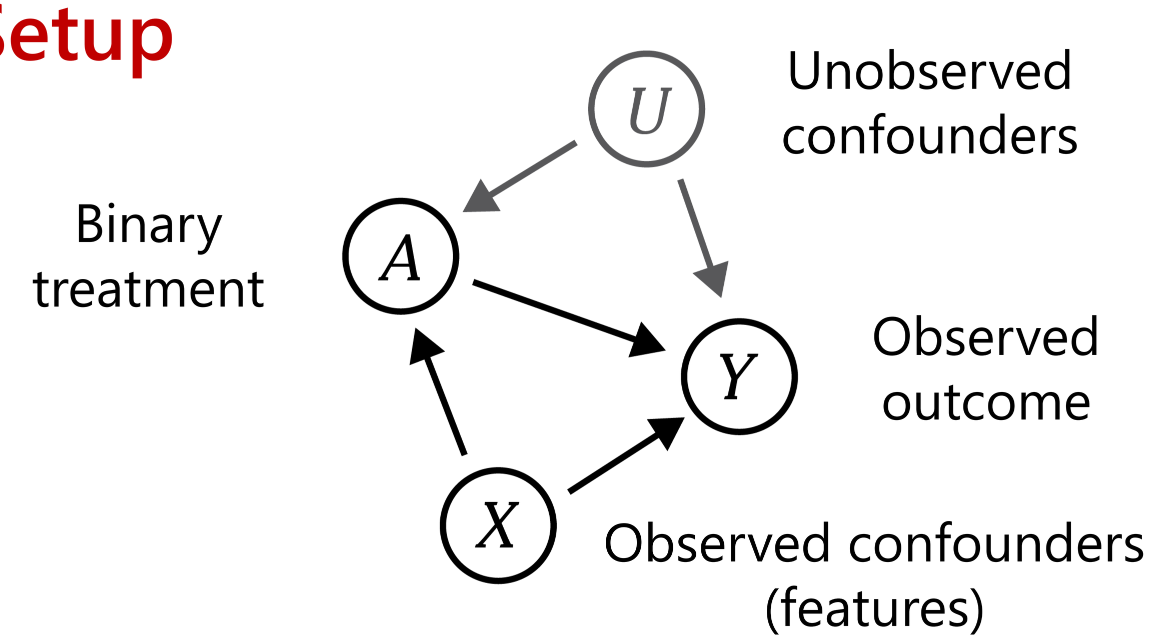
B-Learner: Quasi-Oracle Bounds on Heterogeneous Causal Effects Under Hidden Confounding

Cornell University
Princeton University
University of Oxford
Technion Israel Institute of Technology

Miruna Oprescu[1] Jacob Dorn[2] Marah Ghoummaid[3] Andrew Jesson[4]
Nathan Kallus[1] Uri Shalit[3]

Setup



We want the **bounds** $\tau^+(x), \tau^-(x)$ on the CATE function:

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

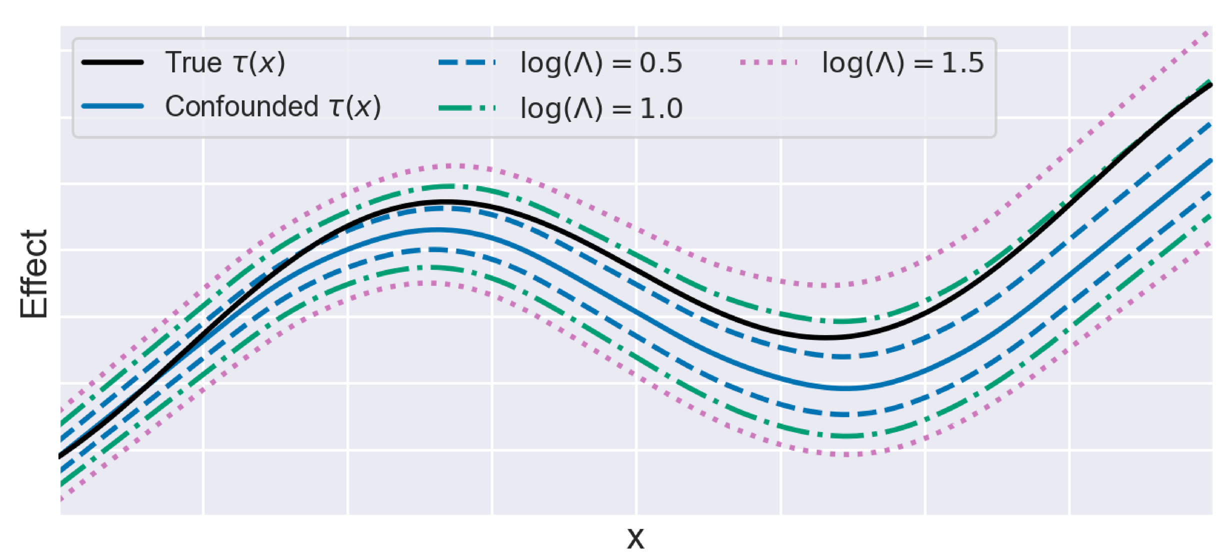
Sensitivity Analysis

Assumption: Marginal Sensitivity Model (MSM).

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

where $e(x) = P(A = 1 \mid X = x)$

$e(x,u) = P(A = 1 \mid X = x, U = u)$



Example of CATE bounds under different values of Λ.

Identification of CATE Bounds

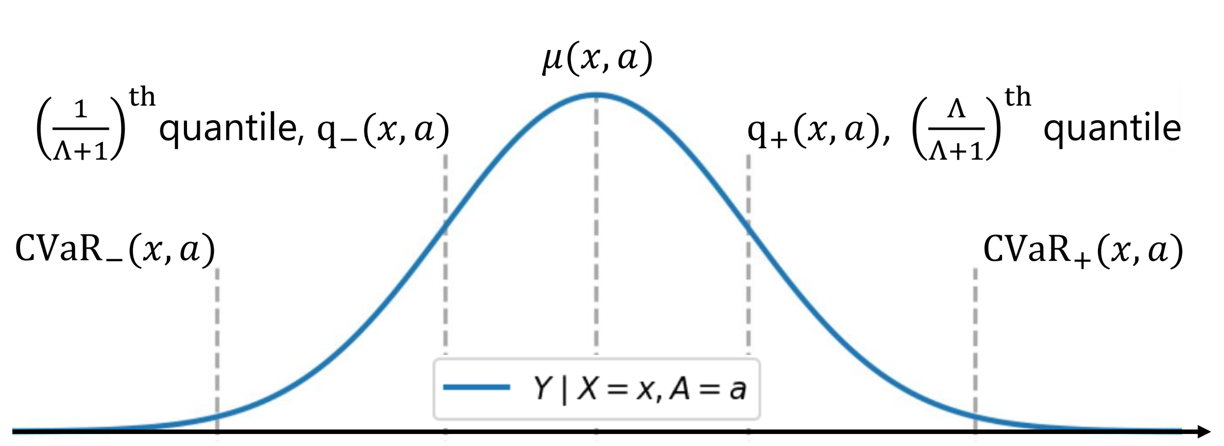
Result 1 (Dorn et al., 2021). $\mu(x,a) = \mathbb{E}[Y \mid X = x, A = a]$ and $Y^{\pm}(x,a)$ is the upper (+) / lower (-) sharp bound of $\mathbb{E}[Y(a) \mid X = x]$ (not identifiable). Then:

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

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

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

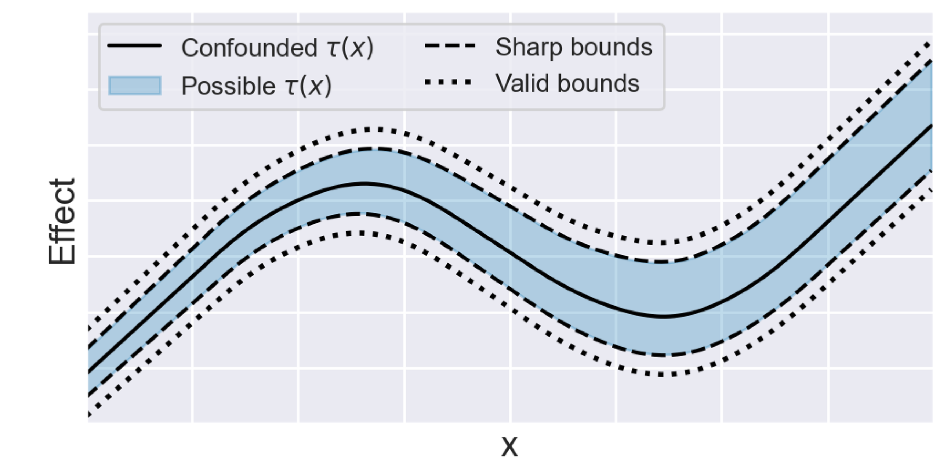
where $\rho_{\pm}(x,a) = \Lambda^{-1}\mu(x,a) + (1 - \Lambda^{-1})CVaR_{\pm}(x,a)$.



Bound Estimates Should Be...

- **Valid:** $\hat{\tau}(x) \geq \tau(x) + o_P(1)$.
- **Sharp:** $\hat{\tau}(x) = \tau(x) + o_P(1)$.
- **Efficient and Robust:** Bounds should converge at desirable rates and have multiple chances at sharp or valid limits.

Previous works do not achieve all!



Example of sharp and valid bounds for CATE estimates.

Estimation of CATE Bounds

Naïve "Plug-in" Estimator

Estimate $e(x), \mu(x,a), \rho_{\pm}(x,a)$ and "plug" them into $Y^{\pm}(x,a)$ to obtain:

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

- Inherits bias from the estimated nuisances $\hat{e}(x), \hat{\mu}(x,a), \hat{\rho}_{\pm}(x,a)$.
- Especially biased when the nuisances are more complex than the CATE bounds.
- Does not yield efficient or robust bounds!

B-Learner

1. Estimate nuisance set $\hat{\eta} = (\hat{e}(x), \hat{q}_{\pm}(x,a), \hat{\rho}_{\pm}(x,a))$ in one sample.
2. Get pseudo-outcomes based on the efficient influence function (EIF):

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

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

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

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

3. 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 $\{(X_i, A_i, Y_i) : i \in \{1, \dots, n\}\}$, folds $K \geq 2$, nuisance estimators, regression learner $\widehat{\mathbb{E}}_n$
1: **for** $k \in \{1, \dots, K\}$ **do**
2: Use data $\{(X_i, A_i, Y_i) : i \not\equiv k - 1 \pmod K\}$ 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 $\widehat{\tau}^+(x) = \widehat{\mathbb{E}}_n[\widehat{\phi}^+_\tau \mid X = x]$

The B-Learner algorithm with K-fold sample splitting.

Theoretical Guarantees

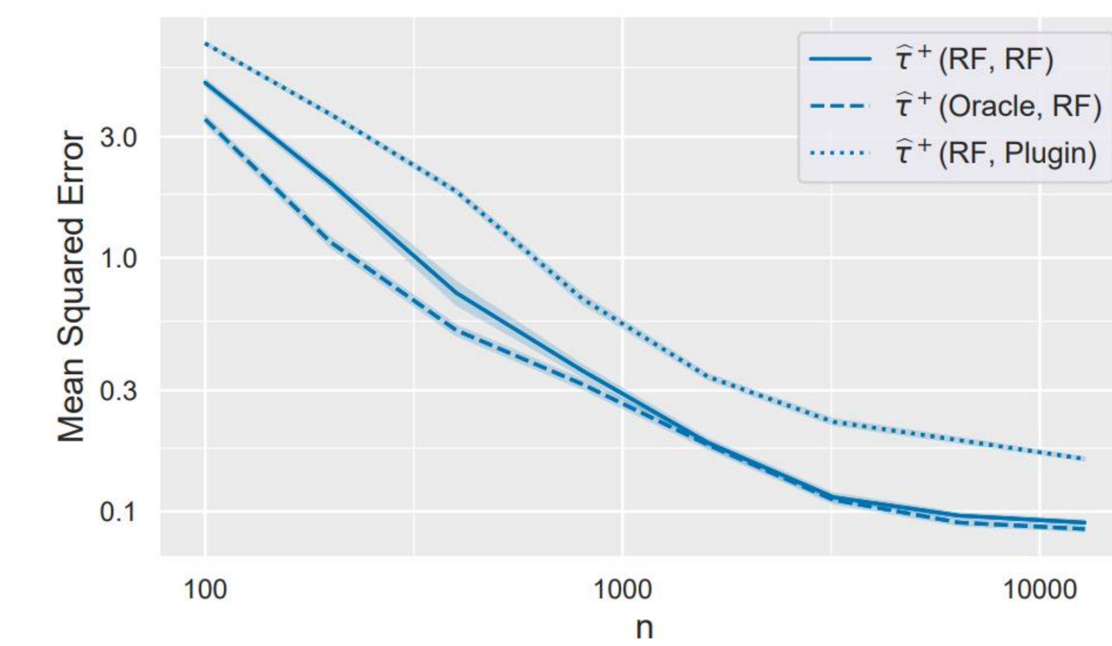
- $\mathbf{L_2}$ validity, sharpness and robustness guarantees for **ERM** second stage estimator $\widehat{\mathbb{E}}_n$:
  1. L_2 bias on the order of $\mathcal{E} = \|\hat{e} - e\|\|\hat{\rho} - \rho\| + \|\hat{q} - q\|^2$.
  2. If $\hat{q}$ and either $\hat{e}$ or $\hat{\rho}$ are consistent, the bounds are **sharp** on average.
  3. If $\hat{q}$ is inconsistent, the bounds are still **valid** in expectation.
  4. The B-Learner has **quasi-oracle efficiency,** i.e. the bounds can be learned at a rate dominated by the complexity of the target class.
- **Pointwise** validity, sharpness and robustness guarantees for **linear smoother** second stage estimator $\widehat{\mathbb{E}}_n$.

Experiments

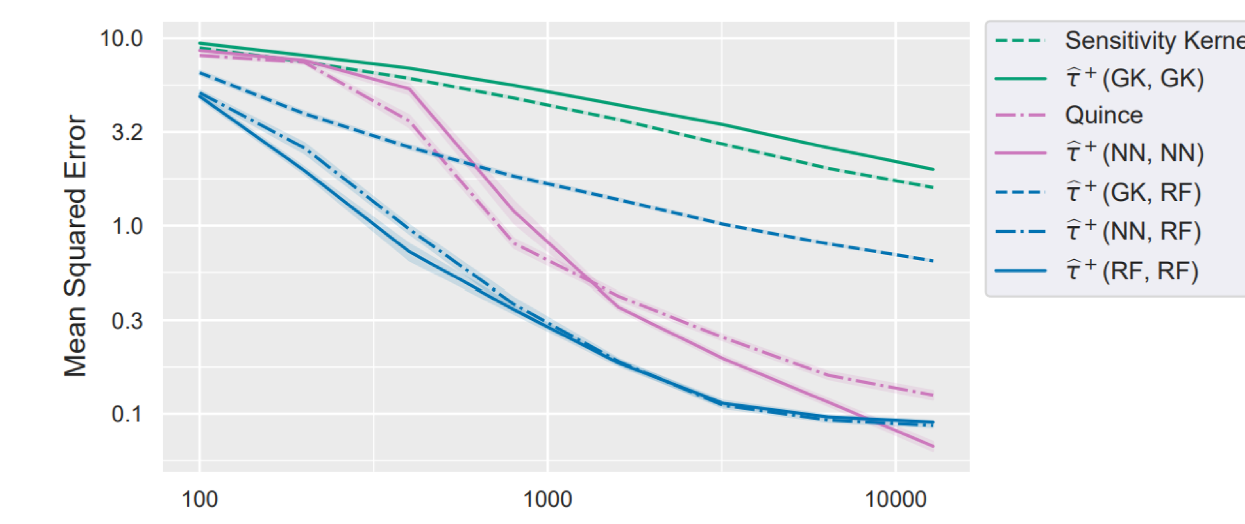
Simulations

$$A \sim \text{Bernoulli}(\text{logit}(0.75X_0 + 0.5))$$

$$Y \sim \mathcal{N}((2A-1)(X_0+1) - 2\sin((4A-2)X_0), 1)$$

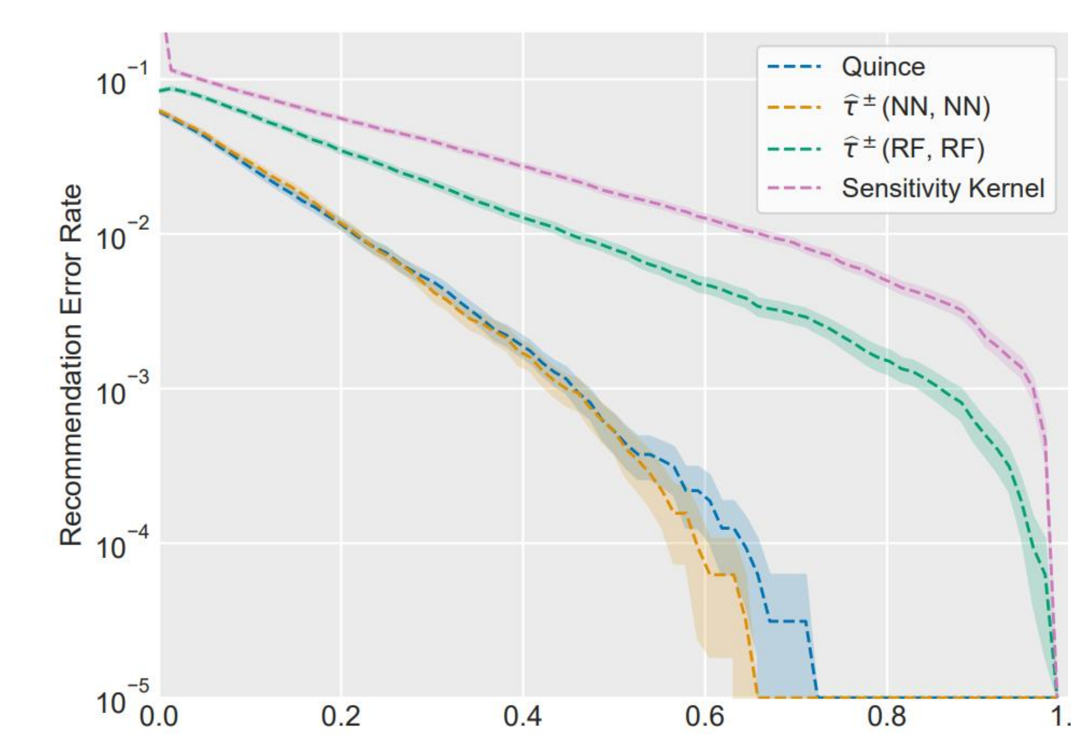


Quasi-oracle property of the B-Learner algorithm. n is the sample size. In $\hat{\tau}^+(x,y)$, x and y are the types of first- and second-stage nuisances.



Performance of the B-Learner compared with the *Sensitivity Kernel* (Kallus et al. 2019) and *Quince* (Jesson et al., 2021). GK=Gaussain Kernel, NN=Neural Network, RF=Random Forest.

IHDP Hidden Confounding



Error recommendation rate for different values of the percentage of deferred points. The x-axis represents different levels of practitioner caution by varying the percentage of recommendations deferred.

[1]Cornell University and Cornell Tech
[2]Princeton University
[3]Technion, Israel Institute of Technology
[4]OATML, University of Oxford
Correspondence to: Miruna Oprescu
miruna@cs.cornell.edu

Paper Code