Causal Inference II
(COMS W4775/Spring 2021)

Lecture 7: Causal Effect Estimation for Arbitrary Functionals

Invited Lecturer: Yonghan Jung

Elias Barenboim
Columbia University
Twitter: @eliasbareinboim

Outline

- **Part 1 [10:10 - 11:40]**: Semiparametric approach

  Jung, Tian, Bareinboim (2021), Estimating Identifiable Causal Effects through Double Machine Learning. In Proceedings of the 35th AAAI Conference on AI, 2021.

- **Break [11:40 - 11:50]**

- **Part 2 [11:50 - 12:40]**: Empirical risk minimization approach

  Jung, Tian, Bareinboim (2020), Learning Causal Effects via Weighted Empirical Risk Minimization. In Proceedings of the 34th Neural Information Processing Systems, 2020.

Part I.

Semiparametric inference

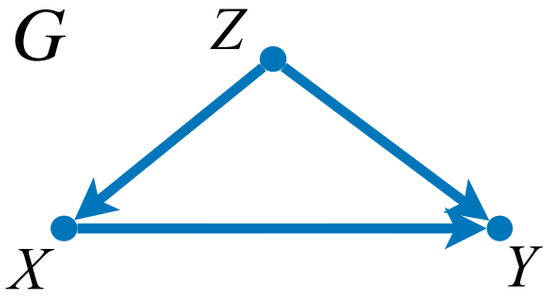
Problem setup

The task of Identification (ID)

1 query

$$Q = P_{\mathbf{x}}(\mathbf{y}) \equiv P(\mathbf{y} \,|\, do(\mathbf{x}))$$

2 graph



3 probability

$$P(\mathbf{V})$$

With the current scientific knowledge (encoded as a graph) about the problem (2) and the available distribution (3), can we answer the research question (1)?

ID (G, P, Q)

solution

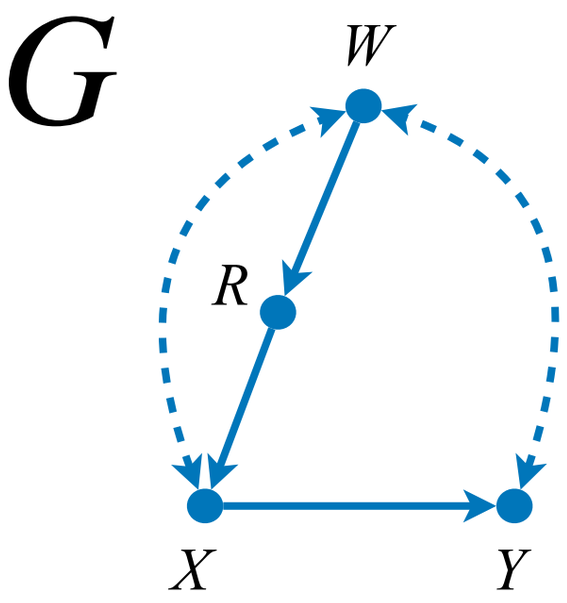
yes / no

Causal Functional

$$P_{\mathbf{x}}(\mathbf{y}) = f(P)$$

J. Pearl, Causal Diagrams for empirical research. Biometrika, Vol. 82, p. 669-709, 1995.

Working example: Napkin



- Y blood pressure;
- X cardiac output;
- R heart rate; and
- W hormone level.

- The goal is to identify $P_x(y)$, the effect of the cardiac output to the blood pressure.

- ID$(G, P, P_x(y))$: $\quad P_x(y) = f(P) = \dfrac{\sum_w P(y, x \mid r, w)P(w)}{\sum_w P(x \mid r, w)P(w)}$

Working example: Napkin

- A causal functional $f(P)$ can be evaluated if $P(v)$ is available.
- In practice, we only have samples D drawn from P, not the distribution itself. In practice, we need to *estimate* the identifiable functional from finite samples.

- The goal is to identify $P_x(y)$, the effect of the cardiac output to the blood pressure.
- ID$(G, P, P_x(y))$: $P_x(y) = f(P) = \dfrac{\sum_w P(y, x \mid r, w)P(w)}{\sum_w P(x \mid r, w)P(w)}$

Task: Causal Effects Estimation

$\text{ID}\left(G, P, P_{\mathbf{x}}(\mathbf{y})\right)$

With (1) a causal functional $f(P)$ such that $P_{\mathbf{x}}(\mathbf{y}) = f(P)$ and (2) a dataset D, can we have a reliable estimate T_N for $P_{\mathbf{x}}(\mathbf{y})$?

Alternative representation for amenable estimation

1 Causal Functional

$P(y \mid do(x)) = f(P)$

2 Data

$D = \{\mathbf{V}_{(i)}\}_{i=1}^{N} \sim P$

Estimation Engine

$\text{EST}(f(P), D)$

Estimand

$g(P) = f(P)$

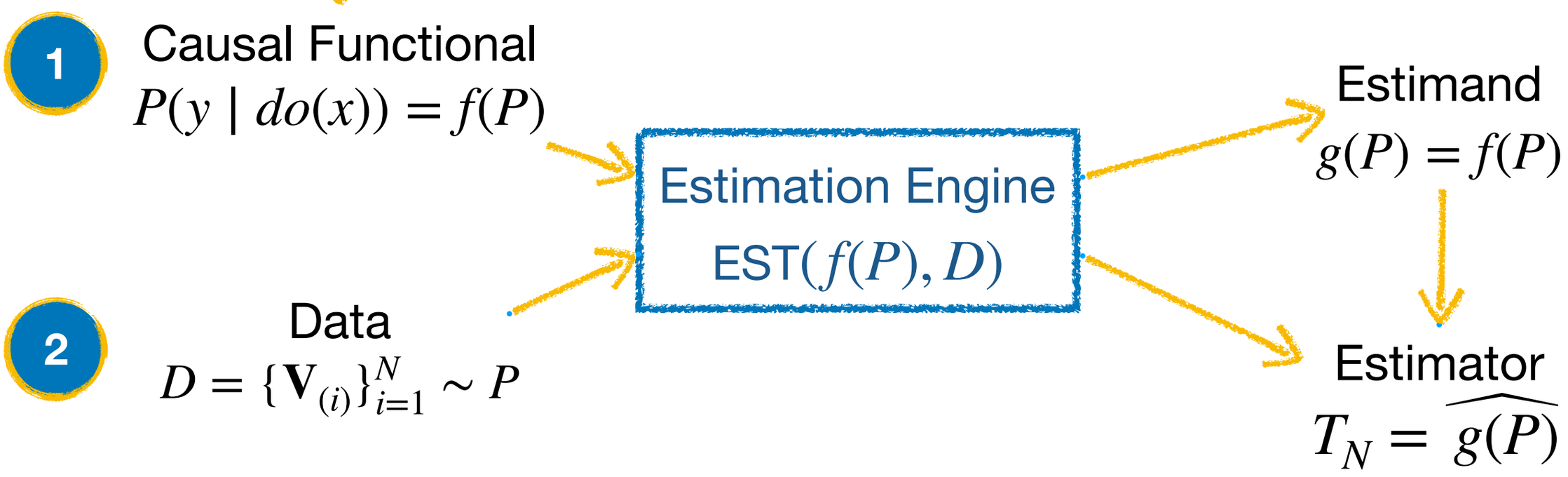
Estimator

$T_N = \widehat{g(P)}$

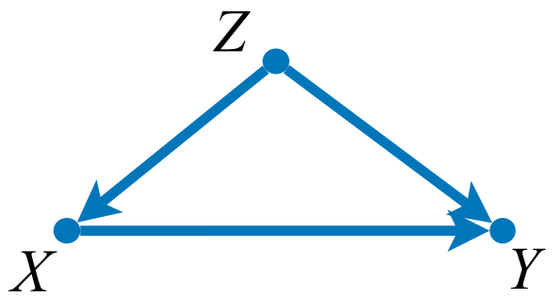
Task 2: Causal Effect Estimation

- What estimands $g(P)$ and estimators T_N are available?
- Among available $(g(P), T_N)$, which one should be preferred?

This problem has been only studied for a small portion of all identifiable scenarios.

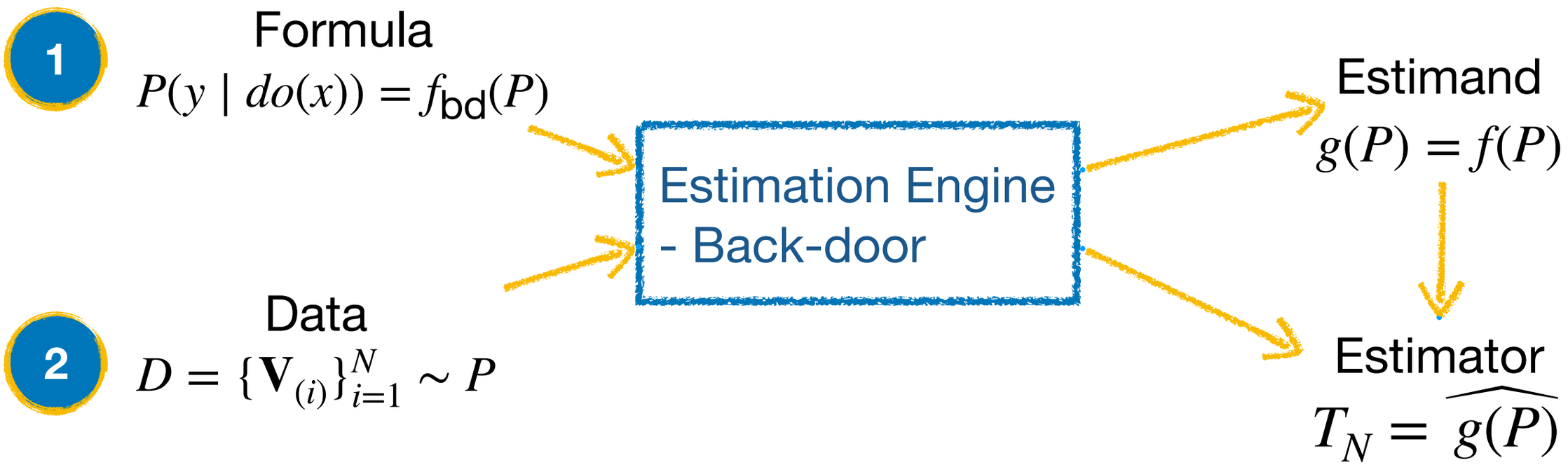


Small portion of Estimation engine — Back-door (a.k.a. ignorability)

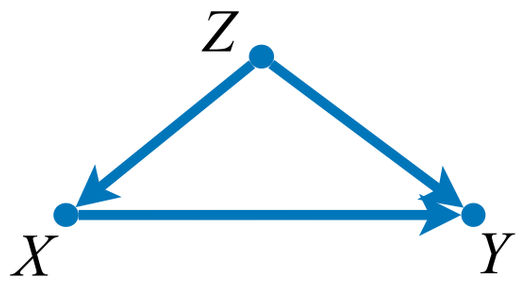


$$P(y \mid do(x)) = f_{\text{bd}}(P) = \sum_{z} P(y \mid x, z)P(z)$$

With (1) a back-door adjustment formula $f_{\text{bd}}(P)$ and (2) a dataset D, can we have a reliable estimate T_N for $P_{\mathbf{x}}(\mathbf{y})$?



Classic BD estimator: Inverse probability weighting (IPW)



$$P_x(y) = \sum_z P(y \mid x, z) P(z)$$

$$\sum_z P(y \mid x, z) P(z) = \sum_z P(y \mid x, z) \frac{P(x \mid z) P(z)}{P(x \mid z)} \frac{1}{}$$

$$= \sum_z P(z, x, y) \frac{1}{P(x \mid z)}$$

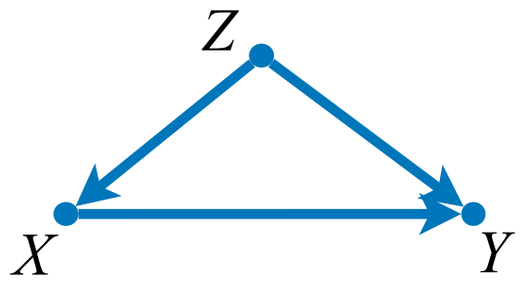
$$= \sum_{z, x', y'} P(z, x', y') \frac{I_x(x')}{P(x \mid z)} I_y(y')$$

$$= \mathbb{E}_P \left[\frac{I_x(X)}{P(X \mid Z)} \cdot I_y(Y) \right]$$

$$= g(P)$$

$I_x(x') = 1$ if $x = x'$; otherwise 0.

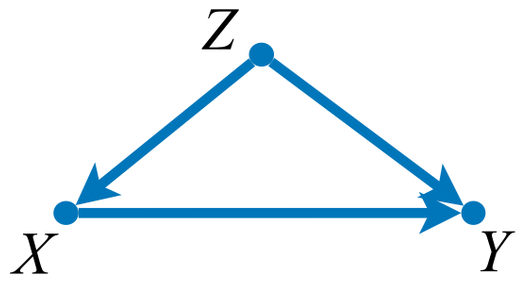
Classic BD estimator:
1. Inverse probability weighting (IPW)



$$P_x(y) = f_{\text{bd}}(P) = \sum_z P(y \mid x, z) P(z)$$

Estimand ($g(P)$)	$\mathbb{E}_P \left[\frac{I_x(X)}{P(X \mid Z)} \cdot I_y(Y) \right]$
Estimator (T_N)	$\mathbb{E}_D \left[\frac{I_x(X)}{\hat{P}(X \mid Z)} \cdot I_y(Y) \right]$
For consistent estimation	$\hat{P}(x \mid z)$ converges to $P(x \mid z)$.
For fast ($N^{-1/2}$ rate) convergence	(1) $\hat{P}(x \mid z) \to P(x \mid z)$ fast; and (2) $\{\hat{P}(x \mid z), P(x \mid z)\}$ in Donsker Function class.

Classic BD estimator:
1. Inverse probability weighting (IPW)

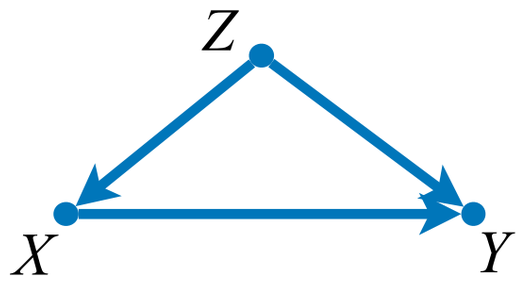


$$P_x(y) = f_{\text{bd}}(P) = \sum_z P(y \mid x, z) P(z)$$

Estimand ($g(P)$)	$\mathbb{E}_P\left[\frac{I_x(X)}{P(X \mid Z)} \cdot I_y(Y)\right]$
Estimator (T_N)	$\mathbb{E}_D\left[\frac{I_x(X)}{\hat{P}(X \mid Z)} \cdot I_y(Y)\right]$
For consistent estimation	$\hat{P}(x \mid z)$ converges to $P(x \mid z)$.
For fast ($N^{-1/2}$ rate) convergence	(1) $\hat{P}(x \mid z) \to P(x \mid z)$ fast; and (2) $\{\hat{P}(x \mid z), P(x \mid z)\}$ in Donsker Function class.

$\mathbb{E}_D[f(\mathbf{W})] \equiv \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{W}_{(i)})$, an empirical expectation of $f(\mathbf{W})$ using samples.

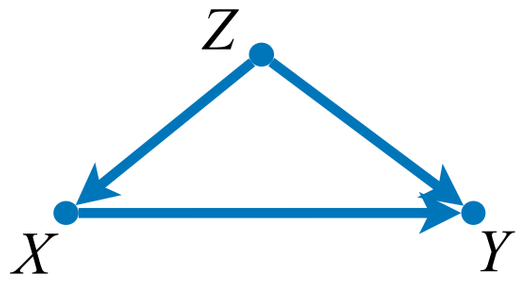
Classic BD estimator:
1. Inverse probability weighting (IPW)



$$P_x(y) = f_{\text{bd}}(P) = \sum_z P(y \mid x, z) P(z)$$

Estimand ($g(P)$)	$\mathbb{E}_P\left[\frac{I_x(X)}{P(X \mid Z)} \cdot I_y(Y)\right]$
Estimator (T_N)	• A function class s.t. complexities of functions are restricted. (e.g., linear/logistic regression, smooth parametric functions). • It's unclear modern flexible/complicated ML methods (e.g., neural networks) are in this class.
For consistent estimation	
For fast ($N^{-1/2}$ rate) convergence	(1) $\hat{P}(x \mid z) \to P(x \mid z)$ fast; and (2) $\{\hat{P}(x \mid z), P(x \mid z)\}$ in Donsker Function class.

Classic BD estimator: 2. Outcome-regression (OR)



$$P_x(y) = f_{\text{bd}}(P) = \sum_z P(y \mid x, z) P(z)$$

Estimand ($g(P)$)	$\mathbb{E}_P[P(y \mid x, Z)]$
Estimator (T_N)	$\mathbb{E}_D\left[\hat{P}(y \mid x, Z)\right]$
For correct estimation	$\hat{P}(y \mid x, z)$ converges to $P(y \mid x, z)$.
For fast ($N^{-1/2}$ rate) convergence	(1) $\hat{P}(y \mid x, z) \to P(y \mid x, z)$ fast; and (2) $\{\hat{P}(y \mid x, z), P(y \mid x, z)\}$ in Donsker Function class.

Expectation over Z.
$= \sum_z P(y \mid x, z) P(z)$

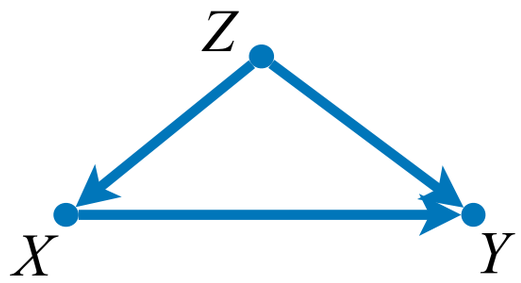
Comparison
Classic BD estimators

	IPW	**OR**
Estimand ($g(P)$)	$\mathbb{E}_P\left[\frac{I_x(X)}{P(X\mid Z)}\cdot I_y(Y)\right]$	$\mathbb{E}_P\left[P(y\mid x,Z)\right]$
Estimator (T_N)	$\mathbb{E}_D\left[\frac{I_x(X)}{\hat{P}(X\mid Z)}\cdot I_y(Y)\right]$	$\mathbb{E}_D\left[\hat{P}(y\mid x,Z)\right]$
For correct estimation	$\hat{P}(x\mid z)\to P(x\mid z)$	$\hat{P}(y\mid x,z)\to P(y\mid x,z)$
For fast ($N^{-1/2}$ rate) convergence	(1) $\hat{P}(x\mid z)\to P(x\mid z)$ fast; (2) $\{\hat{P}(x\mid z), P(x\mid z)\}$ in Donsker class.	(1) $\hat{P}(y\mid x,z)\to P(y\mid x,z)$ fast; (2) $\{\hat{P}(y\mid x,z), P(y\mid x,z)\}$ in Donsker class.

Comparison
Classic BD estimators

	IPW	**OR**
Estimand ($g(P)$)	$\mathbb{E}_P\left[\frac{I_x(X)}{P(X\mid Z)}\cdot I_y(Y)\right]$	$\mathbb{E}_P\left[P(y\mid x,Z)\right]$
Estimator (T_N)	$\mathbb{E}_D\left[\frac{I_x(X)}{\hat{P}(X\mid Z)}\cdot I_y(Y)\right]$	$\mathbb{E}_D\left[\hat{P}(y\mid x,Z)\right]$
For correct estimation	**(Single chance of being correct)** If $\hat{P}(x\mid z)$ ($\hat{P}(y\mid x,z)$) is misspecified, then T_N fails to converge	
For fast ($N^{-1/2}$ rate) convergence	(1) $\hat{P}(x\mid z)\to P(x\mid z)$ fast; (2) $\{\hat{P}(x\mid z), P(x\mid z)\}$ in Donsker class.	(1) $\hat{P}(y\mid x,z)\to P(y\mid x,z)$ fast; (2) $\{\hat{P}(y\mid x,z), P(y\mid x,z)\}$ in Donsker class.

Double chance of being correct! Augmented IPW (IPW + OR)[1]



$$P_x(y) = \sum_z P(y \mid x, z) P(z)$$

- Consider the following estimand: $g(P) \equiv \mathbb{E}_P \left[\frac{I_x(X)}{P(X \mid Z)} (I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z) \right]$

If $P(X \mid Z)$ is misspecified to $\tilde{P}(X \mid Z)$:

$$\mathbb{E}_P \left[\frac{I_x(X)}{\tilde{P}(X \mid Z)} (I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z) \right]$$

(Law of total expectation): Taking expectation to Y and (X, Z) in sequence.

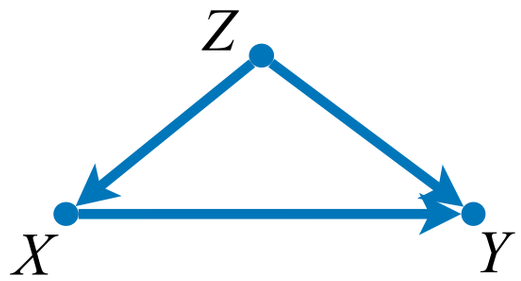
$$= \mathbb{E}_{P(X,Z)} \left\{ \mathbb{E}_{P(Y \mid X,Z)} \left[\frac{I_x(X)}{\tilde{P}(X \mid Z)} (I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z) \,\middle|\, X, Z \right] \right\}$$

Since $\mathbb{E}_{P(Y \mid X,Z)}[I_y(Y) \mid X, Z] = P(y \mid X, Z)$

$$= \mathbb{E}_{P(X,Z)} \left\{ \frac{I_x(X)}{\tilde{P}(X \mid Z)} (\cancel{P(y \mid X, Z) - P(y \mid X, Z)}) + P(y \mid x, Z) \right\}$$

$$= \mathbb{E}_P[P(y \mid x, Z)] = \sum\nolimits_z P(y \mid x, z) P(z) = P_x(y)$$

Double chance of being correct! Augmented IPW (IPW + OR)



$$P_x(y) = \sum_z P(y \mid x, z) P(z)$$

- Consider the following estimand: $g(P) \equiv \mathbb{E}_P \left[\frac{I_x(X)}{P(X \mid Z)} (I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z) \right]$

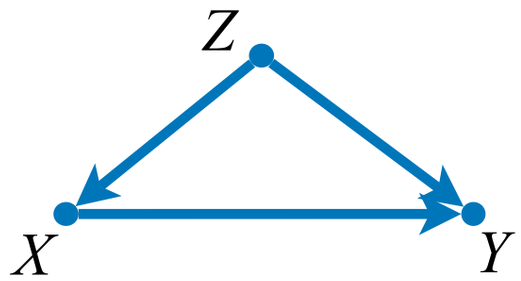
If $P(y \mid X, Z)$ is misspecified to $\tilde{P}(y \mid X, Z)$:

$$\mathbb{E}_P \left[\frac{I_x(X)}{P(X \mid Z)} (I_y(Y) - \tilde{P}(y \mid X, Z)) + \tilde{P}(y \mid x, Z) \right] = \mathbb{E}_{P(X,Z)} \left[\mathbb{E}_{P(Y \mid X,Z)} \left\{ \frac{I_x(X)}{P(X \mid Z)} (I_y(Y) - \tilde{P}(y \mid X, Z)) + \tilde{P}(y \mid x, Z) \,\middle|\, X, Z \right\} \right]$$

$$= \mathbb{E}_{P(X,Z)} \left[\frac{I_x(X)}{P(X \mid Z)} (P(y \mid X, Z) - \tilde{P}(y \mid X, Z)) + \tilde{P}(y \mid x, Z) \right] = \mathbb{E}_{P(Z)} \left[\mathbb{E}_{P(X \mid Z)} \left\{ \frac{I_x(X)}{P(X \mid Z)} (P(y \mid X, Z) - \tilde{P}(y \mid X, Z)) + \tilde{P}(y \mid x, Z) \,\middle|\, Z \right\} \right]$$

$$= \mathbb{E}_{P(Z)} \left[\frac{\cancel{P(x \mid Z)}}{\cancel{P(x \mid Z)}} (P(y \mid x, Z) - \tilde{P}(y \mid x, Z)) + \tilde{P}(y \mid x, Z) \right] = \mathbb{E}_P[P(y \mid x, Z)] = \sum_z P(y \mid x, z) P(z) = P_x(y)$$

Double chance of being correct! Augmented IPW (IPW + OR)



$$P_x(y) = \sum_z P(y \mid x, z) P(z)$$

- Consider the following estimand: $g(P) \equiv \mathbb{E}_P \left[\frac{I_x(X)}{P(X \mid Z)} (I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z) \right]$

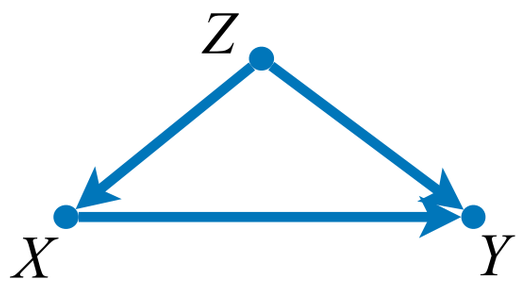
If $P(y \mid X, Z)$ is misspecified to $\tilde{P}(y \mid X, Z)$:

Takeaways:

- $g(P) = P_x(y)$ even if $P(y \mid X, Z)$ is misspecified to $\tilde{P}(y, \mid X, Z)$
- $g(P) = P_x(y)$ even if $P(X \mid Z)$ is misspecified to $\tilde{P}(X \mid Z)$

$\Longrightarrow$ **Doubly robustness:** An estimand $g(P)$ gives a double chance of being correct!

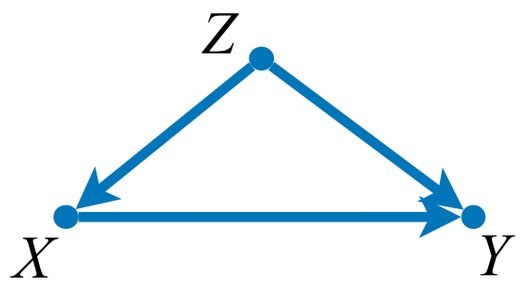
Classic BD estimator: AIPW



$$P_x(y) = \sum_z P(y \mid x, z) P(z)$$

Estimand ($g(P)$)	$\mathbb{E}_P \left[\frac{I_x(X)}{P(X \mid Z)} (I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z) \right]$
Estimator (T_N)	$\mathbb{E}_D \left[\frac{I_x(X)}{\hat{P}(X \mid Z)} (I_y(Y) - \hat{P}(y \mid X, Z)) + \hat{P}(y \mid x, Z) \right]$
For correct estimation	$\hat{P}(x \mid z) \to P(x \mid z)$; *Or* $\hat{P}(y \mid x, z) \to P(y \mid x, z)$.
For fast ($N^{-1/2}$ rate) convergence	

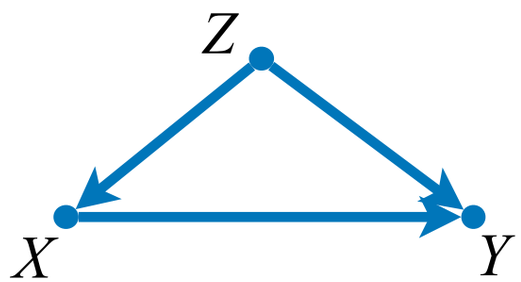
Classic BD estimator: AIPW



$$P_x(y) = \sum_z P(y \mid x, z) P(z)$$

Estimand ($g(P)$)	$\mathbb{E}_P \left[\frac{I_x(X)}{P(X \mid Z)} (I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z) \right]$
Estimator (T_N)	$\mathbb{E}_D \left[\frac{I_x(X)}{\hat{P}(X \mid Z)} (I_y(Y) - \hat{P}(y \mid X, Z)) + \hat{P}(y \mid x, Z) \right]$
For correct estimation	😄 "*Doubly robustness*" — Double chance of being correct!
For fast ($N^{-1/2}$ rate) convergence	😄 "*Debiasedness*" $\{\hat{P}(x \mid z), \hat{P}(y \mid x, z)\} \rightarrow \{P(x \mid z), P(y \mid x, z)\}$ can converge relatively slowly ($N^{-1/4}$ rate). $\{\hat{P}(x \mid z), P(x \mid z), \hat{P}(y \mid x, z), P(y \mid x, z)\}$ in Donsker class.

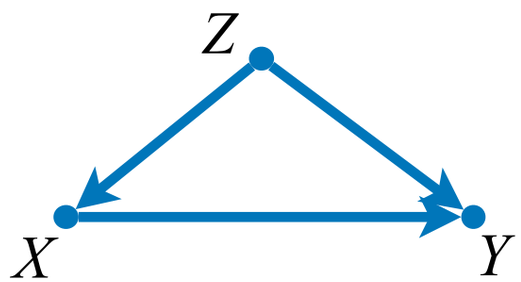
Classic BD estimator: AIPW



$$P_x(y) = \sum_z P(y \mid x, z) P(z)$$

Estimand ($g(P)$)	$\mathbb{E}_P \left[\frac{I_x(X)}{P(X \mid Z)} (I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z) \right]$
Estimator (T_N)	$\mathbb{E}_D \left[\frac{I_x(X)}{\hat{P}(X \mid Z)} (I_y(Y) - \hat{P}(y \mid X, Z)) + \hat{P}(y \mid x, Z) \right]$
For correct estimation	😄 "*Doubly robustness*" — Double chance of being correct!
For fast ($N^{-1/2}$ rate) convergence	😄 "*Debiasedness*" $\{\hat{P}(x \mid z), \hat{P}(y \mid x, z)\} \to \{P(x \mid z), P(y \mid x, z)\}$ can converge relatively slowly ($N^{-1/4}$ rate). 😢 Unclear that modern ML methods are in Donsker.

Recent advance — Double/ Debiased Machine Learning[2]



$$P_x(y) = f_{\mathsf{bd}}(P) = \sum_z P(y \mid x, z) P(z)$$

Estimand ($g(P)$)	$\mathbb{E}_P \left[\frac{I_x(X)}{P(X \mid Z)} (I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z) \right]$
Estimator (T_N)	$\frac{1}{2} \sum_{i \in \{0,1\}} \mathbb{E}_{D_i} \left[\frac{I_x(X)}{\hat{P}_{1-i}(X \mid Z)} (I_y(Y) - \hat{P}_{1-i}(y \mid X, Z)) + \hat{P}_{1-i}(y \mid x, Z) \right]$
For correct estimation	$\hat{P}(x \mid z) \to P(x \mid z)$; *Or* $\hat{P}(y \mid x, z) \to P(y \mid x, z)$.
For fast ($N^{-1/2}$ rate) convergence	

Recent advance — Double/ine Learning

Sample-splitting (a.k.a. Cross-fitting, Cross-validation)

1. (Sample-splitting). Randomly split the sample $D = \{D_0, D_1\}$.
2. Using D_i, learn $\{\hat{P}_i(x \mid z), \hat{P}_i(y \mid x, z)\}$.
3. Evaluate h using samples in D_i with models $\{\hat{P}_{1-i}(x \mid z), \hat{P}_{1-i}(y \mid x, z)\}$trained through D_{1-i} (i.e., samples for evaluation / training are distinct)
4. Take an empirical expectation of each h over D_i (i.e., $\mathbb{E}_{P_{D_i}}$) and divide it half.

$$\text{bd}(P) = \sum_z P(y \mid x, z)P(z)$$

$$P(y \mid X, Z)) + P(y \mid x, Z)\Big]$$

Estimator (T_N)	$\frac{1}{2}\sum_{i\in\{0,1\}} \mathbb{E}_{D_i}\left[\frac{I_x(X)}{\hat{P}_{1-i}(X \mid Z)}(I_y(Y) - \hat{P}_{1-i}(y \mid X, Z)) + \hat{P}_{1-i}(y \mid x, Z)\right] \equiv h$
For correct estimation	😄 "*Doubly robustness*" — Double chance of being correct!
For fast ($N^{-1/2}$ rate) convergence	

Recent advance — Double/ine Learning

Sample-splitting (a.k.a. Cross-fitting, Cross-validation)

1. (Sample-splitting). Randomly split the sample $D = \{D_0, D_1\}$.
2. Using D_i, learn $\{\hat{P}_i(x \mid z), \hat{P}_i(y \mid x, z)\}$.
3. Evaluate h using samples in D_i with models $\{\hat{P}_{1-i}(x \mid z), \hat{P}_{1-i}(y \mid x, z)\}$trained through D_{1-i} (i.e., samples for evaluation / training are distinct)
4. Take an empirical expectation of each h over D_i (i.e., $\mathbb{E}_{P_{D_i}}$) and divide it half.

$$\text{bd}(P) = \sum_z P(y \mid x, z)P(z)$$

$$P(y \mid X, Z)) + P(y \mid x, Z)\Big]$$

Estimator (T_N)	$\frac{1}{2}\sum_{i\in\{0,1\}} \mathbb{E}_{D_i}\left[\frac{I_x(X)}{\hat{P}_{1-i}(X \mid Z)}(I_y(Y) - \hat{P}_{1-i}(y \mid X, Z)) + \hat{P}_{1-i}(y \mid x, Z)\right]$ $\equiv h$
For correct estimation	😄 "*Doubly robustness*" — Double chance of being correct!
For fast ($N^{-1/2}$ rate) convergence	😄 "*Debiasedness*" $\{\hat{P}(x \mid z), \hat{P}(y \mid x, z)\} \to \{P(x \mid z), P(y \mid x, z)\}$ can converge relatively slowly ($N^{-1/4}$ rate).

Recent advance — Double/ine Learning

Sample-splitting (a.k.a. Cross-fitting, Cross-validation)

1. (Sample-splitting). Randomly split the sample $D = \{D_0, D_1\}$.
2. Using D_i, learn $\{\hat{P}_i(x \mid z), \hat{P}_i(y \mid x, z)\}$.
3. Evaluate h using samples in D_i with models $\{\hat{P}_{1-i}(x \mid z), \hat{P}_{1-i}(y \mid x, z)\}$trained through D_{1-i} (i.e., samples for evaluation / training are distinct)
4. Take an empirical expectation of each h over D_i (i.e., $\mathbb{E}_{P_{D_i}}$) and divide it half.

$$\text{bd}(P) = \sum_z P(y \mid x, z)P(z)$$

$$P(y \mid X, Z)) + P(y \mid x, Z)\Big]$$

Estimator (T_N)

$$\frac{1}{2}\sum_{i\in\{0,1\}} \mathbb{E}_{D_i}\left[\frac{I_x(X)}{\hat{P}_{1-i}(X \mid Z)}(I_y(Y) - \hat{P}_{1-i}(y \mid X, Z)) + \hat{P}_{1-i}(y \mid x, Z)\right] \equiv h$$

For correct estimation

ct!

Functions not in Donsker class are prone to overfitting bias. Overfitting bias are mitigated by sample-splitting.

For fast ($N^{-1/2}$ rate) convergence

"*Debiasedness*" $\{\hat{P}(x \mid z), \hat{P}(y \mid x, z)\} \to \{P(x \mid z), P(y \mid x, z)\}$ can converge relatively slowly ($N^{-1/4}$ rate).

Recent advance — Double/ine Learning

Sample-splitting (a.k.a. Cross-fitting, Cross-validation)

1. (Sample-splitting). Randomly split the sample $D = \{D_0, D_1\}$.
2. Using D_i, learn $\{\hat{P}_i(x \mid z), \hat{P}_i(y \mid x, z)\}$.
3. Evaluate h using samples in D_i with models $\{\hat{P}_{1-i}(x \mid z), \hat{P}_{1-i}(y \mid x, z)\}$trained through D_{1-i} (i.e., samples for evaluation / training are distinct)
4. Take an empirical expectation of each h over D_i (i.e., $\mathbb{E}_{P_{D_i}}$) and divide it half.

$$\text{bd}(P) = \sum_z P(y \mid x, z)P(z)$$

$$P(y \mid X, Z)) + P(y \mid x, Z)\Big]$$

Estimator (T_N)

$$\frac{1}{2} \sum \mathbb{E}_{D_i} \left[\frac{I_x(X)}{\hat{P}_{1-i}(X \mid Z)} (I_y(Y) - \hat{P}_{1-i}(y \mid X, Z)) + \hat{P}_{1-i}(y \mid x, Z) \right] \equiv h$$

Key result:

Double/debiased machine learning (DML) estimator for BD enjoys doubly robustness and debiasedness without restrictions on the function class!

Formalizing the Problem - Notation

- Let $h(\mathbf{V}; \eta = \{\eta_a, \eta_b\})$ be a functional of parameters called '*nuisance*' $\eta = \{\eta_a, \eta_b\}$ *(possibly* $\eta_b = \varnothing$*)* such that $\mathbb{E}_P[h(\mathbf{V}; \eta)] = P_x(y)$.

IPW $$h(\mathbf{V}, \eta = \{\eta_a = \{P(x \mid z))\}, \eta_b = \varnothing\} = \frac{I_x(X)}{P(X \mid Z)} \cdot I_y(Y)$$

OR $$h(\mathbf{V}, \eta = \{\eta_a = \{P(y \mid x, z)\}, \eta_b = \varnothing\}) = P(y \mid x, Z)$$

AIPW/DML $$h(\mathbf{V}, \eta = \{\eta_a = \{P(y \mid x, z)\}, \eta_b = \{P(x \mid z)\}\}) = \frac{I_x(X)}{P(X \mid Z)}(I_y(Y) - P(y \mid X, Z)) + P(y \mid x, Z)$$

- Let $T_N(\hat{\eta}_a, \hat{\eta}_b)$ be an estimator constructed based on h: either (1) empirical average; or (2) sample-splitting.

(1) $T_N = \mathbb{E}_D[h(\mathbf{V}; \hat{\eta}_a, \hat{\eta}_b)]$. (2) $T_N = \frac{1}{2} \sum_{i \in \{0,1\}} \mathbb{E}_{D_i}\left[h(\mathbf{V}; \hat{\eta}_a^{1-i}, \hat{\eta}_b^{1-i})\right]$.

Desirable statistical properties

- Let $h(\mathbf{V}; \eta = \{\eta_a, \eta_b\})$ s.t. $\mathbb{E}_P[h(\mathbf{V}; \eta)] = P_x(y)$.
- Let $T_N(\hat{\eta}_a, \hat{\eta}_b) = \mathbb{E}_D[h(\mathbf{V}; \hat{\eta}_a, \hat{\eta}_b)]$ or $T_N = \frac{1}{2}\sum_{i\in\{0,1\}} \mathbb{E}_{D_i}\left[h(\mathbf{V}; \hat{\eta}_a^{1-i}, \hat{\eta}_b^{1-i})\right]$.

AIPW/DML $\quad h(\mathbf{V}, \eta = \{\eta_a = \{P(y\,|\,x,z)\}, \eta_b = \{P(x\,|\,z)\}\}) = \frac{I_x(X)}{P(X\,|\,Z)}(I_y(Y) - P(y\,|\,X,Z)) + P(y\,|\,x,Z)$

Desirable statistical properties

Doubly robust (DR): T_N converges to $P_x(y)$ if $\hat{\eta}_a \to \eta_a$ *or* $\hat{\eta}_b \to \eta_b$.

Double chance of being correct!

Debiasedness (DB): T_N converges fast ($\sqrt{N}$ rate) to $P_x(y)$ even when $\hat{\eta}_a \to \eta_a$ and $\hat{\eta}_b \to \eta_b$ relatively slow ($N^{-1/4}$ rate) without any complexity restriction on $\eta, \hat{\eta}$.

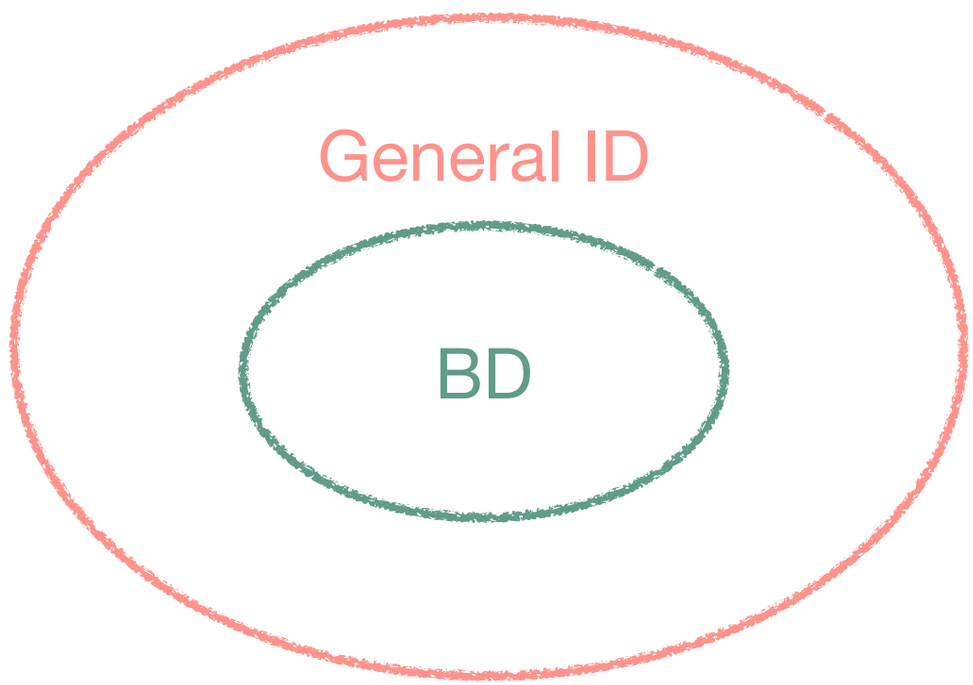
Double speed convergence!

Formalizing the Problem

Under the BD setting,

Formalizing the Problem

Under the general identifiable setting (i.e., general $f(P) = P_x(y)$),



Related works - Estimators for the BD case

	Statistical properties		Causal properties	
	Doubly Robustness	Debiasedness	Beyond BD	General ID
Inverse Probability Weighting (IPW)	✗	✗	✗	✗
Outcome Regression (OR)	✓	✗	✗	✗
AIPW[1]	✗	✗	✗	✗
DML for BD [2]	✓	✓	✗	✗

Related works - Estimators beyond the BD case

	Statistical properties		Causal properties	
	Doubly Robustness	Debiasedness	Beyond BD	General ID
Jung, Tian, Bareinboim (2020a)	✗	✗	✓	✗
Fulcher et al., (2019) Bhattacharya et al (2020)	✓	✗	✓	✗
Jung, Tian, Bareinboim (2020b)	✗	✗	✓	✓
This work (Jung, Tian, Bareinboim (2021))	✓	✓	✓	✓

Summary

- We reviewed the evolution of the BD estimators:

$$\{\text{IPW, OR}\} \rightarrow \text{AIPW} \rightarrow \text{DML}$$

- DML estimators achieve doubly-robustness (DR) and debiasedness (DB) without any restrictions on the complexity of functions.

- No estimators that achieve DR&DB and working for general ID functional exist.

- In this paper, we will construct such estimators.

Strategy

General approach

1. **(Moment condition)** We will choose $h(\mathbf{V};\eta)$ s.t. $\mathbb{E}_P[h(\mathbf{V};\eta)] = P_x(y)$.

e.g., DML $$h(\mathbf{V}, \eta = \{P(y|x,z), P(x|z)\}) = \frac{I_x(X)}{P(X|Z)}(I_y(Y) - P(y|X,Z)) + P(y|x,Z)$$

2. Then, we will construct T_N based on $h(\mathbf{V};\hat{\eta})$ where $\hat{\eta}$ is an estimate.

e.g., DML $$T_N = \frac{1}{2}\sum_{i\in\{0,1\}} \mathbb{E}_{P_{D_i}}\left[\frac{I_x(X)}{\hat{P}_{1-i}(X|Z)}(I_y(Y) - \hat{P}_{1-i}(y|X,Z)) + \hat{P}_{1-i}(y|x,Z)\right]$$

How do we choose $h(\mathbf{V};\eta)$ and T_N to attain DR & DB?

Useful result: If h is constructed using "*Neyman orthogonal score*", then T_N using sample-splitting achieves debiasedness[2]

Neyman Orthogonal Score[2]

Chernozhukov et al., (2018)[2] formalize the idea of developing estimators achieving debiasedness, based on *Neyman orthogonal score.*

Neyman Orthogonal Score: A function $q(\mathbf{V}; \eta, \psi)$, where (η, ψ) are (named 'nuisance', 'target') parameters (e.g., $\eta = \{P(y|x,z), P(x|z)\}, \psi = P_x(y)$), is *Neyman orthogonal score* if,

1. $\mathbb{E}_P[q(\mathbf{V}; \eta, \psi)] = 0$; and — Mean-zero

2. $\frac{\partial}{\partial \eta'} \mathbb{E}_P[q(\mathbf{V}; \eta', \psi)]\,|_{\eta'=\eta} = 0.$ — Invariant to the perturbation of the nuisance η

Double Machine Learning (DML) Estimator[2]

DML Estimator: Suppose an orthogonal score is decomposed as $q(\mathbf{V};\eta,P_x(y)) = h(\mathbf{V};\eta) - P_x(y)$ for some function h. In this setting, an estimator T_N is called *DML estimator* if it is constructed using *sample-splitting*:

1. Randomly split the sample $D = \{D_0, D_1\}$.
2. Estimate η using D_i, denoted $\hat{\eta}^i$ for $i \in \{0,1\}$.
3. $T_N = \frac{1}{2} \sum_{i\in\{0,1\}} \mathbb{E}_{D_i}\left[h(\mathbf{V};\hat{\eta}^{1-i})\right]$.

(Sample-splitting) — Evaluating h with distinct samples used in training $\hat{\eta}$ (just as cross-validation)

Debiasedness property of DML estimator:

- A DML estimator $T_N \to P_x(y)$ fast ($N^{-1/2}$-rate) even when $\hat{\eta} \to \eta$ slow ($N^{-1/4}$ rate), without any restriction on the function class (i.e., Donsker)
- No assumption on functional classes! $\Longrightarrow$ Any ML methods can be employed.

Double Machine Learning (DML) Estimator

DML Estimator: Suppose an orthogonal score is decomposed as $q(\mathbf{V};\eta,P_x(y)) = h(\mathbf{V};\eta) - P_x(y)$ for some function h. In this setting, an estimator T_N is called *DML estimator* if it is constructed using *sample-splitting*:

1. Randomly split the sample $D = \{D_0, D_1\}$.
2. Estimate η using D_i, denoted $\hat{\eta}^i$ for $i \in \{0,1\}$

So far:

- We want to find h s.t. $\mathbb{E}_P[h(\mathbf{V};\eta)] = P_x(y)$ to construct T_N.
- If $q(\mathbf{V};\eta,P_x(y)) = h(\mathbf{V};\eta) - P_x(y)$ is an orthogonal function, then an estimator T_N constructed using sample-splitting achieves *debiasedness*.

Example: Neyman Orthogonal Score

Neyman Orthogonal Score:

1. $\mathbb{E}_P[q(\mathbf{V};\eta,\psi)] = 0$; and 2. $\frac{\partial}{\partial \eta'}\mathbb{E}_P[q(\mathbf{V};\eta',\psi)]\,|_{\eta'=\eta} = 0.$

Example: AIPW/DML estimand for BD.

- Recall DML estimand $h(\mathbf{V},\eta=\{P(y\,|\,x,z),P(x\,|\,z)\}) = \frac{I_x(X)}{P(X\,|\,Z)}(I_y(Y)-P(y\,|\,X,Z))+P(y\,|\,x,Z)$.
- We showed $\mathbb{E}_P[h(\mathbf{V};\eta)] = P_x(y)$ in AIPW/DML part.
- $\mathbb{E}_P[h(\mathbf{V};\eta)] = P_x(y) \Longrightarrow \mathbb{E}_P[h(\mathbf{V};\eta)-P_x(y)] = 0$

$$\equiv q(\mathbf{V};\eta,P_x(y)) = \frac{I_x(X)}{P(X\,|\,Z)}(I_y(Y)-P(y\,|\,X,Z))+P(y\,|\,x,Z)-P_x(y)$$

Fact: $q(\mathbf{V};\eta,P_x(y))$ is a Neyman orthogonal score.

Example: Neyman Orthogonal Score

Neyman Orthogonal Score:

1. $\mathbb{E}_P[q(\mathbf{V};\eta,\psi)] = 0$; and
2. $\frac{\partial}{\partial \eta'}\mathbb{E}_P[q(\mathbf{V};\eta',\psi)]\,|_{\eta'=\eta} = 0.$

$$q(\mathbf{V};\eta,P_x(y)) = \frac{I_x(X)}{P(X|Z)}(I_y(Y) - P(y|X,Z)) + P(y|x,Z) - P_x(y)$$

Condition 1:

$$\mathbb{E}_P[q(\mathbf{V};\eta,\psi)] = \mathbb{E}_P\left[\frac{I_x(X)}{P(X|Z)}(I_y(Y) - P(y|X,Z)) + P(y|x,Z) - P_x(y)\right]$$

$$= \mathbb{E}_P\left[\frac{I_x(X)}{P(X|Z)}(P(y|X,Z) - P(y|X,Z)) + P(y|x,Z)\right] - P_x(y)$$

$= \mathbb{E}_P[h(\mathbf{V};\eta)] = P_x(y)$ since this part is an estimand of AIPW/DML

$$= 0$$

Example: Neyman Orthogonal Score

Neyman Orthogonal Score:

1. $\mathbb{E}_P[q(\mathbf{V};\eta,\psi)] = 0$; and
2. $\frac{\partial}{\partial \eta'}\mathbb{E}_P[q(\mathbf{V};\eta',\psi)]\,|_{\eta'=\eta} = 0.$

$$q(\mathbf{V};\eta,P_x(y)) = \frac{I_x(X)}{P(X|Z)}(I_y(Y) - P(y|X,Z)) + P(y|x,Z) - P_x(y)$$

Condition 2:

$$\frac{\partial}{\partial P'(y|x,z)}\mathbb{E}_P\left[\frac{I_x(X)}{P(X|Z)}(I_y(Y) - P'(y|X,Z)) + P'(y|x,Z) - P_x(y)\right]\Bigg|_{P'=P}$$

$$= \underbrace{\frac{\partial}{\partial P'(y|x,z)}\mathbb{E}_P\left[\frac{I_x(X)I_y(Y)}{P(X|Z)}\right]}_{=0} - \underbrace{\frac{\partial}{\partial P'(y|x,z)}\mathbb{E}_P\left[\frac{I_x(X)P'(y|X,Z)}{P(X|Z)}\right]}_{=\mathbb{E}_P\left[\frac{I_x(X)}{P(X|Z)}\right]} + \underbrace{\frac{\partial}{\partial P'(y|x,z)}\mathbb{E}_P\left[P'(y|x,Z)\right]}_{=1} - \underbrace{\frac{\partial}{\partial P'(y|x,z)}\mathbb{E}_P\left[P_x(y)\right]}_{=0}$$

$$= -\underbrace{\mathbb{E}_P\left[\frac{I_x(X)}{P(X|Z)}\right]}_{} + 1 \quad = -1 + 1 = 0$$

$$= \sum_{x',z}\frac{I_x(x')}{P(x'|z)}P(x'|z)P(z) = \sum_{x',z} I_x(x')P(z) = \sum_{z} P(z) = 1$$

Example: Neyman Orthogonal Score

Neyman Orthogonal Score:

1. $\mathbb{E}_P[q(\mathbf{V};\eta,\psi)] = 0$; and
2. $\dfrac{\partial}{\partial\eta'}\mathbb{E}_P[q(\mathbf{V};\eta',\psi)]\,|_{\eta'=\eta} = 0.$

$$q(\mathbf{V};\eta,P_x(y)) = \frac{I_x(X)}{P(X|Z)}(I_y(Y) - P(y|X,Z)) + P(y|x,Z) - P_x(y)$$

Condition 2:

$$\frac{\partial}{\partial P'(x|z)}\mathbb{E}_P\left[\frac{I_x(X)}{P'(X|Z)}(I_y(Y) - P(y|X,Z)) + P(y|x,Z) - P_x(y)\right]\Bigg|_{P'=P}$$

$$= \frac{\partial}{\partial P'(x|z)}\mathbb{E}_P\left[\frac{I_x(X)}{P'(X|Z)}(I_y(Y) - P(y|X,Z))\right] + \underbrace{\frac{\partial}{\partial P'(x|z)}\mathbb{E}_P\left[P(y|x,Z)\right]}_{=0} - \underbrace{\frac{\partial}{\partial P(x|z)}\mathbb{E}_P\left[P_x(y)\right]}_{=0}$$

$$= \mathbb{E}_{P(X,Z)}\left\{\mathbb{E}_{P(Y|X,Z)}\left[\frac{I_x(X)}{P'(X|Z)}(I_y(Y) - P(y|X,Z)\middle|X,Z\right]\right\} = \mathbb{E}_{P(X,Z)}\left\{\frac{I_x(X)}{P'(X|Z)}(P(y|X,Z) - P(y|X,Z)\right\} = 0$$

$$= 0.$$

Example: Neyman Orthogonal Score

Neyman Orthogonal Score:

1. $\mathbb{E}_P[q(\mathbf{V};\eta,\psi)] = 0$; and
2. $\frac{\partial}{\partial\eta'}\mathbb{E}_P[q(\mathbf{V};\eta',\psi)]\,|_{\eta'=\eta} = 0.$

So far:

- For a AIPW/DML estimand $h(\mathbf{V};\eta)$, $q(\mathbf{V};\eta,P_x(y)) = h(\mathbf{V};\eta) - P_x(y)$ is an orthogonal score.
- Then, T_N constructed using sample-splitting, is a DML estimator achieving *debiasedness* ($\eta = \{P(x\,|\,z), P(y\,|\,x,z)\}$ converges slowly but T_N converges fast).

How can we find such orthogonal score $q(\mathbf{V};\eta,P_x(y)) = h(\mathbf{V};\eta) - P_x(y)$ for the setting beyond the BD?

$= 0.$

Finding Neyman Orthogonal Score via Influence Function

Neyman Orthogonal Score:

1. $\mathbb{E}_P[q(\mathbf{V};\eta,\psi)] = 0$; and

Mean-zero

2. $\frac{\partial}{\partial \eta'}\mathbb{E}_P[q(\mathbf{V};\eta',\psi)]\,|_{\eta'=\eta} = 0.$

Invariant to the small perturbation

We start from formalizing "perturbation" in Condition 2.

How can we find such orthogonal score $q(\mathbf{V};\eta,P_x(y)) = h(\mathbf{V};\eta) - P_x(y)$ for the setting beyond the BD?

- Perturbed distribution: For P and any mean-zero noise function $\alpha(\mathbf{V})$, let $P_\gamma = P(1+\gamma\alpha(\mathbf{V}))$ (a distribution perturbed along the direction $\gamma\alpha(\mathbf{V})$) be the perturbed distribution. (Formally called "parametric submodel")

- Sensitivity of using perturbed distribution: $f(P_\gamma)$ is a causal functional using P_γ. Then, $\frac{\partial}{\partial\gamma} f(P_\gamma)\,|_{\gamma=0}$ is a sensitivity of using the perturbed distribution.

Uncentered Influence Function (UIF)

How can we find such orthogonal score $q(\mathbf{V};\eta,P_x(y)) = h(\mathbf{V};\eta) - P_x(y)$ for the setting beyond the BD?

- Sensitivity: $\frac{\partial}{\partial\gamma} f(P_\gamma)\,|_{\gamma=0}$ is a sensitivity of using the perturbed distribution.

Influence Function (IF) — Formally capturing 'sensitivity': A mean-zero function $q(\mathbf{V};\eta,P_x(y))$ is called *influence function* if it satisfies

$$\frac{\partial}{\partial\gamma} f(P_\gamma)\,|_{\gamma=0} = \mathbb{E}_P\left[q(\mathbf{V};\eta,P_x(y)) \cdot \frac{\partial}{\partial\gamma}\log P_\gamma(\mathbf{V})\,|_{\gamma=0}\right].$$

Uncentered IF (UIF): If $q(\mathbf{V};\eta,P_x(y)) = h(\mathbf{V};\eta) - P_x(y)$ for some function h, then such $h(\mathbf{V};\eta)$ is called *uncentered IF*.

Uncentered Influence Function (UIF)

How can we find such orthogonal score $q(\mathbf{V};\eta,P_x(y)) = h(\mathbf{V};\eta) - P_x(y)$ for the setting beyond the BD?

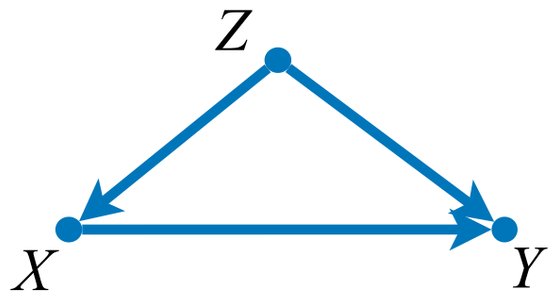
- Sensitivity: $\frac{\partial}{\partial\gamma} f(P_\gamma)\,|_{\gamma=0}$ is a sensitivity of using the perturbed distribution.

So far,

- Orthogonal score based estimator achieves debiasedness.
- Orthogonal score is a function invariant to the local perturbation.
- Influence function (IF) formally captures the local perturbation.

How can we use IF to derive the orthogonal score? / What is a relationship b/w IF and an orthogonal score?

Derivation of IF/UIF for the BD adjustment (example)



$$P_x(y) = f(P) = \sum_z P(y \mid x, z) P(z)$$

Derivative of $P(a \mid b)$

$$\frac{\partial}{\partial \gamma} P_\gamma(a \mid b)\,|_{\gamma=0} = \mathbb{E}_P \left[\left(\frac{I_b(B)}{P(B)} (I_a(A) - P(a \mid b)) \right) \cdot \frac{\partial}{\partial \gamma} P_\gamma(a, b)\,|_{\gamma=0} \right].$$

$$f(P_\gamma) = \sum\nolimits_z P_\gamma(y \mid x, z) P_\gamma(z)$$

$$\frac{\partial}{\partial \gamma} f(P_\gamma)\,|_{\gamma=0} = \sum_z \left(\frac{\partial}{\partial \gamma} P_\gamma(y \mid x, z)\,|_{\gamma=0} \right) P(z) + \sum_z P(y \mid x, z) \left(\frac{\partial}{\partial \gamma} P_\gamma(z)\,|_{\gamma=0} \right)$$

Derivative for $P(y \mid x, z)$

$$= \sum_z \mathbb{E}_P \left[\left(\frac{I_{x,z}(X, Z)}{P(X, Z)} (I_y(Y) - P(y \mid x, z)) \right) \cdot \frac{\partial}{\partial \gamma} P_\gamma(\mathbf{V})\,|_{\gamma=0} \right] P(z)$$

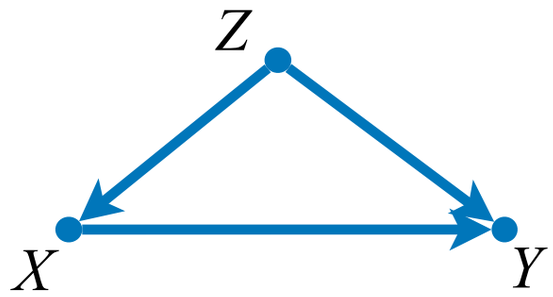
Derivative for $P(z)$

$$+ \sum_z P(y \mid x, z) \mathbb{E}_P \left[\left((I_z(Z) - P(z)) \right) \cdot \frac{\partial}{\partial \gamma} P_\gamma(\mathbf{V})\,|_{\gamma=0} \right]$$

IF $q(\mathbf{V}; \eta, P_x(y))$

$$= \mathbb{E}_P \left[\left(\frac{I_x(X)}{P(X \mid Z)} (I_y(Y) - P(y \mid x, Z)) + P(y \mid x, Z) - P_x(y) \right) \cdot \frac{\partial}{\partial \gamma} P_\gamma(\mathbf{V})\,|_{\gamma=0} \right]$$

Derivation of IF/UIF for the BD adjustment (example)



$$P_x(y) = f(P) = \sum_z P(y \mid x, z) P(z)$$

Derivative of $P(a \mid b)$

$$\frac{\partial}{\partial \gamma} P_\gamma(a \mid b)\,|_{\gamma=0} = \mathbb{E}_P \left[\left(\frac{I_b(B)}{P(B)} (I_a(A) - P(a \mid b)) \right) \cdot \frac{\partial}{\partial \gamma} P_\gamma(a, b)\,|_{\gamma=0} \right].$$

$$f(P_\gamma) = \sum\nolimits_z P_\gamma(y \mid x, z) P_\gamma(z)$$

$$\frac{\partial}{\partial \gamma} f(P_\gamma)\,|_{\gamma=0} = \sum_z \left(\frac{\partial}{\partial \gamma} P_\gamma(y \mid x, z)\,|_{\gamma=0} \right) P(z) + \sum_z P(y \mid x, z) \left(\frac{\partial}{\partial \gamma} P_\gamma(z)\,|_{\gamma=0} \right)$$

Derivative for $P(y \mid x, z)$

$$= \sum_z \mathbb{E}_P \left[\left(\frac{I_{x,z}(X, Z)}{P(X, Z)} (I_y(Y) - P(y \mid x, z)) \right) \cdot \frac{\partial}{\partial \gamma} P_\gamma(\mathbf{V})\,|_{\gamma=0} \right] P(z)$$

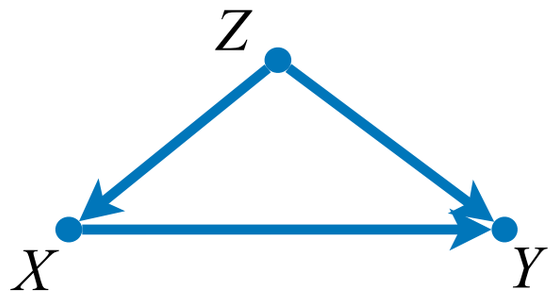
Derivative for $P(z)$

$$+ \sum_z P(y \mid x, z) \mathbb{E}_P \left[\left((I_z(Z) - P(z)) \right) \cdot \frac{\partial}{\partial \gamma} P_\gamma(\mathbf{V})\,|_{\gamma=0} \right]$$

UIF $h(\mathbf{V}; \eta)$

$$= \mathbb{E}_P \left[\left(\frac{I_x(X)}{P(X \mid Z)} (I_y(Y) - P(y \mid x, Z)) + P(y \mid x, Z) - P_x(y) \right) \cdot \frac{\partial}{\partial \gamma} P_\gamma(\mathbf{V})\,|_{\gamma=0} \right]$$

Derivation of IF/UIF for the BD adjustment (example)



$$P_x(y) = f(P) = \sum_z P(y \mid x, z) P(z)$$

Implication

- We showed that AIPW/DML estimand h is an orthogonal score.
- We just showed that the h is also a UIF!

What is a relationship b/w IF and an orthogonal score?

Orthogonal score & IF

What is the relationship b/w an IF and orthogonal score?

Neyman Orthogonal Score:

1. $\mathbb{E}_P[q(\mathbf{V};\eta,\psi)] = 0$; and
2. $\frac{\partial}{\partial \eta'}\mathbb{E}_P[q(\mathbf{V};\eta',\psi)]\,|_{\eta'=\eta} = 0.$

Influence Function (IF): A mean-zero function $q(\mathbf{V};\eta,P_x(y))$ is called *influence function* if it satisfies

$$\frac{\partial}{\partial \gamma} f(P_\gamma)\,|_{\gamma=0} = \mathbb{E}_P\left[q(\mathbf{V};\eta,P_x(y)) \cdot \frac{\partial}{\partial \gamma}\log P_\gamma(\mathbf{V})\,|_{\gamma=0}\right].$$

IF is an orthogonal function: An IF $q(\mathbf{V};\eta,P_x(y))$ is an orthogonal score[3]

Summary: Orthogonal score & IF

Neyman Orthogonal Score:

1. $\mathbb{E}_P[q(\mathbf{V};\eta,\psi)] = 0$; and 2. $\frac{\partial}{\partial\eta'}\mathbb{E}_P[q(\mathbf{V};\eta',\psi)]\,|_{\eta'=\eta} = 0.$

DML estimator (Under the setting $q(\mathbf{V};\eta,\psi) = h(\mathbf{V};\eta) - P_x(y)$)

$T_N = (1/2)\sum_{i\in\{0,1\}} \mathbb{E}_{D_i}\left[h(\mathbf{V};\hat{\eta}^{1-i})\right]$ constructed using sample-splitting.

Debiasedness: $T_N \to P_x(y)$ fast ($N^{-1/2}$-rate) even when $\hat{\eta} \to \eta$ slow ($N^{-1/4}$ rate).

Influence Function (IF): $\frac{\partial}{\partial\gamma} f(P_\gamma)\,|_{\gamma=0} = \mathbb{E}_P\left[q(\mathbf{V};\eta,P_x(y)) \cdot \frac{\partial}{\partial\gamma}\log P_\gamma(\mathbf{V})\,|_{\gamma=0}\right].$

Uncentered IF (UIF): $q(\mathbf{V};\eta,P_x(y)) = h(\mathbf{V};\eta) - P_x(y)$

IF is an orthogonal function $\Longrightarrow$ UIF based T_N is a DML estimator.

Summary: Orthogonal score & IF

Neyman Orthogonal Score:

1. $\mathbb{E}_P[q(\mathbf{V};\eta,\psi)] = 0$; and 2. $\frac{\partial}{\partial\eta'}\mathbb{E}_P[q(\mathbf{V};\eta',\psi)]\,|_{\eta'=\eta} = 0.$

DML estimator (Under the setting $q(\mathbf{V};\eta,\psi) = h(\mathbf{V};\eta) - P_x(y)$)

$T_N = (1/2)\sum_{i\in\{0,1\}} \mathbb{E}_{D_i}\left[h(\mathbf{V};\hat{\eta}^{1-i})\right]$ constructed using sample-splitting.

So far:

- We want to find h s.t. $\mathbb{E}_P[h(\mathbf{V};\eta)] = P_x(y)$ to construct T_N.
- If $q(\mathbf{V};\eta,P_x(y)) = h(\mathbf{V};\eta) - P_x(y)$ is an orthogonal function, then an estimator T_N constructed using sample-splitting achieves *debiasedness*.
- We can find such h by finding an IF.

Refined strategy

1. We will derive a UIF $h(\mathbf{V};\eta)$ for $f(P) = P_x(y)$.

- IF: A function $q(\mathbf{V};\eta, P_x(y))$ satisfies the following:

$$\frac{\partial}{\partial\gamma} f(P_\gamma)\,|_{\gamma=0} = \mathbb{E}_P\left[q(\mathbf{V};\eta, P_x(y)) \cdot \frac{\partial}{\partial\gamma}\log P_\gamma(\mathbf{V})_{\gamma=0}\right]$$

- UIF: If an IF $q(\mathbf{V};\eta, P_x(y)) \equiv h(\mathbf{V};\eta) - P_x(y)$, h is a UIF.

2. Then, we will construct a DML estimator T_N based on $h(\mathbf{V};\eta)$ and sample splitting.

$$T_N = \frac{1}{2}\sum_{i\in\{0,1\}} \mathbb{E}_{P_{D_i}}\left[h(\mathbf{V};\hat{\eta}^{1-i})\right].$$

Beyond-BD

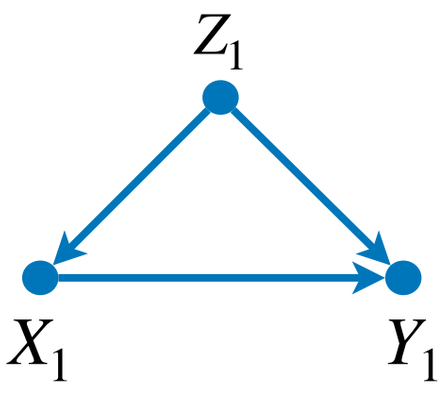
Easy case — Multi-outcome sequential BD (mSBD)

Multi-outcome sequential BD (mSBD)[4]

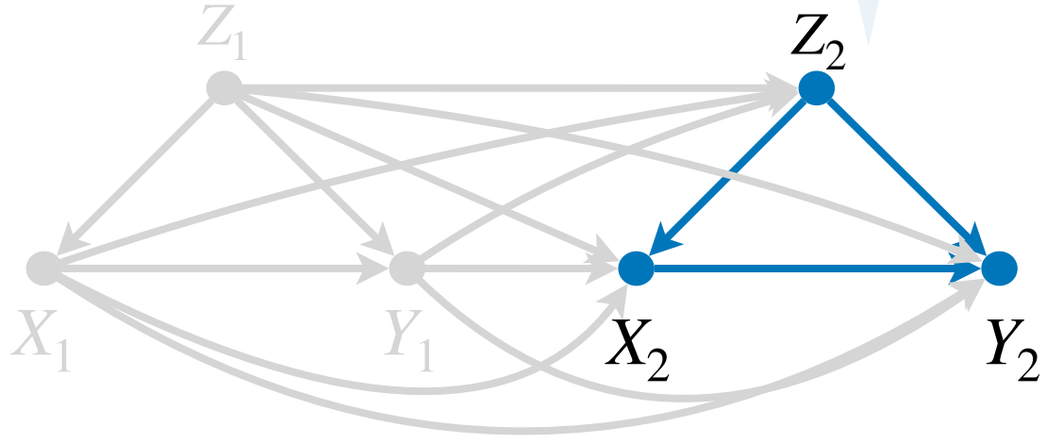
(informal) Multi-outcome sequential BD (mSBD): An extension of BD, where, at ith round, Z_i satisfies the BD criterion relative to $\{X_i, Y_i\}$ conditioned on previous variables $\{(X_j, Y_j, Z_j)\}_{j=1}^{i-1}$.

- This implies that there are no unmeasured co
  $\mathbf{X} = \{X_1, \cdots, X_n\}$ and $\mathbf{Y} = \{Y_1, \cdots, Y_n\}$.

Z_2 satisfies BD criterion relative to (X_2, Y_2) conditioned on $\{Z_1, X_1, Y_1\}$.



$i = 1$



$i = 2$

Multi-outcome sequential BD (mSBD)

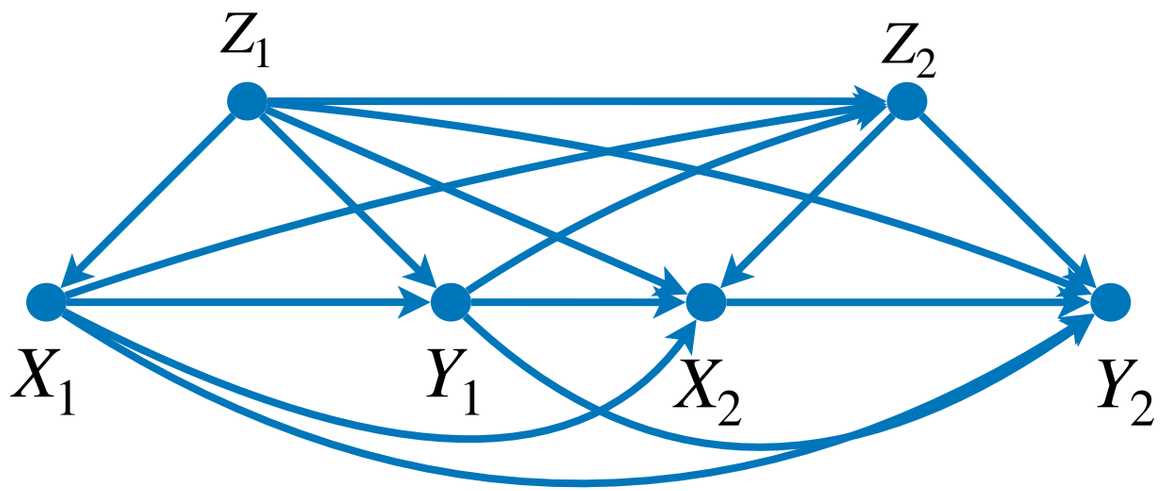
(informal) Multi-outcome sequential BD (mSBD): An extension of BD, where, at ith round, Z_i satisfies the BD criterion relative to $\{X_i, Y_i\}$ conditioned on previous variables $\{(X_j, Y_j, Z_j)\}_{j=1}^{i-1}$.

mSBD adjustment: If $\mathbf{Z} = \{Z_1, \cdots, Z_n\}$ satisfies the mSBD criterion relative to $(\mathbf{X}, \mathbf{Y})$,

$\mathbf{X}^{(i)} = \{X_1, \cdots, X_i\}$

$$P(\mathbf{y} \mid do(\mathbf{x})) = \sum_{\mathbf{z}} \prod_{Y_i \in \mathbf{Y}} P(y_i \mid \mathbf{x}^{(i)}, \mathbf{z}^{(i)}, \mathbf{y}^{(i-1)}) \prod_{Z_i \in \mathbf{Z}} P(z_i \mid \mathbf{x}^{(i-1)}, \mathbf{z}^{(i-1)}, \mathbf{y}^{(i-1)})$$

mSBD - example

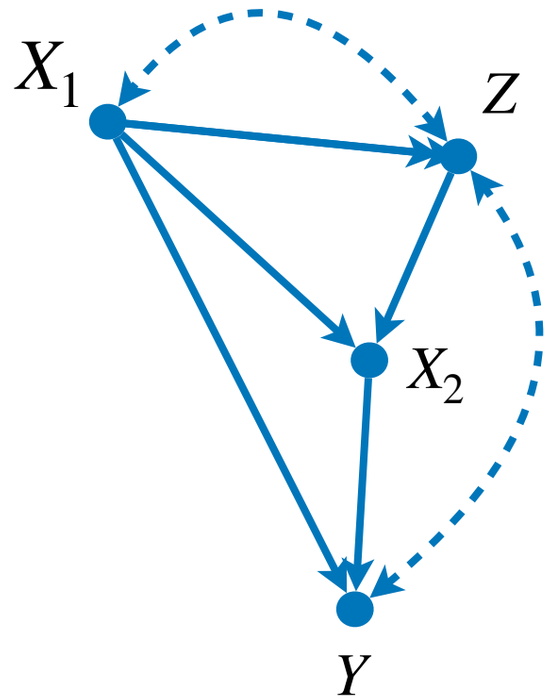


$\{Z_1, Z_2\}$ satisfies mSBD criterion relative to $(\{X_1, X_2\}, \{Y_1, Y_2\})$.

$$P_{x_1,x_2}(y_1, y_2) = \sum_{z_1,z_2} P(z_1)P(y_1 \mid x_1, z_1)P(z_2 \mid z_1, x_1, y_1)P(y_2 \mid z_1, x_1, y_1, z_2, x_2)$$

mSBD - example 2

mSBD permits a variable ∅.



$Z = \{\varnothing, Z\}$ satisfies mSBD criterion relative to $(\{X_1, X_2\}, \{\varnothing, Y\}) = (\{X_1, X_2\}, Y)$.

$$P_{x_1,x_2}(y) = \sum_z P(y \mid x_1, x_2, z) P(z \mid x_1)$$

IF/UIF for mSBD

1. UIF for mSBD

If $f(P) = P_x(y)$ is given as mSBD adjustment, a UIF is given as

$$h(\mathbf{V}; \eta = \{\mathbf{H}, \mathbf{W}\}) = H_2 + \sum_{i=2}^{n+1} W_i(H_{i+1} - H_1), \text{ where}$$

$$H_i = P_{\mathbf{x}}(\mathbf{y}^{\geq i-1} \,|\, \mathbf{Z}^{(i-1)}, \mathbf{y}^{(i-2)}) I_{y^{(i-2)}}(Y^{(i-2)})$$

$$= I_{y^{(i-2)}}(Y^{(i-2)}) \sum\nolimits_{z^{\geq i+1}} \prod_{k=i-1}^{n} P(y_k \,|\, x^{(k)}, y^{(k-1)}, z^{(k)}) P(z_k \,|\, x^{(k-1)}, y^{(k-1)}, z^{(k-1)})$$

$$W_i = \prod_{p=1}^{i} \frac{I_{x_p}(X_p)}{P(x_p \,|\, \mathbf{Z}^{(p)}, \mathbf{x}^{(p-1)}, \mathbf{y}^{(p-1)})}.$$

IF/UIF for mSBD

2. Mean of UIF

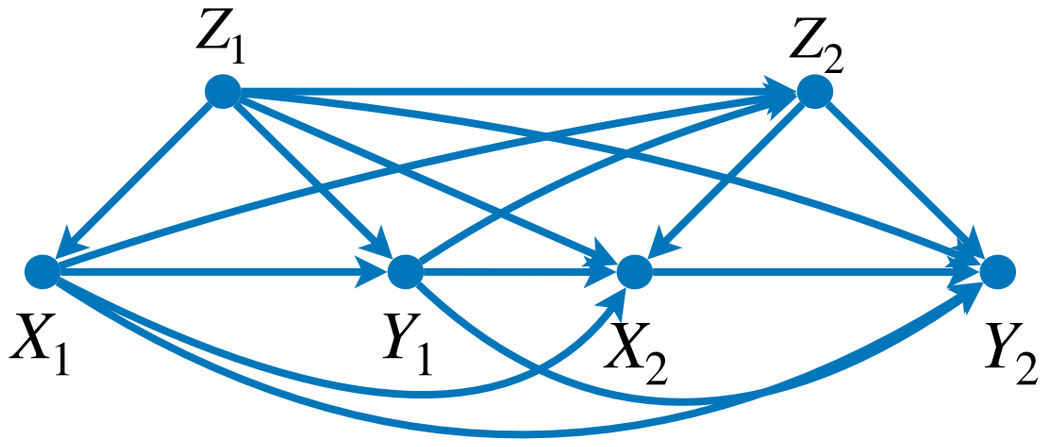
Note $\mathbb{E}_P[h(\mathbf{V};\eta)] = f(P) = P_x(y)$, by definition of the UIF.

3. IF for mSBD

An IF is given as $q(\mathbf{V};\eta, P_x(y)) = h(\mathbf{V};\eta) - \mathbb{E}_P[h(\mathbf{V};\eta)]$

Since IF is defined as $q(\mathbf{V};\eta, P_x(y)) = h(\mathbf{V};\eta) - P_x(y)$ and $\mathbb{E}_P[h(\mathbf{V};\eta)] = P_x(y)$

UIF for mSBD - Examples

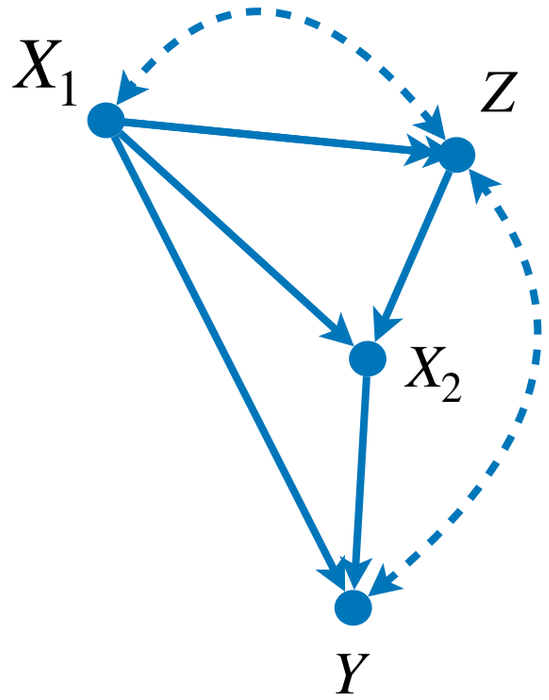


$$P_{x_1,x_2}(y_1,y_2)=\sum_{z_1,z_2}P(z_1)P(y_1\,|\,x_1,z_1)P(z_2\,|\,z_1,x_1,y_1)P(y_2\,|\,z_1,x_1,y_1,z_2,x_2)$$

$$h=H_2+\frac{I_{x_1}(X_1)}{P(X_1\,|\,Z_1)}(H_3-H_2)+\frac{I_{x_1,x_2}(X_1,X_2)}{P(X_1\,|\,Z_1)P(X_2\,|\,Z_1,X_1,Y_1,Z_1)}(H_4-H_3),\ \text{where}$$

$$H_2=P_{x_1,x_2}(y_1,y_2\,|\,Z_1),\ H_3=P_{x_1,x_2}(y_1,y_2\,|\,y_1,Z_1,Z_2)I_{y_1}(Y_1),\ H_4=I_{y_1,y_2}(Y_1,Y_2).$$

UIF for mSBD - Examples 2



$$P_{x_1,x_2}(y) = \sum_z P(y \mid x_1, x_2, z) P(z \mid x_1)$$

$$h = H_2 + \frac{I_{x_1}(X_1)}{P(X_1)}(H_3 - H_2) + \frac{I_{x_1,x_2}(X_1, X_2)}{P(X_1)P(X_2 \mid Z, X_1)}(H_4 - H_3), \text{ where}$$

$$H_2 = P_{x_1,x_2}(y),\ H_3 = P_{x_1,x_2}(y \mid Z),\ H_4 = I_{y_1,y_2}(Y_1, Y_2).$$

UIF for mSBD - Properties

Recall: UIF for mSBD is given as

$$h(\mathbf{V};\eta=\{\mathbf{H},\mathbf{W}\})=H_2+\sum_{i=2}^{n+1}W_i(H_{i+1}-H_1)$$

Properties of DML estimator T_N based on h

Let $\eta_a=\{\mathbf{H}\}$, $\eta_b=\{\mathbf{W}\}$, $T_N=\frac{1}{2}\sum_{i\in\{0,1\}}\mathbb{E}_{P_{D_i}}\left[h(\mathbf{V};\hat{\eta}_a^{1-i},\hat{\eta}_b^{1-i})\right]$.

Doubly robust (DR): T_N converges to $P_x(y)$ if $\hat{\eta}_a\to\eta_a$ or $\hat{\eta}_b\to\eta_b$.

Debiasedness (DB): T_N converges fast ($\sqrt{N}$ rate) to $P_x(y)$ even when $\hat{\eta}_a\to\eta_a$ and $\hat{\eta}_b\to\eta_b$ relatively slow ($N^{-1/4}$ rate).

UIF for mSBD - Properties

Recall: UIF for mSBD is given as

$$h(\mathbf{V}; \eta = \{\mathbf{H}, \mathbf{W}\}) = H_2 + \sum_{i=2}^{n+1} W_i(H_{i+1} - H_1)$$

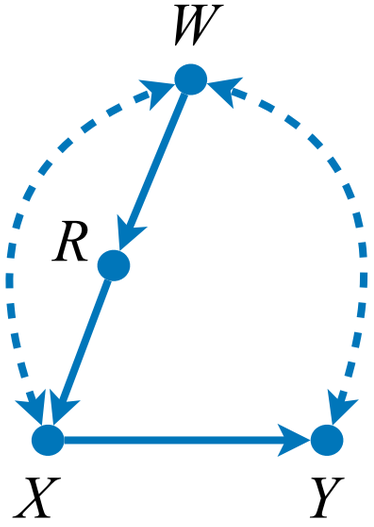
So far,

- mSBD (a sequence of BD) gives a graphical criterion for the case that there are no unmeasured confounders b/w $\mathbf{X}$ and $\mathbf{Y}$.
- We give mSBD adjustment (will call $M[\mathbf{y} \mid \mathbf{x}; \mathbf{z}]$) and its IF/UIF.
- Based on the UIF, we give a DML estimator for $M[\mathbf{y} \mid \mathbf{x}; \mathbf{z}]$.

General case — Causal functional represented as a function of mSBDs

Examples for general case

mSBD operator $M = M[\mathbf{y} \mid \mathbf{x}; \mathbf{z}]$: mSBD adjustment for the case where $\mathbf{Z}$ satisfies mSBD adjustment w.r.t. $(\mathbf{X}, \mathbf{Y})$.



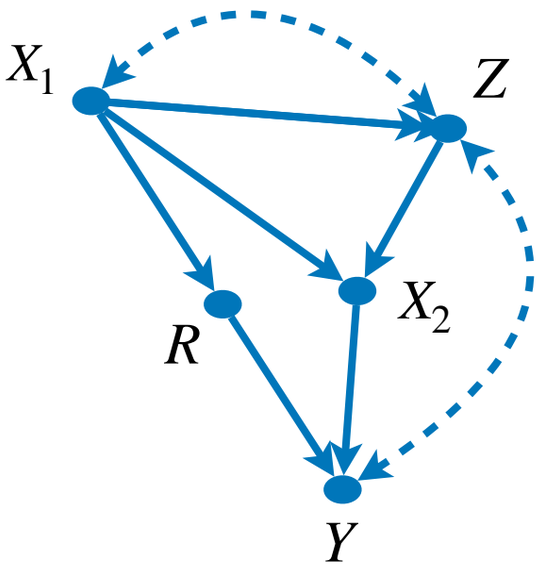
$$P_x(y) = \frac{\sum_w P(y, x \mid r, w) P(w)}{\sum_w P(x \mid r, w) P(w)} = \frac{M_2}{M_1}$$

W is BD admissible w.r.t. $(R, \{X, Y\})$; $M_2 \equiv M[(x, y) \mid r; w]$

W is BD admissible w.r.t. (R, X) $M_1 \equiv M[x \mid r; w]$

Examples for general case

mSBD operator $M[\mathbf{y} \mid \mathbf{x}; \mathbf{z}]$: mSBD adjustment when $\mathbf{Z}$ satisfies mSBD adjustment w.r.t. $(\mathbf{X}, \mathbf{Y})$.



$\varnothing$ is BD admissible w.r.t. (X_1, R);
$M_1 \equiv M[r \mid x_1; \varnothing]$

$\{Z, X_1\}$ is BD admissible w.r.t. (X_2, Y);
$M_2 \equiv M[y \mid x_2; (z, x_1)]$

$$P_x(y) = \sum_r P(r \mid x_1) \sum_{x_1', z} P(y \mid r, x_1', x_2, z) P(z, x_1')$$

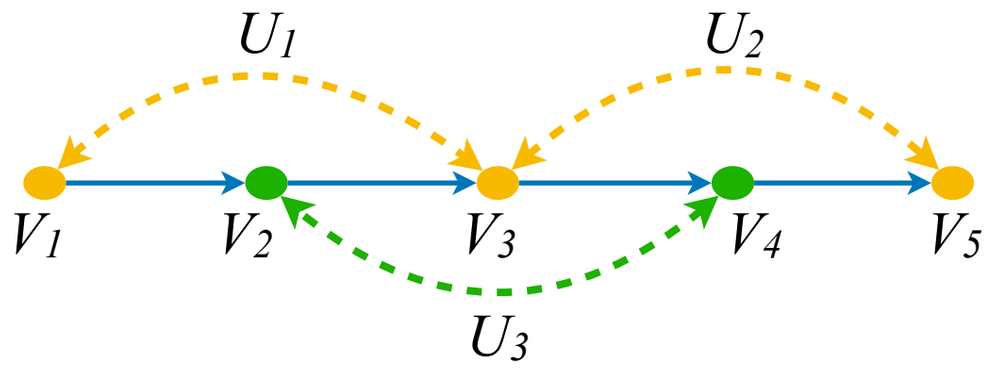
$$= \sum_r M_1 \cdot M_2$$

Can any identifiable functional be encoded as a function of mSBD?

A little detour

Review (CI 1, Lecture 7) — Algorithmic approach to Identifiability

C-factors



- **C-component**: A set of variables connected by bi-directed edges (e.g., $\{V_1, V_3, V_5\}$ and $\{V_2, V_4\}$.

- **C-factor**: $Q[\mathbf{C}](\mathbf{c}, pa_{\mathbf{c}}) = \sum_{u(\mathbf{C})} P(u(\mathbf{C})) \prod_{V_i \in \mathbf{C}} P(v_i \mid pa_i, u_i)$ where $U(\mathbf{C}) = \bigcup_{V_i \in \mathbf{C}} U_i$

- A distribution can be factorized w.r.t. C-factors.

$$P(\mathbf{v}) = \left(\sum_{u_3} P(u_3) P(v_2 \mid v_1, u_3) P(v_4 \mid v_3, u_3) \right) \left(\sum_{u_1, u_2} P(u_1, u_2) P(v_1 \mid u_1) P(v_3 \mid v_2, u_1, u_2) P(v_5 \mid v_4, u_2) \right)$$

$$= Q[V_2, V_4](v_2, v_4, v_1, v_3) Q[V_1, V_3, V_5](v_1, v_3, v_5, v_2, v_4)$$

C-factor Algebra - Summary

We have two basic operations over c-factors

1. Reduce to an ancestral set

$$Q[\mathbf{W}] = \sum_{\mathbf{c}\setminus\mathbf{w}} Q[\mathbf{C}]$$

If $\boldsymbol{W}$ is ancestral in $G(\mathbf{C})$

2. Factorize into c-components

$$Q[\mathbf{H}] = \prod_{j} Q[H_j]$$

Where $\mathbf{H}_1, \ldots, \mathbf{H}_k$, are the c-components in $\mathcal{G}[\mathbf{H}]$

3. Identification of c-factor:

$$Q[\mathbf{C}] = \prod_{V_i \in \mathbf{C}} P(v_i \mid v^{(i-1)})$$

where $\mathbf{C}$ is a C-component in G

Recap: Tian's ID algorithm

ID$(\mathbf{X}, \mathbf{Y}, G)$

1. Let $\mathbf{S}_1, \mathbf{S}_2, \ldots$ be the C-components of G.
2. Let $Q[\mathbf{S}_i] = \prod_{V_k \in \mathbf{S}_i} P(v_k \mid v^{(k-1)})$.
3. Let $\mathbf{D}_1, \mathbf{D}_2, \ldots$ be C-components of $G(\mathbf{D})$ where $\mathbf{D} = An(\mathbf{Y})_{G(\mathbf{V}\setminus\mathbf{X})}$.
4. Identify $Q[\mathbf{D}_j]$ from $Q[\mathbf{S}]$ by recursively applying C-factor algebra
5. $P_{\mathbf{x}}(\mathbf{y}) = \sum_{\mathbf{d}\setminus\mathbf{y}} \prod_j Q[\mathbf{D}_j]$ if all $Q[\mathbf{D_j}]$ is defined, FAIL otherwise.

Recap: Tian's ID algorithm

ID$(\mathbf{X}, \mathbf{Y}, G)$

1. Let $\mathbf{S}_1, \mathbf{S}_2, \ldots$ be the C-compo

2. Let $Q[\mathbf{S}_i] = \prod_{V_k \in \mathbf{S}_i} P(v_k \mid v^{(k-1)}$

3. Let $\mathbf{D}_1, \mathbf{D}_2, \ldots$ be C-components of $G(\mathbf{D})$ where $\mathbf{D} = An(\mathbf{Y})_{G(\mathbf{V}\setminus\mathbf{X})}$.

4. Identify $Q[\mathbf{D}_j]$ from $Q[\mathbf{S}]$ by recursively applying C-factor algebra

C-factor algebra:

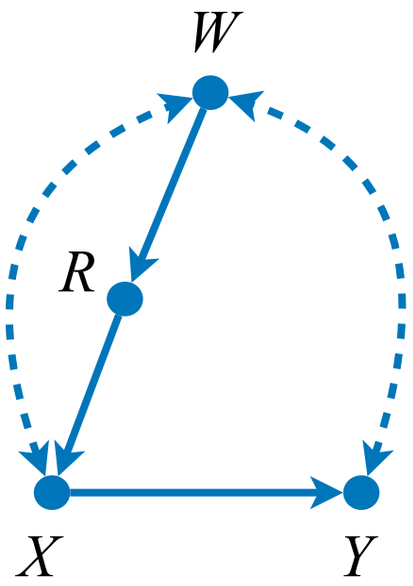
Identification	$Q[\mathbf{C}] = \prod_{V_i \in \mathbf{C}} P(v_i \mid v^{(i-1)})$	where $\mathbf{C}$ is a C-component in G
Marginalization	$Q[\mathbf{W}] = \sum_{\mathbf{c}\setminus\mathbf{w}} Q[\mathbf{C}]$	If $\boldsymbol{W}$ is ancestral in $G[\mathbf{C}]$
Factorization	$Q[\mathbf{H}] = \prod_{j} Q[H_j]$	Where $\mathbf{H}_1, \ldots, \mathbf{H}_k$, are the c-components in $G[\mathbf{H}]$

Key results:

1. A unit for identification is C-factor.
2. An identification functional as an arithmetic (marginalization, ratio, multiplication) of C-factors.

ID example: the Napkin

- Recall the Napkin graph

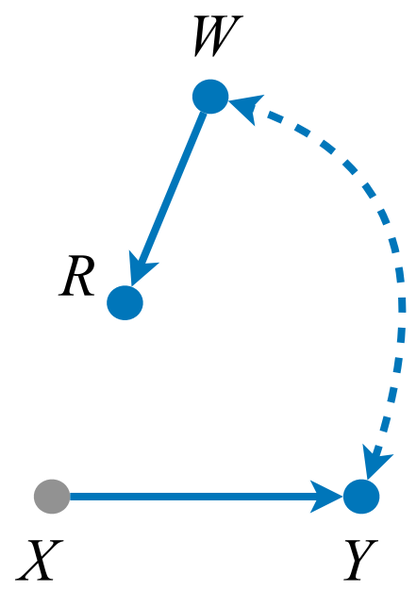


C-components:

$$\mathbf{S}_1 = \{W, X, Y\}, \mathbf{S}_2 = \{R\}.$$

C-factors:

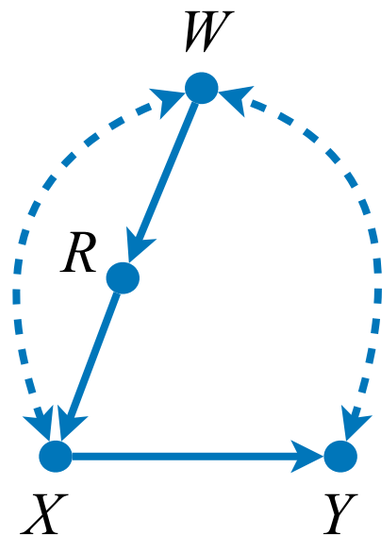
$$Q[\mathbf{S}_1] = P(w)P(x, y \mid r, w),\ Q[\mathbf{S}_2] = P(r \mid w)$$



$$\mathbf{D} \equiv An(\mathbf{Y})_{G(\mathbf{V}\setminus\mathbf{X})}\colon \{Y\}$$

$$P_x(y) = Q[Y]$$

ID example: the Napkin



$$P_x(y) = Q[Y]$$

- (Marginalization) $\{X, Y\}$ is an ancestral set in $G(\mathbf{S}_1)$.

$$Q[X, Y] = \sum_w Q[\mathbf{S}_1] = \sum_w P(x, y \mid r, w)P(w)$$

- (Factorization)

$$Q[X, Y] = Q[X]Q[Y] \implies Q[Y] = \frac{Q[X, Y]}{Q[X]}$$

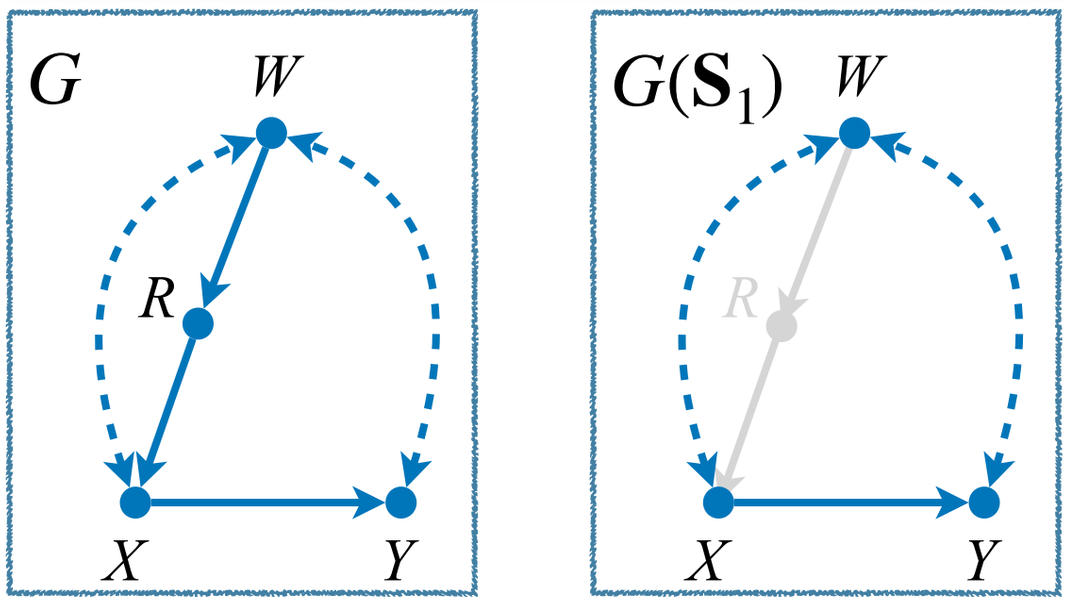
Can any identifiable functional be encoded as a function of mSBD?

C-factor as mSBD

C-factor identification

Let $\mathbf{C}$ be a C-component in G, $\mathbf{W}$ denote the ancestral set of $\mathbf{C}$ (i.e., $\mathbf{W} = An(\mathbf{W})_{G(\mathbf{C})}$) and $\mathbf{R} \equiv Pa(\mathbf{W})$. Then, $\mathbf{Z} = (\mathbf{C} \backslash \mathbf{W}) \cap An(\mathbf{R}, \mathbf{W})$ satisfies mSBD adjustment relative to $(\mathbf{R}, \mathbf{W})$, and

$$Q[\mathbf{W}] = P_{\mathbf{r}}(\mathbf{w}) = M[\mathbf{w} \mid \mathbf{r}; \mathbf{z}]$$



$\mathbf{S}_1 = \{W, X, Y\}, \mathbf{S}_2 = \{R\}.$

- $\mathbf{W} = \{X, Y\}$ s.t. $\{X, Y\} = An(\{X, Y\})_{G(\mathbf{S}_1)}$
- $\mathbf{R} = Pa(\{X, Y\}) = \{R\}.$
- $\mathbf{Z} = (\mathbf{S}_1 \backslash \{X, Y\}) \cap An(\{R, X, Y\}) = \{W\}.$

$$Q[X, Y] = M[(x, y) \mid r; w]$$

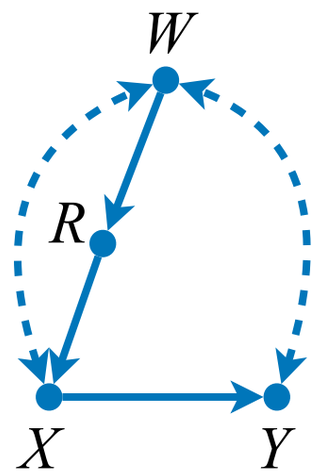
Marginalization of mSBD operators

Marginalization of mSBD operators

Let $M[\mathbf{y} \mid \mathbf{x}; \mathbf{z}]$ be a mSBD operator (where $\mathbf{Z}$ satisfies mSBD).

- If $\mathbf{A}$ is an ancestral set in $\mathbf{Y}$, $\sum_{\mathbf{a}} M[\mathbf{y} \mid \mathbf{x}; \mathbf{z}] = M[\mathbf{y} \backslash \mathbf{a} \mid \mathbf{x}; \mathbf{z} \cup \mathbf{a}]$.

- If $\mathbf{W}$ is a descendent set in $\mathbf{Y}$, $\sum_{\mathbf{w}} M[\mathbf{y} \mid \mathbf{x}; \mathbf{z}] = M[\mathbf{y} \backslash \mathbf{w} \mid \mathbf{x} \cap Pre(\mathbf{y} \backslash \mathbf{w}); \mathbf{z} \cap Pre(\mathbf{y} \backslash \mathbf{w})]$.



$\mathbf{S}_1 = \{W, X, Y\}, \mathbf{S}_2 = \{R\}.$

$Q[X, Y] = M[(x, y) \mid r; w]$

- $\{X\}$ is an ancestral set in $G(\{X, Y\})$.

$$Q[X] = \sum_y Q[X, Y] = \sum_y M[(x, y) \mid r; w]$$

- $\{Y\}$ is a descendant set in $G(\{X, Y\})$:

$$Q[X] = M[x \mid r; w]$$

DML-ID: Representing a causal effect as a function of mSBDs

DML-ID$(\mathbf{X}, \mathbf{Y}, G)$

1. Let $\mathbf{S}_1, \mathbf{S}_2, \ldots$ be the c-components of G.
2. Let $Q[\mathbf{S}_i] = M[\mathbf{s}_i \mid Pa(\mathbf{s}_i); \varnothing]$.
3. Let $\mathbf{D}_1, \mathbf{D}_2, \ldots$ be the c-components of $G(\mathbf{D})$ where $\mathbf{D} = An(\mathbf{Y})_{G(\mathbf{V}\setminus\mathbf{X})}$.
4. $Q[\mathbf{D}_j] =$ Identify$(\mathbf{D}_j, \mathbf{S}_j, Q[\mathbf{S}_j])$.
4. $Q[\mathbf{D}_j] = A^j(\{M^j_\ell\}) =$ Identify$(\mathbf{D}_j, \mathbf{S}_j, Q[\mathbf{S}_j])$. $Q[\mathbf{D}_j]$ is a function of mSBD operators
5. $P(\mathbf{y} \mid do(\mathbf{x})) = \sum_{\mathbf{d}\setminus\mathbf{y}} \prod_j A^j(\{M^j_\ell\})$ if all $A^j(\{M^j_\ell\})$ have been defined, FAIL otherwise.

DML-ID: Representing a causal effect as a function of mSBDs

DML-ID$(\mathbf{X}, \mathbf{Y}, G)$

1. Let $\mathbf{S}_1, \mathbf{S}_2, \ldots$ be the c-compone

2. Let $Q[\mathbf{S}_i] = M[\mathbf{s}_i \mid Pa(\mathbf{s}_i); \varnothing]$.

3. Let $\mathbf{D}_1, \mathbf{D}_2, \ldots$ be the c-components of $G(\mathbf{D})$ where $\mathbf{D} = An(\mathbf{Y})_{G(\mathbf{V}\setminus\mathbf{X})}$.

4. $Q[\mathbf{D}_j] =$ Identify$(\mathbf{D}_j, \mathbf{S}_j, Q[\mathbf{S}_j])$.

4. $Q[\mathbf{D}_j] = A^j(\{M^j_\ell\}) =$ Identify$(\mathbf{D}_j, \mathbf{S}_j, Q[\mathbf{S}_j])$. $Q[\mathbf{D}_j]$ is a function of mSBD operators

5. $P(\mathbf{y} \mid do(\mathbf{x})) = \sum_{\mathbf{d}\setminus\mathbf{y}} \prod_j A^j(\{M^j_\ell\})$ if all $A^j(\{M^j_\ell\})$ have been defined, FAIL otherwise.

C-factor & mSBD algebra:

Identification	$Q[\mathbf{W}] = M[\mathbf{w} \mid \mathbf{r}; \mathbf{z}]$	If $\mathbf{W}$ is ancestral in a C-component $\mathbf{C}$ in G.
Marginalization	$Q[\mathbf{W}] = \sum_{\mathbf{c}\setminus\mathbf{w}} Q[\mathbf{C}]$	If $\boldsymbol{W}$ is ancestral in $G[\mathbf{C}]$.
Factorization	$Q[\mathbf{H}] = \prod_j Q[H_j]$	Where $\mathbf{H}_1, \ldots, \mathbf{H}_k$, are the c-components in $G[\mathbf{H}]$

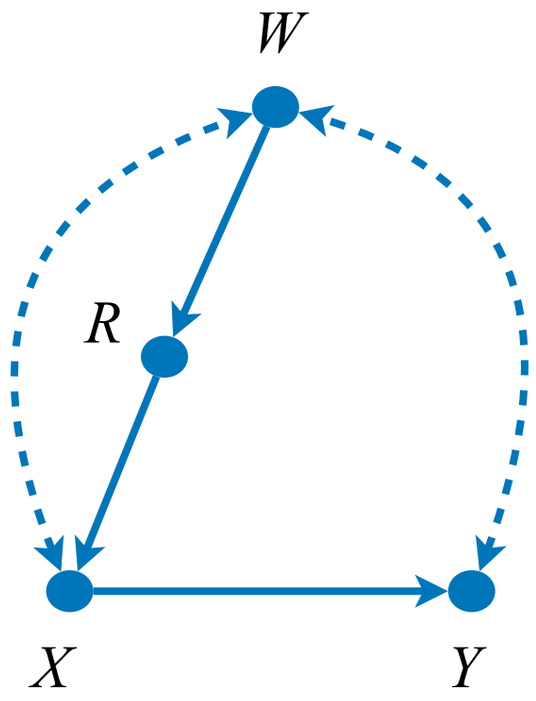
Soundness and Completeness of DML-ID

Soundness and Completeness of DML-ID

A causal effect $P_{\mathbf{x}}(\mathbf{y})$ is identifiable if and only if DML-ID$(\mathbf{X}, \mathbf{Y}, G)$ returns $P_{\mathbf{x}}(\mathbf{y})$ as an arithmetic (marginalization, multiplication, ratios) combination of mSBD operators M:

$$P_{\mathbf{x}}(\mathbf{y}) = \sum_{\mathbf{d} \setminus \mathbf{y}} \prod_{j:\mathbf{D}_j \in \mathbf{D}} A^j(\{M^j_\ell\})$$

Example of DML-ID: Napkin



$\mathbf{S}_1 = \{W, X, Y\}, \mathbf{S}_2 = \{R\}.$

- $Q[\mathbf{S}_1] \equiv P_r(w, x, y) = M[\mathbf{s}_1 \,|\, r; \varnothing]$
- $\{W\}$ is an ancestral set in $G(\mathbf{S}_1)$,

$$Q[X, Y] = \sum_w M[\mathbf{s}_1 \,|\, r; \varnothing] = M[(x, y) \,|\, r; w]$$

- $\{Y\}$ is an descendent set in $G(\{X, Y\})$.

$$Q[X] = \sum_y M[(x, y) \,|\, r; w] = M[x \,|\, r; w]$$

- $$Q[Y] = \frac{Q[X, Y]}{Q[X]} = \frac{M[x, y \,|\, r; w]}{M[x \,|\, r; w]}$$

Task of causal effect estimation

ID $\left(G, P, P_{\mathbf{x}}(\mathbf{y})\right)$

With (1) a causal functional $f(P)$ such that $P_{\mathbf{x}}(\mathbf{y}) = f(P)$ and (2) a dataset D, can we have a reliable estimate T_N for $P_{\mathbf{x}}(\mathbf{y})$?

1 Causal Functional

$P(y \mid do(x)) = f(P)$

2 Data

$D = \{\mathbf{V}_{(i)}\}_{i=1}^{N} \sim P$

Estimation Engine

EST$(f(P), D)$

Estimand

$g(P) = f(P)$

Estimator

$T_N = \widehat{g(P)}$

Task of causal effect estimation with DML-ID

DML-ID $\left(G, P, P_{\mathbf{x}}(\mathbf{y})\right)$

With (1) a causal functional $f(P)$ such that $P_{\mathbf{x}}(\mathbf{y}) = f(P)$ and (2) a dataset D, can we have a reliable estimate T_N for $P_{\mathbf{x}}(\mathbf{y})$?

1 Causal Functional

$P(y \mid do(x)) = A(\{M\})$

2 Data

$D = \{\mathbf{V}_{(i)}\}_{i=1}^{N} \sim P$

Estimation Engine

EST$(A(\{M\}), D)$

UIF h s.t. $\mathbb{E}_P[h(P)] = f(P)$

DML Estimator T_N

Derivation of UIF & Construction of DML estimator

What we learned so far

IF for mSBD adjustment $M(P)$: $q_M \equiv h(\mathbf{V}; \eta = \{\mathbf{H}, \mathbf{W}\}) - M$, where $h = h_M$ is a UIF for M.

IF for mSBD

$$h(\mathbf{V}; \eta = \{\mathbf{H}, \mathbf{W}\}) = H_2 + \sum_{i=2}^{n+1} W_i(H_{i+1} - H_1)$$

Estimand & DML estimator for mSBD adjustment:

Recall that $\mathbb{E}_P[h_M] = M$ since $0 = \mathbb{E}_P[q_M] = \mathbb{E}_P[h_M] - M$.

We will call μ_M

Result of DML-ID: Any identifiable causal effect can be represented as follow:

$$P_{\mathbf{x}}(\mathbf{y}) = f(P) \equiv \sum_{\mathbf{d}\backslash\mathbf{y}} \prod_{j:\mathbf{D}_j \in \mathbf{D}} Q[\mathbf{D}_j] = \sum_{\mathbf{d}\backslash\mathbf{y}} \prod_{j:\mathbf{D}_j \in \mathbf{D}} A^j(\{M^j_\ell\}).$$

$= Q[\mathbf{D}_j]$

Recall: Refined strategy

1. We will derive a UIF $h(\mathbf{V};\eta)$ for $f(P) = P_x(y)$.

- IF: A function $q(\mathbf{V};\eta, P_x(y))$ satisfies the following:

$$\frac{\partial}{\partial\gamma} f(P_\gamma)\,|_{\gamma=0} = \mathbb{E}_P\left[q(\mathbf{V};\eta, P_x(y)\cdot\frac{\partial}{\partial\gamma}\log P_\gamma(\mathbf{V})_{\gamma=0}\right]$$

- UIF: If an IF $q(\mathbf{V};\eta, P_x(y)) \equiv h(\mathbf{V};\eta) - P_x(y)$, h is a UIF.

2. Then, we will construct a DML estimator T_N based on $h(\mathbf{V};\hat{\eta})$ and sample splitting.

$$T_N = \frac{1}{2}\sum_{i\in\{0,1\}} \mathbb{E}_{P_{D_i}}\left[h(\mathbf{V};\hat{\eta}^{1-i})\right].$$

Approach for deriving UIF

Result of DML-ID: Any identifiable causal effect can be represented as follow:

$$P_{\mathbf{x}}(\mathbf{y}) = f(P) \equiv \sum_{\mathbf{d}\backslash\mathbf{y}} \prod_{j:\mathbf{D}_j \in \mathbf{D}} Q[\mathbf{D}_j] = \sum_{\mathbf{d}\backslash\mathbf{y}} \prod_{j:\mathbf{D}_j \in \mathbf{D}} A^j(\{M^j_\ell\}).$$

To derive an IF/UIF for $f(P)$, for $P_\gamma = P(1+\gamma g)$ (where g is a mean-zero function), we consider a ***derivative*** of $f(P_\gamma) = \sum_{\mathbf{d}\backslash\mathbf{y}} \prod_{j:\mathbf{D}_j \in \mathbf{D}} A^j(\{M^j_\ell(P_\gamma)\})$.

$$\frac{\partial}{\partial\gamma} f(P_\gamma)\big|_{\gamma=0} = \sum_{\mathbf{d}\backslash\mathbf{y}} \sum_{j:\mathbf{D}_j \in \mathbf{D}} \left(\frac{\partial}{\partial\gamma} A^j(\{M^j_\ell(P_\gamma)\})\big|_{\gamma=0} \right) \prod_{p\neq j} A^p(\{M^p_\ell\})$$

$$= \sum_{\mathbf{d}\backslash\mathbf{y}} \sum_{j:\mathbf{D}_j \in \mathbf{D}} \left(\frac{\partial}{\partial\gamma} A^j(\{M^j_\ell(P_\gamma)\})\big|_{\gamma=0} \right) \prod_{p\neq j} A^p(\{\mu_{M^p_\ell}\})$$

Replace M to μ_M (because mean of the UIF of M is M; i.e., $\mu_M \equiv \mathbb{E}_P[h_M] = M$.

We replace for statistical benefits.

A directional derivative along the direction Pg.

Denoted $\nabla Q[\mathbf{D}_j]$.

Algorithm for deriving UIF

$$\nabla Q[\mathbf{D}_j] = \mathbb{E}_P\left[\left(\sum_{\ell} J_{A^j,M_\ell^j}\right)\cdot S(\mathbf{V})\right] \quad \text{where } J_{A^j,M_\ell^j} = \text{ComponentUIF}(A^j, M_\ell^j)$$

IF for $Q[\mathbf{D}_j]$

ComponentUIF(A, M_r)

Run $J_{A,r} \equiv$ UIF(A, M_r); and Replace arguments M_r of $J_{A,r}$ to μ_{M_r} and **return** $J_{A,r}$.

By a mechanism of UIF( ), it returns a function of $\{\mu_M\}$ and q_{M_r}

i.e., $J_{A,r} = J_{A,r}\left(\{\mu_M\}, q_{M_r}\right)$

UIF(A, M_r)

If $A = C$ for some constant w.r.t M_r, then **return** 0.

If $A = M_r$, then **return** q_{M_r}, an IF for the mSBD adjustment M_r.

If $A = C \cdot A'$ for some function A' of M_r, then **return** UIF(A', M_r) (constant multiplication)

If $A = A' \cdot A''$ for some function A', A'' of M_r, then **return** $A'' \cdot$ UIF$(A', M_r) + A' \cdot$ UIF(A'', M_r) (product rule)

If $A = 1/A'$. then **return** $-1/(A')^2 \cdot$ UIF(A', M_r). (quotient rule)

If $A = \sum A'$, then **return** $\sum$ UIF(A', M_r). (interchange rule)

IF/UIF for identifiable functional

Derivation of IF/UIF for identifiable causal functional.

For $f(P) = P_x(y)$, let $q(\mathbf{V}; \eta, P_x(y))$ denote an IF for $f(P)$, and $h(\mathbf{V}; \eta)$ denote a UIF. Let $J_{A^j, M^j_\ell} = \text{ComponentUIF}(A^j, M^j_\ell)$. Then,

$$q(\mathbf{V}; \eta, P_x(y)) = \sum_{\mathbf{d}\backslash\mathbf{y}} \sum_{j:\mathbf{D}_j \in \mathbf{D}} \left(\sum_{\ell} J_{A^j, M^j_\ell} \right) \prod_{p \neq j} A^p(\{\mu_{M^p_\ell}\})$$

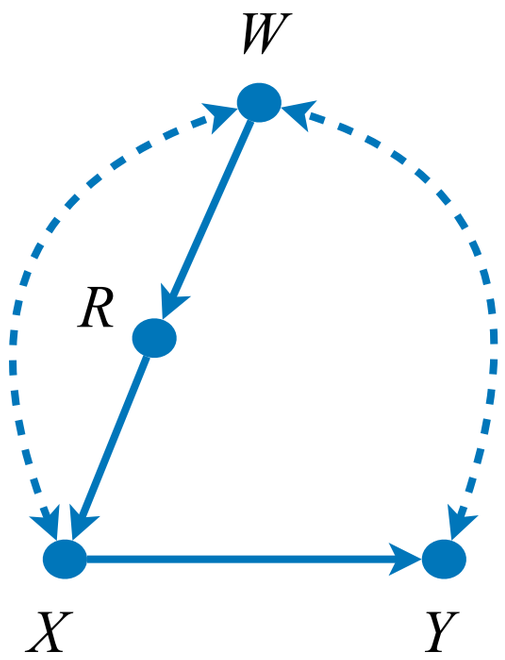
$$J_{A,r} = J_{A,r}\left(\{\mu_{M^i_r}\}, q_{M^i_\ell}\right)$$

$$h(\mathbf{V}; \eta) = \sum_{\mathbf{d}\backslash\mathbf{y}} \sum_{\substack{j:\mathbf{D}_j \in \mathbf{D} \\ j \neq 1}} \left(\sum_{\ell} J_{A^j, M^j_\ell} \right) \prod_{p \neq j} A^p(\{\mu_{M^p_\ell}\})$$

$$+ \sum_{\mathbf{d}\backslash\mathbf{y}} \left(\sum_{\ell \neq 1} J_{A^1, M^1_\ell} \right) \prod_{p \neq 1} A^p(\{\mu_{M^p_r}\}) + \sum_{\mathbf{d}\backslash\mathbf{y}} A^1(h_{M^1_1}, \{\mu_{M^1_r}\}_{r \neq 1}) \cdot \prod_{p \neq 1} A^p(\{\mu_{M^p_r}\})$$

- Given that q, h are a function of $\mu_{M^j_\ell}$, nuisances η are $\{\mathbf{H}^j_\ell, \mathbf{W}^j_\ell\}$, nuisances for an IF for mSBD adjustment M^j_ℓ.

Example for deriving IF



$$\mathbf{S}_1 = \{W, X, Y\}, \mathbf{S}_2 = \{R\}, \mathbf{D} = \{Y\}.$$

$$P_x(y) = Q[Y] = \frac{Q[X,Y]}{Q[Y]} = \frac{M[x, y \mid r; w]}{M[x \mid r; w]}$$

$$= A(M_1, M_2) = \frac{M_1}{M_2}$$

$$q(\mathbf{V}; \eta, P_x(y)) = \sum_{\mathbf{d}\backslash\mathbf{y}} \sum_{j:\mathbf{D}_j \in \mathbf{D}} \left(\sum_{\ell} J_{A^j, M^j_\ell} \right) \prod_{p \neq j} A^p(\{\mu_{M^p_\ell}\})$$

$$q(\mathbf{V}; \eta, P_x(y)) = \left(\sum_{\ell=1}^{2} J_{A, M_\ell} \right),$$

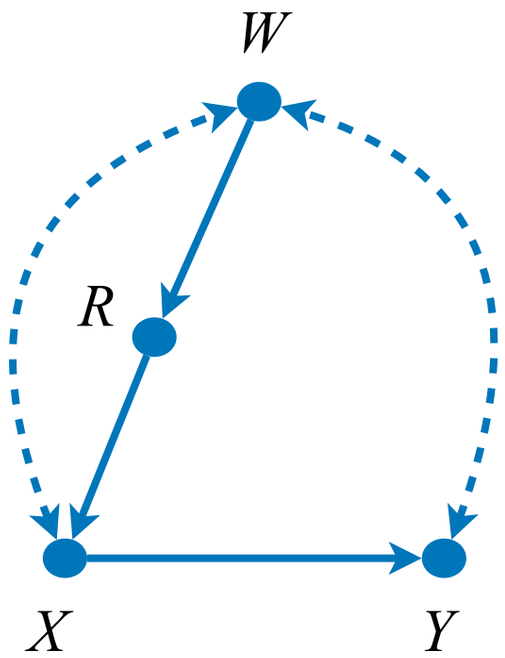
where $J_{A,M_1} = \frac{q_{M_1}}{\mu_{M_2}}$; $J_{A,M_2} = -q_{M_2} \cdot \frac{\mu_{M_1}}{\mu_{M_2}} \cdot \frac{1}{\mu_{M_2}}$, where q_M is an IF for mSBD; and $\mu_M = \mathbb{E}_P[h_M]$, h_M is a UIF.

$$q(\mathbf{V}; \eta, P_x(y)) = \frac{1}{\mu_{M_2}} \left(q_{M_1} - q_{M_2} \frac{\mu_{M_1}}{\mu_{M_2}} \right).$$

$$h_{M_1} = \frac{I_r(R)}{P(R \mid X)} (I_{x,y}(X, Y) - P(x, y \mid R, W)) + P(x, y \mid r, W)$$

$$h_{M_2} = \frac{I_r(R)}{P(R \mid X)} (I_x(X, Y) - P(x \mid R, W)) + P(x \mid r, W)$$

Example for deriving UIF



$$\mathbf{S}_1 = \{W, X, Y\}, \mathbf{S}_2 = \{R\}, \mathbf{D} = \{Y\}.$$

$$Q[Y] = \frac{Q[X,Y]}{Q[Y]} = \frac{M[x,y \mid r;w]}{M[x \mid r;w]}$$

$$= A(M_1, M_2) = \frac{M_1}{M_2}$$

$$h(\mathbf{V};\eta) = \sum_{\mathbf{d}\backslash\mathbf{y}} \sum_{\substack{j:\mathbf{D}_j \in \mathbf{D} \\ j \neq 1}} \left(\sum_{\ell} J_{A^j, M_\ell^j} \right) \prod_{p \neq j} A^p(\{\mu_{M_\ell^p}\})$$

$$+ \sum_{\mathbf{d}\backslash\mathbf{y}} \left(\sum_{\ell \neq 1} J_{A^1, M_\ell^1} \right) \prod_{p \neq 1} A^p(\{\mu_{M_r^p}\})$$

$$+ \sum_{\mathbf{d}\backslash\mathbf{y}} A^1(h_{M_1^1}, \{\mu_{M_r^1}\}_{r \neq 1}) \cdot \prod_{p \neq 1} A^p(\{\mu_{M_r^p}\})$$

$$h(\mathbf{V};\eta) = J_{A,M_2} + A(h_{M_1}, \mu_{M_2}) = \frac{h_{M_1}}{\mu_{M_2}} - q_{M_2} \frac{\mu_{M_1}}{\mu_{M_2}} \cdot \frac{1}{\mu_{M_2}}$$

$$h_{M_1} = \frac{I_r(R)}{P(R \mid X)} (I_{x,y}(X,Y) - P(x,y \mid R,W)) + P(x,y \mid r,W)$$

$$h_{M_2} = \frac{I_r(R)}{P(R \mid X)} (I_x(X,Y) - P(x \mid R,W)) + P(x \mid r,W)$$

Construction of DML estimator

1. We will derive a UIF $h(\mathbf{V};\eta)$ for $f(P) = P_x(y)$.

$$h(\mathbf{V};\eta) = \sum_{\mathbf{d}\backslash\mathbf{y}} \sum_{\substack{j:\mathbf{D}_j \in \mathbf{D} \\ j \neq 1}} \left(\sum_{\ell} J_{A^j, M^j_\ell} \right) \prod_{p \neq j} A^p(\{\mu_{M^p_\ell}\})$$
$$+ \sum_{\mathbf{d}\backslash\mathbf{y}} \left(\sum_{\ell \neq 1} J_{A^1, M^1_\ell} \right) \prod_{p \neq 1} A^p(\{\mu_{M^p_r}\}) + \sum_{\mathbf{d}\backslash\mathbf{y}} A^1(h_{M^1_1}, \{\mu_{M^1_r}\}_{r \neq 1}) \cdot \prod_{p \neq 1} A^p(\{\mu_{M^p_r}\})$$

2. Then, we will construct a DML estimator T_N based on $h(\mathbf{V};\hat{\eta})$ and sample splitting.

$$T_N = \frac{1}{2} \sum_{i \in \{0,1\}} \mathbb{E}_{D_i} \left[h(\mathbf{V}; \hat{\eta}^{1-i}) \right].$$

Properties of DML estimator

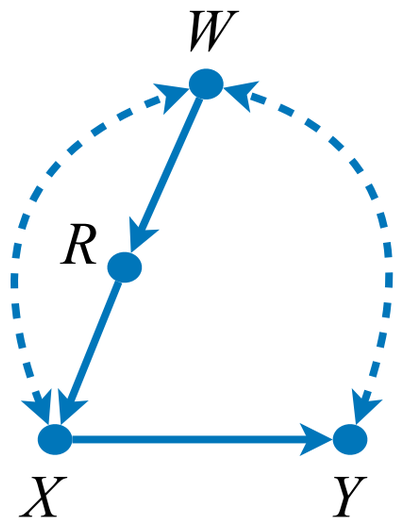
For η are $\{\mathbf{H}^j_\ell, \mathbf{W}^j_\ell\}$, nuisances for $q_{M^j_\ell}$, UIF h and DML estimator T_N is

- $$h(\mathbf{V};\eta) = \sum_{\mathbf{d}\backslash\mathbf{y}} \sum_{\substack{j:\mathbf{D}_j \in \mathbf{D} \\ j \neq 1}} \left(\sum_{\ell} J_{A^j, M^j_\ell} \right) \prod_{p \neq j} A^p(\{\mu_{M^p_\ell}\}) + \sum_{\mathbf{d}\backslash\mathbf{y}} \left(\sum_{\ell \neq 1} J_{A^1, M^1_\ell} \right) \prod_{p \neq 1} A^p(\{\mu_{M^p_r}\})$$
  $$+ \sum_{\mathbf{d}\backslash\mathbf{y}} A^1(h_{M^1_1}, \{\mu_{M^1_r}\}_{r \neq 1}) \cdot \prod_{p \neq 1} A^p(\{\mu_{M^p_r}\})$$
- $$T_N = \frac{1}{2} \sum_{i \in \{0,1\}} \mathbb{E}_{D_i} \left[h(\mathbf{V}; \hat{\eta}^{1-i}) \right]$$

Debiasedness: $\hat{\eta} \to \eta$ at rate $N^{-1/4}$, then $T_N \to P_x(y)$ in $N^{-1/2}$.

Doubly robustness: If estimation models for $\hat{\mathbf{H}}^j_\ell$ or $\hat{\mathbf{W}}^j_\ell$ converges to $\mathbf{H}^j_\ell$ or $\mathbf{W}^j_\ell$, then $T_N \to P_x(y)$.

Properties of DML estimator: Example



$$h(\mathbf{V};\eta) = \frac{h_{M_1}}{\mu_{M_2}} - q_{M_2}\frac{\mu_{M_1}}{\mu_{M_2}} \cdot \frac{1}{\mu_{M_2}} \text{ where}$$

$$h_{M_1} = \frac{I_r(R)}{P(R\,|\,X)}(I_{x,y}(X,Y) - P(x,y\,|\,R,W)) + P(x,y\,|\,r,W)$$

$$h_{M_2} = \frac{I_r(R)}{P(R\,|\,X)}(I_x(X,Y) - P(x\,|\,R,W)) + P(x\,|\,r,W)$$

$$P_x(y) = \frac{M[x,y\,|\,r;w]}{M[x\,|\,r;w]} = \frac{M_1}{M_2}$$

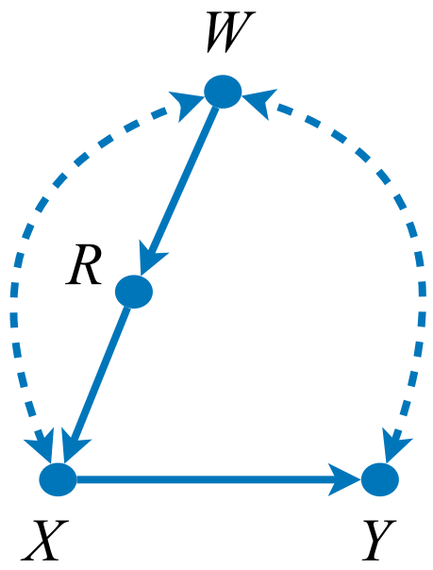
$$\eta = \{P(r\,|\,w), P(x\,|\,r,w), P(x,y\,|\,r,w)\}$$

Debiasedness: $\hat{\eta} \to \eta$ at rate $N^{-1/4}$, then $T_N \to P_x(y)$ in $N^{-1/2}$.

Doubly robustness: If $\hat{P}(r\,|\,w)$ or $\{\hat{P}(x,y\,|\,r,w), \hat{P}(x\,|\,r,w)\}$ converges correctly, then $T_N \to P_x(y)$.

Experimental Results

Experimental setup



$$P_x(y) = f(P) = \frac{\sum_w P(x, y \mid r, w) P(w)}{\sum_w P(x \mid r, w) P(w)}$$

$$f(\hat{P}) = \frac{\sum_w \hat{P}(x, y \mid r, w) \hat{P}(w)}{\sum_w \hat{P}(x \mid r, w) \hat{P}(w)}$$

Debiasedness: $\hat{\eta} \to \eta$ at rate $N^{-1/4}$, then $T_N \to P_x(y)$ in $N^{-1/2}$.

Doubly robustness: If $\hat{P}(r \mid x)$ or $\{\hat{P}(x, y \mid r, w), \hat{P}(x \mid r, w)\}$ converge correctly, then $T_N \to P_x(y)$.

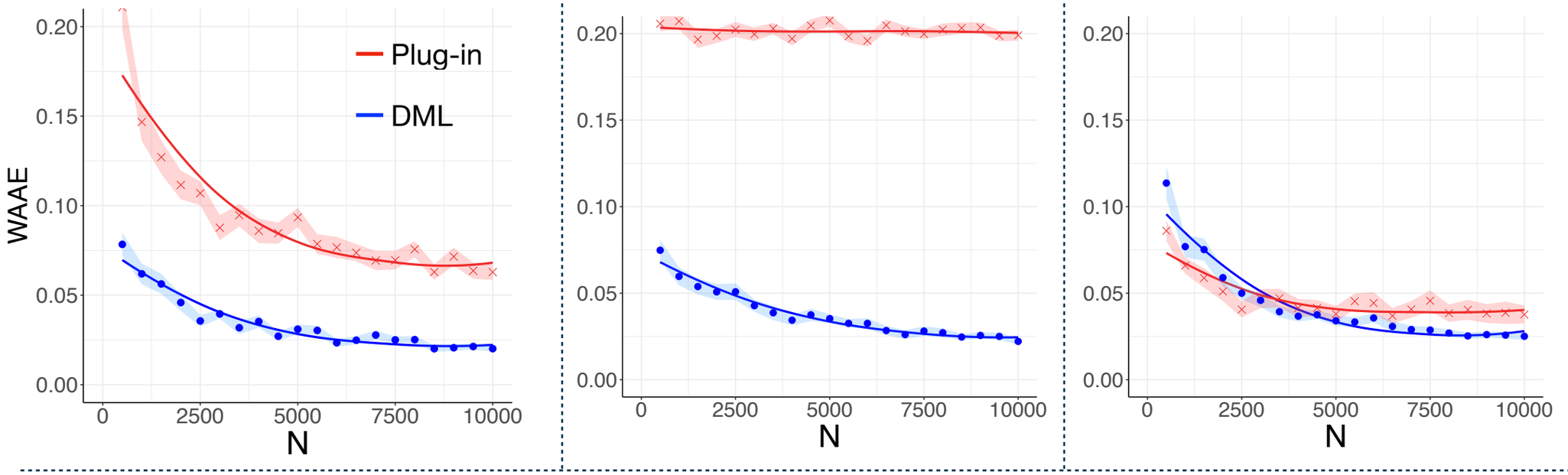
- T_N vs. $f(\hat{P})$: A proposed DML estimator is compared with the plug-in estimator $f(\hat{P})$, only viable estimator working for identifiable causal functional.

Simulations

Both converge at rate $N^{-1/4}$

$\{P(x, y \mid r, w), P(x \mid r, w)\}$ misspecified.

$P(r \mid w)$ misspecified.

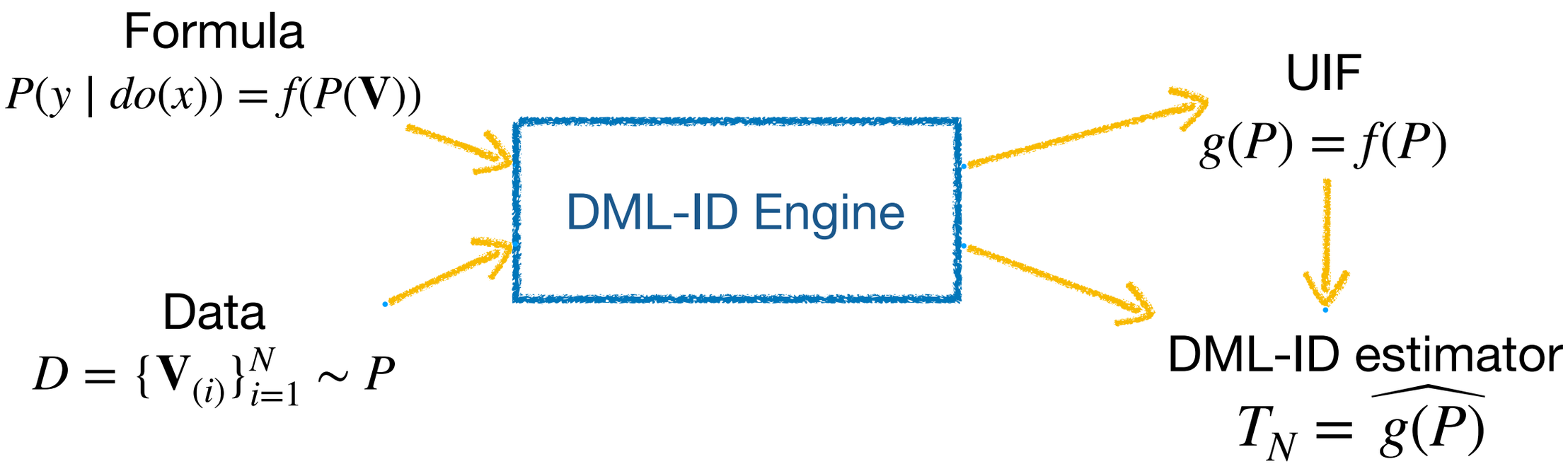


- **(**Debiasedness**;** Left**)** DML converges (i.e., the error 'WAAE' decreases) faster even when nuisances converge slower rate ($N^{-1/4}$).

- (Doubly Robustness; (Center, Right)) DML converges even when models for either $P(x, y \mid r, w)$ (center) or $P(r \mid w)$ (right) is misspecified.

Conclusions

- We develop a systematic procedure for deriving IF/UIF for estimands of any identifiable causal effects.
- A DML estimator for any identifiable causal effect, which enjoy *debiasedness* and *doubly robustness* against model misspecification and slow convergence rate, is developed.



Reference

1. Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. "Estimation of regression coefficients when some regressors are not always observed." Journal of the American statistical Association 89.427 (1994): 846-866.

2. Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

3. Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., & Robins, J. M. (2016). Locally robust semiparametric estimation. *arXiv preprint arXiv:1608.00033*.

4. Jung, Tian, Bareinboim (2020), Estimating Causal Effects Using Weighting-Based Estimators. In Proceedings of the 34th AAAI Conference on Artificial Intelligence, 2020.

Part II.

Weighted Empirical Risk Minimization

Necessity of other thread of estimators

Despite the power of DML-ID estimator, we need to develop a new family of estimators…

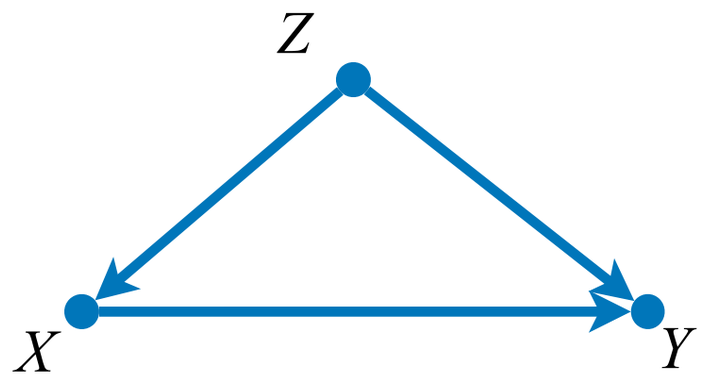
- "*two wrong models are worse than one*" — For doubly robust estimators, '*errors get doubled*' if all nuisances are misspecified.

- Computational difficulty — Evaluating a DML-ID estimator with high-dimensional variables in polynomial time is difficult.

> In this work, we present an estimator inspired by weighting-based methods, which addresses the computational issue.

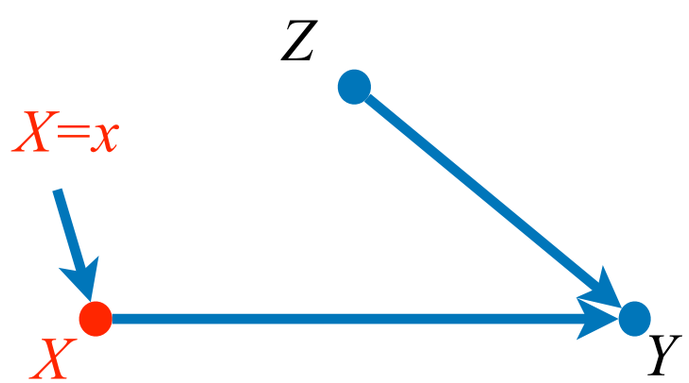
IPW as Domain adaptation

Observational (Source)



$$P(z, x, y) = P(z)P(x \mid z)P(y \mid x, z)$$

Interventional (Target)



$$P(z, x', y) = P(z)I_x(x')P(y \mid x, z)$$

$$P(y \mid do(x)) = P(y) = \mathbb{E}_P[I_y(Y)]$$

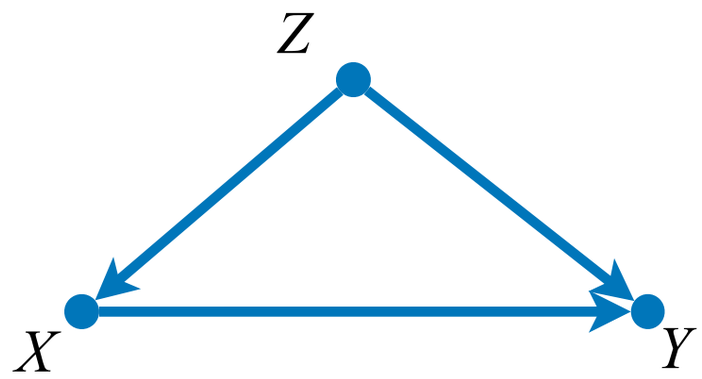
How can we bridge?

- Let $P^W(z, x', y) \equiv W \cdot P(z, x', y)$ be the distribution weighted by $W = \dfrac{I_x(x')}{P(x' \mid z)}$.

$$P^W(z, x', y) = P(z, x', y),$$

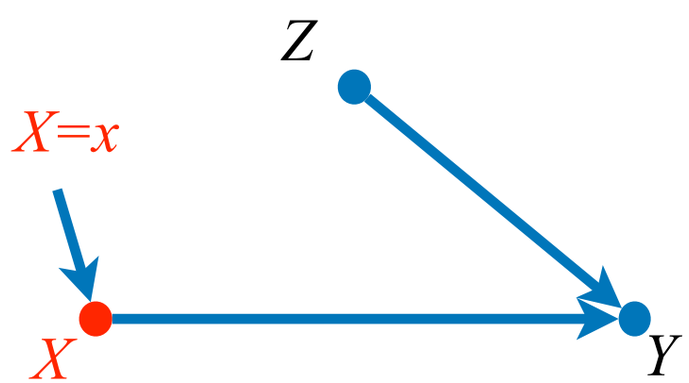
IPW as Domain adaptation

Observational (Source)



$$P(z,x,y) = P(z)P(x\,|\,z)P(y\,|\,x,z)$$

Interventional (Target)



$$P(z,x',y) = P(z)I_x(x')P(y\,|\,x,z)$$

$$I_y(Y)]$$

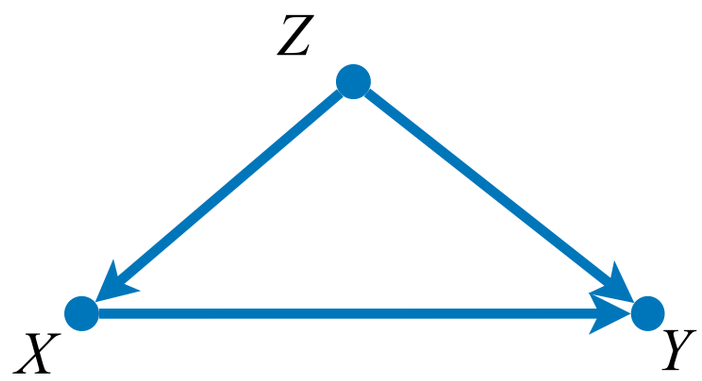
$$P^W(z,x',y) = \frac{I_x(x')}{P(x'\,|\,z)}P(z,x',y) = P(z)\cancel{P(x'\,|\,z)}\frac{I_x(x')}{\cancel{P(x'\,|\,z)}}P(y\,|\,x',z) = P(z,x',y)$$

- Let $P^W(z,x',y) \equiv W \cdot P(z,x',y)$ be the distribution weighted by $W = \frac{I_x(x')}{P(x'\,|\,z)}$.

$$P^W(z,x',y) = P(z,x',y),$$

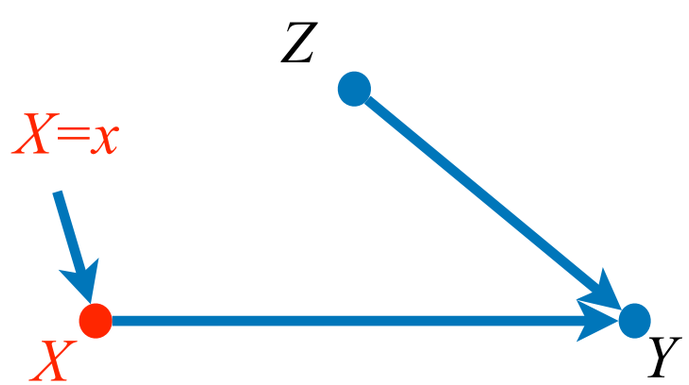
Interpretation of IPW

Observational



$$P(z,x,y) = P(z)P(x\,|\,z)P(y\,|\,x,z)$$

Interventional



$$P(z,x',y) = P(z)I_x(x')P(y\,|\,x,z)$$

$$P(y\,|\,do(x)) = P(y) = \mathbb{E}_P[I_y(Y)]$$

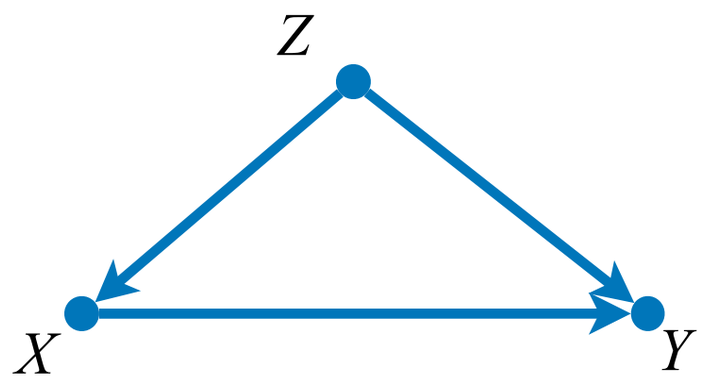
- $P^W(z,x',y) = P(z,x',y)$. Therefore, $P(y\,|\,do(x)) = \mathbb{E}_P[I_y(Y)] = \mathbb{E}_{P^W}[I_y(Y)]$. Then,

$$\mathbb{E}_{P^W}[I_y(Y)] = \mathbb{E}_P\left[W \cdot I_y(Y)\right], \text{ where } W = \frac{I_x(X)}{P(X\,|\,Z)}$$

IPW!

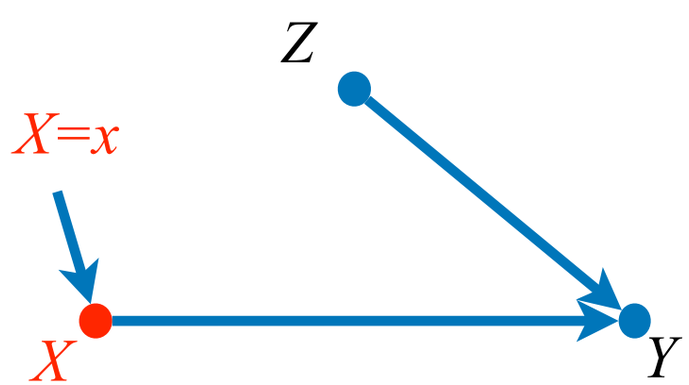
Interpretation of IPW

Observational



$$P(z, x, y) = P(z)P(x \mid z)P(y \mid x, z)$$

Interventional



$$P(z, x', y) = P(z)I_x(x')P(y \mid x, z)$$

$$\mathbb{E}_{P^W}[I_y(Y)] = \sum_{x',z,y} P^W(y', x', z)I_x(x')I_y(y') = \sum_{x',z,y'} \frac{I_x(x')I_y(y')}{P(x' \mid z)} P(z, x', y) = \mathbb{E}_P\left[\frac{I_x(X)}{P(X \mid Z)} I_y(Y)\right]$$

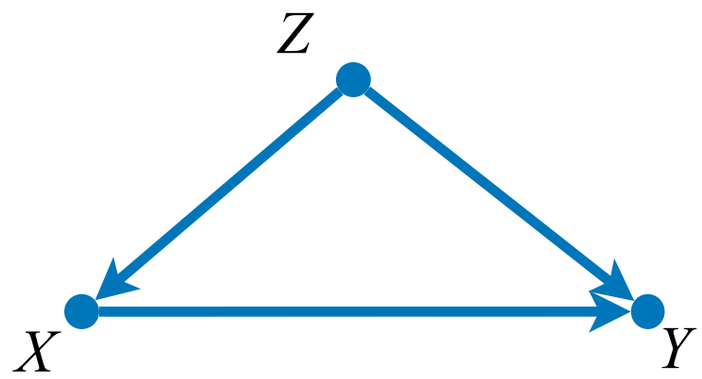
- P

$$\mathbb{E}_{P^W}[I_y(Y)] = \mathbb{E}_P\left[W \cdot I_y(Y)\right], \text{ where } W = \frac{I_x(X)}{P(X \mid Z)}$$

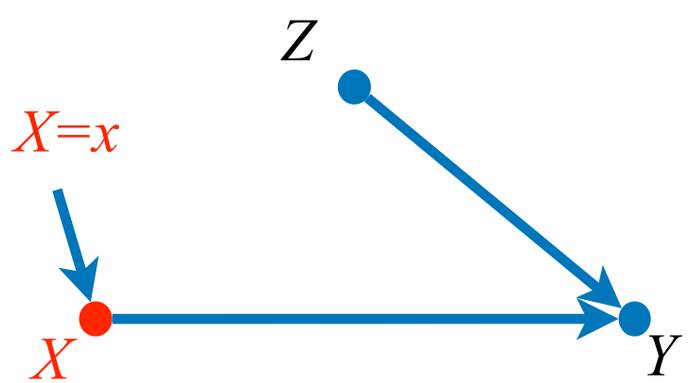
IPW!

Interpretation of IPW

Observational



Interventional



- An IPW estimator can be interpreted as an expectation on a weighted distribution (Domain adaptation)
- A weight $W = \dfrac{I_x(x')}{P(x'|z)}$ is a bridge connecting observational & interventional distribution.

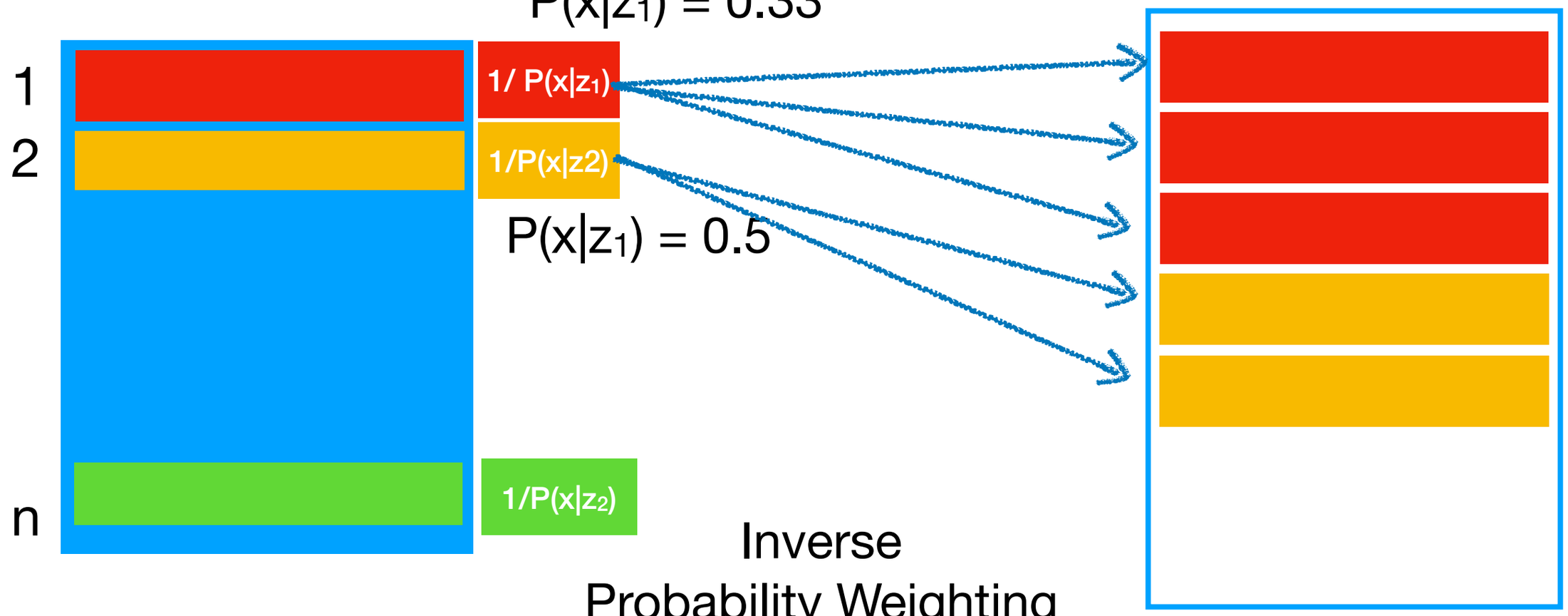
$$\mathbb{E}_{P^W}[I_y(Y)] = \mathbb{E}_P\left[W \cdot I_y(Y)\right], \text{ where } W = \frac{}{P(X|Z)}$$

IPW!

Visualization of IPW

- In practice, evaluating the expr. can be seen as:

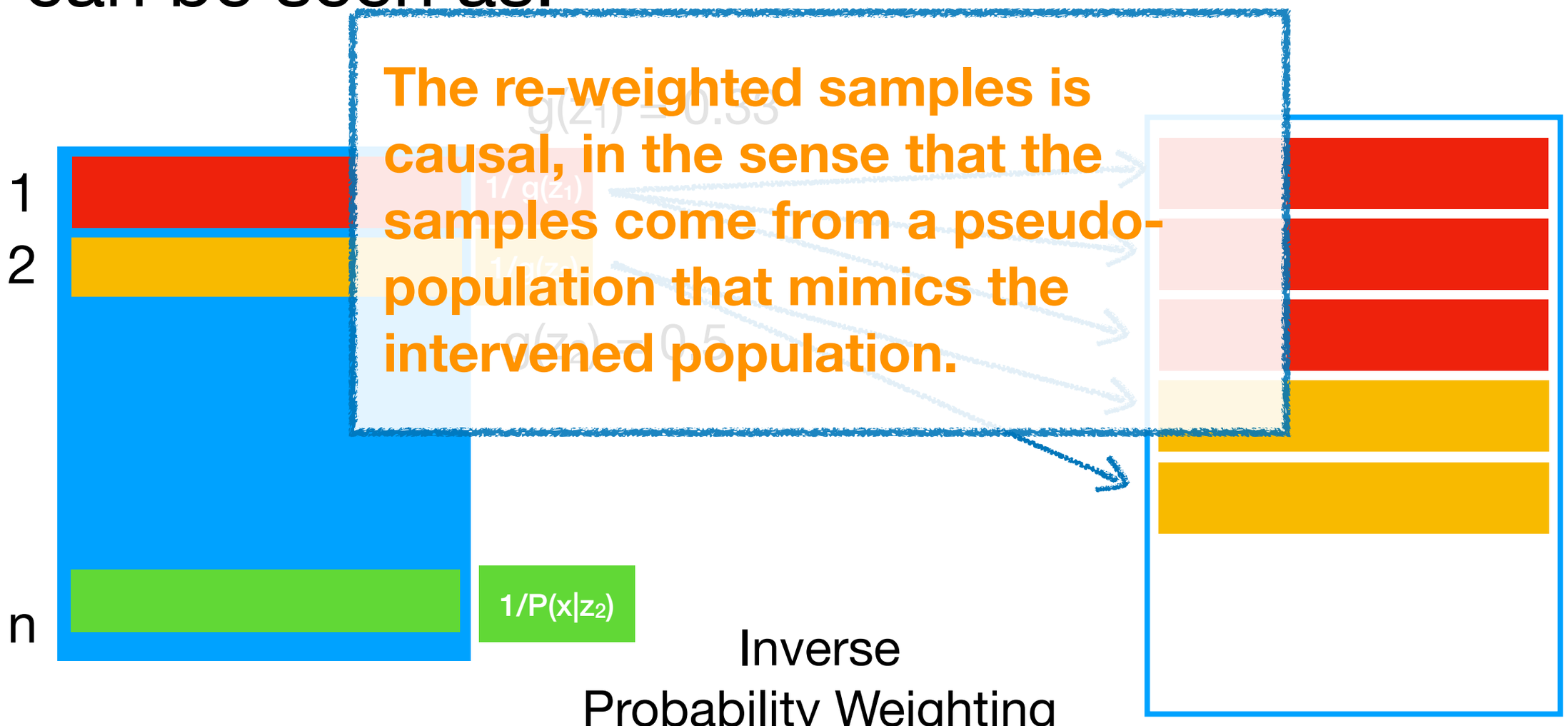
$$\mathbb{E}_P\left[\frac{I_x(X)}{P(X|Z)}I_y(Y)\right]$$



Visualization of IPW

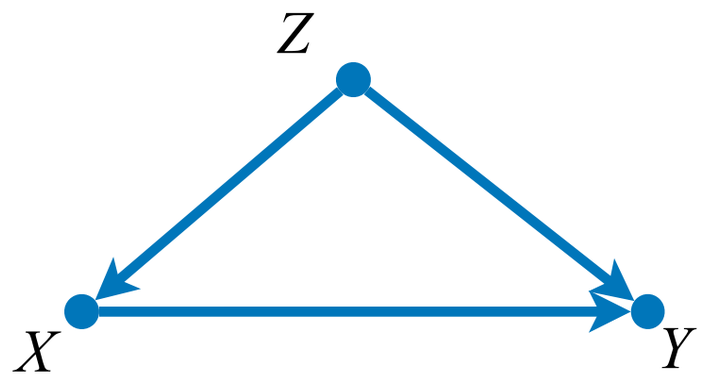
- In practice, evaluating the expr. can be seen as:

$$\mathbb{E}_P\left[\frac{I_x(X)}{P(X|Z)}I_y(Y)\right]$$



Stabilized IPW (SW)

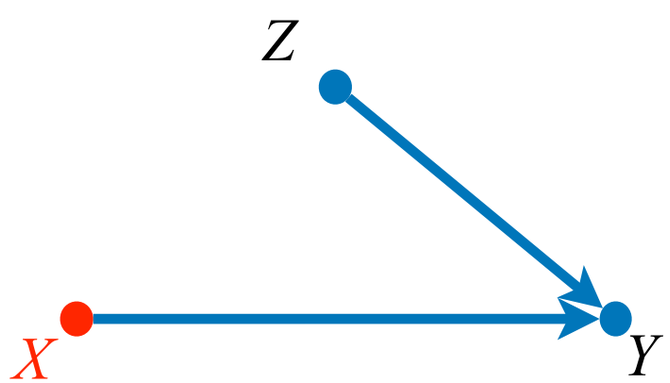
Observational (Source)



$$P(z,x,y) = P(z)P(x\,|\,z)P(y\,|\,x,z)$$

$$P(y\,|\,do(x)) = \sum_z P(y\,|\,x,z)P(z)$$

Interventional (Target)



$$P(z,x,y) = P(z)P(x)P(y\,|\,x,z)$$

$$P(y\,|\,do(x)) = P(y\,|\,x) = \mathbb{E}_P[I_y(Y)\,|\,x]$$

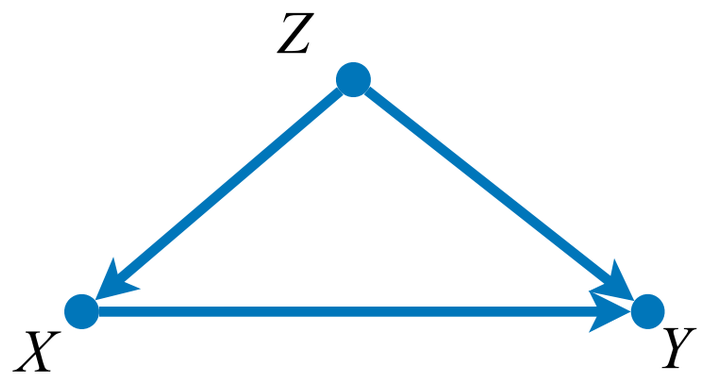
How can we bridge?

- Let $P^W(z,x,y) \equiv W \cdot P(z,x,y)$ be the distribution weighted by $W = \dfrac{P(x)}{P(x\,|\,z)}$, Then,

$$P^W(z,x,y) = P(z,x,y),$$

Stabilized IPW (SW)

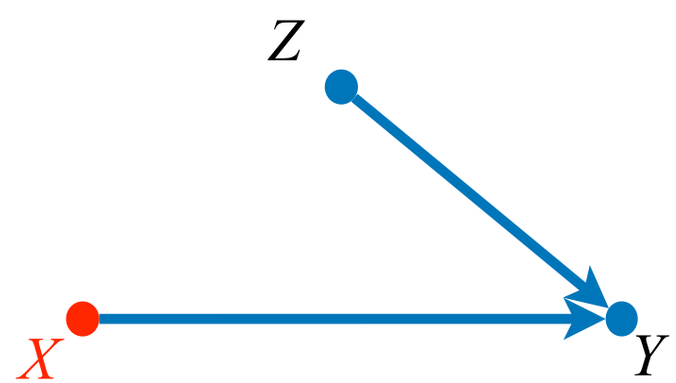
Observational (Source)



$$P(z,x,y) = P(z)P(x\,|\,z)P(y\,|\,x,z)$$

$$P(y\,|\,do(x)) = \sum_z P(y\,|\,x,z)P(z)$$

Interventional (Target)



$$P(z,x,y) = P(z)P(x)P(y\,|\,x,z)$$

$$P(y\,|\,do(x)) = P(y\,|\,x) = \mathbb{E}_P[I_y(Y)\,|\,x]$$

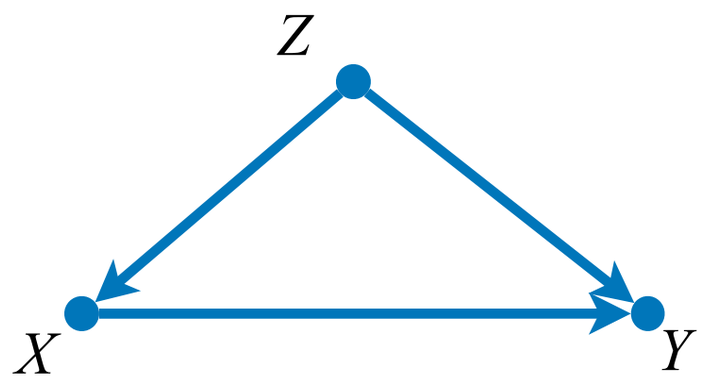
$$P^W(z,x,y) = \frac{P(x)}{P(x\,|\,z)}P(z,x,y) = P(z)\cancel{P(x\,|\,z)}\frac{P(x)}{\cancel{P(x\,|\,z)}}P(y\,|\,x,z) = P(z,x,y).$$

- Let $P^W(z,$ …hen, $P(x\,|\,z)$

$$P^W(z,x,y) = P(z,x,y),$$

Stabilized IPW (SW)

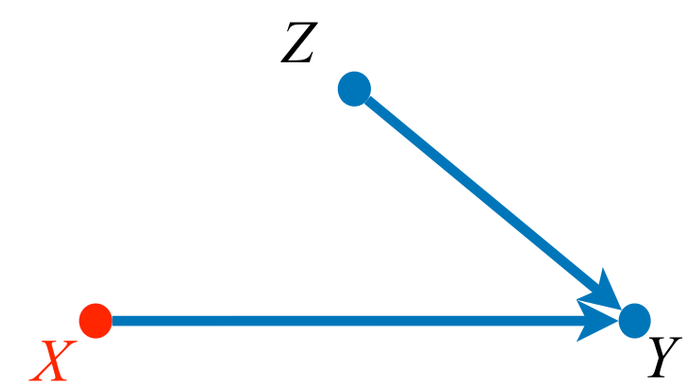
Observational



$$P(z,x,y) = P(z)P(x\,|\,z)P(y\,|\,x,z)$$

$$P(y\,|\,do(x)) = \sum_z P(y\,|\,x,z)P(z)$$

Interventional



$$P(z,x,y) = P(z)P(x)P(y\,|\,x,z)$$

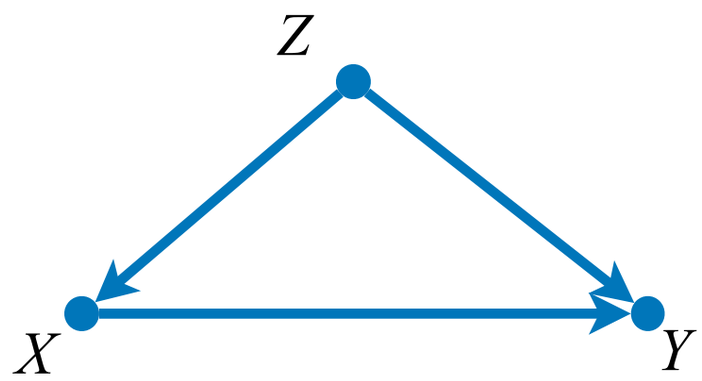
$$P(y\,|\,do(x)) = P(y\,|\,x) = \mathbb{E}_P[I_y(Y)\,|\,x]$$

- $P^W(z,x,y) = P(z,x,y)$. Therefore, $P(y\,|\,do(x)) = \mathbb{E}_P[I_y(Y)\,|\,x] = \mathbb{E}_{P^W}[I_y(Y)\,|\,x]$. Then,

$$\mathbb{E}_{P^W}[I_y(Y)\,|\,x] = \mathbb{E}_P\left[\frac{P(x)}{P(x\,|\,Z)}I_y(Y)\,\middle|\,x\right]$$

Stabilized IPW (SW)

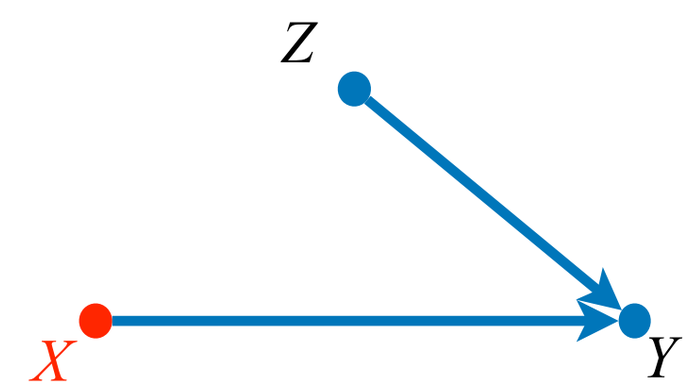
Observational



$$P(z,x,y) = P(z)P(x\,|\,z)P(y\,|\,x,z)$$

$$P(y\,|\,do(x)) = \sum_z P(y\,|\,x,z)P(z)$$

Interventional



$$P(z,x,y) = P(z)P(x)P(y\,|\,x,z)$$

$$P(y\,|\,do(x)) = P(y\,|\,x) = \mathbb{E}_P[I_y(Y)\,|\,x]$$

$$\mathbb{E}_{P^W}[I_y(Y)\,|\,x] = \sum_{z,y'} P^W(y',z\,|\,x)I_y(y') = \sum_{z,y'} P^W(y',z,x)\frac{1}{P^W(x)}I_y(y') = \sum_{z,y'} \frac{P(x)I_y(y')}{P(x\,|\,z)}\frac{P(z,x,y')}{P(x)} = \mathbb{E}_P\left[\frac{P(x)}{P(x\,|\,Z)}I_y(Y)\,\middle|\,x\right]$$

$$\mathbb{E}_{P^W}[I_y(Y)\,|\,x] = \mathbb{E}_P\left[\frac{P(x)}{P(x\,|\,Z)}I_y(Y)\,\middle|\,x\right]$$

Stabilized IPW (SW)

Observational

Interventional

- A weight $\frac{P(x)}{P(x \mid z)}$ is a bridge connecting observational & interventional distribution.

- An SW estimator can be interpreted as an estimate of conditional probability in a weighted distribution.

$$\mathbb{E}_{P^W}[I_y(Y) \mid x] = \sum_{z,y'} P^W(y', z \mid x) I_y(y') = \sum_{z,y'} P^W(y', z, x) \frac{1}{P^W(x)} I_y(y') = \sum_{z,y'} \frac{P(x) I_y(y')}{P(x \mid z)} \frac{P(z, x, y')}{P(x)} = \mathbb{E}_P \left[\frac{P(x)}{P(x \mid Z)} I_y(Y) \middle| x \right]$$

$$\mathbb{E}_{P^W}[I_y(Y) \mid x] = \mathbb{E}_P \left[\frac{P(x)}{P(x \mid Z)} I_y(Y) \middle| x \right]$$

Weighting based method in ML

- In BD setting, there is a weight that makes observational distribution (weighted distribution) matched with interventional distribution.
  - We may view this weighting as *switching a domain* from observational (source) to interventional (target) distribution (i.e., Domain adaptation).

- IPW/SW has been broadly employed/formalized in ML in the framework of "(weighted) empirical risk minimization (WERM)" (with many other names: transfer learning, domain adaptation, importance sampling, policy evaluation, counterfactual risk minimization, counterfactual regression)

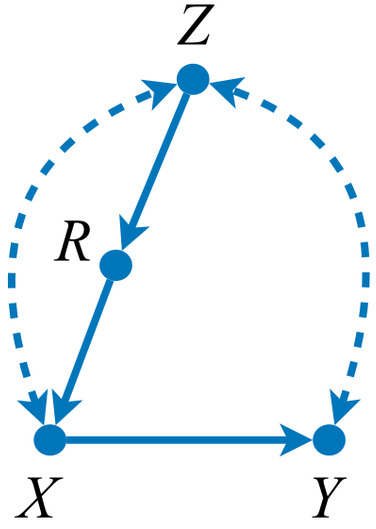
- However, there have been no works employing those recent ML frameworks to the causal effect estimation problem for general ID functionals.

Weighting based method in ML

- In BD setting, there is a weight that makes observational distribution (weighted distribution) matched with interventional distribution.
  - We may view this weighting as *switching a domain* from observational (source) to interventional (target) distribution (i.e., Domain adaptation).
- IPW/SW has been broadly employed/formalized in ML in the framework of "(weighted) empirical risk minimization (WERM)" (with many other names: transfer learning, domain adaptation, importance sampling, policy evaluation, counterfactual risk minimization, counterfactual regression)

Can we generalize this weighting-based method to estimate causal functionals beyond BD?

Napkin as domain adaptation



$$P_x(y) = \frac{\sum_z P(y, x \mid r, z)P(z)}{\sum_z P(x \mid r, z)P(z)}$$

Z is BD admissible w.r.t. $(R, \{X, Y\})$;

Z is BD admissible w.r.t. (R, X)

🤷 Is this a valid input for WERM?

Using SW in BD,

$$\sum_z P(y, x \mid r, z)P(z) = P^W(y, x \mid r) \text{ where } W = \frac{P(r)}{P(r \mid z)} \text{ and } P^W(y, x, r, z) = W \cdot P(y, x, r, z)$$

$$\sum_z P(x \mid r, z)P(z) = P^W(x \mid r) \text{ where } W = \frac{P(r)}{P(r \mid w)} \text{ and } P^W(y, x, r, z) = W \cdot P(y, x, r, z)$$

$$P_x(y) = \frac{P^W(y, x \mid r)}{P^W(x \mid r)} = P^W(y \mid x, r)$$

😄 Yes!

Problem setup

- In the BD setting, weighting-based ML methods (i.e., WERM) have been developed.

- We found a non-BD instance (Napkin, where the functional is given in a quotient form) that can be encoded as a valid input for WERM.

Can we generalize this weighting-based method to estimate causal functional beyond BD?

- Is there a systematic procedure to encode non-BD instance as a weighted distribution?

- Can we have a procedure of learning causal effect time-efficiently?

- Can we have learning guarantee and sample-complexity of the proposed method?

Weighted distribution

- Weighted distribution: Given a distribution $P(\mathbf{v})$ and a *weight function* $0 < W(\mathbf{v}) < \infty$ s.t. $\mathbb{E}_P[W(\mathbf{V})] = 1$ and $\mathbb{E}_P[W^2(\mathbf{V})] < \infty$, a *weighted distribution* $P^W(\mathbf{v})$ is given as

$$P^W(\mathbf{v}) \equiv W(\mathbf{v})P(\mathbf{v})$$

- Conditional weighted distribution: Given $P^W(\mathbf{v})$, a conditional distribution $P^W(\mathbf{a} \mid \mathbf{b})$ is induced as

$$P^W(\mathbf{a} \mid \mathbf{b}) = \frac{P^W(\mathbf{a}, \mathbf{b})}{P^W(\mathbf{b})} = \frac{\sum_{\mathbf{v} \setminus \{\mathbf{a},\mathbf{b}\}} P^W(\mathbf{v})}{\sum_{\mathbf{v} \setminus \mathbf{b}} P^W(\mathbf{v})}$$

Preliminary - mSBD

mSBD adjustment as a weighted distribution

mSBD adjustment: If $\mathbf{Z} = \{Z_1, \cdots, Z_n\}$ satisfies the mSBD criterion relative to $(\mathbf{X}, \mathbf{Y})$,

$$P(\mathbf{y} \mid do(\mathbf{x})) = \sum_{\mathbf{z}} \prod_{Y_i \in \mathbf{Y}} P(y_i \mid \mathbf{x}^{(i)}, \mathbf{z}^{(i)}, \mathbf{y}^{(i-1)}) \prod_{Z_i \in \mathbf{Z}} P(z_i \mid \mathbf{x}^{(i-1)}, \mathbf{z}^{(i-1)}, \mathbf{y}^{(i-1)})$$

mSBD adjustment can be encoded as a weighted distribution.

Let $W(\mathbf{Z}, \mathbf{X}) = \dfrac{P(\mathbf{X})}{\prod_{i=1}^{n} P(X_i \mid \mathbf{Z}^{(i)}, \mathbf{X}^{(i-1)}, \mathbf{Y}^{(i-1)})}$ be a weight function. Then,

$$P(\mathbf{y} \mid do(\mathbf{x})) = P^{W}(\mathbf{y} \mid \mathbf{x})$$

mSBD adjustment as a weighted distribution — Proof

$$P(\mathbf{y} \mid do(\mathbf{x})) = \sum_{\mathbf{z}} \prod_{Y_i \in \mathbf{Y}} P(y_i \mid \mathbf{x}^{(i)}, \mathbf{z}^{(i)}, \mathbf{y}^{(i-1)}) \prod_{Z_i \in \mathbf{Z}} P(z_i \mid \mathbf{x}^{(i-1)}, \mathbf{z}^{(i-1)}, \mathbf{y}^{(i-1)})$$

mSBD as a weighted distribution

$W(\mathbf{Z}, \mathbf{X}, \mathbf{Y}) = \dfrac{P(\mathbf{X})}{\prod_{i=1}^{n} P(X_i \mid \mathbf{Z}^{(i)}, \mathbf{X}^{(i-1)}, \mathbf{Y}^{(i-1)})}$. Then, $P(\mathbf{y} \mid do(\mathbf{x})) = P^W(\mathbf{y} \mid \mathbf{x})$

Let $q(X_i) \equiv P(X_i \mid \mathbf{Z}^{(i)}, \mathbf{X}^{(i-1)}, \mathbf{Y}^{(i-1)})$; $q(Z_i) = P(Z_i \mid \mathbf{X}^{(i-1)}, \mathbf{Z}^{(i-1)}, \mathbf{Y}^{(i-1)})$; $q(Y_i) = P(Y_i \mid \mathbf{X}^{(i)}, \mathbf{Z}^{(i)}, \mathbf{Y}^{(i-1)})$.

$P(\mathbf{Z}, \mathbf{X}, \mathbf{Y}) = \prod_{i=1}^{n} q(X_i) \prod_{i=1}^{n} q(Z_i) q(Y_i)$; $P(\mathbf{y} \mid do(\mathbf{x}) = \sum_{\mathbf{z}} \prod_{i=1}^{n} q(Z_i) q(Y_i)$; and

$W = P(\mathbf{X}) / \prod_{i=1}^{n} q(X_i)$. Then, $P^W(\mathbf{z}, \mathbf{x}, \mathbf{y}) = W \cdot P = P(\mathbf{x}) \prod_{i=1}^{n} q(Z_i) q(Y_i)$.

Since $P^W(\mathbf{y}, \mathbf{x}) = P(\mathbf{x}) \sum_{\mathbf{z}} \prod_{i=1}^{n} q(Z_i) q(Y_i)$; $P^W(\mathbf{x}) = P(\mathbf{x})$,

$$P^W(\mathbf{y} \mid \mathbf{x}) = P^W(\mathbf{y}, \mathbf{x}) / P^W(\mathbf{x}) = \sum_{\mathbf{z}} \prod_{i=1}^{n} q(Z_i) q(Y_i) = P(\mathbf{y} \mid do(\mathbf{x})).$$

mSBD adjustment as a weighted distribution — Proof

$$P(\mathbf{y} \mid do(\mathbf{x})) = \sum_{\mathbf{z}} \prod_{Y_i \in \mathbf{Y}} P(y_i \mid \mathbf{x}^{(i)}, \mathbf{z}^{(i)}, \mathbf{y}^{(i-1)}) \prod_{Z_i \in \mathbf{Z}} P(z_i \mid \mathbf{x}^{(i-1)}, \mathbf{z}^{(i-1)}, \mathbf{y}^{(i-1)})$$

mSBD as a weighted distribution

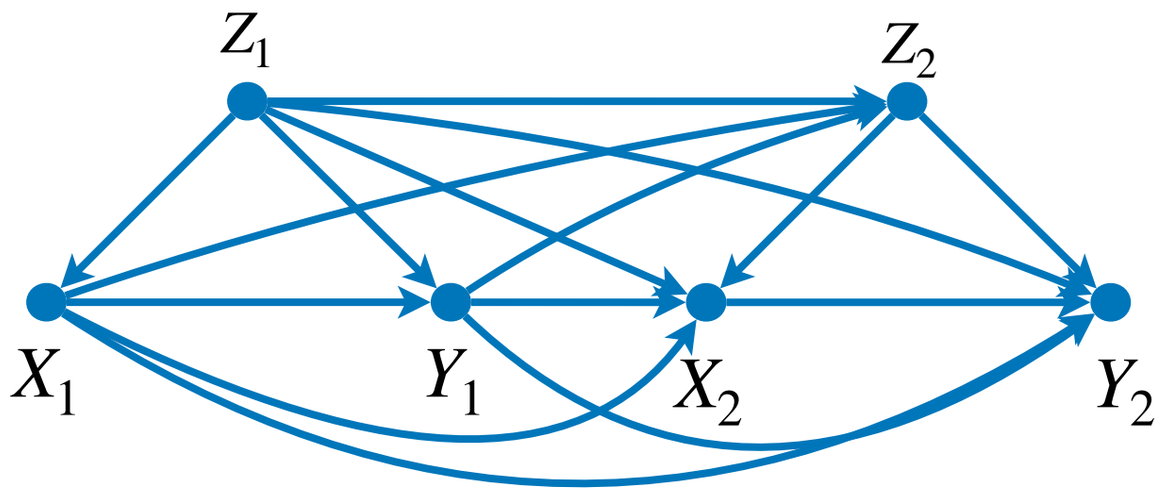
$W(\mathbf{Z}, \mathbf{X}, \mathbf{Y}) = \dfrac{P(\mathbf{X})}{\prod_{i=1}^{n} P(X_i \mid \mathbf{Z}^{(i)}, \mathbf{X}^{(i-1)}, \mathbf{Y}^{(i-1)})}$. Then, $P(\mathbf{y} \mid do(\mathbf{x})) = P^{W}(\mathbf{y} \mid \mathbf{x})$

Let $q(X_i) \equiv P(X_i \mid \mathbf{Z}^{(i)}, \mathbf{X}^{(i-1)}, \mathbf{Y}^{(i-1)})$; $q(Z_i) = P(Z_i \mid \mathbf{X}^{(i-1)}, \mathbf{Z}^{(i-1)}, \mathbf{Y}^{(i-1)})$; $q(Y_i) = P(Y_i \mid \mathbf{X}^{(i)}, \mathbf{Z}^{(i)}, \mathbf{Y}^{(i-1)})$.

$P(\mathbf{Z}, \mathbf{X}, \mathbf{Y}) = \prod_{i=1}^{n} q(X_i) \prod_{i=1}^{n} q(Z_i) q(Y_i)$; $P(\mathbf{y} \mid do(\mathbf{x}) = \sum_{\mathbf{z}} \prod_{i=1}^{n} q(Z_i) q(Y_i)$; and

Takeaway: mSBD adjustment can be represented as a weighted distribution.

mSBD adjustment as a weighted distribution — Example

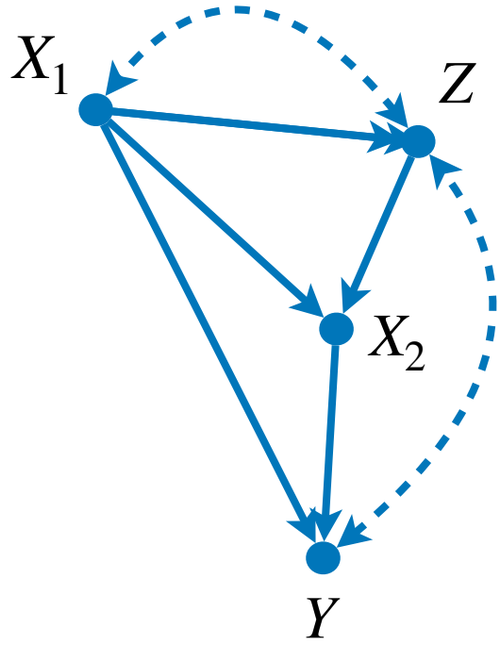


$\{Z_1, Z_2\}$ satisfies mSBD criterion relative to $(\{X_1, X_2\}, \{Y_1, Y_2\})$.

$$P_{x_1,x_2}(y_1, y_2) = \sum_{z_1,z_2} P(z_1)P(y_1 \mid x_1, z_1)P(z_2 \mid z_1, x_1, y_1)P(y_2 \mid z_1, x_1, y_1, z_2, x_2)$$

$$= P^W(y_1, y_2 \mid x_1, x_2), \text{ where } W = \frac{P(X_1, X_2)}{P(X_1 \mid Z_1)P(X_2 \mid Z_2, Y_1, X_1, Z_1)}$$

mSBD - example 2

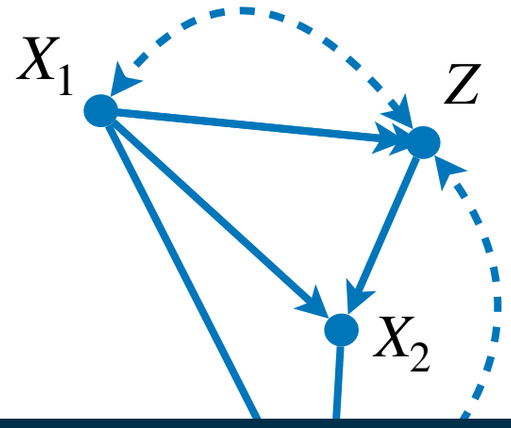


$Z = \{\emptyset, Z\}$ satisfies mSBD criterion relative to $(\{X_1, X_2\}, \{\emptyset, Y\}) = (\{X_1, X_2\}, Y)$.

$$P_{x_1,x_2}(y) = \sum_z P(y \mid x_1, x_2, z) P(z \mid x_1)$$

$$= P^W(y \mid x_1, x_2) \text{ where } W = \frac{P(X_1, X_2)}{P(X_1) P(X_2 \mid Z, X_1)}$$

mSBD - example 2



So far,

- mSBD (a sequence of BD) gives a graphical criterion for the case that there are no unmeasured confounders b/w $\mathbf{X}$ and $\mathbf{Y}$.
- We showed that mSBD adjustment can be encoded as a weighted distribution.

$$= P^W(y \mid x_1, x_2) \text{ where } W = \frac{P(X_1, X_2)}{P(X_1)P(X_2 \mid Z, X_1)}$$

General ID

Connection of C-factor & Weighted distribution

C-factor algebra as a weighted distribution

(Revisit) C-factor as mSBD adjustment

Let $\mathbf{C}$ be a C-component in G, $\mathbf{W}$ denote the ancestral set of $\mathbf{C}$ (i.e., $\mathbf{W} = An(\mathbf{W})_{G(\mathbf{C})}$) and $\mathbf{R} \equiv Pa(\mathbf{W})$. Then, $\mathbf{Z} = (\mathbf{S} \backslash \mathbf{W}) \cap An(\mathbf{R}, \mathbf{W})$ satisfies mSBD adjustment relative to $(\mathbf{R}, \mathbf{W})$, and

$$Q[\mathbf{W}] = P_{\mathbf{r}}(\mathbf{w}) = M[\mathbf{w} \mid \mathbf{r}; \mathbf{z}]$$

C-factor as weighted distribution

(Using the fact that mSBD can be represented as a weighted-distribution)

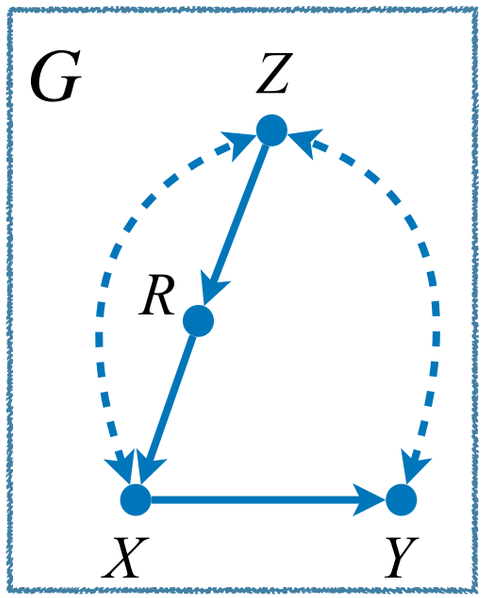
$Q[\mathbf{W}] = P_{\mathbf{r}}(\mathbf{w}) = P^{W}(\mathbf{w} \mid \mathbf{r})$ for the weight:

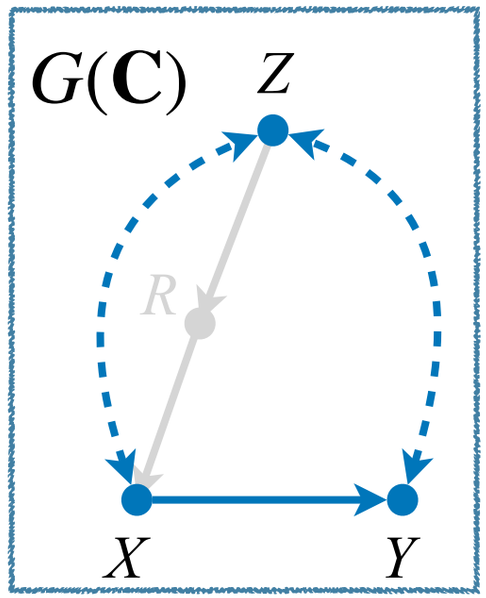
$$W = W(\mathbf{Z}, \mathbf{R}, \mathbf{W}) = \frac{P(\mathbf{R})}{\prod P(r_i \mid \mathbf{r}^{(i-1)}, \mathbf{z}^{(i)}, \mathbf{w}^{(i-1)})}$$

C-factor algebra as a weighted distribution: Example

Let $\mathbf{C}$ be a C-component in G, $\mathbf{W}$ denote the ancestral set of $\mathbf{C}$ (i.e., $\mathbf{W} = An(\mathbf{W})_{G(\mathbf{C})}$) and $\mathbf{R} \equiv Pa(\mathbf{W})$. Then, $\mathbf{Z} = (\mathbf{S} \backslash \mathbf{W}) \cap An(\mathbf{R}, \mathbf{W})$ satisfies mSBD adjustment relative to $(\mathbf{R}, \mathbf{W})$, and

$$Q[\mathbf{W}] = P_{\mathbf{r}}(\mathbf{w}) = P^{W}(\mathbf{w} \,|\, \mathbf{r}) \text{ for the weight } W = \frac{P(\mathbf{R})}{\prod P(r_i \,|\, \mathbf{r}^{(i-1)}, \mathbf{z}^{(i)}, \mathbf{w}^{(i-1)})}$$





$$\mathbf{C} = \{Z, X, Y\}$$

$\mathbf{W} = \{X, Y\}$ is an ancestral set in $G(\mathbf{C})$.

$\mathbf{R} = Pa(\mathbf{W}) = \{R\}$

$\mathbf{Z} = (\mathbf{C} \backslash \mathbf{W}) \cap An(\mathbf{R}, \mathbf{W}) = \{Z\}$ satisfies mSBD adjustment relative to $(\mathbf{R}, \mathbf{W})$.

$$Q[\mathbf{W}] = Q[X, Y] = P^{W}(x, y \,|\, r), \quad W = \frac{P(r)}{P(r \,|\, z)}$$

Connection to general ID

- (So far) For a C-component in G (denoted $\mathbf{C}$) and its ancestral set $\mathbf{W}$, a corresponding C-factor $Q[\mathbf{W}]$ can be encoded as a *weighted distribution* P^W by weighting P.

- (Recursive operation in ID algo.) ID algo. runs a C-factor algebra for an arbitrary subgraph $G(\mathbf{S})$ with the corresponding C-factor $Q[\mathbf{S}]$.

- We will extend the weighting method to encode $Q[\mathbf{W}]$ defined in an arbitrary subgraph $G(\mathbf{S})$ as a weighted distribution.

Weighted distribution algebra - Recursive operation

Let $\mathbf{C}$ be a C-component in G, $\mathbf{W}$ denote the ancestral set of $\mathbf{C}$ (i.e., $\mathbf{W} = An(\mathbf{W})_{G(\mathbf{C})}$) and $\mathbf{R} \equiv Pa(\mathbf{W})$. Then, $\mathbf{Z} = (\mathbf{S} \backslash \mathbf{W}) \cap An(\mathbf{R}, \mathbf{W})$ satisfies mSBD adjustment relative to $(\mathbf{R}, \mathbf{W})$:

$$Q[\mathbf{W}] = P_{\mathbf{r}}(\mathbf{w}) = P^{W}(\mathbf{w} \,|\, \mathbf{r}) \text{ for the weight } W = \frac{P(\mathbf{r})}{\prod P(r_i \,|\, \mathbf{r}^{(i-1)}, \mathbf{z}^{(i)}, \mathbf{w}^{(i-1)})}$$

Weighed distribution algebra

Suppose $Q[\mathbf{S}] = P^{W}(\mathbf{s} \,|\, \mathbf{r})$ for some known $\{W, \mathbf{R}\}$. Let $\mathbf{A} \subseteq \mathbf{S}$.

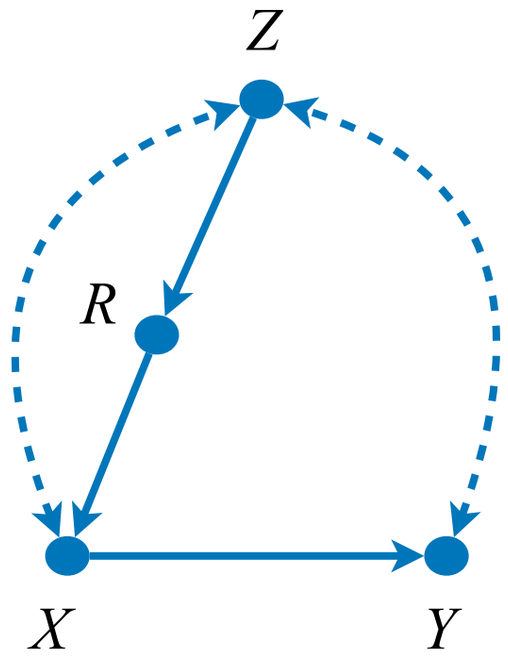
- If $\mathbf{A}$ is an ancestral set in $G(\mathbf{S})$, then $Q[\mathbf{A}] = P^{W}(\mathbf{a} \,|\, \mathbf{r})$.
- If $\mathbf{A}$ is a C-component of $G(\mathbf{S})$, then $Q[\mathbf{A}] = P^{W \times W'}(\mathbf{a} \,|\, \mathbf{r}')$ where
  - $\mathbf{R}' = \mathbf{R} \cup \mathbf{R}''$ where $\mathbf{R}'' \equiv \{\text{Variables in } \mathbf{S} \backslash \mathbf{A} \text{ that is an ancestor of } \mathbf{A}\}$; and
  - $W' = \dfrac{P^{W}(\mathbf{r}'' \,|\, \mathbf{r})}{\prod_{R_i'' \in \mathbf{R}''} P^{W}(r_i'' \,|\, \mathbf{r}, Pre_{\mathbf{A}}(r_i''))}$

C-factor as a weighted distribution: Example

C-factor algebra

Suppose $Q[\mathbf{S}] = P^W(\mathbf{s}\,|\,\mathbf{r})$. If $\mathbf{A}$ is an ancestral set in $G(\mathbf{S})$, then $Q[\mathbf{A}] = P^W(\mathbf{a}\,|\,\mathbf{r})$. If $\mathbf{A}$ is a C-component of $G(\mathbf{S})$, then $Q[\mathbf{A}] = P^{W\times W'}(\mathbf{a}\,|\,\mathbf{r}')$ where

$\mathbf{R}' = \mathbf{R} \cup \mathbf{R}''$ where $\mathbf{R}'' \equiv \{\text{Variables in } \mathbf{S}\backslash\mathbf{A} \text{ that is an ancestor of } \mathbf{A}\}$; and $W' = \dfrac{P^W(\mathbf{r}''\,|\,\mathbf{r})}{\prod_{R_i''\in\mathbf{R}''} P^W(r_i''\,|\,\mathbf{r}, Pre_{\mathbf{S}}(r_i''))}$



$\mathbf{C} = \{Z, X, Y\}$

$\mathbf{W} = \{X, Y\}$ is an ancestral set in $G(\mathbf{C})$.

$\mathbf{R} = Pa(\mathbf{W}) = \{R\}$

$\mathbf{Z} = (\mathbf{C}\backslash\mathbf{W}) \cap An(\mathbf{R}, \mathbf{W}) = \{Z\}$ satisfies mSBD adjustment relative to $(\mathbf{R}, \mathbf{W})$.

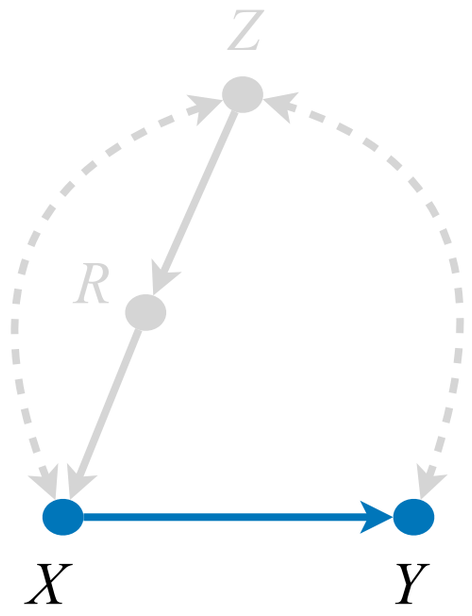
$$Q[\mathbf{W}] = Q[X, Y] = P^W(x, y\,|\,r),\ W = \frac{P(r)}{P(r\,|\,z)}$$

C-factor as a weighted distribution: Example

C-factor algebra

Suppose $Q[\mathbf{S}] = P^W(\mathbf{s}\,|\,\mathbf{r})$. If $\mathbf{A}$ is an ancestral set in $G(\mathbf{S})$, then $Q[\mathbf{A}] = P^W(\mathbf{a}\,|\,\mathbf{r})$. If $\mathbf{A}$ is a C-component of $G(\mathbf{S})$, then $Q[\mathbf{A}] = P^{W\times W'}(\mathbf{a}\,|\,\mathbf{r}')$ where

$\mathbf{R}' = \mathbf{R} \cup \mathbf{R}''$ where $\mathbf{R}'' \equiv \{\text{Variables in } \mathbf{S}\backslash\mathbf{A} \text{ that is an ancestor of } \mathbf{A}\}$; and $W' = \dfrac{P^W(\mathbf{r}'\,|\,\mathbf{r})}{\prod_{R'_i\in\mathbf{R}''} P^W(r'_i\,|\,\mathbf{r}, Pre_{\mathbf{S}}(r'_i))}$



$\mathbf{W} = \mathbf{S} = \{X, Y\}$

$$Q[\mathbf{S}] = Q[X, Y] = P^W(x, y\,|\,r), \quad W = \frac{P(r)}{P(r\,|\,z)}$$

$\mathbf{A} = \{Y\}$ is a C-component in $G(\mathbf{S})$.

$$\mathbf{R}'' = \{X\} \text{ and } \mathbf{R}' = \{R, X\}. \quad W' = \frac{P^W(x\,|\,r)}{P^W(x\,|\,r)} = 1.$$

$$Q[Y] = P^{W\times W'}(y\,|\,\mathbf{r}') = P^W(y\,|\,r, x).$$

wID: Representing a causal effect as a weighted distribution

wID$(\mathbf{X}, \mathbf{Y}, G)$

1. Let $\mathbf{S}_1, \mathbf{S}_2, \ldots$ be the C-components of G.
2. Let $Q[\mathbf{S}_i] = P^{W_{\mathbf{S}_i}}(\mathbf{s}_i \mid \mathbf{r}_{\mathbf{S}_i})$ where $\mathbf{R}_{\mathbf{S}_i} \equiv Pa(\mathbf{S}_i)$
3. Let $\mathbf{D}_1, \mathbf{D}_2, \ldots$ be the c-components of $G[\mathbf{D}]$ where $\mathbf{D} = An(\mathbf{Y})_{G(\mathbf{V}\setminus\mathbf{X})}$.
4. $Q[\mathbf{D}_j] = P^{W_{\mathbf{D}_i}}(\mathbf{d}_i \mid \mathbf{r}_{\mathbf{D}_i}) =$ wIdentify$(\mathbf{D}_j, \mathbf{S}_j, Q[\mathbf{S}_j])$
5. If there is only one C-component in $G(\mathbf{D})$, then return $P^{W_{\mathbf{D}_1}}(\mathbf{y} \mid \mathbf{r}_{\mathbf{D}_1})$.
   Otherwise, set $W = \prod_i P^{W_{\mathbf{D}_i}}(\mathbf{d}_i \mid \mathbf{r}_{\mathbf{d}_i}) / P(\mathbf{d} \mid \mathbf{r})$ and return $P^{W}(\mathbf{y} \mid \mathbf{v}\setminus\mathbf{d})$.

Representation of $Q[\mathbf{D}_j]$ as a weighted distribution, by recursively applying weighting operation for C-factor.

wID is sound & complete — Any ID functional can be represented as as weighted distribution.

wID: Representing a causal effect as a weighted distribution

wID$(\mathbf{X}, \mathbf{Y}, G)$

1. Let $\mathbf{S}_1, \mathbf{S}_2, \ldots$ be the C-components of G.
2. Let $Q[\mathbf{S}_i] = P^{W_{\mathbf{S}_i}}(\mathbf{s}_i \mid \mathbf{r}_{\mathbf{S}_i})$ where $\mathbf{R}_{\mathbf{S}_i} \equiv Pa(\mathbf{S}_i)$
3. Let $\mathbf{D}_1, \mathbf{D}_2, \ldots$ be the c-components of $G[\mathbf{D}]$ where $\mathbf{D} = An(\mathbf{Y})_{G(\mathbf{V}\setminus\mathbf{X})}$.
4. $Q[\mathbf{D}_j] = P^{W_{\mathbf{D}_i}}(\mathbf{d}_j \mid \mathbf{r}_{\mathbf{D}_j}) =$ wIdentify$(\mathbf{D}_j, \mathbf{S}_j, Q[\mathbf{S}_j])$

Representation of $Q[\mathbf{D}_j]$ as a weighted distribution, by recursively applying weighting operation for C-factor.

- **So far**, We have a sound & complete algorithm for representing any identifiable functional as a weighted distribution.
- **Now**, We will provide a learning framework, given the weighted distribution.

Learning procedure & guarantee

Learning Framework — Weighted Empirical Risk Minimization

- So far, we learned how to derive $(W^*, \mathbf{r})$ s.t. $P(\mathbf{y} \mid do(\mathbf{x})) = P^{W^*}(\mathbf{y} \mid \mathbf{r})$.

We use W^* as a true weight s.t. $P(\mathbf{y} \mid do(\mathbf{x})) = P^{W^*}(\mathbf{y} \mid \mathbf{r})$; and W for an arbitrary weight function.

Weighted Empirical Risk Minimization (WERM)

- Given a loss function $\ell(y, y')$ (e.g., $\ell(y, y') = (y - y')^2$), $h(\mathbf{r}) \in \mathscr{H}$ is a hypothesis for estimating $\mathbf{Y}$. (h can be viewed as an estimate for $P^{W^*}(\mathbf{y} \mid \mathbf{r})$)
- We attempt to learn h that minimizes an expected loss (a.k.a. risk) on the *weighted domain*:

$$R^{W^*}(h) \equiv \underbrace{\mathbb{E}_{PW^*}}_{= W^* \cdot P}\left[\ell(h(\mathbf{R}), \mathbf{Y})\right] = \mathbb{E}_P\left[W^*(\mathbf{V})\ell(h(\mathbf{R}), \mathbf{Y})\right].$$

- For (W, h), let an *Weighted Empirical Risk (WER)* using samples D is

$$\hat{R}^W(h) = \mathbb{E}_D\left[W(\mathbf{V})\ell(h(\mathbf{R}), \mathbf{Y})\right].$$

Learning Framework — Weighted Empirical Risk Minimization

- We attempt to learn h that approximates $\mathbb{E}_{PW^*}[Y|\mathbf{r}]$ by minimizing

$$R^{W*}(h) \equiv \mathbb{E}_{PW^*}\left[\ell(h(\mathbf{R}), \mathbf{Y})\right] = \mathbb{E}_P\left[W*(\mathbf{V})\ell(h(\mathbf{R}), \mathbf{Y})\right]$$

- For $(\dot{W}, h)$, let an *Weighted Empirical Risk (WER)* be denoted as

$$\hat{R}^W(h) = \mathbb{E}_D\left[W(\mathbf{V})\ell(h(\mathbf{R}), \mathbf{Y})\right].$$

- $W*$ (e.g., $W* = P(X)/P(X|Z)$ in BD example) is not accessible, since we are only given a sample D, not a distribution P itself.
- Using the direct estimates $\widehat{W*}$ (e.g., $\widehat{W*} = \hat{P}(X)/\hat{P}(X|Z)$ in BD example) is not desirable, since it might yield high variance.

Learning Framework — Weighted Empirical Risk Minimization

- We attempt to learn h that approximates $\mathbb{E}_{PW^*}[Y \mid \mathbf{r}]$ by minimizing

$$R^{W*}(h) \equiv \mathbb{E}_{PW^*}\left[\ell(h(\mathbf{R}), \mathbf{Y})\right] = \mathbb{E}_P\left[W^*(\mathbf{V})\ell(h(\mathbf{R}), \mathbf{Y})\right]$$

- For $(\dot{W}, h)$, let an *Weighted Empirical Risk (WER)* be denoted as

$$\hat{R}^W(h) = \mathbb{E}_D\left[W(\mathbf{V})\ell(h(\mathbf{R}), \mathbf{Y})\right].$$

- W^* (e.g., $W^* = P(X)/P(X \mid Z)$ in BD example) is not accessible, since we are only given a sample D, not a distribution P itself.

How do we mitigate this issue?

Learning Framework — Weighted Empirical Risk Minimization

- We attempt to learn h that approximates $\mathbb{E}_{PW^*}[Y|\mathbf{r}]$ by minimizing

$$R^{W*}(h) \equiv \mathbb{E}_{PW^*}\left[\ell(h(\mathbf{R}), \mathbf{Y})\right] = \mathbb{E}_P\left[W^*(\mathbf{V})\ell(h(\mathbf{R}), \mathbf{Y})\right]$$

- For $(\dot{W}, h)$, let an *Weighted Empirical Risk (WER)* be denoted as

$$\hat{R}^W(h) = \mathbb{E}_D\left[W(\mathbf{V})\ell(h(\mathbf{R}), \mathbf{Y})\right].$$

  - We will use other W as a proxy of W^*, which is estimable from given dataset, and expected to yield the lower variance.

- We use W, instead of W^*. We use a finite sample D, instead of a distribution P.
- What will be a gap b/w performances of using (W^*, P) vs. (W, D)?

$R^{W*}(h)$ $\hat{R}^W(h)$

Learning Framework — Weighted Empirical Risk Minimization

- We use W, instead of W^*. We use a finite sample D, instead of a distribution P.
- What will be a gap b/w performances of using (W^*, P) vs. (W, D)?

$R^{W^*}(h)$ $\hat{R}^{W}(h)$

$$\left| R^{W^*}(h) - \hat{R}^{W}(h) \right| \leq (a) + (b) \cdot (c) \text{ in high probability } (1-\delta)^{[1]}$$

- $(a) = \mathbb{E}_P \left[\left| W^*(\mathbf{V}) - W(\mathbf{V}) \right| \right]$ A gap b/w two weights
- $(b) = c \max \left(\sqrt{\mathbb{E}_P \left[W^2 \ell_h^2 \right]}, \sqrt{\mathbb{E}_D \left[W^2 \ell_h^2 \right]} \right)$ where $\ell_h \equiv \ell(h(\mathbf{V}), \mathbf{Y})$

  Second moment of W
- $(c) = F(p, N, \delta)$, a function of a complexity (p) of the hypothesis class of $h \in \mathscr{H}$; a sample size N; and a probability δ. Complexity of hypothesis class.

Learning Framework — Weighted Empirical Risk Minimization

$$\left| R^{W*}(h) - \hat{R}^{W}(h) \right| \leq (a) + (b) \cdot (c) \text{ in high probability } (1-\delta)$$

- $(a) = \mathbb{E}_P \left[\left| W^*(\mathbf{V}) - W(\mathbf{V}) \right| \right]$
- $(b) = c \max \left(\sqrt{\mathbb{E}_P \left[W^2 \ell_h^2 \right]}, \sqrt{\mathbb{E}_D \left[W^2 \ell_h^2 \right]} \right)$ where $\ell_h \equiv \ell(h(\mathbf{V}), \mathbf{Y})$

This motivates to find W such that (a) close to the true W^*; while (b) having lower variance.

A gap b/w (W^*, P) vs. (W, D) are controlled by

(a) A gap b/w two weights

(b) The variance term (second moment) of W;

(c) Complexity of hypothesis class $h \in \mathscr{H}$.

Learning objective

$$\left| R^{W*}(h) - \hat{R}^{W}(h) \right| \leq (a) + (b) \cdot (c) \text{ in high probability } (1 - \delta)$$

$(a) = \mathbb{E}_P \left[\left| W^*(\mathbf{V}) - W(\mathbf{V}) \right| \right]$ A gap b/w two weights

$(b) = c \max \left(\sqrt{\mathbb{E}_P \left[W^2 \ell_h^2 \right]}, \sqrt{\mathbb{E}_D \left[W^2 \ell_h^2 \right]} \right)$ The variance term (second moment) of W;

$(c) = F(p, N, \delta)$ Complexity of a function class for h

Learning objective

- Note $R^{W*}(h) \leq \hat{R}^{W}(h) + (a) + (b) \cdot (c)$. We use this upper bound as a proxy of the true risk $R^{W*}(h)$. Principle of structural risk minimization (Vapnik)
- To minimize this proxy, a learning objective is given as:

$$L\left(W, h\right) = \hat{R}^{W}(h) + (\lambda_h / N) C(h) + \sqrt{\mathbb{E}_D \left[(W(\mathbf{V}) - W^*(\mathbf{V}))^2 \right]} + (\lambda_W / N) \|W\|^2$$

Empirical risk at (W, h) | Penalty on complexity (c) | Penalty on a gap b/w weights (a) | Penalty on 2nd moment of W (b)

Learning guarantee

Learning objective

$$L\left(W,h\right)=\hat{R}^{W}(h)+(\lambda_h/N)C(h)+\sqrt{\mathbb{E}_D\left[(W(\mathbf{V})-W^*(\mathbf{V}))^2\right]+(\lambda_W/N)\|W\|^2}$$

Empirical risk at (W,h)

Penalty on complexity (c)

Penalty on a gap b/w weights (a)

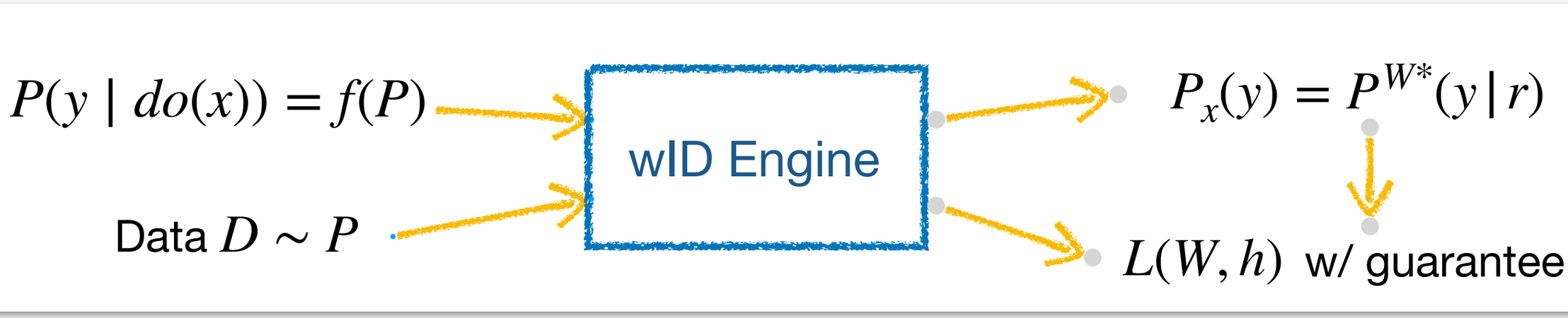
Penalty on 2nd moment of W (b)

Learning guarantee

For a minimizer $(W_N,h_N)\equiv\arg\min_{W,h}L(W,h)$, h_N converges to $h^*=\arg\min R^{W^*}(h)$ under a correct choice of hypothesis class, at a rate $N^{-1/4}$.

$\Rightarrow$ By minimizing $L(W,h)$, we can obtain an asymptotic true minimizer for $R^{W^*}(h)$.

Learning procedure: WERM-ID

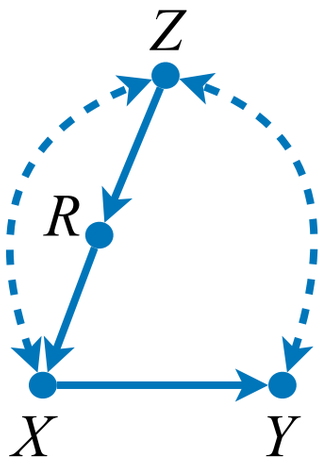


$$L\left(W,h\right) = \underbrace{\hat{R}^W(h) + (\lambda_h/N)C(h)}_{= L_h(h; W, \lambda_h)} + \underbrace{\sqrt{\mathbb{E}_D\left[(W(\mathbf{V}) - W^*(\mathbf{V}))^2\right] + (\lambda_W/N)\|W\|^2}}_{= L_W(W; W^*, \lambda_W)}$$

A procedure for solving $L(W, h)$

1. Evaluate $\widehat{W^*}$ using D.
2. Learn $W = \arg\min_{W'} L_W(W'; \widehat{W^*}, \lambda_W)$
3. For W in step 2, learn $h = \arg\min L_h(h; W, \lambda_h)$

Example: Napkin



$$P_x(y) = \frac{\sum_z P(y, x \mid r, w) P(z)}{\sum_z P(x \mid r, z) P(z)} = P^{W*}(y \mid x, r) \text{ for } W* = \frac{P(r)}{P(r \mid z)}.$$

$$L\left(W, h\right) = \underbrace{\hat{R}^W(h) + (\lambda_h / N) C(h)}_{= L_h(h; W, \lambda_h)} + \underbrace{\sqrt{\mathbb{E}_D\left[(W(\mathbf{V}) - W*(\mathbf{V}))^2\right] + (\lambda_W / N)\|W\|^2}}_{= L_W(W; W*, \lambda_W)}$$

1. Evaluate $\widehat{W*} = \hat{P}(r)/\hat{P}(r \mid z)$ using D.
2. Learn $W = W(R, W) = \arg\min_{W' \in \mathcal{H}_W} \mathbb{E}_D \left[(W'(R, W) - \widehat{W*}(R, W))^2 + (\lambda_W / N)\|W\|^2\right]$.
3. Given W, $h = \arg\min_{h \in \mathcal{H}} - \mathbb{E}_D \left[W(R, W) \cdot \left\{Y \log h' + (1 - Y) \log(1 - h')\right\} + (\lambda_h / N) C(h')\right]$.

Cross-entropy loss (where $h \in (0,1)$)

Time complexity for running WERM-ID

1. Evaluate $\widehat{W^*}$ using D.
2. Learn $W = \arg\min_{W'} L_W(W'; \widehat{W^*}, \lambda_W)$
3. For W in step 2, learn $h = \arg\min L_h(h; W, \lambda_h)$

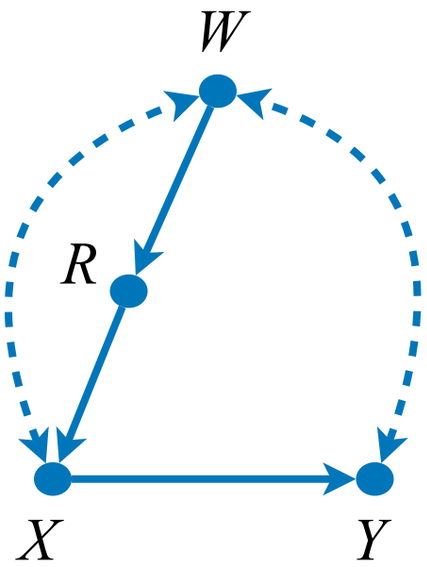
Time complexity

Let $N = |D|$ and $n = |\mathbf{V}|$. Let $T_1(N)$ denote a complexity for estimating $\hat{P}(a|b)$ from sample D. Let $T_2(N)$ denote a time for solving objectives L_W and L_h. Then, the time complexity for learning a causal effect via WERM-ID is

WERM-ID runs in polynomial w.r.t. (N, n) — Any ID functional can be estimated through WERM-ID framework in polynomial time.

Simulation

Simulation results



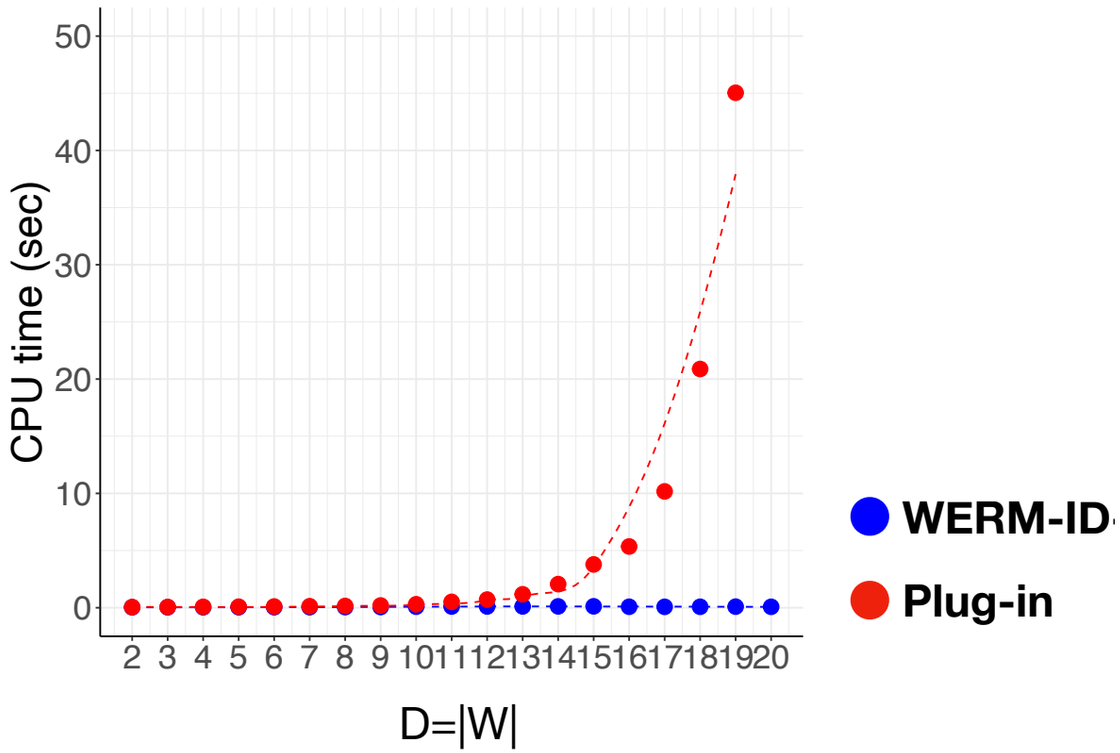
$$P_x(y) = f(P) = \frac{\sum_w P(x, y \mid r, w)P(w)}{\sum_w P(x \mid r, w)P(w)}$$

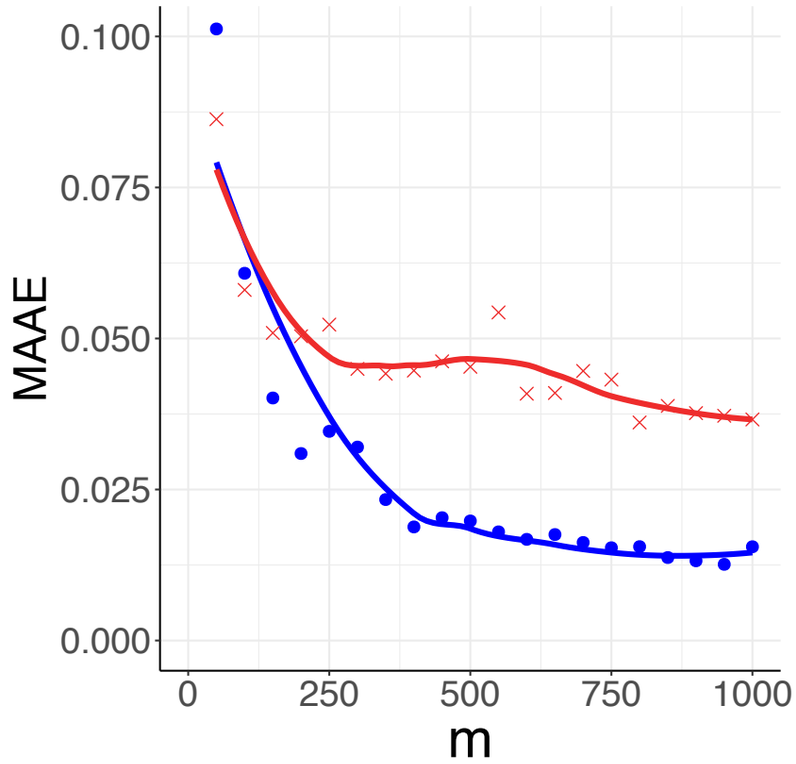
$$f(\hat{P}) = \frac{\sum_w \hat{P}(x, y \mid r, w)\hat{P}(w)}{\sum_w \hat{P}(x \mid r, w)\hat{P}(w)}$$

WERM-ID runs in polynomial w.r.t. (N, n) — Any ID functional can be estimated through WERM-ID framework in polynomial time.

- To verify this, we compare with the plug-in estimator with varying the dimension for the variable W.

Simulation results





- **(**Time complexity**;** Left**)** WERM-ID runs in polynomial time w.r.t. the dimension of variables, while the plug-in in exponential time.

- **(**Sample complexity; (Center, Right)**)** A simulation provides an empirical result that WERM-ID performs better than the plug-in given finite samples.

Conclusions

We develop time & sample efficient estimator working for any identifiable causal functional based on weighted ERM theory.

- (wID) We devised wID algorithm to represent any identifiable causal functional as a weighted distribution
- (WERM-ID) We introduce WERM based learning guarantee & procedure to learn a causal effect.

Reference

1. Cortes, Corinna, Yishay Mansour, and Mehryar Mohri. "Learning bounds for importance weighting." Proceedings of the 23rd International Conference on Neural Information Processing Systems-Volume 1. 2010.

Thank you!