Estimation

Brady Neal

causalcourse.com

Estimation portion of the flowchart

Preliminaries

Conditional average treatment effects (CATEs):

$$\tau(x) \triangleq \mathbb{E}[Y(1) - Y(0) \mid X = x]$$

Preliminaries

Conditional average treatment effects (CATEs):

$$\tau(x) \triangleq \mathbb{E}[Y(1) - Y(0) \mid X = x]$$

Always assuming unconfoundedness and positivity

Preliminaries

Conditional average treatment effects (CATEs):

$$\tau(x) \triangleq \mathbb{E}[Y(1) - Y(0) \mid X = x]$$

Always assuming unconfoundedness and positivity

$$\tau \triangleq \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}_W \left[\mathbb{E}[Y \mid T = 1, W] - \mathbb{E}[Y \mid T = 0, W]\right]$$

Given W is a sufficient adjustment set

Preliminaries

Conditional average treatment effects (CATEs):

$$\tau(x) \triangleq \mathbb{E}[Y(1) - Y(0) \mid X = x]$$

Always assuming unconfoundedness and positivity

$$\tau \triangleq \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}_W \left[\mathbb{E}[Y \mid T = 1, W] - \mathbb{E}[Y \mid T = 0, W]\right]$$

Given W is a sufficient adjustment set

$$\tau(x) \triangleq \mathbb{E}[Y(1) - Y(0) \mid X = x] = \mathbb{E}_W[\mathbb{E}[Y \mid T = 1, X = x, W] - \mathbb{E}[Y \mid T = 0, X = x, W]]$$

Given $W \cup X$ is a sufficient adjustment set

Conditional Outcome Modeling

Increasing Data Efficiency

Propensity Scores and IPW

Other Methods

Conditional Outcome Modeling

Increasing Data Efficiency

Propensity Scores and IPW

Other Methods

Conditional outcome modeling (COM)

$$\tau = \mathbb{E}_W \left[\mathbb{E}[Y \mid T = 1, W] - \mathbb{E}[Y \mid T = 0, W] \right]$$

Conditional outcome modeling (COM)

$$\tau = \mathbb{E}_W \left[\underbrace{\mathbb{E}[Y \mid T = 1, W]}_{\text{model}} - \underbrace{\mathbb{E}[Y \mid T = 0, W]}_{\text{model}} \right]$$

Conditional outcome modeling (COM)

$$\tau = \mathbb{E}_W \left[\underbrace{\mathbb{E}[Y \mid T = 1, W]}_{\text{model}} - \underbrace{\mathbb{E}[Y \mid T = 0, W]}_{\text{model}} \right]$$

Conditional outcome modeling (COM)

$$\tau = \mathbb{E}_W \left[\mathbb{E}[Y \mid T = 1, W] - \mathbb{E}[Y \mid T = 0, W] \right]$$

$$\tau = \mathbb{E}_W \left[\overbrace{\mu(1, W)}^{\text{model}} - \overbrace{\mu(0, W)}^{\text{model}} \right]$$

Conditional outcome modeling (COM)

$$\tau = \mathbb{E}_W \left[\mathbb{E}[Y \mid T = 1, W] - \mathbb{E}[Y \mid T = 0, W] \right]$$

$$\tau = \mathbb{E}_W \left[\overbrace{\mu(1, W)}^{\text{model}} - \overbrace{\mu(0, W)}^{\text{model}} \right]$$

Model-assisted estimator:

$$\hat{\tau} = \frac{1}{n} \sum_i \left(\hat{\mu}(1, w_i) - \hat{\mu}(0, w_i) \right)$$

COM estimation of CATEs

ATE COM Estimator: $\hat{\tau} = \frac{1}{n} \sum_i (\hat{\mu}(1, w_i) - \hat{\mu}(0, w_i))$

COM estimation of CATEs

ATE COM Estimator: $\hat{\tau} = \frac{1}{n} \sum_i (\hat{\mu}(1, w_i) - \hat{\mu}(0, w_i))$

CATE Estimand:

$$\tau(x) \triangleq \mathbb{E}[Y(1) - Y(0) \mid X = x] = \mathbb{E}_W[\mathbb{E}[Y \mid T = 1, X = x, W] - \mathbb{E}[Y \mid T = 0, X = x, W]]$$

COM estimation of CATEs

ATE COM Estimator: $\hat{\tau} = \frac{1}{n} \sum_i (\hat{\mu}(1, w_i) - \hat{\mu}(0, w_i))$

CATE Estimand:

$$\tau(x) \triangleq \mathbb{E}[Y(1) - Y(0) \mid X = x] = \mathbb{E}_W[\mathbb{E}[Y \mid T = 1, X = x, W] - \mathbb{E}[Y \mid T = 0, X = x, W]]$$

$$\mu(t, w, x) \triangleq \mathbb{E}[Y \mid T = t, W = w, X = x]$$

COM estimation of CATEs

ATE COM Estimator: $\hat{\tau} = \frac{1}{n} \sum_i (\hat{\mu}(1, w_i) - \hat{\mu}(0, w_i))$

CATE Estimand:

$$\tau(x) \triangleq \mathbb{E}[Y(1) - Y(0) \mid X = x] = \mathbb{E}_W[\mathbb{E}[Y \mid T = 1, X = x, W] - \mathbb{E}[Y \mid T = 0, X = x, W]]$$

$$\mu(t, w, x) \triangleq \mathbb{E}[Y \mid T = t, W = w, X = x]$$

CATE COM Estimator: $\hat{\tau}(x) = \frac{1}{n_x} \sum_{i:x_i=x} (\hat{\mu}(1, w_i, x) - \hat{\mu}(0, w_i, x))$

$$\hat{\tau}_i = \hat{\tau}(x_i) = \hat{\mu}(1, w_i, x_i) - \hat{\mu}(0, w_i, x_i)$$

Question:
What could go wrong with this estimator?

COM estimation's many faces

COM estimation's many faces

- G-computation estimators

COM estimation's many faces

- G-computation estimators

COM estimation's many faces

- G-computation estimators
- Parametric G-formula

COM estimation's many faces

- G-computation estimators
- Parametric G-formula

COM estimation's many faces

- G-computation estimators
- Parametric G-formula
- Standardization

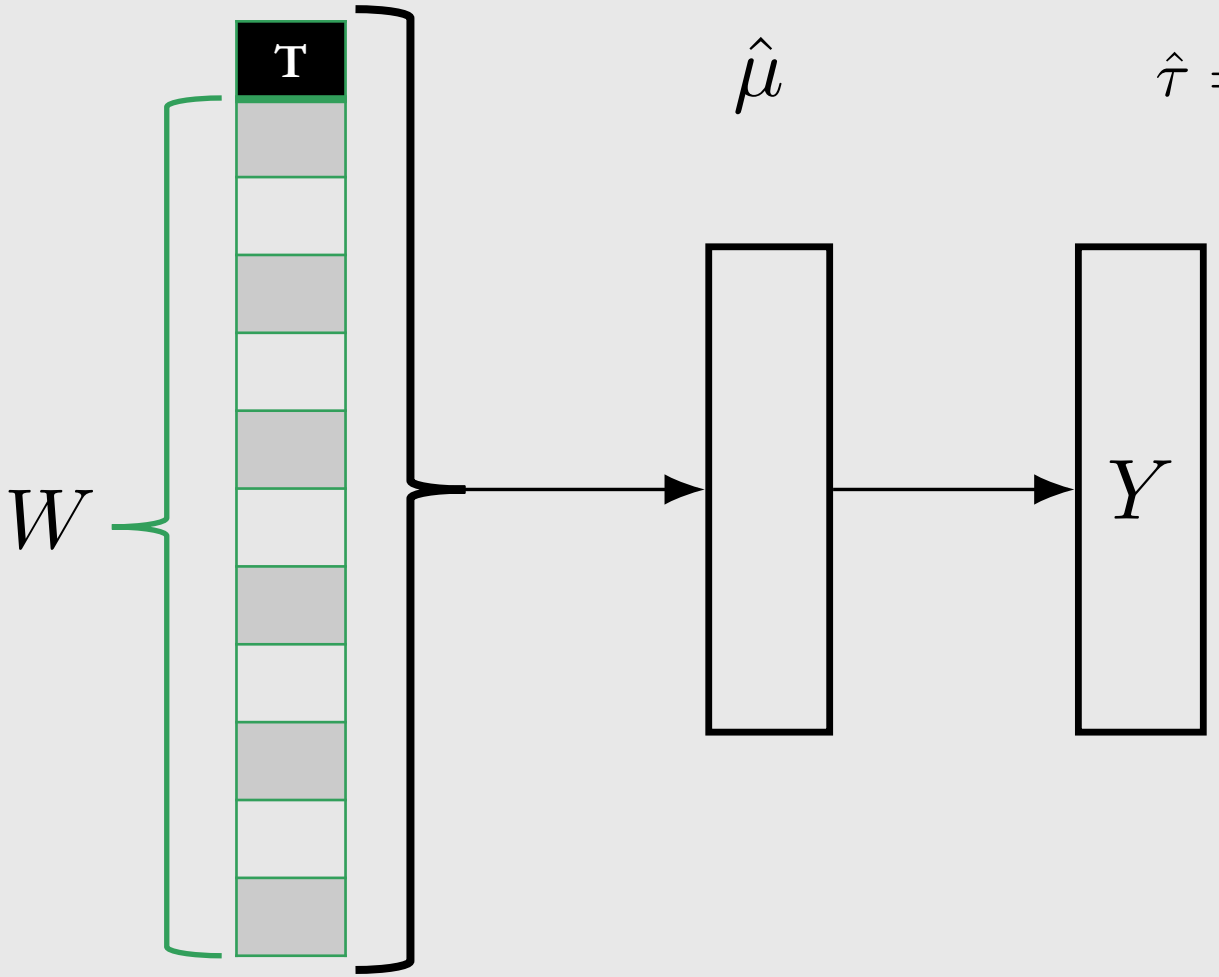
COM estimation's many faces

- G-computation estimators
- Parametric G-formula
- Standardization
- S-learner where "S" is for "Single"

Problem with COM estimation in high dimensions

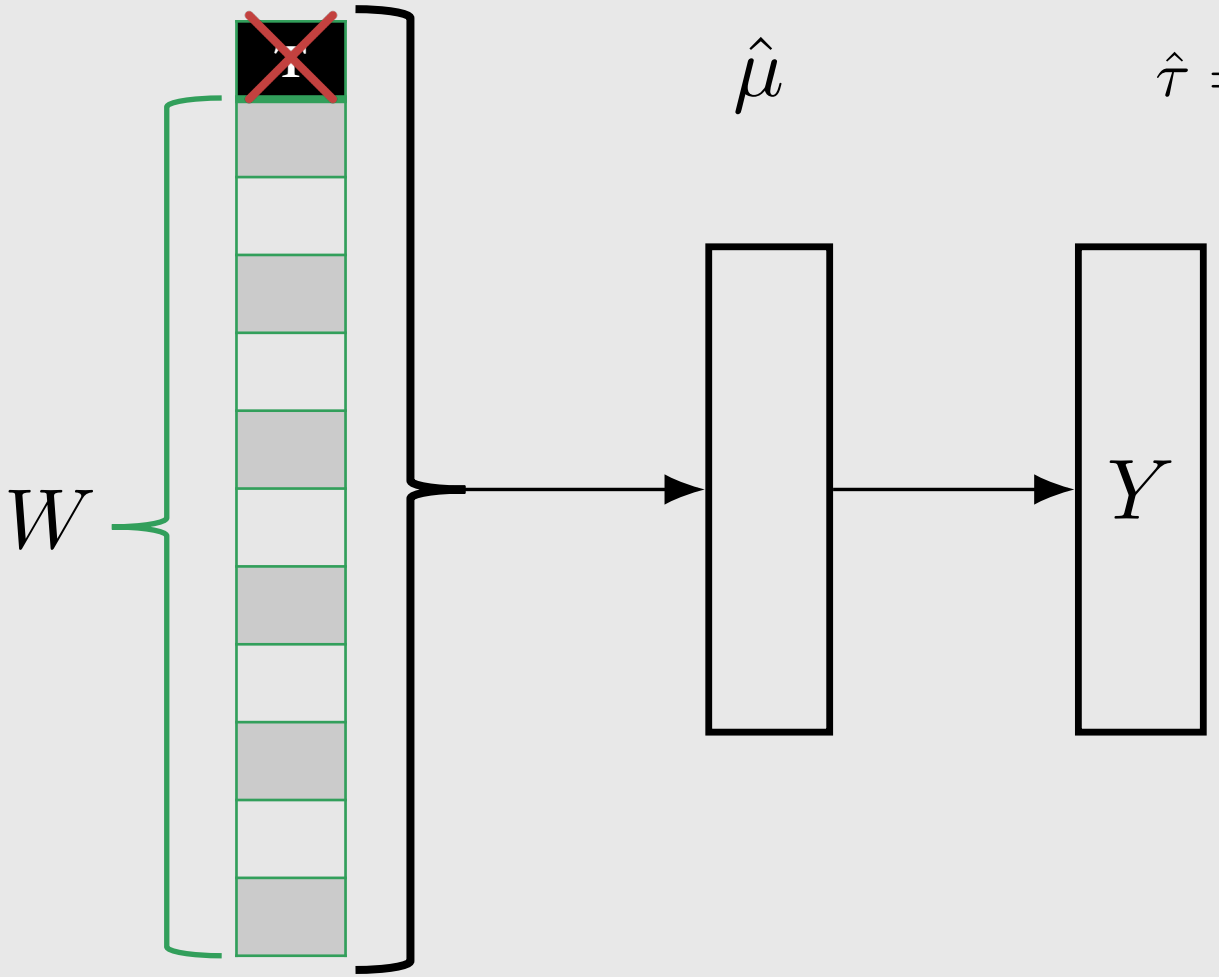
$$\hat{\tau} = \frac{1}{n} \sum_i (\hat{\mu}(1, w_i) - \hat{\mu}(0, w_i))$$

Problem with COM estimation in high dimensions



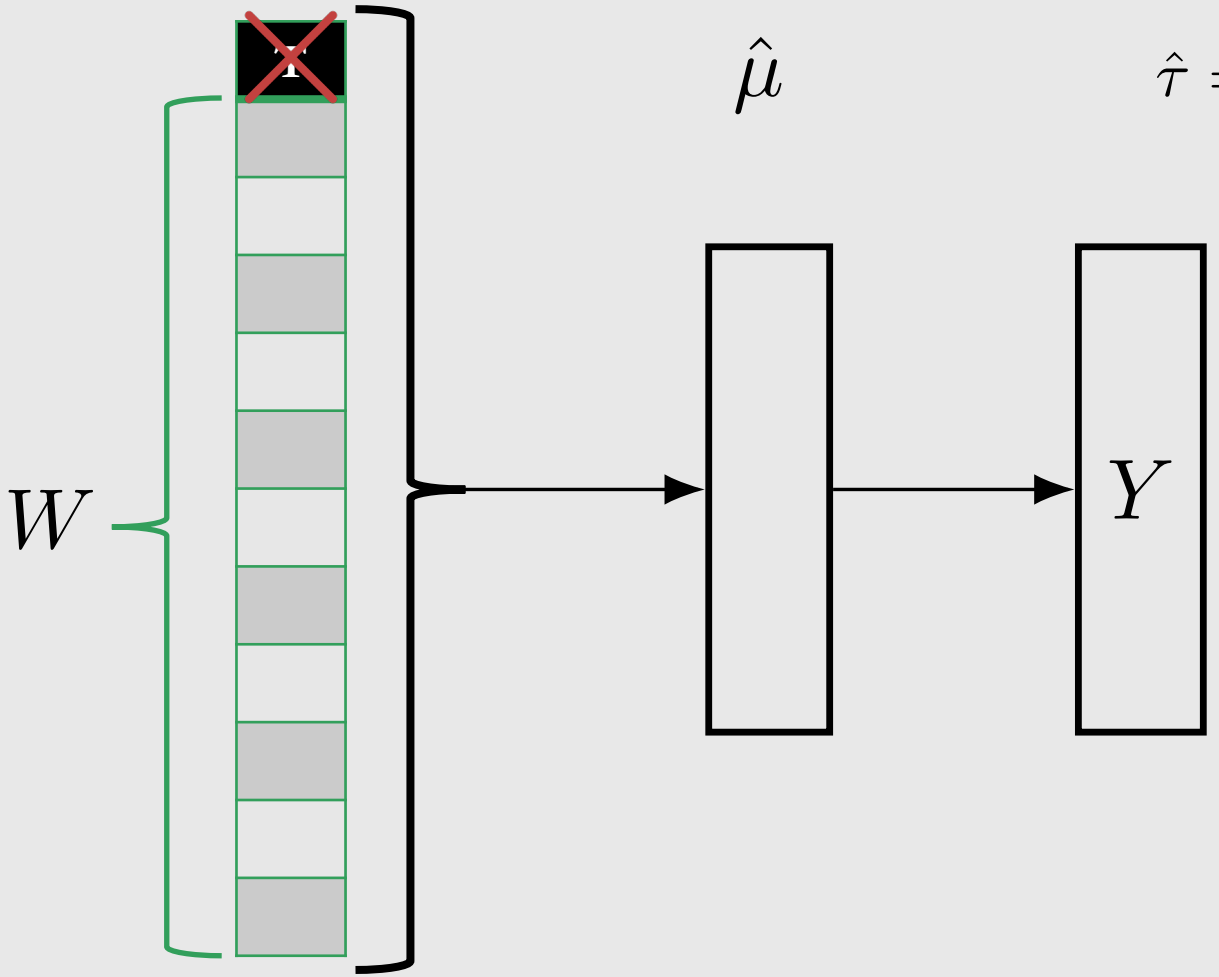
$$\hat{\tau} = \frac{1}{n}\sum_i (\hat{\mu}(1, w_i) - \hat{\mu}(0, w_i))$$

Problem with COM estimation in high dimensions



$$\hat{\tau} = \frac{1}{n}\sum_i (\hat{\mu}(\cancel{1}, w_i) - \hat{\mu}(\cancel{0}, w_i))$$

Problem with COM estimation in high dimensions



$$\hat{\tau} = \frac{1}{n}\sum_i (\hat{\mu}(\cancel{1}, w_i) - \hat{\mu}(\cancel{0}, w_i))$$

Problem: estimate can be biased toward zero (Künzel et al., 2019)

How can we ensure that the model doesn't ignore T?

Grouped COM (GCOM) estimation

$$\text{COM: } \hat{\tau} = \frac{1}{n} \sum_i (\hat{\mu}(1, w_i) - \hat{\mu}(0, w_i))$$

Grouped COM (GCOM) estimation

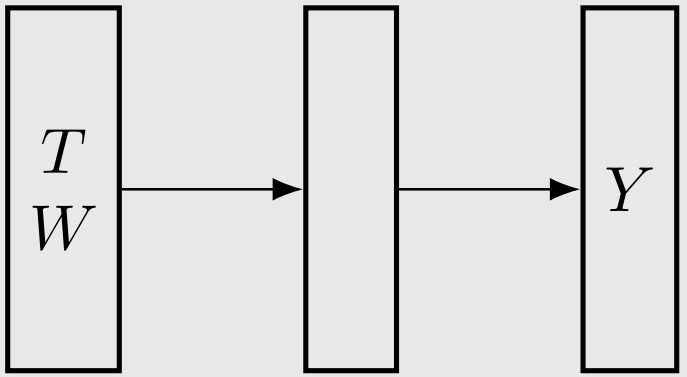
$$\text{COM: } \hat{\tau} = \frac{1}{n} \sum_i (\hat{\mu}(1, w_i) - \hat{\mu}(0, w_i))$$

$$\text{GCOM: } \hat{\tau} = \frac{1}{n} \sum_i (\hat{\mu}_1(w_i) - \hat{\mu}_0(w_i))$$

Grouped COM (GCOM) estimation

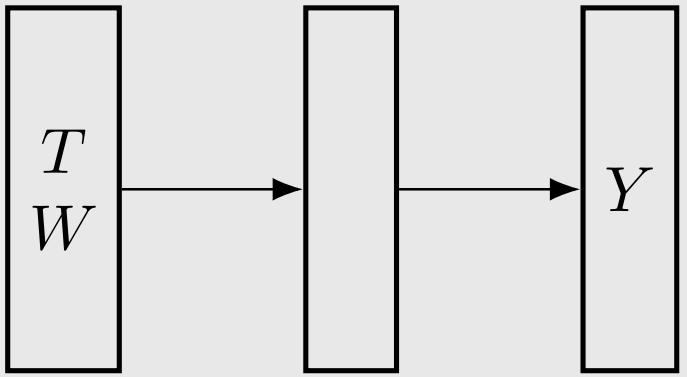
COM: $\hat{\tau} = \frac{1}{n} \sum_i (\hat{\mu}(1, w_i) - \hat{\mu}(0, w_i))$

GCOM: $\hat{\tau} = \frac{1}{n} \sum_i (\hat{\mu}_1(w_i) - \hat{\mu}_0(w_i))$



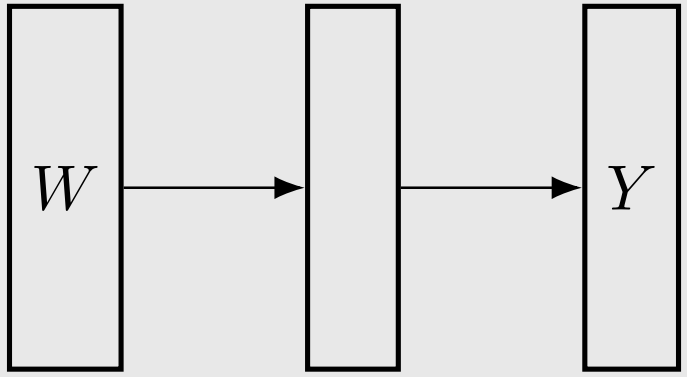
Grouped COM (GCOM) estimation

COM: $\hat{\tau} = \frac{1}{n} \sum_i (\hat{\mu}(1, w_i) - \hat{\mu}(0, w_i))$

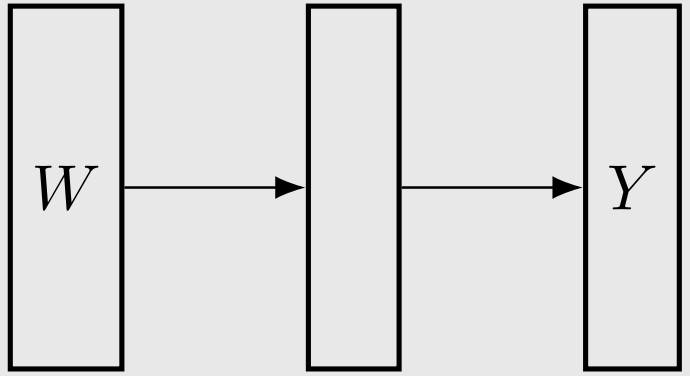


GCOM: $\hat{\tau} = \frac{1}{n} \sum_i (\hat{\mu}_1(w_i) - \hat{\mu}_0(w_i))$

T = 1 network

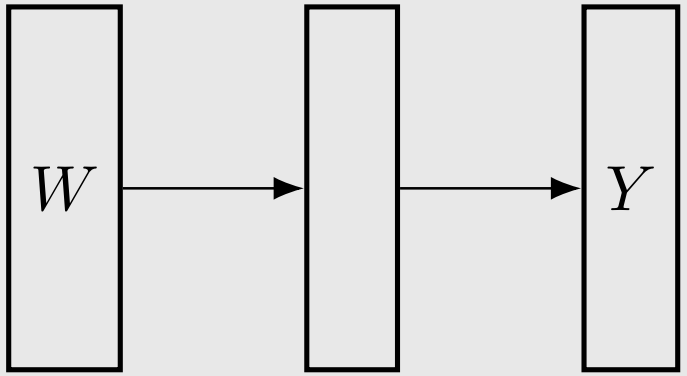


T = 0 network



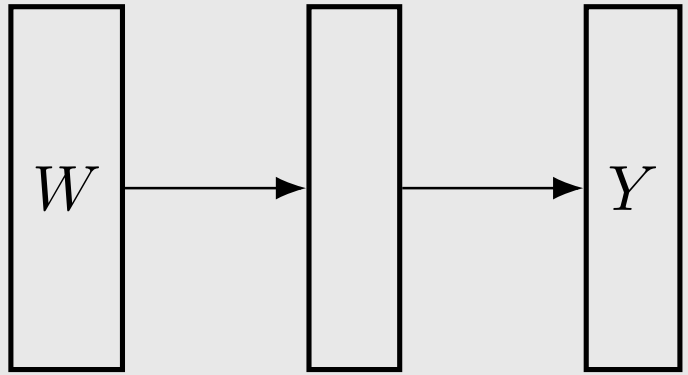
Trained with treatment group data

T = 1 network



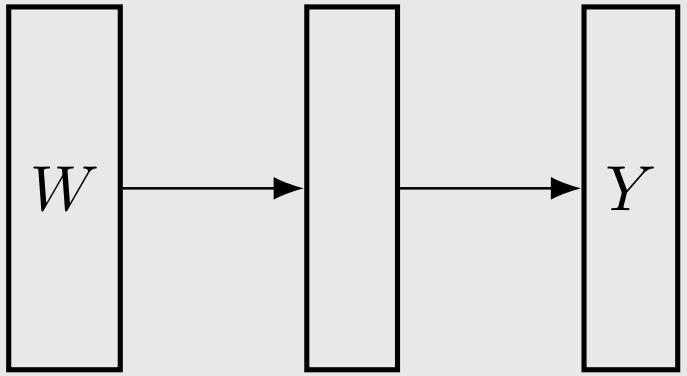
Trained with control group data

T = 0 network



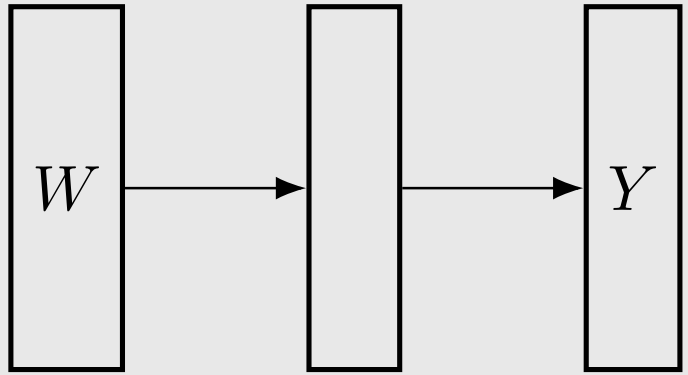
Trained with treatment group data

T = 1 network



Trained with control group data

T = 0 network



Problem: networks have higher variance than they would if they were trained with all the data (not efficient)

Question:
Write down the general form of a COM estimator and a GCOM estimator.

Conditional Outcome Modeling

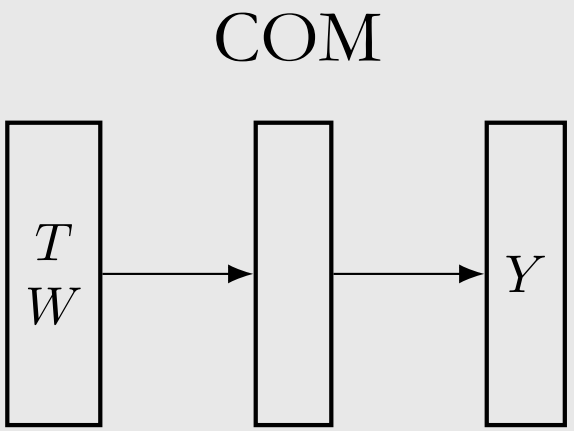
Increasing Data Efficiency

Propensity Scores and IPW

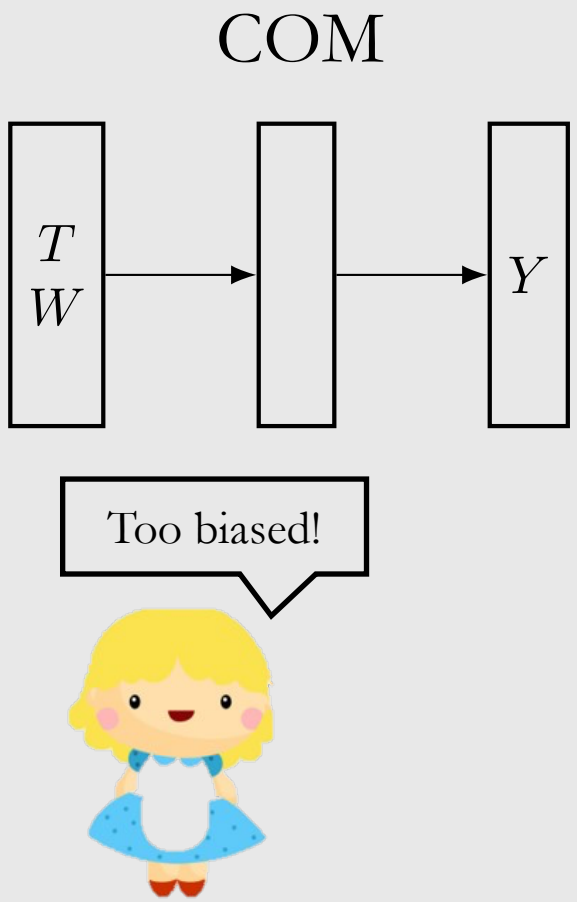
Other Methods

TARNet

TARNet

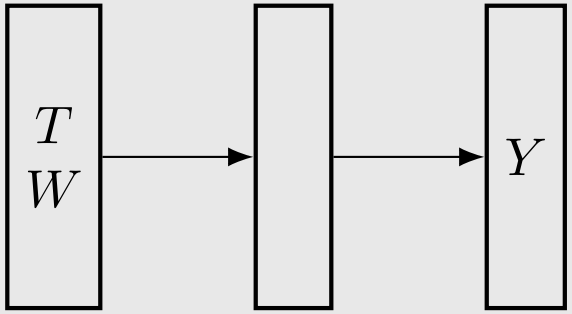


TARNet



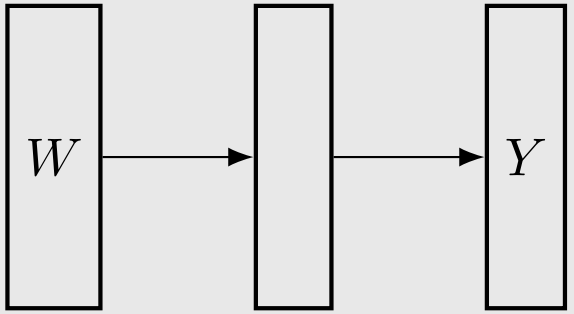
TARNet

COM

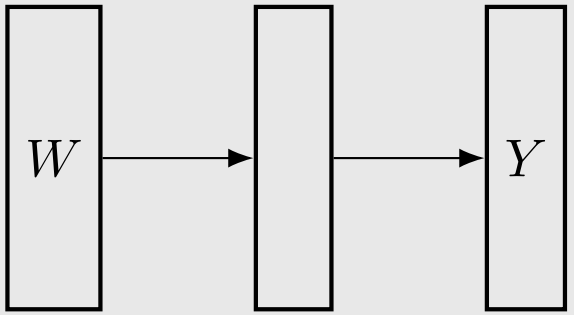


GCOM

$T = 1$ network

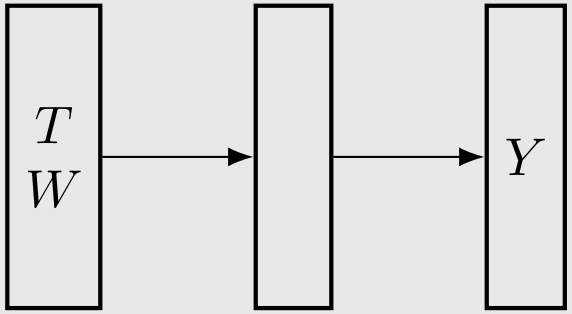


$T = 0$ network



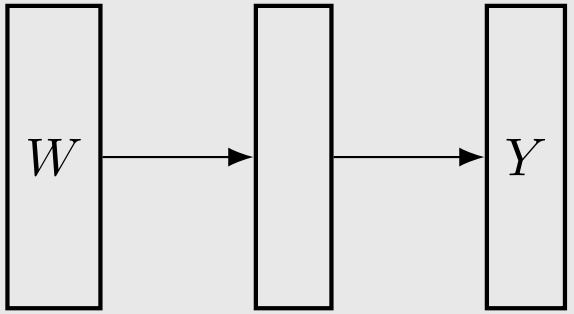
TARNet

COM

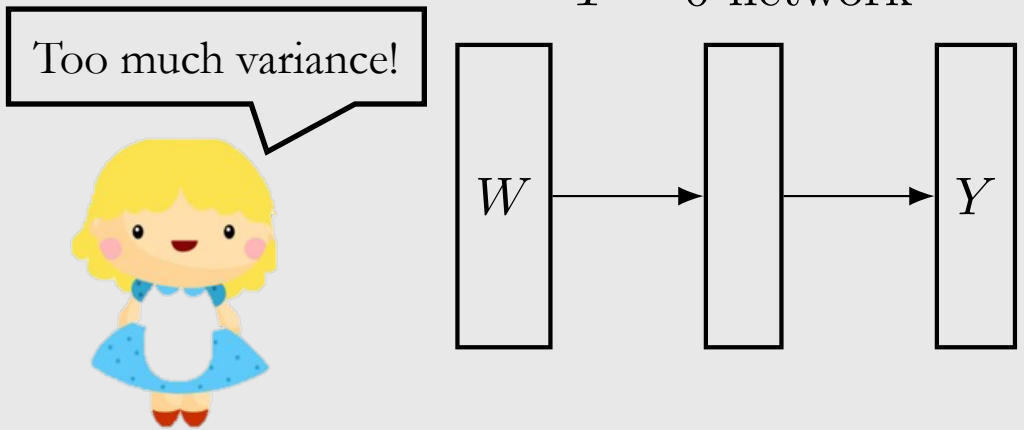


GCOM

$T = 1$ network

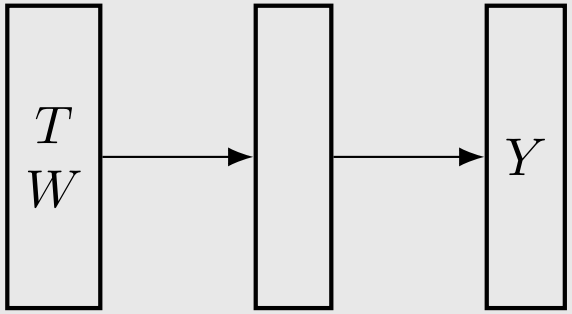


$T = 0$ network



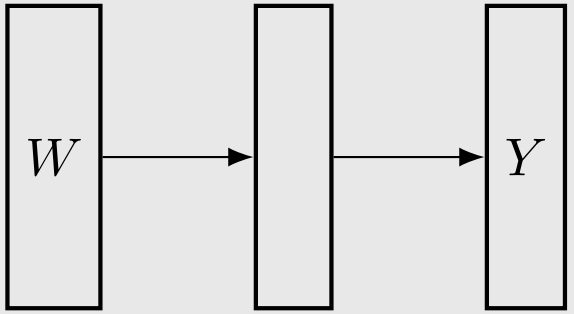
TARNet

COM

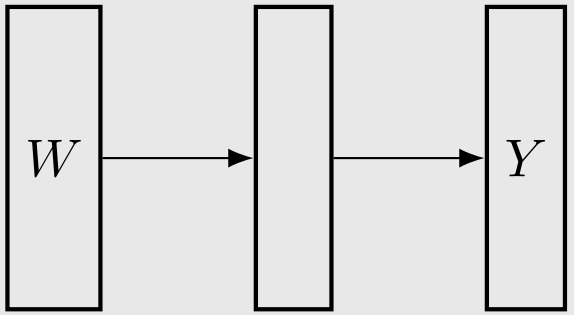


GCOM

$T = 1$ network

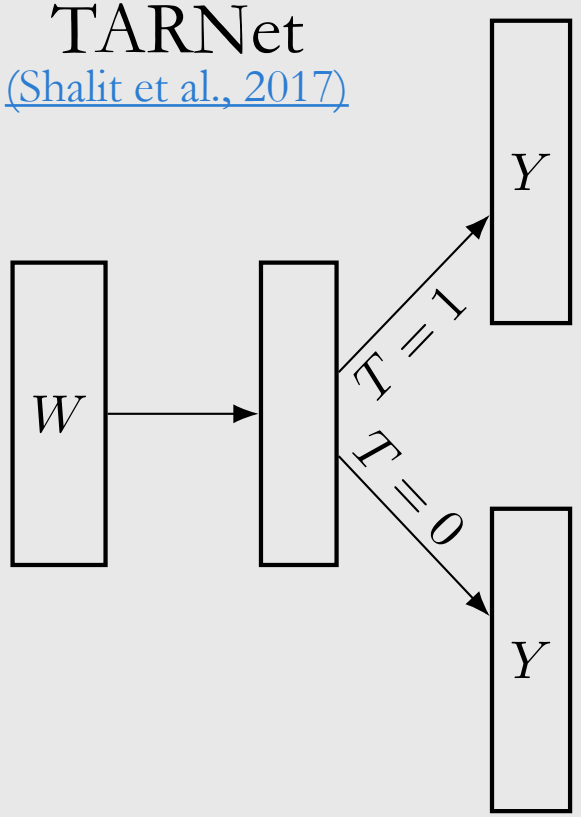


$T = 0$ network



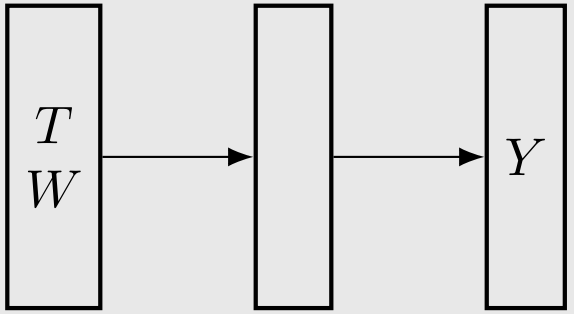
TARNet

(Shalit et al., 2017)



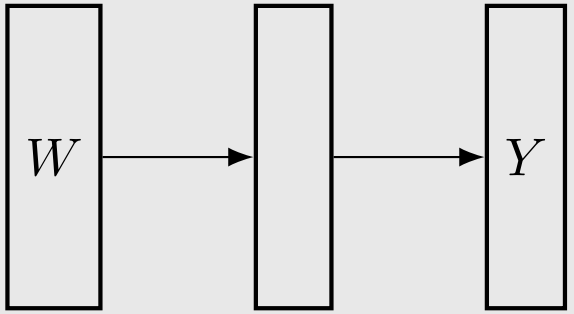
TARNet

COM

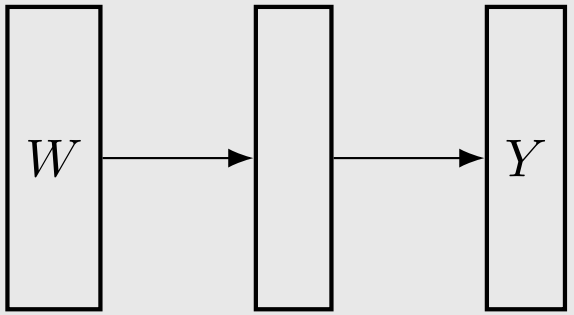


GCOM

$T = 1$ network

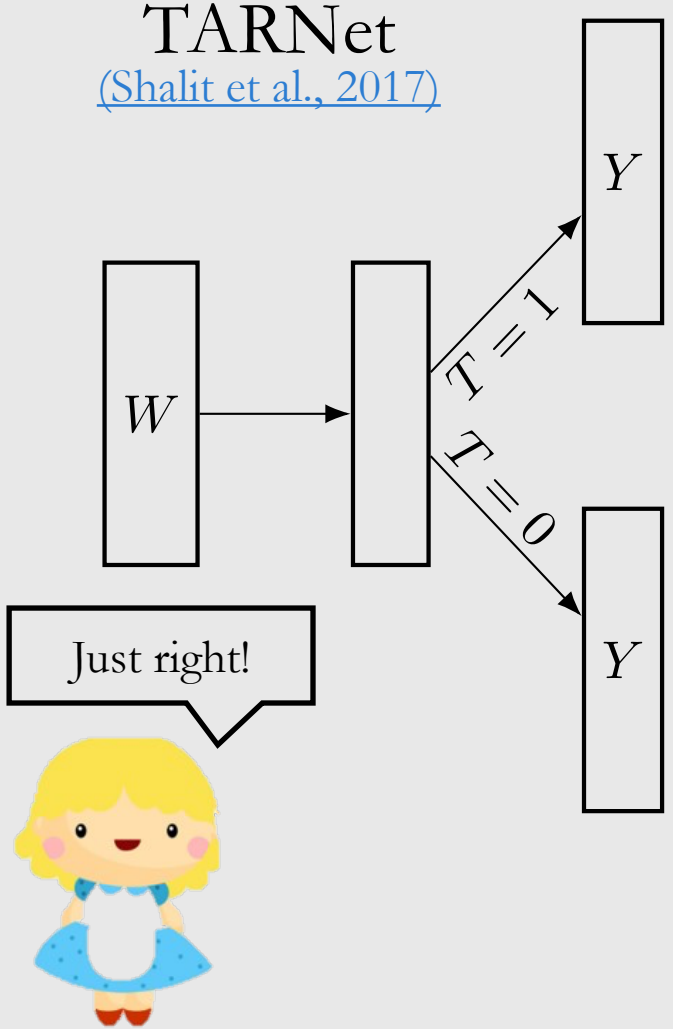


$T = 0$ network



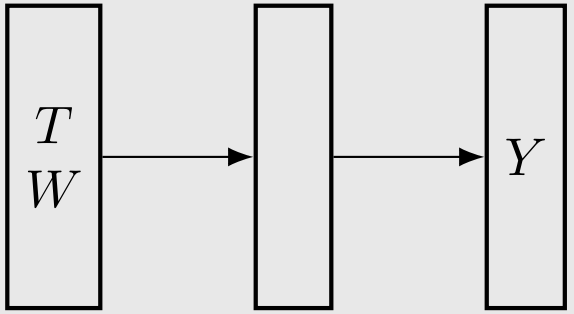
TARNet

(Shalit et al., 2017)



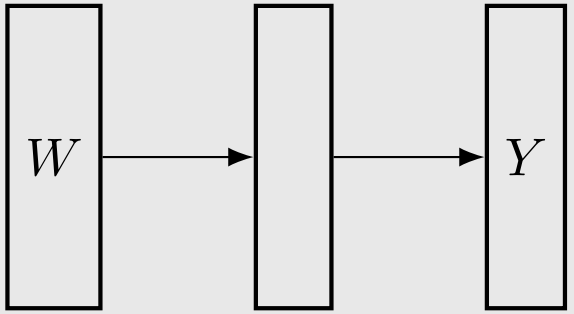
TARNet

COM

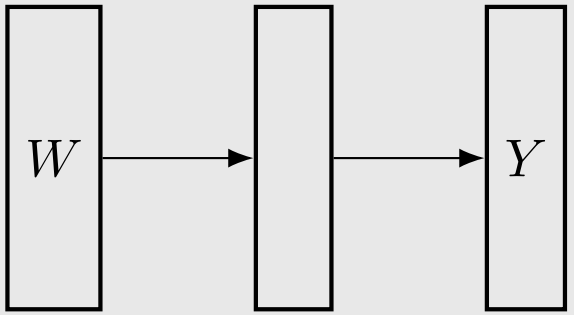


GCOM

$T = 1$ network

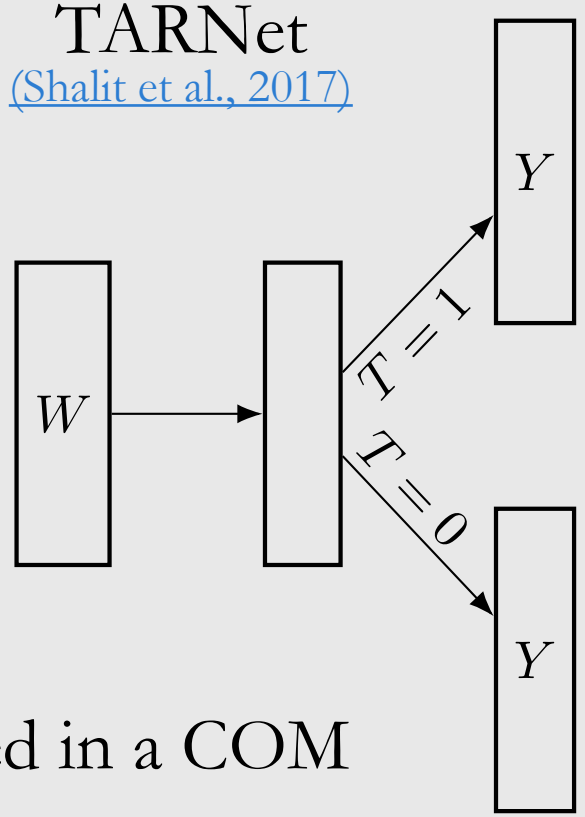


$T = 0$ network



TARNet

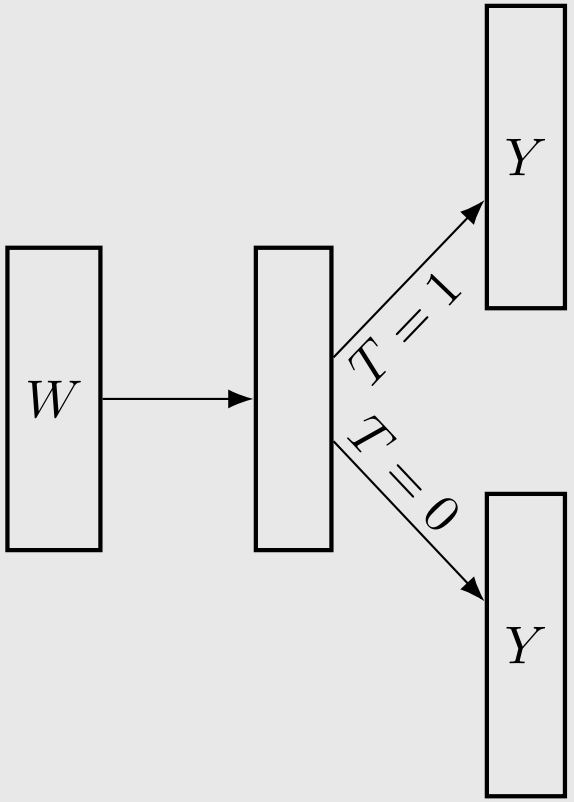
(Shalit et al., 2017)



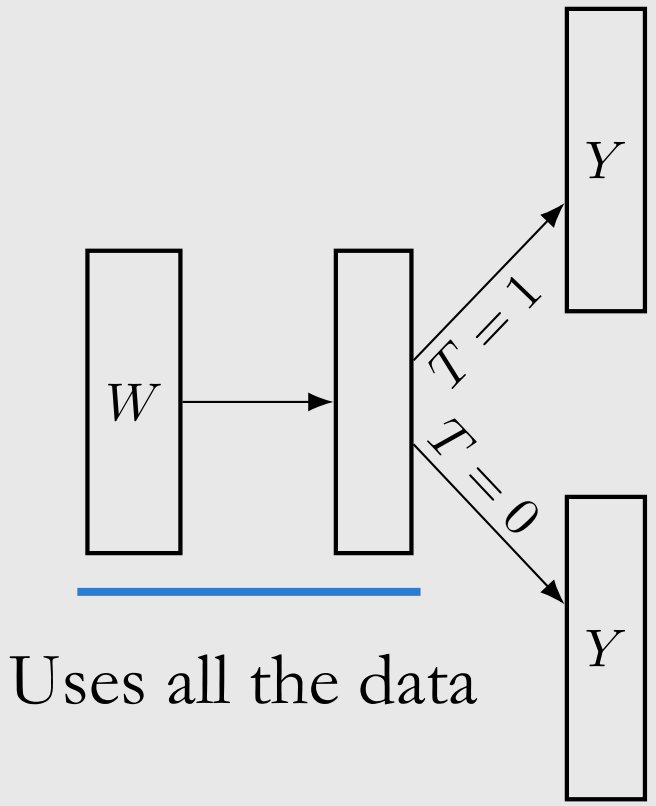
Used in a COM estimator:

$$\hat{\tau} = \frac{1}{n} \sum_i (\hat{\mu}(1, w_i) - \hat{\mu}(0, w_i))$$

TARNet inefficiency

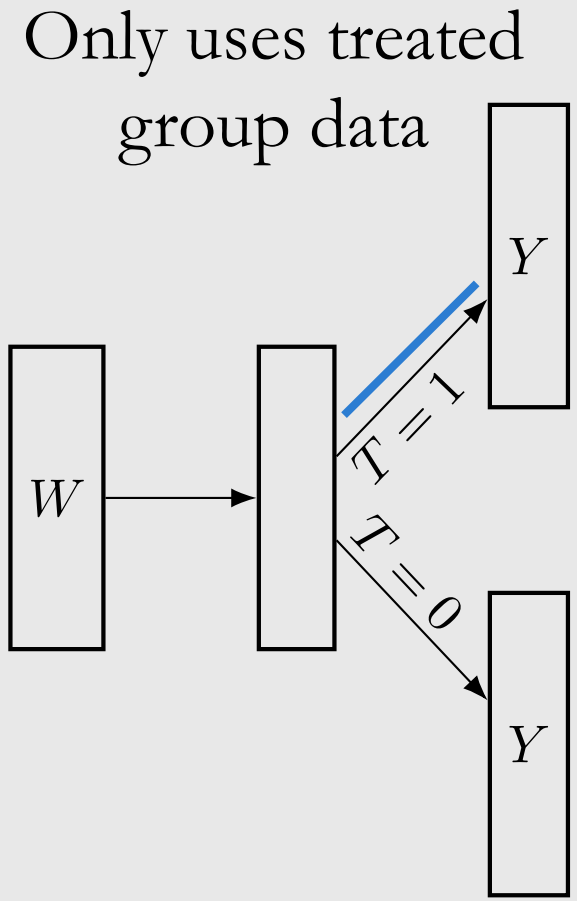


TARNet inefficiency

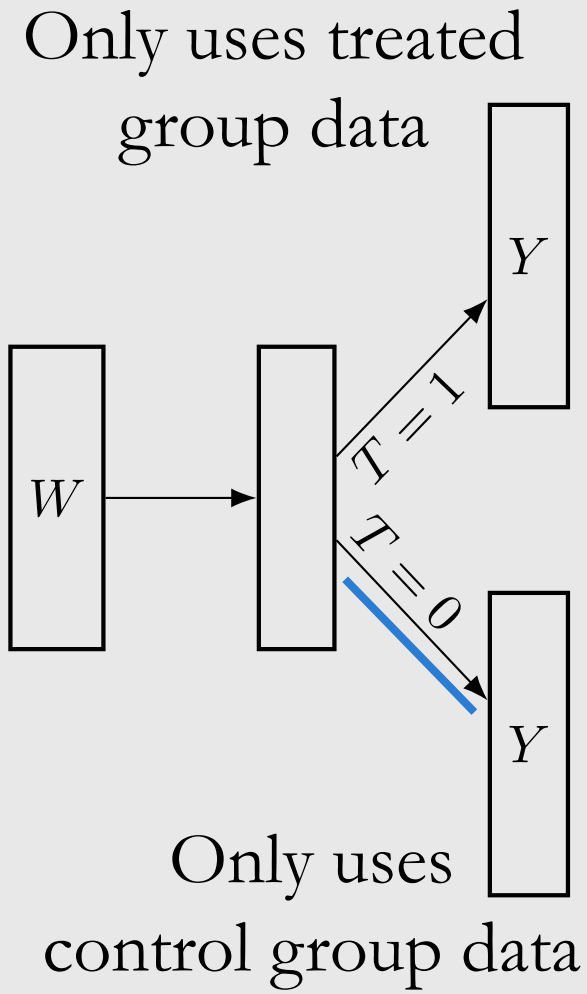


Uses all the data

TARNet inefficiency



TARNet inefficiency



X-Learner

(Künzel et al., 2019)

X-Learner

(Künzel et al., 2019)

1. Estimate $\hat{\mu}_1(x)$ and $\hat{\mu}_0(x)$

X-Learner

(Künzel et al., 2019)

1. Estimate $\hat{\mu}_1(x)$ and $\hat{\mu}_0(x)$

Assume X is a sufficient adjustment set and is all observed covariates

X-Learner

(Künzel et al., 2019)

1. Estimate $\hat{\mu}_1(x)$ and $\hat{\mu}_0(x)$

Assume X is a sufficient adjustment set and is all observed covariates

2a. Impute ITEs

X-Learner

(Künzel et al., 2019)

1. Estimate $\hat{\mu}_1(x)$ and $\hat{\mu}_0(x)$

Assume X is a sufficient adjustment set and is all observed covariates

2a. Impute ITEs

Treatment group:

$$\hat{\tau}_{1,i} = Y_i(1) - \hat{\mu}_0(x_i)$$

X-Learner

(Künzel et al., 2019)

1. Estimate $\hat{\mu}_1(x)$ and $\hat{\mu}_0(x)$

Assume X is a sufficient adjustment set and is all observed covariates

2a. Impute ITEs

Treatment group:

$$\hat{\tau}_{1,i} = Y_i(1) - \hat{\mu}_0(x_i)$$

Control group:

$$\hat{\tau}_{0,i} = \hat{\mu}_1(x_i) - Y_i(0)$$

X-Learner

(Künzel et al., 2019)

1. Estimate $\hat{\mu}_1(x)$ and $\hat{\mu}_0(x)$

Assume X is a sufficient adjustment set and is all observed covariates

2a. Impute ITEs

Treatment group: $\hat{\tau}_{1,i} = Y_i(1) - \hat{\mu}_0(x_i)$

Control group: $\hat{\tau}_{0,i} = \hat{\mu}_1(x_i) - Y_i(0)$

2b. Fit a model $\hat{\tau}_1(x)$ to predict $\hat{\tau}_{1,i}$ from x_i in treatment group

X-Learner

(Künzel et al., 2019)

1. Estimate $\hat{\mu}_1(x)$ and $\hat{\mu}_0(x)$

Assume X is a sufficient adjustment set and is all observed covariates

2a. Impute ITEs

Treatment group: $\hat{\tau}_{1,i} = Y_i(1) - \hat{\mu}_0(x_i)$

Control group: $\hat{\tau}_{0,i} = \hat{\mu}_1(x_i) - Y_i(0)$

2b. Fit a model $\hat{\tau}_1(x)$ to predict $\hat{\tau}_{1,i}$ from x_i in treatment group

Fit a model $\hat{\tau}_0(x)$ to predict $\hat{\tau}_{0,i}$ from x_i in control group

X-Learner

(Künzel et al., 2019)

1. Estimate $\hat{\mu}_1(x)$ and $\hat{\mu}_0(x)$

Assume X is a sufficient adjustment set and is all observed covariates

2a. Impute ITEs

Treatment group: $\hat{\tau}_{1,i} = Y_i(1) - \hat{\mu}_0(x_i)$

Control group: $\hat{\tau}_{0,i} = \hat{\mu}_1(x_i) - Y_i(0)$

2b. Fit a model $\hat{\tau}_1(x)$ to predict $\hat{\tau}_{1,i}$ from x_i in treatment group

Fit a model $\hat{\tau}_0(x)$ to predict $\hat{\tau}_{0,i}$ from x_i in control group

3. $$\hat{\tau}(x) = g(x)\,\hat{\tau}_0(x) + (1 - g(x))\,\hat{\tau}_1(x)$$

where $g(x)$ is some weighing function between 0 and 1. Example: propensity score

Question:
What would motivate someone to consider a more complex type of estimation than COM/GCOM?

Conditional Outcome Modeling

Increasing Data Efficiency

Propensity Scores and IPW

Other Methods

Propensity scores

Propensity scores

$$P(T = 1 \mid W)$$

Propensity scores

$$e(W) \triangleq P(T = 1 \mid W)$$

Propensity scores

$$e(W) \triangleq P(T = 1 \mid W)$$

Given positivity, unconfoundedness given W implies unconfoundedness given the propensity score $e(W)$.

Propensity scores

$$e(W) \triangleq P(T = 1 \mid W)$$

Given positivity, unconfoundedness given W implies unconfoundedness given the propensity score $e(W)$.

Even if W is high-dimensional, $e(W)$ is only 1-dimensional!

Propensity Score Theorem

Given positivity, unconfoundedness given W implies unconfoundedness given the propensity score $e(W)$.

Propensity Score Theorem

Given positivity, unconfoundedness given W implies unconfoundedness given the propensity score $e(W)$. Equivalently,

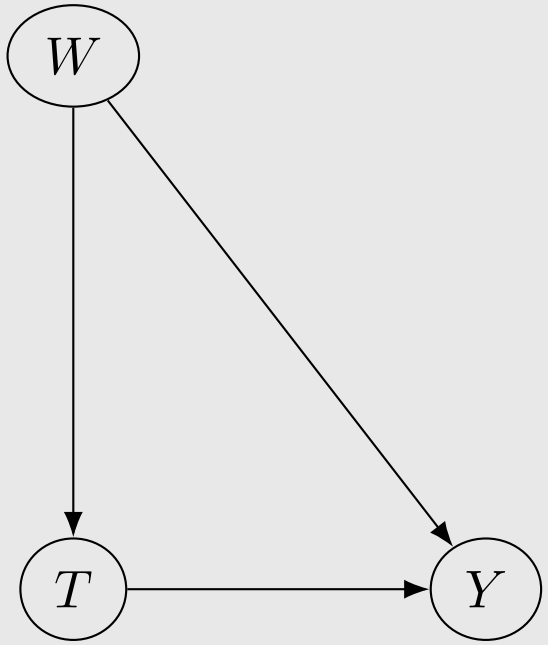
$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid W \implies (Y(1), Y(0)) \perp\!\!\!\perp T \mid e(W)$$

Propensity Score Theorem

Given positivity, unconfoundedness given W implies unconfoundedness given the propensity score $e(W)$. Equivalently,

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid W \implies (Y(1), Y(0)) \perp\!\!\!\perp T \mid e(W)$$

Graphical Proof:

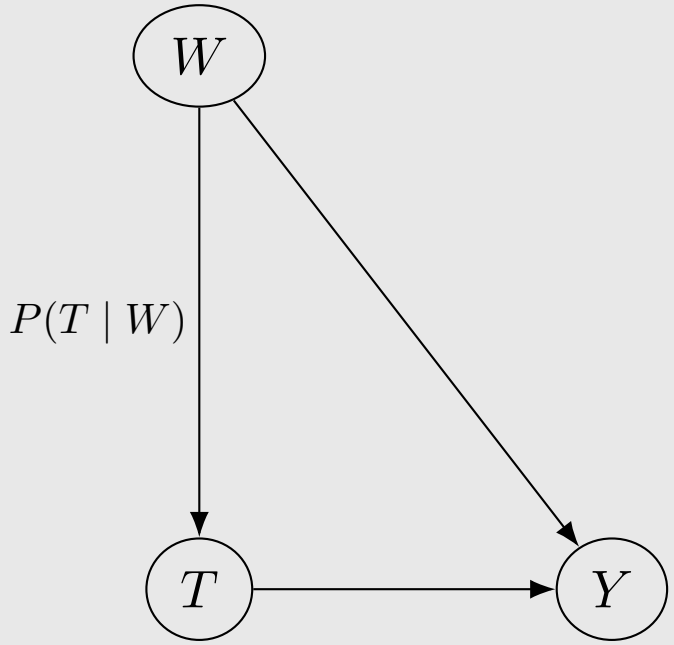


Propensity Score Theorem

Given positivity, unconfoundedness given W implies unconfoundedness given the propensity score $e(W)$. Equivalently,

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid W \implies (Y(1), Y(0)) \perp\!\!\!\perp T \mid e(W)$$

Graphical Proof:

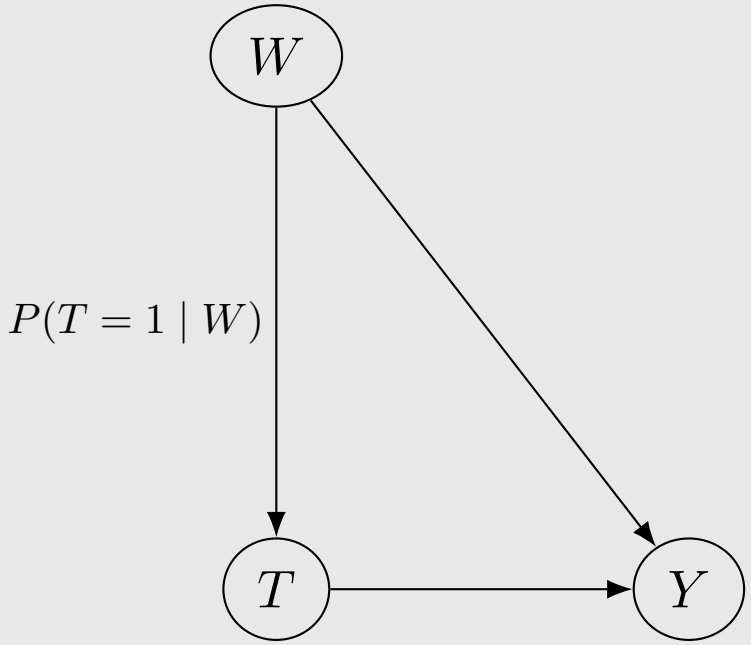


Propensity Score Theorem

Given positivity, unconfoundedness given W implies unconfoundedness given the propensity score $e(W)$. Equivalently,

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid W \implies (Y(1), Y(0)) \perp\!\!\!\perp T \mid e(W)$$

Graphical Proof:

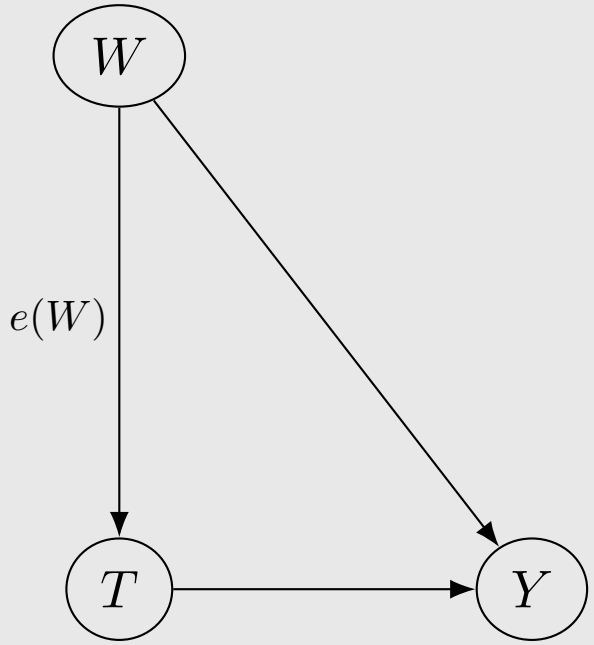


Propensity Score Theorem

Given positivity, unconfoundedness given W implies unconfoundedness given the propensity score $e(W)$. Equivalently,

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid W \implies (Y(1), Y(0)) \perp\!\!\!\perp T \mid e(W)$$

Graphical Proof:

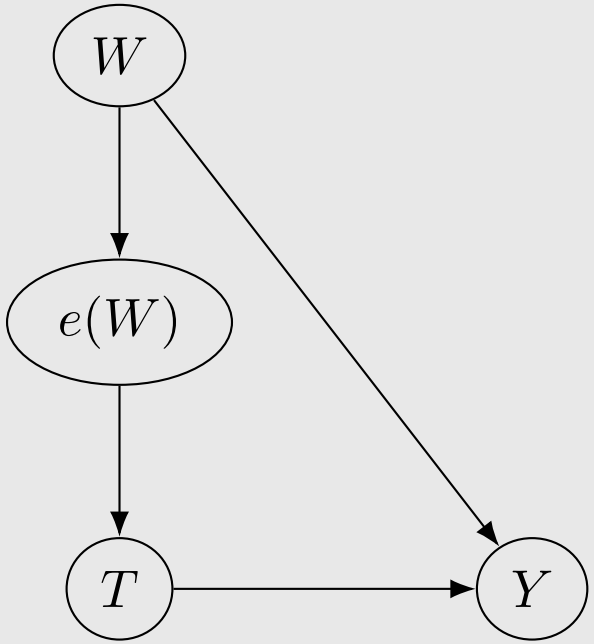


Propensity Score Theorem

Given positivity, unconfoundedness given W implies unconfoundedness given the propensity score $e(W)$. Equivalently,

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid W \implies (Y(1), Y(0)) \perp\!\!\!\perp T \mid e(W)$$

Graphical Proof:

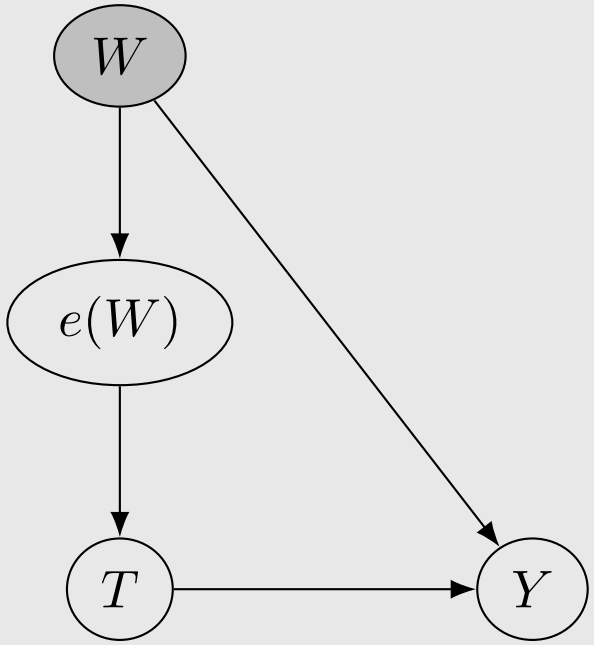


Propensity Score Theorem

Given positivity, unconfoundedness given W implies unconfoundedness given the propensity score $e(W)$. Equivalently,

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid W \implies (Y(1), Y(0)) \perp\!\!\!\perp T \mid e(W)$$

Graphical Proof:

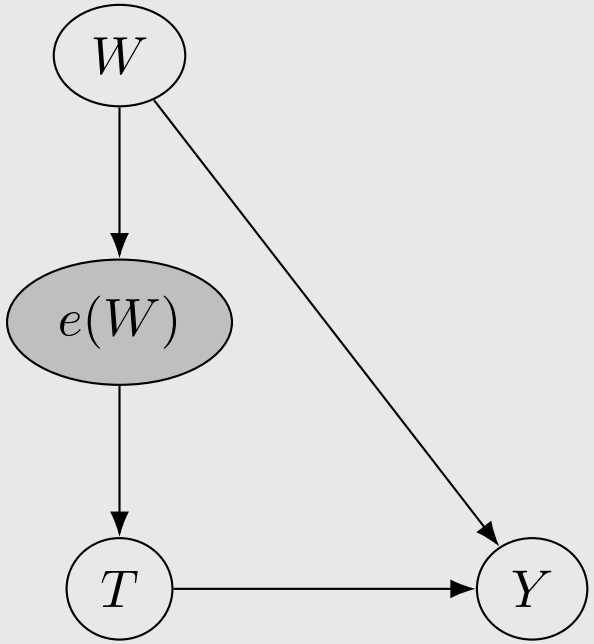


Propensity Score Theorem

Given positivity, unconfoundedness given W implies unconfoundedness given the propensity score $e(W)$. Equivalently,

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid W \implies (Y(1), Y(0)) \perp\!\!\!\perp T \mid e(W)$$

Graphical Proof:

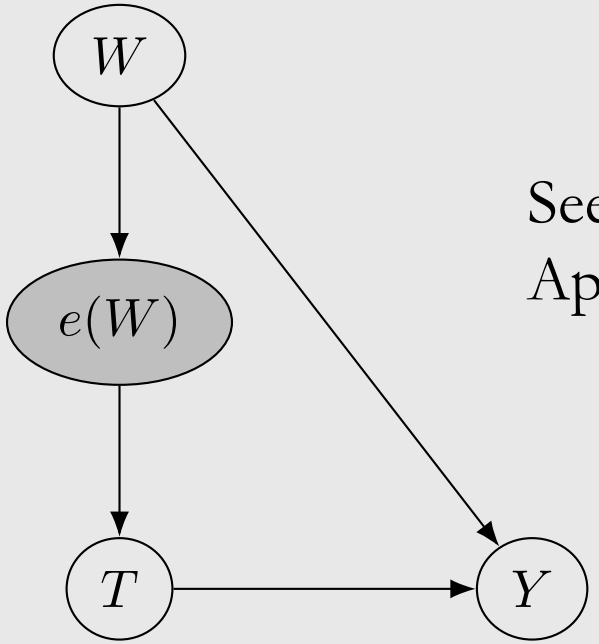


Propensity Score Theorem

Given positivity, unconfoundedness given W implies unconfoundedness given the propensity score $e(W)$. Equivalently,

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid W \implies (Y(1), Y(0)) \perp\!\!\!\perp T \mid e(W)$$

Graphical Proof:



See non-graphical proof in Appendix A.2 of the course book

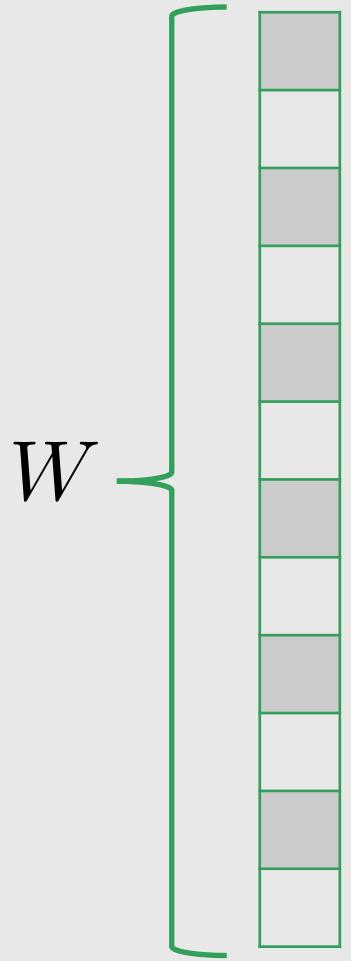
Implications for the Positivity-Unconfoundedness Tradeoff

Implications for the Positivity-Unconfoundedness Tradeoff

Recall that overlap decreases with the dimensionality of the adjustment set

Implications for the Positivity-Unconfoundedness Tradeoff

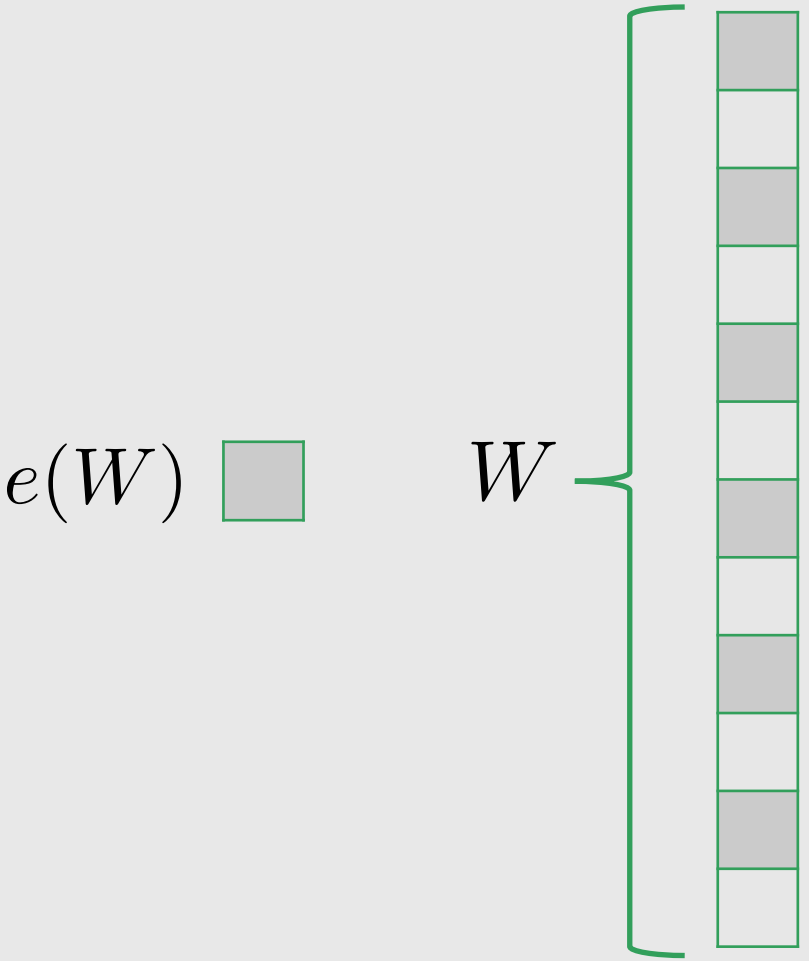
Recall that overlap decreases with the dimensionality of the adjustment set



Implications for the Positivity-Unconfoundedness Tradeoff

Recall that overlap decreases with the dimensionality of the adjustment set

The propensity score magically reduces the dimensionality of the adjustment set done to 1!

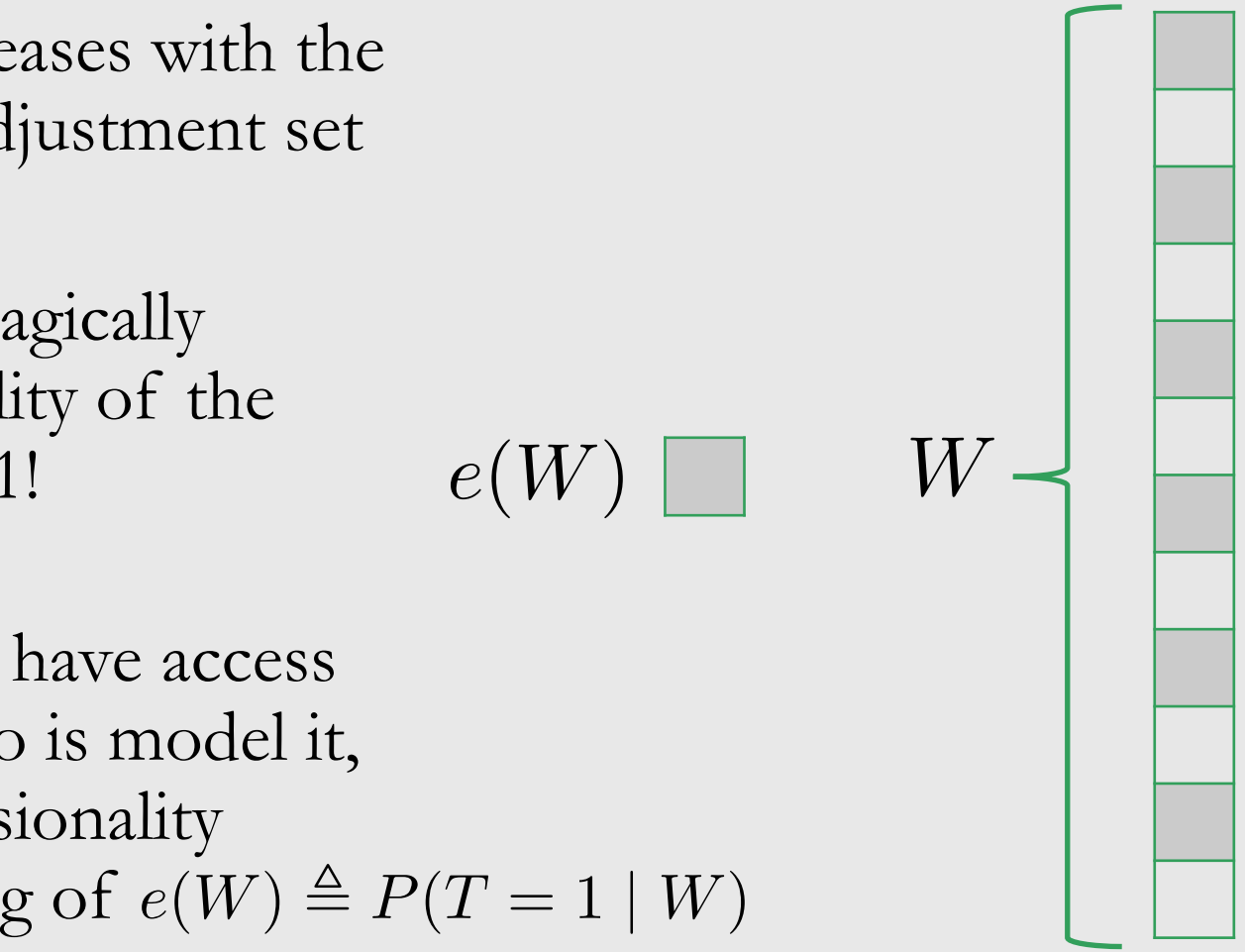


Implications for the Positivity-Unconfoundedness Tradeoff

Recall that overlap decreases with the dimensionality of the adjustment set

The propensity score magically reduces the dimensionality of the adjustment set done to 1!

Unfortunately, we don't have access to it. The best we can do is model it, shifting the high-dimensionality problem to the modeling of $e(W) \triangleq P(T = 1 \mid W)$



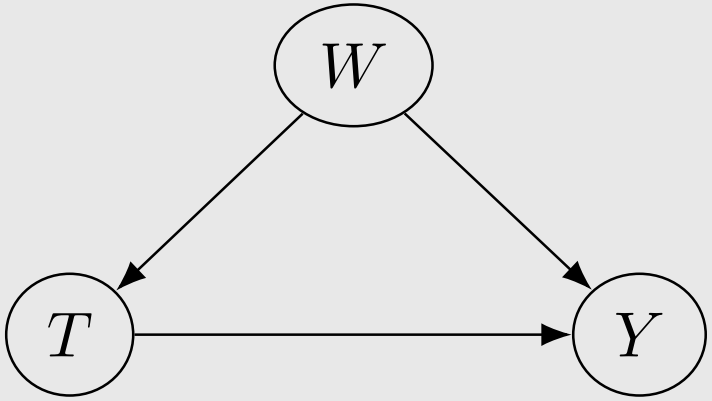
Questions:

1. What is the intuition behind why we can condition on e(W) instead of W?
2. What is attractive about conditioning on e(W) as opposed to W?
3. Why does this not solve positivity issues when W is high-dimensional?

Pseudo-populations

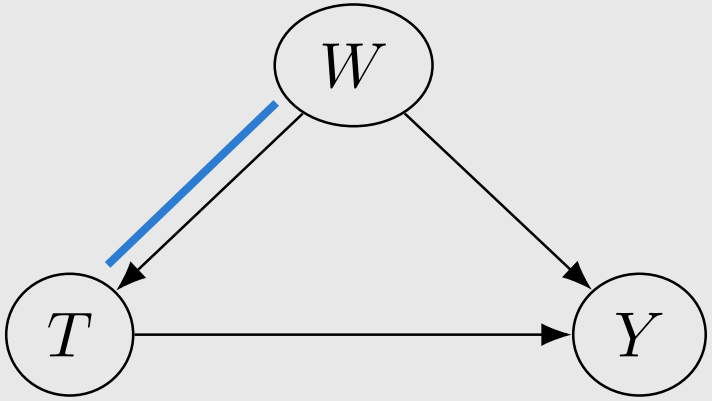
Pseudo-populations

Regular population



Pseudo-populations

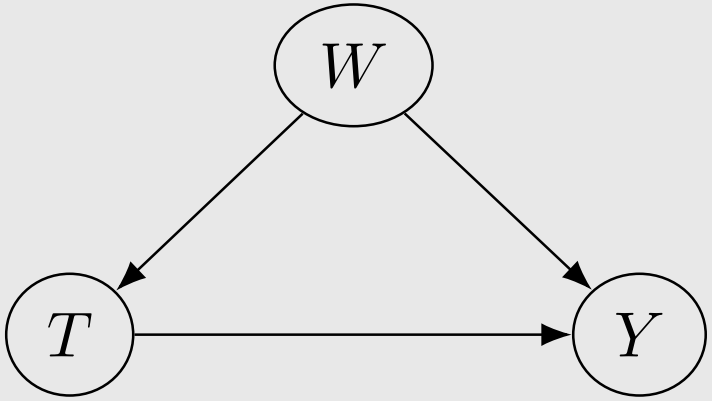
Regular population



$$P(T \mid W) \neq P(T)$$

Pseudo-populations

Regular population



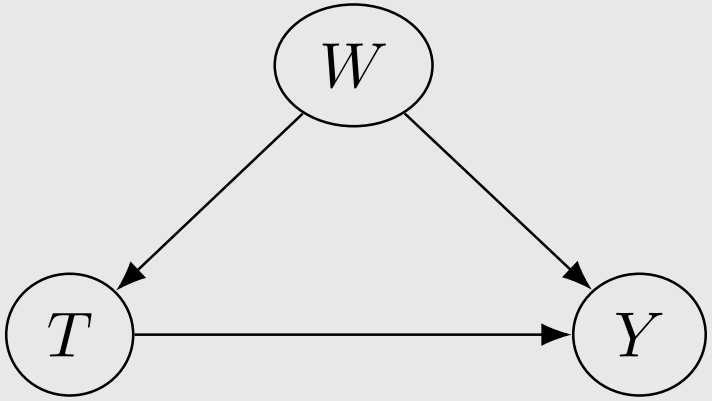
$$P(T \mid W) \neq P(T)$$

Reweighted population
(Pseudo-population)

$$P(T \mid W) = P(T) \text{ or}$$
$$P(T \mid W) = 1$$

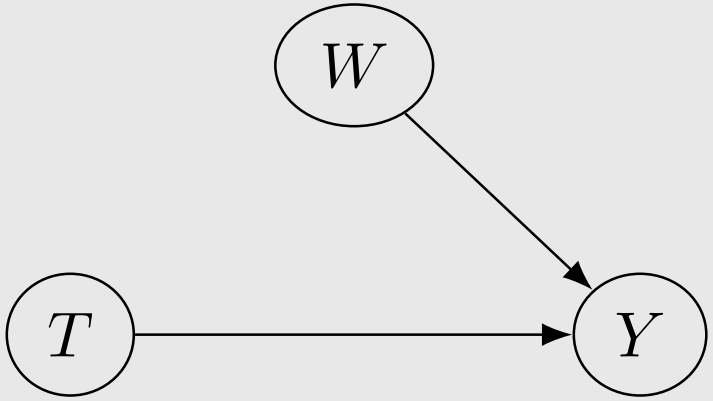
Pseudo-populations

Regular population



$$P(T \mid W) \neq P(T)$$

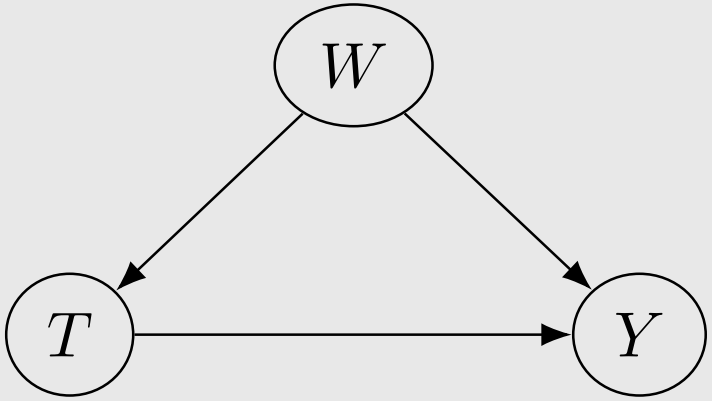
Reweighted population
(Pseudo-population)



$$P(T \mid W) = P(T) \text{ or}$$
$$P(T \mid W) = 1$$

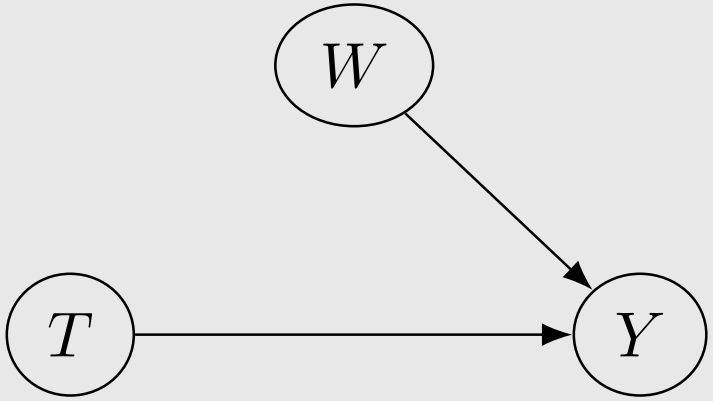
Pseudo-populations

Regular population



$$P(T \mid W) \neq P(T)$$

Reweighted population
(Pseudo-population)

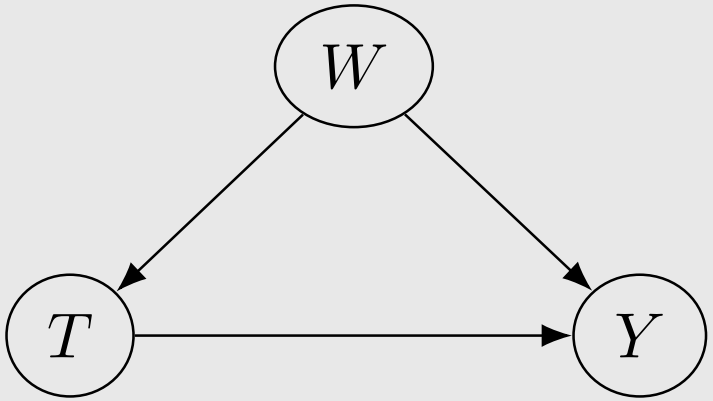


$$P(T \mid W) = P(T) \text{ or } P(T \mid W) = 1$$

Reweighting intuition: $P(T \mid W) \cdot \dfrac{1}{P(T \mid W)} = 1$

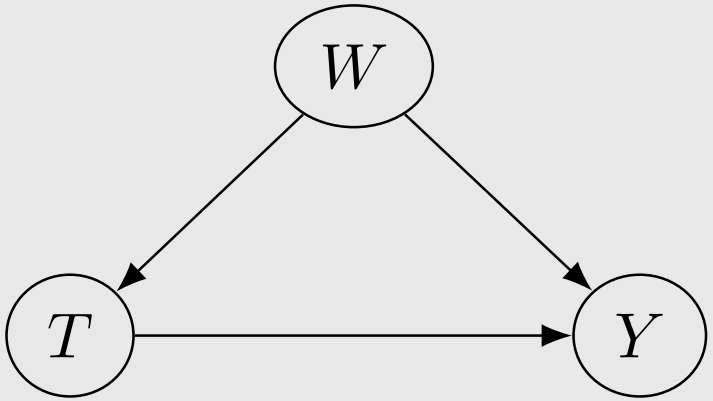
Inverse probability weighting (IPW)

Inverse probability weighting (IPW)



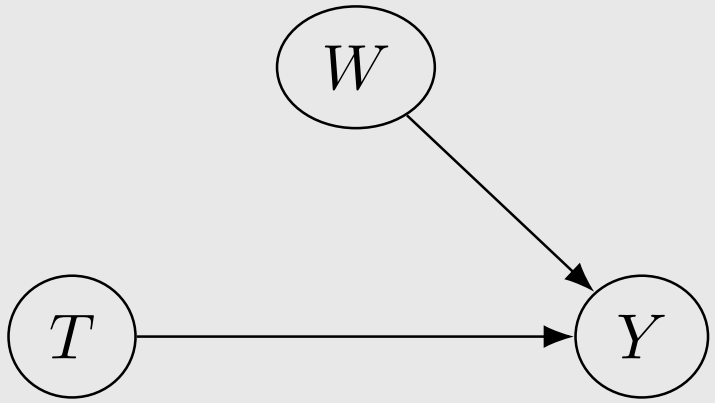
Inverse probability weighting (IPW)

Y



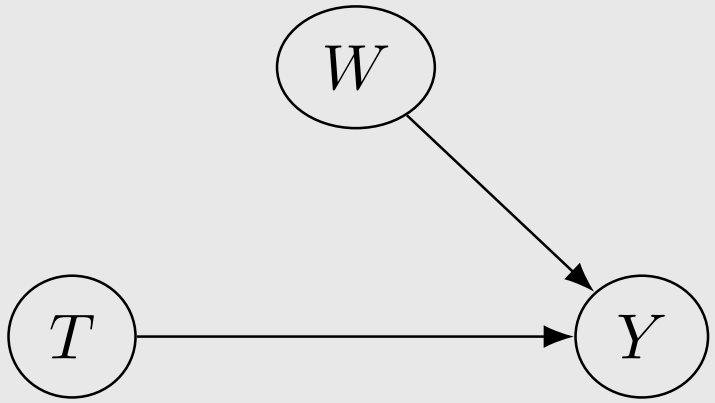
Inverse probability weighting (IPW)

$$\frac{Y}{P(t \mid W)}$$



Inverse probability weighting (IPW)

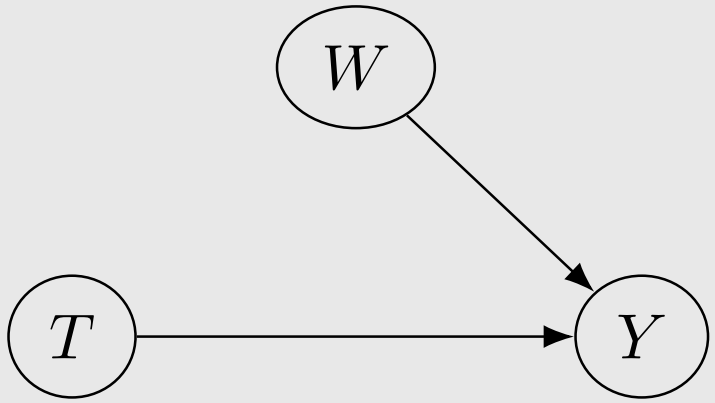
$$\mathbb{E}[Y(t)] = \mathbb{E}\left[\frac{\mathbb{1}(T=t)Y}{P(t \mid W)}\right]$$



Inverse probability weighting (IPW)

$$\mathbb{E}[Y(t)] = \mathbb{E}\left[\frac{\mathbb{1}(T=t)Y}{P(t \mid W)}\right]$$

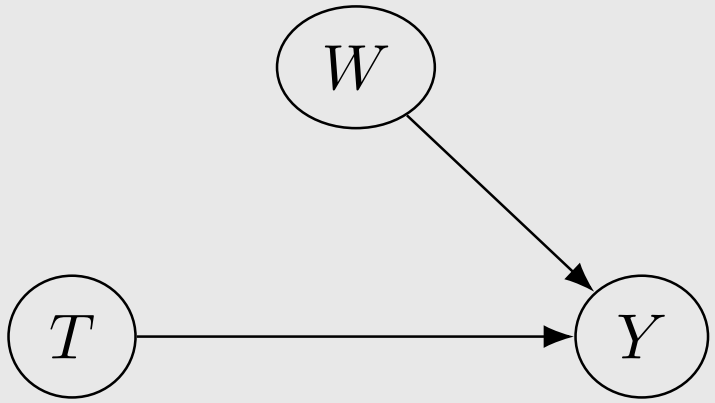
See proof in Appendix A.3 of the course book



Inverse probability weighting (IPW)

$$\mathbb{E}[Y(t)] = \mathbb{E}\left[\frac{\mathbb{1}(T=t)Y}{P(t \mid W)}\right]$$

See proof in Appendix A.3 of the course book

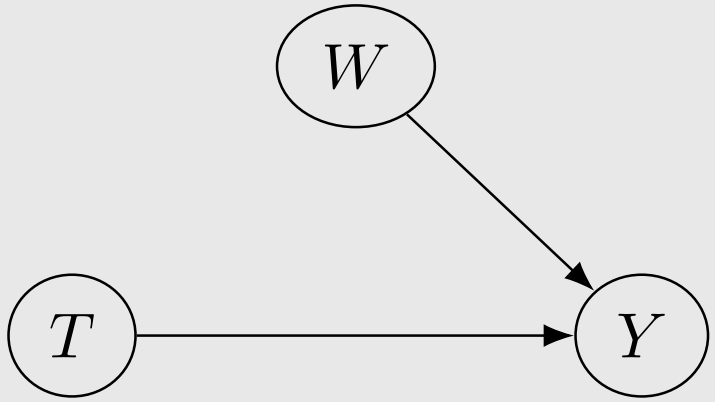


$$\tau \triangleq \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}\left[\frac{\mathbb{1}(T=1)Y}{e(W)}\right] - \mathbb{E}\left[\frac{\mathbb{1}(T=0)Y}{1-e(W)}\right]$$

Inverse probability weighting (IPW)

$$\mathbb{E}[Y(t)] = \mathbb{E}\left[\frac{\mathbb{1}(T=t)Y}{P(t \mid W)}\right]$$

See proof in Appendix A.3 of the course book



$$\tau \triangleq \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}\left[\frac{\mathbb{1}(T=1)Y}{e(W)}\right] - \mathbb{E}\left[\frac{\mathbb{1}(T=0)Y}{1-e(W)}\right]$$

$$\hat{\tau} = \frac{1}{n_1}\sum_{i:t_i=1}\frac{y_i}{\hat{e}(w_i)} - \frac{1}{n_0}\sum_{i:t_i=0}\frac{y_i}{1-\hat{e}(w_i)}$$

Questions:

1. What happens if the estimated propensity score for some unit is 1 or 0?
2. What happens if the estimated propensity score is near 1 or 0?

IPW CATE estimation

IPW CATE estimation

Not quite as natural with IPW as with COM, so beyond scope of course

IPW CATE estimation

Not quite as natural with IPW as with COM, so beyond scope of course

Simple extension:

$$\hat{\tau}(x) = \frac{1}{n_x} \sum_{i:x_i=x} \left(\frac{\mathbb{1}(t_i = 1) y_i}{\hat{e}(w_i)} - \frac{\mathbb{1}(t_i = 0) y_i}{1 - \hat{e}(w_i)} \right)$$

IPW CATE estimation

Not quite as natural with IPW as with COM, so beyond scope of course

Simple extension:

$$\hat{\tau}(x) = \frac{1}{n_x} \sum_{i:x_i=x} \left(\frac{\mathbb{1}(t_i = 1) y_i}{\hat{e}(w_i)} - \frac{\mathbb{1}(t_i = 0) y_i}{1 - \hat{e}(w_i)} \right)$$

See, e.g., Abrevaya et al. (2015) and references therein

Questions:

1. What is the graphical intuition for how inverse probability weighting deals with confounding?
2. What do we model in IPW? What did we model in COM/GCOM estimation?

Conditional Outcome Modeling

Increasing Data Efficiency

Propensity Scores and IPW

Other Methods

Using both conditional outcome models and propensity score models

Model both $\mu(t, w)$ and $e(w)$

Using both conditional outcome models and propensity score models

Model both $\mu(t, w)$ and $e(w)$

Example:

$$\hat{\tau} = \frac{1}{n} \sum_i [\hat{\mu}(1, w_i) - \hat{\mu}(0, w_i)]$$

Using both conditional outcome models and propensity score models

Model both $\mu(t, w)$ and $e(w)$

Example:

$$\hat{\tau} = \frac{1}{n} \sum_i [\hat{\mu}(1, \hat{e}(w_i)) - \hat{\mu}(0, 1 - \hat{e}(w_i))]$$

Doubly robust methods

Doubly robust methods

- Model both $\mu(t, w)$ and $e(w)$

Doubly robust methods

- Model both $\mu(t, w)$ and $e(w)$
- Consistent if either $\hat{\mu}(t, w)$ or $\hat{e}(w)$ is consistent

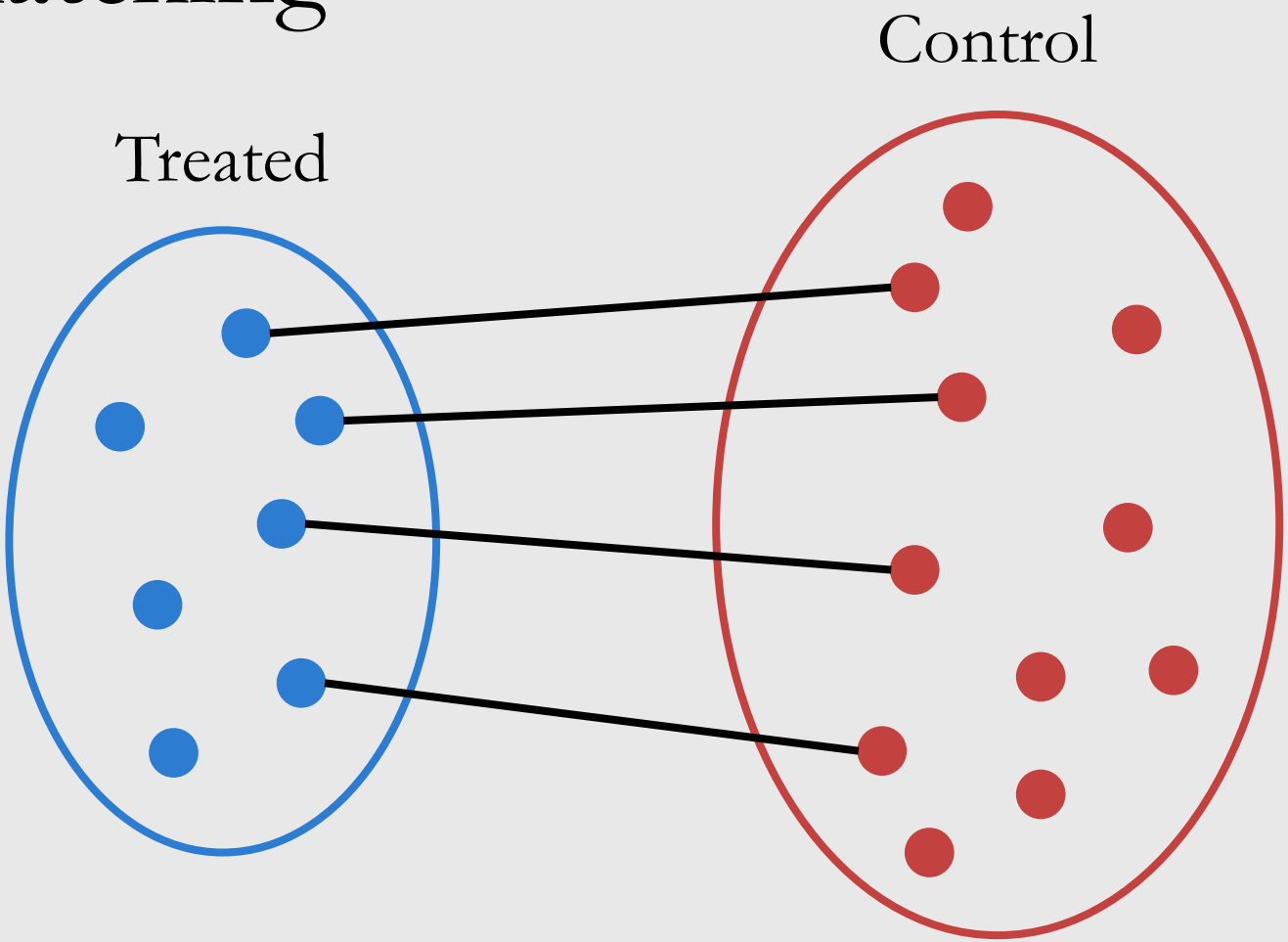
Doubly robust methods

- Model both $\mu(t, w)$ and $e(w)$
- Consistent if either $\hat{\mu}(t, w)$ or $\hat{e}(w)$ is consistent
- Theoretically converge to the estimand at a faster rate than COM/IPW

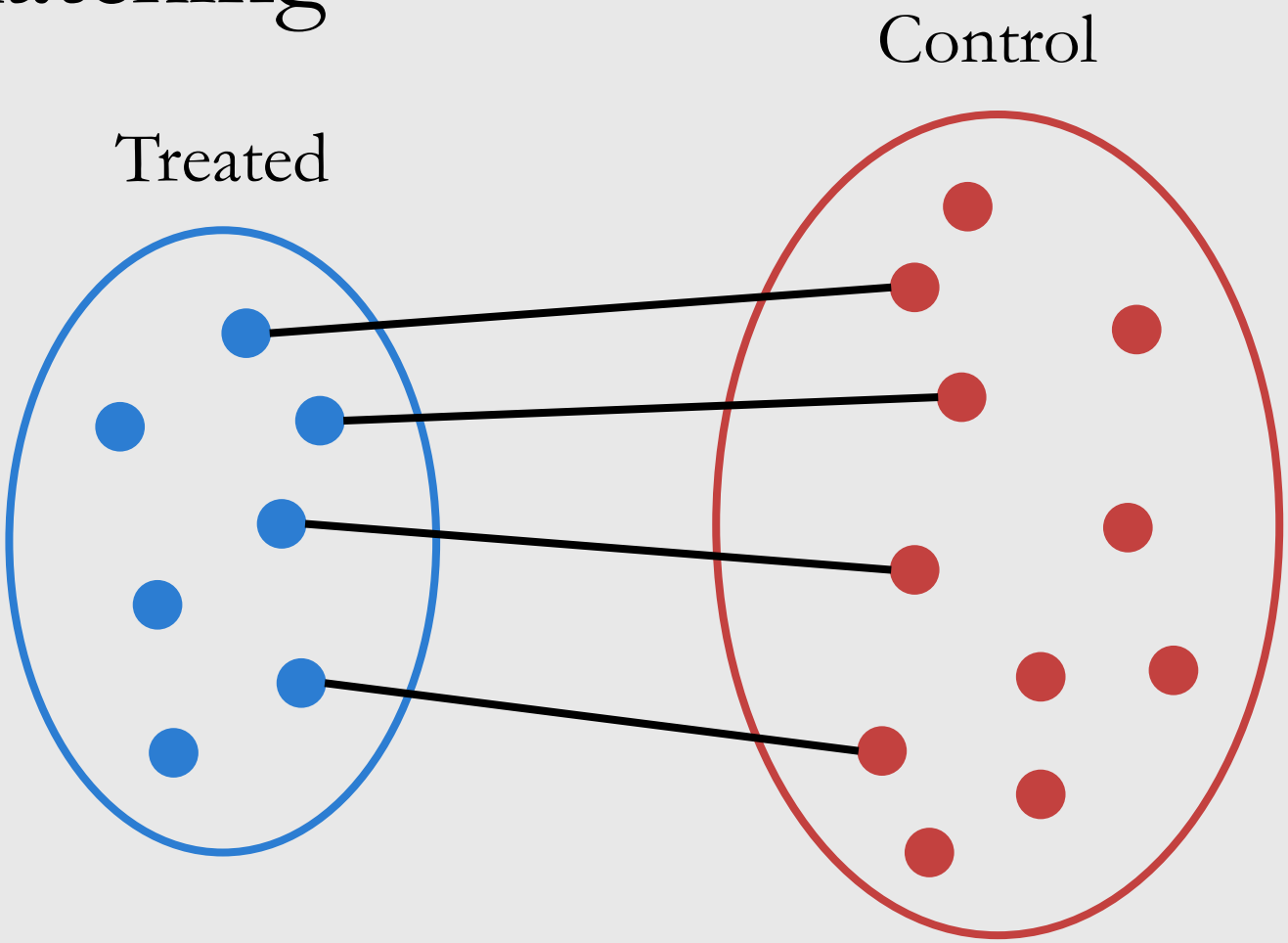
Doubly robust methods

- Model both $\mu(t, w)$ and $e(w)$
- Consistent if either $\hat{\mu}(t, w)$ or $\hat{e}(w)$ is consistent
- Theoretically converge to the estimand at a faster rate than COM/IPW
- See Section 7.7 of the course book for references to relevant papers

Matching

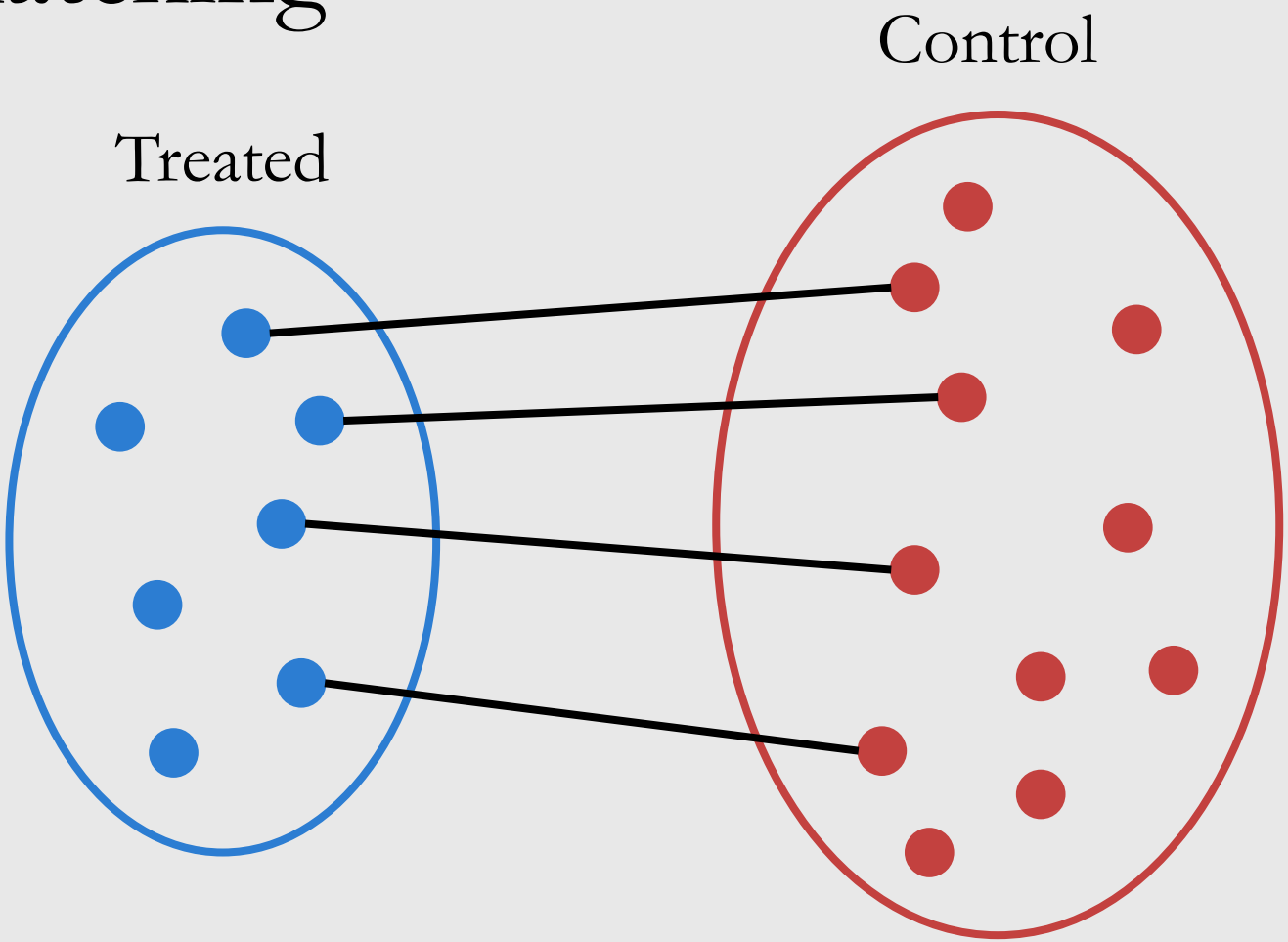


Matching



- Space: raw, coarsened, propensity score

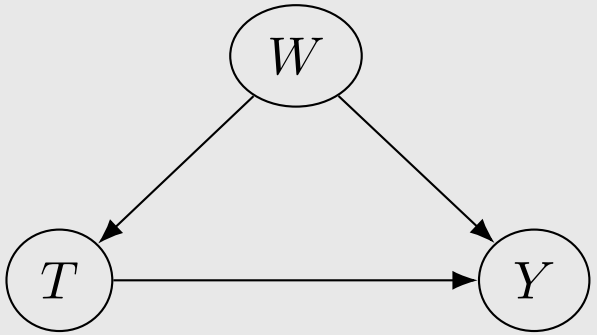
Matching



- Space: raw, coarsened, propensity score
- Different criteria for "close enough"

Double machine learning

Stage 1:

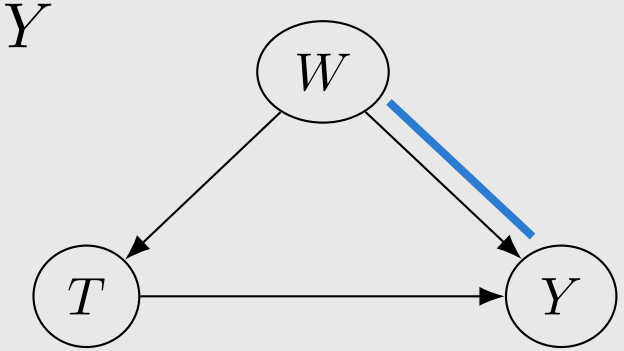


Stage 2:

Double machine learning

Stage 1:

- Fit a model to predict Y from W to get the predicted $\hat{Y}$

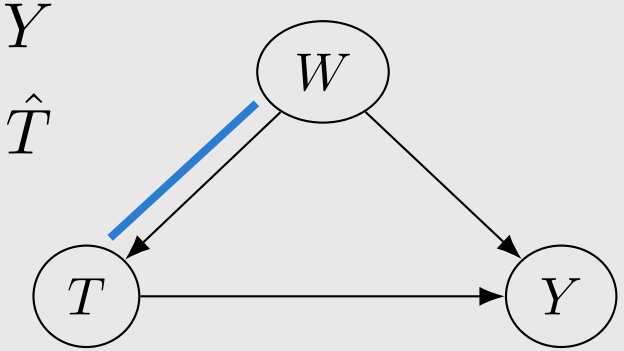


Stage 2:

Double machine learning

Stage 1:

- Fit a model to predict Y from W to get the predicted $\hat{Y}$
- Fit a model to predict T from W to get the predicted $\hat{T}$

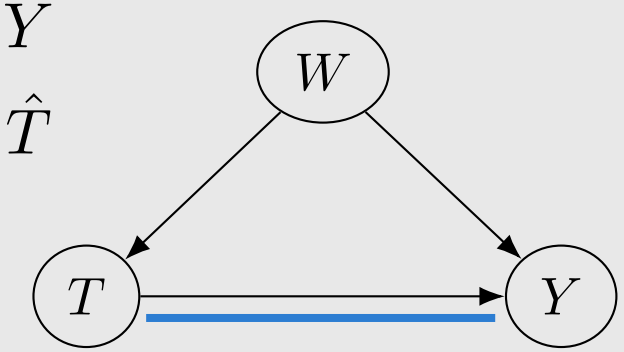


Stage 2:

Double machine learning

Stage 1:

- Fit a model to predict Y from W to get the predicted $\hat{Y}$
- Fit a model to predict T from W to get the predicted $\hat{T}$



Stage 2:

Partial out W by fitting a model to predict $Y - \hat{Y}$ from $T - \hat{T}$

Causal trees and forests

Flexible and yield valid confidence intervals (for sampling variability)