

An Automated Approach to Causal Inference in Discrete Settings*

Guilherme Duarte
gjduarte@upenn.edu

Noam Finkelstein
noam@jhu.edu

Dean Knox
dcknox@upenn.edu

Jonathan Mummolo
jummolo@princeton.edu

Ilya Shpitser
ilyas@cs.jhu.edu

First draft: February 10, 2021
This draft: December 23, 2022

Abstract

Applied research conditions often make it impossible to point-identify causal estimands without untenable assumptions. *Partial identification*—bounds on the range of possible solutions—is a principled alternative, but the difficulty of deriving bounds in idiosyncratic settings has restricted their application. We present a general, automated numerical approach to causal inference in discrete settings. We show causal questions with discrete data reduce to polynomial programming problems, then present an algorithm to automatically bound causal effects using efficient dual relaxation and spatial branch-and-bound techniques. The user declares an estimand, states assumptions, and provides data—however incomplete or mismeasured. The algorithm then searches over admissible data-generating processes and outputs the most precise possible range consistent with available information—i.e., *sharp* bounds—including a point-identified solution if one exists. Because this search can be computationally intensive, our procedure reports and continually refines non-sharp ranges guaranteed to contain the truth at all times, even when the algorithm is not run to completion. Moreover, it offers an ε -sharpness guarantee, characterizing the worst-case looseness of the incomplete bounds. These techniques are implemented in our Python package, **autobounds**. Analytically validated simulations show the method accommodates classic obstacles including confounding, selection, measurement error, noncompliance, and nonresponse.

Keywords: causal inference, partial identification, constrained optimization, linear programming, polynomial programming

*Guilherme Jardim Duarte is a Ph.D. student in the Operations, Information and Decisions Department, the Wharton School of the University of Pennsylvania. Noam Finkelstein is a Ph.D. student in the Department of Computer Science, Johns Hopkins University. Dean Knox is an Andrew Carnegie Fellow and an assistant professor in the Operations, Information and Decisions Department, the Wharton School of the University of Pennsylvania. Jonathan Mummolo is an assistant professor of Politics and Public Affairs, Princeton University. Ilya Shpitser is the John C. Malone Assistant Professor in the Department of Computer Science, Whiting School of Engineering at the Johns Hopkins University. Authors listed in alphabetical order. For helpful feedback, we thank Peter Aronow, Justin Grimmer, Kosuke Imai, Luke Keele, Gary King, Christopher Lucas, Fredrik Sävje, Brandon Stewart, Eric Tchetgen Tchetgen, and participants in the Harvard Applied Statistics Workshop, the New York University Data Science Seminar, University of Pennsylvania Causal Inference Seminar, PolMeth 2021, and the Yale Quantitative Research Methods Workshop. We gratefully acknowledge financial support from AI for Business and the Analytics at Wharton Data Science and Business Analytics Fund. This research was made possible in part by a grant from the Carnegie Corporation of New York. The statements made and views expressed are solely the responsibility of the authors.

Contents

1	Introduction	1
2	Related literature	4
3	Preliminaries	5
3.1	Canonical DAGs	6
3.2	Potential outcomes	7
3.3	Generalized principal stratification	7
4	Formulating the polynomial program	11
4.1	Example program for outcome-based selection	17
4.2	Simplifying the polynomial program	19
5	Computing ε-sharp bounds in polynomial programs	20
6	Statistical inference	25
7	Simulated examples	27
7.1	Instrumental variables	27
7.2	Coverage of confidence bounds	28
7.3	More complex bounding problems	29
8	Potential critiques of the approach	32
9	Conclusion	33
A	Glossary of terms	37
A.1	Bounding terms	37
A.2	Computational terms	38
A.3	Causal terms	38
B	Algorithms	42
B.1	Algorithm 1: Canonicalization of DAGs, with discussion	42
B.2	Algorithm 2: Constructing polynomial programs	43
B.3	Algorithm 3: Computing ε -sharp Bounds	44
C	Examples and detailed discussion	45
C.1	Step-by-step illustration of polynomial program construction	45
C.1.1	Step 1: Canonicalization	45
C.1.2	Step 2: Principal stratification	45
C.1.3	Step 3: Estimand polynomialization	46
C.1.4	Step 4: Constraint polynomialization	48
C.1.5	Example code	51
C.2	Example of deterministic relationships	51
C.3	Example of program simplification	52
C.4	Discussion of Nested Markov models	56

D	Details on program simplifications	61
D.1	Reducing polynomial degree by exploiting graph structure	63
D.2	Eliminating variables by solving equality constraints	64
D.3	Refactorizing empirical constraints to eliminate redundant information	64
D.4	Eliminating additional constraints and parameters	70
E	Details on statistical inference	70
F	Proofs	74
F.1	Proof of Proposition 1	74
F.2	Proof of Proposition 2	76
F.3	Proof of Proposition 3, with discussion	76
F.4	Proof of Proposition 4	77
G	Details of simulated models	77
G.1	Noncompliance simulation	79
G.2	Outcome-based selection simulation	79
G.3	Measurement error simulation	80
G.4	Outcome missingness simulation	81
G.5	Joint missingness simulation	82
G.6	Scaling of computation time with variable cardinality	84

1 Introduction

When causal quantities cannot be point identified, researchers often pursue partial identification to quantify the range of possible answers. These solutions are tailored to specific settings (e.g. [Lee, 2009](#); [Sjölander et al., 2014](#); [Kennedy et al., 2019](#); [Knox et al., 2020](#); [Gabriel et al., 2022](#)), but the idiosyncrasies of applied research can render prior results unusable if even slightly differing scenarios are encountered. This piecemeal approach to deriving causal bounds presents a major obstacle to scientific progress. To increase the pace of discovery, researchers need a more general solution.

In this paper, we present an automated approach to causal inference in discrete settings which applies to all causal graphs, as well as all standard observed quantities and domain assumptions. Users declare an estimand, state assumptions, and provide available data—however incomplete or mismeasured. The algorithm numerically computes *sharp bounds*, the most precise possible answer to the causal query given these inputs, including a unique point estimate if one exists. Our approach accommodates any classic threat to inference, including missing data, selection, measurement error, and noncompliance. It can fuse information from numerous sources—including observational and experimental data, datasets that are unlinkable due to anonymization, or even summary statistics from other studies. The method allows for sensitivity analyses on any assumption by relaxing or removing it entirely. Moreover, it alerts users when assumptions conflict with observed data, indicating faulty causal theory. We also develop techniques for drawing statistical inferences about estimated bounds. We implement these methods in a Python package, `autobounds`, and demonstrate them with a host of analytically validated simulations.

Our work advances a rich literature on partial identification in causal inference ([Manski,](#)

1990; Zhang and Rubin, 2003; Cai et al., 2008; Swanson et al., 2018; Gabriel et al., 2022; Molinari, 2020), outlined in Section 2, which has sometimes cast partial identification as a constrained optimization problem. In pioneering work, Balke and Pearl (1997) provided an automatic sharp bounding method for causal queries that can be expressed as linear programming problems. However, numerous estimands and empirical obstacles do not fit this description, and a complete and feasible computational solution has remained elusive.

When feasible, sharp bounding represents a principled and transparent method that makes maximum use of available data while acknowledging its limitations. Claims outside the bounds can be immediately rejected, and claims inside the bounds must be explicitly justified by additional assumptions or new data. But several obstacles still preclude widespread use. For one, analytic bounds—which can be derived once and then applied repeatedly, unlike our numerical bounds which must be recomputed each time—remain intractable for many problems. Within the subclass of linear problems, Balke and Pearl’s (1997) simplex method offers an efficient analytic approach, but analytic nonlinear solutions are still derived case by case (e.g. Kennedy et al., 2019; Knox et al., 2020; Gabriel et al., 2022). Moreover, though general sharp bounds can in theory be obtained by standard nonlinear optimization techniques (Geiger and Meek, 1999; Zhang and Bareinboim, 2021), in practice, such approaches are often computationally infeasible. This is because without exhaustively exploring a vast model space to avoid local optima, they can inadvertently report *invalid* bounds that may fail to contain the truth.

To address these limitations, we first show in Sections 3–4 that—using a generalization of principal strata (Frangakis and Rubin, 2002)—causal estimands, modeling assumptions, and observed information can be rewritten as polynomial objective functions and polynomial constraints with no loss of information. We extend results from Geiger and Meek (1999) and Wolfe et al. (2019) to show that essentially all discrete partial identification problems reduce to polynomial programs, a well-studied class of optimization tasks that nest linear programming as a special case.¹ However, it is well known that solving polynomial programs to global

¹Specifically, our results apply to elementary arithmetic functionals or monotonic transformations thereof—a broad set that essentially includes all causal assumptions, observed quantities, and estimands in standard use.

optimality is in general NP-hard, highlighting the need for efficient bounding techniques that remain valid even under time constraints (Belotti et al., 2009; Vigerske and Gleixner, 2018).

To ameliorate these computational difficulties, Section 4.2 shows how causal graphs can be restated as equivalent canonical models, further simplifying the polynomial program. Next, Section 5 develops an efficient optimization procedure, based on dual relaxation and spatial branch-and-bound relaxation techniques, that provides bounds of arbitrary sharpness. We show this procedure is guaranteed to achieve complete sharpness with sufficient computation time; in the problems we examine here, this occurs in a matter of seconds. However, in cases where the time needed is prohibitive, our algorithm is *anytime* (Dean and Boddy, 1988), meaning it can be interrupted to obtain non-sharp bounds that are nonetheless guaranteed to be valid. Crucially, our technique offers an additional guarantee we term “ ε -sharpness,” a *worst-case looseness factor* that quantifies how much the current non-sharp bounds could potentially be improved with additional computation. In Section 6, we provide two approaches for characterizing uncertainty in the estimated bounds. We demonstrate our technique in a series of analytically validated simulations in Section 7, showing the flexibility of our approach and the ease with which assumptions can be modularly imposed or relaxed. Moreover, we demonstrate how it can improve over widely used bounds (Manski, 1990) and recover a recent point-identification result in the literature on nonrandom missingness (Miao et al., 2016).

In short, our approach offers a complete and computationally feasible approach to causal inference in discrete settings. Given a well-defined causal query, valid assumptions, and data, researchers now have a general and automated process to draw causal inferences that are guaranteed to be valid and, with sufficient computation time, provably optimal.

For example, the average treatment effect and the log odds ratio can be sharply bounded with our approach, but non-analytic functionals (which are rarely if ever encountered) cannot. Functionals that do not meet these conditions can be approximated to arbitrary precision, if they have convergent power series.

2 Related literature

Researchers have long sought to automate partial identification by recasting causal bounding problems as constrained optimization problems that can be solved computationally. Our work is most closely related to [Balke and Pearl \(1997\)](#), which showed that certain bounding problems in discrete settings—generally, when interventions and outcomes are fully observed—could be formulated as the minimization and maximization of a linear objective function subject to linear equality and inequality constraints. Such programming problems admit both symbolic solutions and highly efficient numerical solutions. Subsequent studies have proven that the bounds produced by this technique are sharp ([Bonet, 2001](#); [Ramsahai, 2012](#); [Sachs et al., 2020](#)). These results were extended by [Geiger and Meek \(1999\)](#), who showed that a much broader class of discrete problems can be formulated in terms of polynomial relations when analysts have precise information about the kinds of disturbances or confounders that may exist.² In addition to the well-known conditional independence constraints implied by d-separation, these can include generalized equality constraints ([Verma and Pearl, 1990](#); [Tian and Pearl, 2002](#)) and generalizations of the instrumental inequality constraints ([Pearl, 1995](#); [Bonet, 2001](#)).

[Geiger and Meek \(1999\)](#) note that in theory, quantifier elimination algorithms can provide symbolic bounds. However, the time required for quantifier elimination grows as a doubly exponential function of the number of parameters, rendering it infeasible for all but the simplest cases. At the core of this issue is that symbolic methods provide a general solution, meaning that they must explore the space of all possible inputs. In contrast, numerical methods such as ours can accelerate computation by eliminating irrelevant portions of the model space.

Even so, computation can be time-consuming.³ In practice, many optimizers can rapidly find reasonably good values but cannot guarantee optimality without exhaustively searching

²A subtle point in nonlinear settings is that the region of possible values for the estimand—i.e., estimand values associated with models in the model subspace that is consistent with available data and assumptions—may be disconnected. That is, while the sharp lower and upper bounds correspond to minimum and maximum possible values of the estimand, not all estimand values between these extremes are necessarily possible.

³Sharp bounds can always be obtained by exhaustively searching the model space. But the computation time required to do so—i.e., to solve the polynomial programming problem—can explode with the number of variables (principal strata sizes).

the model space. This approach poses a challenge for obtaining causal bounds, which are global minimum and maximum values of the estimand across all models that are *admissible*, or consistent with observed data and modeling assumptions. If a local optimizer operates on the original problem (the *primal*), proceeding from the interior and widening bounds as more extreme models are discovered, then failing to reach global optimality will result in *invalid bounds*—ranges narrower than the true sharp bounds, failing to contain all possible solutions.

In the following sections, we detail our approach to addressing each of these outstanding obstacles to automating the discovery of sharp bounds for discrete causal problems.

3 Preliminaries

We now define notation and introduce key concepts. A technical glossary is given in Appendix A. We first review how any causal model represented by a directed acyclic graph (DAG) can be “canonicalized,” or reduced into simpler form, without loss of generality (w.l.o.g.; Evans, 2016). We describe how these graphs give rise to potential outcomes and a generalization of principal strata (Frangakis and Rubin, 2002), two key building blocks in our analytic strategy.

We follow the convention that bold letters denote collections of variables; uppercase and lowercase letters denote random variables and their realizations, respectively. Consider a structured system in which random vectors $\mathbf{V} = \{V_1, \dots, V_J\}$ represent observable *main variables* and $\mathbf{U} = \{U_1, \dots, U_K\}$ represent *unobserved disturbances*. We will assume each observed variable V_j is discrete and its space $\mathcal{S}(V_j)$ has finite cardinality; the spaces of unobserved variables are unrestricted. Observed data for each unit $i \in \{1, \dots, N\}$ is an i.i.d. draw from \mathbf{V} . Further suppose that causal relationships between all variables in \mathbf{V} and \mathbf{U} are represented by a nonparametric structural equation model with independent errors (NPSEM-IE, Pearl, 2000).⁴ Here, we concentrate on deriving results for the NPSEM-IE model, but our approach is also

⁴The NPSEM-IE model states that each $V_j \in \mathbf{V}$ and each $U_k \in \mathbf{U}$ is a deterministic function of (i) variables in $\mathbf{V} \cup \mathbf{U}$ corresponding to its parents in \mathcal{G} and (ii) an additional disturbance term, ϵ_{V_j} or ϵ_{U_k} . The crucial assumption in the NPSEM-IE is that these ϵ terms are mutually independent. Note that throughout this paper, we keep the presence of ϵ variables kept implicit; we will prove that each V_j can equivalently viewed as a deterministic function of its parents in $\mathbf{V} \cup \mathbf{U}$, absorbing the variation induced by ϵ terms into \mathbf{U} .

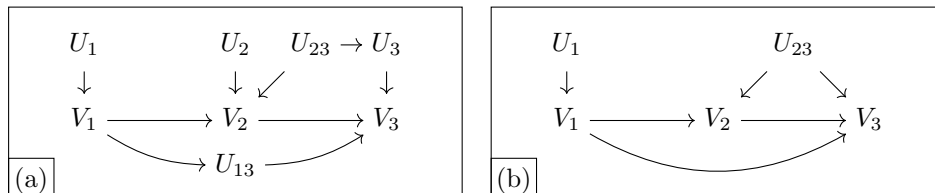
applicable to the model of [Robins \(1986\)](#) and [Richardson and Robins \(2013\)](#) without change.⁵

Figure 1 presents a DAG \mathcal{G} representing relationships between $\mathbf{V} \cup \mathbf{U}$. Note that fully observing these variables would be sufficient to identify every quantity we consider in this paper. However, since disturbances \mathbf{U} are unobserved, and since information about main variables \mathbf{V} may be incomplete, partial identification techniques are needed.

3.1 Canonical DAGs

We now discuss how canonicalizing DAGs—reformulating them w.l.o.g. into a simpler form—simplifies the bounding task. A DAG is said to be in canonical form if (i) no disturbance U_k has a parent in \mathcal{G} ; and (ii) there exists no pair of disturbances, U_k and $U_{k'}$, such that U_k influences a subset of variables influenced by $U_{k'}$. [Evans \(2016\)](#) showed that any non-canonical DAG \mathcal{G}' has a canonical form \mathcal{G} with an identical distribution governing all variables in \mathbf{V} ; an algorithm for obtaining this canonical form is given in Appendix B.1. In short, canonicalization distills the data-generating process (DGP) to its simplest form by eliminating potentially complex networks of irrelevant disturbances. Figure 1 shows a non-canonical DAG in panel (a); panel (b) gives the canonicalized version. Note that disturbances affecting only a single variable, such as U_1 , are often left implicit; here, we depict them explicitly for clarity.

Figure 1: **Canonicalization of a mediation graph.** Non-canonical and canonicalized forms are given in panels (a) and (b), respectively. Both are fully equivalent with respect to their full data law. Canonicalization proceeds as follows: (i) the dependent disturbance U_3 is absorbed into its parent U_{23} ; (ii) the superfluous U_2 is eliminated as it influences a subset of U_{23} ’s children; and (iii) the irrelevant U_{13} is absorbed into the $V_1 \rightarrow V_3$ arrow as it is neither observed nor of interest. A complete guide to canonicalization is given in Appendix B.1.



⁵See Appendix F.3 for further discussion of the the finest fully randomized causally interpretable structured tree graph (FFRCISTG).

3.2 Potential outcomes

The notation of potential outcome functions allows us to compactly express the effects of manipulating variable V_j 's main parents, $\mathbf{pa}_V(V_j)$, or other ancestors that are also main variables. Similarly, $\mathbf{pa}_U(V_j)$ denotes parents of V_j that are disturbances. Let $\mathbf{A} \subset \mathbf{V}$ be intervention variables that will be fixed to $\mathbf{A} = \mathbf{a}$. When $\mathbf{A} = \emptyset$, so no intervention occurs, then define $V_j(\mathbf{a}) = V_j$, the natural value. When $\mathbf{A} \subseteq \mathbf{pa}_V(V_j)$, so only immediate parents are manipulated, then the potential outcome function is given by its structural equation, $V_j(\mathbf{a}) = f_j[\mathbf{A} = \mathbf{a}, \mathbf{pa}_V(V_j) \setminus \mathbf{A}, \mathbf{pa}_U(V_j)]$. For example, in Figure 1(b), the effect of intervention $V_2 = v_2$ on outcome V_3 is defined in terms of $V_3(V_2 = v_2) = f_3(V_2 = v_2, V_1, U_{23})$. Here, the intervention set is $\mathbf{A} = V_2$, and the remaining parents of V_3 —the non-intervened main parent, $\mathbf{pa}_V(V_3) \setminus \mathbf{A} = V_1$, and the disturbance parent, $\mathbf{pa}_U(V_3) = U_{23}$ —are allowed to follow their natural distributions. We now define more general potential outcome functions by *recursive substitution* (Richardson and Robins, 2013; Shpitser, 2018). For arbitrary interventions on $\mathbf{A} \subset \mathbf{V}$, let $V_j(\mathbf{a}) = V_j(\{a_\ell : \mathbf{A}_\ell \in \mathbf{pa}_V(V_j)\} \cup \{V_{j'}(\mathbf{a}) : V_{j'} \in \mathbf{pa}_V(V_j) \setminus \mathbf{A}\})$; here, ℓ is a generic index that sweeps over main variables in the graph. That is, if a parent of V_j is in \mathbf{A} , it is set to the corresponding value in \mathbf{a} . Otherwise, the parent takes its potential value after intervention on causally prior variables, or its natural value otherwise. To obtain the parent's potential value, apply the same definition recursively.⁶ For example, in Figure 1(b), potential outcomes for V_3 include (i) $V_3(\emptyset) = V_3(V_1, V_2)$, the observed distribution; (ii) $V_3(v_1) = V_3[v_1, V_2(v_1)]$, relating to total effects; and (iii) $V_3(v_1, v_2)$, relating to controlled effects.

3.3 Generalized principal stratification

In this section, we show how any DAG and any causal quantity can be represented w.l.o.g. using a generalization of *principal strata*. Roughly speaking, principal strata on a variable V_j are groups of units that would respond to counterfactual interventions in the same way (Greenland and Robins, 1986; Frangakis and Rubin, 2002). Formally, let $\mathbf{A} = \mathbf{pa}_V(V_j)$ be an

⁶When defining potential outcomes for V_j , intervention on V_j itself is ignored.

intervention set for which all main parents of V_j are jointly set to some \mathbf{a} , and consider unit i 's collection of potential outcomes $\{V_{i,j}(\mathbf{A} = \mathbf{a}) : \mathbf{a} \in \mathcal{S}(\mathbf{A})\}$. Each principal stratum of V_j then represents a subset of units in which this collection is identical.

The NPSEM of a graph is closely related to its principal stratification. This is because each potential outcome in the collection above is given by $V_{i,j}(\mathbf{A} = \mathbf{a}) = f_j[\mathbf{A} = \mathbf{a}, \mathbf{pa}_{i,U}(V_j)]$, in which the only source of random variation is unit i 's realization of the relevant disturbances. After fixing these disturbances, all structural equations become deterministic, meaning that a realization of \mathbf{U}_i must fix every potential outcome for every variable under every intervention. For example, consider the simple DAG $U_1 \rightarrow V_1 \rightarrow V_2 \leftarrow U_2$, in which V_1 and V_2 are binary. This relationship is governed by the structural equations $V_1 = f_1(U_1)$ and $V_2 = f_2(V_1, U_2)$, where the functions $f_1 : \mathcal{S}(U_1) \rightarrow \mathcal{S}(V_1)$ and $f_2 : \mathcal{S}(V_1) \times \mathcal{S}(U_2) \rightarrow \mathcal{S}(V_2)$ are deterministic and shared across all units. Thus, the only source of randomness is in $\mathbf{U} = \{U_1, U_2\}$.

Analysts generally do not have direct information about these disturbances. For example, U_1 could potentially take on any value in $(-\infty, \infty)$. However, as Proposition 1 will state in greater generality, this variation is irrelevant because V_1 has only two possible values: 0 and 1. The space of U_1 can therefore be divided into two *canonical partitions* (Balke and Pearl, 1997)—those that deterministically lead to $V_1 = 0$ and those that lead to $V_1 = 1$ —and thus treating U_1 as if it were binary is w.l.o.g.

Strata for V_2 are similar but more involved. After U_2 is realized, it induces the *partially applied* response function $V_2 = f_2(V_1, U_2 = u_2) = f_2^{(U_2=u_2)}(V_1)$, which deterministically governs how V_2 counterfactually responds to V_1 . Regardless of how many are in $\mathcal{S}(U_2)$, this response function must fall into one of only four possible strata, each a mapping of the form $f_2^{(U_2=u_2)} : \mathcal{S}(V_1) \rightarrow \mathcal{S}(V_2)$ (Angrist et al., 1996). These groups are (i) $V_2 = 1$ regardless of V_1 , “always takers”; (ii) $V_2 = 0$ regardless of V_1 , “never takers”; (iii) $V_2 = V_1$, “compliers”; and (iv) $V_2 = 1 - V_1$, “defiers”. Thus, from the perspective of V_2 , any finer-grained variation in $\mathcal{S}(U_2)$ beyond the canonical partitions is irrelevant. These partitions are in one-to-one correspondence with principal strata, which in turn allow causal quantities to be expressed in simple algebraic

expressions. For example, the average treatment effect (ATE) is equal to the proportion of compliers minus that of defiers.⁷ As Proposition 2 will show, by writing down all information in terms of these strata, essentially any causal inference problem can be converted into an equivalent optimization problem involving polynomials of variables that represent strata sizes.

Finally, consider the more complex mediation DAG of Figure 2(a). Response functions for V_1 and V_2 remain as above. In contrast, V_3 is caused by $\mathbf{pa}_{\mathbf{V}}(V_3) = \{V_1, V_2\}$ via the structural equation $V_3 = f_3(V_1, V_2, U_{23})$. Substituting in disturbance $U_{23} = u_{23}$ produces one of 16 response functions of the form $f_3^{(U_{23}=u_{23})} : \mathcal{S}(V_1) \times \mathcal{S}(V_2) \rightarrow \mathcal{S}(V_3)$, yielding 16 strata.⁸

In turn, the number of principal strata determines the minimum complexity of a reduced but non-restrictive alternative model in which the full data law, or joint distribution over every potential outcome, is preserved. This means the reduction is w.l.o.g. for every possible factual or counterfactual quantity involving \mathbf{V} . Specifically, the number of principal strata in the graph determines the minimum cardinalities of each $U_k \in \mathbf{U}$ that are needed to represent the original model w.l.o.g., if we were to redefine U_k in terms of a categorical distribution over principal strata. For example, to capture the joint response patterns that a unit may have on V_2 and V_3 , a reduced version of U_{23} can express any full data law if it has a cardinality of $|\mathcal{S}(U_{23})| = 4 \times 16$, because V_2 has four possible response functions and V_3 has 16.

Below, Proposition 1 states that a generalization of this approach can produce non-restrictive models w.l.o.g. for any discrete-variable DAG and any full data law. Crucially, this also holds for (1) graphs where a variable V_j is influenced by multiple disturbances U_k and $U_{k'}$, as in Figure 2(b); and (2) the challenging case of *non-gear*ed graphs (Evans, 2018) such as Figure 2(c)—roughly speaking, when disturbances U_k , $U_{k'}$, and $U_{k''}$ touch overlapping combinations of main variables to create cycles of confounding. Formalization is provided later.

⁷To see this, note that the ATE is given by $\mathbb{E}[V_2(V_1 = 1) - V_2(V_1 = 0)] = \sum_{\text{strata}} \mathbb{E}[V_2(V_1 = 1) - V_2(V_1 = 0) \mid \text{strata}] \cdot \Pr(\text{strata}) = 0 \cdot \Pr(\text{always taker}) + 0 \cdot \Pr(\text{never taker}) + 1 \cdot \Pr(\text{complier}) - 1 \cdot \Pr(\text{defier})$.

⁸More generally, the number of unique response functions grows with (i) the cardinality of the variable, (ii) the number of causal parents it has, and (iii) the parents' cardinalities. Specifically, V_j has $|\mathcal{S}(V_j)|^{|\mathcal{S}(\mathbf{pa}_{\mathbf{V}}(V_j))|}$ possible mappings: given a particular input from V_j 's parents, the number of possible outputs for V_j is $|\mathcal{S}(V_j)|$; the number of possible inputs from V_j 's parents is $|\mathcal{S}(\mathbf{pa}_{\mathbf{V}}(V_j))| = \prod_{V_{j'} \in \mathbf{pa}_{\mathbf{V}}(V_j)} |\mathcal{S}(V_{j'})|$, the product of the parents' cardinalities.

Proposition 1. *Suppose \mathcal{G} is a canonical DAG over discrete main variables \mathbf{V} and disturbances \mathbf{U} with infinite cardinality. The model over the full data law implied by \mathcal{G} is unchanged by assuming that the disturbances have sufficiently large finite cardinalities.*

A proof can be found in Appendix F.1, along with details on how to obtain a lower bound on non-restrictive cardinalities for the disturbances. Briefly, Proposition 1 extends a result from Finkelstein et al. (2021), which showed there are reductions of $\mathcal{S}(\mathbf{U})$ that do not restrict the model over the factual \mathbf{V} . We build on this result to show that there are reductions that do not restrict the full data law, or model over all factual and counterfactual versions of \mathbf{V} .

Though the theory of principal stratification is well understood when each main variable V_j is influenced by only one disturbance U_k , complications arise when V_j is influenced by multiple disturbances U_k and $U_{k'}$. For each such main variable, any one of the associated disturbances can be allocated to take primary responsibility—i.e. to be the input for which the response function is partially applied. For the purposes of defining this response function, all remaining disturbances are treated as if they were main variables.⁹ For example, in Figure 2(b), V_2 is influenced by both U_{12} and U_{23} ; we will allocate V_2 to U_{23} for illustration, but allocating it to U_{12} would produce identical bounds. Next, we compute the cardinality of remaining disturbances as usual. Here, U_{12} is left only to determine V_1 , meaning that it has a cardinality of two. Finally, we return to the primary disturbance and determine its cardinality based on main variables and remaining disturbances. In this example, after fixing U_{23} , the variable V_2 is a function of V_1 and U_{12} , both binary, meaning that U_{23} has a cardinality of sixteen.

Finally, Proposition 1 extends Evans (2018) by allowing us to develop generalized principal strata for graphs that are *non-geared*, meaning that disturbances do not satisfy the running intersection property.¹⁰ These cases differ only in that they contain *cycles* of confounding; after

⁹Note that if any main variable V has multiple parents in \mathbf{U} , there may be multiple valid parameterizations—i.e., methods for constructing generalized principal strata—depending on which disturbance is assigned primary responsibility for determining which main variable. If each main variable has only a single parent in \mathbf{U} , there is only a single functional parameterization.

¹⁰Here, the running intersection property requires that there exists a total ordering of disturbances such that the main variables touched by any U_k at most overlap with those touched by only one additional $U_{k'} < U_k$. For example, in Figure 2(c), if the ordering is $U_{13} < U_{12} < U_{23}$, then U_{23} touches both V_2 and V_3 . Thus, V_2 and V_3 together can be influenced by at most one additional disturbance that is earlier in the ordering. This is not

breaking the cycle at any point, they can be dealt with in the same manner as geared graphs. An example of a non-geared graph is given in Figure 2(c). Finkelstein et al. (2021) presents an algorithm for constructing a generalized principal stratification for non-geared graphs. In brief, the algorithm breaks the confounding cycle by selecting an arbitrary disturbance—e.g., U_{13} —and fixing its cardinality at a value that is guaranteed to be non-restrictive of the model over factual random variables, by Carathéodory’s theorem. In this case, based on U_{13} ’s district,¹¹ $\{V_1, V_2, \text{ and } V_3\}$, U_{13} can be analyzed w.l.o.g. as if it had a cardinality of $|\mathcal{S}(V_1) \times \mathcal{S}(V_2) \times \mathcal{S}(V_3)| - 2$. In all subsequent analysis of Figure 2(c), U_{13} would then be treated as a main variable, allowing the graph to be analyzed as if it were geared. As in Figure 2(b), U_{12} then determines the response of V_1 to U_{13} . Finally, U_{23} jointly determines (i) the responses of V_2 to V_1 and U_{12} as well as (ii) V_3 to $V_1, V_2, \text{ and } U_{13}$. We note that the number of parameters involved in non-geared graphs can quickly become intractable. In these cases, valid but possibly non-sharp bounds can always be obtained by solving a relaxed problem in which a single disturbance is connected to each main variable in a district, absorbing multiple disturbances that influence only a subset of those variables (for example, adding a U_{123} that absorbs $U_{12}, U_{13}, \text{ and } U_{23}$).

In sum, all classes of discrete-variable DAGs can be parameterized in terms of *generalized principal strata*. In what follows, we show how this representation allows us to reformulate causal bounding problems in terms of polynomial programs that can be optimized over the sizes of these strata, subject to constraints implied by assumptions and available data.

4 Formulating the polynomial program

We now turn to the central problem of this paper: sharply bounding causal quantities with incomplete information. Our approach is to (i) rewrite the causal estimand into a polynomial expression, (ii) rewrite modeling assumptions and empirical information into polynomial con-

the case, because U_{12} touches V_2 , U_{13} touches V_3 , and both U_{12} and U_{13} are prior to U_{23} in the ordering; thus, the children of U_{23} are influenced by multiple prior disturbances. Furthermore, there exists no other ordering that satisfies the requirement, so Figure 2(c) is non-geared. For further discussion, see Finkelstein et al. (2021)

¹¹Districts of a canonical graph are components that remain connected after removing arrows among \mathbf{V} .

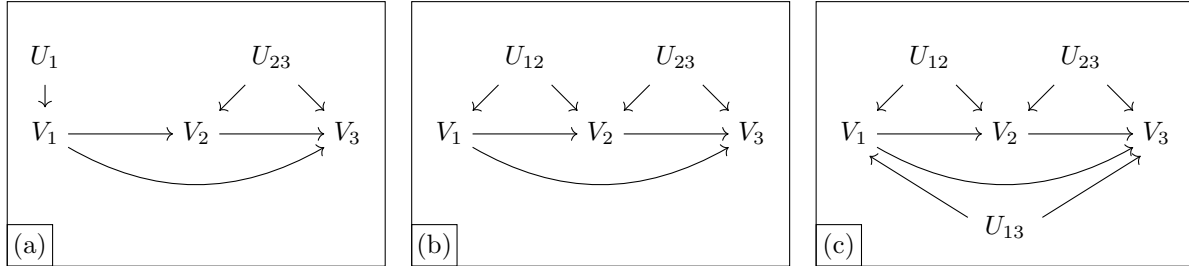
straints, and (iii) thereby transform the task into a constrained optimization problem that can be solved computationally. Appendix C.1 provides a detailed walkthrough of this process with a concrete instrumental variable problem, along with example code that illustrates how the above steps are automated by our software in merely eight lines of code.

Our goal is to obtain *sharp bounds* on the estimand, or the narrowest range that contain all admissible values consistent with available information: structural causal knowledge in the form of a canonical DAG, \mathcal{G} ; empirical evidence, \mathcal{E} ; and modeling assumptions, \mathcal{A} , formalized below. Importantly, our definition of “empirical evidence” flexibly accommodates essentially any data about the joint, marginal, or conditional distributions of the main variables.

We will suppose the main variables take on values in a known, discrete set, $\mathcal{S} = \mathcal{S}(\mathbf{V})$. In this section, we will demonstrate (i) that $\{\mathcal{G}, \mathcal{E}, \mathcal{A}, \mathcal{S}\}$ restricts the admissible values of the target quantity, and (ii) this range of observationally indistinguishable values can be recovered by polynomial programming. The causal graph and variable space, \mathcal{G} and \mathcal{S} , together imply an infinite set of possible structural equation models, each capable of producing the same full data laws. By Proposition 1, w.l.o.g., we can consider a simple model in which (i) each counterfactual main variable is a deterministic function of exogeneous, discrete disturbances; (ii) there are a relatively small number of such disturbances; and (iii) disturbances take on a finite number of possible values, corresponding to principal strata of the main variables. When repeatedly sampling units (along with each unit’s random disturbances, \mathbf{U}), the k -th disturbance thus follows the categorical distribution with parameters $\mathcal{P}_{U_k} = \{\Pr(U_k = u_k) : u_k \in \mathcal{S}(U_k)\}$. By the properties of canonical DAGs, these disturbances are independent. It follows that the parameters $\mathcal{P}_{\mathbf{U}}$ of the joint disturbance distribution $\Pr(\mathbf{U} = \mathbf{u}) = \prod_k \Pr(U_k = u_k)$ not only fully determine the distribution of each factual main variable under no intervention, $V_j(\emptyset)$ —they also determine the counterfactual distribution of $V_j(\mathbf{a})$ under any intervention \mathbf{a} , as well as its joint distribution with other counterfactual variables $V_{j'}(\mathbf{a}')$ under possibly different interventions \mathbf{a}' . This leads to the following proposition, proven in Appendix F.2.

Proposition 2. *Suppose \mathcal{G} is a canonical DAG and $\mathcal{C}_\ell = \{V_\ell(\mathbf{a}_\ell) = v_\ell\}$ are counterfactual*

Figure 2: **Any discrete-variable DAG can be represented in terms of principal strata.** Panels (a–b) depict geared graphs. In (a), each main variable is influenced by only one disturbance. In (b), V_2 is influenced by both U_{12} and U_{23} . In (c), a non-geared graph with cyclical confounding by U_{12} , U_{23} , and U_{13} is shown. For each case, the functional parameterizations—representations of each graph in terms of principal strata—are illustrated below.



Functional parameterization of (a)

V_1 has no main parents—it is deterministically assigned a value by the disturbance U_1 . Therefore, the possible values of V_1 's main parents are the empty set. V_2 has one binary main parent and takes on binary values, for a total of four possible response functions of the form $\{0,1\} \mapsto \{0,1\}$. Finally, V_3 takes in two binary parents and produces a binary outcome, with sixteen possible response patterns of the form $\{0,1\}^2 \mapsto \{0,1\}$. The disturbance U_{23} determines the 4×16 possible joint response functions of V_2 and V_3 and therefore must have a cardinality of 64.

Structural Eq.	Response Func.	Response Form	Cardinality
$V_1 = f_1(U_1)$	$f_1^{(U_1=u_1)}(\emptyset)$	$\emptyset \mapsto \{0,1\}$	$ \mathcal{S}(U_1) = 2^1$
$V_2 = f_2(V_1, U_{23})$	$f_2^{(U_{23}=u_{23})}(v_1)$	$\{0,1\} \mapsto \{0,1\}$	$ \mathcal{S}(U_{23}) = 2^2 \times 2^{2^2}$
$V_3 = f_3(V_1, V_2, U_{23})$	$f_3^{(U_{23}=u_{23})}(v_1, v_2)$	$\{0,1\}^2 \mapsto \{0,1\}$	

Functional parameterization of (b)

When a main variable (here, V_2) is influenced by multiple disturbances (U_{12} and U_{23}), an arbitrary disturbance is selected to represent its response function, while the remaining disturbances are treated as main variables. We begin by allocating U_{12} to determine the response of V_1 , then treat U_{12} as a main variable when analyzing V_2 . This leaves U_{23} to determine the response of V_2 to both V_1 and U_{12} . In addition, as before, U_{23} also determines the response of V_3 to V_1 and V_2 . Different disturbance allocations result in identical bounds.

Structural Eq.	Response Func.	Response Form	Cardinality
$V_1 = f_1(U_{12})$	$f_1^{(U_{12}=u_{12})}(\emptyset)$	$\emptyset \mapsto \{0,1\}$	$ \mathcal{S}(U_{12}) = 2^1$
$V_2 = f_2(V_1, U_{12}, U_{23})$	$f_2^{(U_{23}=u_{23})}(v_1, u_{12})$	$\{0,1\} \times \mathcal{S}(U_{12}) \mapsto \{0,1\}$	$ \mathcal{S}(U_{23}) = 2^{2 \times 2} \times 2^{2^2}$
$V_3 = f_3(V_1, V_2, U_{23})$	$f_3^{(U_{23}=u_{23})}(v_1, v_2)$	$\{0,1\}^2 \mapsto \{0,1\}$	

Functional parameterization of (c)

In non-geared graphs, confounding cycles are broken by selecting an arbitrary disturbance (here, U_{13}) and fixing its cardinality at a non-restrictive value, $|\mathcal{S}(V_1) \times \mathcal{S}(V_2) \times \mathcal{S}(V_3)| - 2$, based on the size of its district (V_1 , V_2 , and V_3). U_{13} is then treated as a main variable for all subsequent analysis. Following (b), U_{12} then determines the response of V_1 to U_{13} . Finally, U_{23} jointly determines (i) the responses of V_2 to V_1 and U_{12} and (ii) V_3 to V_1 , V_2 , and U_{13} .

Structural Eq.	Response Func.	Response Form	Cardinality
$U_{13} = g_{13}(\emptyset)$	—	$\emptyset \mapsto \mathcal{S}(U_{13})$	$ \mathcal{S}(U_{13}) = 2^3 - 2$
$V_1 = f_1(U_{12}, U_{13})$	$f_1^{(U_{12}=u_{12})}(u_{13})$	$\mathcal{S}(U_{13}) \mapsto \{0,1\}$	$ \mathcal{S}(U_{12}) = 2^{2^3-2}$
$V_2 = f_2(V_1, U_{12}, U_{23})$	$f_2^{(U_{23}=u_{23})}(v_1, u_{12})$	$\{0,1\} \times \mathcal{S}(U_{12}) \mapsto \{0,1\}$	$ \mathcal{S}(U_{23}) = 2^{2 \times 2^{2^3-2}} \times 2^{2^2 \times (2^3-2)}$
$V_3 = f_3(V_1, V_2, U_{13}, U_{23})$	$f_3^{(U_{23}=u_{23})}(v_1, v_2, u_{13})$	$\{0,1\}^2 \times \mathcal{S}(U_{13}) \mapsto \{0,1\}$	

statements, indexed by ℓ , that variable V_ℓ will take on value v_ℓ under manipulation \mathbf{a}_ℓ . Let $\mathbb{1}\{\mathbf{u} \Rightarrow \{\mathcal{C}_\ell : \ell\}\}$ be an indicator function that evaluates to 1 if and only if disturbance realizations $\mathbf{u} = \{u_1, \dots, u_K\}$ deterministically lead to \mathcal{C}_ℓ being satisfied for every ℓ . Then under the structural equation model \mathcal{G} ,

$$\Pr\left(\bigcap_{\ell} \mathcal{C}_\ell\right) = \sum_{\mathbf{u} \in \mathcal{S}(\mathcal{U})} \mathbb{1}\{\mathbf{u} \Rightarrow \{\mathcal{C}_\ell : \ell\}\} \prod_{u_k \in \mathbf{u}} \Pr(U_k = u_k), \quad (1)$$

which is a polynomial equation in $\mathcal{P}_{\mathcal{U}}$, the probabilities $\Pr(U_k = u_k)$.

For example, in the mediation setting of Figure 1(b), Proposition 2 implies that the joint distribution of the factual variables— $V_1(\emptyset)$, $V_2(\emptyset)$, and $V_3(\emptyset)$ —is given by

$$\Pr(V_1(\emptyset) = v_1, V_2(\emptyset) = v_2, V_3(\emptyset) = v_3) = \sum_{\{u_1, u_{23}\} \in \mathcal{U}} \Pr(U_1 = u_1) \Pr(U_{23} = u_{23}),$$

where $\mathcal{U} = \{\mathbf{u} : \mathbf{u} \Rightarrow \mathbf{v}\}$ is the set of disturbance realizations that are consistent with a particular $\mathbf{V}(\emptyset) = \mathbf{v}$, in other words,

$$\mathcal{U} = \left\{ \{u_1, u_{23}\} : f_1^{(U_1=u_1)}(\emptyset) = v_1, f_2^{(U_{23}=u_{23})}(v_1) = v_2, f_3^{(U_{23}=u_{23})}(v_1, v_2) = v_3 \right\}.$$

Alternatively, analysts may be interested in the probability that a randomly drawn unit i has a positive controlled direct effect when fixing the mediator to $V_2 = 0$. This is given by $\Pr[V_3(V_1 = 0, V_2 = 0) = 0, V_3(V_1 = 1, V_2 = 0) = 1]$ and is similarly expressed in terms of the disturbances as $\sum_{\{u_1, u_{23}\} \in \mathcal{U}'} \Pr(U_1 = u_1) \Pr(U_{23} = u_{23})$, summing over a different subset of the disturbance space, $\mathcal{U}' = \left\{ \{u_1, u_{23}\} : f_3^{(U_{23}=u_{23})}(V_1 = 1, V_2 = 0) = 1, f_3^{(U_{23}=u_{23})}(V_1 = 0, V_2 = 0) = 0 \right\}$.

We now expand this result to include a large class of functionals of marginal probabilities and logical statements about these functionals.

Corollary 1. *Suppose \mathcal{G} is a canonical DAG. Let $\mathcal{P}_{\mathbf{V}}$ denote the full data law and $g_1(\mathcal{P}_{\mathbf{V}})$ denote a functional of $\mathcal{P}_{\mathbf{V}}$ involving elementary arithmetic operations on constants and marginal probabilities of $\mathcal{P}_{\mathbf{V}}$. Then $g_1(\mathcal{P}_{\mathbf{V}})$ can be re-expressed as a polynomial fraction in the parameters of $\mathcal{P}_{\mathcal{U}}$, $g_2(\mathcal{P}_{\mathcal{U}})$, by replacing each marginal probability with its Proposition 2 polynomialization.*

We denote this replacement process with the operation *polynomial-fractionalize* [$g_1(\mathcal{P}_{\mathbf{V}})$]. The corollary has a number of implications, which we discuss briefly. First, it shows that a wide array of single-world and cross-world functionals can be expressed as polynomial fractions.

These include traditional quantities such as the ATE, as well as more complex ones such as the pure direct effect and the probability of causal sufficiency. It also suggests any non-elementary functional of \mathcal{P}_V can be approximated to arbitrary precision by a polynomial fraction, if the functional has a convergent power series. We note that non-elementary functionals rarely arise in practice, apart from logarithmic- or exponential-scale estimands.¹² An example that our approach cannot handle is the non-analytic functional $\mathbb{1}\{\text{ATE is rational}\}$.

A non-obvious implication of Corollary 1 is that when (i) $g_1(\mathcal{P}_V)$ is an elementary arithmetical functional; (ii) $\star \in \{<, \leq, =, >, \geq\}$ is a binary comparison operator; and (iii) α is a finite constant, then any statement of the form $g_1(\mathcal{P}_V) \star \alpha$ can be transformed into a set of equivalent non-fractional relations, $\{h_\ell(\mathcal{P}_U, \mathbf{s}) \star_\ell 0 : \ell\}$. Here, each $h_\ell(\cdot)$ denotes a non-fractional polynomial in the parameters indicated, \star is a possibly different binary comparison; and \mathbf{s} are newly created auxiliary variables that are sometimes necessary. The transformation proceeds as follows. First, $g_1(\mathcal{P}_V) \star \alpha$ can be rewritten as $g_2(\mathcal{P}_U) \star \alpha$, by Proposition 1. Then, note that any fractional $g_2(\mathcal{P}_U)$ can be rewritten as some $\frac{g_3(\mathcal{P}_U)}{h(\mathcal{P}_U)}$ in which $g_3(\mathcal{P}_U)$ has fewer fractions than $g_2(\mathcal{P}_U)$. Regardless of whether $h(\mathcal{P}_U)$ is positive, negative, or of indeterminate sign, it can be shown that $h(\mathcal{P}_U)$ can be cleared to obtain an equivalent relation. The exact procedure differs for each case and, when $h(\mathcal{P}_U)$ is indeterminate, requires a set of auxiliary variables, \mathbf{s} , to be created.¹³ If all fractions have been cleared from $g_3(\mathcal{P}_U)$, then the rewritten statement is also of the promised form and we are done; otherwise, recurse. We denote this transformation of the original statement—i.e., polynomial-fractionalizing its components and then clearing all resulting fractions—as *polynomialize* [$g_1(\mathcal{P}_V) \star \alpha$].

By the same token, any estimand $g(\mathcal{P}_V)$ that is a polynomial-fractional $g'(\mathcal{P}_U)$ in the parameters of \mathcal{P}_U can be re-expressed as a polynomial in the expanded parameter space,

¹²Bounds on a monotonic transform of x can be obtained by bounding x , then applying the transform.

¹³First, consider strictly positive $h(\mathcal{P}_U)$; here, $g_3(\mathcal{P}_U) - \alpha h(\mathcal{P}_U) \star 0$ is equivalent to the original statement. Second, consider strictly negative $h(\mathcal{P}_U)$: clearing the fraction yields $g_3(\mathcal{P}_U) - \alpha h(\mathcal{P}_U) \star 0$, where \star reverses an inequality \star . Finally, in the case when $h(\mathcal{P}_U)$ can take on both positive and negative values, let an auxiliary variable $s \in \mathbf{s}$ be defined such that $s \cdot h(\mathcal{P}_U) - 1 = 0$, which is a polynomial relation of the promised form. It can now be seen that the original statement is equivalent to $s \cdot g_3(\mathcal{P}_U) - \alpha \star 0$. For a concrete example of how auxiliary variables can be used to clear fractions, see Appendix C.1.3.

$h(\mathcal{P}_U, \mathbf{s})$, along with a set of additional polynomial relations. To see this, first define a new estimand, s , which is a monomial (and hence a polynomial). This new estimand can be made equivalent to the original one by imposing a new polynomial-fractional constraint, $s - g(\mathcal{P}_V) = 0$. Any remaining fractions in the new constraint are cleared as above. We will make extensive use of these properties to convert causal queries to polynomial programs.

Algorithm 2 provides a step-by-step procedure for formulating a polynomial programming problem. Solving this program via Algorithm 3 then produces sharp bounds. Both algorithms, given in Appendix B, mirror the discussion here with more formality. We begin by transforming a factual or counterfactual target of inference \mathcal{T} into polynomial form, possibly creating additional auxiliary variables to eliminate fractions. To accomplish this task, the procedure utilizes the possibly non-canonical DAG \mathcal{G} and the variable space $\mathcal{S}(\mathbf{V})$ to re-express \mathcal{T} in terms of functional parameters that correspond to principal strata proportions. The result is the objective function of the polynomial program. The procedure then polynomializes the sets of constraints resulting from empirical evidence and by modeling assumptions, respectively denoted \mathcal{E} and \mathcal{A} . In Figure 2, if only observational data is available, then \mathcal{E} consists of eight pieces of evidence, each represented as a statement corresponding to a cell of the factual distribution $\Pr[V_1(\emptyset) = v_1, V_2(\emptyset) = v_2, V_3(\emptyset) = v_3] = \Pr(V_1 = v_1, V_2 = v_2, V_3 = v_3)$ for observable values in $\{0, 1\}^3$. Modeling assumptions include all other information, such as monotonicity or dose-response assumptions; these can be expressed in terms of principal strata. For example, the assumed unit-level monotonicity of the $V_1 \rightarrow V_2$ relationship (e.g., the “no defiers” assumption of Angrist et al., 1996) can be written as the statement that $\Pr(V_2(V_1 = 0) = 1, V_2(V_1 = 1) = 0) = 0$.¹⁴ Finally, the statement that each disturbance k follows a categorical probability distribution is re-expressed as the polynomial relations $\Pr(U_k = u_k) \geq 0 \ \forall \ u_k$ and $\sum_{u_k} \Pr(U_k = u_k) = 1$.

Algorithm 2 produces an optimization problem with a polynomial objective subject to polynomial constraints. This polynomial programming problem is equivalent to the original

¹⁴Assumed population-level monotonicity is typically written $\mathbb{E}[V_2(V_1 = 1) - V_2(V_1 = 0)] \geq 0$, but can be rewritten in terms of strata as $\Pr[V_2(V_1 = 1) = 1, V_2(V_1 = 0) = 0] - \Pr[V_2(V_1 = 0) = 1, V_2(V_1 = 1) = 0] \geq 0$.

causal bounding problem. This leads directly to the following theorem.

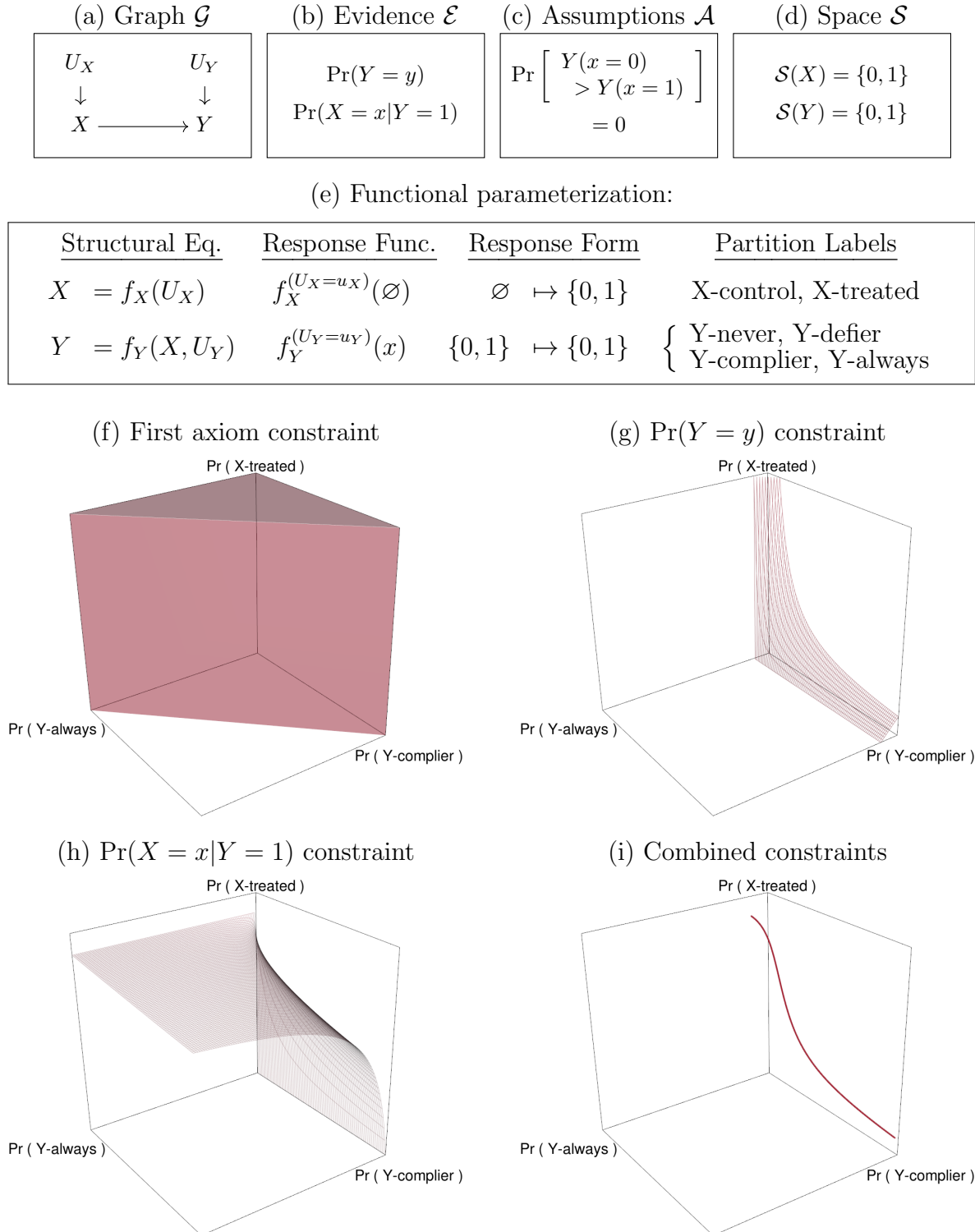
Theorem 1. *Minimization (maximization) of the polynomial program produced by Algorithm 2 produces sharp lower (upper) bounds on \mathcal{T} under the sample space $\mathcal{S}(\mathbf{V})$, structural equation model \mathcal{G} , additional modeling assumptions \mathcal{A} , and empirical evidence \mathcal{E} .*

4.1 Example program for outcome-based selection

For intuition, consider the simple example in Figure 3, motivated by a hypothetical study of discrimination in traffic law enforcement using (1) police data on vehicle stops and (2) traffic-sensor data on overall vehicle volume. For illustrative purposes, suppose all drivers behave identically. Here, $X \in \{0, 1\}$ indicates whether a motorist is a racial minority and $Y \in \{0, 1\}$ whether the motorist is stopped by police. X and Y are assumed to be unconfounded. However, there exists outcome-based selection: we only learn driver race (X) from police records if a stop occurs ($Y = 1$), thus precluding point identification. Panels (a–d) in Figure 3 depict the inputs to the algorithm: (a) the causal graph, \mathcal{G} ; (b) the observed evidence, \mathcal{E} , consisting of the marginal $\Pr(Y = y)$ and the conditional $\Pr(X = x|Y = 1)$; (c) additional assumptions, \mathcal{A} , such as a monotonicity assumption that white drivers are not discriminatorily stopped; and (d) the sample space $\mathcal{S}(X, Y)$. The target \mathcal{T} is the ATE $\mathbb{E}[Y(x = 1) - Y(x = 0)]$, the amount of anti-minority bias in stopping. Next, Figure 3(e) depicts functional parameterization in terms of six disturbance partitions, following Section 3.3. Applying simplifications from Section 4.2 results in elimination of $\Pr(Y\text{-defier})$ by assumption, then elimination of redundant strata that complete the sum to unity, $\Pr(X\text{-control})$ and $\Pr(Y\text{-never})$. The problem can thus be reduced to three dimensions. Next, the ATE is re-written as the probability of an anti-minority stop, minus that of an anti-white stop (which is zero by assumption). Finally, Figure 3(f–i) depict how each constraint narrows the space of potential solutions, leaving the admissible region shown in Figure 3(i), the only part of the model space simultaneously satisfying all constraints.

Once formulated in this way, optimization proceeds by locating the highest and lowest values of \mathcal{T} within this region, which respectively represent the upper and lower bounds on

Figure 3: **Visualization of Algorithm 2.** Constructing the polynomial program for a simple bounding problem with outcome-dependent selection, motivated by a study of discrimination in traffic law enforcement. Panels (a–d) depict inputs to the algorithm. The graph, \mathcal{G} , contains unconfounded treatment X and outcome Y . The evidence \mathcal{E} contains (i) the marginal distribution of Y and (ii) the conditional distribution of X if $Y = 1$. \mathcal{A} consists of a monotonicity assumption. \mathcal{S} states that X and Y are binary. The target \mathcal{T} is the ATE $\mathbb{E}[Y(x = 1) - Y(x = 0)]$. Panel (e) depicts functional parameterization with six disturbance partitions, following Section 3.3. Applying simplifications from Section 4.2 results in elimination of $\Pr(Y\text{-defier})$ by assumption, then elimination of $\Pr(X\text{-control})$ and $\Pr(Y\text{-never})$ by the second axiom. Panels (f–i) show constraints in the simplified model space.



the ATE. A variety of computational solvers can in principle be used to minimize and maximize it.¹⁵ However, in practice, the resulting polynomial programming problem can be much more complex than the simple case shown in Figure 3. For example, even seemingly simple causal problems can result in nonconvex objective functions or constraints; moreover, both the admissible region of the model space and the region of possible objective values can be disconnected.¹⁶ Local solvers thus cannot guarantee valid bounds without exhaustively searching the space; when time is finite, these can fail to discover global extrema for the causal estimand, resulting in invalid intervals that are not guaranteed to contain the quantity of interest.

4.2 Simplifying the polynomial program

The time needed to solve polynomial programs can grow exponentially with the number of variables. To address this, in Appendix D, we employ various techniques that draw on graph theory and probability theory to simplify polynomial programs into forms with fewer variables that generally have identical solutions but are usually faster to solve. At a high level, these simplifications fall into four categories. Appendix D.1 proposes a simplification that reduces the degree of polynomial expressions. Using the graph’s structure, we show how to detect when a disturbance U_k is guaranteed to be irrelevant, meaning its parameters only occur in contexts where $\sum_{u_k \in \mathcal{S}(U_k)} \Pr(U_k = u_k)$ can be factored out and replaced with unity. Appendix D.2 introduces a simplification that reduces the degree of polynomial expressions by exploiting equality constraints like the simple $\Pr(\text{X-control}) + \Pr(\text{X-treated}) = 1$ example above. We note some practical considerations when using symbolic algebra systems such as SageMath (Stein et al., 2019), specifically about the computational efficiency of factoring out complex polynomial expressions and replacing them with constants, as opposed to solving for one

¹⁵Throughout this paper, we will neglect the distinction between minimum (maximum) and infimum (supremum), as is standard practice in numerical optimization.

¹⁶For example, the polynomial constraint $x^3 - x^2 < -0.1$ would produce a disconnected admissible region of $x \in (-\infty, -0.280] \cup [0.413, 0.867]$. Moreover, even connected admissible regions can produce disconnected sets of possible objective values; e.g., with the objective $\frac{1}{x}$ (which can be transformed to a polynomial objective, as discussed on page 15), the constraint $\{-1 \leq x \leq 1\}$ leads to possible objective values of $(-\infty, -1] \cup [1, \infty)$. Note that discontinuity is a merely computational challenge rather than a conceptual issue, as the definition of the bounds in this case would be $[\min_{x \in [-1, 1]} \frac{1}{x}, \max_{x \in [-1, 1]} \frac{1}{x}] = (-\infty, \infty)$.

variable in terms of others. Appendix D.3 discusses a broad class of simplifications that reduce the number of constraints in the program, but with important tradeoffs. We show that assumptions encoded in a DAG, such as the empty binary graph $U_X \rightarrow X \quad Y \leftarrow U_Y$, allow the empirical evidence to be expressed using fewer constraints—here, the reduction uses only two pieces of information, $\Pr(X = 1)$ and $\Pr(Y = 1)$, exploiting the previously mentioned equality constraints and the assumed independence of X and Y . This is a reduction from the three pieces of information needed to convey $\Pr(X = x, Y = y)$,¹⁷ but comes at the cost that analysts can no longer falsify the independence assumption. Finally, Appendix D.4 provides a simplification for detecting when constraints and parameters no longer bind the objective function, meaning they can be safely eliminated from the program.

We caution that the practical application of these techniques remains an important area for future research: applying these techniques in different orders, or even with slightly different software implementations, can produce optimization programs that are mathematically equivalent but can vary substantially in runtime.

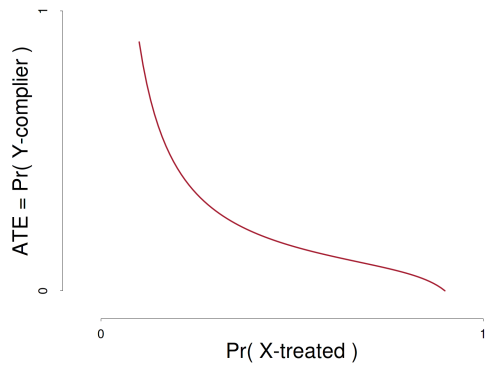
5 Computing ε -sharp bounds in polynomial programs

We now turn to the practical optimization of the polynomial program defined by Algorithm 2, which we refer to as the *primal* program; see Figure 4(a) for an example. Per Theorem 1, minimization and maximization of the polynomialized target, $\mathcal{T}(\mathbf{p})$, is equivalent to the causal bounding problem. (Optimization is implicitly over the admissible region of the model space.) We denote the sharp lower and upper bounds as $\underline{T} \equiv \min_{\mathbf{p}} \mathcal{T}(\mathbf{p})$ and $\overline{T} \equiv \max_{\mathbf{p}} \mathcal{T}(\mathbf{p})$. As we note above, the challenge is that these problems are often nonconvex and high dimensional, meaning globally optimal solutions can be difficult to obtain. Conventional primal optimizers, which iteratively improve suboptimal values, can be trapped in local extrema, failing to produce valid bounds that contain all possible values of the estimand (including global extrema).

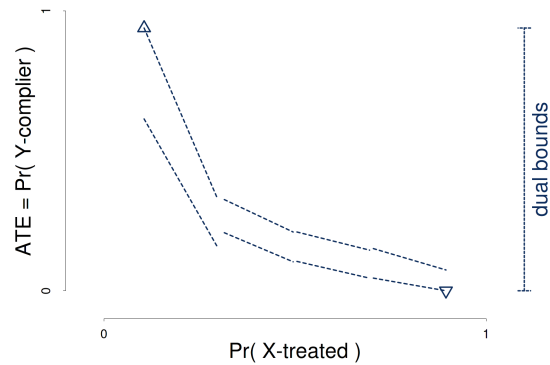
¹⁷Any one of the four cells can be automatically eliminated, as it is redundant given the implied constraint that $\sum_{x,y} \Pr(X = x, Y = y) = 1$ by construction of the principal strata.

Figure 4: **Visualization of Algorithm 3.** Computing ε -sharp bounds for the outcome-based selection problem of Figure 3. Panel (a) shows how the target ATE varies over the feasible region of the model space (reparameterized in terms of possible $\mathcal{P}_{\mathbf{U}}$ distributions) depicted in Figure 3(i). Panel (b) depicts the first step of our method, partitioning of the model space into *branches* within which computationally tractable, piecewise linear *dual relaxations* are obtained. Panel (c) shows how suboptimal values of the primal function, obtained with standard local optimizers, can be combined with the dual envelope to *prune* large regions of the model space that cannot possibly contain the global extrema. In panel (d), the procedure applied recursively. The pruned model space is rebranched and heuristic primal optimization is repeated, potentially yielding narrower dual bounds and wider primal bounds, respectively. The looseness factor, ε , narrows until reaching zero (sharpness) or a specified threshold.

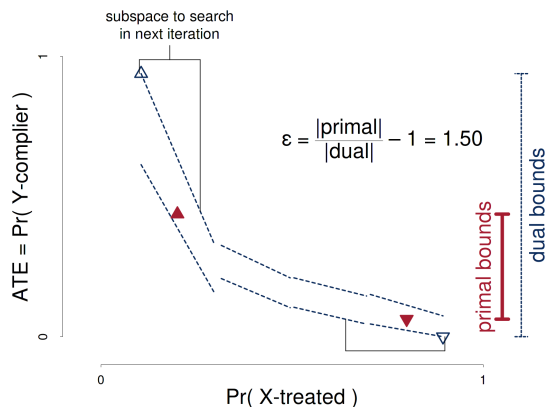
- (a) Primal function. Possible target estimand values as function of feasible disturbance distributions. Global extrema are sharp bounds. Because problems are often nonconvex, standard optimization can lead to local optima and invalid bounds.



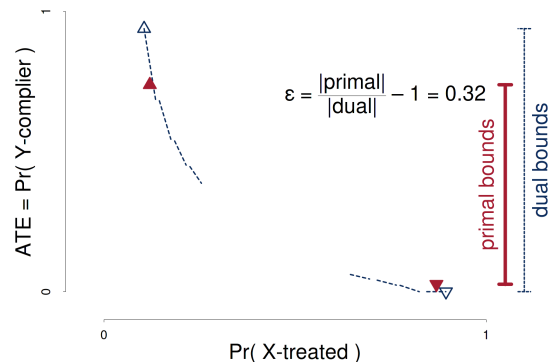
- (b) Dual envelope. Model space is divided into branches. Computationally efficient piecewise linear relaxations (curvewise bounds on primal) are obtained in each branch. Extreme dual values (hollow blue triangles) are valid but possibly loose bounds on target.



- (c) Primal refinement and pruning. Heuristic optimization produces suboptimal primal values (solid red triangles). Dual envelope then identifies regions where higher and lower primal values might be found.



- (d) Dual refinement and recursion. Remaining model space is rebranched and dual envelope is recomputed, potentially leading to narrower reported bounds. Heuristic primal optimization is repeated, potentially widening primal bounds. ε is updated.



To address this challenge and guarantee the validity of reported bounds, our approach incorporates *dual* techniques that do not directly optimize the original primal objective function, $\mathcal{T}(\mathbf{p})$. Instead, these techniques construct alternative objective functions that are easier to optimize; solutions to the easier dual problems can then be related back to the original primal problems. In particular, we will construct piecewise linear *dual envelope* functions $\underline{\mathcal{D}}(\mathbf{p})$ and $\overline{\mathcal{D}}(\mathbf{p})$ that satisfy $\underline{\mathcal{D}}(\mathbf{p}) \leq \mathcal{T}(\mathbf{p}) \leq \overline{\mathcal{D}}(\mathbf{p})$ for every \mathbf{p} in the admissible region. An illustration is given in Figure 4(b). In statistics, related techniques have found use in variational inference, an approach that constructs an analytically tractable dual function that can be maximized in place of the likelihood function (Jordan et al., 1999; Blei et al., 2017).¹⁸

A key property of this envelope is that the easier-to-compute *dual bounds*, $\underline{D} \equiv \min_{\mathbf{p}} \underline{\mathcal{D}}(\mathbf{p})$ and $\overline{D} \equiv \max_{\mathbf{p}} \overline{\mathcal{D}}(\mathbf{p})$, will always bracket the unknown true sharp bounds. This is because $\underline{\mathcal{D}}(\mathbf{p})$ and $\overline{\mathcal{D}}(\mathbf{p})$ are downward- and upward-shifted relaxations of the original objective function, which can only lead to a lower minimum and higher maximum, respectively. The dual bounds are thus guaranteed to be valid causal bounds. Viewed differently, the dual bounds $[\underline{D}, \overline{D}]$ also represent outer bounds (where bounding addresses the computationally difficult task of computing the global extrema) on the unknown sharp causal bounds $[\underline{T}, \overline{T}]$ (where bounding addresses the fundamental unknowability of the DGP). Here, a key consideration is that the choice of a dual envelope determines the looseness, or the *duality gaps* $\underline{T} - \underline{D}$ and $\overline{D} - \overline{T}$. Our task therefore reduces to the question of how to evaluate the looseness of the dual bounds and, if needed, to refine the envelope so that it leads to tighter dual bounds.

We now discuss our procedure for assessing the looseness of the dual bounds. To start, note that for any admissible point in the model space, \mathbf{p} , the corresponding value of the target quantity, $\mathcal{T}(\mathbf{p})$, must satisfy $\underline{T} \leq \mathcal{T}(\mathbf{p}) \leq \overline{T}$ by definition, even when the true sharp bounds are unknown. This immediately suggests that for any collection of points $\{\mathbf{p}, \mathbf{p}', \mathbf{p}'', \dots\}$ within the

¹⁸Variational inference uses an analytic relaxation to obtain a dual that lower-bounds the likelihood function everywhere in the model space. Our approach diverges in that (i) we conduct two simultaneous dual relaxations to obtain an envelope—both lower and upper—for the original primal function; (ii) we computationally generate piecewise dual functions, rather than analytically deriving smooth duals, and (iii) instead of working with a fixed dual function, we generate a sequence of dual envelopes that iteratively tighten the duals.

admissible region for we choose to evaluate $\mathcal{T}(\cdot)$, the lowest and highest values discovered—which we denote \underline{P} and \overline{P} —must also be contained within the sharp bounds. In other words, $[\underline{P}, \overline{P}]$ represents an inner bound on the unknown sharp bounds $[\underline{T}, \overline{T}]$. Therefore, for any choice of dual envelope and any collection of evaluated points, we have $\underline{D} \leq \underline{T} \leq \underline{P} \leq \overline{P} \leq \overline{T} \leq \overline{D}$. We evaluate the looseness of the reported dual bounds by taking the ratio of the outer bounds’ excess width to the width of the inner bounds, $\varepsilon \equiv (\overline{D} - \underline{D}) / (\overline{P} - \underline{P}) - 1$. It can be seen that when $\underline{P} = \underline{D}$ and $\overline{P} = \overline{D}$, then the reported dual bounds have provably attained sharpness and $\varepsilon = 0$. However, $\varepsilon > 0$ does not necessarily imply that the dual bounds are not sharp; for example, it may simply be that $\underline{D} = \underline{T}$, so the lower bound is sharp, but the collection of points evaluated is insufficiently large, so that $\underline{T} < \underline{P}$ and this sharpness cannot be proven. For this reason, we refer to ε as the *worst-case looseness factor*.

We are now ready to discuss how bounds are iteratively refined; a step-by-step procedure is given in Algorithm 3 in Appendix B. Note that at the outset of the procedure, the initial dual envelope may lie far from the true objective function, meaning ε will be large. We employ the spatial branch-and-bound approach to recursively subdivide the model space and efficiently search for regions in which the bounds may be improved. A variety of mature optimization frameworks can be used to implement the proposed methods, including Couenne and SCIP (Belotti et al., 2009; Vigerske and Gleixner, 2018); the key to Algorithm 3 is that the upper- and lower-bounding optimization problems must be executed in parallel, so that the relative looseness ε can be tracked over time. In addition to the polynomial program produced by Algorithm 2, our procedure accepts two stopping parameters: $\varepsilon^{\text{thresh}}$, the desired level of provable sharpness; and θ^{thresh} , the desired width of the bounds.¹⁹

Figure 4 illustrates the procedure for the outcome-based selection problem of Figure 3. The algorithm receives the primal objective function, $\mathcal{T}(\mathbf{p})$, shown in Figure 4(a), as input. It then partitions the parameter space into a series of *branches*, or connected subsets of the parameter space. Separate partitions, $\underline{\mathcal{B}}$ and $\overline{\mathcal{B}}$, are used for lower and upper bounding, respectively.

¹⁹We include θ^{thresh} to address the possibility of point identification, in which case $\overline{P} - \underline{P} = 0$, finite $\varepsilon^{\text{thresh}}$ cannot be achieved, and algorithms based on this stopping criteria alone will not terminate.

Within each branch b , a linear function $\delta_0 + \mathbf{p}^\top \boldsymbol{\delta}$ is constructed; easily computed properties such as derivatives and boundary values are used to ensure that this plane lies above or below $\mathcal{T}(\mathbf{p})$ for all admissible points in the branch.²⁰ We collect these branch-specific bounds in the piecewise functions $\underline{\mathcal{D}}(\mathbf{p}) \equiv \{\underline{\mathcal{D}}_b(\mathbf{p}) \text{ if } \mathbf{p} \in \underline{\mathcal{B}}_b : b\}$ and $\overline{\mathcal{D}}(\mathbf{p}) \equiv \{\overline{\mathcal{D}}_b(\mathbf{p}) \text{ if } \mathbf{p} \in \overline{\mathcal{B}}_b : b\}$, which define the initial dual envelope shown with dashed blue lines in Figure 4(b). Because each piece is linear, it is straightforward to compute the extreme points of the dual envelope within each branch, $\underline{D}_b = \min \{\underline{\mathcal{D}}_b(\mathbf{p}) : \mathbf{p} \in \underline{\mathcal{B}}_b\}$ and $\overline{D}_b = \max \{\overline{\mathcal{D}}_b(\mathbf{p}) : \mathbf{p} \in \overline{\mathcal{B}}_b\}$. The overall dual (outer) bounds are then $\underline{D} = \min_b \underline{D}_b$ and $\overline{D} = \max_b \overline{D}_b$, depicted with hollow blue triangles.

Next, the algorithm seeks to expand the primal (inner) bounds. Recall that these bounds, $[\underline{P}, \overline{P}]$, are the minimum and maximum values of the target function that have been encountered in any set of admissible DGPs—regardless of how that set was constructed. We can therefore use standard constrained optimization techniques to optimize the primal problem. Various heuristics—e.g. initializing optimizers in regions that appear promising based on the duals—can also be used. The fact that these techniques are only guaranteed to produce local optima is not of concern, because primal bounds are used only for computational convenience. Examples of two admissible primal points are shown with red triangles in Figure 4(c). These primal bounds represent the narrowest possible causal bounds: the (unknown) sharp lower bound \underline{T} must satisfy $\underline{T} \leq \underline{P}$, and similarly the sharp upper bound must satisfy $\overline{P} \leq \overline{T}$. This means entire swaths of the parameter space can now be ignored, greatly accelerating the search. For example, in Figure 4(c), the upper dual function (upper dashed blue lines) indicates that the rightmost three-quarters of the parameter space cannot possibly produce a target value that is higher than \overline{P} , the upper primal bound that has already been found (upper solid red triangle). Therefore, optimization of the upper dual bound can focus on the bracketed “subspace to search in next iteration.” Optimization of the lower dual bound only

²⁰For example, consider the objective function $\mathcal{T}(x) = x^2$. Any tangent line is a valid lower dual function. Moreover, within any interval $[x_a, x_b]$, the secant line from $(x_a, \mathcal{T}(x_a))$ to $(x_b, \mathcal{T}(x_b))$ is a valid upper dual function. A piecewise linear envelope can thus be constructed by proceeding one branch at a time, computing derivatives (for example, at the branch midpoint) to obtain a branch-specific lower dual function $\underline{\mathcal{D}}_b(x)$ and boundary values to obtain a branch-specific upper dual function $\overline{\mathcal{D}}_b(x)$.

need consider regions that $\underline{D}(\mathbf{p})$ indicates can produce lower values than \underline{P} .

A new refined dual envelope can now be constructed by subdividing the remaining space and recomputing tighter dual functions, as shown in Figure 4(d). The procedure is then repeated recursively—the algorithm heuristically selects branches in the model space that appear promising, then refines primal and dual bounds in turn. If a more extreme admissible target value is found, it is stored as a new primal bound. Finally, the algorithm prunes branches of $\underline{\mathcal{B}}$ and $\overline{\mathcal{B}}$ that cannot improve dual bounds or that wholly violate constraints. Optimization terminates when either ε reaches $\varepsilon^{\text{thresh}}$ or $\theta \equiv \overline{D} - \underline{D}$ reaches θ^{thresh} . For complex problems, the time to convergence may be prohibitive. But because the dual function is always guaranteed to contain the true objective function, the algorithm is *anytime*—the user can halt the program at any point and obtain valid (but potentially loose) bounds.

6 Statistical inference

We now briefly discuss how to modify Algorithm 3 to account for sampling error in the empirical evidence used to construct bounds. A more rigorous formalization is provided in Appendix E.

Consider a simulated binary $X \rightarrow Y$ graph with confounding $X \leftarrow U_{XY} \rightarrow Y$. Up until now, when discussing how empirical evidence constrains the admissible DGPs, we have only considered population distributions of observable quantities—here, $\mathcal{E} = \{ \Pr(X = 0, Y = 0) = 0.121, \Pr(X = 1, Y = 0) = 0.346, \Pr(X = 0, Y = 1) = 0.349, \Pr(X = 1, Y = 1) = 0.184 \}$. When these *population constraints* are input to the algorithm, we refer to the results as the *population bounds*. In practice, however, analysts only have access to noisily estimated versions; with $N = 1,000$, the sample analogues might respectively be 0.113, 0.352, 0.357, and 0.178. By the plug-in principle, *estimated bounds* are obtained by supplying *estimated constraints* instead. In other words, we apply the algorithm *as if* $\Pr(X = x, Y = y) = \widehat{\Pr}(X = x, Y = y)$.

Next, we propose two easily polynomializable methods to account for uncertainty from sampling error. Our general approach is to relax empirical-evidence constraints: we say that

$\Pr(X = x, Y = y)$ must be *near* $\widehat{\Pr}(X = x, Y = y)$, rather than equaling it. Our first method is based on the “Bernoulli-KL” approach of [Malloy et al. \(2020\)](#), which constructs separate confidence regions for each observable $\Pr(X = x, Y = y)$. For example, rather than constraining Algorithm 3 to only consider DGPs exactly satisfying $\Pr(X = 0, Y = 0) = 0.121$ as in the population bounds, or $\Pr(X = 0, Y = 0) = 0.113$ as in the estimated bounds, we instead allow it to consider any DGP in which $0.073 \leq \Pr(X = 0, Y = 0) \leq 0.163$. Thus, each equality constraint in the original empirical evidence is replaced with two linear inequality constraints; this is equivalent to constraining $\Pr(X = x, Y = y)$ to lie within a hypercube.

Our second method is based on the multivariate Gaussian limiting distribution of the multinomial proportion ([Bienaymé, 1838](#)). This approach will essentially say that $\Pr(X = x, Y = y)$ is constrained to lie within an ellipsoid, rather than a hypercube. Let $\hat{\mathbf{E}}$ be a vector collecting $\left[\widehat{\Pr}(X = 0, Y = 0), \widehat{\Pr}(X = 1, Y = 0), \widehat{\Pr}(X = 0, Y = 1)\right]$.²¹ We then compute a confidence region for the distribution $\mathcal{N}\left(\hat{\mathbf{E}}, \frac{1}{N}\text{diag}(\hat{\mathbf{E}}) - \frac{1}{N}\hat{\mathbf{E}}\hat{\mathbf{E}}^\top\right)$. This replaces all of the original equality constraints with a single quadratic inequality constraint of the form $\left(\hat{\mathbf{E}} - \mathbf{E}\right)^\top \left(\frac{1}{N}\text{diag}(\hat{\mathbf{E}}) - \frac{1}{N}\hat{\mathbf{E}}\hat{\mathbf{E}}^\top\right)^{-1} \left(\hat{\mathbf{E}} - \mathbf{E}\right) \leq z$, where $\mathbf{E} = \left[\Pr(X = 0, Y = 0), \Pr(X = 1, Y = 0), \Pr(X = 0, Y = 1)\right]$ and z is some critical value of the χ^2 distribution.

Specifics of the calculations are given in Appendix E. These confidence regions for the empirical quantities aim to jointly cover $\Pr(X = x, Y = y)$ for every x and y in at least $1 - \alpha$ of repeated samples (the Bernoulli-KL method guarantees conservative coverage in finite samples, whereas the Gaussian method offers only asymptotic guarantees). When this holds, *confidence bounds* obtained by optimizing subject to the relaxed empirical constraints are guaranteed to have at least $1 - \alpha$ coverage of the population bounds. In Section 7.2, we show that empirically, confidence bounds obtained from both methods are conservative.

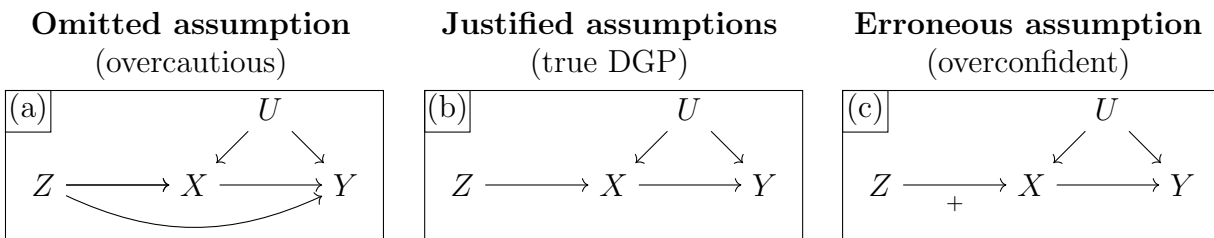
²¹To avoid degeneracy issues, one empirical quantity is excluded, as it must sum to unity.

7 Simulated examples

7.1 Instrumental variables

Noncompliance with randomized treatment assignment is a common obstacle to causal inference. [Balke and Pearl \(1997\)](#) showed that bounds on the ATE under noncompliance can be obtained via linear programming. However, that approach cannot be used to bound the local ATE (LATE) among “compliers” because this quantity is nonlinear. [Angrist et al. \(1996\)](#) shows the LATE can be point identified if certain conditions hold—including, notably, (i) the absence of a direct effect of treatment assignment Z on the outcome Y ; and (ii) monotonicity, or the absence of “defiers” in which actual treatment X is the inverse of Z .²² Because these may not be satisfied in practice, [Figure 5](#) shows three possible sets of assumptions that analysts may make: (a) neither; (b) the former but not the latter; and (c) both. We simulate a true DGP corresponding to panel (b), in which no-direct-effect holds but monotonicity is violated. The true ATE is -0.25 and the true LATE is -0.36 . We will suppose analysts have access to the population distribution $\Pr(Z = z, X = x, Y = y)$; inference is discussed in [Section 7.2](#).

Figure 5: **DGPs with noncompliance.** Three possible scenarios involving encouragement Z , treatment X , and outcome Y . Panel (b) represents the true simulation DGP, where $Z \rightarrow X$ monotonicity is violated (indicated by absence of a +). Panel (a) depicts a “overcautious” analyst unwilling to assume away a direct $Z \rightarrow Y$ effect. Panel (c) depicts an “overconfident” analyst that incorrectly assumes monotonicity of $Z \rightarrow X$.



An overcautious analyst might be unwilling to rule out a direct $Z \rightarrow Y$ effect or defiers in $Z \rightarrow X$, making only assumptions shown in panel (a). Applying our method yields bounds of $[-0.63, 0.37]$ and $[-1, 1]$ for the ATE and LATE, respectively—sharp, but uninformative in sign. With an additional no-direct-effect assumption, per panel (b), they would instead

²²Other conditions include ignorability of Z and a non-null effect of Z on X .

obtain ATE bounds of $[-0.55, -0.15]$, revealing a negative effect and correctly containing the true ATE, -0.25 . However, LATE bounds remain at $[-1, 1]$; as compliers cannot be identified experimentally, this quantity is difficult to learn about without strong assumptions. Finally, an overconfident analyst might mistakenly make an additional monotonicity assumption. Helpfully, when asked to produce bounds, Algorithm 3 reports the causal query is *infeasible*—meaning that it cannot locate any DGP consistent with data and assumptions. This clearly warns that the causal theory is deficient. If the analyst had naïvely applied the traditional two-stage least squares estimator, they would have received no such warning. Instead, they would have obtained an erroneous point estimate of -0.74 , roughly double the true LATE.

7.2 Coverage of confidence bounds

We now evaluate the performance of confidence bounds that characterize uncertainty due to sampling error, constructed according to Section 6 and Appendix E, using the instrumental variable model of Figure 5(b). Specifically, we draw samples of $N = 1,000$, $N = 10,000$, or $N = 100,000$ observations from this DGP. For each sample, we then use the empirical proportions $\frac{1}{N} \sum_i \mathbb{1}\{Z_i = z, X_i = x, Y_i = y\}$ for all $x, y, z \in \{0, 1\}$. These eight quantities form the basis of estimated bounds, by the plug-in principle. To quantify uncertainty, we compute 95% confidence regions on the same observed quantities, then convert them to polynomial constraints for inclusion in Algorithm 3. Optimizing subject to these confidence constraints produces confidence bounds, depicted in Figure 6. For each combination of sample size and uncertainty method, we draw 1,000 simulated datasets and run Algorithm 3 on each.

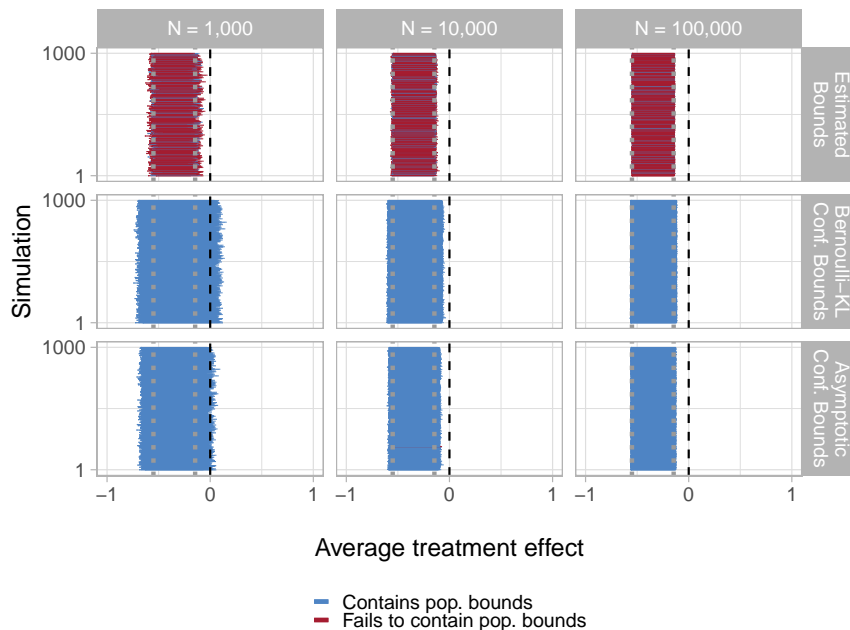
Table 1 reports average values of estimated lower (upper) confidence bounds obtained by Algorithm 3 over 1,000 simulated datasets, for varying N . At all sample sizes, estimated bounds are centered on population bounds. Figure 13 shows confidence bounds obtained across methods and sample sizes. The Bernoulli-KL method produces wider confidence intervals at all N ; at $N = 1,000$, it is generally unable to reject zero, whereas the asymptotic method does so occasionally. Differences in interval width persist but shrink rapidly as sample size

grows and both methods collapse on population bounds. As discussed in Section 6, we find more conservative coverage for confidence bounds on the ATE (100% coverage of population bounds), compared to coverage of the underlying confidence regions on the observed quantities (95% joint coverage of observed population quantities for the asymptotic method).

Table 1: **Bias of estimated bounds.** Average lower (upper) estimated bounds simulated datasets of varying size. Average estimated bounds correspond closely to population bounds.

Quantity	$N = 1,000$	$N = 10,000$	$N = 100,000$	Population
Lower bound	-0.5497	-0.5498	-0.5500	-0.5502
Upper bound	-0.1453	-0.1455	-0.1459	-0.1460

Figure 6: **Coverage of confidence bounds.** Each of 1,000 simulations is depicted with a horizontal line. For each simulation, a horizontal error bar represents a 95% confidence bound obtained per Section 6. All confidence bounds fully contain the population bounds, indicating 100% coverage. The upper (lower) row of panels reflect confidence bounds obtained with the Bernoulli-KL (asymptotic) method. Columns of panels report confidence bounds obtained using samples of various sizes. Vertical dotted gray lines show true population lower and upper bounds, which contain the true ATE of -0.25 ; vertical dashed black lines indicate zero.



7.3 More complex bounding problems

We now examine four hypothetical DGPs, shown in Figure 7, featuring various threats to inference. Throughout, we target the ATE of X on Y . Panel (a) illustrates outcome-based

selection: we observe unit i only if $S = 1$, where S may be affected by Y . Selection severity, $\Pr(S = 0)$, is known, but no information about $\Pr(X = x, Y = y | S = 0)$ is available. X and Y are also confounded by unobserved U . Bounding in this setting is a nonlinear program, with an analytic solution recently derived in [Gabriel et al. \(2022\)](#). Panel (b) illustrates measurement error: an unobserved confounder U jointly causes Y and its proxy Y^* , but only treatment and the proxy outcome are observed. Bounding in this setting is a linear problem. A number of results for linear measurement error were recently presented in [Finkelstein et al. \(2020\)](#); here, we examine the monotonic errors case, where $Y^*(Y = 1) \geq Y^*(Y = 0)$. Panel (c) depicts missingness in outcomes, i.e. nonresponse or attrition. Here, X affects both the partially observed Y and response indicator R ; if $R = 1$, then $Y^* = Y$, but if $R = 0$, then Y^* takes on the missing value indicator NA. Nonresponse on Y is differentially affected by both X and the value of Y itself (i.e. “missingness not at random,” MNAR); [Manski \(1990\)](#) provides analytic bounds. Finally, panel (d) depicts joint missingness in both treatment and outcome—sometimes a challenge in longitudinal studies with dropout—with MNAR on Y .

Figure 7: **Various threats to inference.** Panels depict (a) outcome-based selection, (b) measurement error, (c) nonresponse and (d) joint missingness. In each graph, X and Y are treatment and outcome, respectively. Dotted red regions represent observed information. In (a), the box around S indicates selection: other variables are only observed conditional on $S = 1$. In (b), Y^* represents a mismeasured version of the unobserved true Y . In (c), R_Y indicates reporting, so that $Y^* = Y$ if $R = 1$ and is missing otherwise. In (d), both treatment and outcome can be missing; and missingness on X can affect missingness on Y .

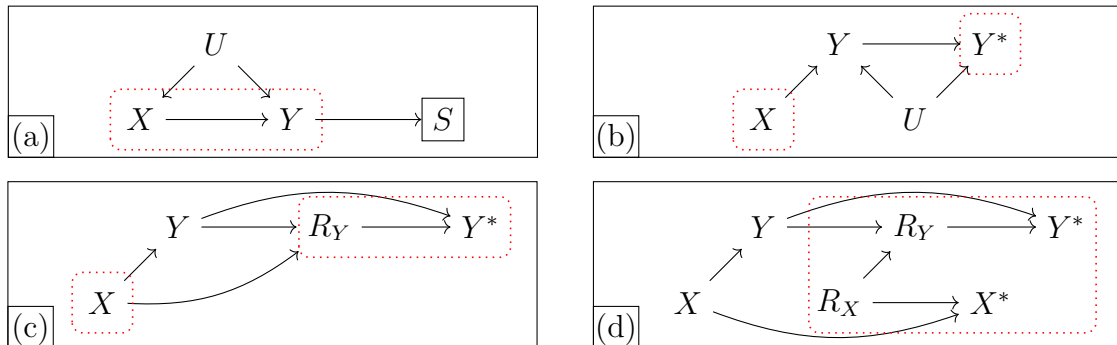
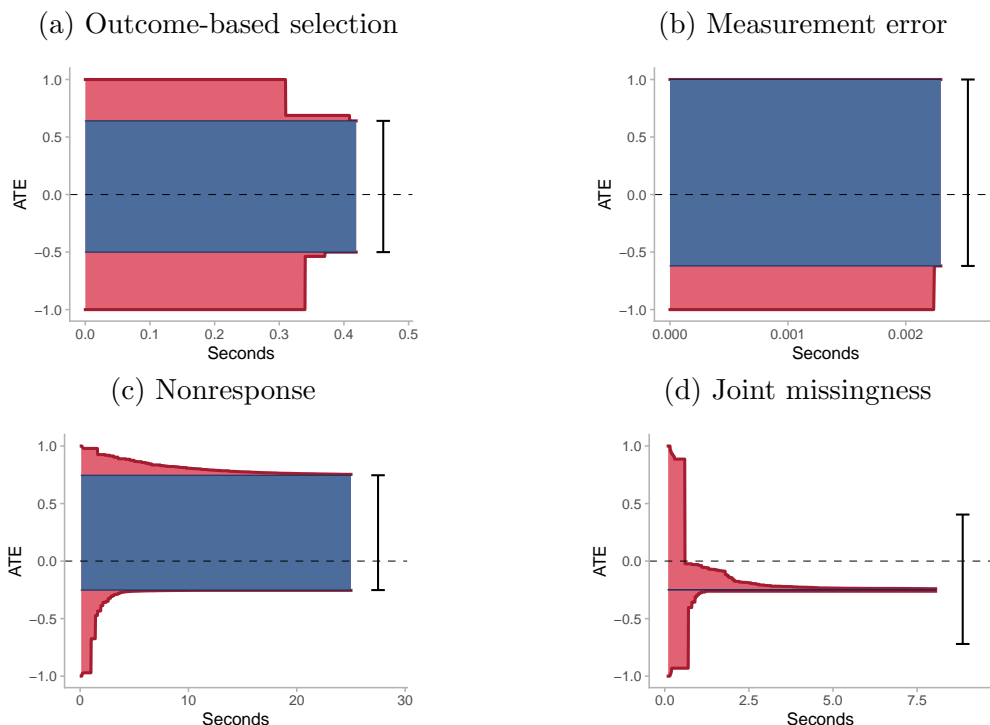


Figure 8 illustrates how Algorithm 3 recovers sharp bounds. Each panel shows progress in time. Primal bounds (blue) can widen over time if more extreme, observationally equivalent models are found. Dual bounds (red) narrow as the outer envelope is tightened. Our method

simultaneously searches for more extreme primal points and narrows the dual envelope. Analysts can terminate the process at any time, reporting guaranteed-valid dual bounds along with their worst-case suboptimality factor, ε —or await complete sharpness, $\varepsilon = 0$.

Figure 8: **Computation of ATE bounds.** Progress of Algorithm 3 for simulation data from DGPs depicted in Figure 7(a–d). Black error bars are known analytic bounds, y -axes are ATE values, and x -axes are runtimes of Algorithm 3. Red regions are dual bounds, which always contain sharp bounds and the unknown true causal effect; these can only narrow over time, converging on optimality. Blue regions are primal bounds, which can only widen over time as more extreme models are found. Optimization stops when primal and dual bounds meet, indicating bounds are sharp. Prior analytic bounds are sharp for problems (a–c). In setting (d), Algorithm 3 achieves point identification, but Manski (1990) bounds do not.



In Figure 8(a–c), the algorithm converges on known analytic results. Ultimately, in the selection simulation (a), Algorithm 3 achieves bounds of $[-0.50, 0.64]$, correctly recovering Gabriel et al.’s (2022) analytic bounds; in (b), measurement error bounds are $[-0.62, 1.00]$, matching Finkelstein et al. (2020); and in (c), outcome missingness bounds are $[-0.25, 0.75]$, equaling Manski (1990) bounds. Somewhat counterintuitively, Figure 8(d) shows dual bounds collapsing to a point, eventually point-identifying the ATE at -0.25 despite severe missingness. This surprising result turns out to be a special case of an approach using “shadow variables”

recently developed by Miao et al. (2016).²³ This example illustrates the algorithm is general enough to recover results even when they are not widely known in a particular model; note the commonly used approach of Manski (1990) produces far looser bounds of $[-0.72, 0.40]$, failing to exploit causal structure given in Figure 7(d). This result suggests our approach enables an empirical investigation of complex models where general identification results are not yet available. Situations where bounds converge suggest models where point identification via an explicit functional may be possible, potentially enabling new identification theory.

8 Potential critiques of the approach

“The user must know the true causal model.” This is false; users do not need to assert a faulty “complete” model, but rather only what they know or believe. Our approach simply derives the conclusions that follow from data and those transparently stated assumptions.

“The bounds will be too wide to be informative.” This is no tradeoff: faulty point estimates based on faulty assumptions are also uninformative. When sharp bounds incorporating all defensible assumptions are wide, it merely means progress will require more information.

“What about continuous variables?” Discrete approximations often suffice in applied work. If continuous treatments only affect discrete outcomes when exceeding a threshold, discretization is lossless. Future work may study discrete approximations when effects are smooth.

“The bounds will take too long to compute.” Achieving $\varepsilon = 0$ may sometimes take prohibitive time, but our approach remains faster than manual derivation. Figure 8 shows that several recently published results were recovered in mere seconds. Moreover, our anytime guarantee ensures that premature termination will still produce valid bounds.

²³Specifically, it can be shown the ATE is identified for the Figure 7(d) graph only among faithful distributions where $X \rightarrow Y$ is non-null—i.e. almost everywhere in the model space.

9 Conclusion

Causal inference is a central goal of science, and many established techniques can point-identify causal quantities under ideal conditions. But in many applications, these conditions are simply not satisfied, necessitating partial identification—yet few tools for obtaining these bounds exist. For knowledge accumulation to proceed in the messy world of applied statistics, a general solution is needed. We present a tool to automatically produce sharp bounds on causal quantities in settings involving discrete data. Our approach involves a reduction of all such causal queries to polynomial programming problems, enables efficient search over observationally indistinguishable DGPs, and produces sharp bounds on arbitrary causal estimands. This approach is sufficiently general to accommodate essentially every classic inferential obstacle.

Beyond providing a general tool for causal inference, our approach aligns closely with recent calls to improve research transparency by explicitly declaring estimands, identifying assumptions, and causal theory (Miguel et al., 2014; Lundberg et al., 2021). Only with a common understanding of goals and premises can scholars have meaningful debates over the credibility of research. When aspects of a theory are contested, our approach allows for a fully modular exploration of how assumptions affect empirical conclusions. Scholars can learn whether assumptions are empirically consequential, and if so, craft a targeted line of inquiry to probe its validity. Our approach can also act as a safeguard for analysts, flagging assumptions as infeasible when they conflict with observed information.

Among other challenges, our method is not immune to universal issues in causal inference such as the difficulty of knowing the correct causal structure. Other obstacles relate to computation time for complex problems, an important avenue for future research. While we do supply a method to characterize the looseness of non-sharp bounds, future work should seek to reduce computation time for sharp bounds, especially when incorporating point-identified sub-quantities or additional semi-parametric modeling approaches. Causal inference scholars may also use this method as an exploratory tool to aid in the discovery of new identification theory. These lines of inquiry now represent the major open questions in discrete causal inference.

References

- Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 91(434), 444–455.
- Balke, A. and J. Pearl (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association* 92(439), 1171–1176.
- Belotti, P., J. Lee, L. Liberti, F. Margot, and A. Wächter (2009). Branching and bounds tightening techniques for non-convex MINLP. *Optimization Methods and Software* 24(4-5).
- Bienaymé, I. J. (1838). *Mémoire sur la probabilité des résultats moyens des observations: démonstration directe de la règle de Laplace*. Imprimerie Royale.
- Blei, D. M., A. Kucukelbir, and J. D. McAuliffe (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association* 112(518), 859–877.
- Bonet, B. (2001). Instrumentality tests revisited. In J. S. Breese and D. Koller (Eds.), *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, pp. 48–55.
- Cai, Z., M. Kuroki, J. Pearl, and J. Tian (2008). Bounds on direct effects in the presence of confounded intermediate variables. *Biometrics* 64(3), 695–701.
- Carathéodory, C. (1907, March). Über den variabilitätsbereich der koeffizienten von potenzreihen, die gegebene werte nicht annehmen. *Mathematische Annalen* 64(1), 95–115.
- Dean, T. L. and M. Boddy (1988). An analysis of time-dependent planning. pp. 49–54. American Association for Artificial Intelligence.
- Evans, R. (2018). Margins of discrete Bayesian networks. *Annals of Statistics* 46(6A).
- Evans, R. J. (2016). Graphs for margins of Bayesian networks. *Scandinavian Journal of Statistics* 43(3), 625–648.
- Evans, R. J. and T. S. Richardson (2019). Smooth, identifiable supermodels of discrete DAG models with latent variables. *Bernoulli* 25(2), 848–876.
- Finkelstein, N., R. Adams, S. Saria, and I. Shpitser (2020). Partial identifiability in discrete data with measurement error. *arXiv preprint arXiv:2012.12449*.
- Finkelstein, N., E. Wolfe, and I. Shpitser (2021). Non-restrictive cardinalities and functional models for discrete latent variable DAGs. *Working Paper*.
- Frangakis, C. E. and D. B. Rubin (2002). Principal stratification in causal inference. *Biometrics* 58(1), 21–29.
- Gabriel, E. E., M. C. Sachs, and A. Sjölander (2022). Causal bounds for outcome-dependent sampling in observational studies. *Journal of the American Statistical Association* 117.
- Geiger, D. and C. Meek (1999). Quantifier elimination for statistical problems. In *Proceedings of Fifteenth Conference on Uncertainty in Artificial Intelligence*, pp. 226–235.

- Greenland, S. and J. Robins (1986). Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology* 15, 413–419.
- Guo, F. R. and T. S. Richardson (2021, jan). Chernoff-type concentration of empirical probabilities in relative entropy. *IEEE Transactions on Information Theory* 67(1), 549–558.
- Jordan, M. I., Z. Ghahramani, T. S. Jaakkola, and L. K. Saul (1999). An introduction to variational methods for graphical models. *Machine learning* 37(2), 183–233.
- Kennedy, E. H., S. Harris, and L. J. Keele (2019). Survivor-complier effects in the presence of selection on treatment, with application to a study of prompt ICU admission. *Journal of the American Statistical Association* 114(525), 93–104.
- Knox, D., W. Lowe, and J. Mummolo (2020). Administrative records mask racially biased policing. *American Political Science Review* 114, 619–637.
- Lee, D. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies* 76(3), 1071–1102.
- Lundberg, I., R. Johnson, and B. M. Stewart (2021). What is your estimand? Defining the target quantity connects statistical evidence to theory. *American Sociological Review* 86(3).
- Malloy, M. L., A. Tripathy, and R. D. Nowak (2020). Optimal confidence regions for the multinomial parameter. *arXiv preprint arXiv:2002.01044*.
- Manski, C. (1990). Nonparametric bounds on treatment effects. *The American Economic Review* 80(2), 319–323.
- Miao, W., L. Liu, E. T. Tchetgen, and Z. Geng (2016). Identification, doubly robust estimation, and semiparametric efficiency theory of nonignorable missing data with a shadow variable. *Biometrika* 103, 475–482.
- Miguel, E., C. Camerer, K. Casey, J. Cohen, K. Esterling, A. Gerber, R. Glennerster, D. Green, M. Humphreys, G. Imbens, and D. Laitin (2014). Promoting transparency in social science research. *Science* 343(6166), 30–31.
- Molinari, F. (2020). Microeconometrics with partial identification. [arXiv:2004.11751](https://arxiv.org/abs/2004.11751).
- Pearl, J. (1995). On the testability of causal models with latent and instrumental variables. *Uncertainty in Artificial Intelligence II*. San Francisco, CA: Morgan Kaufmann Publishers.
- Pearl, J. (2000). *Causality*. New York: Cambridge University Press.
- Ramsahai, R. R. (2012). Causal bounds and observable constraints for non-deterministic models. *Journal of Machine Learning Research* 13(3), 829–848.
- Richardson, T. (2003). Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics* 30(1), 145–157.
- Richardson, T. S., R. J. Evans, J. M. Robins, and I. Shpitser (2017). Nested Markov properties for acyclic directed mixed graphs. Working paper.

- Richardson, T. S. and J. M. Robins (2013). Single world intervention graphs (SWIGs) : A unification of the counterfactual and graphical approaches to causality. *Working Paper, Center for Stat. & Soc. Sci., U. Washington* 128(30).
- Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling* 7(9-12), 1393–1512.
- Sachs, M., E. Gabriel, and A. Sjölander (2020). Symbolic computation of tight causal bounds.
- Shpitser, I. (2013). Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive Science (Rumelhart special issue)* 37, 1011–1035.
- Shpitser, I. (2018). Identification in graphical causal models. In M. Maathuis, M. Drton, S. Lauritzen, and M. Wainwright (Eds.), *Handbook of Graphical Models*. CRC Press.
- Shpitser, I., R. J. Evans, and T. S. Richardson (2018). Acyclic linear SEMs obey the nested Markov property. In *Proc. of the 34th Conf. on Uncertainty in Artificial Intelligence*.
- Sjölander, A., W. Lee, H. Källberg, and Y. Pawitan (2014). Bounds on causal interactions for binary outcomes. *Biometrics* 70(3), 500–505.
- Stein, W. et al. (2019). *Sage Mathematics Software (Version 9.0)*. The Sage Development Team. www.sagemath.org.
- Swanson, S., M. Hernán, M. Miller, J. Robins, and T. Richardson (2018). Partial identification of the average treatment effect using instrumental variables. *Journal of the American Statistical Association* 113(522), 933–947.
- Tian, J. and J. Pearl (2002). On the testable implications of causal models with hidden variables. In *Proceedings of the 18th Conference in Uncertainty in Artificial Intelligence*.
- Tukey, J. (1986). Sunset salvo. *The American Statistician* 40(1), 72–76.
- Verma, T. and J. Pearl (1990). Equivalence and synthesis of causal models. In P. P. Bonissone, M. Henrion, L. N. Kanal, and J. F. Lemmer (Eds.), *Proc. of the Conf. on Uncertainty in Artificial Intelligence*, pp. 255–268. Morgan Kaufmann.
- Vigerske, S. and A. Gleixner (2018). SCIP: Global optimization of mixed-integer nonlinear programs in a branch-and-cut framework. *Optimization Methods and Software* 33(3).
- Wolfe, E., R. W. Spekkens, and T. Fritz (2019). The inflation technique for causal inference with latent variables. *Journal of Causal Inference* 7(2).
- Zhang, J. and E. Bareinboim (2021, Feb). Non-parametric methods for partial identification of causal effects. Technical Report R-72, Causal AI Lab, Columbia University.
- Zhang, J. L. and D. B. Rubin (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by “death”. *Journal of Educational and Behavioral Statistics* 28(4), 353–368.

A Glossary of terms

A.1 Bounding terms

- **Sharp bounds.** The narrowest range that contain all values of the target quantity that are admissible, or consistent with available information.
- **Primal function.** The target quantity, expressed as a function of the unknown parameters, or principal strata sizes.
- **Primal bounds.** The minimal and maximal admissible values of the target quantity discovered thus far in the bounding process (corresponding to possible parameter values, or principal strata sizes, that are consistent with available information).
- **Dual function (lower and upper).** A relaxation of the primal function; i.e., an auxiliary function that provably takes on lower (higher) values than the minimum (maximum) values of the primal function at all points in the model space.
- **Dual envelope.** The region enclosed by the lower and upper dual functions.
- **Dual bounds.** The minimal point on the lower dual function and the maximal point on the upper dual function.
- **ε -sharpness.** One minus the ratio of (i) the dual bounds width to (ii) the primal bounds width. This represents a worst-case looseness factor, quantifying how much the current (potentially non-sharp) bounds could potentially be improved with additional computation.
- **Empirical evidence.** A set of constraints, one for each observable quantity, that relates the value of that quantity to a function of the unknown parameters (principal strata sizes).
- **Modeling assumptions.** Assumed restrictions, such as monotonicity, on how a main variable responds to counterfactual manipulation of its causal parents.
- **Elementary arithmetic operations.** Addition, subtraction, multiplication, and division.
- **Population bounds.** Bounds for a quantity assuming infinite data, i.e., without assuming sampling.
- **Estimated bounds.** Calculated bounds assuming finite data (sampling).
- **Confidence bounds.** Estimated bounds that must cover population bounds with a probability α .
- **Population constraints.** Constraints of a causal problem when one assumes infinite data.
- **Estimated constraints.** Constraints of a causal problem when one assumes finite data (sampling).
- **Confidence constraints.** Constructed constraints that must cover population constraints with probability α .

A.2 Computational terms

- **Auxiliary variable.** Parameters added to an optimization problem with the goal of aiding simplification. They are used here for the purpose of eliminating fractions to produce polynomials.
- **Co-occurrence (of parameters).** In a polynomial program, two parameters x and y are said to co-occur if they appear in the same constraint.
- **Constraint set.** Inequalities and equalities that constrain the admissible values of the optimization variables.
- **Interaction (of parameters).** In a polynomial program, two parameters x and z are said to interact if there exists a sequence of parameters starting with x and ending in z in which every adjacent pair co-occurs (defined above). For example, x and z interact if x co-occurs with y and y co-occurs with z .
- **Linear programming.** A class of optimization problems in which the objective functions and all constraints are expressed as linear functions of the optimization variables.
- **NP.** The class of problems solvable by a nondeterministic Turing machine in polynomial time (see below).
- **NP-hard.** The class of problems that are as hard as the hardest problems in *NP*.
- **Numerical optimization.** An approach to optimization that returns numerical solutions (in contrast to symbolic optimization).
- **Polynomial time.** Property of an algorithm indicating that worst-case computation time grows at most as a polynomial function of the size of the number of inputs.
- **Polynomial programming.** A class of optimization problems in which the objective function and all constraints are expressed in terms of polynomials of the optimization variables.
- **Relaxation.** The transformation of an original problem to another one which is easier to solve. For example, under certain conditions, parts of a polynomial program can be approximated by a linear one.
- **Symbolic optimization.** An approach to optimization that returns symbolic solutions in terms of the original parameters.
- **Objective function.** The function to maximize or minimize in an optimization problem.

A.3 Causal terms

- **Always-taker.** See principal strata.
- **Ancestor.** Given a vertex V in a graph \mathcal{G} containing directed edges, an ancestor of V is either V itself or any vertex W with a directed path to V .
- **Blocking.** A disturbance U_k is blocked from a main variable V_j by an intervention set \mathbf{A} if all directed paths from U_k to V_j pass through \mathbf{A} .
- **C-component.** A synonym for “district,” defined below.

- **Canonical model.** A canonical model is a probabilistic model where the stochastic part of each variable is partitioned into partially applied response functions, or principal strata, w.l.o.g. For example, canonicalization of $X \rightarrow Y$ indicates transformation into an alternative model where the stochastic part of X is partitioned into $\Pr(X = 0)$ and $\Pr(X = 1)$ and of Y into $\Pr[Y(X = 0) = 0, Y(X = 1) = 0]$, $\Pr[Y(X = 0) = 1, Y(X = 1) = 0]$, $\Pr[Y(X = 0) = 0, Y(X = 1) = 1]$, and $\Pr[Y(X = 0) = 1, Y(X = 1) = 1]$
- **Canonical partition.** See canonical model.
- **Complier.** See principal strata.
- **Cross-world (event or distribution).** An event or distribution involving counterfactual values of variables under an inconsistent set of interventions, such as $\Pr[Y(X = 0) = 0, Y(X = 1) = 0]$. See also single-world marginal distribution.
- **Defier.** See principal strata.
- **Directed path (from W to V).** A sequence of connected directed edges exclusively pointing away from W and toward V .
- **District.** Given a graph \mathcal{G} containing main variables \mathbf{V} and disturbances \mathbf{U} , and a modified graph \mathcal{G}' in which all edges from any V_j to any other $V_{j'}$ are deleted, a district is a maximal connected set of vertices in \mathcal{G}' . In other words, a district consists of a maximal set of vertices forming a spanning tree in \mathcal{G}' . Districts always form a partition of vertices $\mathbf{V} \cup \mathbf{U}$. A synonym for “C-component.”
- **Full data law.** In causal or missing-data models, the joint distribution over all factual and counterfactual variables relevant to the problem. This is contrasted with the observed data distribution, which is a joint distribution over only observed variables. For example, given binary $X \rightarrow Y$, the full data law consists of $\Pr[X = x, Y(X = 0) = y, Y(X = 1) = y']$. The observed data law is $\Pr(X = x, Y = y)$.
- **Geared graph.** A geared canonical graph \mathcal{G} is a canonical graph that satisfies the running intersection property, defined below. Loosely speaking, geared graphs are those that lack particular kinds of cyclical confounding. In Figure 2, panels (a–b) are geared, whereas panel (c) is non-geared because it contains the cycle $U_{13} \rightarrow V_1 \leftarrow U_{12} \rightarrow V_2 \leftarrow U_{23} \rightarrow V_3 \leftarrow U_{13}$. In the present context, the running intersection property requires that there exists a total ordering of disturbances such that if a disturbance U_k touches any main variables that are also touched by any earlier disturbance in the ordering—letting $\mathbf{V}^{\text{intersection}} \subseteq \mathbf{V}$ denote these main variables—then the entire collection $\mathbf{V}^{\text{intersection}}$ can be influenced by at most one prior disturbance. That is, the main variables touched by U_k can at most overlap with those touched by only one additional $U_{k'} < U_k$. The existence of this ordering is what allows the cycle of confounding to be broken and a simple principal stratification to be constructed.

For example, in Figure 2(c), if the ordering is $U_{13} < U_{12} < U_{23}$, then U_{23} touches both V_2 and V_3 . Thus, V_2 and V_3 together can be influenced by at most one additional disturbance that is earlier in the ordering. This is not the case, because U_{12} touches V_2 , U_{13} touches V_3 , and both U_{12} and U_{13} are prior to U_{23} in the ordering; thus, the children of U_{23} are influenced by multiple prior disturbances. Furthermore, there exists no other ordering that satisfies the requirement, so Figure 2(c) is non-geared.
- **Generalized equality constraints.** See Verma constraints.

- **Instrumental inequality.** A restriction imposed on the observed marginal distribution by a hidden-variable DAG model corresponding to the instrumental variable model, shown in Figure 5(b). Valid instruments must satisfy this inequality. This inequality may be stated as follows: $\max_x \sum_y [\max_z \Pr(X = x, Y = y \mid Z = z)] \leq 1$.
- **Nested Markov parameterization.** An alternative representation of the observed-data distribution from a DAG with hidden disturbances. This alternative representation is specifically designed to eliminate redundant information. An example of redundancy is providing both $\Pr(A_1 = 0)$ and $\Pr(A_1 = 1)$ for binary A_1 ; because the second piece of information is already implied by the fact that A_1 is binary, it can be omitted from the polynomial program w.l.o.g. The nested Markov parameterization is obtained by factorizing the observed-data distribution, such as the example $\Pr(A_1 = a_1, L_1 = l_1, A_2 = a_2, L_2 = l_2)$ of Appendix C.4, into a minimal number of observed quantities such as $\Pr(A_1 = 0)$, $\Pr(L_1 = 0 \mid A_1 = a_1)$, $\sum_{a_1} \Pr(A_2 = 0 \mid L_1 = l_1, A_1 = 0) \Pr(A_1 = 0)$, and so on. When complete, this factorization can be used to fully reconstruct the original joint distribution of the observed data.

The factorization is based on the Möbius parameterization of the nested Markov model. These parameterizations are known for nested Markov models for categorical and Gaussian data. The procedure for obtaining the nested Markov factorization when all variables are binary is described in Appendix C.4. A detailed illustration of its use is provided in Appendix C.3.

- **Never-taker.** See principal strata.
- **Non-geared graph.** Any graph that is not geared. See geared graph.
- **Nonrestrictive disturbance cardinality.** Consider a disturbance U_{12} with unrestricted sample space $\mathcal{S}(U_{12})$ that influences main variables V_1 and V_2 . A core goal of this paper is to show that $\mathcal{S}(U_{12})$ can be partitioned according to the principal stratum assignments for V_1 and V_2 that it implies, or equivalently, that U_{12} can be reduced into a categorical random variable with categories corresponding to these partitions. A nonrestrictive disturbance cardinality is a number of partitions or categories such that this transformation of U_{12} is w.l.o.g. for the resulting joint distribution of factual and counterfactual values for V_1 and V_2 . For example, if V_1 has two principal strata and V_2 has four, U_{12} must have a cardinality of eight in order to reproduce any possible joint distribution of principal strata for V_1 and V_2 .
- **Parent.** Given a vertex V in a graph \mathcal{G} containing directed edges, a parent of V is any vertex W with a directed edge from W to V .
- **Partial application.** For a function $f(x, y)$, the partial application of a parameter x fixes one input and yields an alternative function $f^{(x)}(y)$ that maps the reduced domain y to the range of the original $f(x, y)$. For example, if $f(x, y) = xy$, then the partial application of $x = 2$ yields $f^{(x=2)}(y) = 2y$.
- **Principal strata.** Classifications of units on the basis of their joint counterfactual responses. For example, given a single treatment binary treatment X and a single binary response Y , there are four such counterfactual responses:
 - Never-takers: $Y(X = 0) = 0, Y(X = 1) = 0$.
 - Defiers: $Y(X = 0) = 1, Y(X = 1) = 0$.

- Compliers: $Y(X = 0) = 0, Y(X = 1) = 1$.
- Always-takers: $Y(X = 0) = 1, Y(X = 1) = 1$.

These names were given in the context of analysis of compliance in randomized trials, where X represents assignment of an individual to either treatment or control group, and Y represents whether the person in fact took the treatment or not. Note that since these four groups represent joint counterfactual responses, membership of particular individuals in these groups is not identifiable from observed data in typical settings.

In the presence of multiple treatments, principal strata may be appropriately generalized into a set consisting of all possible joint counterfactual responses. Principal strata have many applications in causal modeling. In the context of this paper, principal strata or their generalizations may be used to formulate the minimum cardinality of hidden disturbances in a causal model that does not impose any restrictions on any observed margin of the model.

- **Response function.** A partially applied structural equation. Synonym for “principal strata.”
 - **Single-world marginal distribution.** A marginal of the full data distribution where every variable either is not affected by treatments, or is a response to a consistent assignment of values to treatments by means of an intervention operation. For example, in the binary instrumental variable model of Figure 5(b), the full data law over all factual and counterfactual variables is $\Pr[Z = z, X(Z = 0) = x, X(Z = 1) = x', Y(X = 0) = y, Y(X = 1) = y']$. Two single world marginal distributions are $\Pr[Z = z, Y(X = 0) = y]$ and $\Pr[Z = z, Y(X = 1) = y]$, both involving intervention on X . Note that both of these distributions involve either variables that occur prior to intervention, such as Z , or responses to a single intervention. The full data distribution itself is not a single world distribution, since it contains both $Y(X = 0)$ and $Y(X = 1)$, which are responses of Y to different, inconsistent value assignments to the treatment X .
 - **Structural causal model.** A causal model of a data-generating process. A structural causal model is a structural equation equipped with the additional assumption that all ϵ_{U_k} and ϵ_{V_j} terms are mutually independent. See also structural equation model.
 - **Structural equation model.** A causal model of a data-generating process. A structural equation model is a structured system composed of a set of variables $\mathbf{U} \cup \mathbf{V}$ in which every $U_k \in \mathbf{U}$ and every $V_j \in \mathbf{V}$ is determined from (i) values of its parents, another subset $\mathbf{pa}(U_k) \subseteq \mathbf{U} \cup \mathbf{V} \setminus U_k$ or $\mathbf{pa}(V_j) \subseteq \mathbf{U} \cup \mathbf{V} \setminus V_j$, and (ii) an additional ϵ_{U_k} or ϵ_{V_j} term. This determination is done by means of functions of the form $U_k = f_{U_k}[\mathbf{pa}(U_k), \epsilon_{U_k}]$ and $V_j = f_{V_j}[\mathbf{pa}(V_j), \epsilon_{V_j}]$. The functions $f_{U_k}(\cdot)$ and $f_{V_j}(\cdot)$ are called *structural equations*. It is assumed that some of these functions may be altered to instead be constant functions, outputting values v_j for V_j . Such a replacement operation is called an intervention, or manipulation.
- Many structural models are assumed to be acyclic, meaning that there exists a total order on \mathbf{V} such that for every $V_j \in \mathbf{V}$, no variable later in the order than V_j takes part in $\mathbf{pa}(V_j)$.
- **Verma constraints.** A synonym for *generalized independence constraints*. These constraints are additional equality relationships that can implied by hidden-variable DAG models, beyond the conditional independences encoded in the DAG.

A standard example of a Verma constraint is that the function $\sum_{l_1} \Pr(L_2 = l_2 \mid A_2 = a_2, L_1 = l_1, A_1 = a_1) \Pr(L_1 = l_1 \mid A_1 = a_1)$ is not a function of A_1 if $\Pr(A_1 = a_1, L_1 = l_1, A_2 = a_2, L_2 = l_2)$ is a marginal distribution obtained from a joint that factorizes with respect to a DAG in Figure 11 (a). The nested Markov model is defined by conditional independences and Verma constraints.

B Algorithms

B.1 Algorithm 1: Canonicalization of DAGs, with discussion

In Algorithm 1, we formally state a procedure for obtaining a canonical hidden variable DAG (Definition 4.6 in Evans, 2016). Theorem 4.13 of the same work shows that the marginal model of any hidden variable DAG is the same as that of its canonical hidden variable DAG, and Proposition 7.4 shows that the same holds for the model for post-intervention distributions, when interventions are restricted to the main variables. For simplicity, we will use the generic X_j to refer to any node in a hidden-variable DAG \mathcal{G} , including both disturbances \mathbf{U} or main variables \mathbf{V} .

Before proceeding to Algorithm 1, we will state the canonicalization procedure informally and provide some intuition. Canonicalization proceeds in three steps. First, we take any indirect effect $X_j \rightarrow \dots \rightarrow X_{j'}$ that flows solely through hidden disturbances \mathbf{U} , then collapse it into the direct effect $X_j \rightarrow X_{j'}$. This is without consequence because by definition, analysts are not interested in reasoning about hidden disturbances—including how effects are mediated through them—except insofar as these disturbances affect the main variables. Second, we remove any effects that nodes have on disturbances. This can be done, even with complex networks of effects between disturbances, because the first step means that the role of any disturbance can be subsumed by its oldest ancestor. Thus, by construction, all latent variables in the resulting canonical DAG will be exogenous. Finally, as a consequence of the first and second steps, any disturbance can be eliminated without consequence if its role can be subsumed by another, more expressive disturbance.

Algorithm 1 DAG Canonicalization

Input: graph \mathcal{G}

Output: canonical graph \mathcal{G}'

Collapse effects that flow through disturbances

- 1: **for** variable pair $\{X_j, X_{j'}\} \in \mathbf{U} \cup \mathbf{V}$ **do**
- 2: **if** there exists a path $X_j \rightarrow \cdots \rightarrow X_{j'}$ such that all intermediate variables are disturbances
- 3: (e.g., $X_j \rightarrow U_k \rightarrow X_{j'}$, $X_j \rightarrow U_k \rightarrow U_{k'} \rightarrow X_{j'}$) **then**
 add edge $X_j \rightarrow X_{j'}$

Exogenize disturbances

- 4: **for** disturbance $U_k \in \mathbf{U}$ and node $X_j \in \mathbf{U} \cup \mathbf{V}$ **do**
- 5: **if** edge $X_j \rightarrow U_k$ exists **then**
 remove edge $X_j \rightarrow U_k$

Eliminate extraneous disturbances

- 6: **for** disturbance pair $\{U_k, U_{k'}\} \in \mathbf{U}$ **do**
 define children $\mathbf{ch}(U_k) = \{V_j : U_k \rightarrow V_j \text{ exists}\}$ and $\mathbf{ch}(U_{k'}) = \{V_j : U_{k'} \rightarrow V_j \text{ exists}\}$
- 7: **if** $\mathbf{ch}(U_k) \subset \mathbf{ch}(U_{k'})$ **then**
 delete U_k

Obtain canonicalized graph

- 8: **return** modified \mathcal{G}
-

B.2 Algorithm 2: Constructing polynomial programs

Algorithm 2 Constructing a polynomial program

Input: graph \mathcal{G} , evidence \mathcal{E} , assumptions \mathcal{A} , sample space $\mathcal{S}(\mathbf{V})$, target \mathcal{T}

Output: polynomial program in parameters $\mathcal{P}_{\mathbf{U}}$ or $\mathcal{P}_{\mathbf{U}} \cup s$

Initialization

initialize empty constraint set $\mathcal{C} \leftarrow \emptyset$

$\mathcal{G} \leftarrow$ canonicalize \mathcal{G}

$\mathcal{P}_{\mathbf{U}} \leftarrow$ parameters of functional model for \mathcal{G}

Polynomialize objective function

$\mathcal{T} \leftarrow$ polynomial-fractionalize(\mathcal{T})

- 1: **if** \mathcal{T} contains fractions **then**
 polynomialize($\mathcal{T} = s$) and append to \mathcal{C}
 $\mathcal{T} \leftarrow s$

Polynomialize constraints

- 2: **for** $(g(\mathcal{P}_{\mathbf{V}}) \star \alpha) \in (\mathcal{E} \cup \mathcal{A})$ **do**
 polynomialize($g(\mathcal{P}_{\mathbf{V}}) \star \alpha$) and append to \mathcal{C}
- 3: **for** $U_k \in \mathbf{U}$ **do**
 append (\mathcal{P}_{U_k} is a distribution) to \mathcal{C}

Optimize

- 4: **return** optimize \mathcal{T} subject to \mathcal{C}
-

B.3 Algorithm 3: Computing ε -sharp Bounds

Algorithm 3 Computing ε -sharp Bounds

Input: parameter space \mathcal{P} , target $\mathcal{T}(\mathbf{p})$ and constraint set $\{\mathcal{C}_\ell(\mathbf{p}) : \ell\}$ in parameters $\mathbf{p} \in \mathcal{P}$,
stopping thresholds $\varepsilon^{\text{thresh}}$ and θ^{thresh}

Output: lower bound \underline{D} , upper bound \overline{D} , maximum looseness factor ε

Initialization

Partitions $\underline{\mathcal{B}} \leftarrow \{\mathcal{P}\}$, $\overline{\mathcal{B}} \leftarrow \{\mathcal{P}\}$, initially with $b = 1$ branch each

Dual functions $\underline{\mathcal{D}}(\mathbf{p}) \leftarrow \{\underline{\mathcal{D}}_b(\mathbf{p}) = -\infty \text{ if } \mathbf{p} \in \underline{\mathcal{B}}_b : b\}$, $\overline{\mathcal{D}}(\mathbf{p}) \leftarrow \{\overline{\mathcal{D}}_b(\mathbf{p}) = +\infty \text{ if } \mathbf{p} \in \overline{\mathcal{B}}_b : b\}$

Primal bounds: $\underline{P} \leftarrow +\infty$, $\overline{P} \leftarrow -\infty$

Bounds width $\theta \leftarrow +\infty$, looseness factor $\varepsilon \leftarrow +\infty$

Spatial branch and bound

while $\varepsilon > \varepsilon^{\text{thresh}}$ **and** $\theta > \theta^{\text{thresh}}$ **do**

Create upper/lower dual functions and find extreme points in each branch

for each branch $\underline{\mathcal{B}}_b$ of the partition $\underline{\mathcal{B}}$ used for lower-bounding, **do**

Find relaxation parameters $\delta_0, \boldsymbol{\delta}$ s.t. $\delta_0 + \mathbf{p}^\top \boldsymbol{\delta} \leq \mathcal{T}(\mathbf{p})$ for all $\mathbf{p} \in \underline{\mathcal{B}}_b$ where $\mathcal{C}_\ell(\mathbf{p})$ is satisfied

Store this as the lower portion of the dual envelope for this branch, $\underline{\mathcal{D}}_b(\mathbf{p}) \leftarrow \delta_0 + \mathbf{p}^\top \boldsymbol{\delta}$

Identify the lowest point on the dual envelope in this branch, $\underline{D}_b \leftarrow \min \{\underline{\mathcal{D}}_b(\mathbf{p}) : \mathbf{p} \in \underline{\mathcal{B}}_b\}$

for each branch $\overline{\mathcal{B}}_b$ of the partition $\overline{\mathcal{B}}$ used for upper-bounding, **do**

Find relaxation parameters $\delta_0, \boldsymbol{\delta}$ s.t. $\delta_0 + \mathbf{p}^\top \boldsymbol{\delta} \geq \mathcal{T}(\mathbf{p})$ for all $\mathbf{p} \in \overline{\mathcal{B}}_b$ where $\mathcal{C}_\ell(\mathbf{p})$ is satisfied

Store this as the upper portion of the dual envelope for this branch, $\overline{\mathcal{D}}_b(\mathbf{p}) \leftarrow \delta_0 + \mathbf{p}^\top \boldsymbol{\delta}$

Identify the highest point on the dual envelope in this branch, $\overline{D}_b \leftarrow \max \{\overline{\mathcal{D}}_b(\mathbf{p}) : \mathbf{p} \in \overline{\mathcal{B}}_b\}$

Identify most extreme upper/lower branches, then update overall dual bounds on the estimand

Identify extreme branches $\underline{b} \leftarrow \arg \min_b \underline{D}_b$ and $\overline{b} \leftarrow \arg \max_b \overline{D}_b$

Update dual bounds $\underline{D} \leftarrow \underline{D}_{\underline{b}}$ and $\overline{D} \leftarrow \overline{D}_{\overline{b}}$

Update primal bounds

Reinitialize local minimization and maximization of $\mathcal{T}(\mathbf{p})$ within extreme branches $\underline{\mathcal{B}}_{\underline{b}}$ and $\overline{\mathcal{B}}_{\overline{b}}$,
subject to constraints $\mathcal{C}_\ell(\mathbf{p})$. Obtain \underline{P}' and \overline{P}' . Update $\underline{P} \leftarrow \min\{\underline{P}, \underline{P}'\}$ and $\overline{P} \leftarrow \max\{\overline{P}, \overline{P}'\}$.

Subdivide extreme branches

Remove lowest branch $\underline{\mathcal{B}}_{\underline{b}}$ from $\underline{\mathcal{B}}$ and remove highest branch $\overline{\mathcal{B}}_{\overline{b}}$ from $\overline{\mathcal{B}}$

Subpartition $\underline{\mathcal{B}}_{\underline{b}}$ into $\underline{\mathcal{B}}_{\underline{b}'}$, $\underline{\mathcal{B}}_{\underline{b}''}$ and subpartition $\overline{\mathcal{B}}_{\overline{b}}$ into $\overline{\mathcal{B}}_{\overline{b}'}$, $\overline{\mathcal{B}}_{\overline{b}''}$

Reinsert $\underline{\mathcal{B}}_{\underline{b}'}$, $\underline{\mathcal{B}}_{\underline{b}''}$ into $\underline{\mathcal{B}}$ and reinsert $\overline{\mathcal{B}}_{\overline{b}'}$, $\overline{\mathcal{B}}_{\overline{b}''}$ into $\overline{\mathcal{B}}$

Prune branches

for each branch $\underline{\mathcal{B}}_b$ in $\underline{\mathcal{B}}$ **do**

If $\underline{D}_b > \underline{P}$, the branch cannot lead to an improved lower bound; remove $\underline{\mathcal{B}}_b$ from $\underline{\mathcal{B}}$

If there exists no point $\mathbf{p} \in \underline{\mathcal{B}}_b$ satisfying $\mathcal{C}(\mathbf{p})$, remove $\underline{\mathcal{B}}_b$ from $\underline{\mathcal{B}}$

for each branch $\overline{\mathcal{B}}_b$ in $\overline{\mathcal{B}}$ **do**

If $\overline{D}_b < \overline{P}$, the branch cannot lead to an improved upper bound; remove $\overline{\mathcal{B}}_b$ from $\overline{\mathcal{B}}$

If there exists no point $\mathbf{p} \in \overline{\mathcal{B}}_b$ satisfying $\mathcal{C}(\mathbf{p})$, remove $\overline{\mathcal{B}}_b$ from $\overline{\mathcal{B}}$

Check progress

$\theta \leftarrow \overline{D} - \underline{D}$

$\varepsilon \leftarrow \theta / (\overline{P} - \underline{P}) - 1$

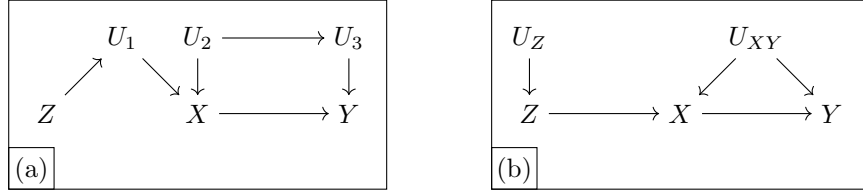
return \underline{D} , \overline{D} , ε

C Examples and detailed discussion

C.1 Step-by-step illustration of polynomial program construction

C.1.1 Step 1: Canonicalization

Panel (a) depicts a non-canonical DAG. Canonicalization proceeds by following Algorithm 1. First, a direct $Z \rightarrow X$ edge is added, collapsing the indirect $Z \rightarrow U_1 \rightarrow X$ effect; similarly, a direct $U_2 \rightarrow Y$ edge is added, collapsing $U_2 \rightarrow U_3 \rightarrow Y$. Second, the $Z \rightarrow U_1$ and $U_2 \rightarrow U_3$ edges are removed. Third, the U_1 and U_3 disturbances are eliminated, as they each affect a subset of the children of U_2 . Finally, for clarity, we rename the disturbance that confounds X and Y as U_{XY} , and we explicitly draw the previously implicit disturbance parent of Z , U_Z .



C.1.2 Step 2: Principal stratification

<u>Structural Eq.</u>	<u>Response Func.</u>	<u>Form</u>	<u>Possible Response Types</u>
$Z = f_Z(U_Z)$	$f_Z^{(U_Z=u_Z)}(\emptyset)$	$\emptyset \mapsto \{0, 1\}$	Z -control, Z -encourage
$X = f_X(Z, U_{XY})$	$f_X^{(U_{XY}=u_{XY})}(z)$	$\{0, 1\} \mapsto \{0, 1\}$	X -never, X -defy, X -comply, X -always
$Y = f_Y(X, U_{XY})$	$f_Y^{(U_{XY}=u_{XY})}(x)$	$\{0, 1\} \mapsto \{0, 1\}$	Y -never, Y -defy, Y -comply, Y -always

The U_Z disturbance is responsible for determining only the response function for Z . Because the structural equation for Z has no inputs besides the disturbance U_Z , supplying a disturbance value, $u_Z = Z$ -control or $u_Z = Z$ -encourage, will deterministically produce $Z = 0$ or $Z = 1$, respectively.

The U_{XY} disturbance is more complex because it determines the response functions for both X and Y simultaneously. We will first define the possible response functions. For clarity, we use “ V -never” and “ V -always” (where V stands in for either X or Y) to respectively

refer to response functions such that $f_V^{(U_k=u_k)}(a) = 0$ and $f_V^{(U_k=u_{k'})}(a) = 1$, i.e. never and always taking on a positive value. Similarly, “V-comply” and “V-defy” respectively refer to disturbance values $u_{k''}$ and $u_{k'''}$ which lead to the response functions $f_V^{(U_k=u_{k''})}(a) = a$ and $f_V^{(U_k=u_{k'''})}(a) = 1 - a$, i.e. taking on response values that comply with or defy the assigned treatment. Next, observe that U_{XY} can produce 16 possible joint response functions for X and Y : $\{X\text{-never}, Y\text{-never}\}$, $\{X\text{-never}, Y\text{-defy}\}$, $\{X\text{-defy}, Y\text{-never}\}$, and so on.

Our simplified program thus involves 18 disturbance values, which are in one-to-one correspondence with the 2 response-function probabilities $\Pr(Z\text{-type})$ for $\text{type} \in \{\text{control}, \text{encourage}\}$ and the 16 joint response-function probabilities $\Pr(X\text{-type}, Y\text{-type}')$ for $\text{type}, \text{type}' \in \{\text{never}, \text{defy}, \text{comply}, \text{always}\}$.

As Section D.3 notes, when parameterizing the problem, two terms can be immediately eliminated because $\Pr(Z\text{-control}) = 1 - \Pr(Z\text{-encourage})$ and, similarly, $\Pr(X\text{-never}, Y\text{-never})$ is unity less the sum of all other joint response-function probabilities for X and Y . For purposes of exposition, we will retain these superfluous parameters rather than eliminating them.

C.1.3 Step 3: Estimand polynomialization

Consider the local ATE among units that comply with encouragement to treatment uptake, i.e. units with $X(Z = 0) = 0$ and $X(Z = 1) = 1$. This estimand, $\mathbb{E}[Y(X = 1) - Y(X = 0) | X(Z = 0) = 0, X(Z = 1) = 1]$, a common quantity of interest in instrumental variable designs. However, if analysts cannot defend an assumption of monotonicity in $Z \rightarrow X$, this quantity cannot be point identified. Moreover, it is a nonlinear function of the principal strata, meaning that the approach of Balke and Pearl (1997) cannot be used. This estimand can be

polynomialized as follows:

$$\begin{aligned}
\mathcal{T} &= \mathbb{E}[Y(x=1) - Y(x=0) | X(z=1)=1, X(z=0)=0] \quad (\text{ATE among compliers}) \\
&= \frac{\Pr[Y(x=1)=1, X(z=1)=1, X(z=0)=0]}{\Pr[X(z=1)=1, X(z=0)=0]} \\
&\quad - \frac{\Pr[Y(x=0)=1, X(z=1)=1, X(z=0)=0]}{\Pr[X(z=1)=1, X(z=0)=0]} \\
&= \frac{\sum_{\{u_Z, u_{XY}\} \in \mathcal{S}(\mathbf{U})} \mathbb{1} \left\{ \begin{array}{l} f_Y^{(U_{XY}=u_{XY})}(x=1)=1, \\ f_X^{(U_{XY}=u_{XY})}(z=1)=1, \\ f_X^{(U_{XY}=u_{XY})}(z=0)=0 \end{array} \right\} \cdot \Pr(U_Z = u_Z) \cdot \Pr(U_{XY} = u_{XY})}{\sum_{\{u_Z, u_{XY}\} \in \mathcal{S}(\mathbf{U})} \mathbb{1} \left\{ \begin{array}{l} f_X^{(U_{XY}=u_{XY})}(z=1)=1, \\ f_X^{(U_{XY}=u_{XY})}(z=0)=0 \end{array} \right\} \cdot \Pr(U_Z = u_Z) \cdot \Pr(U_{XY} = u_{XY})} \\
&\quad - \frac{\sum_{\{u_Z, u_{XY}\} \in \mathcal{S}(\mathbf{U})} \mathbb{1} \left\{ \begin{array}{l} f_Y^{(U_{XY}=u_{XY})}(x=0 \text{ and})=1, \\ f_X^{(U_{XY}=u_{XY})}(z=1)=1, \\ f_X^{(U_{XY}=u_{XY})}(z=0)=0 \end{array} \right\} \cdot \Pr(U_Z = u_Z) \cdot \Pr(U_{XY} = u_{XY})}{\sum_{\{u_Z, u_{XY}\} \in \mathcal{S}(\mathbf{U})} \mathbb{1} \left\{ \begin{array}{l} f_X^{(U_{XY}=u_{XY})}(z=1)=1, \\ f_X^{(U_{XY}=u_{XY})}(z=0)=0 \end{array} \right\} \cdot \Pr(U_Z = u_Z) \cdot \Pr(U_{XY} = u_{XY})}
\end{aligned}$$

Next, note that in statements of the form $\Pr[Y(x)=y]$, U_Z is blocked from Y by the manipulation of x . Similarly, in statements of the form $\Pr[X(z)=x]$, U_Z is blocked from X by the manipulation of z . Therefore, by Proposition 3, U_Z can be eliminated from the expression for \mathcal{T} . This can be verified by observing that no term in the indicator function involves u_Z , so $\sum_{u_Z \in \mathcal{S}(U_Z)} \Pr(U_Z = u_Z) = 1$ can be factored out and eliminated.

$$\begin{aligned}
&\sum_{u_{XY} \in \mathcal{S}(U_{XY})} \mathbb{1} \left\{ \begin{array}{l} f_Y^{(U_{XY}=u_{XY})}(x=1)=1, \\ f_X^{(U_{XY}=u_{XY})}(z=1)=1, \\ f_X^{(U_{XY}=u_{XY})}(z=0)=0 \end{array} \right\} \cdot \Pr(U_{XY} = u_{XY}) \\
&= \frac{\sum_{u_{XY} \in \mathcal{S}(U_{XY})} \mathbb{1} \left\{ \begin{array}{l} f_X^{(U_{XY}=u_{XY})}(z=1)=1, \\ f_X^{(U_{XY}=u_{XY})}(z=0)=0 \end{array} \right\} \cdot \Pr(U_{XY} = u_{XY})}{\sum_{u_{XY} \in \mathcal{S}(U_{XY})} \mathbb{1} \left\{ \begin{array}{l} f_Y^{(U_{XY}=u_{XY})}(x=0)=1, \\ f_X^{(U_{XY}=u_{XY})}(z=1)=1, \\ f_X^{(U_{XY}=u_{XY})}(z=0)=0 \end{array} \right\} \cdot \Pr(U_{XY} = u_{XY})} \\
&\quad - \frac{\sum_{u_{XY} \in \mathcal{S}(U_{XY})} \mathbb{1} \left\{ \begin{array}{l} f_X^{(U_{XY}=u_{XY})}(z=1)=1, \\ f_X^{(U_{XY}=u_{XY})}(z=0)=0 \end{array} \right\} \cdot \Pr(U_{XY} = u_{XY})}{\sum_{u_{XY} \in \mathcal{S}(U_{XY})} \mathbb{1} \left\{ \begin{array}{l} f_X^{(U_{XY}=u_{XY})}(z=1)=1, \\ f_X^{(U_{XY}=u_{XY})}(z=0)=0 \end{array} \right\} \cdot \Pr(U_{XY} = u_{XY})}
\end{aligned}$$

We can now simplify by explicitly selecting the disturbance realizations that lead to the statements within the indicator functions being satisfied.

$$\begin{aligned}
&= \frac{\Pr(X\text{-comply}, Y\text{-comply}) + \Pr(X\text{-comply}, Y\text{-always})}{\sum_{\text{type}} \Pr(X\text{-comply}, Y\text{-type})} \\
&\quad - \frac{\Pr(X\text{-comply}, Y\text{-defy}) + \Pr(X\text{-comply}, Y\text{-always})}{\sum_{\text{type}} \Pr(X\text{-comply}, Y\text{-type})} \\
&= \frac{\Pr(X\text{-comply}, Y\text{-comply}) - \Pr(X\text{-comply}, Y\text{-defy})}{\sum_{\text{type}} \Pr(X\text{-comply}, Y\text{-type})}
\end{aligned}$$

Because this leads to a polynomial fraction, following Algorithm 2, we will define an auxiliary variable, s , and set it equal to \mathcal{T} .

$$s = \frac{\Pr(X\text{-comply}, Y\text{-comply}) - \Pr(X\text{-comply}, Y\text{-defy})}{\sum_{\text{type}} \Pr(X\text{-comply}, Y\text{-type})}$$

We can then manipulate to eliminate the fraction. After this step of Algorithm 2, we obtain a polynomial objective function, $\mathcal{T} = s$ (which is a monomial, and therefore a polynomial, in the expanded parameter space that includes s) and a single polynomial equality constraint that binds s to the target quantity,

$$\mathcal{C}_{\mathcal{T}} = \left\{ \begin{array}{l} \Pr(X\text{-comply}, Y\text{-comply}) - \Pr(X\text{-comply}, Y\text{-defy}) \\ -s \cdot \Pr(X\text{-comply}, Y\text{-never}) - s \cdot \Pr(X\text{-comply}, Y\text{-defy}) \\ -s \cdot \Pr(X\text{-comply}, Y\text{-comply}) - s \cdot \Pr(X\text{-comply}, Y\text{-always}) = 0 \end{array} \right\}$$

C.1.4 Step 4: Constraint polynomialization

Next, we turn to three types of information: (i) the first and second axioms of probability; (ii) modeling assumptions \mathcal{A} , such as the common “no defiers” assumption of Angrist et al. (1996), which states that encouragement would not lead to any unit rejecting treatment if they would otherwise have taken treatment under control; (iii) empirical evidence \mathcal{E} that correspond to observed quantities.

The axiomatic constraints are straightforwardly given by

$$\begin{aligned} \Pr(U_Z = u_Z) &\geq 0 \quad \forall \quad u_Z \in \mathcal{S}(U_Z) & \sum_{u_Z \in \mathcal{S}(U_Z)} \Pr(U_Z = u_Z) &= 1 \\ \Pr(U_{XY} = u_{XY}) &\geq 0 \quad \forall \quad u_{XY} \in \mathcal{S}(U_{XY}) & \sum_{u_{XY} \in \mathcal{S}(U_{XY})} \Pr(U_{XY} = u_{XY}) &= 1 \end{aligned}$$

or, equivalently,

$$\begin{aligned} \Pr(Z\text{-type}) &\geq 0 \quad \forall \quad \text{type} & \sum_{\text{type}} \Pr(Z\text{-type}) &= 1 \\ \Pr(X\text{-type}, Y\text{-type}') &\geq 0 \quad \forall \quad \text{type}, \text{type}' & \sum_{\text{type}, \text{type}'} \Pr(X\text{-type}, Y\text{-type}') &= 1 \end{aligned}$$

The first axiom translates into 18 inequality constraints, and the second into 2 equality constraints. We collect constraints arising from the laws of probability in $\mathcal{C}_{\mathcal{L}}$. As noted above, for simplicity of exposition, we do not exploit equality constraints to eliminate optimization variables and reduce the problem space. However, it is important to note that doing so can speed computation dramatically.

Next, we turn to the polynomialization of modeling assumptions \mathcal{A} . Formally, the “no defiers” assumption is that $\Pr[X(z = 1) < X(z = 0)] = 0$. Following Proposition 2, this is equivalent to

$$0 = \sum_{u_{XY} \in \mathcal{S}(U_{XY})} \mathbb{1} \left\{ \begin{array}{l} f_X^{(U_{XY}=u_{XY})}(z=1) = 0, \\ f_X^{(U_{XY}=u_{XY})}(z=0) = 1 \end{array} \right\} \cdot \Pr(U_{XY} = u_{XY})$$

where we again eliminate the blocked U_Z terms via Proposition 3. This is equivalent to

$$\mathcal{C}_{\mathcal{A}} = \left\{ \begin{array}{l} 0 = \Pr(X\text{-defy}, Y\text{-never}) + \Pr(X\text{-defy}, Y\text{-defy}) \\ \quad + \Pr(X\text{-defy}, Y\text{-comply}) + \Pr(X\text{-defy}, Y\text{-always}) \end{array} \right\},$$

It can be seen that, in conjunction with the first-axiom constraint, this implies

$$0 = \Pr(X\text{-defy}, Y\text{-type}) \quad \forall \quad \text{type}$$

which is, unsurprisingly, a literal statement of the no-defiers assumption on treatment response to encouragement. Finally, we polynomialize the empirical evidence \mathcal{E} . Each piece of evidence is one of eight observed probabilities of the form $\Pr(Z = z, X = x, Y = y)$. Each piece of evidence can be polynomialized by

$$\begin{aligned} & \Pr(Z = z, X = x, Y = y) \\ &= \sum_{\{u_Z, u_{XY}\} \in \mathcal{S}(\mathcal{U})} \mathbb{1} \left\{ \begin{array}{l} f_Z^{(U_Z=u_Z)}(\emptyset) = z, \\ f_X^{(U_{XY}=u_{XY})}(z) = x, \\ f_Y^{(U_{XY}=u_{XY})}(x) = y \end{array} \right\} \cdot \Pr(U_Z = u_Z) \cdot \Pr(U_{XY} = u_{XY}) \end{aligned}$$

We omit the translation to named strata types, which is straightforward but differs for each of the eight pieces of observed evidence. We note that, per Section [D.3](#)), these single-world marginal distributions can be simplified further. We do not pursue this simplification here, but an illustration of its use can be found in Appendix [C.3](#).

C.1.5 Example code

The following code from our `autobounds` software automatically implements the steps described above, obtaining 0.01-sharp bounds in 1.3 seconds.

```
# Algo 1, step 1 (App. B.4.1): provide canonicalized IV DAG
dag = DAG()
dag.from_structure("Z -> X, X -> Y, U -> X, U -> Y", unob = "U")

# Algo 1, step 2 (App. B.4.2): initialize program
# this automatically constructs generalized principal strata parameters
problem = causalProblem(dag) # Z, X, Y binary by default

# Algo 1, step 3 (App. B.4.3): state & automatically polynomialize estimand
problem.set_ate("X", "Y")

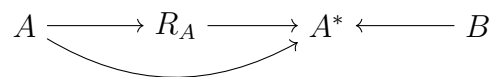
# Algo 1, step 4 (App. B.4.4): add & automatically polynomialize constraints
# this includes both empirical evidence and probability axiom constraints
problem.load_data(datafile, optimize = False) # no Sec. 5 simplifications
problem.add_prob_constraints()

# Algo 2 (Sec. 6): compile program, implement primal-dual optimization
program = problem.write_program()
program.run_couenne() # epsilon = 0.01 by default
```

C.2 Example of deterministic relationships

The general approach for obtaining generalized principal strata for discrete hidden variable DAGs (Evans, 2018; Finkelstein et al., 2021) does not take account of the kind of determinism introduced into the model by missingness indicators, and as such may yield more principal-strata parameters than are strictly needed. Due to the complexity of polynomial programming, it is beneficial to avoid excess parameters where possible. We now briefly explore this issue.

Figure 9: **A graph with determinism.**



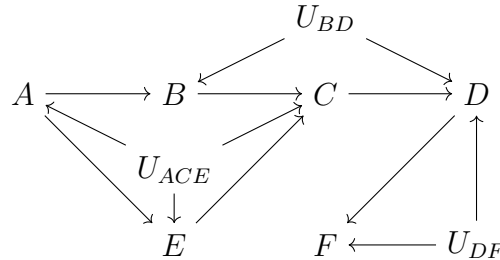
Consider the scenario depicted in Figure 9. In this graph, A^* is a proxy for the unobserved variable A , which is observed with missingness as indicated by R_A . When $R_A = 0$, then A^*

is deterministically equal to a special value indicating missingness (usually denoted with the special value such as “?” or “NA”). In addition, A^* is affected by B . This scenario might arise if A is measured with missingness *and* measurement error, and the nature of the error is affected by B . Of note, A^* is not a fully deterministic function of A and R_A , and cannot simply be removed from the functional parameterization, as in traditional missingness without measurement error. However, we can use the fact that it is a *partially* deterministic function of R_A to reduce the number of parameters needed in the functional model for this graph.

The standard principal stratification for this graph would allocate one value of ϵ_{A^*} —the exogenous noise that determines A^* in terms of its parents—for every combination of possible responses of A^* to its parents. Suppose A^* takes values in $\{0, 1, ?\}$, and A , R_A and B take values in $\{0, 1\}$. This would correspond to $3^8 = 6561$ possible values of ϵ_{A^*} . However, any such value that maps $R_A = 0$ to $A^* \in \{0, 1\}$ or $R_A = 1$ to $A^* = ?$ is ruled out by the deterministic relationship. As a result, ϵ_{A^*} need only specify the response of A^* in $\{0, 1\}$ to A and B when $R_A = 1$. This yields only $2^4 = 16$ possible values for ϵ_{A^*} . This example demonstrates that incorporating known deterministic relationships can yield a non-restrictive parameterization with fewer parameters.

C.3 Example of program simplification

Figure 10: A graph with conditional independence and Verma constraints.



Consider the graph presented in Figure 10. We will use this graph to illustrate a number of points raised in the main body of the paper. Suppose we are interested in the ATE of E on C . First, we will explicitly construct the generalized principal stratification of this graph, then use it to generate a simple polynomial program that bounds a causal target. Next, we will employ

several of the strategies described in Section 4.2 to simplify the program, demonstrating the importance of these strategies in obtaining tractable program formulations. Finally, we will observe that a broader class of partial identification problems than previously recognized can be formulated as linear programs.

Suppose all observed variables in the graph above are binary. In stratifying, we first note that U_{ACE} is responsible for determining the values of A , C and E in response to their parents. A has no parents, E has one parent, and C has two parents. Therefore U_{ACE} takes values in a state space of size $2^1 \times 2^2 \times 2^4 = 128$. Next, we suppose U_{BD} is responsible for determining the value of B in response to A , and therefore has size $2^2 = 4$. U_{DF} is left to determine the value of F in response to D , and of D in response to U_{BD} and C . It therefore takes values in space of size $2^8 \times 2^2 = 1024$.²⁴

To construct the polynomial program, we begin with the non-negativity and linear marginalization constraints on the parameters of the distributions of the disturbances (for simplicity, we abstain from eliminating one parameter per distribution using the sum-to-unity constraint):

$$\begin{aligned} \Pr(U_k = u_k) &\geq 0 && \text{for all } k \in \{ACE, BD, DE\} \text{ and } u_k \in \mathcal{S}(U_k) \\ \sum_{u_k \in \mathcal{S}(U_k)} \Pr(U_k = u_k) &= 1 && \text{for all } k \in \{ACE, BD, DE\}. \end{aligned}$$

We then add constraints encoding the empirical evidence \mathcal{E} . For simplicity, we assume that we observe the full joint distribution $\Pr(A = a, B = b, C = c, D = d, E = e, F = f)$, which is a vector of size $2^6 = 64$, corresponding to 64 equality constraints in the program. There are 3 disturbance variables in this graph, leading to polynomials in these equality constraints with terms of degree 3. Given the cardinalities of the disturbances, there are $2^4 \times 2^7 \times 2^{10} = 2,097,152$ possible combinations of disturbance assignments. By a simple exchangeability argument, the same number of possible combinations lead to each outcome in the state space. As there are 2^6 outcomes, each of the 64 polynomial equality constraints for \mathcal{E} will have $\frac{2^{21}}{2^6} = 2^{15}$ terms,

²⁴It is also possible stratify by first taking U_{DF} to be responsible for determining F in response to D , and then U_{BD} to be responsible for determining B in response to A and D in response to C and U_{DF} . By a simple symmetry argument, the two approaches yield the same number of parameters.

each with the following form, each of degree 3. This is a very large program.

$$\begin{aligned} & \Pr(A = a, B = b, C = c, D = d, E = e, F = f) \\ &= \sum_{\substack{u_{ACE} \in \\ \mathcal{S}(U_{ACE})}} \sum_{\substack{u_{BD} \in \\ \mathcal{S}(U_{BD})}} \sum_{\substack{u_{DF} \in \\ \mathcal{S}(U_{DF})}} \Pr(U_{ACE} = u_{ACE}) \Pr(U_{BD} = u_{BD}) \Pr(U_{DF} = u_{DF}) \times \mathbb{1}\{u_{ACE}, u_{BD}, u_{DF} \Rightarrow a, b, c, d, e, f\} \end{aligned}$$

where $\mathbb{1}\{\mathbf{u} \Rightarrow \mathbf{v}\}$ is an indicator function applied to realizations $\mathbf{u} = \{u_{ACE}, u_{BD}, u_{DF}\}$ and $\mathbf{v} = \{a, b, c, d, e, f\}$ of random variables \mathbf{U} and \mathbf{V} , which evaluates to 1 if and only if the values \mathbf{u} deterministically imply—together with the structural equations that are omitted for compactness—that $\mathbf{V} = \{A, B, C, D, E, F\}$ will attain values $\mathbf{v} = \{a, b, c, d, e, f\}$.

We now consider the strategies described in Section 4.2. First, observe that there are only 31 nested Markov quantities for this graph, corresponding to 31 polynomial equality constraints encoding \mathcal{E} : a substantial savings over the 64 quantities of the naïve approach. These quantities are listed below.

$$\begin{aligned} \theta_A &\equiv \Pr(A = 0) \\ \theta_B(a) &\equiv \Pr(B = 0 \mid A = a) \\ \theta_E(a) &\equiv \Pr(E = 0 \mid A = a) \\ \theta_C(a, b, e) &\equiv \Pr(C = 0 \mid A = 0, B = b, E = e) \\ \theta_{\{B, D\}}(a, c) &\equiv \Pr(D = 0 \mid C = c, B = 0, A = a) \Pr(B = 0 \mid A = a) \\ \theta_D(c) &\equiv \sum_b \Pr(D = 0 \mid C = c, B = b, A = a) \Pr(B = b \mid A = a) \\ \theta_{\{B, F\}}(a, c, d) &\equiv \Pr(F = 0 \mid D = d, C = c, B = 0, A = a) \Pr(B = 0 \mid A = a) \\ \theta_F(d, c) &\equiv \frac{\sum_b \Pr(F = 0, D = d \mid C = c, B = b, A = a) \Pr(B = b \mid A = a)}{\sum_b \Pr(D = d \mid C = c, B = b, A = a) \Pr(B = b \mid A = a)} \end{aligned}$$

For a review of nested Markov parameterizations for binary models, see Appendix C.4.

This reduced parameterization is possible because it encodes standard conditional independences, such as $F \perp A \mid D$. In addition, it encodes Verma constraints, which emerge either (i) from independences in post-intervention distributions or (ii) from the irrelevance of

an intervention to a particular distribution. In this case, $A(C = c) \perp \{D(C = c), F(C = c)\}$. As discussed in the main text, each equality constraint can be used to reduce the number of parameters needed in a non-restrictive reduction that can express every possible distribution in the model.

Recall that each nested Markov parameter corresponds to the identified probability of a single-world event, where the event is specified in terms of variables in a single district, and the intervention is on all parents of the district relevant to those variables. For example, in this case, one of the nested Markov parameters is $\Pr [B(A = 1, C = 1) = 1, F(A = 1, C = 1) = 1 | D = 1]$. We can now make use of Proposition 3 to reason that each of these polynomial constraints must involve only disturbances from a single district. Therefore in the equations corresponding to nested Markov parameters for the district corresponding to U_{ACE} , parameters of the distributions of U_{BD} and U_{DF} can all be factored out as terms that will sum to unity, meaning we will be left with equations that are linear in the parameters of U_{ACE} . Likewise, in equations corresponding to nested Markov parameters for the district containing descendants of U_{BD} and U_{DF} , parameters for the distribution of U_{ACE} will factor out, and we will be left with a quadratic equation.

Finally, we can make use of Proposition 4 to note that constraints involving nested Markov parameters corresponding to the $\{U_{BD}, U_{DF}\}$ district can be dropped from the program. This is because they only involve parameters for the distributions of U_{BD} and U_{DF} , which do not appear in any constraint involving parameters for the distribution of U_{ACE} . The target, by contrast, involves only parameters for the distribution of U_{ACE} .

As a result of taking the three steps described in Section 4.2, we have taken this problem from a *polynomial* program involving 1156 parameters to a *linear* program involving only $2^7 = 128$ parameters and fewer constraints. This example also motivates the following corollary, which expands the class of partial identification problems that can be formulated as linear programs relative to known results (Balke and Pearl, 1997; Finkelstein et al., 2020; Wolfe et al., 2019).

Corollary 2. *Suppose \mathcal{G} is a hidden variable DAG with observed variables \mathbf{V} , $\mathcal{C}_\ell = \{V_\ell(\mathbf{a}_\ell) =$*

$v_\ell\}$ are counterfactual statements indexed by $\ell \in \mathcal{L}$, and $\Pr(\bigcap_\ell \mathcal{C}_\ell)$ is the target of interest. Further suppose that the full joint distribution $\Pr(\mathbf{V} = \mathbf{v})$ is observed. Then $\Pr(\bigcap_\ell \mathcal{C}_\ell)$ can be sharply bounded given the observed data by optimizing a linear program if all $\{V_\ell : \ell \in \mathcal{L}\}$ are in the same single-latent-variable district.

Proof. Because the common district of \mathcal{C} contains only a single latent variable, by Proposition 3 the objective will be linear in the parameters of the distribution of that latent variable. By Proposition 4, the constraints will not involve parameters corresponding to other districts. By Algorithm 2, no single term in a constraint will involve multiple parameters for the same latent distribution, meaning that all constraints involving only parameters corresponding to a single-variable district will be linear. The non-negativity and sum-to-unity constraints on the parameters of the latent-variable distribution are also linear. It follows that the objective and all constraints are linear. \square

C.4 Discussion of Nested Markov models

The nested Markov model is a set of distributions $\Pr(\mathbf{V})$ associated with an acyclic directed mixed graph (ADMG) \mathcal{G} . This model is notable in the study in hidden variable DAGs because it has been shown (Evans, 2018) that given a set of distributions $\Pr(\mathbf{V} \cup \mathbf{H})$ that factorize with respect to a DAG with vertices $\mathbf{V} \cup \mathbf{H}$, the nested Markov model associated with a latent projection $\mathcal{G}(\mathbf{V})$ of the DAG captures all equality constraints implied by this factorization on the marginal distribution $\Pr(\mathbf{V})$. To define this model, we will need to introduce a number of definitions.

A conditional ADMG (CADMG) is a graph $\mathcal{G}(\mathbf{V}, \mathbf{W})$ with random and fixed vertices with directed and bidirected edges, with the property that no edge may have an arrowhead into an element of \mathbf{W} , and no directed cycles exist. Note that an ADMG is a special case of a CADMG where \mathbf{W} is empty. The notion of a district (bidirected connected set) generalizes to CADMGs, but only applies to elements in \mathbf{V} .

A Markov kernel $q_{\mathbf{V}}(\mathbf{V}|\mathbf{W})$ is a mapping from values of \mathbf{W} to distributions over \mathbf{V} . While conditional distributions are Markov kernels, not all Markov kernels are conditional distribu-

tions. For example, interventional distributions arising in causal inference are Markov kernels, but are generally not conditional distributions. Marginalization and conditioning are defined in the same way for Markov kernels as for conditional distributions. That is, for any $\mathbf{A} \subset \mathbf{V}$,

$$q_{\mathbf{V}}(\mathbf{A}|\mathbf{W}) = \sum_{\mathbf{V} \setminus \mathbf{A}} q_{\mathbf{V}}(\mathbf{V}|\mathbf{W})$$

$$q_{\mathbf{V}}(\mathbf{V} \setminus \mathbf{A}|\mathbf{A} \cup \mathbf{W}) = \frac{q_{\mathbf{V}}(\mathbf{V}|\mathbf{W})}{q_{\mathbf{V}}(\mathbf{A}|\mathbf{W})}.$$

Given a CADMG $\mathcal{G}(\mathbf{V}, \mathbf{W})$, a variable $V \in \mathbf{V}$ is said to be *fixable* if there does not exist a variable $Z \in \mathbf{V}$ such that $Z \neq V$, and which is a descendant of V and lies in the same district as V . Given V fixable in $\mathcal{G}(\mathbf{V}, \mathbf{W})$, define the fixing operator $\phi_V(\mathcal{G})$ that outputs a new CADMG $\mathcal{G}(V \setminus \{V\}, \mathbf{W} \cup \{V\})$ which inherits all vertices and edges from $\mathcal{G}(\mathbf{V}, \mathbf{W})$ with the following two exceptions. First, V is treated as a fixed variable, and second all edges with an arrowhead into V are removed.

Given a pair of a CADMG $\mathcal{G}(\mathbf{V}, \mathbf{W})$ and kernel $q_{\mathbf{V}}(\mathbf{V}|\mathbf{W})$, if V is fixable in $\mathcal{G}(\mathbf{V}, \mathbf{W})$, we define the fixing operator $\phi_V(q_{\mathbf{V}}; \mathcal{G})$ that yields the following kernel:

$$q_{\mathbf{V} \setminus \{V\}}(\mathbf{V} \setminus \{V\}|\{V\} \cup \mathbf{W}) \equiv \frac{q_{\mathbf{V}}(\mathbf{V}|\mathbf{W})}{q_{\mathbf{V}}(V|\text{mb}_{\mathcal{G}}(V) \cup \mathbf{W})},$$

where $\text{mb}_{\mathcal{G}}(V)$, the *Markov blanket* of V in \mathcal{G} is defined to be the district of V , along with all parent variables of this district (this may include elements of \mathbf{W}).

A sequence V_1, V_2, \dots is said to be fixable in a CADMG $\mathcal{G}(\mathbf{V}, \mathbf{W})$ if either it is the empty sequence, or V_1 is fixable in $\mathcal{G}(\mathbf{V}, \mathbf{W})$, and V_2, \dots is fixable in $\phi_{V_1}(\mathcal{G}(\mathbf{V}, \mathbf{W}))$. We inductively define the following natural extensions of the fixing operator to sequences, as follows:

$$\begin{aligned} \phi(\mathcal{G}) &\equiv \mathcal{G} \\ \phi_{V_1, V_2, \dots}(\mathcal{G}) &\equiv \phi_{V_2, \dots}(\phi_{V_1}(\mathcal{G})) \\ \phi(q_{\mathbf{V}}; \mathcal{G}) &\equiv q_{\mathbf{V}} \\ \phi_{V_1, V_2, \dots}(q_{\mathbf{V}}; \mathcal{G}) &\equiv \phi_{V_2, \dots}(\phi_{V_1}(q_{\mathbf{V}}; \mathcal{G}); \phi_{V_1}(\mathcal{G})) \end{aligned}$$

For graphs any fixable sequence for a set \mathbf{S} yields the same result, thus we define the operator $\phi_{\mathbf{S}}(\mathcal{G})$ w.l.o.g. to mean “apply the fixing operator in order according to any fixing sequence for elements in \mathbf{S} .”

A set \mathbf{R} is said to be *reachable* in $\mathcal{G}(\mathbf{V}, \mathbf{W})$ if there exists a fixable sequence for $\mathbf{V} \setminus \mathbf{R}$ in $\mathcal{G}(\mathbf{V}, \mathbf{W})$. A reachable set \mathbf{R} is called *intrinsic* in $\mathcal{G}(\mathbf{V}, \mathbf{W})$ if $\phi_{\mathbf{V} \setminus \mathbf{R}}(\mathcal{G}(\mathbf{V}, \mathbf{W}))$ has a single district containing all elements of \mathbf{R} .

Given a CADMG $\mathcal{G}(\mathbf{V}, \mathbf{W})$, define an ADMG $\mathcal{G}^{\mathbf{W}}$ to be a graph which contains vertices $\mathbf{V} \cup \mathbf{W}$ and all edges in $\mathcal{G}(\mathbf{V}, \mathbf{W})$, and in addition a bidirected edge between any pair of vertices in \mathbf{W} .

The nested Markov model may be (equivalently) defined by means of a factorization, or a global Markov or local Markov properties. We reproduce the global Markov property here, with the other model definitions, along with extensive discussion, may be found in (Richardson et al., 2017).

The global Markov property is defined using m-separation, a generalization of d-separation applicable to graphs with both directed and bidirected edges. Like d-separation, a vertex set \mathbf{A} is said to be m-separated from the vertex set \mathbf{B} given a vertex set \mathbf{C} if all paths from \mathbf{A} to \mathbf{B} are “blocked” by \mathbf{C} . A path is considered blocked if any consecutive triplet of vertices on the path is blocked. Any non-collider triplet is blocked if the middle vertex is in \mathbf{C} . Any collider triplet is blocked if neither the middle vertex, nor any descendant of this vertex, is in \mathbf{C} . In mixed graphs like ADMGs, colliders may be formed using either directed or bidirected edges. M-separation is discussed in more detail in (Richardson, 2003). A distribution $\Pr(\mathbf{V})$ is said to obey the m-separation criterion in an ADMG \mathcal{G} if whenever \mathbf{A} is m-separated from \mathbf{B} given \mathbf{C} in \mathcal{G} , then \mathbf{A} is conditionally independent of \mathbf{B} given \mathbf{C} in $\Pr(\mathbf{V})$.

A distribution $\Pr(\mathbf{V})$ is said to obey the global nested Markov property with respect to an ADMG \mathcal{G} if for every reachable set \mathbf{R} , the kernel $\phi_{\mathbf{V} \setminus \mathbf{R}}(\Pr(\mathbf{V}); \mathcal{G})$ obeys the m-separation criterion with respect to $\mathcal{G}^{\mathbf{V} \setminus \mathbf{R}}$ obtained from $\phi_{\mathbf{V} \setminus \mathbf{R}}(\mathcal{G})$.

It is known that if $\Pr(\mathbf{V})$ is nested Markov with respect to an ADMG \mathcal{G} , then given a reachable set \mathbf{R} , any fixable sequence for $\mathbf{V} \setminus \mathbf{R}$ yields the same kernel if applied to $\Pr(\mathbf{V})$ and \mathcal{G} . Thus, we will define $\phi_{\mathbf{V} \setminus \mathbf{R}}(\Pr(\mathbf{V}); \mathcal{G})$ for any \mathbf{R} reachable in \mathcal{G} , w.l.o.g. for distribution

$\Pr(\mathbf{V})$ in the nested Markov model, to mean “apply the fixing operator to any fixable sequence for $\mathbf{V} \setminus \mathbf{R}$.”

Parameterizations for distributions $\Pr(\mathbf{V})$ nested Markov with respect to an ADMG \mathcal{G} have been derived for multivariate normal (Shpitser et al., 2018), and categorical (Evans and Richardson, 2019) data.

Here we describe the parameterization for binary nested Markov models. More details may be found in (Evans and Richardson, 2019).

Given an ADMG \mathcal{G} , let $\mathcal{I}(\mathcal{G})$ be the set of all intrinsic sets in \mathcal{G} . For each such set \mathbf{S} , define $\text{head}(\mathbf{S})$ to be the subset of \mathbf{S} with no children in $\phi_{\mathbf{V} \setminus \mathbf{S}}(\mathcal{G})$. Similarly, define $\text{tail}(\mathbf{S})$ to be the set of parents of \mathbf{S} in \mathcal{G} . Note that the head and the tail for each $\mathbf{S} \in \mathcal{I}(\mathcal{G})$ are disjoint.

The parameterization of the binary nested Model is given by the set of parameters of the form

$$\{\theta_{\mathbf{S}}(\mathbf{t}) : \mathbf{S} \in \mathcal{I}(\mathcal{G}), \mathbf{t} \text{ values of } \text{tail}(\mathbf{S})\}.$$

Each parameter $\theta_{\mathbf{S}}(\mathbf{t})$ is obtained from $q_{\mathbf{S}}(\mathbf{S} | \mathbf{V} \setminus \mathbf{S}) = \phi_{\mathbf{V} \setminus \mathbf{S}}(p(\mathbf{V}); \mathcal{G})$ as follows:

$$\theta_{\mathbf{S}}(\mathbf{t}) = \frac{q_{\mathbf{S}}(\mathbf{S} | \mathbf{V} \setminus \mathbf{S})}{\sum_{\text{head}(\mathbf{S})} q_{\mathbf{S}}(\mathbf{S} | \mathbf{V} \setminus \mathbf{S})} \Bigg|_{\text{head}(\mathbf{S})=0, \text{tail}(\mathbf{S})=\mathbf{t}}.$$

Conversely, the observed data distribution $\Pr(\mathbf{V})$ may be obtained from these parameters using the Möbius inversion formula, as described in (Evans and Richardson, 2019).

Consider a hidden variable DAG shown in Figure 11 (a) where variables H_1 and H_2 are unobserved, with the corresponding latent projection ADMG shown in Figure 11 (b). The global nested Markov property allows us to read off conditional independences from $\Pr(\mathbf{V} = \mathbf{v})$, where $\mathbf{V} = \{A_1, L_1, A_2, L_2\}$, and from Markov kernels associated with sets reachable in this ADMG, derived from $\Pr(\mathbf{V})$ by means of the fixing operator. For example, the set $\{L_2, L_1\}$ is reachable since the set $\mathbf{V} \setminus \{L_2, L_1\} = \{A_1, A_2\}$ is fixable in the sequence A_2, A_1 . Applying this fixing sequence to the ADMG in Figure 11 (b) yields the CADMG in Figure 11 (c), and applying this fixing sequence to the distribution $\Pr(\mathbf{V} = \mathbf{v})$ along with the graph in Figure 11

(b), yields the kernel $q_{\{L_1, L_2\}}(L_1, L_2 \mid A_1, A_2) = \Pr(L_2 \mid A_1, L_1, A_2) \Pr(L_1 \mid A_1)$.

Any m-separation in the graph $\mathcal{G}^{\{A_1, A_2\}}$, shown in Figure 11 (d), corresponds to a conditional independence that holds in the corresponding kernel. In particular, since L_2 is m-separated from A_1 given A_2 in Figure 11 (d), we conclude that $\sum_{L_1} \Pr(L_2 \mid A_1, L_1, A_2) \Pr(L_1 \mid A_1)$ is not a function of A_1 . This is an example of a generalized independence constraint or Verma constraint, a type of equality constraint that is captured by the global nested Markov property. Note that L_2 is not independent of A_1 conditional on A_2 in the original graph in Fig 11 (b), due to the existence of a path $A_1 \rightarrow L_1 \leftrightarrow L_2$ that is m-connected since a descendant A_2 of a collider at L_1 is conditioned on.

The binary parameterization of the nested Markov model associated with the graph in Figure 11 (b) is as follows:

$$\begin{aligned}\theta_{A_1} &\equiv \Pr(A_1 = 0) \\ \theta_{L_1}(a_1) &\equiv \Pr(L_1 = 0 \mid a_1) \\ \theta_{A_1, A_2}(l_1) &\equiv \Pr(A_2 = 0 \mid l_1, A_1 = 0) \Pr(A_1 = 0) \\ \theta_{A_2}(l_1) &\equiv \sum_{a_1} \Pr(A_2 = 0 \mid l_1, a_1) \Pr(a_1) \\ \theta_{L_2, L_1}(a_1, a_2) &\equiv \Pr(L_2 = 0 \mid L_1 = 0, a_1, a_2) \Pr(L_1 = 0 \mid a_1) \\ \theta_{L_2}(a_2) &\equiv \sum_{l_1} \Pr(L_2 = 0 \mid l_1, a_1, a_2) \Pr(l_1 \mid a_1).\end{aligned}$$

The total parameter count is $1 + 2 + 2 + 2 + 4 + 2 = 13$, which is two less than the saturated parameterization of a four variable binary model, which has dimension 15. The 2 missing parameters are due to the missing edge between A_1 and L_2 and are associated with the Verma constraint discussed above. Had that edge been present, the last parameter family $\theta_{L_2}(a_2)$ consisting of two parameters would have instead taken the form $\theta_{L_2}(a_2, a_1)$ for all values of a_1, a_2 , yielding 4 rather than 2 parameters.

These parameters are algebraic functions of the observed data distribution $\Pr(\mathbf{V} = \mathbf{v})$. Conversely the observed data distribution may be obtained from these parameters by means of the Möbius inversion formula, with the details given in (Evans and Richardson, 2019).

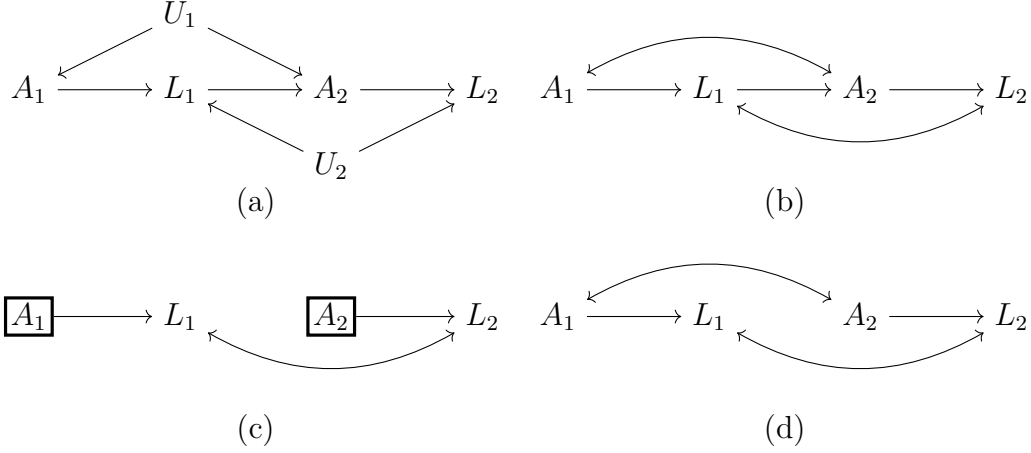
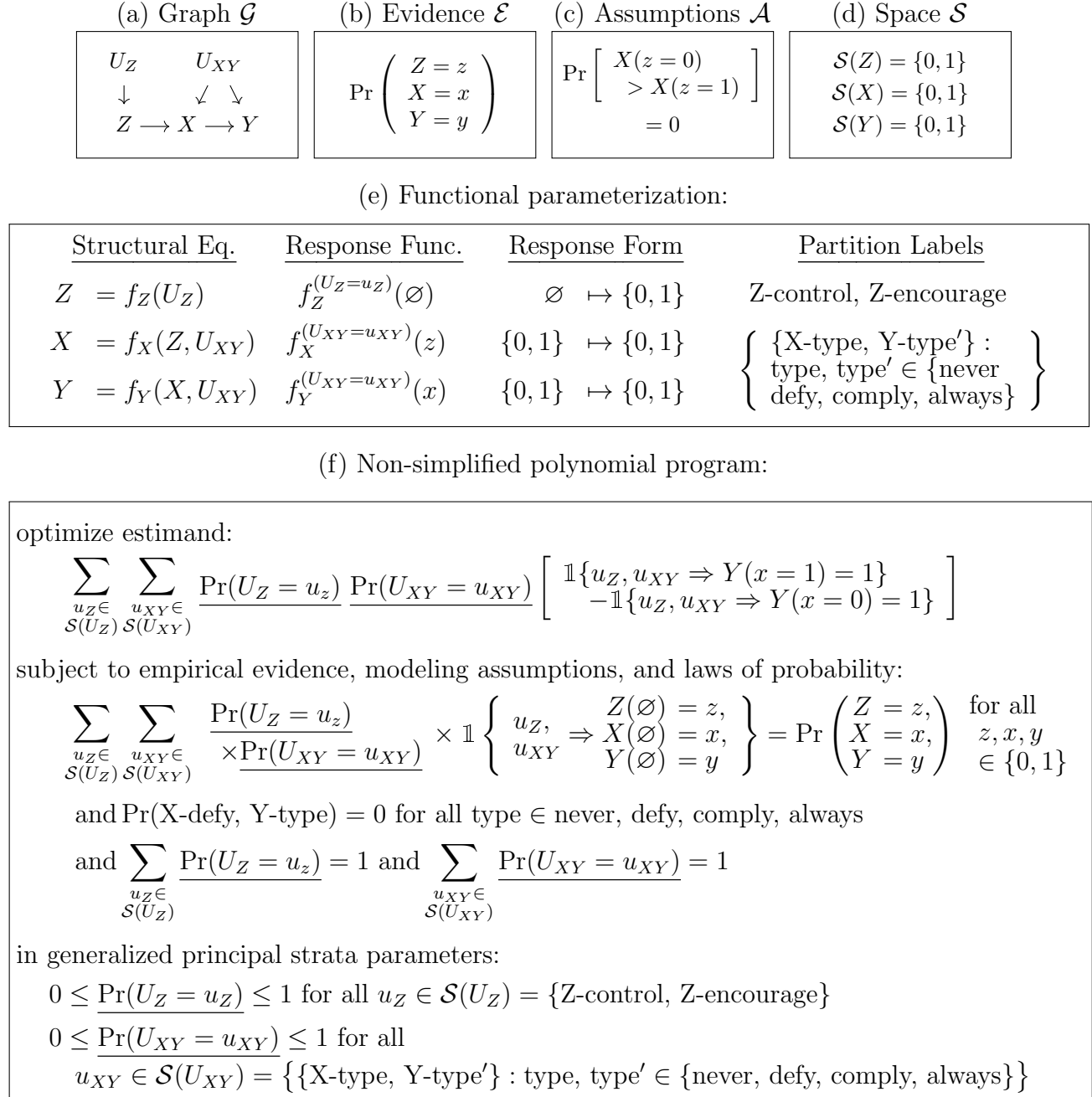


Figure 11: (a) A hidden variable DAG where only variables A_1, L_1, A_2 and L_2 are observed. (b) A latent projection mixed graph obtained from the DAG in (a). (c) A CADMG obtained from (b) by fixing A_2 . (d) The graph $\mathcal{G}^{\{A_1, A_2\}}$ obtained from (c) which is used to encode the global nested Markov property for the reachable set $\{L_1, L_2\}$.

D Details on program simplifications

To make the proposed simplifications more concrete, we will illustrate them in the context of the instrumental variable problem shown in Figure 12. As the figure shows, the problem involves a randomized encouragement, Z , with associated disturbance U_Z , and confounded treatment and outcome, X and Y , with shared disturbance U_{XY} . Applying Algorithm 2 yields an initial, non-simplified program in 18 parameters, each taking on possible values in $[0, 1]$: the 2 parameters of $\Pr(U_Z = u_Z)$, representing $\Pr(\text{Z-type})$, and the 16 parameters of $\Pr(U_{XY} = u_{XY})$, representing $\Pr(\text{X-type}, \text{Y-type}')$. We will suppose that the available empirical evidence \mathcal{E} consists of eight pieces of information representing the joint observational distribution $\Pr(Z = z, X = x, Y = y)$ for each $\{x, y, z\} \in \{0, 1\}^3$, producing eight constraints. We further suppose that the only modeling assumption in \mathcal{A} is the monotonicity or “no defiers” assumption on $Z \rightarrow X$, which translates into four additional constraints: $\Pr(\text{X-defy}, \text{Y-type}) = 0$ for each of the four principal strata of Y . Finally, both $\Pr(U_Z = u_Z)$ and $\Pr(U_{XY} = u_{XY})$ are constrained to sum to unity. Together, these comprise 14 constraints in the initial, non-simplified polynomial program.

Figure 12: **A polynomial program produced by Algorithm 2.** Construction of a polynomial program for an instrumental variable problem. Panels (a–d) depict inputs to the algorithm. The graph, \mathcal{G} , contains randomized encouragement, Z , as well as confounded treatment and outcome, X and Y . The evidence \mathcal{E} consists of the joint distribution of Z , X , and Y . \mathcal{A} consists of a monotonicity assumption for $Z \rightarrow X$. \mathcal{S} states that Z , X , and Y are binary. The target \mathcal{T} is the ATE $\mathbb{E}[Y(x=1) - Y(x=0)]$. Panel (e) depicts functional parameterization with 18 disturbance partitions, following Section 3.3. Panel (f) shows the polynomial program output by Algorithm 2, with optimization parameters indicated by underlining.



D.1 Reducing polynomial degree by exploiting graph structure

We begin by showing how the graphical structure of causal queries reveals simplifications that can be automatically detected and exploited. As an example, consider the polynomialized estimand of Figure 12(f), the ATE

$$\sum_{\substack{u_Z \in \\ \mathcal{S}(U_Z)}} \sum_{\substack{u_{XY} \in \\ \mathcal{S}(U_{XY})}} \Pr(U_Z = u_Z) \Pr(U_{XY} = u_{XY}) \left[\begin{array}{c} \mathbb{1}\{u_Z, u_{XY} \Rightarrow Y(x=1) = 1\} \\ - \mathbb{1}\{u_Z, u_{XY} \Rightarrow Y(x=0) = 1\} \end{array} \right].$$

The initial objective function produced by Algorithm 2 has this form because it was automatically generated according to Proposition 2. However, it can be seen that the $\sum_{u_Z \in \mathcal{S}(U_Z)} \Pr(U_Z = u_Z)$ term can be factored out of the expression and eliminated, since (i) neither $Y(x=1)$ nor $Y(x=0)$ are affected by the disturbance U_Z and (ii) the factored-out term is a sum over all possible disturbance realizations for U_Z and thus evaluates to unity. This simplification, and others like it, is easily implemented using symbolic algebra systems such as SageMath (Stein et al., 2019). This reduces the degree of the objective polynomial from quadratic to linear, which can greatly accelerate computation.

More generally, a disturbance U_k can be identified as irrelevant whenever the interventions of interest, $\mathbf{A} \subset \mathbf{V}$, *block* the relevant outcomes, $\mathbf{Y} \subset \mathbf{V}$ —meaning that all directed paths from U_k to \mathbf{Y} pass through \mathbf{A} . At a high level, Proposition 3 states that in any constraint or objective function in which this occurs, the parameters $\Pr(U_k = u_k)$ can always be factored out and eliminated.

Proposition 3. *Consider the polynomialization of a probability $\Pr\left(\bigcap_{\ell} \mathcal{C}_{\ell}\right)$ in which each $\mathcal{C}_{\ell} = \{V_{\ell}(\mathbf{a}_{\ell}) = v_{\ell}\}$. When U_k is blocked from V_{ℓ} by \mathbf{a}_{ℓ} for every ℓ , the parameters \mathcal{P}_{U_k} —representing the probabilities $\Pr(U_k = u_k)$ —can be eliminated from the polynomialization.*

The proposition provides some additional guidance on when this basic intuition can be extended to more complex scenarios involving multiple treatment and outcome sets. A proof and additional intuition is given in Appendix F.3.

D.2 Eliminating variables by solving equality constraints

Next, we make the somewhat obvious observation that equalities can be used to eliminate optimization variables and thus simplify the system of constraints. In the running IV example, consider the constraint $\Pr(\text{X-defy}, \text{Y-never}) = 0$ which arises from the monotonicity modeling assumption. The original polynomial program can be simplified by (i) deleting this constraint and (ii) replacing every occurrence of the $\Pr(\text{X-defy}, \text{Y-never})$ parameter with zero in every other constraint. The resulting simplified program is equivalent in that it produces exactly the same minimum and maximum, but more computationally efficient by virtue of having one less variable and one less constraint. The second-axiom constraint $\Pr(\text{Z-control}) + \Pr(\text{Z-encourage}) = 1$ can similarly be used to eliminate one variable and one constraint. Here, an important consideration is that in practice, the resulting program can be substantially more efficient when using a two-step simplification process. We recommend first using a symbolic algebra solver to factorize out the left-hand-side polynomial, such as $\Pr(\text{Z-control}) + \Pr(\text{Z-encourage})$, and replace it with the right-hand-side scalar (here, unity) where possible. In a clean-up step, any remaining occurrences of the variable can be eliminated—here, by substituting any leftover $\Pr(\text{Z-control})$ terms for $1 - \Pr(\text{Z-encourage})$.

A natural extension of this technique is to use the empirical constraints arising from the observed $\Pr(Z = z, X = x, Y = y)$ quantities to eliminate additional variables. However, we recommend first implementing the Section D.3 simplifications, which can substantially reorganize the empirical evidence, before using this information to eliminate variables.

D.3 Refactorizing empirical constraints to eliminate redundant information

In Section 4, we defined the empirical evidence \mathcal{E} in general terms, emphasizing how it can flexibly accommodate published summary statistics from administrative records or prior research. In practice, however, the most common form of empirical information is a *single-world marginal distribution*, or a marginal distribution over the full joint distribution of counterfactual potential outcomes in which the same intervention (or lack of intervention) is applied for each

variable of interest. The number of useful constraints provided by each such distribution is at most the number of outcomes in the state space, minus one. For example, in the observational IV case of Figure 12, the empirical evidence shown in panel (b) consists of a single-world marginal distribution $\Pr(Z = z, X = x, Y = y) = \Pr[Z(\emptyset) = z, X(\emptyset) = x, Y(\emptyset) = y]$ in which the same lack of intervention applies to Z , X , and Y . It can immediately be seen that one of the eight constraints must be redundant since the program already implicitly requires that $\sum_{z,x,y} \Pr(Z = z, X = x, Y = y) = 1$; thus, any constraint—for example, $\Pr(Z = 0, X = 0, Y = 0)$ —can be dropped w.l.o.g.

It can further be seen that there are numerous equivalent ways of reformulating the same information: for example, analysts could provide one constraint for $\Pr(Y = 1)$, two for $\Pr(X = 1|Y = y)$ for $y \in \{0, 1\}$, and four for $\Pr(Z = 1|X = x, Y = y)$ for $x, y \in \{0, 1\}$, also totaling seven constraints.²⁵ With these equivalent inputs, Algorithm 2 would yield an alternative program that obtains identical bounds to the one shown in Figure 12(f). In this subsection, we show that one family of techniques for reformulating the empirical evidence—*district factorization* and an extension known as the *nested Markov* formulation Richardson et al. (2017), both defined below—lead to particularly simple and efficient polynomial programs. However, as we elaborate below, some of these simplifications come at a cost: by eliminating information that is superfluous, according to the assumed model, analysts lose the ability to test certain observable implications of those assumptions.

To describe the techniques, we first require some additional notation. Let $\mathbf{A} \subset \mathbf{V}$ be a subset of main variables that are intervened upon by setting them jointly to \mathbf{a} ; this subsumes the special case when $\mathbf{A} = \emptyset$ and no intervention is made. Conversely, let $\mathbf{B} \equiv \mathbf{V} \setminus \mathbf{A}$ represent a subset of main variables that are of interest. Formally, a single-world marginal distribution is a marginal distribution of the full-data law in which (i) every variable in \mathbf{B} is either unaffected by the intervention or is a response to that intervention and (ii) every possible outcome $\mathbf{b} \in \mathcal{S}(\mathbf{B})$ is observed. That is, the single-world marginal distribution is comprised of a set of observed quantities $\{\Pr[\mathbf{B}(\mathbf{a}) = \mathbf{b}] : \mathbf{b} \in \mathcal{S}(\mathbf{B})\}$. As we note above, the

²⁵This implicitly fixes $\Pr(Y = 0)$, $\Pr(X = 0|Y = y)$, and $\Pr(Z = 0|X = x, Y = y)$ by the properties of the generalized principal strata and the second-axiom constraints.

single-world marginal distribution of Figure 12 is $\Pr[Z(\emptyset) = z, X(\emptyset) = x, Y(\emptyset) = y]$ for all $\{z, x, y\} \in \{0, 1\}^3$, so \mathbf{A} is the empty set and \mathbf{B} is $\{Z, X, Y\}$.²⁶

Next, we introduce the key concept of canonical graph *districts*—components of \mathcal{G} that remain connected after removing arrows between the main variables—which will form the basis of this simplification.²⁷ For example, in Figure 12, one district contains X and Y , because they are connected by U_{XY} ; a second district contains only Z .²⁸ It has been shown that single-world marginal distributions arising from causal models associated with hidden-variable DAGs, such as those studied here, can be factorized into district-specific terms (Tian and Pearl, 2002; Richardson, 2003). A particularly useful simplification that arises from this fact is that information equivalent to a single-world marginal distribution can be introduced, district by district, in a particular form that substantially limits the maximum degree of the polynomials involved.

As an illustration, consider the observational distribution of Figure 12(b), $\Pr[Z(\emptyset) = z, X(\emptyset) = x, Y(\emptyset) = y]$ —which is converted by Algorithm 2 into eight quadratic polynomial equality constraints shown in panel (f), one of which is redundant by construction and can be dropped. An equivalent way to introduce this information is in terms of $\Pr[Z(x, y) = z]$ and $\Pr[X(z) = x, Y(z) = y]$, which are guaranteed to be identified because the effects of Z on $\{X, Y\}$ and $\{X, Y\}$ on Z are guaranteed to be unconfounded by the definition of districts. To elaborate, the distribution $\Pr[Z(x, y) = z]$ reduces to $\Pr[Z(\emptyset) = z] = \Pr(Z = z)$, because neither X nor Y are ancestors of Z ; one constraint is sufficient to represent this information, as constraining $\Pr(Z = 1)$ immediately fixes $\Pr(Z = 0)$. An additional three constraints are sufficient to fully specify $\Pr[X(z = 0) = x, Y(z = 0) = y]$ since, for example, the three other

²⁶If analysts subsequently conducted a randomized intervention on X , this would create two additional single-world marginal distributions that could be incorporated into the program: $\Pr[Z(x = 0) = z, Y(x = 0) = y]$ and $\Pr[Z(x = 1) = z, Y(x = 1) = y]$. Note that (i) $Z(x)$ would then reduce to $Z(\emptyset)$, as X is not an ancestor of Z , and (ii) these experimental distributions point-identify the ATE with no other information, and they point-identify the local ATE among compliers in conjunction with the observational distribution and the monotonicity assumption shown in Figure 12.

²⁷Note that districts can also be defined for non-canonical graphs, which we do not examine; for those cases, an extended definition is required.

²⁸As additional examples, in Figure 2(a), V_1 lies in one district while V_2 and V_3 , which share the common parent U_{23} , lie in another district. In Figure 2(b–c), all nodes lie in the same district. Note in Figure 2(b), V_1 is connected through U_{12} to V_2 , which in turn is connected through U_{23} to V_3 ; as a result, V_1 and V_3 are indirectly connected and thus lie in the same district.

possible x, y outcome probabilities jointly fix $\Pr[X(z = 0) = 0, Y(z = 0) = 0]$. Yet another three constraints fully specify $\Pr[X(z = 1) = x, Y(z = 1) = y]$, yielding seven reformulated constraints that contain information equivalent to the original formulation. However, in reformulated constraints of the form $\Pr[X(z) = x, Y(z) = y]$, U_Z is now blocked from X and Y by intervention $Z = z$. This means that all U_Z terms can be eliminated as shown in Section D.1, producing linear constraints instead of the original quadratic ones in Figure 12. Moreover, the reformulated $\Pr(Z = z)$ constraints immediately point-identify the U_Z optimization variables and can thus be eliminated as described in Section D.2.

More generally, let m index districts $\{1, \dots, M\}$. Let $\mathbf{A}^{(m)}$ denote the union of (i) all variables intervened upon, and (ii) all variables outside district m , which by definition cannot be confounded with variables inside district m . Let $\mathbf{B}^{(m)}$ be all variables inside district m that are not intervened upon; thus, $\mathbf{A}^{(m)}$ and $\mathbf{B}^{(m)}$ are mutually exclusive and collectively contain all main variables in the graph. The *district-factorized* quantities are a set of empirical quantities that are guaranteed to be identified: $\{ \Pr [\mathbf{B}^{(m)}(\mathbf{A}^{(m)} = \mathbf{a}^{(m)}) = \mathbf{b}^{(m)}] : \mathbf{a}^{(m)} \in \mathcal{S}(\mathbf{A}^{(m)}), \mathbf{b}^{(m)} \in \mathcal{S}(\mathbf{B}^{(m)}) \}$. In other words, for any hypothetical intervention $\mathbf{a}^{(m)}$, the probability of any possible outcome $\mathbf{b}^{(m)}$ is observed, because by definition $\mathbf{a}^{(m)}$ was either exogeneously set or is unconfounded with $\mathbf{b}^{(m)}$. Note that any $\mathbf{a}^{(m)}$ that is not an ancestor of $\mathbf{B}^{(m)}$ can be immediately dropped, as intervening on these variables has no consequence.

A particularly effective simplification is therefore to polynomialize $\Pr[\mathbf{B}(\mathbf{A} = \mathbf{a}) = \mathbf{b}]$ in terms of the principal strata parameters for each such $\mathbf{a} \in \mathcal{S}(\mathbf{A})$ and $\mathbf{b} \in \mathcal{S}(\mathbf{B})$, then define a constraint in which the polynomial is set equal to $\Pr(\mathbf{B} = \mathbf{b} | \mathbf{A} = \mathbf{a})$. Proposition 4 offers a guarantee on the complexity of these reformulated constraints. A proof is given in Appendix F.4.

Proposition 4. *Every district-factorized constraint can be reformulated in terms of polynomials with a degree bounded from above by the number of disturbances in the corresponding district.*

We conclude with a brief discussion of alternative conceptualizations, tradeoffs, and extensions for this technique. District factorization can be thought of as a way to exploit

certain *conditional independence* and *generalized equality* constraints (or Verma constraints, Verma and Pearl, 1990; Tian and Pearl, 2002) on the observed single-world marginal distributions to simplify the polynomial program. For example, consider a $U_X \rightarrow X \quad Y \leftarrow U_Y$ graph in which binary X and Y variables are completely independent, with no confounding and no causal relationship. In this case, the original empirical evidence would consist of $\Pr(X = x, Y = y)$, in which each constraint is quadratic (the product of U_X and U_Y parameters) and three constraints are required to convey this information (after excluding one redundant constraint since $\sum_{x,y} \Pr(X = x, Y = y) = 1$). After factorizing this single-world marginal distribution by district, however, we obtain two types of reformulated constraints. The first is $\Pr[X(Y = y) = x] = \Pr(X = x)$, where intervention $Y = y$ is dropped as it is not an ancestor of X ; the second is $\Pr[Y(x) = y] = \Pr(Y = y)$, for the same reason. Each of these binary distributions requires only one linear constraint to express, for a total of two constraints. However, the simplification comes at a cost. Essentially, this reformulation of the empirical evidence can be thought of as using the assumed structure of the graph to identify implicit equalities, such as $\Pr(X = x|Y = 0) = \Pr(X = x|Y = 1)$, that are implicitly encoded in the polynomialization. Then, these assumed equalities are used to eliminate constraints that are redundant if that structure in fact holds—here, constraints that relate to dependence between X and Y , which the assumptions imply cannot exist. The tradeoff is that the resulting simplified program is no longer capable of detecting violations of the assumptions, because the reformulated constraints no longer contain information about dependence of X and Y . In other words, unlike the simplifications in Sections D.1–D.2 and D.4, applying the techniques discussed here will result in a simplified program that is not strictly identical, in that it is blind to some violations of the assumptions’ observable implications, whereas the unsimplified program may report that no admissible solution (i.e., no DGP in the model space) exists. However, when the assumptions are in fact satisfied and both programs identify admissible solutions, the resulting bounds are guaranteed to be identical.

Note that the approach described above does not fully exploit every equality constraint implied by the graph. For example, consider a hypothetical modification of the Figure 12 graph in which the $X \rightarrow Y$ edge is removed (i.e., $U_Z \rightarrow Z \rightarrow X \leftarrow U_{XY} \rightarrow Y$), where the districts are

$\{Z\}$ and $\{X, Y\}$ as before. In this case, the district factorization approach still results in seven constraints: one for $\Pr(Z = z)$, three for $\Pr[X(Z = 0) = x, Y = y] = \Pr(X = x, Y = y|Z = 0)$, and three for $\Pr[X(Z = 1) = x, Y = y] = \Pr(X = x, Y = y|Z = 1)$. In contrast, the *nested Markov* factorization of Evans and Richardson (2019) can re-express the same observed single-world marginal distributions using even fewer constraints, by fully exploiting every conditional independence and generalized equality implied by the graph. For example, here the three constraints for $\Pr[X(Z = 1) = x, Y = y|Z = 1]$ can be replaced with a single constraint for $\Pr[X(Z = 1)] = \Pr(X|Z = 1)$, because the assumed graph implies that the distribution of Y is unaffected by intervention on Z , meaning that the now-omitted constraints are duplicative of the previous $\Pr[X(Z = 0) = x, Y = y|Z = 0]$ constraints if the graphical assumptions hold. Note that this simplification exploits a conditional independence constraint implied by the graph; for an example of how generalized independence constraints are also exploited by the nested Markov formulation, see Appendix C.3.

The procedure in Appendix C.4 shows how to employ the nested Markov simplification for an arbitrary single-world marginal distribution by (i) computing empirical quantities; (ii) polynomializing each quantity; (iii) setting it equal to its observed value, and (iv) adding the constraint to the program. This approach can result in considerable savings; for example, Appendix C.3 shows how 64 constraints can be reduced to 31 constraints using the nested Markov simplification.

Finally, when certain deterministic relationships exist between variables in \mathbf{V} , as in the missing-data setting of Figure 7(c-d),²⁹ these relationships may imply additional equality constraints not exploited by the nested Markov approach. In such cases, it may be possible to further reduce the number of constraints. We caution that in general, the simplifications proposed in this should be viewed as a spectrum of tradeoffs. Fully implementing them will result in the simplest and fastest-running polynomial programs, but prevent analysts from testing many observable implications of their theories. Alternatively, providing the most complete possible form of information—one constraint for every $\mathbf{v} \in \mathcal{S}(\mathbf{V})$ —will result in more complex

²⁹In this graph, a latent variable Y has an observed version Y^* that deterministically inherits $Y^* = Y$ when a reporting variable $R = 1$, but takes on the missing-value indicator $Y^* = \text{NA}$ otherwise.

programs that retain the ability to falsify every observable implication of an analyst’s theories.

D.4 Eliminating additional constraints and parameters

Finally, we describe when constraints and parameters can be safely eliminated from a program. We say that parameters x and y *co-occur* in a polynomial system if they appear in the same constraint; they *interact* if there exists a sequence of parameters from x to y such that every adjacent pair co-occurs. For example, consider the constraints $x + y = a$, $y + z = b$. Here, x and y co-occur; x and z interact. If a constraint’s parameters do not interact with the objective’s parameters, that constraint may be dropped. If a parameter exists only in constraints that have been eliminated, then the parameter has also been eliminated, simplifying the system. This is frequently of use after employing the factorization techniques of Section D.3, as disturbance parameters from one district rarely interact with those from another district. In the instrumental variable example of Figure 12, this results in the constraint for $\Pr(Z = 1)$ being eliminated from the program entirely.

E Details on statistical inference

Here, we elaborate on the procedure for statistical inference that was briefly introduced in Section 6. As in the main text, we will use the running example of a binary $X \rightarrow Y$ graph with confounding $X \leftarrow U \rightarrow Y$ to illustrate.³⁰ As previously shown in Figure 3, bounds on the causal estimand are implied by admissible values of the constraints our algorithm seeks to optimize, which, in this simple example, can be represented in three-dimensional space. In other words, if we can characterize the uncertainty in our estimates of the relevant constraints, such as $\Pr(X = 1, Y = 0)$ here, we can obtain confidence intervals on the causal quantity of interest.

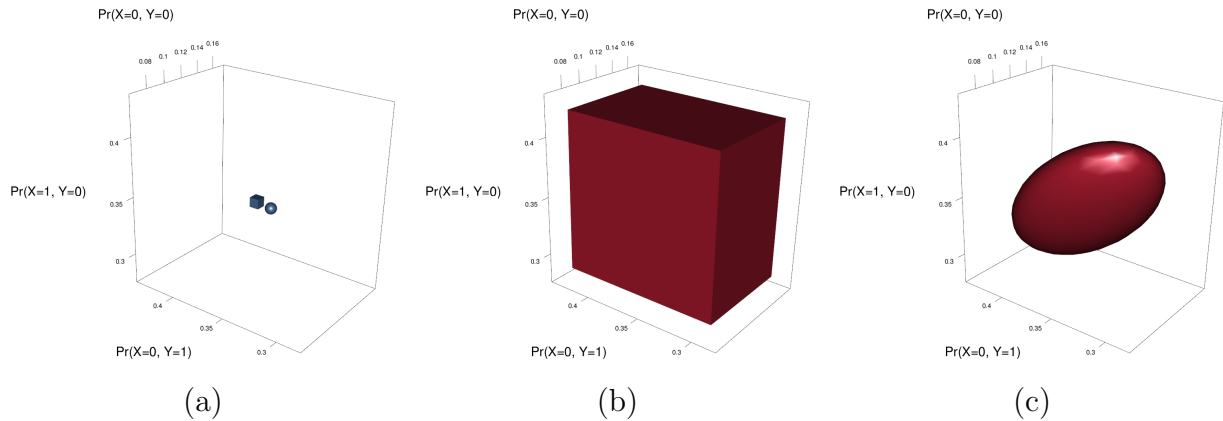
³⁰In what follows, we will assume that empirical evidence arises from a single multinomial distribution, such as $\Pr(X = x, Y = y)$; if multiple independent sets of empirical evidence about differing quantities are available, the procedure generalizes straightforwardly by repeating the procedure within each set and combining the results appropriately.

In this example, the population empirical constraints are $\mathcal{E} = \{g_\ell(\mathcal{P}_V) = E_\ell : \ell\} = \{\Pr(X = 0, Y = 0), \Pr(X = 1, Y = 0), \Pr(X = 0, Y = 1), \Pr(X = 1, Y = 1)\}$. For compactness, we will collect observable population quantities in the vector $\mathbf{E} = [E_\ell]$; in the simulation population here, this is $\mathbf{E} = [0.121, 0.346, 0.349, 0.184]$. When these constraints are input to the algorithm, we refer to the results as the *population bounds*. In practice, the empirical quantities used in these constraints are estimated from finite samples—for example, rather than \mathbf{E} , we may only have $\hat{\mathbf{E}} = [0.113, 0.352, 0.357, 0.178]$ from a sample of $N = 1,000$. By the plug-in principle, the algorithm then relies on the *estimated constraints*, $\hat{\mathcal{E}} = \{g_\ell(\mathcal{P}_V) = \hat{E}_\ell : \ell\}$, and produces the *estimated bounds*.

To construct confidence bounds in this example using Algorithm 3, we then replace the $\hat{\mathcal{E}}$ equality constraints with a set of loosened *confidence constraints* $\text{CR}_\alpha(\hat{\mathcal{E}})$. In other words, suppose the population bounds would be obtained by optimizing subject to a set of exact equality constraints contained in \mathcal{E} , in which the ℓ -th observable quantity imposes the constraint $\{g_\ell(\mathcal{P}_V) = \mathbf{E}_\ell\}$. The vector of observable population quantities, \mathbf{E} , is depicted with a sphere in Figure 13(a). Because these population quantities are unknown, to estimate the bounds, analysts instead use the plug-in principle to construct the an alternate set of equality constraints, $\hat{\mathcal{E}}$, containing $\{g_\ell(\mathcal{P}_V) = \hat{\mathbf{E}}_\ell\}$. The estimated quantities used here, $\hat{\mathbf{E}}$, are shown as a small cube in Figure 13(a). Finally, the confidence bounds will incorporate the loosened constraint $\{g_\ell(\mathcal{P}_V) \in \text{CR}_\alpha(\hat{\mathbf{E}}_\ell)\} \in \text{CR}_\alpha(\hat{\mathcal{E}})$ where $\text{CR}_\alpha(\hat{\mathcal{E}})$ is designed to contain $\hat{\mathcal{E}}$. Two possible confidence regions for the estimated quantities are depicted in Figure 13(b–c). Confidence bounds on the estimand are constructed by minimizing and maximizing the target quantity subject to this loosened constraint.³¹ For example, we will show that one possible confidence region—depicted in Figure 13(b)—is $\Pr(X = 0, Y = 0) \in [0.084, 0.147]$, $\Pr(X = 1, Y = 0) \in [0.305, 0.401]$, $\Pr(X = 0, Y = 1) \in [0.310, 0.406]$, and $\Pr(X = 1, Y = 1) \in [0.142, 0.219]$, a region that contains the sample proportions $\widehat{\Pr}(X = 0, Y = 0) = 0.121$, $\widehat{\Pr}(X = 1, Y = 0) = 0.346$, $\widehat{\Pr}(X = 0, Y = 1) = 0.349$, and $\widehat{\Pr}(X = 1, Y = 1) = 0.184$. Therefore, every DGP that is admissible for the estimated bounds (which require the sample proportions to be exactly

³¹Note that the resulting confidence bounds also incorporate the user-specified looseness threshold ε , and larger values will result in wider confidence bounds.

Figure 13: **Polynomial confidence regions in a binary graph.** We consider a confounded $X \rightarrow Y$ graph with $X \leftarrow U_{XY} \rightarrow Y$. In panel (a), a small blue sphere depicts the population distribution $\Pr(X = x, Y = y)$ along three dimensions: $\Pr(X = 0, Y = 0)$, $\Pr(X = 0, Y = 1)$, and $\Pr(X = 1, Y = 0)$; the final category, $\Pr(X = 1, Y = 1)$ (not depicted), sums to unity. Observed proportions, shown with a small blue cube, differ slightly due to sampling error. Panel (b) shows the Bernoulli-KL confidence region, which is conservative in finite samples and can be polynomialized as a set of linear inequalities. Panel (c) shows the Gaussian confidence region, which is asymptotically valid and can be polynomialized as a single convex quadratic inequality. Both regions are centered on the sample proportions (blue cube), but will contain the population proportions (blue dot) with $\geq 95\%$ (Bernoulli-KL) and asymptotically 95% coverage (Gaussian) over repeated samples.



satisfied) is also admissible in the confidence bounds (which only require them to be approximately satisfied). However, the confidence bounds also consider many additional DGPs that are inadmissible for the estimated bounds. As a result, the confidence bounds must contain the estimated bounds, because taking a minimum or maximum over a superset can only result in more extreme values.

By the same logic, if the confidence region for the observable quantities, $\text{CR}_\alpha(\hat{\mathcal{E}})$, covers the true population values, \mathcal{E} —as in the example above—then the resulting confidence bounds will also cover the population bounds. This means that if the confidence region for the observable quantities has coverage in $1 - \alpha$ of repeated samples, then the resulting confidence bounds for the estimand will also cover the population bounds with probability $\geq 1 - \alpha$.³²

In discrete settings, the task of obtaining confidence bounds thus reduces to the problem

³²However, when the confidence region does not fully contain population quantities due to sampling error, it is still the case that confidence bounds may contain population bounds. This can occur if the non-covered quantity corresponds to a constraint that is irrelevant to the bounds. For example, consider a scenario in which an irrelevant third variable Z is added to the confounded X - Y example, with no connection to either of the original variables. In that case, failure to cover by overestimating $\Pr(Z = 1)$ will have no bearing on the resulting bounds.

of constructing regions $\text{CR}_\alpha(\hat{\mathbf{E}})$ for the multinomial proportion, such that $\Pr(\mathbf{E} \in \text{CR}_\alpha(\hat{\mathbf{E}})) \geq 1 - \alpha$. We discuss two methods for doing so that are easily polynomialized and can thus be incorporated into Algorithm 3. As we show in Section 7.2, coverage is substantially higher than nominal. We note that obtaining nominal coverage of bounds is a notoriously difficult problem, and improving these confidence bounds—for example, by incorporating refinements such as Guo and Richardson (2021)—remains an important direction for future research.

Our first method for constructing confidence regions for the observable quantities is based on the “Bernoulli-KL” approach of Malloy et al. (2020) that constructs separate confidence regions for each observable event in the data generating process, such as $X = 0, Y = 0$. When combined, these event-specific confidence regions form a hypercube that is guaranteed to contain the full vector of population event probabilities, \mathbf{E} , at conservative rates in repeated samples.

Formally, let $k \in \{1, \dots, K\}$ index possible atomic events, such as $X = 0, Y = 0$, which, when considered jointly, represent the true population region of the constraints corresponding to the bounds on the causal estimand. We denote the probability of the k -th event as $p_k = \Pr(\mathbf{V} = \mathbf{v}_k)$. These proportions can be estimated noisily from data, denoted \hat{p}_k . We will develop a confidence region of the form $\text{CR}_\alpha(\hat{\mathbf{E}}) = \bigotimes_{k=1}^K [\underline{p}_k, \bar{p}_k]$ —the aforementioned hypercube around the estimated sample proportions. In our running example, this means we first construct the region $\Pr(X = 0, Y = 0) \in [\underline{\text{CR}}_\alpha(\widehat{\Pr}(X = 0, Y = 0)), \overline{\text{CR}}_\alpha(\widehat{\Pr}(X = 0, Y = 0))]$, proceeding to $\Pr(X = 0, Y = 1) \in [\underline{\text{CR}}_\alpha(\widehat{\Pr}(X = 0, Y = 1)), \overline{\text{CR}}_\alpha(\widehat{\Pr}(X = 0, Y = 1))]$, and so on before combining. A visualization of the resulting region is given in Figure 13(b).

We now summarize how \underline{p}_k and \bar{p}_k are calculated to ensure that they jointly cover the population proportions with probability of at least $1 - \alpha$, using a result on the Kullback-Leibler divergence of sampling distributions. Taking each \hat{p}_k estimate as given, we identify regions of the unknown p_k from which the observed \hat{p}_k diverge substantially. Equation 11 of Malloy et al. (2020) provides bounds on the sampling probability of observing $\text{KL}([1 - \hat{p}_k, \hat{p}_k], [1 - p_k, p_k]) = \hat{p}_k \log \frac{\hat{p}_k}{p_k} + (1 - \hat{p}_k) \log \frac{1 - \hat{p}_k}{1 - p_k}$ in excess of some threshold. In turn, these bounds imply regions of p_k that can be conservatively rejected. Let \underline{p}_k be given by the solution to $\text{KL}([1 - \hat{p}_k, \hat{p}_k], [1 - \underline{p}_k, \underline{p}_k]) = \frac{1}{N} \log \frac{2K}{1 - \alpha}$ subject to $\underline{p}_k \in [0, \hat{p}_k]$. Similarly, let \bar{p}_k be given by $\text{KL}([1 - \hat{p}_k, \hat{p}_k], [1 - \bar{p}_k, \bar{p}_k]) = \frac{1}{N} \log \frac{2K}{1 - \alpha}$ subject to $\bar{p}_k \in [\hat{p}_k, 1]$.

$\bar{p}_k, \bar{p}_k]) = \frac{1}{N} \log \frac{2K}{1-\alpha}$ subject to $\bar{p}_k \in [\hat{p}_k, 1]$. It can be seen from [Malloy et al. \(2020\)](#) that when constructing \underline{p}_k and \bar{p}_k in this way, $\Pr \left(\bigotimes_{k=1}^K p_k \in [\underline{p}_k, \bar{p}_k] \right) \geq \alpha$ over repeated samples.

The Bernoulli-KL method produces a confidence region for single-world distributions that is guaranteed to have conservative coverage for the multinomial proportion in finite samples. The region can be represented as a system of linear inequality constraints, then incorporated into the polynomial program. For example, rather than using the equality constraint $\Pr(X\text{-control}, Y\text{-never}) + \Pr(X\text{-control}, Y\text{-defy}) = \Pr(X = 0, Y = 0) = 0.121$ to obtain population bounds, or $\widehat{\Pr}(X = 0, Y = 0) = 0.113$ for estimated bounds, we instead optimize subject to the inequality constraints $0.073 \leq \Pr(X\text{-control}, Y\text{-never}) + \Pr(X\text{-control}, Y\text{-defy}) \leq 0.163$, the values obtained from the procedure described above.

Our second approach uses an asymptotic confidence region based on the multivariate Gaussian limiting distribution of the multinomial proportion, $\mathcal{N} \left(\mathbf{p}, \frac{1}{N} \text{diag}(\mathbf{p}) - \frac{1}{N} \mathbf{p} \mathbf{p}^\top \right)$ ([Bienaymé, 1838](#)). Because the multinomial proportion must sum to unity, this distribution is degenerate, and it is often more convenient to work with its first $K - 1$ elements, $\mathbf{p}_{\setminus K}$. We construct the asymptotic confidence region as $(\hat{\mathbf{p}}_{\setminus K} - \mathbf{p}_{\setminus K})^\top \left(\frac{1}{N} \text{diag}(\hat{\mathbf{p}}_{\setminus K}) - \frac{1}{N} \hat{\mathbf{p}}_{\setminus K} \hat{\mathbf{p}}_{\setminus K}^\top \right)^{-1} (\hat{\mathbf{p}}_{\setminus K} - \mathbf{p}_{\setminus K}) \leq z$, where z is an appropriate critical value of the χ^2 distribution. A visualization of the resulting region is given in [Figure 13\(c\)](#). Each element in \mathbf{p} is polynomializable, leading to a single quadratic confidence constraint that can be straightforwardly incorporated into the optimization routine.

Simulations reported in [Section 7.2](#) evaluate coverage of the methods for various sample sizes.

F Proofs

F.1 Proof of Proposition 1

Proof. We adapt the proof of [Finkelstein et al. \(2021\)](#) to account for counterfactuals as follows. First, we define *one-step-ahead* counterfactuals, $V_j(\mathbf{p} \mathbf{a}_V(V_j) = \mathbf{a})$, to be those where all main parents of a variable are subject to intervention $\mathbf{p} \mathbf{a}_V(V_j) = \mathbf{a}$. Next, we note that all

other counterfactuals and factials in the full data law are deterministic functions of one-step-ahead variables, after fixing \mathbf{U} . Therefore it is sufficient to reason about only one-step-ahead variables; intervention on other variables is irrelevant to the full data law.

Because the likelihoods of multi-district graphs factorize as the likelihoods of the districts after intervention on their parents (Richardson et al., 2017), we can consider single-district graphs w.l.o.g. In multi-district graphs, the bound obtained below can be applied within each district.

Each main variable V_j has $|\mathcal{S}(\mathbf{pa}_{\mathbf{V}}(V_j))|$ one-step-ahead counterfactuals, corresponding to possible manipulations of its parents. Each one-step-ahead counterfactual $V_j(\mathbf{pa}_{\mathbf{V}}(V_j) = \mathbf{a})$ has a cardinality equal to those of the corresponding main variable $|\mathcal{S}(V_j)|$. Therefore, the collection of a single variable's one-step-ahead counterfactuals $\{V_j(\mathbf{pa}_{\mathbf{V}}(V_j) = \mathbf{a}), V_j(\mathbf{pa}_{\mathbf{V}}(V_j) = \mathbf{a}'), \dots\}$ can take on $|\mathcal{S}(V_j)|^{|\mathcal{S}(\mathbf{pa}_{\mathbf{V}}(V_j))|}$ possible values, and there are $d \equiv \prod_{V_j \in \mathbf{V}} |\mathcal{S}(V_j)|^{|\mathcal{S}(\mathbf{pa}_{\mathbf{V}}(V_j))|}$ values that the full collection of all one-step-ahead variables can take. Any model over this full collection must be a subset of the $d - 1$ simplex. We let $\mathbf{V}(\mathbf{pa}_{\mathbf{V}}(\mathbf{V}))$ denote the collection of one-step-ahead variables.

Suppose the disturbances \mathbf{U} are enumerated as $\{U_1, \dots, U_K\}$. We will now show that each U_k can be assumed to be discrete without altering the model for $\mathbf{V}(\mathbf{pa}_{\mathbf{V}}(\mathbf{V}))$ and therefore the full data law. First, for each value u_k in the domain of U_k , we define the distribution $P_{u_k}(\mathbf{V}(\mathbf{pa}_{\mathbf{V}}(\mathbf{V}))) = \int_{\mathbf{u}_{\setminus k}} P(\mathbf{V}(\mathbf{pa}_{\mathbf{V}}(\mathbf{V})) \mid \mathbf{u}_{\setminus k}, u_k) P(\mathbf{u}_{\setminus k})$, where $\mathbf{u}_{\setminus k}$ denotes all disturbances other than u_k . This fixes U_k at the value u_k , modifying the distribution over $\mathbf{V}(\mathbf{pa}_{\mathbf{V}}(\mathbf{V}))$.

We now make two observations. First, the model for $\mathbf{V}(\mathbf{pa}_{\mathbf{V}}(\mathbf{V}))$ contains P_{u_k} for any u_k , because U_k is not restricted by the model and is therefore permitted to have a point-mass distribution at u_k . Second, the expected value of P_{u_k} with respect to U_k recovers the original marginal distribution $P(\mathbf{V}(\mathbf{pa}_{\mathbf{V}}(\mathbf{V})))$, which is therefore in the convex hull of the set of distributions $\mathcal{S}(P_{u_k}) \equiv \{P_{u_k} \mid u_k \in \mathcal{S}(U_k)\}$.

Carathéodory's Theorem (1907) states that for any point P in the convex hull of a set \mathcal{S} in a space of dimension $d - 1$, there exists a set of $d - 1$ points $\{P_{u_{k_1}}, \dots, P_{u_{k_{d-1}}}\}$ and weights $\{w_1, \dots, w_{d-1}\}$ such that $P = \sum_{\ell=1}^{d-1} w_{\ell} P_{u_{k_{\ell}}}$. It then follows directly that any distribution in

the marginal model over $\mathbf{V}(\mathbf{pa}_{\mathbf{V}}(\mathbf{V}))$ when latent variables have unrestricted cardinality is also in the marginal model over $\mathbf{V}(\mathbf{pa}_{\mathbf{V}}(\mathbf{V}))$ when latent variables have cardinality restricted to $\prod_{V_j \in \mathbf{V}} |\mathcal{S}(V_j)|^{|\mathcal{S}(\mathbf{pa}_{\mathbf{V}}(V_j))|} - 1$ or higher. \square

F.2 Proof of Proposition 2

Proof. Using the approach developed in Evans (2018) and generalized to arbitrary graphs in Finkelstein et al. (2021), we can obtain a generalized principal stratification that is non-restrictive of the causal model of \mathcal{G} over observed variables. In such a model, each $V_\ell(\mathbf{a}_\ell)$ is determined by values of the disturbances \mathbf{U} . By assumption, \mathcal{G} is in canonical form, rendering all disturbances marginally independent. The proposition then follows from standard probability calculus. \square

F.3 Proof of Proposition 3, with discussion

Proof. Under the conditions specified, no \mathcal{C}_ℓ involves a function of U_k . It follows that whether the disturbances lead to \mathcal{C}_ℓ being jointly satisfied is not a function of the value of U_k . As a result, a sum over all parameters of the distribution of U_k can be factored out of the product in Equation 1. By the definition of probability distributions, this sum will be equal to unity, rendering the parameters irrelevant to the polynomial. \square

Discussion. To see why this independence must hold, consider that all disturbances, including U_k , are by construction exogenous in canonical DAGs. Therefore there can be no path from a common ancestor to both U_k and a set of outcomes $\mathbf{Y} \subset \mathbf{V}$, and there can be no path from \mathbf{Y} to U_k . It follows that the only kinds of paths that are possible are (i) paths from U_k to \mathbf{Y} and (ii) collider paths between U_k and \mathbf{Y} . Of these, only the former can induce dependence. If all paths from U_k to \mathbf{Y} are blocked by \mathbf{A} , then it has been shown that in the FFRCISTG model—and therefore also in the stronger NPSEM-IE model that we use here—that \mathbf{U} must be independent of $\mathbf{Y}(\mathbf{A})$ (Richardson and Robins, 2013). The intuition behind this observation is simple: once \mathbf{A} have been intervened upon, they are no longer random and therefore cannot propagate statistical dependence along a causal pathway. If U_k is independent of $\mathbf{Y}(\mathbf{A})$, and our target of inference is a functional of $\Pr(\mathbf{Y}(\mathbf{A}) = \mathbf{y})$, then by definition the distribution of

U_k contains no information about our target.

We briefly remark on the relationship between the NPSEM-IE and FFRCISTG models. It has been shown that independence of disturbance terms in the NPSEM-IE model implies seemingly counterintuitive restrictions on “cross-world independences,” governing how sets of random variables counterfactually behave under inconsistent “worlds” or sets of treatment assignments. These restrictions arise in identification theory in mediation analysis ([Shpitser, 2013](#)). The FFRCISTG model is a weaker causal model which only implies independences on counterfactual variables with a consistent set of interventions—the so-called “single-world independences.” It has been noted that by imposing fewer assumptions on counterfactual variables, FFRCISTG leads to wider bounds than NPSEM-IE.

F.4 Proof of Proposition 4

Proof. Each of the nested Markov parameters corresponds to the probability that random variables in a single district take certain values after an intervention on parents of the district. It follows from Proposition 3 that no disturbances outside the district corresponding to the nested Markov parameter will appear in the polynomialization of that parameter. From this, it then follows that no disturbances in different districts will interact in constraints corresponding to nested Markov parameters. By Proposition 2, the degree of a polynomialization of the probability of the event is at most the number of relevant disturbances. \square

G Details of simulated models

In this section, we detail all models presented in Section 7. For simplicity, all main variables in these models are binary. Simulation parameters are described in terms of principal strata. Principal strata can take one of three forms, depending on the number of parents of the relevant variable. Below, we provide compact notation for referring to these principal strata. Subsequent sections report strata probabilities for each simulation, including joint distributions over strata for multiple variables where confounding exists.

1. **Variables with no parents, which have two strata.** Consider a hypothetical variable

X with no parents, as in Figure 7(a). We use x_0 to denote units with $X(\emptyset) = 0$ and x_1 to denote $X(\emptyset) = 1$.

2. **Variables with a single parent, which have four strata.** Consider a hypothetical variable Y influenced by parent X , also depicted in Figure 7(a). For compactness, we adopt the convention that counterfactual manipulations of parent variables are presented in the form $y_{Y(X=0),Y(X=1)}$. For example, (i) we use y_{00} to denote “never takers” with example, $Y(X = 0) = 0$ and $Y(X = 1) = 0$. Similarly, (ii) y_{01} denotes “compliers” with $Y(X = 0) = 0$ and $Y(X = 1) = 1$, (iii) y_{10} denotes “defiers” with $Y(X = 0) = 1$ and $Y(X = 1) = 0$, and y_{11} denotes “always takers” with $Y(X = 0) = 1$ and $Y(X = 1) = 1$.
3. **Variables with two parents, which have sixteen strata.** Consider a hypothetical variable Y influenced by parents Z and X , as in Figure 5(a). Extending the convention described above, we denote these in compact forms ranging from y_{0000} to y_{1111} . Specific definitions are provided in Table 2.

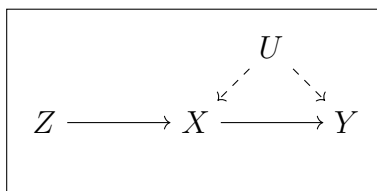
Table 2: **Principal strata for a variable Y with two parents, Z and X .** Each row corresponds to a strata, with compact names given in the first column. For each strata, counterfactual values of Y are given in subsequent columns.

	$Y(Z = 0, X = 0)$	$Y(Z = 0, X = 1)$	$Y(Z = 1, X = 0)$	$Y(Z = 1, X = 1)$
y_{0000}	0	0	0	0
y_{1000}	1	0	0	0
y_{0100}	0	1	0	0
y_{1100}	1	1	0	0
y_{0010}	0	0	1	0
y_{1010}	1	0	1	0
y_{0110}	0	1	1	0
y_{1110}	1	1	1	0
y_{0001}	0	0	0	1
y_{1001}	1	0	0	1
y_{0101}	0	1	0	1
y_{1101}	1	1	0	1
y_{0011}	0	0	1	1
y_{1011}	1	0	1	1
y_{0111}	0	1	1	1
y_{1111}	1	1	1	1

G.1 Noncompliance simulation

In this section, we describe the DGP for our noncompliance simulation analyzed in Section 7.1. The DGP follows the model of Figure 5(b), reproduced below for ease of reference. Simulation parameters are reported in terms of the joint distribution over principal strata.

Figure 14: **DGP with noncompliance.**



Strata for Z :

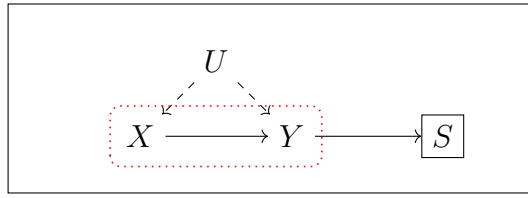
z_0	0.649335
z_1	0.350665

Strata for X and Y :

	y_{00}	y_{10}	y_{01}	y_{11}
x_{00}	0.0	0.0172	0.00541	0.0
x_{10}	0.0	0.549	0.173	0.0
x_{01}	0.0	0.0305	0.0971	0.0
x_{11}	0.0	0.0304	0.0968	0.0

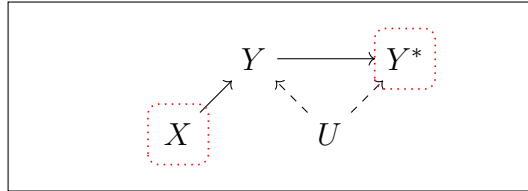
G.2 Outcome-based selection simulation

In this section, we describe the DGP for our outcome-based selection simulation, analyzed in Section 7.3 and Figure 8(a). The DGP follows the model of Figure 7(a), reproduced below for ease of reference. For this case, we employed the DGP of [Gabriel et al. \(2022\)](#).



G.3 Measurement error simulation

In this section, we describe the DGP for our measurement error simulation, analyzed in Section 7.3 and Figure 8(b). The DGP follows the model of Figure 7(b), reproduced below for ease of reference. Simulation parameters are reported in terms of the joint distribution over principal strata.



Strata for X

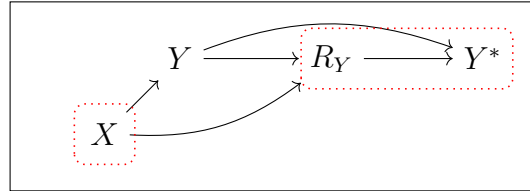
x_0	0.499442
x_1	0.500558

Strata for Y and Y^*

	Y_{00}^*	Y_{10}^*	Y_{01}^*	Y_{11}^*
y_{00}	0	0.167269	0	0
y_{10}	0	0	0	0
y_{01}	0	0.165838	0.500388	0
y_{11}	0	0.166505	0	0

G.4 Outcome missingness simulation

In this section, we describe the DGP for our outcome missingness simulation, analyzed in Section 7.3 and Figure 8(c). The DGP follows the model of Figure 7(c), reproduced below for ease of reference. Simulation parameters are reported in terms of the joint distribution over principal strata.



Strata for X

x_0	0.499159
x_1	0.500841

Strata for Y

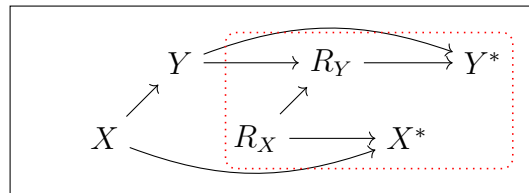
y_{00}	0.166371
y_{10}	0
y_{01}	0.666851
y_{11}	0.166778

Strata for R

r_{0000}	0
r_{1000}	0
r_{0100}	0.250368
r_{1100}	0.249910
r_{0010}	0
r_{1010}	0
r_{0110}	0
r_{1110}	0
r_{0001}	0
r_{1001}	0
r_{0101}	0.250154
r_{1101}	0
r_{0011}	0
r_{1011}	0
r_{0111}	0
r_{1111}	0.249568

G.5 Joint missingness simulation

In this section, we describe the DGP for our joint missingness simulation, analyzed in Section 7.3 and Figure 8(d). The DGP follows the model of Figure 7(d), reproduced below for ease of reference. Simulation parameters are reported in terms of the joint distribution over principal strata.



Strata for X

x_0	0.43464
x_1	0.56536

Strata for Y

y_{00}	0.485336
y_{10}	0.253616
y_{01}	0.003768
y_{11}	0.257279

Strata for R_x

$r_{x,0}$	0.470201
$r_{x,1}$	0.529798

Strata for R_y

$r_{y,0000}$	0
$r_{y,1000}$	0
$r_{y,0100}$	0.162045
$r_{y,1100}$	0
$r_{y,0110}$	0.177470
$r_{y,0001}$	0.107010
$r_{y,1001}$	0.120311
$r_{y,0101}$	0.255778
$r_{y,1101}$	0.081733
$r_{y,0011}$	0
$r_{y,1011}$	0
$r_{y,0111}$	0.095652
$r_{y,1111}$	0

G.6 Scaling of computation time with variable cardinality

The computation time needed to solve a causal problem depends on three considerations: (i) the structure of the graph; (ii) the form of the estimand and constraints; and (iii) the cardinality, or number of categories, for each main variable.

First, when no simplifications are applied, the polynomial degree of empirical constraints from an observed single-world marginal distribution will grow with the number of districts in the graph. For example, consider an observational distribution in which $\Pr(\mathbf{V} = \mathbf{v})$ is fully observed. If all variables in the graph are confounded by a single U , the polynomialization of the empirical evidence will immediately produce linear constraints; if the objective function is also linear, then we will obtain a linear program, which can generally be solved quickly. However, as Proposition 4 shows, even when a graph has multiple districts, the resulting evidence can be refactorized into a series of district-specific constraints in which the polynomial degree is upper-bounded by the number of disturbances in the district. If there are multiple districts, but each contains only one disturbance, then the empirical evidence can be reorganized into a series of linear constraints.

Second, the estimand itself also determines the polynomial degree, as certain estimands can result in more complex objective functions. For example, cross-world conditional interventions—such as the local average treatment effect among compliers in an instrumental variable problem—can result polynomial fractions that require auxiliary variables to clear.

Finally, the number of levels in each variable can increase optimization time exponentially. To show this, we solved several instrumental variable bounding problems, avoiding the use of Section 4.2 simplifications to ensure that results are comparable. We varied the number of levels for encouragement Z , treatment X , and outcome Y from binary to ternary, both independently and in combination. In each problem, we used a uniform distribution over $\Pr(Z = z, X = x, Y = y)$ as the sole empirical evidence. No monotonicity assumption was used.

The resulting computation time are given in Table 3. As the table shows, runtime grows with the number of strata, increases the number of optimization variables in the problem. For

the IV problem, in the all-binary case, the problem was formulated and solved in 1.3 seconds without simplifications, whereas the case where every variable is ternary was solved in 313.1 seconds. Intermediary cases where X , Z , and Y has differing cardinalities, were solved in less than 4 seconds. We emphasize that these computation times do not reflect the Section 4.2 simplifications, which can improve runtime drastically.

Table 3: **Runtime for ATE bounds in instrumental variable problems.** The number of levels of each variable in Figure 5(b) was varied from binary to ternary.

Variable cardinality			Disturbance parameters			Seconds
Z	X	Y	U_Z	U_{XY}	Total	
binary	binary	binary	2	$2^2 \cdot 2^2 = 16$	18	1.30
ternary	binary	binary	3	$2^3 \cdot 2^2 = 32$	35	3.91
binary	ternary	binary	2	$3^2 \cdot 2^3 = 72$	74	3.63
binary	binary	ternary	2	$2^2 \cdot 3^2 = 36$	38	1.18
ternary	ternary	ternary	3	$3^3 \cdot 3^3 = 729$	732	313.12