Non-Parametric Methods for Partial Identification of Causal Effects

Junzhe Zhang [1] **Elias Bareinboim** [1]

Abstract

This paper investigates the problem of learning *non-identifiable* causal effects (Pearl, 2000, Def. 3.2.4) from a combination of observational data and qualitative assumptions about the underlying data-generating model, represented in the form of a directed acyclic causal diagram. Some prominent work exists on bounding causal effects for specific causal diagrams, such as the instrumental variable (IV) setting (Balke & Pearl, 1994). Still, there is no systematic way for bounding effects for almost any other non-identifiable settings, beyond IV models. This paper fills the gap in this area. Specifically, we introduce a novel family of canonical causal models accompanied by a systematic procedure that allows for the replacement of unspecified domains (possibly continuous) of any unobserved variable with discrete variables taking on a countable set of values. This construction's importance stems from the fact that the resulting model is equivalent in all observational and interventional distributions over finite observed variables, i.e., it is complete. Building on this new characterization, we develop an efficient algorithm for bounding causal effects from observational data in arbitrary causal diagrams.

1. Introduction

This paper studies the problem of inferring causal effects of interventions from a combination of non-experimental data (e.g., observational studies) and qualitative assumptions about the data-generating process. This problem arises in diverse fields such as artificial intelligence, statistics, cognitive science, economics, and the health and social sciences. For example, investigators in health sciences are interested in the effects of treatments on diseases; policymakers in healthcare want to evaluate the effectiveness of policy decisions during a pandemic; developers of an online advertisement engine are concerned with the effects of ad-placements in order to increase the click-through rate; and so on.

We consider the settings where the underlying causal mechanisms are represented in the form of a *causal diagram* (Pearl, 2000), which is a directed acyclic graph where arrows indicate the potential existence of functional relationships among corresponding variables and some variables are not observed. The problem of deciding the feasibility of uniquely discerning values of a causal query from the non-experimental data provided with the causal diagram, called the identification of causal effects, has been studied in the causal inference literature. Several criteria and algorithms have been developed to solve this problem (Pearl, 2000; Spirtes et al., 2000; Tian & Pearl, 2002; Shpitser & Pearl, 2006; Huang & Valtorta, 2006; Bareinboim & Pearl, 2016), which means that the conditions under which the target causal effect could be point-identified from the observational data are, at least theoretically, well-understood.

The combination of quantitative knowledge and observational data, however, does not always permit one to uniquely determine the target causal effect. Such settings, called *non-identifiable*, indicate that there exist more than one parametrization of the target effect that are compatible with the same observational data and causal diagram (Pearl, 2000, Def. 3.2.4). The problem of partial identification, which concerns learning causal effects in non-identifiable settings, has been a subject of growing interest in the domains of causal inference (Robins, 1989; Manski, 1990; Balke & Pearl, 1995; Chickering & Pearl, 1996; Balke & Pearl, 1997; Evans, 2012; Richardson et al., 2014; Cinelli et al., 2019), and more recently, in machine learning (Zhang & Bareinboim, 2017; Kallus & Zhou, 2018; 2020; Kilbertus et al., 2020; Zhang & Bareinboim, 2021). Among these works, two approaches are often employed: (1) bounds are derived for the target causal effect under a minimal set of assumptions; or (2) additional assumptions are invoked under which the causal effect could be identified, and then the sensitivity analysis is performed to assess how the target effect varies as these additional assumptions are changed.

This paper focuses on the bounding approach. (Robins, 1989; Manski, 1990) derived the first informative bounds over the causal effects from studies with imperfect compliance, under a set of non-parametric assumptions called *instrumental variables* (IV), graphically described in Fig. 1a.

[1]Causal Artificial Intelligence Lab, Columbia University. Correspondence to: Junzhe Zhang <junzhez@cs.columbia.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, PMLR 108, 2021.

(Zhang & Bareinboim, 2019; 2020) extended this bounding strategy to estimate system dynamics in sequential decision-making settings. Another line of research, taken by (Balke & Pearl, 1994; 1997), proposed a family of canonical models[1] with finite unobserved states, which sufficiently represent all observations and consequences of interventions in IV models (Fig. 1a) with binary X, Y, Z. Based on this canonical characterization, (Balke & Pearl, 1994) reduced the bounding problem to a series of equivalent linear programs; the resulting bounds improved over previous results (Robins, 1989; Manski, 1990). (Chickering & Pearl, 1997) further used Bayesian techniques to investigate the sharpness of these bounds with regard to the observational sample size.

Despite all these advances, there still exist outstanding challenges for bounding causal effects from observational data when an arbitrary causal diagram is provided. Almost all strategies described so far are applicable only for a limited collection of causal diagrams (Robins, 1989; Manski, 1990; Balke & Pearl, 1994; Zhang & Bareinboim, 2019). While simple generalizations to an arbitrary causal diagram might be feasible (Balke & Pearl, 1994), these strategies often fail to account for all constraints imposed by the diagram (to be shown later on), resulting in loose bounds. A systematic approach for bounding in general settings is still missing.

Our goal in this paper is to overcome these challenges. We study the partial identification of causal effects from observational data, provided with an arbitrary causal diagram. We focus on structural causal models (Pearl, 2000, Ch. 7.1) where observed variables are categorical; we *make no parametric assumption* about the underlying causal mechanisms, neither the structural functions nor the domains of unobserved variables. More specifically, our contributions are as follows. (1) We propose a new family of canonical models for arbitrary causal diagrams. (2) We show that for any structural causal model, there exists a canonical model such that it is equivalent in all observational and interventional distributions over endogenous (observed) variables. (3) Based on this complete characterization, we develop a systematic algorithm to bound interventional distributions from finite observational data provided with an arbitrary causal diagram. Simulation results confirm that our algorithm consistently improves over existing methods. Due to space constraints, the proofs are provided in Appendix B.

1.1. Preliminaries

We introduce in this section some basic notations and definitions used throughout the paper. We use capital letters to denote variables (X), small letters for their values (x) and Ω_X for their domains. For an arbitrary set $\boldsymbol{X}$, let $|\boldsymbol{X}|$ be its cardinality. We will consistently use $P(\boldsymbol{x})$ to represent probabilities $P(\boldsymbol{X} = \boldsymbol{x})$. Finally, $\mathbb{1}\{\boldsymbol{Z} = \boldsymbol{z}\}$ is an indicator function that returns 1 if $\boldsymbol{Z} = \boldsymbol{z}$ holds true; otherwise 0.

[1] A family of causal models is said to be canonical if any other structural causal model could be reduced to it.

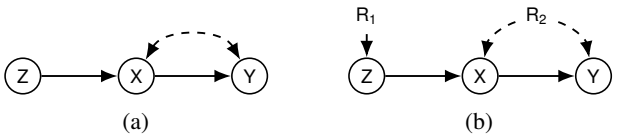


Figure 1: (a) the causal diagram $\mathcal{G}$ containing a treatment X, an outcome Y and an instrument Z; (b) the simple canonical diagram $\mathcal{H}$ where exogenous variables R_1, R_2 are shown.

The basic semantical framework of our analysis rests on *structural causal models* (SCMs) (Pearl, 2000, Ch. 7). An SCM M is a tuple $\langle \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{F}, P \rangle$ where $\boldsymbol{V}$ is a set of endogenous variables and $\boldsymbol{U}$ is a set of exogenous variables. $\boldsymbol{F}$ is a set of functions where each $f_V \in \boldsymbol{F}$ decides values of an endogenous variable $V \in \boldsymbol{V}$ taking as argument a combination of other variables in the system. That is, $v \leftarrow f_V(pa_V, u_V), Pa_V \subseteq \boldsymbol{V}, U_V \subseteq \boldsymbol{U}$. Exogenous variables in $\boldsymbol{U}$ are mutually independent; P is an exogenous distribution deciding values of $\boldsymbol{U}$. Naturally, each SCM M induces an joint distribution $P(\boldsymbol{v})$ over endogenous variables $\boldsymbol{V}$, called the *endogenous distribution*. An intervention on an arbitrary subset $\boldsymbol{X} \subseteq \boldsymbol{V}$, denoted by $\text{do}(\boldsymbol{x})$, is an operation where values of $\boldsymbol{X}$ are set to constants $\boldsymbol{x}$, regardless of how they are ordinarily determined. For an SCM M, let $M_{\boldsymbol{x}}$ denote a submodel of M induced by intervention $\text{do}(\boldsymbol{x})$. The interventional distribution $P(\boldsymbol{v}|\text{do}(\boldsymbol{x}))$ induced by $\text{do}(\boldsymbol{x})$ is defined as the distribution over $\boldsymbol{V}$ in $M_{\boldsymbol{x}}$, namely, $P_M(\boldsymbol{v}|\text{do}(\boldsymbol{x})) \triangleq P_{M_{\boldsymbol{x}}}(\boldsymbol{v})$. For a detailed survey on SCMs, we refer readers to (Pearl, 2000, Ch. 7).

Each SCM is associated with a causal diagram $\mathcal{G}$ (e.g., Fig. 1a), which is a directed acyclic graph (DAG) where nodes represent endogenous variables $\boldsymbol{V}$ and arrows represent the arguments Pa_V of each function f_V. By convention, exogenous variables $\boldsymbol{U}$ are often not explicitly shown. Instead, a bi-directed arrow between nodes V_i and V_j indicates the presence of unobserved confounders (UCs) affecting both V_i and V_j, i.e., $U_{V_i} \cap U_{V_j} \neq \emptyset$. We will use standard graph-theoretic family abbreviations for graphical relationships such as parents and children. For example, the set of parents of $\boldsymbol{X}$ in $\mathcal{G}$ is denoted by $pa(\boldsymbol{X})_{\mathcal{G}} = \cup_{X \in \boldsymbol{X}} pa(X)_{\mathcal{G}}$; ch are similarly defined. We write Pa, Ch if arguments are included as well, e.g. $Pa(\boldsymbol{X})_{\mathcal{G}} = pa(\boldsymbol{X})_{\mathcal{G}} \cup \boldsymbol{X}$. A path consisting of only bi-directed arrows is called a bi-directed path. A pair of nodes $V_i, V_j \in \boldsymbol{V}$ belong to the same *c-component* in a causal diagram $\mathcal{G}$ (Tian & Pearl, 2002) if they are connected by a bi-directed path in $\mathcal{G}$ (e.g., $X \leftrightarrow Y$ in Fig. 1a). A c-component $\boldsymbol{C}$ in $\mathcal{G}$ is maximal if there exists no other c-component that contains $\boldsymbol{C}$. We denote by $\mathcal{C}(\mathcal{G})$ the collection of all maximal c-components in $\mathcal{G}$ which, in turn, forms a partition over endogenous variables $\boldsymbol{V}$.

2. Characterizing Interventional Distributions of Causal Diagrams

We are interested in computing an interventional distribution $P(\boldsymbol{y}|\text{do}(\boldsymbol{x}))$ of an *unknown* SCM M^*. We only have access to the causal diagram $\mathcal{G}$ and the observational distribution $P(\boldsymbol{v})$ associated with M^*; endogenous variables $\boldsymbol{V}$ are assumed to be categorical. Let $\mathscr{M} = \{\forall M : \mathcal{G}_M = \mathcal{G}\}$, i.e., the set of all the SCMs of $\mathcal{G}$ and let $\mathscr{M}_{\text{o}} = \{\forall M \in \mathscr{M} : P_M(\boldsymbol{v}) = P(\boldsymbol{v})\}$ be all the SCMs in $\mathscr{M}$ consistent with observations $P(\boldsymbol{v})$. For any $\boldsymbol{X} \subseteq \boldsymbol{V}$, we denote by $P_{\mathscr{M}_{\text{o}}}(\boldsymbol{y}|\text{do}(\boldsymbol{x}))$ the set of all interventional distributions $P(\boldsymbol{y}|\text{do}(\boldsymbol{x}))$ induced by candidate SCMs in family $\mathscr{M}_{\text{o}}$, i.e.,

$$P_{\mathscr{M}_{\text{o}}}(\boldsymbol{y}|\text{do}(\boldsymbol{x})) = \{P_M(\boldsymbol{y}|\text{do}(\boldsymbol{x})) : \forall M \in \mathscr{M}_{\text{o}}\}. \quad (1)$$

We will refer to $P_{\mathscr{M}_{\text{o}}}(\boldsymbol{v}|\text{do}(\boldsymbol{x}))$ as the *parameter space* of $P(\boldsymbol{y}|\text{do}(\boldsymbol{x}))$ with regard to $\mathcal{G}$ and $P(\boldsymbol{v})$. Our goal in this paper is to learn such parameter space. Since we do not have parametric knowledge about exogenous variables $\boldsymbol{U}$ and functions $\boldsymbol{F}$, it is infeasible to directly infer candidate models in $\mathscr{M}_{\text{o}}$. We will circumvent this issue by studying an alternative, parametrized set of SCMs $\mathscr{N}_{\text{o}}$, which generates the same parameter space of interventional distributions.

Definition 1 (do-Equivalence)**.** Let $\mathscr{M}, \mathscr{N}$ be sets of SCMs with endogenous variables $\boldsymbol{V}$. $\mathscr{M}$ and $\mathscr{N}$ are *do-equivalent* if for any $M \in \mathscr{M}$ (or $N \in \mathscr{N}$), there exists $N \in \mathscr{N}$ ($M \in \mathscr{M}$) such that $\forall \boldsymbol{X} \subset \boldsymbol{V}$, $P_M(\boldsymbol{v}|\text{do}(\boldsymbol{x})) = P_N(\boldsymbol{v}|\text{do}(\boldsymbol{x}))$.

As a special case, for any do-equivalent pair $\mathscr{M}$ and $\mathscr{N}$, they coincide in parameters of the observational distribution, i.e., $P_{\mathscr{M}}(\boldsymbol{v}) = P_{\mathscr{N}}(\boldsymbol{v})$. A parametrized set of SCMs $\mathscr{N}$ is said to *completely characterize* interventional distributions of a causal diagram $\mathcal{G}$ if $\mathscr{N}$ is do-equivalent to the set of all the SCMs $\mathscr{M}$ associated with $\mathcal{G}$. The following corollary could be derived based on the definition of do-equivalence.

Corollary 1. *For an observational distribution $P(\boldsymbol{v})$, let $\mathscr{M}, \mathscr{N}$ be two sets of SCMs with endogenous variables $\boldsymbol{V}$. Let subsets $\mathscr{M}_{\text{o}} = \{\forall M \in \mathscr{M} : P_M(\boldsymbol{v}) = P(\boldsymbol{v})\}$ and $\mathscr{N}_{\text{o}} = \{\forall N \in \mathscr{N} : P_N(\boldsymbol{v}) = P(\boldsymbol{v})\}$. If $\mathscr{M}$ and $\mathscr{N}$ are do-equivalent, then $\mathscr{M}_{\text{o}}$ and $\mathscr{N}_{\text{o}}$ are do-equivalent.*

Corol. 1 suggests a systematic approach for learning the parameter space of $P(\boldsymbol{y}|\text{do}(\boldsymbol{x}))$ from the combination of the causal diagram $\mathcal{G}$ and the observational distribution $P(\boldsymbol{v})$. One could (1) derive a complete characterization $\mathscr{N}$ of $\mathcal{G}$; (2) find $\mathscr{N}_{\text{o}} \subseteq \mathscr{N}$ consistent with $P(\boldsymbol{v})$; and (3) compute interventional distribution $P(\boldsymbol{y}|\text{do}(\boldsymbol{x}))$ of candidate models in $\mathscr{N}_{\text{o}}$. The induced parameter space $P_{\mathscr{N}_{\text{o}}}(\boldsymbol{y}|\text{do}(\boldsymbol{x}))$ is ensured to coincide with actual parameters in $P_{\mathscr{M}_{\text{o}}}(\boldsymbol{y}|\text{do}(\boldsymbol{x}))$.

2.1. A Simple Canonical Model

To realize this goal, we introduce a family of canonical causal models that could represent interventional distributions in a causal diagram. For each endogenous node $V \in \boldsymbol{V}$, we denote by $\mathscr{H}_V = \{h_V : \Omega_{Pa_V} \mapsto \Omega_V\}$ a hypothesis class of functions mapping from domains of observed parents Pa_V to domains of V. Let $\mathscr{H}_V$ be ordered by $h_V^{(1)}, \dots, h_V^{(m_V)}$, where $m_V = |\mathscr{H}_V|$. For each configuration $U_V = u_V$, the induced function $f_V(\cdot, u_V)$ must correspond to a unique element in $\mathscr{H}_V$. We could thus divide the exogenous domains Ω_{U_V} into partitions $\mathcal{U}_V^{(1)}, \dots, \mathcal{U}_V^{(m_V)}$ so that for any $u_V \in \mathcal{U}_V^{(r_V)}$, function $f_V(\cdot, u_V) = h_V^{(r_V)}$.

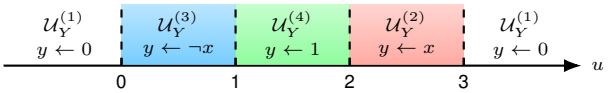


Figure 2: A graphical representation of partitions $\mathcal{U}_Y^{(i)}$.

As an example, consider an SCM M associated with the causal diagram of Fig. 1a where the endogenous variables $X, Y, Z \in \{0, 1\}$; exogenous variable U is drawn from a normal distribution $\mathcal{N}(0, 1)$; values of Y are given by

$$y \leftarrow f_Y(x, u) = \mathbb{1}\{u \in [x, 2 + x]\}. \quad (2)$$

Let functions in $\mathscr{H}_Y$ be ordered by $h_Y^{(1)}(x) = 0$, $h_Y^{(2)}(x) = x$, $h_Y^{(3)}(x) = \neg x$ and $h_Y^{(4)}(x) = 1$. These functions correspond to population partitions as follows (Balke & Pearl, 1994; Heckerman & Shachter, 1995; Imbens & Rubin, 1997): "*never-recover*" $\mathcal{U}_Y^{(1)} = (-\infty, 0) \cup (3, +\infty)$, "*helped*" $\mathcal{U}_Y^{(2)} = (2, 3]$, "*hurt*" $\mathcal{U}_Y^{(3)} = [0, 1)$, and "*always-recover*" $\mathcal{U}_Y^{(4)} = [1, 2]$ [2]. We show in Fig. 2 the graphical representation of partitions $\mathcal{U}_Y^{(i)}$. As U_Y varies along its domain, regardless of how complex the variation is, its only effect is to switch the functional relationship between Y and X among elements in the class $\mathscr{H}_Y$. Formally,

Lemma 1. *For an SCM $\langle \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{F}, P \rangle$, for each $V \in \boldsymbol{V}$, function $f_V \in \boldsymbol{F}$ could be written as follows:*

$$f_V(pa_V, u_V) = \sum_{r_V=1}^{m_V} h_V^{(r_V)}(pa_V) \mathbb{1}\left\{u_V \in \mathcal{U}_V^{(r_V)}\right\}. \quad (3)$$

Consider again the SCM M of Fig. 1a. The decomposition in Lem. 1 implies that function f_Y of Eq. (2) could be decomposed over population partitions $\mathcal{U}_Y^{(i)}$ as follows:

$$\begin{aligned} f_Y(x, u) = {} & \mathbb{1}\{u \in (2, 3]\}x \\ & + \mathbb{1}\{u \in [0, 1)\}\neg x + \mathbb{1}\{u \in [1, 2]\}. \end{aligned} \quad (4)$$

Note that when consider intervention $\text{do}(X = 0)$ in M, the outcome $Y = 1$ if and only if $U \in [0, 1]$ or $U \in [1, 2]$, i.e.,

$$P(Y = 1|\text{do}(X = 0)) = P(U \in [0, 1)) + P(U \in [1, 2]).$$

[2] For instance, $y \leftarrow 1$ for all $x = 0, 1$ (i.e., $h_Y^{(4)}$) if and only if $0 \le u \le 2$ and $1 \le u \le 3$; their intersection $\mathcal{U}_Y^{(4)} = [1, 2]$.

Importantly, despite the continuous nature of U, for SCMs described in Fig. 1a with finite X, Y, Z, one could always discretize domains of U into finite partitions while preserving all interventional distributions. We next describe a simple procedure that formalizes such discretization step.

Definition 2. For a causal diagram $\mathcal{G}$, the simple canonical diagram $\mathcal{H}$ is given by INVERSEPROJECT$(\mathcal{G}, \mathcal{C}(\mathcal{G}))$.

The procedure INVERSEPROJECT is described in Alg. 1. In words, a *simple canonical diagram* $\mathcal{H}$ is obtained from a causal diagram $\mathcal{G}$ by (1) removing bi-directed arrows; (2) for each c-component $\boldsymbol{C}_i \in \mathcal{C}(\mathcal{G})$, adding an exogenous node R_i; and (3) adding an arrow $R_i \to V$ for each $V \in \boldsymbol{C}_i$. For example, the causal diagram $\mathcal{G}$ in Fig. 1a consists of two c-compoments $\boldsymbol{C}_1 = \{Z\}$ and $\boldsymbol{C}_2 = \{X, Y\}$; Fig. 1b shows the corresponding simple canonical diagram $\mathcal{H}$.

Definition 3. Given a simple canonical diagram $\mathcal{H}$, a *simple canonical causal model* (for short, s-CCM) M associated with $\mathcal{H}$ is an SCM $\langle \boldsymbol{V}, \boldsymbol{R}, \boldsymbol{F}, P \rangle$ where:

1. Each $R_i \in \boldsymbol{R}$ is a set $\{R_V : \forall V \in ch(R_i)_{\mathcal{H}}\}$.
2. For each $V \in \boldsymbol{V}$, R_V is an index in $\{1, \dots, m_V\}$.
3. Values of each $V \in \boldsymbol{V}$ are given by a function $v \leftarrow f_V(pa_V, r_V) = h_V^{(r_V)}(pa_V)$.

In a simple canonical model, each exogenous variable R_V is an index for the hypothesis class $\mathscr{H}_V$. Once $R_V = r_V$ is fixed, the functional relationship between V and Pa_V are determined. Consider now an s-CCM N associated with $\mathcal{H}$ of Fig. 1b, where $X, Y, Z \in \{0, 1\}$ and $R_1 = \{R_Z\}$, $R_2 = \{R_X, R_Y\}$. Def. 3 implies that in such model N,

$$P(Y = 1|\text{do}(X = 0)) = P(R_Y = 3) + P(R_Y = 4).$$

Now let $P(R_Y = 3) = P(U \in [0, 1))$ and $P(R_Y = 4) = P(U \in [1, 2])$. It follows immediately from above equations that M and N induce the same $P(Y = 1|\text{do}(X = 0))$.

Lemma 2. *Given a causal diagram $\mathcal{G}$, let $\mathcal{H}$ be the simple canonical diagram of $\mathcal{G}$. For any SCM M associated with $\mathcal{G}$, there exists an s-CCM N associated with $\mathcal{H}$ such that for any $\boldsymbol{X} \subset \boldsymbol{V}$, $P_M(\boldsymbol{v}|do(\boldsymbol{x})) = P_N(\boldsymbol{v}|do(\boldsymbol{x}))$.*

Lem. 2 implies that any SCM M could be reduced to a s-CCM N with finite latent states that is equivalent with regard to all interventional distributions. Thm. 1 summarizes the characterization of simple canonical causal models while simplifying some of its detailed, explicit parametric forms.

Theorem 1. *For a causal diagram $\mathcal{G}$ and its simple canonical diagram $\mathcal{H}$, consider the following conditions:*

1. *$\mathscr{M}$ is the set of all SCMs associated with $\mathcal{G}$.*
2. *$\mathscr{N}$ is the set of all SCMs associated with $\mathcal{H}$ where for each $R_i \in \boldsymbol{R}$, $|\Omega_{R_i}| = \prod_{V \in ch(R_i)} |\mathscr{H}_V|$.*
3. *$V_i \leftrightarrow V_j \in \mathcal{G}$ whenever a path $V_i \leftarrow R_k \to V_j \in \mathcal{H}$.*

Then $\mathscr{M}$ and $\mathscr{N}$ are do-equivalent.

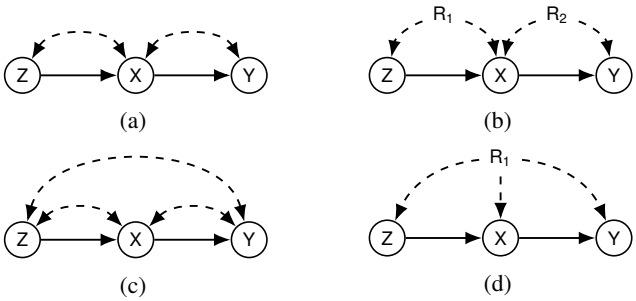


Figure 3: (a, c) causal diagrams $\mathcal{G}$; (b, d) canonical diagrams $\mathcal{H}$ where exogenous variables $\boldsymbol{R}$ are explicitly shown.

Algorithm 1 INVERSEPROJECT

Input: $\mathcal{G}$ and $\{\boldsymbol{C}_1, \dots, \boldsymbol{C}_K\}$ where $\boldsymbol{C}_k \subseteq \boldsymbol{V}$.
Output: A canonical diagram $\mathcal{H}$ where all exogenous variables $\boldsymbol{R}$ are shown explicitly.
For each node $V \in \mathcal{G}$, add a node V in $\mathcal{H}$.
For each arrow $V_i \to V_j \in \mathcal{G}$, add $V_i \to V_j$ in $\mathcal{H}$.
For each $\boldsymbol{C}_k$, add an empty node R_i in $\mathcal{H}$.
For each $V \in \boldsymbol{C}_k$, add an arrow $R_i \to V$.

Consider again the causal diagram $\mathcal{G}$ and the simple canonical diagram $\mathcal{H}$ described in Fig. 1. Since $X \leftrightarrow Y \in \mathcal{G}$ and $X \leftarrow R_2 \to Y \in \mathcal{H}$, it follows from Thm. 1 that the set of SCMs $\mathscr{N}$ with finite exogenous states completely characterizes interventional distributions of Fig. 1a, which confirms the canonical partitioning approah in (Balke & Pearl, 1994).

3. Refined Canonical Characterization

Despite its power, the representation of simple canonical diagrams does not always capture all constraints on interventional distributions imposed by a causal diagram, especially when Condition 3 in Thm. 1 is violated. For concreteness, consider the causal diagram $\mathcal{G}$ of Fig. 3a. For any SCM M associated with $\mathcal{G}$, its interventional distribution $P(y, z|\text{do}(x))$ satisfies the following relationship:

$$\begin{aligned} P(z, y|\text{do}(x)) &= \sum_{u_1} P(z|u_1)P(u_1) \sum_{u_2} P(y|x, u_2)P(u_2) \\ &= P(z|\text{do}(x))P(y|\text{do}(x)) \end{aligned}$$

That is, variables Z and Y are independent under the intervention $\text{do}(x)$. We show in Fig. 3d the simple canonical diagram $\mathcal{H}$ obtained from $\mathcal{G}$. Nodes Z and Y are now connected by a backdoor path $Z \leftarrow R_1 \to Y$. This suggests that there exists an SCM such that Z and Y no longer independent due to the presence of confounder R_1. For instance, consider an SCM N where R_1 is a binary variable uniformly drawn from $\{0, 1\}$; values of X, Y, Z are decided by $z \leftarrow r_1$, $x \leftarrow z \wedge r_1$ and $y \leftarrow x \wedge r_1$ respectively. It is verifiable that $P(Z = Y|\text{do}(X = 1)) = 1$ in N, i.e., Z and

Y are perfectly correlated under $do(X = 1)$. This means that the characterization of simple canonical diagram $\mathcal{H}$ in Fig. 3d is not complete for the causal diagram $\mathcal{G}$ in Fig. 3a.

To address this issue, we will introduce a refined canonical diagram that preserves all do-implications in a causal diagram. We build on the notation of c-components and define a novel type of clustering called the *confounded clique*.

Definition 4 (c-clique)**.** For a causal diagram $\mathcal{G}$, a subset $\boldsymbol{C} \subseteq \boldsymbol{V}$ is a c-clique if any pair $V_i, V_j \in \boldsymbol{C}$ is connected by a *bi-directed arrow* in $\mathcal{G}$.

A c-clique $\boldsymbol{C}$ in $\mathcal{G}$ is *maximal* if there exists no other c-clique that contains $\boldsymbol{C}$. We denote by $c(\mathcal{G})$ the set of all maximal c-cliques in a causal diagram $\mathcal{G}$. For instance, the causal diagram $\mathcal{G}$ of Fig. 3a has a single c-component $\boldsymbol{C} = \{X, Y, Z\}$; however, there exist in $\mathcal{G}$ two c-cliques $\boldsymbol{C}_1 = \{Z, X\}$ and $\boldsymbol{C}_2 = \{X, Y\}$. On the other hand, if we connect nodes Z, Y with a bi-directed arrow (e.g., see Fig. 3c), c-components $\mathcal{C}(\mathcal{G})$ and c-cliques $c(\mathcal{G})$ now coincide, equating to a single clustering $\boldsymbol{C} = \{X, Y, Z\}$.

Definition 5. For a causal diagram $\mathcal{G}$, the canonical diagram $\mathcal{H}$ is given by INVERSEPROJECT$(\mathcal{G}, c(\mathcal{G}))$ (Alg. 1).

As an example, Fig. 3b shows the canonical diagram corresponding to Fig. 3a where new exogenous nodes R_1, R_2 correspond to c-cliques $\boldsymbol{C}_1 = \{Z, X\}$ and $\boldsymbol{C}_2 = \{X, Y\}$ respectively. When nodes are fully connected by bi-directed arrows (see Fig. 3c), the canonical diagram coincides with the simple canonical diagram, which is shown in Fig. 3d.

Lemma 3. *For a causal diagram $\mathcal{G}$ and its canonical diagram $\mathcal{H}$, consider the following conditions:*

1. *$\mathscr{M}$ is the set of all SCMs associated with $\mathcal{G}$.*
2. *$\mathscr{N}$ is the set of all SCMs associated with $\mathcal{H}$.*

Then $\mathscr{M}$ and $\mathscr{N}$ are do-equivalent.

Lem. 3 says that to estimate interventional distributions in a causal diagram $\mathcal{G}$, it is sufficient to consider only SCMs associated with the corresponding canonical diagram $\mathcal{H}$. However, the domains of exogenous variables $\boldsymbol{R}$ and the forms of structural functions $\boldsymbol{F}$ in $\mathcal{H}$ are still unspecified. The remainder of this section grounds this specification.

3.1. Canonical Causal Models

We first consider SCMs where each exogenous variable $U \in \boldsymbol{U}$ is a real in $\mathbb{R}$. Let exogenous variables $\boldsymbol{U}$ be ordered by $U_1, \ldots, U_n$, where $n = |\boldsymbol{U}|$. Each canonical type $\mathcal{U}_V^{(j)}$, $j = 1, \ldots, m_V$, could be decomposed into a countable set of (almost) disjoint rectangles $\mathcal{R}_V^{(1)}, \mathcal{R}_V^{(2)}, \ldots$, i.e, $\mathcal{U}_V^{(j)} = \cup_{i \in \mathbb{N}} \mathcal{R}_V^{(i)}$. Each rectangle $\mathcal{R}_V^{(i)}$ is of the form

$$\mathcal{R}_V^{(i)} = \times_{U_k \in U_V} \mathcal{R}_{V,k}^{(i)}, \text{ where } \mathcal{R}_{V,k}^{(i)} = [a_k, b_k] \subset \mathbb{R}. \quad (5)$$

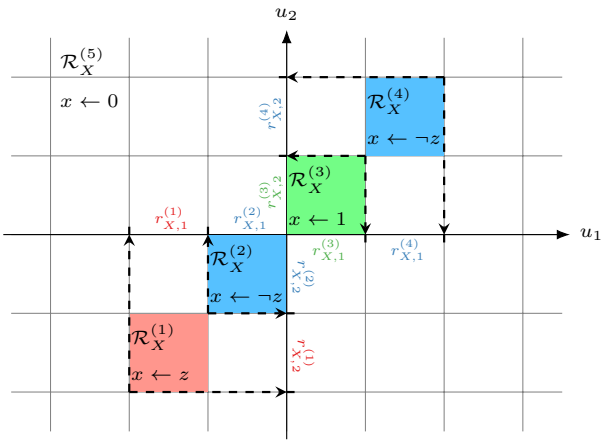


Figure 4: A graphical representation of rectangles $\mathcal{R}_X^{(i)}$.

For $|U_V| = 1$, $\mathcal{R}_V^{(i)}$ are bounded intervals in $\mathbb{R}$; for $|U_V| = 2$, they are the usual four-sided rectangles in $\mathbb{R}^2$, and so on. Each function f_V in Eq. (3) could be further written as:

$$f_V(pa_V, u_V) = \sum_{i \in \mathbb{N}} h_V^{\left(j_V^{(i)}\right)}(pa_V) \mathbb{1}\left\{u_V \in \mathcal{R}_V^{(i)}\right\}. \quad (6)$$

Among quantities in the above equation, $j_V^{(i)}$ is an index in $\{1, \ldots, m_V\}$ for any $i \in \mathbb{N}$. $\mathbb{1}\left\{u_V \in \mathcal{R}_V^{(i)}\right\}$ could be written as a product of indicator variables as follows,

$$\mathbb{1}\left\{u_V \in \mathcal{R}_V^{(i)}\right\} = \prod_{U_k \in U_V} \mathbb{1}\left\{u_k \in \mathcal{R}_{V,k}^{(i)}\right\} \quad (7)$$

The decomposition of Eqs. (6) and (7) permits us to discretize the domains of exogenous variables $\boldsymbol{U}$ while preserving independence relationships among them.

To illustrate, consider an SCM M inducing the causal diagram $\mathcal{G}$ of Fig. 3a. Exogenous variables $\boldsymbol{U} = \{U_1, U_2\}$ where U_1, U_2 are i.i.d. draws from the normal distribution $\mathcal{N}(0, 1)$. Endogenous X, Y, Z are binary variables in $\{0, 1\}$. Let functions in $\mathscr{H}_X$ be ordered by $h_X^{(1)}(z) = 0$, $h_X^{(2)}(z) = z$, $h_X^{(3)}(z) = \neg z$ and $h_X^{(4)}(z) = 1$, which corresponds to the population of "*never-taker*", "*complier*", "*defier*" and "*always-taker*". Function f_X is defined as:

$$\begin{aligned} x &\leftarrow f_X(z, u_1, u_2) \\ &= \mathbb{1}\{u_1, u_2 \in [-2, -1]\}\, z + \mathbb{1}\{u_1, u_2 \in [-1, 0]\}\, \neg z \\ &+ \mathbb{1}\{u_1, u_2 \in [0, 1]\} + \mathbb{1}\{u_1, u_2 \in [1, 2]\}\, \neg z. \end{aligned}$$

We show in Fig. 4 a graphical representation of the above decomposition. Similar to previous discussion around Fig. 2, $P(X = 1|\text{do}(Z = 0))$ is decomposable as follows:

$$\begin{aligned} &P(X = 1|\text{do}(Z = 0)) \\ &= \sum_{i=2}^{4} P(U_1 \in [i-3, i-2]) P(U_2 \in [i-3, i-2]). \end{aligned}$$

In the above equation, the domains of U_1, U_2 are discretized into a set of intervals in $\mathbb{R}$ respectively. We next generalize this observation and define a family of canonical causal models taking on a countable set of exogenous states.

Definition 6. Given a canonical diagram $\mathcal{H}$, a *canonical causal model* (for short, CCM) M associated with $\mathcal{H}$ is an SCM $\langle \boldsymbol{V}, \boldsymbol{R}, \boldsymbol{F}, P \rangle$ where:

1. Each $R_k \in \boldsymbol{R}$ is a set $\{R_{V,k} : \forall V \in de(R_k)_{\mathcal{H}}\}$.
2. Values of each $R_{V,k}$ is binary sequence $\left(r^{(i)}_{V,k}\right)_{i\in\mathbb{N}} \in \{0,1\}^{\mathbb{N}}$ drawn from a countable set $\Omega_{R_{V,k}}$.
3. For each $V \in \boldsymbol{V}$, $R_V = pa(V)_{\mathcal{H}} \cap \boldsymbol{R}$.
4. Values of each $V \in \boldsymbol{V}$ are given by a function $f_V(pa_V, r_V) \in \boldsymbol{F}$ defined as, for $j^{(i)}_V \in \{1, \ldots, m_V\}$,

$$f_V(pa_V, r_V) = \sum_{i\in\mathbb{N}} h_V^{\left(j^{(i)}_V\right)}(pa_V) \prod_{R_k \in R_V} r^{(i)}_{V,k}, \qquad (8)$$

where $\sum_i \prod_{R_k} r^{(i)}_{V,k} \leq 1, \forall r_{V,k} \in \Omega_{R_{V,k}}, \forall R_k \in R_V$.

Consider an CCM N associated with Fig. 3b. Fix a sequence $j_X = \{2,3,4,3\}$. Function f_X of N is defined as:

$$f_X(z, r_{X,1}, r_{X,2}) = r^{(1)}_{X,1} r^{(1)}_{X,2} z + r^{(2)}_{X,1} r^{(2)}_{X,2} \neg z + r^{(3)}_{X,1} r^{(3)}_{X,2} + r^{(4)}_{X,1} r^{(4)}_{X,2} \neg z,$$

where $R_{X,k}$, $k = 1,2$, is a binary sequence $\left(r^{(i)}\right)_{i=1}^{4}$ and $\sum_i r^{(i)} \leq 1$. The construction of Def. 6 implies that in N,

$$P(X = 1|\text{do}(Z = 0)) = \sum_{i=2}^{4} P(R^{(i)}_1 = 1) P(R^{(i)}_2 = 1).$$

Now let $P(R^{(i)}_{X,k} = 1) = P(U_k \in [i-3, i-2])$, $k = 1,2$. It immediately follows that M and N induce the same $P(X = 1|\text{do}(Z = 0))$. Our next result shows that the expressive power of CCMs in representing interventional distributions is indeed universal, not limited to SCMs satisfying Eq. (6).

Lemma 4. *Given a causal diagram $\mathcal{G}$, let $\mathcal{H}$ be the canonical diagram of $\mathcal{G}$. For any SCM M associated with $\mathcal{G}$, there exists a CCM N associated with $\mathcal{H}$ such that for any $\boldsymbol{X} \subset \boldsymbol{V}$, $P_M(\boldsymbol{v}|do(\boldsymbol{x})) = P_N(\boldsymbol{v}|do(\boldsymbol{x}))$.*

A mental image for understanding Lem. 4 is that any open set in $\mathbb{R}^n$ could be decomposed into a countable set of (almost) disjoint rectangles. Any distribution $P(u)$ could be projected into probabilities over a real line $\mathbb{R}$. We can now state one of the main results of this paper:

Theorem 2. *For a causal diagram $\mathcal{G}$ and its canonical diagram $\mathcal{H}$, consider the following conditions:*

1. *$\mathscr{M}$ is the set of all SCMs associated with $\mathcal{G}$.*
2. *$\mathscr{N}$ is the set of all SCMs associated with $\mathcal{H}$ where each $R_i \in \boldsymbol{R}$ is a discrete variable drawn from $\mathbb{N}$.*

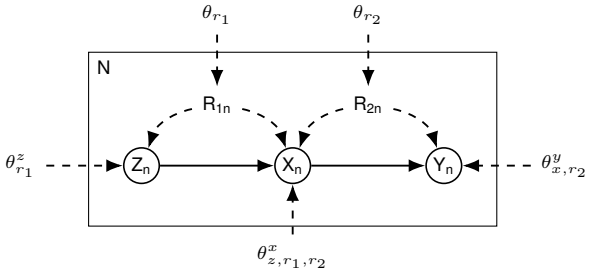


Figure 5: The data-generating process for observational data $\{Z_n, X_n, Y_n\}_{n=1}^N$ in SCMs associated with Fig. 3b.

Then $\mathscr{M}$ and $\mathscr{N}$ are do-equivalent.

Thm. 2 says that for any SCM, we could always replace its exogenous domains with a finite set of discrete variables $\boldsymbol{R}$ taking on values in countable set. The resulting model preserves all causal relationships and is equivalent in all observational and interventional distributions over $\boldsymbol{V}$.

4. Bounding Interventional Distributions

We now apply the canonical characterization introduced so far and develop an efficient algorithm for the partial identification of interventional distributions. Given a causal diagram $\mathcal{G}$ and finite observations $\{\boldsymbol{V}_i = \boldsymbol{v}_i\}_{i=1}^N$ drawn from $P(\boldsymbol{v})$, our goal is to obtain causal bounds $[l, h]$ containing the actual interventional distribution $P(\boldsymbol{y}|\text{do}(\boldsymbol{x}))$.

Let $\mathcal{H}$ be the canonical diagram obtained from $\mathcal{G}$ and let variables $\boldsymbol{Z} = \boldsymbol{V} \setminus (\boldsymbol{Y} \cup \boldsymbol{X})$. The interventional distribution $P(\boldsymbol{y}|\text{do}(\boldsymbol{x}))$ of any SCM associated with $\mathcal{H}$ is given by:

$$P(\boldsymbol{y}|\text{do}(\boldsymbol{x})) = \sum_{\boldsymbol{r},\boldsymbol{z}} \prod_{V \in \boldsymbol{V}\setminus\boldsymbol{X}} P(v|pa_V, r_V) \prod_{R\in\boldsymbol{R}} P(r) \qquad (9)$$

For instance, in the canonical diagram of Fig. 3b, it follows from Eq. (9) that $P(y|\text{do}(x)) = \sum_{r_2} P(y|x, r_2) P(r_2)$.

The characterization of Thm. 2 permits us to consider only discrete distributions $P(r)$ and $P(v|pa_V, r_V)$. For every $R \in \boldsymbol{R}$, let θ_r be parameters of $P(r)$, i.e., $\theta_r = P(r)$. We assume θ_r are drawn from a Dirichlet process (Ferguson, 1973). We follow the stick-breaking construction (Sethuraman, 1994) which successively breaks pieces off a unit-length stick; θ_r are proportions of each of the infinite pieces relative to its original size. Formally, for $r = 1, 2, \ldots,$

$$\theta_r = \beta_r \prod_{i=1}^{r-1}(1 - \beta_i), \qquad \beta_r \sim \texttt{Beta}(1, \alpha_R), \qquad (10)$$

where α_R is a hyperparameter. For every endogenous $V \in \boldsymbol{V}$, let parameters $\theta^v_{pa_V, r_V} = P(v|pa_V, r_V)$. We assume that $\theta^v_{pa_V, r_V}$ are drawn from a Dirichlet prior given by

$$\theta^v_{pa_V, r_V} \sim \texttt{Dir}(\lambda_V), \qquad \forall pa_V, r_V, \qquad (11)$$

where λ_V is a hyperparameter. We will use $\boldsymbol{R}_n = \{R_n : \forall R \in \boldsymbol{R}\}$ to represent exogenous variables generating the n-th observation $\boldsymbol{V}_n$. For example, Fig. 5 shows the graphical representation of the data-generating process associated with Fig. 3b, spanning over observations $\{\boldsymbol{V}_n = \boldsymbol{v}_n\}_{n=1}^N$.

Let $\boldsymbol{\lambda} = \{\lambda_V : \forall V \in \boldsymbol{V}\}$ and $\boldsymbol{\alpha} = \{\alpha_R : \forall R \in \boldsymbol{R}\}$. We denote by parameters $\theta^{\boldsymbol{y}}_{\text{do}(\boldsymbol{x})} = P(\boldsymbol{y}|\text{do}(\boldsymbol{x}))$. The problem of bounding $P(\boldsymbol{y}|\text{do}(\boldsymbol{x}))$ is reducible to the estimation of the *posterior distribution* $P\left(\theta^{\boldsymbol{y}}_{\text{do}(\boldsymbol{x})} \mid \bar{\boldsymbol{v}}; \boldsymbol{\alpha}, \boldsymbol{\lambda}\right)$ given observational data $\bar{\boldsymbol{v}} = \{\boldsymbol{V}_i = \boldsymbol{v}_i\}_{i=1}^N$ (Chickering & Pearl, 1996). The remainder of this section describes a sampling procedure to approximate this posterior distribution.

Posterior Sampling Since each $R \in \boldsymbol{R}$ could take on infinitely many values, it is difficult to directly compute parameters $\theta^{\boldsymbol{y}}_{\text{do}(\boldsymbol{x})}$ from $\theta_r, \theta^v_{pa_V, u_V}$. We will introduce an efficient estimator to address this issue. For each $R \in \boldsymbol{R}$, let a sequence $\bar{R} = \{R_1, \ldots, R_N\}$ and let $\Omega^*_R = \{r^1, \ldots, r^K\}$ denote unique values which $\bar{R}$ take on. We approximate $\theta^{\boldsymbol{y}}_{\text{do}(\boldsymbol{x})}$ by summing over finite subdomains $\Omega^* = \times_R \Omega^*_R$,

$$\theta^{\boldsymbol{y}}_{\text{do}(\boldsymbol{x}),N} = \sum_{\boldsymbol{r} \in \Omega^*} \sum_{\boldsymbol{z}} \prod_{V \in \boldsymbol{V} \setminus \boldsymbol{X}} \theta^v_{pa_V, r_V} \prod_{R \in \boldsymbol{R}} \theta_r. \tag{12}$$

As the number of observational samples $N \to \infty$, each subdomain Ω^* converges to the complete Ω_R. It follows from Eqs. (9) and (12) that $\theta^{\boldsymbol{y}}_{\text{do}(\boldsymbol{x}),N}$ converges to parameters $\theta^{\boldsymbol{y}}_{\text{do}(\boldsymbol{x})}$ in probability, i.e., $\theta^{\boldsymbol{y}}_{\text{do}(\boldsymbol{x}),N}$ is a consistent estimator.

Lemma 5. *Given observations* $\{\boldsymbol{V}_i = \boldsymbol{v}_i\}_{i=1}^N$, *for* $\forall \epsilon > 0$,

$$\lim_{N \to \infty} P\left(\sup_{\boldsymbol{x}, \boldsymbol{y}} \left|\theta^{\boldsymbol{y}}_{do(\boldsymbol{x}),N} - \theta^{\boldsymbol{y}}_{do(\boldsymbol{x})}\right| > \epsilon \mid \bar{\boldsymbol{v}}; \boldsymbol{\alpha}, \boldsymbol{\lambda}\right) = 0.$$

Lem. 5 implies that it is sufficient to sample estimators $\theta^{\boldsymbol{y}}_{\text{do}(\boldsymbol{x}),N}$ given observational data $\bar{\boldsymbol{v}}$. We first describe a collapsed Gibbs sampler to draw exogenous variables $\bar{\boldsymbol{r}} = \{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N\}$ from the posterior distribution $P(\bar{\boldsymbol{r}} \mid \bar{\boldsymbol{v}}; \boldsymbol{\alpha}, \boldsymbol{\lambda})$. In particular, the algorithm iteratively samples each $R_n \in \boldsymbol{R}_n$ from the complete conditional:

$$P(r_n | \bar{\boldsymbol{v}}, \bar{\boldsymbol{r}} \setminus \{r_n\}; \boldsymbol{\alpha}, \boldsymbol{\lambda}) \propto P(\boldsymbol{v}_n | \bar{\boldsymbol{v}}_{-n}, \bar{\boldsymbol{r}}; \boldsymbol{\lambda}) \, P(r_n | \bar{r}_{-n}; \alpha_R), \tag{13}$$

where $\bar{\boldsymbol{v}}_{-n} = \bar{\boldsymbol{v}} \setminus \{\boldsymbol{v}_n\}$ and $\bar{r}_{-n} = \bar{r} \setminus \{r_n\}$. Among quantities in the above equation, the first term $P(\boldsymbol{v}_n | \bar{\boldsymbol{v}}_{-n}, \bar{\boldsymbol{r}}; \boldsymbol{\lambda})$ is a product of Dirichlet-multinomial distributions, marginalizing over parameters $\theta^v_{pa_V, u_V}$. With a slight abuse of notation, we write $\{r^1, \ldots, r^K\}$ as K unique values which $\bar{r}_{-n}$ take on. The second term $P(r_n | \bar{r}_{-n}; \alpha_R)$ is given by:

$$P(r_n = r^k | \bar{r}_{-n}; \alpha_R) = \frac{\sum_{i \neq n} \mathbb{1}\{r_i = r^k\}}{N - 1 + \alpha_R}, \qquad P(r_n \notin \bar{r}_{-n} | \bar{r}_{-n}; \alpha_R) = \frac{\alpha_R}{N - 1 + \alpha_R}, \tag{14}$$

Algorithm 2 BOUNDCAUSALEFFECT

1: **Input:** Observations $\{\boldsymbol{V}_n = \boldsymbol{v}_n\}_{n=1}^N$, a canonical diagram $\mathcal{H}$, outcomes $\boldsymbol{y}$, treatments $\boldsymbol{x}$.
2: **Output:** A causal bound $[l, h]$ over $P(\boldsymbol{y}|\text{do}(\boldsymbol{x}))$.
3: **Initialization:** set $l = 1$, $h = 0$.
4: **while** $[l, h]$ has not converged **do**
5: Sample $\bar{\boldsymbol{r}} \mid \bar{\boldsymbol{v}}$ through Gibbs sampling (Eq. (13)).
6: For every $R \in \boldsymbol{R}$, sample $\theta_r \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{r}}$ (Eq. (15)).
7: For every $V \in \boldsymbol{V}$, sample $\theta^v_{pa_V, r_V} \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{r}}$ (Eq. (16)).
8: Compute a bound $[l_N, h_N]$ over parameters $\theta^{\boldsymbol{y}}_{\text{do}(\boldsymbol{x})}$ from θ_r and $\theta^v_{pa_V, r_V}$ (Eq. (17)).
9: Let $l = \min\{l, l_N\}$, $h = \max\{h, h_N\}$.
10: **end while**

We next describe methods to sample parameters $\theta_r, \theta^v_{pa_V, u_V}$ given observations $\bar{\boldsymbol{v}}$ and exogenous states $\bar{\boldsymbol{r}}$. For each $R \in \boldsymbol{R}$, the posterior $P\left(\theta_{r^k}, \text{for all } r^k \in \Omega^*_R \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{r}}; \boldsymbol{\alpha}, \boldsymbol{\lambda}\right)$ is a Dirichlet distribution given by (Sethuraman, 1994):

$$\left(\theta_{r^1}, \ldots, \theta_{r^K}, \theta_{r \notin \Omega^*_R}\right) \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{r}} \sim \texttt{Dir}(n_R), \quad \text{where} \tag{15}$$

$$n_R^{(k)} = \sum_{n=1}^N \mathbb{1}\{r_n = r^k\} \; (\forall k \leq K), \text{ and } n_R^{(K+1)} = \alpha_R.$$

For every endogenous $V \in \boldsymbol{V}$, the posterior distribution $P\left(\theta^v_{pa_V, r_V} \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{r}}; \boldsymbol{\alpha}, \boldsymbol{\lambda}\right)$ for any pa_V, r_V is given by:

$$\theta^v_{pa_V, r_V} \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{r}} \sim \texttt{Dir}(\lambda_V + n_{pa_V, r_V}), \quad \text{where} \tag{16}$$

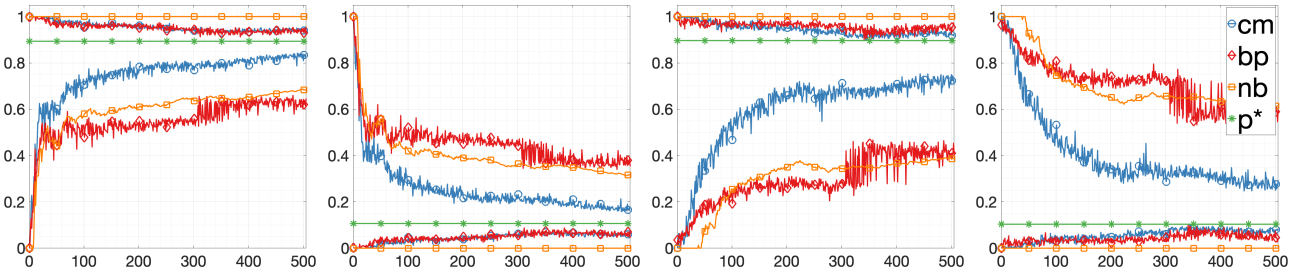
$$n^{(k)}_{pa_V, r_V} = \sum_{n=1}^N \mathbb{1}\{v_n = v^k, pa_{V_n} = pa_V, r_{V_n} = r_V\}.$$

We could then compute the estimator $\theta^{\boldsymbol{y}}_{\text{do}(\boldsymbol{x}),N}$ from sampled parameters $\theta_r, \theta^v_{pa_V, u_V}$ following Eq. (12).

We are now ready to introduce an algorithm that bounds an interventional distribution $P(\boldsymbol{y}|\text{do}(\boldsymbol{x}))$ from observational data $\{\boldsymbol{V}_n = \boldsymbol{v}_n\}_{n=1}^N$. Details of our algorithm are summarized in Alg. 2. It repeatedly samples form the posterior distribution $P\left(\theta^{\boldsymbol{y}}_{\text{do}(\boldsymbol{x})} \mid \bar{\boldsymbol{v}}; \boldsymbol{\alpha}, \boldsymbol{\lambda}\right)$ and update the current bound estimates (Step 9). More specifically, it first draws exogenous states $\bar{\boldsymbol{r}}$ from $P(\bar{\boldsymbol{r}}|\bar{\boldsymbol{v}}; \boldsymbol{\alpha}, \boldsymbol{\lambda})$ using the collapsed Gibbs sampler (Step 5). At Steps 6-7, it then samples parameters θ_r and $\theta^v_{pa_V, r_V}$ from the posterior distribution given $\bar{\boldsymbol{v}}, \bar{\boldsymbol{r}}$, only for exogenous states visited by previously drawn $\bar{\boldsymbol{r}}$. To address the deviation due to finite observational samples, Step 8 computes from parameters $\theta_r, \theta^v_{pa_V, r_V}$ a bound $[l_N, h_N]$ over $\theta^{\boldsymbol{y}}_{\text{do}(\boldsymbol{x})} \sim P\left(\theta^{\boldsymbol{y}}_{\text{do}(\boldsymbol{x})} \mid \bar{\boldsymbol{v}}; \boldsymbol{\alpha}, \boldsymbol{\lambda}\right)$ given by:

$$l_N = \theta^{\boldsymbol{y}}_{\text{do}(\boldsymbol{x}),N}, \qquad h_N = l + 1 - \prod_{R \in \boldsymbol{R}} \sum_{r \in \Omega^*_R} \theta_r, \tag{17}$$

where $\theta^{\boldsymbol{y}}_{\text{do}(\boldsymbol{x}),N}$ is defined in Eq. (12). The validity of the above bound follows from the decomposition of Eq. (9).



(a) $X = 0, Y = 0$ (b) $X = 0, Y = 1$ (c) $X = 1, Y = 0$ (d) $X = 1, Y = 1$

Figure 6: Causal bounds over the effect $P(y|\text{do}(x))$ of aspirin X on the death Y in the International Stroke Trial (IST). The x-axle represents the number of observational samples. *cm* are new bounds derived by Alg. 2 (blue); *bp* are derived using simple canonical partitions (red); *nb* are the natural bounds (yellow). The actual effect $P(y|do(x))$ is labeled as p^* (green).

	$P(y\|\text{do}(x))$	l_{cm}	h_{cm}	l_{bp}	h_{bp}	l_{nb}	h_{nb}
$X = 0, Y = 0$	0.8934	**0.8157**	**0.9070**	0.5495	0.9206	0.6984	0.9478
$X = 0, Y = 1$	0.1066	**0.0930**	**0.1843**	0.0794	0.4505	0.0522	0.3016
$X = 1, Y = 0$	0.8964	**0.7391**	**0.9185**	0.2944	0.9591	0.4416	0.9970
$X = 1, Y = 1$	0.1036	**0.0815**	**0.2609**	0.0409	0.7056	0.0030	0.5584

Table 1: Causal bounds $[l, h]$ over the interventional distribution $P(y|\text{do}(x))$. The optimal bounds are marked in **bold**.

5. Experiments

We demonstrate our algorithms on the patient data collected from the International Stroke Trial (IST) (Carolei et al., 1997). In all experiments, we test Alg. 2 (*cm*) with the characterization of canonical causal models (Thm. 2) and a sampling procedure (Chickering & Pearl, 1996) using the canonical partitioning introduced in (Balke & Pearl, 1994) (*bp*), which is equivalent to the characterization of simple canonical models (Thm. 1). As a baseline, we also include the natural bounds (Robins, 1989; Manski, 1990) estimated at the 95% confidence level (*nb*) and the actual interventional distribution $P(\boldsymbol{y}|\text{do}(\boldsymbol{x}))$ (p^*). We refer readers to Appendix A for details on experiment setups.

IST was a large, randomized, open trial of up to 14 days of antithrombotic therapy after stroke onset (Carolei et al., 1997). We are interested in evaluating the effect of aspirin X (0 for no aspirin allocation, 1 otherwise) on the death of the patient by the end of treatment Y (1 for dead, 0 otherwise). To emulate the unobserved confounding described in Fig. 3a, we introduce an instrumental variable $Z \in \{0, 1\}$, filter data following a selection rule taking Z, X as input, and hide all columns except for X, Y, Z. Details on the data selection procedure are described in Appendix A. We show in Table 1 the causal bounds $[l, h]$ estimated from 1×10^3 observational samples. The actual interventional distribution $P(y|\text{do}(x))$ are estimated from the randomized trial data. We also plot causal bounds computed from first 500 observational samples in Fig. 6; x-axle represents the number of observational data. Our analysis reveals that *cm*, *bp* and *nb* all obtain valid causal bounds containing the target interventional distribution $P(y|\text{do}(x))$. Our new bounding strategy (*cm*) consistently dominate state-of-art natural bounds (*nb*, *bp*). The characterization of simple canonical partitions (*bp*) performs the worst among all algorithms. This is probably due to the fact that it does not capture the independence relationship between Z and Y under intervention $\text{do}(x)$.

Overall, we found that our algorithm utilizing canonical causal models could obtain efficient bounds over interventional distributions; the derived bounds consistently dominate state-of-art methods. Simulation results corroborate the completeness of the canonical characterization. It captures all constraints on interventional distributions associated with a causal diagram, thus leading to tight causal bounds.

6. Conclusion

We introduced a canonical characterization of interventional implications imposed by qualitative assumptions of the data-generating process, represented in the form of a directed acyclic causal diagram. Such characterization permits the replacement of unspecified domains (possibly continuous) of unobserved variables, in any structural causal model, with discrete variables taking on a countable set of values. This is important since all non-identifiable results are due to the presence of unobserved confounders. We showed that the resulting model is equivalent in all observational and interventional distributions over categorical observed variables. Using this novel representation, we developed an efficient algorithm to derive bounds over causal effects from finite observational data in an arbitrary causal diagram.

References

Balke, A. and Pearl, J. Counterfactual probabilities: Computational methods, bounds and applications. In *Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence*, UAI'94, pp. 46–54, 1994.

Balke, A. and Pearl, J. Counterfactuals and policy analysis in structural models. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 11–18, 1995.

Balke, A. and Pearl, J. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1172–1176, September 1997.

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

Bauer, H. Probability theory and elements of measure theory. *Holt*, 1972.

Bauer, H. *Measure and integration theory*, volume 26. Walter de Gruyter, 2011.

Carolei, A. et al. The international stroke trial (ist): a randomized trial of aspirin, subcutaneous heparin, both, or neither among 19435 patients with acute ischaemic stroke. *The Lancet*, 349:1569–1581, 1997.

Chickering, D. and Pearl, J. A clinician's apprentice for analyzing non-compliance. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, volume Volume II, pp. 1269–1276. MIT Press, Menlo Park, CA, 1996.

Chickering, D. and Pearl, J. A clinician's tool for analyzing non-compliance. *Computing Science and Statistics*, 29 (2):424–431, 1997.

Cinelli, C., Kumor, D., Chen, B., Pearl, J., and Bareinboim, E. Sensitivity analysis of linear structural causal models. In *International Conference on Machine Learning*, pp. 1252–1261, 2019.

Evans, R. J. Graphical methods for inequality constraints in marginalized dags. In *2012 IEEE International Workshop on Machine Learning for Signal Processing*, pp. 1–6. IEEE, 2012.

Ferguson, T. S. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pp. 209–230, 1973.

Heckerman, D. and Shachter, R. Decision-theoretic foundations for causal reasoning. *Journal of Artificial Intelligence Research*, 3:405–430, 1995.

Huang, Y. and Valtorta, M. Pearl's calculus of intervention is complete. In Dechter, R. and Richardson, T. (eds.), *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pp. 217–224. AUAI Press, Corvallis, OR, 2006.

Imbens, G. W. and Rubin, D. B. Bayesian inference for causal effects in randomized experiments with noncompliance. *The annals of statistics*, pp. 305–327, 1997.

Kallus, N. and Zhou, A. Confounding-robust policy improvement. In *Advances in neural information processing systems*, pp. 9269–9279, 2018.

Kallus, N. and Zhou, A. Confounding-robust policy evaluation in infinite-horizon reinforcement learning. *Advances in Neural Information Processing Systems*, 2020.

Kilbertus, N., Kusner, M. J., and Silva, R. A class of algorithms for general instrumental variable models. In *Advances in Neural Information Processing Systems*, 2020.

Manski, C. Nonparametric bounds on treatment effects. *American Economic Review, Papers and Proceedings*, 80: 319–323, 1990.

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

Richardson, A., Hudgens, M. G., Gilbert, P. B., and Fine, J. P. Nonparametric bounds and sensitivity analysis of treatment effects. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 29(4):596, 2014.

Robins, J. The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. In Sechrest, L., Freeman, H., and Mulley, A. (eds.), *Health Service Research Methodology: A Focus on AIDS*, pp. 113–159. NCHSR, U.S. Public Health Service, Washington, D.C., 1989.

Sethuraman, J. A constructive definition of dirichlet priors. *Statistica sinica*, pp. 639–650, 1994.

Shpitser, I. and Pearl, J. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence*, pp. 1219–1226. 2006.

Spirtes, P., Glymour, C. N., and Scheines, R. *Causation, prediction, and search*. MIT press, 2000.

Tian, J. and Pearl, J. A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence*, pp. 567–573. AAAI Press/The MIT Press, Menlo Park, CA, 2002.

Zhang, J. and Bareinboim, E. Transfer learning in multi-armed bandits: a causal approach. In *Proceedings of the 26th IJCAI*, pp. 1340–1346, 2017.

Zhang, J. and Bareinboim, E. Near-optimal reinforcement learning in dynamic treatment regimes. In *Advances in Neural Information Processing Systems*, pp. 13401–13411, 2019.

Zhang, J. and Bareinboim, E. Designing optimal dynamic treatment regimes: A causal reinforcement learning approach. In *Proceedings of the Thirty-seventh International Conference on Machine Learning (ICML-20)*, 2020.

Zhang, J. and Bareinboim, E. Bounding causal effects on continuous outcomes. In *Proceedings of the 35nd AAAI Conference on Artificial Intelligence*, 2021.