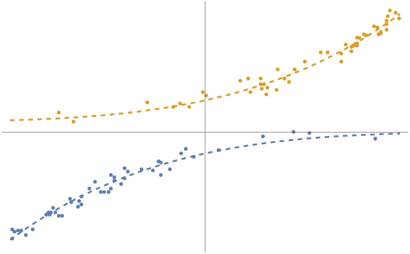


# Estimating Individual Treatment Effect:

## Generalization Bounds and Algorithms

Uri Shalit · Fredrik D Johansson · David Sontag

Presenters: Vahid & Tom

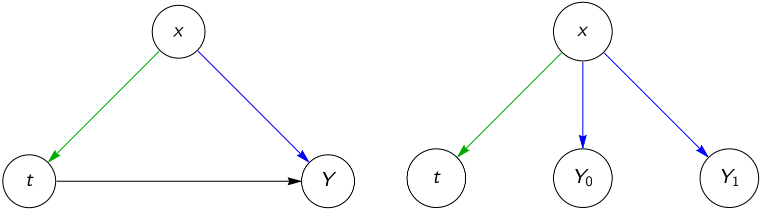
November 19, 2021

# Recall: Potential Outcomes

**Potential Outcomes**

For covariates $X$ and binary treatment $T$, potential outcomes $Y_0(x)$ and $Y_1(x)$ are defined as

$$Y_i(x) = \mathbb{E}[Y|X = x, do(T = i)]$$




(a) Causal graph for treatment-effect

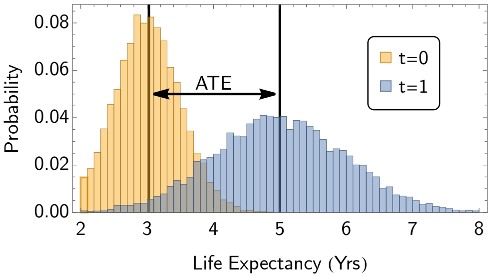
(b) Causal graph after replacing $Y$ with $Y_0$ and $Y_1$

# Recall: Average Treatment Effect

## Average Treatment Effect (ATE)

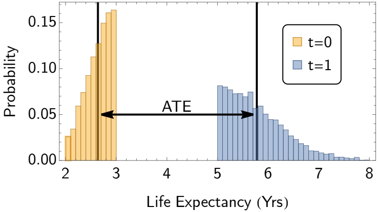
ATE is the difference in mean potential outcomes, i.e.,

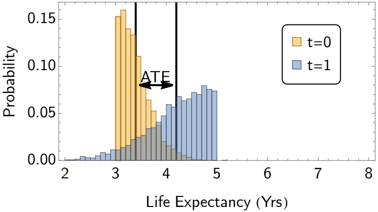
$$\text{ATE} \triangleq \mathbb{E}_x[Y_1(x)] - \mathbb{E}_x[Y_0(x)]$$



# ATE with Selection Bias

Selection bias may occur when treatment and control groups are not chosen randomly



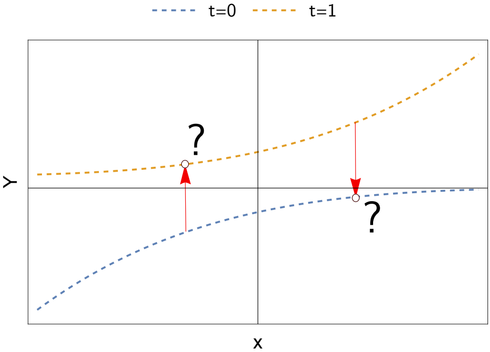



# Individual Treatment Effect

## Individual Treatment Effect (ITE)

Answers the question of how well does an **individual** $x$ respond to a treatment

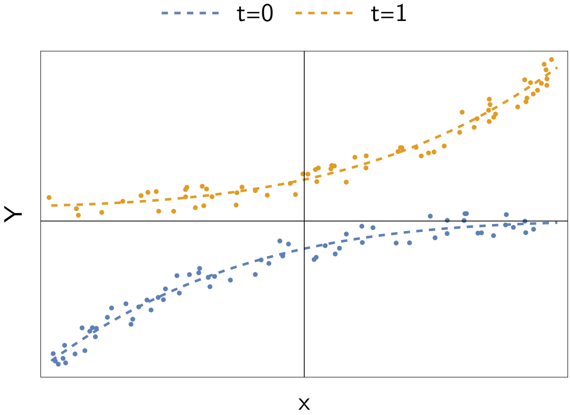
$$\text{ITE} \triangleq \mathbb{E}[Y_1|x] - \mathbb{E}[Y_0|x]$$



# Learning ITE

Learn $m_1(x) = \hat{Y}_1(x)$ and $m_0(x) = \hat{Y}_0(x)$ using supervised learning:

$$\boxed{\hat{\text{ITE}} = m_1(x) - m_0(x)}$$



# Learning ITE: Problems

- An individual $x$ is either treated or not

# Learning ITE: Problems

- An individual $x$ is either treated or not
- Observed data (factuals): $(x_i, t_i, Y_{t_i})$

# Learning ITE: Problems

- An individual $x$ is either treated or not
- Observed data (factuals): $(x_i, t_i, Y_{t_i})$
- Counterfactuals: $(x_i, t_i, Y_{1-t_i})$

# Learning ITE: Problems

- An individual $x$ is either treated or not
- Observed data (factuals): $(x_i, t_i, Y_{t_i})$
- Counterfactuals: $(x_i, t_i, Y_{1-t_i})$
- Need to know both to estimate ITE

# Learning ITE: Problems

- An individual $x$ is either treated or not
- Observed data (factuals): $(x_i, t_i, Y_{t_i})$
- Counterfactuals: $(x_i, t_i, Y_{1-t_i})$
- Need to know both to estimate ITE

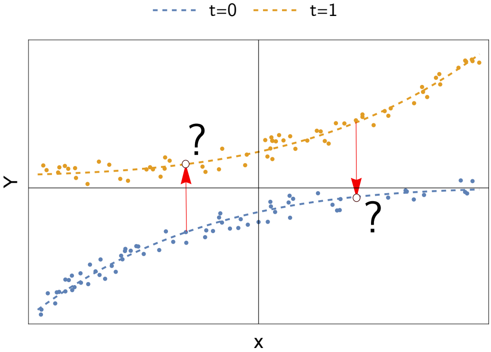



Figure: For each $x$, only one potential outcome is observed. One can use similar samples to estimate the other.

# Learning ITE: Problems

- In observational datasets $x$ and $t$ may not be independant

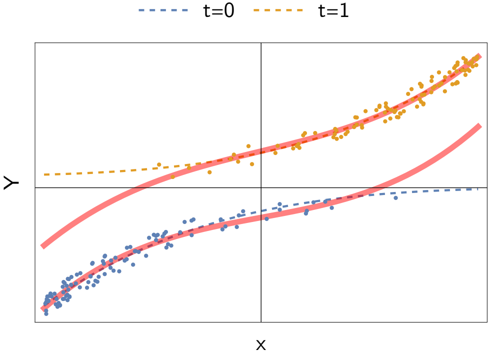
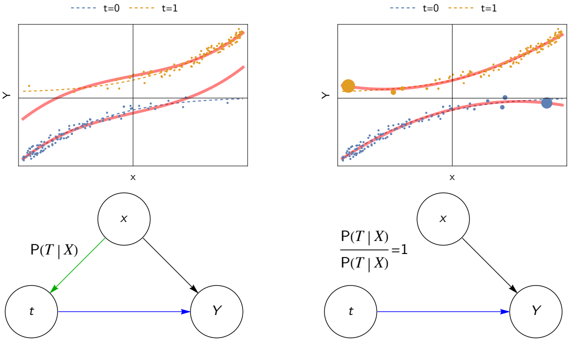



Figure: Induced selection bias from dependence between $x$ and $t$

# Inverse Propensity Score Weighting

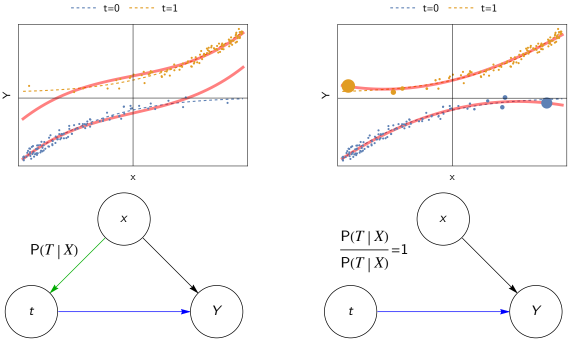
Turn observational data into pseudo-randomized trial data by re-weighting samples[1]




[1] Austin, 2011.

# Inverse Propensity Score Weighting

Turn observational data into pseudo-randomized trial data by re-weighting samples[1]

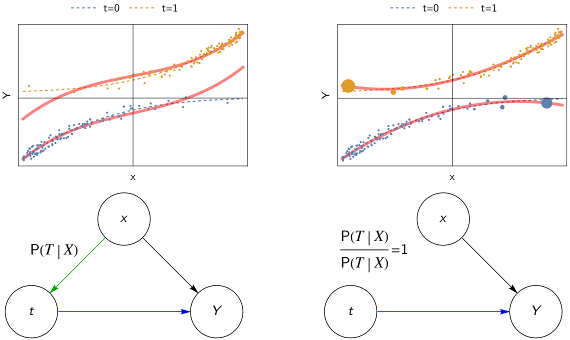



- Need to estimate $P(T|X)$, which is difficult for high-dim $X$

[1] Austin, 2011.

# Inverse Propensity Score Weighting

Turn observational data into pseudo-randomized trial data by re-weighting samples[1]

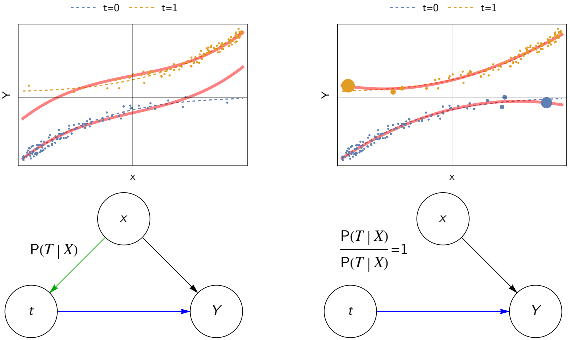



- Need to estimate $P(T|X)$, which is difficult for high-dim $X$
- Small $P(T|X)$ creates large variance

[1] Austin, 2011.

# Inverse Propensity Score Weighting

Turn observational data into pseudo-randomized trial data by re-weighting samples[1]




- Need to estimate $P(T|X)$, which is difficult for high-dim $X$
- Small $P(T|X)$ creates large variance
- Works primarily for ATE

[1] Austin, 2011.

# Unbiased Representation Learning

Learning a representation of data that removes the treatment bias[2]

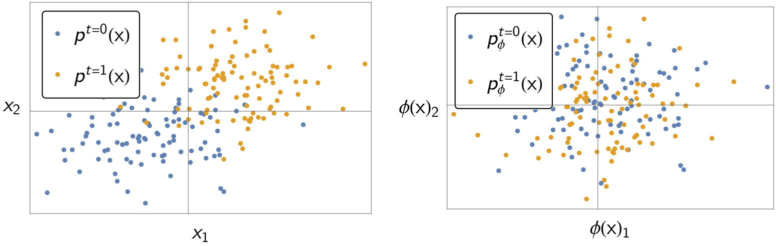



Figure: A 2D example of unbiased representation learning. Different colored dots represent treatment groups. (Left) shows sample locations in original feature space (Right) shows a possible unbiased representation encoded by some function $\phi$

[2] Johansson, Shalit, and Sontag, 2016.

# Neural network architecture for ITE estimation

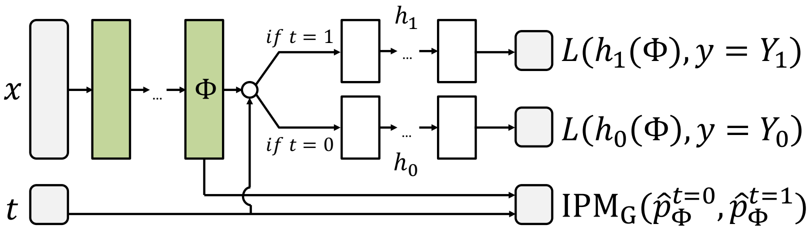
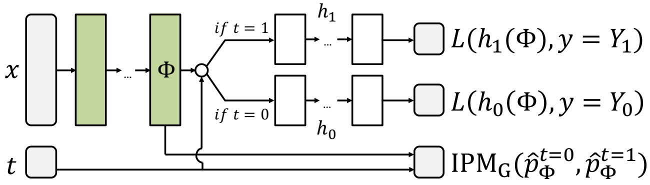



Figure: Neural network architecture for ITE estimation. $L$ is a loss function, $\mathrm{IPM}$ is an integral probability metric. Note that only one of $h_0$ and $h_1$ is updated for each sample during training[3].

[3]Shalit, Johansson, and Sontag, 2016.

# Integral Probability Metric (IPM) Regularizer

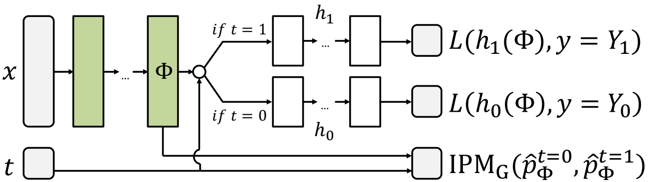



An IPM *is* a distance function between distributions[a]

- Minimizing the IPM between treatment groups encourages an unbiased representation

[a]Sriperumbudur et al., 2012.

# Integral Probability Metric (IPM) Regularizer




**Definition:** $\mathrm{IPM}_G$

$$\mathrm{IPM}_G(p_1, p_2) = \sup_{g \in C} \left| \int_s g(s)\,(p_1(s) - p_2(s))\,ds \right|$$

$G$ is some family of functions which defines the metric.

# Possible Metrics

For $\mathrm{IPM}_G$ to be a valid metric we must use a *sufficiently diverse* family of functions $G$ s.t

# Possible Metrics

For $\mathrm{IPM}_G$ to be a valid metric we must use a *sufficiently diverse* family of functions $G$ s.t

1. $\mathrm{IPM}_G(p,q) = 0 \Leftrightarrow p = q$ — identity of indiscernibles
2. $\mathrm{IPM}_G(p,q) = \mathrm{IPM}_G(q,p)$ — symmetry
3. $\mathrm{IPM}_G(p,q) \leq \mathrm{IPM}_G(p,r) + \mathrm{IPM}_G(r,q)$ — triangle inequality

# Possible Metrics

For $\text{IPM}_G$ to be a valid metric we must use a *sufficiently diverse* family of functions $G$ s.t

| | |
|---|---|
| 1. $\text{IPM}_G(p,q) = 0 \Leftrightarrow p = q$ | identity of indiscernibles |
| 2. $\text{IPM}_G(p,q) = \text{IPM}_G(q,p)$ | symmetry |
| 3. $\text{IPM}_G(p,q) \leq \text{IPM}_G(p,r) + \text{IPM}_G(r,q)$ | triangle inequality |

It is know that:

- $G := \{$ family of $1$-Lipschitz functions$\}$ then $\text{IPM}_G$ becomes the Wasserstein metric (earth movers distance)

# Possible Metrics

For $\text{IPM}_G$ to be a valid metric we must use a *sufficiently diverse* family of functions $G$ s.t

1. $\text{IPM}_G(p,q) = 0 \Leftrightarrow p = q$ — identity of indiscernibles
2. $\text{IPM}_G(p,q) = \text{IPM}_G(q,p)$ — symmetry
3. $\text{IPM}_G(p,q) \leq \text{IPM}_G(p,r) + \text{IPM}_G(r,q)$ — triangle inequality

It is know that:

- $G := \{$ family of 1-Lipschitz functions$\}$ then $\text{IPM}_G$ becomes the Wasserstein metric (earth movers distance)
- $G := \{$ any reproducing kernel Hilbert space with bounded norm $\}$ then $\text{IPM}_G$ becomes the Maximum Mean Discrepancy (MMD) metric

# Possible Metrics

For $\text{IPM}_G$ to be a valid metric we must use a *sufficiently diverse* family of functions $G$ s.t

1. $\text{IPM}_G(p,q) = 0 \Leftrightarrow p = q$ — identity of indiscernibles
2. $\text{IPM}_G(p,q) = \text{IPM}_G(q,p)$ — symmetry
3. $\text{IPM}_G(p,q) \leq \text{IPM}_G(p,r) + \text{IPM}_G(r,q)$ — triangle inequality

It is know that:

- $G :=$ { family of 1-Lipschitz functions} then $\text{IPM}_G$ becomes the Wasserstein metric (earth movers distance)
- $G :=$ { any reproducing kernel Hilbert space with bounded norm } then $\text{IPM}_G$ becomes the Maximum Mean Discrepancy (MMD) metric

Authors experiment with both Wasserstein and MMD as candidate metrics

# Error Bounds

Error in ITE Estimation:

$$\epsilon_{\text{ITE}} = \mathbb{E}_{p(x)}\left[(\hat{\text{ITE}}(x) - \text{ITE}(x))^2\right]$$

**Problem**: The true ITE is not known in real-world datasets

# Error Bounds

Error in ITE Estimation:

$$\epsilon_{\text{ITE}} = \mathbb{E}_{p(x)} \left[ (\hat{\text{ITE}}(x) - \text{ITE}(x))^2 \right]$$

**Problem**: The true ITE is not known in real-world datasets
All we know are factual errors

$$\epsilon_F^t = \mathbb{E}_{p(x)}[(Y_t(x) - \hat{Y}_t(x))^2]$$

# Error Bounds

Error in ITE Estimation:

$$\epsilon_{\text{ITE}} = \mathbb{E}_{p(x)}\left[(\hat{\text{ITE}}(x) - \text{ITE}(x))^2\right]$$

**Problem**: The true ITE is not known in real-world datasets
All we know are factual errors

$$\epsilon_F^t = \mathbb{E}_{p(x)}[(Y_t(x) - \hat{Y}_t(x))^2]$$

But not counterfactual errors

$$\epsilon_{CF}^t = \mathbb{E}_{p(x)}[(Y_{1-t}(x) - \hat{Y}_{1-t}(x))^2]$$

# Error Bounds

Error in ITE Estimation:

$$\epsilon_{\text{ITE}} = \mathbb{E}_{p(x)} \left[ (\hat{\text{ITE}}(x) - \text{ITE}(x))^2 \right]$$

**Problem**: The true ITE is not known in real-world datasets
All we know are factual errors

$$\epsilon_F^t = \mathbb{E}_{p(x)} [(Y_t(x) - \hat{Y}_t(x))^2]$$

But not counterfactual errors

$$\epsilon_{CF}^t = \mathbb{E}_{p(x)} [(Y_{1-t}(x) - \hat{Y}_{1-t}(x))^2]$$

**Factual Error Bound (Shalit et al.)**

$$\underbrace{\epsilon_{\text{ITE}}}_{\text{Effect error}} \leq \epsilon_F + \epsilon_{CF} \leq \boxed{2\underbrace{\left(\epsilon_F^{t=0} + \epsilon_F^{t=1}\right)}_{\text{Prediction error}} + \underbrace{\alpha \text{IPM}_G \left(p^{t=1}, p^{t=0}\right)}_{\text{Treatment/control distance}}}^{Loss}$$

# ITE Evaluation

- No ground truth: $Y_0$ and $Y_1$ are never observed for the same $x$

# ITE Evaluation

- No ground truth: $Y_0$ and $Y_1$ are never observed for the same $x$
- Use synthetic data, where structural equations are known

# ITE Evaluation

- No ground truth: $Y_0$ and $Y_1$ are never observed for the same $x$
- Use synthetic data, where structural equations are known
- Real-world data + randomized controlled trial
  - Still no ground truth
  - Evaluate the risk of induced policy $\pi_f(x) = \mathbb{I}(f(x,1) - f(x,0) > \lambda)$

# ITE Evaluation: IHDP

- Dataset: Semi-synthetic IHDP[4]
- Real-world features and treatment
- Synthetic outcome

**Out-of-sample**

| | IHDP | |
|---|---|---|
| | $\epsilon_{\text{ITE}}$ | $\epsilon_{\text{ATE}}$ |
| OLS/LR-1 | $5.8 \pm .3$ | $.94 \pm .06$ |
| OLS/LR-2 | $2.5 \pm .1$ | $.31 \pm .02$ |
| BLR | $5.8 \pm .3$ | $.93 \pm .05$ |
| $k$-NN | $4.1 \pm .2$ | $.79 \pm .05$ |
| BART | $2.3 \pm .1$ | $.34 \pm .02$ |
| RAND.FOR. | $6.6 \pm .3$ | $.96 \pm .06$ |
| CAUS.FOR. | $3.8 \pm .2$ | $.40 \pm .03$ |
| BNN | $2.1 \pm .1$ | $.42 \pm .03$ |
| TARNET | $\mathbf{.95 \pm .0}$ | $.28 \pm .01$ |
| CFR MMD | $\mathbf{.78 \pm .0}$ | $.31 \pm .01$ |
| CFR WASS | $\mathbf{.76 \pm .0}$ | $.27 \pm .01$ |

Table: Results on IHDP. Lower is better.

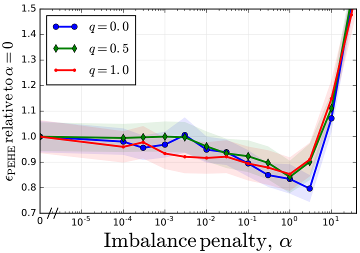
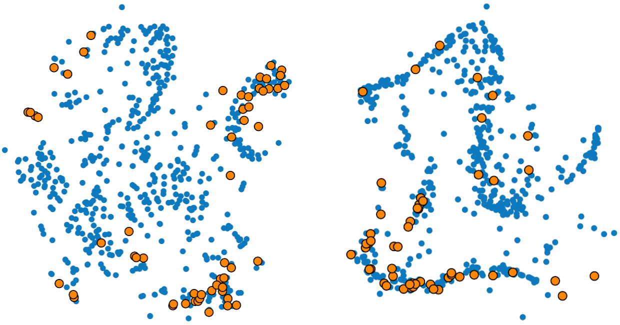



Figure: Out-of-sample ITE error versus IPM regularization for CFR Wass, with high (q = 1), medium and low (artificial) imbalance between control and treated.

[4] Hill, 2011.

# ITE Evaluation: Balanced Representation



(a) Original dataset (b) Wasserstein regularizer

Figure: t-SNE visualizations of the balanced representations of IHDP learned by the algorithm. Blue (orange) points represent control (treatment) group.

## Implications

+ state of the art method for estimating ITE from observational data; can use general functions to estimate $Y_0$ and $Y_1$

**Implications**

- state of the art method for estimating ITE from observational data; can use general functions to estimate $Y_0$ and $Y_1$
- Useful for discovering new outcomes. For example:

## Implications

- + state of the art method for estimating ITE from observational data; can use general functions to estimate $Y_0$ and $Y_1$
- + Useful for discovering new outcomes. For example:
  - Is a drug intended for cancer patients useful for treating alzheimer's patients?

## Implications

- state of the art method for estimating ITE from observational data; can use general functions to estimate $Y_0$ and $Y_1$
- Useful for discovering new outcomes. For example:
  - Is a drug intended for cancer patients useful for treating alzheimer's patients?
- Provides the first generalization-error bound for the expected ITE estimation error; useful for developing further theory

## Implications

+ state of the art method for estimating ITE from observational data; can use general functions to estimate $Y_0$ and $Y_1$
+ Useful for discovering new outcomes. For example:
  - Is a drug intended for cancer patients useful for treating alzheimer's patients?
+ Provides the first generalization-error bound for the expected ITE estimation error; useful for developing further theory

## Implications

+ state of the art method for estimating ITE from observational data; can use general functions to estimate $Y_0$ and $Y_1$
+ Useful for discovering new outcomes. For example:
  - Is a drug intended for cancer patients useful for treating alzheimer's patients?
+ Provides the first generalization-error bound for the expected ITE estimation error; useful for developing further theory

## Limitations

- Assumes no hidden confounding variables

**Implications**

+ state of the art method for estimating ITE from observational data; can use general functions to estimate $Y_0$ and $Y_1$
+ Useful for discovering new outcomes. For example:
  - Is a drug intended for cancer patients useful for treating alzheimer's patients?
+ Provides the first generalization-error bound for the expected ITE estimation error; useful for developing further theory

**Limitations**

- Assumes no hidden confounding variables
- No results on tightness of bound

## Implications

+ state of the art method for estimating ITE from observational data; can use general functions to estimate $Y_0$ and $Y_1$
+ Useful for discovering new outcomes. For example:
  - Is a drug intended for cancer patients useful for treating alzheimer's patients?
+ Provides the first generalization-error bound for the expected ITE estimation error; useful for developing further theory

## Limitations

- Assumes no hidden confounding variables
- No results on tightness of bound
  - Does minimizing the bound always results in smaller error?

**Implications**

+ state of the art method for estimating ITE from observational data; can use general functions to estimate $Y_0$ and $Y_1$
+ Useful for discovering new outcomes. For example:
  - Is a drug intended for cancer patients useful for treating alzheimer's patients?
+ Provides the first generalization-error bound for the expected ITE estimation error; useful for developing further theory

**Limitations**

- Assumes no hidden confounding variables
- No results on tightness of bound
  - Does minimizing the bound always results in smaller error?
- Generalization to continuous treatments is not obvious

Austin, Peter (May 2011). "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies". In: *Multivariate behavioral research* 46, pp. 399–424. DOI: 10.1080/00273171.2011.568786.

Johansson, Fredrik D., Uri Shalit, and David Sontag (May 2016). "Learning Representations for Counterfactual Inference". In: *arXiv e-prints*, arXiv:1605.03661, arXiv:1605.03661. arXiv: 1605.03661 [stat.ML].

Shalit, Uri, Fredrik D. Johansson, and David Sontag (June 2016). "Estimating individual treatment effect: generalization bounds and algorithms". In: *arXiv e-prints*, arXiv:1606.03976, arXiv:1606.03976. arXiv: 1606.03976 [stat.ML].

Sriperumbudur, Bharath K et al. (2012). "On the empirical estimation of integral probability metrics". In: *Electronic Journal of Statistics* 6.none, pp. 1550–1599. DOI: 10.1214/12-EJS722. URL: https://doi.org/10.1214/12-EJS722.

Hill, Jennifer L (2011). "Bayesian nonparametric modeling for causal inference". In: *Journal of Computational and Graphical Statistics* 20.1, pp. 217–240.