Scalable Computation of Causal Bounds

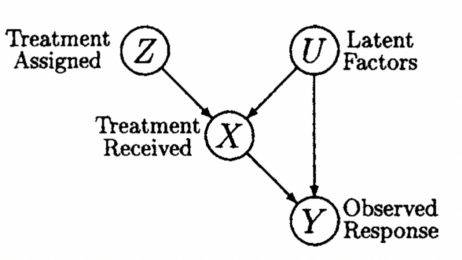
Madhumitha Shridharan
Joint Work with Garud Iyengar

Department of Industrial Engineering and Operations Research
Columbia University

June 24, 2022

Problem

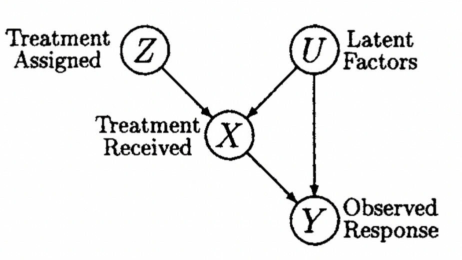
- Consider the following instrumental variable setting with $X, Y, Z \in \{0, 1\}$ i.e. binary:



- **Goal**: Bounds for query $\mathbb{P}(Y|do(X))$ given distribution $\mathbb{P}(X, Y|Z)$ from observational data.

Prior Work: Modeling unobserved confounders

- For fixed U: X is a function of Z, and Y is a function of X.



- **Insight**: Only need to model impact of U on the **dependence** between variables.

Prior Work: Modeling unobserved confounders (contd)

- Define
  - $\mathcal{F} = \{f : f \text{ is a function from } Z \to X\}$
  - $\mathcal{G} = \{g : g \text{ is a function from } X \to Y\}$

- U effectively selects one function each from $\mathcal{F}$ and $\mathcal{G}$

Prior Work: Modeling unobserved confounders (contd)

- Define
  - $\mathcal{F} = \{f : f \text{ is a function from } Z \to X\}$
  - $\mathcal{G} = \{g : g \text{ is a function from } X \to Y\}$
- U effectively selects one function each from $\mathcal{F}$ and $\mathcal{G}$
- Index elements in $\mathcal{F}$ and $\mathcal{G}$ can be indexed by $r = (r_X, r_Y) \in R = \{1, \ldots, 4\}^2$
  - f_{r_X} denotes the r_X-th function from $\mathcal{F}$
  - g_{r_Y} denotes the r_Y-th function from $\mathcal{G}$
- $|R|$ exponential in the number of arcs

Prior work: Bounds via linear programming

- Variable for the LP are $q_{r_X r_Y} = \mathbb{P}(r_X, r_Y)$.

Prior work: Bounds via linear programming

- Variable for the LP are $q_{r_X r_Y} = \mathbb{P}(r_X, r_Y)$.

- Constraints of the LP:

$$\mathbb{P}(X = x, Y = y | Z = z) = \sum_{(r_X, r_Y) \in R_{xy.z}} q_{r_X r_Y}$$

where

$$R_{xy.z} = \{(r_X, r_Y) : f_{r_X}(z) = x, g_{r_Y}(x) = y\}, \quad (1)$$

denote the set of r-values that map $z \mapsto (x, y)$.

Prior work: Bounds via linear programming

- Variable for the LP are $q_{r_X r_Y} = \mathbb{P}(r_X, r_Y)$.
- Constraints of the LP:

$$\mathbb{P}(X = x, Y = y | Z = z) = \sum_{(r_X, r_Y) \in R_{xy.z}} q_{r_X r_Y}$$

where

$$R_{xy.z} = \{(r_X, r_Y) : f_{r_X}(z) = x, g_{r_Y}(x) = y\}, \quad (1)$$

denote the set of r-values that map $z \mapsto (x, y)$.

- Objective of the LP: $\mathbb{P}(Y = 1 | do(X = 1)) = \sum_{(r_X, r_Y) \in R_Q} q_{r_X r_Y}$ where

$$R_Q = \{(r_X, r_Y) : g_{r_Y}(1) = 1\} \quad (2)$$

Prior work: Bounds via linear programming

The bounds α_L/α_U of $\mathbb{P}(Y = 1 | do(X = 1))$ are given by:

$$
\begin{aligned}
\alpha_L/\alpha_U = \quad \min_q / \ \max_q \quad & \sum_{(r_X, r_Y) \in R_{\mathcal{Q}}} q_{r_X r_Y} \\
\text{s.t.} \quad & \sum_{(r_X, r_Y) \in R_{xy.z}} q_{r_X r_Y} = \mathbb{P}(X = x, Y = y | Z = z), \ \forall (x, y, z) \\
& q \geq 0,
\end{aligned}
$$

Prior work: Bounds via linear programming

The bounds α_L/α_U of $\mathbb{P}(Y=1|do(X=1))$ are given by:

$$
\begin{aligned}
\alpha_L/\alpha_U = \quad \min_q / \ \max_q \quad & \textstyle\sum_{(r_X,r_Y)\in R_{\mathcal{Q}}} q_{r_X r_Y} \\
\text{s.t.} \quad & \textstyle\sum_{(r_X,r_Y)\in R_{xy.z}} q_{r_X r_Y} = \mathbb{P}(X=x, Y=y|Z=z), \ \ \forall (x,y,z) \\
& q \geq 0,
\end{aligned}
$$

Number of variables grows exponentially with number of edges!

Aggregating Variables

- Fix a function $h : Z \to (X, Y)$. Let

$$R_h = \left\{ (r_X, r_Y) \in R : \begin{array}{l} \big(f_{r_X}(0), g_{r_Y}(f_{r_X}(0))\big) = h(0) \\ \big(f_{r_X}(1), g_{r_Y}(f_{r_X}(1))\big) = h(1) \end{array} \right\} \quad (3)$$

denote the set of (r_X, r_Y) values consistent with h.

Aggregating Variables

- Fix a function $h : Z \to (X, Y)$. Let

$$R_h = \left\{ (r_X, r_Y) \in R : \begin{array}{l} \big(f_{r_X}(0), g_{r_Y}(f_{r_X}(0))\big) = h(0) \\ \big(f_{r_X}(1), g_{r_Y}(f_{r_X}(1))\big) = h(1) \end{array} \right\} \tag{3}$$

denote the set of (r_X, r_Y) values consistent with h.

- All $q_{r_X r_Y}$ variables with $(r_X, r_Y) \in R_h$ contribute to the same two constraints:

$$(x, y, z) = \big(h(0), 0\big) \qquad (x, y, z) = \big(h(1), 1\big)$$

Aggregating Variables

- Fix a function $h : Z \to (X, Y)$. Let

$$R_h = \left\{ (r_X, r_Y) \in R : \begin{array}{l} \left(f_{r_X}(0), g_{r_Y}(f_{r_X}(0))\right) = h(0) \\ \left(f_{r_X}(1), g_{r_Y}(f_{r_X}(1))\right) = h(1) \end{array} \right\} \quad (3)$$

denote the set of (r_X, r_Y) values consistent with h.

- All $q_{r_X r_Y}$ variables with $(r_X, r_Y) \in R_h$ contribute to the same two constraints:

$$(x, y, z) = \left(h(0), 0\right) \qquad (x, y, z) = \left(h(1), 1\right)$$

- Therefore, the LP can be reformulated as

$$\begin{array}{rl} \min_q & \sum_{h \in H} c_h q_h \\ \text{s.t.} & \sum_{h \in H : h(z) = (x,y)} q_h = p_{xy.z}, \ \forall (x, y, z) \\ & q \geq 0, \end{array} \quad (4)$$

H denotes the set of *valid* hyperarcs for which $R_h \neq \emptyset$. Not all hyperarcs are valid!

Main contributions

- Savings realized only if set H and c_h can be efficiently computed without enumerating over the set R, and we show how to do so.

Main contributions

- Savings realized only if set H and c_h can be efficiently computed without enumerating over the set R, and we show how to do so.

- Aggregated formulation allows closed form expression for bounds in special cases.

Closed Form Bounds: Multi-Cause Setting with Unobserved Confounders

- T_i, $i = 1, \ldots, 5$, indicates whether the patient was prescribed treatment i
- Y indicates the progression of the disease in the patient.
- C_i, $i = 1, \ldots, 2$, indicates the presence of pre-existing condition i in the patient
- U_A is an unobserved confounder (e.g. a patient characteristic)
- U_B is an unobserved confounder (e.g. doctor biases, treatment preferences)

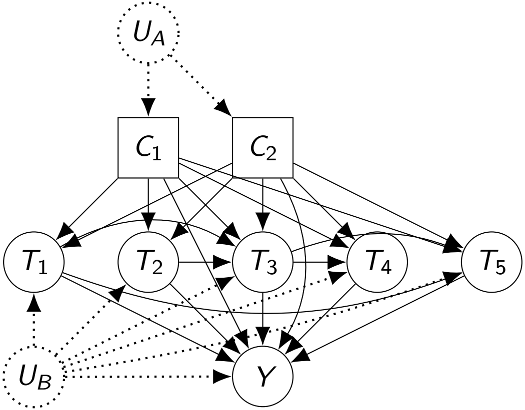


Figure: Computing $E[Y|do(\mathbf{T}, \mathbf{C})]$ in Closed Form