

# Causality (cont.)

<https://www.cambridge.org/core/books/causal-inference-for-statistics-social-and-biomedical-sciences/71126BE90C58F1A431FE9B2DD07938AB>

**<https://onlinelibrary.wiley.com/doi/abs/10.1111/ectj.12097>**

**<https://www.pnas.org/content/116/10/4156>**

**<https://arxiv.org/pdf/1712.04912.pdf>**

# Potential outcomes

- Framework for explicitly modeling counterfactuals
- $A$ : binary treatment assignment (1: treated, 0: control)
- $Y(1)$  and  $Y(0)$  are potential outcomes
- $X$  is observed covariates

**First goal:** Estimate average treatment effect

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

**Problem:** We only observe  $Y := Y(A)$

# No unobserved confounding

- Previous regression-based direct method still works if there are no unobserved confounders (also called ignorability)

**Assumption.**  $Y(1), Y(0) \perp A \mid X$

- Observed treatment assignments are based on covariate information alone (+ random noise)
  - Treatment assignment does not use information about counterfactuals
- Strong assumption. Often violated in practice.
  - e.g. doctors often use unrecorded info to prescribe treatments

# Overlap

- We need enough samples for both control and treatment throughout the covariate space
  - This governs the effective sample size
- Propensity score  $e^*(X) := \mathbb{P}(A = 1 | X)$
- Assume that there exists  $\epsilon > 0$  such that  $\epsilon \leq e^*(X) \leq 1 - \epsilon$  almost surely
- This means I have at least  $\epsilon n$  number of samples for fitting the two outcome models

# Overlap

- This breaks if data is generated by a deterministic policy
  - e.g. always assign the drug (treatment) when  $\text{age} > 50$
- We need sufficient amount of randomness in treatment assignment in all covariate regions
- Governs difficulty of estimation. Often violated in practice.

# Direct method

- By no unobserved confounding,  
$$\begin{aligned}\mu_a^\star(X) &:= \mathbb{E}[Y(a) \mid X] = \mathbb{E}[Y(a) \mid X, A = a] \\ &= \mathbb{E}[Y \mid X, A = a]\end{aligned}$$
  **observable**
- Fit  $\mu_a^\star(X)$  via the loss minimization problem  
$$\text{minimize}_{\mu_a \in \mathfrak{M}_a} \mathbb{E}[(Y - \mu_a(X))^2 \mid A = a]$$
- ATE estimator  $\hat{\tau}_{\text{DM}} := \frac{1}{n} \sum_{i=1}^n \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)$
- Good if the outcome models are easy to learn

# Inverse propensity weighting

- What if the outcome models are very complex and difficult to estimate?
- A natural approach is to reweight samples to correct for confounding bias
  - Essentially importance sampling
- First, estimate the propensity score  $e^*(X) := \mathbb{P}(A = 1 | X)$ 
  - e.g. run logistic regression to predict A given X

# Inverse propensity weighting

$$\hat{\tau}_{\text{IPW}} := \frac{1}{n} \sum_{i=1}^n \left( \frac{A_i}{\hat{e}(X_i)} Y_i - \frac{1 - A_i}{1 - \hat{e}(X_i)} Y_i \right)$$

- Can work well if propensity score is simple to estimate
- But estimating this well over the entire covariate space can be difficult
  - Calibration is hard, especially in high-dimensions
- When overlap doesn't hold, importance weights blow up

# Augmented IPW

- Can we combine the best of both worlds?
  - Direct method + IPW
- Propensity weight residuals to debias the direct method

$$\begin{aligned}\hat{\tau}_{\text{AIPW}} := & \frac{1}{n} \sum_{i=1}^n (\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)) \\ & + \frac{1}{n} \sum_{i=1}^n \left( \frac{A_i}{\hat{e}(X_i)} (Y_i - \hat{\mu}_1(X_i)) - \frac{1 - A_i}{1 - \hat{e}(X_i)} (Y_i - \hat{\mu}_0(X_i)) \right)\end{aligned}$$

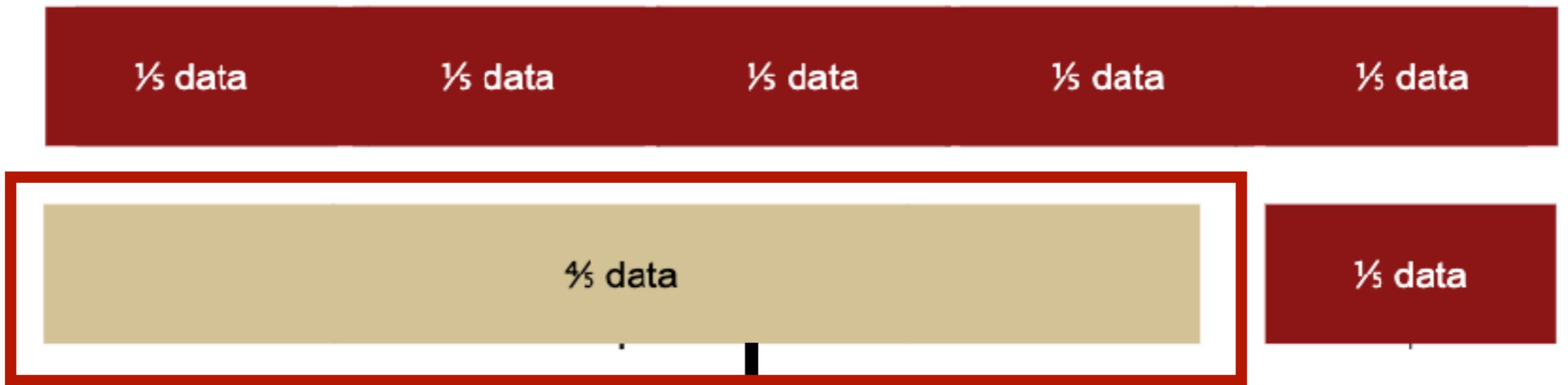
# Augmented IPW

- Best asymptotic variance; semiparametrically efficient
- Doubly robust: asymptotically consistent as long as either outcome model or the propensity score model is well-specified
- Insensitive to errors in nuisance parameters  $\mu_a^\star, e^\star$ 
  - Neyman orthogonality gives central limit behavior so long as  $\|\hat{e} - e^\star\|_{P,2}(\|\hat{\mu}_1 - \mu_1^\star\|_{P,2} + \|\hat{\mu}_0 - \mu_0^\star\|_{P,2}) = o_p(n^{-1/2})$

# Cross-fitting

- Instead of sample-splitting, we can alternate the role of main and auxiliary samples over multiple splits

**Cross-fitting**  
[Chernozhukov '18]



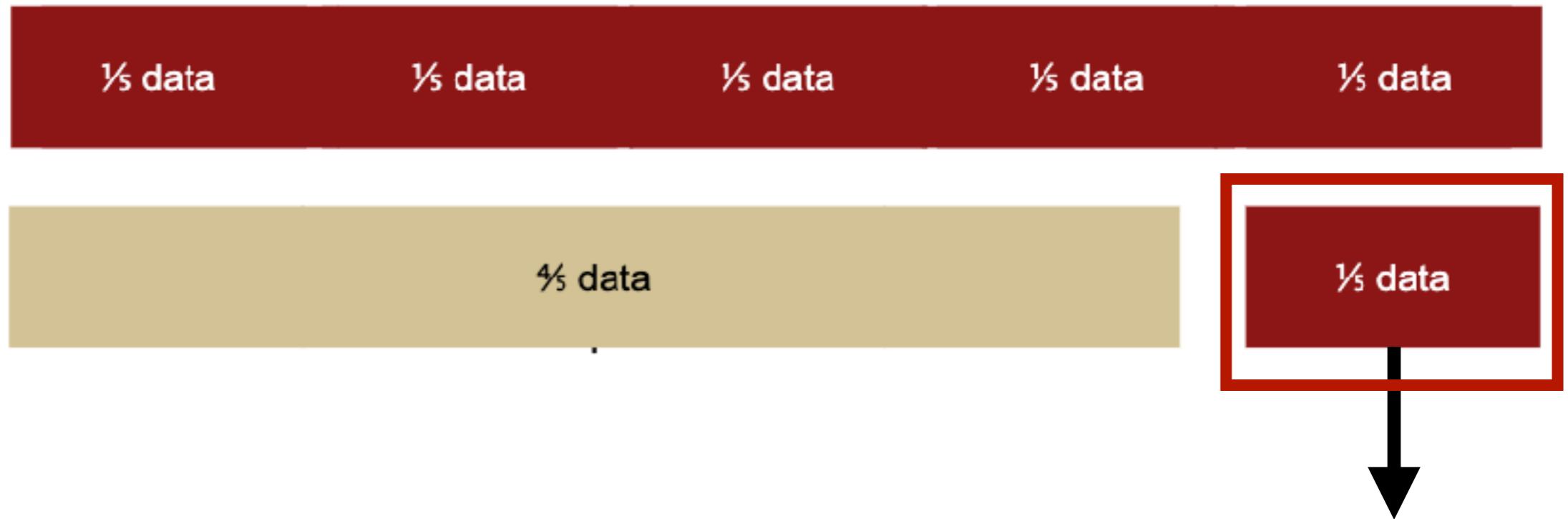
$$\hat{\mu}_a(X) \approx \mathbb{E}[Y(a) \mid X = x], \quad a \in \{0, 1\}$$

$$\hat{e}(X) \approx \mathbb{P}(A = 1 \mid X)$$

- Estimate nuisance parameters on the auxiliary sample

# Cross-fitting

Cross-fitting  
[Chernozhukov '18]



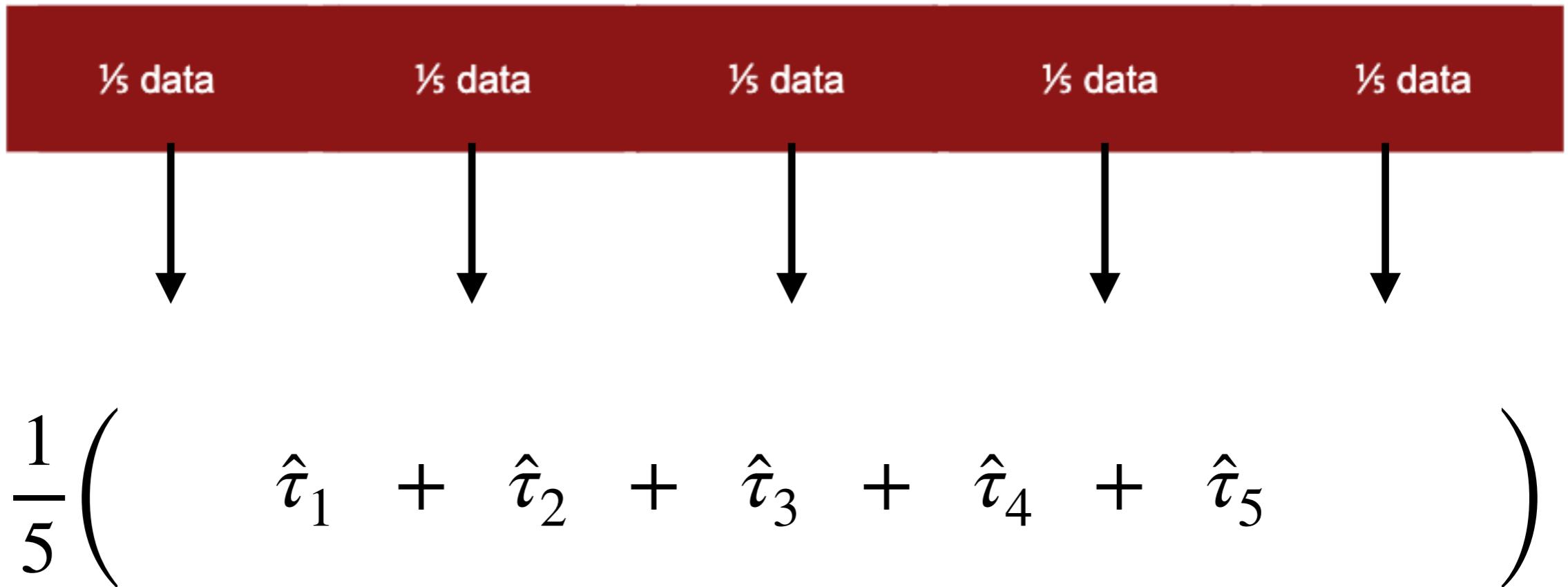
$$\hat{\tau}_1 := \frac{1}{n} \sum_{i=1}^n \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + \frac{A_i}{\hat{e}(X_i)}(Y - \mu_1(X_i)) - \frac{1 - A_i}{1 - \hat{e}(X_i)}(Y - \mu_0(X_i))$$

- Estimate ATE by plugging in nuisance estimates

# Cross-fitting

## Cross-fitting

[Chernozhukov '18]



- Same procedure for direct method, IPW
- Similar central limit result follows as before

# SUTVA

- Throughout we implicitly assumed there is only a single version of the treatment that gets applied to all treated units
  - This may not be true if drugs go stale in storage, or dosages differ
- We also assumed there is *no interference between units*
  - Whether or not individual i is treated has no impact on the treatment effect of another individual j
  - This can also fail in many real-world scenarios
- Together these assumptions are called stable unit treatment value assumption (SUTVA)

# Interference

- Any two-sided platform faces interference between units
- Consider the following scenario:
  - Lyft A/B tests a new promotion strategy for drivers
  - Each driver is randomized into treatment or control
  - It is observed that drivers finish a lot more rides with the promotion
  - So they decide this promotion is worth spending resources on
- But the estimate turned out to be an **overestimate**, not worth the cost of the promotion. Why?

# Interference

- Both treated and control drivers see the same set of demand
- If promotion incentivizes treated drivers to work more for less nominal fares, this cannibalizes demand that would usually go to control drivers
- Interference occurs in a number of different settings
  - Two-sided platforms: Airbnb, ridesharing, ad auctions
  - Network effects: e.g. adoption of new education technology
- When this happens, the potential outcomes now depend on all possible  $2^n$  treatment assignments
  - Very active area of research

# Assessing overlap

- “If the covariate distributions are similar, as they would be, in expectation, in the setting of a completely randomized experiment, there is less reason to be concerned about the sensitivity of estimates to the specific method chosen than if these distributions are substantially different.”
- “On the other hand, even if unconfoundedness holds, it may be that there are regions of the covariate space with relatively few treated units or relatively few control units, and, as a result, inferences for such regions rely largely on extrapolation and are therefore less credible than inferences for regions with substantial overlap in covariate distributions.”
- Imbens and Rubin

# Assessing overlap

- Overlap governs effective sample size
  - Even approaches that don't require propensity weighting is affected under this fundamental restriction
- Causal inference literature has developed various “supplementary analysis” tools for assessing credibility of empirical claims
- One of the most common conventions is to plot the propensity scores of treated and control groups

# Assessing overlap

- Difference in covariate distributions between treatment and control group is summarized by the propensity score
- Let  $f_1(X)$  be the density of  $X$  in the treatment group (similarly  $f_0(X)$ )
- Let  $p := \mathbb{P}(A = 1)$

$$\text{Var}(e^{\star}(X)) = p(1 - p)(\mathbb{E}[e^{\star}(X) | A = 1] - \mathbb{E}[e^{\star}(X) | A = 0])$$

$$= p^2(1 - p)^2 \cdot \mathbb{E} \left[ \left( \frac{f_1(X) - f_0(X)}{pf_1(X) + (1 - p)f_0(X)} \right)^2 \right]$$

# Assessing overlap

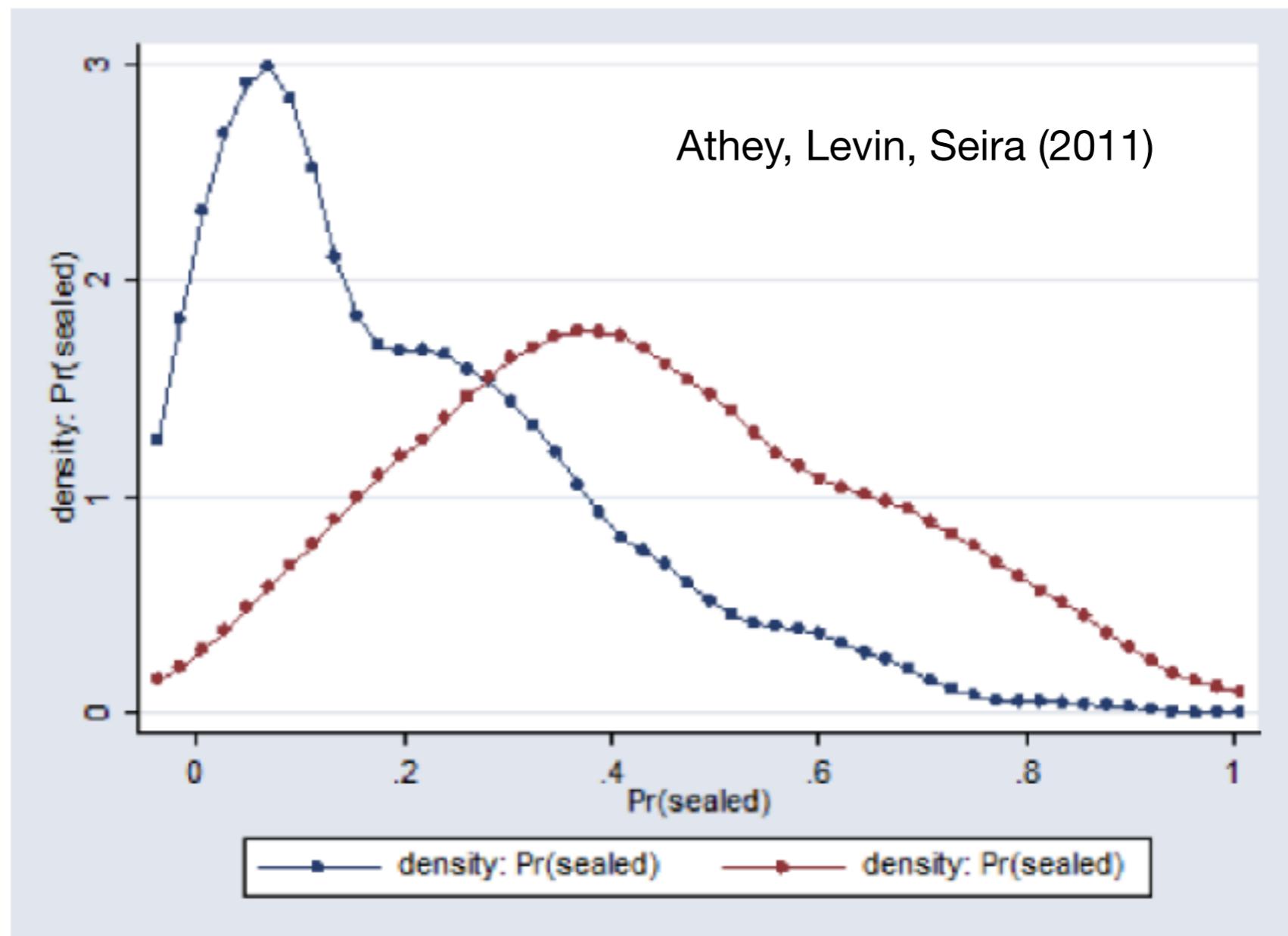
- A common visualization is to look at the pdf of the propensity score across treatment groups
- Plot approximates pdfs of the distribution  
 $\mathbb{P}(e^*(X) \in \cdot | A = a)$
- For each  $q \in (0,1)$ , plot fraction of observations in the treatment group with  $e^*(x) = q$  (and similarly for control)

# Assessing overlap

- Athey, Levin, Seira (2011) studied timber auctions
  - Award timber harvest contracts via first price sealed auction or open ascending auction
- Idaho: randomized with different probabilities across different regions
- California: determined by small vs. large sales volume; cutoff varies by region

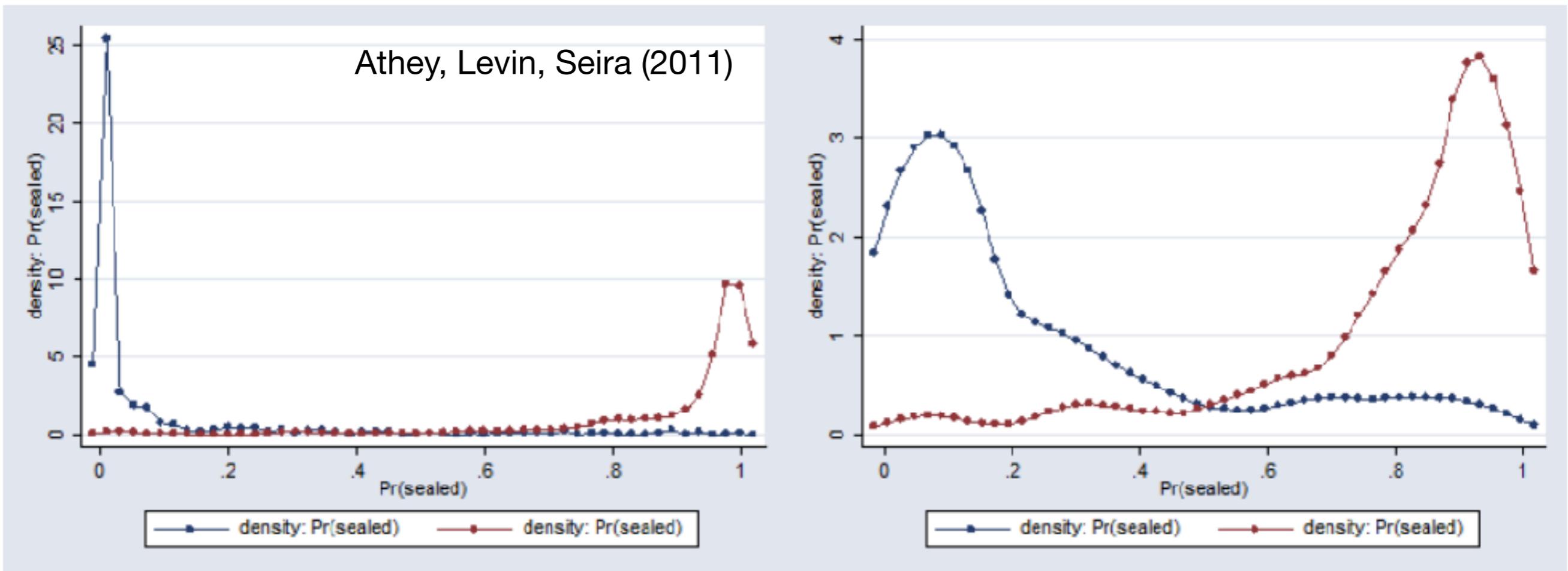
# Idaho

Very few observations with extreme propensity scores



# California

Untrimmed v. trimmed so that  $e(x) \in [.025, .975]$



# Heterogenous Treatment Effects

# CATE

- Treatment effect often varies with user / patient / agent characteristics (covariates)
- To estimate personalized treatment effects, we want to estimate the **conditional average treatment effect (CATE)**

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

- Few different ways to estimate this using black-box ML models
- Again, key challenging is missing data
  - We never observed counterfactuals

# S-Learner

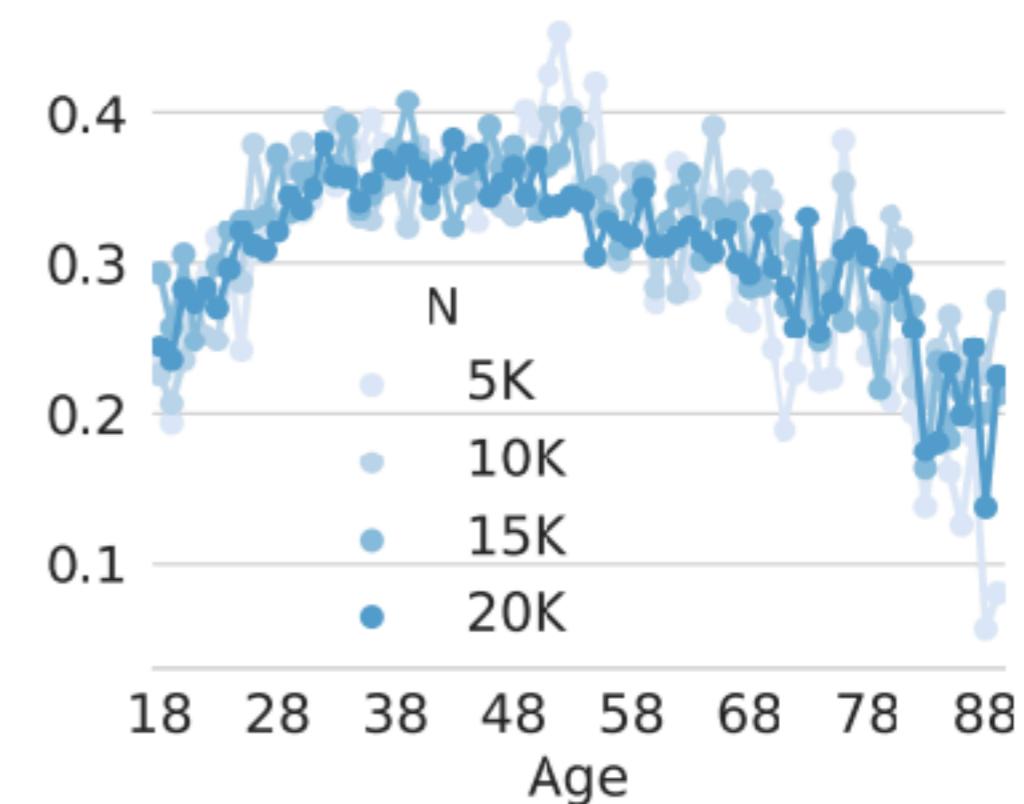
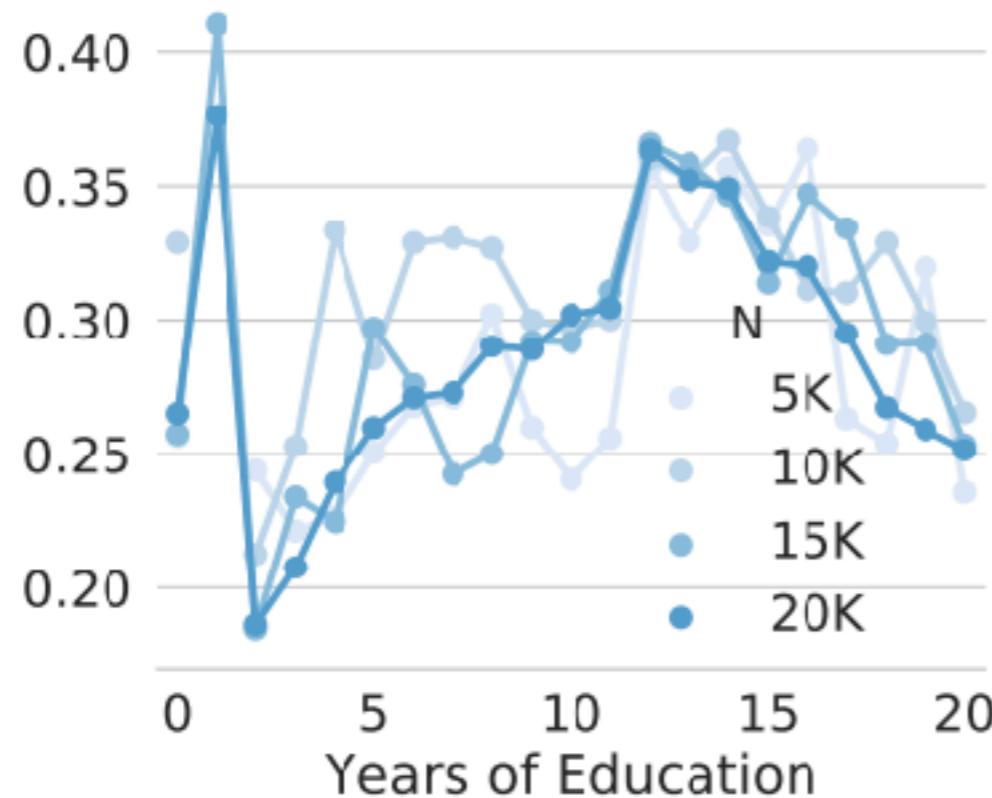
- By no unobserved confounding,  
$$\begin{aligned}\mu^*(a, x) &:= \mathbb{E}[Y(a) \mid X = x] = \mathbb{E}[Y(a) \mid X = x, A = a] \\ &= \mathbb{E}[Y \mid X = x, A = a]\end{aligned}$$
- Fit  $\mu^*(a, x)$  via the loss minimization problem  
$$\text{minimize}_{\mu \in \mathfrak{M}} \quad \mathbb{E}[(Y - \mu(A, X))^2]$$
- $\hat{\tau}(X) := \hat{\mu}(1, X) - \hat{\mu}(0, X)$
- Shared feature representation, assuming similar model class for both treatment and control

# T-Learner

- By no unobserved confounding,  
$$\begin{aligned}\mu_a^*(X) &:= \mathbb{E}[Y(a) \mid X] = \mathbb{E}[Y(a) \mid X, A = a] \\ &= \mathbb{E}[Y \mid X, A = a]\end{aligned}$$
- Fit  $\mu_a^*(X)$  via the loss minimization problem  
$$\text{minimize}_{\mu_a \in \mathfrak{M}_a} \mathbb{E}[(Y - \mu_a(X))^2 \mid A = a]$$
- $\hat{\tau}(X) := \hat{\mu}_1(X) - \hat{\mu}_0(X)$
- Can fit different models over treatment options

# Welfare attitudes experiment

- Evaluate effect of wording on survey results (“welfare” vs “assistance to the poor”)
- Resoundingly positive treatment effects, but significant heterogeneity across covariates



# X-Learner

Kunzel et al. (2018)

- Regress on the imputed treatment effect  $Y(1) - Y(0)$
- Fit T-learner models and compute imputed treatment effects

$$Y_i - \hat{\mu}_{\theta,0}(X_i) \text{ if } A_i = 1, \hat{\mu}_{\theta,1}(X_i) - Y_i \text{ if } A_i = 0$$

- Fit another set of models  $\hat{\tau}_1, \hat{\tau}_0$  on the two category of imputed values, take

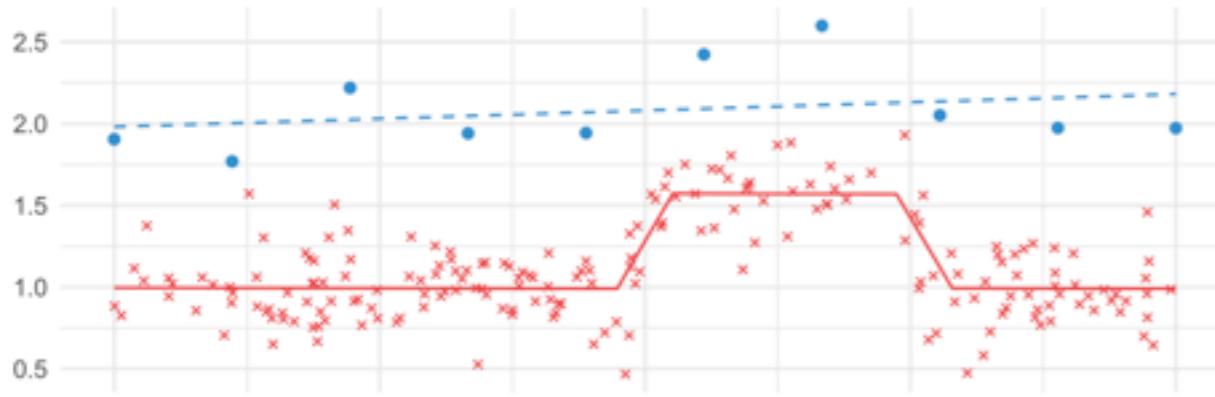
$$\hat{\tau}(X) := \hat{e}(X)\hat{\tau}_0(X) + (1 - \hat{e}(X))\hat{\tau}_1(X)$$

# X-Learner

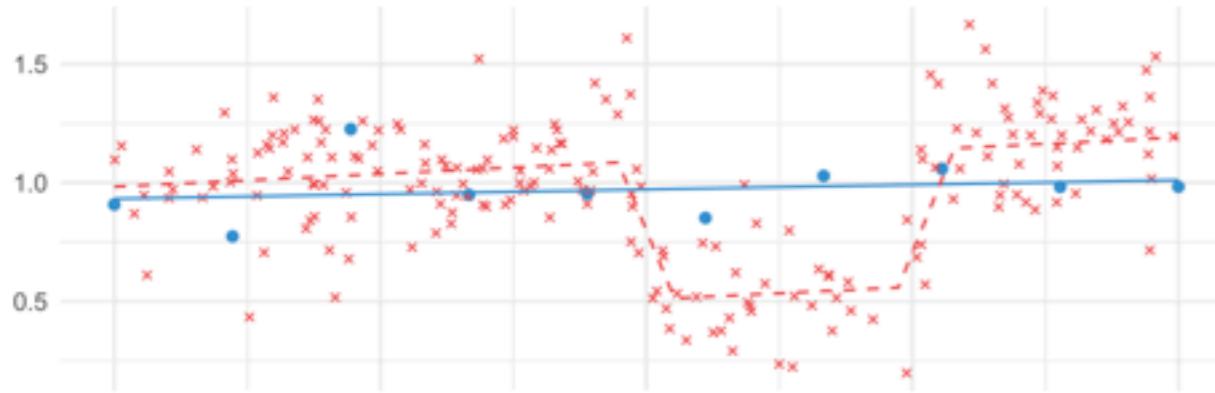
Kunzel et al. (2018)

- Usually, number of samples in treatment >> those in control
- Advantageous if CATE is much smoother than individual outcome functions

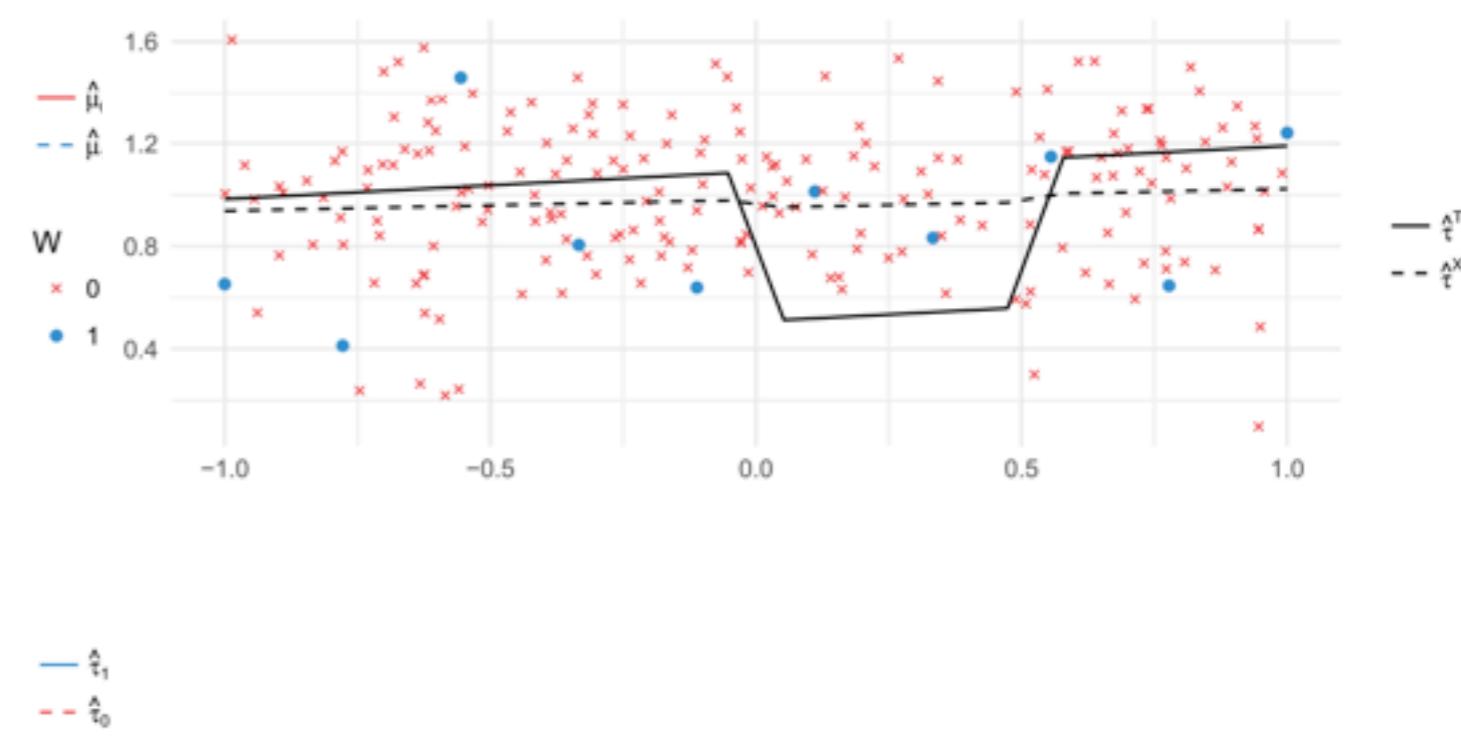
A Observed Outcome & First Stage Base Learners



B Imputed Treatment Effects & Second Stage Base Learners



C Individual Treatment Effects & CATE Estimators



# R-Learner

$$\begin{cases} M_0^* & \text{if } a=0 \\ M_1^* & \text{if } a=1 \end{cases}$$

Robinson's decomposition

Define  $\varepsilon(a) := Y(a) - (M_a^*(x) + aZ(x))$ ,  $\varepsilon := \varepsilon(A)$ .

From no unobserved confounding,

$$\begin{aligned} E[\varepsilon | x] &= e^*(x) E[\varepsilon(1) | x, A=1] + (1 - e^*(x)) E[\varepsilon(0) | x, A=0] \\ &= e^*(x) E[Y(a) - M_a^*(x) | x, A=1] + (1 - e^*(x)) E[Y(a) - M_a^*(x) | x, A=0] \\ &= e^*(x) E[Y(a) - M_a^*(x) | x] + (1 - e^*(x)) E[Y(a) - M_a^*(x) | x] \\ &= 0 \end{aligned}$$

By def of  $\varepsilon$ ,  $Y = M_0(x) + AZ(x) + \varepsilon$ .  $\therefore (*)$

Define  $m^*(x) := E[Y | x]$ . Taking  $E[\cdot | x]$  on both sides

$$m^*(x) = M_0(x) + e^*(x)Z(x).$$

# R-Learner

Subtracting this from (\*), we arrive at

$$Y - \hat{m}(x) = (A - \hat{e}(x))z(x) + \xi.$$

So if we fit  $\hat{m}, \hat{e}$  on heldout data,  
we can now solve

$$\min_{\theta} \mathbb{E}[(Y - \hat{m}(x) - (A - \hat{e}(x))z_\theta(x))^2]$$

to get  $\hat{z}_\theta(x)$ .

# Sensitivity Analysis

# Observational studies

- When experimentation is risky, crucial to leverage collected data
- Historically, many important findings from observational data
  - “citrus fruit curing scurvy described in the 1700s or insulin as a treatment for diabetes in the 1920s long preceded the advent of the modern randomized clinical trial.”
  - “these methods had in common a reliable method of diagnosis, a predictable clinical course, and a large and obvious effect of the treatment.” [Corrigan-Curay et al. 2018]
- These results need to be contextualized and viewed with more skepticism than RCTs

# Unobserved confounding

- So far, we assumed that there are no unobserved confounders that simultaneously affect potential outcomes and treatment assignments
- What if there's a hidden variable  $U$  that wasn't observed?

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

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

Other examples: Antioxidant vitamin beta carotene [Willett '90, ATBC CPSG '94](#)

Hormone replacement therapy [Pedersen '03 WHI, Lawlor '04  
\[Rutter '07\]](#)

- Even in tech, important features are unrecorded due to privacy or data management issues

# Unobserved confounding

- Clinicians use visual observations or discussions with patients to inform treatment decisions (e.g. admission to NICU)
- Drugs are preferentially prescribed to patients for which it will be effective, or those who can tolerate them
- These factors are not properly recorded even at the resolution of large databases.
- Example: Patients in emergency departments often do not have an existing record in the hospital's electronic health system. This leaves important information unobserved in subsequent observational analysis.

# Bounded unobserved confounding

- What if there's a hidden variable  $U$  that wasn't observed
  - Estimates can be arbitrarily bad under general confounding
- Often it is reasonable to assume an unobserved confounder has bounded effect on observed treatment assignments
  - Odds ratio of treatment can only vary by up to a factor of  $\Gamma > 1$

**Relaxed assumption: Bounded unobserved confounding**

$$\frac{1}{\Gamma} \leq \frac{\mathbb{P}(A = 1 \mid X, \textcolor{red}{U} = u)}{\mathbb{P}(A = 0 \mid X, \textcolor{red}{U} = u)} \frac{\mathbb{P}(A = 0 \mid X, \textcolor{red}{U} = u')}{\mathbb{P}(A = 1 \mid X, \textcolor{red}{U} = u')} \leq \Gamma$$

and  $Y(1), Y(0) \perp\!\!\!\perp A \mid X, U$

[Rosenbaum '02]

- Such  $U$  always exists since we can set  $U = (Y(1), Y(0))$

# Equivalence

Let there exist a random variable  $U$  such that  $Y(1), Y(0) \perp\!\!\!\perp A \mid X, U$

There exists a  $\Gamma > 1$  such that

$$\frac{1}{\Gamma} \leq \frac{\mathbb{P}(A = 1 \mid X, U = u)}{\mathbb{P}(A = 0 \mid X, U = u)} \frac{\mathbb{P}(A = 0 \mid X, U = u')}{\mathbb{P}(A = 1 \mid X, U = u')} \leq \Gamma \text{ a.s.}$$

if and only if there exists  $f(X), g(X, U)$  s.t.  $g(X, U) \in [0,1]$  a.s. and

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

Odds ratio of treatment can only vary by up to a factor of  $\Gamma$



Bounded influence of  $U$  in a nonparametric logistic regression model

# Equivalence

$\Rightarrow$

Define  $\underline{e}(x) := \text{essinf}_u \text{IP}(A=1 | X, u)$ .

cf.  $\text{essinf } f = \sup \{a : \text{IP}(\text{IP}(A=1 | X, u) \leq a | X) = 0\}$

Let  $f(x) := \log \frac{\underline{e}(x)}{1 - \underline{e}(x)}$  and  $g(x, u) := \frac{1}{\log \Gamma} \left( \log \frac{\text{IP}(A=1 | X, u)}{\text{IP}(A=0 | X, u)} - f(x) \right)$

By construction,  $g(x, u) \geq 0$  a.s. and  $\log \frac{\text{IP}(A=1 | X, u)}{\text{IP}(A=0 | X, u)} = f(x) + g(x, u) \cdot \log \Gamma$

From the  $\Gamma$ -bdd confounding condition,

$$\frac{\text{IP}(A=1 | X, u=u)}{\text{IP}(A=0 | X, u=u)} \underset{\text{essinf } u'}{=} \frac{\text{IP}(A=1 | X, u=u')}{\text{IP}(A=0 | X, u=u')} = \frac{\text{IP}(A=1 | X, u=u)}{\text{IP}(A=0 | X, u=u')} \underset{\text{cp}(f(x))}{\leq} \Gamma$$

so we have  $g(x, u) \leq 1$  a.s.

# Equivalence

$$\begin{aligned} \Leftarrow & \frac{P(A=1 | X, U=u)}{P(A=0 | X, U=u)} \frac{P(A=0 | X, U=u')}{P(A=1 | X, U=u')} = \frac{\exp(f(x) + g(x, u) \cdot \log \Gamma)}{\exp(f(x) + g(x, u') \cdot \log \Gamma)} \\ & = \exp(\log \Gamma \cdot (g(x, u) - g(x, u'))) \leq \Gamma \\ & \geq 1/\Gamma \end{aligned}$$

since  $g(x, u), g(x, u') \in [0, 1]$

# FAQs

## Bounded unobserved confounding

$$\frac{1}{\Gamma} \leq \frac{\mathbb{P}(A = 1 \mid X, \textcolor{red}{U} = u)}{\mathbb{P}(A = 0 \mid X, \textcolor{red}{U} = u)} \frac{\mathbb{P}(A = 0 \mid X, \textcolor{red}{U} = u')}{\mathbb{P}(A = 1 \mid X, \textcolor{red}{U} = u')} \leq \Gamma$$

and  $Y(1), Y(0) \perp\!\!\!\perp A \mid X, U$  [Rosenbaum '02]

- How do I choose  $\Gamma$ ?
  - Domain expertise (e.g. clinical intuition)
  - Reverse thinking: what would be a clinically significant result? what value of  $\Gamma$  would change its significance?
  - Sensitivity of a study: at what level of  $\Gamma$  is the conclusion of the study invalidated?
- Is this the only natural confounding model?
  - No. Today we discuss a modern semiparametric framework under this model; the framework may be developed under different models.

## Goal

Derive population-level bounds on the ATE  $\mathbb{E}[Y(1) - Y(0)]$  and CATE  $\mathbb{E}[Y(1) - Y(0)|X]$ .

We consider each potential outcome separately, and later combine upper/lower bounds on  $\mathbb{E}[Y(1)|X]$  &  $\mathbb{E}[Y(0)|X]$ .

W.l.o.g. focus on lower bounds for  $Y(1)$ .

Begin with the familiar decomposition

$$\begin{aligned}\mathbb{E}[Y(1)|X] &= \mathbb{E}[Y(1)|X, A=1] P(A=1|X) + \mathbb{E}[Y(1)|X, A=0] P(A=0|X) \\ &= \mathbb{E}[Y(1)|X, A=1] e^*(x) + \mathbb{E}[Y(1)|X, A=0] (\perp e^*(x))\end{aligned}$$

*unobserved counterfactual*

We bound the unobserved component over distributions satisfying

$$\Gamma\text{-bounded unobserved confounding } \frac{1}{F} \leq \frac{P(A=1|X, U=u)}{P(A=0|X, U=u)} \frac{P(A=0|X, U=u')}{P(A=1|X, U=u')} \leq \Gamma \text{ a.s.}$$

(\*)

where  $U$  is s.t.  $Y(1), Y(0) \perp\!\!\!\perp A | X, U$ .

Idea If  $L(y_1|X) := \frac{dP(Y(1) \in \cdot | X, A=0)}{dP(Y(1) \in \cdot | X, A=1)}(y_1)$  exists, then

$$\mathbb{E}[Y(1)|X, A=0] = \mathbb{E}[Y(1)L(Y(1)|X) | X, A=1].$$

observable

Since  $L$  is unknown, take a worst-case approach over  $P$  satisfying (\*)

↳ This will be observable

Lemma Under (\*),  $L_1$  exists and satisfies  $0 \leq \frac{L_1(y_1|X)}{L_1(y_i|X)} \leq \Gamma$   $y, y_i, x \text{ a.s.} \dots$  (\*)

Converse also holds:

For any  $L_1$  satisfying (\*),  $\exists P$  satisfying (\*) s.t.  $L_1$  is the likelihood ratio

$$\frac{dP(Y(1) \in \cdot | X, A=0)}{dP(Y(1) \in \cdot | X, A=1)}(y_1) = L_1(y_1|X).$$

So we take the lower bound

$$\Theta_i^*(x) := \inf_{\tilde{L} \geq 0} \mathbb{E}\left[\tilde{Y}_i(1) \tilde{L}_i(Y_i(1)) \mid X, A=1\right]$$

$$\text{st. } \mathbb{E}[\tilde{L}_i(Y_i(1)) \mid X, A=1] = 1,$$

$0 \leq \tilde{L}_i(y_i) / \tilde{L}_i(\tilde{y}_i) \leq \Gamma$   $y_i, \tilde{y}_i$ -a.s. according to  $\mathbb{P}(Y_i(1) \in \cdot \mid X, A=1)$

$$\text{Then, } \mathbb{E}[Y_i(1) \mid X] = e^*(x) \mathbb{E}[Y_i(1) \mid X, A=1] + (1 - e^*(x)) \mathbb{E}[Y_i(1) \mid X, A=0]$$

$$\geq e^*(x) \mathbb{E}[Y_i(1) \mid X, A=1] + (1 - e^*(x)) \Theta_i^*(x)$$

↳ We estimate this observable lower bound

Proof of Lemma, main direction

From  $(*)$ , we get

$$\mathbb{P}(A=1 \mid X, u=u) / \mathbb{P}(A=0 \mid X, u=u') \leq \Gamma \mathbb{P}(A=0 \mid X, u=u) / \mathbb{P}(A=1 \mid X, u=u')$$

Integrating over  $u'$  and  $u \in S$ , where  $S$  is an arbitrary measurable set,

$$\frac{\mathbb{P}(A=1 \mid X, u \in S)}{\mathbb{P}(A=0 \mid X, u \in S)} \cdot \frac{\mathbb{P}(A=0 \mid X)}{\mathbb{P}(A=1 \mid X)} = \frac{\mathbb{P}(u \in S \mid A=0, X)}{\mathbb{P}(u \in S \mid A=1, X)} \in [\Gamma^{-1}, \Gamma]$$

Let  $g_\alpha$  the conditional density of  $U \mid A=\alpha, X$ , this implies  $\frac{g_0(u \mid X)}{g_1(u \mid X)} \in [\Gamma^{-1}, \Gamma]$ .

Now note that for any  $S' \subseteq \mathcal{Y}$ , Bayes rule gives

$$\frac{\mathbb{P}(Y_i(1) \in S' \mid A=0, X)}{\mathbb{P}(Y_i(1) \in S' \mid A=1, X)} = \frac{\int \mathbb{P}(Y_i(1) \in S' \mid A=0, X, u=u) g_0(u \mid X) d\mathbb{M}(u)}{\int \mathbb{P}(Y_i(1) \in S' \mid A=1, X, u=u) g_1(u \mid X) d\mathbb{M}(u)} = \text{Ratio}$$

Quasi-convexity:  $f(\lambda z_1 + (1-\lambda)z_2) \leq \max(f(z_1), f(z_2)) \quad \forall \lambda \in [0, 1] \Rightarrow f(\mathbb{E}z) \leq \text{esssup } f(z)$

Since  $(a, b) \mapsto \frac{a}{b}$  is quasi-convex when  $a, b > 0$ , Ratio  $\in [\Gamma^{-1}, \Gamma]$ .  $\square$

Next, we derive a representation for  $\Theta_i^*(x)$  that is efficiently learnable.

Next, we derive a dual form of  $\theta_i^+$ .

Lemma Let  $\Psi_\theta(y) := (y-\theta)_+ - \Gamma(y-\theta)_+$ .

If  $|\theta_i'(x)| < \infty$ , then  $\theta_i^+(x) = \sup_M \{ M : \mathbb{E}[\Psi_M(Y(1)) | A=1, X=x] \geq 0 \}$

Pf Everything is conditional on  $X$ , so we abuse notation and omit it.

Since  $\tilde{L}_i \equiv 1$  is in the relative interior of the primal constraints, Slater's condition holds.

$$\begin{aligned} & \inf_{\tilde{L}_i \geq 0} \left\{ \mathbb{E}[\tilde{L}_i(Y(1)) \tilde{L}_i(Y(1))' | A=1] : \mathbb{E}[\tilde{L}_i(Y(1)) | A=1] = 1, \quad \tilde{L}_i(y_1)/\tilde{L}_i(y_1') \leq \Gamma \quad y_1, y_1' - \text{a.s.} \right\} \\ &= \sup_M \inf_{\tilde{L}_i \geq 0} \left\{ \mathbb{E}[(Y(1)-M) \tilde{L}_i(Y(1))' | A=1] + M : \tilde{L}_i(y_1)/\tilde{L}_i(y_1') \leq \Gamma \quad y_1, y_1' - \text{a.s.} \right\} \end{aligned}$$

By inspection, the inner infimum is attained at

$$\tilde{L}_i^*(y_1) = \begin{cases} c \Gamma & \text{if } Y(1)-M \leq 0 \quad \text{for some } c \geq 0 \\ c & \text{o/w} \end{cases}$$

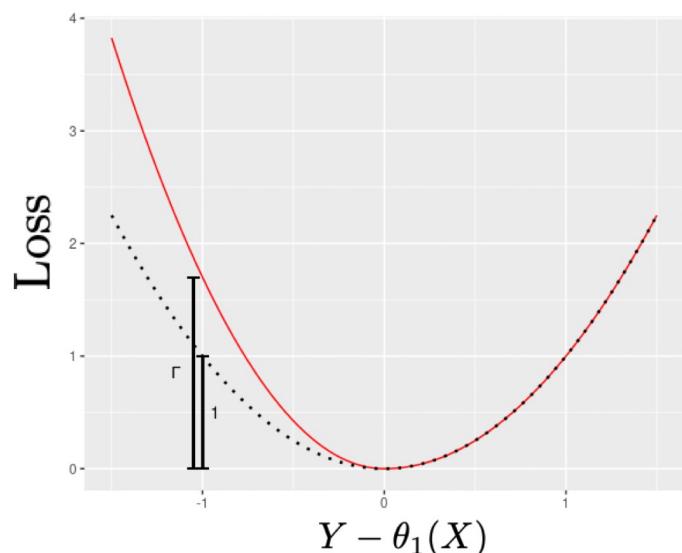
Plugging this in,

$$\begin{aligned} &= \sup_M \inf_{c \geq 0} \{ c \mathbb{E}[\Psi_M(Y(1)) | A=1] + M \} \\ &= \sup_M \{ M : \mathbb{E}[\Psi_M(Y(1)) | A=1] \geq 0 \}. \quad \square. \end{aligned}$$

Or can be learned by solving a loss minimization problem.

Let  $lr(\theta, y) := \frac{1}{2} ((y-\theta)_+^2 + \Gamma(y-\theta)_-^2)$ .

$$\min_{\theta(x)} \mathbb{E}[lr(\theta(x), Y(1)) | A=1]$$



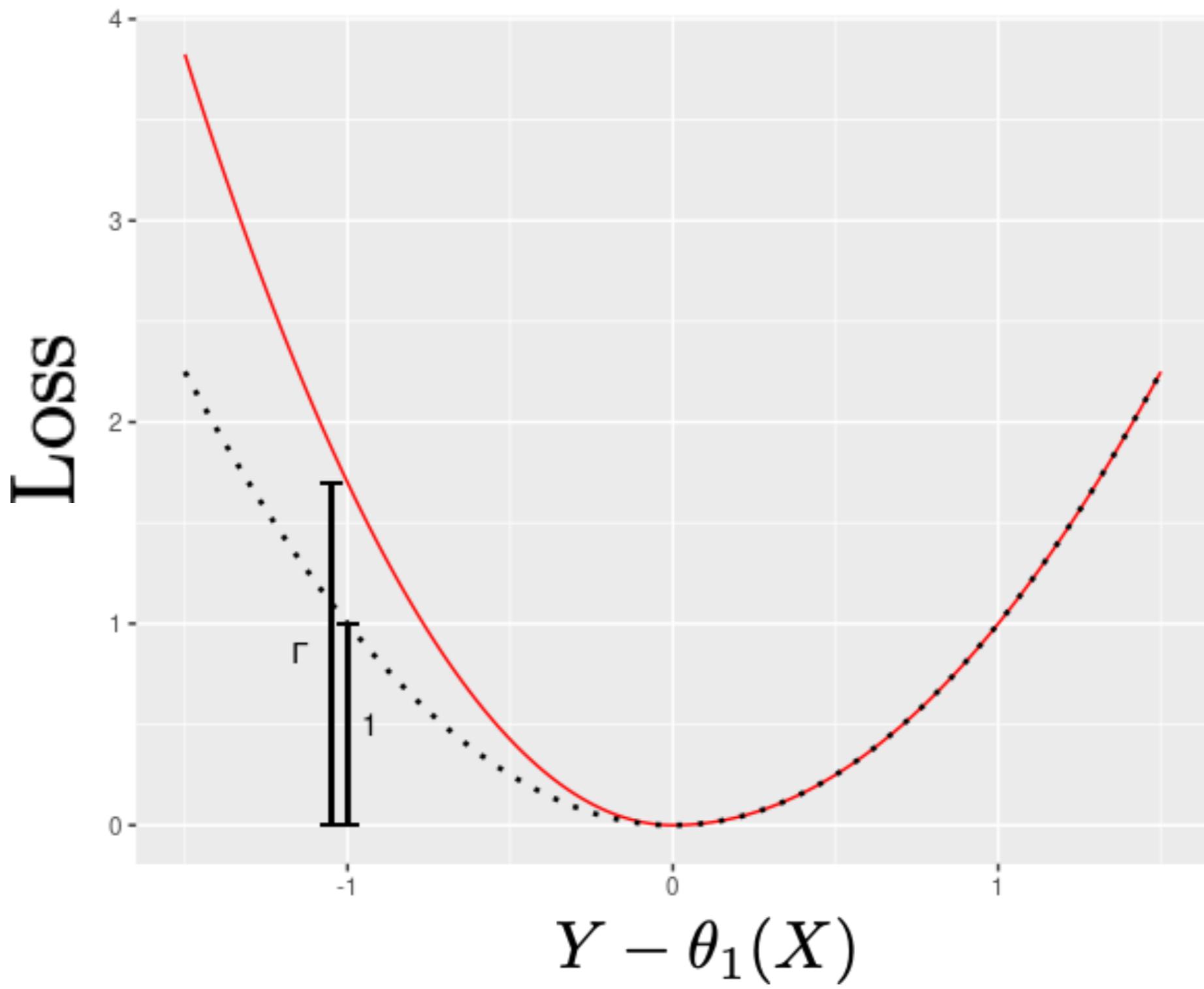
Lemma If  $\mathbb{E}[\ell_r(\theta_i^*(x); \gamma_{(1)}) | A=1] < \infty$ , then  $\theta_i^*(\cdot)$  is the (unique) minimizer of

$$\min_{\theta_i(\cdot): \text{measurable}} \mathbb{E}[\ell_r(\theta_i(x); \gamma_{(1)}) | A=1]$$

Pf The optimal solution  $\bar{\theta}_i(\cdot)$  satisfies the first order optimality condition

$$\begin{aligned} & \mathbb{E}[-(\gamma_{(1)} - \bar{\theta}_i(x))_+ - \Gamma(\gamma_{(1)} - \bar{\theta}_i(x))_- | x, A=1] \\ &= -\mathbb{E}[\psi_{\bar{\theta}_i(x)}(\gamma_{(1)}) | x, A=1] = 0. \end{aligned}$$

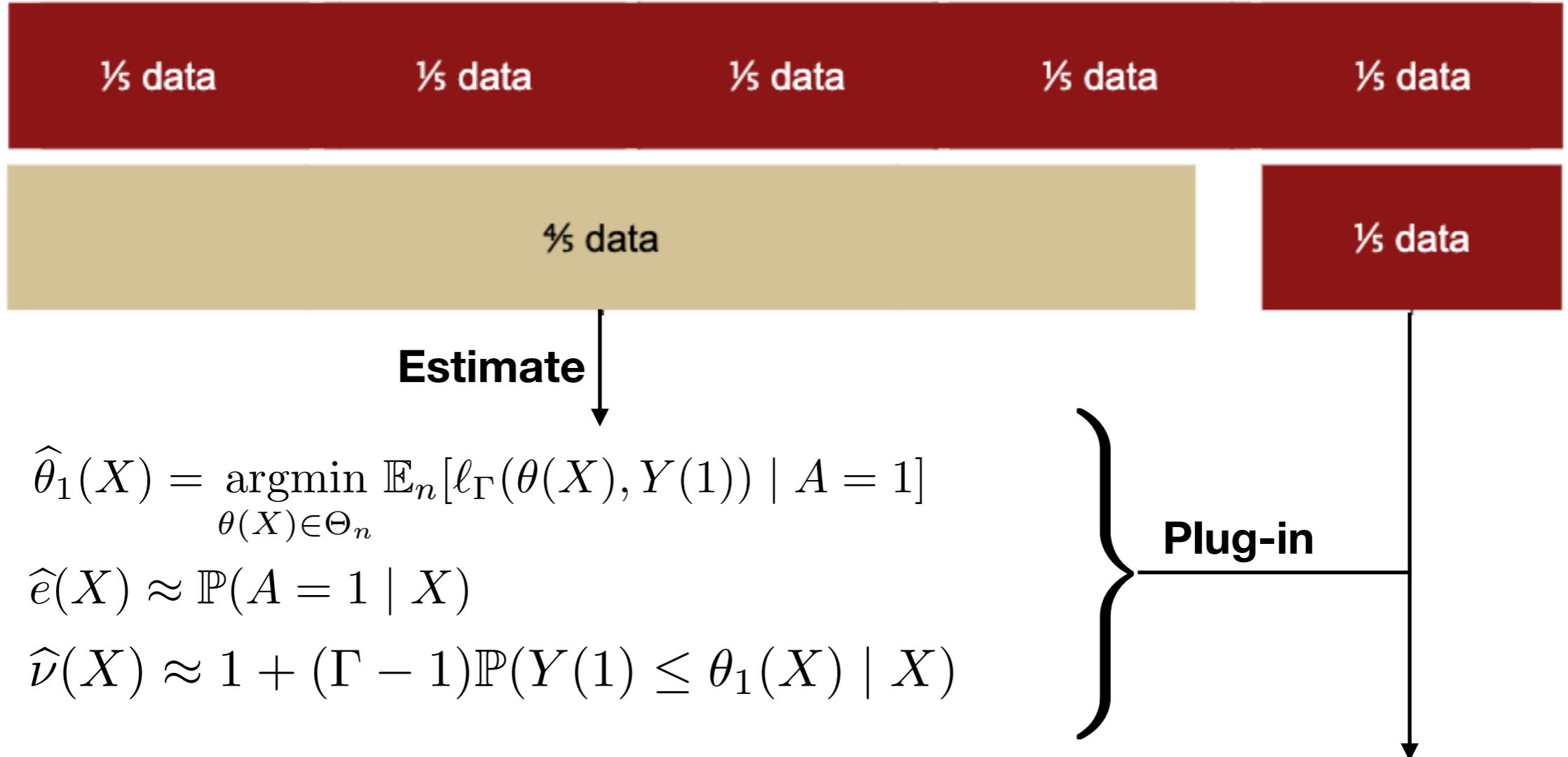
Since  $M \mapsto \gamma_M(y)$  is strictly monotone, this is obtained by the unique  $\sup_M \{M: \mathbb{E}[\gamma_M(\gamma_{(1)}) | A=1, x=x] \geq 0\}$ . \(\square\)



# Estimate lower bound on ATE

Estimate  $\mu_1^- = \mathbb{E}[AY(1) + (1 - A)\theta_1(X)] \leq \mathbb{E}[Y(1)]$

Cross-fitting  
[Chernozhukov '18]



$$\hat{\mu}_1^- = \frac{1}{n} \sum_{i=1}^n A_i Y_i + (1 - A_i) \hat{\theta}_1(X_i) + \frac{A_i}{\hat{e}(X_i)} \frac{(Y_i - \hat{\theta}_1(X_i))_+ + \Gamma(Y_i - \hat{\theta}_1(X_i))_-}{\hat{\nu}_1(X_i)}$$

Reduces to AIPW when  $\Gamma = 1$

# Asymptotics

Assume nuisance variables can be estimated reasonably well

$$\left\| \widehat{\theta}_1(\cdot) - \theta_1(\cdot) \right\|_{2,P} = o_p(n^{-1/4}), \quad \left\| \widehat{e}(\cdot) - P(A = 1 \mid X = \cdot) \right\|_{2,P} = o_p(n^{-1/4})$$

$$\left\| \widehat{\nu}(\cdot) - 1 - (\Gamma - 1)\mathbb{P}(Y(1) \leq \theta_1(\cdot) \mid X = \cdot) \right\|_{2,P} = o_p(n^{-1/4})$$

$\widehat{\mu}^-$ : cross-fitting estimator for  $\mu_1^- = \mathbb{E}[AY(1) + (1 - A)\theta_1(X)] \leq \mathbb{E}[Y(1)]$

**Theorem** Under regularity conditions,

$$\frac{\sqrt{n}}{\widehat{\sigma}_n} (\widehat{\mu}^- - \mu^-) \xrightarrow{d} N(0, 1)$$

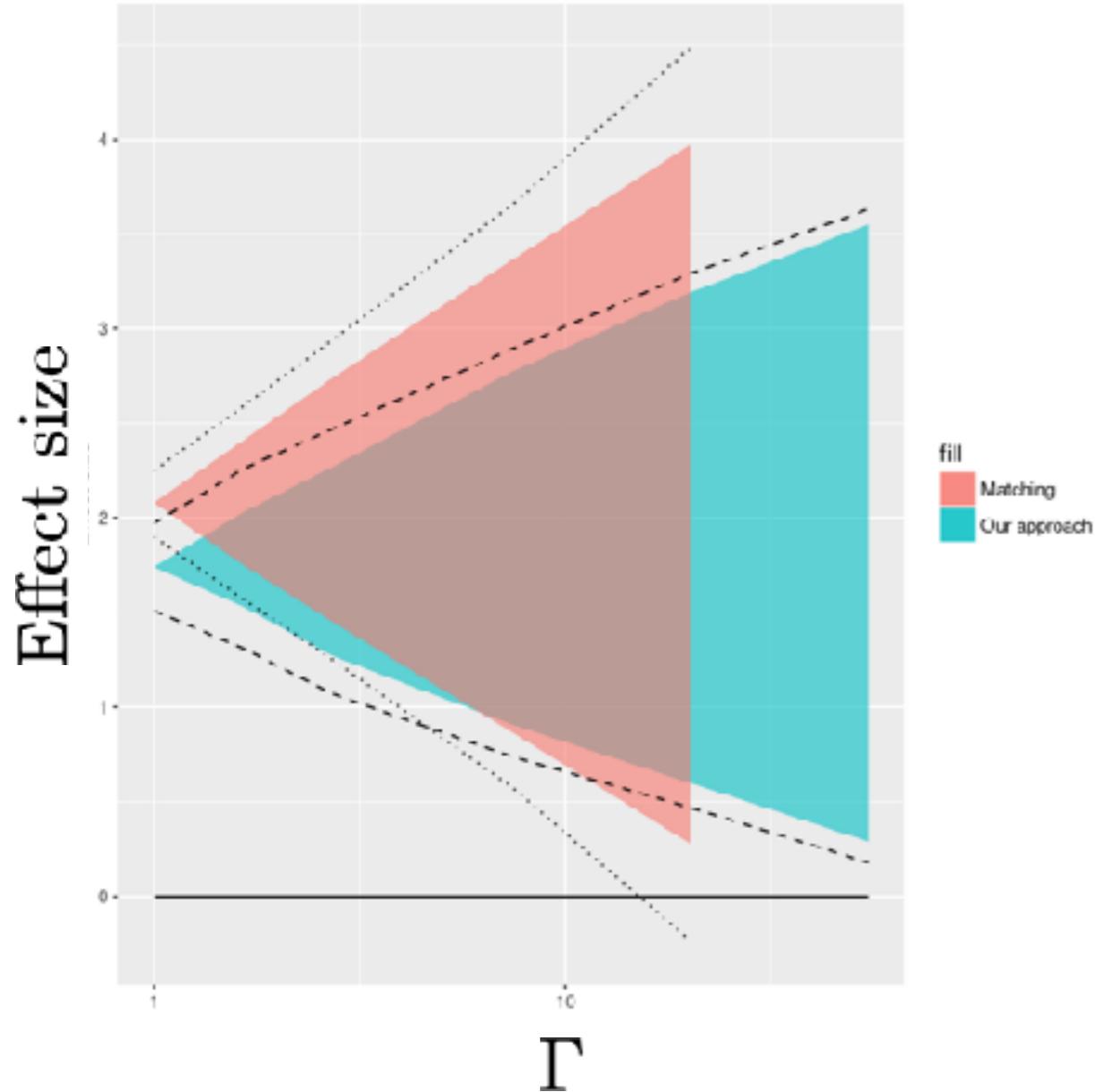
for some specified (known)  $\widehat{\sigma}_n^-$ .

**Combining, we can develop a central limit theorem for the bound on ATE**

# Example: fish consumption

- Study analyzing the impact of fish consumption on total blood mercury concentration
- N = 2,512 adult participants in 2013-14 NHANES survey in US
- Treatment is high fish consumption, >12 servings of fish or shellfish in the previous month
- Control is low fish consumption, 0 or 1 servings of fish
- Outcome as  $\log_2$  of total blood mercury concentration (ug/L)
- Covariates: gender, age, income, missing income, race, education, ever smoked, and number of cigarettes smoked last month)

# Example: fish consumption



- Filled areas are estimated bounds
- Dashed lines represent 95% confidence intervals around filled area
- Differences in centers due to statistical bias
- Tighter CIs under this approach consistent with theoretical results