Symbolic Computation of Tight Causal Bounds

Preprint, Under Review as of March 2020

Biometrika (2020), **103**, 1, *pp*. 1–19
Printed in Great Britain Advance Access publication on 31 July 2018

Symbolic Computation of Tight Causal Bounds

BY M.C. SACHS
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet
Box 281, 17177 Stockholm, Sweden.
michael.sachs@ki.se

E.E. GABRIEL
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet
Box 281, 17177 Stockholm, Sweden.
erin.gabriel@ki.se

A. SJÖLANDER
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet
Box 281, 17177 Stockholm, Sweden.
arvid.sjolander@ki.se

SUMMARY

Causal inference involves making a set of assumptions about the nature of things, defining a causal query, and attempting to find estimators of the query based on the distribution of observed variables. When causal queries are not identifiable from the observed data, it still may be possible to derive bounds for these quantities in terms of the distribution of observed variables. We develop and describe a general approach for computation of bounds, proving that if the problem can be stated as a linear program, then the true global extrema result in tight bounds. Building upon previous work in this area, we characterize a class of problems that can always be stated as a linear programming problem; we describe a general algorithm for constructing the linear objective and constraints based on the causal model and the causal query of interest. These problems therefore can be solved using a vertex enumeration algorithm. We develop an R package implementing this algorithm with a user friendly graphical interface using directed acyclic graphs, which only allows for problems within this class to be depicted. We have implemented additional features to help with interpreting and applying the bounds that we illustrate in examples.

Some key words: Causal Inference; Identifiability; Bounds; Computation.

1. INTRODUCTION

In many fields of research, a common goal is to determine the causal effect of a particular exposure, event or circumstance on a particular outcome. Unless one is able to experimentally intervene on the exposure, this investigation is typically complicated by the fact that there are common causes of the exposure and the outcome. Often, these common causes are at least partly

unknown, in which case the causal effect of interest is generally not identifiable in the sense that it cannot be computed uniquely from the probability distribution of observed variables because any observable association could be due to the uncontrolled common causes. When the causal effects of interest, which we will call causal queries, cannot be identified, it still may be possible to derive bounds, i.e. a range of possible values, for these quantities in terms of the true distribution of the observed variables. Such bounds have been derived in a variety of settings (Robins, 1989; Manski, 1990; Zhang & Rubin, 2003; Cai et al., 2008; Sjölander, 2009; Sjölander et al., 2014). Although numeric optimization is possible, symbolic bounds can provide useful information with which to draw conclusions about a study design or form of data collection in the absence of data.

Twenty-five years ago in his PhD dissertation, Alexander Balke illustrated a method for translating a causal theory, represented by a directly acyclic graph (DAG), and a certain type of causal query into a constrained optimization problem (Balke & Pearl, 1994a). Balke & Pearl (1994b) develop the representation of a causal theory in terms of latent response function variables that describe the probability distribution of counterfactual quantities. That representation permits one to write the causal query in terms of the distribution of the response function variables and also to derive a system of equations relating observed probabilities to the distribution of the response function variables. Taken together and combined with standard probabilistic constraints, this defines a constrained optimization problem. If the problem is linear then a vertex enumeration algorithm can be used to find the global maximum and minimum of the causal query in terms of the true probability distribution of the observed variables (Dantzig, 1963).

Balke & Pearl (1994a) states, but does not prove, that the resulting extrema yield tight bounds for the causal query of his example in terms of the true probability distribution of the observed variables. "Tight" here means that all values inside the bounds are logically compatible with the true distribution of the observed variables. To the knowledge of the authors, no one has proven this is true in general for all problems that define a linear program, although some have proven this result in specific settings (Ramsahai, 2012; Bonet, 2013; Heckman & Vytlacil, 2001). Regardless of the tightness of the bounds, vertex enumeration is only guaranteed to produce global extrema in linear optimization problems, i.e., linear objective and linear constraints. To the knowledge of the authors, there has been no attempt in the literature to describe a class of problems that are always linear or an approach for determining whether a problem is linear, given the DAG and target causal query.

Balke wrote a program in C++ to take a linear programming problem as text file input, perform variable reduction, conversion of equality constraints into inequality constraints, and perform the vertex enumeration algorithm of Mattheiss (1973). This program has been used by researchers in the field of causal inference with great success (Balke & Pearl, 1997; Cai et al., 2008; Sjölander, 2009; Sjölander et al., 2014) but it is not particularly accessible to other researchers because of the technical challenge of translating the DAG plus causal query into the constrained optimization problem and to determine whether it is linear. Thus, applications of this approach have been limited to a small number of settings.

In this paper, we generalize and extend Balke and Pearl's approach for computation of bounds, proving that if the problem can be stated as a linear program, then the true global extrema result in tight bounds for the DAG and causal query and additional constraints in question. We characterize a class of problems that can always be stated as a linear programming problem; we describe a general algorithm for constructing the linear objective and constraints based on the DAG and the causal query of interest. These problems therefore, at least theoretically, can be solved using a vertex enumeration algorithm. We develop an R package called `causaloptim`, available on the Comprehensive R Archive Network (CRAN), that implements this algorithm

with a user friendly interface for setting up such problems via DAGs, which only allows for problems within this class to be depicted. The user can then define the target causal quantity and optionally linear constraints using standard causal notation.

We illustrate the steps of the algorithm by using it to derive bounds in a simple example with two confounded variables. Then we apply the method to derive bounds in a novel setting, where there are two instrumental variables that are correlated with each other.

2. NOTATION AND PRELIMINARIES

2.1. *Response functional expression of a causal theory*

Let the set of variables of interest be denoted $\mathcal{W} = \{W_1, \ldots, W_n\}$, with observed values represented by the vector w with elements $w_1, \ldots, w_n$. We assume that all of these variables are binary and can take values in $\{0, 1\}$. Each variable of interest W_i has an associated irreducible error term that is latent, not necessarily binary, and denoted ε_i. There may be additional variables that are latent and not necessarily binary. We also need to describe variables in the potential outcome world, and this will be denoted using brackets: i.e., $W_1(W_2 = w_2)$. Thus a counterfactual probability such as $\text{pr}\{W_1(W_2 = 0) = 1\}$ will be read as 'if W_2 were intervened upon to have value 0, what is the probability that W_1 would have been equal to 1?'.

(a) Example with error terms. The dashed ellipse outlines the latent causal influences of W_2 and W_3. Since they both contain U, the common cause, their response function variables are dependent.

(b) Example with response function variables. The response function variables denoted are categorical, with R_1 taking two possible values, R_2 taking 4 possible values, R_3 taking 16 possible values.

Fig. 1: Example DAG to illustrate the concepts and notation. In this example, the variables W_1, W_2, and W_3 are of interest, assumed to be binary, and the others are latent errors. Since the variables of interest are binary, the assumptions can be represented using categorical response function variables.

The DAG encodes assumptions regarding the variables in the model. For each i, we have the functional expression for w_i, the value of W_i: $w_i = f_{W_i}(\text{pa}_i, \text{U}_{w_i}, \varepsilon_i)$, where pa_i denotes the values of variables in $\mathcal{W}$ that are parents of W_i in the DAG, and U_{W_i} represents the (possibly empty) vector of latent variables that are parents of W_i in the DAG, and ε_i the independent errors due to omitted factors that may influence W_i but no other variables. The U_{W_i} variables are not assumed independent, and they generally will represent unmeasured confounders. Since all variables of interest in the graph are assumed to be binary, we can, without loss of generality, recode the assumptions by defining a series of categorical variables R_{W_i}, one for each variable in $\mathcal{W}$, which specifies how W_i is determined from its parents. In this response function variable form of the DAG, if the binary variable W_i has k_i parents (including the response function variables), then there are $2^{2^{k_i-1}}$ possible response patterns of W_i with respect to pa_i. Thus, we may represent each R_{W_i} as a categorical random variable that takes on $2^{2^{k_i-1}}$ possible values, one for each response pattern, for $i = 1, \ldots, n$. Let r_{W_i} denote an arbitrary value that the random

variable R_{W_i} can take, i.e., each category of r_{W_i} occurs with a certain prior probability $\text{pr}\{R_{W_i} = r_{W_i}\}$ such that

$$\sum_{i=1}^{2^{2^{k_i-1}}} \text{pr}\{R_{W_i} = r_{W_i}\} = 1.$$

Let R denote the vector of response function variables $(R_1, \ldots, R_n)$ and $r = (r_{W_1}, \ldots, r_{W_n})$ an arbitrary value of the random vector R. The vector r can take on

$$\aleph = \prod_{i=1}^{n} 2^{2^{k_i-1}}$$

possible values.

The joint distribution of the response function variables $\text{pr}\{R = r\}$ together with the response functions fully characterize the causal model. To see this, note that given the value r, all variables $W_i \in \mathcal{W}$ have values that are functionally determined. For a given W_i and fixed r, we define a procedure for determining its value by recursively evaluating the functional expression. We will use nested subscripts to denote parents of W_i that are in $\mathcal{W}$, i.e., $W_{i1}, \ldots, W_{ik_i}$ are variables in $\mathcal{W}$ that are parents of W_i. Then w_i, the value of W_i can be obtained by recursively evaluating

$$f_{W_i}(r) = f_{W_i}(f_{W_{i1}}(r), \ldots, f_{W_{ik_i}}(r), r_{W_i}).$$

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

$$\text{pr}\{w_1 = W_1; \ldots; w_n = W_n\} = \sum_{r:\forall_{j\in 1,\ldots,n}[w_j = f_{W_j}(r)]} \text{pr}\{R = r\}.$$

As an example, Figure 1a shows the DAG for a model in which the outcome W_3 has two parents W_2 and W_1, which both have an effect on W_3 and where W_1 also has a direct effect on W_2. Figure 1b the equivalent DAG with response functional variables in place of the errors. The variables that have a latent common cause have response function variables that are dependent, as indicated by the dashed ellipse that outlines the latent causal influences of W_2 and W_3. Since they both contain U, the common cause, then their response function variables are dependent and thus connected by an undirected edge. As can be seen, in Figure 1b W_2 has two parents, W_1 and R_{W_2}, and then we can define R_{W_2} so the values $0, 1, 2, 3$ of R_{W_2} correspond to the response patterns $f_{W_2}(\text{pa}_2 = w_1, r_{W_2} = 0) = 0, f_{W_2}(w_1, r_{W_2} = 1) = w_1, f_{W_2}(w_1, r_{W_2} = 2) = 1 - w_1, f_{W_2}(w_1, r_{W_2} = 3) = 1$, respectively. Under this model shown in Figure 1b, with $r_{W_1} = 0$, $r_{W_2} = 1$, $r_{W_3} = 3$, we can evaluate the function to determine w_2:

$$f_{W_2}(r = (0, 1, 3)) = f_{W_2}(f_{W_1}(0), 1) = f_{W_2}(0, 1) = 0.$$

For W_3, we need to define response patterns for each of the 2^2 possible combinations of values of (w_1, w_2), i.e., $2^{2^2} = 16$, while W_1 has only 2 possible response patterns. Then, to evaluate the probability $\text{pr}\{W_1 = 1; W_2 = 0; W_3 = 1\}$ in terms of R, we can follow the same procedure as above for all $2^1 \cdot 2^2 \cdot 2^4 = 128$ possible combinations of r, keeping track of the resulting values w. It can be shown that the variable value $w = (1, 0, 1)$ is consistent with 16 values of r. Thus the probability of this event is the sum over the set of these 16 values of the probability that R equals them. See Balke & Pearl (1994a) or Pearl (2009), Chapter 8 for another example and further interpretation.

3. RESULTS

3.1. *Class of Problems*

Next, we describe a general class of problems in terms of conditions on the DAG and the query such that the problem is guaranteed to be a linear programming problem, and the algorithm to obtain the bounds. By "problem", we mean the assumptions encoded in a DAG together with the causal query, and optionally additional linear constraints. In causal inference problems, latent common causes (confounding) make the causal effect non-identifiable, which motivates the use of bounds. However, bounds can be improved upon by having a variable that is unconfounded with the outcome of interest (Pearl, 2009). To generalize this idea we make a separation into sets of variable indices $\mathcal{L}$ and $\mathcal{R}$, where the $\mathcal{L}$ variables are unconfounded with the $\mathcal{R}$ variables.

The set of variables $\mathcal{W}$ in the graph can be partitioned into two groups $\mathcal{W} = \{\mathcal{W}_\mathcal{L}, \mathcal{W}_\mathcal{R}\}$, where $\mathcal{L}$ may be empty. We will likewise write $R = \{R_\mathcal{L}, R_\mathcal{R}\}$ for the corresponding response function variables, and the values of the vectors variables in lowercase. We assume without loss of generality that the indices of the variables are ordered in such a way that $\mathcal{L} = \{1, \ldots, K\}$ and $\mathcal{R} = \{K+1, \ldots, n\}$, where K may be 0. The graph must meet all of the following conditions:

Assumption 1. Edges that connect two variables, one from $\mathcal{L}$ and one from $\mathcal{R}$, must be directed from $\mathcal{L}$ to $\mathcal{R}$.

Assumption 2. 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 $\mathcal{L}$ are confounded with each other.

Assumption 3. 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 $\mathcal{R}$ are confounded with each other.

Assumption 4. There exists no unmeasured variable U such that U is a parent of W_i and W_j for any $i \in \mathcal{L}$ and any $j \in \mathcal{R}$. That is, the variables in $\mathcal{L}$ and $\mathcal{R}$ are not confounded with each other.

We introduce some additional notation before stating and proving the results. Let $\text{pr}\{W_\mathcal{R} = w_\mathcal{R} | W_\mathcal{L} = w_\mathcal{L}\}$ denote the observed probabilities of all variables in $\mathcal{R}$ conditional on all variables in $\mathcal{L}$. Let p denote the vector of length 2^n of all possible observed probabilities of that form, the elements of which will be denoted p_b.

As we will soon show, it suffices to consider only the response function variables in $\mathcal{R}$. Thus we will denote $\text{pr}(R_\mathcal{R} = r_\gamma) = q_\gamma$ for each of the r_γ in the domain of $R_\mathcal{R}$ which number

$$\aleph_\mathcal{R} = \prod_{j=K+1}^{n} 2^{2^{k_j - 1}}.$$

That is, q_γ indexes the parameters of the joint probability distribution of the response function variables $R_\mathcal{R}$, such that

$$\sum_{\gamma=1}^{\aleph_\mathcal{R}} q_\gamma = 1,$$

and q denotes the vector.

For $i \in \mathcal{R}$ and for a fixed value of $w_\mathcal{L}$, we will write $f_{W_i}(w_\mathcal{L}, r_\gamma)$ to denote the function

$$f_{W_i}(w_{i1}, \ldots, w_{il}, f_{W_{il+1}}(r_\gamma), \ldots, f_{W_{ik_i}}(r_\gamma), r_{W_i}),$$

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

An overview of the algorithm is as follows: (1) For each observed probability conditional on variables in $\mathcal{L}$, convert to linear combination of joint probabilities of the response function variables. (2) Allow for additional linear constraints. (3) Convert the causal query to a linear combination of joint probabilities of the response function variables. (4) Enumerate the vertices of the dual of the linear programming problem. (5) Return bounds in terms of observed probabilities. We describe 1-3 in more detail in turn, and the description of the algorithm serves as a constructive proof that this class of problems is a linear programming problem.

3.2. *Obtaining linear constraints on observed probabilities*

THEOREM 1. *In DAGs that satisfy the Assumptions 1 - 4, conditional probabilities pr$\{\mathcal{W}_\mathcal{R} = w_\mathcal{R}|\mathcal{W}_\mathcal{L} = w_\mathcal{L}\}$ for all possible combinations of $w_\mathcal{R}$ and $w_\mathcal{L}$ are linear in response function variable probabilities. This defines a system of linear equations that can be written $p = Pq$ for a matrix P.*

Proof of Theorem 1. By Assumption 3, all variables in $(R_\mathcal{R})$ are mutually dependent and thus the probabilities cannot be factorized, i.e., pr$(R_{W_i} = r_{W_i}, R_{W_j} = r_{W_j}) \neq$ pr$(R_{W_i} = r_{W_i})$pr$(R_{W_j} = r_{W_j})$ for any pair or tuple.

Algorithm 1 allows us to determine the linear equations that specify the relationship between the observed conditional probabilities and q. Recall p denotes the vector of length B of all possible observed probabilities of the form above, the elements of which will be denoted p_b and the corresponding variable values are denoted $(w_\mathcal{R}, w_\mathcal{L})_b$.

Algorithm 1. Algorithm to determine linear system of equations relating p to q.

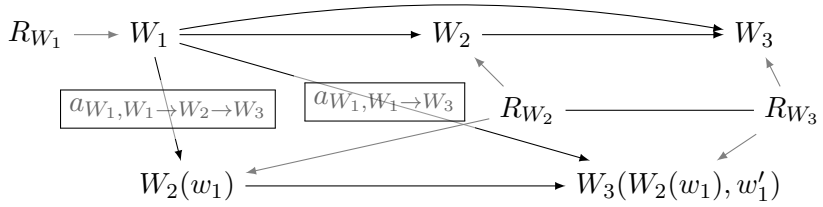
Result: System of linear equations relating p to q
Initialize P a B by $\aleph_\mathcal{R}$ matrix of 0s;
for $b \in 1, \ldots, B$ **do**
  Set $w = (w_\mathcal{R}, w_\mathcal{L})_b$;
  for $\gamma \in 1, \ldots, \aleph_\mathcal{R}$ **do**
    Initialize w^*;
    for $j \in \mathcal{R}$ **do**
      Compute $w_j^* = f_{W_j}(w_{\mathcal{L},b}, r_\gamma)$;
    if $(w^*, w_{\mathcal{L},b}) = w$ **then**
      $P_{b,\gamma} = 1$;

In the recursive function f_{W_i} in this algorithm, the parents of W_i that are in $\mathcal{L}$ are fixed at the values determined by the conditional probability statement, which is why in this case, f_{W_i} only depends on r_γ and not the full vector of response function variables. Thus, Algorithm 1 yields a system of equations $Pq = p$ where p is the vector of conditional probabilities of observed variables, proving Theorem 1. □

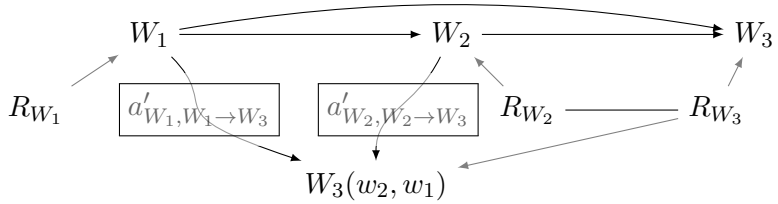
3.3. *Functional expression incorporating interventions*

In order to determine the values of variables of interest for counterfactual quantities that incorporate interventions, we must also define a procedure for evaluating the 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 counterfactual quantities of interest as in Balke & Pearl (1994b). These are called twin networks in Pearl (2009), Chapter 7. Two examples are shown in Figure 2a and 2b. The factual nodes remain as they are, and for each counterfactual quantity of interest, nodes are added. The corresponding factual and counterfactual nodes share the same response function variables. Edges that connect factual nodes to counterfactual nodes

are labelled with letters that denote intervention sets indexed by the child variable of that edge. These sets define the variables being externally set, and the values that they are being set to.



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



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

Fig. 2: Extended DAGs to illustrate that multiple intervention sets are needed to define certain counterfactual quantities.

Balke & Pearl (1994a) considered cases where we externally force a single set of the variables to some fixed values. This construction suffices for the examples they consider, which are to derive bounds for the noncompliance example and the 'party example'. This formulation, however, does not suffice for defining and bounding effects like the natural direct effect of W_1 in the graph in Figure 2a whose first term is $\text{pr}\{W_3(W_2(W_1 = 0), W_1 = 1) = 1\}$. 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. Sjölander (2009) extended the method to work for the natural direct effect, but not more generally than that. As another example, the causal query $\text{pr}\{W_3(W_2(W_1 = 0)) = 1; W_2(W_1 = 1) = 1\}$ 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 one assigns to a value 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 the extended function expression, which "remembers" the path of edges taken to get the value that is being determined at each call.

Let A be a n by J matrix that encodes the interventions and variables on which to intervene, with rows indexed by i corresponding to the variables and the columns indexed by j corresponding to all possible interventional paths, the entries can be 0, 1, or $\emptyset$. The desired interventions within the causal query then define the entries of A which are denoted a_{ij}. In our procedure for evaluating counterfactuals, there is a distinct interventional matrix A_p corresponding to each outcome variable with a single index p. We define the procedure for evaluating the interventional response functional for an outcome variable such that there is no intervention on the outcome as

$$f_{W_i}\big(w_{i1} = f^A_{W_{i1}}(r, W_{i1} \to W_i), \ldots, w_{ik_i} = f^A_{W_{ik_i}}(r, W_{ik_i} \to W_i), R_{W_i}\big),$$

where for all variables W_i, $f^A_{W_i}(r, j)$ is defined recursively as:

$$f^A_{W_i}(r,j) = \begin{cases} a_{ij} & \text{if } a_{ij} \neq \emptyset \\ f_{W_i}(r_{W_i}) & \text{if } a_{ij} = \emptyset \text{ and } \text{pa}(W_i) = \emptyset \\ f_{W_i}(f^A_{W_{i1}}(r, W_{i1} \to j), \ldots, f^A_{W_{ik_i}}(r, W_{ik_i} \to j), R_{W_i}) & \text{otherwise,} \end{cases}$$

where $\{W_{i1}, \ldots, W_{ik_i}\} = \text{pa}(W_i)$ are the the parents of W_i and k_i are their number in the causal model and the notation $i \to j$ means that $i \to$ is appended to the front of whatever is included in 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 first part of the causal query $\text{pr}\{W_3(W_2(W_1 = 0), W_1 = 1) = 1\}$, we have

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

Thus, evaluating the functional expression results in

$$w_3 = f_{W_3}(f^A_{W_1}(r, W_1 \to W_3), f^A_{W_2}(r, W_2 \to W_3), r_{W_3}).$$

For the first element of that function call we have $f^A_{W_1}(r, W_1 \to W_3) = 1$. Then for the second element, we recurse, giving

$$f^A_{W_2}(r, W_2 \to W_3) = f_{W_2}(f^A_{W_1}(r, W_1 \to W_2 \to W_3), r_{W_2}).$$

Now $f^A_{W_1}(r, W_1 \to W_2 \to W_3) = 0$, giving the result $w_3 = f_{W_3}(w_1 = 1, w_2 = W_2(W_1 = 0), r_{W_3})$.

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

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

Thus, evaluating the functional expression results in

$$w_3 = f_{W_3}(f^A_{W_1}(r, W_1 \to W_3), f^A_{W_2}(r, W_2 \to W_3), r_{W_3}).$$

For the first element of that function call we have $f^A_{W_1}(r, W_1 \to W_3) = 1$. Then for the second element $f^A_{W_2}(r, W_2 \to W_3) = 0$, giving the result $w_3 = f_{W_3}(w_1 = 1, w_2 = 0, r_{W_3})$.

The procedures for evaluating the functions f and f^A are sufficient to translate any factual or counterfactual joint probability statement into probability statements involving only the response function variables R. Using our response function formulation, any counterfactual or factual joint probability statement can be written

$$\begin{aligned} Q_v = \text{pr}\{f_{W_{i_1}}(f^{A_{i_1}}_{W_{i_1\cdot}}(r), r_{W_{i_1}}) = w_{i_1}, \ldots, f_{W_{i_P}}(f^{A_{i_P}}_{W_{i_P\cdot}}(r), r_{W_{i_P}}) = w_{i_P}, \\ f_{W_{j_1}}(r) = w_{j_1}, \ldots, f_{W_{j_O}}(r) = w_{j_O}\}, \end{aligned} \quad (1)$$

where $\mathcal{P} = \{i_1, \ldots, i_P\}$ denote the indices of counterfactual outcomes, and $\mathcal{O} = \{j_1, \ldots, j_O\}$ the indices of the factual outcomes. The sets may be overlapping, and each set may contain

duplicates. Viewing the vector R as a random variable, it is clear that

$$Q_v = \sum_{r \in \Gamma} \text{pr}\{R = r\}, \text{ where } \Gamma = \{r : w_{i_p} = f^{A_i}_{W_{i_p}}(r, i_p) \text{ and } w_{j_o} = f_{W_{j_o}}(r)\},$$

for all $i_p \in \mathcal{P}$ and all $j_o \in \mathcal{O}$. This form is completely general, and allows arbitrarily nested counterfactuals, and combinations with observational quantities. Causal contrasts such as the risk difference are constructed by defining Q to be sums and differences of a set of Q_v indexed by $v \in \{1, \ldots, V\}$.

3.4. *Obtaining causal query as linear function of causal parameters*

Causal queries must satisfy:

Assumption 5. $Q = \sum_{v=1}^{V} \alpha_v Q_v$, where each $\alpha_v \in \{-1, 1\}$.

Assumption 6. Each Q_v is a counterfactual probability as given in (1) where $i_1, \ldots, i_p, j_1, \ldots, j_o \in \mathcal{R}$ and if $\mathcal{L}$ is not empty: (1) none of the variables in $\mathcal{L}$ that are intervened upon can have children in $\mathcal{L}$, (2) all variables in $\mathcal{L}$ 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 $\emptyset$, (3) No observations are allowed in Q_v.

THEOREM 2. *Under DAGs that satisfy Assumptions 1 - 4, causal queries that satisfy Assumptions 5 - 6 are linear functions of the joint probabilities of the response function variables. That is, we can write $Q = \alpha^T q$, for some vector α.*

Proof of Theorem 2. Algorithm 2 describes the manner in which the causal query is converted into a linear function of the response function variable probabilities.

Algorithm 2. Converting Q to a linear combination of q.

Result: Q in terms of qs
for $v \in 1, \ldots, V$ **do**
  Set $Q_v = 0$;
  Set A according to Q_v;
  for $\gamma \in 1, \ldots, \aleph_{\mathcal{R}}$ **do**
    Set $\mathcal{P} = \{i_1, \ldots, i_p\}$ to the indices of variables intervened upon, and
     $\mathcal{O} = \{j_1, \ldots, j_o\}$ the indices of the variables not intervened upon in in Q_v;
    Initialize w^* ;
    for $l \in \mathcal{P}$ **do**
      Compute $w^*_l = f^A_{W_l}(r_\gamma)$;
    for $l \in \mathcal{O}$ **do**
      Compute $w^*_l = f_{W_l}(r_\gamma)$;
    if $w^* = w$ **then**
      $Q_v = Q_v + \alpha_j q_k$;
Compute $Q = \sum_{v=1}^{V} \alpha_v Q_v$, and reduce the q variables.

By Assumptions 6, all of the variables in w are in $\mathcal{R}$. Then by Assumptions 1 - 4, we know that changing the values of the response function variables in $\mathcal{L}$ does not influence the possible values of w. Then, since $R_{\mathcal{L}}$ is independent of $R_{\mathcal{R}}$, each match in the final if statement of Algorithm 2 leads to a sum over all possible values of $r_{\mathcal{L}}$ that can be factored out and is equal to 1, thereby leaving only the sum of distinct parameters for $R_{\mathcal{R}}$.

Therefore, we have $Q = \alpha^T q$ for some vector $\overline{\alpha}$. This is the objective function in terms of the counterfactual probabilities thereby proving Theorem 2. □

3.5. *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:

$$\sum_{w_{\mathcal{R}}} \text{pr}(\mathcal{W}_{\mathcal{R}} = w_{\mathcal{R}} | \mathcal{W}_{\mathcal{L}} = w_{\mathcal{L}}) = 1, \text{for all possible values of the vector } w_{\mathcal{L}},$$

$$\sum_{\gamma=1}^{\aleph_{\mathcal{R}}} q_\gamma = 1.$$

Additional user specified constraints on q can be optionally specified as $Bq + h \geq 0$ where h is a vector of constants. We now can state the following linear programming problem (linear objective with linear constraints).

Minimize (maximize): Q

Subject to:

$$\begin{aligned} \sum_{\gamma=1}^{\aleph_{\mathcal{R}}} q_\gamma &= 1 \\ Pq &= p \\ Bq + h &\geq 0 \\ q_j, p_i &\geq 0 \\ \sum_{\mathcal{R}} P(\mathcal{W}_{\mathcal{R}} = w_{\mathcal{R}} | \mathcal{W}_{\mathcal{L}} = w_{\mathcal{L}}) &\geq 0, \text{ for all levels of } w_{\mathcal{L}}. \end{aligned}$$

Global solutions to this problem can be found symbolically by applying Balke's implementation of a vertex enumeration algorithm (Balke & Pearl, 1994a; Mattheiss, 1973). In brief, this algebraically reduces the variables in the optimization problem, then adds slack variables so that all constraints are converted into inequality constraints. The dual of this problem is to maximize (minimize) $y^T p$, for some vector y subject to a set of constraints. Thus the extremum of the causal query as stated in terms of q is equal to the extremum in the p space defined by the dual constraints. Then, by noting that those constraints describe a convex polytope in the p space, the global extrema can be found by enumerating all of the vertices of the polytope. This gives the bounds on the causal effect of interest as the minimum (maximum) of a list of terms involving only observable probabilities, each of which corresponds to a vertex of this polytope. This demonstrates that for this class of problems, tight bounds can be derived symbolically according to this algorithm.

3.6. *Conditional probabilities are sufficient*

THEOREM 3. *Under Assumptions 1 - 6, the bounds obtained by solving the linear programming problem are valid and tight.*

Proof of Theorem 3. In order to completely exhaust the relations between observed probabilities and response function variable probabilities, one would have to consider the joint probabilities $\text{pr}(\mathcal{W}_{\mathcal{R}} = w_{\mathcal{R}}, \mathcal{W}_{\mathcal{L}} = w_{\mathcal{L}})$. By Assumption 1, we have $R_{\mathcal{L}}$ is independent of $R_{\mathcal{R}}$. Thus, applying the same procedure as in Algorithm 1 to those joint probabilities would yield a system of equations of the form $PqK = p^*$, where K is a diagonal matrix of response function variable probabilities for variables in $\mathcal{L}$ and p^* is the vector of joint observed probabilities. Thus,

$Pq = K^{-1}p^* = p$. That is, the constraints implied by the joint observed probabilities do not contain any information beyond what is implied by the conditional observed probabilities.

Therefore, the set of equations relating the conditional probabilities to the response function variable probabilities exhausts the relations between observed probabilities and counterfactual probabilities. That is, Algorithm 1 completely describes the relationship between observed probabilities and response function variable probabilities, and in such a way that the relationships are linear. Therefore, the global extrema for the constrained optimization problem yields global extrema for the causal model, hence the bounds are the tightest possible assumption-free bounds.□

3.7. *A note on the equivalence class of causal problems for which the bounds are tight*

The algorithm is formulated so that bounds are derived in terms of the true probabilities of the observed variables in $\mathcal{R}$ conditional on the variables in $\mathcal{L}$. Provided one is not intervening on any of the variables in $\mathcal{L}$, this implies that the direction of the edges within the left side cannot be informative. 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 of interest on the left side. For example, the bounds computed for a query such as pr$\{Y(X = 1) = 1\}$ 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, and as shown above, conditional probabilities are sufficient.

Alternatively, if the desired query was pr$\{Y(X(Z = 1)) = 1\}$, 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 Assumption 6, if we intervene upon a variable in $\mathcal{L}$, then the direction of edges within $\mathcal{L}$ matters, and in fact if the intervened upon variable has a child also in $\mathcal{L}$, the condition will not be met.

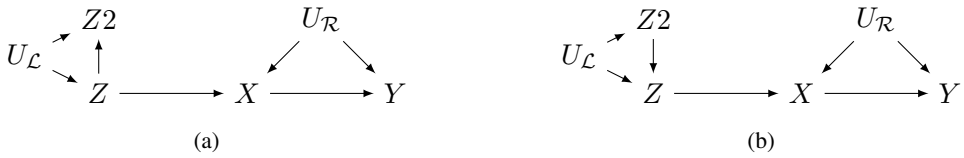


Fig. 3: An equivalence class of DAGs defined by arbitrary connections on the leftside. Bounds for causal parameters that involve intervening on X that meet our conditions are equivalent and tight for these two graphs in (a) and (b).

4. EXAMPLES

4.1. *User interface*

The `R` package `causaloptim`, now available on CRAN, has a graphical user interface which allows users to define a DAG that is constrained by design to be in the class of problems that we describe (R Core Team, 2019; Sachs et al., 2020). The graph is divided into a left side, which corresponds to the $\mathcal{L}$ set, and a right side, which corresponds to the $\mathcal{R}$ set, as in Figure 4. The left side is displayed as a violet (dary grey) box, and the right side a yellow (light grey) box. The program is constrained so that only DAGs that meet Assumptions 1-4 may be drawn. Unmeasured common causes on each side are added to the DAG automatically by the program.

After the user draws the DAG interactively using a web browser, they specify the causal effect of interest using the same notation we have used in this paper in a text interface. Additional user-

specified constraints are optional and are also specified using a text interface. The program then applies our algorithm in order to find the symbolic bounds in terms of the observed conditional probabilities.

In cases where there is only a single intervention set that applies for all possible paths, the program allows for a shorthand notation. For example, in the graph in Figure 1b, if the query of interest is pr$\{Y(X(Z=1), Z=1)=1\}$, then the user may instead write pr$\{Y(Z=1)=1\}$. This is understood by the program to indicate a single intervention set on Z, and the interventions are propagated through all possible paths to the outcome Y. This is useful in situations where there may be a large number of possible paths between a single intervention of interest and the outcome.

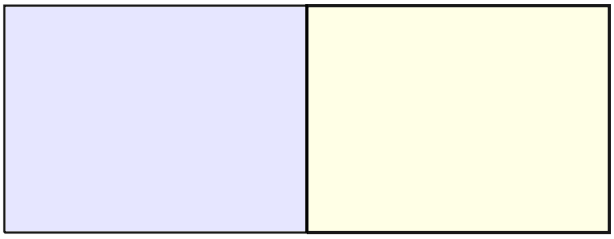


Fig. 4: Depiction of the graphical user interface available in the R package. The left side in violet (dark grey) defines the variables in $\mathcal{L}$ and the right side in yellow (light grey) defines variables in $\mathcal{R}$. The interface and algorithm is set up so that only DAGs that meet our Assumptions 1 - 4 are allowed.

4.2. *Confounded exposure and outcome*

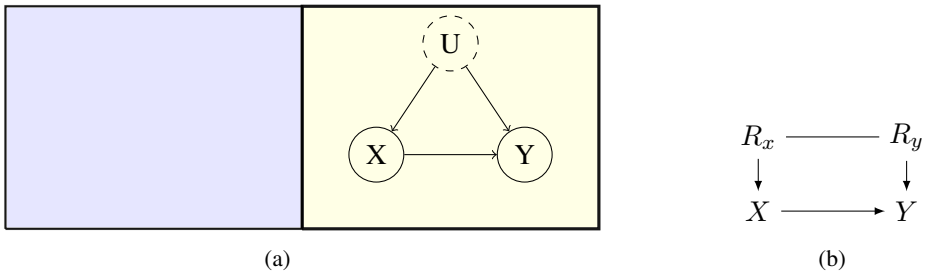


Fig. 5: Simple confounded example as drawn in the program, and the equivalent response function variable graph.

The basic DAG with two variables that are confounded as shown in Figure 5a 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. Our causal effect of interest is the risk difference pr$\{Y(X=1)=1\} - P\{Y(X=0)=1\}$, and we have no additional constraints to specify.

Here we have two variables and therefore two response functions. The response function variable formulation of the graph in Figure 5b is an equivalent representation of the causal model. We have

$$
\begin{aligned}
y &= f_Y(x, r_Y) \\
x &= f_X(r_X) \\
A_1 &= \left[\begin{array}{c|c} & X \to Y \\ \hline X & 1 \\ Y & \emptyset \end{array}\right], \text{ for the first term of the query} \\
A_2 &= \left[\begin{array}{c|c} & X \to Y \\ \hline X & 0 \\ Y & \emptyset \end{array}\right], \text{ for the second term of the query.}
\end{aligned}
$$

R_Y is a random variable that can take on 4 possible values, and R_X is a random variable that can take on 2 possible values. Thus, the joint distribution of (R_X, R_Y) is characterized by 8 parameters, say $q_{i,j}$, where $i \in \{0, 1\}$ and $j \in \{0, 1, 2, 3\}$. Applying Algorithm 1, we can relate the 4 observed probabilities to the parameters of the response function variable distribution as follows

$$
\begin{aligned}
\text{pr}(Y = 0, X = 0) &= q_{0,0} + q_{0,2} \\
\text{pr}(Y = 0, X = 1) &= q_{1,0} + q_{1,1} \\
\text{pr}(Y = 1, X = 0) &= q_{0,1} + q_{0,3} \\
\text{pr}(Y = 1, X = 1) &= q_{1,2} + q_{1,3}.
\end{aligned}
$$

Applying Algorithm 2, we find the relation

$$
\text{pr}\{Y(X = 1) = 1\} - \text{pr}\{Y(X = 0) = 1\} = (q_{0,2} + q_{1,2}) - (q_{0,1} + q_{1,1}).
$$

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

$$
\begin{aligned}
-\text{pr}(X = 1, Y = 0) - \text{pr}(X = 0, Y = 1) \leq\ & \text{pr}\{Y(X = 1) = 1\} - \text{pr}\{Y(X = 0) = 1\} \\
\leq\ & 1 - \text{pr}(X = 1, Y = 0) - \text{pr}(X = 0, Y = 1),
\end{aligned}
$$

which coincide with the bounds derived in (Robins, 1989).

4.3. *Two instruments*

Our next example is shown in the DAG in Figure 6. This extends the instrumental variable example to the case where there are two variables on the left side that may be correlated 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 on directly on the outcome.

The bounds on risk difference $\text{pr}\{Y(X = 1)\} - \text{pr}\{Y(X = 0)\}$ under this DAG can be computed using our algorithm. In this problem, there are 16 constraints involving the conditional probabilities, the distribution of the response function variables 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 Supplemental Appendix along with code to reproduce the results using our algorithm.

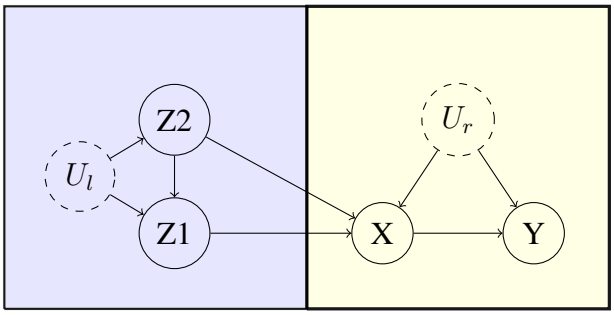


Fig. 6: Two instrumental variables example

We can also use the package to create `R` functions from the bounds, and compute them for specific values of the observed probabilities. We did this, simulating a large number of distributions that satisfy the DAG, and compare the bounds for when there are two instruments to the bounds we get when we assume that only one of the instruments is observed.

Specifically, we generated probability distributions $\text{pr}\{U_l, U_r, Z1, Z2, X, Y\}$ under the causal diagram in Figure 6 from the model

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

where $\Phi(x)$ is the cumulative distribution function of a standard normal random variable.

The results are shown in Figure 7 for 5,000 simulated distributions. The bounds with two instruments are never wider than those with only one instrument.

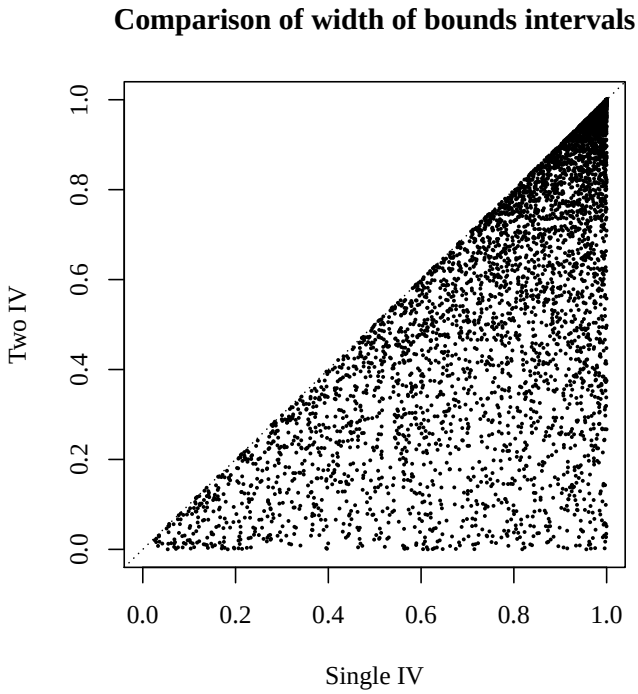


Fig. 7: Under a DAG with two instruments, this is a comparison of the width of the bounds intervals for the causal risk difference assuming only one of the instruments is observed to the width of the bounds assuming both are observed.

4.4. *Measurement error in the outcome*

Our final example illustrates some additional features of our method. In Figure 8, we have the variable X that is a cause of Y, but Y is not observed. Instead, $Y2$ which is a child of Y is observed that is also confounded with the true Y. In our `R` package, users can indicate that variables on the right side are unobserved (indicated by dashed circles) by selecting the node and typing 'u'. Additionally, we would like to include a user-specified constraint that can be specified in a text box of the web browser. The constraint is $Y2(Y = 1) \geq Y2(Y = 0)$, 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 in unobserved implies that we have 4 possible conditional probabilities to work with $\text{pr}(Y2 = y2|X = x)$, for $y2, x \in \{0, 1\}$. There are 12 parameters that characterize the distribution of the response function variables, and 5 constraints. The bounds for the risk difference $\theta = \text{pr}\{Y(X = 1) = 1\} - \text{pr}\{Y(X = 0) = 1\}$ computed using our method are

$$\begin{aligned}\max\{-1, 2\,\text{pr}(Y2 = 0|X = 0) - 2\,\text{pr}(Y2 = 0|X = 1) - 1\} \leq \theta \leq \\ \min\{1, 2\,\text{pr}(Y2 = 0|X = 0) - 2\,\text{pr}(Y2 = 0|X = 1) + 1\}.\end{aligned}$$

Except in cases where $\text{pr}(Y2 = 0|X = 0) = \text{pr}(Y2 = 0|X = 1)$, these bounds are informative; meaning they give an interval that is shorter than the widest possible interval for θ which is $[-1, 1]$.

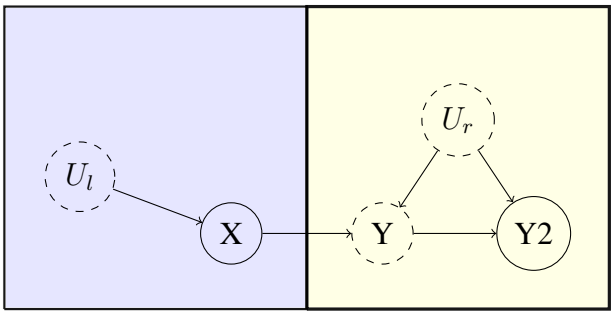


Fig. 8: Example with measurement error in the outcome. Dashed circles indicate unobserved variables.

5. CONCLUSION AND DISCUSSION

In this paper, we have described a general approach to symbolic computation of bounds on causal queries that are not identified from the true probability distribution of the observed variables. We described an algorithm that applies to a broad class of graphs combined with causal queries for which we have proven that the bounds are valid and tight. This has been implemented in the `R` package `causaloptim` with a user-friendly interface that allows for graphical description of DAGs and description of causal queries and constraints in a natural way (Sachs et al., 2020). All in a web browser, users can draw DAGs, describe causal targets, describe constraints, compute bounds, and output them as text, LaTeXformulas, or `R` functions. Advanced users can

interface with the algorithm directly using code, to ensure reproducibility or for more complex situations.

Our approach is useful is several novel scenarios, as we have illustrated with our examples. Additional applications of this method to unsolved problems in causal inference are now much more accessible to researchers. Our basic example and previously described bounds such as the instrumental variable (Balke & Pearl, 1994a), controlled direct effect (Cai et al., 2008), and natural direct effect (Sjölander et al., 2014) all run in a matter of seconds using our software on a modern laptop computer. The multiple instrumental variable problem takes approximately 6 hours, which involved enumerating 112 vertices twice (once for the upper bound and once for the lower). There is no theoretical upper limit to the number of vertices that can be enumerated using this approach. Modern vertex enumeration algorithms and implementations using parallel processing may allow currently unfeasible problems to be solved.

We cannot rule out that there exist problems outside of our class that can be stated as linear, so one suggestion for future work would be to identify a broader class of problems or a different algorithm that may apply on a case-by-case basis. Causal quantities such as the relative risk or odds ratio clearly imply a nonlinear optimization problem. Measured confounding, or knowledge about the absence of confounding often implies nonlinear constraints. We have assumed that all variables are binary, which is uncommon in real scientific problems that usually involve categorical or continuous variables. Extensions and insights into solving these sorts of problems would be useful in the causal inference community.

SUPPLEMENTAL MATERIAL

Supplementary material available at *Biometrika* online includes additional and more detailed results for the two instruments example. 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, and is also available at `https://sachsmc.github.io/causaloptim/articles/example-code.html`.

REFERENCES

BALKE, A. & PEARL, J. (1994a). Counterfactual probabilities: Computational methods, bounds and applications. In *Proceedings of the Tenth international conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc.

BALKE, A. & PEARL, J. (1994b). Probabilistic evaluation of counterfactual queries. In *Proceedings of the twelfth national conference on artificial intelligence*. The AAAI Press, Menlo Park, California.

BALKE, A. & PEARL, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association* **92**, 1171–1176.

BONET, B. (2013). Instrumentality tests revisited , arXiv: 1301.2258.

CAI, Z., KUROKI, M., PEARL, J. & TIAN, J. (2008). Bounds on direct effects in the presence of confounded intermediate variables. *Biometrics* **64**, 695–701.

DANTZIG, G. B. (1963). *Linear Programming and Extensions*. Princeton University Press.

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

MANSKI, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review* **80**, 319–323.

MATTHEISS, T. H. (1973). An algorithm for determining irrelevant constraints and all vertices in systems of linear inequalities. *Operations Research* **21**, 247–260.

PEARL, J. (2009). *Causality*. Cambridge university press.

R CORE TEAM (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RAMSAHAI, R. R. (2012). Causal Bounds and Observable Constraints for Non-deterministic Models. *Journal of Machine Learning Research* **13**, 829–848.

ROBINS, J. M. (1989). The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. *Health service research methodology: a focus on AIDS* , 113–159.

SACHS, M. C., SJÖLANDER, A. & GABRIEL, E. E. (2020). *causaloptim: An Interface to Specify Causal Graphs and Compute Bounds on Causal Effects*. R package version 0.6.4.

SJÖLANDER, A. (2009). Bounds on natural direct effects in the presence of confounded intermediate variables. *Statistics in Medicine* **28**, 558–571.

SJÖLANDER, A., LEE, W., KÄLLBERG, H. & PAWITAN, Y. (2014). Bounds on causal interactions for binary outcomes. *Biometrics* **70**, 500–505.

ZHANG, J. L. & RUBIN, D. B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *Journal of Educational and Behavioral Statistics* **28**, 353–368.

[*Received on* 8 *April* 2020. *Editorial decision on* 1 *April* 2021]