Computational and Algebraic Questions in Causal Inference

Alex Markham[1,3][0000-0002-5495-1077], Erin E. Gabriel[2,3][0000-0002-0504-8404], and Michael C. Sachs[2,3][0000-0002-1279-8676]

[1] Department of Mathematical Sciences, University of Copenhagen
awm@math.ku.dk
[2] Section of Biostatistics, University of Copenhagen
[3] Pioneer Centre for SMARTbiomed

Abstract. We revisit a classic causal inference problem—computing symbolic bounds on causal effects that are only partially identifiable—introducing it to an audience in algebraic statistics. The problem is typically approached as a linear program and subsequently solved by vertex enumeration. Subject to certain assumptions on the underlying causal model, this leads to provably valid and sharp bounds, however this approach quickly becomes computationally intractable as the complexity of the underlying causal model increases. We ask the community: what computational and algebraic tools could lead to methods that are more efficient or that require less restrictive assumptions?

Keywords: Causal inference · Algebraic statistics · Linear programming · Vertex enumeration · Discrete DAG models.

1 Introduction

Causal models extend statistical models by specifying how distributions change under intervention. Formally, a *statistical model* [10] is a collection $\mathcal{P} = \{P, P', \dots\}$ of probability distributions over a sample space $\mathcal{X}$, while a *causal model* $\mathcal{C} = \{\mathbf{P}_\mathcal{I} = \{P_I\}_{I \in \mathcal{I}}, \mathbf{P}'_\mathcal{I} = \{P'_I\}_{I \in \mathcal{I}}, \dots\}$ is a collection of sets of distributions over $\mathcal{X}$ for a fixed set of interventions indexed by $\mathcal{I}$. Common frameworks include potential outcome models [14,13] and graphical causal models [18,12], which allow for probabilistic reasoning and statistical estimation to be extended to account for interventions, such as taking medication or knocking down a gene.

Causal inference concerns the study of *causal estimands*, which are functionals $f : \mathcal{C} \to \mathbb{R}$. A causal estimand f is said to be *identifiable* from an observable set of interventions indexed by $\mathcal{I}_\text{obs} \subseteq \mathcal{I}$ if there exists an $f^* : \{\mathbf{P}_{\mathcal{I}_\text{obs}}, \mathbf{P}'_{\mathcal{I}_\text{obs}}, ...\} \to \mathbb{R}$ such that for all $\mathbf{P}_\mathcal{I} \in \mathcal{C}$ we have $f^*(\mathbf{P}_{\mathcal{I}_\text{obs}}) = f(\mathbf{P}_\mathcal{I})$. Roughly, if a causal estimand is identifiable, this means it is possible to reason from the observed setting about how the system would respond and behave in settings that have not been observed.

In graphical causal models, Shpitser and Pearl [16] provide a forbidden graph characterization of identifiability, showing that a large class of causal estimands is identifiable if and only if a certain structure (called a "hedge", and representing a problematic kind of unobserved confounding) is absent in the underlying graph. In this paper, we focus on a subclass of causal graphs introduced by Sachs et al. [15] in which the causal estimand of interest is not, in general, identifiable, but in which the unobserved confounding imposes linear constraints on the observable distributions, defining a polytope over which sharp bounds on some specific estimands can be computed by linear programming.

The rest of the paper is organized as follows: in Section 2, we formally introduce graphical causal models, including example causal estimands and the problem of unobserved confounding; in Section 3, after presenting the approach of Sachs et al. [15] to computing causal bounds and the approach of Shridharan and Iyengar [17] to reduce the number of variables in the bounds computation, we provide a characterization of exactly when such a reduction is possible and a method for directly constructing the reduced variable set; and finally, in Section 4, we pose open questions about the computational tractability and algebraic structure of causal bounds.

2 Graphical causal models

Consider discrete random variables $X_{[d]} \coloneqq \{X_1, \dots, X_d\}$ taking values in the finite joint sample space $\mathcal{X}_{[d]} \coloneqq \prod_{i \in [d]} \mathcal{X}_i$. Associate with these random variables a directed acyclic graph (DAG) $\mathcal{G} = ([d], E)$ with nodes $[d]$ and edges E, and denote the parent set of node $j \in [d]$ as $\mathrm{pa}(j) \coloneqq \{i \mid i \to j \in E\}$. Define a *causal discrete DAG model* $\mathcal{D} = (\mathcal{C}, \mathcal{G})$ to be a causal model whose corresponding probability mass functions (PMFs) factorize according to the DAG[4] for each intervention; that is, with each $I \in \mathcal{I}$ indicating which random variables $X_I \subseteq X_{[d]}$ are targeted by the corresponding intervention, the associated PMF factorizes as

$$p^I(X_{[d]}) = \prod_{i \in I} p^I(X_i \mid X_{\mathrm{pa}(i)}) \prod_{j \in [d] \setminus I} p^{\emptyset}(X_j \mid X_{\mathrm{pa}(j)}), \tag{1}$$

where $p^{\emptyset}$ denotes the observational[5] PMF. In this way, the DAG encodes which conditionals remain invariant across different interventions [20].

As a consequence of this factorization, the causal discrete DAG model can be represented compactly via conditional probability tables (CPTs) [6], as in the following example.

Example 1. A causal discrete DAG model $\mathcal{D} = (\mathcal{C}, \mathcal{G})$ for three binary random variables is shown in Figure 1. The factorization in Equation (1) implies that, for example, the probability of the observational joint outcome $x_{[3]} = (1, 0, 1)$ is obtained from the CPTs as $p^{\emptyset}(X_{[3]} = x_{[3]}) = p_{1**}p_{10*}p_{*01}$. $\diamond$

$p^{\emptyset}(X_1)$:

$X_1 = 0$	$X_1 = 1$
p_{0**}	p_{1**}

$p^{\emptyset}(X_2 \mid X_1)$:

	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	p_{00*}	p_{01*}
$X_1 = 1$	p_{10*}	p_{11*}

$p^{\emptyset}(X_3 \mid X_2)$:

	$X_3 = 0$	$X_3 = 1$
$X_2 = 0$	p_{*00}	p_{*01}
$X_2 = 1$	p_{*10}	p_{*11}

$p^{\{2\}}(X_2 \mid X_1)$:

	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	p'_{00*}	p'_{01*}
$X_1 = 1$	p'_{10*}	p'_{11*}

Fig. 1: Example causal discrete DAG model, defined by the DAG $\mathcal{G}$ on the left and the corresponding CPTs for the set of interventions $\mathcal{I} = \{\emptyset, \{2\}\}$ on the right.

From an algebraic statistical point of view, the CPTs in Figure 1 give a polynomial parameterization of a subset of the probability simplex. Writing

$$\theta \in \Theta \coloneqq \big\{(p_{0**}, p_{1**}, p_{00*}, \dots, p_{*11}) \in (0,1)^m : \text{row sums equal } 1\big\},$$

the observational joint probabilities define a polynomial map

$$\phi_{\emptyset} : \Theta \longrightarrow \Delta(\mathcal{X}_{[3]}) \subset \mathbb{R}^{|\mathcal{X}_{[3]}|}, \qquad \theta \longmapsto \big(p^{\emptyset}(X_{[3]} = x_{[3]})\big)_{x_{[3]} \in \mathcal{X}_{[3]}},$$

where each component $p^{\emptyset}(X_{[3]} = x_{[3]})$ is given by the factorization in (1). In this way, the discrete DAG model is a semialgebraic subset of the probability simplex, cut out by polynomial

---

[4] This is known as the *Markov factorization* [18,12,7] and relates separation in the graph to conditional independence in the corresponding distributions.

[5] Note that there is a distinction between the *observational* distribution and *observed* (or observable) distributions: 'observational' is in contrast to 'interventional', and it describes a causal system in its natural or unperturbed state; 'observed' is in contrast to 'unobserved' and simply describes what sorts of data are available, which could be observational or interventional.

equalities (normalization, factorization) and inequalities (nonnegativity). In the fully observed case, the DAG's conditional independence relations give binomial generators for the model's vanishing ideal [19]; latent variables complicate this ideal, motivating the reparameterization in Section 3 using response functions. More generally, for a discrete DAG model $\mathcal{D} = (\mathcal{C}, \mathcal{G})$ on d nodes and with intervention set $\mathcal{I}$, we obtain a set of polynomial maps $\{\phi_I\}_{I\in\mathcal{I}}$ whose images describe the observational and interventional distributions compatible with $\mathcal{G}$.

Several different causal estimands are of practical interest in graphical causal models [11], for example the *average treatment effect (ATE)*:

Example 2. Continuing with the model in Figure 1, consider an intervention $\mathrm{do}(X_2 = t)$ in which the conditional factor generating X_2 is replaced by a point mass at $t \in \{0, 1\}$. We write $p^{\mathrm{do}(X_2=t)}$ for the resulting interventional distribution. Let $\mathcal{I} = \{\emptyset, \mathrm{do}(X_2 = t)\}$ and $\mathcal{I}_{\mathrm{obs}} = \{\emptyset\}$. By the factorization in Equation (1),

$$p^{\mathrm{do}(X_2=t)}(X_{[3]} = x_{[3]}) = p^{\emptyset}(X_1 = x_1)\,\mathbf{1}\{X_2 = t\}\,p^{\emptyset}(X_3 = x_3 \mid X_2 = t).$$

Then, since X_3 is binary,

$$\begin{aligned}\mu_t \coloneqq \mathbb{E}\big[X_3 \mid \mathrm{do}(X_2 = t)\big] &= \sum_{x_3\in\{0,1\}} x_3\,p^{\emptyset}(X_3 = x_3 \mid X_2 = t)\\ &= p^{\emptyset}(X_3 = 1 \mid X_2 = t),\end{aligned}$$

so the ATE is $f_{\mathrm{ATE}}(p^{\emptyset}) = \mu_1 - \mu_0$. In this discrete causal DAG model, μ_t is determined by $p^{\emptyset}$, so the ATE is identifiable. ⋄

Causal estimands such as the ATE can be viewed as rational functions of the underlying parameters θ. We collect all interventional and observational distributions into a single map

$$\Phi : \Theta \longrightarrow \prod_{I\in\mathcal{I}} \Delta(\mathcal{X}_{[d]}), \qquad \theta \longmapsto \big(\phi_I(\theta)\big)_{I\in\mathcal{I}},$$

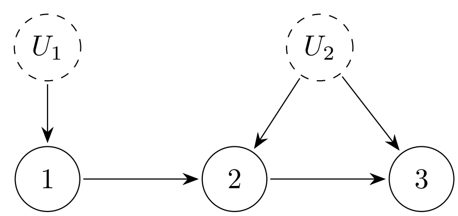

Fig. 2: DAG representing unobserved confounding with latent parent U_2 of X_2 and X_3.

and let $\Phi_{\mathcal{I}_{\mathrm{obs}}}$ denote its restriction to the observable coordinates $\{\phi_I\}_{I\in\mathcal{I}_{\mathrm{obs}}}$. Identifiability from $\mathcal{I}_{\mathrm{obs}}$ then asks whether the value of such a rational function is determined solely by the joint image of $\Phi_{\mathcal{I}_{\mathrm{obs}}}$. Algebraically, this is an elimination problem: the equations $\Phi_{\mathcal{I}_{\mathrm{obs}}}(\theta) = \pi_{\mathrm{obs}}$ express each observable probability as a polynomial in the CPT parameters θ; eliminating θ from these equations—for instance via Gröbner basis methods—yields polynomial constraints on π_{obs} alone that cut out the model's image. Identifiability of an estimand then asks whether a rational function of π_{obs} exists that agrees with the estimand on this image—equivalently, whether the estimand is constant along the fibers of $\Phi_{\mathcal{I}_{\mathrm{obs}}}$. Here, for a given observable configuration π_{obs}, the *fiber* of $\Phi_{\mathcal{I}_{\mathrm{obs}}}$ over π_{obs} is

$$\Phi^{-1}_{\mathcal{I}_{\mathrm{obs}}}(\pi_{\mathrm{obs}}) \coloneqq \{\theta \in \Theta : \Phi_{\mathcal{I}_{\mathrm{obs}}}(\theta) = \pi_{\mathrm{obs}}\},$$

the set of all parameter points—including different specifications of the unobserved variables in U—that induce the same distributions of the observed variables in V.

Example 3. Now modify the model in Figure 1 by introducing an unobserved variable U_2 that is a common parent of X_2 and X_3, resulting in the canonical instrumental variable DAG in Figure 2 (with X_1 as instrument, X_2 as treatment, and X_3 as outcome [1]). We still consider $\mu_t \coloneqq \mathbb{E}[X_3 \mid \mathrm{do}(X_2 = t)]$ and $f_{\mathrm{ATE}} = \mu_1 - \mu_0$.

Assume for the moment that U_2 is discrete (we return to the general case in Example 4). With U_2 present, the interventional distribution is

$$p^{\mathrm{do}(X_2=t)}\big(X_3\big) = \sum_{u_2} p^{\emptyset}\big(X_3 \mid X_2 = t, U_2 = u_2\big)\,p^{\emptyset}(U_2 = u_2),$$

so μ_t depends on the *marginal* distribution $p^{\emptyset}(U_2)$. In contrast, the observational conditional is

$$p^{\emptyset}(X_3 \mid X_2 = t) = \sum_{u_2} p^{\emptyset}\big(X_3 \mid X_2 = t, U_2 = u_2\big)\, p^{\emptyset}(U_2 = u_2 \mid X_2 = t),$$

which involves the *conditional* $p^{\emptyset}(U_2 \mid X_2 = t)$, not the marginal. Observational data determine $p^{\emptyset}(X_3 \mid X_2 = t)$ but not the marginal and conditional separately: different choices of $p^{\emptyset}(U_2)$ and $p^{\emptyset}(U_2 \mid X_2 = t)$ can yield the same observational distribution but different values of μ_t, so the ATE is not identifiable. $\diamond$

Algebraically, latent variables project the parameterized variety onto the observable coordinates, and the nontrivial fiber structure of this projection makes the confounded ATE nonidentifiable: in Example 2, the ATE is constant along fibers of $\phi_{\emptyset}$ and hence a rational function of the observable coordinates; whereas in Example 3, fibers contain parameter points with different ATE values.

3 Causal bounds

When a causal estimand is not identifiable, one can still seek sharp bounds on its value. Balke and Pearl [1] show that in an instrumental variable setting (like Figure 2), the set of observable distributions compatible with the causal model forms a polytope, so sharp bounds reduce to optimizing a linear functional over it. Sachs et al. [15] generalize this approach to a larger class of graphical causal models, namely those with the *response-linear* property:

Property 1 (Proposition 2, [15]). Let $\mathcal{G} = (U \cup V, E)$ be a DAG whose nodes are partitioned into unobserved U and observed V, with V further partitioned into left and right parts $V = V_L \cup V_R$. Fix $\mathcal{I}_{\text{obs}} = \{\emptyset\}$ and let $\mathcal{I}$ additionally consist of all $\text{do}(\cdot)$ interventions on subsets of nodes in V_R. A causal discrete DAG model with $\mathcal{G}, \mathcal{I}$, and $\mathcal{I}_{\text{obs}}$ is called *response-linear* if it satisfies:

1. for all $v \in V_L$ and $w \in V_R$, there is no edge $w \to v \in E$;
2. for all $u \in U$, either $\text{ch}(u) \cap V_L = \emptyset$ or $\text{ch}(u) \cap V_R = \emptyset$;
3. there exists a $u \in U$ such that $\text{ch}(u) = V_L$; and
4. there exists a $u \in U$ such that $\text{ch}(u) = V_R$.

The first step in constructing the linear program is to reparameterize the causal model in terms of deterministic *response functions*. For each variable X_j such that $j \in V_R$ and each configuration $x_{\text{pa}(j)\cap V}$ of its observed parents, a local response function r_j specifies a deterministic output $x_j = r_j(x_{\text{pa}(j)\cap V})$—equivalently, r_j encodes the potential outcomes [14] of X_j: the value X_j would take under each possible configuration of its observed parents. Let

$$\mathcal{R}_j \coloneqq \big\{ r_j : \mathcal{X}_{\text{pa}(j)\cap V} \to \mathcal{X}_j \big\}$$

denote the finite set of local response functions for node j. Define the global response function set

$$R \coloneqq \bigtimes_{j \in V_R} \mathcal{R}_j.$$

Each $\rho = (r_j)_{j \in V_R} \in R$ corresponds to a deterministic assignment of X_{V_R} as a function of its observed parents.

The key insight is that fixing the latent variables with children in V_R determines a specific response function $\rho \in R$, so any distribution over U induces a distribution $w \in \Delta^{|R|-1}$ over R:

Example 4. We make this concrete for the confounded model of Example 3 (Figure 2), with $V_L = \{1\}$, $V_R = \{2, 3\}$, and a latent confounder U_2 with arbitrary (possibly continuous or multivariate) distribution. Dropping the assumption of discrete U_2 in Example 3, a CPT-like parameterization would specify the marginals $p^{\emptyset}(X_1)$ and $p^{\emptyset}(U_2)$ together with the conditionals $p^{\emptyset}(X_2 \mid X_1, U_2)$ and $p^{\emptyset}(X_3 \mid X_2, U_2)$, where marginalizing over U_2 yields the observable joint as an expectation over U_2 of a product of conditionals. In contrast, the response function reparameterization makes the observable map linear in a mixture weight vector w.

The latent sample space of U_2 admits a *canonical partition* into at most $|R|$ equivalence classes [15, Prop. 1][12, Sect. 8.2.2]: two values are equivalent if and only if they induce the same response function pair $(r_2, r_3) \in R$, i.e., the same deterministic outputs of X_2 and X_3 as functions of their observed parents. Let $\rho(u_2) \in R$ denote the label of the class containing u_2. The local response functions for node 2 are all functions $r_2 : \mathcal{X}_1 \to \mathcal{X}_2$, of which there are $|\mathcal{R}_2| = 2^2 = 4$; for node 3, $r_3 : \mathcal{X}_2 \to \mathcal{X}_3$, giving $|\mathcal{R}_3| = 2^2 = 4$. So $|R| = 4 \times 4 = 16$. The pushforward of U_2 through $\rho(\cdot)$ is a distribution on R with masses

$$w_\rho \coloneqq p^{\emptyset}\big(\{u_2 : \rho(u_2) = \rho\}\big), \qquad w \in \Delta^{|R|-1}.$$

Conditional on the class $\rho = (r_2, r_3)$, X_2 and X_3 are deterministic functions of their observed parents, so the mixture over classes gives the observable joint as

$$p^{\emptyset}(X_{[3]} = x_{[3]}) = p^{\emptyset}(X_1 = x_1) \sum_{\substack{\rho \in R: \\ r_2(x_1) = x_2,\ r_3(x_2) = x_3}} w_\rho,$$

linear in w. Similarly, under the $\mathrm{do}(X_2 = t)$ intervention, $X_3 = r_3(t)$ is deterministic conditional on ρ, so the causal estimand from Examples 2 and 3 is

$$\mu_t = p^{\mathrm{do}(X_2 = t)}(X_3 = 1) = \sum_{\substack{\rho \in R: \\ r_3(t) = 1}} w_\rho,$$

also linear in w. $\diamond$

Conditions 3 and 4 of Property 1 make this work: they ensure that a single latent variable mediates all dependence within V_L and all dependence within V_R; together with conditions 1 and 2, this forces each response function to depend only on observed parents, guaranteeing the linearization.

In general, given a causal discrete DAG model $\mathcal{D}$ satisfying Property 1, any parameter specification in terms of CPTs can equivalently be expressed as a *finite mixture model* [19] with $|R|$ deterministic components—the response functions—and mixing weights $w \in \Delta^{|R|-1}$. Conditional on ρ, the observational distribution is deterministic in X_{V_R}, and marginalizing over ρ yields the vector of observable probabilities π_{obs} as linear functions of w. The response function formulation induces a linear map

$$A : \Delta^{|R|-1} \to \mathbb{R}^m,$$

where m is the number of observable probabilities, such that $\pi_{\mathrm{obs}} = Aw$ for some w in the simplex. The feasible set for w is

$$\mathcal{F} \coloneqq \big\{ w \in \Delta^{|R|-1} : Aw = \pi_{\mathrm{obs}} \big\},$$

a polytope (an affine section of the simplex).

Causal estimands such as the ATE can be written as linear functionals of the weights: there exists a vector $c \in \mathbb{R}^{|R|}$ such that $\mathrm{ATE} = c^\top w$ for all w in the simplex consistent with the model. Computing sharp bounds thus reduces to solving the pair of linear programs

$$\underline{\mathrm{ATE}} = \min_{w \in \mathcal{F}} c^\top w, \qquad \overline{\mathrm{ATE}} = \max_{w \in \mathcal{F}} c^\top w.$$

Geometrically, $\mathcal{F} = \Delta^{|R|-1} \cap A^{-1}(\{\pi_{\mathrm{obs}}\})$ is the affine fiber of A over π_{obs} inside the simplex. The image $P = A(\Delta^{|R|-1}) \subset \mathbb{R}^m$ is the polytope of observable probability vectors compatible with the model—a *marginal polytope* [19] in the sense that it is the convex hull of the $|R|$ deterministic distributions indexed by response functions, analogous to those arising in exponential family models. In practice, the standard approach constructs an H-representation of the LP feasibility region from the constraints on w (nonnegativity, normalization, and $Aw = \pi_{\mathrm{obs}}$) and enumerates the vertices of the associated dual polytope via double description [4]; each dual vertex yields a linear expression in π_{obs}, and the sharp bounds are obtained by taking the maximum (respectively minimum) over these expressions.

3.1 The hyperarc encoding

The cardinality $|R| = \prod_{j\in V_R} |\mathcal{X}_j|^{|\mathcal{X}_{\mathrm{pa}(j)\cap V}|}$ grows exponentially in the sizes of observed parent sets; in the running example, already $|R| = 16$. Shridharan and Iyengar [17] introduce a *hyperarc* encoding that, assuming V_L is nonempty, reduces the number of LP variables from $|R|$ (the *local* encoding) to $|H| \le |R|$, while still producing the same valid and sharp bounds.

A *valid hyperarc* [17, Definition 4] is a map $h : \mathcal{X}_{V_L} \to \mathcal{X}_{V_R}$ satisfying, for every $j \in V_R$ and every pair $v, v' \in \mathcal{X}_{V_L}$, with $\oplus$ representing concatenation,

$$(v \oplus h(v))_{\mathrm{pa}(j)\cap V} = (v' \oplus h(v'))_{\mathrm{pa}(j)\cap V} \implies (v \oplus h(v))_j = (v' \oplus h(v'))_j. \tag{2}$$

That is, the output of each node in V_R is determined by its observed parents, mirroring the local encoding that defines each $\mathcal{R}_j$. Let H denote the set of all valid hyperarcs.

Given $\rho = (r_j)_{j\in V_R} \in R$, define $h_\rho \colon \mathcal{X}_{V_L} \to \mathcal{X}_{V_R}$ by processing $j \in V_R$ in a topological order and setting

$$(h_\rho(v))_j \coloneqq r_j\big((v \oplus h_\rho(v))_{\mathrm{pa}(j)\cap V}\big). \tag{3}$$

The right-hand side is well-defined because every coordinate it reads has already been assigned: parents of j in V_L come from v, while parents of j in V_R precede j in topological order. By construction, any two v, v' agreeing on $(v \oplus h_\rho(v))_{\mathrm{pa}(j)\cap V}$ yield the same r_j output, so h_ρ is a valid hyperarc. Conversely, the assignment $\rho \mapsto h_\rho$ is surjective: given $h \in H$, for each $j \in V_R$ and each parent configuration $y \in \mathcal{X}_{\mathrm{pa}(j)\cap V}$ realized as $(v \oplus h(v))_{\mathrm{pa}(j)\cap V}$ for some $v \in \mathcal{X}_{V_L}$, set $r_j(y) \coloneqq (v \oplus h(v))_j$; this is well-defined by Equation (2), and r_j can be extended arbitrarily to unrealized configurations, giving $h_{(r_j)_j} = h$. Hence, using $R_h \coloneqq \{\rho \in R : h_\rho = h\}$, the sets $\{R_h\}_{h\in H}$ partition R into nonempty parts [17, Lemma 14], so $|H| \le |R|$. Setting $q_h = \sum_{\rho\in R_h} w_\rho$ and substituting into the linear program from Section 3 gives an equivalent LP in $|H|$ variables with the same feasible set and the same sharp bounds. Denote the constraint matrix of this LP by A_H (one column per valid hyperarc).

We point out that, instead of iterating over all $|\mathcal{X}_{V_R}|^{|\mathcal{X}_{V_L}|}$ candidate maps—which can be larger than $|R|$ and is doubly exponential in $|V_L|$—and selecting valid ones, the set of valid hyperarcs can be enumerated directly by processing V_R in topological order. At each node j, partition $\mathcal{X}_{V_L} \times \mathcal{X}_{i\,:\,i\preceq j\in V_R}$ into groups G_j of configurations that realize the same values of $\mathrm{pa}(j) \cap V$ under the assignments already made, then independently assign X_j a constant within each group; the number of valid choices at j is $|\mathcal{X}_j|^{|G_j|}$.

Fibers R_h of size greater than one arise precisely when the hyperarc h leaves some parent configuration of a node in V_R unrealized: two local response functions that agree on every realized configuration but differ on an unrealized one produce the same observable distribution and hence the same hyperarc. The following proposition characterizes exactly when this collapse occurs.

Proposition 1. $|H| = |R|$ *if and only if* $\mathrm{pa}(j) \cap V_R = \emptyset$ *for every* $j \in V_R$.

Proof. If $\mathrm{pa}(j) \cap V_R = \emptyset$ for all j, then the projection $\mathcal{X}_{V_L} \to \mathcal{X}_{\mathrm{pa}(j)\cap V}$ is surjective for every j (all observed parents lie in V_L), so every parent configuration of j is realized and $|G_j| = |\mathcal{X}_{\mathrm{pa}(j)\cap V}|$ for every j and every partial hyperarc. The choices per node are independent, giving $|H| = \prod_{j\in V_R} |\mathcal{X}_j|^{|G_j|} = \prod_{j\in V_R} |\mathcal{X}_j|^{|\mathcal{X}_{\mathrm{pa}(j)\cap V}|} = |R|$.

For the other direction of the proof, suppose $k \in \mathrm{pa}(j_0) \cap V_R$ for some $j_0 \in V_R$; since k is a parent of j_0, k precedes j_0 in any topological order. Fix $\rho^* \in R$ whose r_k component is the constant function 0. By Equation (3), $(h_{\rho^*}(v))_k = 0$ for every $v \in \mathcal{X}_{V_L}$, so every realized parent configuration of j_0 has 0 at the k-th coordinate. Now let r'_{j_0} agree with the r_{j_0} component of ρ^* on all configurations y with $y_k = 0$, but differ on some $y^\dagger$ with $y^\dagger_k \neq 0$ (possible since $|\mathcal{X}_{j_0}| \ge 2$). Setting $\rho' = \rho^*$ except with r'_{j_0} in place of r_{j_0}, an induction along the topological order in Equation (3) shows that $h_{\rho'}$ and h_{ρ^*} agree at every node: ρ' and ρ^* share the same local functions at every $j \neq j_0$, and at j_0 the parent configurations $(v \oplus h_{\rho^*}(v))_{\mathrm{pa}(j_0)\cap V}$ all have k-coordinate 0, on which r'_{j_0} and r_{j_0} agree by construction. Hence $h_{\rho'} = h_{\rho^*}$ even though $\rho' \neq \rho^*$. Hence $|R_{h_{\rho^*}}| \ge 2$; combined with surjectivity of $\rho \mapsto h_\rho$, this gives $|H| < |R|$. $\square$

Example 5. We continue with the confounded model of Example 3 (Figure 2), with $V_L = \{1\}$, $V_R = \{2, 3\}$, latent confounder U_2, all observed variables binary, and $|R| = 16$ as computed in Example 4.

We enumerate H directly. Node 2 has observed parent $X_1 \in V_L$. The two values of X_1 are always distinct, so the two configurations of V_L form separate groups, $|G_2| = 2$, and there are $2^2 = 4$ valid choices for X_2—the same count as the local encoding.

Node 3 has observed parent $X_2 \in V_R$. Given the choice for node 2, the groups for node 3 are determined by which values of X_2 are realized across the two configurations of V_L. Two of the four assignments for node 2 map both X_1 values to the same X_2 value (the constant assignments $X_2 \equiv 0$ and $X_2 \equiv 1$); these give $|G_3| = 1$ and $2^1 = 2$ valid choices for X_3. The other two assignments (the identity and flip maps) realize both X_2 values, giving $|G_3| = 2$ and $2^2 = 4$ choices for X_3.

Thus $|H| = 2 \cdot 2 + 2 \cdot 4 = 12 < 16 = |R|$. By the canonical partitioning of [15], any distribution over U_2—regardless of cardinality—induces a distribution over the $|R| = 16$ response types; the LP over $w \in \Delta^{|R|-1}$ is therefore without loss of generality. The four pruned response functions from the local encoding are those in which r_2 is constant but r_3 disagrees on the unrealized X_2 value: since that value is never realized, the two choices of r_3 that agree on the realized input but differ on the unrealized one collapse into a single hyperarc. Companion code enumerating H directly and verifying that both encodings yield identical sharp bounds is provided in the `causaloptim` package vignette[6]. $\diamond$

4 Open questions

We have outlined how discrete graphical causal models give rise to polynomial parameterizations, how nonidentifiable causal effects naturally lead to bound computation problems, how response functions convert these problems into linear optimization over a polytope, and how the hyperarc encoding reduces the LP dimension relative to the local encoding from $|R|$ to $|H| \leq |R|$, with equality if and only if V_R contains no internal edges. This analysis leads to several questions, which we pose alongside connections to tools from algebraic statistics.

Question 1. Can the hyperarc aggregation be improved further? Is there an encoding with $|E| \leq |H|$ variables for all graphs, with $|E| < |H|$ for some? And is there a generalization that prunes R even when V_L is empty?

The constraint matrix A_H has one column per valid hyperarc, and distinct valid hyperarcs produce distinct columns (since the observation $(v \oplus h(v))$ includes v). Thus, further reduction cannot come from column merging alone and requires a fundamentally different aggregation. Lee [9] proposes a complementary, query-specific approach: by merging response functions $\rho \in R$ that are indistinguishable to every linear functional in both the LP objective and the observable constraints, the reduced domain can have $|E| < |H|$ variables for some queries.

Question 2. Property 1 gives sufficient, but not necessary, conditions for a model to be response-linear. Can we characterize broader classes of graphs for which a response function formulation still yields linear observable constraints, and are there computationally tractable approaches to bounding causal estimands in models whose DAG implies nonlinear constraints?

For models that are *not* response-linear, the observable constraints are polynomial rather than linear, and the image of the parameterization is a semialgebraic set rather than a polytope. In this setting, tools from computational algebra—Gröbner bases, elimination ideals, cylindrical algebraic decomposition—become necessary rather than optional. There are complementary algebraic approaches to related problems: quantifier elimination in the first-order theory of the reals can derive all equality and inequality constraints implied by graphical models with hidden variables [5], and Gröbner basis methods can characterize feasibility and identifiability for causal structures with two binary observed variables [8]. However, these methods quickly become intractable for larger graphs.

[6] https://sachsmc.github.io/causaloptim/articles/hyperarc-example.html

Question 3. Can Gröbner basis methods on the polynomial parameterizations arising from discrete DAG models yield bounds or identifiability results for models where linear programming does not apply?

Recent work by Boushehrian et al. [2] shows that some of the complexity we encounter is unavoidable: in discrete instrumental variable settings, any sharp analytical bound for the ATE must be expressible as a maximum or minimum over exponentially many linear terms. Hence, heuristics—also discussed by Shridharan and Iyengar [17]—though they cannot guarantee sharp bounds, could nonetheless be necessary for complex causal models:

Question 4. Can symbolic, polyhedral, or sampling-based methods reduce the cost of searching the feasible set when exact vertex enumeration is too expensive?

The constraint matrix A_H determines a *toric ideal* I_{A_H}, generated by binomials corresponding to integer vectors in $\ker(A_H)$ [19]. By Diaconis and Sturmfels [3], these generators form a *Markov basis*: a set of moves connecting all integer points in any fiber $\{q \geq 0 : A_H q = \hat{p},\, q \in \mathbb{Z}^{|H|}\}$, enabling MCMC-based conditional tests in the integer setting.

Question 5. For the probability simplex relevant to causal bounds, what is the appropriate analog of the Diaconis and Sturmfels conditional test, and does the Markov basis structure of I_{A_H} offer computational advantages over generic polytope samplers?

Acknowledgments. We thank Pratik Misra and Cecilie Olesen Recke for helpful discussions and feedback on an earlier draft. All authors were supported by the Pioneer Centre for Statistical and computational Methods for Advanced Research to Transform Biomedicine (SMARTbiomed), DNRF grant number P4. A.M. was also supported by Novo Nordisk Foundation grant NNF20OC0062897.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Balke, A., Pearl, J.: Counterfactual probabilities: Computational methods, bounds and applications. In: de Mantaras, R.L., Poole, D. (eds.) Uncertainty in Artificial Intelligence, pp. 46–54. Morgan Kaufmann, San Francisco (CA) (1994). https://doi.org/10.1016/B978-1-55860-332-5.50011-0
2. Boushehrian, A., Badri, M.R., Akbari, S., Kiyavash, N.: Fundamental Limits and Optimal Methods for Sharp Analytical Causal Bounds in Instrumental Variable Models (2026). https://doi.org/10.48550/arXiv.2604.12802
3. Diaconis, P., Sturmfels, B.: Algebraic algorithms for sampling from conditional distributions. The Annals of Statistics **26**(1), 363–397 (1998). https://doi.org/10.1214/aos/1030563990
4. Fukuda, K., Prodon, A.: Double description method revisited. In: Combinatorics and Computer Science, Lecture Notes in Computer Science, vol. 1120, pp. 91–111. Springer (1996). https://doi.org/10.1007/3-540-61576-8_77
5. Geiger, D., Meek, C.: Quantifier elimination for statistical problems. In: Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI). pp. 226–235 (1999). https://doi.org/10.48550/arXiv.1301.6698
6. Koller, D., Friedman, N.: Tabular CPDs. In: Probabilistic Graphical Models: Principles and Techniques, chap. 5.1, pp. 157–158. MIT Press (2009)
7. Lauritzen, S.L.: Graphical Models. Oxford University Press (1996). https://doi.org/10.1093/oso/9780198522195.001.0001
8. Lee, C.M., Spekkens, R.W.: Causal inference via algebraic geometry: Feasibility tests for functional causal structures with two binary observed variables. Journal of Causal Inference **5**(2) (2017). https://doi.org/10.1515/jci-2016-0013
9. Lee, S.: Canonical domain reduction for computationally efficient partial causal identification (2026), https://www.sanghacklee.me/assets/canonical-domain-reduction-pcid.pdf
10. McCullagh, P.: What is a statistical model? The Annals of Statistics **30**(5) (2002). https://doi.org/10.1214/aos/1035844977

11. Pearl, J.: Causal inference in statistics: An overview. Statistics Surveys **3**, 96–146 (2009). https://doi.org/10.1214/09-SS057
12. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press (2009). https://doi.org/10.1017/cbo9780511803161
13. Pfister, N.: Lecture notes – causality (2024), https://niklaspfister.github.io/download/lecture_notes_causality.pdf
14. Rubin, D.B.: Causal inference using potential outcomes: Design, modeling, decisions. Journal of the American Statistical Association **100**(469), 322–331 (2005). https://doi.org/10.1198/016214504000001880
15. Sachs, M.C., Jonzon, G., Sjölander, A., Gabriel, E.E.: A general method for deriving tight symbolic bounds on causal effects. Journal of Computational and Graphical Statistics **32**(2), 567–576 (2023). https://doi.org/10.1080/10618600.2022.2071905
16. Shpitser, I., Pearl, J.: Complete identification methods for the causal hierarchy. Journal of Machine Learning Research (JMLR) **9**(64), 1941–1979 (2008), http://jmlr.org/papers/v9/shpitser08a.html
17. Shridharan, M., Iyengar, G.: Scalable computation of causal bounds. Journal of Machine Learning Research **24**(237), 1–35 (2023), http://jmlr.org/papers/v24/22-1081.html
18. Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search. The MIT Press (2001). https://doi.org/10.7551/mitpress/1754.001.0001
19. Sullivant, S.: Algebraic statistics, vol. 194. American Mathematical Society (2023)
20. Yang, K., Katcoff, A., Uhler, C.: Characterizing and learning equivalence classes of causal DAGs under interventions. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 5541–5550. PMLR (10–15 Jul 2018), https://proceedings.mlr.press/v80/yang18a.html