Beyond Conditional Averages:
Robust and Agnostic Learning of Conditional Distributional Treatment Effects

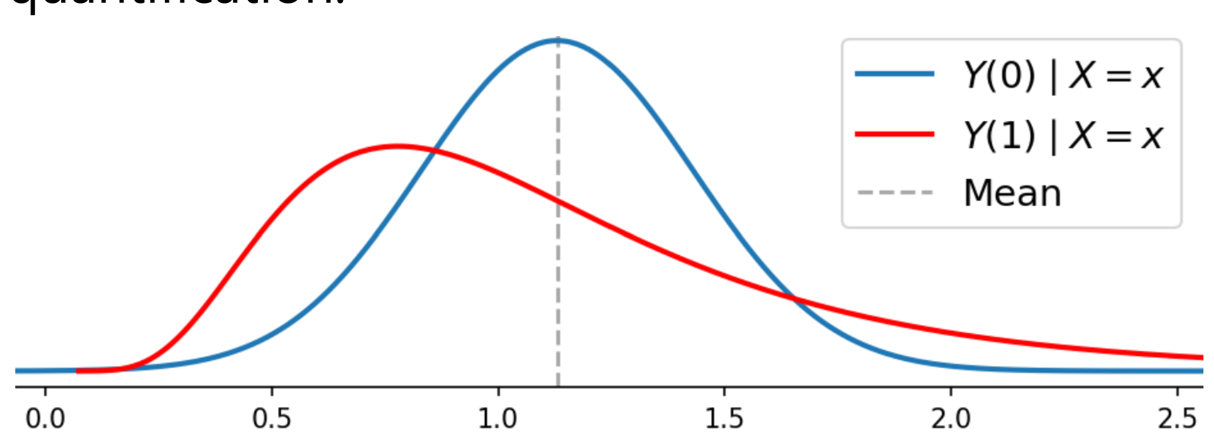
Miruna Oprescu (1st Year Fellow) and Nathan Kallus

Treatment Effects

- Binary intervention ("treatment") $A \in \{0,1\}$, features $X \in \mathcal{X}$, potential outcomes $Y(0), Y(1) \in \mathbb{R}$ under A.
- We want to describe differences in the outcome distributions $F_{Y(1)|X}$ and $F_{Y(0)|X}$.
- Problem #1. For a X_i, we only observe $Y(0)$ or $Y(1)$, not both. Data is:
  $$Z_i = (X_i, A_i, Y_i) \sim (X, A, Y(A))$$
- Problem #2. Correlation ≠ Causation. Selection bias:
  $$e^*(X) = \mathbb{P}(A = 1 \mid X)$$
- Problem #3. Literature focuses mainly on averages:
  $$\mathbb{E}_F\,[Y(1) \mid X = x] - \mathbb{E}_F[Y(1) \mid X = x]$$

Beyond Averages: Motivation

- Skewed outcome functions (e.g., income) and risk quantification.



Outcomes with the same conditional mean, but different tail effects.

- Need to look beyond the conditional mean effect: Conditional **Distributional** Treatment Effects (**CDTEs**)

CDTEs

Definition:

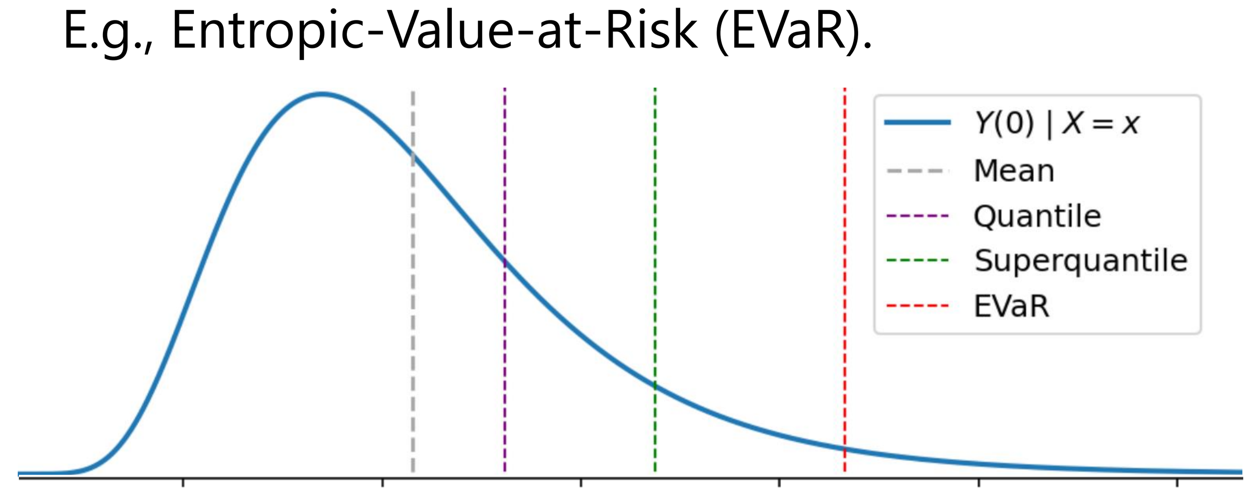
$$CDTE(X) = \kappa^*(F_{Y(1)|X}) - \kappa^*(F_{Y(0)|X})$$

where $\kappa^*(F)$ is any distribution statistic.

Examples of statistics and corresponding CDTEs:

- Mean (CATE)
- Quantiles (CQTE)
- Superquantiles, i.e., tail averages (CSQTE)
- f-risk measures from f-divergences (CfRTE)

E.g., Entropic-Value-at-Risk (EVaR).



Different distribution statistics (quantile, superquantile, EVaR) at level 0.75.

TL;DR

We propose an algorithm for learning (conditional) **distributional causal effects**. Our method is **robust** and **model agnostic** in that we can learn these effects at fast rates, and we can conduct valid inference on coefficients of linear projections.

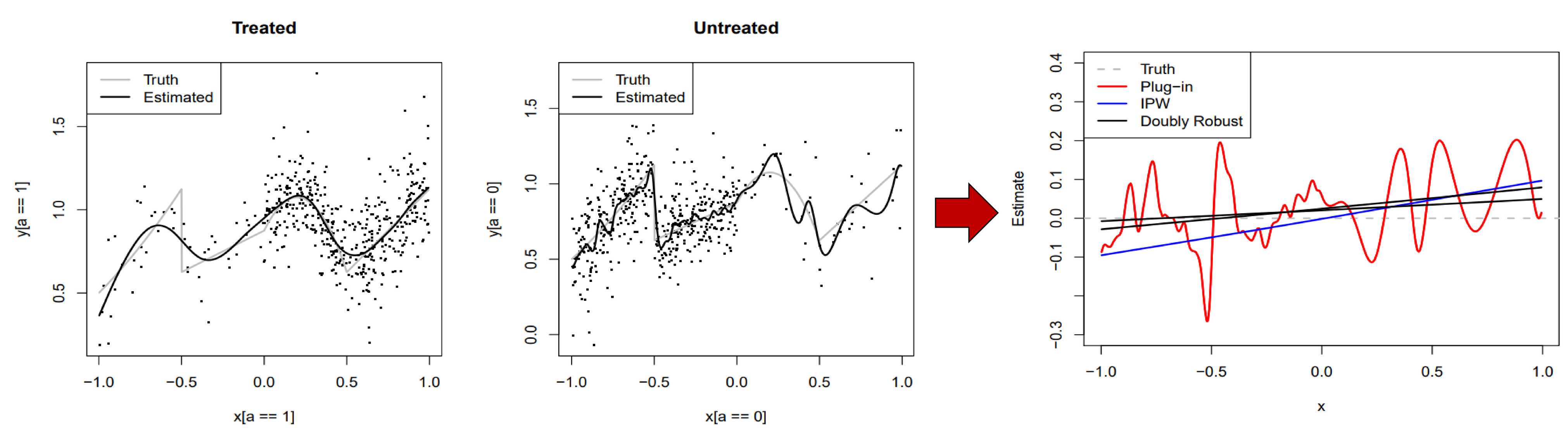
Why Is This a Challenging Problem?

Consider the naïve ("plug-in") estimator:

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

where $\hat{\kappa}_a(X)$ are estimates for $\kappa_a^*(X)$.

- Inherits bias from the nuisances $\hat{\kappa}_a(X)$ (so-called "plug-in" bias).
- Can wash out the signal when the nuisances are more complex than the CDTE.
- Not robust: difference of linear predictors ≠ best linear predictor of difference.



Plugin bias illustration for CATE estimators (from "Towards optimal doubly robust estimation of heterogeneous causal effects", Kennedy, 2020). For means, our method reduces to "Doubly Robust" above.

Debiased CDTE Estimation Algorithm

General Framework: Moment Statistics

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

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

- Mean: $\rho(y, \mu) = y - \mu$
- Quantiles (level τ): $\rho(y, q) = \tau - \mathbb{I}[y \le q]$
- Superquantiles (level τ): $\rho(y, \mu, q) = ((1-\tau)^{-1} y \mathbb{I}[y \ge q], \tau - \mathbb{I}[y \le q])$

Debiased Regression Estimator

1. We derive a debiasing term for the plug-in estimator:

$$\psi(Z, e, \alpha, \nu) = \underbrace{\kappa_1(X) - \kappa_0(X)}_{\text{"plug-in" estimator}} - \underbrace{\frac{A - e(X)}{e(X)(1 - e(X))} \alpha_A(X)^T \rho(Y, \nu_A(X))}_{\text{bias correction}}$$

where $\nu_a = (\kappa_a, h_a)$ and the $\alpha_a\,(X)$'s are additional nuisances to estimate.

2. We regress $\psi(Z, e, \alpha, \nu)$ on features $X \in \mathcal{X}$.

Algorithm 1 CDTE Learner

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

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

  Use data $\{(X_i, A_i, Y_i) : i \ne k - 1 \pmod K\}$ 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 $\widehat{\mathrm{CDTE}}(x) = \widehat{\mathbb{E}}_n[\widehat{\psi} \mid X = x]$

Learning and Inference Guarantees

⚠ (Machine Learning jargon)

Robustness:

- Our algorithm's error (RMSE) has a product structure so small errors in the nuisances lead to second-order errors in the CDTE estimates.
- E.g., if nuisances are estimated at a rate of at least $O(n^{-1/4})$ (nonparametric), CDTEs can be estimated at the rate $O(n^{-1/2})$ (parametric).
- There are many chances at convergence when some of the nuisances are misspecified.

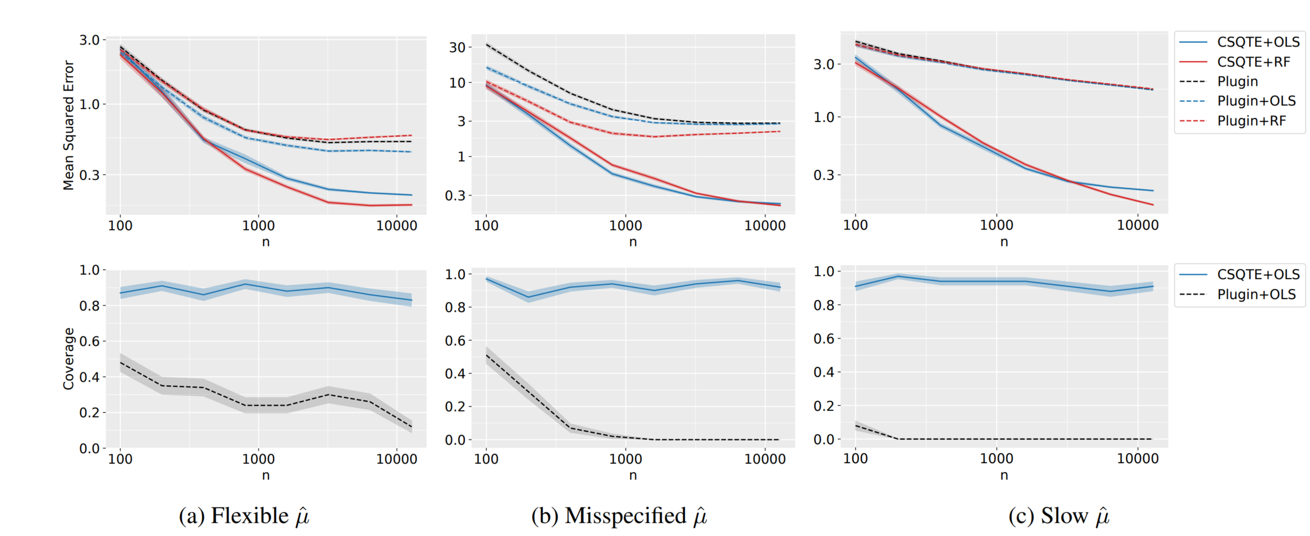
Model Agnostic:

- Linear regression parameters are asymptotically normal with oracle variance.
- E.g., if we use OLS as the final regression, the confidence intervals are valid.

Simulation Study:

- Tail averages for DGP:

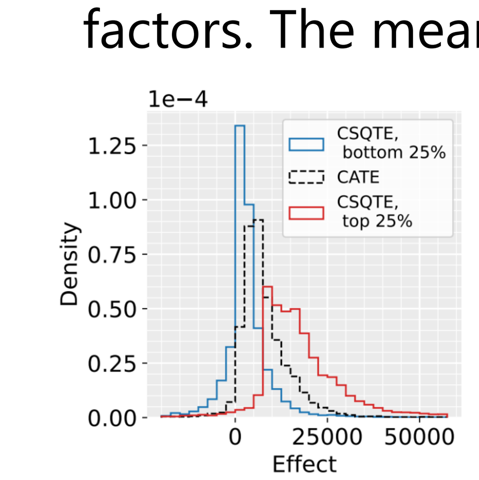
$$A \sim \text{Bernoulli}(\text{logit}(6X_0 - 3))$$
$$Y \mid X, A \sim \text{Lognormal}(X_0 + AX_1, 0.2)$$



Performance of CSQTE learner with a flexible, misspecified or slow converging $\hat{\mu}_a(X)$ estimator. Either a flexible learner (random forest) or OLS is used for the final stage,

Case Study: Effect of 401(k) Eligibility

- Effect of 401(k) eligibility on net worth.
- Effect on the tail averages (CSQTE) of bottom and top 25% of asset holders.
- Conclusion: Tail effects are driven by different factors. The mean does not capture this variation.



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.

This work was published in the International Conference on Artificial Intelligence and Statistics (2023).