TECHNICAL REPORT
R-78
First Version: Jun, 2021
Last Version: Dec, 2022

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>.

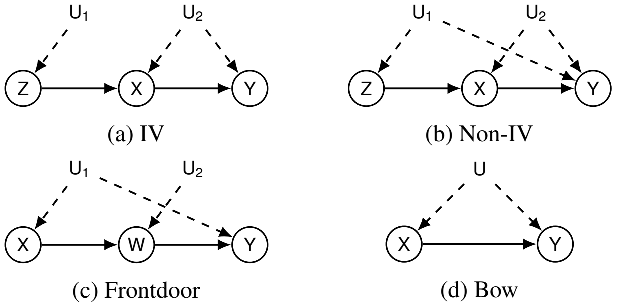
Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

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 Appendices A and B.



(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 Ω_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 $\text{do}(\boldsymbol{x})$, 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 $\text{do}(\boldsymbol{x})$. 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 Ω_V.
2. For every exogenous $U \in \boldsymbol{U}$, its values u are drawn from a finite domain Ω_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 in the c-component $\boldsymbol{C}(U)$ covering U.

[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 Ω_{PA_V} to Ω_V.

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 $\boldsymbol{Y}_{\boldsymbol{x}}, \dots, \boldsymbol{Z}_{\boldsymbol{w}}$, $P_M\left(\boldsymbol{Y}_{\boldsymbol{x}}, \dots, \boldsymbol{Z}_{\boldsymbol{w}}\right) = P_N\left(\boldsymbol{Y}_{\boldsymbol{x}}, \dots, \boldsymbol{Z}_{\boldsymbol{w}}\right)$.*

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 $P\left(\boldsymbol{Y}_{\boldsymbol{x}}, \dots, \boldsymbol{Z}_{\boldsymbol{w}}\right)$ 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 $\boldsymbol{Y}_{\boldsymbol{x}}, \dots, \boldsymbol{Z}_{\boldsymbol{w}}$ be an arbitrary set of counterfactual variables. The distribution $P\left(\boldsymbol{Y}_{\boldsymbol{x}}, \dots, \boldsymbol{Z}_{\boldsymbol{w}}\right)$ decomposes as*

$$
\begin{aligned}
&P\left(\boldsymbol{y}_{\boldsymbol{x}}, \dots, \boldsymbol{z}_{\boldsymbol{w}}\right) \\
&= \sum_{U \in \boldsymbol{U}} \sum_{u=1}^{d_U} \mathbb{1}_{\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{u}) = \boldsymbol{y}, \dots, \boldsymbol{Z}_{\boldsymbol{w}}(\boldsymbol{u}) = \boldsymbol{z}} \prod_{U \in \boldsymbol{U}} P(u),
\end{aligned} \tag{5}
$$

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 $\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{u}) = \{Y_{\boldsymbol{x}}(\boldsymbol{u}) \mid \forall Y \in \boldsymbol{Y}\}$ 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 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 $\boldsymbol{X}, \boldsymbol{Y} \subseteq \boldsymbol{V}$, $P(\boldsymbol{Y}_{\boldsymbol{x}})$ decomposes as follows:*

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

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 Appendix A for details about 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 Ω_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 θ_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 θ_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:

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

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 $\{0, 1\}$; θ_{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. 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.

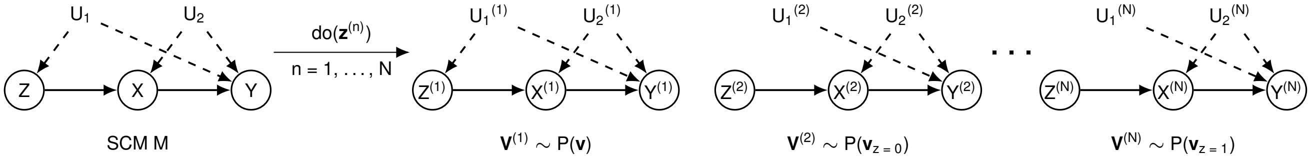


Figure 2: The data-generating process for a finite dateset $\{x^{(n)}, y^{(n)}, z^{(n)}\}_{n=1}^N$ in an SCM associated with in Fig. 1b; the set $\mathbb{Z} = \{\emptyset, z = 0, z = 1\}$ where the idle intervention do($\emptyset$) corresponds to the observational distribution.

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 $\{P(\boldsymbol{V_z}) \mid \boldsymbol{z} \in \mathbb{Z}\}$, provided with prior distributions over parameters θ_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 $P(\boldsymbol{V_z})$ 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 $P(\theta_{\text{ctf}} \mid \bar{\boldsymbol{v}})$ over an arbitrary counterfactual probability $\theta_{\text{ctf}} = P(\boldsymbol{y_x}, \dots, \boldsymbol{z_w})$. 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 Ω_V. For every $U \in \boldsymbol{U}$, exogenous parameters θ_u are drawn from a Dirichlet distribution, i.e.,

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

where the cardinality $d_U = \prod_{V \in \boldsymbol{C}(U)} |\Omega_{PA_V} \mapsto \Omega_V|$ 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:

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

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 Appendix B.1.

- **Sampling** $P(\bar{\boldsymbol{u}} \mid \bar{\boldsymbol{v}}, \boldsymbol{\theta}, \boldsymbol{\mu})$**.** 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** $P(\boldsymbol{\mu}, \boldsymbol{\theta} \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{u}})$**.** 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 $V \notin \boldsymbol{Z}^{(n)}$ 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 Ω_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 θ_u is,

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

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 $\theta_{\text{ctf}} = P(\boldsymbol{y_x}, \dots, \boldsymbol{z_w})$ following the three-step algorithm in (Pearl, 2000) which consists of abduction, action, and prediction. Thus computing θ_{ctf} from each draw $\boldsymbol{\theta}, \boldsymbol{\mu}, \bar{\boldsymbol{U}}$ eventually gives us the draw from the posterior distribution $P(\theta_{\text{ctf}} \mid \bar{\boldsymbol{v}})$.

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

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

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 $\left\{u^{(1)}, \dots, u^{(N)}\right\} \setminus \left\{u^{(n)}\right\}$. 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 θ_{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 θ_{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 θ_{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 θ_{ctf} that is compatible with observational data $\bar{\boldsymbol{v}}$ lies between the interval l_α and r_α with probability $1-\alpha$. The 100% credible interval $[l_0, r_0]$ is the smallest closed set (i.e., the closure) containing the union of all credible intervals $[l_\alpha, r_\alpha]$, $\forall \alpha \in (0, 1]$. For consistency, we also define $l_\alpha \triangleq l_0$ and $r_\alpha \triangleq r_0$ if $\alpha < 0$. Credible intervals have been applied in the literature for computing bounds over partially identifiable parameters provided with finite observational data, including in artificial intelligence (Chickering & Pearl, 1997; Richardson et al., 2011) and in econometrics (Imbens & Rubin, 1997; Poirier, 1998; Imbens & Manski, 2004; Vansteelandt et al., 2006; Romano & Shaikh, 2008; Stoye, 2009; Bugni, 2010; Todem et al., 2010; Moon & Schorfheide, 2012).

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 $\theta_{ctf} = P\left(\boldsymbol{y}_{\boldsymbol{x}}, \dots, \boldsymbol{z}_{\boldsymbol{w}}\right)$, 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 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 $\boldsymbol{z} \in \mathbb{Z}$.*

In words, Thm. 3.2 formalizes the sense where the 100% credible interval $[l_0, r_0]$ contains the optimal counterfactual bound $[l, r]$, and asymptotically converges to the optimal $[l, r]$ as the number of samples $N_{\boldsymbol{z}}$ from every $\boldsymbol{z} \in \mathbb{Z}$ grows.

Let $\left\{\theta^{(t)}\right\}_{t=1}^T$ be T samples drawn from $P\left(\theta_{\text{ctf}} \mid \bar{\boldsymbol{v}}\right)$. One

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 $\{\theta^{(1)}, \dots, \theta^{(T)}\}$
    from the posterior distribution $P(\theta_{\text{ctf}} \mid \bar{\boldsymbol{v}})$.
Return interval $\left[\hat{l}_\alpha(T), \hat{r}_\alpha(T)\right]$ (Eq. (21)).
```

could compute the $100(1-\alpha)\%$ credible interval for θ_{ctf} using following estimators (Sen & Singer, 1994):

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

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 $\{\theta^{(t)}\}$[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*

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

We summarize our algorithm, CREDIBLEINTERVAL, in Alg. 1. It takes a credible level α and tolerance levels δ, ϵ as inputs. In particular, CREDIBLEINTERVAL repeatedly draw $T \geq \lceil 2\epsilon^{-2} \ln(4/\delta) \rceil$ samples from $P(\theta_{\text{ctf}} \mid \bar{\boldsymbol{v}})$. 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 θ_{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

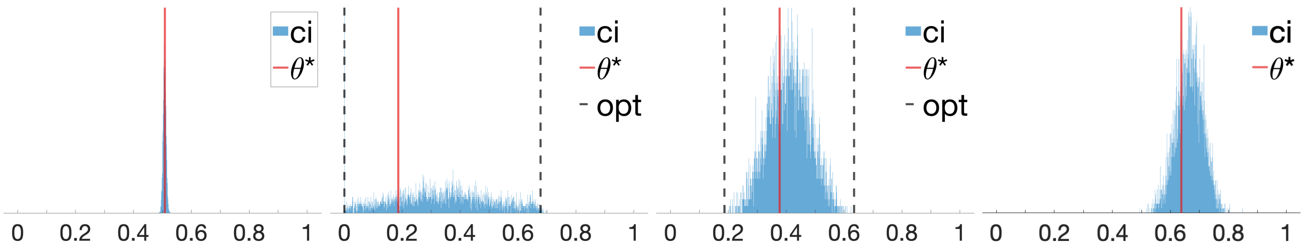
We demonstrate our algorithms on various synthetic and real datasets in different causal diagrams. Overall, we found that simulation results support our findings and the proposed bounding strategy consistently dominates state-of-art algorithms. When target probabilities are identifiable (Experiment 1), 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 4×10^3 samples from the posterior distribution $P(\theta_{\text{ctf}} \mid \bar{\boldsymbol{v}})$ over the target counterfactual. This allows us to compute 100% credible interval over θ_{ctf} within error $\epsilon = 0.05$, with probability at least $1 - \delta = 0.95$. As the baseline, we include the true counterfactual probability θ^*. We refer readers to 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 $(P(Y_{x=1} = 1, Y_{x=0} = 0) \mid \bar{\boldsymbol{v}})$.

[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 α. Similarly, $\lfloor \alpha \rfloor = \max\{n \in \mathbb{Z} \mid n \leq \alpha\}$ is the largest integer $n \in \mathbb{Z}$ smaller than α.



(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; θ^* is the actual counterfactual probability; and *opt* is the optimal asymptotic bounds.

The analysis reveals that the 100% credible interval (*ci*) matches the optimal PNS bound $l = 0, r = 0.6775$ over the actual PNS probability $P(Y_{x=1} = 1, Y_{x=0} = 0) = 0.1867$, 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 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.

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

Blackwell, D. A. and Girshick, M. A. *Theory of games and statistical decisions*. Courier Corporation, 1979.

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.

Durrett, R. *Probability: theory and examples*, volume 49. Cambridge university press, 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.

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.

Grendár, M. and Judge, G. Asymptotic equivalence of empirical likelihood and bayesian map. *The Annals of Statistics*, pp. 2445–2457, 2009.

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.

Moon, H. R. and Schorfheide, F. Bayesian and frequentist inference in partially identified models. *Econometrica*, 80(2):755–782, 2012.

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.

Poirier, D. J. Revising beliefs in nonidentified models. *Econometric theory*, 14(4):483–509, 1998.

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.

Rubin, H. and Wesler, O. A note on convexity in euclidean n-space. *Proceedings of the American Mathematical Society*, 9(4):522–523, 1958.

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.

Silva, R. and Evans, R. Causal inference through a witness protection program. *Journal of Machine Learning Research*, 17, 2016.

Stoye, J. More on confidence intervals for partially identified parameters. *Econometrica*, 77(4):1299–1315, 2009.

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.

A. On the Expressive Power of Canonical Structural Causal Models

We will provide proofs for the partial counterfactual identification algorithm presented in Sec. 2, which establishes the expressive power of canonical SCMs in representing counterfactual distributions in an arbitrary causal diagram containing observed variables with finite domains.

We start the discussion by introducing some necessary notations and concepts. The probability distribution for every exogenous variable $U \in \boldsymbol{U}$ is characterized with a probability space. It is frequently designated $\langle \Omega_U, \mathcal{F}_U, P_U \rangle$ where Ω_U is a sample space containing all possible outcomes; $\mathcal{F}_U$ is a σ-algebra containing subsets of Ω_U; P_U is a probability measure on $\mathcal{F}_U$ normalized by $P_U(\Omega_U) = 1$. Elements of $\mathcal{F}_U$ are called *events*, which are closed under operations of set complement and unions of countably many sets. By means of P_U, a real number $P_U(\mathcal{A}) \in [0, 1]$ is assigned to every event $\mathcal{A} \in \mathcal{F}_U$; it is called the probability of event $\mathcal{A}$.

For an arbitrary set of exogenous variables $\boldsymbol{U}$, its realization $\boldsymbol{U} = \boldsymbol{u}$ is an element in the Cartesian product $\bigtimes_{U \in \boldsymbol{U}} \Omega_U$, represented by a sequence $(u)_{U \in \boldsymbol{U}}$. If now $\mathcal{A}_U \in \Omega_U$, $\forall U \in \boldsymbol{U}$, we may be interested in inferring whether a sequence of events $U \in \mathcal{A}_U$ for every $U \in \boldsymbol{U}$ occurs. Such an event is represented by a subset $\bigtimes_{U \in \boldsymbol{U}} \mathcal{A}_U \subseteq \bigtimes_{U \in \boldsymbol{U}} \Omega_U$. The products $\times_{U \in \boldsymbol{U}} \mathcal{A}_U$ with $\mathcal{A}_U$ running through $\mathcal{F}_U$ generate precisely the product σ-algebra $\bigotimes_{U \in \boldsymbol{U}} \mathcal{F}_U$. The product measure $\bigotimes_{U \in \boldsymbol{U}} P_U$ is the only probability measure P with restrictions to $\bigotimes_{U \in \boldsymbol{U}} \mathcal{F}_U$ that satisfies the following consistency condition

$$P\left(\bigtimes_{U \in \boldsymbol{U}} \mathcal{A}_U\right) = \prod_{U \in \boldsymbol{U}} P_U\left(\mathcal{A}_U\right), \tag{23}$$

for arbitrary $\mathcal{A}_U \in \mathcal{F}_U$. It is obvious that P is a probability measure. Consequently,

$$\left\langle \bigtimes_{U \in \boldsymbol{U}} \Omega_U, \bigotimes_{U \in \boldsymbol{U}} \mathcal{F}_U, \bigotimes_{U \in \boldsymbol{U}} P_U \right\rangle \tag{24}$$

defines the product of probability spaces $\langle \Omega_U, \mathcal{F}_U, P_U \rangle$, $U \in \boldsymbol{U}$. It is adequate to describe all "measurable events" occurring to exogenous variables $\boldsymbol{U}$.

Recall that for subsets $\boldsymbol{X}, \boldsymbol{Y} \subseteq \boldsymbol{V}$, counterfactual random variables (or potential responses) $\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{u})$ is defined as the solution of $\boldsymbol{Y}$ in the submodel $M_{\boldsymbol{x}}$ induced by intervention do$(\boldsymbol{x})$ given the configuration $\boldsymbol{U} = \boldsymbol{u}$. For any $\boldsymbol{y} \in \Omega_{\boldsymbol{Y}}$, let the inverse image $\boldsymbol{Y}_{\boldsymbol{x}}^{-1}(\boldsymbol{y})$ be the set of values $\boldsymbol{u}$ generating the event $\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{u}) = \boldsymbol{y}$, i.e.,

$$\boldsymbol{Y}_{\boldsymbol{x}}^{-1}(\boldsymbol{y}) = \left\{\boldsymbol{u} \in \Omega_{\boldsymbol{U}} \mid \boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{u}) = \boldsymbol{y}\right\}. \tag{25}$$

Evidently, we are dealing with a $\bigotimes_{U \in \boldsymbol{U}} \mathcal{F}_U$-measurable mapping $\boldsymbol{Y}_{\boldsymbol{x}} : \boldsymbol{u} \mapsto \boldsymbol{y}$. Because of this measurability, the inverse image $\boldsymbol{Y}_{\boldsymbol{x}}^{-1}(\boldsymbol{y})$ is an event in $\bigotimes_{U \in \boldsymbol{U}} \mathcal{F}_U$ for any realization $\boldsymbol{y}$. Thus $P\left(\boldsymbol{Y}_{\boldsymbol{x}}^{-1}(\boldsymbol{y})\right)$ is defined as the probability of $\boldsymbol{Y}_{\boldsymbol{x}}$ taking on a value $\boldsymbol{y}$. Similarly, for any $\boldsymbol{Y}, \dots, \boldsymbol{Z}$, $\boldsymbol{X}, \dots, \boldsymbol{W} \subseteq \boldsymbol{V}$, the probability of a sequence of counterfactual events $\boldsymbol{Y}_{\boldsymbol{x}} = \boldsymbol{y}, \dots, \boldsymbol{Z}_{\boldsymbol{w}} = \boldsymbol{z}$ is defined as:

$$P\left(\boldsymbol{y}_{\boldsymbol{x}}, \dots, \boldsymbol{z}_{\boldsymbol{w}}\right) = P\left(\boldsymbol{Y}_{\boldsymbol{x}}^{-1}(\boldsymbol{y}) \cap \dots \cap \boldsymbol{Z}_{\boldsymbol{w}}^{-1}(\boldsymbol{z})\right).$$

We refer readers to (Durrett, 2019; Bauer, 1972) for a detailed discussion on measure-theoretic probability concepts.

A.1. Proof for Theorem 2.4

We first provide the construction for canonical SCMs in Thm. 2.4, showing that they could generate all counterfactual distributions in an arbitrary causal diagram. The validity and tightness of bounds $[l, r]$ in Eq. (9) naturally follows.

For every endogenous $V \in \boldsymbol{V}$, let $\Omega_{Pa_V} \mapsto \Omega_V$ denote the hypothesis class containing all functions mapping from domains of PA_V to V. Since $\boldsymbol{V}$ are discrete variables with finite domains, the cardinality of the class $\Omega_{Pa_V} \mapsto \Omega_V$ must be also finite. Given any configuration $U_V = u_V$, the induced function $f_V(\cdot, u_V)$ must correspond to a unique element in the hypothesis class $\Omega_{PA_V} \mapsto \Omega_V$. Such mappings lead to a finite partition over the exogenous domain Ω_{U_V}.

Definition A.1. For an SCM $M = \langle \boldsymbol{V}, \boldsymbol{U}, \mathscr{F}, P \rangle$, for every $V \in \boldsymbol{V}$, let functions in $\Omega_{PA_V} \mapsto \Omega_V$ be ordered by $\left\{h_V^{(i)} \mid i \in \boldsymbol{I}_V\right\}$ where $\boldsymbol{I}_V = \{1, \dots, m_V\}, m_V = |\Omega_{PA_V} \mapsto \Omega_V|$. A *equivalence class* $\mathcal{U}_V^{(i)}$ for function $h_V^{(i)}$, $i = 1, \dots, m_V$, is a subset in Ω_{U_V} such that

$$\mathcal{U}_V^{(i)} = \left\{u_V \in \Omega_{U_V} \mid f_V(\cdot, u_V) = h_V^{(i)}\right\}. \tag{26}$$

Definition A.2 (Canonical Partition)**.** For an SCM $M = \langle \boldsymbol{V}, \boldsymbol{U}, \mathscr{F}, P \rangle$, $\left\{\mathcal{U}_V^{(i)} \mid i \in \boldsymbol{I}_V\right\}$ is the *canonical partition* over exogenous domain Ω_{U_V} for every $V \in \boldsymbol{V}$.

Def. A.2 extends the canonical partition in (Balke & Pearl, 1994) which was designed for binary variables $X, Y, Z \in \{0, 1\}$ in the "IV" diagram of Fig. 1a.

As exogenous variables U_V vary along its domain, regardless of how complex the variation is, its only effect is to switch the functional relationship between Pa_V and V among elements in class $\Omega_{PA_V} \mapsto \Omega_V$. Formally,

Lemma A.3. *For an SCM $M = \langle \boldsymbol{V}, \boldsymbol{U}, \mathscr{F}, P \rangle$, for each $V \in \boldsymbol{V}$, function $f_V \in \mathscr{F}$ could be decomposed as:*

$$f_V(pa_V, u_V) = \sum_{i \in \boldsymbol{I}_V} h_V^{(i)}(pa_V) \mathbb{1}_{u_V \in \mathcal{U}_V^{(i)}}. \tag{27}$$

Proof. By the definition of canonical partitions (Def. A.2), for every $i = 1, \dots, m_V$, fix any $u_V \in \mathcal{U}_V^{(i)}$.

We must have $f_V(\cdot, u_V) = h_V^{(i)}(\cdot)$. This implies $f_V(pa_V, u_V) = h_V^{(i)}(pa_V)$ for any $PA_V = pa_V$. Recall that $\left\{\mathcal{U}_V^{(i)} \mid i = 1, \dots, m_V\right\}$ forms a partition over the domain Ω_{U_V}. Given the same pa_V, u_V, the r.h.s. of Eq. (27) must equate to $h_V^{(i)}(pa_V)$, which completes the proof. □

As an example, consider an SCM M associated with the "IV" graph of Fig. 1a where X, Y, Z are binary variables contained in $\{0, 1\}$; U_1, U_2 are continuous variables drawn uniformly from the interval $[0, 3]$. Values of X, Y, Z are decided by functions defined as follows, respectively,

$$\begin{aligned} x &\leftarrow f_X(z, u_2) = \mathbb{1}_{z \leq u_2 \leq z+2}, \\ y &\leftarrow f_Y(x, u_2) = \mathbb{1}_{u_2 < x} + \mathbb{1}_{u_2 > x+2}, \\ z &\leftarrow f_Z(u_1) = \mathbb{1}_{u_1 \leq 1.5}, \end{aligned} \tag{28}$$

We show in Fig. 4 the graphical representation of canonical partitions induced by functions f_X, f_Y and f_Z respectively. A detailed description is provided in Table 1. It follows from the decomposition of Lem. A.3 that functions f_X, f_Y, f_Z in Eq. (28) could be written as follows:

$$\begin{aligned} f_X(z, u_2) &= \mathbb{1}_{u_2 \in [0,1)} \neg z + \mathbb{1}_{u_2 \in [1,2]} 1 + \mathbb{1}_{u_2 \in (2,3]} z \\ f_Y(x, u_2) &= \mathbb{1}_{u_2 \in [0,1)} x + \mathbb{1}_{u_2 \in [1,2]} 0 + \mathbb{1}_{u_2 \in (2,3]} \neg x, \\ f_Z(u_1) &= \mathbb{1}_{u_1 \in [0,1.5]} 1 + \mathbb{1}_{u_1 \in (1.5,3]} 0. \end{aligned}$$

Let $\boldsymbol{I}$ denote the product of indexing sets $\times_{V \in \boldsymbol{V}} \boldsymbol{I}_V$. For any index $\boldsymbol{i} \in \boldsymbol{I}$, we use i_V to represent the element in $\boldsymbol{i}$ restricted to $V \in \boldsymbol{V}$. We omit the subscript V when it is obvious; therefore, $\mathcal{U}_V^{(i)} = \mathcal{U}_V^{(i_V)}$, $h_V^{(i)} = h_V^{(i_V)}$. Our next result establishes a universal decomposition of counterfactual distributions in any SCM using canonical partitions.

Lemma A.4. *For an SCM $M = \langle \boldsymbol{V}, \boldsymbol{U}, \mathscr{F}, P \rangle$, for any $\boldsymbol{Y}, \dots, \boldsymbol{Z}, \boldsymbol{X}, \dots, \boldsymbol{W} \subseteq \boldsymbol{V}$[5],*

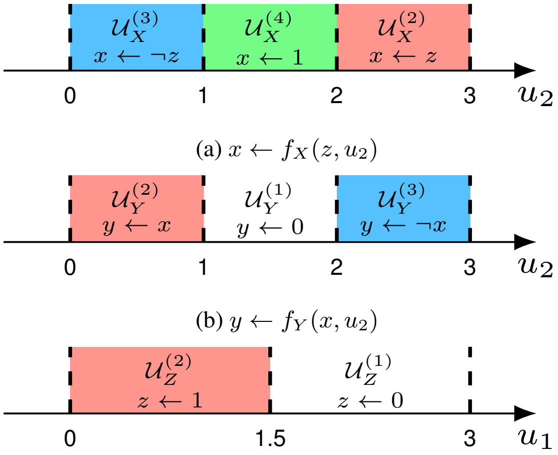
$$\begin{aligned} & P\left(\boldsymbol{y_x}, \dots, \boldsymbol{z_w}\right) \\ &= \sum_{\boldsymbol{i} \in \boldsymbol{I}} \mathbb{1}_{\boldsymbol{Y_x}(\boldsymbol{i}) = \boldsymbol{y}, \dots, \boldsymbol{Z_w}(\boldsymbol{i}) = \boldsymbol{z}} P\left(\bigcap_{V \in \boldsymbol{V}} \mathcal{U}_V^{(i)}\right), \end{aligned} \tag{29}$$

where variables of the form $\mathbf{Y}_{\boldsymbol{x}}(\boldsymbol{i}) = \{Y_{\boldsymbol{x}}(\boldsymbol{i}) \mid \forall Y \in \boldsymbol{Y}\}$; every $Y_{\boldsymbol{x}}(\boldsymbol{i})$ is recursively defined as:

$$Y_{\boldsymbol{x}}(\boldsymbol{i}) = \begin{cases} \boldsymbol{x}_Y & \text{if } Y \in \boldsymbol{X} \\ h_Y^{(i)}\left((PA_Y)_{\boldsymbol{x}}(\boldsymbol{i})\right) & \text{otherwise} \end{cases} \tag{30}$$

Proof. We will first prove the following claims: for arbitrary subsets $\boldsymbol{Y}, \boldsymbol{X} \subseteq \boldsymbol{V}$, for any $\boldsymbol{u}, \boldsymbol{x}, \boldsymbol{y}$,

$$\mathbb{1}_{\boldsymbol{Y_x}(\boldsymbol{u}) = \boldsymbol{y}} = \sum_{\boldsymbol{i} \in \boldsymbol{I}} \mathbb{1}_{\boldsymbol{Y_x}(\boldsymbol{i}) = \boldsymbol{y}} \prod_{V \in \boldsymbol{V}} \mathbb{1}_{u_V \in \mathcal{U}_V^{(i)}}. \tag{31}$$



(a) $x \leftarrow f_X(z, u_2)$

(b) $y \leftarrow f_Y(x, u_2)$

(c) $z \leftarrow f_Z(u_1)$

Figure 4: Canonical partitions for exogenous domains over U_1, U_2 induced by functions of X, Y, Z defined in Eq. (28).

	$0 \leq U_2 < 1$	$1 \leq U_2 \leq 2$	$2 < U_2 \leq 3$
$Z = 0$	$X = 1$	$X = 1$	$X = 0$
$Z = 1$	$X = 0$	$X = 1$	$X = 1$

(a) $x \leftarrow f_X(z, u_2)$

	$0 \leq U_2 < 1$	$1 \leq U_2 \leq 2$	$2 < U_2 \leq 3$
$X = 0$	$Y = 0$	$Y = 0$	$Y = 1$
$X = 1$	$Y = 1$	$Y = 0$	$Y = 0$

(b) $y \leftarrow f_Y(x, u_2)$

$0 \leq U_1 < 1.5$	$1.5 \leq U_1 \leq 3$
$Z = 1$	$Z = 0$

(c) $z \leftarrow f_Z(u_1)$

Table 1: Canonical partitions for exogenous domains over U_1, U_2 induced by functions of X, Y, Z defined in Eq. (28).

Let $\mathcal{G}_{\overline{\boldsymbol{X}}}$ be a subgraph obtained from the causal diagram $\mathcal{G}$ by removing all incoming arrows of $\boldsymbol{X}$. We will prove Eq. (31) by induction on $n = \max_{Y \in \boldsymbol{Y}} \left|An(Y)_{\mathcal{G}_{\overline{\boldsymbol{X}}}}\right|$.

Base Case $n = 1$**.** Recall that an intervention do($\boldsymbol{x}$) set values of variables $\boldsymbol{X}$ as constants $\boldsymbol{x}$. For any $Y \in \boldsymbol{X} \cap \boldsymbol{Y}$, let $\boldsymbol{x}_Y$ be the values assigned to Y in $\boldsymbol{x}$. It is verifiable that

$$\mathbb{1}_{Y_{\boldsymbol{x}}(\boldsymbol{u}) = y} = \mathbb{1}_{y = \boldsymbol{x}_Y} \tag{32}$$

As for every variable $Y \in \boldsymbol{Y} \setminus \boldsymbol{X}$, we must have its parent nodes $PA_Y = \emptyset$ since $n = 1$. This implies

$$\mathbb{1}_{Y_{\boldsymbol{x}}(\boldsymbol{u}) = y} = \mathbb{1}_{f_Y(u_Y) = y} = \sum_{i \in \boldsymbol{I}_Y} \mathbb{1}_{h_Y^{(i)} = y} \mathbb{1}_{u_Y \in \mathcal{U}_Y^{(i)}} \tag{33}$$

[5] For an arbitrary subset $\mathcal{U} \subseteq \Omega_{\boldsymbol{U}}$, we will consistently use $P(\mathcal{U})$ as a shorthand for the probability $P(\boldsymbol{U} \in \mathcal{U})$.

The last step follows from the decomposition in Lem. A.3. Eqs. (32) and (33) together imply that

$$\begin{aligned}&\mathbb{1}_{\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{u})=\boldsymbol{y}}\\ &=\sum_{i\in\boldsymbol{I}}\prod_{Y\in\boldsymbol{Y}\cap\boldsymbol{X}}\mathbb{1}_{y=\boldsymbol{x}_Y}\prod_{Y\in(\boldsymbol{Y}\setminus\boldsymbol{X})}\mathbb{1}_{h_Y^{(i)}=y}\prod_{V\in\boldsymbol{V}}\mathbb{1}_{u_V\in\mathcal{U}_V^{(i)}}\\ &=\sum_{i\in\boldsymbol{I}}\mathbb{1}_{\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{i})=\boldsymbol{y}}\prod_{V\in\boldsymbol{V}}\mathbb{1}_{u_V\in\mathcal{U}_V^{(i)}}.\end{aligned}$$

The last step follows from the definition of variables $\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{i})$ in Eq. (30) given an index $\boldsymbol{i}\in\boldsymbol{I}$.

Induction Case $n=k+1$. Assume that Eq. (31) holds for $n=k$. We will prove for the case $n=k+1$. For every $Y\in\boldsymbol{X}\cap\boldsymbol{Y}$, $\mathbb{1}_{Y_{\boldsymbol{x}}(\boldsymbol{u})=y}$ is given in Eq. (32). For every $Y\in\boldsymbol{Y}\setminus\boldsymbol{X}$, the decomposition in Lem. A.3 implies:

$$\begin{aligned}&\mathbb{1}_{Y_{\boldsymbol{x}}(\boldsymbol{u})=y}\\ &=\mathbb{1}_{f_Y((PA_Y)_{\boldsymbol{x}}(\boldsymbol{u}),u_Y)=y}\\ &=\mathbb{1}\left\{y=\sum_{i\in\boldsymbol{I}_Y}h_Y^{(i)}((PA_Y)_{\boldsymbol{x}}(\boldsymbol{u}))\mathbb{1}_{u_Y\in\mathcal{U}_Y^{(i)}}\right\}\\ &=\sum_{i\in\boldsymbol{I}_Y}\mathbb{1}_{h_Y^{(i)}((PA_Y)_{\boldsymbol{x}}(\boldsymbol{u}))=y}\mathbb{1}_{u_Y\in\mathcal{U}_Y^{(i)}}\\ &=\sum_{i\in\boldsymbol{I}_Y}\sum_{pa_Y}\mathbb{1}_{h_Y^{(i)}(pa_Y)=y}\mathbb{1}_{(PA_Y)_{\boldsymbol{x}}(\boldsymbol{u})=pa_Y}\mathbb{1}_{u_Y\in\mathcal{U}_Y^{(i)}}.\end{aligned}$$

The last step hold by conditioning on events $(PA_Y)_{\boldsymbol{x}}(\boldsymbol{u})=pa_Y, \forall pa_Y\in\Omega_{PA_Y}$. Since we assume Eq. (31) holds for Case $n=k$, the above equation could be further written as

$$\begin{aligned}\mathbb{1}_{Y_{\boldsymbol{x}}(\boldsymbol{u})=y}=&\sum_{i\in\boldsymbol{I}_Y}\sum_{pa_Y}\mathbb{1}_{h_Y^{(i)}(pa_Y)=y}\mathbb{1}_{u_Y\in\mathcal{U}_Y^{(i)}}\\ &\cdot\sum_{i\in\boldsymbol{I}}\mathbb{1}_{(PA_Y)_{\boldsymbol{x}}(\boldsymbol{u})=pa_Y}\prod_{V\in\boldsymbol{V}}\mathbb{1}_{u_V\in\mathcal{U}_V^{(i)}}\end{aligned}$$

A few simplification gives:

$$\begin{aligned}&\mathbb{1}_{Y_{\boldsymbol{x}}(\boldsymbol{u})=y}\\ &=\sum_{i\in\boldsymbol{I}}\sum_{pa_Y}\mathbb{1}_{h_Y^{(i)}(pa_Y)=y}\mathbb{1}_{(PA_Y)_{\boldsymbol{x}}(\boldsymbol{u})=pa_Y}\prod_{V\in\boldsymbol{V}}\mathbb{1}_{u_V\in\mathcal{U}_V^{(i)}}\\ &=\sum_{i\in\boldsymbol{I}}\mathbb{1}_{h_Y^{(i)}((PA_Y)_{\boldsymbol{x}}(\boldsymbol{u}))=y}\prod_{V\in\boldsymbol{V}}\mathbb{1}_{u_V\in\mathcal{U}_V^{(i)}}.\end{aligned}\tag{34}$$

Eqs. (32) and (34) together imply that

$$\begin{aligned}&\mathbb{1}_{\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{u})=\boldsymbol{y}}\\ &=\sum_{i\in\boldsymbol{I}}\left(\prod_{Y\in\boldsymbol{Y}\cap\boldsymbol{X}}\mathbb{1}_{y=\boldsymbol{x}_Y}\prod_{Y\in(\boldsymbol{Y}\setminus\boldsymbol{X})}\mathbb{1}_{h_Y^{(i)}((PA_Y)_{\boldsymbol{x}}(\boldsymbol{u}))=y}\right)\\ &\cdot\prod_{V\in\boldsymbol{V}}\mathbb{1}_{u_V\in\mathcal{U}_V^{(i)}}\\ &=\sum_{i\in\boldsymbol{I}}\mathbb{1}_{\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{i})=\boldsymbol{y}}\prod_{V\in\boldsymbol{V}}\mathbb{1}_{u_V\in\mathcal{U}_V^{(i)}}.\end{aligned}$$

Again, the last step follows from the definition of variables $\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{i})$ in Eq. (30) given an index $\boldsymbol{i}\in\boldsymbol{I}$.

We are now ready to prove Eq. (29). The statement of Eq. (31) implies that for any $\boldsymbol{Y},\dots,\boldsymbol{Z},\boldsymbol{X},\dots,\boldsymbol{W}\subseteq\boldsymbol{V}$,

$$\begin{aligned}&P\left(\boldsymbol{y}_{\boldsymbol{x}},\dots,\boldsymbol{z}_{\boldsymbol{w}}\right)\\ &=\int_{\Omega_{\boldsymbol{U}}}\mathbb{1}_{\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{u})=\boldsymbol{y},\dots,\boldsymbol{Z}_{\boldsymbol{w}}(\boldsymbol{u})=\boldsymbol{z}}dP(\boldsymbol{u})\\ &=\int_{\Omega_{\boldsymbol{U}}}\left(\sum_{\boldsymbol{i}\in\boldsymbol{I}}\mathbb{1}_{\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{i})=\boldsymbol{y}}\prod_{V\in\boldsymbol{V}}\mathbb{1}_{u_V\in\mathcal{U}_V^{(i)}}\right)\wedge\\ &\cdots\cap\left(\sum_{\boldsymbol{i}\in\boldsymbol{I}}\mathbb{1}_{\boldsymbol{Z}_{\boldsymbol{w}}(\boldsymbol{i})=\boldsymbol{z}}\prod_{V\in\boldsymbol{V}}\mathbb{1}_{u_V\in\mathcal{U}_V^{(i)}}\right)dP(\boldsymbol{u})\end{aligned}$$

Simplifying the above equation gives:

$$\begin{aligned}&P\left(\boldsymbol{y}_{\boldsymbol{x}},\dots,\boldsymbol{z}_{\boldsymbol{w}}\right)\\ &=\int_{\Omega_{\boldsymbol{U}}}\sum_{\boldsymbol{i}\in\boldsymbol{I}}\mathbb{1}_{\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{i})=\boldsymbol{y}}\wedge\cdots\wedge\mathbb{1}_{\boldsymbol{Z}_{\boldsymbol{w}}(\boldsymbol{i})=\boldsymbol{z}}\prod_{V\in\boldsymbol{V}}\mathbb{1}_{u_V\in\mathcal{U}_V^{(i)}}dP(\boldsymbol{u})\\ &=\sum_{\boldsymbol{i}\in\boldsymbol{I}}\mathbb{1}_{\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{i})=\boldsymbol{y}}\wedge\cdots\wedge\mathbb{1}_{\boldsymbol{Z}_{\boldsymbol{w}}(\boldsymbol{i})=\boldsymbol{z}}\int_{\Omega_{\boldsymbol{U}}}\prod_{V\in\boldsymbol{V}}\mathbb{1}_{u_V\in\mathcal{U}_V^{(i)}}dP(\boldsymbol{u})\\ &=\sum_{\boldsymbol{i}\in\boldsymbol{I}}\mathbb{1}_{\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{i})=\boldsymbol{y},\dots,\boldsymbol{Z}_{\boldsymbol{w}}(\boldsymbol{i})=\boldsymbol{z}}P\left(\bigcap_{V\in\boldsymbol{V}}\mathcal{U}_V^{(i)}\right).\end{aligned}$$

In the above equations, the last two steps hold since variables $\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{i}),\dots,\boldsymbol{Z}_{\boldsymbol{w}}(\boldsymbol{i})$ are not functions of exogenous variables $\boldsymbol{U}$. This completes the proof. □

Let $\mathcal{C}(\mathcal{G})$ denote the collection of all maximal c-components (Def. 2.2) in a causal diagram $\mathcal{G}$. For instance, in the "IV" diagram $\mathcal{G}$ of Fig. 1a, $\mathcal{C}(\mathcal{G})$ contains c-components $\boldsymbol{C}(U_1)=\{Z\}, \boldsymbol{C}(U_2)=\{X,Y\}$. The following proposition shows that probabilities over canonical partitions factorize over c-components in a causal diagram.

Lemma A.5. *For an SCM $M=\langle\boldsymbol{V},\boldsymbol{U},\mathscr{F},P\rangle$, let $\mathcal{G}$ be the associated causal diagram. For any $\boldsymbol{i}\in\boldsymbol{I}$,*

$$P\left(\bigcap_{V\in\boldsymbol{V}}\mathcal{U}_V^{(i)}\right)=\prod_{\boldsymbol{C}\in\mathcal{C}(\mathcal{G})}P\left(\bigcap_{V\in\boldsymbol{C}}\mathcal{U}_V^{(i)}\right).\tag{35}$$

Proof. For any c-compoment $\boldsymbol{C}\in\mathcal{C}(\mathcal{G})$, let $U_{\boldsymbol{C}}=\cup_{V\in\boldsymbol{C}}U_V$ the set of exogenous variables affecting (at least one of) endogenous variables in $\boldsymbol{C}$. By the definition of c-components (Def. 2.2), it is verifiable that for two different c-compoments $\boldsymbol{C}_1,\boldsymbol{C}_2\in\mathcal{C}(\mathcal{G})$, their corresponding exogenous variables $U_{\boldsymbol{C}_1},U_{\boldsymbol{C}_2}$ do not share any element, i.e., $U_{\boldsymbol{C}_1}\cap U_{\boldsymbol{C}_2}=\emptyset$. We complete the proof by noting that exogenous variables in $\boldsymbol{U}$ are mutually independent. □

As an example, consider again the SCM M compatible with Fig. 1a defined in Eq. (28). The event $Z=1,X_{z=0}=$

$1, Y_{x=1} = 0$ occurs if any only if $U_1 \in \mathcal{U}_Z^{(2)}$ and $U_2 \in \left(\mathcal{U}_X^{(3)} \cup \mathcal{U}_X^{(4)}\right) \cap \left(\mathcal{U}_Y^{(1)} \cup \mathcal{U}_Y^{(3)}\right)$. This implies

$$\begin{aligned}&P\left(Z=1, X_{z=0}=1, Y_{x=1}=0\right)\\&= P\left(\mathcal{U}_Z^{(2)} \cap (\mathcal{U}_X^{(3)} \cup \mathcal{U}_X^{(4)}) \cap (\mathcal{U}_Y^{(1)} \cup \mathcal{U}_Y^{(3)})\right)\\&= P\left(\mathcal{U}_Z^{(2)}\right) P\left(\mathcal{U}_X^{(3)} \cup \mathcal{U}_X^{(4)}) \cap (\mathcal{U}_Y^{(1)} \cup \mathcal{U}_Y^{(3)})\right).\end{aligned}$$

The last step holds since $\{Z\}$ and $\{X, Y\}$ are two different c-components. It is verifiable from Fig. 4 that $\mathcal{U}_Z^{(2)} = \{u_1 \in [0, 1.5]\}$, $\left(\mathcal{U}_X^{(3)} \cup \mathcal{U}_X^{(4)}\right) \cap \left(\mathcal{U}_Y^{(1)} \cup \mathcal{U}_Y^{(3)}\right) = \{u_2 \in [1, 2]\}$. The above equation could be further written as:

$$\begin{aligned}&P\left(Z=1, X_{z=0}=1, Y_{x=1}=0\right)\\&= P\left(U_1 \in [0, 1.5]\right) P\left(U_2 \in [1, 2]\right) = \frac{1}{6}.\end{aligned}$$

The last step follows since variables U_1, U_2 are drawn uniformly at random over the interval $[0, 3]$.

Lems. A.4 and A.5 together allow us to write any counterfactual distribution in an SCM as a function of products of probabilities assigned to the intersections of canonical partitions in every c-component. To prove Thm. 2.4, it is thus sufficient to construct a canonical SCM N from an arbitrary SCM M such that (1) M, N are compatible with the same causal diagram $\mathcal{G}$; and (2) M, N generate the same probabilities over canonical partitions. This section will describe how to construct such a discrete SCM.

For a c-component $\boldsymbol{C}$ in a causal diagram $\mathcal{G}$, we denote by $U_{\boldsymbol{C}} = \cup_{V \in \boldsymbol{C}} U_V$ the union of exogenous variables U_V affecting an endogenous variable V for every $V \in \boldsymbol{C}$. , $m = |U_{\boldsymbol{C}}|$. For convenience, we consistently write $\langle \Omega_i, \mathcal{F}_i, \rho_i \rangle$ as the probability space of U_i, $i = 1, \dots, m$. The product of these probability spaces is thus written as

$$\left\langle \bigtimes_{i=1}^{m} \Omega_i, \bigotimes_{i=1}^{m} \mathcal{F}_i, \bigotimes_{i=1}^{m} \rho_i \right\rangle. \tag{36}$$

For any SCM M compatible with the diagram $\mathcal{G}$, the joint distribution over events defined by canonical partitions $\mathcal{U}_V^{(i)}$ associated with variables $V \in \boldsymbol{C}$ is given by

$$\begin{aligned}&P\left(\bigcap_{V \in \boldsymbol{C}} \mathcal{U}_V^{(i)}\right)\\&= \int_{\times_{i=1}^{m} \Omega_i} \prod_{V \in \boldsymbol{C}} \mathbb{1}_{u_V \in \mathcal{U}_V^{(i)}} d\left(\bigotimes_{i=1}^{m} \rho_i\right).\end{aligned} \tag{37}$$

Our goal is to show that all correlations among events $\mathcal{U}_V^{(i)}$, $V \in \boldsymbol{V}$, induced by exogenous variables described by arbitrary probability spaces could be produced by a "simpler" generative process with discrete exogenous domains.

Lemma A.6. *Any distribution $P\left(\bigcap_{V \in \boldsymbol{C}} \mathcal{U}_V^{(i)}\right)$ in Eq. (37) could be reproduced with a generic model of the form:*

$$\begin{aligned}&P\left(\bigcap_{V \in \boldsymbol{C}} \mathcal{U}_V^{(i)}\right)\\&= \sum_{j=1}^{m} \sum_{u_j=1}^{d} \prod_{V \in \boldsymbol{C}} \mathbb{1}_{u_V \in \mathcal{U}_V^{(i)}} \prod_{j=1}^{m} P(u_j),\end{aligned} \tag{38}$$

where every exogenous variable $U_j \in \boldsymbol{U}$ takes values in a finite domain $\{1, \dots, d\}$, $d = \prod_{V \in \boldsymbol{C}} |\Omega_{PA_V} \mapsto \Omega_V|$.

(Rosset et al., 2018, Proposition 2) applied a classic result of Carathéodory theorem in convex geometry (Carathéodory, 1911) and showed that the observational distribution in any causal diagram could be generated using discrete exogenous variables, assuming that exogenous variables are drawn from distributions with probability density functions. We here present a constructive proof that applies to the general framework of measure-theoretic probability theory.

Proof of Lemma A.6. Let $\vec{P}$ be a vector representing probabilities of $\left(P\left(\bigcap_{V \in \boldsymbol{C}} \mathcal{U}_V^{(i)}\right)\right)_{\boldsymbol{i} \in \boldsymbol{I}}$. Note that for every $V \in \boldsymbol{V}$, there are $|\Omega_{PA_V} \mapsto \Omega_V|$ equivalence classes $\mathcal{U}_V^{(i)}$. $\vec{P}$ is thus a vector with $d = \prod_{V \in \boldsymbol{C}} |\Omega_{PA_V} \mapsto \Omega_V|$ elements. Since $\sum_{\boldsymbol{i}} P\left(\bigcap_{V \in \boldsymbol{C}} \mathcal{U}_V^{(i)}\right) = 1$, it only takes a vector with $d-1$ dimensions to determine $\vec{P}$. We could thus see $\vec{P}$ as a point in the $(d-1)$-dimensional real space. Similarly, $\left(\vec{P}, 1\right)$ is vector in $\mathbb{R}^d$ where the d-th element is equal to 1

Fix an exogenous variable $U_1 \in U_{\boldsymbol{C}}$. We define function $P_{u_1}\left(\bigcap_{V \in \boldsymbol{C}} \mathcal{U}_V^{(i)}\right)$ as the distribution over canonical partitions when U_1 is fixed as a constant $u_1 \in \Omega_1$. That is,

$$\begin{aligned}&P_{u_1}\left(\bigcap_{V \in \boldsymbol{C}} \mathcal{U}_V^{(i)}\right)\\&= \left[\int_{\times_{j=2}^{m} \Omega_i} \prod_{V \in \boldsymbol{C}} \mathbb{1}_{u_V \in \mathcal{U}_V^{(i)}} d\left(\bigotimes_{j=2}^{m} P_j\right)\right]_{U_1 = u_1}\end{aligned} \tag{39}$$

The associativity of the product of probability spaces (Bauer, 1972, Ch. 3.3) generally implies:

$$\begin{aligned}\bigotimes_{j=1}^{m} \mathcal{F}_j &= \mathcal{F}_1 \otimes \left(\bigotimes_{j=2}^{m} \mathcal{F}_j\right),\\ \bigotimes_{j=1}^{m} P_j &= P_1 \otimes \left(\bigotimes_{j=2}^{m} P_j\right).\end{aligned} \tag{40}$$

Let $\vec{P}_{u_1}$ be a vector in $\mathbb{R}^{d-1}$ representing probabilities of $P_{u_1}\left(\bigcap_{V \in \boldsymbol{C}} \mathcal{U}_V^{(i)}\right)$ and let $\left(\vec{P}_{u_1}, 1\right)$ be vector in $\mathbb{R}^d$ where the d-th element is equal to 1. Applying Fubini's Theorem (Durrett, 2019, Thm. 1.7.2) implies that function $u_1 \mapsto \left(\vec{P}_{u_1}, 1\right)$ is $\mathcal{F}_1$-measurable. That is, $\langle \Omega_1, \mathcal{F}_1, P_1 \rangle$ yields a probability measure for a set $\left\{\left(\vec{P}_{u_1}, 1\right) \mid \forall u_1 \in \Omega_1\right\}$ with respective to Borel sets in real space $\mathbb{R}^d$ with average

$$\left(\vec{P}, 1\right) = \int_{\Omega_1} \left(\vec{P}_{u_1}, 1\right) dP_1. \tag{41}$$

It can be shown that the probability vector $\left(\vec{P}, 1\right)$ is a point lying in the convex hull of a set $\left\{\left(\vec{P}_{u_1}, 1\right) \mid \forall u \in \Omega_U\right\}$ (see (Blackwell & Girshick, 1979, Thm. 2.4.1) and its extension to arbitrary probability measures in (Rubin & Wesler, 1958)). This means that there exists a finite set of vectors $\left(\vec{P}_{u_1^{(1)}}, 1\right), \dots, \left(\vec{P}_{u_1^{(n)}}, 1\right)$ and a sequence of positive coefficients $\alpha_1, \dots, \alpha_n > 0$ such that

$$\left(\vec{P}, 1\right) = \sum_{k=1}^{n} \alpha_k \left(\vec{P}_{u_1^{(k)}}, 1\right). \tag{42}$$

The above equation implies

$$\vec{P} = \sum_{k=1}^{n} \alpha_k \vec{P}_{u_1^{(k)}}, \qquad \text{and } \sum_{k=1}^{n} \alpha_k = 1 \tag{43}$$

Indeed, we could further reduce the number of coefficients n by removing linearly dependent vectors. If vectors $\left(\vec{P}_{u_1^{(k)}}, 1\right)$ are not linearly independent, there exists a non-trivial solution $\lambda_1, \dots \lambda_n$ such that $\sum_k \lambda_k \left(\vec{P}_{u_1^{(k)}}, 1\right) = \vec{0}$. It is verifiable that for any real value $\beta > 0$

$$\sum_{k=1}^{n} (\alpha_k - \beta\lambda_k) \left(\vec{P}_{u_1^{(k)}}, 1\right) \tag{44}$$

$$= \sum_{k=1}^{n} \alpha_k \left(\vec{P}_{u_1^{(k)}}, 1\right) - \beta \sum_{k=1}^{n} \lambda_k \left(\vec{P}_{u_1^{(k)}}, 1\right) \tag{45}$$

$$= \sum_{k=1}^{n} \alpha_k \left(\vec{P}_{u_1^{(k)}}, 1\right). \tag{46}$$

The last step holds since $\sum_k \lambda_k \left(\vec{P}_{u_1^{(k)}}, 1\right) = \vec{0}$. Therefore, coefficients $\alpha_k - \beta\lambda_k$, $k = 1, \dots, n$, satisfy

$$\sum_{k=1}^{n} (\alpha_k - \beta\lambda_k) \left(\vec{P}_{u_1^{(k)}}, 1\right) = \left(\vec{P}, 1\right). \tag{47}$$

Let β be the largest value such that $\alpha_k - \beta\lambda_k \geq 0$ for all k. Consequently, there must exist a coefficient $\alpha_k - \beta\lambda_k = 0$. We could then remove the corresponding vector $\left(\vec{P}_{u_1^{(k)}}, 1\right)$ from the base. This procedure continues until all remaining vectors are linearly independent. Since $\left(\vec{P}_{u_1}, 1\right) \in \mathbb{R}^d$, there are at most d linearly independent vectors, i.e., $n \leq d$.

Finally, we replace the probability measure P_1 with a discrete distribution $P\left(U_1 = u_1^{(k)}\right) = w_k$ over a finite discrete domain $\Omega_1^* = \left\{u_1^{(1)}, \dots, u_1^{(d)}\right\}$. Doing so generated a new SCM N^*, with cardinality $|\Omega_1| \leq d$, that reproduces probabilities $P\left(\bigcap_{V \in \boldsymbol{C}} \mathcal{U}_V^{(i)}\right)$ over canonical partitions in the original SCM M. Repeatedly applying this procedure for every exogenous $U_2, \dots, U_m$ completes the proof. □

Lems. A.4 to A.6 together yield a natural constructive proof for Thm. 2.4 in an arbitrary causal diagram $\mathcal{G}$.

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 $\boldsymbol{Y}_{\boldsymbol{x}}, \dots, \boldsymbol{Z}_{\boldsymbol{w}}$, $P_M\left(\boldsymbol{Y}_{\boldsymbol{x}}, \dots, \boldsymbol{Z}_{\boldsymbol{w}}\right) = P_N\left(\boldsymbol{Y}_{\boldsymbol{x}}, \dots, \boldsymbol{Z}_{\boldsymbol{w}}\right)$.*

Proof. By the definition of c-components (Def. 2.2), it is verifiable that for two different c-compoments $\boldsymbol{C}_1, \boldsymbol{C}_2 \in \mathcal{C}(\mathcal{G})$, their corresponding exogenous variables $U_{\boldsymbol{C}_1}, U_{\boldsymbol{C}_2}$ do not share any element, i.e., $U_{\boldsymbol{C}_1} \cap U_{\boldsymbol{C}_2} = \emptyset$. Therefore, we could repeatedly apply the construction of Lem. A.6 for every c-component $\boldsymbol{C} \in \mathcal{C}(\mathcal{G})$. Doing so generates a discrete SCM N satisfying conditions as follows:

1. N and M are compatible with the same diagram $\mathcal{G}$;
2. N and M share the same set of structural functions $\mathscr{F}$;
3. N and M generate the same joint distribution over the intersections of canonical partitions associated with every c-component.

It follows from Lems. A.4 and A.5 that M and N must coincide in all counterfactual distributions over endogenous variables. This completes the proof. □

A.2. Proofs for Other Results in Section 2

More generally, Thm. 2.4 implies that counterfactual distributions $P\left(\boldsymbol{Y}_{\boldsymbol{x}}, \dots, \boldsymbol{Z}_{\boldsymbol{w}}\right)$ 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 $\boldsymbol{Y}_{\boldsymbol{x}}, \dots, \boldsymbol{Z}_{\boldsymbol{w}}$ be an arbitrary set of counterfactual vari-*

ables. 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:*

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

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* $\{V_{\boldsymbol{x}}(\boldsymbol{u})\mid\forall V\in PA_Y\}$*.*

Proof. The finite-state decomposition of counterfactual distribution $P(\boldsymbol{Y_x},\dots,\boldsymbol{Z_w})$ in Eq. (5) follows immediately from the construction of canonical SCM N in Thm. 2.4. □

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

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

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

Proof. For an arbitrary set $\boldsymbol{C}\subseteq\boldsymbol{V}$, let $\boldsymbol{PA_C}=Pa(\boldsymbol{C})\setminus\boldsymbol{C}$, i.e., the set of all direct parents of nodes in $\boldsymbol{C}$ except themselves. The observational distribution $P(\boldsymbol{V})$ in any causal diagram $\mathcal{G}$ could be decomposed over c-components as follows (Tian & Pearl, 2002, Lem. 1):

$$P(\boldsymbol{V}=\boldsymbol{v})=\prod_{\boldsymbol{C}\in\mathcal{C}(\mathcal{G})}P\left(\boldsymbol{C}_{\boldsymbol{pa_C}}=\boldsymbol{c}\right), \tag{48}$$

where $\mathcal{C}(\mathcal{G})$ denote the set of all c-components in diagram $\mathcal{G}$. Next we could apply a similar discretization procedure in Lem. A.6 to construct a canonical SCM that represents every interventional distribution $P\left(\boldsymbol{C}_{\boldsymbol{pa_C}}\right)$ while maintaining the same causal diagram. The total number of model parameters of $P\left(\boldsymbol{C}_{\boldsymbol{pa_C}}\right)$ is given by $d=\prod_{V\in Pa(\boldsymbol{C}(U))}|\Omega_V|$. Repeatedly discretizing every exogenous variables in $\boldsymbol{U}$ gives the canonical SCM N. □

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

Proof. It follows from the ancestral set factorization in (Correa et al., 2021, Thm. 1) that any interventional distribution $P(\mathbf{Y_x})$ could be written as a function of a collection of counterfactual distributions given by

$$\left\{P\left(\tilde{\boldsymbol{C}}_*\right)\mid\forall\boldsymbol{C}\in\mathcal{C}(\mathcal{G})\right\} \tag{49}$$

where for every c-component $\boldsymbol{C}\in\mathcal{C}(\mathcal{G})$ in diagram $\mathcal{G}$, $\tilde{\boldsymbol{C}}_*$ is a set of counterfactual variables defined as:

$$\tilde{\boldsymbol{C}}_*=\left\{V_{\boldsymbol{pa}_V}\mid\forall V\in\boldsymbol{C},\exists\boldsymbol{pa}_V\in\Omega_{\boldsymbol{PA}_V}\right\}. \tag{50}$$

It is thus sufficient to construct a canonical SCM N to represent counterfactual distributions $P\left(\tilde{\boldsymbol{C}}_*\right)$ in Eq. (49). The number of parameters for each $P\left(\tilde{\boldsymbol{C}}_*\right)$ is bounded by $d=\prod_{V\in\boldsymbol{C}(U)}|\Omega_{\boldsymbol{PA}_V}||\Omega_V|$. Repeatedly discretizing every exogenous variables in $\boldsymbol{U}$ following the procedure in Lem. A.6 gives the canonical SCM N. □

Thm. 2.4 implies that it is sufficient to search over the family of canonical SCMs when bounding counterfactual probabilities in an arbitrary causal diagram, without loss of generality.

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

Proof. It follows from Thm. 2.4 that for any SCM $M\in\mathscr{M}(\mathcal{G})$, there exists a canonical SCM $N\in\mathscr{N}(\mathcal{G})$ such that

$$P_M(\boldsymbol{Y_x},\dots,\boldsymbol{Z_w})=P_N(\boldsymbol{Y_x},\dots,\boldsymbol{Z_w}), \tag{51}$$

and for any $\boldsymbol{z}\in\mathbb{Z}$,

$$P_M(\boldsymbol{V_z})=P_N(\boldsymbol{V_z}). \tag{52}$$

The reverse direction of the above equations also holds since $\mathscr{N}(\mathcal{G})\subset\mathscr{M}(\mathcal{G})$. This means that solutions for optimization problems in Eqs. (2) and (9) must coincide. □

B. Markov Chain Monte Carlo for Partial Counterfactual Identification

In this section, we will show derivations for complete conditional distributions utilized in our proposed Gibbs samplers. We will also provide proofs for non-asymptotic bounds for empirical estimates of credible intervals used in Alg. 1.

B.1. Derivations of Complete Conditionals

Sampling $P\left(\bar{\boldsymbol{u}}\mid\bar{\boldsymbol{v}},\boldsymbol{\theta},\boldsymbol{\mu}\right)$**.** It is verifiable that variables $\boldsymbol{U}^{(n)},\boldsymbol{V}^{(n)}$, $n=1,\dots,N$, are mutually independent given parameters $\boldsymbol{\theta}, \boldsymbol{\mu}$. This implies

$$\begin{aligned} P\left(\bar{\boldsymbol{u}} \mid \bar{\boldsymbol{v}}, \boldsymbol{\theta}, \boldsymbol{\mu}\right) &= \prod_{U\in\boldsymbol{U}} P\left(\boldsymbol{u}^{(n)} \mid \bar{\boldsymbol{v}}, \boldsymbol{\theta}, \boldsymbol{\mu}\right) \\ &= \prod_{U\in\boldsymbol{U}} P\left(\boldsymbol{u}^{(n)} \mid \boldsymbol{v}^{(n)}, \boldsymbol{\theta}, \boldsymbol{\mu}\right) \end{aligned}$$

The complete conditional over $\left(\boldsymbol{U}^{(n)} \mid \boldsymbol{V}^{(n)}, \boldsymbol{\theta}, \boldsymbol{\mu}\right)$, $n = 1, \dots, N$, is given by

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

Among quantities in the above equation,

$$P\left(v^{(n)} \mid pa_V^{(n)}, u_V^{(n)}, \boldsymbol{\theta}, \boldsymbol{\mu}\right) = \mu_{v^{(n)}}^{\left(pa_V^{(n)}, u^{(n)}\right)},$$

and

$$P\left(u_V^{(n)} \mid \boldsymbol{\theta}, \boldsymbol{\mu}\right) = \theta_u \;\text{ for }\; u = u_V^{(n)}.$$

Sampling $P\left(\boldsymbol{\mu}, \boldsymbol{\theta} \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{u}}\right)$. For every exogenous variable $U \in \boldsymbol{U}$, we denote by $\boldsymbol{\theta}_U$ the set of parameters $\{\theta_u \mid \forall u\}$. Similarly, for every endogenous variable $V \in \boldsymbol{V}$, let $\boldsymbol{\mu}_V = \left\{\mu_V^{(pa_V, u_V)} \mid \forall pa_V, u_V\right\}$. Obviously, parameters $\boldsymbol{\mu}_V$ and $\boldsymbol{\theta}_U$ are mutually independent, and they do not directly determine values of a variable (exogenous or endogenous) simultaneously. We must have

$$P\left(\boldsymbol{\mu}, \boldsymbol{\theta} \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{u}}\right) = \prod_{V\in\boldsymbol{V}} P\left(\mu_V \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{u}}\right) \prod_{U\in\boldsymbol{U}} P\left(\theta_U \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{u}}\right).$$

The above independence relationship implies that to draw samples from the posterior $P\left(\boldsymbol{\mu}, \boldsymbol{\theta} \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{u}}\right)$, we could sample distributions over $\left(\mu_V \mid \bar{\boldsymbol{V}}, \bar{\boldsymbol{U}}\right)$ and $\left(\theta_U \mid \bar{\boldsymbol{V}}, \bar{\boldsymbol{U}}\right)$ for every $V \in \boldsymbol{V}$ and every $U \in \boldsymbol{U}$ separately.

Recall that for every $V \in \boldsymbol{V}$, any pa_V, u_V, $\mu_V^{(pa_V, u_V)} = \left(\mu_v^{(pa_V, u_V)} \mid \forall v \in \Omega_V\right)$ is an indicator vector such that

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

The complete conditional distribution over $\left(\boldsymbol{\mu}_V \mid \bar{\boldsymbol{V}}, \bar{\boldsymbol{U}}\right)$, given by Eq. (15), follows from the fact that in any discrete SCM, the n-th observation of $V \in \boldsymbol{V} \setminus \boldsymbol{Z}^{(n)}$ is decided by

$$v^{(n)} \leftarrow f_V\left(pa_V^{(n)}, u_V^{(n)}\right) = v,$$

where v is a unique element in Ω_V such that $\mu_v^{(pa_V, u_V)} = 1$.

The complete conditional distribution over $\left(\theta_U \mid \bar{\boldsymbol{V}}, \bar{\boldsymbol{U}}\right)$, given by Eq. (16), follows from the conjugacy of Dirichlet distributions with regard to categorical distributions (e.g., see (Ishwaran & James, 2001, Sec. 5.2)).

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

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

Among quantities in the above equation, by expanding on valus of parameters $\mu_V^{(pa_V, u_V)}$, one could rewrite the posterior distribution $P\left(v^{(n)} \mid pa_V^{(n)}, u_V^{(n)}, \bar{\boldsymbol{v}}_{-n}, \bar{\boldsymbol{u}}_{-n}\right)$ for every $V \in \boldsymbol{V} \setminus \boldsymbol{Z}^{(n)}$ as follows

$$\begin{aligned} &P\left(v^{(n)} \mid pa_V^{(n)}, u_V^{(n)}, \bar{\boldsymbol{v}}_{-n}, \bar{\boldsymbol{u}}_{-n}\right) \\ &= \sum_{pa_V, u_V} \sum_{\mu_V^{(pa_V, u_V)}} \mu_{v^{(n)}}^{(pa_V, u_V)} \mathbb{1}_{pa_V = pa_V^{(n)}} \mathbb{1}_{u_V = u_V^{(n)}} \\ &\cdot P\left(\mu_V^{(pa_V, u_V)} \mid \bar{\boldsymbol{v}}_{-n}, \bar{\boldsymbol{u}}_{-n}\right). \end{aligned} \tag{53}$$

The complete conditional over $\left(\mu_V^{(pa_V, u_V)} \mid \bar{\boldsymbol{V}}_{-n}, \bar{\boldsymbol{V}}_{-n}\right)$, $\forall pa_V, u_V$, follows from the definition of discrete SCMs. The n-th observation of $V \in \boldsymbol{V} \setminus \boldsymbol{Z}^{(n)}$ is decided by

$$v^{(n)} \leftarrow f_V\left(pa_V^{(n)}, u_V^{(n)}\right) = v,$$

for a unique $v \in \Omega_V$ such that $\mu_v^{(pa_V, u_V)} = 1$. Formally, if there exists a sample $i \neq n$ such that $V \not\in \boldsymbol{Z}^{(i)}$ and $pa_V^{(i)} = pa_V, u_V^{(i)} = u_V$, the posterior over $\mu_V^{(pa_V, u_V)}$ is given by

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

Otherwise,

$$P\left(\mu_V^{(pa_V, u_V)} \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{u}}\right) = \frac{1}{|\Omega_V|}.$$

Marginalizing probabilities $P\left(\mu_V^{(pa_V, u_V)} \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{u}}\right)$ over the domain Ω_V in Eq. (53) gives the complete conditional distribution over $\left(V^{(n)} \mid PA_V^{(n)}, U_V^{(n)}, \bar{\boldsymbol{U}}_{-n}, \bar{\boldsymbol{U}}_{-n}\right)$.

For every $U \in \boldsymbol{U}$, the complete conditional over $\left(U^{(n)} \mid \bar{\boldsymbol{V}}_{-n}, \bar{\boldsymbol{U}}_{-n}\right)$, given by Eq. (16), follows immediately from the Pólya urn characterization of Dirichlet distributions (e.g., see (Ishwaran & James, 2001, Sec. 4)).

B.2. Monte Carlo Estimation of Credible Intervals

Next we show the efficacy of the Bayesian approach in approximating the optimal bounds over unknown counterfactual probabilities from the observational and experimental data. We will show that 100% credible interval converges to the optimal bound as the total number of data increases (to infinite), provided with "proper" prior distributions over model parameters (to be defined).

We first introduce some necessary notations. Let $\boldsymbol{\Theta}$ denote the collection of parameters $\boldsymbol{\theta}, \boldsymbol{\mu}$ of canonical SCMs in the family $\mathscr{N}(\mathcal{G})$ that generate interventional distributions $P(\boldsymbol{V_z})$ for every $\boldsymbol{z} \in \mathbb{Z}$. Formally,

$$\boldsymbol{\Theta} = \Big\{ (\boldsymbol{\theta}_N, \boldsymbol{\mu}_N) \,|\forall N \in \mathscr{N}(\mathcal{G}), \\ P_N(\boldsymbol{V_z}) = P(\boldsymbol{V_z}), \forall \boldsymbol{z} \in \mathbb{Z} \Big\}. \tag{54}$$

We assume that the parameter space $\boldsymbol{\Theta}$ has positive probability with regard to the prior distribution ρ, i.e.,

$$P\left(\boldsymbol{\theta}, \boldsymbol{\mu} \in \boldsymbol{\Theta}\right) = \int_{\boldsymbol{\Theta}} \rho(\boldsymbol{\theta})\rho(\boldsymbol{\mu}) d\boldsymbol{\theta} d\boldsymbol{\mu} > 0. \tag{55}$$

We also assume that prior ρ has full support over domains of $\boldsymbol{\theta}$ and $\boldsymbol{\mu}$. That is, the probability density function $\rho\left(\boldsymbol{\theta}\right) > 0$ and $\rho\left(\boldsymbol{\mu}\right) > 0$ for every possible realization of $\boldsymbol{\theta}, \boldsymbol{\mu}$.

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 $\theta_{\text{ctf}} = P\left(\boldsymbol{y_x}, \dots, \boldsymbol{z_w}\right)$, and let $[l, r]$ be the optimal bound over $P\left(\boldsymbol{y_x}, \dots, \boldsymbol{z_w}\right)$ 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 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 $\boldsymbol{z} \in \mathbb{Z}$.*

Proof. By the definition of Eq. (54), for every pair of parameter $(\boldsymbol{\theta}, \boldsymbol{\mu}) \in \boldsymbol{\Theta}$, it must be compatible with the dataset $\bar{\boldsymbol{v}}$, i.e., $P\left(\bar{\boldsymbol{v}} \mid \boldsymbol{\theta}, \boldsymbol{\mu}\right) > 0$. Let $\lambda = \inf_{\boldsymbol{\Theta}} P\left(\bar{\boldsymbol{v}} \mid \boldsymbol{\theta}, \boldsymbol{\mu}\right) > 0$. The posterior distribution over $(\boldsymbol{\theta}, \boldsymbol{\mu}) \in \boldsymbol{\Theta}$ given finite samples $\bar{\boldsymbol{v}}$ could thus be written as

$$\begin{aligned} P\left(\boldsymbol{\theta}, \boldsymbol{\mu} \in \boldsymbol{\Theta} \mid \bar{\boldsymbol{v}}\right) &= \alpha P\left(\boldsymbol{\theta}, \boldsymbol{\mu} \in \boldsymbol{\Theta}, \bar{\boldsymbol{v}}\right) \\ &= \alpha \int_{\boldsymbol{\Theta}} P\left(\bar{\boldsymbol{v}} \mid \boldsymbol{\theta}, \boldsymbol{\mu}\right) \rho(\boldsymbol{\theta})\rho(\boldsymbol{\mu}) d\boldsymbol{\theta} d\boldsymbol{\mu} \\ &\geq \alpha\lambda \int_{\boldsymbol{\Theta}} \rho(\boldsymbol{\theta})\rho(\boldsymbol{\mu}) d\boldsymbol{\theta} d\boldsymbol{\mu} \\ &\geq \alpha\lambda P\left(\boldsymbol{\theta}, \boldsymbol{\mu} \in \boldsymbol{\Theta}\right) > 0 \end{aligned}$$

where $\alpha > 0$ is a normalizing constant. Note that parameters $(\boldsymbol{\theta}, \boldsymbol{\mu}) \in \boldsymbol{\Theta}$ fully determine the optimal bound $\theta_{\text{ctf}} \in [l, r]$. The above equation implies that

$$P\left(\theta_{\text{ctf}} \in [l, r] \mid \bar{\boldsymbol{v}}\right) > 0. \tag{56}$$

Since the prior ρ has full support over domains of $\boldsymbol{\theta}, \boldsymbol{\mu}$, it follows that the 100% credible interval $[l_0, r_0]$ given $\bar{\boldsymbol{v}}$ must contain the optimal bound $[l, r]$.

We next show that the credible interval $[l_0, r_0]$ converges to the optimal bound $[l, r]$ when the sample size $N_{\boldsymbol{z}} \to \infty$ for every $\boldsymbol{z} \in \mathbb{Z}$. Let $\theta_{\boldsymbol{z}}$ denote probabilities of $P(\boldsymbol{V_z})$ computed from parameters $\boldsymbol{\theta}, \boldsymbol{\mu}$. By the Bayesian law of large numbers (Grendár & Judge, 2009), we must have, when the number of samples $N_{\boldsymbol{z}} \to \infty, \forall \boldsymbol{z} \in \mathbb{Z}$,

$$P\left(\theta_{ctf} \mid \bar{\boldsymbol{v}}\right) \xrightarrow{a.s} P\left(\theta_{ctf} \mid \boldsymbol{\theta}, \boldsymbol{\mu} \in \boldsymbol{\Theta}\right) \tag{57}$$

The above equation, together with the definition of optimal bounds $[l, r]$ in Eq. (2), implies that

$$P\left(\theta_{ctf} \in [l, r] \mid \bar{\boldsymbol{v}}\right) \xrightarrow{a.s} 1, \quad \text{when } N_{\boldsymbol{z}} \to \infty, \forall \boldsymbol{z} \in \mathbb{Z}.$$

That is, the 100% credible interval $[l_0, r_0]$ converges to the optimal bound $[l, r]$ as the sample size $N_{\boldsymbol{z}}$ gets larger for every intervention do$(\boldsymbol{z})$ in the collection $\mathbb{Z}$. □

Recall that for samples $\left\{\theta^{(t)}\right\}_{t=1}^{T}$ drawn from $P\left(\theta_{\text{ctf}} \mid \bar{\boldsymbol{v}}\right)$, the empirical estimates for $100(1-\alpha)\%$ credible interval over θ_{ctf} are defined as:

$$\hat{l}_\alpha(T) = \theta^{(\lceil(\alpha/2)T\rceil)}, \quad \hat{r}_\alpha(T) = \theta^{(\lceil(1-\alpha/2)T\rceil)}, \tag{58}$$

where $\theta^{(\lceil(\alpha/2)T\rceil)}, \theta^{(\lceil(1-\alpha/2)T\rceil)}$ are the $\lceil(\alpha/2)T\rceil$th smallest and the $\lceil(1-\alpha/2)T\rceil$th smallest of $\left\{\theta^{(t)}\right\}$. One could apply standard concentration inequalities to determine a sufficient number of draws T required for obtaining accurate estimates of a $100(1-\alpha)\%$ credible interval.

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

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

Proof. Fix $\epsilon > 0$. If $\hat{l}_\alpha(T) > l_{\alpha+\epsilon}$, this means that there are at most $\lceil(\alpha/2)T\rceil - 1$ instances in $\left\{\theta^{(t)}_{\text{ctf}}\right\}_{t=1}^{T}$ that are smaller than or equal to $l_{\alpha+\epsilon}$. That is,

$$\begin{aligned} &P\left(\hat{l}_\alpha(T) > l_{\alpha+\epsilon}\right) \\ &\leq P\left(\sum_{t=1}^{T} \mathbb{1}_{\theta^{(t)}_{\text{ctf}} \leq l_{\alpha+\epsilon}} \leq \lceil(\alpha/2)T\rceil - 1\right) \\ &\leq P\left(\sum_{t=1}^{T} \mathbb{1}_{\theta^{(t)}_{\text{ctf}} \leq l_{\alpha+\epsilon}} \leq (\alpha/2)T\right) \\ &\leq P\left(\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{\theta^{(t)}_{\text{ctf}} \leq l_{\alpha+\epsilon}} \leq \frac{\alpha+\epsilon}{2} - \frac{\epsilon}{2}\right) \\ &\leq \exp\left(-\frac{T\epsilon^2}{2}\right). \end{aligned}$$

The last step in the above equation follows from the standard Hoeffding's inequality.

If $\hat{l}_\alpha(T) < l_{\alpha-\epsilon}$, this implies that there are at least $\lceil(\alpha/2)T\rceil$ instances in $\left\{\theta^{(t)}_{\text{ctf}}\right\}_{t=1}^{T}$ that are larger than or equal to $l_{\alpha+\epsilon}$. That is,

$$\begin{aligned}
&P\left(\hat{l}_\alpha(T) < l_{\alpha-\epsilon}\right)\\
&\leq P\left(\sum_{t=1}^{T} \mathbb{1}_{\theta^{(t)}_{\text{ctf}} \leq l_{\alpha-\epsilon}} \geq \lceil(\alpha/2)T\rceil\right)\\
&\leq P\left(\sum_{t=1}^{T} \mathbb{1}_{\theta^{(t)}_{\text{ctf}} \leq l_{\alpha-\epsilon}} \geq (\alpha/2)T\right)\\
&\leq P\left(\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{\theta^{(t)}_{\text{ctf}} \leq l_{\alpha-\epsilon}} \geq \frac{\alpha-\epsilon}{2} + \frac{\epsilon}{2}\right)\\
&\leq \exp\left(-\frac{T\epsilon^2}{2}\right).
\end{aligned}$$

The last step follows from the standard Hoeffding's inequality. Similarly, we could also show that

$$\begin{aligned}
P\left(\hat{h}_\alpha(T) < h_{\alpha+\epsilon}\right) \leq \exp\left(-\frac{T\epsilon^2}{2}\right),\\
P\left(\hat{h}_\alpha(T) > h_{\alpha-\epsilon}\right) \leq \exp\left(-\frac{T\epsilon^2}{2}\right).
\end{aligned}$$

Finally, bounding the error rate by $\delta/4$ gives:

$$\exp\left(-\frac{T\epsilon^2}{2}\right) = \frac{\delta}{4} \Rightarrow \epsilon = \sqrt{2T^{-1}\ln(4/\delta)}. \tag{59}$$

Replacing the error rate ϵ with $f(T,\delta) = \sqrt{2T^{-1}\ln(4/\delta)}$ completes the proof. □

As a corollary, it immediately follows from Lem. 3.3 that Algorithm CREDIBLEINTERVAL (Alg. 1) is guaranteed to from a sufficient estimate of $100(1-\alpha)\%$ credible intervals within the specified margin of errors.

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

Proof. The statement follows immediately from Lem. 3.3 by setting $\sqrt{2T^{-1}\ln(4/\delta)} \leq \epsilon$. □

C. Simulation Setups and Additional Experiments

In this section, we will provide details on the simulation setups and preprocessing of datasets. We also conduct additional experiments on other more involved causal diagrams and using skewed hyperparameters for prior distributions. For all experiments, we will focus on Dirichlet priors in Eq. (12) with hyperparameters $\alpha^{(u)}_U = \alpha_U/d_U$ for some real $\alpha_U > 0$. This is equivalent to drawing probabilities θ_u from a Dirichlet distribution defined as follows:

$$(\theta_1,\dots,\theta_{d_U}) \sim \texttt{Dirichlet}\left(\frac{\alpha_U}{d_U},\cdots,\frac{\alpha_U}{d_U}\right), \tag{60}$$

All experiments were performed on a computer with 32GB memory, implemented in MATLAB. We are in the process of migrating the source code to other open-source platforms (e.g., Julia). We will release them once the code migration is done and the manuscript is accepted.

Experiment 1: Frontdoor We study the problem of evaluating interventional probabilities $P(y_x)$ from the observational distribution $P(X,Y,W)$ in the "Frontdoor" diagram of Fig. 1c. We collect $N = 10^4$ samples $\bar{\boldsymbol{v}} = \{x^{(n)}, y^{(n)}, w^{(n)}\}_{n=1}^{N}$ from an SCM compatible with Fig. 1c. Detailed parametrization of the SCM is provided in the following:

$$\begin{aligned}
&U_1 \sim \texttt{Unif}(0,1),\\
&U_2 \sim \texttt{Normal}(0,1),\\
&X \sim \texttt{Binomial}(1,\rho_X),\\
&W \sim \texttt{Binomial}(1,\rho_W),\\
&Y \sim \texttt{Binomial}(1,\rho_Y),
\end{aligned} \tag{61}$$

where probabilities ρ_X, ρ_W, ρ_Y are given by

$$\begin{aligned}
\rho_X &= U_1,\\
\rho_W &= \frac{1}{1+\exp(-X-U_2)},\\
\rho_Y &= \frac{1}{1+\exp(W-U_1)}.
\end{aligned}$$

Each observation $\left(x^{(n)}, y^{(n)}, w^{(n)}\right)$ is an independent draw from the observational distribution $P(X,Y,W)$. We set hyperparameters $\alpha_{U_1} = d_{U_1} = 8$, $\alpha_{U_1} = d_{U_2} = 4$.

Experiment 2: PNS We study the problem of evaluating the counterfactual probability $P(y_x, y'_{x'}) \equiv P(Y_x = y, Y_{x'} = y')$ for any $x \neq x', y \neq y'$ from the observational distribution $P(X,Y)$ in the "Bow" diagram of Fig. 1d. We collect $N = 10^3$ observational samples $\bar{\boldsymbol{v}} = \{x^{(n)}, y^{(n)}\}_{n=1}^{N}$ from an SCM compatible with Fig. 1d. Detailed parametrization of the SCM is defined as follows:

$$\begin{aligned}
&U \sim \texttt{Normal}(0,1),\\
&X \sim \texttt{Binomial}(1,\rho_X),\\
&E \sim \texttt{Logistic}(0,1),\\
&Y \leftarrow \mathbb{1}_{X-U+E+0.1>0},
\end{aligned} \tag{62}$$

where probabilities ρ_X are given by

$$\rho_X = \frac{1}{1 + \exp(U)}.$$

Each observation $\left(x^{(n)}, y^{(n)}\right)$ is an independent draw from the observational distribution $P(X, Y)$. In this experiment, we set hyperparameters $\alpha_U = d_U = 8$.

Experiment 3: IST International Stroke Trials (IST) was a large, randomized, open trial of up to 14 days of antithrombotic therapy after stroke onset (Carolei et al., 1997). The aim was to provide reliable evidence on the efficacy of aspirin and of heparin. The dataset is released under Open Data Commons Attribution License (ODC-By). In particular, the treatment X is a pair (i, j) where $i = 0$ stands for no aspirin allocation, 1 otherwise; $j = 0$ stands for no heparin allocation, 1 for median-dosage, and 2 for high-dosage. 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 with selection rules $f_X(\boldsymbol{u})$, following a procedure introduced in (Zhang & Bareinboim, 2021). More specifically, we are provided with a collection of IST samples $\left\{x^{(n)}, y^{(n)}, u^{(n)}\right\}_{n=1}^{N}$ where $u^{(n)}$ is the age of the n-th patient. For each data point $\left(x^{(n)}, y^{(n)}, u^{(n)}\right)$, we check if values of $x^{(n)}$ satisfy the following condition

$$x^{(n)} = f_X\left(u^n\right) = \left\lfloor 6 \times \left(\frac{u^{(n)}}{100}\right)^2 \right\rfloor. \tag{63}$$

If the above condition is satisfied, we keep the data point $\left(x^{(n)}, y^{(n)}, u^{(n)}\right)$ in the dataset; otherwise, the data point is dropped. After this data selection process is complete, we hide columns of variables $u^{(n)}$. Doing so allows us to obtain $N = 1 \times 10^3$ synthetic observational samples $\bar{\boldsymbol{v}} = \left\{x^{(n)}, y^{(n)}\right\}_{n=1}^{N}$ that are compatible with the "Bow" diagram in Fig. 1d.

In this experiment, we set hyperparameters $\alpha_{U_1} = 10$ and $\alpha_{U_2} = 1$. As a baseline, we estimate the actual treatment effect $P\left(Y_{x=(1,0)} \geq 2\right) = 0.3775$ for only assigning aspirin $X = (1, 0)$ on the recovery of patients $Y \geq 2$ from randomized trial data containing 1.9285×10^4 subjects.

Experiment 4: Non-IV We study the problem of evaluating counterfactual probabilities $P(z, x_{z'}, y_{x'})$ from the combination of the observational distribution $P(X, Y, Z)$ and interventional distributions $P(X_z, Y_z)$, $\forall z \in \Omega_Z$, in the causal diagram of Fig. 1b. We collect $N = 10^3$ samples $\bar{\boldsymbol{v}} = \{x^{(n)}, y^{(n)}, z^{(n)}\}_{n=1}^{N}$ from an SCM compatible with Fig. 1b, which we define as follows:

$$\begin{aligned}
&U_1 \sim \texttt{Unif}(0, 1),\\
&U_2 \sim \texttt{Unif}(0, 1),\\
&Z \leftarrow \min\left\{\lfloor 15 \cdot U_1 \rfloor, 9\right\},\\
&X \sim \texttt{Binomial}(9, \rho_X),\\
&Y \sim \texttt{Binomial}(9, \rho_Y),
\end{aligned} \tag{64}$$

where for any real $\alpha \in \mathbb{R}$, the operator $\lfloor \alpha \rfloor$ denotes the largest integer $n \in \mathbb{Z}$ smaller than α, i.e., $\lfloor \alpha \rfloor = \min\{n \in \mathbb{Z} \mid n \geq \alpha\}$; probabilities ρ_X, ρ_Y are given by

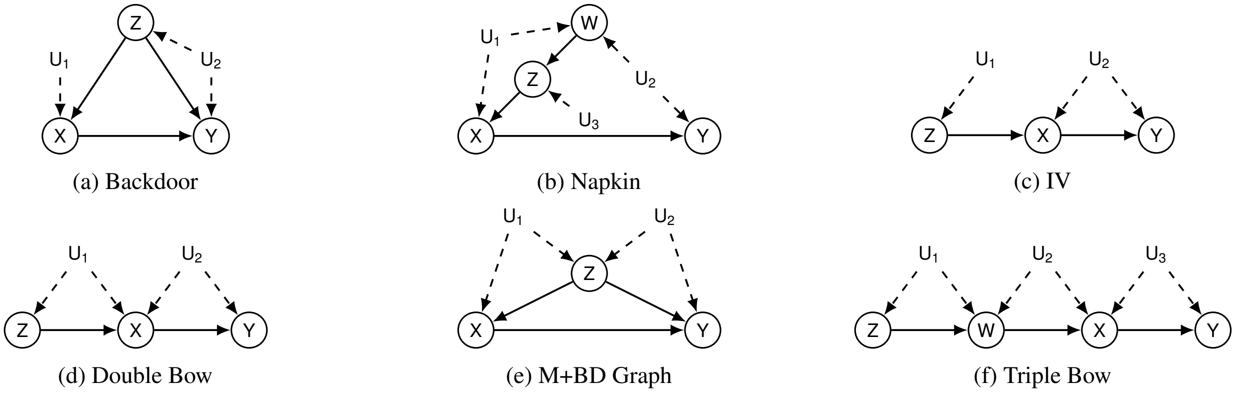
$$\begin{aligned}
\rho_X &= \frac{1}{1 + \exp(-Z - U_2)},\\
\rho_Y &= \frac{1}{1 + \exp(X/10 - U_1 \cdot U_2)}.
\end{aligned}$$

Each sample $\left(x^{(n)}, y^{(n)}, z^{(n)}\right)$ is an independent draw from the observational distribution $P(X, Y, Z)$ or an interventional distribution $P(X_z, Y_z)$. To obtain a sample from $P(X_z, Y_z)$, we pick a constant $z \in \Omega_Z$ uniformly at random, perform intervention do$(Z = z)$ in the SCM described in Eq. (64) and observed subsequent outcomes. In this experiment, we set hyperparameters $\alpha_{U_1} = 10$ and $\alpha_{U_2} = 10$.

C.1. Additional Simulation Results

We also evaluate our algorithms on various simulated SCM instances in other more involved causal diagrams. Overall, we found that simulation results match our findings in the main manuscript. For identifiable settings (Experiments 5 & 6), our algorithms are able to recover the actual, unknown counterfactual probabilities. For non-identifiable settings, our algorithm consistently dominates existing bounding strategies: it achieves sharp bounds if closed-formed solutions exist (Experiments 7 & 8); otherwise, it improves over state-of-art bounds (Experiment 9). Finally, for other more challenging non-identifiable settings where existing strategies do not apply (Experiment 10), our algorithm is able to achieve effective bounds over unknown counterfactual probabilities.

Experiment 5: Backdoor Consider the "Backdoor" graph described in Fig. 5a where X, Y, Z are binary variables in $\{0, 1\}$; $U_1, U_2 \in \mathbb{R}$. In this case, any interventional probability $P(Y_x = y)$ is identifiable from the observational distribution $P(X, Z, Y)$ through the backdoor criterion (Pearl, 2000, Def. 3.3.1). We collect $N = 10^4$ observational samples $\bar{\boldsymbol{v}} = \{x^{(n)}, y^{(n)}, z^{(n)}\}_{n=1}^{N}$ from a synthetic SCM instance compatible with Fig. 5a. Detailed



(a) Backdoor

(b) Napkin

(c) IV

(d) Double Bow

(e) M+BD Graph

(f) Triple Bow

Figure 5: Causal diagrams for additional experiments. Each diagram contains (not exclusively) a treatment X, an outcome Y, ancestors Z, W, and exogenous variables U_i, $i = 1, 2, 3$.

parametrization of the SCM is provided in the following:

$$\begin{aligned} &U_1 \sim \texttt{Unif}(0,1), \\ &U_2 \sim \texttt{Unif}(0,1), \\ &Z \sim \texttt{Binomial}(1, \rho_Z), \\ &X \sim \texttt{Binomial}(1, \rho_X), \\ &Y \sim \texttt{Binomial}(1, \rho_Y), \end{aligned} \tag{65}$$

where probabilities ρ_Z, ρ_X, ρ_Y are given by

$$\begin{aligned} \rho_Z &= U_2, \\ \rho_X &= \frac{1}{1 + \exp(-Z - U_1)}, \\ \rho_Y &= \frac{1}{1 + \exp(X + Z + U_2 + 1)}. \end{aligned}$$

Each observation $\left(x^{(n)}, y^{(n)}, z^{(n)}\right)$ is an independent draw from the observational distribution $P(X, Y, Z)$. In this experiment, we set hyperparameters $\alpha_{U_1} = d_{U_1} = 8$, $\alpha_{U_1} = d_{U_2} = 4$. Fig. 6a 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 interventional probability $P(Y_{x=0} = 1) = 0.1401$, which confirms the identifiability of $P(Y_x = y)$ in the "backdoor" graph.

Experiment 6: Napkin Graph Consider the "Napkin" graph in Fig. 5b where X, Y, Z, W are binary variables in $\{0, 1\}$; U_1, U_2, U_3 take values in real $\mathbb{R}$. Interventional probabilities $P(y_x)$ is identifiable from $P(X, Y, Z, W)$ by iteratively applying inference rules of "do-calculus" (Pearl, 2000, Thm. 4.3.1). We collect $N = 10^4$ observational samples $\bar{\boldsymbol{v}} = \{x^{(n)}, y^{(n)}, z^{(n)}, w^{(n)}\}_{n=1}^N$ from an SCM compatible with Fig. 5b, defined as follows:

$$\begin{aligned} &U_i \sim \texttt{Normal}(0,1), \quad i = 1, 2, 3, \\ &W \sim \texttt{Binomial}(1, \rho_W), \\ &Z \sim \texttt{Binomial}(1, \rho_Z), \\ &X \sim \texttt{Binomial}(1, \rho_X), \\ &Y \sim \texttt{Binomial}(1, \rho_Y), \end{aligned} \tag{66}$$

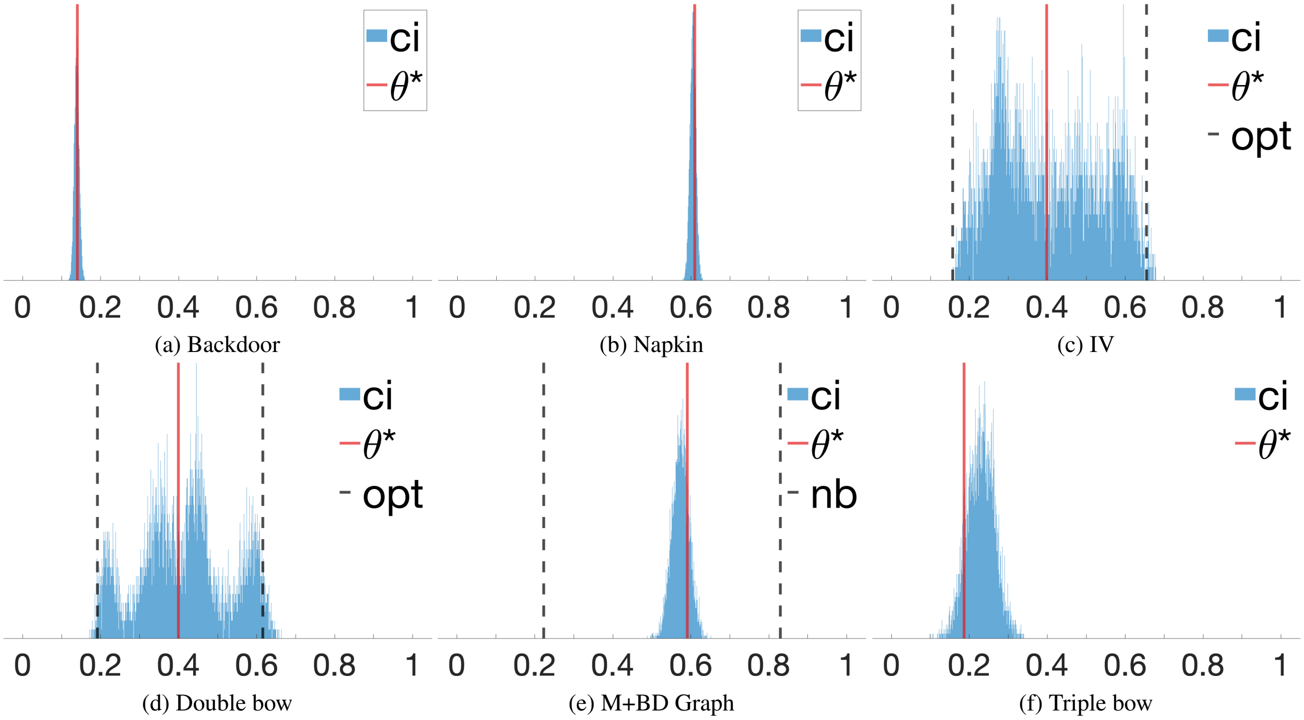
where probabilities $\rho_W, \rho_Z, \rho_X, \rho_Y$ are given by:

$$\begin{aligned} \rho_W &= \frac{1}{1 + \exp(U_1 - U_2)}, \\ \rho_Z &= \frac{1}{1 + \exp(W - U_3)}, \\ \rho_X &= \frac{1}{1 + \exp(-Z - U_1)}, \\ \rho_Y &= \frac{1}{1 + \exp(X - U_2 - 0.5)}. \end{aligned}$$

Each observation $\left(x^{(n)}, y^{(n)}, z^{(n)}, w^{(n)}\right)$ is an independent draw from the observational distribution $P(X, Y, Z, W)$.

In this experiment, we set hyperparameters $\alpha_{U_1} = d_{U_1} = 32$, $\alpha_{U_2} = d_{U_1} = 32$, and $\alpha_{U_3} = d_{U_3} = 4$. Fig. 6b shows a histogram containing samples drawn from the posterior distribution of $(P(Y_{x=0} = 1) \mid \bar{\boldsymbol{v}})$. Our analysis reveals that these samples converges to the actual interventional probability $P(Y_{x=0} = 1) = 0.6098$, which confirms the identifiability of $P(y_x)$ in the napkin graph.

Experiment 7: IV Consider the "IV" diagram in Fig. 5c where X, Y, Z are binary variables taking values in $\{0, 1\}$. The non-identifiability of $P(Y_x = y)$ from the instrumental variable Z and the unobserved confounding between X and Y has been shown in (Bareinboim & Pearl, 2012; Lee et al., 2019). We study the problem of bounding interventional probabilities $P(y_x)$ from the observational distribution $P(X, Y, Z)$. We collect $N = 10^3$ observational



(a) Backdoor

(b) Napkin

(c) IV

(d) Double bow

(e) M+BD Graph

(f) Triple bow

Figure 6: Histogram plots for samples drawn from the posterior distribution over target counterfactual probabilities. For all plots (a - f), *ci* represents our proposed algorithms; θ^* is the actual counterfactual probability; *opt* is the optimal asymptotic bounds (if exists); *nb* stands for the natural bounds (Manski, 1990).

samples $\bar{\boldsymbol{v}} = \left\{x^{(n)}, y^{(n)}, z^{(n)}\right\}_{n=1}^{N}$ from a synthetic SCM compatible with Fig. 5c. Detailed parametrization of the SCM is provided the following:

$$\begin{aligned}
U_1 &\sim \texttt{Normal}(0,1),\\
U_2 &\sim \texttt{Normal}(0,1),\\
Z &\sim \texttt{Binomial}(1,\rho_Z),\\
X &\sim \texttt{Binomial}(1,\rho_X),\\
Y &\sim \texttt{Binomial}(1,\rho_Y),
\end{aligned} \tag{67}$$

where probabilities ρ_Z, ρ_X, ρ_Y are given by

$$\begin{aligned}
\rho_Z &= \frac{1}{1+\exp(-U_1)},\\
\rho_X &= \frac{1}{1+\exp(-Z-U_2)},\\
\rho_Y &= \frac{1}{1+\exp(X-U_2+0.5)}.
\end{aligned}$$

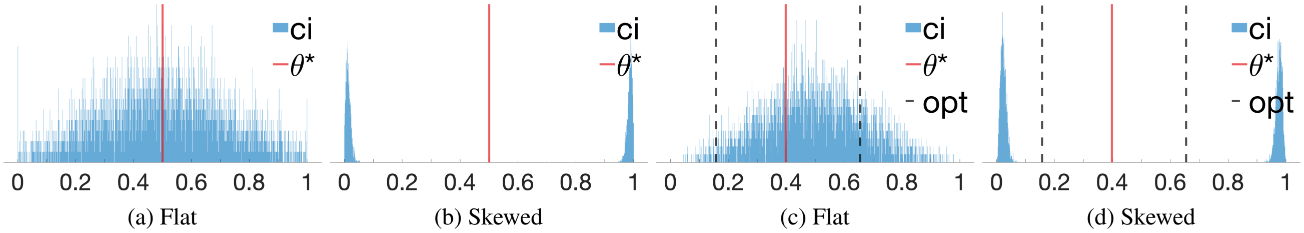
(Balke & Pearl, 1997) introduced a closed-form bound over $P(y_x)$ from the observational distribution $P(X,Y,Z)$ for the "IV" diagram with binary $X,Y,Z \in \{0,1\}$, which is provably optimal (labeled as *opt*). To obtain a 100% credible intervals, we apply the collapsed Gibbs sampler with hyperparameters $\alpha_{U_1} = d_{U_1} = 2$ and $\alpha_{U_2} = d_{U_1} = 16$. Fig. 6c shows samples drawn from the posterior distribution of $(P(Y_{x=0}=1) \mid \bar{\boldsymbol{v}})$. The analysis reveals that our algorithm derives a valid bound over the actual probability $P(Y_{x=0}=1) = 0.3954$; the 100% credible interval converges to the optimal IV bound $l = 0.1980, r = 0.6258$.

Experiment 8: Double Bow Consider the "Double Bow" diagram in Fig. 5d where $X,Y,Z \in \{0,1\}$ and $U_1, U_2 \in \mathbb{R}$. We study the problem of evaluating interventional probabilities $P(y_x)$ from the observational distribution $P(X,Y,Z)$. We collect $N = 10^3$ observational samples $\bar{\boldsymbol{v}} = \{x^{(n)}, y^{(n)}, z^{(n)}\}_{n=1}^{N}$ from an SCM compatible with Fig. 5d, defined as the following:

$$\begin{aligned}
U_i &\sim \texttt{Normal}(0,1), \quad i = 1,2,\\
Z &\sim \texttt{Binomial}(1,\rho_Z),\\
X &\sim \texttt{Binomial}(1,\rho_X),\\
Y &\sim \texttt{Binomial}(1,\rho_Y),
\end{aligned} \tag{68}$$

where probabilities ρ_Z, ρ_X, ρ_Y are given by:

$$\begin{aligned}
\rho_Z &= \frac{1}{1+\exp(-U_1)},\\
\rho_X &= \frac{1}{1+\exp(-Z-U_1-U_2)},\\
\rho_Y &= \frac{1}{1+\exp(X-U_2+0.5)}.
\end{aligned}$$



(a) Flat (b) Skewed (c) Flat (d) Skewed

Figure 7: Prior distributions for (a, b) Experiment 9 and (c, d) Experiment 10.

(Balke & Pearl, 1997) introduced a closed-form bound over $P(y_x)$ from the observational distribution $P(X, Y, Z)$ for the "IV" diagram in Fig. 5c with binary $X, Y, Z \in \{0, 1\}$. It is verifiable that such a bound is also applicable in the "Double-bow" diagram of Fig. 5d with binary endogenous domains, and is provably optimal (labeled as *opt*). To obtain a 100% credible intervals, we apply the collapsed Gibbs sampler with hyperparameters $\alpha_{U_1} = d_{U_1} = 32$ and $\alpha_{U_2} = d_{U_1} = 32$. Fig. 6d shows samples drawn from the posterior distribution of $(P(Y_{x=0} = 1) \mid \bar{\boldsymbol{v}})$. The analysis reveals that our algorithm derives a valid bound over the actual probability $P(Y_{x=0} = 1) = 0.3954$; the 100% credible interval converges to the optimal IV bound $l = 0.1980, r = 0.6258$, confirming the efficacy of the proposed approach.

Experiment 9: M+BD Graph Consider the "M+BD" graph in Fig. 5e where $X, Y, Z \in \{0, 1\}$ and $U_1, U_2 \in \mathbb{R}$. In this case, interventional probabilities $P(y_x)$ are non-identifiable from the observational distribution $P(X, Y, Z)$ due to the presence of the collider path $X \leftarrow U_1 \rightarrow Z \leftarrow U_2 \rightarrow Y$. We collect $N = 10^3$ observational samples $\bar{\boldsymbol{v}} = \{x^{(n)}, y^{(n)}, z^{(n)}\}_{n=1}^{N}$ from an SCM compatible with Fig. 5e. Detailed parametrization of the SCM is given by:

$$\begin{aligned} &U_i \sim \texttt{Normal}(0, 1), \quad i = 1, 2, \\ &Z \sim \texttt{Binomial}(1, \rho_Z), \\ &X \sim \texttt{Binomial}(1, \rho_X), \\ &Y \sim \texttt{Binomial}(1, \rho_Y), \end{aligned} \tag{69}$$

where probabilities ρ_Z, ρ_X, ρ_Y are given by:

$$\begin{aligned} \rho_Z &= \frac{1}{1 + \exp(-U_1)}, \\ \rho_X &= \frac{1}{1 + \exp(-Z - U_1 - U_2)}, \\ \rho_Y &= \frac{1}{1 + \exp(X - Z - U_2)}. \end{aligned}$$

Each observation $\left(x^{(n)}, y^{(n)}, z^{(n)}\right)$ is an independent draw from the observational distribution $P(X, Y, Z)$.

In this experiment, we set hyperparameters $\alpha_{U_1} = d_{U_1} = 32$ and $\alpha_{U_2} = d_{U_1} = 32$. Fig. 6e shows samples drawn from the posterior distribution of $(P(Y_{x=0} = 1) \mid \bar{\boldsymbol{v}})$. As a baseline, we also include the natural bounds introduced in (Robins, 1989; Manski, 1990) (*nb*). The analysis reveals that all algorithms achieve bounds that contain the actual, target causal effect $P(Y_{x=0} = 1) = 0.5910$. Our algorithm obtains a 100% credible interval $l_{ci} = 0.4884, r_{ci} = 0.6519$, which improves over the existing bounding strategy ($l_{nb} = 0.2230, r_{nb} = 0.8296$).

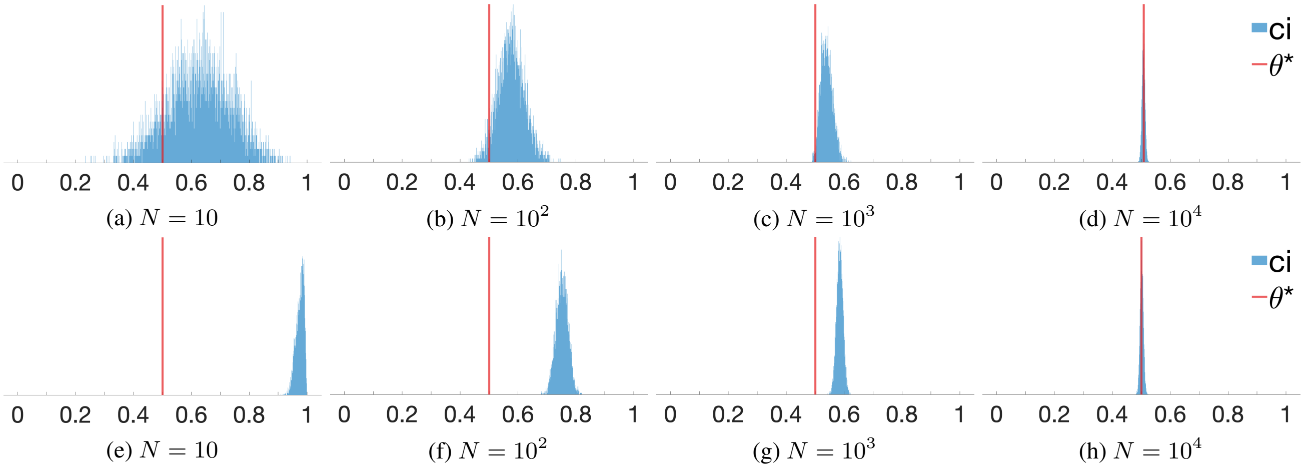
Experiment 10: Triple Bow Consider the "Triple Bow" diagram in Fig. 5f where $X, Y, Z \in \{0, 1\}$ and $U_1, U_2, U_3 \in \mathbb{R}$. We are interested in evaluating the counterfactual probability $P(Y_{x=1} = 1, Y_{x=0} = 0)$ from the combination of the observational distribution $P(X, Y, Z, W)$ and interventional distributions $P(X_z, Y_z, W_z)$. To our best knowledge, existing bounding strategies are not applicable to this setting. We collect $N = 10^3$ samples $\bar{\boldsymbol{v}} = \{x^{(n)}, y^{(n)}, z^{(n)}, w^{(n)}\}_{n=1}^{N}$ from an SCM compatible Fig. 5f. The detailed parametrization of the SCM is provided in the following:

$$\begin{aligned} &U_1 \sim \texttt{Unif}(0, 1), \\ &U_i \sim \texttt{Normal}(0, 1), \quad i = 2, 3, \\ &Z \sim \lfloor 1.5 \cdot U_1 \rfloor, \\ &W \sim \texttt{Binomial}(1, \rho_W), \\ &X \sim \texttt{Binomial}(1, \rho_X), \\ &E \sim \texttt{Logistic}(0, 1), \\ &Y \leftarrow \mathbb{1}_{X - U_3 + E + 0.1 > 0}, \end{aligned} \tag{70}$$

where probabilities ρ_Z, ρ_W, ρ_X are given by:

$$\begin{aligned} \rho_Z &= \frac{1}{1 + \exp(-U_1)}, \\ \rho_W &= \frac{1}{1 + \exp(-Z - U_1 - U_2)}, \\ \rho_X &= \frac{1}{1 + \exp(-W - U_2 - U_3)}. \end{aligned}$$

Each sample $\left(x^{(n)}, y^{(n)}, z^{(n)}, w^{(n)}\right)$ is an independent draw from the observational distribution $P(X, Y, Z, W)$ or an interventional distribution $P(X_z, Y_z, W_z)$. To obtain a sample from $P(x_z, y_z, w_z)$, we pick a constant $z \in \Omega_Z$ uniformly at random, perform intervention $\operatorname{do}(Z = z)$ in the SCM in Eq. (70) and observed subsequent outcomes.



(a) $N = 10$ (b) $N = 10^2$ (c) $N = 10^3$ (d) $N = 10^4$

(e) $N = 10$ (f) $N = 10^2$ (g) $N = 10^3$ (h) $N = 10^4$

Figure 8: Histogram plots for samples drawn from the posterior distribution over probability $P(Y_{x=0} = 0)$ in "Frontdoor" graph of Fig. 1c using two priors. (a - d) shows the posteriors using the flat prior and observational data of size $N = 10, 10^2, 10^3$ and 10^4 respectively; (e - h) shows the posetriors using the skewed prior and the same observational datasets.

In this experiment, we set hyperparameters $\alpha_{U_1} = d_{U_1} = 32$ and $\alpha_{U_2} = d_{U_1} = 32$. Fig. 6f shows samples drawn from the posterior distribution of $(P(Y_{x=0} = 1) \mid \bar{\boldsymbol{v}})$. The analysis reveals that our proposed approach is able to achived an effective bound that contain the actual counterfactual probability $P(Y_{x=1} = 1, Y_{x=0} = 0) = 0.1867$. The 100% credible interval (*ci*) is equal to $l = 0.1150, r = 0.3686$.

C.2. The Effect of Sample Size and Prior Distributions

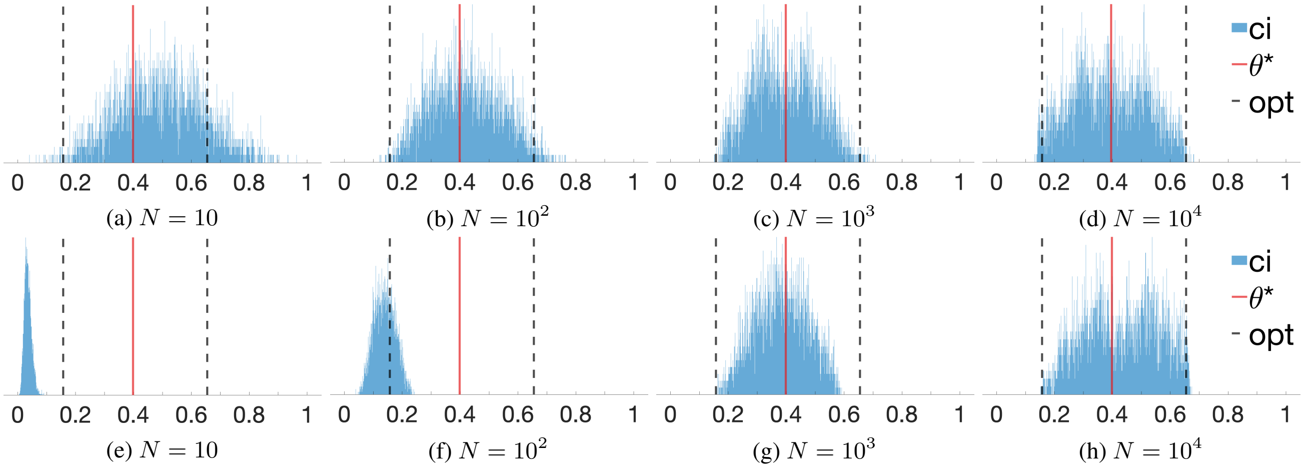
We will evaluate our algorithms using skewed prior distributions. We found that increasing the size of observational samples was able to wash away the bias introduced by prior distributions. That is, despite the influence of prior distributions, our algorithms eventually converge to sharp bounds over unknown counterfactual probabilities as the number of observational sample grows (to infinite).

Experiment 11: Frontdoor Consider first the "Frontdoor" graph in Fig. 1d where interventional probabilities $P(y_x)$ is identifiable from the observational distribution $P(X, Y, W)$. The detailed parametrization of the underlying SCM is described in Eq. (61). We present our results using two different priors. The first is a flat (uniform) distribution over probabilities of U_1 and U_2 respectively, i.e., $\alpha_{U_1} = d_{U_1} = 8$ and $\alpha_{U_1} = d_{U_2} = 4$. The second is skewed to present a strong preference on the deterministic relationships between X and Y; in this case, $\alpha_1 = 300 \times d_{U_i}$, $i = 1, 2$, for prior distributions associated with both U_1 and U_2. Figs. 7a and 7b shows the distribution of $P(Y_{x=0})$ induced by these two priors (in the absence of any observational data). We see that the skewed prior of Fig. 7b assigns almost all weights to deterministic events $P(Y_{x=0} = 1) = 1$ or $P(Y_{x=0} = 0) = 1$.

Fig. 5 shows posterior samples obtained by our Gibbs sampler when applied to observational data of various sizes, using both the flat prior (Figs. 8a to 8d) and the skewed prior (Figs. 8e to 8h). Both priors eventually collapse to the actual, unknown probability $P(Y_{x=0} = 1) = 0.5085$. As expected, more observational data are needed for the skewed prior before the posterior distribution converges, since the skewed prior is concentrated further away from the value 0.5085 than the uniform prior.

Experiment 12: IV Consider the "IV" diagram in Fig. 1a where X, Y, Z are binary variables taking values in $\{0, 1\}$. Detailed parametrization of the underlying SCM is described in Eq. (67). We present our results using two different priors. The first is a flat (uniform) distribution over probabilities of U_1 and U_2 respectively, i.e., $\alpha_{U_1} = d_{U_1} = 2$ and $\alpha_{U_1} = d_{U_2} = 16$. The second is skewed to present a strong preference on the deterministic relationships between X and Y; in this case, $\alpha_1 = 300 \times d_{U_i}$, $i = 1, 2$, for prior distributions associated with both U_1 and U_2. Figs. 7c and 7d shows distributions of $P(Y_{x=0})$ induced by these two prior distributions (in the absence of any observational data). We see that the skewed prior of Fig. 7d assigns almost all weights to deterministic events $P(Y_{x=0} = 1) = 1$ or $P(Y_{x=0} = 0) = 1$.

Fig. 9 shows posterior samples obtained by our Gibbs sampler when applied to observational data of various sizes, using both the flat prior (Figs. 9a to 9d) and the skewed prior (Figs. 9e to 9h). Our analysis reveals that 100% credible intervals of both priors eventually converge to the sharp IV bound $l = 0.1468, r = 0.6617$ over the unknown interventional probability $P(Y_{x=0} = 1) = 0.3954$. It is interesting to note that, in this experiment, while the choice of prior



(a) $N = 10$ (b) $N = 10^2$ (c) $N = 10^3$ (d) $N = 10^4$

(e) $N = 10$ (f) $N = 10^2$ (g) $N = 10^3$ (h) $N = 10^4$

Figure 9: Histogram plots for samples drawn from the posterior distribution over probability $P(Y_{x=0} = 0)$ in "IV" graph of Fig. 1a using two priors. (a - d) shows the posteriors using the flat prior and observational data of size $N = 10, 10^2, 10^3$ and 10^4 respectively; (e - h) shows the posteriors using the skewed prior and the same respective observational datasets.

distribution does not influence the final bound, it still has an effect on the shape of posterior distributions given finite samples of the observational data.

D. Polynomial Optimization for Bounding Counterfactual Probabilities

In this section, we will demonstrate through examples how to translate the optimization problem in Eq. (9) into equivalent polynomial programs in various causal diagrams.

Example 1: IV Consider the "IV" diagram $\mathcal{G}$ in Fig. 1a. We study the problem of bounding counterfactual probabilities $P(y'_{x'}, x, y) \equiv P(Y_{x'} = y', X = x, Y = y)$ from the observational distribution $P(X, Y, Z)$. Formally, let $\mathscr{M}(\mathcal{G})$ denote the set of all SCMs compatible with the diagram $\mathcal{G}$. One could obtain the tight bound over $P(y'_{x'}, x, y)$ from $P(X, Y, Z)$ by solving the optimization problem as follows:

$$\begin{aligned} \min_{M \in \mathscr{M}(\mathcal{G})} / \max \quad & P_M\left(y'_{x'}, x, y\right) \\ \text{s.t.} \quad & P_M(x, y, z) = P(x, y, z), \quad \forall x, y, z. \end{aligned} \tag{71}$$

In the above optimization problem, it follows from Thm. 2.4 that the objective function could be written as

$$\begin{aligned} & P_M(y'_{x'}, x, y) \\ & = \sum_{u_1=1}^{d_1} \sum_{u_2=1}^{d_2} \mu_{y'}^{(x',u_2)} \mu_y^{(x,u_2)} \sum_z \mu_x^{(z,u_2)} \mu_z^{(u_1)} \theta_{u_1} \theta_{u_2}. \end{aligned}$$

Similarly, the observational constraints could be written as:

$$P_M(x, y, z) = \sum_{u_1=1}^{d_1} \sum_{u_2=1}^{d_2} \mu_z^{(u_1)} \mu_x^{(z,u_2)} \mu_y^{(x,u_2)} \theta_{u_1} \theta_{u_2}.$$

The above equations imply that Eq. (71) could be reducible to an equivalent polynomial program as follows:

$$\begin{aligned} \min / \max \quad & \sum_{u_1=1}^{d_1} \sum_{u_2=1}^{d_2} \mu_{y'}^{(x',u_2)} \mu_y^{(x,u_2)} \sum_z \mu_x^{(z,u_2)} \mu_z^{(u_1)} \theta_{u_1} \theta_{u_2} \\ \text{subject to} \quad & \sum_{u_1=1}^{d_1} \sum_{u_2=1}^{d_2} \mu_z^{(u_1)} \mu_x^{(z,u_2)} \mu_y^{(x,u_2)} \theta_{u_1} \theta_{u_2} \\ & = P(x, y, z), \quad \forall x, y, z \\ & \forall z, u_1, \quad \mu_z^{(u_1)}\left(1 - \mu_z^{(u_1)}\right) = 0, \quad \sum_z \mu_z^{(u_1)} = 1 \\ & \forall x, z, u_2, \quad \mu_x^{(z,u_2)}\left(1 - \mu_x^{(z,u_2)}\right) = 0 \\ & \forall x, z, u_2, \quad \sum_x \mu_x^{(z,u_2)} = 1 \\ & \forall y, x, u_2, \quad \mu_y^{(x,u_2)}\left(1 - \mu_y^{(x,u_2)}\right) = 0 \\ & \forall y, x, u_2 \quad \sum_y \mu_y^{(x,u_2)} = 1 \\ & \forall u_1, \quad 0 \leq \theta_{u_1} \leq 1, \quad \sum_{u_1} \theta_{u_1} = 1 \\ & \forall u_2, \quad 0 \leq \theta_{u_2} \leq 1, \quad \sum_{u_2} \theta_{u_2} = 1 \end{aligned}$$

where cardinalities d_1, d_2 are equal to

$$d_1 = |\Omega_Z|, \qquad d_2 = |\Omega_Z \mapsto \Omega_X| \times |\Omega_X \mapsto \Omega_Y|.$$

Example 2 Consider the causal diagram in Fig. 1b. We study the problem of bounding counterfactual probabilities $P(z, x_{z'}, y_{x'})$ from a combination of the observational distribution $P(X, Y, Z)$ and the interventional distribution

$\{P(X_z, Y_z) \mid \forall z \in \Omega_Z\}$. That is,

$$\begin{aligned} \min_{M \in \mathscr{M}(\mathcal{G})} / \max \quad & P_M\left(z, x_{z'}, y_{x'}\right) \\ \text{s.t.} \quad & P_M(x,y,z) = P(x,y,z), \quad \forall x,y,z \\ & P_M(x_z, y_z) = P(x_z, y_z), \quad \forall x,y,z \end{aligned} \tag{72}$$

Among quantities in the above equation, it follows from Thm. 2.4 that the objective function could be written as

$$\begin{aligned} & P_M(z, x_{z'}, y_{x'}) \\ & = \sum_{u_1,u_2=1}^{d} \mu_z^{(u_1)} \mu_x^{(z',u_2)} \mu_y^{(x',u_1,u_2)} \theta_{u_1}\theta_{u_2}. \end{aligned}$$

Similarly, the observational constraints could be written as:

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

and the interventional constraints imply:

$$P_M(x_z, y_z) = \sum_{u_1,u_2=1}^{d} \mu_x^{(z,u_2)} \mu_y^{(x,u_1,u_2)} \theta_{u_1}\theta_{u_2}.$$

The above equations imply that one could obtain an optimal bound over $P(z, x_{z'}, y_{x'})$ given by Eq. (72) by solving an equivalent polynomial program as follows:

$$\begin{aligned} \min / \max \quad & \sum_{u_1,u_2=1}^{d} \mu_z^{(u_1)} \mu_x^{(z',u_2)} \mu_y^{(x',u_1,u_2)} \theta_{u_1}\theta_{u_2} \\ \text{s.t.} \quad & \sum_{u_1,u_2=1}^{d} \mu_z^{(u_1)} \mu_x^{(z,u_2)} \mu_y^{(x,u_1,u_2)} \theta_{u_1}\theta_{u_2} \\ & = P(x,y,z), \quad \forall x,y,z \\ & \sum_{u_1,u_2=1}^{d} \mu_x^{(z,u_2)} \mu_y^{(x,u_1,u_2)} \theta_{u_1}\theta_{u_2} \\ & = P(x_z,y_z), \quad \forall x,y,z \\ & \forall z, u_1, \quad \mu_z^{(u_1)}\left(1 - \mu_z^{(u_1)}\right) = 0 \\ & \forall z, u_1, \quad \sum_z \mu_z^{(u_1)} = 1 \\ & \forall x, z, u_1, u_2, \quad \mu_x^{(z,u_1,u_2)}\left(1 - \mu_x^{(z,u_1,u_2)}\right) = 0 \\ & \forall x, z, u_1, u_2, \quad \sum_x \mu_x^{(z,u_1,u_2)} = 1 \\ & \forall y, x, u_2, \quad \mu_y^{(x,u_2)}\left(1 - \mu_y^{(x,u_2)}\right) = 0 \\ & \forall y, x, u_2, \quad \sum_y \mu_y^{(x,u_2)} = 1 \\ & \forall u_1, \quad 0 \leq \theta_{u_1} \leq 1, \quad \sum_{u_1} \theta_{u_1} = 1 \\ & \forall u_2, \quad 0 \leq \theta_{u_2} \leq 1, \quad \sum_{u_2} \theta_{u_2} = 1 \end{aligned}$$

where the cardinality d equates to

$$d = |\Omega_Z| \times |\Omega_Z \mapsto \Omega_X| \times |\Omega_X \mapsto \Omega_Y|.$$

Example 3: Frontdoor Consider the "Frontdoor" diagram in Fig. 1c. We are interested in evaluating interventional probabilities $P(y_x)$ from the observational distribution $P(X, Y, W)$. That is,

$$\begin{aligned} \min_{M \in \mathscr{M}(\mathcal{G})} / \max \quad & P_M\left(y_x\right) \\ \text{s.t.} \quad & P_M(x,y,w) = P(x,y,w), \quad \forall x,y,w \end{aligned} \tag{73}$$

It follows from Thm. 2.4 that the objective function in the above optimization problem could be further written as

$$P_M(y_x) = \sum_{u_1=1}^{d_1} \sum_{u_1=1}^{d_2} \sum_w \mu_y^{(w,u_1)} \mu_w^{(x,u_2)} \theta_{u_1}\theta_{u_2}.$$

Similarly, the observational constraints could be written as:

$$P_M(x,y,w) = \sum_{u_1=1}^{d_1} \sum_{u_1=1}^{d_2} \mu_x^{(u)} \mu_y^{(w,u_1)} \mu_w^{(x,u_2)} \theta_{u_1}\theta_{u_2}.$$

The above equations imply that one could obtain the optimal solution in Eq. (73) by solving an equivalent polynomial optimization problem defined as follows:

$$\begin{aligned} \min / \max \quad & \sum_{u_1=1}^{d_1} \sum_{u_1=1}^{d_2} \sum_w \mu_y^{(w,u_1)} \mu_w^{(x,u_2)} \theta_{u_1}\theta_{u_2} \\ \text{subject to} \quad & \sum_{u_1=1}^{d_1} \sum_{u_1=1}^{d_2} \mu_x^{(u)} \mu_y^{(w,u_1)} \mu_w^{(x,u_2)} \theta_{u_1}\theta_{u_2} \\ & = P(x,y,w), \forall x,y,w \\ & \forall x, u_1, \quad \mu_x^{(u)}\left(1 - \mu_x^{(u)}\right) = 0, \\ & \forall x, u_1, \quad \sum_x \mu_x^{(u)} = 1, \\ & \forall y, w, u_1, \quad \mu_y^{(w,u_1)}\left(1 - \mu_y^{(w,u_1)}\right) = 0, \\ & \forall y, w, u_1, \quad \sum_y \mu_y^{(w,u_1)} = 1, \\ & \forall w, x, u_2, \quad \mu_w^{(x,u_2)}\left(1 - \mu_w^{(x,u_w)}\right) = 0, \\ & \forall w, x, u_2, \quad \sum_w \mu_w^{(x,u_w)} = 1, \\ & \forall u_1, \quad 0 \leq \theta_{u_1} \leq 1, \quad \sum_{u_1} \theta_{u_1} = 1, \\ & \forall u_2, \quad 0 \leq \theta_{u_2} \leq 1, \quad \sum_{u_2} \theta_{u_2} = 1, \end{aligned}$$

where cardinalities d_1, d_2 equate to

$$d_1 = |\Omega_X| \times |\Omega_W \mapsto \Omega_Y|, \qquad d_2 = |\Omega_X \mapsto \Omega_W|.$$

Example 4: Bow Consider the "Bow" diagram in Fig. 1d. We study the problem of bounding counterfactual probabilities $P(y_x, y'_{x'}) \equiv P\left(Y_x = y, Y_{x=x'} = y'\right)$ from a combination of the observational distribution $P(X, Y, Z)$ and the interventional distribution $\{P(Y_x) \mid \forall x \in \Omega_X\}$, i.e.,

$$\begin{aligned} \min_{M\in\mathscr{M}(\mathcal{G})} / \max \quad & P_M\left(y_x, y'_{x'}\right) \\ \text{s.t.} \quad & P_M(x, y) = P(x, y), \;\; \forall x, y \\ & P_M(y_x) = P(y_x), \;\; \forall x, y \end{aligned} \tag{74}$$

Among quantities in the above equation, it follows from Thm. 2.4 that the objective function could be written as

$$P_M(y_x, y'_{x'}) = \sum_{u=1}^{d} \mu_y^{(x,u)} \mu_{y'}^{(x',u)} \theta_u.$$

Similarly, observational and interventional constraints could be written as the following, respectively,

$$P_M(x, y) = \sum_{u=1}^{d} \mu_x^{(u)} \mu_y^{(x,u)} \theta_u, \quad P_M(y_x) = \sum_{u=1}^{d} \mu_y^{(x,u)} \theta_u.$$

The optimization problem defined in Eq. (74) is thus reducible to an equivalent polynomial program as follows:

$$\begin{aligned} \min / \max \quad & \sum_{u=1}^{d} \mu_y^{(x,u)} \mu_{y'}^{(x',u)} \theta_u \\ \text{subject to} \quad & \sum_{u=1}^{d} \mu_x^{(u)} \mu_y^{(x,u)} \theta_u = P(x, y), \;\; \forall x, y \\ & \sum_{u=1}^{d} \mu_y^{(x,u)} \theta_u = P(y_x), \;\; \forall x, y \\ & \forall x, u, \;\; \mu_x^{(u)}\left(1 - \mu_x^{(u)}\right) = 0 \\ & \forall x, u, \;\; \sum_x \mu_x^{(u)} = 1 \\ & \forall y, x, u, \;\; \mu_y^{(x,u)}\left(1 - \mu_y^{(x,u)}\right) = 0 \\ & \forall y, x, u, \;\; \sum_y \mu_y^{(x,u)} = 1 \\ & \forall u, \;\; 0 \leq \theta_u \leq 1, \;\; \sum_u \theta_u = 1 \end{aligned}$$

where the cardinality d is equal to $|\Omega_Z \mapsto \Omega_X|$.

E. A Generalization of (Balke & Pearl, 1994)

In this section, we will describe a naïve generalization of the discretization procedure introduced in (Balke & Pearl, 1994) to the causal diagram of Fig. 10a. In particular, given any SCM M compatible with Fig. 10a, we will construct a discrete SCM N compatible with a different causal diagram described in Fig. 10b such that M and N coincide in all counterfactual distributions $P(\boldsymbol{Y_x}, \dots, \boldsymbol{Z_w})$.

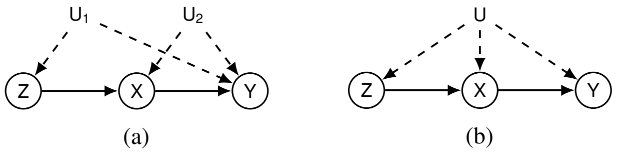


Figure 10: Causal diagrams (a-b) containing a treatment X, an outcome Y, an ancestor Z, and unobserved Us.

We first introduce some useful notations. Let f_Z, f_X, f_Y denote functions associated with Z, X, Y in SCM M. Let constants $h_Z^{(1)} = 0$ and $h_Z^{(2)} = 1$. Note that given any $U_1 = u_1$, $f_Z(u_1)$ must equate to a binary value in $\{0, 1\}$. Therefore, we could define a partition $\mathcal{U}_Z^{(i)}$, $i = 1, 2$, over domains of U_1 such that $u_1 \in \mathcal{U}_Z^{(i)}$ if and only if

$$f_Z(u_1) = h_Z^{(i)}. \tag{75}$$

Given any u_2, $f_X(\cdot, u_2)$ defines a function mapping from domains of Z to X. Let functions in the hypothesis class $\Omega_Z \mapsto \Omega_X$ be ordered by

$$\begin{aligned} h_X^{(1)}(z) &= 0, & h_X^{(2)}(z) &= z, \\ h_X^{(3)}(z) &= \neg z, & h_X^{(4)}(z) &= 1. \end{aligned} \tag{76}$$

Similarly, we could define a partition $\mathcal{U}_X^{(i)}, i = 1, 2, 3, 4$ over the domain Ω_{U_2} such that $u_2 \in \mathcal{U}_X^{(i)}$ if and only if the induced function $f_X(\cdot, u_2) = h_X^{(i)}$. Finally, let functions in $\Omega_X \mapsto \Omega_Y$ be ordered by

$$\begin{aligned} h_Y^{(1)}(x) &= 0, & h_Y^{(2)}(x) &= x, \\ h_Y^{(3)}(x) &= \neg x, & h_Y^{(4)}(x) &= 1. \end{aligned} \tag{77}$$

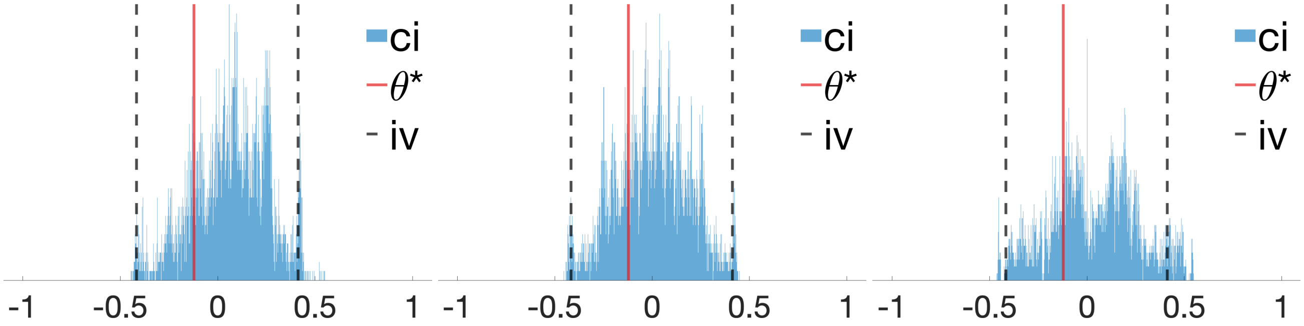
For any u_1, u_2, the induced function $f_Y(\cdot, u_1, u_2)$ must coincide with only of the above elements in the hypothesis class $\Omega_X \mapsto \Omega_Y$. Let $\mathcal{U}_Y^{(i)}, i = 1, 2, 3, 4$ be a subset of the product domain $\Omega_{U_1} \times \Omega_{U_2}$ such that $(u_1, u_2) \in \mathcal{U}_Y^{(i)}$ if any only if $f_Y(\cdot, u_1, u_2) = h_Y^{(i)}$. It is verifiable that $\mathcal{U}_Y^{(i)}, i = 1, 2, 3, 4$ must form a partition over $\Omega_{U_1} \times \Omega_{U_2}$.

We now construct a discrete SCM N compatible with the causal diagram of Fig. 10b. Let the exogenous variable U in N be a tuple (U_Z, U_X, U_Y), where $U_Z \in \{1, 2\}$, $U_X \in \{1, 2, 3, 4\}$ and $U_Y \in \{1, 2, 3, 4\}$. For any u_Z, values of Z are decided by a function as follows:

$$z \leftarrow f_Z(u_z) = h_Z^{(u_Z)}, \tag{78}$$

where $h_Z^{(1)} = 0$ and $h_Z^{(2)} = 1$. Given any input z, u_X, values of X are given by

$$x \leftarrow f_X(z, u_X) = h_X^{(u_X)}(z), \tag{79}$$



(a) $\alpha = 0.1$ (b) $\alpha = 1$ (c) $\alpha = 10$

Figure 11: Simulation results for the experiment in (Silva & Evans, 2016, Section 4.4). For all plots (a - c), *ci* represents our proposed algorithm; θ^* is the actual ACE; and *iv* stands for the optimal IV bounds (Balke & Pearl, 1997).

where $h_X^{(i)}(z)$, $i = 1, 2, 3, 4$, are defined in Eq. (76). Similarly, given any x, u_Y, values of Y are given by

$$y \leftarrow f_Y(x, u_Y) = h_Y^{(u_Y)}(x), \tag{80}$$

where $h_Y^{(i)}(x)$, $i = 1, 2, 3, 4$, are functions defined in Eq. (77). Finally, we define the exogenous distribution $P(u_Z, u_X, u_Y)$ in the discrete SCM N as the joint probability over partitions $\mathcal{U}_Z^{(i)}, \mathcal{U}_X^{(j)}, \mathcal{U}_Y^{(k)}$, $i = 1, 2$, $j = 1, 2, 3, 4$, $k = 1, 2, 3, 4$. That is,

$$\begin{aligned} &P_N\left(U_Z = i, U_X = j, U_Y = k\right) \\ &= P_M\left((U_1, U_2) \in \mathcal{U}_Z^{(i)} \wedge \mathcal{U}_X^{(j)} \wedge \mathcal{U}_Y^{(k)}\right). \end{aligned} \tag{81}$$

It follows from the decomposition in Lem. A.4 that N and M must coincide in all counterfactual distributions over binary X, Y, Z. The total cardinality of the exogenous domains in N is $|\Omega_{U_Z}| \times |\Omega_{U_X}| \times |\Omega_{U_Y}| = 2 \times 4 \times 4 = 32$.

However, the construction for the reverse direction does not hold true. That is, given an arbitrary discrete N compatible with the causal diagram in Fig. 10b, one may not be able to construct an SCM M compatible with the causal diagram in Fig. 10a such that M and N coincide in all counterfactual distributions. To witness, consider a discrete SCM N where $P(U_Z = U_X) = 1$, i.e., variables U_Z and U_X are always the same, taking values in $\{1, 2\}$. Since in SCM N, values of $Z(u_Z)$ and $X_{z=1}(u_X)$ are given by

$$\begin{aligned} Z(u_Z) &= h_Z^{(u_Z)} = 0 \times \mathbb{1}_{u_Z=1} + 1 \times \mathbb{1}_{u_Z=2}, \\ X_{z=1}(u_X) &= h_X^{(u_X)}(1) = 0 \times \mathbb{1}_{u_X=1} + 1 \times \mathbb{1}_{u_X=2}. \end{aligned}$$

This means that values of counterfactual variables Z and $X_{z=0}$ must always coincide, i.e., $P(Z = X_{x=1}) = 1$. However, for any SCM M compatible with Fig. 10a, counterfactual variables Z and X_z must be independent due to the independence restriction (Pearl, 2000, Ch. 7.3.2), i.e., $Z \perp\!\!\!\perp X_z$, which is a contradiction.

F. Comparison with (Silva & Evans, 2016)

(Silva & Evans, 2016) argued against the Bayesian approach for inferring about unknown counterfactual probabilities even when correct cardinalities of exogenous domains are known. More specifically, the authors consider the "IV" diagram in Fig. 1a with binary variables $X, Y, Z \in \{0, 1\}$. The goal is to bound the average causal effect ACE $= E[Y_{X=1}] - E[Y_{X=0}]$ from the observational distribution $P(X, Y, Z)$. The authors assigned a few different choices of Dirichlet priors over the exogenous distribution $P(U)$. Simulation results, shown in (Silva & Evans, 2016, Figure 3), found that "the posterior over the ACE covers a much narrower area than" the optimal IV bound in (Balke & Pearl, 1997), "and its behavior is erratic".

First, we would like to point out that there is a critical difference between our proposed Bayesian approach and the one discussed in (Silva & Evans, 2016). Our algorithm does not simply "put priors on the latent variable model to get a point estimate, such as the posterior expected value of the ACE.". Indeed, we are aware that when the counterfactual probability (e.g., ACE) is not identifiable in the causal diagram, the posterior expected value does not necessarily encode any true knowledge about the underlying model. Overall, it is infeasible to obtain an accurate point-estimate of a non-identifiable counterfactual probability from the observational data, regardless of how sophisticated the prior distributions are.

Therefore, our algorithm focuses on deriving the support of the posterior distribution, which is a 100% credible interval containing all possible values of the target counterfactual probability. It has been shown in (Chickering & Pearl, 1997; Imbens & Rubin, 1997) that the credible interval of the posterior distribution could effectively approximate the IV bounds in (Balke & Pearl, 1997).

Nevertheless, we notice that the supports of the posterior distributions in Figures 3(a-c) in (Silva & Evans, 2016) do not match the IV bound. We believe this is primarily due to implementations of Gibbs samplers rather than the technicality of the credible interval approach. For instance, the Gibbs sampler of (Chickering & Pearl, 1997) tends to get stuck in the local optima, thus failing to travel the full support of the posterior distribution. There are several ways to address this issue in practice. For example, one could run multiple Monte Carlo Markov Chains (MCMCs) with random initialization. Another approach is to use priors that concentrate more on the boundary of the domains of the target ACE. We could observe this phenomenon by comparing Figures 3(a) and 3(c) in (Silva & Evans, 2016), where the flat prior in Figure 3(a) achieves better approximation to the IV bound compared to the narrow prior in Figure 3(c).

Finally, we also implement the simulation in Figures 3(a-c) of (Silva & Evans, 2016) using our proposed Gibbs sampler. We randomly draw an IV model following the procedure described in (Silva & Evans, 2016, Section 4.4), generate the observational data, and compute the IV bound in (Balke & Pearl, 1997). We generate the posterior distributions using Dirhchlet distributions $\texttt{Dir}(\alpha)$ with hyperparameters $\alpha = 0.1, 1, 10$, respectively. We show the updated simulation results in Fig. 11. Our analysis reveals that all three priors could effectively approximate the optimal IV bound. The shape of posterior distributions may vary based on the hyperparameter α. However, the support of posterior distributions remains invariant, matching the optimal IV bound. This confirms the efficacy of our proposed approach.