Causality

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

https://scholar.harvard.edu/imbens/files/efficient_estimation_of_average_treatment_effects_using_the_estimated_propensity_score.pdf

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

Prediction and causality

- A central goal of ML is to predict an outcome given variables describing a situation
  - Given patient characteristics, will their outcome improve?

- Most decision-making problems revolve around a decision / intervention / treatment
  - What would happen if we changed the system?
  - Given patient characteristics, will their outcome improve if **they follow a new diet**?

- We want to develop a scientific understanding of a decision
  - If you predict housing demand based on price, then a prediction model will say high price means high demand

Prediction and causality

- Causal inference is a multi-disciplinary field built across economics, epidemiology, and statistics

- Focus is on questions about **counterfactuals**
  - What structure of data do we need to answer this question?
  - How do we interpret the key estimands?

- ML models can predict outcomes; when can it predict counterfactuals?
  - How can we leverage flexible ML models to infer causality?

Binary actions

- Today we will focus on the setting with two actions
  - One action represents treatment (1), the other is control (0)

- This is still foundational
  - Key difficulties still persist here despite the simplicity
  - Core technical insights will translate to more general settings

- In complex problems, this is often the de facto standard
  - Control is status quo, treatment is a new elaborate program
  - Throughout economics, medicine, and tech, it requires a tremendous amount of domain knowledge and effort to come up with an alternative to the current system

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.

WIRED Technology | Science | Culture | Video | Reviews | Magazine

Liking curly fries on Facebook reveals your high IQ

What you Like on Facebook could reveal your race, age, IQ, sexuality and other personal data, even if you've set that information to "private".

By PHILIPPA WARR

Causality

- You came up with a new diet regimen that you believe will alleviate symptoms of rheumatism (e.g. chronic joint pain)
- To test it, you recruit people to try the diet
- You find that
  - Small fraction on the diet experience chronic pain
  - Large fraction not on the diet (aka all rheumatism patients outside your volunteer pool) experience chronic pain
  - Awesome! Everyone should try this diet
- But after years of adoption, you realize the diet does not affect chronic pain

Causality

- What could have gone wrong?
  - Volunteers to the diet may have been people with healthy predispositions, and affluent socioeconomic backgrounds

- **Fundamental problem:** we don’t observe counterfactuals

- How do we model this?

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

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

Don't observe Y(1-A)

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

ATE

First goal: Estimate average treatment effect

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

- We only observe Y := Y(A)
- What could go wrong?
  - Volunteers to the diet (A = 1) may have been people with healthy predispositions, and affluent socioeconomic backgrounds

Person	A	Y(0)	Y(1)	Y(1) - Y(0)
1	1	0	0	0
2	1	0	0	0
3	1	0	0	0
4	1	0	0	0
5	0	1	1	0
6	0	1	1	0
7	0	1	1	0
8	0	1	1	0

Randomized control trials

also called A/B testing, (randomized) experiments

- First try: let's **randomize** treatment assignments

$$Y(1), Y(0) \perp A$$

$Y := Y(A)$

- By virtue of randomized assignments, we have

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

$$= \mathbb{E}[Y \mid A = 1] - \mathbb{E}[Y \mid A = 0]$$ ← **observable**

- We can estimate final line from i.i.d. data (Y_i, A_i)

Randomized control trials

(Y_i, A_i) iid

$$\hat{\tau} = \frac{1}{n_1}\sum_{A_i=1} Y_i - \frac{1}{n_0}\sum_{A_i=0} Y_i \quad \text{where} \quad n_1 = |\{i : A_i = 1\}|, \quad n_0 = |\{i : A_i = 0\}|$$

$$= \frac{\sum_1^n A_i Y_i}{\sum_1^n A_i} - \frac{\sum_1^n (1-A_i) Y_i}{\sum_1^n (1-A_i)}$$

Unbiased estimators of $\mathbb{E}[Y \mid A=1]$, $\mathbb{E}[Y \mid A=0]$

Randomized control trials

By the CLT, $\sqrt{n}(\hat{\tau} - \tau) \Rightarrow N(0, \sigma^2)$

where $\sigma^2 = Var(Y|A=1) + Var(Y|A=0) + 2cov(\cdots)$

$= Var(Y(1)) + Var(Y(0)) + //$

So for any estimator $\hat{\sigma}_n^2 \xrightarrow{p} \sigma^2$,

$$P\left(\hat{\tau} - \frac{\hat{\sigma}_n}{\sqrt{n}} 1.96 \leq \tau \leq \hat{\tau} + \frac{\hat{\sigma}_n}{\sqrt{n}} 1.96\right) \to 0.95$$

cf. Construct $\hat{\sigma}_n^2$ by sample variance of $Y|A=a$'s.

RCT with covariates

- If by randomness more treatments get assigned to young patients with a better prognosis, then we will exaggerate the treatment effect
  - Problem goes away in large samples, but matters for small samples

- If you have access to covariates X, and can estimate $\mathbb{E}[Y \mid X, A]$ accurately, then we can improve this

- Using any regression model, we can estimate $\mathbb{E}[Y \mid X, A = 1], \mathbb{E}[Y \mid X, A = 0]$ ← **observable**
  - Random forests, boosted decision trees, kernels, NNs etc

Estimator

Define $\mu_a^*(x) := \mathbb{E}[Y(a) \mid X]$, for $a \in \{0,1\}$

By the tower law, $\tau = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[\mu_1^*(x) - \mu_0^*(x)]$.

Note that $\mathbb{E}[Y \mid X, A=a] = \mathbb{E}[Y(a) \mid X, A=a] = \mathbb{E}[Y(a) \mid X] = \mu_a^*(x)$

$\mathbb{E}[Y \mid X, A=a]$: Observable

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

So if $\hat{\mu}_a$ is an estimator of μ_a^* then

$\hat{\tau} = \frac{1}{n} \sum \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)$ is an estimator of τ

Fitting outcome models

How do we learn μ_a^*? Loss minimization

$$\min_{\theta \in \Theta} \mathbb{E}\left[(Y - \mu_a(X;\theta))^2 \mid A = a \right]$$

If $\mu_a^*(\cdot) = \mu_a(\cdot;\theta^*)$ $\exists \theta^* \in \Theta$, then solving above via SGD finds a good estimator of μ_a^*.

CLT for covariate adjustments

If $\hat{\mu}_a = \mu_a^*$, then $\hat{\tau}$ is a good estimator

For simplicity, if $\hat{\mu} = \mu^*$,

$$\mathbb{E}\hat{\tau} = \mathbb{E}\left[\frac{1}{n}\sum_i \mu_1^*(X_i) - \mu_0^*(X_i)\right] = \tau$$

By the CLT, $\frac{\sqrt{n}}{s_n}(\hat{\tau} - \tau) \Rightarrow N(0,1)$

where s_n^2 is the sample variance of $(\mu_1^* - \mu_0^*)(X)$.

Beyond RCTs

- What if clean randomization is not possible?
- Randomization sometimes affected by the site
  - Oxford / AstraZeneca trial made a dosage mistake at a location
  - Turned out to be more effective
- Ignoring variables that affect treatment assignment leads to biases

Beyond RCTs

- Run large-scale experiment, randomized for each sex

	Men		Women	
	No disease ($Y=1$)	Disease ($Y=0$)	No disease ($Y=1$)	Disease ($Y=0$)
Treatment ($A=1$)	0.1500	0.2250	0.1000	0.0250
Control ($A=0$)	0.0375	0.0875	0.2625	0.1125

(Here the numbers are the fractions of individuals in each category.)

- $\mathbb{P}(Y=1 \mid A=1)=0.5$ vs $\mathbb{P}(Y=1 \mid A=0)=0.6$
  - So maybe treatment is not effective?

Slide by Ramesh Johari

Simpson's paradox

- But if you compute treatment effect for each sexes,

$$\mathbb{E}[Y(1) - Y(0) \mid X = m] = \mathbb{E}[Y(1) - Y(0) \mid X = w] = 0.1$$

- So ATE = 0.1. What happened?

- Women are more likely to be in control than treatment; men are more likely to be in treatment than control. And women have higher potential outcomes on average than men.

Simpson's paradox

- Issue here is that

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

- If you ignore sex as a confounding variable, you create a **omitted variable bias** in estimating the ATE

Slide by Ramesh Johari

Berkeley admissions

- Berkeley was sued for gender bias in admissions based on 1973 numbers: 44% of men were admitted but only 35% of women

- But individual department's admissions record showed no evidence of such gender-based discrimination

- Turns out women systematically applied to more competitive majors

Observational studies

- Randomization is sometimes infeasible or prohibitively expensive
  - e.g. post-market drug surveillance, effect of air pollution on long-term health outcomes

- Experimentation can be risky in high-stakes scenarios
  - operational scenarios: new inventory policy for Amazon, new pricing algorithm for Uber

- May want to use existing large-scale data collected under some data-generating policy (e.g. legacy system)

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

No unobserved confounding

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

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

↑ No unobserved conf.

$$= \mathbb{E}[Y \mid X, A=1] - \mathbb{E}[Y \mid X, A=0] \quad \text{since } Y := Y(A)$$

observable

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

As before estimate $\mu_a^*(x) := \mathbb{E}[Y(a) | X]$

By no unobs. confounding,

$$\mathbb{E}[Y(a) | X] = \mathbb{E}[Y | X, A=a]$$

so we can use any black-box ML model to fit

So split data:

On first chunk, fit $\hat{\mu}_a(x)$ $a=0,1$

On second chunk, use $\hat{\mu}_a$ to estimate τ

Direct method

$$\hat{\tau}_{DM,1} = \frac{1}{n_2} \sum_{\text{second chunk}} \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)$$

$$\hat{\tau}_{DM,2} = \frac{1}{n_2} \sum_{\text{second chunk}} \left\{ A_i (Y_i - \hat{\mu}_0(X_i)) + (1 - A_i)(\hat{\mu}_1(X_i) - Y_i) \right\}$$

Works well if μ_a is easy to estimate

Inverse propensity weighting

- What if the outcome models are very complex and difficult to estimate?

- A natural approach is to reweight samples, to change the distribution $\mathbb{E}[\,\cdot \mid A=1,X]$ to $\mathbb{E}[\,\cdot \mid X]$
  - Essentially importance sampling

$e^*(x) := P(A=1 \mid X)$: prop score

$$\hat{\tau}_{IPW} = \frac{1}{n}\sum_i \frac{A_i Y_i}{e^*(X_i)} - \frac{(1-A_i)Y_i}{1-e^*(X_i)} \quad \text{if } e^* \text{ known.}$$

If e^* unknown, replace e^* with an estimate $\hat{e}$.

Unbiasedness

$e^*(x) := P(A=1 \mid X)$: prop score

$$\hat{\tau}_{IPW} = \frac{1}{n}\sum_i \frac{A_i Y_i}{e^*(X_i)} - \frac{(1-A_i)Y_i}{1-e^*(X_i)}$$

$$\mathbb{E}\left[\frac{AY}{e^*(X)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{AY}{e^*(X)} \mid X\right]\right] = \mathbb{E}\left[P(A=1\mid X)\cdot \mathbb{E}\left[\frac{AY}{e^*(X)} \mid X, A=1\right]\right]$$

$$= \mathbb{E}\left[\mathbb{E}[Y \mid X, A=1]\right] = \mathbb{E}\left[\mathbb{E}[Y(1) \mid X, A=1]\right], \quad Y := Y(A)$$

$$= \mathbb{E}\left[\mathbb{E}[Y(1) \mid X]\right] \quad \text{by no unobs. conf. } Y(1), Y(-) \perp\!\!\!\perp A \mid X$$

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

CLT for IPW

By CLT $\quad \sqrt{n}(\hat{\tau}_{IPW} - \tau) \Rightarrow N(0, \sigma^2_{IPW})$

$$\sigma^2_{IPW} = Var\left(\frac{AY}{e(X)} + \frac{(1-A)Y}{1-e(X)}\right)$$

$$= Var\left(\frac{AY}{e(X)}\right) + Var\left(\frac{(1-A)Y}{1-e(X)}\right) - 2\,\mathbb{E}[Y(1)]\,\mathbb{E}[Y(0)]$$

Take s_n^2 as sample variance of $\frac{AY}{e(X)} - \frac{(1-A)Y}{1-e(X)}$

then $\quad P\left(\tau \in \left[\hat{\tau}_{IPW} \pm \frac{s_n}{\sqrt{n}} 1.96\right]\right) \to 95\%.$

Estimating propensity score

But e^* is often unknown.

Logistic regression $\log \frac{e(x)}{1-e(x)} = f(x;\theta)$

e.g. $f(x;\theta) = x^\top \theta$

then solve

$$\min_{\theta \in \Theta} \mathbb{E} \log\left(1 + \exp\left((\underbrace{2A-1}_{\in \{-1,1\}}) \cdot f(x;\theta)\right)\right)$$

If model is well-specified, SGD on $\mathcal{P}$ gives $\hat{e} \approx e^*$

since this is just MLE.

Inverse propensity weighting

- 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

$$\hat{\tau}_{AIPW} := \frac{1}{n}\sum_i \left\{ \underbrace{\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)}_{DM} + \underbrace{\frac{A_i}{\hat{e}(X_i)}\left(Y_i - \hat{\mu}_1(X_i)\right) - \frac{1-A_i}{1-\hat{e}(X_i)}\left(Y_i - \hat{\mu}_0(X_i)\right)}_{\text{control variate}} \right\}$$

$$= \frac{1}{n}\sum_i \left\{ \underbrace{\frac{A_i Y_i}{\hat{e}(X_i)} - \frac{(1-A_i)Y_i}{(1-\hat{e}(X_i))}}_{IPW} + \underbrace{\left(1 - \frac{A_i}{\hat{e}(X_i)}\right)\hat{\mu}_1(X_i) - \left(1 - \frac{1-A_i}{1-\hat{e}(X_i)}\right)\hat{\mu}_0(X_i)}_{\text{control variate}} \right\}$$

Unbiasedness

(X_i, A_i, Y_i)

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

$$= \frac{1}{n}\sum_i^n \left\{ \underbrace{\frac{A_i Y_i}{\hat{e}^{-i}(X_i)} - \frac{(1-A_i)Y_i}{(1-\hat{e}^{-i}(X_i))}}_{IPW} + \underbrace{\left(1 - \frac{A_i}{\hat{e}^{-i}(X_i)}\right)\hat{\mu}_1^{-i}(X_i) - \left(1 - \frac{1-A_i}{1-\hat{e}^{-i}(X_i)}\right)\hat{\mu}_0^{-i}(X_i)}_{control\ variate} \right\}$$

$$\mathbb{E}[\hat{\mu}_1^{-i}(X) - \hat{\mu}_0^{-i}(X)] = 2$$

$$\mathbb{E}\left[\frac{A}{\hat{e}^{-i}(X)}(Y - \hat{\mu}_1^{-i}(X)) \mid X\right] = P(A=1 \mid X)\,\mathbb{E}\left[\frac{A}{\hat{e}^{-i}(X)}(Y - \hat{\mu}_1^{-i}(X)) \mid X, A=1\right]$$

$$= \mathbb{E}[Y - \hat{\mu}_1^{-i}(X) \mid X, A=1] = \mathbb{E}[Y(1) - \hat{\mu}_1^{-i}(X) \mid X, A=1]$$

$$= \mathbb{E}[Y(1) - \hat{\mu}_1^{-i}(X) \mid X] \quad \text{by no unobs. conf.} \quad = 0$$

CLT for AIPW

$$\frac{\sqrt{n}}{s_n}\left(\hat{\tau}_{AIPW} - \tau\right) \Rightarrow N(0,1)$$

where s_n^2 is the sample variance of

$$\frac{A_i Y_i}{e^a(X_i)} - \frac{(1-A_i)Y_i}{(1-e^a(X_i))} + \left(1 - \frac{A_i}{e^a(X_i)}\right)\mu_1^a(X_i) - \left(1 - \frac{1-A_i}{1-e^a(X_i)}\right)\mu_0^a(X_i)$$

Control variate

Another interpretation of AIPW is that the aug. term is a control variate.

$$\mathrm{Var}\left(\frac{AY}{e^*(X)} + \left(1 - \frac{A}{e^*(X)}\right)\mu_1^*(X) \right)$$

$$= \mathrm{Var}\left(\frac{AY}{e^*(X)}\right) + \underbrace{\mathrm{Var}\left(\left(1 - \frac{A}{e^*(X)}\right)\mu_1^*(X)\right)} + 2\underbrace{\mathrm{Cov}\left(\frac{AY}{e^*(X)}, \left(1 - \frac{A}{e^*(X)}\right)\mu_1^*(X)\right)}$$

$$= \mathrm{Var}\left(\frac{AY}{e^*(X)}\right) - \mathrm{Var}\left(\left(1 - \frac{A}{e^*(X)}\right)\mu_1^*(X)\right)$$

Control variate

Efficiency

- In fact, this is the best asymptotic variance we can get
- AIPW has optimal asymptotic variance, regardless of whether the propensity score is known or not
- Formalizing this requires a lot of work

Nuisance parameters

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

No unobs. conf.

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

- If a good parametric model exists, then can estimate at the usual $1/\sqrt{n}$ rates

- In general, these are infinite dimensional objects. Can be difficult to estimate.

Semiparametrics

- We only care about estimating the ATE
  - One-dimensional estimand, infinite dimensional nuisance parameters

- Estimation accuracy of nuisance parameters is good only insofar as it helps with estimating the ATE

- Due to its high-dimensional nature, often difficult to estimate nuisances at parametric rates

- Goal: semiparametric estimators that are insensitive to errors in nuisance estimates

Doubly robust

- One main advantage of AIPW is that even if one of the nuisance parameter models are **misspecified,** you can still get correct asymptotic behavior

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

Assume $\hat{e} \in [\varepsilon, 1-\varepsilon]$ a.s.

① If $\hat{e}$ is learned from a well-specified model class, then $\|\hat{e} - e^*\|_{P,2} \xrightarrow{p} 0$

Let's say outcome models are mis-specified so $\|\hat{\mu}_a - \mu_a\|_{P,2} \xrightarrow{p} 0$, $\mu_a \neq \mu_a^*$.

$\hat{\tau}_{AIPW} \xrightarrow{p} \tau$

② If $\hat{\mu}_a$ is well-specified, then $\|\hat{\mu}_a - \mu_a^*\|_{P,2} \xrightarrow{p} 0$

Let's say prop score is misspecified so $\|\hat{e} - e\|_{P,2} \xrightarrow{p} 0$, $e \neq e^*$.

$\hat{\tau}_{AIPW} \xrightarrow{p} \tau$.

Doubly robust

① $\frac{1}{n}\sum \left[\frac{A_i Y_i}{e(X_i)}\right] \xrightarrow{p} \mathbb{E}[Y(1)] = \mathbb{E}[\mathbb{E}[Y \mid A=1, X]]$

$-\frac{1}{n}\sum \frac{(1-A_i)Y_i}{1-e(X_i)} \xrightarrow{p} \mathbb{E}[Y(0)]$

$\frac{1}{n}\sum_1^n \left(1-\frac{A_i}{e(X_i)}\right)\hat{\mu}_1(X_i) \xrightarrow{p} \mathbb{E}\left[\left(1-\frac{A}{e(X)}\right)\mu_1(X)\right] = 0$

$-\frac{1}{n}\sum_1^n \left(1-\frac{1-A_i}{1-e(X_i)}\right)\hat{\mu}_0(X_i) \xrightarrow{p} 0$

② $\frac{1}{n}\sum_1^n \hat{\mu}_1(X_i) \to \mathbb{E}Y(1)$, $\frac{1}{n}\sum \hat{\mu}_0(X_i) \to \mathbb{E}Y(0)$

$\frac{1}{n}\sum \frac{A_i}{e(X_i)}(Y_i - \hat{\mu}_1(X_i)) \xrightarrow{p} \mathbb{E}\left[\frac{A}{e(X)}(Y-\mu_1(X))\right] = 0$

$\frac{1}{n}\sum \frac{1-A_i}{1-e(X_i)}(Y_i - \hat{\mu}_0(X_i)) \xrightarrow{p} 0$

Doubly robust

Orthogonality

- When is a semiparametric estimator insensitive to errors in nuisance estimates?

- Directional derivative of functional wrt nuisance parameters at true value is near-zero

- Ensures that a little perturbation in nuisance parameters near the truth values does not affect functional

Orthogonality

Def

Let $\eta = (M_0, M_1, e)$ be the tuple of nuisance parameters.

A statistical function $\mathbb{E}_P \psi(D; \eta)$ is Neyman orthogonal over Λ if

$$\frac{d}{dr} \mathbb{E}_P \psi(D; \eta^* + r(\eta - \eta^*)) \Big|_{r=0} = 0 \qquad \forall \eta \in \Lambda$$

↳ Directional deriv at true nuisance param η^* ⌟

$D = (X, Y, A) \qquad \eta = (M_0, M_1, e)$

$$\psi_{AIPW}(D; \eta) = M_1(X) - M_0(X) + \frac{A}{e(X)}(Y - M_1(X)) - \frac{1-A}{1-e(X)}(Y - M_0(X))$$

$$\frac{d}{dr} \mathbb{E}_{D \sim P} \psi_{AIPW}(D; \eta^* + r(\eta - \eta^*)) \Big|_{r=0} = 0 \qquad \text{WTS}$$

Orthogonality of AIPW

Assume $\frac{d}{dr}$ & $\mathbb{E}_P$ are interchangeable throughout

$$\frac{d}{dr}\mathbb{E}_P\left(\mu_1^r + r(\mu_1 - \mu_1^r)\right)(x)\Big|_{r=0} = \mathbb{E}\left[(\mu_1 - \mu_1^r)(x)\right]$$ First

$$\frac{d}{dr}\mathbb{E}_P\left[\frac{A}{(e^r + r(e-e^r))(x)}\left(Y - (\mu_1^r + r(\mu_1 - \mu_1^r))(x)\right)\right]\Big|_{r=0}$$ Second

$$= -\mathbb{E}\left[\frac{A(e-e^r)}{(e^r + r(e-e^r))^2}\left(Y - \mu_1^r(x)\right)\right]\Big|_{r=0} - \mathbb{E}\left[\frac{A}{e^r(x)}(\mu_1 - \mu_1^r)(x)\right]$$

$$= -\mathbb{E}\left[\frac{A[e-e^r](x)}{e^r(x)^2}\left(Y - \mu_1^r(x)\right)\right] - \mathbb{E}\left[\frac{A}{e^r(x)}(\mu_1 - \mu_1^r)(x)\right]$$

$$= -\mathbb{E}\left[e^r(x)\cdot\mathbb{E}\left[\frac{A(e-e^r)(x)}{e^r(x)^2}\left(Y - \mu_1^r(x)\right)\Big| X, A=1\right]\right]$$

$$- \mathbb{E}\left[(\mu_1 - \mu_1^r)(x)\right]$$

$$= -\mathbb{E}\left[\frac{(e-e^r)(x)}{e^r(x)}\,\mathbb{E}\left[\left(Y(1) - \mu_1^r(x)\right)\Big| X\right]\right] - \mathbb{E}\left[(\mu_1 - \mu_1^r)(x)\right]$$

$$= -\mathbb{E}\left[(\mu_1 - \mu_1^r)(x)\right]$$

Orthogonality of AIPW

Why orthogonality?

- Allows getting central limit rates on ATE estimation even when we can only estimate nuisance parameters at slower rates

- In addition to no unobserved confounding, $e^\star(X), \hat{e}(X) \in [\epsilon, 1-\epsilon]$, we assume the following rate condition

$$\left\{ \|\hat{e} - e^\star\|_{P,2}(\|\hat{\mu}_1 - \mu_1^\star\|_{P,2} + \|\hat{\mu}_0 - \mu_0^\star\|_{P,2}) = o_p(n^{-1/2}) \right\}$$

- This allows us to trade-off errors between nuisance parameters. Only their product needs to go down at this rate!

Central limit result

- CLT for the semiparametric AIPW, even when nuisance estimates converge at slower-than-parametric rates

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\psi_{\mathrm{AIPW}}(X_i,Y_i,A_i;\hat{\mu}_0,\hat{\mu}_1,\hat{e})-\tau\right)\Rightarrow N(0,\sigma^2_{\mathrm{AIPW}})$$

where $\sigma^2_{\mathrm{AIPW}} := \mathrm{Var}\left(\psi_{\mathrm{AIPW}}\left(X,Y,A;\mu_0^\star,\mu_1^\star,e^\star\right)\right)$

- This is the oracle asymptotic variance; when the true nuisance parameters are known

- AIPW achieves optimal asymptotic efficiency

Sketch of asymptotics

$D=(X,T,A)$ $\quad \eta=(\mu_1,\mu_0,e)$

$$\sqrt{n}\left(\frac{1}{n}\sum \psi(O_i;\hat{\eta}) - \mathbb{E}\psi(O;\eta^*)\right) \Rightarrow N(0,\sigma^2_{AIPW})$$

$$\underbrace{\frac{1}{n}\sum \psi(O_i;\hat{\eta}) - \mathbb{E}_{Dn}\psi(O;\hat{\eta})}_{①} + \underbrace{\mathbb{E}_{Dn}\psi(O;\hat{\eta}) - \mathbb{E}\psi(O;\eta^*)}_{②}$$

$\sqrt{n}\cdot ① \Rightarrow N(0,\sigma^2_{AIPW})$ $\qquad$ $\sqrt{n}\cdot ② \xrightarrow{p} 0.$

Define $R(r) := \mathbb{E}_{Dn}\psi(O;\eta^* + r(\hat{\eta}-\eta^*))$

② $= R(1)-R(0)$ $\qquad$ If $r \mapsto R(r)$ is conti diff,

$= R'(r)\cdot(1-0)$ $\quad$ for some $r\in[0,1]$

So now we want to argue that $\sqrt{n}\sup_{r\in[0,1]} R'(r) \xrightarrow{p} 0$

Sketch of asymptotics

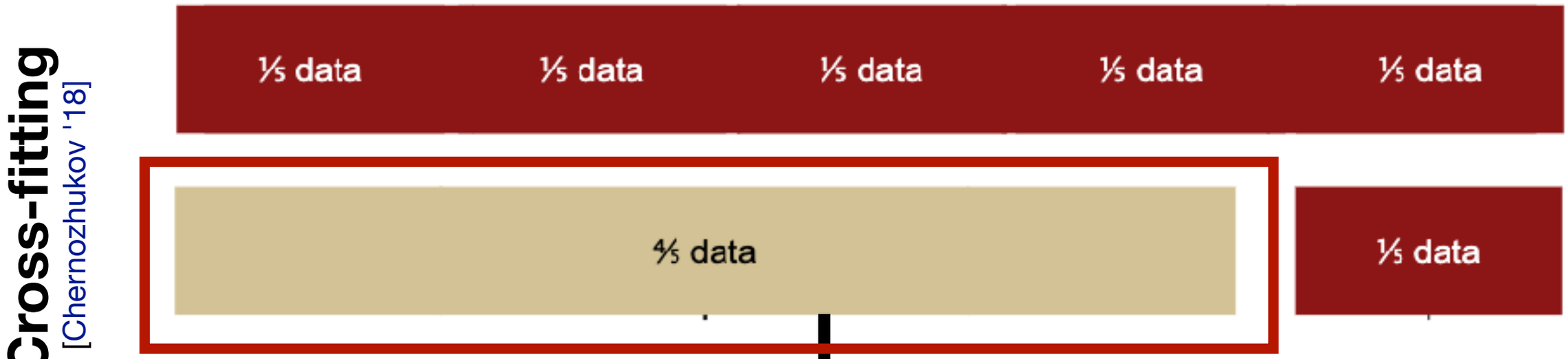
By orthogonality, $\mathcal{R}'(0) = 0$.

If $\mathcal{R}'$ is smooth enough, then

$\sqrt{n} \sup_{r \in [0,1]} \mathcal{R}'(r) \xrightarrow{p} 0$ as long as rate condition.

Cross-fitting

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



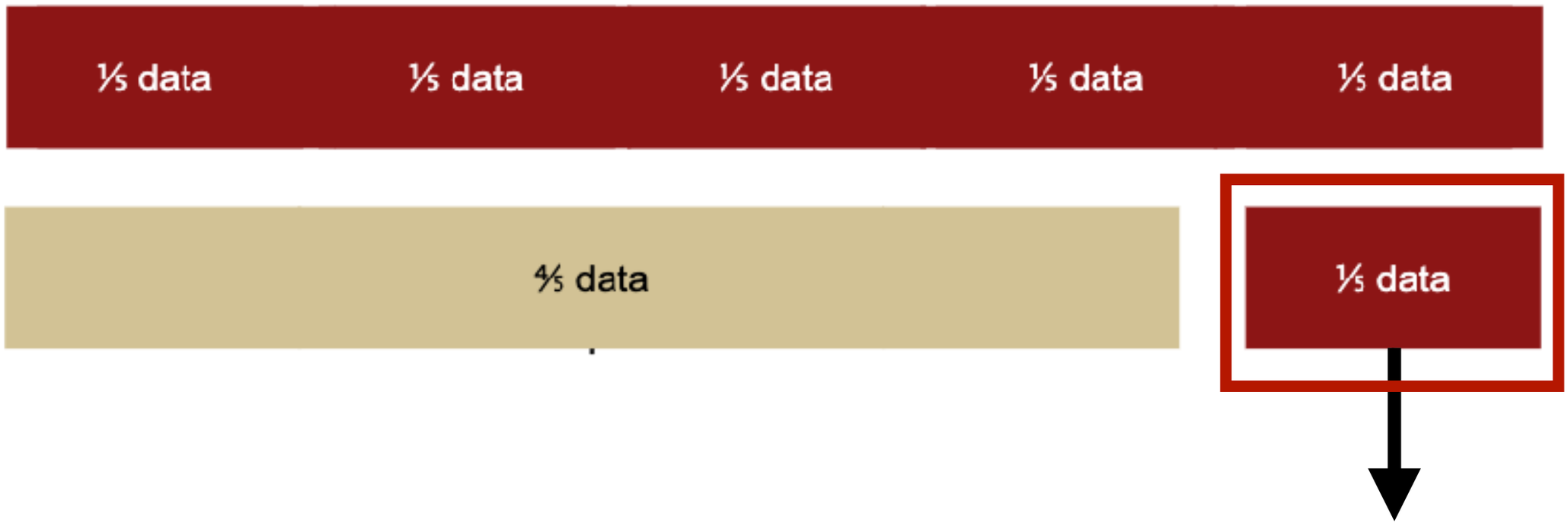
$$\widehat{\mu}_a(X) \approx \mathbb{E}[Y(a) \mid X = x], \ a \in \{0, 1\}$$

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

- Estimate nuisance parameters on the auxiliary sample

Cross-fitting

Cross-fitting
[Chernozhukov '18]

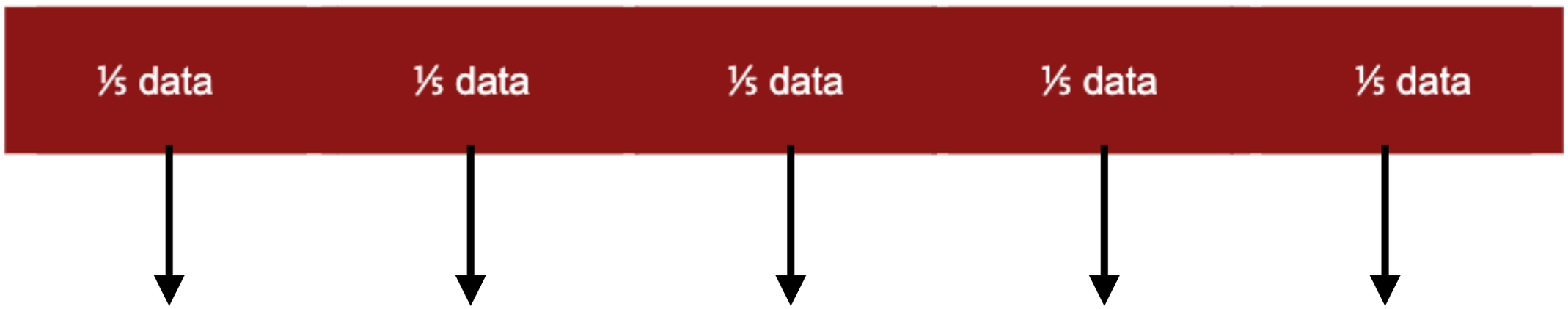


$$\hat{\tau}_1 := \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + \frac{A_i}{\hat{e}(X_i)}(Y - \mu_1(X_i)) - \frac{1 - A_i}{1 - \hat{e}(X_i)}(Y - \mu_0(X_i))$$

- Estimate ATE by plugging in nuisance estimates

Cross-fitting

Cross-fitting [Chernozhukov '18]



$$\hat{\tau} = \frac{1}{5}\Big(\quad \hat{\tau}_1 \; + \; \hat{\tau}_2 \; + \; \hat{\tau}_3 \; + \; \hat{\tau}_4 \; + \; \hat{\tau}_5 \quad \Big)$$

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