Analyzing the sensitivity of causal findings under distribution shifts

Hongseok Namkoong

Columbia Business School
namkoong@gsb.columbia.edu

Based on joint works with Steve Yadlowsky, Sookyo Jeong, and others.

Causality and decision-making

- Causal understanding is crucial for reliable decision-making
- Counterfactuals: what would've happened if the person didn't see the ad, or didn't get the drug?
- Striking progress in machine learning based prediction models
- Today: Leverage scalable ML methods to infer causal effect

Secret to life

Secret to life

The New York Times

Another Benefit to Going to Museums? You May Live Longer

Researchers in Britain found that people who go to museums, the theater and the opera were less likely to die in the study period than those who didn't.

Heavy selection bias based on unobservables (wealth): decision / treatment is intimately connected with health outcomes

Potential outcomes

- A feature vector $X \in \mathbb{R}^k$
- A treatment assignment $Z \in \{0,1\}$
- Potential outcomes: $Y(1), Y(0)$
- **Observe $Y := Y(Z)$, never $Y(1-Z)$**

Average Treatment Effect (ATE)

$$ATE = \mathbb{E}[Y(1) - Y(0)]$$

$$= \mathbb{E}_{X \sim P_X}\left[\mathbb{E}[Y(1) \mid X] - \mathbb{E}[Y(0) \mid X]\right]$$

$$= \mathbb{E}_{X \sim P_X}\left[\mu_1^\star(X) - \mu_0^\star(X)\right] =: \mathbb{E}_{X \sim P_X}\left[\mu^\star(X)\right]$$

Conditional Average Treatment Effect

- P_X is the data generating distribution for X

Randomized trials

a.k.a. A/B testing, experiments

- **Randomize** treatments: $Y(1), Y(0) \perp Z$

$$\text{ATE} = \mathbb{E}[Y(1) - Y(0)]$$

observable

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

- Estimable from data (Y_i, Z_i)

Observational studies

- When experimentation is risky, crucial to leverage collected data
- Historically, many important observational findings: citrus for scurvy, insulin for diabetes
- Must be contextualized and viewed with more skepticism

Observational studies

Recap

- Covariates: X, Treatment: Z
- Potential outcome: $Y(0)$, $Y(1)$
- Response $Y := Y(Z)$

Assumption: $Y(1), Y(0) \perp Z \mid X$ **no unobserved confounding**

- Treatment only based on observed information
- Approach 1 (Importance sampling; IPW): Reweight treated units w.r.t. X so that they look like the overall population. Importance weight is estimated using $\widehat{\mathbb{P}}(Z = 1 \mid X)^{-1}$.

Approach 2: Direct method

Recap

- Covariates: X, Treatment: Z
- Potential outcome: $Y(0),\ Y(1)$
- Response $Y := Y(Z)$

- Under **no unobserved confounding,**

$$\psi(P) := \mathbb{E}_P[Y(1)] = \mathbb{E}[\mathbb{E}_P[Y(1) \mid X, Z = 1]]$$

- Directly regress Y on X for treated units (Z=1) to get $\mathbb{E}_{\hat{P}}[Y \mid X, Z = 1]$

- **What is the statistical error of the plug-in estimator $\psi(\hat{P})$?**

Debiasing

Recap

$\psi(P) := \mathbb{E}_P[Y(1)] = \mathbb{E}[\mathbb{E}_P[Y \mid X, Z = 1]]$

- **Idea: Correct plug-in estimator using the first-order error**

$$\psi(\hat{P}) - \psi(P) = \nabla\psi(\hat{P})^\top(\hat{P} - P) + \text{Rem}_2 \qquad \text{[finite-dimensional]}$$

$$\psi(\hat{P}) - \psi(P) = \int \nabla\psi(\text{data}; \hat{P}) d(\hat{P} - P) + \text{Rem}_2 \qquad \text{[}\infty\text{-dimensional]}$$

- Debiased estimator

$$\psi(\hat{P}) - \int \nabla\psi(\text{data}; \hat{P}) d(\hat{P} - P)$$

- Debiased estimator automatically only has second-order error Rem_2

Debiased estimator

Recap

- Covariates: X, Treatment: Z
- Potential outcome: $Y(0)$, $Y(1)$
- Response $Y := Y(Z)$

- Outcome model $\mu_1^\star(X) := \mathbb{E}[Y \,|\, X, Z = 1]$, propensity score $e^\star(X) := \mathbb{P}(Z = 1 \,|\, X)$
- Debiasing gives **doubly robust** estimator

$$\mathbb{E}[Y(1)] = \mathbb{E}\left[\mu_1^\star(X) + \frac{Z}{\mathbb{P}(Z = 1 \mid X)}(Y - \mu_1^\star(X))\right]$$

- Accurate if you can do either well; **insensitive** to errors in nuisance estimates

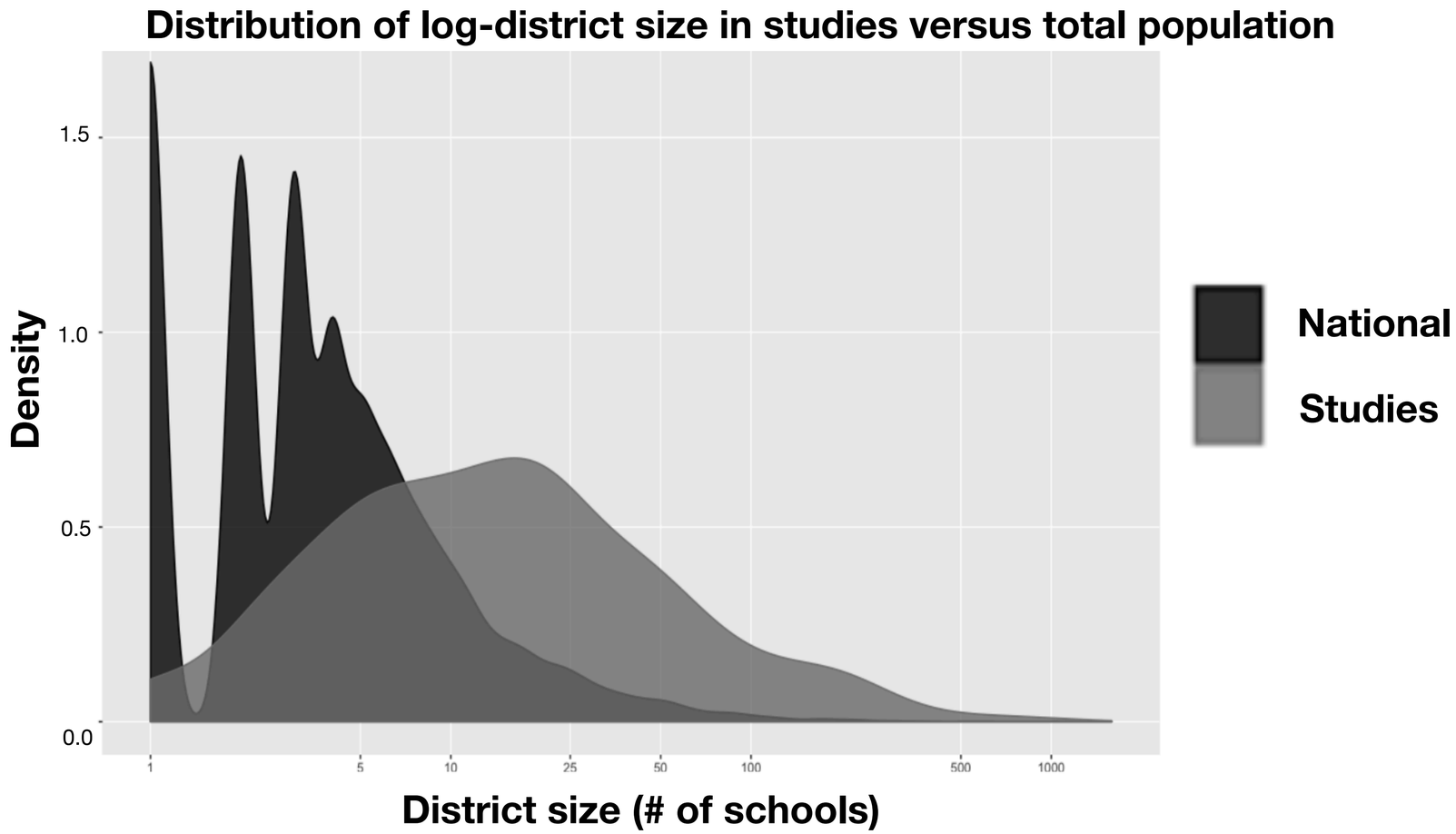
Scalable estimator of causal effect and cumulative rewards

Problem I: population shifts

a.k.a. X-shift, covariate shit

Problem I: what if P_X changes?

- Even for carefully designed randomized trials, “statistics” starts only at treatment assignment, with big biases in selection into study



[Tipton et al. 2019] The convenience of large urban school districts: a study of recruitment practices in 37 randomized trials

Problem I: what if P_X changes?

- "Clinical trials for new drugs **skew heavily white"** [Oh et al. '15, Burchard et al. '15, SA Editors '18]
  - Out of 10,000+ cancer trials, less than 5% of participants were non-white
- Even large clinical trials suffer from these biases. Recently, two large trials with n = 5K-10K had opposite findings on a treatment to lower blood pressure on cardiovascular disease [Leigh et al. '16, Imai et al. '13, Gijsberts et al. '15, Basu et al. '17, Baum et al. '17, Duan et al. '19]

Problem II: unobserved confounders

a.k.a. Y | X shift

Unobserved confounders

- There always exists unobserved confounders that simultaneously affect potential outcomes and treatment assignments

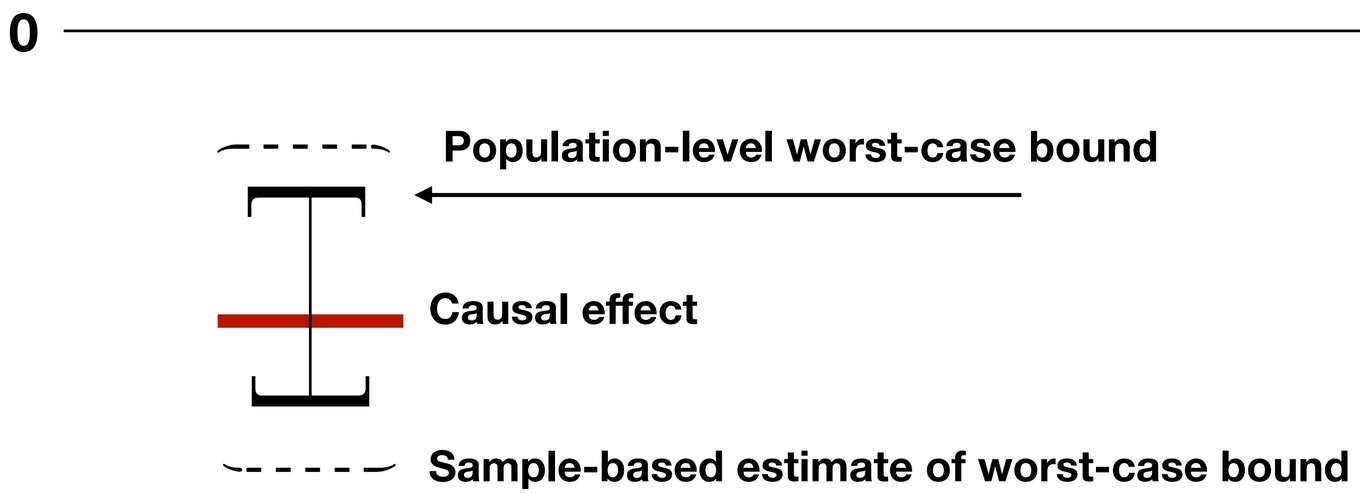
Judges are more lenient after taking a break, study finds theguardian [Danziger '11]

Overlooked factors in the analysis of parole decisions [Weinshall-Margel '11]

- Visual observations used in clinical decisions and drugs preferentially prescribed those who can tolerate them
  - Not properly recorded even at the resolution of large databases

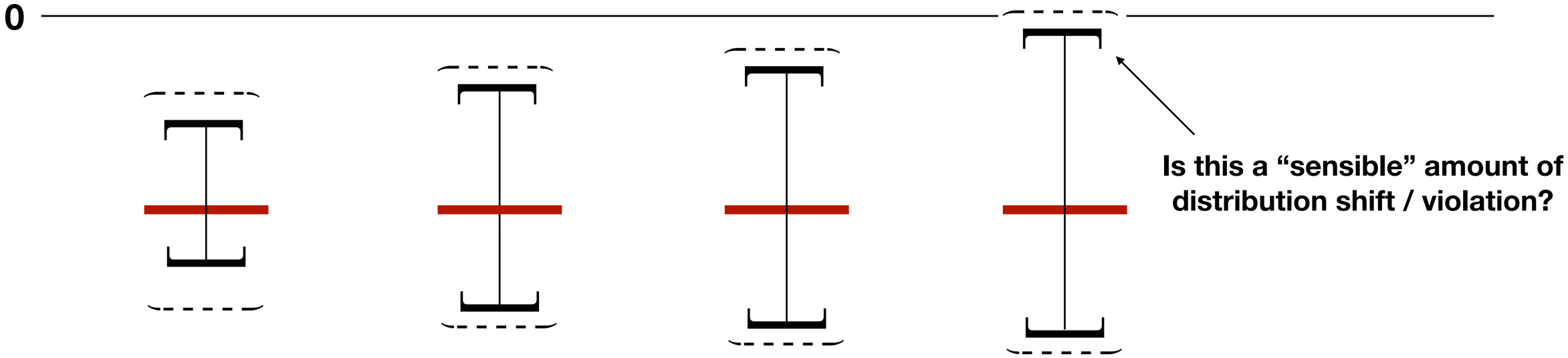
Sensitivity analysis

- Posit a set of “plausible” distribution shifts, and take worst-case over them
- If effects are still valid under plausible shifts, we can certify robustness
- Sensitivity of a finding: magnitude of shift when endpoint crosses a threshold
- Today: worst-case bounds on the Doubly Robust estimator



Sensitivity analysis

- Posit a set of “plausible” distribution shifts, and take worst-case over them
- If effects are still valid under plausible shifts, we can certify robustness
- Sensitivity of a finding: magnitude of shift when endpoint crosses a threshold
- Today: worst-case bounds on the Doubly Robust estimator

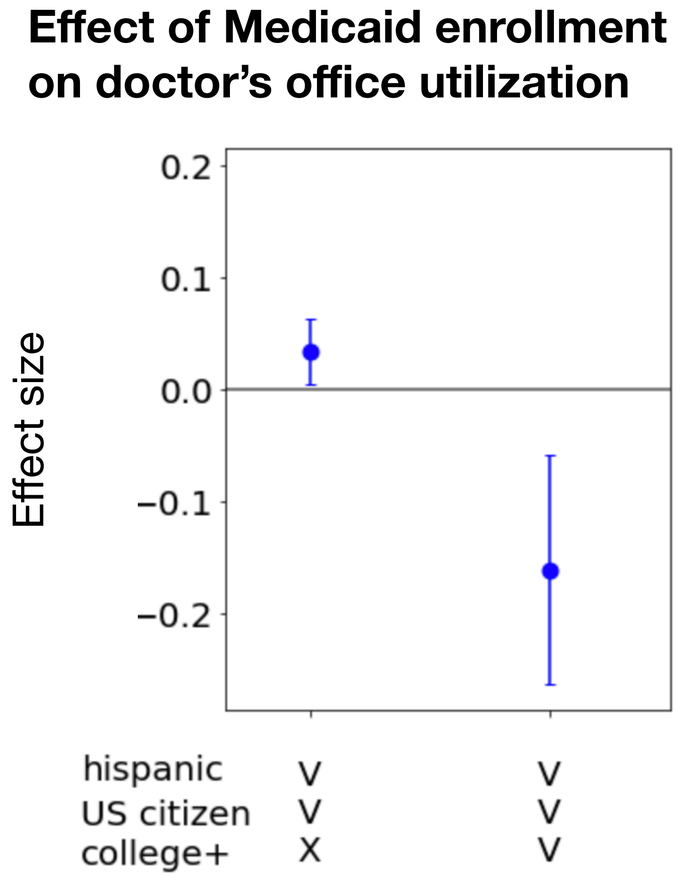


Part I: External validity

a.k.a. X-shift, covariate shit

Challenges

- X-shift problematic when treatment effect is heterogeneous
  - Healthcare: across demographics, comorbidities, and concomitant drugs
- Option 1: Directly estimate conditional average treatment affect (CATE)?
  - ML models unstable on underrepresented groups; resulting inference underpowered
- Option 2: Subgroup analysis?
  - Difficult due to intersectionality

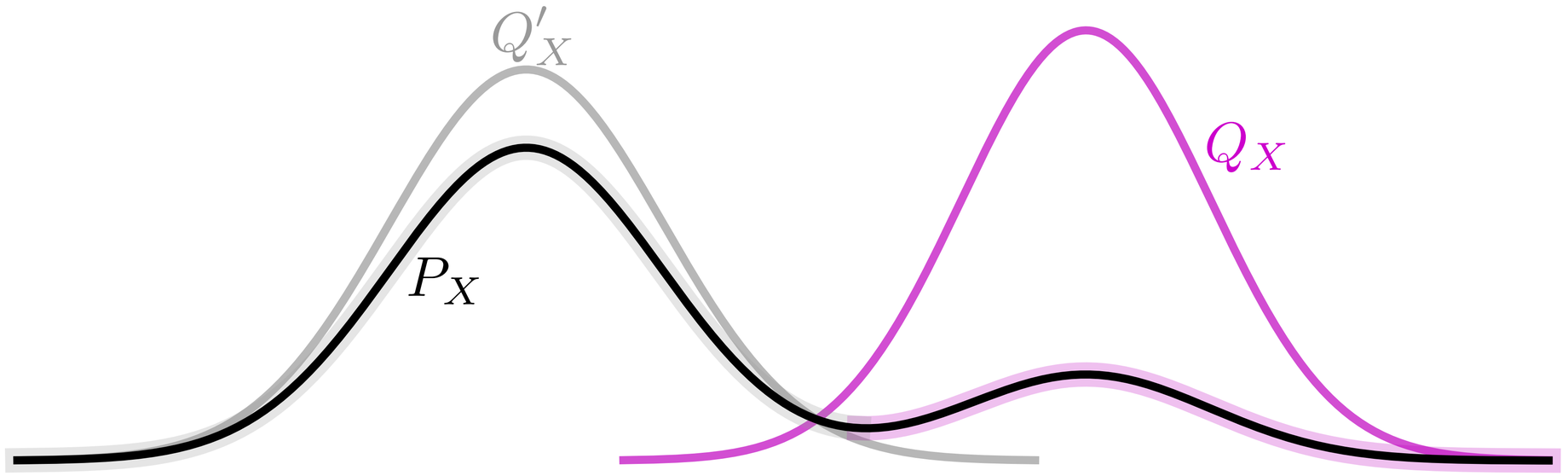


[Leigh et al. ’16, Imai et al. ‘13, Gijsberts et al. ’15, Basu et al. ’17, Baum et al. ’17, Duan et al. ’19, Nie and Wager ’20]

Subpopulations

Automatically find **worst-off subpopulations**
and measure **treatment effect** on them

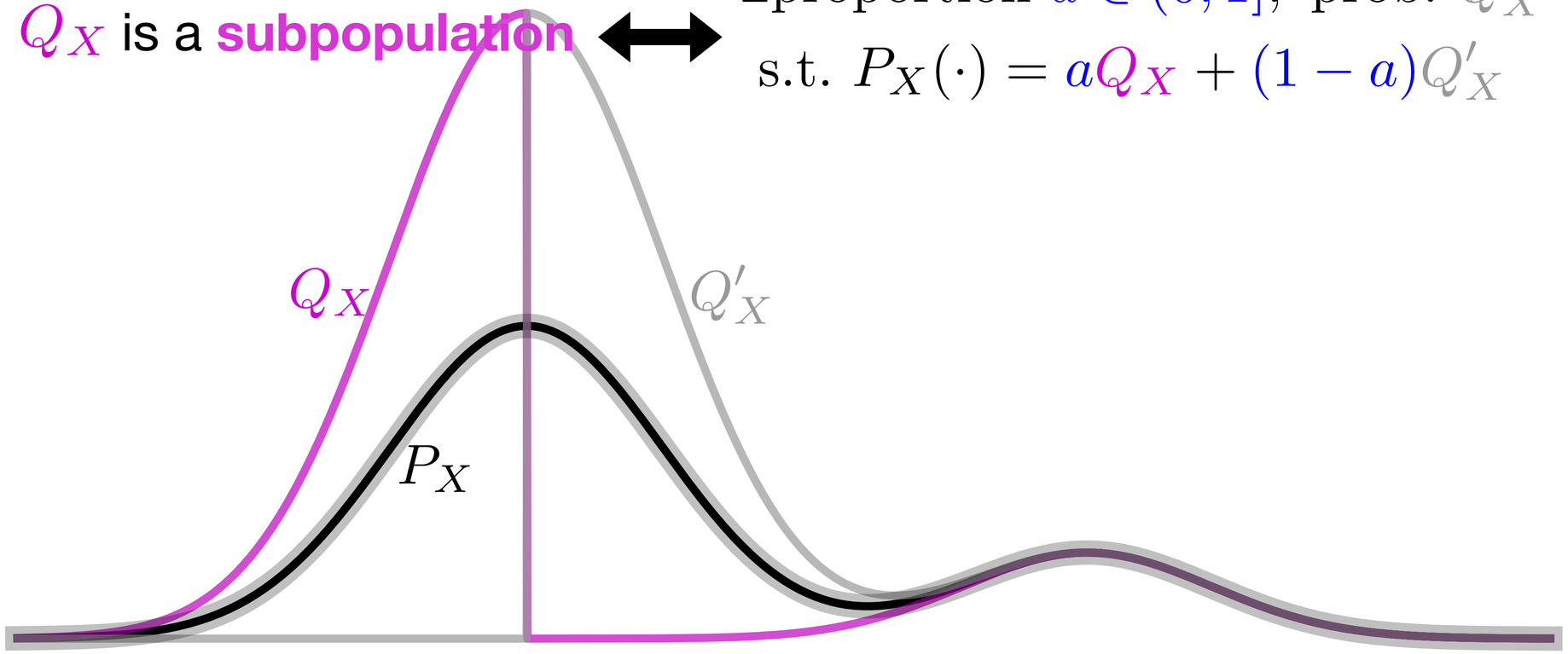
Q_X is a **subpopulation** $\Longleftrightarrow$ $\exists$proportion $a \in (0, 1]$, prob. Q'_X
s.t. $P_X(\cdot) = aQ_X + (1-a)Q'_X$



Subpopulations

Automatically find **worst-off subpopulations**
and measure **treatment effect** on them

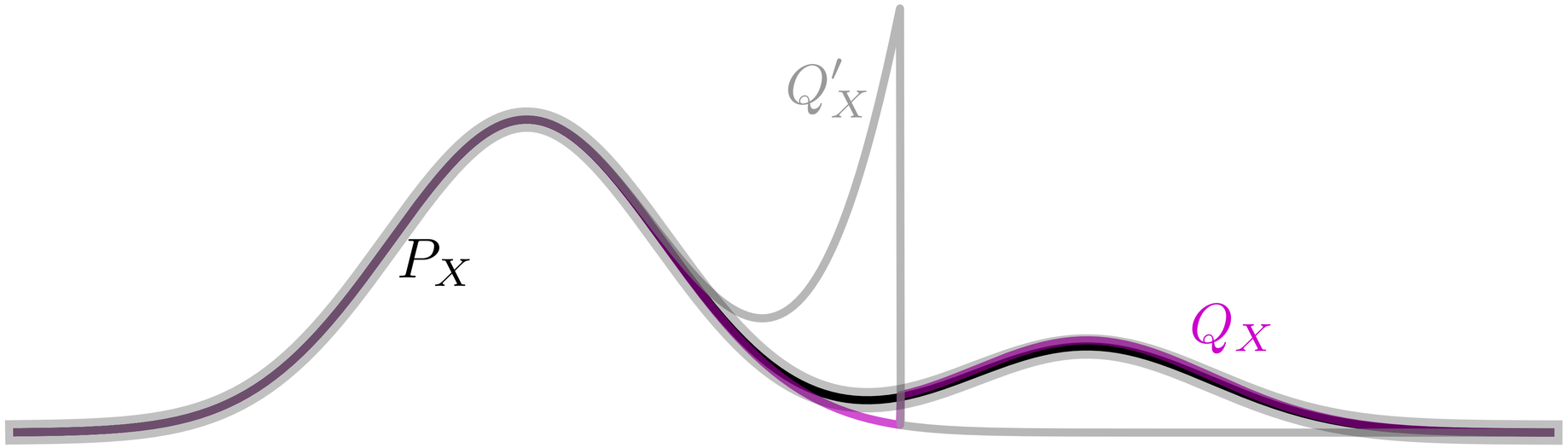
Q_X is a **subpopulation** $\longleftrightarrow$ $\exists$proportion $a \in (0, 1]$, prob. Q'_X
s.t. $P_X(\cdot) = aQ_X + (1-a)Q'_X$



Subpopulations

Automatically find **worst-off subpopulations**
and measure **treatment effect** on them

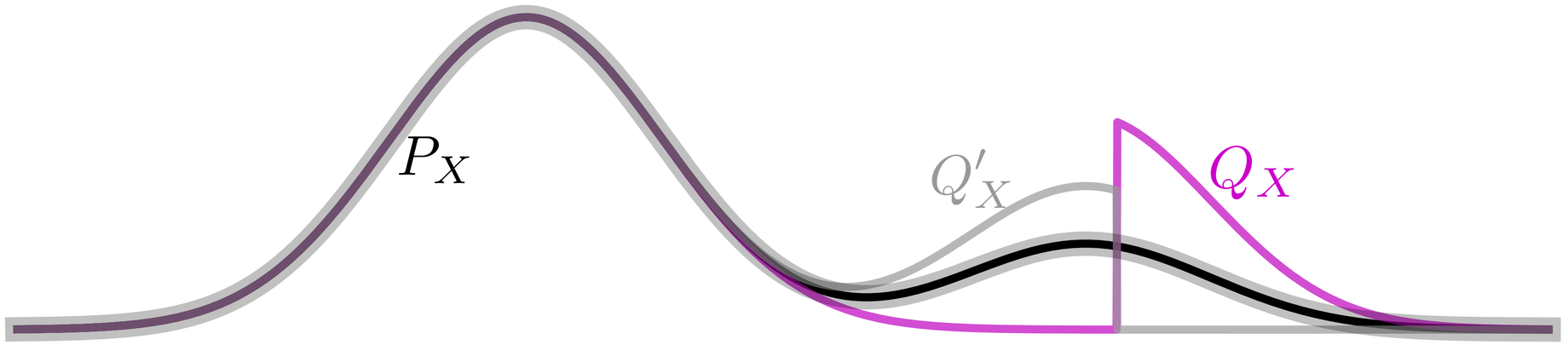
Q_X is a **subpopulation** $\Longleftrightarrow$ $\exists$proportion $a \in (0, 1]$, prob. Q'_X
s.t. $P_X(\cdot) = aQ_X + (1-a)Q'_X$



Subpopulations

Automatically find **worst-off subpopulations**
and measure **treatment effect** on them

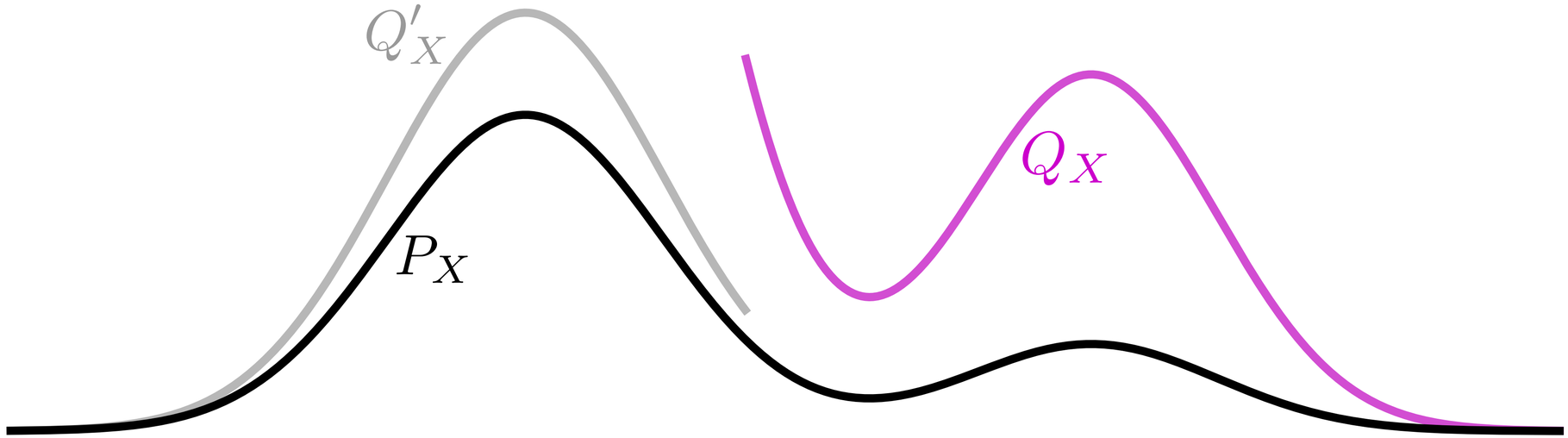
Q_X is a **subpopulation** $\longleftrightarrow$ $\exists$proportion $a \in (0, 1]$, prob. Q'_X
s.t. $P_X(\cdot) = aQ_X + (1-a)Q'_X$



Subpopulations

Automatically find **worst-off subpopulations**
and measure **treatment effect** on them

Q_X is a **subpopulation** $\Longleftrightarrow$ $\exists$proportion $a \in (0, 1]$, prob. Q'_X
s.t. $P_X(\cdot) = aQ_X + (1-a)Q'_X$



Worst-case subpopulation

Recap

- Covariates: X
- Treatment assignment: Z
- Potential outcome: $Y(0)$, $Y(1)$
- Response $Y := Y(Z)$

$Q_X \succeq \alpha$ ⟷ **subpopulation** with **proportion** larger than $\alpha \in (0, 1]$

Worst-case Subpopulation Treatment Effect

$$\mathrm{WTE}_\alpha := \sup_{Q_X \succeq \alpha} \mathbb{E}_{Q_X}[\mu^\star(X)]$$

where $\mu^\star(X) := \mathbb{E}[Y(1) - Y(0) \mid X]$

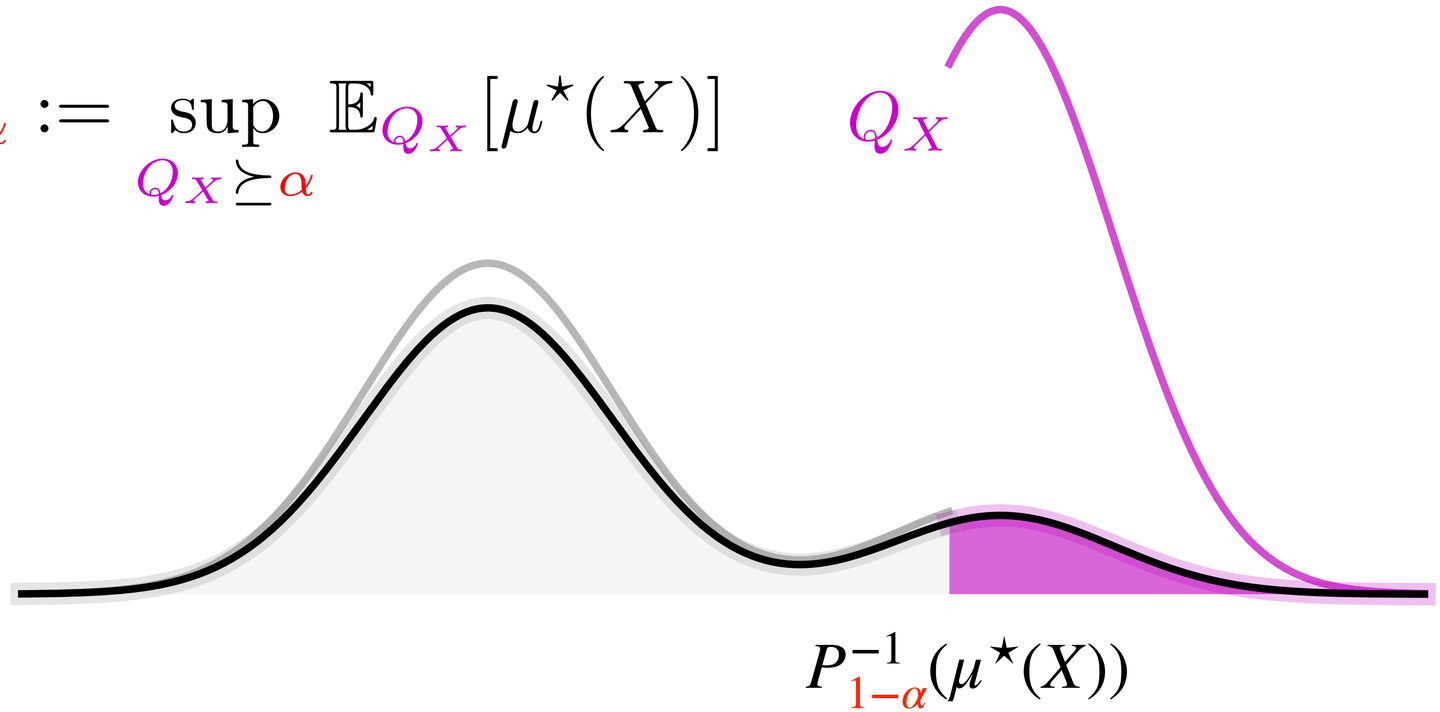
is the conditional average treatment effect (CATE).

WTE = Tail-average

Recap

- Covariates: X
- Treatment assignment: Z
- Potential outcome: $Y(0)$, $Y(1)$
- CATE $\mu^\star(X) = \mathbb{E}[Y(1) - Y(0) \mid X]$

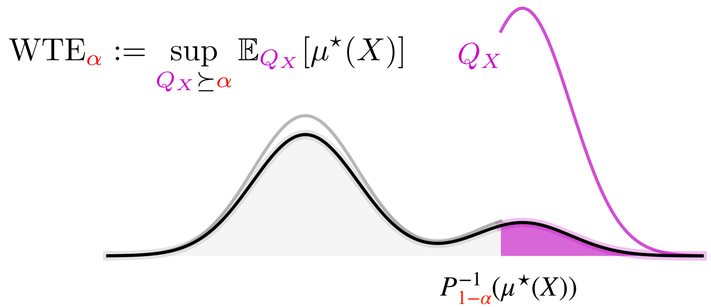
$$\mathrm{WTE}_\alpha := \sup_{Q_X \succeq \alpha} \mathbb{E}_{Q_X}[\mu^\star(X)]$$



WTE = Tail-average

Recap

- Covariates: X
- Treatment assignment: Z
- Potential outcome: $Y(0)$, $Y(1)$
- CATE $\mu^\star(X) = \mathbb{E}[Y(1) - Y(0) \mid X]$



Lemma (Shapiro et al. '09)

$$\sup_{Q_X \succeq \alpha} \mathbb{E}_{Q_X}[\mu^\star(X)] = \mathbb{E}[\mu^\star(X) h^\star(X)]$$

where $h^\star(x) := \frac{1}{\alpha} \mathbf{1}\left\{\mu^\star(x) \geq P^{-1}_{1-\alpha}(\mu^\star)\right\}$

$(1-\alpha)$-**quantile of** $\mu^\star(X)$

Main Result

- Use any ML method to fit nuisance parameters

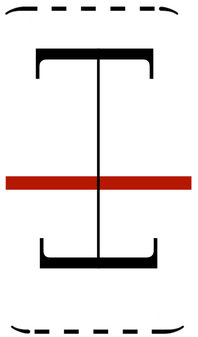
$$\mathbb{E}[Y \mid X = x, Z = z], \quad \mathbb{P}(Z = 1 \mid X = x),$$

- **Debiased** estimator $\hat{w}_\alpha$: generalizes Doubly Robust under population shifts

Theorem (Jeong & N. '20)

1. Even when nuisance parameters converge more slowly,
   $\sqrt{n}(\hat{w}_\alpha - \mathrm{WTE}_\alpha) \Rightarrow N(0, \sigma_\alpha^2)$

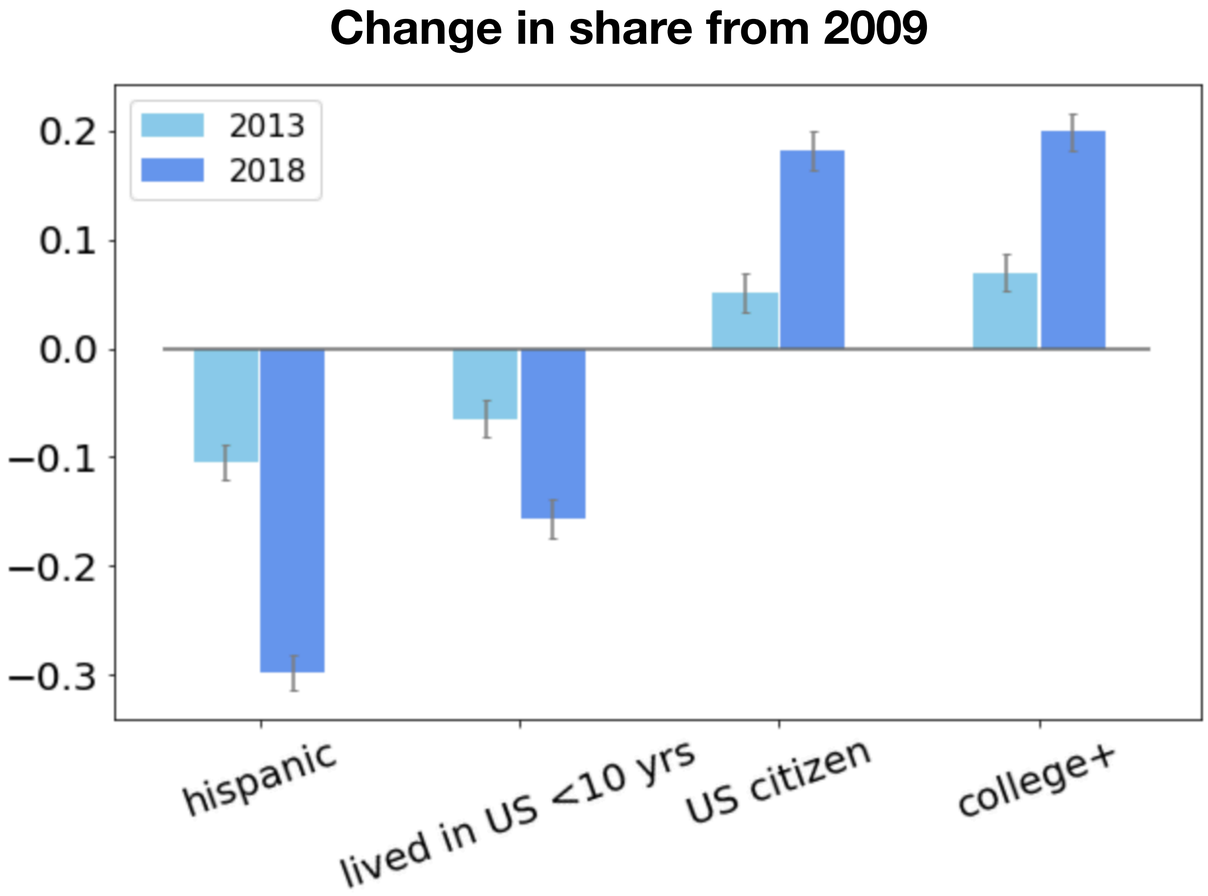
2. σ_α^2 is the ***optimal*** asymptotic variance



Effect of Medicaid on doctor visits over time

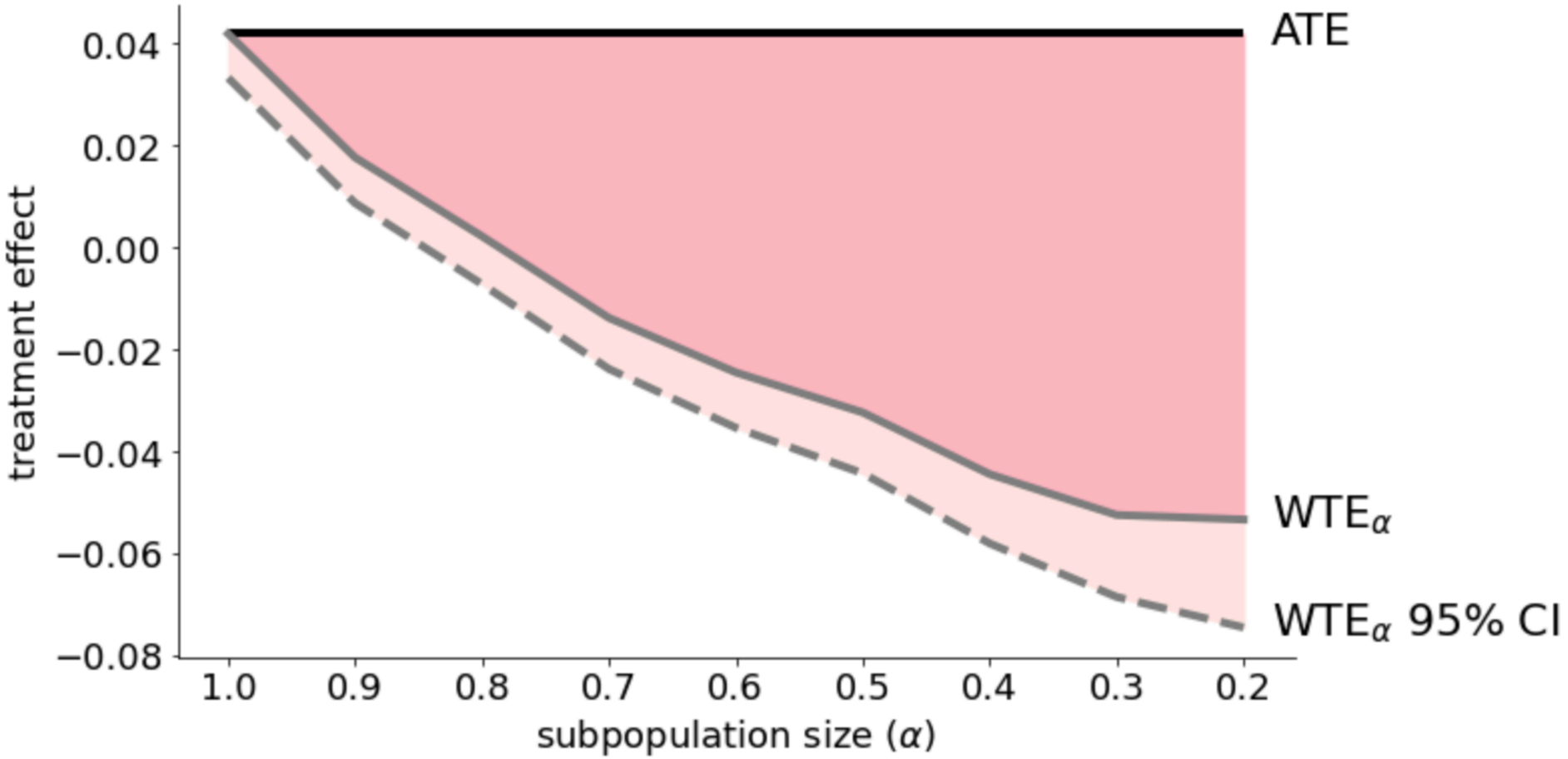
- Evaluate effect of Medicaid enrollment on doctors' office utilization
- Medicaid costs **$553 billion/yr;** need to ensure valid effects through time
- Outcome: visit to doctors in the two-weeks prior to a random survey date
- Control for demographics, medical history, employment, earnings, insurance, government assistance etc (d = 396)
- Take the viewpoint of an analyst in 2009 (n = 82,993)

Demographic compositions shift over time



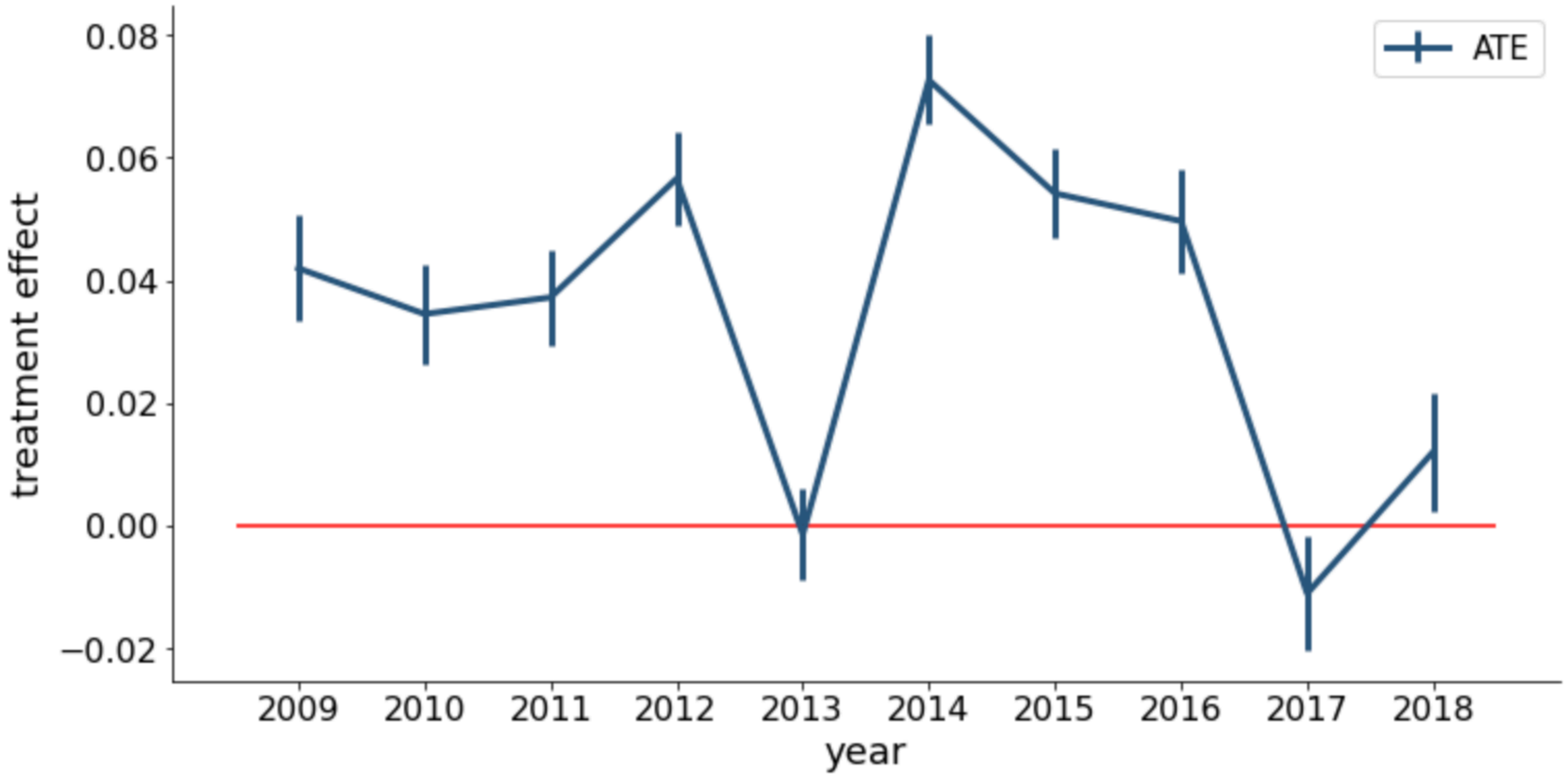
Effect of Medicaid on doctor visits over time

- Evaluate effect of Medicaid enrollment on doctors' office utilization **in 2009**



Effect of Medicaid on doctor visits over time

- Evaluate effect of effect of Medicaid enrollment on doctors' office utilization



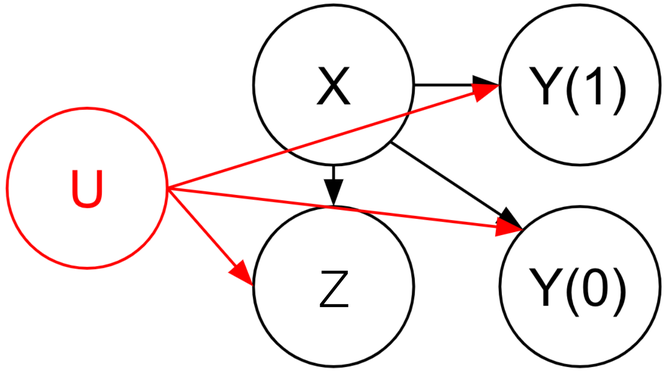
Part II: unobserved confounders

a.k.a. Y | X shift

Bounded unobserved confounding

- What if there's a hidden variable U that wasn't observed?

e.g. partial knowledge of counterfactuals



Recap

- Covariates: X
- Treatment assignment: Z
- Potential outcome: $Y(0),\ Y(1)$
- Response $Y := Y(Z)$

Relaxed assumption: Bounded unobserved confounding

There exists $\Gamma > 1$, and U such that $Y(1), Y(0) \perp Z \mid X, U,$

$$u \mapsto \frac{\mathbb{P}(Z = 1 \mid X, U = u)}{\mathbb{P}(Z = 0 \mid X, U = u)}$$ can vary by at most a factor of Γ [Rosenbaum '02]

Bounded unobserved confounding

Relaxed assumption: Bounded unobserved confounding

There exists $\Gamma > 1$, and U such that $Y(1), Y(0) \perp Z \mid X, U,$

$$u \mapsto \frac{\mathbb{P}(Z = 1 \mid X, U = u)}{\mathbb{P}(Z = 0 \mid X, U = u)}$$ can vary by at most a factor of Γ [Rosenbaum '02]

- Equivalent to a logit model: for some function $\kappa(\cdot) \in [0,1]$, $g(\cdot)$,

$$\log \frac{\mathbb{P}(Z = 1 \mid X, U)}{\mathbb{P}(Z = 0 \mid X, U)} = g(X) + \log \Gamma \cdot \kappa(U)$$

FAQs

Relaxed assumption: Bounded unobserved confounding

There exists $\Gamma > 1$, and U such that $Y(1), Y(0) \perp Z \mid X, U,$

$$u \mapsto \frac{\mathbb{P}(Z = 1 \mid X, U = u)}{\mathbb{P}(Z = 0 \mid X, U = u)}$$ can vary by at most a factor of Γ [Rosenbaum '02]

- How do I choose Γ?

➡ Domain expertise (e.g. clinical intuition)

➡ Sensitivity: what would be a clinically significant result? what value of Γ would change its significance?

- Is this the only natural confounding model?

➡ No. Today: modern semiparametric framework.

Lower bound for $\mathbb{E}[Y(1) \mid X]$

Recap

- Treatment assignment: Z
- Potential outcome: $Y(0)$, $Y(1)$
- Response $Y := Y(Z)$

- Lower bound unobservables under Γ-bounded unobserved confounding

$$\mathbb{E}[Y(1) \mid X, Z = 0] = \mathbb{E}[YL(Y \mid X) \mid X, Z = 1]$$

$$L(\cdot \mid X) := \frac{dP(Y(1) \in \cdot \mid X, Z = 1)}{dP(Y(1) \in \cdot \mid X, Z = 0)}$$

Lemma Under Γ-confounding, $y \mapsto L(y \mid x)$ can vary by at most a factor of Γ

- Minimizing over the above set of likelihood ratios,

$$\mathbb{E}[Y(1) \mid X, Z = 0] \geq \inf_{L \in \mathscr{L}_1} \mathbb{E}[YL(Y \mid X) \mid X, Z = 1] =: \theta_1^\star(X)$$

Bound is tight

Lower bound for $\mathbb{E}[Y(1) \mid X]$

unobservable

- Lower bound unobservables under Γ-bounded unobserved confounding

$$\mathbb{E}[Y(1) \mid X, Z = 0] = \mathbb{E}[YL(Y \mid X) \mid X, Z = 1]$$

$$L(\cdot \mid X) := \frac{dP(Y(1) \in \cdot \mid X, Z = 1)}{dP(Y(1) \in \cdot \mid X, Z = 0)}$$

Lemma Under Γ-confounding, $y \mapsto L(y \mid x)$ can vary by at most a factor of Γ

- Minimizing over the above set of likelihood ratios,

$$\mathbb{E}[Y(1) \mid X, Z = 0] \geq \inf_{L \in \mathscr{L}_1} \mathbb{E}[YL(Y \mid X) \mid X, Z = 1] =: \theta_1^\star(X)$$

Bound is tight

Lower bound for $\mathbb{E}[Y(1) \mid X]$

unobservable

- Lower bound unobservables under Γ-bounded unobserved confounding

$$\mathbb{E}[Y(1) \mid X, Z = 0] = \mathbb{E}[YL(Y \mid X) \mid X, Z = 1]$$

observable

$$L(\cdot \mid X) := \frac{dP(Y(1) \in \cdot \mid X, Z = 1)}{dP(Y(1) \in \cdot \mid X, Z = 0)}$$

Lemma Under Γ-confounding, $y \mapsto L(y \mid x)$ can vary by at most a factor of Γ

- Minimizing over the above set of likelihood ratios,

$$\mathbb{E}[Y(1) \mid X, Z = 0] \geq \inf_{L \in \mathscr{L}_1} \mathbb{E}[YL(Y \mid X) \mid X, Z = 1] =: \theta_1^\star(X)$$

Bound is tight

Lower bound for $\mathbb{E}[Y(1) \mid X]$

unobservable

- Lower bound unobservables under Γ-bounded unobserved confounding

$$\mathbb{E}[Y(1) \mid X, Z = 0] = \mathbb{E}[YL(Y \mid X) \mid X, Z = 1]$$

observable

$$L(\cdot \mid X) := \frac{dP(Y(1) \in \cdot \mid X, Z = 1)}{dP(Y(1) \in \cdot \mid X, Z = 0)}$$

Lemma Under Γ-confounding, $y \mapsto L(y \mid x)$ can vary by at most a factor of Γ

- Minimizing over the above set of likelihood ratios,

$$\mathbb{E}[Y(1) \mid X, Z = 0] \geq \inf_{L \in \mathscr{L}_1} \mathbb{E}[YL(Y \mid X) \mid X, Z = 1] =: \theta_1^\star(X)$$

Bound is tight

Convex Duality

Recap

- Treatment assignment: Z
- Potential outcome: $Y(0),\ Y(1)$
- Response $Y := Y(Z)$

Lemma (YNDBT'22) $\quad \theta_1^\star(X) = \sup\left\{\mu : \mathbb{E}[(Y(1)-\mu)_+ - \Gamma(Y(1)-\mu)_- \mid X, Z=1] \geq 0\right\}$

Recap

- Treatment assignment: Z
- Potential outcome: $Y(0),\ Y(1)$
- Response $Y := Y(Z)$

Convex Duality

Lemma Under Γ-confounding, $y \mapsto L(y \mid x)$ can vary by at most a factor of Γ

- Minimizing over the above set of likelihood ratios,

$$\mathbb{E}[Y(1) \mid X, Z = 0] \geq \inf_{L \in \mathscr{L}_1} \mathbb{E}[YL(Y \mid X) \mid X, Z = 1] =: \theta_1^\star(X)$$ **Bound is tight**

- **One-dimensional dual for each X**

Lemma (YNDBT'22) $\theta_1^\star(X) = \sup\left\{\mu : \mathbb{E}[(Y(1) - \mu)_+ - \Gamma(Y(1) - \mu)_- \mid X, Z = 1] \geq 0\right\}$

Recap

- Treatment assignment: Z
- Potential outcome: $Y(0),\ Y(1)$
- Response $Y := Y(Z)$

Convex Duality

Lemma Under Γ-confounding, $y \mapsto L(y \mid x)$ can vary by at most a factor of Γ

- Minimizing over the above set of likelihood ratios,

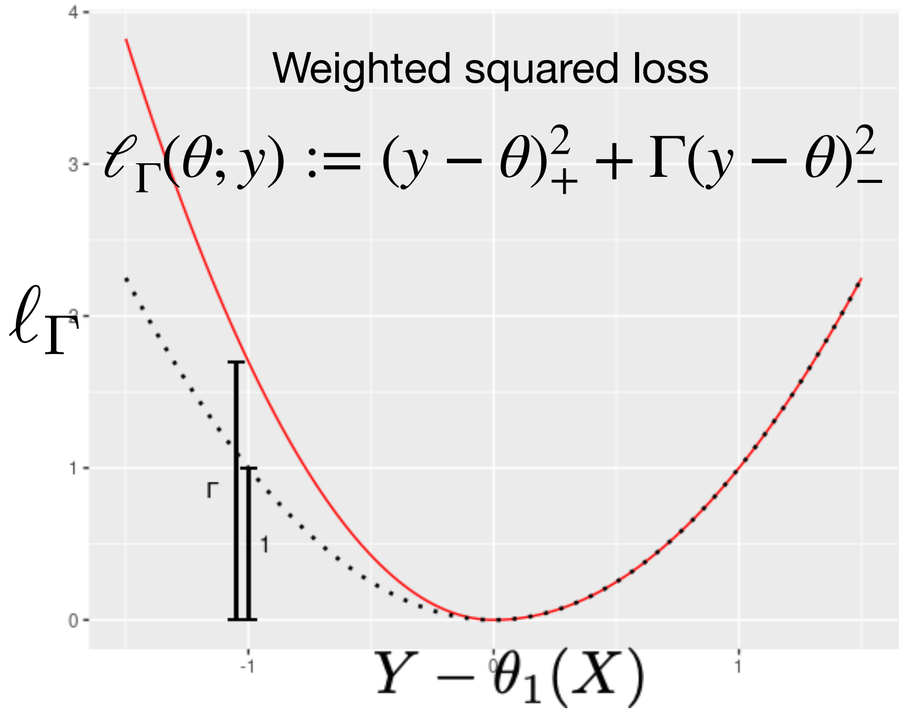
$$\mathbb{E}[Y(1) \mid X, Z = 0] \geq \inf_{L \in \mathscr{L}_1} \mathbb{E}[YL(Y \mid X) \mid X, Z = 1] =: \theta_1^\star(X)$$ **Bound is tight**

- **One-dimensional dual for each X**

Lemma (YNDBT'22) $\theta_1^\star(X) = \sup\left\{\mu : \mathbb{E}[(Y(1) - \mu)_+ - \Gamma(Y(1) - \mu)_- \mid X, Z = 1] \geq 0\right\}$

What can ML do?

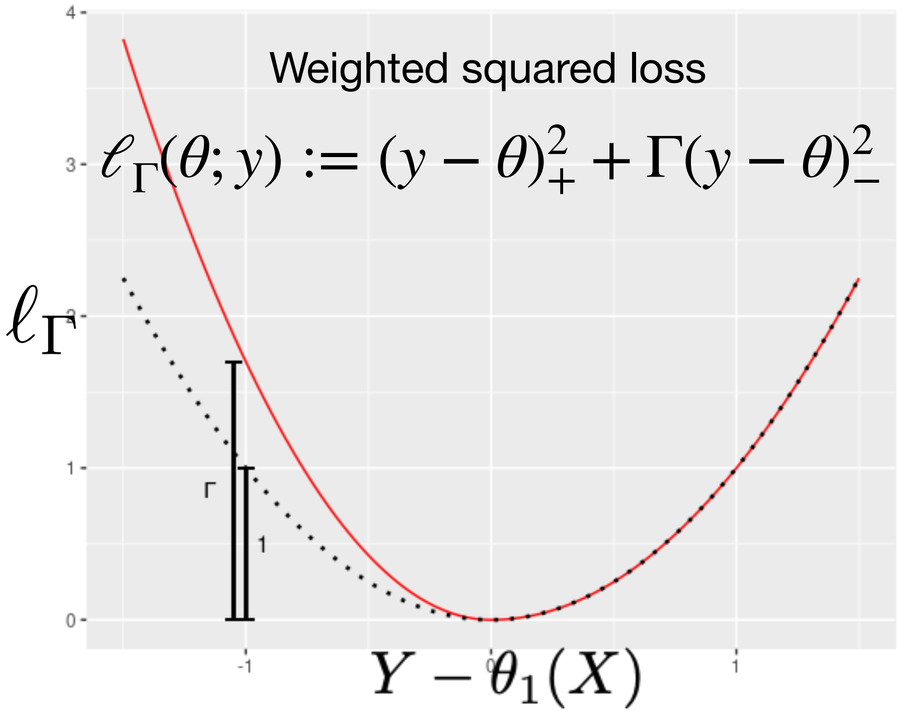
- Tremendous empirical success is **curve-fitting tools** in high-dimensions, under noisy data
- Key ingredients: stochastic optimization & model selection



Sensitivity of CATE via loss minimization

- $\mathbb{E}[Y(1) \mid X, Z = 0] \geq \theta_1^\star(X) = \sup\left\{\mu : \mathbb{E}[(Y(1) - \mu)_+ - \Gamma(Y(1) - \mu)_- \mid X, Z = 1] \geq 0\right\}$

Main result 1: $\theta_1^\star$ is the unique solution to **minimize**$_{\theta(.)}$ $\mathbb{E}[\ell_\Gamma(\theta(X); Y(1)) \mid Z = 1]$

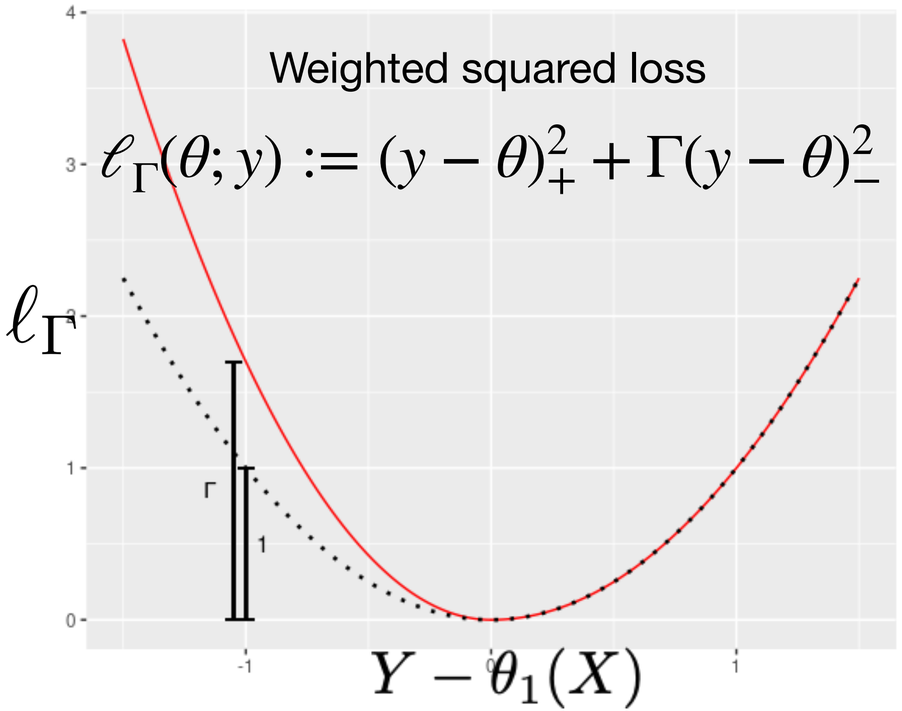


- Estimate lower bound using flexible ML models
- Solve weighted regression problem using any black-box ML approach
- e.g., random forests, boosted trees, NNs

Sensitivity of CATE via loss minimization

- $\mathbb{E}[Y(1) \mid X, Z = 0] \geq \theta_1^\star(X) = \sup\left\{\mu : \mathbb{E}[(Y(1) - \mu)_+ - \Gamma(Y(1) - \mu)_- \mid X, Z = 1] \geq 0\right\}$

Main result 1: $\theta_1^\star$ is the unique solution to **minimize**$_{\theta(.)}$ $\mathbb{E}[\ell_\Gamma(\theta(X); Y(1)) \mid Z = 1]$



- Estimate lower bound using flexible ML models
- Solve weighted regression problem using any black-box ML approach
- e.g., random forests, boosted trees, NNs

Lower bound for $\mathbb{E}[Y(1)]$

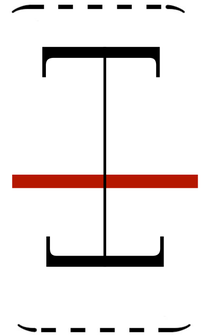
Recap

- Treatment assignment: Z
- Potential outcome: $Y(0)$, $Y(1)$
- Response $Y := Y(Z)$

- Similarly as before, we derive a **debiased estimator** for $\mu_1^- = \mathbb{E}[ZY(1) + (1 - Z)\theta_1^\star(X)] \leq \mathbb{E}[Y(1)]$
- Bounds the doubly robust estimator for the ATE; equal when $\Gamma = 1$
- **Value of prediction:** DR estimator close to worst-case bound μ_1^- (a.k.a. robust to confounding) when residuals $Y - \hat{\theta}_1(X)$ are small

Theorem Even when ML-based nuisance estimators converge at slower rates,

$$\sqrt{n}(\hat{\mu}_1^- - \mu_1^-) \Rightarrow N(0, \sigma^2)$$

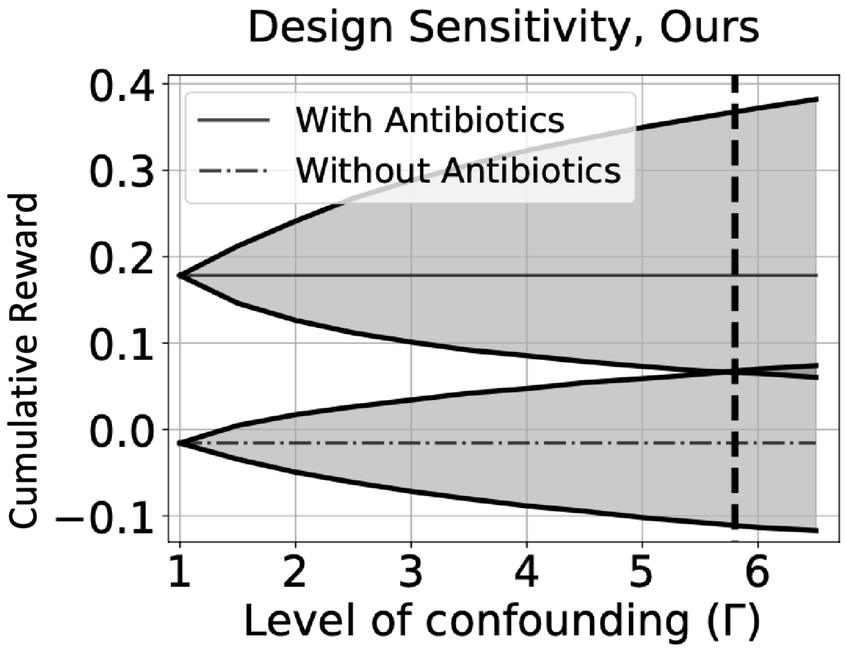


Sepsis management in the ICU

- Sepsis in ICU patients accounts for 1/3 of deaths in hospitals [Howell and Davis ’17]
- Automated approaches can manage important medication for sepsis [Futoma ’18; Komorowski 18; Raghu 17]
- ICU data suffers from unobserved confounders
- ED physician: “initial treatment of antibiotics at admission to the hospital are often confounded by unrecorded factors that affect the eventual outcome (death or discharge from the ICU).”

Proof of concept

- Whether to quickly begin antibiotic treatment is a topic of much discussion: balance early treatment vs. risks of over-prescription [Seymour '17; Sterling '15]
- Two policies: with or without antibiotics in the first step
- We use simulator developed by Oberst and Sontag (2019)



Our approach allows certifying robustness under realistic values of confounding

Summary

- Worst-case bounds on the causal effect estimated through ML models
- Debiasing: CLT even when nuisance estimates converge slower; **optimal**
- Guard against brittle findings that do not hold under distribution shift

Assessing External Validity Over Worst-case Subpopulations.
Jeong & N. Under review. Short version appeared in COLT 2020.
Bounds on the conditional and average treatment effect with unobserved confounding factors.
Yadlowsky, N., Basu, Duchi, and Tian. Annals of Statistics, 2022.
Off-policy policy evaluation for sequential decisions under unobserved confounding.
N., Keramati, Yadlowsky, and Brunskill. NeurIPS 2020.