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 $e^\star(X) := \mathbb{P}(A = 1 \mid X)$

- Assume that there exists $\epsilon > 0$ such that $\epsilon \leq e^\star(X) \leq 1 - \epsilon$ almost surely

- This means I have at least ϵn 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 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,
$$\mu_a^\star(X) := \mathbb{E}[Y(a) \mid X] = \mathbb{E}[Y(a) \mid X, A = a]$$
$$= \mathbb{E}[Y \mid X, A = a]$$ ← **observable**

- Fit $\mu_a^\star(X)$ via the loss minimization problem
$$\text{minimize}_{\mu_a \in \mathfrak{M}_a} \quad \mathbb{E}[(Y - \mu_a(X))^2 \mid A = a]$$

- ATE estimator $\hat{\tau}_{\mathrm{DM}} := \frac{1}{n} \sum_{i=1}^{n} \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)$

- 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 $e^{\star}(X) := \mathbb{P}(A = 1 \mid X)$
  - 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

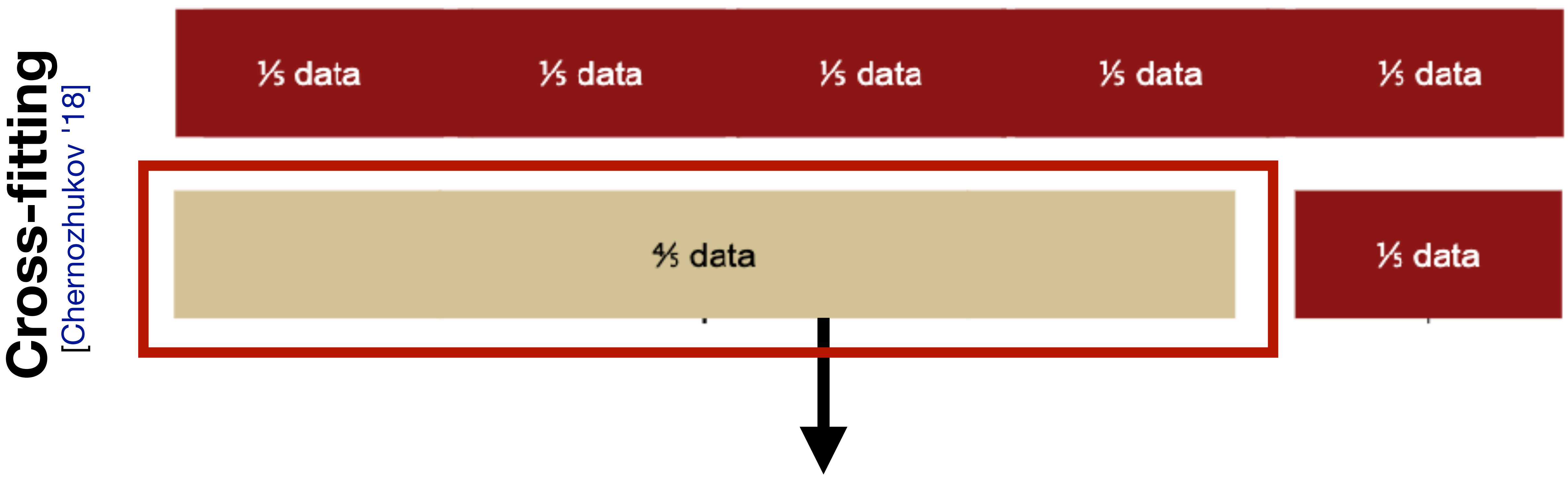
$$\hat{\tau}_{\mathrm{AIPW}} := \frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i))$$
$$+\frac{1}{n}\sum_{i=1}^{n}\left(\frac{A_i}{\hat{e}(X_i)}(Y_i - \hat{\mu}_1(X_i)) - \frac{1 - A_i}{1 - \hat{e}(X_i)}(Y_i - \hat{\mu}_0(X_i))\right)$$

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

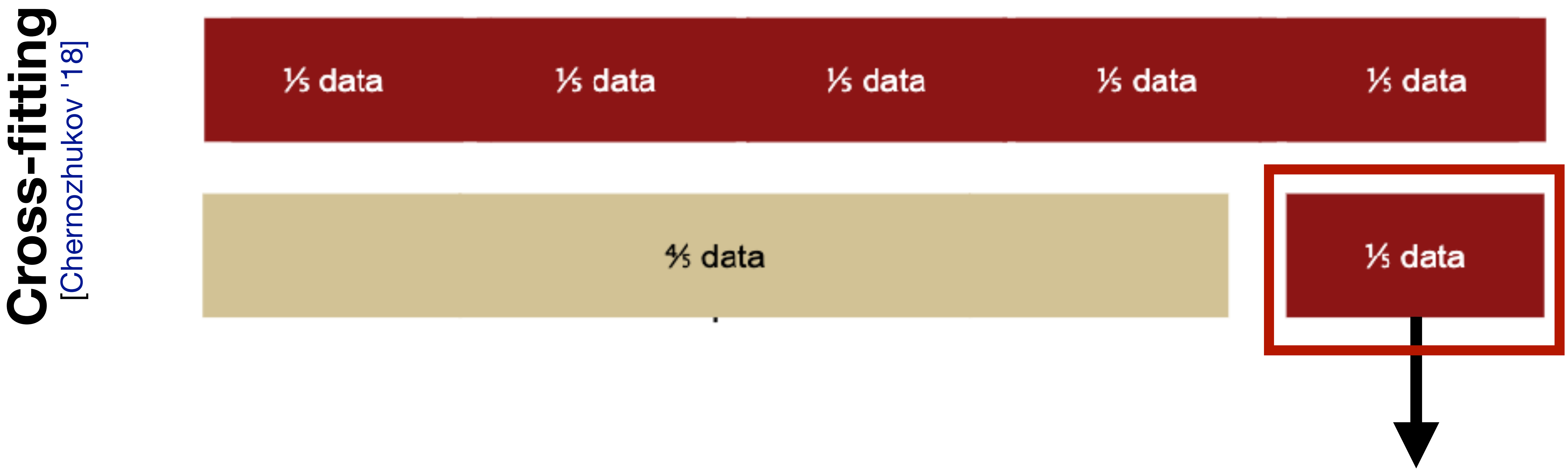


$$\widehat{\mu}_a(X) \approx \mathbb{E}[Y(a) \mid X = x], \; a \in \{0, 1\}$$

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

- Estimate nuisance parameters on the auxiliary sample

Cross-fitting

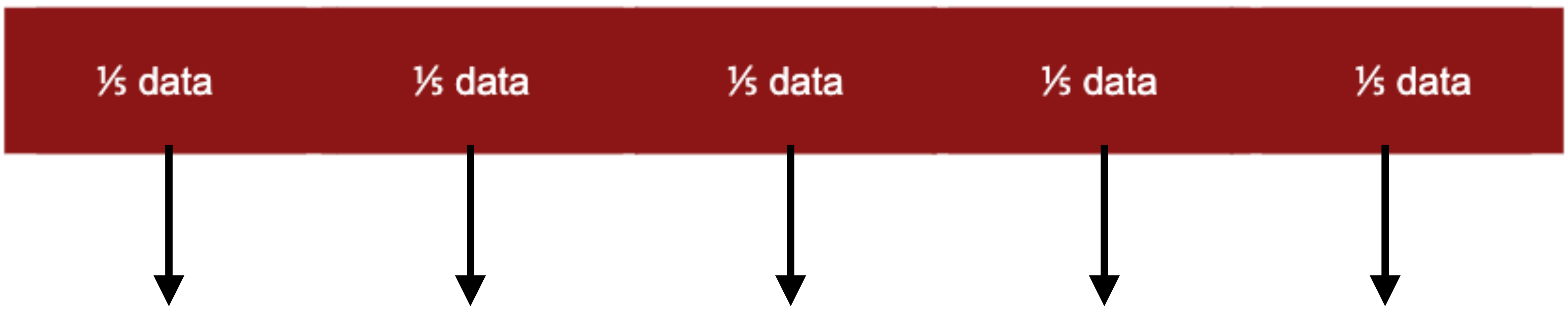


$$\hat{\tau}_1 := \frac{1}{n} \sum_{i=1}^{n} \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + \frac{A_i}{\hat{e}(X_i)}(Y - \mu_1(X_i)) - \frac{1 - A_i}{1 - \hat{e}(X_i)}(Y - \mu_0(X_i))$$

- Estimate ATE by plugging in nuisance estimates

Cross-fitting

Cross-fitting
[Chernozhukov '18]



$$\widehat{\tau} = \frac{1}{5}\left(\hat{\tau}_1 + \hat{\tau}_2 + \hat{\tau}_3 + \hat{\tau}_4 + \hat{\tau}_5 \right)$$

- 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 $p := \mathbb{P}(A = 1)$

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

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

Assessing overlap

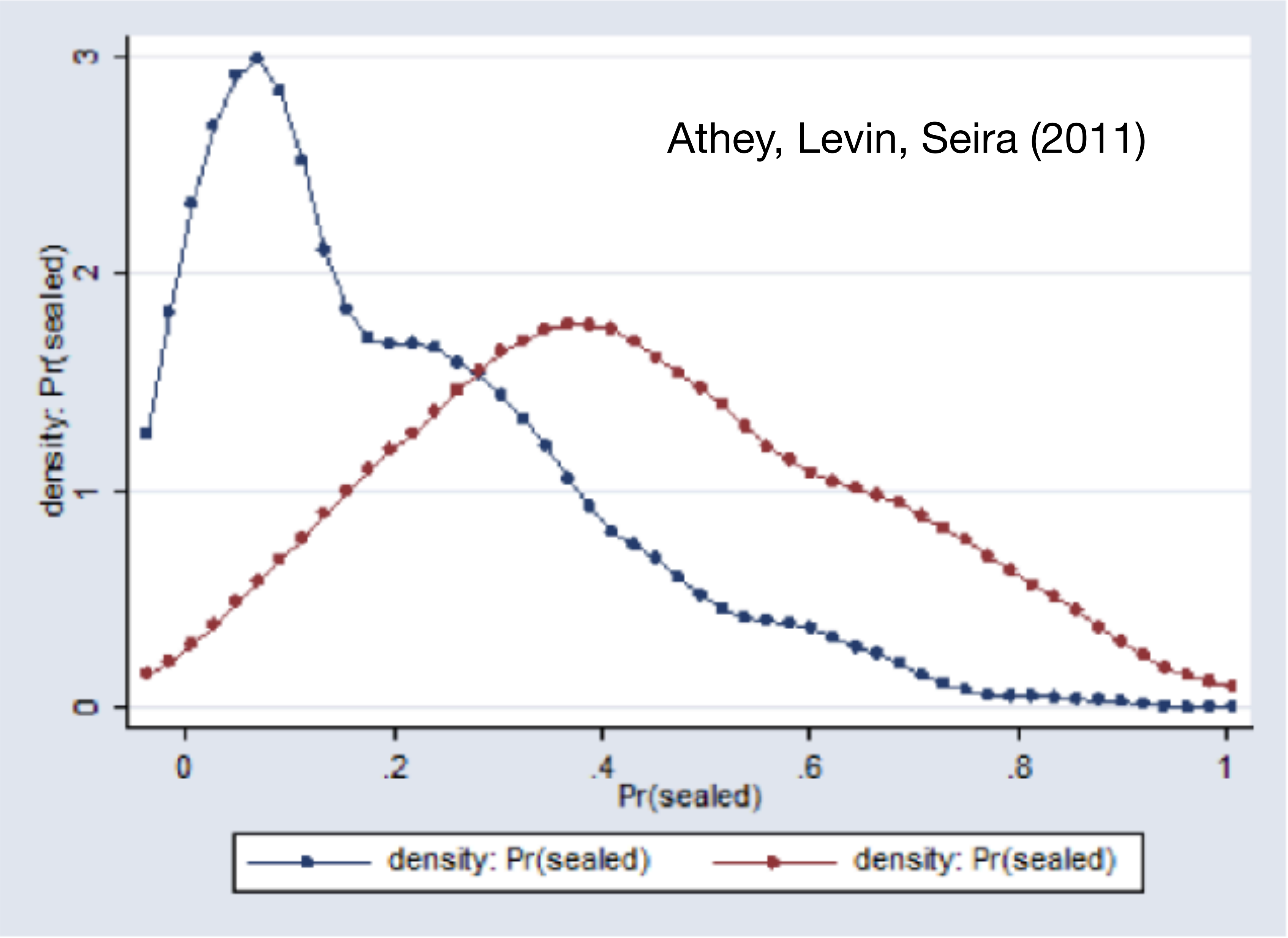
- A common visualization is to look at the pdf of the propensity score across treatment groups
- Plot approximates pdfs of the distribution $\mathbb{P}(e^\star(X) \in \cdot \mid A = a)$
- For each $q \in (0,1)$, plot fraction of observations in the treatment group with $e^\star(x) = q$ (and similarly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

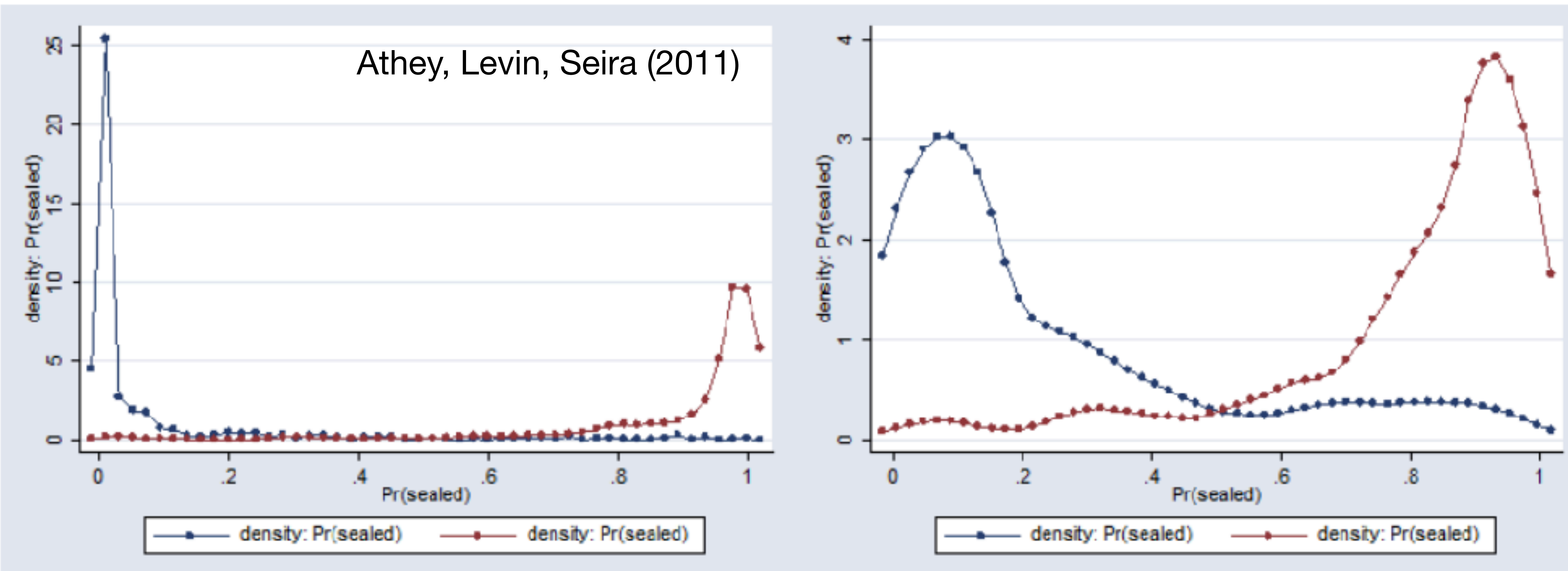
Very few observations with extreme propensity scores



Slide by Susan Athey and Stefan Wager

California

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



Slide by Susan Athey and Stefan Wager

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)**

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

- 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,
$$\mu^\star(a, x) := \mathbb{E}[Y(a) \mid X = x] = \mathbb{E}[Y(a) \mid X = x, A = a]$$
$$= \mathbb{E}[Y \mid X = x, A = a]$$

- Fit $\mu^\star(a, x)$ 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,
  $$\mu_a^\star(X) := \mathbb{E}[Y(a) \mid X] = \mathbb{E}[Y(a) \mid X, A = a]$$
  $$= \mathbb{E}[Y \mid X, A = a]$$

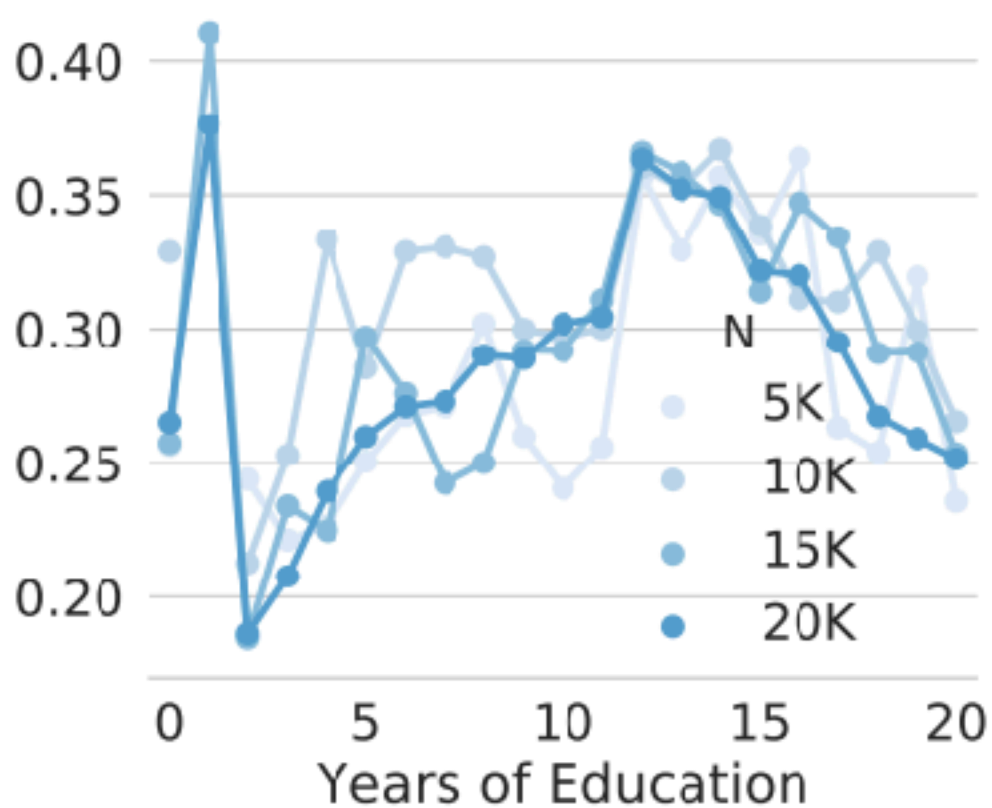
- Fit $\mu_a^\star(X)$ via the loss minimization problem
  $$\text{minimize}_{\mu_a \in \mathfrak{M}_a} \quad \mathbb{E}[(Y - \mu_a(X))^2 \mid A = a]$$

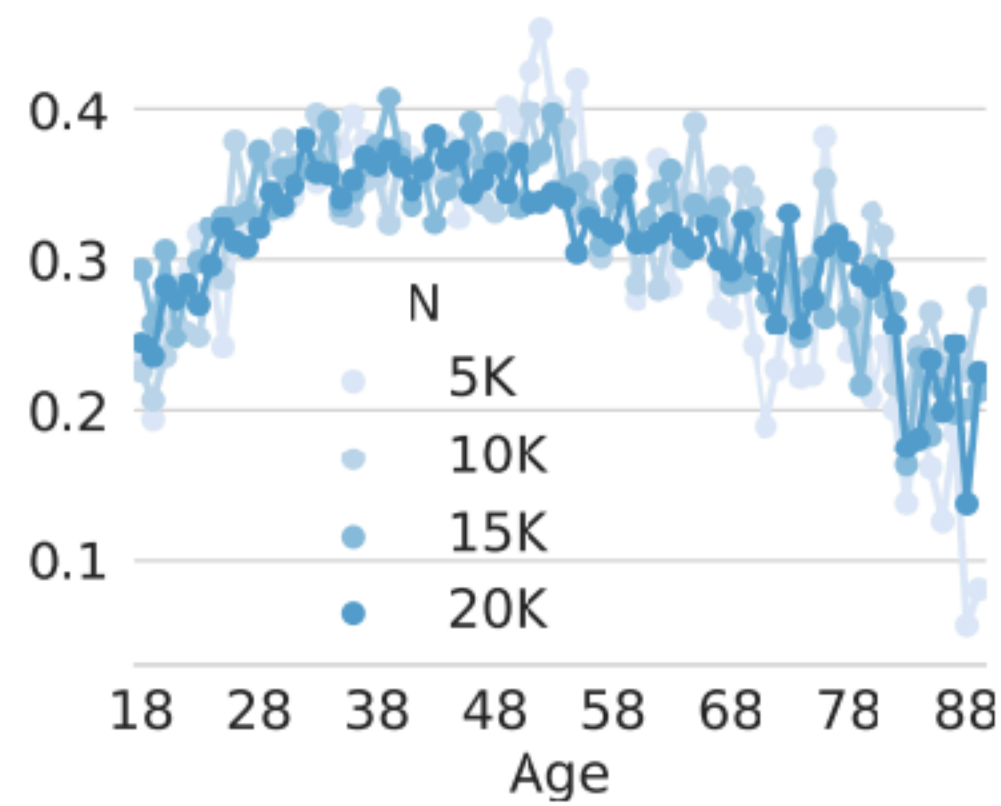
- $\hat{\tau}(X) := \hat{\mu}_1(X) - \hat{\mu}_0(X)$

- 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

$$Y_i - \hat{\mu}_0(X_i) \text{ if } A_i = 1, \hat{\mu}_1(X_i) - Y_i \text{ if } A_i = 0$$

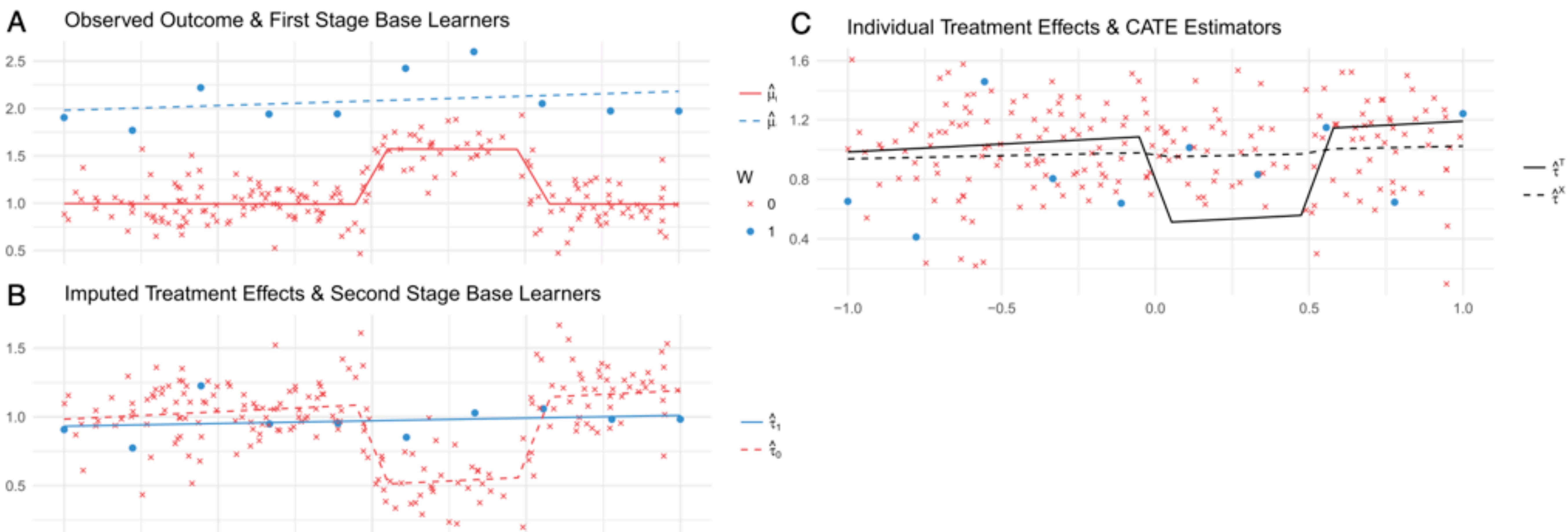
- Fit another set of models $\hat{\tau}_1, \hat{\tau}_0$ on the two category of imputed values, take

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

X-Learner

Kunzel et al. (2018)

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



R-Learner

Nie and Wager (2020)

R-Learner

Nie and Wager (2020)

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 assigments

- 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

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

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 Γ

⇕

Bounded influence of U in a nonparametric logistic regression model

Equivalence

FAQs

Bounded unobserved confounding

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

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

- How do I choose Γ?

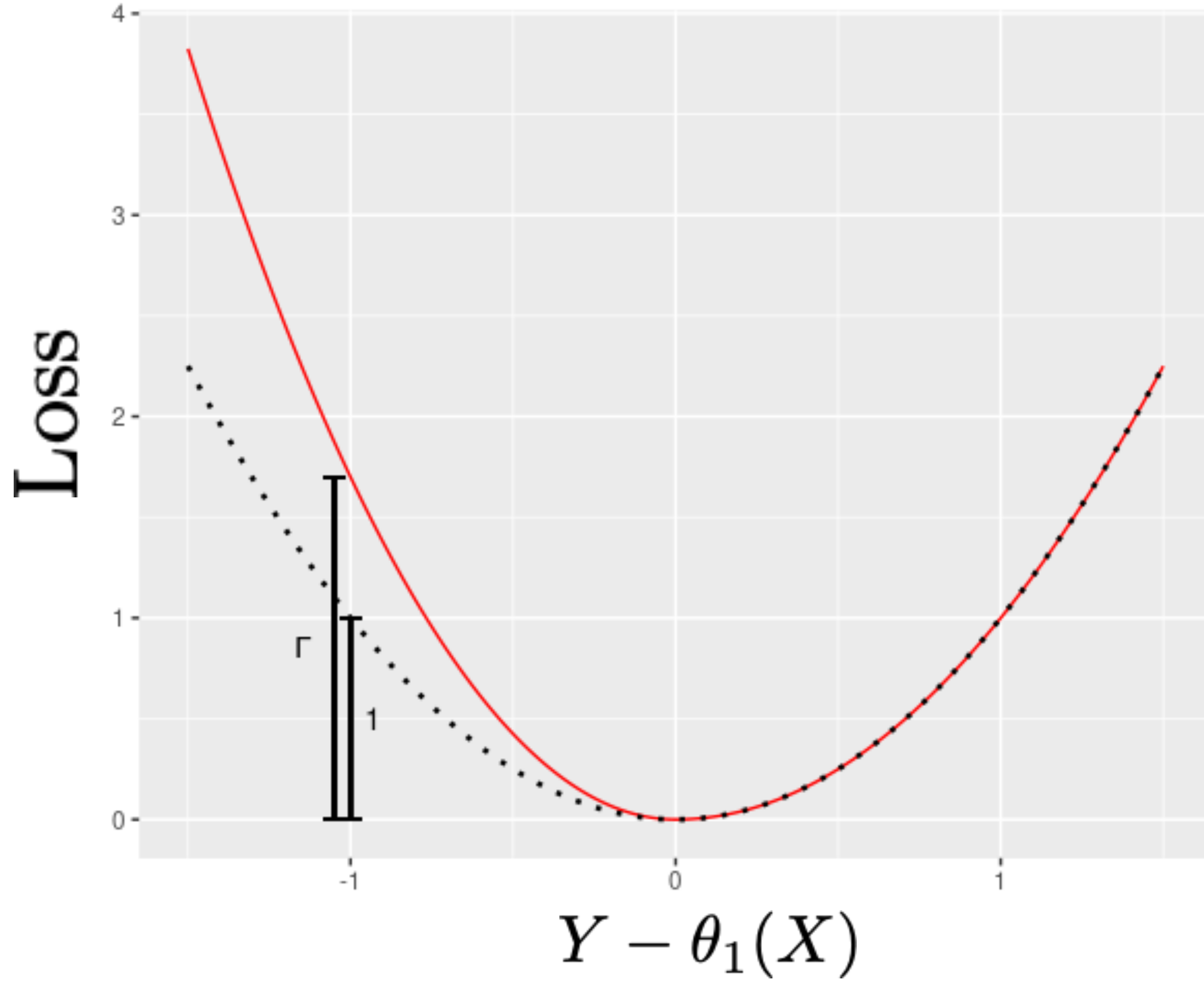
➡ Domain expertise (e.g. clinical intuition)

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

➡ Sensitivity of a study: at what level of Γ is the conclusion of the study invalidated?

- Is this the only natural confounding model?

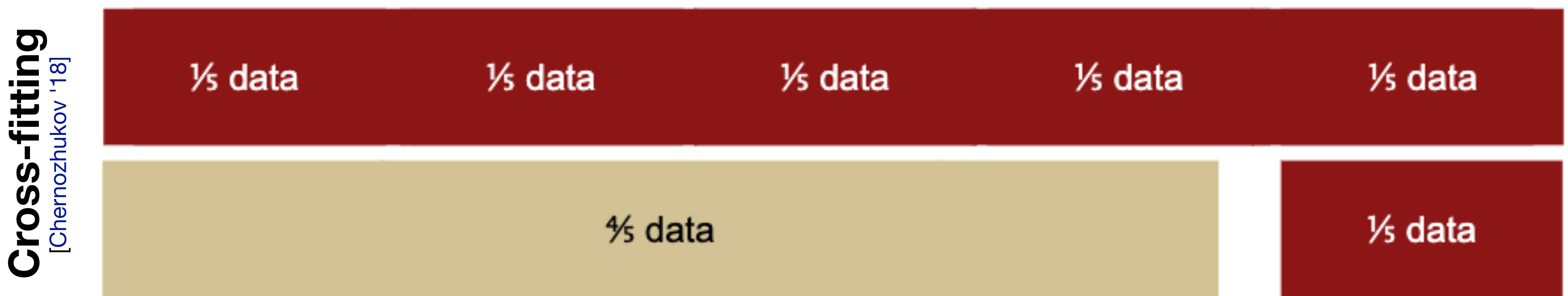
➡ No. Today we discuss a modern semiparametric framework under this model; the framework may be developed under different models.



Loss
4
3
2
1
0
Γ
1
-1
0
1
$Y - \theta_1(X)$

Estimate lower bound on ATE

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



Estimate

$$\widehat{\theta}_1(X) = \underset{\theta(X)\in\Theta_n}{\operatorname{argmin}}\ \mathbb{E}_n[\ell_\Gamma(\theta(X), Y(1)) \mid A = 1]$$

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

$$\widehat{\nu}(X) \approx 1 + (\Gamma - 1)\mathbb{P}(Y(1) \leq \theta_1(X) \mid X)$$

Plug-in

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

Reduces to AIPW when $\Gamma = 1$

Asymptotics

Assume nuisance variables can be estimated reasonably well

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

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

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

Theorem Under regularity conditions,

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

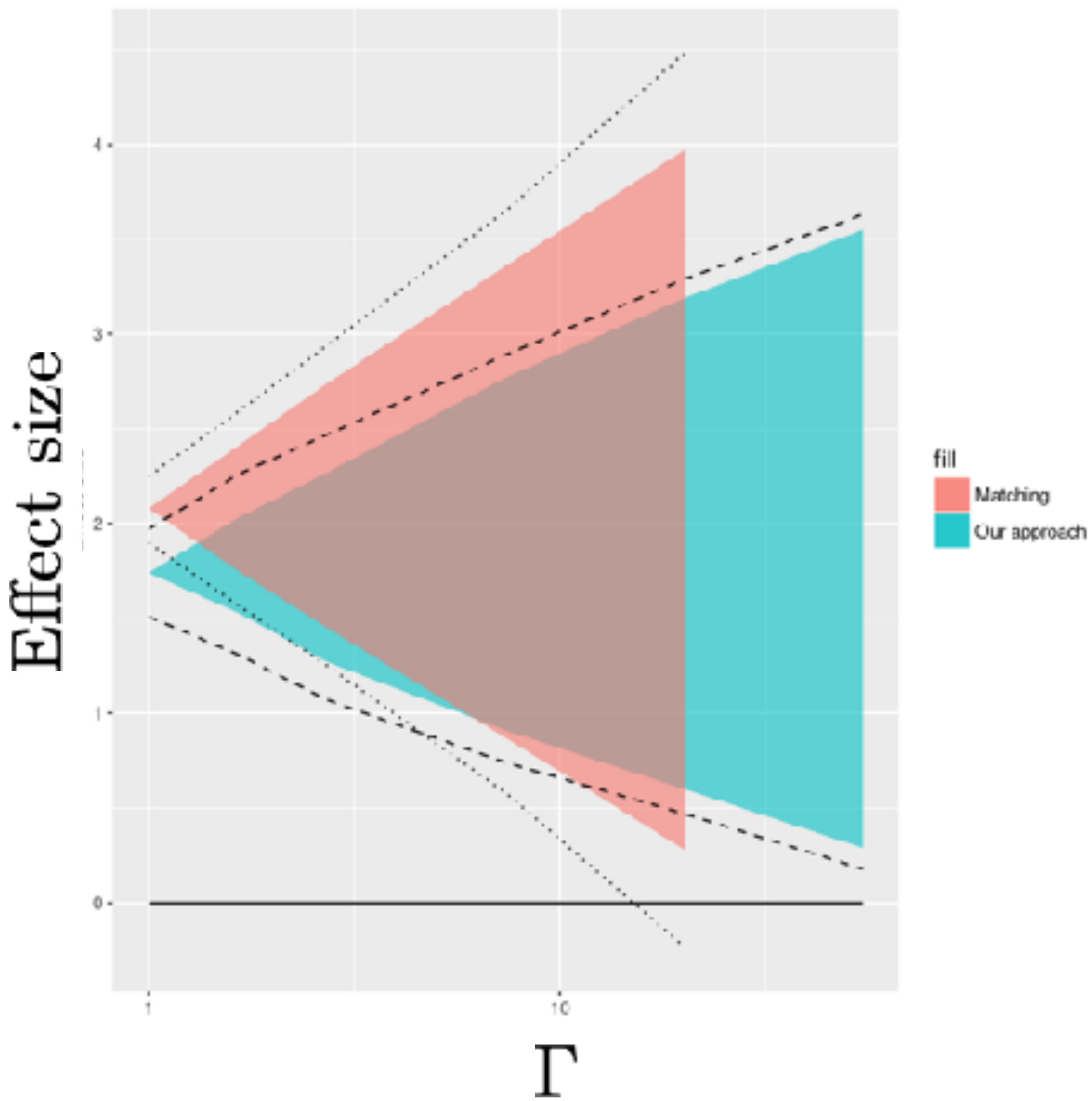
for some specified (known) $\widehat{\sigma}_n^-$.

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