Canonical Domain Reduction for Partial Counterfactual Identification

Yesong Choe[1] **Yeahoon Kwon**[1,*] **Minwoo Park**[1,*] **Sanghack Lee**[1,†]

[1]Graduate School of Data Science, Seoul National University, Seoul, Republic of Korea
[*]Equal contribution (co-second authors).
[†]Corresponding author.

Abstract

Many counterfactual and causal queries are only partially identified from data, especially under unmeasured confounding. A common approach represents compatible nonparametric structural causal models (SCMs) on a finite canonical domain and computes sharp bounds via linear programming (LP) over the induced simplex. However, the canonical full counterfactual state space grows exponentially even for small graphs, making naive LP-based bounding computationally heavy. We propose a *constraint-aware* reduction that quotients out degrees of freedom irrelevant to the optimization problem. Because sharp bounds are determined jointly by the query functional and the data-implied information set, we aggregate full states into equivalence classes that are indistinguishable to every linear functional appearing in the LP objective and constraints. We show that optimizing over the induced push-forward distribution on the reduced domain preserves feasibility and yields the same sharp bounds as the full-domain.

1 INTRODUCTION

Identifying causal effects clarifies the consequences of interventions and supports informed decision-making. When standard causal assumptions hold (e.g., no unmeasured confounding), many causal effects are identifiable from observational data. In contrast, in the presence of unmeasured confounders—as in semi-Markovian (latent-variable) causal models—the target causal query may fail to be point-identified. In such cases, partial identification methods are employed to derive sharp bounds for causal estimands and counterfactual queries [Manski, 1990, Balke, 1995, Tian and Pearl, 2000, Evans, 2012].

A standard way to characterize the identified set is to represent compatible nonparametric SCMs on a finite canonical domain and compute sharp bounds via linear programming (LP) over the induced simplex. A growing literature derives informative counterfactual bounds by combining observational and interventional information [Finkelstein and Shpitser, 2020, Zhang and Bareinboim, 2021]; see, e.g., Zhang et al. [2022] building on Balke and Pearl [1994]. To improve scalability, recent work either automates bounding through nonconvex formulations with relaxations and branch-and-bound [Duarte et al., 2024], relaxes global SCM structure to enforce only local marginal consistency (the causal marginal polytope) [Zeitler and Silva, 2022], or leverages complementary geometric and latent-space viewpoints such as adversarial separations and equivalence-class enumeration [Botosaru et al., 2024, Gu et al., 2025]. However, the canonical *full* counterfactual state space still grows exponentially even for small graphs—since it must encode structural-function outputs across many parent configurations and counterfactual worlds—making naive LP-based bounding computationally prohibitive.

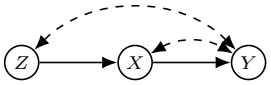


Figure 1: IV with a confounding

For instance, in Fig. 1, the canonical domain size for $\mathcal{L}_1$ observational distributions is 8. If we instead consider an $\mathcal{L}_2$ target such as the interventional distribution $P(\mathbf{Y}_{\mathbf{x}})$, the required canonical domain size increases to 32 (see Sec. B for details). This comparison highlights that the canonical domain is not a fixed object determined solely by the graph: the effective state-space size—and thus the computational burden—depends on the query and the information set used to define feasibility. In particular, once a specific query is fixed, many degrees of freedom in the full parameterization can be safely quotiented out, yielding a smaller optimization problem without changing the sharp bounds.

Our goal is to make partial identification computationally efficient without approximation by reducing the canonical

state space in a principled way. The key point is that sharp bounds are defined jointly by the query functional and the information set: both the LP objective and the data-implied constraints determine feasibility and hence the identified set. Accordingly, the reduction cannot be purely "query-only". Instead, we quotient the canonical domain by merging full states that are indistinguishable to every linear functional appearing in the LP objective or constraints; optimizing over the induced push-forward distribution on the reduced domain yields the same sharp bounds as optimizing over the full domain.

Contributions. We make two primary contributions. First, we provide a criterion for **safe, constraint-aware reduction**. We show that if a state-space quotient preserves both the query functional and the information set defining the feasible region, the reduced LP yields exactly the same sharp bounds as the full LP without approximation. Second, we propose a **query-specific structural counting methodology**. We establish that the required canonical domain is determined not just by the causal diagram, but jointly by the specific (query, information set) pair. By quantifying only the strictly necessary parent–world contexts, we prevent combinatorial explosion and provide a principled route to significant computational savings.

Organization. Sec. 2 reviews canonical domains. Sec. 3 introduces our safe state-space reduction framework and demonstrates it through a running example. Sec. 4 develops the structural counting methodology to quantify these reduced cardinalities. Finally, Sec. 5 evaluates the computational savings empirically, and Sec. 6 concludes.

2 PRELIMINARIES

We introduce in this section some basic notations and definitions that will be used throughout the paper.

Notations We use capital letters to denote variables (X), small letters for their values (x), bold letters to denote a set of those variables ($\mathbf{X}$) and values ($\mathbf{x}$). The domain of a variable X is denoted by Ω_X, and for a set of variables $\mathbf{X}$ we denote the joint domain by $\Omega_{\mathbf{X}}$. For an arbitrary set $\mathbf{X}$, let $|\mathbf{X}|$ be its cardinality (i.e., the size of the domain). The probability distribution over a set of variables $\mathbf{X}$ is denoted by $P(\mathbf{X})$. We use $P(\mathbf{x})$ as a shorthand for the probability $P(\mathbf{X} = \mathbf{x})$. The indicator function $\mathbb{1}\{X = x\}$ returns 1 if the event $X = x$ holds and 0 otherwise.

A structural causal model (SCM) [Pearl, 2009] $\mathcal{M}$ is our basic semantic framework, which is a tuple $\langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{U}) \rangle$ where $\mathbf{V}$ is a set of endogenous variables and $\mathbf{U}$ is a set of exogenous variables. $\mathcal{F}$ is a set of functions where each $f_V \in \mathcal{F}$ determines values of an endogenous variable $V \in \mathbf{V}$ taking as argument a combination of other variables in the system. That is, $v \leftarrow f_V(\mathbf{pa}_V, \mathbf{u}_V)$, $\mathbf{PA}_V \subseteq \mathbf{V}, \mathbf{U}_V \subseteq \mathbf{U}$. Exogenous variables $U \in \mathbf{U}$ are mutually independent and the values of which are drawn from the exogenous distribution $P(\mathbf{U})$. Each SCM $\mathcal{M}$ is associated with a causal diagram $\mathcal{G}$ where solid nodes represent endogenous variables $\mathbf{V}$ and bi-directed edges encode unobserved confounders. Edges represent the argument $\mathbf{PA}_V, \mathbf{U}_V$ of each structural function f_V. We write $pa(V)$ for the set of strict directed parents of V in $\mathcal{G}$.

Given an SCM, we can define distributions under different interventional regimes as follows. An SCM $\mathcal{M}$ naturally induces a joint distribution $P(\mathbf{V})$ over endogenous variable $\mathbf{V}$. This distribution is called observational distribution. An SCM $\mathcal{M}$ induces a sub-model $\mathcal{M}_{\mathbf{x}}$ where an intervention on an arbitrary subset $\mathbf{X} \subseteq \mathbf{V}$, denoted by $do(\mathbf{X} = \mathbf{x}) = do(\mathbf{x})$, is conducted. $P_{\mathbf{x}}(\mathbf{V})$ denotes the interventional distribution induced by $\mathcal{M}_{\mathbf{x}}$. For any subset $\mathbf{Y} \subseteq \mathbf{V}$, the potential response $\mathbf{Y}_{\mathbf{x}}(\mathbf{u})$ is defined as the solution of $\mathbf{Y}$ in the sub-model $\mathcal{M}_{\mathbf{x}}$ under $\mathbf{U} = \mathbf{u}$. The potential response defines the counterfactual variable $\mathbf{Y}_{\mathbf{x}}$, whose distribution is obtained by averaging over the $\mathbf{U}$. The event $\mathbf{Y}_{\mathbf{x}} = \mathbf{y}$ is interpreted as "$\mathbf{Y}$ would be $\mathbf{y}$ had $\mathbf{X}$ been $\mathbf{x}$." For subsets $\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \mathbf{W}, \ldots \subseteq \mathbf{V}$, $P(\mathbf{y}_{\mathbf{x}}, \ldots, \mathbf{z}_{\mathbf{w}})$ is the distribution over counterfactuals.

Definition 2.1 (C-component)**.** For a causal diagram $\mathcal{G}$, a non-empty subset $\mathbf{C} \subseteq \mathbf{V}$ is a c-component if (i) any two nodes V_i, V_j in $\mathbf{C}$ are connected by a path of bi-directed edges in $\mathcal{G}$, and (ii) $\mathbf{C}$ is maximal w.r.t. this property.

Definition 2.2 (Canonical SCM)**.** An SCM $\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{U}) \rangle$ is said to be a canonical SCM if

1. For every endogenous $V \in \mathbf{V}$, its values v are given by a function $v \leftarrow f_V(\mathbf{pa}_V, \mathbf{u}_V)$ where for any $\mathbf{pa}_V, \mathbf{u}_V, f_V(\mathbf{pa}_V, \mathbf{u}_V)$ is contained in a finite domain Ω_v.
2. For every exogenous $U \in \mathbf{U}$, its values u are drawn from a finite domain Ω_U; its cardinality is bounded by

$$|\Omega_U| = \prod_{V \in \mathbf{C}(U)} |\Omega_{\mathbf{PA}_V} \mapsto \Omega_V|$$

i.e., the total number of functions mapping from domains of input $\mathbf{PA}_V$ to V for every endogenous V in the c-component $\mathbf{C}(U)$ covering U.

Based on this definition, any SCM with finite discrete domains admits a canonical representation that preserves all of its distributional properties. Crucially, this representation enables optimization over a finite simplex, making LP-based bounding computationally tractable.

Theorem 2.3. *(SCM and canonical SCM) [Balke and Pearl, 1994, Balke, 1995] For an arbitrary SCM $\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{U}) \rangle$ where endogenous variables have finite categorical domains, there exists a canonical SCM $\mathcal{N}$ such that*

1. *$\mathcal{M}$ and $\mathcal{N}$ are associated with the same causal diagram $\mathcal{G}$. That is, $\mathcal{G}_{\mathcal{M}} = \mathcal{G}_{\mathcal{N}}$.*

2. *For any set of counterfactual variables $\mathbf{Y_x}, \cdots, \mathbf{Z_w}$, $P_{\mathcal{M}}(\mathbf{Y_x}, \cdots, \mathbf{Z_w}) = P_{\mathcal{N}}(\mathbf{Y_x}, \cdots, \mathbf{Z_w})$.*

Computational challenge of canonical domains. To compute sharp bounds via LP, solvers must optimize over a distribution defined on a finite canonical domain. However, the worst-case size of this domain (d_U) scales combinatorially depending on the layer of Pearl's Causal Hierarchy (PCH) [Pearl, 2009] that the query belongs to. Understanding this exponential growth—especially at the counterfactual $\mathcal{L}_3$ layer—is crucial, as it directly mathematically motivates the need for the computationally efficient, constraint-aware reductions we develop in the subsequent sections.

Canonical domain cardinalities across layers. Let $\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{U}) \rangle$ and let $U \in \mathbf{U}$ index a c-component $\mathbf{C}(U)$. As one moves up the layers of Pearl's causal hierarchy (PCH), the parameterization must encode responses to progressively more hypothetical parent configurations, and the sufficient canonical domain size grows accordingly: (1) for $\mathcal{L}_1$ (observational distributions), $d_U = \prod_{V \in Pa(\mathbf{C}(U))} |\Omega_V|$; (2) for $\mathcal{L}_2$ (single-world interventional distributions), $d_U = \prod_{V \in \mathbf{C}(U)} |\Omega_{\mathbf{PA}_V} \times \Omega_V|$; and (3) for $\mathcal{L}_3$ (counterfactual distributions), $d_U = \prod_{V \in \mathbf{C}(U)} |\Omega_{\mathbf{PA}_V} \mapsto \Omega_V|$, where $|\Omega_{\mathbf{A}} \mapsto \Omega_{\mathbf{B}}| = |\Omega_{\mathbf{B}}|^{|\Omega_{\mathbf{A}}|}$ [Zhang et al., 2022]. As summarized above, moving upward in the causal hierarchy systematically increases the computational complexity. The naive application of $\mathcal{L}_3$ counterfactual bounds effectively makes standard LP solvers computationally demanding even for modestly sized diagrams, highlighting the pressing need for the query-specific geometric reductions proposed next.

3 CONSTRAINT-AWARE REDUCTION PRINCIPLE

While Sec. 2 outlines the worst-case canonical domains required for arbitrary queries, many partial identification problems do not need the full $\mathcal{L}_3$ parameterization. To reduce the dimension of a partial-identification problem, retaining only the degrees of freedom that influence the query is insufficient; the reduction must simultaneously preserve the *information set*—i.e., the observational distribution or summary-data constraints that define the feasible set. In this section, we provide theoretical guarantees for when a state-space reduction is "safe"—meaning it preserves the original sharp bounds. The fundamental condition requires that the quotient operation simultaneously preserve the linear functionals defining the query and those defining the data-induced constraints.

3.1 A SUFFICIENT CONDITION FOR "SAFE" REDUCTION

Let Ω_{full} be a finite full counterfactual state space and let $\psi \in \Delta(\Omega_{\text{full}})$. Suppose we seek sharp bounds of a linear query $c^\top \psi$ subject to linear constraints

$$M\psi = p, \qquad \psi \in \Delta(\Omega_{\text{full}}),$$

where $M \in \mathbb{R}^{m \times |\Omega_{\text{full}}|}$ encodes the observational/summarized information identified from data and $p \in \mathbb{R}^m$ is the corresponding identified quantity. To formalize the reduction, let $\pi : \Omega_{\text{full}} \mapsto \Omega_{\text{red}}$ be a surjection that maps each full state to a reduced state retaining only the relevant degrees of freedom, and let $\Pi \in \{0,1\}^{|\Omega_{\text{red}}| \times |\Omega_{\text{full}}|}$ be the pushforward matrix $\Pi_{t,u} = \mathbb{1}\{\pi(u) = t\}$, so that $\phi = \Pi\psi$ is the induced distribution on Ω_{red}. Because π is a function, each column of Π contains exactly one 1; hence $\phi \geq 0$ and $\mathbf{1}^\top \phi = \mathbf{1}^\top \Pi \psi = \mathbf{1}^\top \psi = 1$, so $\phi \in \Delta(\Omega_{\text{red}})$ whenever $\psi \in \Delta(\Omega_{\text{full}})$.

Proposition 3.1 (Constraint-aware reduction preserves sharp bounds)**.** *Assume there exist a vector $\tilde{c} \in \mathbb{R}^{|\Omega_{\text{red}}|}$ and a matrix $\tilde{M} \in \mathbb{R}^{m \times |\Omega_{\text{red}}|}$ such that*

$$c = \Pi^\top \tilde{c}, \qquad M = \tilde{M}\Pi.$$

Then the full-space and reduced-space partial-identification problems are equivalent:

$$\min_{\psi \in \Delta(\Omega_{\text{full}}):\, M\psi = p} c^\top \psi \;=\; \min_{\phi \in \Delta(\Omega_{\text{red}}):\, \tilde{M}\phi = p} \tilde{c}^\top \phi,$$

and likewise with min *replaced by* max*. In particular, the sharp bounds computed on the reduced simplex equal those computed on the full simplex.*

In words, the quotient map must preserve both the query objective and the data-imposed constraints; otherwise, optimizing on the smaller domain may alter the sharp bounds. Notably, a "query-only" reduction can easily fail. Even if a coarsening π preserves the query (i.e., $c = \Pi^\top \tilde{c}$), it may distort the feasible set if it does not also preserve the constraints (i.e., if there is no $\tilde{M}$ with $M = \tilde{M}\Pi$). For example, suppose the information set includes a joint constraint such as $P(A = 1, B = 1)$ (or any moment depending on (A, B)), but π keeps only A (or only (A, C)). Then the quantity $P(A = 1, B = 1)$ cannot be expressed as a linear functional of the reduced distribution $\phi = \Pi\psi$, so there is no factorization $M = \tilde{M}\Pi$. As a result, two full-state distributions ψ, ψ' can induce the same reduced ϕ (and hence the same query value), while having different values of $P(A = 1, B = 1)$. In this case the reduced problem admits spurious ϕ that are infeasible in the full problem, thereby altering the sharp bounds.

A canonical way to construct π, representing the coarsest safe quotient where irrelevant variables cancel out, is to

(a) NDE (b) Realizable NDE for $\mathcal{Q} = P(Y_{x,W_{x'}})$

Figure 2: NDE diagrams.

declare $u \sim u'$ whenever *every* function that appears in either the objective or the constraints takes the same value at u and u'. The reduced space $\Omega_{\text{red}} = \Omega_{\text{full}}/\sim$ is then the coarsest quotient that simultaneously preserves both the query and the feasible set. Algorithm 1 summarizes this construction and the resulting reduced LP that preserves the sharp bounds under the coarsest safe quotient.

3.2 DEMONSTRATING CONSTRAINT-AWARENESS: 64 VS. 8 STATES

To make the theoretical guarantees of Prop. 3.1 concrete, we consider a running example on the Natural Direct Effect (NDE) graph to demonstrate that sharp bounds for a specific counterfactual query do not generally require the full canonical parameterization of the SCM. We examine two distinct information regimes on the NDE graph.

We first consider a binary nested causal query

$$\mathcal{Q} = P\big(Y_{x,W_{x'}} = 1\big),$$

and, without loss of generality, fix $(x, x') = (0, 1)$ so that $\mathcal{Q} = P\big(Y_{0,W_1} = 1\big)$. Since the query depends on only three counterfactual variables, we rename them as $A := W_1$, $B := Y_{0,0}$, and $C := Y_{0,1}$, so that $(A, B, C) \in \{0, 1\}^3$. Under this notation, the outcome variable of interest can be expressed as

$$Y_{0,W_1} = (1 - A)B + AC.$$

Query-Specific (Quotient) State Space Define the reduced state $t = (a, b, c) \in \Omega_{\text{query}} := \{0, 1\}^3$, and let ϕ denote the induced distribution on Ω_{query}:

$$\phi_{abc} := P(A = a, B = b, C = c).$$

Fix an ordering of Ω_{query} and identify ϕ with the corresponding vector in $\mathbb{R}^8$. Accordingly, write

$$\Delta(\Omega_{\text{query}}) := \big\{\phi \in \mathbb{R}^8_{\geq 0} : \mathbf{1}_8^\top \phi = 1\big\}, \tag{1}$$

where $\mathbf{1}_8$ is the all-ones vector (we also write Δ_8).

Having formalized the reduced state space for our target query, we now set up the two essential components of our LP formulation: the objective function we wish to bound, and the feasible region dictated by the data. We first demonstrate how the query can be expressed as a linear objective over the reduced space, and subsequently detail how the available observational data impose linear constraints on these states.

Query Indicator and Linearity in ϕ Define the indicator $g(t) := (1 - a)b + ac \in \{0, 1\}$ on the reduced state space and its vector representation $\mathbf{g} \in \mathbb{R}^8$. This expresses the query as a linear functional:

$$\mathcal{Q} = \mathbb{E}_\phi[g(t)] = \mathbf{g}^\top \phi. \tag{2}$$

Table 2 lists our fixed ordering of the eight query-states and the corresponding values of $g(a, b, c)$.

Case 1: Observation Model and Constraints from P_{obs} ($128 \to 64$ States) In partial identification, the information set is typically the full observational distribution

$$P_{\text{obs}}(x, w, y) := P(X = x, W = w, Y = y), (x, w, y) \in \{0, 1\}^3,$$

which induces linear constraints on the full-state distribution. In the binary $\mathcal{L}_3$ (function-space) representation of the NDE graph, each full state $u \in \Omega_{\text{full}}$ determines the values

$$X(u), W_0(u), W_1(u), Y_{0,0}(u), Y_{0,1}(u), Y_{1,0}(u), Y_{1,1}(u),$$

so $|\Omega_{\text{full}}| = 2^7 = 128$ and $\psi \in \Delta(\Omega_{\text{full}})$ assigns mass to these states. By consistency, if $X = x$ and $W = w$ are observed then $W = W_x$ and $Y = Y_{x,w}$. Equivalently, $W = W_X$ and $Y = Y_{X,W} = Y_{X,W_X}$. Hence each cell probability is a linear functional of ψ:

$$P_{\text{obs}}(x, w, y) = \sum_{u \in \Omega_{\text{full}}} \mathbb{1}\left\{ \begin{array}{l} X(u) = x,\ W_x(u) = w, \\ Y_{x,w}(u) = y \end{array} \right\} \psi_u. \tag{3}$$

Equivalently, stacking the eight probabilities $\{P_{\text{obs}}(x, w, y)\}$ produces a vector p_{obs} and a matrix M_{obs} such that $M_{\text{obs}}\psi = p_{\text{obs}}$.

With both the query and the constraints now expressed as linear functionals of ψ, we can identify which counterfactual variables actually influence the optimization and which can be safely collapsed. The query $\mathcal{Q} = P(Y_{0,W_1} = 1)$ depends on u only through $(W_1, Y_{0,0}, Y_{0,1})$. However, the observational constraints Eq. (3) also depend on how (X, W, Y) are generated, namely through $(X, W_0, W_1, Y_{0,0}, Y_{0,1}, Y_{1,0}, Y_{1,1})$. Crucially, when $X = 1$ is observed, the outcome uses only the *selected* potential outcome Y_{1,W_1}, never both $(Y_{1,0}, Y_{1,1})$ simultaneously. When $X = 0$, neither $Y_{1,0}$ nor $Y_{1,1}$ appears in the constraint, but capturing Y_{1,W_1} in the reduced state (needed for the $X = 1$ stratum) still leaves exactly $Y_{1,1-W_1}$ free. Thus, regardless of X, one of these two bits is always unused by (3) and can be collapsed and quotiented out.

This observation motivates the following reduced parameterization, which drops the unused bit while retaining all information that the LP requires. Define the reduced map

$$\pi_{\text{obs}}(u) := \big(X(u),\ W_0(u),\ W_1(u),\ Y_{0,0}(u),\ Y_{0,1}(u),\ Y_{1,W_1}(u)\big) \in \Omega_{\text{red}} := \{0, 1\}^6.$$

Let $\phi = \Pi_{\text{obs}}\psi$ be the induced distribution on Ω_{red}. Given $(X, W_0, W_1, Y_{00}, Y_{01}, Y_{1,W_1})$, when $X = 1$, the remaining bit $Y_{1,1-W_1}$ never affects (X, W, Y) and can vary freely within each fiber, so the fiber size is 2 and $|\Omega_{\text{red}}| = 2^7/2 = 2^6 = 64$. By Prop. 3.1, optimizing over $\Delta(\Omega_{\text{red}})$ yields identical sharp bounds while easing the computational burden.

Case 2: Identified Marginals from "$\mathcal{L}_{2.5}$" Information ($128 \to 8$ States) Now suppose the identified data is restricted to marginal summaries. In this case, the reduced space coincides with the safe-quotient reduced domain in Prop. 3.1, i.e., $\Omega_{\text{red}} = \Omega_{\text{query}} = \{0, 1\}^3$. Crucially, the available information is *itself* a function of (A, B, C), so the same reduction that makes the query linear in ϕ also preserves the constraints used for partial identification. In our canonical experiments, the data identify (or are modeled as informing) only the marginals:

$$p_{w1} := \mathbb{E}_\phi[A], \qquad p_{y00} := \mathbb{E}_\phi[B], \qquad p_{y01} := \mathbb{E}_\phi[C].$$

Here, $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^8$ are the evaluation vectors $\mathbf{A}_t := a, \mathbf{B}_t := b, \mathbf{C}_t := c$ for $t = (a, b, c) \in \Omega_{\text{query}}$, so that $\mathbb{E}_\phi[A] = \mathbf{A}^\top\phi$, etc. These three marginals, together with the simplex constraint on ϕ, define the feasible set for the reduced optimization problem. Let $p = (p_{w1}, p_{y00}, p_{y01})$. The feasible set of query-level distributions is

$$\begin{aligned}\mathcal{F}(p) := \ &\big\{\phi \in \Delta(\Omega_{\text{query}}) : \\ &\mathbf{A}^\top\phi = p_{w1},\ \mathbf{B}^\top\phi = p_{y00},\ \mathbf{C}^\top\phi = p_{y01}\big\}.\end{aligned} \tag{4}$$

Since $\mathcal{Q} = \mathbf{g}^\top\phi$ is linear in ϕ, the image of $\mathcal{F}(p)$ under $\mathbf{g}^\top$ is a closed interval whose endpoints are exactly the sharp bounds. Formally, the identified set for $\mathcal{Q}$ given p is $\mathcal{I}_\mathcal{Q}(p) := \{\mathbf{g}^\top\phi : \phi \in \mathcal{F}(p)\} = \big[\underline{\mathcal{Q}}(p), \overline{\mathcal{Q}}(p)\big]$. The sharp bounds are the extrema of a linear functional over $\mathcal{F}(p)$:

$$\underline{\mathcal{Q}}(p) := \min_{\phi \in \mathcal{F}(p)} \mathbf{g}^\top\phi, \qquad \overline{\mathcal{Q}}(p) := \max_{\phi \in \mathcal{F}(p)} \mathbf{g}^\top\phi. \tag{5}$$

Equivalently, (5) are two linear programs over $\phi \in \Delta_8$ with three linear equality constraints.

Consequently, these distinct reductions (to 64 or to 8 states) show that safe reductions depend critically on what the constraints require. The full-state LP operates on 128 variables, whereas the constraint-aware reduction guarantees exact identification with significantly fewer variables. Table 1 summarizes the comparison between the required counterfactual parameters depending on the available information regimes.

4 QUERY-SPECIFIC CANONICAL DOMAIN CARDINALITIES ACROSS LAYERS

Having established the theoretical soundness of constraint-aware reduction, a natural and practical question arises: *how much computational savings can we achieve for a given query and information regime?* In the previous section, we observed that geometric insights allow a 128-state formulation to collapse into exactly 64 or 8 states. We now formalize this intuition by presenting a systematic counting methodology to explicitly quantify these minimal canonical-domain cardinalities. The underlying principle is that we count only the strictly necessary parent–world contexts in which structural functions must be evaluated, avoiding the combinatorial growth caused by unreachable counterfactual branches.

4.1 CANONICAL DOMAIN CARDINALITIES IN THE STANDARD PCH: $\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3$

We consider Fig. 2a with binary variables, $|\Omega_X| = |\Omega_W| = |\Omega_Y| = 2$. The bi-directed arc $X \leftrightarrow W$ induces two c-components in $\mathcal{L}_1$: $\mathbf{C}_1 = \{X, W\}$, and $\mathbf{C}_2 = \{Y\}$.

$\mathcal{L}_1$ (Observational Factorization) At the observational level, the canonical domain simply enumerates the joint value tuples within each c-component. Under $\mathcal{L}_1$, $\mathbf{C}_1$ contributes the marginal table of (X, W) and $\mathbf{C}_2$ contributes the conditional table of Y given (X, W):

$$d_{U_1} = |\Omega_X|\,|\Omega_W| = 4, \qquad d_{U_2} = |\Omega_X|\,|\Omega_W|\,|\Omega_Y| = 8.$$

$\mathcal{L}_2$ (Interventional / Single-World Mechanisms) Moving to the interventional level, each structural function must additionally record which parent configuration generated each value, expanding the domain. Under $\mathcal{L}_2$, the $\{X, W\}$ c-component is parameterized by an entry-space description of $P(X)$ and $P(W \mid X)$, while the $\{Y\}$ c-component is parameterized by $P(Y \mid X, W)$:

$$\begin{aligned}d_{U_1} &= |\Omega_X| \cdot |\Omega_X|\,|\Omega_W| = |\Omega_X|^2|\Omega_W| = 8,\\ d_{U_2} &= |\Omega_X|\,|\Omega_W|\,|\Omega_Y| = 8.\end{aligned}$$

$\mathcal{L}_3$ (Counterfactual / Function Space) At the counterfactual level, the parameterization must enumerate all possible structural functions—not just input-output pairs—leading to an exponential blow-up. Under $\mathcal{L}_3$, we count functions. For $\mathbf{C}_1 = \{X, W\}$, $f_X : \Omega_\emptyset \to \Omega_X$ and $f_W : \Omega_X \to \Omega_W$, and for $\mathbf{C}_2 = \{Y\}$, $f_Y : \Omega_{X,W} \to \Omega_Y$ hence

$$d_{U_1} = |\Omega_X|^{|\Omega_\emptyset|} \cdot |\Omega_W|^{|\Omega_X|} = |\Omega_X| \cdot |\Omega_W|^{|\Omega_X|} = 2 \cdot 2^2 = 8,$$
$$d_{U_2} = |\Omega_Y|^{|\Omega_X|\,|\Omega_W|} = 2^{2\cdot 2} = 16.$$

Thus, on the same graph, moving upward in the hierarchy generally increases the worst-case number of free "mechanism degrees of freedom" (especially at $\mathcal{L}_3$).

4.2 QUERY-DEPENDENT COUNTS FOR REALIZABLE MIXED-WORLD QUERIES: $\mathcal{L}_{2.5}$ AND $\mathcal{L}_{2.25}$

We now focus on the realizable mixed-world query $\mathcal{Q} = P(Y_{x,W_{x'}})$, which belongs to $\mathcal{L}_{2.5}$ [Yang and Bareinboim,

Table 1: Comparison of canonical domain sizes across varying information constraints for the target query $\mathcal{Q} = P(Y_{0,W_1} = 1)$ on the NDE graph. The set of required parameters directly dictates the number of active states. Factors decompose by variable blocks: 2^1 for X, 2^2 for (W_0, W_1), and 2^4 for $(Y_{1,0}, Y_{1,1}, Y_{0,0}, Y_{0,1})$ (or 2^3 when only $(Y_{0,0}, Y_{0,1}, Y_{1,W_1})$ is needed).

Information Regime	Data-Implied Constraints	Required Counterfactual Parameters	State Size ($\lvert\Omega_{\text{red}}\rvert$)
Worst-case ($\mathcal{L}_3$)	None (baseline)	$X, W_0, W_1, Y_{00}, Y_{01}, Y_{10}, Y_{11}$	$2^1 \times 2^2 \times 2^4 = 128$
Case 1	Full Observational: $P_{\text{obs}}(X, W, Y)$	$X, W_0, W_1, Y_{00}, Y_{01}, Y_{1,W_1}$	$2^1 \times 2^2 \times 2^3 = 64$
Case 2	Marginals: $P(W_1), P(Y_{00}), P(Y_{01})$	W_1, Y_{00}, Y_{01}	$1 \times 2^1 \times 2^2 = 8$

2025]. In the realizable graph Fig. 2b, X is fixed when feeding into Y, while W is evaluated only under x'. The function-space degrees required by this *specific* query shrink because each node's structural function need only be evaluated on a subset of its parent configurations. We capture this via a *context multiplicity* $m_V(\mathcal{Q})$, counting the number of distinct parent–world contexts actually visited:

$$d_{U_1} = |\Omega_X|^{m_X(\mathcal{Q})}|\Omega_W|^{m_W(\mathcal{Q})} = |\Omega_X|^0|\Omega_W|^1 = 2,$$
$$d_{U_2} = |\Omega_Y|^{m_Y(\mathcal{Q})} = |\Omega_Y|^2 = 4.$$

Here $m_X(\mathcal{Q}) = 0$ because f_X is never evaluated (the value is set by intervention); $m_W(\mathcal{Q}) = 1$ because only $W_{x'}$ is needed; and because $W_{x'} \in \{0, 1\}$ and the query must accommodate both possibilities across admissible SCMs, f_Y must be evaluated at $(X = x, W = 0_{x'})$ and $(X = x, W = 1_{x'})$.

Definition 4.1 (Parent–world context set)**.** Let $\mathcal{G}$ be a causal diagram over observed nodes $\mathbf{V}$ and let $\mathcal{Q}$ be a realizable, conflict-free mixed-world query. For $V \in \mathbf{V}$, let $pa(V)$ be the parent set of V in $\mathcal{G}$, and let $\lambda_{\mathcal{Q}}(p) \in \Lambda(\mathcal{Q})$ denote the world/regime label assigned to node p by $\mathcal{Q}$. Let $\mathcal{T}_V(\mathcal{Q})$ be the set of labeled parent-assignment tuples to $pa(V)$ that occur as arguments of f_V in the structural expansion of $\mathcal{Q}$.

Define the *parent–world context set* by

$$S_V(\mathcal{Q}) := \Big\{ x = \big((p = a_p)_{\lambda_p}\big)_{p \in pa(V)} : x \in \mathcal{T}_V(\mathcal{Q}) \ \wedge \forall p \in pa(V),\ \lambda_p = \lambda_{\mathcal{Q}}(p) \Big\}.$$

Define the context multiplicity $m_V(\mathcal{Q}) := |S_V(\mathcal{Q})|$.

Remark. Equivalently, $S_V(\mathcal{Q})$ indexes the distinct parent–world inputs at which f_V is queried when computing $\mathcal{Q}$.

Example 1 ($\mathcal{Q}_1 = P(Y_{x,W_{x'}}) \in \mathcal{L}_{2.5}$)**.** Given Fig. 2b, for binary variables,

$$S_Y(\mathcal{Q}_1) = \{(x, 0_{x'}), (x, 1_{x'})\}, S_W(\mathcal{Q}_1) \quad = \{(x')\},$$
$$S_X(\mathcal{Q}_1) = \varnothing.$$

so $m_Y = 2, m_W = 1, m_X = 0$.

These context multiplicities translate directly into the number of function evaluations—and hence the required canonical domain size.

Proposition 4.2 (Context-count function-space cardinality for realizable mixed-world queries)**.** *Let $\mathcal{G}$ and $\mathcal{Q}$ be as in Def. 4.1. Form the ancestor subgraph of variables appearing in $\mathcal{Q}$ after deleting incoming edges into intervened nodes. For each canonical exogenous U indexing a (maximal) c-component in this subgraph, let $\mathbf{C}(U)$ denote the corresponding c-component. Then a sufficient function-space cardinality to realize $\mathcal{Q}$ is*

$$d_U \ = \prod_{V \in \mathbf{C}(U)} |\Omega_V|^{m_V(\mathcal{Q})}.$$

In words, each node V contributes a factor $|\Omega_V|$ once for each distinct parent–world context in which f_V must be evaluated to answer $\mathcal{Q}$; c-component domains multiply these contributions across nodes. If a node V is never evaluated, then $m_V(\mathcal{Q}) = 0$ and its contribution is $|\Omega_V|^0 = 1$.

While Prop. 4.2 establishes the cardinality formula, Lem. 4.3 below provides a practical decomposition for computing $S_V(\mathcal{Q})$ by breaking the query into individual terms.

Lemma 4.3 (Union representation of context sets)**.** *Let $\mathcal{Q}$ be as in Def. 4.1. For a node V, let $\mathsf{Terms}_V(\mathcal{Q})$ denote the set of distinct V-terms that are evaluated in the recursive substitution of structural functions of $\mathcal{Q}$ (including those appearing implicitly as ancestors of the counterfactuals in $\mathcal{Q}$ (e.g., Y_x, $Y_{x'}$, $Y_{x,W_{x'}}$; duplicates with identical labels/structure are merged). For each $T \in \mathsf{Terms}_V(\mathcal{Q})$, let $S_V(T)$ be the set of labeled parent assignments to $pa(V)$ under which f_V is evaluated when answering the single term T. Then*

$$S_V(\mathcal{Q}) = \bigcup_{T \in \mathsf{Terms}_V(\mathcal{Q})} S_V(T),$$

In particular, always $m_V(\mathcal{Q}) \le \sum_{T \in \mathsf{Terms}_V(\mathcal{Q})} |S_V(T)|$, with equality if the sets are disjoint.

Crucially, the practical implication of Prop. 4.2 and Lem. 4.3 is straightforward. We do not need to expand the parameter space for counterfactual variables or parent assignments that are never triggered by the target query and the given constraints. By modeling only the specific structural mechanisms that are effectively invoked, we prevent the unnecessary exponential growth characteristic of naive canonical SCMs.

To verify consistency with $\mathcal{L}_3$, observe that when a query forces each structural function f_V to be evaluated over *all* labeled parent assignments in its full parent domain (i.e., the worst-case $\mathcal{L}_3$ setting), the context-count rule in Prop. 4.2 recovers the standard function-space cardinality:

$$d_U \;=\; \prod_{V \in \mathbf{C}(U)} |\Omega_V|^{|\Omega_{\mathbf{PA}_V}|}.$$

For many realizable mixed-world queries, the multiplicities $m_V(\mathcal{Q})$ are much smaller because only a strict subset of parent–world contexts is ever visited.

Example 2 (worst-case $\mathcal{L}_3$ contexts on the NDE graph). Consider the $\mathcal{L}_3$ query $\mathcal{Q}_2 = P\big(X,\, Y_x,\, Y_{x'}\big)$. On the NDE graph with parents $Pa(Y) = \{X, W\}$, evaluating both Y_x and $Y_{x'}$ requires f_Y over the four labeled (X, W) contexts

$$S_Y(\mathcal{Q}_2) = \{(x, 0_x), (x, 1_x), (x', 0_{x'}), (x', 1_{x'})\},$$

hence $m_Y(\mathcal{Q}_2) = 4$ (binary case). Similarly,

$$S_W(\mathcal{Q}_2) = \{(x)_x, (x')_{x'}\}, \qquad S_X(\mathcal{Q}_2) = \{()\},$$

so $m_W(\mathcal{Q}_2) = 2$ and $m_X(\mathcal{Q}_2) = 1$. This illustrates how multiple regimes can force the full set of parent–world contexts and thereby recover the worst-case $\mathcal{L}_3$ count.

Furthermore, requiring the same node in multiple regimes increases context multiplicities. Intuitively, this is because the number of distinct labeled parent–world contexts that must be visited increases. When the regimes are disjoint in labels (e.g., x versus x'), the required context sets typically add (i.e., the union is close to a disjoint union), leading to larger $m_V(\mathcal{Q})$.

$\mathcal{L}_{2.25}$ Illustration Consider a "globally consistent" query (single global regime label propagation) such as

$$\mathcal{Q} = P\big(X,\, W_x,\, Y_x\big) \in \mathcal{L}_{2.25} \setminus \mathcal{L}_2.$$

Then $m_X(\mathcal{Q}) = 1$ (the natural X is present), $m_W(\mathcal{Q}) = 1$ (only W_x is needed), and $m_Y(\mathcal{Q}) = 2$ because Y_x must be evaluated at $(x, 0_x)$ and $(x, 1_x)$. Thus,

$$d_{U_1} = |\Omega_X|^1\, |\Omega_W|^1 = 4, \qquad d_{U_2} = |\Omega_Y|^2 = 4.$$

Comparing $\mathcal{L}_{2.25}$ and $\mathcal{L}_{2.5}$ through the context-set lens, we find that both layers use the same counting rule (Prop. 4.2) but differ in which mixed labels are admissible. Specifically, $\mathcal{L}_{2.25}$ enforces a single global regime label per intervened variable that is propagated to all descendants, whereas $\mathcal{L}_{2.5}$ permits different labels to flow to different children (realizable split/edge-style interventions), subject to conflict-freeness. When a query is admissible under both layers, the set of visited parent–world contexts is identical, so the resulting $m_V(\mathcal{Q})$ (and hence d_U) coincides. However, $\mathcal{L}_{2.5}$ strictly contains $\mathcal{L}_{2.25}$: there exist realizable mixed-world queries allowed under $\mathcal{L}_{2.5}$ but excluded under $\mathcal{L}_{2.25}$ (e.g., $\mathcal{Q} = P(Y_{x,W_{x'}})$ on the NDE graph). In these cases, $\mathcal{L}_{2.25}$ does not admit a comparable parameterization for the query.

Table 2: Query-specific 8-state simplex $\phi_{abc} = P(W_1 = a, Y_{0,0} = b, Y_{0,1} = c)$ for $\mathcal{Q} = P(Y_{0,W_1})$, with $g(a, b, c) = Y_{0,W_1}$.

$a = W_1$	$b = Y_{0,0}$	$c = Y_{0,1}$	ϕ_{abc}	$Y_{0,W_1} = g(a,\, b,\, c)$
0	0	0	ϕ_{000}	0
0	0	1	ϕ_{001}	0
0	1	0	ϕ_{010}	1
0	1	1	ϕ_{011}	1
1	0	0	ϕ_{100}	0
1	0	1	ϕ_{101}	1
1	1	0	ϕ_{110}	0
1	1	1	ϕ_{111}	1

Query-Induced Equivalence (Quotient View) Let Ω_{full} be the full $\mathcal{L}_3$ function space for the binary NDE model ($|\Omega_{\text{full}}| = 2^7 = 128$), and consider the query map for $\mathcal{Q} = P(Y_{x,W_{x'}})$ with $(x, x') = (0, 1)$:

$$\pi : \Omega_{\text{full}} \mapsto \Omega_{\text{query}}, \qquad \pi(u) = \big(W_1(u),\, Y_{0,0}(u),\, Y_{0,1}(u)\big).$$

Declare $u \sim u'$ iff $\pi(u) = \pi(u')$. Then $\sim$ partitions Ω_{full} into fibers (equivalence classes) that the query cannot distinguish. Since exactly three bits are used by π and the remaining four bits are unused, each fiber has size $2^4 = 16$ and $|\Omega_{\text{query}}| = \frac{|\Omega_{\text{full}}|}{\text{fiber size}} = \frac{2^7}{2^4} = 2^3 = 8$.

Interpretation. The reduction $128 \to 8$ is a formal quotient: full mechanisms indistinguishable under π collapse into a single query-state.

Ultimately, the context multiplicity m_V quantifies reduction sizes without manual partitioning: the constraint-aware quotient expands parameters only as required by the evidence, yielding memory and runtime gains over the full $\mathcal{L}_3$ (128-state) formulation.

5 EXPERIMENTS

This section empirically validates our query-specific reduction on the running NDE example under the $\mathcal{L}_{2.5}$ marginal information set: (i) the 8-state domain $\Omega_{\text{query}} = \{0, 1\}^3$ suffices for sharp partial identification, and (ii) solving the sharp-bound LP on Ω_{query} is faster than on the lifted 128-state domain.

5.1 SETUP: TARGET, SHARP BOUNDS, AND GROUND TRUTH

Let $t = (A, B, C) \in \Omega_{\text{query}} = \{0, 1\}^3$ and $\phi \in \Delta(\Omega_{\text{query}})$. The target query is $Q = P(Y_{0,W_1} = 1) = g^\top \phi$ with $g(a, b, c) = (1 - a)b + ac$. Under $\mathcal{L}_{2.5}$, the available data enter the LP through three marginal summaries, which serve as the constraint vector. Define the marginal summary vector $p = (p_{w1}, p_{y00}, p_{y01})$ by $p_{w1} = A^\top \phi$, $p_{y00} = B^\top \phi$, and $p_{y01} = C^\top \phi$, where $A, B, C \in \{0, 1\}^8$ are the stacked coordinate vectors over Ω_{query} (fixed ordering), i.e., for $t = (a, b, c)$ we set $A_t = a$, $B_t = b$, and $C_t = c$.

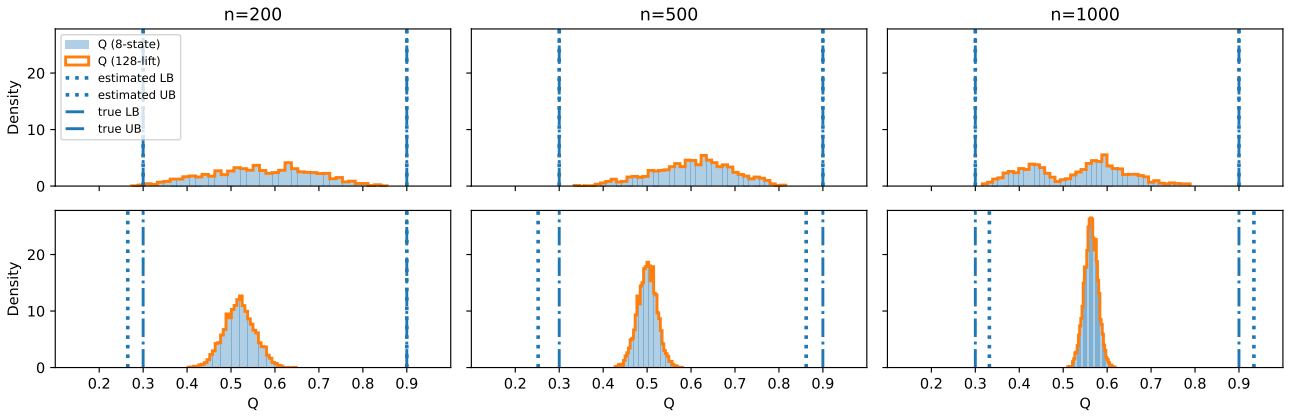


Figure 3: Posterior histograms of Q under binomial summaries (top) and multinomial t-data (bottom) for $n \in \{200, 500, 1000\}$. The 8-state and lifted 128-state evaluations overlap up to numerical precision.

Given p, the sharp bounds are obtained by the linear programs

$$\underline{Q}(p) = \min_{\phi \in \Delta(\Omega_{\text{query}})} g^\top \phi \quad \text{s.t.} \quad A^\top \phi = p_{w1}, \;\; B^\top \phi = p_{y00}, \;\; C^\top \phi = p_{y01}. \tag{6}$$

and $\overline{Q}(p)$ is defined analogously by replacing min with max. We fix $p^\star = (0.6, 0.3, 0.7)$ and consider $n \in \{200, 500, 1000\}$. The interval $[\underline{Q}(p^\star), \overline{Q}(p^\star)]$ is treated as ground-truth sharp bounds in the plots. For the multinomial regime, we instantiate $\phi^\star$ matching $p^\star$ via the product form $\phi^\star(a, b, c) = P(A = a)P(B = b)P(C = c)$ with $P(A{=}1) = p^\star_{w1}$, $P(B{=}1) = p^\star_{y00}$, and $P(C{=}1) = p^\star_{y01}$.

5.2 OBSERVATION REGIMES AND POSTERIOR SAMPLING ON ϕ

Regime I (binomial summaries). We observe three binomial counts $K_r \sim \text{Binomial}(n, p_r)$ with $p_r := \mathbf{r}^\top \phi$ for $r \in \{A, B, C\}$, generated using $p^\star$. We use $\hat{p} = (K_A/n, K_B/n, K_C/n)$ to compute plug-in bounds $[\underline{Q}(\hat{p}), \overline{Q}(\hat{p})]$. With $\phi \sim \text{Dirichlet}(\alpha \mathbf{1}_8)$ ($\alpha = 1$), we draw posterior samples of ϕ via an allocation-augmented Gibbs sampler (Appendix Sec. E).

Regime II (multinomial query-state samples). We observe $t = (A, B, C)$ directly, i.e., $(N_t)_{t \in \Omega_{\text{query}}} \sim \text{Multinomial}(n, \phi)$. Under the same prior, conjugacy yields $\phi \mid (N_t) \sim \text{Dirichlet}(\alpha \mathbf{1}_8 + N)$, so posterior sampling is immediate.

5.3 POSTERIOR PROPAGATION AND 8-VS-128 EQUIVALENCE

For each posterior draw $\phi^{(s)}$, we compute $Q_8^{(s)} = g^\top \phi^{(s)}$ and its full-space counterpart $Q_{128}^{(s)} = q^\top \psi^{(s)}$, where $\psi^{(s)}$ is any lift of $\phi^{(s)}$ to $\Delta(\Omega_{\text{full}})$ and $q(u) := g(\pi(u))$ with $\pi : \Omega_{\text{full}} \mapsto \Omega_{\text{query}}$. Since q is constant on fibers of π, $Q_{128}^{(s)} = Q_8^{(s)}$ for any such lift. Fig. 3 confirms this equality (numerical tolerance) across all n and both regimes, alongside ground-truth and plug-in sharp bounds.

Table 3: Solver-reported optimization time (`solve_time`) per LP solve on the reduced 8-state domain vs. the lifted 128-state domain (speedup = ratio of medians).

	8-state solve time (ms)			128-state solve time (ms)			
n	med	p90	max	med	p90	max	speedup
200	0.154	0.202	0.284	1.810	1.894	2.104	11.72
500	0.152	0.153	0.158	1.807	1.841	1.856	11.85
1000	0.153	0.159	0.166	1.819	1.867	1.947	11.92

5.4 RUNTIME EVALUATION

For each n, we select $K = 10$ posterior draws and solve the sharp-bound LP (i) on the reduced 8-variable form (6) and (ii) on the lifted 128-variable form on Ω_{full}. Table 3 reports median, 90th percentile, and maximum solve times, and the empirical speedup $\text{time}_{128}/\text{time}_8$.

6 CONCLUSION

We proposed a constraint-aware reduction framework to address the exponential growth of state spaces in partial causal identification. By recognizing that canonical domains are tightly coupled with a target query and an available information set, we provided theoretical guarantees for safe state-space quotients and introduced a structural counting methodology to characterize minimal canonical dimensions. Empirically, the query-specific 8-state formulation reproduces the optimal bounds and posterior propagation of the lifted 128-state formulation within numerical precision, while achieving a consistent runtime speedup (median $\approx 11.7\times$ in `solve_time`). Future work will explore extending these geometric reductions to continuous variables and more complex missing-data topologies.

References

Alexander Balke. *Probabilistic Counterfactuals: Semantics, Computation, and Applications*. PhD thesis, Computer Science Department, University of California, Los Angeles, CA, 11 1995.

Alexander Balke and Judea Pearl. Counterfactual probabilities: Computational methods, bounds and applications. In Ramón López de Mántaras and David Poole, editors, *Proceedings of the Tenth Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 46–54. Morgan Kaufmann, 1994.

Irene Botosaru, Isaac Loh, and Chris Muris. An adversarial approach to identification. *arXiv preprint arXiv:2411.04239*, 2024.

Guilherme Duarte, Noam Finkelstein, Dean Knox, Jonathan Mummolo, and Ilya Shpitser. An automated approach to causal inference in discrete settings. *Journal of the American Statistical Association*, 119(547):1778–1793, 2024.

Robin 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.

Noam Finkelstein and Ilya Shpitser. Deriving bounds and inequality constraints using logical relations among counterfactuals. In *Conference on uncertainty in artificial intelligence*, pages 1348–1357. PMLR, 2020.

Jiaying Gu, Thomas M Russell, and Thomas Stringham. Counterfactual identification and latent space enumeration in discrete outcome models. *Review of Economic Studies*, page rdaf058, 2025.

Charles F Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323, 1990.

Piotr Mikusinski, Howard Sherwood, and Michael D. Taylor. The fréchet bounds revisited. *Real Analysis Exchange*, 17(2):759–764, 1991–1992.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jin Tian and Judea Pearl. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1):287–313, 2000.

Hongshuo Yang and Elias Bareinboim. A hierarchy of graphical models for counterfactual inferences. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL `https://causalai.net/r130.pdf`.

Jakob Zeitler and Ricardo Silva. The causal marginal polytope for bounding treatment effects. *arXiv preprint arXiv:2202.13851*, 2022.

Junzhe Zhang and Elias Bareinboim. Bounding causal effects on continuous outcome. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12207–12215, 2021.

Junzhe Zhang, Jin Tian, and Elias Bareinboim. Partial counterfactual identification from observational and experimental data. In *International Conference on Machine Learning*, pages 26548–26558. PMLR, 2022.

Canonical Domain Reduction for Partial Counterfactual Identification (Supplementary Material)

A PRELIMINARIES

We introduce in this section some basic notations and definitions that will be used throughout the paper.

Definition A.1 (C-clique)**.** For a causal diagram $\mathcal{G}$, a c-clique $\mathbf{K} \subseteq \mathbf{V}$ is a c-component such that $\mathbf{K}$ is connected by a bi-directed edge in $\mathcal{G}$ (i.e., $\mathbf{K}$ induces a complete subgraph of bi-directed edges). $\mathcal{K}(\mathcal{G})$ denotes the set of all c-cliques in a causal diagram $\mathcal{G}$.

By definition, every c-clique $\mathbf{K}$ is a c-component, but the reverse direction does not hold. For example, the causal diagram of Fig. 1 contains a single c-component $\mathbf{C} = \{X, Y, Z\}$; yet, it contains two c-cliques $\mathbf{K} = \{Z, Y\}$ and $\mathbf{K} = \{X, Y\}$.

Definition A.2 (C-collection)**.** For a causal diagram $\mathcal{G}$ and a set of queries $\mathcal{P} \subseteq \mathcal{L}_1 \cup \mathcal{L}_2$ over observed variables $\mathbf{V}$, we define a c-collection $\mathcal{C}$ for $(\mathcal{G}, \mathcal{P})$ if

$$\mathcal{C} = \bigcup_{\mathbf{Z} \in \mathcal{Z}} \mathbf{C}(\mathcal{G}[\mathbf{Z}]),$$

where

$$\mathcal{Z} = \{\mathbf{Z} = An(\mathbf{Y})_{\mathcal{G}\setminus\mathbf{X}} \mid P_{\mathbf{x}}(\mathbf{Y}) \in \mathcal{P}\},$$

and $\mathbf{C}(\mathcal{G}[\mathbf{Z}])$ denotes the family of (maximal) c-components of the induced subgraph $\mathcal{G}[\mathbf{Z}]$, i.e., each element is a maximal set of vertices that are connected by a path of bi-directed edges within $\mathcal{G}[\mathbf{Z}]$.

Definition A.3 (Minimal c-collection)**.** Let $\mathcal{G}$ be a causal diagram over $\mathbf{V}$, let $\mathcal{P} \subseteq \mathcal{L}_1 \cup \mathcal{L}_2$ be a set of target queries, and let $\mathcal{C} = \mathcal{C}(\mathcal{G}, \mathcal{P})$ be the c-collection constructed from $(\mathcal{G}, \mathcal{P})$. A subfamily $\mathcal{C}' \subseteq \mathcal{C}$ is a *reduction* of $\mathcal{C}$ if it can be obtained by successively removing identifiable c-components, where identifiability is certified whenever there exists $\mathbf{C}' \in \mathcal{C}$ with $\mathbf{C} \subset \mathbf{C}'$ and $\textsc{Identify}(\mathbf{C}, \mathbf{C}', \mathcal{G}) \neq \text{FAIL}$. A reduction $\mathcal{C}^*$ is *minimal* if no proper subset of $\mathcal{C}^*$ is a reduction. We call such a minimal reduction $\mathcal{C}^*$ the *minimal c-collection for* $(\mathcal{G}, \mathcal{P})$.

The MINCOLLECT procedure has several important properties. First, every $\mathcal{C}^*$ is a $\mathcal{C}'$ (i.e., a reduction), but not every $\mathcal{C}'$ is $\mathcal{C}^*$. Regarding identifiability, for any two structural models compatible with the same diagram that agree on all c-factors in $\mathcal{C} \setminus \{\mathbf{C}\}$, the value of $Q[\mathbf{C}]$ is the same; equivalently, $Q[\mathbf{C}]$ is a (single-valued) functional of $\{Q[\mathbf{C}'] : \mathbf{C}' \in \mathcal{C} \setminus \{\mathbf{C}\}\}$. A reduction is obtained by successively removing c-components that are identifiable from the remaining c-factors. In particular, whenever there exists $\mathbf{C}' \in \mathcal{C}$ with $\mathbf{C} \subset \mathbf{C}'$ and an identification routine (e.g., $\textsc{Identify}(\mathbf{C}, \mathbf{C}', \mathcal{G})$) returns a valid expression for $Q[\mathbf{C}]$, we may safely remove $\mathbf{C}$. Finally, the family of reductions is closed under intersection, hence there exists a unique minimal reduction. The procedure MINCOLLECT returns this minimal reduction, which we call the minimal c-collection for $(\mathcal{G}, \mathcal{P})$, and whose c-factors suffice to represent every target query in $\mathcal{P}$.

Definition A.4 (Explicitization)**.** The explicitization of a causal diagram $\mathcal{G}$ over nodes $\mathbf{V}$, denoted by $\mathcal{H} = \textsc{Explicit}(\mathcal{G})$ is a causal diagram over nodes $\mathbf{V} \cup \mathbf{U}$ constructed as follows.

1. Add each variable in $\mathbf{V}$ as a node of $\mathcal{H}$;
2. For each c-clique $\mathbf{K}_i$ in $\mathcal{G}$, add an exogenous node U_i in $\mathcal{H}$;
3. For each variable V in a c-clique $\mathbf{K}_i$, add an edge $U_i \to V$ in $\mathcal{H}$.

For instance, consider the causal diagram of Fig. 4b, U_1 can be explicitly shown and connected to Z and Y. U_2 can be also explicitly drawn and connected to X and Y.

Proposition A.5. *(Canonical domain cardinality for $\mathcal{L}_1$ distribution) For any SCM $\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{U}) \rangle$ and $U \in \mathbf{U}$, the canonical domain cardinality for $P(\mathbf{V})$ is as follows:*

$$d_U = \prod_{V \in Pa(\mathbf{C}(U))} |\Omega_V|,$$

where

$$P(\mathbf{v}) = \sum_{U \in \mathbf{U}} \sum_{u=1}^{d_U} \mathbb{1}\{\mathbf{V}(\mathbf{u}) = \mathbf{v}\} \prod_{U \in \mathbf{U}} P(u).$$

Proposition A.6. *(Canonical domain cardinality for $\mathcal{L}_2$ distribution) For any SCM $\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{U}) \rangle$, $U \in \mathbf{U}$, and any subset $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$, the canonical domain cardinality for $P(\mathbf{Y_x})$ is as follows:*

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

where

$$P(\mathbf{y_x}) = \sum_{U \in \mathbf{U}} \sum_{u=1}^{d_U} \mathbb{1}\{\mathbf{Y_x}(\mathbf{u}) = \mathbf{y}\} \prod_{U \in \mathbf{U}} P(u).$$

Proposition A.7. *(Canonical domain cardinality for $\mathcal{L}_3$ distribution) For any SCM $\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{U}) \rangle$, let $\mathbf{Y_x}, \cdots, \mathbf{Z_w}$ be an arbitrary set of counterfactual variables. Then, the canonical domain cardinality for the distribution $P(\mathbf{Y_x}, \cdots, \mathbf{Z_w})$ is as follows:*

$$d_U = \prod_{V \in \mathbf{C}(U)} |\Omega_{PA_V} \mapsto \Omega_V| \textit{with } |\Omega_A \mapsto \Omega_B| = |\Omega_B|^{|\Omega_A|},$$

where

$$P(\mathbf{y_x}, \cdots, \mathbf{z_w}) = \sum_{U \in \mathbf{U}} \sum_{u=1}^{d_U} \mathbb{1}\{\mathbf{Y_x} = \mathbf{y}, \cdots, \mathbf{Z_w} = \mathbf{z}\} \prod_{U \in \mathbf{U}} P(u).$$

Counterfactual variables $\mathbf{Y_x}(\mathbf{u}) = \{Y_x(\mathbf{u}) \mid \forall Y \in \mathbf{Y}\}$ are recursively defined as:

$$Y_x(\mathbf{u}) = \begin{cases} \mathbf{x}_Y, & \textit{if } Y \in \mathbf{X} \\ f_Y((Pa_Y)_{\mathbf{x}}(\mathbf{u}), u_Y), & \textit{otherwise} \end{cases}$$

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

Proposition A.8. *(Canonical domain cardinality for multiple distributions from $\mathcal{L}_1$ and $\mathcal{L}_2$) For any SCM $\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{U}) \rangle$, $\mathcal{P} = \{P^{(k)} \equiv P(\mathbf{V}^{(k)}_{\mathbf{x}^{(k)}}) : k = 1, \cdots, K\}$ is a finite collection of $\mathcal{L}_1$ and $\mathcal{L}_2$ targets (each $\mathbf{x}^{(k)}$ possibly $\emptyset$; no nested counterfactuals). Then, the canonical cardinality for $\mathcal{P}$ is as follows:*

$$d_U \equiv |\Omega_U| = \sum_{\mathbf{C} \in \mathcal{C}^*(U)} \prod_{V \in Pa(\mathbf{C}(U))} |\Omega_V|,$$

where

$$P^{(k)}(\mathbf{v}^{(k)}) = \sum_{U \in \mathbf{U}} \sum_{u=1}^{d_U} \mathbb{1}\{\mathbf{V}^{(k)}_{\mathbf{x}^{(k)}}(u) = \mathbf{v}^{(k)}\} \prod_{U \in \mathbf{U}} P(u).$$

We can observe that cardinalities of exogenous domains rely on the total number of c-components in $\mathcal{C}$.

Proposition A.9 (Natural bounds from marginals via a truncated-sum argument)**.** *[Mikusinski et al., 1991–1992] Let $S, T \in \{0, 1\}$ be binary random variables. Suppose only the marginals*

$$p := P(S = 1), \qquad q := P(T = 1)$$

are known. Then the joint probability $r := P(S = 1, T = 1)$ is bounded as

$$\max\{0, p + q - 1\} \leq r \leq \min\{p, q\}.$$

Moreover, both endpoints are attainable by some joint distribution of (S, T) with the given marginals.

(a) IV (b) Non-IV

Figure 4: IV and Non-IV diagrams

Definition A.10 (Layer 2.25 and 2.5 ($\mathcal{L}_{2.25}$ and $\mathcal{L}_{2.5}$))**.** [Yang and Bareinboim, 2025] An SCM $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{u}) \rangle$ induces a family of joint distributions over $\mathbf{V}$, indexed by each interventional value set $\mathbf{x}$. For each $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$ and $\mathbf{x} \in \Omega_{\mathbf{X}}$:

$$
\begin{aligned}
&P^{\mathcal{M}}\left(\bigwedge_{V_i \in \mathbf{Y} \setminus \mathbf{X}} V_{i_{[\mathbf{x}_i]}} = v_i, \bigwedge_{V_i \in \mathbf{Y} \cap \mathbf{X},\ v_i = V_i \cap \mathbf{x}} V_{i_{[\mathbf{x}_i \setminus v_i]}} = v_i\right) \\
&\quad = \sum_u \mathbb{1}\left[\bigwedge_{V_i \in \mathbf{Y} \setminus \mathbf{X}} V_{i_{[\mathbf{x}_i]}}(\mathbf{u}) = v_i, \bigwedge_{V_i \in \mathbf{Y} \cap \mathbf{X},\ v_i = V_i \cap \mathbf{x}} V_{i_{[\mathbf{x}_i \setminus v_i]}}(\mathbf{u}) = v_i\right] P(\mathbf{u}).
\end{aligned} \tag{37}
$$

subject to the following conditions if $P^{\mathcal{L}_{2.25}}$:

(i) $\mathbf{x}_i \subseteq \mathbf{x}$ and $\bigcup_i \mathbf{x}_i = \mathbf{x}$;

(ii) for any $v_i \in \mathbf{x}$ and all $V_j \in \mathbf{Y}$, if $V_i \in An(V_j)$ in $\mathcal{M}_{\mathbf{x} \setminus V_j}$, then $v_i \in \mathbf{x}_j$.

and if $P^{\mathcal{L}_{2.5}}$:

(i) $\mathbf{X}_i \subseteq \mathbf{X}, \mathbf{x}_i \in \Omega_{\mathbf{X}_i}$ and $\bigcup_i \mathbf{X}_i = \mathbf{X}$;

(ii) for any V_i and any $B \in \mathbf{X} \cap \mathbf{Pa}(V_i)$, and for all $V_j \in \mathbf{Y}$: if $V_i \notin \mathbf{X}_j$ and $V_i \in An(V_j)$ in $\mathcal{M}_{\mathbf{x}_j}$, then $\mathbf{x}_i \cap B = \mathbf{x}_j \cap B$.

B MOTIVATING EXAMPLE

B.1 CANONICAL DOMAIN CARDINALITY IN $\mathcal{L}_1, \mathcal{L}_2$ DISTRIBUTIONS

Consider the "IV" diagram in Fig. 4a with binary $X, Y, Z \in \{0, 1\}$. By Def. A.4, if we explicitly draw exogenous variables U_1, U_2, there are two c-components $\mathbf{C}_1 = \{Z\}, \mathbf{C}_2 = \{X, Y\}$ with their exogenous parents $\mathbf{U}_{\mathbf{C}_1} = \{U_1\}, \mathbf{U}_{\mathbf{C}_2} = \{U_2\}$. By Prop. A.5, $P(U_1)$ and $P(U_2)$ are discrete distributions over finite domains $\{1, \cdots, d_{U_1}\}$ and $\{1, \cdots, d_{U_2}\}$. These cardinalities d_{U_1} and d_{U_2} are given by

$$d_{U_1} = |\Omega_Z| = 2$$

$$d_{U_2} = |\Omega_X| \times |\Omega_Y| \times |\Omega_Z| = 8.$$

Similarly, for Fig. 4b, the cardinality can be calculated as $d_U = |\Omega_U| = 8$ since it contains one c-component $\mathbf{C} = \{X, Y, Z\}$. So far, we calculate the cardinalities for $\mathcal{L}_1$ distributions. Now, consider $\mathcal{L}_2$ distributions—i.e., $P(\mathbf{Y}_{\mathbf{x}})$ query can be considered. By Prop. A.6, cardinalities d_{U_1} and d_{U_2} are given by

$$d_{U_1} = |\Omega_{Pa_Z} \times \Omega_Z| = |\Omega_Z| = 2$$

$$d_{U_2} = |\Omega_{Pa_X} \times \Omega_X| \cdot |\Omega_{Pa_Y} \times \Omega_Y| = |\Omega_Z \times \Omega_X| \cdot |\Omega_X \times \Omega_Y| = 16.$$

Similarly, for Fig. 4b, U_1, U_2 are over by the same c-component—i.e., there is only one c-component $\mathbf{C} = \{X, Y, Z\}$. Thus, the cardinality is calculated as

$$d_U = |\Omega_Z| \cdot |\Omega_Z \times \Omega_X| \cdot |\Omega_X \times \Omega_Y| = 32.$$

B.2 CANONICAL DOMAIN CARDINALITY IN MULTIPLE DISTRIBUTIONS

We now consider a problem of finding canonical domain cardinalities for a collection of distributions $\mathcal{P} = \{P_{\mathbf{x}}(\mathbf{Y}) \mid \forall \mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}\}$ in $\mathcal{G}$. For instance, consider the problem of representing distributions $\mathcal{P} = \{P(X, Y, Z), P_x(Y, Z)\}$ in Fig. 4a. The c-collection consists of c-components of the observational distribution $P(X, Y, Z)$ and interventional distribution $P_x(Y, Z)$: $\mathcal{C}_{IV} = \{\{Z\}, \{X, Y\}, \{Y\}\}$. By c-coverings $\mathcal{C}(U)$—the subset of c-components in $\mathcal{C}$ covering an exogenous variable U, exogenous node U_1 is covered by $\mathcal{C}_{IV}(U_1) = \{\{Z\}\}$ and U_2 is covered by $\mathcal{C}_{IV}(U_2) = \{\{X, Y\}, \{Y\}\}$. Then, by Prop. A.8, we can obtain two cardinalities of d_{U_1} and d_{U_2}:

$$d_{U_1} = |\Omega_Z| = 2$$

$$d_{U_2} = |\Omega_X| \cdot |\Omega_Y| \cdot |\Omega_Z| + |\Omega_X| \cdot |\Omega_Y| = 12.$$

Similarly, for Fig. 4b, c-collection $\mathcal{C}_{NIV} = \{\{X, Y, Z\}, \{Y, Z\}\}$ is obtained from $\mathcal{G}$. Then, its c-coverings can be written as $\mathcal{C}_{NIV}(U_1) = \mathcal{C}_{NIV}(U_2) = \{\{X, Y, Z\}, \{Y, Z\}\}$. We have one canonical domain cardinality obtained by Prop. A.8: $d_U = |\Omega_X| \cdot |\Omega_Y| \cdot |\Omega_Z| + |\Omega_X| \cdot |\Omega_Y| \cdot |\Omega_Z| = 16$. As one can see that we require fewer model parameters (i.e., smaller cardinality) if we consider multiple distributions in $\mathcal{G}$ rather than a single distribution. Another insight we get from these examples is that the sparser the network structure of $\mathcal{G}$, the fewer model parameters are required to represent target distributions in $\mathcal{P}$. Concretely, comparing Fig. 4a to Fig. 4b, "IV" diagram contains less bi-directed edges than "Non-IV" diagram. "Sparser" refers to edge-wise sparsity—fewer directed parents into each node and fewer bi-directed (latent confounding) edges. In our construction, the canonical exogenous size satisfies Prop. A.5, Prop. A.6 and Prop. A.8; thus, edge sparsity reduces (i) the number of c-components that survive in the target-specific ancestor subgraphs (shrinking $\mathcal{C}^*$), and (ii) the size of each component's parent set $Pa(\mathbf{C})$. Both effects strictly decrease the summed parent-domain term and therefore the total number of model parameters needed to realize the family $\mathcal{P}$.

C DERIVATION OF THE CLOSED-FORM SHARP BOUNDS FOR RUNNING EXAMPLE

Proposition C.1 (Closed-form sharp bounds under $\mathcal{L}_{2.5}$ marginals)**.** *Let $p = (p_{w1}, p_{y00}, p_{y01})$. For this particular query and constraint set, the bounds admit the following closed form:*

$$\underline{\mathcal{Q}}(p) = \max\{0, p_{y00} - p_{w1}\} + \max\{0, p_{y01} - (1 - p_{w1})\}, \tag{7}$$

$$\overline{\mathcal{Q}}(p) = \min\{p_{y00}, 1 - p_{w1}\} + \min\{p_{y01}, p_{w1}\}. \tag{8}$$

Both endpoints are attainable by some $\phi \in \mathcal{F}(p)$; hence the bounds are sharp.

To connect this back to Eq. (7), observe that in the running example, $\mathcal{Q} = \mathbb{E}[g(A, B, C)] = P(A = 0, B = 1) + P(A = 1, C = 1)$. Applying Prop. A.9 to $(S, T) = (\mathbb{1}\{A = 0\}, \mathbb{1}\{B = 1\})$, we have $p = P(A = 0) = 1 - p_{w1}$ and $q = P(B = 1) = p_{y00}$, giving

$$\max\{0, p_{y00} - p_{w1}\} \le P(A = 0, B = 1) \le \min\{1 - p_{w1}, p_{y00}\}.$$

Similarly, applying it to $(S, T) = (\mathbb{1}\{A = 1\}, \mathbb{1}\{C = 1\})$ yields

$$\max\{0, p_{y01} - (1 - p_{w1})\} \le P(A = 1, C = 1) \le \min\{p_{w1}, p_{y01}\}.$$

Summing the lower (resp. upper) bounds successfully recovers Eq. (7).

D POSTERIOR INFERENCE FOR THE RUNNING EXAMPLE

We consider two complementary observation regimes. In both cases, posterior inference targets the identified distribution ϕ on $\Omega_{\text{query}} = \{0, 1\}^3$ and the induced posterior of the query

$$\mathcal{Q} = \mathbf{g}^\top \phi, \qquad g(a, b, c) = (1 - a)b + ac. \tag{9}$$

Given a posterior draw $\phi^{(s)}$, we compute

$$\mathcal{Q}^{(s)} = \mathbf{g}^\top \phi^{(s)}, \qquad p^{(s)} = \big(\mathbf{A}^\top \phi^{(s)}, \mathbf{B}^\top \phi^{(s)}, \mathbf{C}^\top \phi^{(s)}\big), \qquad [L^{(s)}, U^{(s)}] = \big[\underline{\mathcal{Q}}(p^{(s)}), \overline{\mathcal{Q}}(p^{(s)})\big], \tag{10}$$

where $\underline{\mathcal{Q}}(\cdot), \overline{\mathcal{Q}}(\cdot)$ are the sharp bounds computed from (7) (equivalently, the 8-variable LP over Ω_{query}). We summarize $\{\mathcal{Q}^{(s)}\}_{s=1}^S$ and $\{L^{(s)}, U^{(s)}\}_{s=1}^S$ using posterior quantiles.

Containment diagnostic. As an internal diagnostic, we compute the empirical containment frequency

$$\widehat{P}\big(\mathcal{Q} \in [L, U]\big) = \frac{1}{S} \sum_{s=1}^{S} \mathbb{1}\big\{ L^{(s)} \le \mathcal{Q}^{(s)} \le U^{(s)} \big\}. \tag{11}$$

D.1 MARGINAL-ONLY BINOMIAL DATA

Observation model. We observe three independent binomial summaries

$$K_A \sim \mathrm{Binomial}(n_A, p_{w1}), \quad K_B \sim \mathrm{Binomial}(n_B, p_{y00}), \quad K_C \sim \mathrm{Binomial}(n_C, p_{y01}), \tag{12}$$

where

$$(p_{w1}, p_{y00}, p_{y01}) = (\mathbf{A}^\top \phi, \mathbf{B}^\top \phi, \mathbf{C}^\top \phi).$$

Hence the likelihood depends on ϕ only through these three linear functionals (the "marginals-only" regime).

Prior and sampling strategy. We place a symmetric Dirichlet prior on ϕ,

$$\phi \sim \mathrm{Dirichlet}(\alpha \mathbf{1}_8). \tag{13}$$

Because (12) aggregates coordinates of ϕ, the posterior is not conjugate in closed form. We sample from the posterior via a standard data-augmentation Gibbs sampler in which each binomial count in (12) is represented through latent allocations over compatible query-states; conditional on these allocations, ϕ admits a conjugate Dirichlet update. We refer to Appendix E for the explicit allocation step and full conditional distributions. In experiments we use burn-in $5{,}000$, retain $S = 20{,}000$ draws, and thinning $= 1$.

D.2 DIRECT QUERY-STATE SAMPLES (MULTINOMIAL DATA)

As a complementary regime with richer information, we observe i.i.d. samples of the query-state $t = (A, B, C)$. Let N_t be the counts over $t \in \Omega_{\mathrm{query}}$ and write

$$N = (N_t)_{t \in \Omega_{\mathrm{query}}} \in \mathbb{N}^8, \qquad \sum_{t \in \Omega_{\mathrm{query}}} N_t = n.$$

Under the same prior (13), conjugacy yields the closed-form posterior

$$\phi \mid N \sim \mathrm{Dirichlet}\big(\alpha \mathbf{1}_8 + N\big), \tag{14}$$

from which we draw $\{\phi^{(s)}\}_{s=1}^{S}$ directly.

Visualization used in the main text. For both regimes, we report posterior histograms of $\mathcal{Q}$ across $n \in \{200, 500, 1000, 5000\}$, overlaying the 8-state evaluation $\mathcal{Q}_8^{(s)} = \mathbf{g}^\top \phi^{(s)}$ with the full-space evaluation $\mathcal{Q}_{128}^{(s)} = \mathbf{q}^\top \psi^{(s)}$ obtained by lifting $\phi^{(s)}$ uniformly within each fiber of the quotient map. The near-perfect overlap of these histograms illustrates that the query is fully determined by the reduced 8-state representation. Runtime comparisons between the 8-variable and 128-variable LPs are reported in Table 3 (main text).

D.3 PSEUDOCODE FOR THE SAFE-QUOTIENT LP (ALGORITHM 1)

For completeness, we provide pseudocode for the constraint-aware reduction and reduced LP solve described in Sec. 3.1. Phase 1 (constructing Ω_{full} without unnecessary enumeration) can be instantiated using the structural counting procedure in Sec. 4.

E POSTERIOR INFERENCE FOR THE RUNNING EXAMPLE: EXPLICIT ALLOCATION STEP

This appendix provides the latent-allocation construction used for the *marginal-only binomial* regime in Section D.1. Throughout, $t = (A, B, C) \in \Omega_{\mathrm{query}} = \{0, 1\}^3$ and $\phi \in \Delta(\Omega_{\mathrm{query}})$ denotes its distribution, with prior (13).

Algorithm 1 Constraint-Aware Bounding via Query-Specific Canonical Domains

Require: Causal graph $\mathcal{G}$; target linear query $\mathcal{Q}$; linear information constraints $\mathcal{C}$ with observed values p.
Ensure: Sharp bounds $[L, U]$ for $\mathcal{Q}$.

% Phase 1: Canonical-domain construction (constraint-aware reduction)
Determine the minimal set of response-function / counterfactual components required to *evaluate* $\mathcal{Q}$ and all constraints in $\mathcal{C}$ (cf. Sec. 4).
Construct a full canonical domain Ω_{full} as the state space of these required components. {"Active-variable" filtering is only a heuristic; correctness is guaranteed by the safe-quotient criterion below.}
% Phase 2: Full-domain coefficient evaluation (objective + constraints)
Initialize $c \in \mathbb{R}^{|\Omega_{\text{full}}|}$ and $M \in \mathbb{R}^{m \times |\Omega_{\text{full}}|}$.
for each full canonical state $u \in \Omega_{\text{full}}$ **do**
  $c_u \leftarrow \textsc{EvaluateQuery}(\mathcal{Q}, u)$
  for each constraint $c_i \in \mathcal{C}$ (with value p_i) **do**
    $M_{i,u} \leftarrow \textsc{EvaluateConstraint}(c_i, u)$
  end for
end for{Full LP: $\min / \max_{\psi}\ c^\top \psi$ s.t. $M\psi = p,\ \psi \geq 0,\ \mathbf{1}^\top \psi = 1$.}
% Phase 3: Coarsest safe quotient and reduced LP parameters
Define the *signature* $\sigma(u) := (c_u,\ M_{\cdot,u})$ for each $u \in \Omega_{\text{full}}$.
Define the *coarsest safe* equivalence relation on Ω_{full}:

$$u \sim u' \iff \sigma(u) = \sigma(u') \quad \text{(i.e., } c_u = c_{u'} \ \wedge\ M_{\cdot,u} = M_{\cdot,u'}\text{)}.$$

Let $\Omega_{\text{red}} := \Omega_{\text{full}}/\!\sim$ and let $\pi : \Omega_{\text{full}} \to \Omega_{\text{red}}$ be the quotient map.
Form the pushforward matrix $\Pi \in \{0,1\}^{|\Omega_{\text{red}}| \times |\Omega_{\text{full}}|}$ with $\Pi_{t,u} = \mathbb{1}\{\pi(u) = t\}$.
Initialize reduced objective coefficients $\mathbf{q} \leftarrow \mathbf{0}_{|\Omega_{\text{red}}|}$ and reduced constraint matrix $\mathbf{A} \leftarrow \mathbf{0}_{m \times |\Omega_{\text{red}}|}$.
for each class $t \in \Omega_{\text{red}}$ **do**
  pick any representative $u(t) \in \pi^{-1}(t)$
  $\mathbf{q}[t] \leftarrow c_{u(t)}, \qquad \mathbf{A}_{\cdot,t} \leftarrow M_{\cdot,u(t)}$
end for{Well-defined since $\sigma(u)$ is constant within each class; equivalently, $c = \Pi^\top \mathbf{q}$ and $M = \mathbf{A}\Pi$ (Prop. 3.1).}
% Phase 4: Sharp-bound LP on the reduced canonical domain
Define decision variable $\phi \in \Delta(\Omega_{\text{red}})$ over Ω_{red}.
Solve the reduced LPs:

$$L \leftarrow \min_{\phi \in \Delta(\Omega_{\text{red}})} \mathbf{q}^\top \phi \ \text{ s.t. } \mathbf{A}\phi = p,\ \phi \geq \mathbf{0},\ \mathbf{1}^\top \phi = 1, \qquad U \leftarrow \max_{\phi \in \Delta(\Omega_{\text{red}})} \mathbf{q}^\top \phi \ \text{ s.t. } \mathbf{A}\phi = p,\ \phi \geq \mathbf{0},\ \mathbf{1}^\top \phi = 1.$$

{By Prop. 3.1, $[L, U]$ equals the full-domain sharp bounds.}
Solve using an LP solver (e.g., CVXPY, Gurobi, MOSEK).
return $[L, U]$

E.1 BINOMIAL SUMMARIES AS PARTIALLY OBSERVED CATEGORICAL DATA

Define the index sets

$$S_A^1 = \{t : A(t) = 1\}, \quad S_A^0 = \{t : A(t) = 0\}, \qquad S_B^1, S_B^0,\ S_C^1, S_C^0 \text{ analogously.}$$

Regime I observes (K_A, K_B, K_C) as in (12). Equivalently, for each $r \in \{A, B, C\}$, introduce latent i.i.d. categorical samples

$$T_r^{(i)} \overset{\text{i.i.d.}}{\sim} \text{Categorical}(\phi), \qquad i = 1, \ldots, n_r,$$

and suppose we only observe the binary label $Y_r^{(i)} := r(T_r^{(i)}) \in \{0, 1\}$. Then $K_r = \sum_{i=1}^{n_r} Y_r^{(i)}$ satisfies

$$P(Y_r^{(i)} = 1 \mid \phi) = \sum_{t \in S_r^1} \phi_t, \qquad K_r \mid \phi \sim \text{Binomial}\left(n_r, \sum_{t \in S_r^1} \phi_t\right),$$

which matches (12).

E.2 EXPLICIT ALLOCATION STEP AND DIRICHLET UPDATE

Because the binomial likelihood aggregates multiple coordinates of ϕ, the posterior is not conjugate in ϕ directly. We restore conjugacy by introducing latent *allocations* over Ω_{query}.

Allocation variables. For each $r \in \{A, B, C\}$ and $t \in \Omega_{\text{query}}$, define

$$Z^1_{r,t} := \#\{i : T^{(i)}_r = t,\ r(T^{(i)}_r) = 1\}, \qquad Z^0_{r,t} := \#\{i : T^{(i)}_r = t,\ r(T^{(i)}_r) = 0\}.$$

Then $Z^1_{r,t} = 0$ for $t \notin S^1_r$ and $Z^0_{r,t} = 0$ for $t \notin S^0_r$, with

$$\sum_{t \in S^1_r} Z^1_{r,t} = K_r, \qquad \sum_{t \in S^0_r} Z^0_{r,t} = n_r - K_r.$$

Conditional allocation distributions. Given ϕ and K_r, the latent categories for the K_r "successes" are i.i.d. on S^1_r with

$$P(T = t \mid r(T) = 1, \phi) = \frac{\phi_t}{\sum_{u \in S^1_r} \phi_u} \quad (t \in S^1_r),$$

hence

$$(Z^1_{r,t})_{t \in S^1_r} \mid \phi, K_r \sim \text{Multinomial}\left(K_r,\ \left(\frac{\phi_t}{\sum_{u \in S^1_r} \phi_u}\right)_{t \in S^1_r}\right). \tag{15}$$

Similarly, for the $(n_r - K_r)$ "failures",

$$(Z^0_{r,t})_{t \in S^0_r} \mid \phi, n_r - K_r \sim \text{Multinomial}\left(n_r - K_r,\ \left(\frac{\phi_t}{\sum_{u \in S^0_r} \phi_u}\right)_{t \in S^0_r}\right). \tag{16}$$

Pseudo-count aggregation and Dirichlet update. Combine allocations across marginals into pseudo-counts

$$N_t := \sum_{r \in \{A,B,C\}} \big(Z^1_{r,t} + Z^0_{r,t}\big), \qquad t \in \Omega_{\text{query}}. \tag{17}$$

Under the prior (13), conjugacy yields

$$\phi \mid Z \sim \text{Dirichlet}\big(\alpha \mathbf{1}_8 + N\big). \tag{18}$$

E.3 GIBBS SAMPLER FOR REGIME I

Algorithm 2 summarizes the data-augmentation Gibbs sampler used in experiments.

E.4 POSTERIOR PROPAGATION AND 8-VS-128 EVALUATION

For each retained draw $\phi^{(s)}$, we compute $Q^{(s)}_8 = \mathbf{g}^\top \phi^{(s)}$ and $p^{(s)} = (\mathbf{A}^\top \phi^{(s)}, \mathbf{B}^\top \phi^{(s)}, \mathbf{C}^\top \phi^{(s)})$ as in (10). We then evaluate sharp bounds $[L^{(s)}, U^{(s)}]$ via (7) (or equivalently the 8-variable LP). To validate the "8-state suffices" claim, we also lift $\phi^{(s)}$ to $\psi^{(s)} \in \Delta(\Omega_{\text{full}})$ by distributing mass uniformly within each fiber of the quotient map, and compute $Q^{(s)}_{128} = \mathbf{q}^\top \psi^{(s)}$. In all experiments, $Q^{(s)}_{128}$ matches $Q^{(s)}_8$ up to numerical precision.

F ADDITIONAL FIGURES

The main text reports two 4-panel grids of posterior histograms (binomial and multinomial) across $n \in \{200, 500, 1000, 5000\}$. Here we provide compact summaries of posterior uncertainty and diagnostic agreement.

Figures 5–7 provide complementary summaries that are not shown in the main text. They (i) visualize how posterior uncertainty propagates to the *sharp bounds* themselves, (ii) summarize posterior uncertainty for the query value Q, and (iii) verify numerical agreement between the reduced 8-state and lifted 128-state LP solves.

Algorithm 2 Data-augmentation Gibbs sampler for marginal-only binomial summaries

Input: (K_A, K_B, K_C), (n_A, n_B, n_C), prior α
Initialize $\phi^{(0)} \in \Delta(\Omega_{\text{query}})$
for $\ell = 1, 2, \ldots, (\text{burn} + S)$ **do**
  for $r \in \{A, B, C\}$ **do**
    Sample $(Z^1_{r,t})_{t \in S^1_r}$ from (15)
    Sample $(Z^0_{r,t})_{t \in S^0_r}$ from (16)
  end for
  Form N via (17)
  Sample $\phi^{(\ell)}$ from (18)
  if $\ell >$ burn **then**
    Store $\phi^{(\ell)}$
  end if
end for
Output: posterior draws $\{\phi^{(s)}\}_{s=1}^{S}$

	8-state wall (s)			128-state wall (s)			
n	med	p90	max	med	p90	max	spd
200	0.00835	0.00865	0.01055	0.01055	0.01068	0.01071	1.26
500	0.00827	0.00834	0.00839	0.01051	0.01057	0.01077	1.27
1000	0.00826	0.00829	0.00834	0.01046	0.01052	0.01055	1.27
5000	0.00823	0.00826	0.00827	0.01046	0.01052	0.01062	1.27

Table 4: End-to-end wall-clock runtime per LP solve (speedup = ratio of medians).

Figure 5 (posterior bands for sharp bounds). This figure reports posterior uncertainty for the *identified interval endpoints* $[\underline{Q}(p), \overline{Q}(p)]$ as a function of sample size n, separately for the binomial (marginal-only) and multinomial (direct t-data) regimes. For each posterior draw $\phi^{(s)}$, we form $p^{(s)} = (A^\top \phi^{(s)}, B^\top \phi^{(s)}, C^\top \phi^{(s)})$ and compute the induced sharp bounds $[L^{(s)}, U^{(s)}] = [\underline{Q}(p^{(s)}), \overline{Q}(p^{(s)})]$ using the closed form in Appendix C (equivalently, the 8-variable LP). The shaded regions summarize central posterior bands (e.g., 2.5%–97.5%) of $\{L^{(s)}\}$ and $\{U^{(s)}\}$, while the solid lines show posterior medians. The horizontal reference lines correspond to the ground-truth sharp bounds at $p^\star$, i.e., $[\underline{Q}(p^\star), \overline{Q}(p^\star)]$. As n increases, the multinomial regime yields tighter bands because it directly informs ϕ, whereas the binomial regime remains relatively wider since it only informs the three marginals.

Figure 6 (posterior bands for the query value). This figure summarizes the posterior distribution of $Q = g^\top \phi$ across n under the two observation regimes. The central band and median are computed from $\{Q^{(s)}\}_{s=1}^{S}$ with $Q^{(s)} = g^\top \phi^{(s)}$. In the multinomial regime, posterior concentration is rapid because ϕ is directly observed through counts N_t (conjugate Dirichlet update), while in the binomial regime posterior uncertainty is larger because the likelihood depends on ϕ only through $(A^\top \phi, B^\top \phi, C^\top \phi)$ and the remaining degrees of freedom are only weakly informed by the data. The dashed horizontal reference marks the true query value $Q^\star = g^\top \phi^\star$ used in the simulation setup.

Figure 7 (8-state vs. 128-state LP agreement). This diagnostic quantifies absolute discrepancies between sharp bounds computed by (i) the reduced LP on Ω_{query} (8 variables) and (ii) the lifted LP on Ω_{full} (128 variables). For each n, we sample K posterior draws and compute $(\underline{Q}^{(s)}_8, \overline{Q}^{(s)}_8)$ and $(\underline{Q}^{(s)}_{128}, \overline{Q}^{(s)}_{128})$, then plot $\max_{s \le K} |\underline{Q}^{(s)}_8 - \underline{Q}^{(s)}_{128}|$ and $\max_{s \le K} |\overline{Q}^{(s)}_8 - \overline{Q}^{(s)}_{128}|$ as functions of n. The discrepancies remain at solver tolerance (on the order of 10^{-5}), while the *theoretical* discrepancy is exactly zero under Prop. 3.1. The mild variation across n reflects numerical tolerances and the finite max-over-K summary rather than a systematic dependence on sample size.

Wall-clock vs. solver-time runtimes. Table 4 reports end-to-end wall-clock times for solving the sharp-bound LPs on the reduced 8-state domain and on the lifted 128-state domain. Because the running-example LPs are small, wall-clock times include non-negligible modeling overhead (e.g., CVXPY canonicalization and Python-level problem construction), which limits the apparent speedup. For this reason, the main text reports solver-reported optimization time (`solve_time`); we include the wall-clock table here for completeness.

Stress scaling with larger lifted domains. We synthetically increase the fiber size of the lifted full-space LP (thus increasing the full dimension) while keeping the reduced 8-state LP fixed. Table 5 reports solver-reported optimization time

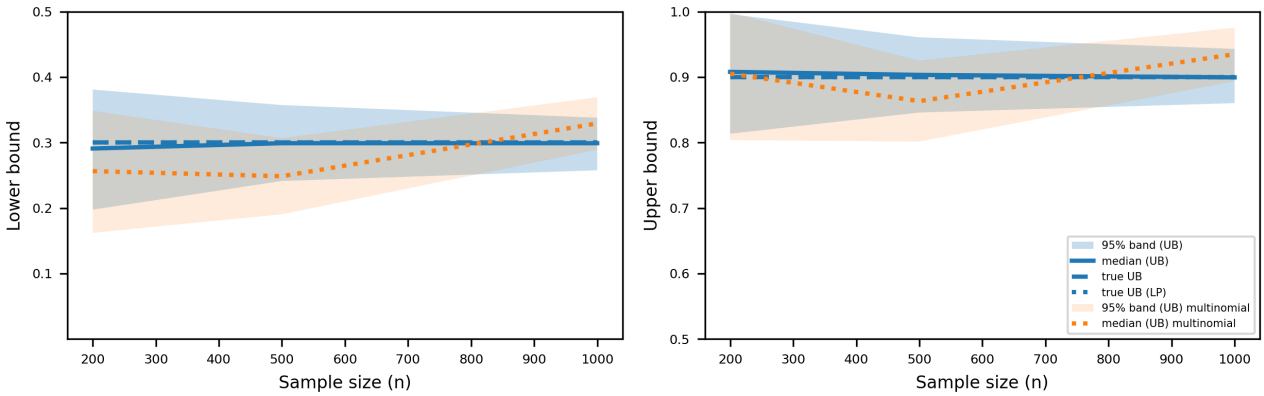


Figure 5: Posterior bands for the sharp lower and upper bounds across n under the binomial (marginal-only) and multinomial (direct query-state) observation regimes. Shaded regions denote central posterior bands and solid lines denote posterior medians; horizontal lines show the ground-truth sharp bounds.

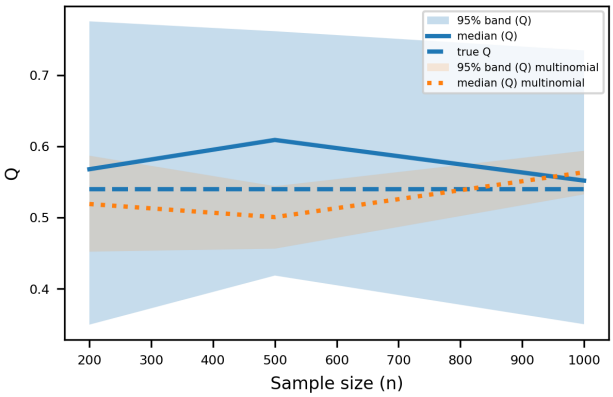


Figure 6: Posterior bands of the query value Q across n under the two observation regimes.

(`solve_time`); the speedup grows rapidly as the lifted domain increases.

G ADDITIONAL EXPERIMENT: P_{obs} INFORMATION SET (64 VS 128 STATES)

We additionally evaluate our *information-set-aware* canonical reduction under the full observational information set $I = P_{\text{obs}}(X, W, Y)$ ("P_{obs} regime"). For each sample size $n \in \{200, 500, 1000, 5000\}$, we generate $S_{\text{pobs}} = 400$ independent realizations of the empirical observational distribution and compute sharp bounds for the same target query using linear programming (LP) on: (i) the full 128-state canonical domain (reference), (ii) the safe quotient domain implied by (Q, I) in this regime (64 states), and (iii) a naive *query-only* reduction baseline (8 states). Since a query-only domain does not preserve the full set of Pobs constraints, we implement the baseline by dropping non-representable constraints ("queryonly-drop"), which reflects a common but unsafe simplification.

Exactness of the safe quotient. Fig. 8 reports the worst-case absolute discrepancies between 64-state and 128-state bounds across the S_{pobs} replications. Across all n, the maximum discrepancies are on the order of 10^{-5} (e.g., $\max |LB_{64} - LB_{128}| \leq 1.35 \times 10^{-5}$ and $\max |UB_{64} - UB_{128}| \leq 1.28 \times 10^{-5}$), consistent with numerical solver tolerances. This confirms that the 64-state safe quotient preserves the feasible set induced by Pobs and reproduces the same sharp bounds as the 128-state formulation.

Runtime. Fig. 10 compares median LP solver times. The 64-state LP is consistently faster, yielding a $1.95\times$–$2.08\times$ median solver-time speedup over the 128-state LP for all tested n.

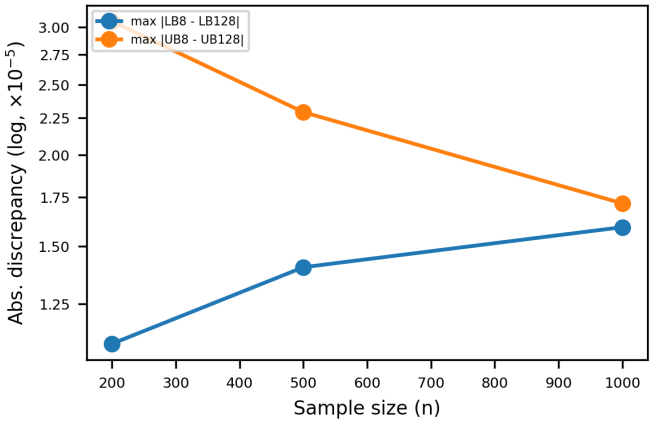
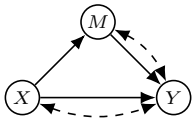


Figure 7: Agreement of sharp bounds computed on the reduced 8-state domain and on the full 128-state domain (after lifting) across posterior draws. Absolute discrepancies are at solver tolerance.

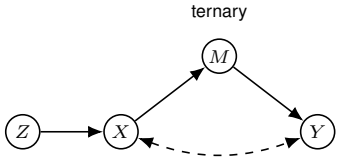
		8-state (ms)			full (ms)			
fib	dim	med	p90	max	med	p90	max	spd
16	128	0.152	0.156	0.158	1.823	1.871	1.929	12.0
64	512	0.152	0.156	0.158	6.828	6.981	7.008	44.8
256	2048	0.152	0.156	0.158	33.224	33.673	34.480	218.0

Table 5: Stress scaling using solver-reported `solve_time` (speedup = ratio of medians).

Query-only baseline. Fig. 9 and Table 6 quantify how the query-only baseline can deviate from the reference 128-state bounds. In our mixture2 configuration, the Pobs-regime bounds are near-vacuous (median LB_{128} is close to 0 and median $UB_{128} = 1$), so the observed deviations are modest; nonetheless, the worst-case discrepancies reach 10^{-6}–10^{-5}, and the fraction of instances with discrepancy exceeding 10^{-6} ranges from 1.5% to 9.25% depending on n. These results illustrate that query-only reductions are not guaranteed to preserve sharp bounds once the information set includes constraints beyond what the reduced representation can encode.



(a) Confounded mediator graph with query $\mathcal{Q} = P(Y_x = 1, Y_{x'} = 0)$.



(b) IV with ternary mediator and query $\mathcal{Q} = P(Y_{1,M_0} = 1)$.

Figure 11: Left: confounded mediator graph for the PNS-type query. Right: IV graph with a ternary mediator for the NDE-type query.

H ADDITIONAL NON-IDENTIFIABLE EXAMPLES: CONTEXT-AWARE CANONICAL CARDINALITIES

To further illustrate that the effective canonical domain depends on the *(query, information-set)* pair rather than on the graph alone, we consider two additional non-identifiable examples beyond the running NDE setting in the main text. Throughout, we use the same counting logic as Def. 4.1 and Prop. 4.2: we count only those parent–world contexts at which each structural function is actually queried. Thus, the resulting sizes should be interpreted in the same sense as Table 1 in the main text, namely as *sufficient reduced canonical sizes* induced by the target query together with the chosen information regime.

Two notions of reduced size. In the additional examples below, we distinguish between two closely related but conceptually different notions. First, Def. 4.1 and Prop. 4.2 yields a *query-only sufficient function-space size*, obtained by counting only those parent–world contexts at which each structural function must be evaluated in order to realize the target query. Second, Prop. 3.1 justifies a *constraint-aware reduction* that preserves both the target query and the selected information constraints.

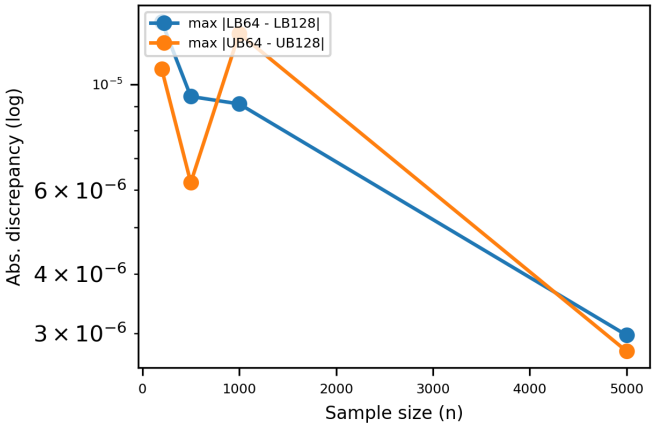


Figure 8: Pobs regime: worst-case (over S_{pobs} replications) absolute discrepancy between 64-state (safe quotient) and 128-state (full) sharp bounds.

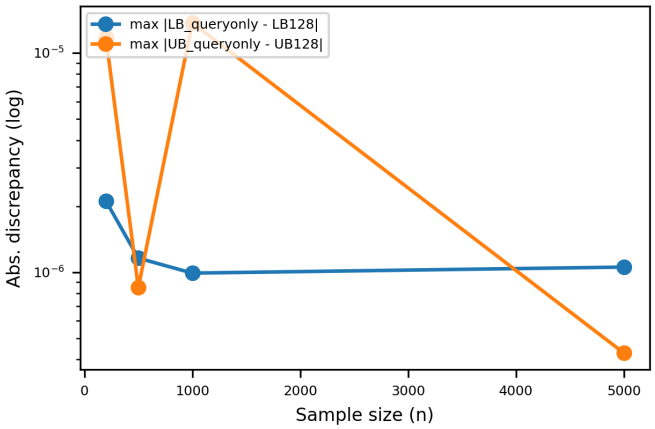


Figure 9: Pobs regime: worst-case absolute discrepancy between the query-only baseline (8 states; dropping non-representable Pobs constraints) and the 128-state reference bounds.

Graph 1: Confounded mediator with a PNS query. (Fig. 11a) Consider the cross-world query

$$\mathcal{Q}_{\text{PNS}} = \Pr(Y_x,\ Y'_{x'}),$$

which, for binary outcomes, is equivalently

$$\mathcal{Q}_{\text{PNS}} = \Pr(Y_x = 1,\ Y_{x'} = 0), \qquad x \neq x',$$

where $Y_x := Y_{x,M_x}$. This graph is non-identifiable because treatment–outcome confounding is present through $X \leftrightarrow Y$, while the front-door pathway is invalidated by mediator–outcome confounding through $M \leftrightarrow Y$.

Graph 2: IV with a ternary mediator and an NDE-type query. (Fig. 11b) Consider Z, X, Y binary and $M \in \{0, 1, 2\}$ (i.e., ternary), and the mixed-world query

$$\mathcal{Q}_{\text{NDE}} = \Pr(Y_{1,M_0} = 1).$$

This example is useful for two reasons: first, the target is genuinely cross-world; second, the ternary mediator increases the worst-case $\mathcal{L}_3$ function-space size substantially, making the difference between worst-case and query/constraint-aware representations more visible.

Why these counts arise. For Fig. 11a (Graph 1), the worst-case $\mathcal{L}_3$ representation keeps the full response-function table $(X, M_0, M_1, Y_{00}, Y_{01}, Y_{10}, Y_{11})$, giving 128 states. For the query $\mathcal{Q}_{\text{PNS}} = \Pr(Y_x = 1, Y_{x'} = 0)$ alone, the context-count rule of Def. 4.1 and Prop. 4.2 yields a sufficient function-space cardinality of 64: X is fixed by intervention, both M_x and $M_{x'}$ are needed, and the outcome mechanism must be evaluated at the four parent–world contexts $(x, 0^x), (x, 1^x), (x', 0^{x'}), (x', 1^{x'})$. Hence the query-only sufficient size is $2^2 \times 2^4 = 64$. If, in addition, one requires

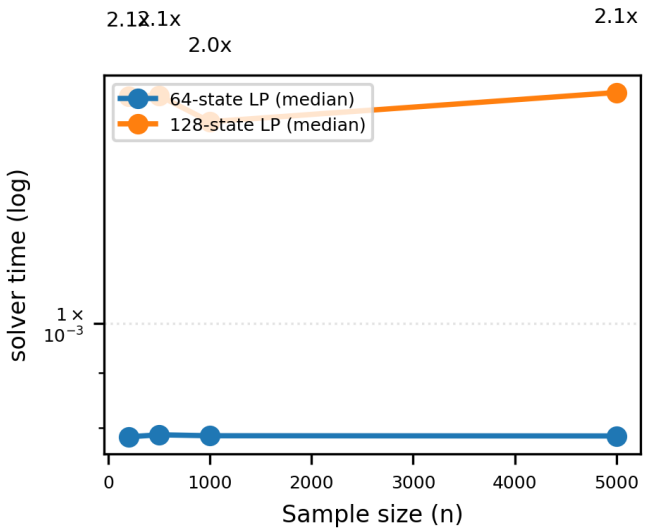


Figure 10: Pobs regime: median LP solver time (log scale) for 64-state vs 128-state formulations. Numbers above markers denote median speedup (128/64).

n	med gap$_{128}$	med gap$_q$	med widen	max widen	speedup	frac diff$> \tau$
200	0.999999	1.000000	5.96e-07	1.20e-05	2.07	0.092
500	0.999999	1.000000	6.70e-07	1.72e-06	2.06	0.015
1000	0.999999	1.000000	5.45e-07	1.43e-05	1.95	0.015
5000	0.999999	1.000000	1.17e-06	1.44e-06	2.08	0.070

Table 6: Query-only reduction can relax (widen) sharp bounds under Pobs constraints. Here we summarize the posterior draws of p_{obs}; widen = gap$_q$ - gap$_{128}$, speedup = median solver-time ratio (128/64), τ is the numerical tolerance.

preservation of the full observational constraints $P_{\text{obs}}(X, M, Y)$, then a smaller *constraint-aware quotient* can be used. Indeed, both the query and every observational cell $\mathbb{1}\{X = x, M_x = m, Y_{x,M_x} = y\}$ depend on a full state only through

$$\big(X,\ M_0,\ M_1,\ Y_{0,M_0},\ Y_{1,M_1}\big).$$

Thus, by the same constraint-aware quotient logic as Prop. 3.1, the active reduced representation has $2^1 \times 2^2 \times 2^2 = 32$ states.

For Fig. 11b (Graph 2), the ternary mediator makes the worst-case $\mathcal{L}_3$ domain much larger. The full response-function representation contains the binary instrument Z, the treatment response table (X_0, X_1), the ternary mediator response table (M_0, M_1), and the six outcome entries $(Y_{00}, Y_{01}, Y_{02}, Y_{10}, Y_{11}, Y_{12})$, leading to 4608 states. Under the query $\mathcal{Q}_{\text{NDE}} = \Pr(Y_{1,M_0} = 1)$ alone, one only needs the mediator response under $X = 0$, namely M_0, together with the three outcome values (Y_{10}, Y_{11}, Y_{12}), so the query-only sufficient function-space size is 24, by the context-count rule of Def. 4.1 and Prop. 4.2. If one additionally requires full observational consistency through $P_{\text{obs}}(Z, X, M, Y)$, then the realized treatment/mediator/outcome paths induced by $Z = 0$ and $Z = 1$ must also be represented. A valid constraint-aware reduced parameterization is obtained from

$$\pi(u) = (Z,\ X_0,\ X_1,\ M_0,\ M_1,\ Y_{1,M_0},\ Y_{X_0,M_{X_0}},\ Y_{X_1,M_{X_1}}).$$

Both the query and every observational cell $\mathbb{1}\{Z = z, X_z = x, M_x = m, Y_{x,m} = y\}$ are functions of $\pi(u)$, so Prop. 3.1 applies. The ambient product-space count of this reduced parameterization is 576, as summarized in Table 7.

No universal monotone ordering across information regimes. Unlike the running NDE example in the main text, these additional examples do not obey a universal monotone pattern such as

$$|\Omega_{\mathcal{L}_3}| \geq |\Omega_{\text{obs}}| \geq |\Omega_{\text{query}}|.$$

This is not a contradiction, but rather a direct consequence of the main point of the paper: the effective canonical size is determined by the specific $(\mathcal{Q}, \mathcal{I})$ pair. In particular, Def. 4.1 and Prop. 4.2 gives a sufficient function-space size for realizing

Table 7: Additional non-identifiable examples analyzed using the context-count logic of Def. 4.1 and Prop. 4.2. The state size is computed from the counterfactual quantities actually invoked by the target query and the corresponding information regime.

Graph / Query	Information Regime	Required Counterfactual Parameters	State Size
Graph 1: Confounded mediator $X \to M \to Y,\ X \leftrightarrow Y,\ M \leftrightarrow Y$, binary X, M, Y, query $\mathcal{Q}_{\mathrm{PNS}} = \Pr(Y_x = 1, Y_{x'} = 0),\ x \neq x'$			
Worst-case ($\mathcal{L}_3$)	None (baseline)	$X,\ M_0, M_1,\ Y_{00}, Y_{01}, Y_{10}, Y_{11}$	$2^1 \times 2^2 \times 2^4 = 128$
Query-only	Only $\mathcal{Q}_{\mathrm{PNS}}$	$M_0, M_1,\ Y_{00}, Y_{01}, Y_{10}, Y_{11}$	$2^2 \times 2^4 = 64$
Full observational	$P_{\mathrm{obs}}(X, M, Y)$	$X,\ M_0, M_1,\ Y_{0,M_0},\ Y_{1,M_1}$	$2^1 \times 2^2 \times 2^2 = 32$
Graph 2: IV + ternary mediator $Z \to X \to M \to Y,\ X \leftrightarrow Y$, binary Z, X, Y, ternary M, query $\mathcal{Q}_{\mathrm{NDE}} = \Pr(Y_{1,M_0} = 1)$			
Worst-case ($\mathcal{L}_3$)	None (baseline)	$Z,\ X_0, X_1,\ M_0, M_1,\ Y_{00}, Y_{01}, Y_{02}, Y_{10}, Y_{11}, Y_{12}$	$2^1 \times 2^2 \times 3^2 \times 2^6 = 4608$
Query-only	Only $\mathcal{Q}_{\mathrm{NDE}}$	$M_0,\ Y_{10}, Y_{11}, Y_{12}$	$3^1 \times 2^3 = 24$
Full observational	$P_{\mathrm{obs}}(Z, X, M, Y)$	$Z,\ X_0, X_1,\ M_0, M_1,\ Y_{1,M_0},\ Y_{X_0,M_{X_0}},\ Y_{X_1,M_{X_1}}$	$2^1 \times 2^2 \times 3^2 \times 2^3 = 576$

a query, whereas Prop. 3.1 justifies reductions that preserve both the query and the selected information constraints. These notions need not be uniformly ordered across examples. For instance, in Fig. 11a, the cross-world PNS query alone requires 64 function-space states, while the full observational constraints factor through a smaller 32-state constraint-aware quotient. By contrast, in Fig. 11b the query-only sufficient size is only 24, but preserving the full observational model requires a substantially richer constraint-aware reduced representation of size 576.

Interpretation. These two examples reinforce the paper's main message. The graph alone does not determine the effective canonical domain size. Instead, the size is governed by which parent–world contexts are actually visited by the *target query* and by the *constraints one insists on preserving*. In particular, the query-only size can be dramatically smaller than the worst-case $\mathcal{L}_3$ size (e.g., 24 instead of 4608 in the IV + ternary mediator example), while adding full observational constraints can increase the required size again (e.g., $24 \to 576$) because more structural evaluations must be retained. Even so, the resulting constraint-aware domain remains substantially smaller than the naive worst-case canonical representation.

H.1 ADDITIONAL AGREEMENT EXPERIMENTS

We next validate the *agreement* claims for the two additional graphs in Fig. 11, using the same evaluation protocol as in the main text (cf. Fig. 3 and Table 3): for each problem instance $(\mathcal{Q}, \mathcal{I})$, we compare (i) the reduced formulation and (ii) a full-domain lift, and verify that they coincide up to numerical precision, while the reduced formulation yields faster LP solves.

Problem instances. We consider four settings:

- **Graph 1 (PNS; Fig. 11a).**
  - *Query-only agreement:* full 128 vs. reduced 64.
  - *Full observational agreement:* full 128 vs. reduced 32 under $P_{\mathrm{obs}}(X, M, Y)$.
- **Graph 2 (NDE-type with ternary mediator; Fig. 11b).**
  - *Query-only agreement:* full 4608 vs. reduced 24.
  - *Full observational agreement:* full 4608 vs. reduced 576 under $P_{\mathrm{obs}}(Z, X, M, Y)$.

Density plots (Figure 12–13). For each setting and each $n \in \{200, 500, 1000\}$, we generate S draws of a feasible reduced canonical state distribution and evaluate the target $\mathcal{Q}$ in two ways: (i) directly in the reduced representation and (ii) via a full-domain *lift* of the reduced draw. In each panel, the reduced evaluation (filled histogram) and the full-lift evaluation (outline) overlap up to numerical precision.

Runtime tables (Table 8–9). We additionally evaluate sharp bounds by repeatedly solving the LP pair (min/max) under each information regime, and record solver-reported optimization time `solve_time` *per LP solve*. We report median / p90 / max (ms) and the speedup defined as the ratio of medians (full / reduced), matching the style of Table 3 in the main text.

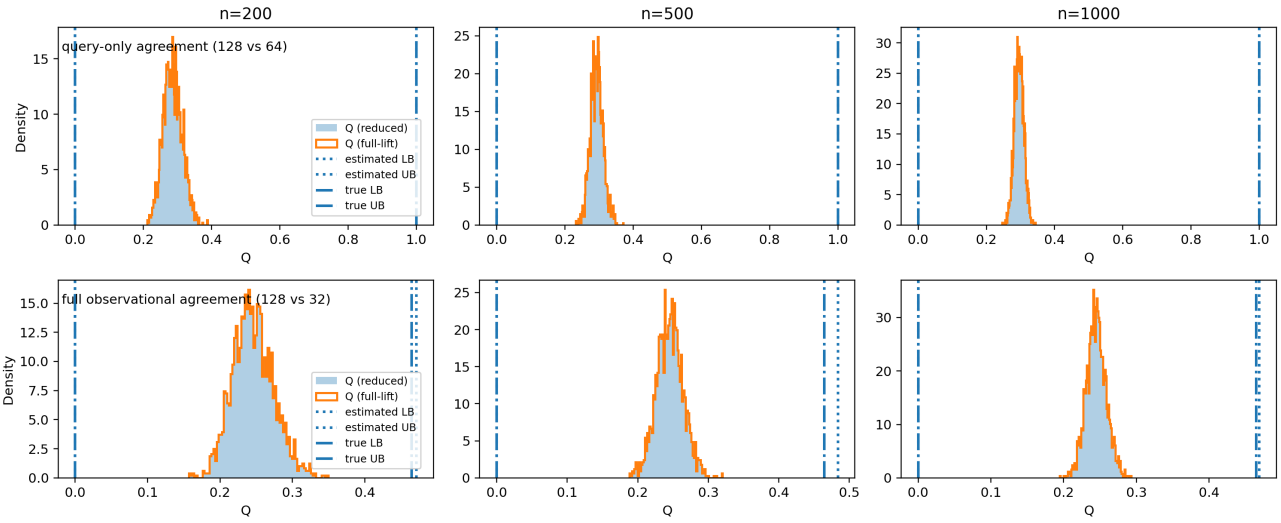


Figure 12: **Graph 1 (PNS).** Density plots of $\mathcal{Q}$ for $n \in \{200, 500, 1000\}$. **Top row:** query-only agreement (full 128 vs reduced 64). **Bottom row:** full observational agreement (full 128 vs reduced 32). Reduced (filled) and full-lift (outline) evaluations overlap up to numerical precision.

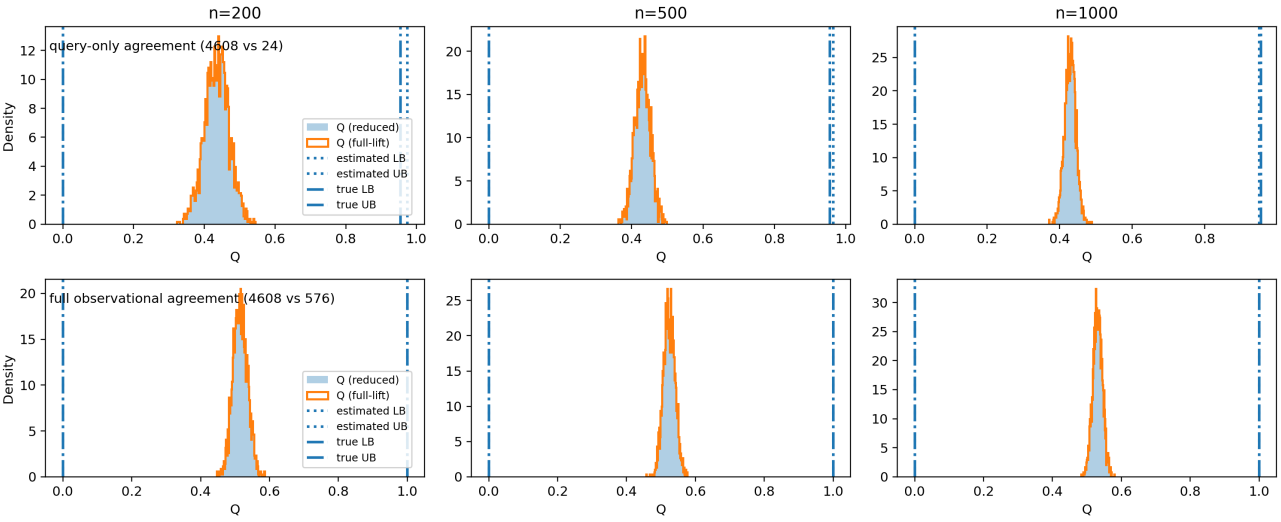


Figure 13: **Graph 2 (NDE-type; ternary mediator).** Density plots of $\mathcal{Q}$ for $n \in \{200, 500, 1000\}$. **Top row:** query-only agreement (full 4608 vs reduced 24). **Bottom row:** full observational agreement (full 4608 vs reduced 576). Reduced (filled) and full-lift (outline) evaluations overlap up to numerical precision.

Numerical agreement (maximum bound discrepancy). In addition to the visual overlap in the density plots, we quantify the maximum absolute discrepancy between sharp bounds computed on the reduced formulation and on the full formulation across repeated LP solves. For Graph 1 (query-only; full 128 vs reduced 64), the worst-case discrepancies are $\max|\mathrm{LB}_{\mathrm{red}} - \mathrm{LB}_{\mathrm{full}}| \leq 4.97 \times 10^{-8}$ and $\max|\mathrm{UB}_{\mathrm{red}} - \mathrm{UB}_{\mathrm{full}}| \leq 1.50 \times 10^{-7}$ (over $n \in \{200, 500, 1000\}$). For Graph 2 (query-only; full 4608 vs reduced 24), the discrepancies are slightly larger, $\max|\mathrm{LB}_{\mathrm{red}} - \mathrm{LB}_{\mathrm{full}}| \leq 1.12 \times 10^{-4}$ and $\max|\mathrm{UB}_{\mathrm{red}} - \mathrm{UB}_{\mathrm{full}}| \leq 1.29 \times 10^{-4}$, which is consistent with solver accuracy limitations for the large full-domain LP.

Summary. Across both graphs and both information regimes, the reduced and full-lift evaluations overlap up to numerical precision in the density plots (Figures 12–13). The runtime results (Tables 8–9) show consistent computational gains: for Fig. 11a (Graph 1), the reduced formulations yield approximately 1.6–2.4× speedups in the query-only regime and

Table 8: **Graph 1 (PNS):** solver-reported optimization time (`solve_time`) per LP solve on the reduced domain vs the full domain (speedup = ratio of medians). Times are in milliseconds.

n	Reduced solve time (ms) med	p90	max	Full solve time (ms) med	p90	max	speedup
Query-only agreement (full 128 vs reduced 64)							
200	0.885	1.846	6.097	1.601	3.890	25.164	1.81
500	0.853	4.826	11.228	1.402	9.888	52.849	1.64
1000	0.637	1.124	2.414	1.260	2.443	2.636	1.98
Full observational agreement (full 128 vs reduced 32)							
200	0.228	0.233	0.241	0.762	0.774	0.864	3.34
500	0.227	0.235	0.244	0.762	0.771	0.787	3.35
1000	0.201	0.202	0.209	0.671	0.686	0.702	3.34

Table 9: **Graph 2 (NDE-type; ternary mediator):** solver-reported optimization time (`solve_time`) per LP solve on the reduced domain vs the full domain (speedup = ratio of medians). Times are in milliseconds.

n	Reduced solve time (ms) med	p90	max	Full solve time (ms) med	p90	max	speedup
Query-only agreement (full 4608 vs reduced 24)							
200	0.234	0.542	0.815	73198.007	75536.322	75572.177	312197
500	0.249	0.473	1.158	72363.037	75454.984	75832.393	290681
1000	0.248	0.299	0.322	74590.155	75357.981	75805.011	301122
Full observational agreement (full 4608 vs reduced 576)							
200	4.105	4.137	4.189	49.692	54.860	55.344	12.11
500	4.088	4.103	4.192	48.342	54.898	54.967	11.83
1000	4.078	4.112	4.191	51.200	55.308	56.017	12.55

$\approx 3.3\times$ under full observational constraints; for Fig. 11b (Graph 2), the reduced formulations yield substantial speedups, including $\approx 12\times$ under full observational constraints (full 4608 vs reduced 576). In particular, for Fig. 11b (Graph 2) in the query-only regime (full 4608 vs reduced 24), the median speedups are extremely large—approximately 2.9×10^5 to 3.1×10^5 across $n \in \{200, 500, 1000\}$—reflecting the fact that each full-domain solve takes on the order of 7.2×10^4 to 7.6×10^4 milliseconds, whereas the reduced-domain solve time is below one millisecond. For completeness, in the query-only regime for Fig. 11a (Graph 1) (full 128 vs reduced 64), the median speedups range from $1.64\times$ to $1.98\times$ across the same sample sizes.

I OMITTED PROOFS

This appendix provides the proofs that were omitted from the main text.

I.1 PROOFS FOR SECTION 3

Proposition 3.1 (Constraint-aware reduction preserves sharp bounds)**.** *Assume there exist a vector $\tilde{c} \in \mathbb{R}^{|\Omega_{\mathrm{red}}|}$ and a matrix $\tilde{M} \in \mathbb{R}^{m\times|\Omega_{\mathrm{red}}|}$ such that*

$$c = \Pi^\top \tilde{c}, \qquad M = \tilde{M}\Pi.$$

Then the full-space and reduced-space partial-identification problems are equivalent:

$$\min_{\psi\in\Delta(\Omega_{\mathrm{full}}):\, M\psi=p} c^\top\psi \;=\; \min_{\phi\in\Delta(\Omega_{\mathrm{red}}):\, \tilde{M}\phi=p} \tilde{c}^\top\phi,$$

and likewise with min *replaced by* max*. In particular, the sharp bounds computed on the reduced simplex equal those computed on the full simplex.*

Proof. Let $\phi := \Pi\psi$. Since $c = \Pi^\top \tilde{c}$,

$$c^\top \psi = (\Pi^\top \tilde{c})^\top \psi = \tilde{c}^\top (\Pi\psi) = \tilde{c}^\top \phi. \tag{19}$$

Moreover, since $M = \tilde{M}\Pi$,

$$M\psi = p \quad \Longleftrightarrow \quad \tilde{M}(\Pi\psi) = p \quad \Longleftrightarrow \quad \tilde{M}\phi = p. \tag{20}$$

Therefore, any feasible ψ for the full problem induces a feasible ϕ for the reduced problem with the same objective value. Hence the reduced optimum cannot exceed the full optimum:

$$\min_{\psi \in \Delta(\Omega_{\text{full}}):\, M\psi = p} c^\top \psi \;\geq\; \min_{\phi \in \Delta(\Omega_{\text{red}}):\, \tilde{M}\phi = p} \tilde{c}^\top \phi.$$

For the reverse inequality, take any feasible ϕ for the reduced problem. Because π is surjective, each fiber $\pi^{-1}(t)$ is nonempty. Choose any selector $u(t) \in \pi^{-1}(t)$ and set

$$\psi_{u(t)} := \phi_t, \qquad \psi_u := 0 \text{ for } u \neq u(t).$$

Then $\psi \geq 0$, $\mathbf{1}^\top \psi = 1$, and for each t,

$$(\Pi\psi)_t = \sum_{u:\pi(u)=t} \psi_u = \psi_{u(t)} = \phi_t,$$

so $\Pi\psi = \phi$. Consequently, (20) gives $M\psi = \tilde{M}(\Pi\psi) = \tilde{M}\phi = p$ and (19) gives $c^\top \psi = \tilde{c}^\top \phi$. Thus every feasible ϕ attains the same objective value as some feasible ψ, implying

$$\min_{\psi \in \Delta(\Omega_{\text{full}}):\, M\psi = p} c^\top \psi \;\leq\; \min_{\phi \in \Delta(\Omega_{\text{red}}):\, \tilde{M}\phi = p} \tilde{c}^\top \phi.$$

Combining both inequalities yields equality. The argument for $\max$ is identical. □

I.2 PROOFS FOR SECTION 4

Lemma 4.3 (Union representation of context sets)**.** *Let $\mathcal{Q}$ be as in Def. 4.1. For a node V, let $\mathsf{Terms}_V(\mathcal{Q})$ denote the set of distinct V-terms that are evaluated in the recursive substitution of structural functions of $\mathcal{Q}$ (including those appearing implicitly as ancestors of the counterfactuals in $\mathcal{Q}$ (e.g., Y_x, $Y_{x'}$, $Y_{x,W_{x'}}$; duplicates with identical labels/structure are merged). For each $T \in \mathsf{Terms}_V(\mathcal{Q})$, let $S_V(T)$ be the set of labeled parent assignments to $pa(V)$ under which f_V is evaluated when answering the single term T. Then*

$$S_V(\mathcal{Q}) = \bigcup_{T \in \mathsf{Terms}_V(\mathcal{Q})} S_V(T),$$

In particular, always $m_V(\mathcal{Q}) \leq \sum_{T \in \mathsf{Terms}_V(\mathcal{Q})} |S_V(T)|$, with equality if the sets are disjoint.

Proof. By definition, $S_V(\mathcal{Q})$ is the set of all labeled parent assignments at which f_V is evaluated during the computation of the whole query. Each evaluation of f_V occurs while computing at least one V-term T, hence it belongs to $S_V(T)$ for some T. Conversely, every context in some $S_V(T)$ is used in evaluating $\mathcal{Q}$ and therefore belongs to $S_V(\mathcal{Q})$. This gives the union identity; the inequality is immediate. □

J DISCUSSION

Exact reduction via constraint-aware quotients. Canonical LP formulations for partial identification often encode many counterfactual degrees of freedom that are irrelevant to the sharp-bound problem at hand. Our key point is that relevance is defined *jointly* by the LP objective and the information constraints: reductions that preserve only the query can be unsound, whereas collapsing directions that are invisible to *both* objective and constraints yields an *exact* speed/space gain. Prop. 3.1 formalizes this requirement through pushforward factorization, guaranteeing that the reduced LP recovers the full-domain sharp bounds.

Coarsest quotient and context multiplicity. The coarsest safe quotient merges full states iff they are indistinguishable to every linear functional appearing in the optimization (objective or constraints), and Alg. 1 implements this via the signature $\sigma(u) = (c_u, M_{\cdot,u})$. Complementarily, context multiplicity provides a structural proxy for minimal dimension: for realizable mixed-world queries, Prop. 4.2 shows that a sufficient function-space size factors as $\prod_V |\Omega_V|^{m_V(\mathcal{Q})}$, where $m_V(\mathcal{Q})$ counts distinct labeled parent–world contexts visited in the structural expansion; Lem. 4.3 computes these contexts as a union over distinct V-terms, avoiding duplicate counting.

Empirical takeaway. In the running NDE example under the $\mathcal{L}_{2.5}$ marginal information set, the lifted 128-state formulation collapses to an 8-state simplex without changing sharp bounds: posterior propagation matches under 8-state evaluation and 128-state lifting up to numerical tolerance, while solver-reported `solve_time` improves by about 11–12$\times$ (Table 3). Stress scaling (Table 5) confirms that the gain grows rapidly as the lifted full-domain dimension increases.

Limitations and scope. Our reductions are exact for problems admitting a finite canonical-domain LP with linear objectives and linear information constraints, but the *magnitude* of computational savings is problem-dependent: if the objective/constraints distinguish most full states (i.e., $\sigma(u)$ is nearly injective), the coarsest safe quotient can be close to trivial and yield only modest gains. Extensions to continuous variables or nonlinear/semialgebraic constraints require redefining indistinguishability relative to the relevant function class, and robust computation of $\mathcal{T}_V(\mathcal{Q})$ may require symbolic tooling for large graphs and complex query syntaxes. Finally, while $\mathcal{L}_{2.5}$ is motivated by realizable split/edge-style regimes, interpreting mixed-world labels as physically implementable interventions requires additional assumptions: realizability results indicate that counterfactual-randomization actions are only defined for subsets of *direct children* of a decision variable and cannot bypass a child to directly control a non-child descendant's perception of the decision variable.

Broader outlook. More broadly, constraint-aware reduction suggests treating canonical domains as problem instances indexed by $(\mathcal{Q}, \mathcal{C})$ rather than objects determined solely by the diagram. This perspective is particularly valuable for iterative workflows that repeatedly solve sharp bounds across many posterior draws or candidate decisions, where an exact quotient can amortize the cost of high-dimensional LP solves while preserving the same identified-set semantics.