Partial Identification of Counterfactual Distributions

Anonymous Author(s)
Affiliation
Address
email

Abstract

This paper investigates the problem of bounding counterfactual queries from a combination of observational data and qualitative assumptions about the underlying data-generating model. These assumptions are usually represented in the form of a causal diagram (Pearl, 1995). We show that all counterfactual distributions (over finite observed variables) in an arbitrary causal diagram could be generated by a special family of structural causal models (SCMs), compatible with the same causal diagram, where unobserved (exogenous) variables are discrete, taking values in a finite domain. This entails a reduction in which the space where the original, arbitrary SCM lives can be mapped to a dual, more well-behaved space where the exogenous variables are discrete, and more easily parametrizable. Using this reduction, we translate the bounding problem in the original space into an equivalent optimization program in the new space. Solving such programs leads to optimal bounds over unknown counterfactuals. Finally, we develop effective Monte Carlo algorithms to approximate these optimal bounds from a finite number of observational data. Our algorithms are validated extensively on synthetic datasets.

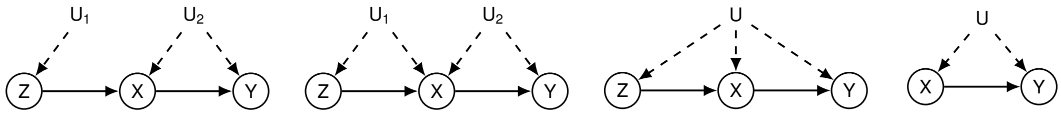
1 Introduction

This paper studies the problem of inferring counterfactual queries from the combination of non-experimental data (e.g., observational studies) and qualitative assumptions about the data-generating process. These assumptions are represented in the form of a *causal diagram* [32], which is a directed acyclic graph where arrows indicate the potential existence of functional relationships among corresponding 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 was not. Therefore, it is not immediately clear how to design effective experimental procedures for evaluating counterfactuals, let alone how to compute them from observations alone.

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 exist a complete proof system for reasoning about counterfactual queries [19]. While such a system, in principle, is sufficient in evaluating any identifiable counterfactual expression, it lacks a proof guideline which determines the feasibility of such evaluation efficiently. There are algorithms to determine whether a counterfactual distribution is inferrable from all possible controlled experiments [41]. There exist also algorithms for identifying path-specific effects from experimental data [1] and observational data [42].

In practice, however, the combination of quantitative knowledge and observed data does not always permit one to point-identify the target counterfactual queries. Partial identification methods concern with deriving informative bounds over the target counterfactual probability, even when the target

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.



(a) IV (b) Double bow (c) (d) Bow

Figure 1: DAGs (a-d) containing a treatment X, an outcome Y, an ancestor Z, and exogenous variables U; Z in (a) is also referred to as an instrumental variable.

itself is non-identifiable. Several algorithms have been developed to bound counterfactuals from the combination of observational and experimental data [30, 36, 3, 4, 14, 35, 23, 24, 16, 25, 49].

In this work, we build on the approach introduced by Balke & Pearl in [3], which involves direct discretization of the exogenous domains, also referred to as the principal stratification [17, 34]. Consider the causal diagram of Fig. 1a, where X, Y, Z are binary variables in $\{0, 1\}$; U is an unobserved variable taking values in an arbitrary continuous domain. [3] showed that domains of U could be discretized into 16 equivalent classes without changing the original counterfactual distributions and the graphical structure in Fig. 1a. For instance, despite it being induced by an arbitrary distribution $P^*(u)$ over a continuous domain of the exogenous variable U, the observational distribution $P(x, y|z)$ must be reproduced by a generative model of the form $P(x, y|z) = \sum_u P(x|u, z)P(y|x, u)P(u)$, where $P(u)$ is a discrete distribution over a finite exogenous domain $\{1, \dots, 16\}$.

Using the finite-state representation of unobserved variables, [4] derived tight bounds on treatment effects under the condition of noncompliance in Fig. 1a. [11, 21] 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 their optimal guarantees, these bounds are only applicable to the specific noncompliance setting in Fig. 1a. For the most general cases, a systematic procedure for bounding counterfactual queries in arbitrary causal diagrams is still missing.

Our goal in this paper is to overcome these challenges. We investigate the expressive power of *discrete structural causal models* (SCMs) [33] where each unobserved variable is drawn from a discrete distribution, takes values in a finite set of states. 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 observation allows us to develop novel partial identification algorithms to bound unknown counterfactual probabilities from the observational data. More specifically, our contributions are as follows. (1) We introduce a special family of discrete SCMs, with finite unobserved domains, and show that it could represent all categorical counterfactual distributions in an arbitrary causal diagram. (2) Using this result, we translate the original partial identification task into equivalent polynomial programs. Solving such programs leads to informative bounds over unknown counterfactual probabilities, which are provably optimal. (3) We develop an effective Monte Carlo algorithm to approximate optimal counterfactual bounds from a finite number of observational data. Finally, our algorithms are validated extensively on synthetic datasets. Given space constraints, all proofs are provided in Appendices A and B.

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. For convenience, we denote by $P(\boldsymbol{x})$ probabilities $P(\boldsymbol{X} = \boldsymbol{x})$; for an arbitrary subdomain $\mathcal{X} \subseteq \Omega_X$, $P(\mathcal{X}) \equiv P(X \in \mathcal{X})$. Finally, the indicator function $\mathbb{1}_{\boldsymbol{X}=\boldsymbol{x}}$ returns 1 if an event $\boldsymbol{X} = \boldsymbol{x}$ holds true; otherwise $\mathbb{1}_{\boldsymbol{X}=\boldsymbol{x}} = 0$.

The basic semantical framework of our analysis rests on *structural causal models* (SCMs) [33, Ch. 7]. An SCM M is a tuple $\langle \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{F}, P \rangle$ where $\boldsymbol{V}$ is a set of endogenous variables and $\boldsymbol{U}$ is a set of exogenous variables. $\boldsymbol{F}$ is a set of functions where each $f_V \in \boldsymbol{F}$ decides values of an endogenous variable $V \in \boldsymbol{V}$ taking as argument a combination of other variables in the system. That is, $v \leftarrow f_V(pa_V, u_V), Pa_V \subseteq \boldsymbol{V}, U_V \subseteq \boldsymbol{U}$. Exogenous variables $U \in \boldsymbol{U}$ 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 is 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 function f_V.

An intervention on an arbitrary subset $\boldsymbol{X} \subseteq \boldsymbol{V}$, denoted by do($\boldsymbol{x}$), is an operation where values of $\boldsymbol{X}$ are set to constants $\boldsymbol{x}$, regardless of how they are ordinarily determined. For an SCM M, let $M_{\boldsymbol{x}}$ denote a submodel of M induced by intervention do($\boldsymbol{x}$). For any subset $\boldsymbol{Y} \subseteq \boldsymbol{V}$, the *potential response* $\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{u})$ is defined as the solution of $\boldsymbol{Y}$ in the submodel $M_{\boldsymbol{x}}$ given $\boldsymbol{U} = \boldsymbol{u}$. Drawing values of exogenous variables $\boldsymbol{U}$ following the probability measure P 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 any 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:

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

Distributions of the form $P(\boldsymbol{y}_{\boldsymbol{x}})$ is called the *interventional distribution*; when the treatment set $\boldsymbol{X} = \emptyset$, $P(\boldsymbol{y})$ coincides with the *observational distribution*. Throughout this paper, we assume that endogenous variables $\boldsymbol{V}$ are discrete and finite; while exogenous variables $\boldsymbol{U}$ could take any (continuous) value. The counterfactual distribution $P\left(\boldsymbol{y}_{\boldsymbol{x}}, \dots, \boldsymbol{z}_{\boldsymbol{w}}\right)$ defined above is thus a categorical distribution. For a more detailed survey on SCMs, we refer readers to [33, Ch. 7].

2 Discretization of Structural Causal Models

For a DAG $\mathcal{G}$ with endogenous $\boldsymbol{V}$ and exogenous variables $\boldsymbol{U}$, let $\boldsymbol{P}^*$ denote the collection of all counterfactual distributions over variables $\boldsymbol{V}$. Formally,

$$\boldsymbol{P}^* = \left\{P\left(\boldsymbol{y}_{\boldsymbol{x}}, \dots, \boldsymbol{z}_{\boldsymbol{w}}\right) \mid \forall \boldsymbol{Y}, \dots, \boldsymbol{Z}, \boldsymbol{X}, \dots, \boldsymbol{W} \subseteq \boldsymbol{V}\right\}. \tag{2}$$

Let $\mathscr{M}$ be the family of all the SCMs compatible with the causal diagram $\mathcal{G}$, i.e., $\mathscr{M} = \{\forall M \mid \mathcal{G}_M = \mathcal{G}\}$[1]. Counterfactual distributions in $\mathcal{G}$ are defined as the collection $\{\boldsymbol{P}^*_M : \forall M \in \mathscr{M}\}$ that contains all counterfactual probabilities induced by SCMs M in the candidate family $\mathscr{M}$. In this section, we will show that counterfactual distributions in any causal diagram $\mathcal{G}$ could be generated by an alternative family of "generic" SCMs compatible with $\mathcal{G}$, which we will define later.

Definition 1 (Counterfactual-Equivalence)**.** For a DAG $\mathcal{G}$, let $\mathscr{M}, \mathscr{N}$ be two sets of SCMs compatible with $\mathcal{G}$. $\mathscr{M}$ and $\mathscr{N}$ are said to be *counterfactually equivalent* (for short, ctf-equivalent) if for any $M \in \mathscr{M}$, there exists an alternative $N \in \mathscr{N}$ such that $\boldsymbol{P}^*_M = \boldsymbol{P}^*_N$, and vice versa.

Our analysis rests on a special family of SCMs where values of each exogenous variable are drawn from a discrete distribution over a finite set of states.

Definition 2. An SCM $M = \langle \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{F}, P\rangle$ is said to be a discrete SCM if

1. Values of every $U \in \boldsymbol{U}$ are drawn from a discrete distribution $P(u)$ over a domain Ω_U; let θ_u denote the probability $P(U = u)$, for any $u \in \Omega_U$.
2. Values of every $V \in \boldsymbol{V}$ are decided by function $v \leftarrow f_V(pa_V, u_V) \equiv \xi_V^{(pa_V, u_V)}$, where for $\forall pa_V, u_V$, $\xi_V^{(pa_V, u_V)}$ is a constant in the finite domain Ω_V.

Given a causal diagram $\mathcal{G}$, our goal is to construct a family of discrete SCMs $\mathscr{N}$ that is counterfactually equivalent to the original family of SCMs $\mathscr{M}$. Our construction utilizes a special type of clustering of nodes in the diagram, called the confounded component [45].

Definition 3. For an DAG $\mathcal{G}$, a subset $\boldsymbol{C} \subseteq \boldsymbol{V}$ is a c-component if any pair $X, Y \in \boldsymbol{C}$ is connected in $\mathcal{G}$ by a *bi-directed path* of the form $V_1 \leftrightarrow V_2 \leftrightarrow \cdots \leftrightarrow V_n$, $n = 1, 2, \dots$, where (1) $V_1 = X$, $V_n = Y$; (2) $\{V_1, \dots, V_n\} \subseteq \boldsymbol{V}$; and (3) each $V_i \leftrightarrow V_j$ is a sequence $V_i \leftarrow U_k \rightarrow V_j$ and $U_k \in \boldsymbol{U}$.

A c-component $\boldsymbol{C}$ in $\mathcal{G}$ is maximal if there exists no other c-component that contains $\boldsymbol{C}$. We denote by $\mathcal{C}(\mathcal{G})$ the collection of all maximal c-components in $\mathcal{G}$. Naturally, c-components in $\mathcal{C}(\mathcal{G})$ form a partition over endogenous variables $\boldsymbol{V}$, which, in turn, defines a partition $\{\cup_{V \in \boldsymbol{C}} U_V \mid \forall \boldsymbol{C} \in \mathcal{C}(\mathcal{G})\}$ over exogenous variables $\boldsymbol{U}$. Therefore, for every $U \in \boldsymbol{U}$, there must exist a unique c-component in $\mathcal{C}(\mathcal{G})$, denoted by $\boldsymbol{C}_U$, such that $U \in \cup_{V \in \boldsymbol{C}_U} U_V$. For example, exogenous variables U_1, U_2 in Fig. 1a corresponds to c-components $\boldsymbol{C}_{U_1} = \{Z\}$ and $\boldsymbol{C}_{U_2} = \{X, Y\}$ respectively; while the causal diagram of Fig. 1b only has a single c-component $\{X, Y, Z\}$.

[1]We will use the subscript M to represent the restriction to a specific SCM M. Therefore, $\mathcal{G}_M$ represents the causal diagram associated with SCM M; so does the collection of counterfactuals $\boldsymbol{P}^*_M$.

Theorem 1. *For a DAG $\mathcal{G}$, consider the following conditions[2]: (1) $\mathscr{M}$ is the set of all SCMs compatible with $\mathcal{G}$; (2) $\mathscr{N}$ is the set of all discrete SCMs compatible with $\mathcal{G}$ where for every $U \in \boldsymbol{U}$, its cardinality $|\Omega_U| = \prod_{V \in \boldsymbol{C}_U} |\Omega_{Pa_V} \mapsto \Omega_V|$, i.e., the number of functions mapping from Pa_V to V for every variable V in the c-component $\boldsymbol{C}_U$. Then, $\mathscr{M}$ and $\mathscr{N}$ are counterfactually equivalent.*

Thm. 1 establishes the expressive power of discrete SCMs in representing counterfactual distributions in a causal diagram $\mathcal{G}$. It implies that the counterfactual distribution $P(\boldsymbol{y_x}, \dots, \boldsymbol{z_w})$ in any SCM M could be generated using a generic model as follows, for $d_U = \prod_{V \in \boldsymbol{C}_U} |\Omega_{Pa_V} \mapsto \Omega_V|$,

$$P(\boldsymbol{y_x}, \dots, \boldsymbol{z_w}) = \sum_{U \in \boldsymbol{U}} \sum_{u=1,\dots,d_U} \mathbb{1}_{\boldsymbol{Y_x}(\boldsymbol{u})=\boldsymbol{y}} \wedge \dots \wedge \mathbb{1}_{\boldsymbol{Z_w}(\boldsymbol{u})=\boldsymbol{z}} \prod_{U \in \boldsymbol{U}} \theta_u. \tag{3}$$

Among above quantities, θ_u are parameters of the exogenous distribution $P(u)$ over a finite domain $\{1, \dots, d_U\}$. Counterfactual variables $\boldsymbol{Y_x}(\boldsymbol{u})$ are recursively defined as follows:

$$\boldsymbol{Y_x}(\boldsymbol{u}) = \{Y_{\boldsymbol{x}}(\boldsymbol{u}) \mid \forall Y \in \boldsymbol{Y}\}, \text{ where } Y_{\boldsymbol{x}}(\boldsymbol{u}) = \begin{cases} \boldsymbol{x}_Y & \text{if } Y \in \boldsymbol{X} \\ \xi_Y^{(\{V_{\boldsymbol{x}}(\boldsymbol{u}) \mid V \in Pa_Y\}, u_Y)} & \text{otherwise} \end{cases} \tag{4}$$

where $\boldsymbol{x}_Y$ is the value assigned to variable Y in constants $\boldsymbol{x}$. As an example, consider the causal diagram $\mathcal{G}$ described in Fig. 1b where X, Y, Z are binary variables in $\{0, 1\}$. Since $\mathcal{G}$ has a single c-component $\{X, Y, Z\}$, exogenous variables U_1, U_2 must share the same cardinality d in the proposed family of discrete SCMs $\mathscr{N}$. It follows from Thm. 1 the counterfactual distribution $P(z, x_{z'}, y_{x'})$ in any SCM compatible with $\mathcal{G}$ could be written as follows:

$$P(z, x_{z'}, y_{x'}) = \sum_{u_1, u_2 = 1}^{d} \mathbb{1}_{\xi_Z^{(u_1)}=z} \wedge \mathbb{1}_{\xi_X^{(z', u_1, u_2)}=x} \wedge \mathbb{1}_{\xi_Y^{(x', u_2)}=y} \theta_{u_1} \theta_{u_2}, \tag{5}$$

where $\xi_Z^{(u_1)}, \xi_X^{(z, u_1, u_2)}, \xi_Y^{(x, u_2)}$ are parameters taking values in $\{0, 1\}$; $\theta_{u_i}, i = 1, 2$, are probabilities of the discrete distribution $P(u_i)$ over the finite domain $\{1, \dots, d\}$. The cardinality $d = |\Omega_Z| \times |\Omega_Z \mapsto \Omega_X| \times |\Omega_X \mapsto \Omega_Y| = 32$. The total cardinalities of domains for U_1, U_2 are thus $2d = 64$.

Comparison with related work One could naïvely apply the discretization procedure in [3] and obtain a family of discrete SCMs that are sufficient in representing distributions in an causal diagram. However, such parametrization is not necessarily complete. To witness, consider again the causal diagram in Fig. 1b with binary X, Y, Z. Applying the discretization in [3] leads to a family of discrete SCMs compatible with a different diagram in Fig. 1c where the cardinality of exogenous variable U is equal to $d = 32$ (see Appendix D for details). However, this parametrization fails to capture some critical constraints over counterfactual distributions since it does not maintain the original structure of the causal diagram. For instance, counterfactual variables Z and Y_x in the original diagram of Fig. 1b are independent due to independence restrictions [33, Ch. 7.3.2]; while Z and Y_x in Fig. 1c are generally correlated due to the presence of unobserved confounder U. Compared with [3], the discretization method in Thm. 1 captures *all* constraints over counterfactual distributions while requiring only a factor of $|\boldsymbol{U}|$ increase in the cardinality of exogenous domains.

More recently, [15] proved a special case of Thm. 1 for interventional distributions in a specific class of causal diagrams that satisfy the running intersection property. When there is no direct arrow between endogenous variables, [38] showed that the observational distribution in a diagram could be represented using finite-state exogenous variables. Thm. 1 generalizes these results by showing that, for the first time, *all* counterfactual distributions in an *arbitrary* causal diagram could be generated using discrete exogenous variables taking values from a finite domain, without any loss of generality.

2.1 Partial identification of Counterfactual Distributions

To demonstrate the expressive power of discrete SCMs, we investigate the problem of partial identification of counterfactual distributions. For an SCM $M^* = \langle \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{F}, P \rangle$, we are interested in evaluating an arbitrary counterfactual probability $P(\boldsymbol{y_x}, \dots, \boldsymbol{z_w})$. The detailed parametrization of M^* is unknown. Instead, the learner only has access to the causal diagram $\mathcal{G}$ and the observational distribution $P(\boldsymbol{v})$ induced by M^*. Our goal is to derive an informative bound $[l, r]$ from the combination of $\mathcal{G}$ and $P(\boldsymbol{v})$ that contains the actual counterfactual probability $P(\boldsymbol{y_x}, \dots, \boldsymbol{z_w})$.

[2]For every $V \in \boldsymbol{V}$, $\Omega_{Pa_V} \mapsto \Omega_V$ is the set of all functions mapping from domains Ω_{Pa_V} to Ω_V.

Let $\mathscr{N}$ denote the family of discrete SCMs defined in Thm. 1 which are compatible with the causal diagram $\mathcal{G}$. We derive a bound $[l, r]$ over $P(\boldsymbol{y_x}, \ldots, \boldsymbol{z_w})$ from the observational data $P(\boldsymbol{v})$ by solving the following optimization problem:

$$[l, r] = \min / \max \left\{ P_N(\boldsymbol{y_x}, \ldots, \boldsymbol{z_w}) \mid \forall N \in \mathscr{N}, P_N(\boldsymbol{v}) = P(\boldsymbol{v}) \right\} \tag{6}$$

For instance, consider again the double-bow diagram $\mathcal{G}$ in Fig. 1b. The observational distribution $P(x, y, z)$ in any discrete SCM in $\mathscr{N}$ could be written as:

$$P(x, y, z) = \sum_{u_1, u_2 = 1}^{d} \mathbb{1}_{\xi_Z^{(u_1)} = z} \wedge \mathbb{1}_{\xi_X^{(z, u_1, u_2)} = x} \wedge \mathbb{1}_{\xi_Y^{(x, u_2)} = y} \theta_{u_1} \theta_{u_2}. \tag{7}$$

One could derive a bound over the counterfactual distribution $P(z, x_{z'}, y_{x'})$ from the observational data $P(x, y, z)$ by solving polynomial programs which optimize the objective Eq. (5) over parameters $\theta_{u_1}, \theta_{u_2}, \xi_Z^{(u_1)}, \xi_X^{(z, u_1, u_2)}, \xi_Y^{(x, u_2)}$, subject to the observational constraints Eq. (7).

As a corollary, it follows immediately from Thm. 1 that the solution $[l, r]$ of the optimization problem Eq. (6) is guaranteed to be a valid bound over the unknown counterfactual $P(\boldsymbol{y_x}, \ldots, \boldsymbol{z_w})$.

Corollary 1 (Soundness)**.** *Given a DAG $\mathcal{G}$ and an observational distribution $P(\boldsymbol{v})$, let $\mathscr{M}$ be the set of all SCMs compatible with $\mathcal{G}$ and let $\mathscr{M}_o = \{\forall M \in \mathscr{M} \mid P_M(\boldsymbol{v}) = P(\boldsymbol{v})\}$. For the solution $[l, r]$ of Eq.* (6)*, $P_M(\boldsymbol{y_x}, \ldots, \boldsymbol{z_w}) \in [l, r]$ for any SCM $M \in \mathscr{M}_o$.*

Since the underlying SCM $M^* \in \mathscr{M}_o$, Corol. 1 implies that the derived bound $[l, r]$ must contain the actual counterfactual probability $P(\boldsymbol{y_x}, \ldots, \boldsymbol{z_w})$. Our next result shows that such a bound $[l, r]$ is provably tight, i.e., it cannot be improved without additional assumptions.

Corollary 2 (Tightness)**.** *Given a DAG $\mathcal{G}$ and an observational distribution $P(\boldsymbol{v})$, let $\mathscr{M}$ be the set of all SCMs compatible with $\mathcal{G}$ and let $\mathscr{M}_o = \{\forall M \in \mathscr{M} \mid P_M(\boldsymbol{v}) = P(\boldsymbol{v})\}$. For the solution $[l, r]$ of Eq.* (6)*, there exist SCMs $M_1, M_2 \in \mathscr{M}_o$ such that $P_{M_1}(\boldsymbol{y_x}, \ldots, \boldsymbol{z_w}) = l$, $P_{M_2}(\boldsymbol{y_x}, \ldots, \boldsymbol{z_w}) = r$.*

Corol. 2 confirms the tightness of the bound $[l, r]$ obtained from Eq. (6). Suppose there exists a valid bound $[l', r']$ strictly contained in $[l, r]$. One could construct from Corol. 2 an SCM M compatible with the causal diagram $\mathcal{G}$ and the observational distribution $P(\boldsymbol{v})$, but its counterfactual probability $P(\boldsymbol{y_x}, \ldots, \boldsymbol{z_w})$ lies outside $[l', r']$, which is a contradiction.

The optimization problem of Eq. (6) is reducible to equivalent polynomial programs (see Appendix E). Despite the soundness and tightness of derived bounds, solving such programs may take exponentially long in the most general case [29]. Our focus here is upon the causal inference aspect of the problem and like earlier discussions we do not specify which solvers are used [3, 4]. In some cases of interest, effective approximate planning methods for polynomial programs do exist. Investigating these methods is an ongoing subject of research [26, 31, 48, 28, 27].

3 Bayesian Approach for Partial Identification

This section describes an effective algorithm to approximate the optimal counterfactual bound in Eq. (6), provided with finite samples $\bar{\boldsymbol{v}} = \left\{\boldsymbol{v}^{(n)}\right\}_{n=1}^{N}$ drawn from the observational distribution $P(\boldsymbol{v})$, and prior distributions over parameters θ_u and $\xi_V^{(pa_V, u_V)}$ (possibly uninformative).

We first introduce Markov Chain Monte Carlo (MCMC) algorithms that sample the posterior distribution $P(\theta_{\text{ctf}} \mid \bar{\boldsymbol{v}})$ over a counterfactual probability $\theta_{\text{ctf}} = P(\boldsymbol{y_x}, \ldots, \boldsymbol{z_w})$. More specifically, for every $V \in \boldsymbol{V}$, $\forall pa_V, u_V$, parameters $\xi_V^{(pa_V, u_V)}$ are drawn uniformly over the finite domain Ω_V. For every $U \in \boldsymbol{U}$, exogenous probabilities θ_u are drawn from a generalized Dirichlet distribution [12]. We will take the view of a stick-breaking construction [40] which successively breaks pieces off a unit-length stick with size proportional to random draws from a Beta distribution. Parameters θ_u are proportions of each of the pieces relative to its original size. Formally,

$$\forall u = 1, 2, \ldots, d_U, \qquad \theta_u = \mu_u \prod_{i=1}^{u-1} (1 - \mu_i), \qquad \mu_u \sim \texttt{Beta}\left(\alpha_U^{(u)}, \beta_U^{(u)}\right), \tag{8}$$

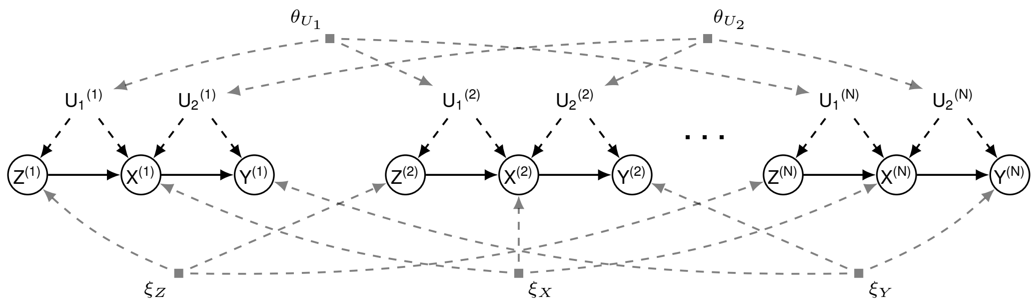


Figure 2: The data-generating process for the observational data $\left\{X^{(n)}, Y^{(n)}, Z^{(n)}\right\}_{n=1}^{N}$ in an SCM associated with the causal diagram in Fig. 1b. For every exogenous variable $U \in \boldsymbol{U}$, $\theta_U = \{\theta_u \mid \forall u\}$. For every endogenous variable $V \in \boldsymbol{V}$, $\xi_V = \left\{\xi_V^{(pa_V, u_V)} \mid \forall pa_V, u_V\right\}$.

where $d_U = \prod_{V \in \boldsymbol{C}_U} |\Omega_{Pa_V} \mapsto \Omega_V|$ and $\alpha_U^{(u)}, \beta_U^{(u)} > 0$ are hyperparameters. Finally, we truncate this construction by setting $\mu_{d_U} = 1$. Note from Eq. (8) that all parameters θ_u for $u > d_U$ are equal to zero. As an example, Fig. 2 shows a graphical representation of the data-generating process over parameters θ_u and $\xi_V^{(pa_V, u_V)}$ associated with SCMs in Fig. 1b, spanning over N observations.

Gibbs sampling is a well-known MCMC algorithm that allows one to sample posterior distributions. For convenience, we introduce the following notations. Let parameters $\boldsymbol{\theta} = \{\theta_u \mid \forall U \in \boldsymbol{U}, \forall u\}$ and $\boldsymbol{\xi} = \left\{\xi_V^{(pa_V, u_V)} \mid \forall V \in \boldsymbol{V}, \forall pa_V, u_V\right\}$. The set $\bar{\boldsymbol{U}} = \left\{\boldsymbol{U}^{(n)}\right\}_{n=1}^{N}$ are exogenous variables affecting N observations $\bar{\boldsymbol{V}} = \left\{V^{(n)}\right\}_{n=1}^{N}$; we use $\bar{\boldsymbol{u}}$ to represent their realizations. Our blocked Gibbs sampler works by iteratively drawing values from the conditional distributions of variables as follows [22]. Detailed derivations of complete conditional distributions are shown in Appendix F.

Sampling $P(\bar{\boldsymbol{u}} \mid \bar{\boldsymbol{v}}, \boldsymbol{\theta}, \boldsymbol{\xi})$. Exogenous variables $\boldsymbol{U}^{(n)}$, $n = 1, \dots, N$, are mutually independent given parameters $\boldsymbol{\theta}, \boldsymbol{\xi}$. We could draw each $\left(\boldsymbol{U}^{(n)} \mid \boldsymbol{\theta}, \boldsymbol{\xi}, \bar{\boldsymbol{V}}\right)$ corresponding to the nth observation independently. The complete conditional for $\boldsymbol{U}^{(n)}$ is given by

$$P\left(\boldsymbol{u}^{(n)} \mid \boldsymbol{v}^{(n)}, \boldsymbol{\theta}, \boldsymbol{\xi}\right) \propto \prod_{V \in \boldsymbol{V}} \mathbb{1}_{\xi_V^{\left(pa_V^{(n)}, u_V^{(n)}\right)} = v^{(n)}} \prod_{U \in \boldsymbol{U}} \theta_u. \tag{9}$$

Sampling $P(\boldsymbol{\xi}, \boldsymbol{\theta} \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{u}})$. Parameters $\boldsymbol{\xi}, \boldsymbol{\theta}$ are independent given $\bar{\boldsymbol{V}}, \bar{\boldsymbol{U}}$. Therefore, we will derive complete conditional $\boldsymbol{\xi}, \boldsymbol{\theta}$ separately. Note that in discrete SCMs, the nth observation of variable $V \in \boldsymbol{V}$ is decided by $v^{(n)} \leftarrow \xi_V^{(pa_V, u_V)}$ given $pa_V^{(n)} = pa_V$, $u_V^{(n)} = u_V$. Thus, draw values of each $\xi_V^{(pa_V, u_V)} \in \boldsymbol{\xi}$ from the complete conditional defined as:

$$P\left(\xi_V^{(pa_V, u_V)} \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{u}}\right) = \begin{cases} \mathbb{1}_{\xi_V^{(pa_V, u_V)} = v^{(i)}} & \text{if } \exists i, \text{ s.t. } pa_V^{(i)} = pa_V, u_V^{(i)} = u_V, \\ 1/|\Omega_V| & \text{otherwise.} \end{cases} \tag{10}$$

Let $n_u = \sum_{n=1}^{N} \mathbb{1}_{u^{(n)} = u}$ records the number of values in $u^{(n)}$ that are equal to u. By the conjugacy of the generalized Dirichlet distribution, the complete conditional of θ_u is given by, for every $U \in \boldsymbol{U}$,

$$\forall u = 1, 2, \dots d_U, \quad \theta_u = \mu_u \prod_{i=1}^{u-1} (1 - \mu_i), \quad \mu_u \sim \texttt{Beta}\left(\alpha_U^{(u)} + n_u, \beta_U^{(u)} + \sum_{k=u+1}^{d_U} n_k\right). \tag{11}$$

Doing so eventually produces values drawn from the posterior distribution over $\left(\boldsymbol{\theta}, \boldsymbol{\xi}, \bar{\boldsymbol{U}} \mid \bar{\boldsymbol{V}}\right)$. Given parameters $\boldsymbol{\theta}, \boldsymbol{\xi}$, we compute the counterfactual probability $\theta_{\text{ctf}} = P(\boldsymbol{y_x}, \dots, \boldsymbol{z_w})$ following the three-step algorithm in [33] which consists of abduction, action, and prediction. Thus computing θ_{ctf} from each draw $\boldsymbol{\theta}, \boldsymbol{\xi}, \bar{\boldsymbol{U}}$ eventually gives us the draw from the posterior distribution $P(\theta_{\text{ctf}} \mid \bar{\boldsymbol{v}})$.

3.1 Collapsed Gibbs Sampling

We also describe an alternative sampler that applies to stick-breaking priors with a known Pólya urn characterization. Formally, consider stick-breaking priors in Eq. (8) with hyperparameters

$\alpha_U^{(u)} = \alpha_U/d_U$ and $\beta_U^{(u)} = (d_U - u)\alpha_U/d_U$ for some real $\alpha_U > 0$. Let $\bar{\boldsymbol{U}}_{-n}$ denote the set difference $\bar{\boldsymbol{U}} \setminus \boldsymbol{U}^{(n)}$; so does $\bar{\boldsymbol{V}}_{-n} = \bar{\boldsymbol{V}} \setminus \boldsymbol{V}^{(n)}$. Our collapsed Gibbs sampler first iteratively draws values from the conditional distribution of $\left(\boldsymbol{U}^{(n)} \mid \bar{\boldsymbol{U}}_{-n}, \bar{\boldsymbol{V}}\right)$, $n = 1, \dots, N$, 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 given by

$$P\left(\boldsymbol{u}^{(n)} \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{u}}_{-n}\right) \propto \prod_{V \in \boldsymbol{V}} P\left(v^{(n)} \mid pa_V^{(n)}, u_V^{(n)}, \bar{\boldsymbol{v}}_{-n}, \bar{\boldsymbol{u}}_{-n}\right) \prod_{U \in \boldsymbol{U}} P\left(u^{(n)} \mid \bar{\boldsymbol{v}}_{-n}, \bar{\boldsymbol{u}}_{-n}\right). \tag{12}$$

Among quantities in the above equation, for every $V \in \boldsymbol{V}$,

$$P\left(v^{(n)} \mid pa_V^{(n)}, u_V^{(n)}, \bar{\boldsymbol{v}}_{-n}, \bar{\boldsymbol{u}}_{-n}\right) = \begin{cases} \mathbb{1}_{v^{(n)} = v^{(i)}} & \text{if } \exists i \neq n, pa_V^{(i)} = pa_V^{(n)}, u_V^{(i)} = u_V^{(n)}, \\ 1/|\Omega_V| & \text{otherwise.} \end{cases} \tag{13}$$

For every $U \in \boldsymbol{U}$, let $\bar{u}_{-n}$ be a set of exogenous samples $\left\{u^{(1)}, \dots, u^{(n-1)}, u^{(n+1)}, \dots, u^{(N)}\right\}$. Let $\{u_1^*, \dots, u_K^*\}$ denote K unique values that samples in $\bar{u}_{-n}$ take on.

$$P\left(u^{(n)} \mid \bar{\boldsymbol{v}}_{-n}, \bar{\boldsymbol{u}}_{-n}\right) = \begin{cases} \dfrac{n_k^* + \alpha_U/d_U}{\alpha_U + N - 1} & \text{if } u^{(n)} = u_k^*, \text{ for } k = 1, \dots, K \\ \dfrac{\alpha_U(1 - K/d_U)}{\alpha_U + N - 1} & \text{if } u^{(n)} \notin \{u_1^*, \dots, u_K^*\} \end{cases}. \tag{14}$$

where $n_k^* = \sum_{i \neq n} \mathbb{1}_{u^{(i)} = u_k^*}$ records the number of values in $u^{(i)} \in \bar{u}_{-n}$ 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 $\left(\boldsymbol{\theta}, \boldsymbol{\xi} \mid \bar{\boldsymbol{U}}, \bar{\boldsymbol{V}}\right)$; the complete conditional $P(\boldsymbol{\xi}, \boldsymbol{\theta} \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{u}})$ are given in Eqs. (10) and (11). Finally, computing θ_{ctf} from each sample $\boldsymbol{\theta}, \boldsymbol{\xi}$ gives us a draw from the posterior distribution $P(\theta_{\text{ctf}} \mid \bar{\boldsymbol{v}})$.

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 $\boldsymbol{\theta}, \boldsymbol{\xi}$ in the high-dimensional space. Instead, the collapsed sampler only draws $\boldsymbol{\theta}, \boldsymbol{\xi}$ 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 [22].

3.2 Credible Intervals over Counterfactual Probabilities

Given a MCMC sampler, one could bound the counterfactual probability θ_{ctf} by computing credible intervals from the posterior distribution $P(\theta_{\text{ctf}} \mid \bar{\boldsymbol{v}})$.

Definition 4. Fix $\alpha \in [0, 1)$. A $100(1-\alpha)\%$ credible interval $[l_\alpha, r_\alpha]$ for θ_{ctf} is given by

$$l_\alpha = \sup\{x \mid P(\theta_{\text{ctf}} \leq x \mid \bar{\boldsymbol{v}}) = \alpha/2\}, \quad r_\alpha = \inf\{x \mid P(\theta_{\text{ctf}} \leq x \mid \bar{\boldsymbol{v}}) = 1 - \alpha/2\}. \tag{15}$$

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 $1 - \alpha$. Credible intervals have been widely applied for computing bounds over counterfactuals provided with finite observations [20, 47, 37, 8, 46]. As the number of observational data N grows (to infinite), the 100% credible interval $[l_0, r_0]$ eventually converges to the optimal asymptotic bound $[l, r]$ in Eq. (6) [11].

Let $\left\{\theta^{(t)}\right\}_{t=1}^T$ be T samples drawn from $P(\theta_{\text{ctf}} \mid \bar{\boldsymbol{v}})$. One could compute the $100(1-\alpha)\%$ credible interval for θ_{ctf} using the following consistent estimators [39]:

$$\hat{l}_\alpha(T) = \theta^{(\lceil(\alpha/2)T\rceil)}, \qquad \hat{r}_\alpha(T) = \theta^{(\lceil(1-\alpha/2)T\rceil)}, \tag{16}$$

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 $\left\{\theta^{(t)}\right\}$[3]. Our next results establish non-asymptotic deviation bounds for the empirical estimates of credible intervals defined in Eq. (16) for finite samples.

Lemma 1. *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*

$$\hat{l}_\alpha(T) \in \left[l_{\alpha - f(T,\delta)}, l_{\alpha + f(T,\delta)}\right], \qquad \hat{r}_\alpha(T) \in \left[r_{\alpha + f(T,\delta)}, r_{\alpha - f(T,\delta)}\right]. \tag{17}$$

[3] For any real $\alpha \in \mathbb{R}$, $\lceil\alpha\rceil$ denotes the smallest integer $n \in \mathbb{Z}$ larger than α, i.e., $\lceil\alpha\rceil = \min\{n \in \mathbb{Z} \mid n \geq \alpha\}$.

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. (16) and return them as the output. It follows immediately from Lem. 1 that such a procedure efficiently approximates a $100(1-\alpha)\%$ credible interval.

Algorithm 1: CREDIBLEINTERVAL

Input: Credible level α, tolerance level δ, ϵ.
Output: An credible interval $[l_\alpha, h_\alpha]$ for θ_{ctf}.
Let $T = \lceil 2\epsilon^{-2} \ln(4/\delta) \rceil$.
Draw samples $\left\{\theta^{(1)}, \dots, \theta^{(T)}\right\}$ 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. (16)).

Corollary 3. *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 implies that any counterfactual parameter θ_{ctf} compatible with observational data $\bar{\boldsymbol{v}}$ falls between $[\hat{l}, \hat{r}] =$ CREDIBLEINTERVAL$(\alpha, \delta, \epsilon)$ with probability $P\left(\theta_{\text{ctf}} \in [\hat{l}, \hat{r}] \mid \bar{\boldsymbol{v}}\right) \approx 1 - \alpha \pm \epsilon$. 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 simulated SCM instances and a real world patient dataset collected from the International Stroke Trial (IST) [10]. Overall, we found that simulation results support our findings and the proposed bounding strategy consistently dominates state-of-art algorithms. When target distributions are identifiable (Experiment 1), our bounds collapse to the actual, unknown counterfactual probabilities. For non-identifiable settings, our algorithm obtains sharp asymptotic bounds when closed-form solutions already exist (Experiments 2 & 3); and improves over state-of-art bounds in other more general cases where the optimal strategy is unknown (Experiment 4).

In all experiments, we evaluate our proposed bounding strategy based on credible intervals (*ci*). In particular, we draw 4×10^3 samples from the posterior distribution over the target counterfactual $\left(\theta_{\text{ctf}} \mid \bar{\boldsymbol{V}}\right)$. 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 also include the actual counterfactual probability θ^*. For details on simulation setups and additional experiments, we refer readers to Appendix C.

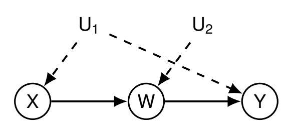
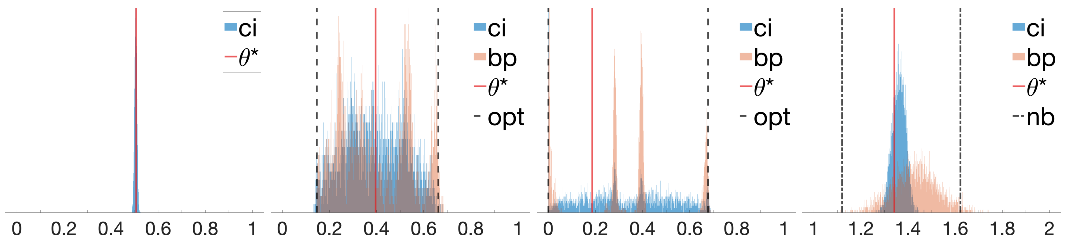


Figure 3: Frontdoor

Experiment 1: Frontdoor Graph This experiment evaluates our sampling algorithm on interventional probabilities that are identifiable from the observational data. Consider the "Frontdoor" graph described in Fig. 3 where X, Y, W are binary variables in $\{0,1\}$; $U_1, U_2 \in \mathbb{R}$. In this case, the interventional distribution $P(y_x)$ is identifiable from $P(x, w, y)$ through the frontdoor adjustment [33, 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 randomly generated SCM. Fig. 4a shows samples drawn from the posterior distribution of the target probability $\left(P(Y_{x=0} = 1) \mid \bar{\boldsymbol{V}}\right)$. The analysis reveals that these samples collapse to the actual interventional probability $P(Y_{x=0} = 1) = 0.5085$, which confirms the identifiability of $P(y_x)$ in Fig. 3.

Experiment 2: Instrumental Variables (IV) This experiment evaluates our bounding strategy in non-identifiable settings, while closed-form solutions for the optimal bounds over target probabilities already exist. Consider first the "IV" diagram in Fig. 1a where $X, Y, Z \in \{0,1\}$ and $U_1, U_2 \in \mathbb{R}$. The non-identifiability of $P(y_x)$ from the observational data $P(x, y, z)$ with the instrument Z and the unobserved confounding between X and Y has been acknowledged in [5]. For binary X, Y, Z, [2] derived closed-form, sharp bounds over $P(y_x)$ (labelled as *opt*). We collect $N = 10^4$ observational samples $\bar{\boldsymbol{V}} = \{X^{(n)}, Y^{(n)}, Z^{(n)}\}_{n=1}^N$ from a randomly generated SCM instance. Fig. 4b shows samples drawn from the posterior distribution of $\left(P(Y_{x=0} = 1) \mid \bar{\boldsymbol{V}}\right)$. As a baseline, we also include the optimal bound *opt*, and posterior samples obtained from the Gibbs sampler of [11], which utilizes the canonical partitions of exogenous domains in [2] (*bp*). The analysis reveals that our algorithm derives the valid bound over the actual probability $P(Y_{x=0} = 1) = 0.3954$; the 100% credible interval converges to the optimal IV bound $l = 0.1468, r = 0.6617$.



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

Figure 4: Histogram plots for samples drawn from the posterior distribution over target counterfactual probabilities. For all plots (a - d), *ci* represents our proposed algorithms; *bp* stands for Gibbs samplers using the representation of canonical partitions [2]; θ^* is the actual counterfactual probability. (b, c) *opt* represents the optimal asymptotic bound, if exists. (d) *nb* stands for the natural bounds [30].

Experiment 3: Probability of Necessity and Sufficiency (PNS) We now study the problem of evaluating the *probability of necessity and sufficiency* $P(Y_{x=1} = 1, Y_{x=0} = 0)$ from the observational data $P(x, y)$ in the "Bow" diagram of Fig. 1d where $X, Y \in \{0, 1\}$ and $U \in \mathbb{R}$. The sharp bound for $P(Y_{x=1} = 1, Y_{x=0} = 0)$ from $P(x, y)$ was introduced in [44] (labelled as *opt*). We collect $N = 10^4$ observational samples $\bar{\boldsymbol{V}} = \{X^{(n)}, Y^{(n)}\}_{n=1}^N$ from an SCM instance. Fig. 4c shows samples drawn from the posterior distribution of $\left(P(Y_{x=1} = 1, Y_{x=0} = 0) \mid \bar{\boldsymbol{V}}\right)$. As a baseline, we also include the optimal bound *opt*, and posterior samples obtained from the Gibbs sampler which discretizes the exogenous domains using canonical partitions [2] (*bp*). The analysis reveals that our 100% credible interval (*ci*) matches the optimal PNS bound $l = 0, r = 0.6775$, i.e., the proposed strategy achieves the sharp bound over the counterfactual probability $P(Y_{x=1} = 1, Y_{x=0} = 0) = 0.1867$.

Experiment 4: International Stroke Trials (IST) IST was a large, randomized, open trial of up to 14 days of antithrombotic therapy after stroke onset [10]. 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 $f_X^{(Z)}, Z \in \{0, \dots, 9\}$, following a procedure in [49]. Doing so allows us to obtain $N = 3 \times 10^3$ synthetic observational samples $\bar{\boldsymbol{V}} = \{X^{(n)}, Y^{(n)}, Z^{(n)}\}_{n=1}^N$ that are compatible with the "Double bow" diagram of Fig. 1b. We are interested in evaluating the treatment effect $E[Y_{x=(1,0)}]$ for only assigning aspirin $X = (1, 0)$. Fig. 4d shows samples drawn from the posterior distribution of $\left(E[Y_{x=(1,0)}] \mid \bar{\boldsymbol{V}}\right)$. As a baseline, we also include a naïve generalization of the discretization procedure (*bp*) [2] (see Appendix D) and the natural bounds [36, 30] estimated at the 95% confidence level (*nb*) [49]. Posterior samples of *ci* and *bp* are drawn using our proposed collapsed sampler due to the high-dimensional latent space. The analysis reveals that all algorithms achieve bounds that contain the actual, target causal effect $E[Y_{x=(1,0)}] = 1.3418$. Our bounding strategy obtains a 100% credible interval $l_{ci} = 1.2604, r_{ci} = 1.4687$, which consistently improves over all the other algorithms ($l_{bp} = 1.1121, r_{bp} = 1.8073, l_{nb} = 1.1195, r_{nb} = 1.6221$).

5 Conclusion

This paper investigated the problem of partial identification of counterfactual distributions, which concerns with bounding unknown counterfactual probabilities from the combination of the observational data and qualitative assumptions of the data-generating process, represented in the form of a directed acyclic causal diagram. We studied a special 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 an arbitrary causal diagram. That is, this new family of discrete SCMs is counterfactual equivalent to the original family of candidate SCMs compatible with the causal diagram. Using this result, we developed a novel algorithm to derive bounds over counterfactual probabilities from finite observations, which are provably tight.

References

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

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

[3] A. Balke and J. Pearl. Counterfactuals and policy analysis in structural models. In P. Besnard and S. Hanks, editors, *Uncertainty in Artificial Intelligence 11*, pages 11–18. San Francisco, 1995.

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

[5] E. Bareinboim and J. Pearl. Causal inference by surrogate experiments: z-identifiability. In N. de Freitas and K. Murphy, editors, *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 113–120, Corvallis, OR, 2012. AUAI Press.

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

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

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

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

[10] A. Carolei 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.

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

[12] R. J. Connor and J. E. Mosimann. Concepts of independence for proportions with a generalization of the dirichlet distribution. *Journal of the American Statistical Association*, 64(325):194–206, 1969.

[13] J. Eckhoff. Helly, radon, and carathéodory type theorems. In *Handbook of convex geometry*, pages 389–448. Elsevier, 1993.

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

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

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

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

[18] D. Galles and J. Pearl. An axiomatic characterization of causal counterfactuals. *Foundation of Science*, 3(1):151–182, 1998.

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

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

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

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

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

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

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

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

[27] J. B. Lasserre. *Moments, positive polynomials and their applications*, volume 1. World Scientific, 2009.

[28] M. Laurent. Sums of squares, moment matrices and optimization over polynomials. In *Emerging applications of algebraic geometry*, pages 157–270. Springer, 2009.

[29] H. R. Lewis. Computers and intractability. a guide to the theory of np-completeness, 1983.

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

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

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

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

[34] J. Pearl. Principal stratification – a goal or a tool? *The International Journal of Biostatistics*, 7(1), 2011. Article 20, DOI: 10.2202/1557-4679.1322. Available at: <http://ftp.cs.ucla.edu/pub/stat_ser/r382.pdf>.

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

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

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

[38] D. Rosset, N. Gisin, and E. Wolfe. Universal bound on the cardinality of local hidden variables in networks. *Quantum Information & Computation*, 18(11-12):910–926, 2018.

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

[40] J. Sethuraman. A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650, 1994.

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

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

[43] J. Tian. *Studies in Causal Reasoning and Learning*. PhD thesis, Computer Science Department, University of California, Los Angeles, CA, November 2002.

[44] J. Tian and J. Pearl. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28:287–313, 2000.

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

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

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

[48] H. Waki, S. Kim, M. Kojima, and M. Muramatsu. Sums of squares and semidefinite program relaxations for polynomial optimization problems with structured sparsity. *SIAM Journal on Optimization*, 17(1):218–242, 2006.

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

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] "Throughout this paper, we assume that endogenous variables $\boldsymbol{V}$ are discrete and finite; while exogenous variables $\boldsymbol{U}$ could take any (continuous) value."
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This work does not present any foreseeable societal consequence.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Sec. 1.1.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendices A and B.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] We are in the process of translating the source code to other open-source platforms (e.g., Julia). We will release them if the paper is accepted.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix C.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix C. "Experiments were performed on a computer with 32GB memory."

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   (a) If your work uses existing assets, did you cite the creators? [Yes] "IST was a large, randomized, open trial of up to 14 days of antithrombotic therapy after stroke onset [10]." See also Appendix C

   (b) Did you mention the license of the assets? [Yes] See Appendix C. The IST dataset is shared under "Open Data Commons Attribution License (ODC-By) v1.0".

   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A On the Expressive Power of Discrete Structural Causal Models

In this section, we provide a detailed proof for Thm. 1 which establishes the expressive power of discrete SCMs in representing counterfactual distributions over finite observed domains. For convenience, we will focus on the following equivalent definition of discrete SCMs which will facilitate the understanding of the proof.

Definition 5. An SCM $M = \langle \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{F}, P \rangle$ is said to be a discrete SCM if

1. For each exogenous $U \in \boldsymbol{U}$, its domain Ω_U is discrete and at most countable;
2. For each endogenous $V \in \boldsymbol{V}$, its domain Ω_V is discrete and finite;
3. Values of each endogenous $V \in \boldsymbol{V}$ are given by $v \leftarrow h_{u_V}(pa_V)$ where h_{u_V} is a function mapping from finite domains of Pa_V to V.

For every $V \in \boldsymbol{V}$, we denote by $\mathscr{H}_V$ a hypothesis class containing all function mapping from domains of Pa_V to V, i.e., $\mathscr{H}_V = \Omega_{Pa_V} \mapsto \Omega_V$.

The main challenge in our proof is to show that given an arbitrary SCM M with arbitrary exogenous domains, one could construct a discrete SCM N, with bounded cardinality of exogenous domains, such that N and M induces the same counterfactual distributions and the causal diagram. To illustrate this idea, consider the sample "Bow" graph in Fig. 1d where X, Y are binary variables in $\{0, 1\}$. Since Y is not a descendant of X, counterfactual variable $X_y = X$ for any $y \in \Omega_Y$, i.e., intervening on Y has no causal effect on X [18]. It is thus sufficient to consider the counterfactual distribution $P(x, y_{x=0}, y_{x=1})$. Let functions in the hypothesis class $\mathcal{H}_X$ be ordered by $h_X^{(1)} = 0$ and $h_X^{(2)} = 1$; and let functions in the hypothesis class $\mathcal{H}_Y$ be ordered by:

$$h_Y^{(1)}(x) = 0, \qquad h_Y^{(2)}(x) = x, \qquad h_Y^{(3)}(x) = \neg x, \qquad h_Y^{(4)}(x) = 1. \tag{18}$$

Let $\mathscr{M}$ be the set of all SCMs compatible with $\mathcal{G}$ and let $\mathscr{N}$ be the set of all discrete SCMs compatible with $\mathcal{G}$ and discrete exogenous domain $|\Omega_U| \leq 8$. To prove the counterfactual equivalence between $\mathscr{M}$ and $\mathscr{N}$, it suffices to show that for any $M \in \mathcal{M}$, one could construct an $N \in \mathcal{N}$ so that $P_M(x, y_{x=0}, y_{x=1}) = P_N(x, y_{x=0}, y_{x=1})$. The construction procedure is described as follows. Let the exogenous U in N be a pair (U_X, U_Y) where $U_X \in \{1, 2\}$ and $U_Y \in \{1, \dots, 4\}$; values of X are given by $x \leftarrow h_X^{(u_X)}$; values of Y are given by $y \leftarrow h_Y^{(u_Y)}(x)$. It is verifiable that in such N, the counterfactual distribution $P(x, y_{x=0}, y_{x=1})$ equates to, for all $i, j, k \in \{0, 1\}$,

$$P_N(X = i, Y_{x=0} = j, Y_{x=1} = k) = P_N(U_X = i + 1, U_Y = 2j + k + 1). \tag{19}$$

For any SCM $M \in \mathcal{M}$, let the exogenous distribution $P_N(u_X, u_Y)$ be, for all $i, j, k \in \{0, 1\}$,

$$P_N(U_X = i + 1, U_Y = 2j + k + 1) = P_M(X = i, Y_{x=0} = j, Y_{x=1} = k). \tag{20}$$

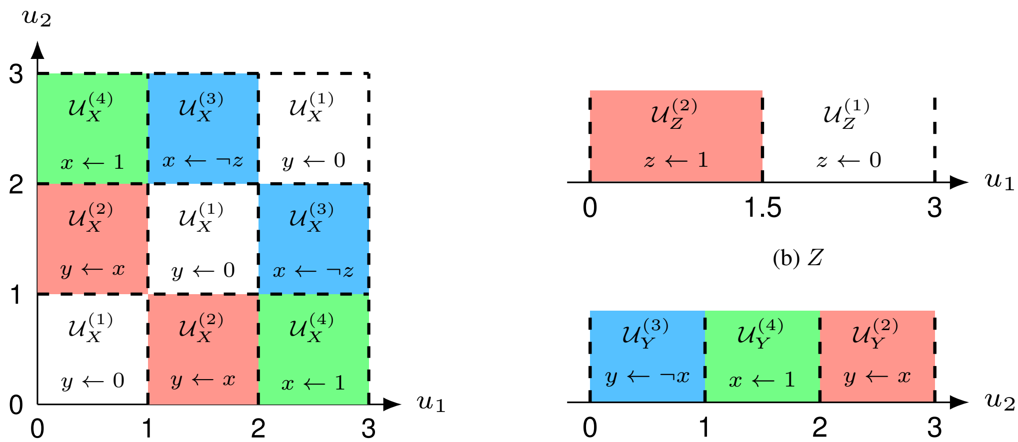
It follows from Eqs. (19) and (20) that M and N coincide in the counterfactual distribution $P(x, y_{x=0}, y_{x=1})$. That is, when inferring counterfactual distributions in Fig. 1d with binary X, Y, we could assume that the exogenous variable U is finite and discrete, without any loss of generality.

For the remainder of this section, we will generalize the construction described above to arbitrary causal diagrams. Our analysis rests on the framework of structural causal models and the measure-theoretic probability theory. Formally, each $U \in \boldsymbol{U}$ is associated with a probability space $\langle \Omega_U, \mathcal{F}_U, P_U \rangle$ where Ω_U is a sample space containing all possible outcomes; $\mathcal{F}_U$ is an event space containing subsets of Ω_U; and P_U is a probability measure mapping from events $\mathcal{F}_U$ to reals in $[0, 1]$. Values of exogenous variables $\boldsymbol{U}$ are drawn following the product measure $P \equiv \otimes_{U \in \boldsymbol{U}} P_U$. We refer readers to [6, 7] for a detailed introduction to the measure-theoretic probability theory.

A.1 Canonical Partitions of Exogenous Domains

Our proof for Thm. 1 relies on a family of canonical models which any SCM could be reduced to while maintaining counterfactual distributions and the network structure encoded in the induced causal diagram. Fix an endogenous $V \in \boldsymbol{V}$. 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 $\mathscr{H}_V$. Naturally, such a mapping leads to a finite partition over the exogenous domain Ω_{U_V}.

Definition 6. For an SCM $M = \langle \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{F}, P \rangle$, for each $V \in \boldsymbol{V}$, let functions in $\mathscr{H}_V$ be ordered by $\{h_V^{(i)}\}_{i \in \boldsymbol{I}_V}$ where $\boldsymbol{I}_V = \{1, \dots, m_V\}, m_V = |\mathscr{H}_V|$. A collection $\left\{\mathcal{U}_V^{(i)}\right\}_{i \in \boldsymbol{I}_V}$ is said to be *canonical partitions* of (exogenous domains of) V if for all $i \in \boldsymbol{I}_V, \mathcal{U}_V^{(i)} = \left\{\forall u_V \mid f_V(\cdot, u_V) = h_V^{(i)}\right\}$.



(a) X (b) Z (c) Y

Figure 5: Canonical partitions of exogenous domains of X, Y and Z. In (a), each canonical partition $\mathcal{U}_X^{(i)}$ is covered by a finite set of (almost) disjoint cells (e.g., $[2,3] \times [0,1]$).

As U_V varies 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 the class $\mathscr{H}_V$. Formally,

Lemma 2. *For an SCM $M = \langle \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{F}, P \rangle$, for each $V \in \boldsymbol{V}$, $f_V \in \boldsymbol{F}$ could be decomposed as:*

$$f_V(pa_V, u_V) = \sum_{i \in \boldsymbol{I}_V} h_V^{(i)}(pa_V) \mathbb{1}_{u_V \in \mathcal{U}_V^{(i)}}. \tag{21}$$

Proof. By the definition of the canonical partitions $\mathcal{U}_V^{(i)}$, $i = 1, \dots, m_V$, for any $u_V \in \mathcal{U}_V^{(r_V)}$, $f_V(\cdot, u_V) = h_V^{(r_V)}(\cdot)$. Fix $Pa_V = pa_V$. We have $f_V(pa_V, u_V) = h_V^{(r_V)}(pa_V)$. Since $\mathcal{U}_V^{(i)}$, $i = 1, \dots, m_V$, form a partition over domains Ω_{U_V}, given the same pa_V, u_V, the r.h.s. of Eq. (21) must equate to $h_V^{(r_V)}(pa_V)$, which completes the proof. □

As an example, consider an SCM M associated with the "Double bow" graph of Fig. 1b where X, Y, Z are binary variables in $\{0, 1\}$; U_1, U_2 are continuous values in $[0, 3]$. More specifically,

$$\begin{aligned} &U_i \sim \text{Unif}(0,3), i = 1, 2, && z \leftarrow f_Z(u_1) = \mathbb{1}_{u_1 \leq 1.5}, \\ &x \leftarrow f_X(z, u_1, u_2) = \mathbb{1}_{z \leq u_1 \leq z+2} \oplus \mathbb{1}_{z \leq u_2 \leq z+2}, && y \leftarrow f_Y(x, u_2) = \mathbb{1}_{x \leq u_2 \leq x+2}, \end{aligned} \tag{22}$$

where $\oplus$ is the "xor" operator. We show in Fig. 5 the canonical partitions induced by functions f_X, f_Y and f_Z respectively. To illustrate, Table 1 describes how the functional mapping between X and Y switches among $\mathcal{H}_Y$ as values of U_2 move across canonical partitions.

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

Table 1: Output of $f_Y(x, u_2)$ in Eq. (22). For any u_2, $f_Y(x, u_2)$ never equates to $h_Y^{(1)}(x) = 0$.

The decomposition of Lem. 2 implies that function f_Y could be written as follows:

$$f_Y(x, u_2) = \mathbb{1}_{u_2 \in [0,1)} x + \mathbb{1}_{u_2 \in [1,2]} \neg x + \mathbb{1}_{u_2 \in (2,3]} 1. \tag{23}$$

A natural question as this point is whether one could (1) discretize the exogenous domains of U_1, U_2 following canonical partitions of X, Y, Z and (2) replace the original U_1, U_2 with a discrete exogenous variable U with cardinality of $2 \times 4 \times 4 = 32$. Fig. 1c shows the causal diagram of the modified discrete SCM. However, such a discretization procedure does not maintain the network structure of the original causal diagram in Fig. 1b, thus failing to encoding some critical constraints over counterfactual distributions. For instance, variables Z and Y_x are independent since they are solutions of exogenous variables U_1 and U_2 respectively; U_1, U_2 are mutually independent. On the other hand, for any discrete SCM of Fig. 1c, such an independence relationship does not necessarily hold: Z and Y_x could be correlated since they are solutions of the same exogenous variable U.

A.2 Decomposing Canonical Partitions

Previous example calls for a more fine-grained decomposition of canonical partitions. To begin the discussion, we introduce a special type of subdomains called cells.

Definition 7 (Cell)**.** For an SCM $M = \langle \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{F}, P\rangle$, for each $V \in \boldsymbol{V}$, $\mathcal{R}_V$ is said to be a *cell* in domain Ω_{U_V} if $\mathcal{R}_V = \times_{U \in U_V} \mathcal{R}_{V,U}$ where $\mathcal{R}_{V,U} \subseteq \Omega_U$, for every $U \in \boldsymbol{U}$.

By definition, for $|U_V| = 1$, any subset of Ω_{U_V} is a cell (e.g., see Fig. 5). However, it is not always the case when $|U_V| \geq 2$. For instance, $\mathcal{U}_Y^{(4)}$ in Fig. 5a is not a cell. To see this, let $\mathcal{R}_{Y,U_1} = \mathcal{R}_{Y,U_2} = [0,1) \cup (2,3]$. It is verifiable that $\mathcal{U}_Y^{(4)} \neq \mathcal{R}_{Y,U_1} \times \mathcal{R}_{Y,U_2}$ since $\mathcal{R}_{Y,U_1} \times \mathcal{R}_{Y,U_2}$ consists of subsets $[0,1)^2$ and $(2,3]^2$ which is contained in $\mathcal{U}_Y^{(1)}$[4].

Arbitrary subsets A, B of an event space are said to be *almost disjoint* if their intersection has measure zero, i.e., $P(A \cap B) = 0$. Our next result shows that each canonical partition could be decomposed into a countable union of almost disjoint cells.

Definition 8 (Covering)**.** For an SCM $M = \langle \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{F}, P\rangle$, for any $V \in \boldsymbol{V}$, let $\mathcal{U}_V$ be an arbitrary subset of Ω_{U_V}. A countable set of cells $\left\{\mathcal{R}_V^{(j)}\right\}_{j \in \boldsymbol{J}_V}$ is said to be a *covering* of $\mathcal{U}_V$ if (1) for any $i \neq j$, $\mathcal{R}_V^{(i)}$ and $\mathcal{R}_V^{(j)}$ are almost disjoint; (2) $\mathcal{U}_V \subseteq \cup_{j \in \boldsymbol{J}_V} \mathcal{R}_V^{(j)}$; (3) $P(\mathcal{U}_V) = \sum_{j \in \boldsymbol{J}_V} P\left(\mathcal{R}_V^{(j)}\right)$.

Lemma 3. *For an SCM $M = \langle \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{F}, P\rangle$, there exists a covering $\left\{\mathcal{R}_V^{(j)}\right\}_{j \in \boldsymbol{J}_V}$ for each canonical partition $\mathcal{U}_V^{(i)}$, for any $i \in \boldsymbol{I}_V$, any $V \in \boldsymbol{V}$.*

Proof. We now consider a stronger statement showing that any subset $\mathcal{U}_V \subseteq \Omega_{U_V}$ has a covering. For any $\mathcal{A} \subseteq \Omega_{U_V}$, define a set of countable collections $\mathcal{C}(\mathcal{A})$ with cells $\mathcal{R}_V \in \Omega_{U_V}$:

$$\mathcal{C}(\mathcal{A}) = \{\mathcal{C} \subseteq \mathcal{F}_{U_V} \mid \mathcal{C} \text{ is at most countable and } \mathcal{A} \subseteq \cup_{\mathcal{R}_V \in \mathcal{C}} \mathcal{R}_V\}. \tag{24}$$

By definition of product measure P [6, Theorem 9.2], we have:

$$P(\mathcal{U}_V) = \inf \left\{ \sum_{\mathcal{R}_V \in \mathcal{C}} P(\mathcal{R}_V) \mid \forall \mathcal{C} \in \mathcal{C}(\mathcal{U}_V) \right\}. \tag{25}$$

We could thus obtain a countable set $\mathcal{C}$ of cells $\mathcal{R}_V \in \Omega_{U_V}$ such that

$$\mathcal{U}_V \subseteq \cup_{\mathcal{R}_V \in \mathcal{C}} \mathcal{R}_V, \qquad P(\mathcal{U}_V) = \sum_{\mathcal{R}_V \in \mathcal{C}} P(\mathcal{R}_V). \tag{26}$$

What remains is to show that every pair $\mathcal{R}_V^{(i)}, \mathcal{R}_V^{(j)} \in \mathcal{C}$ are almost disjoint. This is equivalent to proving the following statement:

$$P(\cup_{\mathcal{R}_V \in \mathcal{C}} \mathcal{R}_V) = \sum_{\mathcal{R}_V \in \mathcal{C}} P(\mathcal{R}_V). \tag{27}$$

It is sufficient to show that

$$P(\cup_{\mathcal{R}_V \in \mathcal{C}} \mathcal{R}_V) \geq \sum_{\mathcal{R}_V \in \mathcal{C}} P(\mathcal{R}_V). \tag{28}$$

Suppose now the above equating does not hold. There must exist a set $\mathcal{C}' \in \mathcal{C}(\cup_{\mathcal{R}_V \in \mathcal{C}} \mathcal{R}_V)$ such that

$$P(\cup_{\mathcal{R}_V \in \mathcal{C}} \mathcal{R}_V) = \sum_{\mathcal{R}_V \in \mathcal{C}'} P(\mathcal{R}_V) < \sum_{\mathcal{R}_V \in \mathcal{C}} P(\mathcal{R}_V). \tag{29}$$

[4] For convenience, we use $[a,b]^2$ to represent the Cartesian product of intervals $[a,b] \times [a,b]$.

By the definition of $\mathcal{C}(\mathcal{U}_V)$ in Eq. (24), we also have $\mathcal{C}' \in \mathcal{C}(\mathcal{U}_V)$. This means that

$$P(\mathcal{U}_V) \leq \sum_{\mathcal{R}_V \in \mathcal{C}'} P(\mathcal{R}_V) < \sum_{\mathcal{R}_V \in \mathcal{C}} P(\mathcal{R}_V), \tag{30}$$

which is a contradiction to Eq. (26). This means that set $\mathcal{C}$ forms a covering $\left\{\mathcal{R}_V^{(j)}\right\}_{j \in \boldsymbol{J}_V}$ over domains of $\mathcal{U}_V$, where $\boldsymbol{J}_V$ is a countable indexing set. □

Consider the partition $\mathcal{U}_X^{(1)}$ in Fig. 5. Let cells $\mathcal{R}_X^{(j)} = [j-1, j]^2$, $j = 1, 2, 3$. It is verifiable that $\mathcal{U}_X^{(1)} \subseteq \cup_{j=1,2,3} \mathcal{R}_X^{(j)}$. Since finite points in $\Omega_{U_1} \times \Omega_{U_2}$ (e.g., $u_1 = u_2 = 1$) has measure zero,

$$P\left(\mathcal{U}_X^{(1)}\right) = P\left((U_1, U_2) \in [0,1)^2 \cup [1,2]^2 \cup (2,3]^2\right) = \sum_{j=1,2,3} P\left(\mathcal{R}_X^{(j)}\right). \tag{31}$$

By Def. 8, $\left\{\mathcal{R}_X^{(1)}, \mathcal{R}_X^{(2)}, \mathcal{R}_X^{(3)}\right\}$ is thus a covering of $\mathcal{U}_X^{(1)}$. The characterization of canonical partitions and coverings permits us to decompose counterfactual distributions in the canonical form as follows.

Lemma 4. *For an SCM $M = \langle \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{F}, P\rangle$, let $\boldsymbol{I} = \times_{V \in \boldsymbol{V}} \boldsymbol{I}_V$. For $\boldsymbol{Y}, \dots, \boldsymbol{Z}, \boldsymbol{X}, \dots, \boldsymbol{W} \subseteq \boldsymbol{V}$*[5]*,*

$$P(\boldsymbol{y}_{\boldsymbol{x}}, \dots, \boldsymbol{z}_{\boldsymbol{w}}) = \sum_{\boldsymbol{i}} \mathbb{1}_{\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{i}) = \boldsymbol{y}} \wedge \cdots \wedge \mathbb{1}_{\boldsymbol{Z}_{\boldsymbol{w}}(\boldsymbol{i}) = \boldsymbol{z}} P\left(\bigwedge_{V \in \boldsymbol{V}} \mathcal{U}_V^{(i)}\right), \tag{32}$$

where variables of the form $\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{i})$ is defined as:

$$\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{i}) = \{Y_{\boldsymbol{x}}(\boldsymbol{i}) \mid \forall Y \in \boldsymbol{Y}\} \text{ where } Y_{\boldsymbol{x}}(\boldsymbol{i}) = \begin{cases} \boldsymbol{x}_Y & \text{if } Y \in \boldsymbol{X} \\ h_Y^{(i)}\left(\{V_{\boldsymbol{x}}(\boldsymbol{i}) \mid V \in Pa_Y\}\right) & \text{otherwise} \end{cases}$$

Moreover, let $\left\{\mathcal{R}_V^{(j)}\right\}_{j \in \boldsymbol{J}_V}$ is a covering of each canonical partition $\mathcal{U}_V^{(i)}$; and let $\boldsymbol{J} = \times_{V \in \boldsymbol{V}} \boldsymbol{J}_V$. The above equation could be further written as, for any $\boldsymbol{i} \in \boldsymbol{I}$,

$$P\left(\bigwedge_{V \in \boldsymbol{V}} \mathcal{U}_V^{(i)}\right) = \sum_{\boldsymbol{j} \in \boldsymbol{J}} P\left(\bigwedge_{V \in \boldsymbol{V}} \mathcal{R}_V^{(j)}\right) = \sum_{\boldsymbol{j} \in \boldsymbol{J}} \prod_{U \in \boldsymbol{U}} P\left(\bigwedge_{V \in ch(U)} \mathcal{R}_{V,U}^{(j)}\right), \tag{33}$$

where $ch(U)$ are child nodes of U in DAG $\mathcal{G}$, i.e., $ch(U) = \{\forall V \in \boldsymbol{V} \mid U \in U_V\}$.

Proof. We first show that for any $\boldsymbol{Y}, \boldsymbol{X} \subseteq \boldsymbol{V}$, given any $\boldsymbol{u}, \boldsymbol{x}, *y$,

$$\mathbb{1}_{\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{u}) = \boldsymbol{y}} = \sum_{\boldsymbol{i} \in \boldsymbol{I}} \mathbb{1}_{\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{i}) = \boldsymbol{y}} \prod_{V \in \boldsymbol{V}} \mathbb{1}_{u_V \in \mathcal{U}_V^{(i)}}. \tag{34}$$

Let $\mathcal{G}_{\overline{\boldsymbol{X}}}$ be a subgraph obtained from the causal diagram $\mathcal{G}$ by removing all incoming arrows of $\boldsymbol{X}$. For any $Y \in \boldsymbol{Y}$, let $An(Y)_{\mathcal{G}}$ be the set of ancestor nodes of Y in a DAG $\mathcal{G}$, including Y itself. We will prove Eq. (34) by induction on $n = \max_{Y \in \boldsymbol{Y}} \left|An(Y)_{\mathcal{G}_{\overline{\boldsymbol{X}}}}\right|$.

Base Case $n = 1$. In this case, for $Y \in \boldsymbol{X} \cap \boldsymbol{Y}$, $\mathbb{1}_{Y_{\boldsymbol{x}}(\boldsymbol{u}) = y} = \mathbb{1}_{y = \boldsymbol{x}_Y}$ where $\boldsymbol{x}_Y$ be the values assigned to Y in $\boldsymbol{x}$. For $Y \in \boldsymbol{Y} \setminus \boldsymbol{X}$, we must have $Pa_Y = \emptyset$. This implies

$$\mathbb{1}_{Y_{\boldsymbol{x}}(\boldsymbol{u}) = y} = \mathbb{1}_{f_Y(u_Y) = y} \tag{35}$$

$$= \mathbb{1}_{y = \sum_{i \in \boldsymbol{I}_Y} h_Y^{(i)} \mathbb{1}_{u_Y \in \mathcal{U}_Y^{(i)}}} \qquad \text{\# By Lem. 2} \tag{36}$$

$$= \sum_{i \in \boldsymbol{I}_Y} \mathbb{1}_{h_Y^{(i)} = y} \mathbb{1}_{u_Y \in \mathcal{U}_Y^{(i)}} \tag{37}$$

[5] For any index sequence $\boldsymbol{i} \in \boldsymbol{I}$, we use i_V to represent the element in $\boldsymbol{i}$ with restriction to $V \in \boldsymbol{V}$. We omit the subscript V when it is obvious; therefore, $\mathcal{U}_V^{(i)} = \mathcal{U}_V^{(i_V)}$, $h_V^{(i)} = h_V^{(i_V)}$. The same applies to $\boldsymbol{j} \in \boldsymbol{J}$.

The above equation implies

$$\mathbb{1}_{\boldsymbol{Y_x}(\boldsymbol{u})=\boldsymbol{y}} = \prod_{Y\in\boldsymbol{Y}\cap\boldsymbol{X}} \mathbb{1}_{y=\boldsymbol{x}_Y} \prod_{Y\in(\boldsymbol{Y}\setminus\boldsymbol{X})} \sum_{i\in\boldsymbol{I}_Y} \mathbb{1}_{h_Y^{(i)}=y} \mathbb{1}_{u_Y\in\mathcal{U}_Y^{(i)}} \tag{38}$$

$$= \sum_{\boldsymbol{i}\in\boldsymbol{I}} \prod_{Y\in\boldsymbol{Y}\cap\boldsymbol{X}} \mathbb{1}_{y=\boldsymbol{x}_Y} \prod_{Y\in(\boldsymbol{Y}\setminus\boldsymbol{X})} \mathbb{1}_{h_Y^{(i)}=y} \prod_{V\in\boldsymbol{V}} \mathbb{1}_{u_V\in\mathcal{U}_V^{(i)}} \tag{39}$$

$$= \sum_{\boldsymbol{i}\in\boldsymbol{I}} \mathbb{1}_{\boldsymbol{Y_x}(\boldsymbol{i})=\boldsymbol{y}} \prod_{V\in\boldsymbol{V}} \mathbb{1}_{u_V\in\mathcal{U}_V^{(i)}}. \tag{40}$$

The last step follows from the definition of variables $\boldsymbol{Y_x}(\boldsymbol{i})$ given index $\boldsymbol{i}\in\boldsymbol{I}$.

Induction Case $n = k+1$. Assume that Eq. (34) hols for $n = k$. We will prove for the case $n = K + 1$. For $Y \in \boldsymbol{X}\cap\boldsymbol{Y}$, $\mathbb{1}_{Y_{\boldsymbol{x}}(\boldsymbol{u})=y} = \mathbb{1}_{y=\boldsymbol{x}_Y}$. For $Y\in\boldsymbol{Y}\setminus\boldsymbol{X}$, the decomposition in Lem. 2 implies:

$$\mathbb{1}_{Y_{\boldsymbol{x}}(\boldsymbol{u})=y} = \mathbb{1}_{f_Y(\{V_{\boldsymbol{x}}(\boldsymbol{u})|V\in Pa_Y\},u_Y)=y} \tag{41}$$

$$= \mathbb{1}_{y=\sum_{i\in\boldsymbol{I}_Y} h_Y^{(i)}(\{V_{\boldsymbol{x}}(\boldsymbol{u})|V\in Pa_Y\})\mathbb{1}_{u_Y\in\mathcal{U}_Y^{(i)}}} \tag{42}$$

$$= \sum_{i\in\boldsymbol{I}_Y} \sum_{pa_Y} \mathbb{1}_{h_Y^{(i)}(pa_Y)=y} \mathbb{1}_{\{V_{\boldsymbol{x}}(\boldsymbol{u})|V\in Pa_Y\}=pa_Y} \mathbb{1}_{u_Y\in\mathcal{U}_Y^{(i)}}. \tag{43}$$

Since Eq. (34) holds for Case $n = k$, the above equation could be further written as

$$\mathbb{1}_{Y_{\boldsymbol{x}}(\boldsymbol{u})=y} = \sum_{i\in\boldsymbol{I}_Y} \sum_{pa_Y} \mathbb{1}_{h_Y^{(i)}(pa_Y)=y} \mathbb{1}_{u_Y\in\mathcal{U}_Y^{(i)}} \sum_{\boldsymbol{i}\in\boldsymbol{I}} \mathbb{1}_{\{V_{\boldsymbol{x}}(\boldsymbol{i})|V\in Pa_Y\}=pa_Y} \prod_{V\in\boldsymbol{V}} \mathbb{1}_{u_V\in\mathcal{U}_V^{(i)}} \tag{44}$$

$$= \sum_{\boldsymbol{i}\in\boldsymbol{I}} \sum_{pa_Y} \mathbb{1}_{h_Y^{(i)}(pa_Y)=y} \mathbb{1}_{\{V_{\boldsymbol{x}}(\boldsymbol{i})|V\in Pa_Y\}=pa_Y} \prod_{V\in\boldsymbol{V}} \mathbb{1}_{u_V\in\mathcal{U}_V^{(i)}} \tag{45}$$

$$= \sum_{\boldsymbol{i}\in\boldsymbol{I}} \mathbb{1}_{h_Y^{(i)}(\{V_{\boldsymbol{x}}(\boldsymbol{i})|V\in Pa_Y\})=y} \prod_{V\in\boldsymbol{V}} \mathbb{1}_{u_V\in\mathcal{U}_V^{(i)}}. \tag{46}$$

We thus have

$$\mathbb{1}_{\boldsymbol{Y_x}(\boldsymbol{u})=\boldsymbol{y}} = \prod_{Y\in\boldsymbol{Y}\cap\boldsymbol{X}} \mathbb{1}_{y=\boldsymbol{x}_Y} \prod_{Y\in(\boldsymbol{Y}\setminus\boldsymbol{X})} \sum_{\boldsymbol{i}\in\boldsymbol{I}} \mathbb{1}_{h_Y^{(i)}(\{V_{\boldsymbol{x}}(\boldsymbol{i})|V\in Pa_Y\})=y} \prod_{V\in\boldsymbol{V}} \mathbb{1}_{u_V\in\mathcal{U}_V^{(i)}} \tag{47}$$

$$= \sum_{\boldsymbol{i}\in\boldsymbol{I}} \prod_{Y\in\boldsymbol{Y}\cap\boldsymbol{X}} \mathbb{1}_{y=\boldsymbol{x}_Y} \prod_{Y\in(\boldsymbol{Y}\setminus\boldsymbol{X})} \mathbb{1}_{h_Y^{(i)}(\{V_{\boldsymbol{x}}(\boldsymbol{i})|V\in Pa_Y\})=y} \prod_{V\in\boldsymbol{V}} \mathbb{1}_{u_V\in\mathcal{U}_V^{(i)}} \tag{48}$$

$$= \sum_{\boldsymbol{i}\in\boldsymbol{I}} \mathbb{1}_{\boldsymbol{Y_x}(\boldsymbol{i})=\boldsymbol{y}} \prod_{V\in\boldsymbol{V}} \mathbb{1}_{u_V\in\mathcal{U}_V^{(i)}}. \tag{49}$$

The last step follows from the definition of variables $\boldsymbol{Y_x}(\boldsymbol{i})$ given index $\boldsymbol{i}\in\boldsymbol{I}$.

We now consider the proof of Eq. (32). The statement of Eq. (34) implies that for any $\boldsymbol{Y},\dots,\boldsymbol{Z},\boldsymbol{X},\dots,\boldsymbol{W}\subseteq\boldsymbol{V}$,

$$P\left(\boldsymbol{y_x},\dots,\boldsymbol{z_w}\right) = \int_{\Omega_{\boldsymbol{U}}} \mathbb{1}_{\boldsymbol{Y_x}(\boldsymbol{u})=\boldsymbol{y}} \wedge \dots \wedge \mathbb{1}_{\boldsymbol{Z_w}(\boldsymbol{u})=\boldsymbol{z}} dP(\boldsymbol{u}) \tag{50}$$

$$= \int_{\Omega_{\boldsymbol{U}}} \left(\sum_{\boldsymbol{i}\in\boldsymbol{I}} \mathbb{1}_{\boldsymbol{Y_x}(\boldsymbol{i})=\boldsymbol{y}} \prod_{V\in\boldsymbol{V}} \mathbb{1}_{u_V\in\mathcal{U}_V^{(i)}}\right) \wedge \dots \wedge \left(\sum_{\boldsymbol{i}\in\boldsymbol{I}} \mathbb{1}_{\boldsymbol{Z_w}(\boldsymbol{i})=\boldsymbol{z}} \prod_{V\in\boldsymbol{V}} \mathbb{1}_{u_V\in\mathcal{U}_V^{(i)}}\right) dP(\boldsymbol{u}) \tag{51}$$

$$= \int_{\Omega_{\boldsymbol{U}}} \sum_{\boldsymbol{i}\in\boldsymbol{I}} \mathbb{1}_{\boldsymbol{Y_x}(\boldsymbol{i})=\boldsymbol{y}} \wedge \dots \wedge \mathbb{1}_{\boldsymbol{Z_w}(\boldsymbol{i})=\boldsymbol{z}} \prod_{V\in\boldsymbol{V}} \mathbb{1}_{u_V\in\mathcal{U}_V^{(i)}} dP(\boldsymbol{u}) \tag{52}$$

$$= \sum_{\boldsymbol{i}\in\boldsymbol{I}} \mathbb{1}_{\boldsymbol{Y_x}(\boldsymbol{i})=\boldsymbol{y}} \wedge \dots \wedge \mathbb{1}_{\boldsymbol{Z_w}(\boldsymbol{i})=\boldsymbol{z}} \int_{\Omega_{\boldsymbol{U}}} \prod_{V\in\boldsymbol{V}} \mathbb{1}_{u_V\in\mathcal{U}_V^{(i)}} dP(\boldsymbol{u}) \tag{53}$$

$$= \sum_{\boldsymbol{i}\in\boldsymbol{I}} \mathbb{1}_{\boldsymbol{Y_x}(\boldsymbol{i})=\boldsymbol{y}} \wedge \dots \wedge \mathbb{1}_{\boldsymbol{Z_w}(\boldsymbol{i})=\boldsymbol{z}} P\left(\bigwedge_{V\in\boldsymbol{V}} \mathcal{U}_V^{(i)}\right). \tag{54}$$

What remains is to prove Eq. (33). We first show that, for any $\mathcal{A} \in \mathcal{F}$,

$$P\left(\mathcal{U}_V^{(i)} \wedge \mathcal{A}\right) = \sum_{j \in \boldsymbol{J}_V} P\left(\mathcal{R}_V^{(i)} \wedge \mathcal{A}\right). \tag{55}$$

Let $\mathcal{A}^{\complement} = \Omega \setminus \mathcal{A}$. Since $\left\{\mathcal{R}_V^{(j)}\right\}_{j \in \boldsymbol{J}_V}$ is a covering of $\mathcal{U}_V^{(i)}$, we have $\mathcal{U}_V^{(i)} \subseteq \cup_{j \in \boldsymbol{J}_V} \mathcal{R}_V^{(j)}$. This implies

$$P\left(\mathcal{U}_V^{(i)} \wedge \mathcal{A}\right) \leq \sum_{j \in \boldsymbol{J}_V} P\left(\mathcal{R}_V^{(j)} \wedge \mathcal{A}\right), \qquad P\left(\mathcal{U}_V^{(i)} \wedge \mathcal{A}^{\complement}\right) \leq \sum_{j \in \boldsymbol{J}_V} P\left(\mathcal{R}_V^{(j)} \wedge \mathcal{A}^{\complement}\right). \tag{56}$$

We will next show that the above inequality relationships are both tight. Suppose say, the inequality in Eq. (55) is strict. We must have

$$P\left(\mathcal{U}_V^{(i)}\right) = P\left(\mathcal{U}_V^{(i)} \wedge \mathcal{A}\right) + P\left(\mathcal{U}_V^{(i)} \wedge \mathcal{A}^{\complement}\right) \tag{57}$$

$$< \sum_{j \in \boldsymbol{J}_V} P\left(\mathcal{R}_V^{(j)} \wedge \mathcal{A}\right) + \sum_{j \in \boldsymbol{J}_V} P\left(\mathcal{R}_V^{(j)} \wedge \mathcal{A}^{\complement}\right). \tag{58}$$

The above equation implies

$$P\left(\mathcal{U}_V^{(i)}\right) < \sum_{j \in \boldsymbol{J}_V} P\left(\mathcal{R}_V^{(j)}\right), \tag{59}$$

which is a contradiction. The property of Eq. (55) implies, for any $\boldsymbol{i} \in \boldsymbol{I}$,

$$P\left(\bigwedge_{V \in \boldsymbol{V}} \mathcal{U}_V^{(i)}\right) = \sum_{\boldsymbol{j} \in \boldsymbol{J}} P\left(\bigwedge_{V \in \boldsymbol{V}} \mathcal{R}_V^{(j)}\right). \tag{60}$$

Since each cell $\mathcal{R}_V^{(j)}$ is a Cartesian product of subsets $\times_{U \in U_V} \mathcal{R}_{V,U}^{(j)}$ of each exogenous domains and exogenous variables in $\boldsymbol{U}$ are mutually independent, we must have, for any $\boldsymbol{j} \in \boldsymbol{J}$,

$$P\left(\bigwedge_{V \in \boldsymbol{V}} \mathcal{R}_V^{(j)}\right) = \prod_{U \in \boldsymbol{U}} P\left(\bigwedge_{V \in ch(U)} \mathcal{R}_{V,U}^{(j)}\right). \tag{61}$$

The above equations together prove Eq. (33). □

Consider again the SCM M described in Eq. (22). Note that the only function in the hypothesis class $\mathcal{H}_Z$ compatible with event $Z = 1$ is $h_Z^{(2)} = 1$. Similarly, event $X_{z=0} = 0, X_{z=1} = 0$ corresponds to the function $h_X^{(1)}(z) = 0$ in $\mathcal{H}_X$. Applying the decomposition of Eq. (32) gives:

$$P\left(Z = 1, X_{z=0} = 0, X_{z=1} = 0\right) = \sum_{i=1,\ldots,4} P\left(\mathcal{U}_Z^{(2)} \wedge \mathcal{U}_X^{(1)} \wedge \mathcal{U}_Y^{(i)}\right) = P\left(\mathcal{U}_Z^{(2)} \wedge \mathcal{U}_X^{(1)}\right). \tag{62}$$

Among above quantities, the canonical partition $\mathcal{U}_Z^{(2)} = \{u_1 \in [0, 1.5]\}$ is a cell. $\mathcal{U}_X^{(1)}$ has a covering of $\left\{(u_1, u_2) \in \mathcal{R}_X^{(j)} \mid j = 1, 2, 3\right\}$ where $\mathcal{R}_X^{(j)} = [j-1, j]^2$. Eq. (33) implies

$$\begin{aligned} P\left(\mathcal{U}_Z^{(2)} \wedge \mathcal{U}_X^{(1)}\right) &= \sum_{j=1,2,3} P\left(U_1 \in [0, 1.5] \wedge (U_1, U_2) \in [j-1, j]^2\right) \\ &= P\left(U_1 \in [0, 1]\right) P\left(U_2 \in [0, 1]\right) + P\left(U_1 \in [1, 1.5]\right) P\left(U_2 \in [1, 2]\right). \end{aligned} \tag{63}$$

Computing Eqs. (62) and (63) gives $P\left(Z = 1, X_{z=0} = 0, X_{z=1} = 0\right) = 1/6$. One could verify this answer from the parametrization of SCM M in Eq. (22) using the three-step algorithm introduced in [33] which consists of abduction, action, and prediction.

A.3 Bounding Cardinalities of Exogenous Domains

The decomposition in Lem. 4 implies a discretization procedure that could reproduce all counterfactual distributions in any SCM $M = \langle \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{F}, P\rangle$. First, we decompose the exogenous domain Ω_{U_V} for each $V \in \boldsymbol{V}$ into the canonical partitions. Second, we further decompose each canonical partition using its covering. By doing so, we obtain a partition over the exogenous domain Ω_{U_V} which consists of countably many (almost) disjoint cells; each cell is assigned with a function (say, h_V) in the hypothesis class $\mathscr{H}_V$. Finally, for each configuration $U_V = u_V$, we find the cell partition containing u_V and generate values of V using the associated function h_V. We formalize this data-generating process using a canonical family of SCMs described as follows.

Definition 9. An SCM $M = \langle \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{F}, P\rangle$ is said to be a *canonical SCM* if for each $V \in \boldsymbol{V}$, let $\left\{\mathcal{R}_V^{(j)}\right\}_{j\in \boldsymbol{J}_V}$ be a covering of Ω_{U_V}; function $f_V \in \boldsymbol{F}$ is given by, for $i_j \in \{1, \dots, m_V\}, j \in \boldsymbol{J}_V$,

$$f_V(pa_V, u_V) = \sum_{j\in \boldsymbol{J}_V} h_V^{(i_j)}(pa_V)\mathbb{1}_{u_V \in \mathcal{R}_V^{(j)}}. \tag{64}$$

Consider the SCM M described in Eq. (22) as an example. Let N be a canonical SCM compatible with the DAG of Fig. 1b; its covering cells (e.g., $\mathcal{R}_X^{(j)}$) and corresponding functions ($h_X^{(j_i)}(z)$) associated with X, Y, Z are graphically described in Fig. 5 respectively. It immediately follows from Lem. 4 that M and N generate the same collection of counterfactual distributions $\boldsymbol{P}^*$.

Lemma 5. *For a DAG $\mathcal{G}$, let M be an arbitrary SCM compatible with $\mathcal{G}$. There exists a canonical SCM N compatible with $\mathcal{G}$ such that $\boldsymbol{P}^*_M = \boldsymbol{P}^*_N$, i.e., they coincide in all counterfactual distributions.*

Proof. For each $V \in \boldsymbol{V}$ in SCM M, let $\left\{\mathcal{R}_V^{(j)}\right\}_{j\in \boldsymbol{J}_V^{(i)}}$ denote a covering for a canonical partition $\mathcal{U}_V^{(i)}$, $i \in \boldsymbol{I}_V$. Since $\{\mathcal{U}_V^{(i)}\}_{i\in \boldsymbol{I}_V}$ forms a partition over the exogenous domain Ω_{U_V}. The collection $\left\{\mathcal{R}_V^{(j)} \mid j \in \boldsymbol{J}_V^{(i)}, V \in \boldsymbol{V}\right\}$ forms a covering over Ω_{U_V}. Let $\boldsymbol{J}_V$ be the union of indexing set $\cup_{i\in \boldsymbol{I}_V} \boldsymbol{J}_V^{(i)}$. Naturally, any element $j \in \boldsymbol{J}_V$ must belong to a subset $\boldsymbol{J}_V^{(i)}$; let i_j denote such index i. We construct a canonical SCM N using coverings $\left\{\mathcal{R}_V^{(j)}\right\}_{j\in \boldsymbol{J}_V}$ and index i_j described previously. Let $\boldsymbol{J} = \times_{V\in \boldsymbol{V}} \boldsymbol{J}_V$. For any $\boldsymbol{Y}, \dots, \boldsymbol{Z}, \boldsymbol{X}, \dots, \boldsymbol{W} \subseteq \boldsymbol{V}$, the counterfactual distribution $P(\boldsymbol{y_x}, \dots, \boldsymbol{z_w})$ in the canonical SCM N is equal to

$$P(\boldsymbol{y_x}, \dots, \boldsymbol{z_w}) = \sum_{\boldsymbol{j}\in \boldsymbol{J}} \mathbb{1}_{\boldsymbol{Y_x}(\boldsymbol{i_j})=\boldsymbol{y}} \wedge \dots \wedge \mathbb{1}_{\boldsymbol{Z_w}(\boldsymbol{i_j})=\boldsymbol{z}} P\left(\bigwedge_{V\in \boldsymbol{V}} \mathcal{R}_V^{(j)}\right), \tag{65}$$

where $\boldsymbol{i_j}$ is the indexing sequence $(i_j)_{j\in \boldsymbol{j}}$. Lem. 4, together with some reordering over indices in $\boldsymbol{i_j}$, implies that M and N induce the same collection of counterfactual distributions. ☐

Given a canonical SCM, one could immediately obtain a discrete SCM by discretizing exogenous domains following the covering cells. Since each cell is a Caresian product of subsets (Def. 7), the resulting discrete model must induce a causal diagram with the same network structure.

Lemma 6. *For a DAG $\mathcal{G}$, consider the following conditions: (1) $\mathscr{M}$ is the set of all SCMs compatible with $\mathcal{G}$; (2) $\mathscr{N}$ is the set of all discrete SCMs compatible with $\mathcal{G}$. Then, $\mathscr{M}$ and $\mathscr{N}$ are counterfactually equivalent.*

Proof. For any cell $\mathcal{R}_V^{(j)} = \times_{U\in U_V} \mathcal{R}_{V,U}^{(j)}$, we call $\mathcal{R}_{V,U}^{(j)}$ the projection of $\mathcal{R}_V^{(j)}$ to domains of U. We will describe a discretization procedure that discretize domains of each $U \in \boldsymbol{U}$ following the intersections of projections $\cap_{V\in ch(U)} \mathcal{R}_{V,U}^{(j)}, \forall j \in \boldsymbol{J}_V$. For each $V \in ch(U)$, for any infinite binary sequence $r_{V,U} \in \{0, 1\}^{\boldsymbol{J}_V}$, let an event $\mathcal{A}_{r_{V,U}^{(j)}} \in \mathcal{F}_{U_k}$ be, for $j \in \boldsymbol{J}_V$,

$$\mathcal{A}_{r_{V,U}^{(j)}} = \begin{cases} \mathcal{R}_{V,U}^{(j)} & \text{if } r_{V,U}^{(j)} = 1 \\ \Omega_U \setminus \mathcal{R}_{V,U}^{(j)} & \text{if } r_{V,U}^{(j)} = 0. \end{cases} \tag{66}$$

For any $r_U = \{r_{V,U} : V \in ch(U)\}$, let a subset $\mathcal{A}_{r_U} \in \Omega_U$ be

$$\mathcal{A}_{r_U} = \bigcap_{V \in ch(U)} \bigcap_{j \in \boldsymbol{J}_V} \mathcal{A}_{r^{(j)}_{V,U}}. \tag{67}$$

Since $\mathcal{A}_{r_{V,U}}$, $\forall r_U$, enumerates all possible intersections of projections $\mathcal{R}^{(j)}_{V,U}$, we could obtain probabilities over any intervention $\cap_{V \in ch(U)} \mathcal{R}^{(j)}_{V,U}$ using the join probability $P(\mathcal{A}_{r_U})$.

It now suffices to show that distribution $P(\mathcal{A}_{r_U})$ has countable support, i.e., the set $\mathcal{A}_U = \{\mathcal{A}_{r_U} : P(\mathcal{A}_{r_U}) > 0\}$ has at most countably elements. Since P is a probability measurable, $P(\mathcal{A}_{r_k}) \in [0, 1]$. By the construction of Eq. (66), we must have $\sum_{r_U} P(\mathcal{A}_{r_U}) = 1$. If the sum over an uncountable set of reals is finite, then there exist at most countable number of events $\mathcal{A}_{r_U}$ such that $P(\mathcal{A}_{r_U}) > 0$, i.e, the set $\mathcal{A}_U$ is countable. $\square$

Lem. 6 implies that one could represent all counterfactual distributions in a causal diagram using a countably infinite number of exogenous states. To prove Thm. 1, what remains is to bound the cardinality of the exogenous domain. More specifically, we will show that any discrete SCM M with cardinality $|\Omega_U| > \prod_{V \in \boldsymbol{C}_U} |\mathscr{H}_V|$, $\forall U \in \boldsymbol{U}$, $\boldsymbol{C}_U$ is the c-component that contains all child nodes of U, can be modified into a discrete SCM N with $|\Omega_U| \leq \prod_{V \in \boldsymbol{C}_U} |\mathscr{H}_V|$, $\forall U \in \boldsymbol{U}$, while maintaining all counterfactual distributions $\boldsymbol{P}^*$ and the same network structure in the causal diagram.

Theorem 1. *For a DAG $\mathcal{G}$, consider the following conditions*[6]*: (1) $\mathscr{M}$ is the set of all SCMs compatible with $\mathcal{G}$; (2) $\mathscr{N}$ is the set of all discrete SCMs compatible with $\mathcal{G}$ where for every $U \in \boldsymbol{U}$, its cardinality $|\Omega_U| = \prod_{V \in \boldsymbol{C}_U} |\Omega_{Pa_V} \mapsto \Omega_V|$, i.e., the number of functions mapping from Pa_V to V for every variable V in the c-component $\boldsymbol{C}_U$. Then, $\mathscr{M}$ and $\mathscr{N}$ are counterfactually equivalent.*

Proof. Lem. 4 implies that it suffices to prove that for any discrete SCM $M \in \mathcal{M}$, there exists a finite SCM $N \in \mathcal{N}$ such that M and N coincide in the joint distribution over canonical partitions $P\left(\bigwedge_{V \in \boldsymbol{V}} \mathcal{U}^{(i)}_V\right)$. C-components in $\mathcal{C}(\mathcal{G})$ implies the following decomposition

$$P\left(\bigwedge_{V \in \boldsymbol{V}} \mathcal{U}^{(i)}_V\right) = \prod_{\boldsymbol{C} \in \mathcal{C}(\mathcal{G})} P\left(\bigwedge_{V \in \boldsymbol{C}} \mathcal{U}^{(i)}_V\right). \tag{68}$$

We now focus on the consistency for the joint probability $P\left(\bigwedge_{V \in \boldsymbol{C}} \mathcal{U}^{(i)}_V\right)$ for each $\boldsymbol{C} \in \mathcal{C}(\mathcal{G})$.

Fix a c-component $\boldsymbol{C}$. Let $\vec{P}$ be a vector representing probabilities of $\left(P\left(\bigwedge_{V \in \boldsymbol{C}} \mathcal{U}^{(i)}_V\right)\right)_{\boldsymbol{i} \in \boldsymbol{I}}$, which could be seen as a point in $d-1$-dimensional real space where $d = \prod_{V \in \boldsymbol{C}} |\mathscr{H}_V|$[7]. Let $U_{\boldsymbol{C}}$ denote the collection $\cup_{V \in \boldsymbol{C}} U_V$. Fix an exogenous $U \in U_{\boldsymbol{C}}$. Let $P_u\left(\bigwedge_{V \in \boldsymbol{C}} \mathcal{U}^{(i)}_V\right)$ denote joint distributions over canonical partitions when U is fixed as a constant $u \in \Omega_U$. More specifically,

$$P_u\left(\bigwedge_{V \in \boldsymbol{C}} \mathcal{U}^{(i)}_V\right) = \sum_{\boldsymbol{u} \setminus u} \prod_{V \in \boldsymbol{C}} \mathbb{1}_{u_v \in \mathcal{U}^{(i)}_V} \prod_{U' \in (\boldsymbol{U} \setminus U)} P(u'). \tag{69}$$

Similarly, let $\vec{P}_u$ be a vector in $\mathbb{R}^{d-1}$ representing probabilities of $P_u\left(\bigwedge_{V \in \boldsymbol{C}} \mathcal{U}^{(i)}_V\right)$. By basic probabilistic operations, we must have $\vec{P} = \sum_u \vec{P}_u P(u)$. That is, $\vec{P} \in \mathbb{R}^{d-1}$ is a point lies in the convex hull of a set $\left\{\vec{P}_u \mid \forall u \in \Omega_U\right\}$. The Carathéodory theorem [9, 13] implies that we could write $\vec{P}$ as a convex combination of at most d points in $\left\{\vec{P}_u \mid \forall u \in \Omega_U\right\}$. That is, for d distinct values $\{u_1, \ldots, u_d\}$ in Ω_U,

$$\vec{P} = \sum_{k=1}^{d} w_d \vec{P}_{u_k}, \qquad \text{where } w_k > 0, \forall k = 1, \ldots, d, \text{ and } \sum_k w_k = 1. \tag{70}$$

[6]For every $V \in \boldsymbol{V}$, $\Omega_{Pa_V} \mapsto \Omega_V$ is the set of all functions mapping from domains Ω_{Pa_V} to Ω_V.

[7]By definition, $\vec{P}$ is a vector with $d = \prod_{V \in \boldsymbol{C}} |\mathscr{H}_V|$ elements. Since $\sum_{\boldsymbol{i}} P\left(\bigwedge_{V \in \boldsymbol{C}} \mathcal{U}^{(i)}_V\right) = 1$, it only takes a vector with $d-1$ dimensions to uniquely determine $\vec{P}$.

We could replace $P(u)$ with a distribution $P'(u_k) = w_k$ over a finite discrete domain $\Omega'_U = \{u_1, \dots, u_d\}$ and obtain a discrete SCM N that reproduce all counterfactual distributions in M with cardinality $|\Omega_U| \leq \prod_{V \in \boldsymbol{C}_U} |\mathscr{H}_V|$ for a fixed $U \in \boldsymbol{U}$. Finally, we complete the proof by repeatedly applying this replacement for every $U \in \boldsymbol{U}$. □

A.4 Partial identification of Counterfactual Distributions

To demonstrate the expressive power of discrete SCMs, we investigate the problem of partial identification of counterfactual distributions. For an SCM $M^* = \langle \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{F}, P \rangle$, we are interested in evaluating an arbitrary counterfactual probability $P(\boldsymbol{y_x}, \dots, \boldsymbol{z_w})$. The detailed parametrization of M^* is unknown. Instead, the learner only has access to the causal diagram $\mathcal{G}$ and the observational distribution $P(\boldsymbol{v})$ induced by M^*. Our goal is to derive an informative bound $[l, r]$ from the combination of $\mathcal{G}$ and $P(\boldsymbol{v})$ that contains the actual counterfactual probability $P(\boldsymbol{y_x}, \dots, \boldsymbol{z_w})$.

Let $\mathscr{N}$ denote the family of discrete SCMs defined in Thm. 1 which are compatible with the causal diagram $\mathcal{G}$. We derive a bound $[l, r]$ over $P(\boldsymbol{y_x}, \dots, \boldsymbol{z_w})$ from the observational data $P(\boldsymbol{v})$ by solving the optimization problem in Eq. (6). It follows immediately from Thm. 1 that the solution $[l, r]$ of the optimization problem Eq. (6) is guaranteed to be a tight bound over the unknown counterfactual $P(\boldsymbol{y_x}, \dots, \boldsymbol{z_w})$.

Corollary 1 (Soundness). *Given a DAG $\mathcal{G}$ and an observational distribution $P(\boldsymbol{v})$, let $\mathscr{M}$ be the set of all SCMs compatible with $\mathcal{G}$ and let $\mathscr{M}_o = \{\forall M \in \mathscr{M} \mid P_M(\boldsymbol{v}) = P(\boldsymbol{v})\}$. For the solution $[l, r]$ of Eq.* (6)*, $P_M(\boldsymbol{y_x}, \dots, \boldsymbol{z_w}) \in [l, r]$ for any SCM $M \in \mathscr{M}_o$.*

Proof. Without loss of generality, we assume $\mathscr{M}_o \neq \emptyset$, i.e., $\mathcal{G}$ and $P(\boldsymbol{v})$ are compatible. For any $M \in \mathscr{M}_o$, Thm. 1 implies that there exists a discrete $N \in \mathscr{N}$ such that $P_N(\boldsymbol{v}) = P_M(\boldsymbol{v}) = P(\boldsymbol{v})$ and $P_N(\boldsymbol{y_x}, \dots, \boldsymbol{z_w}) = P_M(\boldsymbol{y_x}, \dots, \boldsymbol{z_w})$. The optimization problem of Eq. (6) ensures $P_N(\boldsymbol{y_x}, \dots, \boldsymbol{z_w}) \in [l, r]$, which completes the proof. □

Corollary 2 (Tightness). *Given a DAG $\mathcal{G}$ and an observational distribution $P(\boldsymbol{v})$, let $\mathscr{M}$ be the set of all SCMs compatible with $\mathcal{G}$ and let $\mathscr{M}_o = \{\forall M \in \mathscr{M} \mid P_M(\boldsymbol{v}) = P(\boldsymbol{v})\}$. For the solution $[l, r]$ of Eq.* (6)*, there exist SCMs $M_1, M_2 \in \mathscr{M}_o$ such that $P_{M_1}(\boldsymbol{y_x}, \dots, \boldsymbol{z_w}) = l$, $P_{M_2}(\boldsymbol{y_x}, \dots, \boldsymbol{z_w}) = r$.*

Proof. Let $\mathscr{N}_o = \{\forall N \in \mathscr{N} \mid P_N(\boldsymbol{v}) = P(\boldsymbol{v})\}$. The optimization problem of Eq. (6) ensures that there exist discrete SCMs $N_1, N_2 \in \mathscr{N}_o$ such that $P_{N_1}(\boldsymbol{y_x}, \dots, \boldsymbol{z_w}) = l$, $P_{N_2}(\boldsymbol{y_x}, \dots, \boldsymbol{z_w}) = r$. For any $N_i, i = 1, 2$, Thm. 1 implies that one could find an SCM $M_i \in \mathscr{M}_o$ such that $P_{M_i}(\boldsymbol{y_x}, \dots, \boldsymbol{z_w}) = P_{N_i}(\boldsymbol{y_x}, \dots, \boldsymbol{z_w})$. This completes the proof. □

A.5 Acyclic Directed Mixed Graphs

In the causal inference literature [43, 45], a causal diagram could also be represented by an acyclic directed mixed graph (ADMG), where exogenous variables are not explicitly shown. Formally, an ADMG associated with an SCM $M = \langle \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{F}, P \rangle$ is an augmented DAG where nodes represent $\boldsymbol{V}$; arrows represent arguments Pa_V of each function f_V; and a bi-directed arrow between nodes V_i and V_j indicates the presence of unobserved confounders (UCs) affecting both V_i and V_j, i.e., $U_{V_i} \cap U_{V_j} \neq \emptyset$[8]. For instance, Fig. 6b shows an ADMG compatible with SCMs described in Fig. 6a. Similarly, it is also compatible with SCMs graphically described in Fig. 6c. That is, an ADMG describes an equivalence class of DAGs (more than 1). [43, Def. 5] introduce an algorithm to project a DAG to an ADMG which maintains the same causal relationships over endogenous variables.

We will study an inverse algorithm that translates an ADMG into a DAG while maintaining all counterfactual distributions. Our construction rests on a novel object called the *confounded clique*.

Definition 10 (c-clique). For an ADMG $\mathcal{G}$, a subset $\boldsymbol{C} \subseteq \boldsymbol{V}$ is a c-clique if any pair $V_i, V_j \in \boldsymbol{C}$ is connected by a *bi-directed arrow* in $\mathcal{G}$, i.e., $V_i \leftrightarrow V_j \in \mathcal{G}$.

[8]The definition of ADMG used here differs from the one studied in [15]. According to [15], the ADMG in Fig. 6b uniquely corresponds to the DAG in Fig. 6a; the ADMG for the DAG of Fig. 6c is not defined.

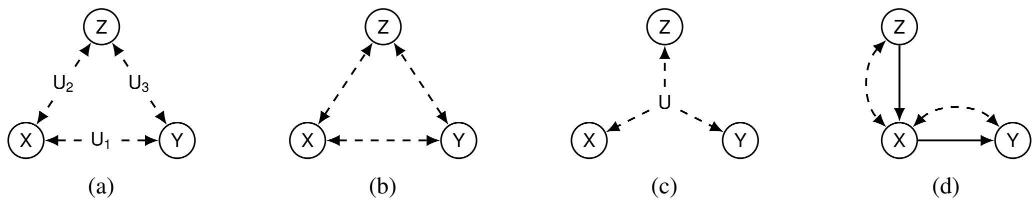


Figure 6: DAGs (a, c) containing a treatment X, an outcome Y, and a covariate Z; and (b) their corresponding ADMG; (d) an ADMG that is counterfactually equivalent to the DAG in Fig. 1b.

A c-clique $\boldsymbol{C}$ in $\mathcal{G}$ is *maximal* if there exists no other c-clique that contains $\boldsymbol{C}$. We denote by $c(\mathcal{G})$ the set of all maximal c-cliques in an ADMG $\mathcal{G}$. For instance, the ADMG of Fig. 6c has a single c-clique $\boldsymbol{C} = \{X, Y, Z\}$. Fig. 6d contains two c-cliques $\boldsymbol{C}_1 = \{X, Z\}$ and $\boldsymbol{C}_2 = \{X, Y\}$; while it only contains a single c-component $\{X, Y, Z\}$.

Algorithm 2: InverseProject

```
Input: An ADMG 𝒢
Output: A DAG ℋ.
Let ℋ = 𝒢.
for each c-clique C in c(𝒢) do
    For every pair V_i, V_j ∈ C, remove
    V_i ↔ V_j from ℋ.
    Add an exogenous node U in ℋ.
    For every V ∈ C, add U → V in ℋ.
end for
```

Our algorithm InverseProject, described in Alg. 2, translates an ADMG into a DAG by replacing bi-directed arrows in each c-clique with arrows from a new exogenous variable. As an example, Fig. 6c shows an DAG obtained from the ADMG of Fig. 6b where exogenous variable U corresponds to the c-clique $\boldsymbol{C} = \{X, Y, Z\}$. Fig. 1b shows a DAG obtained from applying InverseProject to the ADMG of Fig. 6d. The following proposition shows that InverseProject constructs a DAG that generates the same counterfactual distributions in the given ADMG.

Lemma 7. *For an ADMG $\mathcal{G}$, let $\mathcal{H}$ be a DAG obtained from* InverseProject$(\mathcal{G})$, *consider the following conditions: (1) $\mathscr{M}$ is the set of all SCMs associated with $\mathcal{G}$; (2) $\mathscr{N}$ is the set of all SCMs associated with $\mathcal{H}$. Then $\mathscr{M}$ and $\mathscr{N}$ are counterfactually equivalent.*

Proof. By the definition of ADMGs, a backdoor path $V_i \leftarrow U_k \rightarrow V_j \in \mathcal{H}$ indicates the presence of a bi-directed arrow $V_i \leftrightarrow V_j \in \mathcal{G}$. Therefore, any SCM N compatible with the DAG $\mathcal{H}$ is also compatible with the ADMG $\mathcal{G}$. That is, $N \in \mathscr{N}$ implies $N \in \mathscr{M}$.

It suffices to show that for any SCM M compatible with the ADMG $\mathcal{G}$, there exists an SCM N compatible the DAG $\mathcal{H}$ such that for any $\boldsymbol{X} \subset \boldsymbol{V}$, $P_M(\boldsymbol{v}|\text{do}(\boldsymbol{x})) = P_N(\boldsymbol{v}|\text{do}(\boldsymbol{x}))$. Let c^2-components $c(\mathcal{G}) = \{\boldsymbol{C}_1, \dots, \boldsymbol{C}_n\}$. We will construct a partition $\tilde{U}_1, \dots, \tilde{U}_n$ over exogenous variables $\boldsymbol{U}$ in M. Let $\tilde{U}_1 = \cup_{V \in \boldsymbol{C}_i} U_V$ and $\tilde{U}_i = \cup_{V \in \boldsymbol{C}_i} U_V \setminus \left(\cup_{j<i} \tilde{U}_i\right)$ for $i = 2, \dots, n$. By construction, we must have $\tilde{U}_i \subseteq \cup_{V \in \boldsymbol{C}_i} U_V$. Finally, we obtain an SCM N compatible with DAG $\mathcal{H}$ by (1) simply grouping exogenous variables $\boldsymbol{U}$ in M into the partition $\tilde{\boldsymbol{U}} = \{\tilde{U}_1, \dots, \tilde{U}_n\}$ and (2) use $\tilde{\boldsymbol{U}}$ as the exogenous variables in the modified model N. Since structural functions $\boldsymbol{F}$ and exogenous distribution P remain the same, M and N must coincide in all counterfactual distributions. □

To characterize counterfactual distributions in an ADMG $\mathcal{G}$, we could apply procedure InverseProject to obtain a DAG $\mathcal{H}$. Lem. 7 and Thm. 1 imply that one could assume exogenous variables in $\mathcal{G}$ to be exogenous variables in $\mathcal{H}$ with finite domains, without loss of generality.

B Monte Carlo Estimation of Credible Intervals

In this section, we provide proofs for the large deviation bounds for empirical estimates of $100(1-\alpha)\%$ credible intervals introduced in Sec. 3.2.

Lemma 1. *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*

$$\hat{l}_\alpha(T) \in \left[l_{\alpha-f(T,\delta)}, l_{\alpha+f(T,\delta)}\right], \qquad \hat{r}_\alpha(T) \in \left[r_{\alpha+f(T,\delta)}, r_{\alpha-f(T,\delta)}\right]. \tag{17}$$

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 $\left\{\theta_{\text{ctf}}^{(t)}\right\}_{t=1}^{T}$ that are smaller than or equal to $l_{\alpha+\epsilon}$. That is,

$$P\left(\hat{l}_\alpha(T) > l_{\alpha+\epsilon}\right) \leq P\left(\sum_{t=1}^{T} \mathbb{1}_{\theta_{\text{ctf}}^{(t)} \leq l_{\alpha+\epsilon}} \leq \lceil(\alpha/2)T\rceil - 1\right) \tag{71}$$

$$\leq P\left(\sum_{t=1}^{T} \mathbb{1}_{\theta_{\text{ctf}}^{(t)} \leq l_{\alpha+\epsilon}} \leq (\alpha/2)T\right) \tag{72}$$

$$\leq P\left(\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{\theta_{\text{ctf}}^{(t)} \leq l_{\alpha+\epsilon}} \leq \frac{\alpha+\epsilon}{2} - \frac{\epsilon}{2}\right) \tag{73}$$

$$\leq \exp\left(-\frac{T\epsilon^2}{2}\right). \tag{74}$$

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 $\left\{\theta_{\text{ctf}}^{(t)}\right\}_{t=1}^{T}$ that are larger than or equal to $l_{\alpha+\epsilon}$. That is,

$$P\left(\hat{l}_\alpha(T) < l_{\alpha-\epsilon}\right) \leq P\left(\sum_{t=1}^{T} \mathbb{1}_{\theta_{\text{ctf}}^{(t)} \leq l_{\alpha-\epsilon}} \geq \lceil(\alpha/2)T\rceil\right) \tag{75}$$

$$\leq P\left(\sum_{t=1}^{T} \mathbb{1}_{\theta_{\text{ctf}}^{(t)} \leq l_{\alpha-\epsilon}} \geq (\alpha/2)T\right) \tag{76}$$

$$\leq P\left(\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{\theta_{\text{ctf}}^{(t)} \leq l_{\alpha-\epsilon}} \geq \frac{\alpha-\epsilon}{2} + \frac{\epsilon}{2}\right) \tag{77}$$

$$\leq \exp\left(-\frac{T\epsilon^2}{2}\right). \tag{78}$$

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

$$P\left(\hat{h}_\alpha(T) < h_{\alpha+\epsilon}\right) \leq \exp\left(-\frac{T\epsilon^2}{2}\right), \qquad P\left(\hat{h}_\alpha(T) > h_{\alpha-\epsilon}\right) \leq \exp\left(-\frac{T\epsilon^2}{2}\right). \tag{79}$$

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

$$\exp\left(-\frac{T\epsilon^2}{2}\right) = \frac{\delta}{4} \Rightarrow \epsilon = \sqrt{2T^{-1}\ln(4/\delta)}. \tag{80}$$

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

Corollary 3. *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. 1 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 stick-breaking priors in Eq. (8) with hyperparameters $\alpha_U^{(u)} = \alpha_U / d_U$ and $\beta_U^{(u)} = (d_U - u)\alpha_U / d_U$ for some real $\alpha_U > 0$. This is equivalent to drawing probabilities $\theta_U = \{\theta_u \mid \forall u\}$ from a Dirichlet distribution defined as:

$$\theta_U \sim \texttt{Dirichlet}\left(\frac{\alpha_1}{d_U}, \cdots, \frac{\alpha_{d_U}}{d_U}\right), \text{ where } \alpha_i = \alpha_U, \forall i = 1, \ldots, d_U. \tag{81}$$

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

Experiment 1: Frontdoor We collect $N = 10^4$ observational data $\bar{\boldsymbol{V}} = \{X^{(n)}, Y^{(n)}, W^{(n)}\}_{n=1}^N$ from an SCM compatible with the "Frontdoor" graph in Fig. 3, defined as follows:

$$\begin{aligned}
&U_1 \sim \texttt{Unif}(0,1), \quad U_2 \sim \mathcal{N}(0,1), \\
&X \sim \texttt{Binomial}(1, p_X), \text{ where } p_W = U_1, \\
&W \sim \texttt{Binomial}(1, p_W), \text{ where } p_W = \frac{1}{1 + \exp(-X - U_2)}, \\
&Y \sim \texttt{Binomial}(1, p_Y), \text{ where } p_Y = \frac{1}{1 + \exp(W - U_1)}.
\end{aligned} \tag{82}$$

In this experiment, we set hyperparameters $\alpha_{U_1} = d_{U_1} = 8$ and $\alpha_{U_1} = d_{U_2} = 4$.

Experiment 2: Instrumental Variables (IV) We collect $N = 10^4$ observational samples $\bar{\boldsymbol{V}} = \{X^{(n)}, Y^{(n)}, Z^{(n)}\}_{n=1}^N$ from an SCM compatible with the "IV" graph in Fig. 1a, defined as follows:

$$\begin{aligned}
&U_1 \sim \mathcal{N}(0,1), \quad U_2 \sim \mathcal{N}(0,1), \\
&Z \sim \texttt{Binomial}(1, p_Z), \text{ where } p_Z = \frac{1}{1 + \exp(-U_1)}, \\
&X \sim \texttt{Binomial}(1, p_X), \text{ where } p_X = \frac{1}{1 + \exp(-Z - U_2)}, \\
&Y \sim \texttt{Binomial}(1, p_Y), \text{ where } p_Y = \frac{1}{1 + \exp(X - U_2 + 0.5)}.
\end{aligned} \tag{83}$$

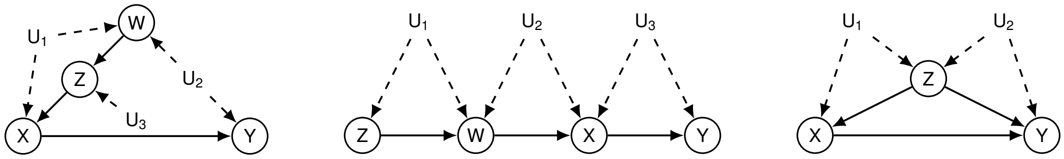
In this experiment, we set hyperparameters $\alpha_{U_1} = d_{U_1} = 2$ and $\alpha_{U_1} = d_{U_2} = 16$.

Experiment 3: Probability of Necessity and Sufficiency (PNS) We collect $N = 10^4$ observational samples $\bar{\boldsymbol{V}} = \{X^{(n)}, Y^{(n)}\}_{n=1}^N$ from an SCM compatible with the "Bow" graph in Fig. 1d, defined as follows:

$$\begin{aligned}
&U \sim \mathcal{N}(0,1), \quad E \sim \texttt{Logistic}(0,1) \\
&X \sim \texttt{Binomial}(1, p_X), \text{ where } p_X = \frac{1}{1 + \exp(U)}, \\
&Y \leftarrow \mathbb{1}_{X - U + E + 0.1 > 0}.
\end{aligned} \tag{84}$$

In this experiment, we set hyperparameters $\alpha_U = d_U = 8$.

Experiment 4: International Stroke Trials (IST) IST was a large, randomized, open trial of up to 14 days of antithrombotic therapy after stroke onset [10]. 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 $Y \in \{0, \ldots, 3\}$ 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.



(a) Napkin (b) Triple Bow (c) M+BD Graph

Figure 7: DAGs for Experiment 5 (a), Experiment 7 (b), and Experiment 8 (d), containing a treatment X, an outcome Y, ancestors Z, W, and exogenous variables U.

To emulate the presence of unobserved confounding, we filter the experimental data with selection rules $f_X^{(Z)}$, $Z \in \{0, \dots, 9\}$, following a procedure in [49]. More specifically, given a collection of IST samples $\{X^{(n)}, Y^{(n)}, U_2^{(n)}\}_{n=1}^N$ where $U_2^{(n)}$ is the age of the nth patient. For each data point $\left(X^{(n)}, Y^{(n)}, U_2^{(n)}\right)$, we introduce an instrumental variable $Z^{(n)} \in \{0, \dots, 9\}$. Values of the instrument $Z^{(n)}$ for nth patient are decided by

$$Z^{(n)} = \lfloor 10 \times U_1 \rfloor, \text{ where } U_1^{(n)} \sim \texttt{Unif}(0, 1). \tag{85}$$

We then check if $X^{(n)}$ satisfies the following condition

$$X^{(n)} = \lfloor 6 \times p_X \rfloor, \text{ where } p_X = \frac{1}{1 + \exp\left(-U_1^{(n)} \times U_2^{(n)}/100 - Z^{(n)}/10\right)} \tag{86}$$

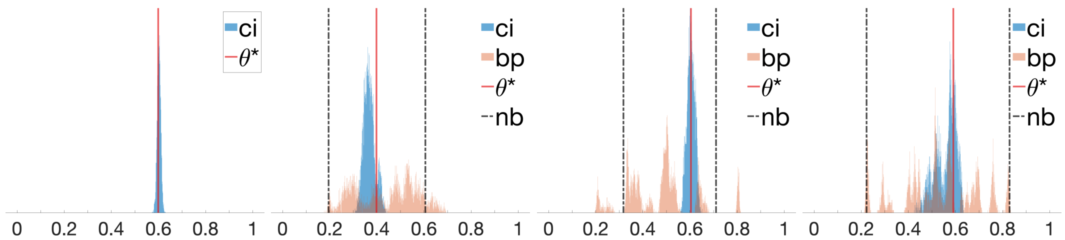
If the above condition is satisfied, we keep the data point $\left(X^{(n)}, Y^{(n)}, Z^{(n)}, U_1^{(n)}, U_2^{(n)}\right)$ in the dataset; otherwise, the data point is dropped. After this data selection process is complete, we hide columns of variables $U_1^{(n)}, U_2^{(n)}$. Doing so allows us to obtain $N = 3 \times 10^3$ synthetic observational samples $\bar{\boldsymbol{V}} = \left\{X^{(n)}, Y^{(n)}, Z^{(n)}\right\}_{n=1}^N$ that are compatible with the "Double bow" diagram of Fig. 1b. In this experiment, we set hyperparameters $\alpha_{U_1} = 10$ and $\alpha_{U_2} = 10$. As a baseline, we estimate the treatment effect $E[Y_{x=(1,0)}] = 1.3418$ for only assigning aspirin $X = (1, 0)$ from the randomized trial data containing 1.9285×10^4 subjects.

C.1 Additional Simulations on Other Causal Diagrams

We also evaluate our algorithms on various simulated SCM instances in other more involved causal diagrams. Overall, we found that simulation results match the findings in the manuscript. For identifiable settings (Experiment 5), our algorithms are able to recover the actual, unknown counterfactual probabilities. For other more general cases where the target distribution is non-identifiable (Experiments 6, 7 and 8), our algorithms consistently dominate state-of-art bounding strategies.

Experiment 5: Napkin Graph This experiment evaluates our sampling 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 "Napkin" graph in Fig. 10a where X, Y, Z, W are binary variables in $\{0, 1\}$; U_1, U_2, U_3 take values in real $\mathbb{R}$. The identifiability of the interventional distribution $P(y_x)$ from the observational data $P(x, y, w, z)$ could be derived by iteratively applying inference rules of "do-calculus" [33, 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 defined as follows:

$$\begin{aligned}
&U_1 \sim \mathcal{N}(0, 1), \quad U_2 \sim \mathcal{N}(0, 1), \quad U_3 \sim \mathcal{N}(0, 1) \\
&W \sim \texttt{Binomial}(1, p_W), \text{ where } p_W = \frac{1}{1 + \exp(U_1 - U_2)}, \\
&Z \sim \texttt{Binomial}(1, p_Z), \text{ where } p_Z = \frac{1}{1 + \exp(W - U_3)}, \\
&X \sim \texttt{Binomial}(1, p_X), \text{ where } p_X = \frac{1}{1 + \exp(-Z - U_1)}, \\
&Y \sim \texttt{Binomial}(1, p_Y), \text{ where } p_Y = \frac{1}{1 + \exp(X - U_2 - 0.5)}.
\end{aligned} \tag{87}$$



(a) Napkin (b) Double bow (c) Triple bow (d) M+BD Graph

Figure 8: Histogram plots for samples drawn from the posterior distribution over target counterfactual probabilities. For all plots (a - d), *ci* represents our proposed algorithms; *bp* stands for Gibbs samplers using the representation of canonical partitions [2]; θ^* is the actual counterfactual probability; *nb* stands for the natural bounds [30].

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. 8a shows a histogram containing samples drawn from the posterior distribution of $\left(P(Y_{x=0} = 1) \mid \bar{\boldsymbol{V}}\right)$. 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 6: Double Bow This experiment evaluates our bounding strategy in non-identifiable settings where the optimal bounding strategy does not exist. In this case, our proposed algorithm should improve over state-of-art bounds. Consider again the "Double Bow" diagram in Fig. 1b where $X, Y, Z \in \{0, 1\}$ and $U_1, U_2 \in \mathbb{R}$. We collect $N = 10^4$ observational samples $\bar{\boldsymbol{V}} = \{X^{(n)}, Y^{(n)}, Z^{(n)}\}_{n=1}^N$ from an SCM instance defined as follows:

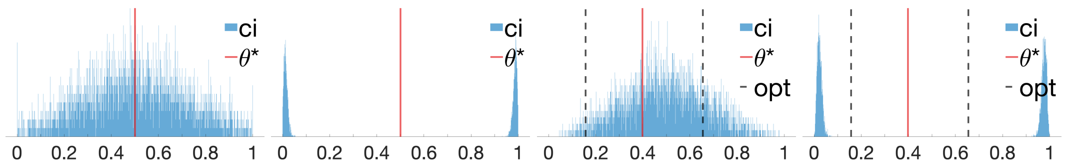
$$
\begin{aligned}
& U_1 \sim \mathcal{N}(0,1), \quad U_2 \sim \mathcal{N}(0,1), \\
& Z \sim \texttt{Binomial}(1, p_Z), \text{ where } p_Z = \frac{1}{1+\exp(-U_1)}, \\
& X \sim \texttt{Binomial}(1, p_X), \text{ where } p_X = \frac{1}{1+\exp(-Z-U_1-U_2)}, \\
& Y \sim \texttt{Binomial}(1, p_Y), \text{ where } p_Y = \frac{1}{1+\exp(X-U_2+0.5)}.
\end{aligned} \tag{88}
$$

In this experiment, we set hyperparameters $\alpha_{U_1} = d_{U_1} = 32$ and $\alpha_{U_2} = d_{U_1} = 32$. Fig. 8b shows samples drawn from the posterior distribution of $\left(P(Y_{x=0} = 1) \mid \bar{\boldsymbol{V}}\right)$. As a baseline, we also include the natural bounds [36, 30] (*nb*), and posterior samples obtained from the Gibbs sampler using a naïve generalization of the discretization procedure (*bp*) in [2]. Our analysis reveals that all algorithms achieve bounds that contain the actual, target causal effect $P(Y_{x=0} = 1) = 0.3954$. Our algorithm obtains a 100% credible interval $l_{ci} = 0.3054, r_{ci} = 0.4456$, which dominates all the other algorithms ($l_{bp} = 0.1778, r_{bp} = 0.6923, l_{nb} = 0.1949, r_{nb} = 0.6061$).

Experiment 7: Triple Bow Consider the "Triple Bow" diagram in Fig. 10b where $X, Y, Z \in \{0, 1\}$ and $U_1, U_2, U_3 \in \mathbb{R}$. We collect $N = 10^4$ observational samples $\bar{\boldsymbol{V}} = \{X^{(n)}, Y^{(n)}, Z^{(n)}\}_{n=1}^N$ from an SCM defined as follows:

$$
\begin{aligned}
& U_1 \sim \mathcal{N}(0,1), \quad U_2 \sim \mathcal{N}(0,1), \quad U_3 \sim \mathcal{N}(0,1), \\
& Z \sim \texttt{Binomial}(1, p_Z), \text{ where } p_Z = \frac{1}{1+\exp(-U_1)}, \\
& W \sim \texttt{Binomial}(1, p_W), \text{ where } p_W = \frac{1}{1+\exp(-Z-U_1-U_2)}, \\
& X \sim \texttt{Binomial}(1, p_X), \text{ where } p_X = \frac{1}{1+\exp(-W-U_2-U_3)}, \\
& Y \sim \texttt{Binomial}(1, p_Y), \text{ where } p_Y = \frac{1}{1+\exp(X-U_3-0.5)}.
\end{aligned} \tag{89}
$$

In this experiment, we set hyperparameters $\alpha_{U_1} = 0.001 \times d_{U_1} = 0.032$ and $\alpha_{U_2} = 0.001 \times d_{U_1} = 0.032$. Fig. 8c shows samples drawn from the posterior distribution of $\left(P(Y_{x=0} = 1) \mid \bar{\boldsymbol{V}}\right)$. As a



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

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

baseline, we also include the natural bounds [36, 30] (*nb*), and posterior samples obtained from the Gibbs sampler using a naïve generalization of the discretization procedure (*bp*) in [2]. Our analysis reveals that while all algorithms achieve valid bounds ($l_{bp} = 0.1964, r_{bp} = 0.8148, l_{nb} = 0.3179, r_{nb} = 0.7105$), our algorithm obtains a 100% credible interval $l_{ci} = 0.5608, r_{ci} = 0.6515$, which is the tightest bound over the target probability $P(Y_{x=0} = 1) = 0.6098$.

Experiment 8: M+BD Graph Consider the "M+BD" graph in Fig. 10c where $X, Y, Z \in \{0, 1\}$ and $U_1, U_2 \in \mathbb{R}$. In this case, the counterfactual distribution $P(y_x)$ is non-identifiable due to the presence of the collider path $X \leftarrow U_1 \rightarrow Z \leftarrow U_2 \rightarrow Y$. We collect $N = 10^4$ observational samples $\bar{\boldsymbol{V}} = \{X^{(n)}, Y^{(n)}, Z^{(n)}\}_{n=1}^N$ from an SCM instance defined as follows:

$$\begin{aligned} & U_1 \sim \mathcal{N}(0,1), \quad U_2 \sim \mathcal{N}(0,1), \\ & Z \sim \texttt{Binomial}(1, p_Z), \text{ where } p_Z = \frac{1}{1+\exp(-U_1)}, \\ & X \sim \texttt{Binomial}(1, p_X), \text{ where } p_X = \frac{1}{1+\exp(-Z-U_1-U_2)}, \\ & Y \sim \texttt{Binomial}(1, p_Y), \text{ where } p_Y = \frac{1}{1+\exp(X-Z-U_2)}. \end{aligned} \tag{90}$$

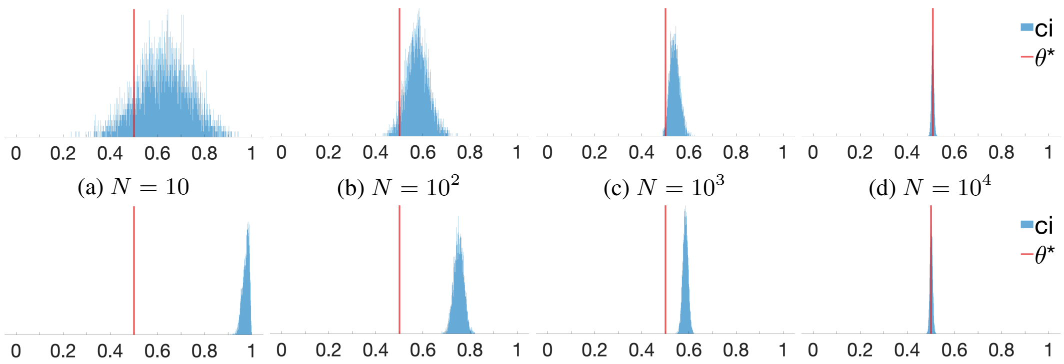
In this experiment, we set hyperparameters $\alpha_{U_1} = 0.01 \times d_{U_1} = 0.32$ and $\alpha_{U_2} = 0.01 \times d_{U_1} = 0.32$. Fig. 8d shows samples drawn from the posterior distribution of $\left(P(Y_{x=0} = 1) \mid \bar{\boldsymbol{V}}\right)$. As a baseline, we also include the natural bounds [36, 30] (*nb*), and posterior samples obtained from the Gibbs sampler using a naïve generalization of the discretization procedure (*bp*) in [2]. Our 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 $l_{ci} = 0.4247, r_{ci} = 0.6345$, which dominates all the other algorithms ($l_{bp} = 0.2140, r_{bp} = 0.8344, l_{nb} = 0.2230, r_{nb} = 0.8296$).

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 9: Frontdoor Consider first the "Frontdoor" graph in Fig. 3 where the counterfactual distribution $P(y_x)$ is identifiable from the observational data $P(x, y, w)$. The detailed parametrization of the underlying SCM is described in Eq. (82). 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. 9a and 9b 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. 9b assigns almost all weights to deterministic probabilities $P(Y_{x=0} = 1) = 1$ or $P(Y_{x=0} = 0) = 1$.

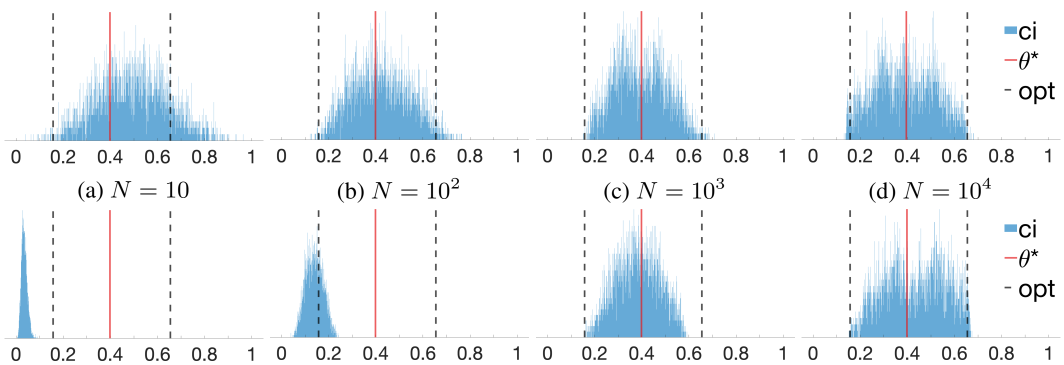
Fig. 10 shows posterior samples obtained by our Gibbs sampler when applied to observational data of various sizes, using both the flat prior (Figs. 10a to 10d) and the skewed prior (Figs. 10e to 10h). Both priors eventually collapse to the actual, unknown probability $P(Y_{x=0} = 1) = 0.5085$. As expected, 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.



(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 10: Histogram plots for samples drawn from the posterior distribution over probability $P(Y_{x=0} = 0)$ in "Frontdoor" graph of Fig. 3 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 respective observational datasets.



(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 11: 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 posetriors using the skewed prior and the same respective observational datasets.

Experiment 10: IV Consider the "IV" graph in Fig. 1b where X, Y, Z are binary variables in $\{0, 1\}$. The detailed parametrization of the underlying SCM is described in Eq. (83). In this case, the counterfactual distribution $P(y_x)$ is not identifiable from the observational data $P(x, y, z)$ [5]. Sharp bounds over $P(y_x)$ from $P(x, y, z)$ were derived in [2] (labelled as *opt*). 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. 9c and 9d shows the distribution 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. 9d assigns almost all weights to deterministic probabilities $P(Y_{x=0} = 1) = 1$ or $P(Y_{x=0} = 0) = 1$.

Fig. 11 shows posterior samples obtained by our Gibbs sampler when applied to observational data of various sizes, using both the flat prior (Figs. 11a to 11d) and the skewed prior (Figs. 11e to 11h). 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 counterfactual probability $P(Y_{x=0} = 1) = 0.3954$. It is interesting to note that, in this experiment, while the choice of prior distribution does not influence the final counterfactual bound, it still has an effect on the shape of posterior distributions.

D Naïve Generalization of (Balke and Pearl, 1995)

In this section, we will describe a naïve generalization of the canonical partitioning approach in [3] to the causal diagram of Fig. 1b. In particular, given any SCM M compatible with Fig. 1b, we will construct a discrete SCM N compatible with the diagram of Fig. 1c such that M and N coincide in all counterfactual distributions $\boldsymbol{P}^*$.

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 $\{0, 1\}$. We 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 $f_Z(u_1) = h_Z^{(i)}$. Given any u_1, u_2, $f_X(\cdot, u_1, u_2)$ defines a function mapping from domains of Z to X. Let functions in the hypothesis class $\Omega_Z \mapsto \Omega_X$ be ordered by

$$h_X^{(1)}(z) = 0, \qquad h_X^{(2)}(z) = z, \qquad h_X^{(3)}(z) = \neg z, \qquad h_X^{(4)}(z) = 1. \tag{91}$$

Similarly, we define a partition $\mathcal{U}_X^{(i)}, i = 1, 2, 3, 4$ over the domain $\Omega_{U_1} \times \Omega_{U_2}$ such that $(u_1, u_2) \in \mathcal{U}_X^{(i)}$ if and only if the induced function $f_X(\cdot, u_1, u_2) = h_X^{(i)}$. Finally, let functions mapping from domains of X to Y be ordered by

$$h_Y^{(1)}(x) = 0, \qquad h_Y^{(2)}(x) = x, \qquad h_Y^{(3)}(x) = \neg x, \qquad h_Y^{(4)}(x) = 1. \tag{92}$$

For any u_2, the induced function $f_Y(\cdot, u_2)$ must coincide with only of the above elements. Let $\mathcal{U}_Y^{(i)}, i = 1, 2, 3, 4$ be a partition over Ω_{U_2} such that $u_2 \in \mathcal{U}_Y^{(i)}$ if any only if $f_Y(\cdot, u_2) = h_Y^{(i)}$.

We now construct a discrete SCM N compatible with the casusal diagram of Fig. 1c. 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 $h_Z^{(u_Z)}$ where $h_Z^{(1)} = 0$, $h_Z^{(2)} = 1$. Given input z, u_X, values of X are given by

$$x \leftarrow \xi_X^{(z, u_X)} = h_X^{(u_X)}(z), \tag{93}$$

where $h_X^{(i)}(z)$, $i = 1, 2, 3, 4$, are functions defined in Eq. (91). Similarly, given input x, u_Y, values of Y are given by

$$y \leftarrow \xi_Y^{(x, u_Y)} = h_Y^{(u_Y)}(x), \tag{94}$$

where $h_Y^{(i)}(x)$, $i = 1, 2, 3, 4$, are functions defined in Eq. (92). Finally, we define the exogenous probability $P(u_Z, u_X, u_Y)$ in 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,

$$P_N\left(U_Z = i, U_X = j, U_Y = k\right) = P_M\left((U_1, U_2) \in \mathcal{U}_Z^{(i)} \wedge \mathcal{U}_X^{(j)} \wedge \mathcal{U}_Y^{(k)}\right). \tag{95}$$

It follows from the decomposition in Lem. 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. 1c, one could not construct an SCM M compatible with the "Double bow" diagram in Fig. 1b such that M and N coincide in all counterfactual distributions. To witness, consider a discrete SCM N where $P(U_Z = U_Y) = 1$, i.e., variables U_Z and U_Y are always the same, taking values in $\{1, 2\}$. Since in SCM N, values of $Z(u_Z)$ and $Y_{x=1}(u_Y)$ are given by

$$\begin{aligned} Z(u_Z) &= h_Z^{(u_Z)} = 0 \times \mathbb{1}_{u_Z = 1} + 1 \times \mathbb{1}_{u_Z = 2}, \\ Y_{x=1}(u_Y) &= h_Y^{(u_Y)}(1) = 0 \times \mathbb{1}_{u_Y = 1} + 1 \times \mathbb{1}_{u_Y = 2}. \end{aligned}$$

This means that counterfactual variables Z and $Y_{x=0}$ must coincide, i.e., $P(Z = Y_{x=1}) = 1$. However, for any SCM M compatible with Fig. 1b, counterfactual variables Z and Y_x must be independent due to the independence restriction [33, Ch. 7.3.2], which is a contradiction.

E Polynomial Optimization for Bounding Counterfactual Probabilities

In this section, we demonstrate how the optimization problem in Eq. (6) could be reduced to an equivalent polynomial program. The main challenge here is to write the counterfactual distribution $P(\boldsymbol{y_x}, \dots, \boldsymbol{z_w})$ in discrete SCMs as a polynomial function of parameters $\xi_V^{(pa_V, u_V)}, \theta_u$. Since for binary $a, b \in \{0, 1\}$, $a \wedge b = ab$, this means that counterfactual distributions $P(\boldsymbol{y_x}, \dots, \boldsymbol{z_w})$ in a discrete SCM could be written as:

$$P(\boldsymbol{y_x}, \dots, \boldsymbol{z_w}) = \sum_{U \in \boldsymbol{U}} \sum_{u=1,\dots,d_U} \mathbb{1}_{\boldsymbol{Y_x}(\boldsymbol{u})=\boldsymbol{y}} \cdots \mathbb{1}_{\boldsymbol{Z_w}(\boldsymbol{u})=\boldsymbol{z}} \prod_{U \in \boldsymbol{U}} \theta_u. \tag{96}$$

For convenience, we will represent parameters $\xi_V^{(pa_V, u_V)}$, for every $V \in \boldsymbol{V}$, any pa_V, u_V, as a binary sequence $\left\{\xi_v^{(pa_V, u_V)} \mid \forall v \in \Omega_V\right\}$ such that $\xi_v^{(pa_V, u_V)} \in \{0, 1\}$ and $\sum_{v \in D_V} \xi_v^{(pa_V, u_V)} = 1$. The following proposition translates indicator functions of the form $\mathbb{1}_{\boldsymbol{Y_x}(\boldsymbol{u})=\boldsymbol{y}}$ into a polynomial function with regard to parameters $\xi_v^{(pa_V, u_V)}, \theta_u$.

Lemma 8. *For a discrete SCM $M = \langle \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{F}, P \rangle$, for any $\boldsymbol{X}, \boldsymbol{Y} \subseteq \boldsymbol{V}$, fix $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{u}$. The indicator function $\mathbb{1}_{\boldsymbol{Y_x}(\boldsymbol{u})=\boldsymbol{y}}$ could be written as*

$$\mathbb{1}_{\boldsymbol{Y_x}(\boldsymbol{u})=\boldsymbol{y}} = \prod_{Y \in \boldsymbol{Y}} \mathbb{1}_{Y_{\boldsymbol{x}}(\boldsymbol{u})=y}, \tag{97}$$

$$\text{where} \quad \mathbb{1}_{Y_{\boldsymbol{x}}(\boldsymbol{u})=y} = \begin{cases} \mathbb{1}_{y=\boldsymbol{x}_Y} & \text{if } Y \in \boldsymbol{X} \\ \sum_{pa_Y} \xi_y^{(pa_Y, u_Y)} \mathbb{1}_{\{V_{\boldsymbol{x}}(\boldsymbol{u}) \mid \forall V \in Pa_Y\} = pa_Y} & \text{otherwise} \end{cases} \tag{98}$$

Proof. By the basic property of indicator function, we must have, for any $\boldsymbol{Y}, \boldsymbol{X} \subseteq \boldsymbol{V}$,

$$\mathbb{1}_{\boldsymbol{Y_x}(\boldsymbol{u})=\boldsymbol{y}} = \prod_{Y \in \boldsymbol{Y}} \mathbb{1}_{Y_{\boldsymbol{x}}(\boldsymbol{u})=y}. \tag{99}$$

Among quantities in the above equation, if $Y \subseteq \boldsymbol{X}$, $\mathbb{1}_{Y_{\boldsymbol{x}}(\boldsymbol{u})=y}$ is equal to $\mathbb{1}_{\boldsymbol{x}_Y=y}$ where $\boldsymbol{x}_Y$ is the assignment to variable Y in constants $\boldsymbol{x}$. Otherwise, for $Y \not\in \boldsymbol{X}$, Eq. (4) implies

$$\mathbb{1}_{Y_{\boldsymbol{x}}(\boldsymbol{u})=y} = \mathbb{1}_{\xi_Y^{(\{V_{\boldsymbol{x}}(\boldsymbol{u}) \mid V \in Pa_Y\}, u_Y)}=y} \tag{100}$$

The indicator $\mathbb{1}_{Y_{\boldsymbol{x}}(\boldsymbol{u})=y}$ could be further written as:

$$\mathbb{1}_{Y_{\boldsymbol{x}}(\boldsymbol{u})=y} = \xi_y^{(\{V_{\boldsymbol{x}}(\boldsymbol{u}) \mid V \in Pa_Y\}, u_Y)} = \sum_{pa_Y \in \Omega_{Pa_Y}} \xi_y^{(pa_Y, u_Y)} \mathbb{1}_{\{V_{\boldsymbol{x}}(\boldsymbol{u}) \mid \forall V \in Pa_Y\} = pa_Y} \tag{101}$$

The last step follows from the fact that values of counterfactual variables $\{V_{\boldsymbol{x}}(\boldsymbol{u}) \mid \forall V \in Pa_Y\}$ given $\boldsymbol{U} = \boldsymbol{u}$ must equate to an element in the domain Ω_{Pa_Y}. □

Recursively applying Lem. 8 to indicator functions $\mathbb{1}_{\boldsymbol{Y_x}(\boldsymbol{u})=\boldsymbol{y}}, \dots, \mathbb{1}_{\boldsymbol{Z_w}(\boldsymbol{u})=\boldsymbol{z}}$ in Eq. (96) allows us to write any counterfactual distribution $P(\boldsymbol{y_x}, \dots, \boldsymbol{z_w})$ as a polynomial function w.r.t. parameters $\theta_u, \xi_v^{(pa_V, u_V)}$. Therefore, the optimization problem in Eq. (6) is reducible to a series of polynomial programs which maximizes the objective $P(\boldsymbol{y_x}, \dots, \boldsymbol{z_w})$ subject to the observational constraints in $P(\boldsymbol{v})$ and other basic parameter constraints over $\theta_u, \xi_v^{(pa_V, u_V)}$. We will illustrate our algorithm using various examples, summarized as follows.

Example 1: Double Bow Consider again the "Double bow" diagram in Fig. 1b. We could derive a tight bound $[l, r]$ over the counterfactual probability $P(z, x_{z'}, y_{x'})$ from the observational distribution

$P(x, y, z)$ by solving the following polynomial program:

$$
\begin{aligned}
\min / \max \;\; & P(z, x_{z'}, y_{x'}) = \sum_{u_1, u_2 = 1}^{d} \xi_z^{(u_1)} \xi_x^{(z', u_1, u_2)} \xi_y^{(x', u_2)} \theta_{u_1} \theta_{u_2} \\
\text{subject to} \;\; & P(x, y, z) = \sum_{u_1, u_2 = 1}^{d} \xi_z^{(u_1)} \xi_x^{(z, u_1, u_2)} \xi_y^{(x, u_2)} \theta_{u_1} \theta_{u_2} \\
& \forall z, u_1, \;\; \xi_z^{(u_1)} \left(1 - \xi_z^{(u_1)}\right) = 0, \;\; \sum_z \xi_z^{(u_1)} = 1, \\
& \forall x, z, u_1, u_2, \;\; \xi_x^{(z, u_1, u_2)} \left(1 - \xi_x^{(z, u_1, u_2)}\right) = 0, \;\; \sum_x \xi_x^{(z, u_1, u_2)} = 1, \\
& \forall y, x, u_2, \;\; \xi_y^{(x, u_2)} \left(1 - \xi_y^{(x, u_2)}\right) = 0, \;\; \sum_y \xi_y^{(x, u_2)} = 1, \\
& \forall u_1, \;\; 0 \leq \theta_{u_1} \leq 1, \;\; \sum_{u_1} \theta_{u_1} = 1, \\
& \forall u_2, \;\; 0 \leq \theta_{u_2} \leq 1, \;\; \sum_{u_2} \theta_{u_2} = 1.
\end{aligned} \tag{102}
$$

where the cardinality $d = |\Omega_Z| \times |\Omega_Z \mapsto \Omega_X| \times |\Omega_X \mapsto \Omega_Y|$.

Example 2: IV Consider the "IV" diagram in Fig. 1a. We could derive a tight bound $[l, r]$ over the counterfactual probability $P(y'_{x'}, x, y) \equiv P\left(Y_{x=x'} = y', X = x, Y = y\right)$ from the observational distribution $P(x, y, z)$ by solving the following polynomial program:

$$
\begin{aligned}
\min / \max \;\; & P(y'_{x'}, x, y) = \sum_{u_1 = 1}^{d_1} \sum_{u_2 = 1}^{d_2} \xi_{y'}^{(x', u_2)} \xi_y^{(x, u_2)} \sum_z \xi_x^{(z, u_2)} \xi_z^{(u_1)} \theta_{u_1} \theta_{u_2} \\
\text{subject to} \;\; & P(x, y, z) = \sum_{u_1 = 1}^{d_1} \sum_{u_2 = 1}^{d_2} \xi_z^{(u_1)} \xi_x^{(z, u_2)} \xi_y^{(x, u_2)} \theta_{u_1} \theta_{u_2} \\
& \forall z, u_1, \;\; \xi_z^{(u_1)} \left(1 - \xi_z^{(u_1)}\right) = 0, \;\; \sum_z \xi_z^{(u_1)} = 1, \\
& \forall x, z, u_2, \;\; \xi_x^{(z, u_2)} \left(1 - \xi_x^{(z, u_2)}\right) = 0, \;\; \sum_x \xi_x^{(z, u_2)} = 1, \\
& \forall y, x, u_2, \;\; \xi_y^{(x, u_2)} \left(1 - \xi_y^{(x, u_2)}\right) = 0, \;\; \sum_y \xi_y^{(x, u_2)} = 1, \\
& \forall u_1, \;\; 0 \leq \theta_{u_1} \leq 1, \;\; \sum_{u_1} \theta_{u_1} = 1, \\
& \forall u_2, \;\; 0 \leq \theta_{u_2} \leq 1, \;\; \sum_{u_2} \theta_{u_2} = 1.
\end{aligned} \tag{103}
$$

where the cardinality $d_1 = |\Omega_Z|$ and $d_2 = |\Omega_Z \mapsto \Omega_X| \times |\Omega_X \mapsto \Omega_Y|$.

Example 3: Bow Consider the "Bow" diagram in Fig. 1d. We could derive a tight bound $[l, r]$ over the counterfactual probability $P(y_x, y'_{x'}) \equiv P\left(Y_x = y, Y_{x=x'} = y'\right)$ from the observational distribution $P(x, y)$ by solving the following polynomial program:

$$
\begin{aligned}
\min / \max \;\; & P(y_x, y'_{x'}) = \sum_{u=1}^{d} \xi_y^{(x,u)} \xi_{y'}^{(x',u)} \theta_u \\
\text{subject to} \;\; & P(x, y) = \sum_{u=1}^{d} \xi_x^{(u)} \xi_y^{(x,u)} \theta_u \\
& \forall x, u, \;\; \xi_x^{(u)} \left(1 - \xi_x^{(u)}\right) = 0, \;\; \sum_x \xi_x^{(u)} = 1, \\
& \forall y, x, u, \;\; \xi_y^{(x,u)} \left(1 - \xi_y^{(x,u)}\right) = 0, \;\; \sum_y \xi_y^{(x,u)} = 1, \\
& \forall u, \;\; 0 \leq \theta_u \leq 1, \;\; \sum_u \theta_u = 1
\end{aligned} \tag{104}
$$

where the cardinality $d = |\Omega_Z \mapsto \Omega_X|$.

Example 4: Frontdoor Consider the "Frontdoor" diagram in Fig. 3. We could derive a tight bound $[l, r]$ over the interventional probability $P(y_x)$ from the observational distribution $P(x, y, z)$ by solving the following polynomial program:

$$
\begin{aligned}
\min / \max \;\; & P(y_x) = \sum_{u_1=1}^{d_1} \sum_{u_1=1}^{d_2} \sum_w \xi_y^{(w,u_1)} \xi_w^{(x,u_2)} \theta_{u_1} \theta_{u_2} \\
\text{subject to} \;\; & P(x, y, w) = \sum_{u_1=1}^{d} \sum_{u_1=1}^{d_2} \sum_w \xi_x^{(u)} \xi_y^{(w,u_1)} \xi_w^{(x,u_2)} \theta_{u_1} \theta_{u_2} \\
& \forall x, u_1, \;\; \xi_x^{(u)} \left(1 - \xi_x^{(u)}\right) = 0, \;\; \sum_x \xi_x^{(u)} = 1, \\
& \forall y, w, u_1, \;\; \xi_y^{(w,u_1)} \left(1 - \xi_y^{(w,u_1)}\right) = 0, \;\; \sum_y \xi_y^{(w,u_1)} = 1, \\
& \forall w, x, u_2, \;\; \xi_w^{(x,u_2)} \left(1 - \xi_w^{(x,u_w)}\right) = 0, \;\; \sum_w \xi_w^{(x,u_w)} = 1, \\
& \forall u_1, \;\; 0 \leq \theta_{u_1} \leq 1, \;\; \sum_{u_1} \theta_{u_1} = 1, \\
& \forall u_2, \;\; 0 \leq \theta_{u_2} \leq 1, \;\; \sum_{u_2} \theta_{u_2} = 1.
\end{aligned} \tag{105}
$$

where the cardinality $d_1 = |\Omega_X| \times |\Omega_W \mapsto \Omega_Y|$ and $d_2 = |\Omega_X \mapsto \Omega_W|$.

F Derivations of Complete Conditional Distributions

In this section, we will provide detailed derivations for complete conditional distributions used in our proposed Gibbs samplers in Sec. 3.

Sampling $P(\bar{\boldsymbol{u}} \mid \bar{\boldsymbol{v}}, \boldsymbol{\theta}, \boldsymbol{\xi})$. Variables $\boldsymbol{U}^{(n)}, \boldsymbol{V}^{(n)}$, $n = 1, \dots, N$, are mutually independent given parameters $\boldsymbol{\theta}, \boldsymbol{\xi}$. This implies

$$P(\bar{\boldsymbol{u}} \mid \bar{\boldsymbol{v}}, \boldsymbol{\theta}, \boldsymbol{\xi}) = \prod_{U \in \boldsymbol{U}} P\left(\boldsymbol{u}^{(n)} \mid \bar{\boldsymbol{v}}, \boldsymbol{\theta}, \boldsymbol{\xi}\right) \tag{106}$$

$$= \prod_{U \in \boldsymbol{U}} P\left(\boldsymbol{u}^{(n)} \mid \boldsymbol{v}^{(n)}, \boldsymbol{\theta}, \boldsymbol{\xi}\right) \tag{107}$$

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

$$P\left(\boldsymbol{u}^{(n)} \mid \boldsymbol{v}^{(n)}, \boldsymbol{\theta}, \boldsymbol{\xi}\right) \propto P\left(\boldsymbol{u}^{(n)} \boldsymbol{v}^{(n)} \mid \boldsymbol{\theta}, \boldsymbol{\xi}\right) \tag{108}$$

$$\propto \prod_{V \in \boldsymbol{V}} P\left(v^{(n)} \mid pa_V^{(n)}, u_V^{(n)}, \boldsymbol{\theta}, \boldsymbol{\xi}\right) \prod_{U \in \boldsymbol{U}} P\left(u_V^{(n)} \mid \boldsymbol{\theta}, \boldsymbol{\xi}\right). \tag{109}$$

Among quantities in the above equation, $P\left(u_V^{(n)} \mid \boldsymbol{\theta}, \boldsymbol{\xi}\right) = \theta_u$ for $u = u_V^{(n)}$; and

$$P\left(v^{(n)} \mid pa_V^{(n)}, u_V^{(n)}, \boldsymbol{\theta}, \boldsymbol{\xi}\right) = \mathbb{1}_{\xi_V^{\left(pa_V^{(n)}, u_V^{(n)}\right)} = v^{(n)}}. \tag{110}$$

Sampling $P(\boldsymbol{\xi}, \boldsymbol{\theta} \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{u}})$. For every exogenous variable $U \in \boldsymbol{U}$, $\theta_U = \{\theta_u \mid \forall u\}$. For every endogenous variable $V \in \boldsymbol{V}$, $\xi_V = \left\{\xi_V^{(pa_V, u_V)} \mid \forall pa_V, u_V\right\}$. Since parameters $\boldsymbol{\xi}_V$, for every $V \in \boldsymbol{V}$, $\boldsymbol{\theta}_U$, for every $U \in \boldsymbol{U}$ are mutually independent, and they do not have common child nodes, we must have

$$P(\boldsymbol{\xi}, \boldsymbol{\theta} \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{u}}) = \prod_{V \in \boldsymbol{V}} P(\xi_V \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{u}}) \prod_{U \in \boldsymbol{U}} P(\theta_U \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{u}}). \tag{111}$$

The above independence relationships imply that we could draw samples of posterior distributions over $\left(\xi_V \mid \bar{\boldsymbol{V}}, \bar{\boldsymbol{U}}\right)$ and $\left(\theta_U \mid \bar{\boldsymbol{V}}, \bar{\boldsymbol{U}}\right)$ for every $V \in \boldsymbol{V}, U \in \boldsymbol{U}$ separately.

The complete conditional over $\left(\xi_V \mid \bar{\boldsymbol{V}}, \bar{\boldsymbol{U}}\right)$, defined in Eq. (10), follows from the fact that in discrete SCMs, the nth observation of variable $V \in \boldsymbol{V}$ is decided by $v^{(n)} \leftarrow \xi_V^{(pa_V, u_V)}$ given $pa_V^{(n)} = pa_V$, $u_V^{(n)} = u_V$. The complete conditional over $\left(\theta_U \mid \bar{\boldsymbol{V}}, \bar{\boldsymbol{U}}\right)$ in Eq. (11), follows from the conjugacy of the generalized Dirichlet distribution to multinomial sampling (e.g., see [22, 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 given by

$$P\left(\boldsymbol{u}^{(n)} \mid \bar{\boldsymbol{v}}, \bar{\boldsymbol{u}}_{-n}\right) \propto \prod_{V \in \boldsymbol{V}} P\left(v^{(n)} \mid pa_V^{(n)}, u_V^{(n)}, \bar{\boldsymbol{v}}_{-n}, \bar{\boldsymbol{u}}_{-n}\right) \prod_{U \in \boldsymbol{U}} P\left(u^{(n)} \mid \bar{\boldsymbol{v}}_{-n}, \bar{\boldsymbol{u}}_{-n}\right). \tag{112}$$

Among quantities in the above equation, for every $V \in \boldsymbol{V}$,

$$P\left(v^{(n)} \mid pa_V^{(n)}, u_V^{(n)}, \bar{\boldsymbol{v}}_{-n}, \bar{\boldsymbol{u}}_{-n}\right) = \sum_{\xi_V^{\left(pa_V^{(n)}, u_V^{(n)}\right)} \in \Omega_V} \mathbb{1}_{\xi_V^{\left(pa_V^{(n)}, u_V^{(n)}\right)} = v^{(n)}} P\left(\xi_V^{\left(pa_V^{(n)}, u_V^{(n)}\right)} \mid \bar{\boldsymbol{v}}_{-n}, \bar{\boldsymbol{u}}_{-n}\right). \tag{113}$$

The complete conditional distribution over $\left(\xi_V^{(pa_V, u_V)} \mid \bar{\boldsymbol{V}}_{-n}, \bar{\boldsymbol{V}}_{-n}\right)$, $\forall pa_V, u_V$, follows from the definition of discrete SCMs, i.e., the nth observation of variable $V \in \boldsymbol{V}$ is decided by $v^{(n)} \leftarrow \xi_V^{(pa_V, u_V)}$ given $pa_V^{(n)} = pa_V$, $u_V^{(n)} = u_V$. Formally,

$$P\left(\xi_V^{(pa_V, u_V)} \mid \bar{\boldsymbol{V}}_{-n}, \bar{\boldsymbol{V}}_{-n}\right) = \begin{cases} \mathbb{1}_{\xi_V^{(pa_V, u_V)} = v^{(i)}} & \text{if } \exists i \neq n, pa_V^{(i)} = pa_V, u_V^{(i)} = u_V, \\ 1/|\Omega_V| & \text{otherwise.} \end{cases} \tag{114}$$

Marginalizing over the domain Ω_V in Eq. (113) gives the complete conditional in Eq. (13). For every $U \in \boldsymbol{U}$, the complete conditional of $P\left(u^{(n)} \mid \bar{\boldsymbol{v}}_{-n}, \bar{\boldsymbol{u}}_{-n}\right)$, defined in Eq. (14), follows from the Pólya urn characterization of generalized Dirichlet distributions (e.g., see [22, Sec. 4]).