Estimating Causal Effects Identifiable from a Combination of Observations and Experiments

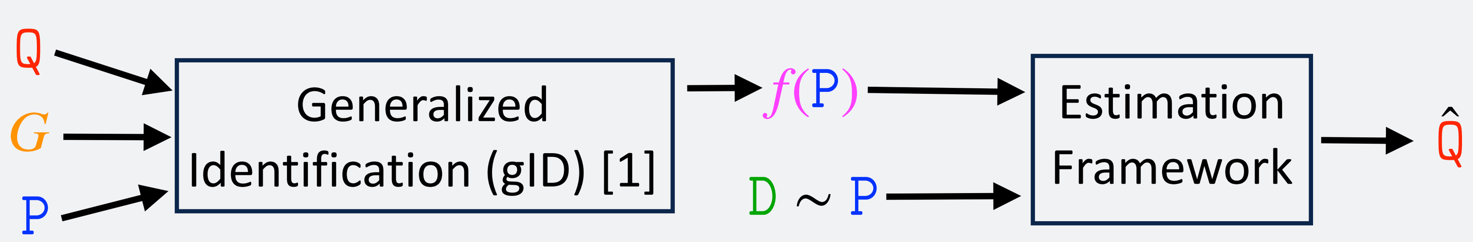
Yonghan Jung[1], Ivan Diaz[2], Jin Tian[3], Elias Bareinboim[4]

[1] Purdue University, [2] New York University,
[3] Iowa State University, [4] Columbia University

Overview

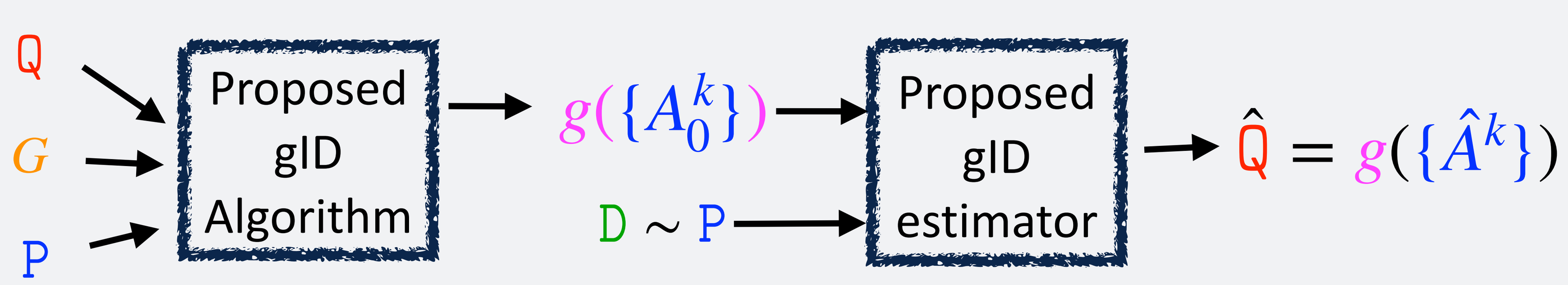
- Causal effect identification represents a causal effect as a function of the input distribution, using on a causal graph.
- A sound and complete identification algorithm (gID) has been developed for cases where the input distributions are a combination of obs & exp distributions [1].
- We have developed an estimation framework for the identifiable causal functional. The proposed estimator exhibits multiply-robustness and fast convergence.

Problem setup



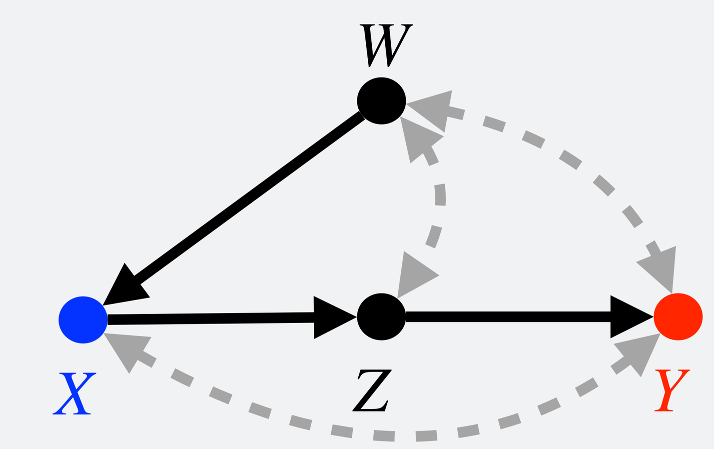
- **Inputs:** Q: Query (:= $P(\mathbf{y} \mid do(\mathbf{x}))$) G: Causal graph, P: Input distributions $\mathrm{P} := \{P(\mathbf{V} \mid do(\mathbf{Z}_1)), \cdots, P(\mathbf{V} \mid do(\mathbf{Z}_m))\}$,
- **gID**: The gID algorithm outputs the causal estimand $f(\mathrm{P})$.
- **Goal:** Given samples $\mathrm{D} := \{D_1, \cdots, D_m\}$ for $D_i \sim P(\mathbf{V} \mid do(\mathbf{Z}_i))$, construct an estimator $\hat{\mathrm{Q}}$ of the query Q.

Proposed estimation pipeline



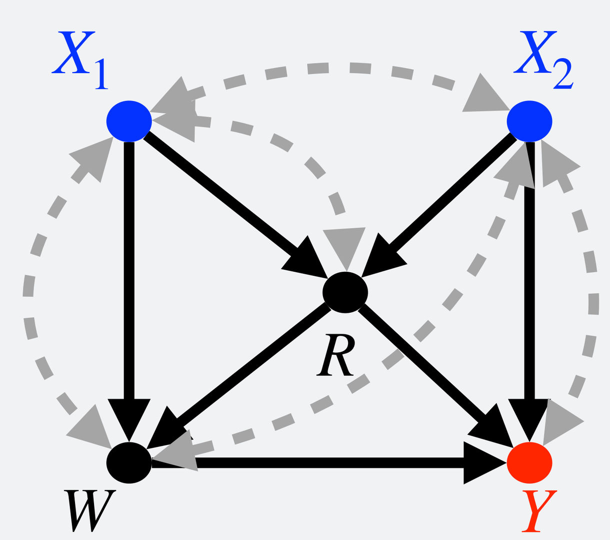
1. Causal estimand $f(\mathrm{P})$ can be represented as a function of g-formula-types estimands (g-mSBD); i.e.,
$$P(\mathbf{y} \mid do(\mathbf{x})) = g(\{A_0^k : k = 1, \cdots, m\})$$
2. The robust estimator can be constructed from $\mathrm{D} \sim \mathrm{P}$ by plugging-in the robust g-mSBD estimator into the function.
$$\widehat{P(\mathbf{y} \mid do(\mathbf{x}))} = g(\{\hat{A}^k : k = 1, \cdots, m\})$$

Practical gID Scenarios



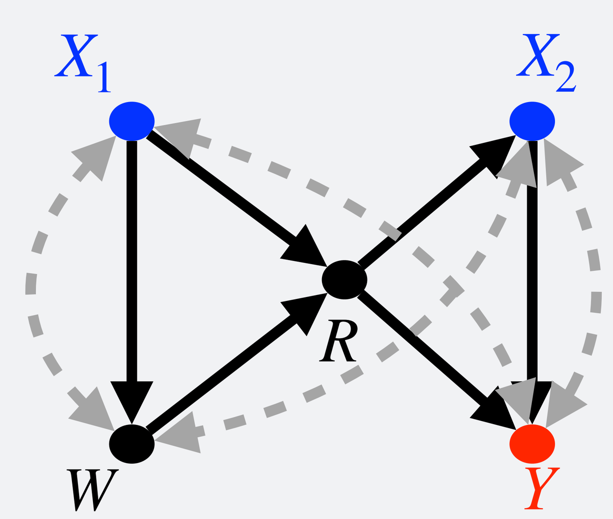
X: Eligibility of the job training
Y: Future salary.
Z: Participation of job training
W: Past average income
$\mathbf{V} := \{W, X, Z, Y\}$

- **Input**: Samples $\{D, D_z\}$ where $D \sim P(\mathbf{V})$ and $D_z \sim P(\mathbf{V} \mid do(Z))$
- **gID**: Identify $P(y \mid do(x)) = f(\{P(\mathbf{V} \mid do(Z)), P(\mathbf{V})\})$ where
$$\sum_z P(y \mid do(z)) \sum_w P(z \mid x, w) P(w).$$



X_1: Antihypertensive drug
X_2: Anti-diabetic drug
W: Blood pressure
Y: Cardiovascular disease.
R: Mediators
$\mathbf{V} := \{X_1, X_2, R, W, Y\}$

- **Input**: Samples $\{D_{x_1}, D_{x_2}\}$ where $D_{x_i} \sim P(\mathbf{V} \mid do(x_i))$
- **gID**: Identify $P(y \mid do(x_1, x_2)) = f(\{P(\mathbf{V} \mid do(x_i)) : i = 1,2\})$ as
$$\sum_{r,w} P(y \mid r, w, do(x_2)) P(r \mid x_2, do(x_1)) \sum_{x_2'} P(w \mid r, x_2', do(x_1)) P(x_2' \mid do(x_1))$$



X_1: Class size at the kindergarten
X_2: Class size at the 3rd grade
R: Academic outcome at 2nd grade.
W: Academic outcome at kindergarten
Y: Academic outcome at 3rd grade
$\mathbf{V} := \{X_1, X_2, R, W, Y\}$

- **Input**: Samples $\{D_{x_1}, D_{x_2}\}$ where $D_{x_i} \sim P(\mathbf{V} \mid do(x_i))$
- **gID**: Identify $P(y \mid do(x_1, x_2)) = f(\{P(\mathbf{V} \mid do(x_i)) : i = 1,2\})$ as
$$\sum_r P(r \mid do(x_1)) \sum_{x_1', w} P(y \mid x_1', w, r, do(x_2)) P(x_1', w \mid do(x_2))$$

Closer look to the pipeline

g-mSBD operator

For an ordered set $\mathbf{W}, \mathbf{C}, \mathbf{R}$ and a set of intervened variables in the input distributions, $\mathbf{Z}_1, \cdots, \mathbf{Z}_m$, the g-formula-type estimand **g-mSBD** A_0 is:

$$A_0[\mathbf{W}, \mathbf{C}, \mathbf{R}; (\mathbf{z}_1, \cdots, \mathbf{z}_m)](\mathbf{w} \backslash \mathbf{c}, \mathbf{r}) := \sum_c \prod_{i: W_i \in \mathbf{W}} P(w_i \mid \mathbf{w}^{(i-1)}, \mathbf{r}^{(i-1)}, do(\mathbf{z}_i))$$

Multiply-robust g-mSBD estimator

For $\mu_0^{m+1} = \mathbb{I}(\mathbf{W} \backslash \mathbf{C} = \mathbf{w} \backslash \mathbf{c})$, and for $i = m-1, \cdots, 1$,

μ_0^{i+1}	$\mathbb{E}\left[\mu_0^{i+2}(\mathbf{W}^{(i+1)}, \mathbf{r}_{i+1}, \mathbf{R}^{(i)}) \mid \mathbf{W}^{(i)}, \mathbf{R}^{(i)}, do(\mathbf{z}_i)\right]$
$\bar{\mu}_0^{i+1}$	$\mu_0^{i+1}(\mathbf{W}^{(i)}, \mathbf{R}^{(i-1)}, \mathbf{r}_i, do(\mathbf{z}_i))$
π_0^i	$\frac{P(\mathbf{W}^{(i)}, \mathbf{R}^{(i-1)} \mid do(\mathbf{z}_i))}{P(\mathbf{W}^{(i)}, \mathbf{R}^{(i-1)} \mid do(\mathbf{z}_{i+1}))} \frac{\mathbb{I}(\mathbf{R}_i = \mathbf{r}_i)}{P(\mathbf{R}_i \mid \mathbf{W}^{(i)}, \mathbf{R}^{(i-1)}, do(\mathbf{z}_{i+1}))}$

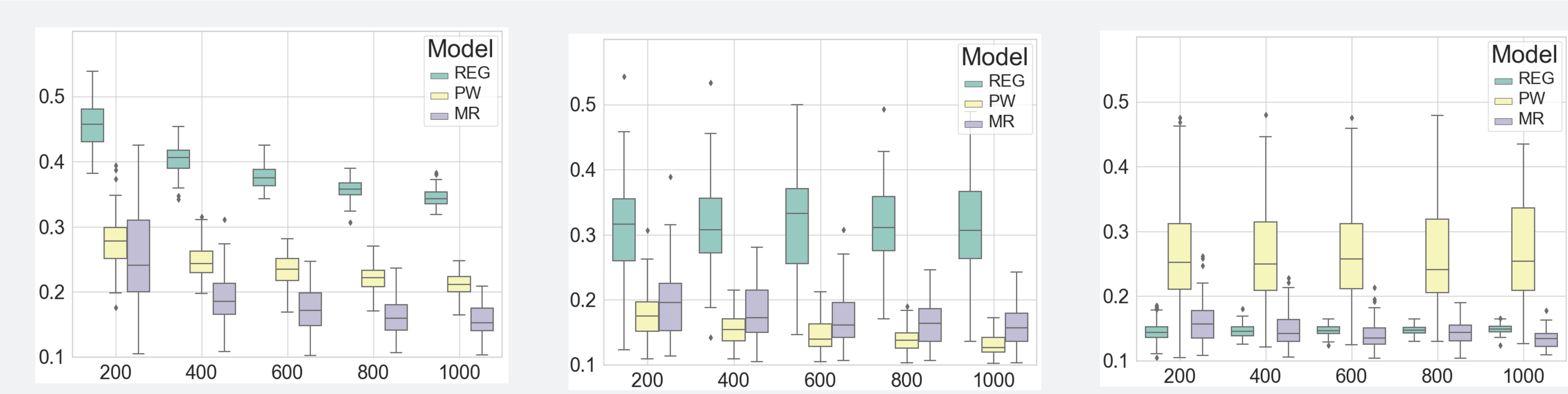
The multiply robust (w.r.t. $\{\mu^i, \pi^i\}$) estimator $\hat{A}$ for A_0 is

$$\hat{A} := \sum_{j=2}^{m} \mathbb{E}_{D_{\mathbf{z}_j}}[\pi^{(j)}\{\bar{\mu}^{i+1} - \mu^i\}] + \mathbb{E}_{D_{\mathbf{z}_1}}[\mu^1].$$

Representation of causal effect and proposed estimator

1. Given inputs (Q, G, P), the proposed gID algorithm represents Q as a function of g-mSBD operators; i.e., $\mathrm{Q} = g(\{A_0^k : k = 1, \cdots, m\})$.
2. The multiply robust (w.r.t. $\{\mu^i, \pi^i\}$) estimator is constructed from $\mathrm{D} \sim \mathrm{P}$ as $\hat{\mathrm{Q}} = g(\{\hat{A}^k\})$.

Simulation (for example 1)



(a) Fast Convergence (b) Multiply Robustness (μ) (c) Multiply Robustness (π)

This result showed the multiply robustness and fast convergence of the proposed estimator.

[1] Lee et al., General Identifiability with Arbitrary Surrogate Experiments, UAI 2019