Reliable Machine Learning for Individualized Treatment Effect Estimation

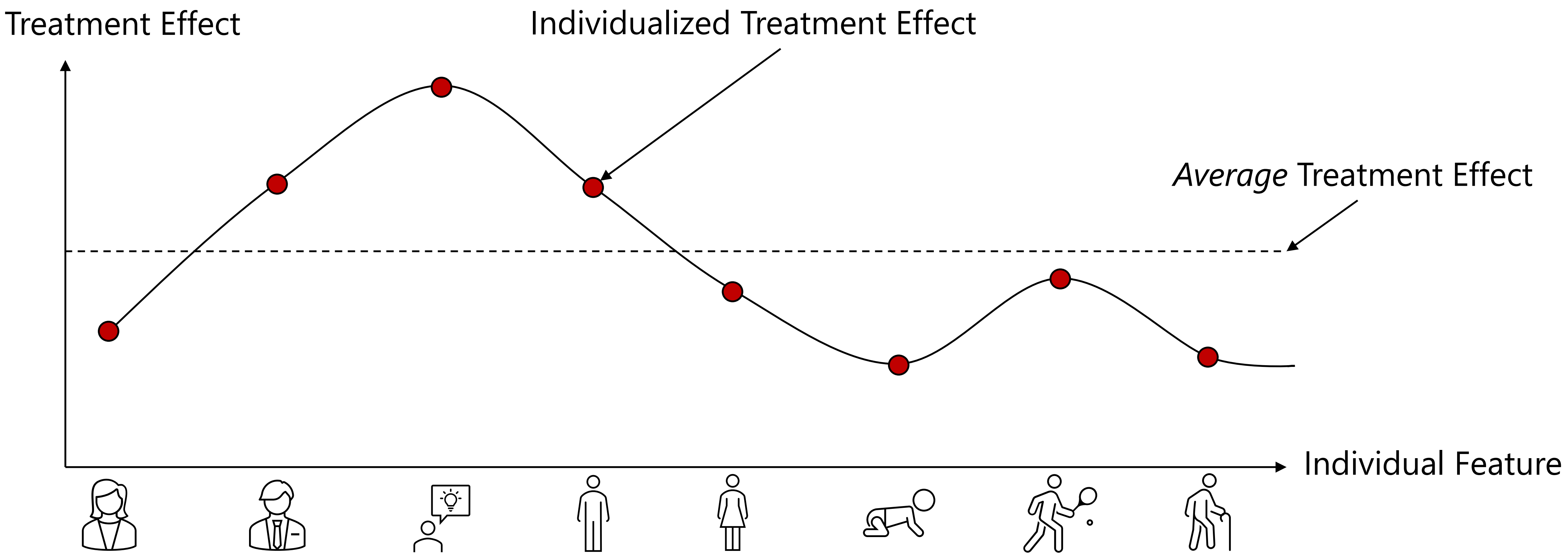
Miruna Oprescu

Cornell University, Cornell Tech

Committee: Nathan Kallus (Chair), Sarah Dean, Peter Frazier, Emma Pierson

Individualized Treatment Effects

- Applications in economics, healthcare, e-commerce, online platforms.
- Example: Treatment effect of 401(k) eligibility on net worth.



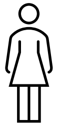
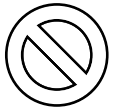
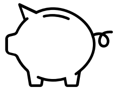
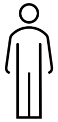
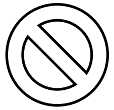
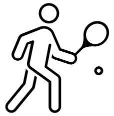
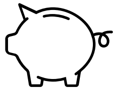
Standard Causal Inference Setting

- Treatment $A \in \{0, 1\}$, covariates $X \in \mathcal{X}$, potential outcomes $Y(0), Y(1) \in \mathbb{R}$.
- We want to estimate the conditional average treatment effect (CATE):

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

- But we only observe data: $Z_i = (X_i, A_i, Y_i) \sim (X, A, Y(A))$.

Example:
Effect of 401(k) eligibility on net worth.

Unit: X	Treatment: A	$Y(0)$ (in $$$$$)	$Y(1)$ (in $$$$$)
		**4**	6
		3	**7**
		**7**	9
		1	**5**

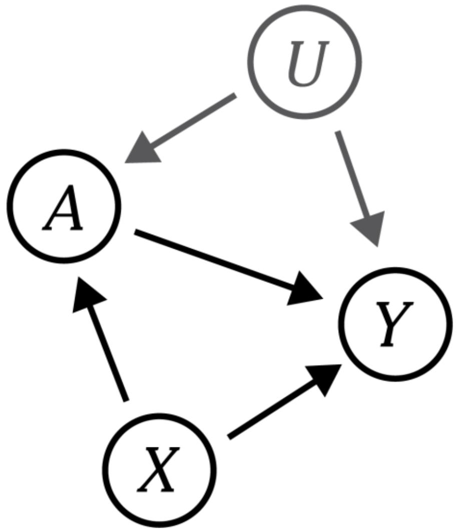
Standard Causal Inference Setting

- Most works assume ignorability (unconfoundedness):

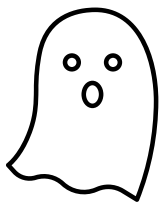
$$Y(0), Y(1) \perp A \mid X, \text{ i.e., } U = \emptyset.$$

Then, they *identify* the CATE $\tau(x)$ from data as:

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



- Two issues with this approach:
  1. Assumes effects are centered around the conditional mean and/or the mean is informative.
  2. Ignorability is an untestable assumption!



Talk Overview

1. Beyond Conditional Averages: Robust and Agnostic Learning of Conditional Distributional Treatment Effects
    - N. Kallus, **M. Oprescu**. AISTATS 2023.

2. Sharp and Efficient Bounds on Heterogeneous Causal Effects Under Hidden Confounding
    - **M. Oprescu**, J. Dorn, M. Ghoummaid, A. Jesson, N. Kallus, U. Shalit. ICML 2023.

3. Research Roadmap: Future Directions and Goals

Talk Overview

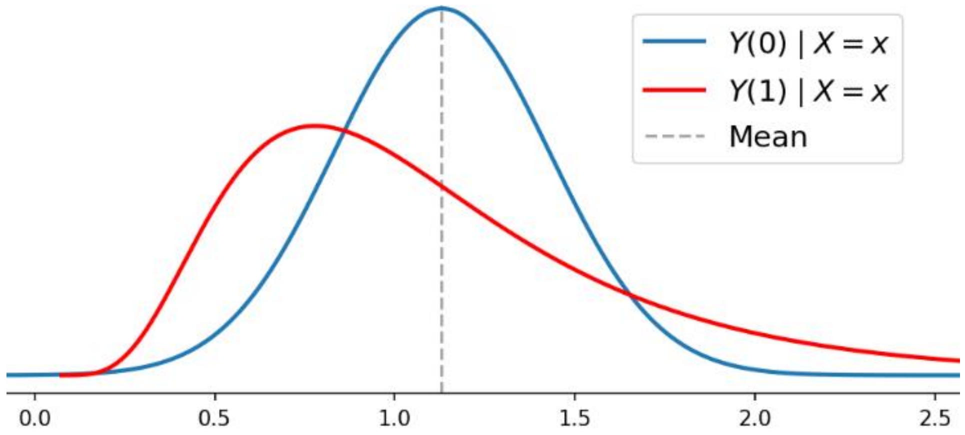
1. Beyond Conditional Averages: Robust and Agnostic Learning of Conditional Distributional Treatment Effects
   - N. Kallus, **M. Oprescu**. AISTATS 2023.

2. Sharp and Efficient Bounds on Heterogeneous Causal Effects Under Hidden Confounding
   - **M. Oprescu**, J. Dorn, M. Ghoummaid, A. Jesson, N. Kallus, U. Shalit. ICML 2023.

3. Research Roadmap: Future Directions and Goals

Beyond Conditional Averages: Motivation

- Skewed outcome functions (e.g., income, platform usage)
- Equity considerations and risk quantification



Potential outcomes with the same conditional mean but different tail effects.

- Beyond the conditional mean effect:
  Conditional Distributional Treatment Effects (**CDTEs**)

Beyond Conditional Averages: CDTEs

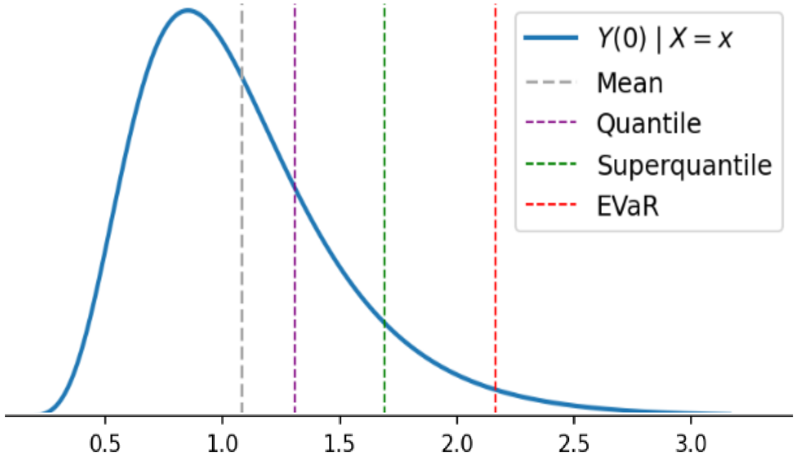
- For any distribution statistic $\kappa^*(F)$:

$$CDTE(X) = \kappa^*\left(F_{Y(1)|X}\right) - \kappa^*\left(F_{Y(0)|X}\right)$$

- Examples:
  - Conditional Average (CATE)
  - Conditional Quantiles (CQTE)
  - Conditional Superquantiles (CSQTE)

    Also known as Conditional-Value-at-Risk (CVaR)
  - f-risk measures from f-divergences (CfRTE)

    E.g., Entropic-Value-at-Risk (EVaR) from the KL divergence

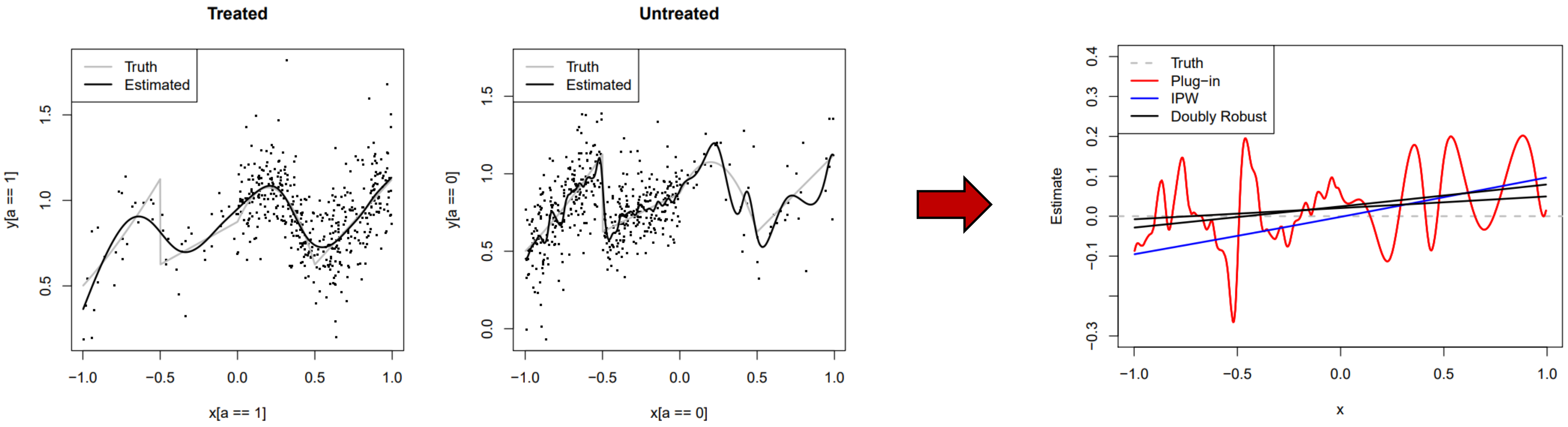


Quantile, superquantile, EVaR ($\tau = 0.75$).

CDTE Plugin Estimator

$$CDTE^{Plugin}(X) = \hat{\kappa}_1(X) - \hat{\kappa}_0(X)$$

- Weaknesses:
  - Can obscure the signal when the $\hat{\kappa}_a(X)$'s are more complex than the CDTE.
  - Not robust: difference of best estimators $\neq$ best estimator of difference.



Plugin bias illustration for CATE estimators (Kennedy, 2020).

CDTEs: General Framework

- Consider statistics that solve moment equations:

$$\mathbb{E}_F[\rho(Y,\kappa,h)] = \mathbf{0}$$

where $h^*(F)$ is a set of nuisances.

- Examples
  - Average: $\rho(y,\mu) = y - \mu$
  - Quantiles (level τ): $\rho(y,q) = \tau - \mathbb{I}[y \leq q]$
  - Superquantiles (level τ):

$$\rho(y,\mu,q) = \left((1-\tau)^{-1}y\mathbb{I}[y \geq q], \tau - \mathbb{I}[y \leq q]\right) \in \mathbb{R}^2$$

A Two-Step Procedure for CDTE Robust Estimation

1. Consider a pseudo-outcome* that targets the effect directly:

$$\psi(Z,\hat{e},\hat{\alpha},\hat{\nu}) = \underbrace{\hat{\kappa}_1(X) - \hat{\kappa}_0(X)}_{\text{plugin estimator}} - \underbrace{\frac{A-\hat{e}(X)}{\hat{e}(X)\big(1-\hat{e}(X)\big)}\hat{\alpha}_A(X)^T\rho(Y,\hat{\nu}_A(X))}_{\text{bias correction}}$$

   where $e(X) = P(A=1 \mid X)$, $\nu_a = (\kappa_a, h_a)$ and $\alpha_a(X)$ are additional nuisances learned on one sample.

2. Regress $\psi(Z,\hat{e},\hat{\alpha},\hat{\nu})$ on features $X \in \mathcal{X}$ in another sample.

Algorithm 1 CDTE Learner

Input: Data $\{(X_i, A_i, Y_i) : i \in \overline{1,n}\}$, folds $K \geq 2$, nuisance estimators, regression learner

for $k \in \overline{1,K}$ **do**
  Use data $\{(X_i, A_i, Y_i) : i \neq k-1 \pmod K\}$ to construct nuisance estimates $\hat{e}^{(k)}, \hat{\alpha}^{(k)}, \hat{\nu}^{(k)}$
  for $i = k-1 \pmod K$ **do** set $\widehat{\psi}_i = \psi(Z_i, \hat{e}^{(k)}, \hat{\alpha}^{(k)}, \hat{\nu}^{(k)})$ **end for**
end for
return $\widehat{\mathrm{CDTE}}(x) = \widehat{\mathbb{E}}_n[\widehat{\psi} \mid X = x]$

* Derived from the efficient influence function (EIF) of $\mathbb{E}_F[CDTE(X)]$.

CDTE Estimator Guarantees

Robustness:

- The error has a product structure so small errors in the nuisances lead to second-order errors in the CDTE estimates.

  E.g., if all nuisances are estimated at a rate of at least $O\left(n^{-1/4}\right)$

  CDTEs are estimated at the rate $O\left(n^{-1/2}\right)$.

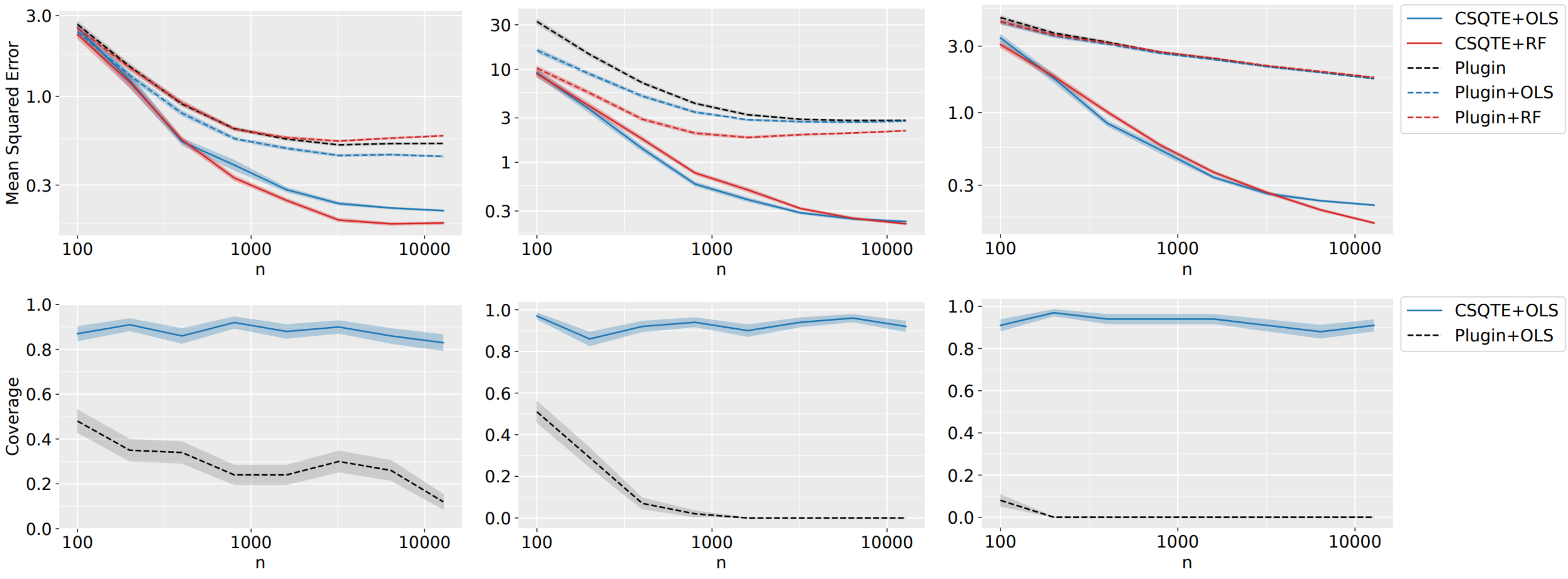
- There are many chances at consistency when some of the nuisances are misspecified.

Model Agnostic:

- Linear regression parameters are asymptotically normal with oracle variance

  I.e., if we use OLS as the final stage, the confidence intervals are valid.

Empirical Example: CSQTE

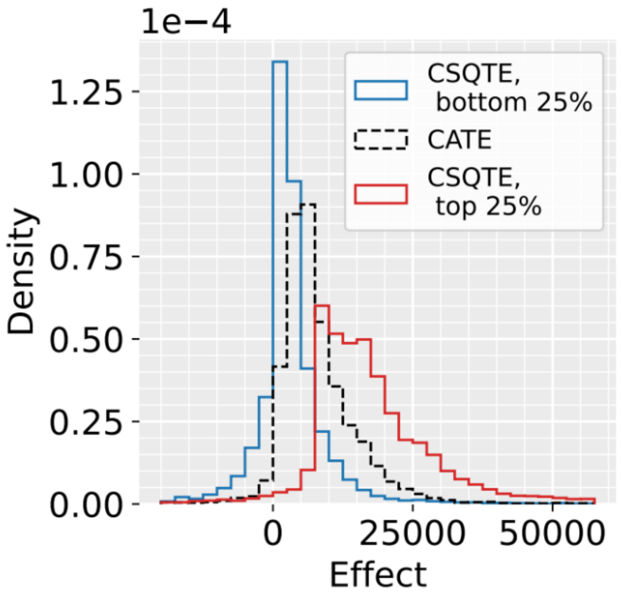


(a) Flexible $\hat{\mu}$ (b) Misspecified $\hat{\mu}$ (c) Slow $\hat{\mu}$

Performance of CSQTE learner with flexible, misspecified or slow converging superquantile estimator $\hat{\mu}$.
Second stages: flexible = Random Forest, misspecified = OLS, slow = Gaussian Kernel.

Case Study: Effect of 401(k) Eligibility

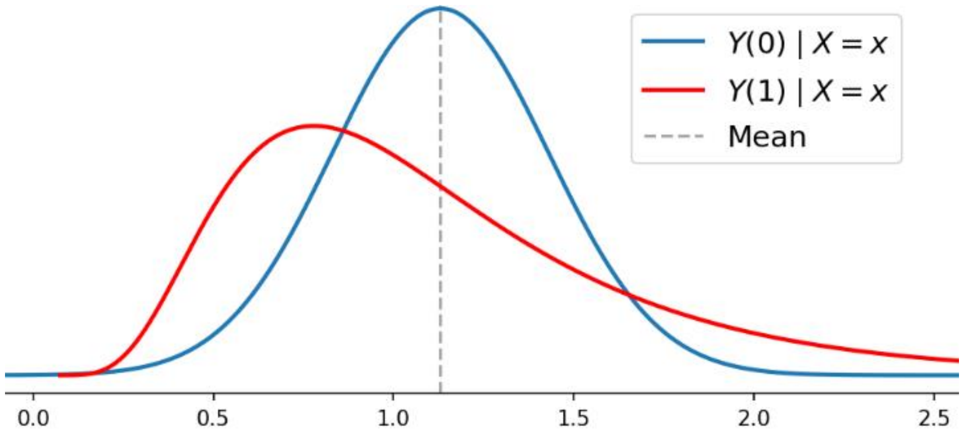
- Effect of 401(k) eligibility on net worth
- CSQTE on bottom and top 25% asset holders



Coefficient	CSQTE Bottom 25%	CATE	CSQTE Top 25%
Intercept ($10,000)	−0.021 (−1.06, 1.02)	−0.95 (−2.42, 0.51)	−2.07 (−7.04, 2.90)
Income	0.25** (0.08, 0.43)	0.21 (−0.08, 0.50)	−0.05 (−1.12, 1.01)
Age	105 (−75, 286)	232** (24, 441)	513 (−182, 1210)
Education	−801** (−1440, −164)	16 (−1050, 1090)	1340 (−2490, 5180)

Left: distribution of CATEs and CSQTEs with random forest last stage.
Right: linear regression coefficients with OLS final stage.

Beyond Conditional Averages: TL;DR



Potential outcomes with the same conditional mean but different tail effects.

- When outcome distributions are skewed, it's essential to consider measures beyond conditional averages.

  E.g.: quantiles, superquantiles, f-risk measures.

- We propose an ML method that enables reliable CDTE estimation by adapting to the complexity of the treatment effects, rather than just baseline functions.

Talk Overview

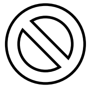
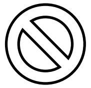
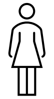
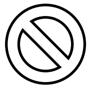
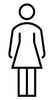
1. Beyond Conditional Averages: Robust and Agnostic Learning of Conditional Distributional Treatment Effects
    - N. Kallus, **M. Oprescu**. AISTATS 2023.

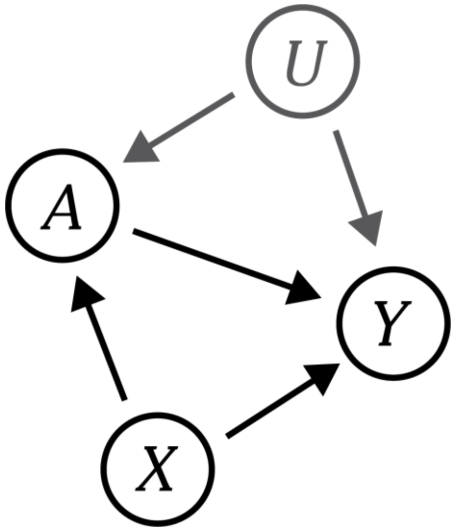
2. Sharp and Efficient Bounds on Heterogeneous Causal Effects Under Hidden Confounding
    - **M. Oprescu**, J. Dorn, M. Ghoummaid, A. Jesson, N. Kallus, U. Shalit. ICML 2023.

3. Research Roadmap: Future Directions and Goals

Hidden Confounding

Example:
- Effect of 401(k) eligibility on net worth.
- $\tau(x) = 0$

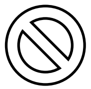
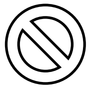
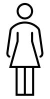
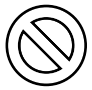
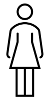
X	A	$Y(0)$	$Y(1)$
		1	**1**
		3	**3**
		**2**	2
		8	**8**
		**9**	9
		**7**	7

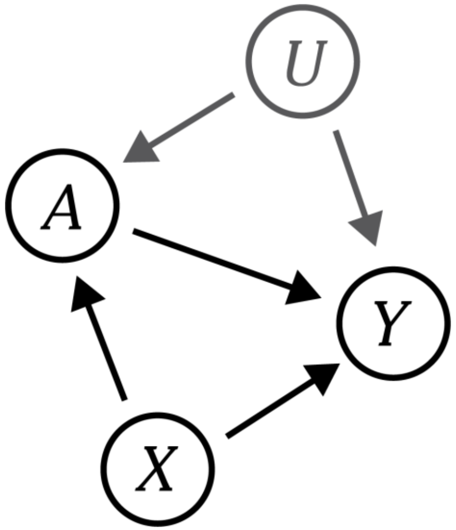


- Assuming no unobserved confounding,

$$\hat{\tau}(x) = \frac{2+9+7}{3} - \frac{1+8+3}{3} = 2$$

Hidden Confounding

U	X	A	$Y(0)$	$Y(1)$
1			1	**1**
1			3	**3**
1			**2**	2
0			8	**8**
0			**9**	9
0			**7**	7



Example:

- Effect of 401(k) eligibility on net worth.
- $\tau(x) = \tau(x, u) = 0$

- Assuming no *other* unobserved confounding,

$$\hat{\tau}(x, 1) = 2 - \frac{1+3}{2} = 0, \qquad \hat{\tau}(x, 0) = 8 - \frac{7+9}{2} = 0$$

Sensitivity Models for Hidden Confounding

What if we make assumptions about the strength of the unobserved confounding U?

- Let $e(x) = P(A = 1 \mid X = x),\ e(x,u) = P(A = 1 \mid X = x, U = u)$.
- Marginal Sensitivity Model (MSM) (Tan, 2006): Assume

$$\Lambda^{-1} \leq \frac{e(x,u)}{1 - e(x,u)} \Big/ \frac{e(x)}{1 - e(x)} \leq \Lambda$$

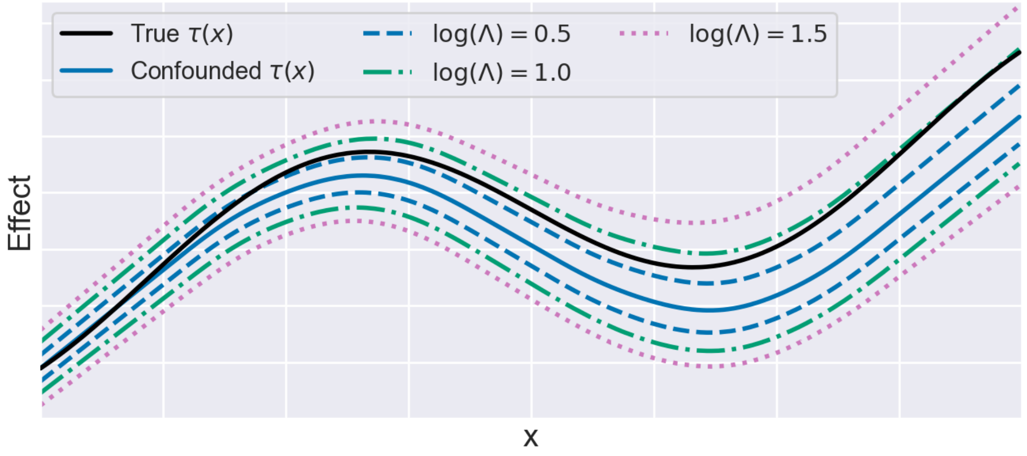
  for a user-specified Λ.

  - Can be seen as an odds ratio.
  - Ratio can be replaced with a divergence between $e(x)$ and $e(x,u)$ to obtain other sensitivity models.

- Under the MSM, we can identify *informative bounds* $\tau^{+}(x), \tau^{-}(x)$ on $\tau(x)$.

MSM How-To Guide

1. Estimate/pick a Λ and obtain bounds on the CATE.



Example of CATE bounds for different values of Λ.

2. Find the value of Λ where the treatment effects change sign.
   - Cornfield et al. (1959): studied the effect of smoking on lung cancer. Found confounding had to be 9 times larger in smokers (Λ=9) than in non-smokers to negate the measured effect.

Bounds Identification Under The MSM

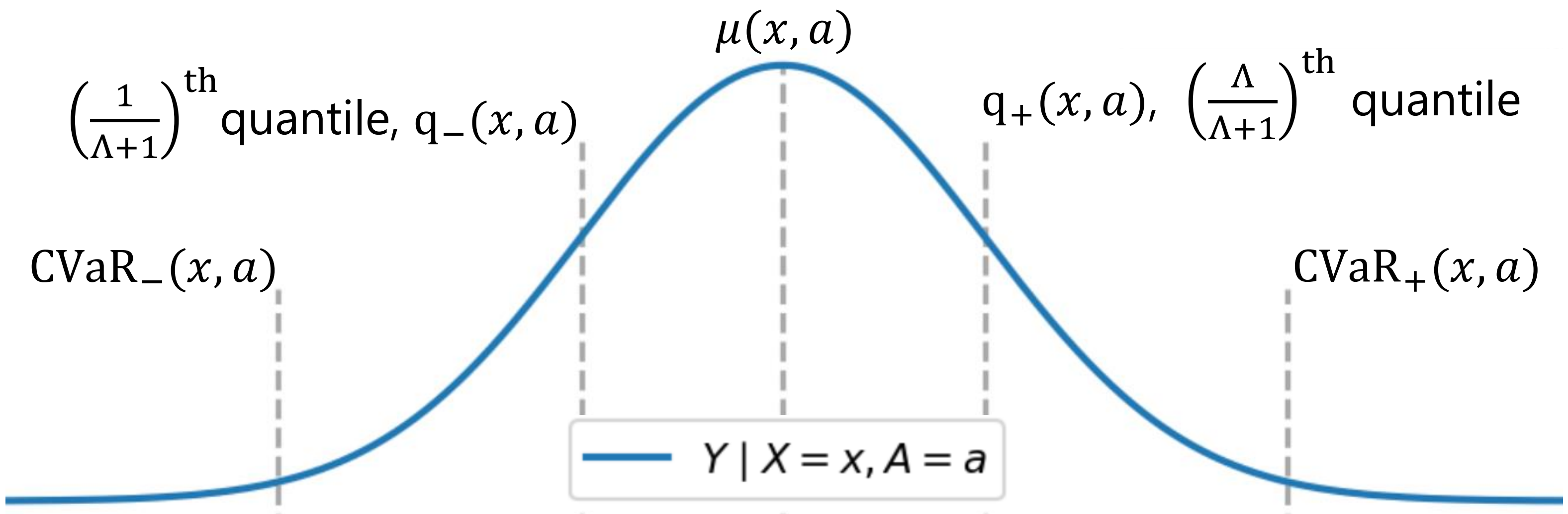
Result 1 (Dorn et al., 2021). $\mu(x,a) = \mathbb{E}[Y \mid X = x, A = a]$ and $Y^{\pm}(x,a)$ is the upper (+)/ lower (-) sharp bound of $\mathbb{E}[Y(a) \mid X = x]$. Then:

$$\boldsymbol{Y^+(x,1)} = e(x)\mu(x,1) + \big(1 - e(x)\big)\rho_+(x,1)$$

$$\boldsymbol{Y^-(x,0)} = \big(1 - e(x)\big)\mu(x,0) + e(x)\rho_-(x,0)$$

$$\boldsymbol{\tau^+(x)} = \boldsymbol{Y^+(x,1)} \text{ - } \boldsymbol{Y^-(x,0)}$$

where $\rho_{\pm}(x,a) = \Lambda^{-1}\mu(x,a) + (1 - \Lambda^{-1})CVaR_{\pm}(x,a)$.



B-Learner: Bound Estimation Under The MSM

- Plug-in estimator: estimate $e(\mathrm{x}), \mu(x,a), \rho_{\pm}(x,a)$ and "plug" them into $Y^{\pm}(x,a)$:

$$\hat{\tau}^{+}_{\mathrm{Plugin}}(x) = \hat{Y}^{+}(x,1) - \hat{Y}^{-}(x,0)$$

B-Learner: Bound Estimation Under The MSM

- A two-step procedure for robust and reliable estimation:
  1. Learn nuisances $\hat{\eta}(x) = (\hat{e}(\mathrm{x}), \hat{\mu}(x,a), \hat{\rho}_{\pm}(x,a))$ on one sample.
  2. Correct bias in another sample using insights from CDTE estimation and regress the pseudo-outcome $\phi_\tau^+(Z,\hat{\eta})$ on features $X \in \mathcal{X}$.

Algorithm 1 The B-Learner

input Data $\{(X_i, A_i, Y_i) : i \in \{1, \ldots, n\}\}$, folds $K \geq 2$, nuisance estimators, regression learner $\widehat{\mathbb{E}}_n$

1: **for** $k \in \{1, \ldots, K\}$ **do**

2: Use data $\{(X_i, A_i, Y_i) : i \neq k-1 \pmod K\}$ to construct nuisance estimates $\widehat{\eta}^{(k)} = (\widehat{e}^{(k)}, \widehat{q}^{(k)}, \widehat{\rho}^{(k)})$

3: **for** $i = k-1 \pmod K$ **do**

4: Set $\widehat{\phi}^+_{\tau,i} = \phi^+_\tau(Z_i, \widehat{\eta}^{(k)})$

5: **end for**

6: **end for**

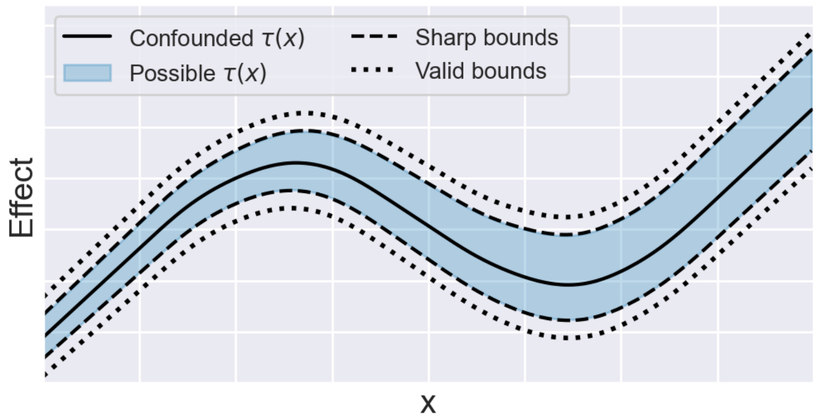
output $\widehat{\tau}^+(x) = \widehat{\mathbb{E}}_n[\widehat{\phi}^+_\tau \mid X = x]$

Theoretical Guarantees

- The (unsigned) bias from the first stage is:

$$\mathcal{E}(x) = \Sigma_{a=0}^{1}(|\hat{e}(x) - e(x)||\hat{\rho}(x,a) - \rho(x,a)| + \left(\hat{q}(x,a) - q(x,a)\right)^2)$$

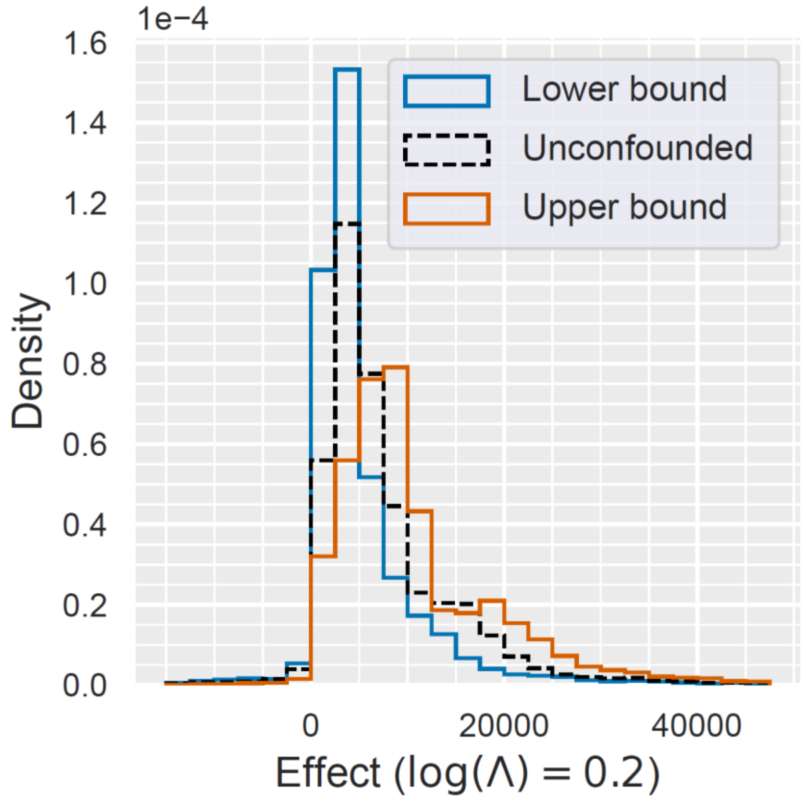
- For an **ERM**-based final stage estimator, the B-Learner deviates from the oracle estimator by $||\mathcal{E}(x)||_2$
- Corollaries:
  1. **Sharpness:** $\hat{q}$ and either $\hat{e}$ or $\hat{\rho}$ are consistent $\Rightarrow \hat{\tau}^{+}(x)$ consistent.
  2. **Validity:** $\hat{q}$ is inconsistent $\Rightarrow$ bounds still **valid** on average.
  2. **Efficiency:** If nuisances are $o_P\left(n^{-1/2(2+r)}\right)$, error is dominated by target class complexity.

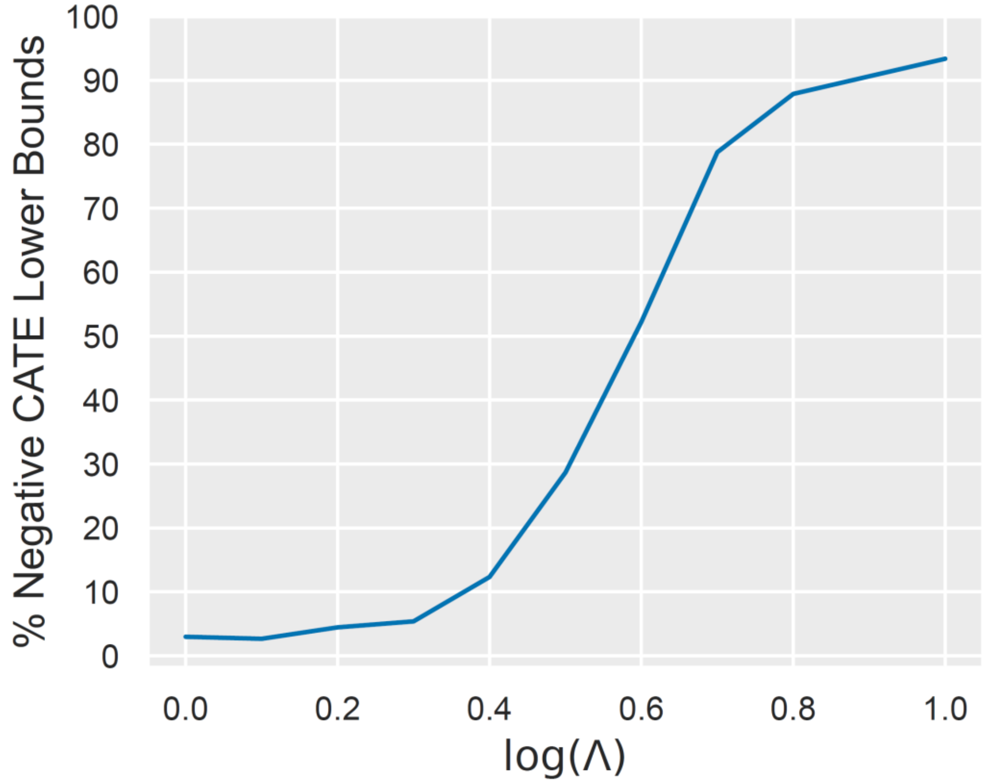


Example of sharp and valid bounds.

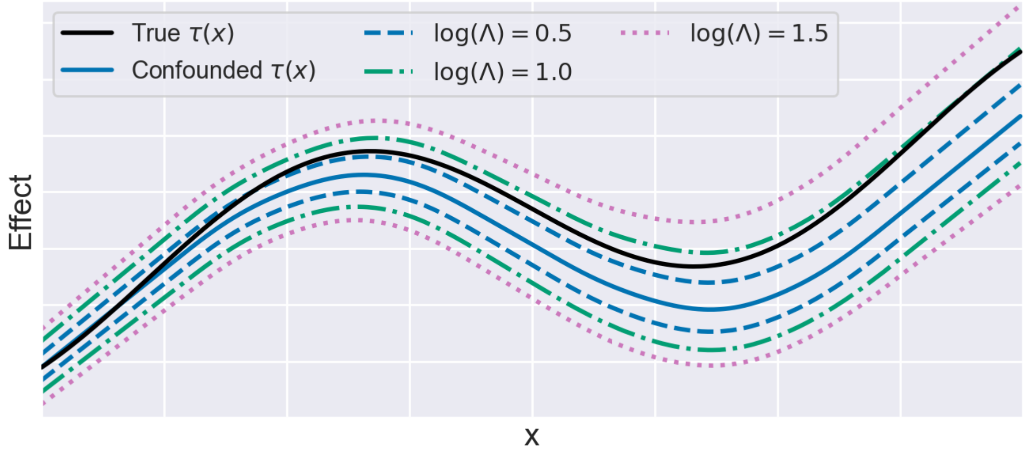
Case Study: Effect of 401(k) Eligibility

- B-Learners for different Λ values.





B-Learner: TL;DR



CATE bounds with unobserved confounding.

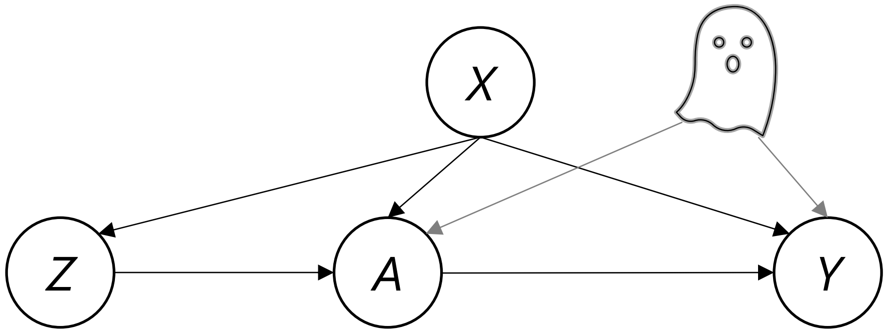
- Lack of unobserved confounding enables causal inference, but it is an untestable assumption.
- Under assumptions about the strength of the unobserved confounding, we can learn *bounds* on $\tau(x)$.
- We propose the B-Learner, a flexible meta-learner that learns **valid**, **sharp** and **efficient** bounds from data.

Talk Overview

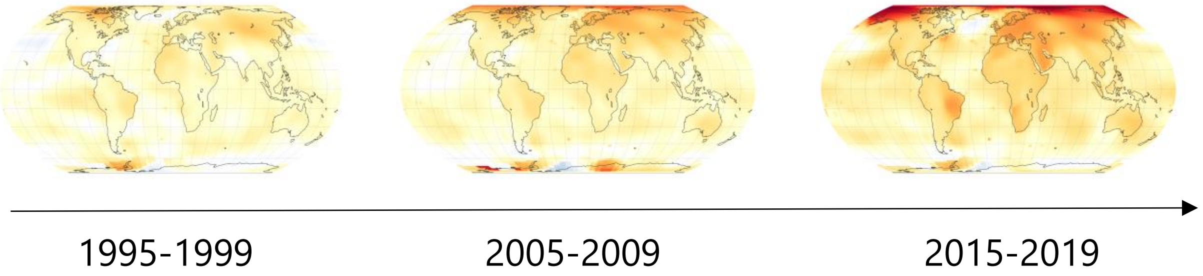
1. Beyond Conditional Averages: Robust and Agnostic Learning of Conditional Distributional Treatment Effects
   - N. Kallus, **M. Oprescu**. AISTATS 2023.
2. Sharp and Efficient Bounds on Heterogeneous Causal Effects Under Hidden Confounding
   - **M. Oprescu**, J. Dorn, M. Ghoummaid, A. Jesson, N. Kallus, U. Shalit. ICML 2023.
3. Research Roadmap: Future Directions and Goals

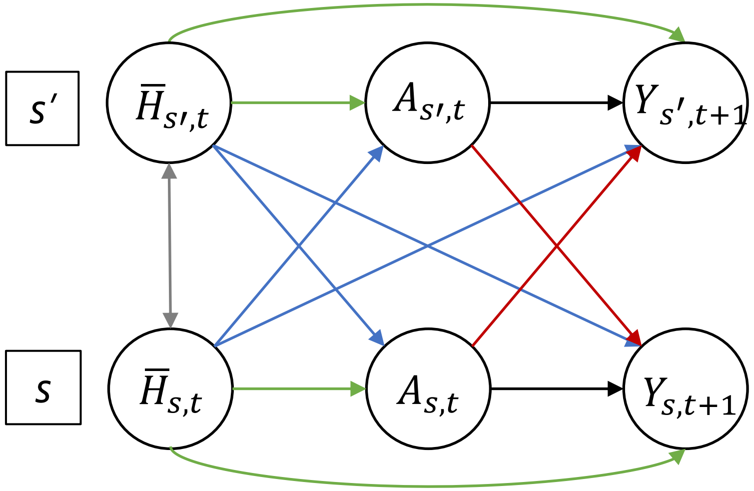
Future Research Directions

1. Causal inference in encouragement designs with weak instruments.



2. Causal inference for spatio-temporal data.
   - E.g. effect of temperature on severe weather events.





Acknowledgments

Special thanks to my committee:

Nathan Kallus (Chair), Sarah Dean, Peter Frazier, Emma Pierson.

Collaborators on this work (in alphabetical order):

Jacob Dorn, Marah Ghoummaid, Andrew Jesson, Nathan Kallus, and Uri Shalit.

Funding support:

The Department of Energy Computational Science Graduate Fellowship.

Appendix

- B-Learner
  - Pseudo-outcome
  - Theoretical guarantees (full)
  - Oracle property
  - Comparison with other works

B-Learner

1. Estimate nuisances $\hat{\eta} = (\hat{e}(x), \hat{q}_{\pm}(x,a), \hat{\rho}_{\pm}(x,a))$ and get pseudo-outcomes:

$$\boldsymbol{Y^+(x,1)} \to \boldsymbol{\phi_1^+(Z,\hat{\eta})} = AY + (1-A)\hat{\rho}_+(X,1) + \frac{(1-\hat{e}(X))A}{\hat{e}(X)}(R_+(Z,\hat{q}_+(X,1)) - \hat{\rho}_+(X,1))$$

$$\boldsymbol{Y^-(x,0)} \to \boldsymbol{\phi_0^-(Z,\hat{\eta})} = (1-A)Y + A\hat{\rho}_-(X,0) + \frac{\hat{e}(X)(1-A)}{1-\hat{e}(X)}(R_-(Z,\hat{q}_-(X,0)) - \hat{\rho}_-(X,0))$$

$$\boldsymbol{\tau^+(x)} \to \boldsymbol{\phi_\tau^+(Z,\hat{\eta})} = \boldsymbol{\phi_1^+(Z,\hat{\eta})} - \boldsymbol{\phi_0^-(Z,\hat{\eta})}$$

where $\mathbb{E}[R_{\pm}(Z,q_{\pm}) \mid X = x, A = a] = \rho_{\pm}(x,a)$.

2. Regress pseudo-outcome $\boldsymbol{\phi_\tau^+(Z,\hat{\eta})}$ on features $X \in \mathcal{X}$ in another sample.

Algorithm 1 The B-Learner

input Data $\{(X_i, A_i, Y_i) : i \in \{1, \ldots, n\}\}$, folds $K \geq 2$, nuisance estimators, regression learner $\widehat{\mathbb{E}}_n$

```
1: for k ∈ {1, ..., K} do
2:    Use data {(X_i, A_i, Y_i) : i ≠ k − 1 (mod K)} to construct nuisance estimates η̂^(k) = (ê^(k), q̂^(k), ρ̂^(k))
3:    for i = k − 1 (mod K) do
4:       Set φ̂⁺_{τ,i} = φ⁺_τ(Z_i, η̂^(k))
5:    end for
6: end for
```

output $\widehat{\tau}^+(x) = \widehat{\mathbb{E}}_n[\widehat{\phi}_\tau^+ \mid X = x]$

Theoretical Guarantees

- The (unsigned) bias from the first stage is:

$$\mathcal{E}(x) = \Sigma_{a=0}^{1}(|\hat{e}(x) - e(x)||\hat{\rho}(x,a) - \rho(x,a)| + \left(\hat{q}(x,a) - q(x,a)\right)^2)$$

- Consider an **ERM**-based second stage estimator $\widehat{\mathbb{E}}_n$ with function class $\mathcal{F}$ bracketing entropy $\log N_{[]}(\mathcal{F}, \epsilon) \leq \epsilon^{-r}$. We have L_2 rate guarantees:

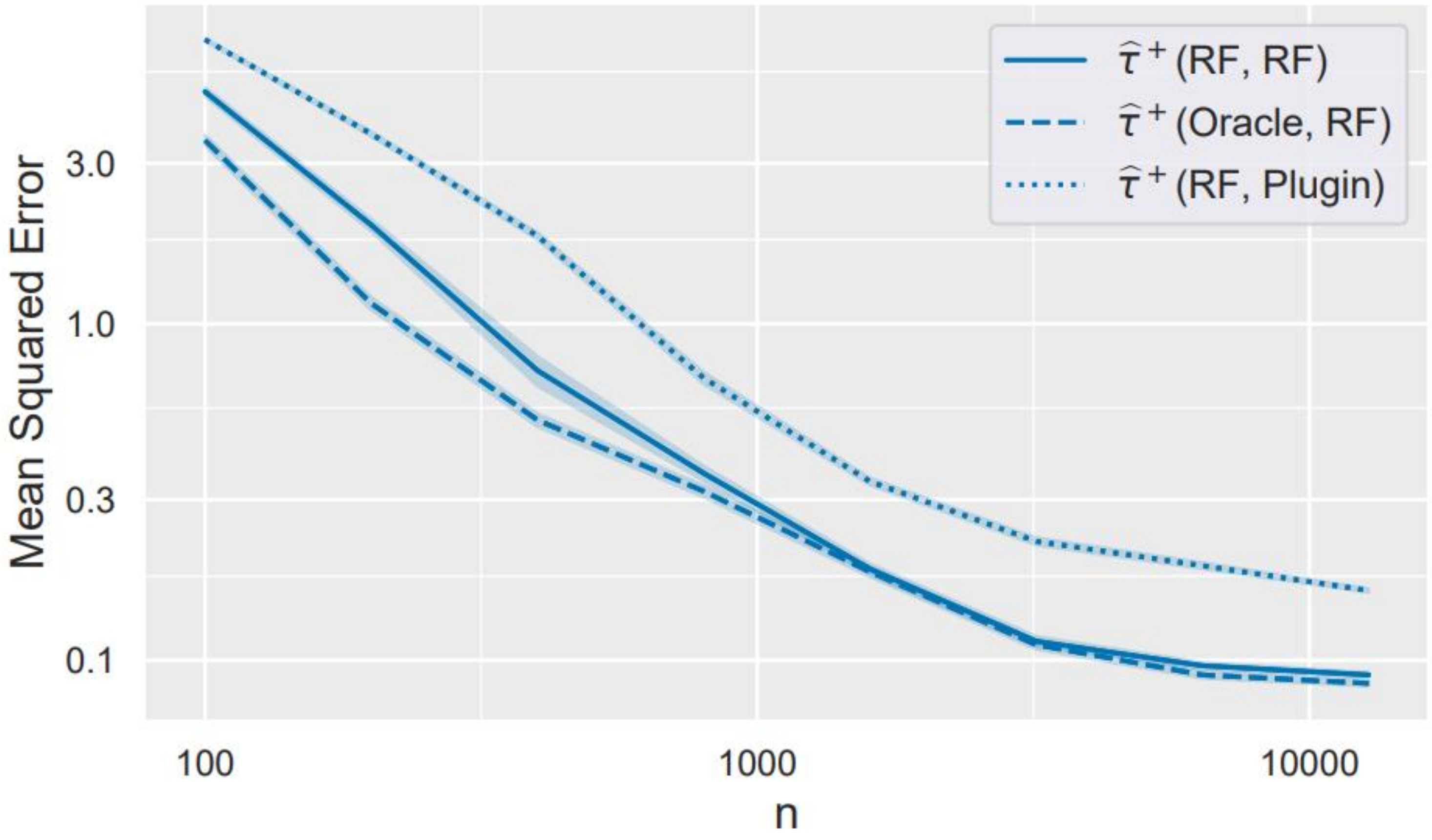
$$\|\hat{\tau}^+(x) - \tau(x)\| \leq O_P\left(n^{-\frac{1}{2+r}}\right) + \|\mathcal{E}(x)\|$$

- Corollaries:

  1. **Sharpness:** If $\hat{q}$ and either $\hat{e}$ or $\hat{\rho}$ are consistent, so is $\hat{\tau}^+(x)$.
  2. **Validity:** If $\hat{q}$ is inconsistent, the bounds are still **valid** on average.
  3. **Quasi-oracle efficiency:** If nuisances are estimated at L_2 rates of $o_P\left(n^{-\frac{1}{2(2+r)}}\right)$, the estimation error is dominated by the complexity of the target class.

Empirical Evidence: Oracle Property

$$A \sim \text{Bernoulli}(\text{logit}(0.75X_0 + 0.5))$$
$$Y \sim \mathcal{N}((2A-1)(X_0+1) - 2\sin\left((4A-2)X_0\right), 1)$$



Empirical Evidence: Comparisons with Other Works