A General Method for Deriving Tight Symbolic Bounds on Causal Effects

Preprint

A General Method for Deriving Tight Symbolic Bounds on Causal Effects

Michael C. Sachs Gustav Jonzon Arvid Sjölander
Erin E. Gabriel*

Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
corresponding author: gustav.jonzon@ki.se

December 10, 2021

Abstract

A causal query will commonly not be identifiable from observed data, in which case no estimator of the query can be contrived without further assumptions or measured variables, regardless of the amount or precision of the measurements of observed variables. However, it may still be possible to derive symbolic bounds on the query in terms of the distribution of observed variables. Bounds, numeric or symbolic, can often be more valuable than a statistical estimator derived under implausible assumptions. Symbolic bounds, however, provide a measure of uncertainty and information loss due to the lack of an identifiable estimand even in the absence of data. We develop and describe a general approach for computation of symbolic bounds and characterize a class of settings in which our method is guaranteed to provide tight valid bounds. This expands the known settings in which tight causal bounds are solutions to linear programs. We also prove that our method can provide valid and possibly informative symbolic bounds that are not guaranteed to be tight in a larger class of problems. We illustrate the use and interpretation of our algorithms in three examples in which we derive novel symbolic bounds.

Keywords: Causal bounds; Causal inference; Unmeasured confounding.

*The authors report there are no competing interests to declare. MCS and GJ are partially supported by Swedish Research Council grant 2019-00227, EEG by Swedish Research Council grant 2017-01898, and AS by Swedish Research Council grant 2016-01267.

1 Introduction

In many fields of research, a common goal is to determine causal relationships or mechanistic pathways. This investigation is often complicated by common causes of the outcome and either the exposure, or other variables of interest along the causal pathway from the exposure to the outcome, causing confounding. When common causes are unmeasured, the causal effect of interest is usually not identifiable. When the causal effect of interest, which we will refer to as a causal query, cannot be identified, one can derive bounds, i.e., a range of possible values for this quantity in terms of the observed data distribution.

In general, arbitrarily wide bounds are trivial to derive, but not informative in the sense that they will not provide further insight into the magnitude of the effect. Deriving narrower bounds that are still valid, i.e. containing all possible values of the true causal effect, can be a complicated task, and in particular, deriving tight bounds, i.e., the narrowest possible given all and only explicit assumptions, may be highly non-trivial. An approach to deriving numeric tight bounds in quite general settings is given in Duarte et al. [2021]. A drawback of the numeric approach, however, is the need for re-computation with each new data set.

Computing bounds symbolically, i.e., as closed form analytic expressions in terms of known observable quantities, rather than numerically, may provide useful information with which to draw conclusions about a study design or form of data collection in the absence of data, in addition to their transparent ease of use in real data once derived. Symbolic tight bounds on a causal query thus, in many ways, provide us with an ideal summary of our effect of interest given our current state of knowledge and/or set of assumptions.

In 1994, in his PhD dissertation, Alexander Balke gave a method for translating a certain type of causal theory, represented by a directed acyclic graph (DAG), and causal query into a constrained optimization problem [Balke and Pearl, 1994a,b] in terms of unmeasured response function variables. The causal query is expressed in terms of the distribution of these variables and the DAG gives rise to linear relationships between this distribution and that of the observed variables. In conjunction with standard probabilistic constraints, this yields a bounded constrained optimization problem. If the problem is linear then a vertex enumeration algorithm can be used to find the global extrema of the causal query in terms of the true probability distribution of the observed variables [Dantzig, 1963].

Balke and Pearl [1994a] states that the resulting extrema give tight bounds for their causal query in the instrumental variable setting. This and related theoretical results have been shown in specific settings that are extensions to the binary instrumental variable problem [Ramsahai, 2012, Bonet, 2013, Heckman and Vytlacil, 2001]. To the knowledge of the authors, there has been no attempt in the literature to characterize the set of causal problems that are always linear or an approach for determining whether a problem is linear, given its DAG and target query.

In this paper, we generalize and extend Balke and Pearl's approach for computation of bounds by characterizing a class of causal problems that always give rise to linear programs and describing a general algorithm for constructing the objective and constraints based on the DAG and query. In Section 2, we introduce the transformation of a causal DAG over categorical variables with unmeasured causal influences into an equivalent one where those influences have been discretized. In Section 3 we characterize a set of DAGs that have linear relations between the distributions of their observed variables and unobserved influences, along with an algorithm that extracts those relations from the DAG. Section 4 develops notation and requirements for general forms of causal queries that are linear in the distribution of the unmeasured discrete influences, and details an algorithm that constructs such relations from a complex causal query expressed in terms of potential outcomes and observable variables. Section 5 then states the final linear program, possible extensions of it and a suitable optimization method. Finally, Section 6 details a few interesting examples using this method. Proofs of the main propositions are given in the Supplementary Materials. The algorithms described herein are implemented in an R [R Core Team, 2019] package called `causaloptim`, available on the Comprehensive R Archive Network (CRAN), with a user friendly interface.

2 Discretization

Let the set of observed variables be denoted $\mathcal{W} = \{W_1, \ldots, W_n\}$, with corresponding vector $\mathbf{W} = (W_1, \ldots, W_n)$ and realized values represented by a vector $\mathbf{w} = (w_1, \ldots, w_n)$. We assume that all of these variables are categorical. Each variable of interest $W_i \in \mathcal{W}$ is affected by a set of unmeasured variables U_{W_i} as well as a subset of the remaining variables. Potential outcomes will be denoted using brackets; e.g., $W_1(W_2 = w_2)$, and the probability that the variable W_1 would have value w_1, if the variable W_2 was intervened upon to have value w_2 will be denoted as $p\{W_1(W_2 = w_2) = w_1\}$. For any given random variable X, we let $\nu(X)$ denote its support, i.e., for discrete variables, the set of all values it can take on with positive probability.

We are assuming the nonparametric structural equation framework, i.e., for each $W_i \in \mathcal{W}$, we assume that there exists a function F_{W_i} such that w_i, the value of W_i is given by $w_i = F_{W_i}(\mathbf{pa}_{W_i}, u_{W_i})$, where $\mathbf{pa}_{W_i}$ denotes the values of variables $\mathbf{Pa}_{W_i}$ in $\mathcal{W}$ that are parents of W_i, and u_{W_i} represents the values of U_{W_i}, the unmeasured causes of W_i. The unmeasured variables U_{W_i} are not assumed independent, unless indicated by the DAG. Since all observed variables of interest in the graph are assumed to be categorical, we can, without loss of generality, recode the assumptions by defining a series of new categorical variables

R_{W_i}, one for each variable $W_i \in \mathcal{W}$, which specifies how the value of W_i is determined from those of its parents. For each $W_i \in \mathcal{W}$, we let R_{W_i} be the variable corresponding to the canonical partition of $\nu(U_{W_i})$ into finite states with respect to the given causal DAG, as stated formally in Proposition 1.

Proposition 1 (Canonical partitions)**.** *Let G be a causal DAG, let $\mathcal{W} := V(G)$ be its vertices and suppose that $\forall W \in \mathcal{W}, |\nu(W)| < \infty$ (i.e. each variable is categorical). Let $\mathcal{D} := \nu(\mathbf{Pa}_W) \times \nu(R_W)$ if $\mathbf{Pa}_W$ is nonempty and $\nu(R_W)$ otherwise. Then there exists a categorical variable R_W (so $|\nu(R_w)| < \infty$) and a mapping $f_W : \mathcal{D} \to \nu(W)$ (called the response function of W) such that for each value $u_W \in \nu(U_W)$ there exists a unique value $r_W \in \nu(R_W)$ for which $F_W(\cdot, u_W) = f_W(\cdot, r_W)$.*

For proof, see the Supplementary Materials.

Regardless of the cardinality of U_{W_i}, we have

$$|\nu(R_{W_i})| = |\nu(W_i)|^{|\nu(\mathbf{Pa}_{W_i})|} = |\nu(W_i)|^{\prod_{V \in \mathbf{Pa}_{W_i}} |\nu(V)|} < \infty,$$

since all variables in $\mathcal{W}$ are assumed categorical. Let $c_{W_i} := |\nu(W_i)|$, so $|\nu(R_{W_i})| = c_{W_i}^{\prod_{V \in \mathbf{Pa}_{W_i}} c_V}$, and note that without loss of generality we can assume that $\nu(W_i) = \{0, \dots, c_{W_i} - 1\}$ and enumerate $\nu(R_{W_i})$ as $\{0, \dots, c_{W_i}^{\prod_{V \in \mathbf{Pa}_{W_i}} c_V} - 1\}$. Let $\mathbf{R} := (R_{W_1}, \dots, R_{W_n})$ and $\aleph := |\nu(\mathbf{R})| = \prod_{i=1}^{n} |\nu(R_{W_i})| = \prod_{i=1}^{n} c_{W_i}^{\prod_{V \in \mathbf{pa}_{W_i}} c_V}$. The joint distribution of $\mathbf{R}$ together with the response functions fully characterize the probabilistic causal model.

For a given $W_i \in \mathcal{W}$ and fixed $\mathbf{r} \in \nu(\mathbf{R})$, we define a procedure for determining its value w_i by recursively evaluating the corresponding functional expression. Using nested subscripts, we let $W_{i1}, \dots, W_{ik_i}$ denote the parents of W_i that are in $\mathcal{W}$. Then w_i, the value of W_i, can be obtained by recursively evaluating

$$w_i = g^*_{W_i}(\mathbf{r}) := f_{W_i}(g^*_{W_{i1}}(\mathbf{r}), \dots, g^*_{W_{ik_i}}(\mathbf{r}), r_{W_i}).$$

Any set of observed probabilities can be related to the distribution of response function variables as follows:

$$p\{\mathbf{W} = \mathbf{w}\} = p\{W_1 = w_1, \dots, W_n = w_n\} = \sum_{\mathbf{r} \in \nu(\mathbf{R}): \forall i \in \{1,\dots,n\}, w_i = g^*_{W_i}(\mathbf{r})} p\{\mathbf{R} = \mathbf{r}\}.$$

As an example, Figure 1 shows a simple setting with three binary variables of interest. Figure 1a shows the DAG for a model in which variables W_1 and W_2 both directly affect an outcome W_3, with W_1 also directly affecting W_2. Figure 1b shows the equivalent DAG with response functional variables in place of the original unmeasured variables. The variables that have an unmeasured common cause have response function variables that are dependent, as indicated by the dashed ellipse that outlines the unmeasured causal influences of W_2 and W_3. Since they both contain U, the common cause, their response function variables are dependent as indicated by an undirected edge. We can encode R_{W_2} so the values $0, 1, 2, 3$ of R_{W_2} correspond to the response patterns $w_2 = f_{W_2}(pa_{W_2}{=}w_1, r_{W_2}{=}0) := 0, w_2 = f_{W_2}(pa_{W_2}{=}w_1, r_{W_2}{=}1) := w_1, w_2 = f_{W_2}(pa_{W_2}{=}w_1, r_{W_2}{=}2) := 1 - w_1, w_2 = f_{W_2}(pa_{W_2}{=}w_1, r_{W_2}{=}3) := 1$, respectively. We encode R_{W_1} taking values 0 and 1 according to $w_1 = f_{W_1}(r_{W_1}{=}0) := 0$ and $w_1 = f_{W_1}(r_{W_1}{=}1) := 1$, respectively. Under the model shown in Figure 1b, with e.g. $r_{W_1} = 0$, $r_{W_2} = 1$, $r_{W_3} = 3$, we can evaluate the function to determine w_2:

$$w_2 = g^*_{W_2}(\mathbf{r}{=}(0, 1, 3)) = f_{W_2}(f_{W_1}(r_{W_1}{=}0), r_{W_2}{=}1)) = f_{W_2}(0, 1) = 0.$$

For W_3, we need to enumerate the response patterns for each of the 2^2 possible combinations of values of (w_1, w_2), i.e., $2^{2^2} = 16$. Then, to evaluate the probability $p\{W_1 =$

$1, W_2 = 0, W_3 = 1\}$ in terms of $\mathbf{R}$, we can follow the same procedure as above for all $2^1 \cdot 2^2 \cdot 2^4 = 128$ possible combinations of $\mathbf{r}$, keeping track of the resulting values $\mathbf{w}$. It can be shown that the variable value $\mathbf{w} = (1, 0, 1)$ is consistent with 16 values of $\mathbf{r}$. Thus the probability of this event is the sum over the set of these 16 values of the probability that $\mathbf{R}$ equals them. See Balke and Pearl [1994a] or Pearl [2009], Chapter 8 for another example and further interpretation.

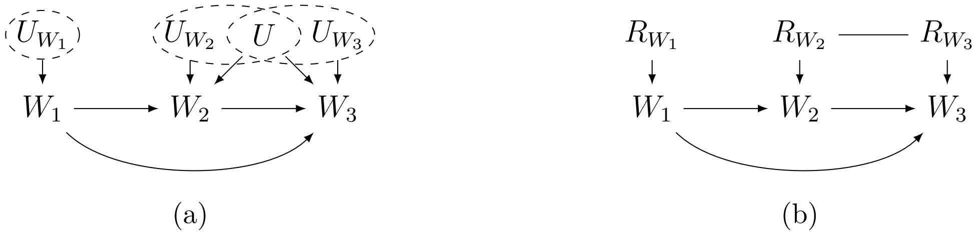


Figure 1: Example DAG to illustrate the concepts and notation. In this example, the measured variables are W_1, W_2, and W_3, and the remaining are unmeasured. Since the measured variables are categorical, an equivalent representation of (a) is given in (b), where $R_{W_1}, R_{W_2}, R_{W_2}$ are categorical response function variables.

Using this discretization, we can enumerate the relationships between the observable probabilities and the distribution of the response function variables. If those relationships are linear, then they define linear constraints in an optimization problem. Next, we describe a general class of DAGs having linear relationships between their distributions of response function variables and distributions of observable variables.

3 A Class of Linear DAGs

To characterize our class of linear problems, the set $\mathcal{W}$ is divided into two subsets $\mathcal{W} = \{\mathcal{W}_{\mathcal{L}}, \mathcal{W}_{\mathcal{R}}\}$, where $\mathcal{W}_{\mathcal{L}}$ may be empty. We assume without loss of generality that the indices of the variables are ordered in such a way that $\mathcal{L} = \{1, \ldots, \beth\}$ and $\mathcal{R} = \{\beth{+}1, \ldots, n\}$, where $\beth$ may be 0 in which case $\mathcal{L}$ is the empty set. We will denote the corresponding subdivisions of the vectors $\mathbf{W}$ and $\mathbf{R}$ by $(\mathbf{W}_{\mathcal{L}}, \mathbf{W}_{\mathcal{R}})$ and $(\mathbf{R}_{\mathcal{L}}, \mathbf{R}_{\mathcal{R}})$, respectively, and likewise for their lowercase value-vector counterparts. $\mathcal{L}$ and $\mathcal{R}$, connote *left* and *right* sides, where the causal paths flow from left to right. We make this division because in our class of problems, the $\mathcal{L}$-side variables are unconfounded with the $\mathcal{R}$-side variables.

Let $B := |\nu(\mathbf{W})| = \prod_{i=1}^{n} |\nu(W_i)| = \prod_{i=1}^{n} c_{W_i}$, and let $\{1, \ldots, B\} \ni b \mapsto \mathbf{w}_b \in \nu(\mathbf{W})$ be an enumeration of $\nu(\mathbf{W})$ that preserves the ordering of the $\mathcal{L}$-indices before the $\mathcal{R}$-indices such that $\forall b \in \{1, \ldots, B\}$, $\mathbf{w}_{b,\mathcal{L}} := (\mathbf{w}_b)_{\mathcal{L}}$ and $\mathbf{w}_{b,\mathcal{R}} := (\mathbf{w}_b)_{\mathcal{R}}$. Let $\mathbf{p}^*, \mathbf{p} \in [0,1]^B$ be given by $\forall b \in \{1, \ldots, B\}$, $p_b^* := p\{\mathbf{W} = \mathbf{w}_b\} = p\{(\mathbf{W}_{\mathcal{L}}, \mathbf{W}_{\mathcal{R}}) = (\mathbf{w}_{b,\mathcal{L}}, \mathbf{w}_{b,\mathcal{R}})\}$ and $p_b := p\{\mathbf{W}_{\mathcal{R}} = \mathbf{w}_{b,\mathcal{R}} \mid \mathbf{W}_{\mathcal{L}} = \mathbf{w}_{b,\mathcal{L}}\}$. Thus, the vector $\mathbf{p}^*$ represents the joint distribution of all observed variables and the vector $\mathbf{p}$ contains the observed conditional distribution of all variables in $\mathcal{W}_{\mathcal{R}}$ given all variables in $\mathcal{W}_{\mathcal{L}}$. As shown in Proposition 2, we will only need to observe the components of $\mathbf{p}$.

We will focus on the response function variables of the $\mathcal{R}$-side, and will provide them a dedicated enumeration. Let

$$\aleph_{\mathcal{R}} := |\nu(\mathbf{R}_{\mathcal{R}})| = \prod_{i=\gimel+1}^{n} |\nu(R_{W_i})| = \prod_{j=\gimel+1}^{n} c_{W_j}^{\prod_{V \in \mathbf{Pa}_{W_j}} c_V}.$$

Let $\{1, \ldots, \aleph_{\mathcal{R}}\} \ni \gamma \mapsto \mathbf{r}_\gamma \in \nu(\mathbf{R}_{\mathcal{R}})$ enumerate $\nu(\mathbf{R}_{\mathcal{R}})$ and $\mathbf{q} \in [0,1]^{\aleph_{\mathcal{R}}}$ be given by $\forall \gamma \in \{1, \ldots, \aleph_{\mathcal{R}}\}$, $q_\gamma := p\{\mathbf{R}_{\mathcal{R}} = \mathbf{r}_\gamma\}$. In particular, the vector $\mathbf{q}$ contains the joint probability distribution of the response function variables $\mathbf{R}_{\mathcal{R}}$. For $i \in \mathcal{R}$ and a fixed value-vector $\mathbf{w}_{\mathcal{L}} \in \nu(\mathbf{W}_{\mathcal{L}})$, we let

$$g_{W_i}(\mathbf{w}_{\mathcal{L}}, \mathbf{r}_\gamma) := f_{W_i}(w_{i1}, \ldots, w_{il_i}, g^*_{W_{il_i+1}}(\mathbf{r}_\gamma), \ldots, g^*_{W_{ik_i}}(\mathbf{r}_\gamma), r_{W_i}) = w_i,$$

where $w_{i1}, \ldots, w_{il_i}$ are the values of the parents of W_i that are in $\mathcal{W}_{\mathcal{L}}$, and $W_{il_i+1}, \ldots, W_{ik_i}$ are the parents of W_i that are in $\mathcal{W}_{\mathcal{R}}$.

Proposition 2. *Let G be a causal DAG satisfying the following Conditions:*

1. *Any edge that connects two variables $W_{\mathcal{L}} \in \mathcal{W}_{\mathcal{L}}$ and $W_{\mathcal{R}} \in \mathcal{W}_{\mathcal{R}}$ must be directed from $W_{\mathcal{L}}$ to $W_{\mathcal{R}}$.*

2. *There exists no unmeasured variable U that has children in both $\mathcal{W}_{\mathcal{L}}$ and $\mathcal{W}_{\mathcal{R}}$. That is, the variables in $\mathcal{W}_{\mathcal{L}}$ and $\mathcal{W}_{\mathcal{R}}$ are not confounded with each other.*

3. *There exists an unmeasured variable $U_{\mathcal{L}}$ such that $U_{\mathcal{L}}$ is a parent of W_i for all $i \in \mathcal{L}$. That is, all variables in $\mathcal{L}$ share an unmeasured common cause.*

4. *There exists an unmeasured variable $U_{\mathcal{R}}$ such that $U_{\mathcal{R}}$ is a parent of W_i for all $i \in \mathcal{R}$. That is, all variables in $\mathcal{R}$ share an unmeasured common cause,*

Then there exist matrices $P \in \{0,1\}^{B \times \aleph_{\mathcal{R}}}$, $P^* \in [0,1]^{B \times \aleph_{\mathcal{R}}}$ and $\Lambda \in [0,1]^{B \times B}$ such that $\mathbf{p} = P\mathbf{q}$, $\mathbf{p}^* = P^*\mathbf{q}$, Λ is diagonal with non-zero diagonal entries, $\Lambda P = P^*$, $\mathbf{p}^* = \Lambda \mathbf{p}$, and there are no other constraints on the distribution of response function variables that are not redundant with these.

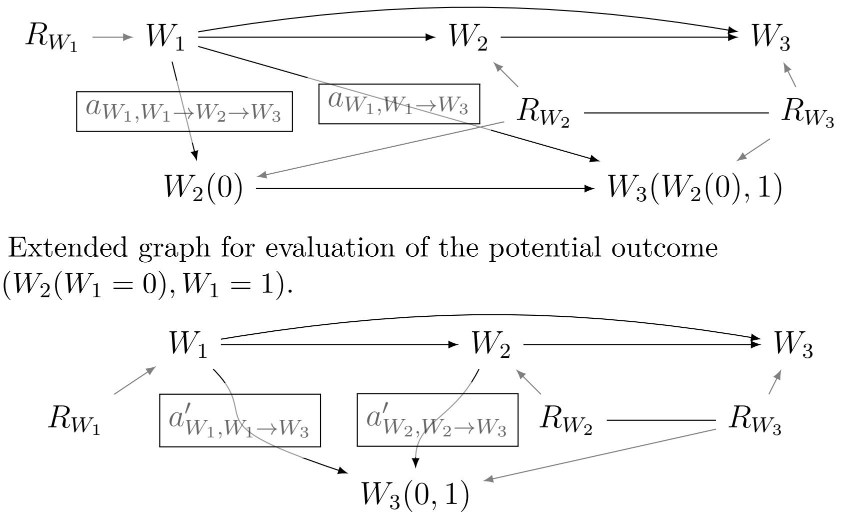
See the Supplementary Materials for proof. Conditions 1 and 2 ensure that the linear relations are necessary for the distribution to be compatible with the causal model, while the additional conditions 3 and 4 ensure that they are also sufficient. Though Proposition 2 guarantees their existence, it may not be trivial to construct these linear relations. Algorithm 1 below details a method for constructing the matrices P, P^* and Λ.

```
Result: Systems of linear equations relating $\mathbf{p}^*$ and $\mathbf{p}$ to $\mathbf{q}$
Initialize $P$ as a $B \times \aleph_{\mathcal{R}}$ matrix of 0s;
Initialize $P^*$ as a $B \times \aleph_{\mathcal{R}}$ matrix of 0s;
Initialize $\Lambda$ as a $B \times B$ matrix of 0s;
for $b \in 1, \ldots, B$ do
    for $\gamma \in 1, \ldots, \aleph_{\mathcal{R}}$ do
        Initialize $\omega$ as an empty vector of length $|\mathcal{R}|$ $(= n - \beth)$;
        for $i \in \mathcal{R}$ do
            Set $\omega_i := g_{W_i}(\mathbf{w}_{b,\mathcal{L}}, \mathbf{r}_\gamma)$;
        end
        if $\omega = \mathbf{w}_{b,\mathcal{R}}$ then
            $P_{b,\gamma} := 1$;
            $\Lambda_{b,b} := p\{\mathbf{W}_{\mathcal{L}} = \mathbf{w}_{b,\mathcal{L}}\}$;
            $P^*_{b,\gamma} := p\{\mathbf{W}_{\mathcal{L}} = \mathbf{w}_{b,\mathcal{L}}\}$;
        end
    end
end
```

Algorithm 1: An algorithm to determine a system of linear equations relating $\mathbf{p}$ and $\mathbf{p}^*$ to $\mathbf{q}$.

4 Functional expressions incorporating interventions

In order to determine the values of variables of interest for potential outcomes that incorporate interventions, we must also define a procedure for evaluating a functional expression that allows for variables to be externally forced to certain values. As a first step, we consider extended DAGs, which add additional nodes for potential outcomes of interest as in Balke and Pearl [1994b]. These are called twin networks in Pearl [2009], Chapter 7. Two examples are shown in Figure 2a and 2b. For each potential outcome of interest, nodes are added such that the corresponding factual and potential outcome nodes share the same response function variables. Edges that connect factual nodes to potential outcome nodes are labelled with letters that denote intervention sets indexed by the tail variable of that edge and the path to the head of that edge sequence. These sets define the variables being externally set, the values that they are being set to, and their indices indicate for which edge sequences they apply.



(a) Extended graph for evaluation of the potential outcome $W_3(W_2(W_1 = 0), W_1 = 1)$.

(b) Extended graph for evaluation of the potential outcome $W_3(W_2 = 0, W_1 = 1)$.

Figure 2: Extended DAGs to illustrate that multiple intervention sets are needed to define certain potential outcomes. In these two examples, the variables are binary.

Balke and Pearl [1994a] considered cases where we externally force a single subset of the variables to some fixed values. This construction suffices for the examples they consider, but not for defining and bounding effects like the natural direct effect of W_1 in the graph in Figure 2a whose first term is $p\{W_3(W_2(W_1 = 0), W_1 = 1) = 1\}$. In that expression, we see that the variable W_1, which is a parent of both W_3 and W_2, is simultaneously being set to 0 and 1, the difference being which child is in question. As another example, the causal query $p\{W_3(W_2(W_1 = 0)) = 1, W_2(W_1 = 1) = 1\}$ is a joint probability statement, and the two events in question are under different fixed values of W_1. Therefore, to be completely general, the variables that one assign to values cannot be a single set; the values that variables are being externally forced to may depend on which children are being considered and also on the term of the probability statement. Thus we define an extended function expression, which "remembers" the path of edges taken to get the value that is being determined at each call.

For $i \in \{\beth + 1, \ldots, n\}$, let A_i be a matrix that encodes the interventions and variables on which to intervene, with rows indexed by l corresponding to the variables in $\mathcal{W}$ and the columns indexed by j corresponding to all possible paths terminating at W_i; the entries in row l are in $\nu(W_l) \cup \{\emptyset\}$. The desired interventions within the causal query then define the entries of A_i which are denoted a_{lj}. In our procedure for evaluating potential outcomes, there is a distinct interventional matrix A_i corresponding to each outcome variable W_i used in the causal query. We define the procedure for evaluating the interventional response functional for an outcome variable W_i as

$$w_i = h_{W_i}^{A_i}\big(\mathbf{r}, W_i\big),$$

where for all $l \in \{1, \ldots, n\}$, all $\mathbf{r} \in \nu(\mathbf{R})$ and all strings j representing paths to W_i, we define $h_{W_i}^{A_i}(\mathbf{r}, j)$ recursively by

$$h^{A_i}_{W_i}(\mathbf{r}, j) := \begin{cases} a_{lj} & \text{if } a_{lj} \neq \emptyset \\ f_{W_i}(r_{W_i}) & \text{if } a_{lj} = \emptyset \text{ and } \mathbf{Pa}_{W_i} = \emptyset \\ f_{W_i}(h^{A_i}_{W_{i1}}(\mathbf{r}, W_{i1} \to j), \ldots, h^{A_i}_{W_{ik_i}}\big(\mathbf{r}, W_{ik_i} \to j), r_{W_i}\big) & \text{otherwise,} \end{cases}$$

where $k_i := |\mathbf{Pa}_{W_i}|$ and $\{W_{i1}, \ldots, W_{ik_i}\} := \mathbf{Pa}_{W_i}$, and the notation $i \to j$ means that $i \to$ is prepended to j. This notation allows us to trace the full path taken from the outcome of interest to the variable being intervened upon.

For example, considering the DAG in Figure 2a and the causal query $p\{W_3(W_2(W_1 = 0), W_1 = 1) = 1\}$, we have the interventional matrix

$$A_3 = \left[\begin{array}{c|cccc} & W_1 \to W_2 \to W_3 & W_1 \to W_3 & W_2 \to W_3 & W_3 \\ \hline W_1 & 0 & 1 & \emptyset & \emptyset \\ W_2 & \emptyset & \emptyset & \emptyset & \emptyset \\ W_3 & \emptyset & \emptyset & \emptyset & \emptyset \end{array}\right].$$

Thus, evaluating the functional expression $w_3 = h^{A_3}_{W_3}(\mathbf{r}, W_3)$ results (since W_3 is not intervened upon and $\mathbf{Pa}_{W_3} = \{W_1, W_2\}$) in

$$w_3 = h^{A_3}_{W_3}(\mathbf{r}, W_3) = f_{W_3}(w_1{=}h^{A_3}_{W_1}(\mathbf{r}, W_1 \to W_3), w_2{=}h^{A_3}_{W_2}(\mathbf{r}, W_2 \to W_3), r_{W_3}).$$

For the first argument of that function call we have $w_1 = h^{A_3}_{W_1}(\mathbf{r}, W_1 \to W_3) = a_{1,W_1 \to W_3} = 1$.

Then for the second argument, $a_{2,W_2 \to W_3} = \emptyset$ and $\mathbf{Pa}_{W_2} = \{W_1\}$, so we recurse, giving

$$w_2 = h^{A_3}_{W_2}(\mathbf{r}, W_2 \to W_3) = f_{W_2}(w_1{=}h^{A_3}_{W_1}(\mathbf{r}, W_1 \to W_2 \to W_3), r_{W_2}).$$

Now, $w_1 = h^{A_3}_{W_1}(\mathbf{r}, W_1 \to W_2 \to W_3) = a_{1,W_1 \to W_2 \to W_3} = 0$, giving $w_2 = f_{W_2}(w_1{=}0, r_{W_2})$, so we get $w_3 = f_{W_3}(w_1{=}1, w_2{=}f_{W_2}(w_1{=}0, r_{W_2}), r_{W_3})$.

For the DAG in Figure 2b and the first part of the causal query $p\{W_3(W_2 = 0, W_1 = 1) = 1\}$, we have

$$A_3 = \left[\begin{array}{c|cccc} & W_1 \to W_2 \to W_3 & W_1 \to W_3 & W_2 \to W_3 & W_3 \\ \hline W_1 & \emptyset & 1 & \emptyset & \emptyset \\ W_2 & \emptyset & \emptyset & 0 & \emptyset \\ W_3 & \emptyset & \emptyset & \emptyset & \emptyset \end{array}\right].$$

Thus, evaluating the functional expression $w_3 = h^{A_3}_{W_3}(\mathbf{r}, W_3)$ results in

$$w_3 = h^{A_3}_{W_3}(\mathbf{r}, W_3) = f_{W_3}(w_1{=}h^{A_3}_{W_1}(\mathbf{r}, W_1 \to W_3), w_2{=}h^{A_3}_{W_2}(\mathbf{r}, W_2 \to W_3), r_{W_3}).$$

For the first argument of that function call we have $w_1 = h^{A_3}_{W_1}(\mathbf{r}, W_1 \to W_3) = a_{1,W_1 \to W_3} = 1$. Then, for the second argument, $w_2 = h^{A_3}_{W_2}(\mathbf{r}, W_2 \to W_3) = a_{2,W_2 \to W_3} = 0$, giving the result $w_3 = f_{W_3}(w_1{=}1, w_2{=}0, r_{W_3})$.

The procedures for evaluating the functions g and h^{A_i} are sufficient to translate any combined factual and/or potential outcome joint probability statement into probability statements involving only the response function variables $\mathbf{R}$. Thus, using our response function formulation, any potential outcome or factual joint probability statement can be written

$$Q := p\{h^{A_{i_1}}_{W_{i_1}}(\mathbf{R}, W_{i_1}) = w_{i_1}, \ldots, h^{A_{i_P}}_{W_{i_P}}(\mathbf{R}, W_{i_P}) = w_{i_P},$$
$$g_{W_{j_1}}(\mathbf{R}) = w_{j_1}, \ldots, g_{W_{j_O}}(\mathbf{R}) = w_{j_O}\}, \tag{1}$$

where $\mathcal{P} = \{i_1, \ldots, i_P\}$ denote the indices of potential outcomes, and $\mathcal{O} = \{j_1, \ldots, j_O\}$ the indices of the factual outcomes (and these sets may be overlapping). Given the functional expressions we have defined and our procedures for evaluating them, we can therefore write

$$Q = \sum_{\mathbf{r} \in \Gamma(Q)} p\{\mathbf{R} = \mathbf{r}\}, \text{ where}$$

$$\Gamma(Q) := \{\mathbf{r} \in \nu(\mathbf{R}) : \forall i_p \in \mathcal{P}, w_{i_p} = h^{A_{i_p}}_{W_{i_p}}(\mathbf{r}, W_{i_p}) \text{ and } \forall j_o \in \mathcal{O}, w_{j_o} = g_{W_{j_o}}(\mathbf{r})\}.$$

We will call an expression of this form an *atomic* query. Their form is completely general, and allows arbitrarily nested potential outcomes, and combinations with observational quantities. We will combine atomic queries to obtain causal contrasts of interest, such as the causal risk difference.

Proposition 3. *Let G be a causal DAG satisfying Conditions 1 and 2, and let Q be an atomic query satisfying the following Conditions:*

5. *Each atomic query is a probability as given in Equation* (1) *where* $i_1, \ldots, i_P, j_1, \ldots, j_O \in \mathcal{R}$ *(i.e., all outcome variables must be in $\mathcal{W}_\mathcal{R}$) and*

6. *if $\mathcal{L} \neq \varnothing$ then: (i) none of the variables in $\mathcal{W}_\mathcal{L}$ that are intervened upon can have any children in $\mathcal{W}_\mathcal{L}$, (ii) all variables in $\mathcal{W}_\mathcal{L}$ must be in the intervention set, or ancestors of the variables in the intervention set (here the intervention set refers to variables*

in the rows of the A matrices that are not $\emptyset$), (iii) no observations are allowed, i.e, $\mathcal{O} = \varnothing$.

Then there exists a constant binary vector $\alpha \in \{0, 1\}^{\aleph_{\mathcal{R}}}$ such that $Q = \alpha^{\top}\mathbf{q}$.

See the Supplementary Materials for proof. A procedure for construction of this α is detailed in Algorithm 2 which converts the atomic query Q into a binary linear combination of probabilities of response function variables of the $\mathcal{R}$-side.

```
Result: $Q$ expressed as a simple sum of a subset of the components of $\mathbf{q}$.
Initialize $\alpha \in \{0,1\}^{\aleph_{\mathcal{R}}}$ by $\forall \gamma \in \{1, \ldots, \aleph_{\mathcal{R}}\}$, $\alpha_\gamma := 1$;
Let $\mathcal{P}, \mathcal{O}$ be the index sets as defined above corresponding to $Q$;
for $\gamma \in 1, \ldots, \aleph_{\mathcal{R}}$ do
    for $l \in \mathcal{P}$ do
        Construct $A_l$ according to $l$;
        Compute $\omega := h^{A_l}_{W_l}(\mathbf{r}_\gamma, W_l)$;
        if $\omega \neq w_l$ then
            Set $\alpha_\gamma := 0$;
            break;
        end
    end
    if $\alpha_\gamma = 0$ then
        break;
    end
    for $l \in \mathcal{O}$ do
        Compute $\omega := g_{W_l}(\mathbf{r}_\gamma)$;
        if $\omega \neq w_l$ then
            Set $\alpha_\gamma := 0$;
            break;
        end
    end
end
```

Algorithm 2: Converting Q to a binary linear combination of $\mathbf{q}$.

The following corollary, which specifies the general form of a causal query, follows immediately since linear combinations of linear combinations again are just linear combinations.

Corollary 1. *Let Q^* be any real linear combination of atomic queries (in particular, Q may be a classic linear causal contrast such as a causal risk difference). Under conditions 1, 2, 5, and 6, there exists a constant vector $\alpha^* \in \mathbb{R}^{\aleph_{\mathcal{R}}}$ such that $Q^* = \alpha^{*\top}\mathbf{q}$.*

The algorithms are formulated so that bounds are derived in terms of the true probabilities of the observed variables in $\mathcal{W}_{\mathcal{R}}$ conditional on the variables in $\mathcal{W}_{\mathcal{L}}$. Provided one is not intervening on any of the variables in $\mathcal{W}_{\mathcal{L}}$, Conditions 1 and 2 imply that the directions of the edges within $\mathcal{W}_{\mathcal{L}}$ cannot influence the bounds. That is, the bounds are tight for the equivalence class of DAGs that contains the set of DAGs for all possible directions of edges among variables in $\mathcal{W}_{\mathcal{L}}$. For example, the bounds computed for a query such as $p\{Y(X = 1) = 1\}$ are tight and equal for both of the DAGs in Figures 3 (a) and (b). In either case, the knowledge of whether Z causes $Z2$ or vice versa does not influence the bounds because both of those variables are conditioned upon in the algorithm.

Alternatively, if the desired query was $p\{Y(X(Z = 1)) = 1\}$, the DAGs in Figures 3 (a) and (b) may not result in the same bounds, and in fact, the causal problem under Figure 3 (a) may not be linear. As required by Conditions 5 and 6, if we intervene upon a variable in $\mathcal{W}_{\mathcal{L}}$, then the direction of edges within $\mathcal{W}_{\mathcal{L}}$ matters, and in fact if the intervened upon variable has a child also in $\mathcal{W}_{\mathcal{L}}$, the condition will not be met.

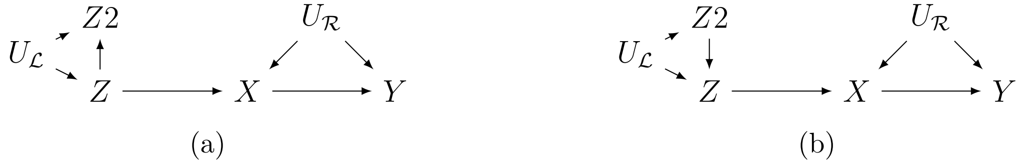


Figure 3: An equivalence class of DAGs defined by arbitrary connections in $\mathcal{W}_{\mathcal{L}}$. Bounds for causal queries that involve intervening on X that meet our conditions are equivalent and tight for these two graphs in (a) and (b).

5 Optimization via vertex enumeration

After applying Algorithms 1 and 2, we have a linear objective and a system of linear constraints. We also have the probabilistic constraints:

$$\forall \mathbf{w}_{\mathcal{L}} \in \nu(\mathbf{W}_{\mathcal{L}}), \qquad \sum_{\mathbf{w}_{\mathcal{R}} \in \nu(\mathbf{W}_{\mathcal{R}})} p\{\mathbf{W}_{\mathcal{R}} = \mathbf{w}_{\mathcal{R}} \mid \mathbf{W}_{\mathcal{L}} = \mathbf{w}_{\mathcal{L}}\} = 1$$

and

$$\sum_{\gamma=1}^{\aleph_{\mathcal{R}}} q_\gamma = \sum_{\gamma=1}^{\aleph_{\mathcal{R}}} p\{\mathbf{R}_{\mathcal{R}} = \mathbf{r}_\gamma\} = \sum_{\mathbf{r}_{\mathcal{R}} \in \nu(\mathbf{R}_{\mathcal{R}})} p\{\mathbf{R}_{\mathcal{R}} = \mathbf{r}_{\mathcal{R}}\} = 1.$$

Additional linear constraints on $\mathbf{q}$ can be optionally given as $B\mathbf{q} \geq \mathbf{d}$ where B and $\mathbf{d}$ are respectively a matrix and vector of real constants. These constraints can be used to encode assumptions about the response functions that are not possible to encode in a DAG, for example, restricting the probabilities of implausible response patterns. We thus arrive at the following linear programming problem for the lower bound; the upper bound is given by the corresponding maximization problem.

$$\begin{aligned} \text{minimize } Q \quad &= \alpha^T \mathbf{q} \\ \text{subject to } P\mathbf{q} \quad &= \mathbf{p}, \\ B\mathbf{q} \quad &\geq \mathbf{d}, \\ \mathbf{q} \geq \mathbf{0}, \quad \text{and } \mathbf{1}^T\mathbf{q} \quad &= 1. \end{aligned}$$

Note that the constraint space constitutes a bounded (due to the probabilistic constraints) convex polytope. By the fundamental theorem of linear programming, the global extrema must occur at one of the vertices of the polytope. We can thus solve this problem symbolically by applying an efficient vertex enumeration algorithm, such as the double description algorithm [Motzkin et al., 1953, Fukuda, 2018] to enumerate the vertices of the polytope of the dual linear program. For instance, the dual of the minimization problem above is given by

$$\begin{aligned} \text{maximize} \quad & \begin{pmatrix} \mathbf{d}^T & 1 & \mathbf{p}^T \end{pmatrix} \mathbf{y} \\ \text{subject to} \quad & \begin{pmatrix} B^T & 1 & P^T \\ I & & 0 \end{pmatrix} \mathbf{y} \leq \begin{pmatrix} \alpha \\ \mathbf{0} \end{pmatrix}. \end{aligned}$$

So by the strong duality theorem, the optimum of the dual, and thus also of the primal problem, is of the form $\begin{pmatrix} \mathbf{d}^T & 1 & \mathbf{p}^T \end{pmatrix} \bar{\mathbf{y}}$ where $\bar{\mathbf{y}}$ is a vertex of the polytope $\{\mathbf{y} : \begin{pmatrix} B^T & 1 & P^T \\ I & & 0 \end{pmatrix} \mathbf{y} \leq \begin{pmatrix} \alpha \\ \mathbf{0} \end{pmatrix}\}$. This gives a lower bound on the causal effect of interest as the maximum of a set of expressions involving only observable probabilities. Similarly, the upper bound is given by reversing the dual inequality and minimizing over the corresponding polytope.

Proposition 4. *Under conditions 1-6 and subject to any additional linear constraints of the form $B\mathbf{q} \geq \mathbf{d}$, the procedure above yields valid and tight symbolic bounds for a causal query that is a linear combination of atomic queries.*

Corollary 2. *If condition 4 does not hold, then the bounds derived using the above procedure are still valid.*

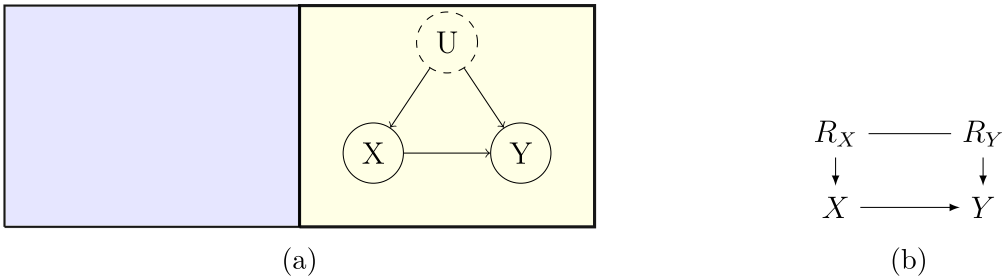
See the Supplementary Materials for proof. The conditions 3 and 4 represent a worst-case scenario of confounding and ensure that the decompositions giving rise to the linear constraints cannot be further factorized to yield more granular but non-linear constraints. If however there is any known (partial) absence of such confounding, then these bounds are still valid, and may be narrow enough to be informative, while not necessarily tight. Such an absence of confounding on the $\mathcal{R}$-side implies some independence among the $\mathbf{R}_{\mathcal{R}}$ variables, and hence additional constraints on their distribution. Thus the true feasible space may be smaller than the one considered in our algorithm, but completely contained inside it.

6 Examples

The graphs in the following examples are divided into a left side, which corresponds to the $\mathcal{W}_{\mathcal{L}}$ set, and a right side, which corresponds to the $\mathcal{W}_{\mathcal{R}}$ set, as in Figure 4a. The left side is displayed as a violet (dark grey) box, and the right side a yellow (light grey) box.

6.1 Confounded exposure and outcome

The basic DAG with two variables that are confounded as shown in Figure 4a conforms to our class of models. In this case, the variable X is the exposure of interest, and Y the outcome of interest. X and Y have a common, unmeasured cause U. We specify X and Y to be ternary and binary respectively, so X takes values in $\{0, 1, 2\}$ and Y in $\{0, 1\}$. Our causal effects of interest are the risk differences $p\{Y(X = 2) = 1\} - P\{Y(X = 0) =$



(a) (b)

Figure 4: Simple confounded example and the equivalent response function variable graph.

$1\}, p\{Y(X=2)=1\} - P\{Y(X=1)=1\}$ and $p\{Y(X=1)=1\} - P\{Y(X=0)=1\}$, and we have no additional constraints to specify.

Here we have two variables and therefore two response function variables. The response function variable formulation of the graph in Figure 4b is an equivalent representation of the causal model. The following tables define the values of the response functions and variables:

$x = f_X(r_X)$	
$r_X = 0$	$x = 0$
$r_X = 1$	$x = 1$
$r_X = 2$	$x = 2$

$y = f_Y(x, r_Y)$	$x = 0$	$x = 1$	$x = 2$
$r_Y = 0$	$y = 0$	$y = 0$	$y = 0$
$r_Y = 1$	$y = 1$	$y = 0$	$y = 0$
$r_Y = 2$	$y = 0$	$y = 1$	$y = 0$
$r_Y = 3$	$y = 1$	$y = 1$	$y = 0$
$r_Y = 4$	$y = 0$	$y = 0$	$y = 1$
$r_Y = 5$	$y = 1$	$y = 0$	$y = 1$
$r_Y = 6$	$y = 0$	$y = 1$	$y = 1$
$r_Y = 7$	$y = 1$	$y = 1$	$y = 1$

R_X is a random variable that can take on 3 possible values, and R_Y is a random variable that can take on $2^3 = 8$ possible values. Thus, the joint distribution of (R_X, R_Y) is characterized by $3 \cdot 8 = 24$ parameters, say $q_{i,j}$, where $i \in \{0, 1, 2\}$ and $j \in \{0, 1, 2, 3, 4, 5, 6, 7\}$.

Applying Algorithm 1, we can relate the $3 \cdot 2 = 6$ observed probabilities to the parameters of the response function variable distribution as follows:

$$
\begin{aligned}
p_{0,0;} := p\{X = 0, Y = 0\} &= q_{0,0} + q_{0,2} + q_{0,4} + q_{0,6} \\
p_{1,0;} := p\{X = 1, Y = 0\} &= q_{1,0} + q_{1,1} + q_{1,4} + q_{1,5} \\
p_{2,0;} := p\{X = 2, Y = 0\} &= q_{2,0} + q_{2,1} + q_{2,2} + q_{2,3} \\
p_{0,1;} := p\{X = 0, Y = 1\} &= q_{0,1} + q_{0,3} + q_{0,5} + q_{0,7} \\
p_{1,1;} := p\{X = 1, Y = 1\} &= q_{1,2} + q_{1,3} + q_{1,6} + q_{1,7} \\
p_{2,1;} := p\{X = 2, Y = 1\} &= q_{2,4} + q_{2,5} + q_{2,6} + q_{2,7}.
\end{aligned}
$$

We get

$$
A = \left[\begin{array}{c|cc} & X \to Y & Y \\ \hline X & 0 & \emptyset \\ Y & \emptyset & \emptyset \end{array}\right], \text{ for } p\{Y(X = 0) = 1\},
$$

$$
A = \left[\begin{array}{c|cc} & X \to Y & Y \\ \hline X & 1 & \emptyset \\ Y & \emptyset & \emptyset \end{array}\right], \text{ for } p\{Y(X = 1) = 1\} \text{ and}
$$

$$
A = \left[\begin{array}{c|cc} & X \to Y & Y \\ \hline X & 2 & \emptyset \\ Y & \emptyset & \emptyset \end{array}\right], \text{ for } p\{Y(X = 2) = 1\}.
$$

Applying Algorithm 2, we get

$$
\begin{aligned}
p\{Y(X=0)=1\} &= q_{0,1}+q_{0,3}+q_{0,5}+q_{0,7} \\
&\quad + q_{1,1}+q_{1,3}+q_{1,5}+q_{1,7} \\
&\quad + q_{2,1}+q_{2,3}+q_{2,5}+q_{2,7},
\end{aligned}
$$

$$
\begin{aligned}
p\{Y(X=1)=1\} &= q_{0,2}+q_{0,3}+q_{0,6}+q_{0,7} \\
&\quad + q_{1,2}+q_{1,3}+q_{1,6}+q_{1,7} \\
&\quad + q_{2,2}+q_{2,3}+q_{2,6}+q_{2,7}, \text{ and} \\
p\{Y(X=2)=1\} &= q_{0,4}+q_{0,5}+q_{0,6}+q_{0,7} \\
&\quad + q_{1,4}+q_{1,5}+q_{1,6}+q_{1,7} \\
&\quad + q_{2,4}+q_{2,5}+q_{2,6}+q_{2,7},
\end{aligned}
$$

hence the contrasts

$$\begin{aligned}
p\{Y(X=1)=1\} - p\{Y(X=0)=1\} &= q_{0,2} + q_{0,6} + q_{1,2} + q_{1,6} + q_{2,2} + q_{2,6} \\
&\quad - q_{0,1} - q_{0,5} - q_{1,1} - q_{1,5} - q_{2,1} - q_{2,5}, \\
p\{Y(X=2)=1\} - p\{Y(X=0)=1\} &= q_{0,4} + q_{0,6} + q_{1,4} + q_{1,6} + q_{2,4} + q_{2,6} \\
&\quad - q_{0,1} - q_{0,3} - q_{1,1} - q_{1,3} - q_{2,1} - q_{2,3} \text{ and} \\
p\{Y(X=2)=1\} - p\{Y(X=1)=1\} &= q_{0,4} + q_{0,5} + q_{1,4} + q_{1,5} + q_{2,4} + q_{2,5} \\
&\quad - q_{0,2} - q_{0,3} - q_{1,2} - q_{1,3} - q_{2,2} - q_{2,3}.
\end{aligned}$$

Together with the probabilistic constraints, we then have the fully specified linear programming problem. The bounds as output by the program are

$$\begin{gathered}
p\{X=0,Y=0\} + p\{X=1,Y=1\} - 1 \\
\leq p\{Y(X=1)=1\} - p\{Y(X=0)=1\} \leq \\
1 - p\{X=1,Y=0\} - p\{X=0,Y=1\},
\end{gathered}$$

$$\begin{gathered}
-p\{X=1,Y=0\} - p\{X=2,Y=0\} - p\{X=0,Y=1\} - p\{X=1,Y=1\} \\
\leq p\{Y(X=2)=1\} - p\{Y(X=0)=1\} \leq \\
1 - p\{X=2,Y=0\} - p\{X=0,Y=1\}
\end{gathered}$$

and

$$\begin{gathered}
-p\{X=0,Y=0\} - p\{X=2,Y=0\} - p\{X=0,Y=1\} - p\{X=1,Y=1\} \\
\leq p\{Y(X=2)=1\} - p\{Y(X=1)=1\} \leq \\
1 - p\{X=2,Y=0\} - p\{X=1,Y=1\}.
\end{gathered}$$

6.2 Two instruments

Our next example is shown in the DAG in Figure 5. This extends the instrumental variable example to the case where there are two binary variables on the left side that may be associated with each other and that both have a direct effect on X, but no direct effect on Y. This situation may arise in Mendelian randomization studies, wherein multiple genes may be known to cause changes in an exposure but not directly on the outcome.

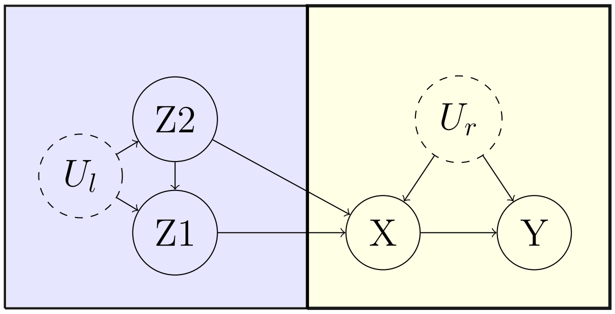


Figure 5: Two instrumental variables example with binary variables

The bounds on risk difference $p\{Y(X = 1)\} - p\{Y(X = 0)\}$ under this DAG can be computed using our method. In this problem, there are 16 constraints involving the conditional probabilities, the distribution of the response function variables of the $\mathcal{R}$-side has 64 parameters, and the causal query is a function of 32 of these parameters. The bounds are the extrema over 112 vertices, and are therefore too long to be presented simply, but they are included in the Supplementary Material along with code to reproduce the results using our method.

To illustrate these bounds, we computed them for specific values of observed probabilities generated from the model in Equation (2) which satisfies the DAG in Figure 5. Using these simulations we compare our bounds to the classic IV bounds from Balke and Pearl [1997] for a single binary instrument and to bounds derived using our method for a single but 4-level categorical instrument.

For each of 50,000 simulations, we generated values pu_l and pu_l of probabilities of the latent influences U_l and U_r from the standard uniform distribution, and each of 12 parameters $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \beta_1, \beta_2, \beta_3, \beta_4, \gamma_1, \gamma_2, \gamma_3$ from the normal distribution with mean 0 and standard deviation 2. Assuming that the conditional distributions of the observed variables follow probit models, we can derive, by Bayesian decomposition according to the diagram in Figure 5, the joint distribution of $p(U_l, U_r, Z1, Z2, X, Y)$. From that, we marginalize out the variables U_l and U_r to get $p\{Z1, Z2, X, Y\}$ and finally compute and divide this by the marginal joint probability $p\{Z1, Z2\}$ of the instruments $Z1$ and $Z2$, to get the conditional probability distribution $p\{X, Y|Z1, Z2\}$ that goes into the symbolic expressions of the tight bounds. We do a similar marginalization of $Z2$ in order to get conditional probabilities $p\{X, Y|Z1\}$ for computation of the single binary IV bounds. In each simulation, we create values of probabilities $p\{Z3 = z3\}, z3 \in \{0, 1, 2, 3\}$ of a 4-level instrument $Z3$ from probabilities $p\{Z1 = z1, Z2 = z2\}, z1, z2 \in \{0, 1\}$ to get appropriate input for the expressions of the tight bound computed in the single 4-level instrument setting.

The the widths of the classic IV bounds and the dual binary instruments are compared for a subsample of the simulations in Figure 6. The bounds with two instruments are never wider than the classic IV bounds with a single binary instrument. The simulations also verify that a single four level instrument yields exactly the same bounds as two binary ones. `R` code for these simulations are provided in the Supplementary Material.

$$
\begin{aligned}
p\{U_l = 1\} &\sim \text{Unif}(0,1) \\
p\{U_r = 1\} &\sim \text{Unif}(0,1) \\
p\{Z2 = 1 | U_l\} &= \Phi(\alpha_1 + \alpha_2 U_l) \\
p\{Z1 = 1 | U_l, Z2\} &= \Phi(\alpha_3 + \alpha_4 U_l + \alpha_5 Z2) \qquad (2) \\
p\{X = 1 | U_r, Z1, Z2\} &= \Phi(\beta_1 + \beta_2 U_r + \beta_3 Z1 + \beta_4 Z2) \\
p\{Y = 1 | U_r, X\} &= \Phi(\gamma_1 + \gamma_2 U_r + \gamma_3 X) \\
(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \beta_1, \beta_2, \beta_3, \beta_4, \gamma_1, \gamma_2, \gamma_3) &\sim N(0,4)
\end{aligned}
$$

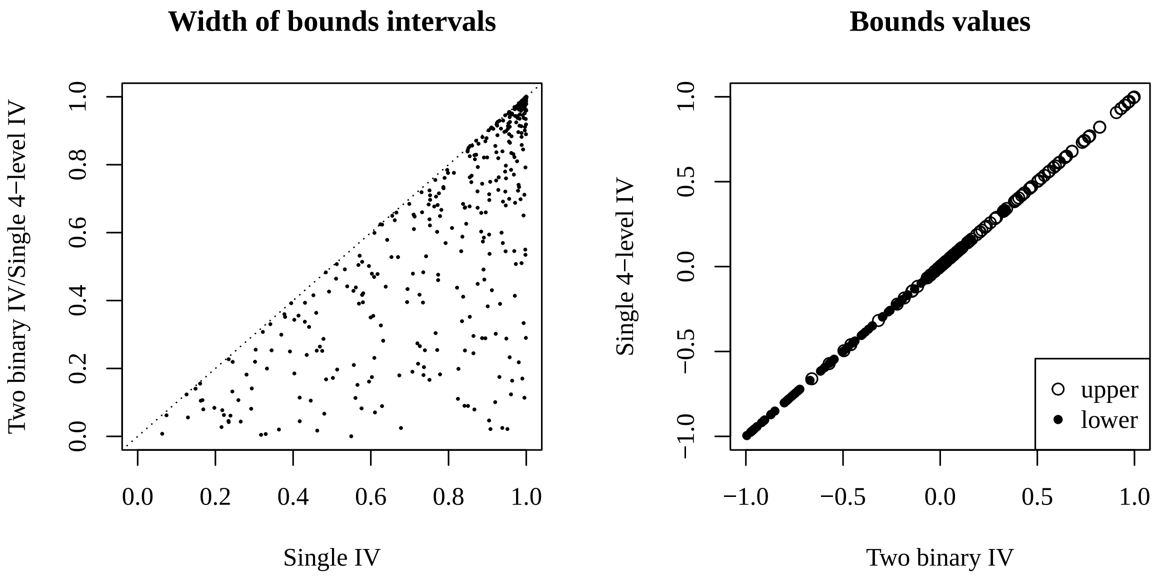


Figure 6: Under a DAG with two instruments, the left panel is a comparison of the width of the bounds intervals for the causal risk difference assuming only one of the instruments is observed to the width of the bounds assuming both are observed. The right panel compares the values of the upper and lower bounds for each replicate for two binary instruments versus a single 4-level instrument.

6.3 Measurement error in the outcome

Our final example illustrates some additional features of our method. In Figure 7, we have a binary variable X affecting a binary variable Y, but Y is not observed. Instead, the binary variable $Y2$ which is a child of Y is observed, and the effect of the true Y on the measured $Y2$ is confounded. Additionally, we would like to include a constraint that $Y2(Y = 1) \geq Y2(Y = 0)$, which is often called the monotonicity constraint. This constraint encodes the assumption that the outcome measured with error would not be equal to 0 unless the true unobserved outcome is also equal to 0. In terms of the response functions, this constraint removes the case where $f_{Y2}(y, r_{Y2}) = 1 - y$, thereby reducing the number of possible values that r_{Y2} can take by 1.

The fact that Y is unobserved implies that we have 4 possible conditional probabilities to work with; $p\{Y2 = y2|X = x\}$, for $y2, x \in \{0, 1\}$. There are 12 parameters that characterize the distribution of the response function variables of the $\mathcal{R}$-side, and 4 constraints involving conditional probabilities. The bounds for the risk difference $p\{Y(X = 1) = 1\} - p\{Y(X = 0) = 1\}$ derived using our method are given by

$$\begin{aligned}
&\max\{-1, 2\,p\{Y2 = 0|X = 0\} - 2\,p\{Y2 = 0|X = 1\} - 1\} \\
&\qquad \leq p\{Y(X = 1) = 1\} - p\{Y(X = 0) = 1\} \leq \\
&\min\{1, 2\,p\{Y2 = 0|X = 0\} - 2\,p\{Y2 = 0|X = 1\} + 1\}.
\end{aligned}$$

Except in cases where $p\{Y2 = 0|X = 0\} = p\{Y2 = 0|X = 1\}$, these bounds are informative; meaning they give an interval that is shorter than the a priori interval $[-1, 1]$.

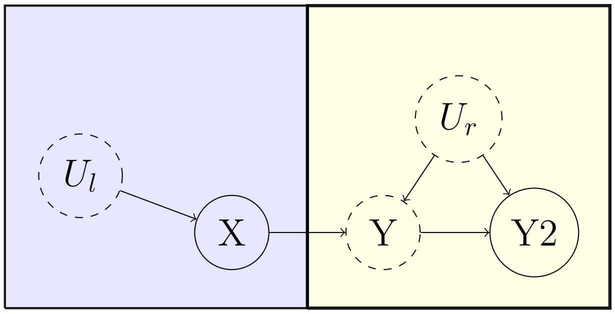


Figure 7: Example with measurement error in the outcome. Dashed circles indicate unobserved variables.

7 Conclusion and Discussion

We have described a general method for the symbolic computation of bounds on causal queries that are not identified from the true probability distribution of the observed variables. For this method, we give two algorithms for deriving the needed constraints and objective to construct such bounds. We describe a class of causal graphs and queries that will always define a linear program, for which we have shown the derived symbolic bounds will always be both valid and tight. We also show that under a broader class of problems our method will provide valid and possibly informative bounds that are not guaranteed to be tight.

Our approach is useful in several novel scenarios, as illustrated in the examples above. Additional applications of this method to unsolved problems in causal inference are now much more accessible to researchers as a class of problems for which linear programming can always be used is well-defined and clear algorithms exist for translating DAGs plus causal queries into linear programs. Our representation of causal estimands as arbitrarily nested counterfactuals and our procedure for translating them into functional expressions provides a significant advance over previous methods. This allows for bounding of cross-world counterfactual quantities which are highly relevant in mediation settings. The generality yet accessibility of the method all but guarantees that practitioners will find novel applications that we have not forseen.

Although our class of problems and method from deriving bounds puts no limit of the number of variables or categories for a given variable, in practice attention must be paid to computational complexity. Since we have $|\nu(R_{W_i})| = \prod_{i=1}^{n} c_{W_i}^{\prod_{V \in \mathbf{pa}_{W_i}} c_V}$ for each variable W_i, the cardinalities of the domains of the response function variables grow exponentially with the those of other variables in the DAG. The exact growth pattern will of course depend on the DAG and its connectivity as well as the number of categorical levels of select influential variables. Thus, the number of variables or levels may be limited by computing power.

It should be noted that our conditions for a class of problems to be linear are sufficient, but not necessary. Thus, we cannot rule out that there exist problems outside of our class that can be stated as linear. It may be possible to identify a broader class of problems or a different algorithm that may apply on a case-by-case basis. Nonlinear causal queries such as the relative risk or odds ratio yield nonlinear optimization problems yet in some cases it may be possible to translate them to equivalent linear problems. Measured confounding, or knowledge about the absence of confounding often implies nonlinear constraints. We have assumed that all variables are categorical, although many real scientific problems involve continuous variables. Extensions and insights into solving these sorts of problems would be useful in the causal inference community and are areas of future research for the authors.

Supplemental material

Supplementary Material available online includes proofs of the propositions. The `R` package `causaloptim`: An Interface to Specify Causal Graphs and Compute Bounds on Causal Effects, is available from CRAN, and from Github at `https://sachsmc.github.io/causaloptim`, with additional documentation and examples. The file `example-code.R` contains the R code used to run the examples and simulations presented in the main text.

References

A. Balke and J. Pearl. Counterfactual probabilities: Computational methods, bounds and applications. In *Proceedings of the Tenth international conference on Uncertainty in artificial intelligence*, pages 46–54. Morgan Kaufmann Publishers Inc., 1994a.

A. Balke and J. Pearl. Probabilistic evaluation of counterfactual queries. In *Proceedings of the twelfth national conference on artificial intelligence*, pages 230–237. The AAAI Press, Menlo Park, California., 1994b.

A. Balke and J. Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997.

B. Bonet. Instrumentality tests revisited. arXiv: 1301.2258, 2013.

G. B. Dantzig. *Linear Programming and Extensions.* Princeton University Press, 1963.

G. Duarte, N. Finkelstein, D. Knox, J. Mummolo, and I. Shpitser. An automated approach to causal inference in discrete settings. arXiv: 2109.13471, 2021.

K. Fukuda. *cdd, cddplus and cddlib homepage.* Swiss Federal Institute of Technology, Zurich., 2018. URL `https://people.inf.ethz.ch/fukudak/cdd_home/`.

J. J. Heckman and E. J. Vytlacil. Instrumental variables, selection models, and tight bounds on the average treatment effect. In *Econometric Evaluations of Active Labor Market Policies in Europe. Physica-Verlag*, 2001.

T. Motzkin, H. Raiffa, G. Thompson, and R. Thrall. The double description method. *Contributions to Theory of Games*, 2, 1953.

J. Pearl. *Causality*. Cambridge University Press, 2009.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL `https://www.R-project.org/`.

R. R. Ramsahai. Causal Bounds and Observable Constraints for Non-deterministic Models. *Journal of Machine Learning Research*, 13(29):829–848, 2012. URL `http://jmlr.org/papers/v13/ramsahai12a.html`.

Appendix A

Proof of Proposition 1. For each $W \in \mathcal{W}$, if $\phi_W : \nu(U_W) \to \{h : \nu(\mathbf{Pa}_W) \to \nu(W)\}$ is given by $u_W \mapsto h_{u_W}$, where the *response function* h_{u_W} is given by $\mathbf{pa}_W \mapsto F_W(\mathbf{pa}_W, u_W)$, then let the set of values of the *response function variable* R_W corresponding to W, $\nu(R_W) := \nu(U_W)/\phi_W$ be the partition of $\nu(U_W)$ induced by the equivalence relation $u_1 \sim u_2 :\iff \phi_W(u_1) = \phi_W(u_2)$. ϕ_W maps $\nu(U_W)$ bijectively to the finite set $\{h : \nu(\mathbf{Pa}_W) \to \nu(W)\}$ of response functions. Thus, for each $u_W \in \nu(U_W)$ there exists a unique $r_W \in \nu(R_W)$ and $f_W(\cdot, r_W) \in \{h : \nu(\mathbf{Pa}_W) \to \nu(W)\}$ such that $F_W(\cdot, u_W) = f_W(\cdot, r_W)$. We will henceforth refer to $f_W(\cdot, r_W)$ as the response function and R_W the response function variable. Note that the set $\{h : \nu(\mathbf{Pa}_W) \to \nu(W)\}$ is finite with cardinality $|\nu(W)|^{|\nu(\mathbf{Pa}_w)|}$ since $|\nu(W)|$ and $|\nu(\mathbf{Pa}_w)|$ are both finite. $\square$

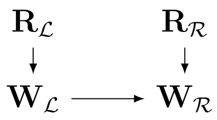


Figure 8: A birds-eye view of G in Proposition 2. G yields the Bayesian decomposition $p\{\mathbf{W}_\mathcal{L} = \mathbf{w}_\mathcal{L}, \mathbf{W}_\mathcal{R} = \mathbf{w}_\mathcal{R}, \mathbf{R}_\mathcal{L} = \mathbf{r}_\mathcal{L}, \mathbf{R}_\mathcal{R} = \mathbf{r}_\mathcal{R}\} = p\{\mathbf{W}_\mathcal{L} = \mathbf{w}_\mathcal{L} \mid \mathbf{R}_\mathcal{L} = \mathbf{r}_\mathcal{L}\}p\{\mathbf{R}_\mathcal{L} = \mathbf{r}_\mathcal{L}\}p\{\mathbf{W}_\mathcal{R} = \mathbf{w}_\mathcal{R} \mid \mathbf{W}_\mathcal{L} = \mathbf{w}_\mathcal{L}, \mathbf{R}_\mathcal{R} = \mathbf{r}_\mathcal{R}\}p\{\mathbf{R}_\mathcal{R} = \mathbf{r}_\mathcal{R}\}$.

Proof of Proposition 2. Conditions 1 and 2 are depicted in Figure 8. Note that this illustrates the setting at a macro-level only, and indicates only the independence relations between the vector-valued variables $\mathbf{W}_\mathcal{L}, \mathbf{W}_\mathcal{R}, \mathbf{R}_\mathcal{L}$ and $\mathbf{R}_\mathcal{R}$ at this level. The internal dependencies among the component variables of $\mathbf{W}_\mathcal{L}$ and $\mathbf{W}_\mathcal{R}$ are further given by the actual ”fine-grained” DAG G. Regarding the internal dependencies among the component variables of the latent $\mathbf{R}_\mathcal{L}$ and $\mathbf{R}_\mathcal{R}$, we make no assumptions whatsoever, which amounts to assuming potential mutual dependency among all component variables within $\mathbf{R}_\mathcal{L}$ and $\mathbf{R}_\mathcal{R}$, respectively (i.e. potential mutual confounding among all variables internal to $\mathcal{W}_\mathcal{L}$ and $\mathcal{W}_\mathcal{R}$, respectively). We have, $\forall \mathbf{r} \in \nu(\mathbf{R})$, $\forall \mathbf{w} \in \nu(\mathbf{W})$ (so in particular, $p\{\mathbf{R}_\mathcal{L} = \mathbf{r}_\mathcal{L}\}, p\{\mathbf{R}_\mathcal{R} = \mathbf{r}_\mathcal{R}\}, p\{\mathbf{W}_\mathcal{L} = \mathbf{w}_\mathcal{L}\}, p\{\mathbf{W}_\mathcal{L} = \mathbf{w}_\mathcal{L}, \mathbf{R}_\mathcal{R} = \mathbf{r}_\mathcal{R}\} = p\{\mathbf{W}_\mathcal{L} = \mathbf{w}_\mathcal{L}\}p\{\mathbf{R}_\mathcal{R} = \mathbf{r}_\mathcal{R}\} > 0$),

$$\begin{aligned} p\{\mathbf{W} = \mathbf{w}, \mathbf{R} = \mathbf{r}\} &= p\{\mathbf{W}_\mathcal{L} = \mathbf{w}_\mathcal{L}, \mathbf{W}_\mathcal{R} = \mathbf{w}_\mathcal{R}, \mathbf{R}_\mathcal{L} = \mathbf{r}_\mathcal{L}, \mathbf{R}_\mathcal{R} = \mathbf{r}_\mathcal{R}\} \\ &= p\{\mathbf{W}_\mathcal{L} = \mathbf{w}_\mathcal{L} \mid \mathbf{R}_\mathcal{L} = \mathbf{r}_\mathcal{L}\}p\{\mathbf{R}_\mathcal{L} = \mathbf{r}_\mathcal{L}\} \\ &\quad\; p\{\mathbf{W}_\mathcal{R} = \mathbf{w}_\mathcal{R} \mid \mathbf{W}_\mathcal{L} = \mathbf{w}_\mathcal{L}, \mathbf{R}_\mathcal{R} = \mathbf{r}_\mathcal{R}\}p\{\mathbf{R}_\mathcal{R} = \mathbf{r}_\mathcal{R}\}. \end{aligned}$$

So $\forall \mathbf{w} \in \nu(\mathbf{W})$,

$$\begin{aligned}
p\{\mathbf{W} = \mathbf{w}\} &= \sum_{\mathbf{r} \in \nu(\mathbf{R})} p\{\mathbf{W} = \mathbf{w}, \mathbf{R} = \mathbf{r}\} \\
&= \sum_{\mathbf{r} \in \nu(\mathbf{R})} p\{\mathbf{W}_{\mathcal{L}} = \mathbf{w}_{\mathcal{L}} \mid \mathbf{R}_{\mathcal{L}} = \mathbf{r}_{\mathcal{L}}\} p\{\mathbf{R}_{\mathcal{L}} = \mathbf{r}_{\mathcal{L}}\} \\
&\qquad p\{\mathbf{W}_{\mathcal{R}} = \mathbf{w}_{\mathcal{R}} \mid \mathbf{W}_{\mathcal{L}} = \mathbf{w}_{\mathcal{L}}, \mathbf{R}_{\mathcal{R}} = \mathbf{r}_{\mathcal{R}}\} p\{\mathbf{R}_{\mathcal{R}} = \mathbf{r}_{\mathcal{R}}\} \\
&= \sum_{\mathbf{r}_{\mathcal{L}} \in \nu(\mathbf{R}_{\mathcal{L}})} \sum_{\mathbf{r}_{\mathcal{R}} \in \nu(\mathbf{R}_{\mathcal{R}})} p\{\mathbf{W}_{\mathcal{L}} = \mathbf{w}_{\mathcal{L}} \mid \mathbf{R}_{\mathcal{L}} = \mathbf{r}_{\mathcal{L}}\} p\{\mathbf{R}_{\mathcal{L}} = \mathbf{r}_{\mathcal{L}}\} \\
&\qquad p\{\mathbf{W}_{\mathcal{R}} = \mathbf{w}_{\mathcal{R}} \mid \mathbf{W}_{\mathcal{L}} = \mathbf{w}_{\mathcal{L}}, \mathbf{R}_{\mathcal{R}} = \mathbf{r}_{\mathcal{R}}\} p\{\mathbf{R}_{\mathcal{R}} = \mathbf{r}_{\mathcal{R}}\} \\
&= \sum_{\mathbf{r}_{\mathcal{L}} \in \nu(\mathbf{R}_{\mathcal{L}})} p\{\mathbf{W}_{\mathcal{L}} = \mathbf{w}_{\mathcal{L}} \mid \mathbf{R}_{\mathcal{L}} = \mathbf{r}_{\mathcal{L}}\} p\{\mathbf{R}_{\mathcal{L}} = \mathbf{r}_{\mathcal{L}}\} \\
&\qquad \sum_{\mathbf{r}_{\mathcal{R}} \in \nu(\mathbf{R}_{\mathcal{R}})} p\{\mathbf{W}_{\mathcal{R}} = \mathbf{w}_{\mathcal{R}} \mid \mathbf{W}_{\mathcal{L}} = \mathbf{w}_{\mathcal{L}}, \mathbf{R}_{\mathcal{R}} = \mathbf{r}_{\mathcal{R}}\} p\{\mathbf{R}_{\mathcal{R}} = \mathbf{r}_{\mathcal{R}}\} \\
&= p\{\mathbf{W}_{\mathcal{L}} = \mathbf{w}_{\mathcal{L}}\} \sum_{\mathbf{r}_{\mathcal{R}} \in \nu(\mathbf{R}_{\mathcal{R}})} p\{\mathbf{W}_{\mathcal{R}} = \mathbf{w}_{\mathcal{R}} \mid \mathbf{W}_{\mathcal{L}} = \mathbf{w}_{\mathcal{L}}, \mathbf{R}_{\mathcal{R}} = \mathbf{r}_{\mathcal{R}}\} p\{\mathbf{R}_{\mathcal{R}} = \mathbf{r}_{\mathcal{R}}\}.
\end{aligned}$$

Hence, $\forall b \in \{1, \dots, B\}$,

$$\begin{aligned}
p_b &= P\{\mathbf{W}_{\mathcal{R}} = \mathbf{w}_{b,\mathcal{R}} \mid \mathbf{W}_{\mathcal{L}} = \mathbf{w}_{b,\mathcal{L}}\} \\
&= \frac{p\{\mathbf{W}_{\mathcal{L}} = \mathbf{w}_{b,\mathcal{L}}, \mathbf{W}_{\mathcal{R}} = \mathbf{w}_{b,\mathcal{R}}\}}{p\{\mathbf{W}_{\mathcal{L}} = \mathbf{w}_{b,\mathcal{L}}\}} \\
&= \frac{p\{\mathbf{W} = \mathbf{w}_b\}}{p\{\mathbf{W}_{\mathcal{L}} = \mathbf{w}_{b,\mathcal{L}}\}} \\
&= \frac{p\{\mathbf{W}_{\mathcal{L}} = \mathbf{w}_{b,\mathcal{L}}\} \sum_{\gamma=1}^{\aleph_{\mathcal{R}}} p\{\mathbf{W}_{\mathcal{R}} = \mathbf{w}_{b,\mathcal{R}} \mid \mathbf{W}_{\mathcal{L}} = \mathbf{w}_{b,\mathcal{L}}, \mathbf{R}_{\mathcal{R}} = \mathbf{r}_\gamma\} p\{\mathbf{R}_{\mathcal{R}} = \mathbf{r}_\gamma\}}{p\{\mathbf{W}_{\mathcal{L}} = \mathbf{w}_{b,\mathcal{L}}\}} \\
&= \sum_{\gamma=1}^{\aleph_{\mathcal{R}}} p\{\mathbf{W}_{\mathcal{R}} = \mathbf{w}_{b,\mathcal{R}} \mid \mathbf{W}_{\mathcal{L}} = \mathbf{w}_{b,\mathcal{L}}, \mathbf{R}_{\mathcal{R}} = \mathbf{r}_\gamma\} p\{\mathbf{R}_{\mathcal{R}} = \mathbf{r}_\gamma\} \\
&= \sum_{\gamma=1}^{\aleph_{\mathcal{R}}} P_{b\gamma} q_\gamma
\end{aligned}$$

where $P \in \{0,1\}^{B \times \aleph_{\mathcal{R}}}$ is given by $\forall b \in \{1, \dots, B\}, \gamma \in \{1, \dots, \aleph_{\mathcal{R}}\}$,

$$\begin{aligned}
P_{b\gamma} &:= p\{\mathbf{W}_{\mathcal{R}} = \mathbf{w}_{b,\mathcal{R}} \mid \mathbf{W}_{\mathcal{L}} = \mathbf{w}_{b,\mathcal{L}}, \mathbf{R}_{\mathcal{R}} = \mathbf{r}_\gamma\} \\
&= \begin{cases} 1 & \text{if } \forall i \in \mathcal{R}, w_i = g_{W_i}(\mathbf{w}_{b,\mathcal{L}}, \mathbf{r}_\gamma) \\ 0 & \text{otherwise} \end{cases} .
\end{aligned}$$

Moreover, $\forall b \in \{1, \ldots, B\}$,

$$\begin{aligned}
p_b^* &= p\{\mathbf{W} = \mathbf{w}_b\} \\
&= p\{\mathbf{W}_{\mathcal{L}} = \mathbf{w}_{b,\mathcal{L}}, \mathbf{W}_{\mathcal{R}} = \mathbf{w}_{b,\mathcal{R}}\} \\
&= p\{\mathbf{W}_{\mathcal{L}} = \mathbf{w}_{b,\mathcal{L}}\} p\{\mathbf{W}_{\mathcal{R}} = \mathbf{w}_{b,\mathcal{R}} \mid \mathbf{W}_{\mathcal{L}} = \mathbf{w}_{b,\mathcal{L}}\} \\
&= p\{\mathbf{W}_{\mathcal{L}} = \mathbf{w}_{b,\mathcal{L}}\} p_b \\
&= p\{\mathbf{W}_{\mathcal{L}} = \mathbf{w}_{b,\mathcal{L}}\} \sum_{\gamma=1}^{\aleph_{\mathcal{R}}} P_{b\gamma} q_\gamma \\
&= P_{b\gamma}^* q_\gamma
\end{aligned}$$

where $P^* \in [0,1]^{B \times \aleph_{\mathcal{R}}}$ is given by $\forall b \in \{1, \ldots, B\}, \gamma \in \{1, \ldots, \aleph_{\mathcal{R}}\}$,

$$\begin{aligned}
P_{b\gamma}^* &:= p\{\mathbf{W} = \mathbf{w}_b \mid \mathbf{R}_{\mathcal{R}} = \mathbf{r}_\gamma\} \\
&= p\{\mathbf{W}_{\mathcal{L}} = \mathbf{w}_{b,\mathcal{L}}, \mathbf{W}_{\mathcal{R}} = \mathbf{w}_{b,\mathcal{R}} \mid \mathbf{R}_{\mathcal{R}} = \mathbf{r}_\gamma\} \\
&= p\{\mathbf{W}_{\mathcal{R}} = \mathbf{w}_{b,\mathcal{R}} \mid \mathbf{W}_{\mathcal{L}} = \mathbf{w}_{b,\mathcal{L}}, \mathbf{R}_{\mathcal{R}} = \mathbf{r}_\gamma\} p\{\mathbf{W}_{\mathcal{L}} = \mathbf{w}_{b,\mathcal{L}} \mid \mathbf{R}_{\mathcal{R}} = \mathbf{r}_\gamma\} \\
&= p\{\mathbf{W}_{\mathcal{L}} = \mathbf{w}_{b,\mathcal{L}}\} p\{\mathbf{W}_{\mathcal{R}} = \mathbf{w}_{b,\mathcal{R}} \mid \mathbf{W}_{\mathcal{L}} = \mathbf{w}_{b,\mathcal{L}}, \mathbf{R}_{\mathcal{R}} = \mathbf{r}_\gamma\} \\
&= p\{\mathbf{W}_{\mathcal{L}} = \mathbf{w}_{b,\mathcal{L}}\} P_{b\gamma} \\
&= \begin{cases} p\{\mathbf{W}_{\mathcal{L}} = \mathbf{w}_{b,\mathcal{L}}\} & \text{if } \forall i \in \mathcal{R}, w_i = g_{W_i}(\mathbf{w}_{b,\mathcal{L}}, \mathbf{r}_\gamma) \\ 0 & \text{otherwise} \end{cases} .
\end{aligned}$$

Since $\forall b \in \{1, \ldots, B\},\ p_b^* = p\{\mathbf{W}_{\mathcal{L}} = \mathbf{w}_{b,\mathcal{L}}\} p_b$, we have $\mathbf{p}^* = \Lambda \mathbf{p}$, where $\Lambda \in [0,1]^{B \times B}$

is given by, $\forall b, c \in \{1, \ldots, B\}$,

$$\Lambda_{bc} := \begin{cases} p\{\mathbf{W}_{\mathcal{L}} = \mathbf{w}_{b,\mathcal{L}}\} & \text{if } b = c \\ \\ 0 & \text{otherwise} \end{cases}$$

Note that $\forall b \in \{1, \ldots, B\}, \forall \gamma \in \{1, \ldots, \aleph_{\mathcal{R}}\}, \sum_{c=1}^{B} \Lambda_{bc} P_{c\gamma} = \Lambda_{bb} P_{b\gamma} = p\{\mathbf{W}_{\mathcal{L}} = \mathbf{w}_{b,\mathcal{L}}\} P_{b\gamma} = P^*_{b\gamma}$, so $\Lambda P = P^*$. Note further that the diagonal entries of Λ all are non-zero (since $\forall b \in \{1, \ldots, B\},\ \mathbf{w}_{b,\mathcal{L}} \in \nu(\mathbf{W}_{\mathcal{L}})$), so Λ is invertible and hence bijectively maps between the conditional probability vector $\mathbf{p} = P\mathbf{q} \in [0,1]^B$ and the corresponding marginal one $\mathbf{p}^* = P^*\mathbf{q} \in [0,1]^B$. Consequently, $\mathbf{p} = P\mathbf{q} \iff \Lambda\mathbf{p} = \Lambda P\mathbf{q} \iff \mathbf{p}^* = P^*\mathbf{q}$.

Since the distribution of the unmeasured influences U, or equivalently the response function variables R, is independent of the DAG, the DAG cannot encode any quantitative constraints in the form of relationships between these variables. Thus, the structural equations encoded by the DAG can only imply constraints (ignoring the distinction between the left and right sides, since this can be considered within each of those sets) based on the following types of independence relations: (i) $W_i \perp\!\!\!\perp U$ for some i, (ii) $W_i \perp\!\!\!\perp U | W_{\mathcal{B}}$ for some i and set of observed variables $W_{\mathcal{B}}$, (iii) $W_i \perp\!\!\!\perp W_j$ for some i, j, (iv) $W_i \perp\!\!\!\perp W_j | W_{\mathcal{A}}$ for some i, j and set of observed variables $W_{\mathcal{A}}$ or (v) $W_i \perp\!\!\!\perp W_j | U$ for some i, j. Cases (i) and (ii) imply that U is not a parent of W_i, in violation of Condition 3 or 4. Cases (iii) and (iv) imply that U is either not a parent of W_i or not of W_j, again in violation of Condition 3 or 4. Case (v) implies that for i, j, we have $p\{W_i, W_j\} = \sum_R p\{W_i, W_j | R\} p\{R\} = \sum_R p\{W_i | R\} p\{W_j | R\} p\{R\}$ which is still linear in $\mathbf{q}$.

Now relating this last point to the enumeration of constraints above, note the vector $\mathbf{p}^*$ enumerates all joint probabilities of all observed variables in the DAG. Hence, constraints relating linear combinations of $\mathbf{q}$ to joint, conditional, or marginal probabilities of subsets of $\mathcal{W}$ can be directly obtained as transformations among rows of the existing constraints $\mathbf{p}^* = P^*\mathbf{q}$. The addition of those are clearly redundant. In other words, the matrix P contains complete information about any and all relationships between the observed joint distribution and the joint distribution of the response function variables of the $\mathcal{R}$-side that are possible under our conditions. By the above, the complete set of constraints on observed probabilities is equivalent to a system that is linear in $\mathbf{q}$. □

Proof of Proposition 3. Let again $\mathcal{P} = \{i_1, \ldots, i_P\}$ and $\mathcal{O} = \{j_1, \ldots, j_O\}$ be respectively the indices of the potential and factual outcomes in Q, and $\Gamma(Q) = \{\mathbf{r} \in \nu(\mathbf{R}) : w_{i_1} = h^{A_{i_1}}_{W_{i_1}}(\mathbf{r}, W_{i_1}), \ldots, w_{i_P} = h^{A_{i_P}}_{W_{i_P}}(\mathbf{r}, W_{i_P}), w_{j_1} = g_{W_{j_1}}(\mathbf{r}), \ldots, w_{j_O} = g_{W_{j_O}}(\mathbf{r})\}$. We have $(\mathbf{R}_\mathcal{L} \perp\!\!\!\perp \mathbf{R}_\mathcal{R})_G$ and, by condition 5, $\mathcal{P} \cup \mathcal{O} \subset \mathcal{R}$ and if $\mathcal{L} \neq \varnothing$, $\Gamma(Q) = \nu(\mathbf{R}_\mathcal{L}) \times \Gamma_\mathcal{R}(Q)$, where $\Gamma_\mathcal{R}(Q) := \{\mathbf{r}_\mathcal{R} \in \nu(\mathbf{R}_\mathcal{R}) : w_{i_1} = h^{A_{i_1}}_{W_{i_1}}(\mathbf{r}_\mathcal{R}, W_{i_1}), \ldots, w_{i_P} = h^{A_{i_P}}_{W_{i_P}}(\mathbf{r}_\mathcal{R}, W_{i_P}), w_{j_1} = g_{W_{j_1}}(\mathbf{r}_\mathcal{R}), \ldots, w_{j_O} = g_{W_{j_O}}(\mathbf{r}_\mathcal{R})\}$. Condition 6 ensures that, if $\mathcal{L}$ is not empty, then all paths from the potential outcomes in Q to any variables in $\mathcal{L}$ must pass through the intervention set, thus negating any influence of $\mathbf{R}_\mathcal{L}$ on any of the variables in Q. Hence, if $\mathcal{L} = \varnothing$, then

$$\begin{aligned}
Q &= p\{h^{A_{i_1}}_{W_{i_1}}(\mathbf{R}, W_{i_1}) = w_{i_1}, \ldots, h^{A_{i_P}}_{W_{i_P}}(\mathbf{R}, W_{i_P}) = w_{i_P}, g_{W_{j_1}}(\mathbf{R}) = w_{j_1}, \ldots, g_{W_{j_O}}(\mathbf{R}) = w_{j_O}\} \\
&= \sum_{\mathbf{r} \in \Gamma(Q)} p\{\mathbf{R} = \mathbf{r}\} \\
&= \sum_{\gamma=1}^{\aleph_\mathcal{R}} \mathbb{I}_{\Gamma(Q)}(\mathbf{r}_\gamma) q_\gamma = \alpha^\top \mathbf{q},
\end{aligned}$$

where $\mathbb{I}(\cdot)$ is the indicator function and $\alpha \in \{0,1\}^{\aleph_{\mathcal{R}}}$ is given by

$$\forall \gamma \in \{1, \ldots, \aleph_{\mathcal{R}}\}, \alpha_\gamma := \begin{cases} 1 & \text{if } \mathbf{r}_\gamma \in \Gamma(Q) \\ \\ 0 & \text{otherwise} \end{cases}.$$

If $\mathcal{L} \neq \varnothing$, we have

$$\begin{aligned} Q &= p\{h^{A_{i_1}}_{W_{i_1}}(\mathbf{R}, W_{i_1}) = w_{i_1}, \ldots, h^{A_{i_P}}_{W_{i_P}}(\mathbf{R}, W_{i_P}) = w_{i_P}\} \\ &= \sum_{\mathbf{r} \in \Gamma(Q)} p\{\mathbf{R} = \mathbf{r}\} \\ &= \sum_{\mathbf{r} \in \Gamma(Q)} p\{\mathbf{R}_{\mathcal{L}} = \mathbf{r}_{\mathcal{L}}, \mathbf{R}_{\mathcal{R}} = \mathbf{r}_{\mathcal{R}}\} \\ &= \sum_{\mathbf{r} \in \Gamma(Q)} p\{\mathbf{R}_{\mathcal{L}} = \mathbf{r}_{\mathcal{L}}\} p\{\mathbf{R}_{\mathcal{R}} = \mathbf{r}_{\mathcal{R}}\} \\ &= \sum_{(\mathbf{r}_{\mathcal{L}}, \mathbf{r}_{\mathcal{R}}) \in \nu(\mathbf{R}_{\mathcal{L}}) \times \Gamma_{\mathcal{R}}(Q)} p\{\mathbf{R}_{\mathcal{L}} = \mathbf{r}_{\mathcal{L}}\} p\{\mathbf{R}_{\mathcal{R}} = \mathbf{r}_{\mathcal{R}}\} \\ &= \sum_{\mathbf{r}_{\mathcal{L}} \in \nu(\mathbf{R}_{\mathcal{L}})} \sum_{\mathbf{r}_{\mathcal{R}} \in \Gamma_{\mathcal{R}}(Q)} p\{\mathbf{R}_{\mathcal{L}} = \mathbf{r}_{\mathcal{L}}\} p\{\mathbf{R}_{\mathcal{R}} = \mathbf{r}_{\mathcal{R}}\} \\ &= \sum_{\mathbf{r}_{\mathcal{L}} \in \nu(\mathbf{R}_{\mathcal{L}})} p\{\mathbf{R}_{\mathcal{L}} = \mathbf{r}_{\mathcal{L}}\} \sum_{\mathbf{r}_{\mathcal{R}} \in \Gamma_{\mathcal{R}}(Q)} p\{\mathbf{R}_{\mathcal{R}} = \mathbf{r}_{\mathcal{R}}\} \\ &= \sum_{\mathbf{r}_{\mathcal{R}} \in \Gamma_{\mathcal{R}}(Q)} p\{\mathbf{R}_{\mathcal{R}} = \mathbf{r}_{\mathcal{R}}\} \\ &= \sum_{\gamma=1}^{\aleph_{\mathcal{R}}} \mathbb{I}_{\Gamma_{\mathcal{R}}(Q)}(\mathbf{r}_\gamma) q_\gamma = \alpha^\top \mathbf{q}, \end{aligned}$$

where $\alpha \in \{0,1\}^{\aleph_{\mathcal{R}}}$ is given by $\forall \gamma \in \{1, \ldots, \aleph_{\mathcal{R}}\}$, $\alpha_\gamma := \begin{cases} 1 & \text{if } \mathbf{r}_\gamma \in \Gamma_{\mathcal{R}}(Q) \\ \\ 0 & \text{otherwise} \end{cases}.$

□

Proof of Proposition 4. Proposition 2 ensures that the linear constraints $\mathbf{p}^* = P^*\mathbf{q}$ are necessary and sufficient for the probability distribution to be compatible with the causal model. Solving the optimization problem with these constraints is equivalent to solving it with the constraints $\mathbf{p} = P\mathbf{q}$ because the relation is obtained by multiplying both sides of the equation by an invertible constant matrix. Proposition 3 demonstrates that the objective function is linear in $\mathbf{q}$. The constraint space is closed and non-empty, and is bounded by the probabilistic constraints. Subject to any additional linear constraints specified in the form of equalities or non-strict inequalities, the constraint space is closed and bounded, hence compact, so by the extreme value theorem and the fact that the objective is linear, hence continuous, the primal problem has an optimal feasible solution. By the strong duality theorem, the dual problem has a global optimum coinciding with that of the primal, and again has a bounded constraint space, so by the fundamental theorem of linear programming, it can be found in terms of $\mathbf{p}$ via vertex enumeration.

□