, x_0, m) P(z \mid w) \big[P(m, x_0 \mid w, z_0) - P(m, x_0 \mid w, z_1) \big] \tag{62}$$

By definition of expectation, we get

$$\text{LATE:} = \frac{\sum_{w,y} y \cdot Q_1(y)P(w)}{\sum_{w,y} Q_1(y)P(w)} - \frac{\sum_{w,y} y \cdot Q_0(y)P(w)}{\sum_{w,y} Q_0(y)P(w)}, \tag{63}$$

□

Causal Graph in Fig. 2b

Proof. We use a similar calculation as in the proof for Fig. 1c.

$$P(Y_{x_1}, X_{z_0} = 0, X_{z_1} = 1) \tag{64}$$
$$= \sum_{w,z,m} P(Y_{x_1 mw}, M_{x_1 zw} = m, X_{z_0 w} = 0, X_{z_1 w} = 1, Z_w = z, w) \tag{65}$$
$$= \sum_{w,z,m} P(Y_{z_1 mw}, X_{z_0 w} = 0, X_{z_1 w} = 1, w) P(M_{x_1 zw} = m, Z_w = z) \quad \text{(Ctf-factorization)} \tag{66}$$
$$= \sum_{w,z,m} \big[P(Y_{z_1 mw}, X_{z_1 w} = 1, w) - P(Y_{z_1 mw}, X_{z_0 w} = 1, w)\big] P(M_{x_1 zw} = m, Z_w = z) \text{ (Diff. Rule)} \tag{67}$$
$$= \sum_{w,z,m} \big[P(y, x_1, w \mid do(z_1, m)) - P(y, x_1, w \mid do(z_0, m))\big] P(m, z \mid do(w, x_0, y)) \tag{68}$$
$$= \sum_{w,z,m} P(w)\big[P(y \mid w, z_1, x_1, m)P(x_1 \mid w, z_1) - P(y \mid w, z_0, x_1, m)P(x_1 \mid w, z_0)\big] P(m \mid w, x_1, z)P(z \mid w) \tag{69}$$

For clarity of notation, let's use:

$$R_1(y) = \sum_{m,z} \big[P(y \mid w, z_1, x_1, m)P(x_1 \mid w, z_1) - P(y \mid w, z_0, x_1, m)P(x_1 \mid w, z_0)\big] P(m \mid w, x_1, z)P(z \mid w) \tag{70}$$

Similarly,

$$R_0(y) = \sum_{m,z} \big[P(y \mid w, z_0, x_0, m)P(x_0 \mid w, z_0) - P(y \mid w, z_1, x_0, m)P(x_0 \mid w, z_1)\big] P(m \mid w, x_0, z)P(z \mid w) \tag{71}$$

By definition of expectation, we get

$$\text{LATE:} = \frac{\sum_{w,y} y \cdot R_1(y)P(w)}{\sum_{w,y} R_1(y)P(w)} - \frac{\sum_{w,y} y \cdot R_0(y)P(w)}{\sum_{w,y} R_0(y)P(w)}, \tag{72}$$

□

B.4 Proof of Prop. 4

Proof. We follow a similar proof technique as employed for Proposition 3.

$$P(Y_{x_1 M_{x_0}}, X_{z_0} = 0, X_{z_1} = 1) \tag{73}$$
$$= \sum_{w,z,m} P(Y_{x_1 mw}, M_{x_0 zw} = m, X_{z_0 w} = 0, X_{z_1 w} = 1, Z_w = z, w) \tag{74}$$
$$= \sum_{w,z,m} P(Y_{z_1 mw}, X_{z_0 w} = 0, X_{z_1 w} = 1, w) P(M_{x_0 zw} = m, Z_w = z) \quad \text{(Ctf-factorization)} \tag{75}$$
$$= \sum_{w,z,m} \big[P(Y_{z_1 mw}, X_{z_1 w} = 1, w) - P(Y_{z_1 mw}, X_{z_0 w} = 1, w)\big] P(M_{x_0 zw} = m, Z_w = z) \text{ (Diff. Rule)} \tag{76}$$
$$= \sum_{w,z,m} \big[P(y, x_1, w \mid do(z_1, m) - P(y, x_1, w \mid do(z_0, m)\big] P(m, z \mid do(w, x_0, y)) \tag{77}$$
$$= \sum_{w,z,m} P(w)\big[P(y \mid w, z_1, x_1, m)P(x_1 \mid w, z_1) - P(y \mid w, z_0, x_1, m)P(x_1 \mid w, z_0)\big] P(m \mid w, x_0, z)P(z \mid w) \tag{78}$$

For clarity, let's call the following quantity $T_1(y)$.

$$T_1(y) = \sum_{m,z} \big[P(y \mid w, z_1, x_1, m)P(x_1 \mid w, z_1) - P(y \mid w, z_0, x_1, m)P(x_1 \mid w, z_0)\big]P(m \mid w, x_0, z)P(z \mid w) \tag{79}$$

Similarly,

$$T_0(y) = \sum_{m,z} \big[P(y \mid w, z_0, x_0, m)P(x_0 \mid w, z_0) - P(y \mid w, z_1, x_0, m)P(x_0 \mid w, z_1)\big]P(m \mid w, x_0, z)P(z \mid w) \tag{80}$$

By definition of expectation, we have

$$\text{LNDE: } = \frac{\sum_{w,y} y \cdot T_1(y)P(w)}{\sum_{w,y} \cdot T_1(y)P(w)} - \frac{\sum_{w,y} y \cdot T_0(y)P(w)}{\sum_{w,y} \cdot T_0(y)P(w)} \tag{81}$$

LNIE can be obtained similarly or simply by subtracting LNDE from LATE. □

In the next part, we show that the addition of bidirected edges to the graph makes LATE non-ID and, as a consequence, LNDE and LNIE.

Proof. **Case 1: Unobserved confounder between** Z, X. Define two SCMs M_1, M_2 as follows:

$$Z = U_Z, \quad P(U_Z = 1) = P(U_Z = 0) = 0.5 \tag{82}$$
$$Y = U_Y, \quad P(U_Y = 1) = P(U_Y = 0) = 0.5 \tag{83}$$
$$X = \begin{cases} Z \cdot U_Z \cdot U_Y + (1 - U_Z \cdot U_Y) \cdot U_X & \text{in } M_1 \\ Z \cdot (1 - U_Z) \cdot (1 - U_Y) + (U_Z + U_Y) \cdot U_X & \text{in } M_2 \\ U_X, \quad P(U_X = 0) = P(U_X = 1) = 0.5 & \text{in other cases} \end{cases} \tag{84}$$

Here $P(U_X = 0) = 0.5$ is there to make the distribution positive. The observational distribution is the same in both the models, while $P^1(Y_{x_1} \mid X_{z_0} = 0, X_{z_1} = 1) = 1$ and $P^1(Y_{x_1} \mid X_{z_0} = 0, X_{z_1} = 1) = 0$.

Case 2: Unobserved confounder between M, Y Define two SCMs M_1, M_2 as follows, where U_M, U_X are of canonical types $\{a, c, n\}$

$$Z = U_Z, \quad P(U_Z = 1) = P(U_Z = 0) = 0.5 \tag{85}$$
$$M = (U_M = a) + Z \cdot (U_M = c) \tag{86}$$
$$X = (U_X = a) + Z \cdot (U_X = c) \tag{87}$$
$$Y = \begin{cases} 0 & \text{if } U_X = c, U_M = c \quad (M_1) \\ M \oplus X & \text{if } U_X = c, U_M = c \quad (M_2) \\ U_Y, \quad P(U_Y = 1) = P(U_Y = 0) = 0.5 & \text{otherwise} \end{cases} \tag{88}$$

The observational distribution is the same in both the models, but $P^1(Y_{x_1} \mid U_X = c, U_M = c) \neq P^2(Y_{x_1} \mid U_X = c, U_M = c)$. Hence, LATE is non-ID.

Case 3: Unobserved confounder between X **and** M. Define two SCMs T_1, T_2 as follows, where U_Z, U_M, U_X, U_Y are binary variables with probability of being 1 is 0.5. Also, M_0 and M_1 are the first and second elements of M, respectively.

$$Z = U_Z \tag{89}$$
$$M = (U_Z, U_M) \tag{90}$$
$$X = Z \cdot U_Y \cdot U_M + (1 - U_Y \cdot U_M) \cdot U_X \tag{91}$$
$$Y = \big\{ M_0 \quad \text{if } M_1 \cdot U_Y = 1 \quad (T_1) X \quad \text{if } M_1 \cdot U_Y = 1 \quad (T_2) \text{Bernoulli}(0.5) \quad \text{otherwise} \tag{92}$$

The observational distribution is the same in both models, since $M_0 = U_Z = Z = X$ when $U_M \cdot U_Y = 1$. However, $P^1(Y_{x_1} = 1 \mid X_{z_0} = 0, X_{z_1} = 1) = 0.5$ and $P^2(Y_{x_1} = 1 \mid X_{z_0} = 0, X_{z_1} = 1) = 1$. Similarly, LATE will be 0 in T_1 and 1 in T_2.

The proof technique can be extended to other cases as well. A general construction is shown in Sec. B.7. □

B.5 Proof of Proposition. 5

When $x = x'$, then we can write the terms as

$$P(Y_x = y \mid m, x) = P(y \mid m, x) \tag{93}$$

$$P(Y_x = y' \mid x, m, y) = \begin{cases} 1 & \text{if } y = y' \\ 0 & \text{otherwise} \end{cases} \tag{94}$$

Proof of ID of $P(Y_x \mid x', m)$

Proof. In this part, we will consider $x \neq x'$.

$$An(Y_x) = \{Y_x, M_x, Z\} \tag{95}$$
$$An(X) = \{X, Z\} \tag{96}$$
$$An(M) = \{M, X, Z\} \tag{97}$$
$$An(Y) = \{Y, X, M, Z\} \tag{98}$$

Now, we can write the first quantity as follows:

$$P(Y_x = y, X = x', M = m) = \sum_{m',z} P(Y_x = y, M_x = m', Z = z, X = x', M = m) \tag{99}$$
$$= \sum_{m',z} P(Y_{xm'z} = y, M_{xz} = m', Z = z, X_z = x', M_{x'z} = m) \tag{100}$$
$$= \sum_{m',z} P(y \mid x, m', z) P(z, x') P(M_{xz} = m', M_{x'z} = m) \tag{101}$$

Now, $P(M_{xz} = m', M_{x'z} = m)$ can be computed by exploiting the monotonicity of $X \to M$ and that M is binary. Without loss of generality, assume $x' \geq x$. Then, by monotonicity, we have

$$P(M_{xz} = 0, M_{x'z} = 0) = P(M_{x'z} = 0) = P(m_0 \mid x', z) \tag{102}$$
$$P(M_{xz} = 0, M_{x'z} = 1) = P(M_{x'z} = 1) - P(M_{xz} = 1) = P(m_1 \mid x', z) - P(m_1 \mid x, z) \tag{103}$$
$$P(M_{xz} = 1, M_{x'z} = 0) = 0 \tag{104}$$
$$P(M_{xz} = 1, M_{x'z} = 1) = P(M_{xz} = 1) = P(m_1 \mid x, z) \tag{105}$$

Thus $P(Y_x = y \mid x, m)$ is identifiable for all x, m, y. □

Proof of ID of $P(Y_x \mid x', y', m)$

Proof. In this part, we will consider the expression $P(Y_x = y \mid x', y', m)$. Similar to the previous proof, we have

$$P(Y_x = y, X = x', M = m, Y = y') \tag{106}$$
$$= \sum_{m',z} P(Y_x = y, M_x = m', Z = z, X = x', M = m, Y = y') \tag{107}$$
$$= \sum_{m',z} P(Y_{xm'z} = y, M_{xz} = m', Z = z, X_z = x', M_{x'z} = m, Y_{x'mz} = y') \tag{108}$$
$$= \sum_{m',z} P(z, x') P(M_{xz} = m', M_{x'z} = m) P(Y_{xm'z} = y, Y_{x'mz} = y') \tag{109}$$

Now, $P(M_{xz} = m', M_{x'z} = m)$ can be computed as shown in the proof of identifiability of $P(Y_x \mid x, m)$. Here, we show how to compute $P(Y_{xm'z} = y, Y_{x'mz} = y')$. If $(x, m') < (x', m)$, then

$$P(Y_{xm'z} = 0, Y_{x'mz} = 0) = P(Y_{x'mz} = 0) = P(y_0 \mid x', m, z) \tag{110}$$
$$P(Y_{xm'z} = 0, Y_{x'mz} = 1) = P(y_1 \mid x', m, z) - P(y_1 \mid x, m', z) \tag{111}$$
$$P(Y_{xm'z} = 1, Y_{x'mz} = 0) = 0 \tag{112}$$
$$P(Y_{xm'z} = 1, Y_{x'mz} = 1) = P(Y_{xm'z} = 1) = P(y_1 \mid x, m', z) \tag{113}$$
$$\tag{114}$$

Similarly, the terms can be derived when $(x, m') > (x', m)$. Note that, for cases $x > x', m' < m$ or $x < x', m' > m$, $P(M_{xz} = m', M_{x'z} = m) = 0$. Hence, $P(Y_{xm'z} = y, Y_{x'mz} = y')$ need not be computed. □

Proof of Necessity of $X \to M$ monotonicity:

Proof. Now, we show that the quantity $P(Y_x, m, x')$ is non-ID without the monotonicity on $X \to M$. Consider the subgraph without Z. Now, we will define two models M_1, M_2, which coincides on Layer 1 and Layer 2 distribution, but not on counterfactual distribution. Define functions such that $P(X = 1) = P(X = 1) = 0.5$ and the distribution of Y as

$$P(y_1 \mid x_0, m_0) = 0.1 \tag{115}$$
$$P(y_1 \mid x_0, m_1) = 0.2 \tag{116}$$
$$P(y_1 \mid x_1, m_0) = 0.3 \tag{117}$$
$$P(y_1 \mid x_1, m_1) = 0.4 \tag{118}$$

In M_1, $M = U_M$ and in $M_2, M = U_M \oplus X$. $P(U_M = 1) = 0.5$. The observational and interventional distribution is the same in all the models as $P(x), P(m \mid x), P(y \mid x, m)$ are the same in both the models. Now, $P(Y_x = y, X = x', M = m)$ can be written as

$$P(Y_x = y, X = x', M = m) = \sum_{m'} P(x')P(y \mid x, m')P(M_x = m', M_{x'} = m) \tag{119}$$

Now, in M_1, the second term is non-zero only when $m' = m$ and in M_2, the second term is non-zero only when $m' \neq m$. Then the counterfactual evaluation varies in the two models

$$P^1(Y_x = y, X = x', M = m) = 0.5 \cdot 0.5 \cdot P(y \mid x, m) \tag{120}$$
$$P^2(Y_x = y, X = x', M = m) = 0.5 \cdot 0.5 \cdot P(y \mid x, m') \tag{121}$$

The quantities cannot be equal by our definition of the models.

□

Proof of necessity of $X \to M, M \to Y, X \to Y$ monotonicity: Now, in the next part, we show that local monotonicity on all the three edges $X \to Y, X \to M, M \to Y$ are necessary. We will show that if either of them is non-monotonic, the query at hand becomes non-ID.

Proof. **Case 1:** $X \to Y$ is non-monotonic. Consider the following SCMs M_1, M_2, where the functions for X, M are the same, and that of Y is different.

$$X = U_X \tag{122}$$
$$M = U_M \tag{123}$$
$$Y = \begin{cases} U_Y & (M_1) \\ X \oplus U_Y & (M_2) \end{cases} \tag{124}$$

Here U_X, U_M, U_Y are binary and $P(U_X = 0) = P(U_M = 0) = P(U_Y = 0) = 0.5$. Note that the monotonicity conditions are satisfied as $X \to M, M \to Y$ trivially. The observational and interventional distribution is the same for both of them since $P(x), P(m \mid x), P(y \mid x, m)$ is the same in both models. However,

$$P^1(Y_x \mid x', y) \neq P^2(Y_x \mid x', y) \quad \text{and} \quad P^1(Y_x \mid m, x', y) \neq P^2(Y_x \mid m, x', y) \tag{125}$$

where P^1, P^2 denotes the distribution in M_1, M_2 respectively.

Case 2: $M \to Y$ is non-monotonic. Consider the following SCMs M_1, M_2, where the functions are given as follows

$$X = U_X \tag{126}$$
$$M = 1 \cdot \{U_M = a_M\} + X \cdot \{U_M = c_M\} + 0 \cdot \{U_M = n_M\} \tag{127}$$
$$Y = \begin{cases} X \cdot \{U_Y = c_Y\} + 1 \cdot \{U_Y = a_Y\} + 0 \cdot \{U_Y = n_Y\} & (M_1) \\ X \cdot \{U_Y = c_Y\} + M \cdot \{U_Y = a_Y\} + (1 - M) \cdot \{U_Y = n_Y\} & (M_2) \end{cases} \tag{128}$$

U_M can be a_M, c_M, n_M with probability of $1/3$ each. U_Y can be a_Y, c_Y, n_Y with probability $1/3$. It is easy to see that the monotonicity constraints are satisfied, and the observational distribution is the same from table 2.

Table 2: $P(Y \mid X, M)$ distribution for Case 2

X	M	Y	M_1	M_2
0	0	0	$c_Y + n_Y$	$c_Y + a_Y$
0	0	1	a_Y	n_Y
0	1	0	$c_Y + n_Y$	$c_Y + n_Y$
0	1	1	a_Y	a_Y
1	0	0	n_Y	a_Y
1	0	1	$c_Y + a_Y$	$c_Y + n_Y$
1	1	0	n_Y	n_Y
1	1	1	$c_Y + a_Y$	$c_Y + a_Y$

However, the quantity $P(Y_{x_0} = 0, Y_{x_1} = 1)$ is not identifiable. In M_1

$$P(Y_{x_1} = 1, Y_{x_0} = 0) = P(c_Y) \tag{129}$$

In M_2, however,

$$P(Y_{x_1} = 1, Y_{x_0} = 0) = P(c_Y) + P(c_M)P(a_Y) \tag{130}$$

Hence,

$$P^1(Y_x \mid x', y) \neq P^2(Y_x \mid x', y) \quad \text{and also} \quad P^1(Y_x \mid m, x', y) \neq P^2(Y_x \mid m, x', y) \tag{131}$$

Case 3: $X \to M$ is non-monotonic. Define the canonical SCM M_1, M_2 as follows:

$$X = U_X \tag{132}$$
$$M = X \cdot \{U_M = c_M\} + (1 - X) \cdot \{U_M = d_M\} \tag{133}$$

Now, define the following distribution for the canonical types of Y

Table 3: Each row represents a canonical type for Y

$Y_{x_0 m_0}$	$Y_{x_1 m_0}$	$Y_{x_0 m_1}$	$Y_{x_1 m_1}$	U_Y	M_1	M_2
0	0	0	0	u_1	p_1	p_1
0	0	0	1	u_2	p_2	$p_2 - \epsilon$
0	0	1	1	u_3	p_3	$p_3 + \epsilon$
0	1	0	1	u_4	p_4	$p_4 + \epsilon$
0	1	1	1	u_5	p_5	$p_5 - \epsilon$
1	1	1	1	u_6	p_6	p_6

Note that the observational distribution is the same in both models since $P(x), P(m \mid x)$ are the same and

$$P(y_1 \mid x_0, m_0) = p_6 \tag{134}$$
$$P(y_1 \mid x_1, m_0) = p_4 + p_5 + p_6 \tag{135}$$
$$P(y_1 \mid x_0, m_1) = p_3 + p_5 + p_6 \tag{136}$$
$$P(y_1 \mid x_1, m_1) = 1 - p_1 \tag{137}$$

However, the distribution $P(Y_{x_0} = 0, Y_{x_1} = 1)$ varies in the two models as shown below:

$$P^1(Y_{x_0} = 0, Y_{x_1} = 1) = P(c_M)(1 - p_1 - p_6) + P(d_M)(p_4) \tag{138}$$
$$P^2(Y_{x_0} = 0, Y_{x_1} = 1) = P(c_M)(1 - p_1 - p_6) + P(d_M)(p_4 + \epsilon) \tag{139}$$

Hence,

$$P^1(Y_x \mid x', y) \neq P^2(Y_x \mid x', y) \quad \text{and} \quad P^1(Y_x \mid m, x', y) \neq P^2(Y_x \mid m, x', y) \tag{140}$$

Cases 1, 2, and 3 demonstrate that all the edges $X \to Y, Y \to M, M \to X$ are necessary.

□

B.6 Proof of Thm. 1

Proof. In this section, we show that M-ID is sound. From (Correa, Lee, and Bareinboim 2021) and (Tian and Pearl 2002), we already know that CTF-FACTOR, CTF-FACTORIZE, IDENTIFY are sound. Here, we need to show that M-REDUCE is sound, that is given a ctf-factor, it returns us an expression over the ctf-factors, which evaluate to the same value, that is

$$P(\mathbf{W}_* = \mathbf{w}_*) = \sum_{c, \mathbf{T}_* = \mathbf{t}_* \in \mathbf{Q}} c \cdot P(\mathbf{T}_* = \mathbf{t}_*) \tag{141}$$

Note that the list $\mathbf{Q}$ in M-REDUCE is a list of ctf-factors, whose sum (along with their signs) evaluates to the given counterfactual query. In the first part, expanding the summation increases the number of items in the list but does not change the evaluation. Simplification Rule of Lem. 2 is sound. Hence, its application also does not change the evaluation of each term. If the algorithm does not return fail, we move on to the application of the Difference Rule.

If the condition of the Difference Rule is satisfied, and it can be applied, then we apply the rule, change the signs accordingly, and add them to the list. Observe that the evaluation still does not change since the Difference Rule is sound.

$$\sum_{c, \mathbf{W}_* = \mathbf{w}_* \in \mathbf{Q}} c \cdot P(\mathbf{W}_* = \mathbf{w}_*) = \sum_{c, \mathbf{W}_* = \mathbf{w}_* \in \mathbf{Q} \setminus \{s, \mathbf{T}_* = \mathbf{t}_*\}} c \cdot P(\mathbf{W}_* = \mathbf{w}_*) + s \cdot P(\mathbf{T}'_* = \mathbf{t}'_*) + (-s) \cdot P(\mathbf{T}''_* = \mathbf{t}''_*) \tag{142}$$

Thus, the final evaluation returned by M-REDUCE evaluates to the same expression as its input, proving that the algorithm is sound. □

B.7 Proof of Thm. 2

We assume that all the impossible ctf-factors have been removed, and hence, every ctf-factor in this part of the proof are possible, that is there exists SCMs where the probability of the ctf-factor is greater than 0. Before, we proceed with the proof, let us introduce a function, by which a variable depends on its parents and exogenous variables. Let's call this c-function.

Definition 10 (c-function)**.** *Suppose $W_{\mathbf{t}_1} = w_1, W_{\mathbf{t}_2} = w_2, \ldots, W_{\mathbf{t}_m} = w_m$ are terms in a ctf-factor. We will call the following construction of the function for W as c-function. U and R are the exogenous variables for W.*

1. ***Case 1:*** *$w_i = 0$, $\forall i \in [m]$, then $W = U$.*
2. ***Case 2:*** *$w_i = 1$, $\forall i \in [m]$, then $W = \sim U$.*
3. ***Case 3:*** *$W = (1 - U) \cdot \mathbb{1}\{\exists j \quad pa(W) \geq t_j, w_j = 1\} + U \cdot R$*

Note that the monotonicity of $pa(W)$ on W holds true by such construction.

Lemma 8. *For any possible ctf-factor $P(\mathbf{W}_* = \mathbf{w}_*)$, we can have an SCM, such that all endogenous variables are c-functions and*

$$P(\mathbf{W}_* = \mathbf{w}_*) = P(\mathbf{U} = \mathbf{0}) \tag{143}$$

Proof. This follows from the definition of the c-function, which guarantees that $\mathbf{U} = 0$ is the only value that satisfies all the variables in the ctf-factor $\mathbf{W}_* = \mathbf{w}_*$. □

Now, we prove Lem 6, which we will call Multi-Parent Inconsistency.

Proof. Without loss of generality, assume $\mathbf{t}_1, \mathbf{t}_2$ are the values of parents that either of the above condition holds. Let all variables in $V(\mathbf{Y}_*)$ follow the c-functions as described earlier. Note that the graph is essentially Markovian.

Let $\mathbf{U}_W$ be the set of all exogenous variables corresponding to the monotonic functions of W, and each one occurs with the same probability in M_1. Let $W_* = w_*$ be an extension of $W_{\mathbf{t}_1} = w_1, W_{\mathbf{t}_2} = w_2, \ldots, W_{\mathbf{t}_m} = w_m$ such that $W_{\mathbf{t}_1} = w_1, W_{\mathbf{t}_2} = w_2, \ldots, W_{\mathbf{t}_m} = w_m, W_* = w_*$ corresponds to exactly one $u \in \mathbf{U}_W$, that is $*$ corresponds to all other assignment of parents of W. Let u_1, u_2, u_3, u_4 be 4 values of U_W which corresponds to the following assignments

$$\begin{aligned}
u_0 &: W_{\mathbf{t}_1} = w_1, W_{\mathbf{t}_2} = w_2, \ldots, W_{\mathbf{t}_m} = w_m, W_* = w_* \\
u_1 &: W_{\mathbf{t}_1} = w'_1, W_{\mathbf{t}_2} = w_2, \ldots, W_{\mathbf{t}_m} = w_m, W_* = w_* \\
u_2 &: W_{\mathbf{t}_1} = w_1, W_{\mathbf{t}_2} = w'_2, \ldots, W_{\mathbf{t}_m} = w_m, W_* = w_* \\
u_3 &: W_{\mathbf{t}_1} = w'_1, W_{\mathbf{t}_2} = w'_2, \ldots, W_{\mathbf{t}_m} = w_m, W_* = w_*
\end{aligned}$$

Note that such an assignment is possible without violating monotonicity since we have already applied simplification. Define model M_2 such that

$$\begin{aligned}
P^2(u_0) &= P^1(u_0) + \epsilon, \qquad & P^2(u_3) &= P^1(u_3) + \epsilon \\
P^2(u_1) &= P^1(u_1) - \epsilon, \qquad & P^2(u_2) &= P^2(u_2) - \epsilon
\end{aligned}$$

Now, consider the observational distribution $P(W = w_j \mid \mathbf{t}_j)$ when $\mathbf{t}_j \neq \mathbf{t}_1, \mathbf{t}_2$

$$P(W = w_j \mid \mathbf{t_j}) = P(u_0) + P(u_1) + P(u_2) + P(u_3) + P(u : W_{t_j}(u) = w_j) \tag{144}$$
$$= P^1(W_{t_j} = w_j) \tag{145}$$
$$= P^2(W_{t_j} = w_j) \tag{146}$$

For $\mathbf{t}_j = \mathbf{t}_1$, $P^1(W_{t_1} = w_1) = P^2(W_{t_1} = w_1) = P(u_0) + P(u_1) + P(u : W_{t_1}(u) = w_1)$
Note that the L_2 distribution is also the same for both models.
How, the counterfactual query, $P(\mathbf{Y}_*, W_{\mathbf{t}_1} = w_1, W_{\mathbf{t}_2} = w_2, \ldots, W_{\mathbf{t}_m} = w_m)$ is given by

$$P^1(Y_*, W_{t_1} = w_1, W_{t_2} = w_2, \ldots, W_{t_m} = w_m) \tag{147}$$
$$= P^1(U_Y^*)(P^1(u_0) + P^1(u : W_{t_1} = w_1, W_{t_2} = w_2, \ldots, W_{t_m} = w_m)) \tag{148}$$
$$< P^2(U_Y^*)(P^2(u_0) + P^2(u : W_{t_1} = w_1, W_{t_2} = w_2, \ldots, W_{t_m} = w_m)) \tag{149}$$

□

Hence, if there is a multi-parent inconsistency, the ctf-factor is non-ID. From here on, we will assume that there is no multi-parent inconsistency in the ctf-factor. Note that if there are no such inconsistencies, then there can be at most two terms corresponding to a variable in the CTF-factor, and if there are two terms, they would have different assignments.

Lemma 9 (Common-Parent Inconsistency)**.** *If after applying Simplification Rule, there exists $W_{i[\mathbf{t}_i]}, W_{j[\mathbf{t}_j]}$ (W_i, W_j in the same c-component of $G[V(\mathbf{W}_*)]$) and $T \in \mathbf{T}_i \cap \mathbf{T}_j$ such that $i \neq j, t \in \mathbf{t}_1, t' \in \mathbf{t}_2, t \neq t'$, then the ctf-factor is non-ID.*

Proof. Before we proceed with the cases, let's define another class of functions called v-functions.

Definition 11 (v-function)**.** *Let W be a node and U and R, its exogenous variables. Let T be a parent where $T \to W$ is monotonic. Given a ctf-factor, define the function as follows:*

1. ***Case 1:*** *If the ctf-factor has only one term with W, then*

$$W = 0 \cdot \{U = n\} + t \cdot \{U = c\} + 1 \cdot \{U = a\}$$

2. ***Case 2:*** *If the ctf-factor has two terms, then either*
   (a) *There exists $W_{\mathbf{t}_1}, W_{\mathbf{t}_2}$ in the ctf-factor, such that $t_0 \in \mathbf{t}_1, t_1 \in \mathbf{t}_2$ then*

$$W = 0 \cdot \{U = n\} + t \cdot \{U = c\} + 1 \cdot \{U = a\} \tag{150}$$

   (b) *There exists $W_{\mathbf{t}_1}, W_{\mathbf{t}_2}$ in the ctf-factor, such that $t_0 \in \mathbf{t}_1, \mathbf{t}_2$ and there exists a monotonic parent X, such that $x_0 \in \mathbf{t}_1, x_1 \in \mathbf{t}_2$. Then*

$$W = R(X + (1 - X) \cdot (0 \cdot \{U = n\} + t \cdot \{U = c\} + 1 \cdot \{U = a\})) \tag{151}$$

   where R is a binary exogenous variable which is 1 with probability 0.5.

Note that this definition will also satisfy the actual counterfactual term since we have only applied the simplification rule for the reduction. The different terms that can be in the ctf-factor corresponding to W and the values of the exogenous variable that satisfies the condition are shown in table 4.

Table 4: v-function and counterfactual terms

Notation	Term	Exogenous variables
W^{cn}	$W_{t_0} = 0$	$U = \{c, n\}$
W^{n}	$W_{t_1} = 0$	$U = \{n\}$
W^{a}	$W_{t_0} = 1$	$U = \{a\}$
W^{ac}	$W_{t_1} = 1$	$U = \{a, c\}$
W^{c}	$W_{t_0} = 0, W_{t_1} = 1$	$U = \{c\}$
W^{acr}	$W_{t_0 x_0} = 0, W_{t_0 x_1} = 1$	$U = \{n, c\}, R = 1$
W^{nr}	$W_{t_1 x_0} = 0, W_{t_1 x_1} = 1$	$U = \{n\}, R = 1$

Let the ctf-factor be $P(\mathbf{Y}_*, W_{1[\mathbf{t}_1]}, W_{m[\mathbf{t}_m]})$. Let $W_2, \ldots, W_{n-1}$ be nodes that lie on the bidirected path from W_1 to W_m. W_i's are in topological ordering. Let T be the parent of W_1, W_m such that $t \in \mathbf{t}_1, t' \in \mathbf{t}_n, t \neq t'$. We will divide the proof into two cases: **Case 1:** There is a bidirected edge between W_1, W_m, that is, $m = 2$. **Case 2:** $m > 2$.

Case 1: $m = 2$. Let all nodes in $V(\mathbf{Y}_*)$ follow c-functions and W_1, W_2 follow v-functions of U given the query, which is the common exogenous variable between W_1, W_2. We will first propose the assignment for the 3 SCMs M_1, M_2, M_3 as shown below (superscript denotes the SCM):

$$P^1(U = u) = 1/9 \quad \forall u \in \{a, c, n\} \times \{a, c, n\} \tag{152}$$
$$P^2(U = u) = 1/9 + \epsilon \quad \forall u \in \{ac, cn, na\} \tag{153}$$
$$P^2(U = u) = 1/9 - \epsilon \quad \forall u \in \{ca, nc, an\} \tag{154}$$
$$P^2(U = u) = 1/9 \quad \forall u \in \{aa, cc, nn\} \tag{155}$$
$$P^3(U = u) = 1/9 + \epsilon \quad \forall u \in \{nc, ca\} \tag{156}$$
$$P^3(U = u) = 1/9 - \epsilon \quad \forall u \in \{cc, na\} \tag{157}$$
$$P^3(U = u) = 1/9 \quad \text{for other } u \tag{158}$$

All other exogenous variables denoted by $\mathbf{U}^*$ are random coin tosses. The counterfactual distributions are then given by:

$$P(\mathbf{Y}_* = \mathbf{y}_*, W_1^c, W_m^c) = P(\mathbf{U}^* = 0)P(U = cc) \tag{159}$$
$$P(\mathbf{Y}_* = \mathbf{y}_*, W_1^c, W_m^{nr}) = P(\mathbf{U}^* = 0)P(U = cn)P(R_2 = 1) \tag{160}$$
$$P(\mathbf{Y}_* = \mathbf{y}_*, W_1^{acr}, W_m^{nr}) = P(\mathbf{U}^* = 0)(P(U = nn) + P(U = cn))P(R_1 = 1)P(R_2 = 1) \tag{161}$$

Similarly, it can be shown that for all combination of queries, either $P^1(\mathbf{Y}_*, W_{1[\mathbf{t}_1]}, W_{m[\mathbf{t}_m]}) \neq P^2(\mathbf{Y}_*, W_{1[\mathbf{t}_1]}, W_{m[\mathbf{t}_m]})$ or $P^1(\mathbf{Y}_*, W_{1[\mathbf{t}_1]}, W_{m[\mathbf{t}_m]}) \neq P^3(\mathbf{Y}_*, W_{1[\mathbf{t}_1]}, W_{m[\mathbf{t}_m]})$. Now, we show that the observational distribution is the same under the three cases.

Subcase 1: W_1, W_2 follows v-function from Eq. 150. Then, the observational distribution is given by,

$$P(V, W_1, W_2) = \sum_{\mathbf{U}^*, U_{12}} P(\mathbf{U}^*)P(U_{12}) \prod_v P(v \mid Pa(V))P(W_1 \mid T, U_{12})P(W_2 \mid t, U_{12}) \tag{162}$$
$$= \sum_{\mathbf{U}^*} Z \cdot \sum_{U_{12}} P(U_{12}) \mathbb{1}(W_1 \mid T, U_{12}) \cdot \mathbb{1}(W_2 \mid T, U_{12}) \tag{163}$$

It follows from the fact that given $\mathbf{U}^*$, all the variables in Z are independent of U_{12}. The values of U_{12} and the assignment of W_1, W_2 given T and U_{12} are given in table 5. Note that in M_1, M_2, M_3, the observational distribution is the same.

Table 5: T, W_1, W_2, U_{12}

T	W_1	W_2	U_{12}
0	0	0	nn, nc, cn, cc
0	0	1	na, ca
0	1	0	an, ac
0	1	1	aa
1	0	0	nn
1	0	1	na, nc
1	1	0	an, cn
1	1	1	aa, ac, ca, cc

Subcase 2: W_1 follows Eq. 150, and W_2 follows Eq. 151. The observational distribution is then given by

$$P(V, W_1, W_2) = \sum_{\mathbf{U}^*, U_{12}, R_2} P(\mathbf{U}^*)P(U_{12})P(R_2) \prod_v P(v \mid Pa(V))P(W_1 \mid T, U_{12})P(W_2 \mid X, T, U_{12}) \tag{164}$$
$$= \sum_{\mathbf{U}^*} Z \cdot \sum_{U_{12}, R_2} P(U_{12})P(R_2) \cdot \mathbb{1}(W_1 \mid T, U_{12}) \cdot \mathbb{1}(W_2 \mid X, T, U_{12}, R_2) \tag{165}$$

Observation: When $R_2 = 0, W_2 = 0$ and for any assignment of W_1, permissible values of U_12 are of the form $U_1 \times \{a, c, n\}$. Hence, when $R_2 = 0$, we have the same distribution in M_1, M_2, M_3. When $R_2 = 1, X = 1$, we have a similar situation. When $R_2 = 1, X = 0$, we have a situation same in Table 5. Note that X can be equal to W_1, but it does not change anything in the analysis. Hence, the observational distribution is the same in the three models.

Subcase 3: Both W_1, W_2 follows the Eq. 151. The observational distribution is given by

$$P(V, W_1, W_2) = \sum_{\mathbf{U}^*, U_{12}, R_1, R_2} P(\mathbf{U}^*)P(U_{12})P(R_1)P(R_2)\prod_v P(v \mid Pa(V))P(W_1 \mid T, U_{12})P(W_2 \mid X, T, U_{12}) \quad (166)$$

$$= \sum_{\mathbf{U}^*} Z \cdot \sum_{U_{12}, R_2} P(U_{12})P(R_1)P(R_2) \cdot \mathbb{1}(W_1 \mid T, X, U_1, R_1) \cdot \mathbb{1}(W_2 \mid Y, T, U_2, R_2) \quad (167)$$

Similar to the observation above, when $R_2 = 0$ or $R_1 = 0$, the distribution remains the same in the three models. Similarly, when $R_2 = 1, Y = 1$ or $R_1 = 1, X = 1$, we have the distribution same in the three models because the permissible values of U_{12} are of the form $U_1 \times \{a, c, n\}$ or $\{a, c, n\} \times U_2$. When $R_1 = 1, X = 0, R_2 = 1, Y = 0$, we have the non-trivial case, which is again similar to Table. 5. Thus, the observational distribution in this case is also the same in the three models M_1, M_2, M_3. Note that if $X = Y$, we can just take it as T, or if X, Y is equal to W_1 or W_2, the analysis does not change.

The interventional distribution is also the same in the three cases, as the permissible values of U_{12} are the same if the intervention is on any other node than W_1, W_2. If it is on W_1 or W_2, the permissible values are of the for $U_1 \times \{a, c, n\}$ or $\{a, c, n\} \times U_2$.

Case 2: $m > 2$. Consider the term $P(Y_*, W_{1*}, W_{2:(m-1)*}, W_{m*})$. Along with $U_1, \ldots, U_{m-1}$ introduce binary exogenous variables $S_1, \ldots, S_m$ such that

1. $S_1 = U_1, S_m = U_{m-1}$
2. $S_i = \mathbb{1}\{U_{i-1} \neq U_i\}$ for $i = 3, \ldots, (n-1)$

W_1, W_n are v-functions of S_1, S_n, and $W_2, \ldots, W_{n-1}$ are c-functions of $S_2, \ldots, S_{n-1}$.

Similar to the previous section, Assignments 1 and 2 are given by

$$P^1(S_2 = 0 \mid u_1, u_2) = 0.5 \quad \forall u_1, u_2 \in \{a, c, n\} \times \{a, c, n\} \quad (168)$$

$$P^2(S_2 = 0 \mid u_1, u_2) = 0.5 + \epsilon \quad \forall u \in \{ac, cn, na\} \quad (169)$$

$$P^2(S_2 = 0 \mid u_1, u_2) = 0.5 - \epsilon \quad \forall u \in \{ca, nc, an\} \quad (170)$$

$$P^2(S_2 = 0 \mid u_1, u_2) = 0.5 \quad \forall u \in \{aa, cc, nn\} \quad (171)$$

$$P^3(S_2 = 0 \mid u_1, u_2) = 0.5 \quad \forall u \in \{a, c, n\} \times \{a, c, n\} \quad (172)$$

$$P^3(S_2 = 0 \mid u_1, u_2) = 0.5 + \epsilon \quad \forall u \in \{nc, ca\} \quad (173)$$

$$P^3(S_2 = 0 \mid u_1, u_2) = 0.5 - \epsilon \quad \forall u \in \{cc, na\} \quad (174)$$

$$P^3(S_2 = 0 \mid u_1, u_2) = 0.5 \quad \text{for other } u \quad (175)$$

All other exogenous variables are random coin toss. Let's compute $P(s_1, \ldots, s_n)$ in the three models. Firstly, compute the number of possible assignments of $u_1, \ldots, u_{m-1}$ given $s_1, \ldots, s_n$. The number can be computed using the following tree, where,

- The root of the tree is u_{m-1}
- If $v_i = 0$, u_i has only child $u_{i-1} = u_i$
- If $v_i = 1$, u_i has two children $u_{i-1} \neq u_i$

Here $i \in [3 \ldots m-1]$.If $u_m = u$, then the number of leaves with value u is given by d_k and those not equal to u is given by e_k.

$$d_k = \frac{2}{3} \cdot (-1)^k + \frac{1}{3} \cdot 2^k \quad (176)$$

$$e_k = \frac{1}{2} \cdot (2^k - c_k) = \frac{1}{3} \cdot 2^k - \frac{1}{3} \cdot (-1)^k = d_k + (-1)^k \quad (177)$$

Let's denote $\epsilon(u_1 \times u_2) = \sum_{u_{12} \in u_1 \times u_2} P^i(S_2 = 0 \mid u_{12}) - P^1(S_2 = 0 \mid u_{12})$ where i is 2 or 3. Now,

$$P^i(u_1, 0, \mathbf{s}, u_m) = \sum_{\mathbf{u}} P^i(u_1, 0, , u_m \mid \mathbf{u})P(\mathbf{u}) \quad (178)$$

$$= P^1(u_1, 0, \mathbf{s}, u_m) + C \cdot d_k(\epsilon(u_1 \times \{a, c, n\}) + Z \cdot (-1)^k(u_1 \times \{a, c, n\} \setminus u_m) \quad (179)$$

The middle term is always 0, and the difference with the model M_1 for different assignments of u_1, u_{m-1} are shown in Tab. 6.

Table 6: $P(s_1, s_2 = 0, \mathbf{s}, s_n)$

$s_1, s_2 = 0, \mathbf{s}, s_n$	ϵ	M_1	M_2	M_3
aa	$\epsilon(ac, an)$	0	0	$-\epsilon$
ac	$\epsilon(aa, an)$	$-\epsilon$	0	ϵ
an	$\epsilon(aa, ac)$	ϵ	0	0
ca	$\epsilon(cc, cn)$	ϵ	$-\epsilon$	0
cc	$\epsilon(ca, cn)$	0	ϵ	0
cn	$\epsilon(cc, ca)$	$-\epsilon$	0	0
na	$\epsilon(nc, nn)$	$-\epsilon$	ϵ	0
nc	$\epsilon(na, nn)$	ϵ	$-\epsilon$	ϵ
nn	$\epsilon(na, nc)$	0	0	$-\epsilon$

Table 7: $P(s_1, s_2 = 1, \mathbf{s}, s_n)$

$s_1, s_2 = 1, \mathbf{s}, s_n$	ϵ	M_1	M_2	M_3
aa	$\epsilon(ac, an)$	0	0	ϵ
ac	$\epsilon(aa, an)$	ϵ	0	$-\epsilon$
an	$\epsilon(aa, ac)$	$-\epsilon$	0	0
ca	$\epsilon(cc, cn)$	$-\epsilon$	ϵ	0
cc	$\epsilon(ca, cn)$	0	$-\epsilon$	0
cn	$\epsilon(ca, cc)$	ϵ	0	0
na	$\epsilon(nc, nn)$	ϵ	$-\epsilon$	0
nc	$\epsilon(na, nn)$	$-\epsilon$	ϵ	$-\epsilon$
nn	$\epsilon(na, nc)$	0	0	ϵ

Similarly, for $v_{n-1} = 1$, it is exactly the same with the signs changed and shown in table 7.

Now the proof follows similar to Case 1, with u_{12} replaced by $s_1 \mathbf{s} s_m$, where $s_1 = u_1, s_m = u_2$. For example, the following counterfactual quantity can be represented by replacing all s with 0 except s_1, s_m in this case as

$$P(\mathbf{Y}_* = \mathbf{y}_*, W_1^{acr}, W_m^{nr}) \tag{180}$$
$$= P(\mathbf{U}^* = 0)(P(s_1 = n, s_{2:m-1} = 0, s_m = n) + P(s_1 = c, s_{2:m-1} = 0, s_m = n))P(R_1 = 1)P(R_2 = 1) \tag{181}$$

Observational distributions can be shown to be equal in a similar fashion:

$$P(V, W_1, W_m) = \sum_{\mathbf{U}^*, s_{2:m}} Z \cdot \sum_{u_1, u_{m-1}, R_1, R_m} P(u_1, u_{m-1})P(R_1)P(R_2) \cdot \mathbb{1}(W_1 \mid T, X, U_1, R_1) \cdot \mathbb{1}(W_m \mid Y, T, u_{m-1}, R_2) \tag{182}$$

Similar to Case 1, the observational and interventional distribution is the same in all three models, M_1, M_2, M_3.

□

Lemma 10 (Parent-Child Inconsistency)**.** *Consider a ctf-factor with a single c-component. If there is no Multi-Parent or Common-Parent Inconsistency in the ctf-factor, and there exists $W_t \in W_*, Z \in \mathbf{T} \cap V(\mathbf{W}_*)$, such that $z \in \mathbf{w}_*, z' \in \mathbf{t}$ and $z \neq z'$, then the ctf-factor is non-ID.*

Proof. Consider the ctf-factor $P(\mathbf{Y}_*, W_{1*}, W_{2:(m-1)*}, W_{m*})$, where $W_2, \ldots, W_m$ lie on the bidirected path from W_1 to W_m. **Observation:** If there are two terms with W_1 in the query then both $w_1, w_1' \in W_{1*}$. Apply the Difference Rule. If it is not applicable, it means w_1, w_1' occur in the counterfactual worlds of other variables.

1. If w_1, w_1' occurs in two different variables, it is a Common-Parent Inconsistency and is taken care of earlier.
2. If w_1, w_1' occurs in W_n, apply the Difference Rule to W_n first so that there are two terms, one with w_1 and another with w_1' and then apply the Difference Rule to W_1, as it is applied in reverse topological order, then the inconsistency is removed.

Hence, if we are returning FAIL because of Parent-Child inconsistency after a Difference Rule, then, $W_{1\mathbf{t}_1} = w_1$ and $W_{n\mathbf{t}_n}$ has $w_1' \in \mathbf{t}_n$.

We will divide these into two cases, as above, when $m = 2$ and $m > 2$.

Case 1: $m = 2$. Let W_1, W_2 are v-functions, the former with no parent and the latter with parent W_1. All other nodes are c-functions, given the query. Define the models M_1, M_2 as follows:

$$P^1(U_{12} = u) = 1/6 \quad \forall u \in \{a, n\} \times \{a, c, n\} \tag{183}$$
$$P^2(U_{12} = u) = 1/6 + \epsilon \quad \forall u \in \{nn, aa\} \tag{184}$$
$$P^2(U_{12} = u) = 1/6 - \epsilon \quad \forall u \in \{nc, ac\} \tag{185}$$
$$P^2(U_{12} = u) = 1/6 \quad \text{for other } u \tag{186}$$

Subcase 1: W_2 follows Eq. 150. Then, the counterfactual distributions are given by

$$P(\mathbf{Y}_*, W_1 = 0, W_{2[w_1=1]} = 1) = P(\mathbf{U}^*)(P(U_{12} = na) + P(U_{12} = nc)) \tag{187}$$
$$P(\mathbf{Y}_*, W_1 = 0, W_{2[w_1=0]} = 0, W_{2[w_1=1]} = 1) = P(\mathbf{U}^*)P(U_{12} = nc) \tag{188}$$

Similarly, it can be shown for combinations that the counterfactual terms vary in models M_1, M_2. Now, we show that the observational distribution is the same.

$$P(V, W_1, W_2) = \sum_{\mathbf{U}^*, U_{12}} P(\mathbf{U}^*)P(U_{12}) \prod_v P(v \mid Pa(V))P(W_1 \mid T, U_{12})P(W_2 \mid X, T, U_{12}) \tag{189}$$
$$= \sum_{\mathbf{U}^*} Z \cdot \sum_{U_{12}} P(U_{12}) \cdot \mathbb{1}(W_1 \mid U_{12}) \cdot \mathbb{1}(W_2 \mid W_1, U_{12}) \tag{190}$$

Now, from Table 8, we can see that the observational distribution is the same in both models.

Table 8: Assignment of W_2

W_1	W_2	U
0	0	nn, nc
0	1	na
1	0	an
1	1	aa, ac

Subcase 2: W_2 follows Eq. 151. Then again, the counterfactual distributions are given by

$$P(\mathbf{Y}_*, W_1 = 0, W_{2[w_1=1,x_0]} = 0, W_{2[w_1=1,x_1]} = 1) = P(\mathbf{U}^*)P(R_1 = 1)(P(U_{12} = nn) \tag{191}$$
$$P(\mathbf{Y}_*, W_1 = 1, W_{2[w_1=0,x_0]} = 0, W_{2[w_1=0,x_1]} = 1) = P(\mathbf{U}^*)P(R_1 = 1)((P(U_{12} = an) + P(U_{12} = ac)) \tag{192}$$

The observational distribution can be shown to be the same by the following:

$$P(V, W_1, W_2) = \sum_{\mathbf{U}^*, U_{12}, R_2} P(\mathbf{U}^*)P(U_{12})P(R_2) \prod_v P(v \mid Pa(V))P(W_1 \mid T, U_{12})P(W_2 \mid X, T, U_{12}, R_2) \tag{193}$$
$$= \sum_{\mathbf{U}^*} Z \cdot \sum_{U_{12}, R_2} P(U_{12})P(R_2) \cdot \mathbb{1}(W_1 \mid U_{12}) \cdot \mathbb{1}(W_2 \mid X, W_1, U_{12}, R_2) \tag{194}$$

We make a similar observation as the last proof. When $R_2 = 0$ or $R_2 = 1, X = 1$, we have the same distribution since the valid values of U_{12} are from $U_1 \times \{a, , c, n\}$. For $R_2 = 1, X = 0$, we have the same assignment as Table 8 for W_1, W_2. Hence, the distribution is the same in models M_1, M_2. The interventional distribution is also the same in the two models, as the permissible values of U_{12} are the same if the intervention is on any other node than W_1, W_2. If it is on W_1 or W_2, the permissible values are of the for $U_1 \times \{a, c, n\}$ or $\{a, c, n\} \times U_2$.

Case 2: We follow the exact same construction as the previous proof and the changes are shown as M_3 in Table 6 and 7. From Case 1, it follows that the observational and interventional distributions are the same in M_3, M_1, and the counterfactual distributions are different. An example calculation is shown below:

$$P(V, W_1, W_m) = \sum_{\mathbf{U}^*, s_{2:m-1}} Z \cdot \sum_{U_1, U_m, R_2} P(U_1, U_m)P(R_2) \cdot \mathbb{1}(W_1 \mid U_1) \cdot \mathbb{1}(W_m \mid X, W_1, U_m, R_2) \tag{195}$$

□

C Discussion and Further Examples

C.1 Monotonicity Reduction Lemma

Testability from observations and interventions. Monotonicity is not testable in general from observational and interventional data. Consider the causal graph shown in Fig. 4. If in the observational or interventional data, we see $P(y_1 \mid x_1) < P(y_1 \mid x_0)$ or $P(Y_{x_1} = 1) < P(Y_{x_0} = 1)$, we can be sure that X is **not positive monotonic** on Y. However, if we have $P(y_1 \mid x_1) \geq P(y_1 \mid x_0)$ or $P(Y_{x_1} = 1) \geq P(Y_{x_0} = 1)$, it is not immediately true that monotonicity holds. To witness, consider two SCMs as follows: M_1, M_2. In both M_1 and M_2, we have

$$X = U_X, \qquad P(U_X = 1) = P(U_X = 0) = 0.5 \tag{196}$$
$$Y = \mathbb{1}\{U_Y = a\} + X \cdot \mathbb{1}\{U_Y = c\} + (1 - X) \cdot \mathbb{1}\{U_Y = d\} \tag{197}$$

In M_1, we have $P^1(U_Y = c) = 0.75, P^1(U_Y = d) = 0.25$ and in M_2, we have $P^2(U_Y = a) = P^2(U_Y = n) = 0.25$ and $P(U_Y = c) = 0.5$. Note that both observational and interventional distributions induced by M_1, M_2 are the same, however, the edge $X \to Y$ is non-monotonic in M_1, but monotonic in M_2.

MRL in Non-binary case: We now explore how Monotonicity Reduction Lemma changes in the non-binary case. First, we show that for the CGMA in Fig. 4, when Y is not binary, nor can be reduced to binary form in the query (that is some set of values can be taken to be 0 and another set of values to be 1 for MRL to be applied), then the query is non-ID. Then we propose a variation of MRL to circumvent this issue. Note that monotonicity for the non-binary case can be defined in many ways, and in this case we follow the one in Def. 1.

Consider SCMs, M_1, M_2, where $P(X = 1) = P(X = 0) = 0.5$ and Y is ternary (values can be $0, 1$ or 2). The join counterfactual distributions in these two models are given by the Table. 9

Y_{x_0}	Y_{x_1}	$P^1(U_Y)$	$P^2(U_Y)$
0	0	p_1	p_1
0	1	p_2	$p_2 + \epsilon$
1	1	p_3	$p_3 - \epsilon$
0	2	p_4	$p_4 - \epsilon$
1	2	p_5	$p_5 + \epsilon$
2	2	p_6	p_6

Table 9: Each row represents the probability of canonical type $Y_{x_0} = y, Y_{x_1} = y'$ in SCM M_1 and M_2

The observational and interventional distribution for the two models are the same, and the following facts follow:

1. $P(Y_{x_0} = y, Y_{x_1} = y')$ is non-ID unless $y = y' = 0$ or $y = y' = 2$, and
2. $P(Y_{x_0} \leq y < Y_{x_1})$ is ID for all y.

With this observation, we propose MRL for non-binary variables as follows:

Lemma 11 (Monotonicity Reduction Lemma (MRL))**.** *Let $\mathbf{T}, \mathbf{S}$ be a partition of the parents of W, such that $\mathbf{T}$ and $\mathbf{S}$ are the set of monotonic and non-monotonic parents of W, respectively. Let $P(Y_*, W_{\mathbf{t},\mathbf{s}} \star w, W_{\mathbf{t}',\mathbf{s}} \star w')$ be a ctf-factor, where $\star$ can be $\leq$ or $>$. Then the following rules can be applied to reduce the query to a simpler ctf-factor.*

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

$$\begin{aligned} &P(Y_*, W_{\mathbf{t},\mathbf{s}} \leq w, W_{\mathbf{t}',\mathbf{s}} > w) \\ &\quad = P(Y_*, W_{\mathbf{t}',\mathbf{s}} > w) - P(Y_*, W_{\mathbf{t},\mathbf{s}} > w) \qquad (198) \\ &\quad = P(Y_*, W_{\mathbf{t},\mathbf{s}} \leq w) - P(Y_*, W_{\mathbf{t}',\mathbf{s}} \leq w) \qquad (199) \end{aligned}$$

Proof. By monotonicity, we have for any unit $\mathbf{u}$, $W_{\mathbf{t},\mathbf{s}} \leq W_{\mathbf{t}',\mathbf{s}}$. Hence, $W_{\mathbf{t}',\mathbf{s}}(\mathbf{u}) \leq w$ implies $W_{\mathbf{t},\mathbf{s}} \leq w'$ for all $\mathbf{u}$. The simplification Rule follows from this. The Difference Rule can be obtained by binarizing the domain and applying MRL. □

C.2 Graphical Conditions

In this section, we discuss how the algebraic conditions for LATE, LNDE and LNIE fail in causal diagrams in Figs. 1c, 2b and then provide a set of assumptions for identification.

Figure 4: 2 variables with monotonicity.

Violation of LATE Assumption: Consider Assumption 1 on the existence of instruments. It is not satisfied in Fig. 1c for the following three reasons:

1. Directed path from Z to Y
2. Unobserved confounder between Z and Y
3. Confounder W between Z and Y
4. Confounder W between Z and X

To illustrate why, consider the following functional forms of Z and Y.

$$Z = f_Z(W, U_Z, U_{WZ}, U_{ZY}) \tag{200}$$
$$Y = f_Y(W, Z, X, M, U_Y, U_{WY}, U_{ZY}) \tag{201}$$
$$Y_x = f_Y(W, Z, x, M, U_Y, U_{WY}, U_{ZY}) \tag{202}$$

Now, since Y_x is still a function of Z, U_{ZY}, W, $Y_x \not\perp\!\!\!\perp Z$.

Similarly, the assumption is violated in Fig. 2b because of:

1. Directed path from Z to Y
2. Confounder W between Z and Y
3. Confounder W between Z and X

To illustrate, consider the following functional forms:

$$Z = f_Z(W, U_Z, U_{ZM}) \tag{203}$$
$$Y = f_Y(W, X, M, U_Y, U_{WY}, U_{XY}) \tag{204}$$
$$M = f_M(W, Z, X, U_M, U_{ZM}) \tag{205}$$
$$Y_x = f_Y(W, x, M, U_Y, U_{WY}, U_{XY}) \tag{206}$$
$$= f_Y(W, x, f_M(W, Z, X, U_M, U_{ZM}), U_Y, U_{WY}, U_{XY}) \tag{207}$$

Since, Y_x is still a function of Z, W, $Y_x \not\perp\!\!\!\perp Z$.

Violation of LNDE/LNIE Assumptions: Consider Fig. 2b and the functional forms shown above:

$$Z = f_Z(W, U_Z, U_{ZM}) \tag{208}$$
$$Y = f_Y(W, X, M, U_Y, U_{WY}, U_{XY}) \tag{209}$$
$$M = f_M(W, Z, X, U_M, U_{ZM}) \tag{210}$$
$$M_x = f_M(W, Z, x, U_M, U_{ZM}) \tag{211}$$
$$\tag{212}$$

Since, M_x is still a function of Z, the assumption of Conditionally Ignorable Treatment Assignment A.2 is violated since $M_x \not\perp\!\!\!\perp Z \mid W$.

Algebraic Assumptions and Graphical Conditions for Identification of LATE: We now provide two sets of assumptions for identifying LATE. We will assume that W, Z, X, M, Y are in strict topological order. Consider Fig. 1c and the following conditions:

Assumption 3. *1. No unobserved confounder between W and X, that is $X_w \perp\!\!\!\perp W$.*
2. No unobserved confounder between W and M, that is $M_w \perp\!\!\!\perp W$.
3. No unobserved confounder between Z and X, that is $X_{zw} \perp\!\!\!\perp Z_w$.
4. No direct effect from Z to M, that is $M_{xzw} = M_{xw}$.
5. No unobserved confounder between Z and M, that is $M_{xw} \perp\!\!\!\perp Z_w$.
6. No unobserved confounder between X and Y, that is $Y_{xmzw} \perp\!\!\!\perp X_{zw}$.
7. No unobserved confounder between M and Y, that is $Y_{xmzw} \perp\!\!\!\perp M_{xzw}$.

It is easy to see that this completely defines the graph in Fig. 1c, as all other edges are present in the graph. We showed in Sec. B that LATE can be identified in this graph. We now distill the set of assumptions satisfied by this causal graph, and implied by Assumptions 3, which uniquely identifies LATE.

Assumption 4. *If for all values z, x, m of Z, X, M and for all values $\mathbf{u}$ of exogenous variables*

1. $Y_{xm}, Z \perp\!\!\!\perp X_z, M_x \mid W$
2. $X_{z_1}(\mathbf{u}) \geq X_{z_0}(\mathbf{u})$ *(Monotonicity)*

Proposition 12. *LATE is identifiable for any structural causal model that satisfies Assumption 4.*

Proof. First, we compute $P(Y_x, X_{z_0} = 0, X_{z_1} = 1)$. Once such terms are available, the LATE can be computed.

$$\text{LATE} := \frac{\sum_y y \cdot P(Y_{x_1} = y, X_{z_0} = 0, X_{z_1} = 1)}{\sum_y P(Y_{x_1} = y, X_{z_0} = 0, X_{z_1} = 1)} - \frac{\sum_y y \cdot P(Y_{x_0} = y, X_{z_0} = 0, X_{z_1} = 1)}{\sum_y P(Y_{x_0} = y, X_{z_0} = 0, X_{z_1} = 1)} \tag{213}$$

By monotonicity, we have

$$P(Y_{x_1} = y, X_{z_0} = 0, X_{z_1} = 1) = P(Y_{x_1} = y, X_{z_1} = 1) - P(Y_{x_1} = y, X_{z_0} = 1) \tag{214}$$

Now, we compute $P(Y_{x_1} = y, X_{z_1} = 1)$. Other terms can be computed similarly.

$$\begin{aligned}
P(Y_{x_1} = 1, X_{z_1} = 1) &= P(Y_{x_1 M_{x_1} Z_{x_1}} = 1, X_{z_1} = 1) && (215)\\
&= \sum_{m,z} P(Y_{x_1 mz} = 1, X_{z_1} = 1, M_{x_1} = 1, Z) \quad \text{(Unnesting)} && (216)\\
&= \sum_{m,z,w} P(Y_{x_1 mz} = 1, X_{z_1} = 1, M_{x_1} = 1, z \mid w) P(w) \quad \text{(Law of total probability)} && (217)\\
&= \sum_{m,z,w} P(Y_{xmz}, z \mid w) P(X_{z_1} = 1, M_{x_1} = 1 \mid w) P(w) \quad \text{(Assumption 4.1)} && (218)\\
&= \sum_{w,z,m} P(y \mid do(x, m), w, z) P(z \mid w) P(w) P(X_{z_1} = 1, M_{x_1} = 1 \mid w) && (219)
\end{aligned}$$

Consider all possible SCMs for which the Assumption 4 holds.

Observation: The assumption $M_x \perp\!\!\!\perp Z \mid W$ implies that all the directed paths from Z to M go through X and all the confounding paths are blocked by W. Hence,

$$P(X_{z_1} = 1, M_{x_1} = 1 \mid w) = P(X_{z_1} = 1, M_{x_1 z_1} = 1 \mid w) = P(x_1, m_1 \mid do(z_1), w) \tag{220}$$

Also, $X_z, M_x \perp\!\!\!\perp Z \mid W$ implies that all confounding paths between by Z and $\{X, M\}$ are blocked by W. Since, Z is an ancestor of $\{X, M\}$, all directed paths are removed in $G_{\underline{Z}}$. Hence, $X, M \perp\!\!\!\perp Z \mid W$ in $G_{\underline{Z}}$, which implies by do-calculus that $P(x_1, m_1 \mid do(z_1), w) = P(x_1, m_1 \mid z_1, w)$.

On the other hand, we have $Y_{mx} \perp\!\!\!\perp X_z, M_x \mid W$, which implies there cannot exist a confounding path between Y and $\{X, M\}$ without W or Z. If the confounding path does not contain Z, then it must be blocked by W, or else $Y_{mx} \perp\!\!\!\perp X_z, M_w \mid W$ will not hold. If the confounding path contains Z, it can be of the form:

- $M \leftarrow \ldots X \leftarrow \ldots \leftarrow Z \leftarrow \ldots \leftarrow C \rightarrow \ldots \rightarrow Y$
- $X/M \leftarrow \ldots \leftarrow C \rightarrow \ldots \rightarrow Z \rightarrow \ldots \rightarrow Y$
- $M \leftarrow \ldots X \leftarrow \ldots \leftarrow Z \rightarrow \ldots \rightarrow Y$
- $X/M \leftarrow \ldots \leftarrow C \rightarrow \ldots \rightarrow Z \leftarrow \ldots \leftarrow C' \rightarrow \ldots \rightarrow Y$

All except the last kind of path are blocked by Z. For the last case, there is a confounding path between Z and X (or M), so it must be blocked by W. Hence, we have $Y \perp\!\!\!\perp X, M \mid Z, W$ in $G_{\underline{XM}}$, which implies $P(y_1 \mid do(x, m), w, z) = P(y_1 \mid w, z, x, m) P(z \mid w)$. From this, we derive the final expression for LATE. □

Next, we consider Fig. 2b and graphical conditions implied by it:

Assumption 5. *1. No unobserved confounder between W and Z, that is $Z_w \perp\!\!\!\perp W$.*
2. No unobserved confounder between W and M, that is $M_w \perp\!\!\!\perp W$.
3. No unobserved confounder between Z and X, that is $X_{zw} \perp\!\!\!\perp Z_w$.
4. No direct effect from Z to Y, that is $Y_{xzmw} = Y_{xmw}$.
5. No unobserved confounder between Z and Y, that is $Y_{xmw} \perp\!\!\!\perp Z_w$.

6. *No unobserved confounder between X and M, that is $M_{xzw} \perp\!\!\!\perp X_{zw}$.*
7. *No unobserved confounder between M and Y, that is $Y_{xmzw} \perp\!\!\!\perp M_{xzw}$.*

Again, it should be noted that these set of assumptions define the graph in Fig. 2b. We now distill another set of assumptions satisfied by this causal graph, and implied by Assumptions 5, which uniquely identifies LATE.

Assumption 6. *If for all values z, x, m of Z, X, M and for all values $\mathbf{u}$ of exogenous variables*

1. $M_x, Z \perp\!\!\!\perp Y_{xm}, X_z \mid W$
2. $X_{z_1}(\mathbf{u}) \geq X_{z_0}(\mathbf{u})$ *(Monotonicity)*

Proposition 13. *LATE is identifiable for any structural causal model that satisfies Assumption 6.*

Proof. First, we compute $P(Y_x, X_{z_0} = 0, X_{z_1} = 1)$. Once, we have computed that LATE can be computed easily. By monotonicity, we have

$$P(Y_{x_1} = y, X_{z_0} = 0, X_{z_1} = 1) = P(Y_{x_1} = y, X_{z_1} = 1) - P(Y_{x_1} = y, X_{z_0} = 1) \tag{221}$$

Now, we compute $P(Y_{x_1} = y, X_{z_1} = 1)$. The rest of the terms can be computed similarly.

$$\begin{aligned}
P(Y_{x_1} = 1, X_{z_1} = 1) &= P(Y_{x_1 M_{x_1}} = 1, X_{z_1} = 1) && (222)\\
&= \sum_m P(Y_{x_1 m} = 1, X_{z_1} = 1, M_{x_1} = 1) \quad \text{(Unnesting)} && (223)\\
&= \sum_{m,w} P(Y_{x_1 m} = 1, X_{z_1} = 1, M_{x_1} = 1 \mid w) P(w) \quad \text{(Law of total probability)} && (224)\\
&= \sum_{m,w} P(M_x = m \mid w) P(Y_{xm} = 1, X_{z_1} = 1 \mid w) P(w) && (225)
\end{aligned}$$

Observation: Since $Z \perp\!\!\!\perp X_z \mid W$, all confounding paths between X and Z are blocked by W. Now, consider the assumption $M_x \perp\!\!\!\perp X_z \mid W$, which implies there are no confounding paths between X and M that do not contain either W, Z. If it does not contain Z it must be blocked by W. If it contains Z, it must be of the following forms:

- $X \leftarrow \ldots \leftarrow Z \leftarrow \ldots \leftarrow C \rightarrow \ldots \rightarrow M$
- $X \leftarrow \ldots \leftarrow C \rightarrow \ldots \rightarrow Z \rightarrow \ldots \rightarrow M$
- $X \leftarrow \ldots \leftarrow Z \rightarrow \ldots \rightarrow M$
- $X \leftarrow \ldots \leftarrow C \rightarrow \ldots \rightarrow Z \leftarrow \ldots \leftarrow C \rightarrow \ldots \rightarrow M$

The first, second, and third types of paths are blocked by Z. For the fourth type, we note that all confounding paths between Z and X are blocked by W. Hence all confounding paths between X, M are blocked by Z, W. Hence, $X \perp\!\!\!\perp M \mid Z, W$ in $G_{\underline{X}}$, which implies

$$P(m \mid do(x), w) = \sum_z P(m \mid x, z, w) P(z \mid w)$$

Since $Y_{xm} \perp\!\!\!\perp Z \mid W$, all directed paths from Z the Y passes through either X or M, and all confounding paths are blocked by W. Using the fact that M is descendant of X, it follows:

$$P(Y_{xm} = 1, X_{z_1} = 1 \mid w) = P(y_1, x_1 \mid do(z, m), w) = P(y_1 \mid do(z, m), x, w) P(x \mid do(z), w)$$

Consider the graph $G_{\underline{ZM}}$. All the directed paths from Z to Y are removed, and all confounding paths between Z and Y are blocked by W. All the confounding paths between M and Y that do not contain Z are also blocked by W. If this confounding path contains Z and is not blocked by Z, Z should be a collider, which means the path is of the form $M \leftarrow \ldots \rightarrow Z \leftarrow \ldots \rightarrow Y$. But we know that confounding paths between Z and Y are blocked by W. Hence, $Y \perp\!\!\!\perp Z, M \mid X, W$

$$P(y_1 \mid do(z, m), x, w) = P(y_1 \mid z, m, x, w), \quad P(x \mid do(z), w) = P(x \mid z, w) \tag{226}$$

Putting the values, we have the expression for LATE.

□

C.3 Algorithmic Identification

In this section, we provide the sub-routines from Algorithm 1, M-ID. CTF-FACTOR is shown in Algorithm 3, CTF-FACTORIZE is shown in Algorithm 4 and IDENTIFY is shown in Algorithm 6. For more details about the algorithm, please refer (Correa, Lee, and Bareinboim 2021) and (Tian and Pearl 2002).

Algorithm 3: CTF-FACTOR

Input: Causal Graph G. $\mathbf{X}_* = \mathbf{x}_*, \mathbf{Y}_* = \mathbf{y}_*$ are two sets of counterfactual variables and their values.
Output: ctf-factor that needs to be computed and $P(\mathbf{W}_* = \mathbf{w}_*)$ and the variables in the summation $\mathbf{d}_*$

Let $\mathbf{A}_1, \mathbf{A}_2, \ldots$ be the ancestral components of $\mathbf{X}_* \cup \mathbf{Y}_*$ given $\mathbf{X}_*$
Let $\mathbf{D}_*$ be the union of the ancestral components containing a variable in $\mathbf{Y}_*$ and $\mathbf{d}_*$ be the corresponding set of values.
$\mathbf{D'}_* = \|\cup_{D_t \in \mathbf{D}_*} D_{pa_d}\|$
If any term in $\mathbf{D'}_*$ is not possible **return** 0
Remove repeated terms in $\mathbf{D'}_*$
$\mathbf{W}_* \leftarrow An(\mathbf{D'}_*)$ and the corresponding values $\mathbf{w}_*$
return $\mathbf{d}_*, P(\mathbf{W}_* = \mathbf{w}_*)$

Algorithm 4: CTF-FACTORIZE

Input: Causal Graph G. Ctf-factor $\mathbf{W}_* = \mathbf{w}_*$
Output: Factors in the factorization of $P(\mathbf{W}_* = \mathbf{w}_*)$

$C_1, C_2, \ldots$ are the c-components of $G[V(\mathbf{W}_*)]$
$\mathbf{C}_{j*} = \{W_{\mathbf{pa}_W} \in \mathbf{W}_* \mid W \in C_j\}$ and $\mathbf{c}_{j*}$ are the values in $\mathbf{w}_*$ corresponding to $\mathbf{C}_{j*}$
return $\mathbf{C}_{1*}, \mathbf{C}_{2*}, \ldots$

D Further Results

D.1 Multiple Instruments

LATE with Multiple Instruments In Sec. 1, we looked at estimating in IV graphs. Here, we consider cases with multiple instruments affecting a binary treatment. IV Graphs with multiple instruments (see Fig. 5) have been studied extensively both in theory and practice (Mogstad, Torgovitsky, and Walters 2019), (Angrist and Imbens 1995). (Angrist and Imbens 1995) studied estimations with 2SLS under monotonicity assumption and (Mogstad, Torgovitsky, and Walters 2019) have focused on causal interpretation under similar assumptions. In this section, we provide a characterization of the queries that are identifiable in the non-parametric setting under the condition that some of the incoming edges are monotonic and some of the edges are not. One approach is to treat multiple parent variables as a single parent whose values are the Cartesian product of their domains. For example, if X has two binary parents, Z_0, Z_1, we can think of it as a single parent $\mathbf{Z}$ with values in $\{0,1\}^2$. This allows us to identify a quantity like $P(Y_x, X_{\mathbf{z}=00} = 0, X_{\mathbf{z}=11})$ using MRL as the domain of $\mathbf{z}$ can be treated as binary for this query. However, complications arise when the values of Z in the counterfactual world are not *comparable*. making estimation of queries like $P(Y_x, X_{\mathbf{z}=01} = 0, X_{\mathbf{z}=10})$ challenging. In the following proposition, we claim that such queries are not identifiable.

Proposition 14 (Multi-Parent LATE Identification)**.** *In the Causal Graph 5, let* $\mathbf{Z} = \{Z_1, \ldots, Z_n\}$ *be the set of monotonic parents of* X *and* $\mathbf{W} = \{W_1, \ldots, W_m\}$ *be the set of non-monotonic parents of* X*. Let* $\mathbf{z_0}, \mathbf{z_1}$ *be two sets of values of* $\mathbf{Z}$ *and* $\mathbf{w_0}, \mathbf{w_1}$ *be two sets of values of* $\mathbf{W}$*. The effect* $P(Y_x \mid X_{\mathbf{z_0},\mathbf{w_0}} = 0, X_{\mathbf{z_1},\mathbf{w_1}} = 1)$ *is identifiable if and only if* $\mathbf{z_0} \leq \mathbf{z_1}$ *and* $\mathbf{w_0} = \mathbf{w_1} = \mathbf{w}$ *and is given by*

$$P(Y_{x_i} \mid X_{\mathbf{z_0},\mathbf{w_0}} = 0, X_{\mathbf{z_1},\mathbf{w_1}} = 1) \tag{227}$$

$$= \frac{P(y, x_i \mid \mathbf{z_i}, \mathbf{w}) - P(y, x_i \mid \mathbf{z_{1-i}}, \mathbf{w})}{P(x_i \mid \mathbf{z_i}, \mathbf{w}) - P(x_i \mid \mathbf{z_{1-i}}, \mathbf{w})} \tag{228}$$

Algorithm 5: ID

Input: Causal Graph G. $C \in T \in \mathbf{V}, Q = Q[T] = P_{\mathbf{V}\setminus T}(T)$. Assume $G_{[C]}, G_{[T]}$ is composed of single c-component.
Output: Expression for $Q[C] = P_{\mathbf{V}\setminus C}P(C)$ in terms of Q or FAIL

$A \leftarrow An(C)_{G_{[T]}}$
if $A = C$, **return** $Q[C] = \sum_{t\setminus c} Q$
if $A = T$, **return** FAIL
if $A = C$ **then**
    T' be the c-component containing C in $G_{[A]}$
    Compute $Q[T']$ from $Q[A] = \sum_{t\setminus a} Q$
    return $\mathrm{ID}(C, T', Q[T'], G)$
end if

Algorithm 6: IDENTIFY

Input: Causal Graph G. A ctf-factor with a single c-component $\mathbf{C}_* = \mathbf{c}_*$. Available distribution $\mathbb{Z}$.
Output: $P_{\mathbf{V}\setminus C)}(C)$

for $\mathbf{Z} \in \mathbb{Z}, \mathbf{Z} \cap C = \emptyset$ **do**
  Let B be the c-component of $G_{\bar{\mathbf{Z}}}$ such that $C \subseteq B$, compute $P_{V\setminus B}(B)$ from $P_{\mathbf{Z}}(V)$
  if ID$(C, B, P_{V\setminus B}(B), G)$ does not FAIL **then**
    return ID$(C, B, P_{V\setminus B}(B), G)$
  end if
end for

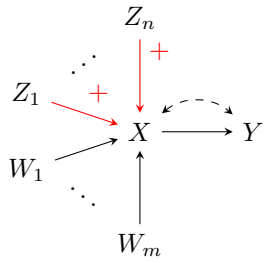


Figure 5: LATE with multiple instruments

This extends the result in (Imbens and Angrist 1994). The proof follows from Multi-Parent Inconsistency in Proof of Thm. 2(Sec. B.7).

E Experiment Details

In this section, we provide some more details about the experiments. The experiments are run on an Apple M2 chip with 16GB of memory.

401k Dataset: We select only a subset of columns, namely income (inc), eligibility in 401k (e401), participation in 401k (p401), net financial asset (net_tfa), and total wealth (tw). Then, we discretize the variables income, net financial assets, and total wealth, based on the quartiles. For each quartile, we take the mean value to be its representative. For example, the mean of the four income groups is approximately $12905, 25388, 39277, 71242$. Then we design a synthetic SCM that has the same distribution as this selected dataset. This is used to generate 30000 samples, from which we evaluate the expression given in Ex. 1. While evaluating the expression in 7, we use add-one smoothing since the dataset is not strictly positive. We use bootstrapping to get the confidence interval. The ground truth is obtained from the synthetic SCM. The code and dataset are available in the accompanying repository.

Fair Machine Learning: The expression for $\mathbb{E}[Y_{x_1} - Y_{x_0} \mid m_1]$ can be obtained as follows:

$$\mathbb{E}[Y_{x_1} - Y_{x_0} \mid m_1] = \frac{P(Y_{x_1} = 1, m_1)}{P(m_1)} - \frac{P(Y_{x_0} = 1, m_1)}{P(m_1)} \tag{229}$$

We first compute $P(Y_{X_1} = 1, m_1)$.

$$\begin{aligned}
&P(Y_{x_1} = 1, m_1) && (230)\\
&= P(Y_{x_1} = 1, x_1, m_1) + P(Y_{x_1} = 1, x_0, m_1) && (231)\\
&= P(y_1, x_1, m_1) + \sum_{m,z} P(Y_{x_1 mz} = 1, M_{x_1 z} = m, M_{x_0 z} = 1, X_z = 0, z) && (232)\\
&= P(y_1, x_1, m_1) + \sum_{z} P(Y_{x_1 m_1 z} = 1, M_{x_1 z} = m_1, M_{x_0 z} = 1, X_z = 0, z) && (233)\\
&= P(y_1, x_1, m_1) + \sum_{z} P(Y_{x_1 m_1 z} = 1, M_{x_0 z} = 1, X_z = 0, z) && (234)\\
&= P(y_1, x_1, m_1) + \sum_{z} P(x_0, z) P(m_1 \mid x_0, z) P(y_1 \mid m_1, x_1, z) && (235)
\end{aligned}$$

Now, we compute $P(Y_{X_0} = 1, m_1)$.

$$P(Y_{x_0} = 1, m_1) \tag{236}$$
$$= P(Y_{x_0} = 1, x_0, m_1) + P(Y_{x_0} = 1, x_1, m_1) \tag{237}$$
$$= P(Y_{x_0} = 1, x_0, m_1) + \sum_{m,z} P(Y_{x_0 mz} = 1, M_{x_0 z} = m, M_{x_1 z} = 1, X_z = 1, z) \tag{238}$$
$$= P(Y_{x_0} = 1, x_0, m_1) + \sum_{z} P(Y_{x_0 m_0 z} = 1, M_{x_0 z} = 0, M_{x_1 z} = 1, X_z = 1, z) \tag{239}$$
$$+ \sum_{z} P(Y_{x_0 m_1 z} = 1, M_{x_0 z} = 1, M_{x_1 z} = 1, X_z = 1, z) \tag{240}$$
$$= P(y_1, x_0, m_1) + \sum_{z} \big[P(Y_{x_0 m_0 z} = 1, M_{x_1 z} = 1, X_z = 1, z) - P(Y_{x_0 mz} = 1, M_{x_0 z} = 1, X_z = 1, z)\big] \tag{241}$$
$$+ \sum_{z} P(Y_{x_0 m_1 z} = 1, M_{x_0 z} = 1, X_z = 1, z) \tag{242}$$
$$= P(y_1, x_0, m_1) + \sum_{z} P(x_1, z) P(y \mid x_0, m_0, z) \big[P(m_1 \mid x_1, z) - P(m_1 \mid x_0, z)\big] \tag{243}$$
$$+ \sum_{z} P(x_1, z) P(y \mid x_0, m_1, z) P(m_1 \mid x_0, z) \tag{244}$$

Now, the final quantity can be calculated using Eq. 229.

The code for running the experiments is provided in the Supplementary Material and also in the following code repository: https://anonymous.4open.science/r/M-ID-Experiments/