# Partial Counterfactual Identification from Observational and Experimental Data

**Junzhe Zhang**[1] **Jin Tian**[2] **Elias Bareinboim**[1]

## Abstract

This paper investigates the problem of bounding counterfactual queries from an arbitrary collection of observational and experimental distributions and qualitative knowledge about the underlying data-generating model represented in the form of a causal diagram. We show that all counterfactual distributions in an arbitrary structural causal model (SCM) with *discrete* observed domains could be generated by a canonical family of SCMs with the same causal diagram where unobserved (exogenous) variables are also *discrete*, taking values in finite domains. Utilizing the canonical SCMs, we translate the problem of bounding counterfactuals into that of polynomial programming whose solution provides optimal bounds for the counterfactual query. Solving such polynomial programs is in general computationally expensive. We then develop effective Monte Carlo algorithms to approximate optimal bounds from a combination of observational and experimental data. Our algorithms are validated extensively on synthetic and real-world datasets.

## 1. Introduction

This paper studies the problem of inferring counterfactual queries from a combination of observations, experiments, and qualitative assumptions about the phenomenon under investigation. The assumptions are represented in the form of a *causal diagram* (Pearl, 1995), which is a directed acyclic graph where arrows indicate the potential existence of functional relationships among variables; some variables are unobserved. This problem arises in diverse fields such as artificial intelligence, statistics, cognitive science, economics, and the health and social sciences. For example, when investigating the gender discrimination in college admission, one may ask "what would the admission outcome be for a female applicant had she been a male?" Such a counterfactual query contains conflicting information: in the real world, the applicant is female; in the hypothetical world, she is not. Formally, counterfactual lies on top of a hierarchy of increasingly expressive languages that also include observations and interventions, which is called *Pearl Causal Hierarchy* (Pearl & Mackenzie, 2018; Bareinboim et al., 2020). In general, counterfactuals are not immediately computable from observational and experimental distributions.

The problem of identifying counterfactual distributions from the combination of data and a causal diagram has been studied in the causal inference literature. First, there exists a sound and complete proof system for reasoning about counterfactual queries (Halpern, 1998). While such a system, in principle, is sufficient in evaluating any identifiable counterfactual expression, it lacks a proof guideline that efficiently determines the feasibility of such evaluation. Further, Shpitser & Pearl (2007) studied an algorithm for the identification of counterfactuals from all possible controlled experiments. There exist also algorithms for identifying path-specific effects from experimental data (Avin et al., 2005) and observational data (Shpitser & Sherman, 2018). More recently, Correa et al. (2021) developed the first sound, complete, and efficient algorithm that decides whether any nested counterfactual distribution is identifiable from an arbitrary combination of observations and experiments.

In practice, the combination of qualitative assumptions and data does not always permit one to uniquely determine the target counterfactual query. In such cases, the counterfactual query is said to be *non-identifiable*. *Partial identification* methods concern with inferring about the target counterfactual probability in non-identifiable settings. Several algorithms have been developed to derive informative bounds over counterfactual probabilities from the combination of observational and experimental data (Manski, 1990; Robins, 1989; Balke & Pearl, 1994; 1997; Tian & Pearl, 2000; Evans, 2012; Richardson et al., 2014; Zhang & Bareinboim, 2017; Kallus & Zhou, 2018; Finkelstein & Shpitser, 2020; Kilbertus et al., 2020; Zhang & Bareinboim, 2021).

In this work, we build on the approach introduced by (Balke & Pearl, 1994), which involves direct discretization of unobserved domains, also referred to as the canonical parti-

[1] Department of Computer Science, Columbia University [2] Department of Computer Science, Iowa State University. Correspondence to: Junzhe Zhang <junzhez@cs.columbia.edu>.

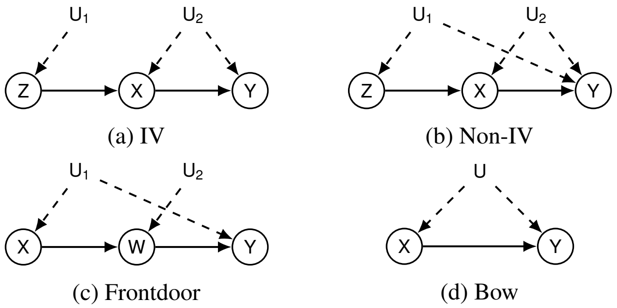
*Proceedings of the 39th International Conference on Machine Learning*, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

tioning or the principal stratification (Frangakis & Rubin, 2002; Pearl, 2011). Consider the causal diagram in Fig. 1a, where $X, Y, Z$ are binary variables in $\{0, 1\}$; $U_2$ is an unobserved variable taking values in an arbitrary continuous domain. Balke & Pearl (1994) showed that domains of $U_2$ could be discretized into 16 equivalent classes without changing the original counterfactual distributions and the graphical structure in Fig. 1a. For instance, suppose that values of $U_2$ are drawn from an arbitrary distribution $P^*(U_2)$ over a continuous domain. It has been shown that the observational distribution $P(x, y, z)$ could be reproduced by a generative model of the form $P(x, y, z) = \sum_u P(x|u_2, z)P(y|x, u_2)P(u_2)P(z)$, where $P(U_2)$ is a discrete distribution over a finite domain $\{1, \dots, 16\}$.

Using the finite-state representation of unobserved variables, Balke & Pearl (1997) derived tight bounds on treatment effects under a set of constraints called *instrumental variables* (e.g., Fig. 1a). Chickering & Pearl (1997); Imbens & Rubin (1997); Richardson et al. (2011) applied the parsimony of finite-state representation in a Bayesian framework, to obtain credible intervals for the posterior distribution of causal effects in noncompliance settings. Despite the optimality guarantees in their treatments, these bounds were only derived for specific settings, but could not be immediately extended to other causal diagrams without loss of generality. A systematic strategy for partial identification in an arbitrary causal diagram is still missing. There are significant challenges in bounding any counterfactual query in an arbitrary causal diagram given an arbitrary collection of observational and experimental data.

The goal of this paper is to overcome these challenges. We show that when inferring about counterfactual distributions (over finite observed variables) in an arbitrary causal diagram, one could restrict domains of unobserved variables to a finite space without loss of generality. This result allows us to develop novel partial identification algorithms to bound unknown counterfactual probabilities from an arbitrary combination of observational and experimental data. In some ways, this paper can be seen as closing a long-standing open problem introduced by (Balke & Pearl, 1994), where they solve a special bounding instance from the observational distribution in the case of the instrumental variable graph.

More specifically, our contributions are summarized as follows. (1) We introduce a special family of *canonical structural causal models*, and show that it could represent all categorical counterfactual distributions in any arbitrary causal diagram. (2) Building on this result, we translate the partial identification task into an equivalent polynomial program. Solving such a program leads to optimal bounds over target counterfactual probabilities. (3) We develop an effective Monte Carlo Markov Chain algorithm to approximate optimal bounds from a finite number of observational and experimental data. Finally, our algorithms are validated on synthetic and real-world datasets. Given the space constraints, all proofs are provided in the complete technical report (Zhang et al., 2021, Appendix A).



(a) IV (b) Non-IV

(c) Frontdoor (d) Bow

Figure 1: Causal diagrams containing treatment $X$, outcome $Y$, ancestor $Z$, mediator $W$, and unobserved variables $U_i$.

### 1.1. Preliminaries

We introduce in this section some basic notations and definitions that will be used throughout the paper. We use capital letters to denote variables ($X$), small letters for their values ($x$) and $\Omega_X$ for their domains. For an arbitrary set $\boldsymbol{X}$, let $|\boldsymbol{X}|$ be its cardinality. The probability distribution over variables $\boldsymbol{X}$ is denoted by $P(\boldsymbol{X})$. For convenience, we consistently use $P(\boldsymbol{x})$ as a shorthand for the probability $P(\boldsymbol{X} = \boldsymbol{x})$. Finally, the indicator function $\mathbb{1}_{\boldsymbol{X}=\boldsymbol{x}}$ returns 1 if an event $\boldsymbol{X} = \boldsymbol{x}$ holds; otherwise, $\mathbb{1}_{\boldsymbol{X}=\boldsymbol{x}}$ is equal to 0.

The basic semantical framework of our analysis rests on *structural causal models* (SCMs) (Pearl, 2000; Bareinboim & Pearl, 2016). An SCM $M$ is a tuple $\langle \boldsymbol{V}, \boldsymbol{U}, \mathscr{F}, P\rangle$ where $\boldsymbol{V}$ is a set of endogenous variables and $\boldsymbol{U}$ is a set of exogenous variables. $\mathscr{F}$ is a set of functions where each $f_V \in \mathscr{F}$ 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 $U \in \boldsymbol{U}$ are mutually independent, values of which are drawn from the exogenous distribution $P(\boldsymbol{U})$. Naturally, $M$ induces a joint distribution $P(\boldsymbol{V})$ over endogenous variables $\boldsymbol{V}$, called the *observational distribution*. Each SCM $M$ is also associated with a causal diagram $\mathcal{G}$ (e.g., Fig. 1), which is a directed acyclic graph (DAG) where solid nodes represent endogenous variables $\boldsymbol{V}$, empty nodes represent exogenous variables $\boldsymbol{U}$, and arrows represent the arguments $PA_V, U_V$ of each structural function $f_V$.

An intervention on an arbitrary subset $\boldsymbol{X} \subseteq \boldsymbol{V}$, denoted by 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 do$(\boldsymbol{x})$. For any subset $\boldsymbol{Y} \subseteq \boldsymbol{V}$, the *potential response* $\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{u})$ is defined as the solution of $\boldsymbol{Y}$ in the submodel $M_{\boldsymbol{x}}$ given $\boldsymbol{U} = \boldsymbol{u}$. Drawing values of exogenous variables $\boldsymbol{U}$ following the distribution $P(\boldsymbol{U})$ induces a *counterfactual variable* $\boldsymbol{Y}_{\boldsymbol{x}}$. Specifically, the event $\boldsymbol{Y}_{\boldsymbol{x}} = \boldsymbol{y}$ (for short, $\boldsymbol{y}_{\boldsymbol{x}}$) can be read as "$\boldsymbol{Y}$ would be $\boldsymbol{y}$ had $\boldsymbol{X}$ been $\boldsymbol{x}$". For subsets $\boldsymbol{Y}, \dots, \boldsymbol{Z}, \boldsymbol{X}, \dots, \boldsymbol{W} \subseteq \boldsymbol{V}$, the distribution over counterfactuals $\boldsymbol{Y}_{\boldsymbol{x}}, \dots, \boldsymbol{Z}_{\boldsymbol{w}}$ is defined as:

$$P\left(\boldsymbol{y}_{\boldsymbol{x}}, \dots, \boldsymbol{z}_{\boldsymbol{w}}\right) = \int_{\Omega_{\boldsymbol{U}}} \mathbb{1}_{\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{u})=\boldsymbol{y}, \dots, \boldsymbol{Z}_{\boldsymbol{w}}(\boldsymbol{u})=\boldsymbol{z}} dP(\boldsymbol{u}). \quad (1)$$

Distributions of the form $P(\boldsymbol{Y}_{\boldsymbol{x}})$ are called *interventional distributions*; when $\boldsymbol{X} = \emptyset$, $P(\boldsymbol{Y})$ coincides with the *observational distribution*. For a more detailed survey on SCMs, we refer readers to (Pearl, 2000; Bareinboim et al., 2020).

## 2. Partial Counterfactual Identification

We introduce the task of partial identification of a counterfactual probability from a combination of observational and interventional distributions, which generalizes the previous partial identifiability settings that assume observational data are given (Balke & Pearl, 1997; Imbens & Rubin, 1997). [1] Throughout this paper, we assume that domains of endogenous variables $\boldsymbol{V}$ are discrete and finite; while exogenous variables $\boldsymbol{U}$ could take values in any (continuous) domains. $P\left(\boldsymbol{Y}_{\boldsymbol{x}}, \dots, \boldsymbol{Z}_{\boldsymbol{w}}\right)$ defined above is thus a categorical distribution. Let $\mathbb{Z} = \{\boldsymbol{z}_i\}_{i=1}^m$ be a finite collection of realizations $\boldsymbol{z}_i$ for sets of variables $\boldsymbol{Z}_i \subseteq \boldsymbol{V}$. The learner has access to data collected from all of the interventional distributions in $\{P(\boldsymbol{V}_{\boldsymbol{z}}) \mid \boldsymbol{z} \in \mathbb{Z}\}$. Note that $\boldsymbol{Z} = \emptyset$ corresponds to the observational distribution $P(\boldsymbol{V})$. Our goal is to find a bound $[l, r]$ for an arbitrary counterfactual probability $P\left(\boldsymbol{y}_{\boldsymbol{x}}, \dots, \boldsymbol{z}_{\boldsymbol{w}}\right)$ from the collection of interventional distributions $\{P(\boldsymbol{V}_{\boldsymbol{z}}) \mid \boldsymbol{z} \in \mathbb{Z}\}$ and the causal diagram $\mathcal{G}$.

**Definition 2.1** (Optimal Counterfactual Bound)**.** For a causal diagram $\mathcal{G}$ and distributions $\{P(\boldsymbol{V}_{\boldsymbol{z}}) \mid \boldsymbol{z} \in \mathbb{Z}\}$, the *optimal bound* $[l, r]$ over a counterfactual probability $P\left(\boldsymbol{y}_{\boldsymbol{x}}, \dots, \boldsymbol{z}_{\boldsymbol{w}}\right)$ is defined as, respectively, the minimum and maximum of the following optimization problem:

$$\begin{aligned} \min_{M \in \mathscr{M}(\mathcal{G})} / \max \quad & P_M\left(\boldsymbol{y}_{\boldsymbol{x}}, \dots, \boldsymbol{z}_{\boldsymbol{w}}\right) \\ \text{s.t.} \quad & P_M(\boldsymbol{V}_{\boldsymbol{z}}) = P(\boldsymbol{V}_{\boldsymbol{z}}) \;\; \forall \boldsymbol{z} \in \mathbb{Z} \end{aligned} \quad (2)$$

where $\mathscr{M}(\mathcal{G})$ is the set of all SCMs associated with the diagram $\mathcal{G}$, i.e., $\mathscr{M}(\mathcal{G}) = \{\forall M \mid \mathcal{G}_M = \mathcal{G}\}$[2].

Among quantities in Eq. (2), $P_M\left(\boldsymbol{Y}_{\boldsymbol{x}}, \dots, \boldsymbol{Z}_{\boldsymbol{w}}\right)$ and $P_M(\boldsymbol{V}_{\boldsymbol{z}})$ are given in the form of Eq. (1). By its formulation, $[l, r]$ must be the tight bound containing all possible values of the target counterfactual $P\left(\boldsymbol{y}_{\boldsymbol{x}}, \dots, \boldsymbol{z}_{\boldsymbol{w}}\right)$.

Since we do not have access to the parametric forms of the underlying structural functions $f_V$ nor the exogenous distribution $P(\boldsymbol{u})$, solving the optimization problem in Eq. (2) is technically challenging. It is not clear how the existing optimization procedures can be used. Next we show the optimization problem in Eq. (2) can be reduced into a polynomial program by constructing a "canonical" SCM that is equivalent to the original SCM in representing the objective $P\left(\boldsymbol{y}_{\boldsymbol{x}}, \dots, \boldsymbol{z}_{\boldsymbol{w}}\right)$ and all constraints $P(\boldsymbol{V}_{\boldsymbol{z}}), \forall \boldsymbol{z} \in \mathbb{Z}$.

### 2.1. Canonical Structural Causal Models

Our construction relies on a special type of clustering of endogenous variables in the causal diagram, which is called *confounded components* (Tian & Pearl, 2002). For convenience, let a *bi-directed arrow* $V_i \leftrightarrow V_j$ between endogenous nodes $V_i, V_j \in \boldsymbol{V}$ be defined as a sequence $V_i \leftarrow U_k \rightarrow V_k$ where $U_k \in \boldsymbol{U}$ is an exogenous parent shared by $V_i, V_j$. A *bi-directed path* is a consecutive sequence of bi-directed arrows. Formally,

**Definition 2.2.** For a causal diagram $\mathcal{G}$, a subset $\boldsymbol{C} \subseteq \boldsymbol{V}$ is said to be a c-component if any pair $V_i, V_j \in \boldsymbol{C}$ is connected by a bi-directed path in $\mathcal{G}$.

A c-component $\boldsymbol{C}$ is maximal if there does not exist any other c-component in the causal diagram $\mathcal{G}$ containing $\boldsymbol{C}$. For an arbitrary exogenous variable $U \in \boldsymbol{U}$, we denote by $\boldsymbol{C}(U)$ the maximal c-component covering $U$ in $\mathcal{G}$, i.e., $U \in \bigcup_{V \in \boldsymbol{C}(U)} U_V$. For instance, Fig. 1a contains two c-components $\boldsymbol{C}(U_1) = \{Z\}$ and $\boldsymbol{C}(U_2) = \{X, Y\}$. On the other hand, exogenous variables $U_1, U_2$ in Fig. 1b are covered by the same c-component $\boldsymbol{C}(U_1) = \boldsymbol{C}(U_2) = \{X, Y, Z\}$ since they share a common child node $Y$.

We are now ready to introduce a parametric family of canonical SCMs where values of each exogenous variable are drawn from a discrete distribution over a finite set of states.

**Definition 2.3.** An SCM $M = \langle \boldsymbol{V}, \boldsymbol{U}, \mathscr{F}, P \rangle$ is said to be a canonical SCM if

1. For every endogenous $V \in \boldsymbol{V}$, its values $v$ are given by a function $v \leftarrow f_V(pa_V, u_V)$ where for any $pa_V, u_V$, $f_V(pa_V, u_V)$ is contained in a finite domain $\Omega_V$.
2. For every exogenous $U \in \boldsymbol{U}$, its values $u$ are drawn from a finite domain $\Omega_U$; its cardinality is bounded by[3]

$$|\Omega_U| = \prod_{V \in \boldsymbol{C}(U)} |\Omega_{PA_V} \mapsto \Omega_V|. \quad (3)$$

   That is, the total number of functions mapping from domains of input $PA_V$ to $V$ for every endogenous $V$

[1] When a combination of observational and experimental data is available, there exist necessary and sufficient conditions and algorithms for deciding point identification (Bareinboim & Pearl, 2012; Lee et al., 2019; Correa et al., 2021).

[2] We will use subscript $M$ to represent the restriction to an SCM $M$. Therefore, $\mathcal{G}_M$ represents the causal diagram associated with $M$; so does counterfactual distributions $P_M\left(\boldsymbol{Y}_{\boldsymbol{x}}, \dots, \boldsymbol{Z}_{\boldsymbol{w}}\right)$.

[3] For every $V \in \boldsymbol{V}$, we denote by $\Omega_{PA_V} \mapsto \Omega_V$ the set of all possible functions mapping from domains $\Omega_{PA_V}$ to $\Omega_V$.

in the c-component $\boldsymbol{C}(U)$ covering $U$.

One may surmise that finite exogenous domains in canonical SCMs are not sufficient in capturing all the uncertainties and randomness introduced by other continuous variables. Perhaps surprisingly, we will show that the SCMs class defined above is indeed "canonical". That is, it could represent all counterfactual distributions in any SCM while maintaining the same structure of its associated causal diagram.

**Theorem 2.4.** *For an arbitrary SCM $M = \langle \boldsymbol{V}, \boldsymbol{U}, \mathscr{F}, P\rangle$, there exists a canonical SCM $N$ such that*

1. *$M$ and $N$ are associated with the same causal diagram, i.e., $\mathcal{G}_M = \mathcal{G}_N$.*
2. *For any set of counterfactual variables $\boldsymbol{Y_x}, \dots, \boldsymbol{Z_w}$, $P_M\left(\boldsymbol{Y_x}, \dots, \boldsymbol{Z_w}\right) = P_N\left(\boldsymbol{Y_x}, \dots, \boldsymbol{Z_w}\right)$.*

Thm. 2.4 establishes the expressive power of canonical SCMs in representing counterfactual distributions in a causal diagram $\mathcal{G}$. As an example, consider the "Non-IV" diagram $\mathcal{G}$ in Fig. 1b where $X, Y, Z$ are binary variables in $\{0, 1\}$. Since $U_1, U_2$ are over by the same c-component $\{X, Y, Z\}$, Eq. (3) implies that they must share the same cardinality $d = |\Omega_Z| \times |\Omega_Z \mapsto \Omega_X| \times |\Omega_X \mapsto \Omega_Y| = 32$ in canonical SCMs compatible with $\mathcal{G}$. It follows from Thm. 2.4 that the counterfactual distribution $P(X_{z'}, Y_{x'})$ in the causal diagram $\mathcal{G}$ could be generated by a canonical SCM associated with $\mathcal{G}$ and be written as follows:

$$
\begin{aligned}
&P(x_{z'}, y_{x'}) \\
&= \sum_{u_1, u_2 = 1}^{d} \mathbb{1}_{f_X(z', u_2) = x} \mathbb{1}_{f_Y(x', u_1, u_2) = y} P(u_1) P(u_2).
\end{aligned} \tag{4}
$$

More generally, Thm. 2.4 implies that counterfactual distributions $P\left(\boldsymbol{Y_x}, \dots, \boldsymbol{Z_w}\right)$ in any SCM could always be decomposed over a finite number of exogenous states. In other words, when inferring about counterfactual probabilities in an arbitrary causal diagram with discrete endogenous domains, one could assume exogenous distributions to be discrete and finite without loss of generality. Formally,

**Proposition 2.5.** *For any SCM $M = \langle \boldsymbol{V}, \boldsymbol{U}, \mathscr{F}, P(\boldsymbol{U})\rangle$, let $\boldsymbol{Y_x}, \dots, \boldsymbol{Z_w}$ be an arbitrary set of counterfactual variables. The distribution $P\left(\boldsymbol{Y_x}, \dots, \boldsymbol{Z_w}\right)$ decomposes as*

$$
\begin{aligned}
&P\left(\boldsymbol{y_x}, \dots, \boldsymbol{z_w}\right) \\
&= \sum_{U \in \boldsymbol{U}} \sum_{u=1}^{d_U} \mathbb{1}_{\boldsymbol{Y_x}(\boldsymbol{u}) = \boldsymbol{y}, \dots, \boldsymbol{Z_w}(\boldsymbol{u}) = \boldsymbol{z}} \prod_{U \in \boldsymbol{U}} P(u),
\end{aligned} \tag{5}
$$

*where for every exogenous $U \in \boldsymbol{U}$, $P(U)$ is a discrete distribution over a finite domain $\{1, \dots, d_U\}$ with cardinality $d_U = \prod_{V \in \boldsymbol{C}(U)} |\Omega_{Pa_V} \mapsto \Omega_V|$. Counterfactual variables $\boldsymbol{Y_x}(\boldsymbol{u}) = \{Y_{\boldsymbol{x}}(\boldsymbol{u}) \mid \forall Y \in \boldsymbol{Y}\}$ are recursively defined as:*

$$
Y_{\boldsymbol{x}}(\boldsymbol{u}) = \begin{cases} \boldsymbol{x}_Y & \text{if } Y \in \boldsymbol{X} \\ f_Y\left(\left(PA_Y\right)_{\boldsymbol{x}}(\boldsymbol{u}), u_Y\right) & \text{otherwise} \end{cases} \tag{6}
$$

*where $\boldsymbol{x}_Y$ is the value assigned to $Y$ in $\boldsymbol{x}$; and $\left(PA_Y\right)_{\boldsymbol{x}}(\boldsymbol{u})$ is a set of potential responses $\{V_{\boldsymbol{x}}(\boldsymbol{u}) \mid \forall V \in PA_Y\}$.*

**Related work** The discretization procedure in (Balke & Pearl, 1994) was originally designed for the "IV" diagram in Fig. 1a, and was extended to causal diagrams satisfying generalized IV constraints (Sachs et al., 2020). However, this procedure is not applicable to a general causal diagram with arbitrary structure without loss of generality; see (Zhang et al., 2021, Appendix E) for a detailed example. More recently, Evans et al. (2018) showed that for a specific class of causal diagrams satisfying a running intersection property among exogenous variables, all equality and inequality constraints over the observational distribution could be generated using discrete unobserved domains. Rosset et al. (2018) applied the classic result of Carathéodory theorem in convex geometry (Carathéodory, 1911) and developed a generic model with finite-state unobserved variables that could represent the observational distribution over discrete domains in an arbitrary causal diagram.

Thm. 2.4 generalizes existing results in several important ways. First, the theorem is applicable to *any* causal diagram, thus not relying on additional graphical conditions, e.g., IV constraints (Balke & Pearl, 1994). Second, we prove that *all* counterfactual distributions could be generated using discrete exogenous variables with finite domains, which subsume both observational and interventional distributions. Indeed, it is possible to show from Thm. 2.4 that there exists a specific subset of canonical SCMs capable of representing observational distributions in an arbitrary causal diagram.

**Proposition 2.6.** *For any SCM $M = \langle \boldsymbol{V}, \boldsymbol{U}, \mathscr{F}, P(\boldsymbol{U})\rangle$, $P(\boldsymbol{V})$ decomposes as follows:*

$$
P\left(\boldsymbol{v}\right) = \sum_{U \in \boldsymbol{U}} \sum_{u=1}^{d_U} \mathbb{1}_{\boldsymbol{V}(\boldsymbol{u}) = \boldsymbol{v}} \prod_{U \in \boldsymbol{U}} P(u), \tag{7}
$$

*where for every $U \in \boldsymbol{U}$, $d_U = \prod_{V \in Pa(\boldsymbol{C}(U))} |\Omega_V|$.*

The above result coincides with the parametrization introduced in (Rosset et al., 2018). Similarly, we also describe a more refined canonical representation for all interventional distributions in a SCM with arbitrary causal relationships.

**Proposition 2.7.** *For any SCM $M = \langle \boldsymbol{V}, \boldsymbol{U}, \mathscr{F}, P(\boldsymbol{U})\rangle$, for any subset $\boldsymbol{X}, \boldsymbol{Y} \subseteq \boldsymbol{V}$, $P(\boldsymbol{Y_x})$ decomposes as follows:*

$$
P\left(\boldsymbol{y_x}\right) = \sum_{U \in \boldsymbol{U}} \sum_{u=1}^{d_U} \mathbb{1}_{\boldsymbol{Y_x}(\boldsymbol{u}) = \boldsymbol{y}} \prod_{U \in \boldsymbol{U}} P(u), \tag{8}
$$

*where for every $U \in \boldsymbol{U}$, $d_U = \prod_{V \in \boldsymbol{C}(U)} |\Omega_{PA_V} \times \Omega_V|$.*

One attractice property of the specific characterization provided in Props. 2.6 and 2.7, when compared to the most general result given by Prop. 2.5 is that the cardinalities of the exogenous variables, $d_U$, are smaller than that in a general canonical SCM (Eq. (3)). This is due to the fact that observational and interventional distributions are strictly contained in the collection of all counterfactual distributions in a causal diagram. The model complexity of canonical SCMs could thus be reduced and will have implication to the tasks downstream. More generally, the discretization procedure in Thm. 2.4 relies on a generalized canonical partitioning over exogenous domains in an arbitrary SCM. Any counterfactual distribution in this SCM could be written as a function of joint probabilities assigned to intersections of generalized canonical partitions. This allows us to discretize exogenous domains while maintaining all counterfactual distributions and structures of the causal diagram. We refer readers to the complete technical report (Zhang et al., 2021, Appendix A.1) for more details about the Thm. 2.4's proof.

### 2.2. Bounding Counterfactual Distributions

The expressive power of canonical SCMs in Thm. 2.4 suggests a natural algorithm for the partial identification of counterfactual distributions. For a causal diagram $\mathcal{G}$, let $\mathscr{N}(\mathcal{G})$ denote the set of all canonical SCM compatible with $\mathcal{G}$ whose exogenous domain $\Omega_U$ for every $U \in \boldsymbol{U}$ is discrete, bounded by Eq. (3). We derive a bound $[l, r]$ over a counterfactual probability $P\left(\boldsymbol{y}_{\boldsymbol{x}}, \dots, \boldsymbol{z}_{\boldsymbol{w}}\right)$ from an arbitrary collection of interventional distributions $\{P(\boldsymbol{V}_{\boldsymbol{z}}) \mid \boldsymbol{z} \in \mathbb{Z}\}$ by solving the following optimization problem:

$$\begin{aligned} \min_{N \in \mathscr{N}(\mathcal{G})} / \max \quad & P_N\left(\boldsymbol{y}_{\boldsymbol{x}}, \dots, \boldsymbol{z}_{\boldsymbol{w}}\right) \\ \text{s.t.} \quad & P_N(\boldsymbol{V}_{\boldsymbol{z}}) = P(\boldsymbol{V}_{\boldsymbol{z}}) \;\; \forall \boldsymbol{z} \in \mathbb{Z} \end{aligned} \tag{9}$$

where the counterfactual probability $P_N\left(\boldsymbol{y}_{\boldsymbol{x}}, \dots, \boldsymbol{z}_{\boldsymbol{w}}\right)$ and interventional distributions $P_N(\boldsymbol{V}_{\boldsymbol{z}})$ are given in the form of Eq. (5). More generally, the optimization problem in Eq. (9) is reducible to an equivalent polynomial program. To witness, for every exogenous variable $U \in \boldsymbol{U}$, let parameters $\theta_u$ represent discrete probabilities $P(U = u)$. For every endogenous variable $V \in \boldsymbol{V}$, we represent the output of structural function $f_V(pa_V, u_V)$ given input $PA_V = pa_V$ and $U_V = u_V$ using an indicator vector $\mu_V^{(pa_V, u_V)} = \left(\mu_v^{(pa_V, u_V)} \mid \forall v \in \Omega_V\right)$ such that

$$\mu_v^{(pa_V, u_V)} \in \{0, 1\}, \qquad \sum_{v \in \Omega_V} \mu_v^{(pa_V, u_V)} = 1.$$

Doing so allows us to write any counterfactual probability $P\left(\boldsymbol{y}_{\boldsymbol{x}}, \dots, \boldsymbol{z}_{\boldsymbol{w}}\right)$ in Eq. (5) as a polynomial function of parameters $\mu_v^{(pa_V, u_V)}$ and $\theta_u$. More specifically, the indicator function $\mathbb{1}_{\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{u}) = \boldsymbol{y}}$ is equal to a product $\prod_{Y \in \boldsymbol{Y}} \mathbb{1}_{Y_{\boldsymbol{x}}(\boldsymbol{u}) = y}$. For every $Y \in \boldsymbol{Y}$, $\mathbb{1}_{Y_{\boldsymbol{x}}(\boldsymbol{u}) = y}$ is recursively given by:

$$\mathbb{1}_{Y_{\boldsymbol{x}}(\boldsymbol{u}) = y} = \begin{cases} \mathbb{1}_{y = \boldsymbol{x}_Y} & \text{if } Y \in \boldsymbol{X} \\ \displaystyle\sum_{pa_Y} \mu_y^{(pa_Y, u_Y)} \mathbb{1}_{(PA_Y)_{\boldsymbol{x}}(\boldsymbol{u}) = pa_Y} & \text{otherwise} \end{cases}$$

For instance, consider again the causal diagram $\mathcal{G}$ in Fig. 1b. The counterfactual distribution $P(X_{z'}, Y_{x'})$ and the observational distribution $P(X, Y, Z)$ of any discrete SCM in $\mathscr{N}(\mathcal{G})$ and be written as following polynomial functions:

$$P(x_{z'}, y_{x'}) = \sum_{u_1, u_2 = 1}^{d} \mu_x^{(z', u_2)} \mu_Y^{(x', u_1, u_2)} \theta_{u_1} \theta_{u_2}, \tag{10}$$

$$P(x, y, z) = \sum_{u_1, u_2 = 1}^{d} \mu_z^{(u_1)} \mu_x^{(z, u_2)} \mu_y^{(x, u_1, u_2)} \theta_{u_1} \theta_{u_2}, \tag{11}$$

where $\mu_z^{(u_1)}, \mu_x^{(z', u_2)}, \mu_y^{(x', u_1, u_2)}$ are parameters taking values in $\{0, 1\}$; $\theta_{u_i}$, $i = 1, 2$, are probabilities of the discrete distribution $P(u_i)$ over the finite domain $\{1, \dots, d\}$. One could derive a bound over $P(x_{z'}, y_{x'})$ from $P(X, Y, Z)$ by solving polynomial programs which optimize the objective Eq. (10) over parameters $\theta_{u_1}, \theta_{u_2}, \mu_z^{(u_1)}, \mu_x^{(z, u_2)}, \mu_y^{(x, u_1, u_2)}$, subject to the constraints in Eq. (11) for all entries $x, y, z$. (Zhang et al., 2021, Appendix D) includes additional examples demonstrating the reduction of partial counterfactual identification to equivalent polynomial programs.

Note that the collection of all counterfactual distributions subsume both observational and interventional ones. It follows immediately from Thm. 2.4 that the solution $[l, r]$ of the optimization program in Eq. (9) is guaranteed to be a valid, tight bound containing the target counterfactual.

**Theorem 2.8.** *Given a causal diagram $\mathcal{G}$ and interventional distributions $\{P(\boldsymbol{V}_{\boldsymbol{z}}) \mid \boldsymbol{z} \in \mathbb{Z}\}$, the solution $[l, r]$ of the polynomial program Eq. (9) is a tight bound over the counterfactual probability $P\left(\boldsymbol{y}_{\boldsymbol{x}}, \dots, \boldsymbol{z}_{\boldsymbol{w}}\right)$.*

The optimization problem in Eq. (2) is generally reducible to an equivalent polynomial program. Investigating effective polynomial optimization methods is an ongoing subject of research (Lasserre, 2001; Parrilo, 2003). Our focus is on the causal inference aspect of the problem, and like earlier works (Balke & Pearl, 1994; 1997), we do not commit to any particular solvers. For instance, in a quasi-Markovian diagram where every endogenous node is affected by at most one exogenous variable, (Zaffalon et al., 2020) showed causal bounds are obtainable by applying variable elimination in credal networks. This corresponds to a mapping between the bounding problem to multilinear programming (De Campos et al., 1994). In some very specific cases, the bounds are obtainable by solving linear programs (e.g., bounding $P(y_x)$ in the "IV" diagram of Fig. 1a). However, it has been shown in (Zaffalon et al., 2021) that the partial counterfactual identification is generally NP-hard and takes exponentially long in some specific diagrams (e.g., a polytree); let alone the most general case. Therefore, this calls for the need of effective algorithms that approximate optimal bounds over unknown counterfactual probabilities.

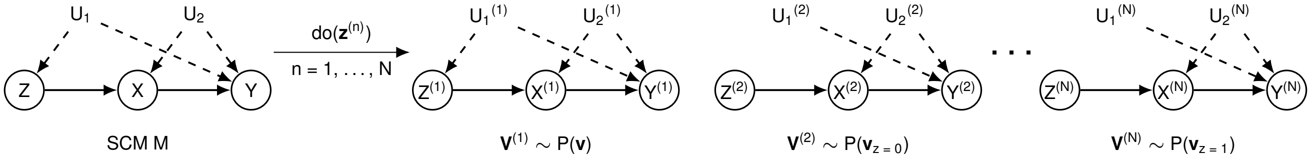


Figure 2: The data-generating process for a finite dateset $\{x^{(n)}, y^{(n)}, z^{(n)}\}_{n=1}^{N}$ in an SCM associated with in Fig. 1b; the set $\mathbb{Z} = \{\emptyset, z = 0, z = 1\}$ where the idle intervention do($\emptyset$) corresponds to the observational distribution.

## 3. Bayesian Approach for Partial Identification

This section describes an algorithm to effectively approximate the optimal counterfactual bound in Eq. (9) from finite samples drawn from interventional distributions $\{P(\boldsymbol{V}_{\boldsymbol{z}}) \mid \boldsymbol{z} \in \mathbb{Z}\}$, provided with prior distributions over parameters $\theta_u$ and $\mu_V^{(pa_V, u_V)}$, possibly uninformative.

More specifically, the learner has access to a finite dataset $\bar{\boldsymbol{v}} = \left\{\boldsymbol{V}^{(n)} = \boldsymbol{v}^{(n)} \mid n = 1, \dots, N\right\}$, where each $\boldsymbol{V}^{(n)}$ is an independent sample drawn from an interventional distribution $P\left(\boldsymbol{V}_{\boldsymbol{z}}\right)$ for some $\boldsymbol{z} \in \mathbb{Z}$. With a slight abuse of notation, we denote by $\boldsymbol{Z}^{(n)}$ the set of variables $\boldsymbol{Z}$ that are intervened for generating the $n$-th sample; therefore, its realization $\boldsymbol{z}^{(n)} = \boldsymbol{z}$. As an example, Fig. 2 shows a graphical representation of the data-generating process for a finite dateset $\{x^{(n)}, y^{(n)}, z^{(n)}\}_{n=1}^{N}$ associated with SCMs in Fig. 1b; the intervention set $\mathbb{Z} = \{\emptyset, z = 0, z = 1\}$.

We first introduce effective Markov Chain Monte Carlo (MCMC) algorithms that sample the posterior distribution $P\left(\theta_{\text{ctf}} \mid \bar{\boldsymbol{v}}\right)$ over an arbitrary counterfactual probability $\theta_{\text{ctf}} = P\left(\boldsymbol{y}_{\boldsymbol{x}}, \dots, \boldsymbol{z}_{\boldsymbol{w}}\right)$. For every $V \in \boldsymbol{V}$, $\forall pa_V, u_V$, endogenous parameters $\mu_V^{(pa_V, u_V)}$ are drawn uniformly over the finite domain $\Omega_V$. For every $U \in \boldsymbol{U}$, exogenous parameters $\theta_u$ are drawn from a Dirichlet distribution, i.e.,

$$\left(\theta_1, \dots, \theta_{d_U}\right) \sim \texttt{Dir}\left(\alpha_U^{(1)}, \dots, \alpha_U^{(d_U)}\right), \tag{12}$$

where the cardinality $d_U = \prod_{V \in \boldsymbol{C}(U)} \left|\Omega_{PA_V} \mapsto \Omega_V\right|$ and hyperparameters $\alpha_1^{(u)}, \dots, \alpha_U^{(d_U)} > 0$.

Gibbs sampling is a well-known MCMC algorithm that allows one to sample posterior distributions. We first introduce the following notations. Let parameters $\boldsymbol{\theta}$ and $\boldsymbol{\mu}$ be:

$$\begin{aligned} \boldsymbol{\theta} &= \left\{\theta_u \mid \forall U \in \boldsymbol{U}, \forall u\right\}, \\ \boldsymbol{\mu} &= \left\{\mu_V^{(pa_V, u_V)} \mid \forall V \in \boldsymbol{V}, \forall pa_V, u_V\right\}. \end{aligned} \tag{13}$$

We denote by $\bar{\boldsymbol{U}} = \left\{\boldsymbol{U}^{(n)} \mid n = 1, \dots, N\right\}$ exogenous variables affecting $N$ endogenous variables $\bar{\boldsymbol{V}} = \left\{\boldsymbol{V}^{(n)} \mid n = 1, \dots, N\right\}$; we use $\bar{\boldsymbol{u}}$ to represent its realization. Our blocked Gibbs sampler works by iteratively drawing values from the conditional distributions of variables as follows (Ishwaran & James, 2001). Detailed derivations of complete conditionals are shown in the technical report (Zhang et al., 2021, Appendix B).

- **Sampling** $P\left(\bar{\boldsymbol{u}} \mid \bar{\boldsymbol{v}}, \boldsymbol{\theta}, \boldsymbol{\mu}\right)$**.** Exogenous variables $\boldsymbol{U}^{(n)}$, $n = 1, \dots, N$, are mutually independent given parameters $\boldsymbol{\theta}, \boldsymbol{\mu}$. We could draw each $\left(\boldsymbol{U}^{(n)} \mid \boldsymbol{\theta}, \boldsymbol{\mu}, \bar{\boldsymbol{V}}\right)$ corresponding to the $n$-th sample induced by do($\boldsymbol{z}^{(n)}$) independently. The complete conditional of $\boldsymbol{U}^{(n)}$ is given by

$$\begin{aligned} &P\left(\boldsymbol{u}^{(n)} \mid \boldsymbol{v}^{(n)}, \boldsymbol{\theta}, \boldsymbol{\mu}\right) \\ &\propto \prod_{V \in \boldsymbol{V} \setminus \boldsymbol{Z}^{(n)}} \mu_{v^{(n)}}^{\left(pa_V^{(n)}, u_V^{(n)}\right)} \prod_{U \in \boldsymbol{U}} \theta_u. \end{aligned} \tag{14}$$

- **Sampling** $P\left(\boldsymbol{\mu}, \boldsymbol{\theta} \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{u}}\right)$**.** Note that parameters $\boldsymbol{\mu}, \boldsymbol{\theta}$ are mutually independent given $\bar{\boldsymbol{V}}, \bar{\boldsymbol{U}}$. Therefore, we will derive complete conditionals over $\boldsymbol{\mu}, \boldsymbol{\theta}$ separately.

  Consider first endogenous parameters $\boldsymbol{\mu}$. For every $V \in \boldsymbol{V}$, fix $pa_V, u_V$. If there exists an instance $n = 1, \dots, N$ such that $V \not\in \boldsymbol{Z}^{(n)}$ and $pa_V^{(n)} = pa_V, u_V^{(n)} = u_V$, the posterior over $\mu_V^{(pa_V, u_V)}$ is given by, for $\forall v \in \Omega_V$,

$$P\left(\mu_v^{(pa_V, u_V)} = 1 \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{u}}\right) = \mathbb{1}_{v = v^{(n)}}. \tag{15}$$

  Otherwise, $\mu_V^{(pa_V, u_V)}$ is drawn uniformly from $\Omega_V$.

  Consider now exogenous parameters $\boldsymbol{\theta}$. For every $U \in \boldsymbol{U}$, fix $u$. Let $n_u = \sum_{n=1}^{N} \mathbb{1}_{u^{(n)} = u}$ be the number of instances in $u^{(n)}$ equal to $u$. By the conjugacy of the Dirichlet distribution, the complete conditional of $\theta_u$ is,

$$\begin{aligned} &\left(\theta_1, \dots, \theta_{d_U}\right) \sim \texttt{Dir}\left(\beta_U^{(1)}, \dots, \beta_U^{(d_U)}\right), \\ &\text{where } \ \beta_U^{(u)} = \alpha_U^{(u)} + n_u \ \text{ for } u = 1, \dots, d_U. \end{aligned} \tag{16}$$

Doing so eventually produces values drawn from the posterior distribution over $\left(\boldsymbol{\theta}, \boldsymbol{\mu}, \bar{\boldsymbol{U}} \mid \bar{\boldsymbol{V}}\right)$. Given parameters $\boldsymbol{\theta}, \boldsymbol{\mu}$, we compute the counterfactual probability $\theta_{\text{ctf}} = P(\boldsymbol{y}_{\boldsymbol{x}}, \dots, \boldsymbol{z}_{\boldsymbol{w}})$ following the three-step algorithm in (Pearl, 2000) which consists of abduction, action, and prediction. Thus computing $\theta_{\text{ctf}}$ from each draw $\boldsymbol{\theta}, \boldsymbol{\mu}, \bar{\boldsymbol{U}}$ eventually gives us the draw from the posterior distribution $P\left(\theta_{\text{ctf}} \mid \bar{\boldsymbol{v}}\right)$.

### 3.1. Collapsed Gibbs Sampling

We describe next an alternative MCMC algorithm that applies to Dirichlet priors in Eq. (12),and which will be advantageous in some other settings. For $n = 1, \dots, N$, let $\bar{\boldsymbol{U}}_{-n}$ denote the set difference $\bar{\boldsymbol{U}} \setminus \boldsymbol{U}^{(n)}$; similarly, we write $\bar{\boldsymbol{V}}_{-n} = \bar{\boldsymbol{V}} \setminus \boldsymbol{V}^{(n)}$. Our collapsed Gibbs sampler first iteratively draws values from the conditional distribution over $\left(\boldsymbol{U}^{(n)} \mid \bar{\boldsymbol{V}}, \bar{\boldsymbol{U}}_{-n}\right)$ for every $n = 1, \dots, N$ as follows.

- **Sampling $P\left(\boldsymbol{u}^{(n)} \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{u}}_{-n}\right)$.** At each iteration, draw $\boldsymbol{U}^{(n)}$ from the conditional distribution given by

$$\begin{aligned} P\left(\boldsymbol{u}^{(n)} \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{u}}_{-n}\right) & \\ \propto \prod_{V \in \boldsymbol{V} \setminus \boldsymbol{Z}^{(n)}} & P\left(v^{(n)} \mid pa_V^{(n)}, u_V^{(n)}, \bar{\boldsymbol{v}}_{-n}, \bar{\boldsymbol{u}}_{-n}\right) \\ & \prod_{U \in \boldsymbol{U}} P\left(u^{(n)} \mid \bar{\boldsymbol{v}}_{-n}, \bar{\boldsymbol{u}}_{-n}\right). \end{aligned} \tag{17}$$

Among quantities in the above equation, for every $V \in \boldsymbol{V} \setminus \boldsymbol{Z}^{(n)}$, if there exists an instance $i \neq n$ such that $V \notin \boldsymbol{Z}^{(i)}$ and $pa_V^{(i)} = pa_V^{(n)}$, $u_V^{(i)} = u_V^{(n)}$,

$$P\left(v^{(n)} \mid pa_V^{(n)}, u_V^{(n)}, \bar{\boldsymbol{v}}_{-n}, \bar{\boldsymbol{u}}_{-n}\right) = \mathbb{1}_{v^{(n)} = v^{(i)}}. \tag{18}$$

Otherwise, the above probability is equal to $1/|\Omega_V|$.

For every $U \in \boldsymbol{U}$, let $\bar{u}_{-n}$ be a set of exogenous samples $\left\{u^{(1)}, \dots, u^{(N)}\right\} \setminus \{u^{(n)}\}$. Let $\{u_1^*, \dots, u_K^*\}$ denote $K$ unique values that samples in $\bar{u}_{-n}$ take on. The conditional distribution over $\left(U^{(n)} \mid \bar{\boldsymbol{V}}_{-n}, \bar{\boldsymbol{U}}_{-n}\right)$ is given by, for hyperparameters $\alpha_U = \sum_{u=1}^{d_U} \alpha_U^{(u)}$,

$$\begin{aligned} & P\left(u^{(n)} \mid \bar{\boldsymbol{v}}_{-n}, \bar{\boldsymbol{u}}_{-n}\right) \\ &= \begin{cases} \dfrac{n_k^* + \alpha_U^{(u_k^*)}}{\alpha_U + N - 1} & \text{if } u^{(n)} = u_k^* \\ \dfrac{\alpha_U^{(u^{(n)})}}{\alpha_U + N - 1} & \text{if } u^{(n)} \notin \{u_1^*, \dots, u_K^*\} \end{cases} \end{aligned} \tag{19}$$

where $n_k^* = \sum_{i \neq n} \mathbb{1}_{u^{(i)} = u_k^*}$, for $k = 1, \dots, K$, records the number of values $u^{(i)} \in \bar{u}_{-n}$ that are equal to $u_k^*$.

Doing so eventually produces exogenous variables drawn from the posterior distribution of $\left(\bar{\boldsymbol{U}} \mid \bar{\boldsymbol{V}}\right)$. We then sample parameters from the posterior distribution of $\left(\boldsymbol{\theta}, \boldsymbol{\mu} \mid \bar{\boldsymbol{U}}, \bar{\boldsymbol{V}}\right)$; complete conditional distributions $P\left(\boldsymbol{\mu}, \boldsymbol{\theta} \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{u}}\right)$ are given in Eqs. (15) and (16). Finally, computing $\theta_{\text{ctf}}$ from each sample $\boldsymbol{\theta}, \boldsymbol{\mu}$ gives a draw from the posterior $P\left(\theta_{\text{ctf}} \mid \bar{\boldsymbol{v}}\right)$.

When the cardinality $d_U$ of exogenous domains is high, the collapsed Gibbs sampler described here is more computational efficient than the blocked sampler, since it does not iteratively draw parameters $\boldsymbol{\theta}, \boldsymbol{\mu}$ in the high-dimensional space. Instead, the collapsed sampler only draws $\boldsymbol{\theta}, \boldsymbol{\mu}$ once after samples drawn from the distribution of $\left(\bar{\boldsymbol{U}} \mid \bar{\boldsymbol{V}}\right)$ converge. On the other hand, when the cardinality $d_U$ is reasonably low, the blocked Gibbs sampler is preferable since it exhibits better convergence (Ishwaran & James, 2001).

### 3.2. Credible Intervals over Counterfactuals

Given a MCMC sampler, one could compute credible intervals over the unknown counterfactual probability $\theta_{\text{ctf}}$ from the posterior distribution $P\left(\theta_{\text{ctf}} \mid \bar{\boldsymbol{v}}\right)$.

**Definition 3.1.** Fix $\alpha \in [0, 1]$. A $100(1-\alpha)\%$ credible interval $[l_\alpha, r_\alpha]$ for $\theta_{\text{ctf}}$ is given by

$$\begin{aligned} l_\alpha &= \sup\left\{x \mid P\left(\theta_{\text{ctf}} \leq x \mid \bar{\boldsymbol{v}}\right) = \alpha/2\right\}, \\ r_\alpha &= \inf\left\{x \mid P\left(\theta_{\text{ctf}} \leq x \mid \bar{\boldsymbol{v}}\right) = 1 - \alpha/2\right\}. \end{aligned} \tag{20}$$

For a $100(1-\alpha)\%$ credible interval $[l_\alpha, r_\alpha]$, any counterfactual probability $\theta_{\text{ctf}}$ that is compatible with observational data $\bar{\boldsymbol{v}}$ lies between the interval $l_\alpha$ and $r_\alpha$ with probability $1-\alpha$. For consistency, we also define $l_\alpha \triangleq l_0$ and $r_\alpha \triangleq r_1$ if $\alpha < 0$. Credible intervals have been widely applied in the literature for computing bounds over unknown counterfactual probabilities provided with finite observational data (Imbens & Manski, 2004; Vansteelandt et al., 2006; Romano & Shaikh, 2008; Bugni, 2010; Todem et al., 2010).

Formally, let $\rho\left(\boldsymbol{\theta}\right)$ and $\rho\left(\boldsymbol{\mu}\right)$ be probability density functions for prior distributions over to model parameters $\boldsymbol{\theta}$ and $\boldsymbol{\mu}$. We say priors over $\boldsymbol{\theta}$ and $\boldsymbol{\mu}$ have *full support* if density functions $\rho\left(\boldsymbol{\theta}\right) > 0$ and $\rho\left(\boldsymbol{\mu}\right) > 0$ for every possible realization of $\boldsymbol{\theta}, \boldsymbol{\mu}$. For any $\boldsymbol{z} \in \mathbb{Z}$, let $N_{\boldsymbol{z}}$ denote the number of samples in $\bar{\boldsymbol{v}}$ drawn from $P\left(\boldsymbol{V}_{\boldsymbol{z}}\right)$; therefore, $\sum_{\boldsymbol{z} \in \mathbb{Z}} N_{\boldsymbol{z}} = N$. Our next result shows that credible intervals from the posterior distribution effectively approximate the optimal counterfactual bounds in Eq. (2) with increasing accuracy as more observational data is obtained.

**Theorem 3.2.** *Given a causal diagram $\mathcal{G}$ and finite samples $\bar{\boldsymbol{v}} = \left\{\boldsymbol{v}^{(n)}\right\}_{n=1}^N$, let $[l_0, r_0]$ be the $100\%$ credible interval for $\theta_{\mathit{ctf}} = P\left(\boldsymbol{y}_{\boldsymbol{x}}, \dots, \boldsymbol{z}_{\boldsymbol{w}}\right)$, and let $[l, r]$ be the optimal bound over $P\left(\boldsymbol{y}_{\boldsymbol{x}}, \dots, \boldsymbol{z}_{\boldsymbol{w}}\right)$ given by Eq. (9). If priors over $\boldsymbol{\theta}$, $\boldsymbol{\mu}$ have full support,*

1. *The credible interval $[l_0, r_0]$ contains the optimal counterfactual bound $[l, r]$, i.e., $[l, r] \subseteq [l_0, r_0]$.*
2. *The credible interval $[l_0, r_0]$ converges almost surely to the tight bound $[l, r]$ as more samples $N_{\boldsymbol{z}}$ are obtained, i.e., $[l_0, r_0] \xrightarrow{a.s.} [l, r]$ when $N_{\boldsymbol{z}} \to \infty$ for every $\boldsymbol{z} \in \mathbb{Z}$.*

Let $\left\{\theta^{(t)}\right\}_{t=1}^T$ be $T$ samples drawn from $P\left(\theta_{\text{ctf}} \mid \bar{\boldsymbol{v}}\right)$. One could compute the $100(1-\alpha)\%$ credible interval for $\theta_{\text{ctf}}$ using following estimators (Sen & Singer, 1994):

$$\hat{l}_\alpha(T) = \theta^{(\lfloor(\alpha/2)T\rfloor + 1)}, \quad \hat{r}_\alpha(T) = \theta^{(\lceil(1-\alpha/2)T\rceil)}, \tag{21}$$

**Algorithm 1** CREDIBLEINTERVAL

```
**Input:** Credible level $\alpha$, tolerance level $\delta, \epsilon$.
**Output:** An credible interval $[l_\alpha, r_\alpha]$ for $\theta_{\text{ctf}}$.
Draw $T = \lceil 2\epsilon^{-2} \ln(4/\delta) \rceil$ samples $\left\{\theta^{(1)}, \dots, \theta^{(T)}\right\}$ from the posterior distribution $P\left(\theta_{\text{ctf}} \mid \bar{\boldsymbol{v}}\right)$.
Return interval $\left[\hat{l}_\alpha(T), \hat{r}_\alpha(T)\right]$ (Eq. (21)).
```

where estimates $\theta^{(\lfloor(\alpha/2)T\rfloor+1)}$ and $\theta^{(\lceil(1-\alpha/2)T\rceil)}$ are the $(\lfloor(\alpha/2)T\rfloor+1)$th smallest and the $\lceil(1-\alpha/2)T\rceil$th smallest samples of $\left\{\theta^{(t)}\right\}$[4]. Our next results establish non-asymptotic deviation bounds for empirical estimates of credible intervals defined in Eq. (21). This allows us to determine the sufficient number of draws $T$ that is required for approximating a $100(1-\alpha)\%$ credible interval.

**Lemma 3.3.** *Fix $T > 0$ and $\delta \in (0, 1)$. Let function $f(T, \delta) = \sqrt{2T^{-1} \ln(4/\delta)}$. With probability at least $1 - \delta$, estimators $\hat{l}_\alpha(T), \hat{r}_\alpha(T)$ for any $\alpha \in [0, 1)$ is bounded by*

$$
\begin{aligned}
l_{\alpha - f(T,\delta)} \leq \hat{l}_\alpha(T) \leq l_{\alpha + f(T,\delta)}, \\
r_{\alpha + f(T,\delta)} \leq \hat{r}_\alpha(T) \leq r_{\alpha - f(T,\delta)}.
\end{aligned}
\tag{22}
$$

We summarize our algorithm, CREDIBLEINTERVAL, in Alg. 1. It takes a credible level $\alpha$ and tolerance levels $\delta, \epsilon$ as inputs. In particular, CREDIBLEINTERVAL repeatedly draw $T \geq \lceil 2\epsilon^{-2} \ln(4/\delta) \rceil$ samples from $P\left(\theta_{\text{ctf}} \mid \bar{\boldsymbol{v}}\right)$. It then computes estimates $\hat{l}_\alpha(T), \hat{h}_\alpha(T)$ from drawn samples following Eq. (21) and return them as the output. It follows immediately from Lem. 3.3 that such a procedure efficiently approximates a $100(1-\alpha)\%$ credible interval.

**Corollary 3.4.** *Fix $\delta \in (0, 1)$ and $\epsilon > 0$. With probability at least $1 - \delta$, the interval $[\hat{l}, \hat{r}] =$ CREDIBLEINTERVAL$(\alpha, \delta, \epsilon)$ for any $\alpha \in [0, 1)$ is bounded by $\hat{l} \in [l_{\alpha-\epsilon}, l_{\alpha+\epsilon}]$ and $\hat{r} \in [r_{\alpha+\epsilon}, r_{\alpha-\epsilon}]$.*

Corol. 3.4 implies that any counterfactual probability $\theta_{\text{ctf}}$ compatible with the dataset $\bar{\boldsymbol{v}}$ falls between $[\hat{l}, \hat{r}] =$ CREDIBLEINTERVAL$(\alpha, \delta, \epsilon)$ with $P\left(\theta_{\text{ctf}} \in [\hat{l}, \hat{r}] \mid \bar{\boldsymbol{v}}\right) \approx 1 - \alpha \pm \epsilon$. As the tolerance rate $\epsilon \to 0$, $[\hat{l}, \hat{r}]$ converges to a $100(1-\alpha)\%$ credible interval with high probability.

## 4. Simulations and Experiments

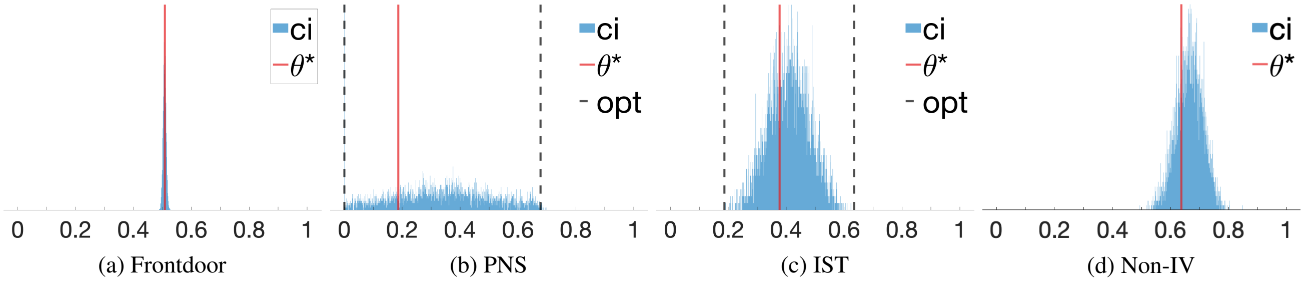
We demonstrate our algorithms on various synthetic and real datasets in different causal diagrams. Overall, we found that simulation results support our findings and the proposed bounding strategy consistently dominates state-of-art algorithms. When target probabilities are identifiable (Experiment 1), our bounds collapse to the true counterfactual probabilities. For non-identifiable settings, our algorithm obtains sharp asymptotic bounds when the closed-form solutions already exist (Experiments 2 & 3); and obtains novel bounds in other more general cases that consistently improve over existing strategies (Experiment 4).

In all experiments, we evaluate our proposed strategy using credible intervals (*ci*). We draw at least $4 \times 10^3$ samples from the posterior distribution $P\left(\theta_{\text{ctf}} \mid \bar{\boldsymbol{v}}\right)$ over the target counterfactual. This allows us to compute $100\%$ credible interval over $\theta_{\text{ctf}}$ within error $\epsilon = 0.05$, with probability at least $1 - \delta = 0.95$. As the baseline, we include the true counterfactual probability $\theta^*$. We refer readers to the complete technical report (Zhang et al., 2021, Appendix C) for more details on the simulation setup and additional experiments on other causal diagrams and datasets.

**Experiment 1: Frontdoor Graph.** In this experiment, we evaluate our algorithm on interventional probabilities that are identifiable from the observational data. In this case, the bounds over the target probability should collapse to a point estimate. Consider the "Frontdoor" graph described in Fig. 1c where $X, Y, W$ are binary variables in $\{0, 1\}$; $U_1, U_2 \in \mathbb{R}$. In this case, any interventional probability $P(y_x)$ is identifiable from the observational distribution $P(X, W, Y)$ through the frontdoor adjustment (Pearl, 2000, Thm. 3.3.4). We collect $N = 10^4$ observational samples $\bar{\boldsymbol{v}} = \{x^{(n)}, y^{(n)}, w^{(n)}\}_{n=1}^N$ from a synthetic SCM instance. Fig. 3a shows samples drawn from the posterior distribution $(P(Y_{x=0} = 1) \mid \bar{\boldsymbol{v}})$. The analysis reveals that these samples collapse to the actual probability $P(Y_{x=0} = 1) = 0.5085$, which confirms the identifiability of $P(y_x)$ in the "frontdoor" graph. This result shows that our sampler is able to draw values from the posterior of identifiable probabilities.

**Experiment 2: Probability of Necessity and Sufficiency.** In this experiment, we compare credible intervals obtained by our algorithm with sharp bounds over unknown counterfactual probabilities derived from the observational data. Consider the "Bow" diagram in Fig. 1d where $X, Y \in \{0, 1\}$ and $U \in \mathbb{R}$. We study the problem of evaluating the *probability of necessity and sufficiency* (for short, PNS) $P(Y_{x=1} = 1, Y_{x=0} = 0)$ from the observational distribution $P(X, Y)$. The non-identifiability of PNS with the unobserved confounding between $X$ and $Y$ has been acknowledged in (Avin et al., 2005). Tian & Pearl (2000) introduced the sharp bound for $P(Y_{x=1} = 1, Y_{x=0} = 0)$ from $P(X, Y)$, labelled as *opt*. We collect $N = 10^3$ observational samples $\bar{\boldsymbol{v}} = \{x^{(n)}, y^{(n)}\}_{n=1}^N$ from a randomly generated SCM instance. Fig. 3c shows samples drawn from the posterior distribution over $(P(Y_{x=1} = 1, Y_{x=0} = 0) \mid \bar{\boldsymbol{v}})$. The analysis reveals that the $100\%$ credible interval (*ci*) matches the optimal PNS bound $l = 0, r = 0.6775$ over the actual PNS probability $P(Y_{x=1} = 1, Y_{x=0} = 0) = 0.1867$,

[4] For any $\alpha \in \mathbb{R}$, let $\lceil \alpha \rceil = \min\{n \in \mathbb{Z} \mid n \geq \alpha\}$ denote the smallest integer $n \in \mathbb{Z}$ larger than $\alpha$. Similarly, $\lfloor \alpha \rfloor = \max\{n \in \mathbb{Z} \mid n \leq \alpha\}$ is the largest integer $n \in \mathbb{Z}$ smaller than $\alpha$.



(a) Frontdoor (b) PNS (c) IST (d) Non-IV

Figure 3: Simulation results for Experiments 1-4 showing posterior samples of target counterfactuals. For all plots (a - d), *ci* represents our proposed algorithm; $\theta^*$ is the actual counterfactual probability; and *opt* is the optimal asymptotic bounds.

which confirms the efficacy of the proposed approach.

**Experiment 3: International Stroke Trials (IST).** In this experiment, we evaluate our algorithm on a real-life dataset and show that it could consistently obtain sharp bounds over unknown counterfactual probabilities. International stroke trials was a large, randomized, open trial of up to 14 days of antithrombotic therapy after stroke onset (Carolei et al., 1997). The aim of the trial was to provide reliable evidence on the efficacy of aspirin and of heparin. In particular, the treatment $X$ is a pair $(i, j)$ where $i \in \{0, 1\}$ stands for aspirin allocation; $j \in \{0, 1, 2\}$ stands for heparin allocation. The primary outcome $Y \in \{0, \dots, 3\}$ is the health of the patient 6 months after the treatment, where 0 stands for death, 1 for being dependent on the family, 2 for the partial recovery, and 3 for the full recovery.

To emulate the presence of unobserved confounding, we filter the experimental data following a procedure in (Kallus & Zhou, 2018). Doing so allows us to obtain $N = 10^3$ synthetic observational samples $\bar{\boldsymbol{v}} = \{x^{(n)}, y^{(n)}\}_{n=1}^N$ that are compatible with the "Bow" diagram of Fig. 1d. We are interested in evaluating the probability $P\left(Y_{x=(1,0)} \geq 2\right)$, i.e., the treatment effect of only assigning aspirin $X = (1, 0)$ for the recovery of patients $Y \geq 2$. As a baseline, we also include the optimal bound for $P(y_x)$ from $P(X, Y)$ (Manski, 1990), labeled as *opt*, which coincides with the solution of the credal network solver (Zaffalon et al., 2020). Simulation results, shown in Fig. 3c, reveal that both algorithms achieve effective bounds containing target interventional probability $P\left(Y_{x=(1,0)} \geq 2\right) = 0.3775$. The $100\%$ credible interval is $l_{ci} = 0.1905, r_{ci} = 0.6239$, which matches the optimal bounding strategy ($l_{opt} = 0.1861, r_{opt} = 0.6343$).

**Experiment 4: Non-IV** This experiment evaluates our algorithm in a novel partial identification setting where the closed-form bounding solution does not exist. Our proposed approach is able to obtain a valid bound over the unknown counterfactual probability. Consider the "Non-IV' diagram in Fig. 1b where $X, Y, Z \in \{0, \dots, 9\}$ and $U_1, U_2 \in \mathbb{R}$. We are interested in evaluating counterfactual probabilities $P\left(z, x_{z'}, y_{x'}\right)$ from the observational distribution $P(X, Y, Z)$ and interventional distributions $P(X_z, Y_z)$ induced by interventions $\text{do}(Z = z)$ for $z = 0, \dots, 9$. We collect $N = 10^3$ samples $\bar{\boldsymbol{v}} = \{x^{(n)}, y^{(n)}, z^{(n)}\}_{n=1}^N$ from a SCM instance of Fig. 1b where each sample $X^{(n)}, Y^{(n)}, Z^{(n)}$ is an independent draw from $P(X, Y, Z)$ or $P(X_z, Y_z)$. To address the challenge of the high-dimensional exogenous domains, we apply the proposed collapsed Gibbs sampler to obtain samples from the posterior distribution $\left(P\left(Z + X_{z=0} + Y_{x=0} \geq 14\right) \mid \bar{\boldsymbol{v}}\right)$. Simulation results, shown in Fig. 3d, reveal that our proposed approach is able to achieve an effective bound that contains the actual counterfactual probability $P\left(Z + X_{z=0} + Y_{x=0} \geq 14\right) = 0.6378$. The $100\%$ credible interval (*ci*) is equal to $l = 0.4949, r = 0.8482$, which is a valid bound containing the target countrefactual. To our best knowledge, no existing bounding strategy is applicable for this setting.

## 5. Conclusion

This paper investigated the problem of partial identification of counterfactual distributions, which concerns with bounding counterfactual probabilities from an arbitrary combination of observational and experimental data, provided with a causal diagram encoding qualitative assumptions about the data-generating process. We introduced a special parametric family of SCMs with discrete exogenous variables, taking values from a finite set of unobserved states, and showed that it could represent *all* counterfactual distributions (over finite observed variables) in *any* causal diagram. Using this result, we reduced the partial identification problem into a polynomial program and developed novel algorithms to approximate the optimal asymptotic bounds over target counterfactual probabilities from finite samples obtained through arbitrary observations and experiments.

## Acknowledgements

The team is supported in part by funding from the Amazon, JP Morgan, and The Alfred P. Sloan Foundation, as well as grants from the NSF, DoE, ONR, NIH.

## References

Avin, C., Shpitser, I., and Pearl, J. Identifiability of path-specific effects. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence IJCAI-05*, pp. 357–363, Edinburgh, UK, 2005. Morgan-Kaufmann Publishers.

Balke, A. and Pearl, J. Counterfactual probabilities: Computational methods, bounds, and applications. In de Mantaras, R. L. and Poole, D. (eds.), *Uncertainty in Artificial Intelligence 10*, pp. 46–54. Morgan Kaufmann, San Mateo, CA, 1994.

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 by surrogate experiments: $z$-identifiability. In de Freitas, N. and Murphy, K. (eds.), *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pp. 113–120, Corvallis, OR, 2012. AUAI Press.

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

Bareinboim, E., Correa, J., Ibeling, D., and Icard, T. On pearl's hierarchy and the foundations of causal inference. *ACM Special Volume in Honor of Judea Pearl*, 2020. forthcoming. Also, Technical Report R-60, Causal AI Lab, Columbia University, https://causalai.net/r60.pdf.

Bugni, F. A. Bootstrap inference in partially identified models defined by moment inequalities: Coverage of the identified set. *Econometrica*, 78(2):735–753, 2010.

Carathéodory, C. Über den variabilitätsbereich der fourier'schen konstanten von positiven harmonischen funktionen. *Rendiconti Del Circolo Matematico di Palermo (1884-1940)*, 32(1):193–217, 1911.

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 tool for analyzing non-compliance. *Computing Science and Statistics*, 29 (2):424–431, 1997.

Correa, J., Lee, S., and Bareinboim, E. Nested counterfactual identification from arbitrary surrogate experiments. In *In Advances in Neural Information Processing Systems*, 2021.

De Campos, L. M., Huete, J. F., and Moral, S. Probability intervals: a tool for uncertain reasoning. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2(02):167–196, 1994.

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.

Evans, R. J. et al. Margins of discrete bayesian networks. *The Annals of Statistics*, 46(6A):2623–2656, 2018.

Finkelstein, N. and Shpitser, I. Deriving bounds and inequality constraints using logical relations among counterfactuals. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1348–1357. PMLR, 2020.

Frangakis, C. and Rubin, D. Principal stratification in causal inference. *Biometrics*, 1(58):21–29, 2002.

Halpern, J. Axiomatizing causal reasoning. In Cooper, G. and Moral, S. (eds.), *Uncertainty in Artificial Intelligence*, pp. 202–210. Morgan Kaufmann, San Francisco, CA, 1998. Also, *Journal of Artificial Intelligence Research* 12:3, 17–37, 2000.

Imbens, G. W. and Manski, C. F. Confidence intervals for partially identified parameters. *Econometrica*, 72(6): 1845–1857, 2004.

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.

Ishwaran, H. and James, L. F. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

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

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.

Lasserre, J. B. Global optimization with polynomials and the problem of moments. *SIAM Journal on optimization*, 11(3):796–817, 2001.

Lee, S., Correa, J., and Bareinboim, E. General identifiability with arbitrary surrogate experiments. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*, Tel Aviv, Israel, 2019. AUAI Press.

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

Parrilo, P. A. Semidefinite programming relaxations for semialgebraic problems. *Mathematical programming*, 96 (2):293–320, 2003.

Pearl, J. Causal diagrams for empirical research. *Biometrika*, 82(4):669–710, 1995.

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

Pearl, J. Principal stratification – a goal or a tool? *The International Journal of Biostatistics*, 7(1), 2011.

Pearl, J. and Mackenzie, D. *The Book of Why*. Basic Books, New York, 2018.

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.

Richardson, T. S., Evans, R. J., and Robins, J. M. Transparent parameterizations of models for potential outcomes. *Bayesian Statistics*, 9:569–610, 2011.

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, 1989.

Romano, J. P. and Shaikh, A. M. Inference for identifiable parameters in partially identified econometric models. *Journal of Statistical Planning and Inference*, 138(9): 2786–2807, 2008.

Rosset, D., Gisin, N., and Wolfe, E. Universal bound on the cardinality of local hidden variables in networks. *Quantum Information & Computation*, 18:910–926, 2018.

Sachs, M. C., Jonzon, G., Sjölander, A., and Gabriel, E. E. A general method for deriving tight symbolic bounds on causal effects. *arXiv preprint arXiv:2003.10702*, 2020.

Sen, P. K. and Singer, J. M. *Large sample methods in statistics: an introduction with applications*, volume 25. CRC press, 1994.

Shpitser, I. and Pearl, J. What counterfactuals can be tested. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, pp. 352–359. AUAI Press, Vancouver, BC, Canada, 2007. Also, *Journal of Machine Learning Research*, 9:1941–1979, 2008.

Shpitser, I. and Sherman, E. Identification of personalized effects associated with causal pathways. In *UAI*, 2018.

Tian, J. and Pearl, J. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28:287–313, 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.

Todem, D., Fine, J., and Peng, L. A global sensitivity test for evaluating statistical hypotheses with nonidentifiable models. *Biometrics*, 66(2):558–566, 2010.

Vansteelandt, S., Goetghebeur, E., Kenward, M. G., and Molenberghs, G. Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statistica Sinica*, pp. 953–979, 2006.

Zaffalon, M., Antonucci, A., and Cabañas, R. Structural causal models are (solvable by) credal networks. In *International Conference on Probabilistic Graphical Models*, pp. 581–592. PMLR, 2020.

Zaffalon, M., Antonucci, A., and Cabañas, R. Causal expectation-maximisation. In *WHY-21 Workshop*, 2021.

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

Zhang, J., Tian, J., and Bareinboim, E. Partial counterfactual identification from observational and experimental data. Technical Report R-78, Causal Artificial Intelligence Lab, Columbia University, Jun 2021.