Chapter 1

Basics of Causal Inference

Zhan Gao, February 11, 2026

Adapted from Zhentao Shi's lecture notes and heavily dependent on Wager (2025).

1.1 Potential Outcome Framework

This a thought experiment. An individual i has two potential outcomes $Y_i(1)$ and $Y_i(0)$, and the inividual treatment effect is the difference $\Delta_i = Y_i(1) - Y_i(0)$. However, no one can step into the same river twice. One and only one of the potential outcomes will be realized, and therefore Δ_i is unobservable. Given the treatment status $D_i \in \{0, 1\}$, in reality what we can be observed is the realized outcome

$$Y_i = Y_i(D_i) = D_i Y_i(1) + (1 - D_i) Y_i(0). \tag{1.1}$$

Stable Unit Treatment Values Assumption. (1.1) is often refered in the literature as the *Stable Unit Treatment Values Assumption* (SUTVA) assumption, that assumes that the treatment on one individual does not interfere the outcome of others. SUTVA may be questionable in the presence of interference through social interactions, for example see Leung (2022). In this note, we maintain the SUTVA assumption, and for now suppres subscript i for simplicity.

We want to use the observable Y to learn the *average treatment effect* (ATE)

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

Randomized controlled trials. If the treatment is randomly assigned by flipping a coin

(the coin does not need to be even), then

$$\begin{pmatrix} Y(1) \\ Y(0) \end{pmatrix} \perp D. \tag{1.2}$$

This is the simplest framework for identifying causal effects, as well as the most reliable level of proof in terms of *interval validity*. Abadie and Cattaneo (2018) refer to RCTs as the "gold standard".

Characterization of ATE. (1.2) implies

$$\mathbb{E}\left[Y|D=1\right] = \mathbb{E}\left[DY(1) + (1-D)Y(0)|D=1\right] = \mathbb{E}\left[Y(1)|D=1\right] = \mathbb{E}\left[Y(1)\right].$$

where the last equality holds by randomized treatment assignment. Since (Y, D) are observable, the LHS is operational. Independence between $Y(1)$ and D ensures that the conditional expectation $\mathbb{E}\left[Y(1)|D=1\right]$ equals the unconditional expectation $\mathbb{E}[Y(1)]$. Similarly,

$$\mathbb{E}\left[Y|D=0\right] = \mathbb{E}\left[Y(0)|D=0\right] = \mathbb{E}[Y(0)].$$

Under RCT, we have an operational formula for ATE:

$$\text{ATE} := \tau = \mathbb{E}\left[Y|D=1\right] - \mathbb{E}\left[Y|D=0\right].$$

Given the data, we mimic the population average to compute

$$\widehat{\text{ATE}} := \widehat{\tau} = \frac{1}{n_1} \sum_{\{i:D_i=1\}} Y_i - \frac{1}{n_0} \sum_{\{i:D_i=0\}} Y_i$$

where $n_1 = \sum_{i=1}^n \mathbf{1}\{D_i = 1\}$ and $n_0 = \sum_{i=1}^n \mathbf{1}\{D_i = 0\}$. This is often refered as the *difference-in-means* estimator.

The above approach uses conditioning, which is intuitive. There is an alternative, yet equivalent way to ATE, using only unconditional quantities. Notice that

$$DY = D^2Y(1) + D(1-D)Y(0) = DY(1)$$

since for either $D \in \{0, 1\}$, we have $D^2 = D$ and $D(1-D) = 0$. Again, Y depends on D but $Y(1)$ is independent of D, and therefore

$$\mathbb{E}\left[DY\right] = \mathbb{E}\left[DY(1)\right] = \mathbb{E}\left[D\right]\mathbb{E}[Y(1)] = \Pr\left[D=1\right]\mathbb{E}[Y(1)]$$

and therefore

$$\mathbb{E}[Y(1)] = \frac{\mathbb{E}[DY]}{\Pr[D=1]} = \frac{\mathbb{E}[\mathbb{I}(D=1)Y]}{\Pr[D=1]}$$

if $\Pr[D=1] \neq 0$. The denominator $\Pr[D=1]$ is called the *propensity score*: the probability that a person is assigned into the treatment group.

Similarly, if $\Pr[D=0] \neq 0$ we have

$$\mathbb{E}[Y(0)] = \frac{\mathbb{E}[(1-D)Y]}{\Pr[D=0]} = \frac{\mathbb{E}[\mathbb{I}(D=0)Y]}{\Pr[D=0]}.$$

Therefore, if $\Pr[D=1] \in (0,1)$ the ATE is

$$\text{ATE} = \frac{\mathbb{E}[DY]}{\Pr[D=1]} - \frac{\mathbb{E}[(1-D)Y]}{\Pr[D=0]}. \tag{1.3}$$

ATE is the difference of the two ratios.

Given data, we can compute $\mathbb{E}[DY] / \Pr[D=1]$ as

$$\frac{(\sum_{i=1}^n D_i Y_i)/n}{n_1/n} = \frac{\sum_{i=1}^n D_i Y_i}{n_1} = \frac{1}{n_1} \sum_{\{i:D_i=1\}} Y_i.$$

It is easy to see that, the sample version is the same either we compute via the conditioning or the propensity score. Their equivalence is analogous to the fact that the conditional density can be written as the ratio of the joint density and the marginal density.

Theorem 1 (Wager (2025) Theorem 1.2)**.** *Suppose $(Y_i(1), Y_i(0))$ are i.i.d. draws from some super-population $\mathbb{P}$. Treatments are assigned based on a Bernoulli trial:*

$$D_i \overset{\text{i.i.d.}}{\sim} \text{Bern}(\pi), \quad 0 < \pi < 1.$$

Then

$$\sqrt{n}(\widehat{\tau} - \tau) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma_1^2}{\pi} + \frac{\sigma_0^2}{1-\pi}\right), \tag{1.4}$$

where $\sigma_d^2 \coloneqq \text{Var}(Y_i(d))$ for $d \in \{0,1\}$.

1.1.1 Stratification

At a granular level, the researcher also observe some confounding factor (causal inference term) (alternatively, in plain stat term it is called covariate) X.

The conditional ATE (CATE)

$$\tau(x) \coloneqq \text{ATE}(x) = \mathbb{E}[\Delta | X = x] = \mathbb{E}[Y(1)|x] - \mathbb{E}[Y(0)|x]$$

can vary across different realizations of X. For example, an vaccine is more effective to children than adults. In RCT, if the random treatment assignment is regardless of the age group, we have $(Y(1), Y(0))$ depend X, while $D \perp X$, then we maintains (1.2).

Stratification. For different age groups, we may want to use different treatment assignment probability. For example, we put 70% of the children into the treatment, while we put 40% of the adults into the treatment. In this case, D also depends on X. RCT can be implemented in a stratified approach: inside each age group, the researcher runs a RCT by flipping a coin to assign treatment.

To formalize the discussion, suppose the covariate $X \in \mathcal{X}$ takes discrete values, where $|\mathcal{X}| = m < \infty$. Assume further that

$$\begin{pmatrix} Y(1) \\ Y(0) \end{pmatrix} \perp D \big| X = x \quad \text{for all } x \in \mathcal{X}. \tag{1.5}$$

Caveat on aggregation. Notice that when D depends on x,

$$\mathbb{E}\left\{\mathbb{E}\left[Y|D=d,x\right]\right\} \neq \mathbb{E}\left[Y|D=d\right]$$

and thus

$$\text{ATE} \neq \mathbb{E}\left[Y|D=1\right] - \mathbb{E}\left[Y|D=0\right].$$

With that being said, the difference-in-means estimator

$$\widehat{\tau} = \frac{1}{n_1} \sum_{\{i:D_i=1\}} Y_i - \frac{1}{n_0} \sum_{\{i:D_i=0\}} Y_i$$

no longer provides a proper estimator for ATE.

Example 1. (Wager, 2025, Chapter 2, p.17) We are interested in evaluating whether providing teenagers with cash incentives can discourage smoking. To this end, an experiment was carried out in two locations: Palo Alto, CA, and Geneva, Switzerland.

Palo Alto	Non-Smoker	Smoker	Ratio (Smoker / Total)
Treatment	152	5	$5 / (152 + 5) = 0.032$
Control	2362	122	$122 / (2362 + 122) = 0.049$

Geneva	Non-Smoker	Smoker	Ratio (Smoker / Total)
Treatment	581	350	$350 / (581 + 350) = 0.376$
Control	2278	1979	$1979 / (2278 + 1979) = 0.465$

The results indicate that the treatment reduces the smoking rate among teenagers in both locations. However, looking at aggreate data can be misleading:

Palo Alto + Geneva	Non-Smoker	Smoker	Ratio (Smoker / Total)
Treatment	733	401	$401 / (733 + 401) = 0.354$
Control	4640	2101	$2101 / (4640 + 2101) = 0.312$

This echos the *Simpson's paradox*. Intuitively, we need to re-weight the two sample taking into consideration that Genevans are both more likely to be treated and more likely to smoke. This results in the following estimation:

$$\begin{aligned}
\hat{\tau}_{\text{PA}} &= \frac{5}{152+5} - \frac{122}{2362+122} \approx -1.7\% \\
\hat{\tau}_{\text{GVA}} &= \frac{350}{350+581} - \frac{1979}{2278+1979} \approx -8.9\% \\
\hat{\tau} &= \frac{2641}{2641+5188}\hat{\tau}_{\text{PA}} + \frac{5188}{2641+5188}\hat{\tau}_{\text{GVA}} \approx -6.5\%
\end{aligned}$$

Stratified Estimation. Example 1 suggests we ought to estimate the ATE by aggregating the estimates of CATE by

$$\widehat{\tau}^{\text{S}} = \sum_{x \in \mathcal{X}} \frac{n_x}{n} \underbrace{\left(\frac{1}{n_{x1}} \sum_{\{X_i = x, D_i = 1\}} Y_i - \frac{1}{n_{x0}} \sum_{\{X_i = x, D_i = 0\}} Y_i \right)}_{\widehat{\tau}(x)} = \sum_{x \in \mathcal{X}} \frac{n_x}{n} \widehat{\tau}(x), \tag{1.6}$$

where $n_x = |\{i : X_i = x\}|$ and $n_{nd} = |\{i : D_i = d, X_i = x\}|$.

Theorem 2 (Wager (2025) Theorem 2.1)**.** *Suppose* $(Y_i(1), Y_i(0), D_i, X_i) \sim_{\text{i.i.d.}} \mathbb{P}$ *for some super-population* $\mathbb{P}$*.* $X_i \in \mathcal{X}$ *where* $|\mathcal{X}| = m < \infty$*.* $\mathbb{E}\left(Y_i(d)^2 | X_i\right) < \infty$*. Assume stratified design* (1.5) *and SUTVA hold, and there is nontrivial treatment variation in* $\mathcal{X}$ *so that* $p(x) = \Pr(D_i = 1 | X_i = x) \in (0, 1)$ *for all* $x \in \mathcal{X}$*. Then*

$$\sqrt{n}\left(\widehat{\tau}^{\text{S}} - \tau\right) = \mathcal{N}\left(0, V^{\text{S}}\right),$$

where

$$\text{Var}\left(\tau(X_i)\right) + \mathbb{E}\left[\frac{\sigma_1^2(X_i)}{p(X_i)} + \frac{\sigma_0^2(X_i)}{1 - p(X_i)}\right]. \tag{1.7}$$

For a straightforward proof, see Wager (2025, pp. 19–20).

1.1.2 Inverse Propensity Weighting

The intuition from the stratified design can be in general formalized as the *unconfoundedness condition* (or *conditional independence assumption*):

$$\begin{pmatrix} Y(1) \\ Y(0) \end{pmatrix} \perp D \big| X. \tag{1.8}$$

Qualitatively, (1.8) requires that we have measured enough covariates to capture the dependence between treatment assignments and potential outcomes so that D cannot "peek" at $(Y(1), Y(0))$. $x \in \mathcal{X}$ is now allowed to be either discrete or continuous. When $\mathcal{X}$ is continuous, we cannot estimate $\widehat{\tau}(x)$ for each $x \in \mathcal{X}$ as in the stratified design, since there will not be enough observations at each value of x to compute the ATE reliably.

Characterization of ATE via Inverse Propensity Weighting. We take another route to do the computation via the propensity score as for (1.3). Denote the (conditional) propensity score as $p(x) = \Pr[D = 1|X = x]$. Then

$$\tau(x) \coloneqq \text{ATE}(x) = \frac{\mathbb{E}[DY|x]}{p(x)} - \frac{\mathbb{E}[(1-D)Y|x]}{1-p(x)}, \tag{1.9}$$

given that $0 < p(x) < 1$. Compared to (1.18), the (1.9) makes the role of $p(x)$ explicit. If we use (1.9) instead of (1.18), then

$$\text{ATE} = \mathbb{E}[\text{ATE}(x)] = \mathbb{E}\left[\frac{\mathbb{E}[DY|x]}{p(x)} - \frac{\mathbb{E}[(1-D)Y|x]}{1-p(x)}\right], \tag{1.10}$$

where the different $p(x)$ is explicitly accounted. Although (1.19) and (1.10) are mathematically equivalent, when we use the sample average to mimic the population average, (1.10) is easier to work with as it conditions only on the random variable X, but not D.

The unconfoundedness condition (1.8) seems impractical at the first glance. however, as shown in Rosenbaum and Rubin (1983), the propensity score can work as a dimension reduction tools which can make (1.8) more tractable: If (1.8) holds, then

$$\begin{pmatrix} Y(1) \\ Y(0) \end{pmatrix} \perp D \Big| p(X). \tag{1.11}$$

It suffices to control for $p(X)$ rather than X to account for the non-random treatment assignment. It is easy to verify that (Wager, 2025, p.21)

$$\Pr[D = d \mid Y(0), Y(1), p(X)] = \Pr[D = d \mid X]. \tag{1.12}$$

Suppose the functional form of $p(x)$ is known, and we can use the *inverse propensity weighting* (IPW) estimator

$$\widehat{\tau}^{\text{IPW}} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{D_i Y_i}{p(X_i)} - \frac{(1-D_i)Y_i}{1-p(X_i)}\right]. \tag{1.13}$$

Notice

$$\begin{aligned}\mathbb{E}\left[\frac{D_i Y_i}{p(X_i)}\right] &= \mathbb{E}\left\{\mathbb{E}\left[\frac{DY}{p(X)}\middle| X\right]\right\} = \mathbb{E}\left\{\frac{1}{p(X)}\mathbb{E}\left[DY|X\right]\right\} = \mathbb{E}\left\{\frac{1}{p(X)}\mathbb{E}\left[DY(1)|X\right]\right\} \\ &= \mathbb{E}\left\{\frac{\mathbb{E}\left[D|x\right]}{p(X)}\mathbb{E}\left[Y(1)|X\right]\right\} = \mathbb{E}\left\{\mathbb{E}\left[Y(1)|X\right]\right\} = \mathbb{E}\left[Y(1)\right],\end{aligned}$$

where we apply the unconfoundedness condition (1.8) from the first to the second line, and similarly

$$\mathbb{E}\left[\frac{(1-D_i)\,Y_i}{1-p(X_i)}\right] = \mathbb{E}\left[Y(0)\right].$$

We have $\widehat{\tau}^{\text{IPW}}$ is an unbiased estimator of ATE:

$$\mathbb{E}\left[\widehat{\tau}^{\text{IPW}}\right] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \text{ATE}.$$

Theorem 3. *Suppose* $(Y_i(1), Y_i(0), D_i, X_i) \sim_{\text{i.i.d.}} \mathbb{P}$. *Assume the unconfoundedness condition* (1.8) *and SUTVA hold.* $\mathbb{E}\left(Y_i\left(d\right)^2|X_i\right) < \infty$ *and there* $\eta > 0$ *s.t.*

$$\epsilon < p(x) < 1-\epsilon \text{ for all } x \in \mathcal{X}. \tag{1.14}$$

Then,

$$\sqrt{n}\left(\widehat{\tau}^{\text{IPW}} - \tau\right) = \mathcal{N}\left(0, V^{\text{IPW}}\right),$$

where

$$V^{\text{IPW}} = \text{Var}\left(\tau\left(X_i\right)\right) + \mathbb{E}\left[\frac{\sigma_1^2\left(X_i\right)}{p\left(X_i\right)} + \frac{\sigma_0^2\left(X_i\right)}{1-p\left(X_i\right)}\right] + \mathbb{E}\left[\frac{\left(\mu_0\left(X_i\right) + \left(1-p\left(X_i\right)\right)\tau\left(X_i\right)\right)^2}{p\left(X_i\right)\left(1-p\left(X_i\right)\right)}\right]. \tag{1.15}$$

Refer to Wager (2025, p.23 - 24) for a proof.

Overlap Assumption. (1.14), together with the finite second moments of the outcomes, ensures all moments involves in (1.15) are finite. This is known as the *overlap* assumption. In general, we need to guarantee there is enough randomness in treatment assignment to justify the treatment effect estimation, i.e. D_i cannot be perfectly predicted by X_i. Technically,

if we strengthen the finite second moment assumption to assume outcomes are uniformly bounded, then (1.14) can be relaxed to

$$\mathbb{E}\left(\frac{1}{p\left(X_i\right)\left(1-p\left(X_i\right)\right)}\right)<\infty,$$

which often referred as weak overlap condition.

Experimental v.s. observational studies. There are two conceptually distinct ways in which potential outcomes can satisfy (1.5). The foregoing discussion primarily addressed the case where data arise from an experiment employing stratified treatment assignment. In this experimental setting, nature first draws $\{Y_i(0), Y_i(1), X_i\} \sim \mathbb{P}$. The experimenter then assigns treatment according to $D_i \sim \text{Bern}\left(p(X_i)\right)$, where the propensity score $p(\cdot)$ is deliberately chosen and thus known. Alternatively, we may encounter data $(Y_i(1), Y_i(0), D_i, X_i) \sim \mathbb{P}$ that do not originate from a controlled experiment; in such cases, the unconfoundedness assumption is invoked to enable causal identification. These latter scenarios are commonly referred to as *observational studies* or *natural experiments*. In such settings, justifying the unconfoundedness assumption is often challenging. Furthermore, the propensity score $p(x)$ is unknown to the researcher and must be estimated from the data.

Feasible IPW. In observational studies, we often don't know the exact propensity score $p\left(x\right)$. To construct the *feasible* IPW estimator, we need to plug in the estimates of the propensity score, $\widehat{p}\left(x\right)$, in (1.13):

$$\widehat{\tau}^{\text{IPW}}=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_iY_i}{\widehat{p}\left(X_i\right)}-\frac{\left(1-D_i\right)Y_i}{1-\widehat{p}\left(X_i\right)}\right). \tag{1.16}$$

In practice, estimating the propensity score accurately can be difficult. Theoretical guarantees as in Theorem 3 can break down if we plug in generic estimates of the propensity without further underlying assumptions. Robustness of methods to errors in the propensity scores is important.

Inefficiency of $\widetilde{\tau}^{\text{IPW}}$. Let's revisit the stratified design, and compare the asymptotic variance of the IPW estimator $\widetilde{\tau}^{IPW}$ and the stratification estimator $\widehat{\tau}^{\text{S}}$, i.e. V^{S} in (1.7) and V^{IPW} in (1.15):

$$V^{\text{IPW}}=V^{\text{S}}+\underbrace{\mathbb{E}\left[\frac{\left(\mu_0\left(X_i\right)+\left(1-p\left(X_i\right)\right)\tau\left(X_i\right)\right)^2}{p\left(X_i\right)\left(1-p\left(X_i\right)\right)}\right]}_{\geq 0}\geq V^{\text{S}},$$

i.e. $\widetilde{\tau}^{\text{IPW}}$ is strictly less efficient than $\widehat{\tau}^{\text{S}}$ unless $\mu_0\left(X_i\right)+\left(1-p\left(X_i\right)\right)\tau\left(X_i\right)=0$ almost surely.

Zooming into $\widehat{\tau}^S$, it turns on $\widehat{\tau}^{\mathrm{S}}$ is actually the *feasible* IPW estimator

$$\widehat{\tau}^{\mathrm{S}} = \widehat{\tau}^{\mathrm{IPW}},$$

where $\widehat{p}(x) = \frac{n_{x1}}{n_x}$. Surprisingly, this yields the seemingly paradoxical result that the *oracle* IPW estimator is actually less efficient than the *feasible* IPW estimator when $\mathcal{X}$ is discrete. Hirano et al. (2003) formalize this insight for the case of continuous covariates, demonstrating that the IPW estimator, based on a nonparametrically estimated propensity score obtained from a sufficiently smooth, growing sieve approximation, attains the semiparametric efficiency bound.

Example 2 (Missing-at-random with binary X: why $\widehat{p}$ can reduce variance)**.** We consider the missing-at-random setup used in Section 3 of Hirano et al. (2003). Let $X \in \{0, 1\}$ be a binary covariate, $D \in \{0, 1\}$ indicate whether Y is observed, and suppose the (observation) propensity score is

$$p(x) = \Pr(D = 1 \mid X = x) = \frac{1}{2},$$

and $D \perp Y \mid X$. The outcome satisfies the structural equation

$$Y = \mu(X) + \varepsilon,\ \mathbb{E}[\varepsilon \mid X] = 0,\ \mathrm{Var}(\varepsilon \mid X) = \mathrm{Var}(Y \mid X).$$

The target is the population mean $\theta = \mathbb{E}[Y] = \mathbb{E}[\mu(X)]$.

Denote

$$n_x = \sum_{i=1}^{n} \mathbf{1}\{X_i = x\}, \quad n_{x1} = \sum_{i=1}^{n} D_i\, \mathbf{1}\{X_i = x\}.$$

Using the true propensity score yields

$$\widetilde{\theta} = \frac{1}{n} \sum_{i=1}^{n} \frac{D_i Y_i}{p(X_i)},$$

which is the Horvitz–Thompson estimator (Horvitz and Thompson, 1952). To see where inefficiency comes from, decompose the estimator using $Y_i = \mu(X_i) + \varepsilon_i$:

$$\begin{aligned}
\widetilde{\theta} &= \frac{1}{n} \sum_{i=1}^{n} \frac{D_i \mu(X_i)}{p(X_i)} + \frac{1}{n} \sum_{i=1}^{n} \frac{D_i \varepsilon_i}{p(X_i)} \\
&= \frac{1}{n} \sum_{x \in \{0,1\}} \mu(x)\, \frac{\sum_{i=1}^{n} D_i \mathbf{1}\{\mathbf{X_i} = \mathbf{x}\}}{p(x)} + \frac{1}{n} \sum_{i=1}^{n} \frac{D_i \varepsilon_i}{p(X_i)}
\end{aligned}$$

$$= \frac{1}{n} \sum_{x \in \{0,1\}} \mu(x) \underbrace{\frac{n_{x1}}{p(x)}}_{:=\widetilde{n}_x} + \frac{1}{n} \sum_{i=1}^{n} \frac{D_i \varepsilon_i}{p(X_i)}$$

$$= \sum_{x \in \{0,1\}} \frac{n_x}{n} \mu(x) + \frac{1}{n} \sum_{x \in \{0,1\}} (\widetilde{n}_x - n_x) \mu(x) + \frac{1}{n} \sum_{i=1}^{n} \frac{D_i \varepsilon_i}{p(X_i)} \tag{1.17}$$

- The first term in (1.17) represents the "full-sample" mean of $\mu(X)$, which can be further expanded as

$$\begin{aligned} \sum_{x \in \{0,1\}} \frac{n_x}{n} \mu(x) &= \frac{1}{n} \sum_{i=1}^{n} \mu(X_i) \\ &= \sum_{x \in \{0,1\}} \Pr(X = x) \mu(x) + \sum_{x \in \{0,1\}} \left(\frac{n_x}{n} - \Pr(X = x) \right) \mu(x) \\ &= \theta + \sum_{x \in \{0,1\}} \left(\frac{n_x}{n} - \Pr(X = x) \right) \mu(x). \end{aligned}$$

  This term appropriately re-weights each group defined by $x \in \{0, 1\}$ in the finite sample according to its sample fraction $\frac{n_x}{n}$. It is centered at θ and its contribution to the variance arises solely from sampling variability in n_x, which in finite samples follows a binomial distribution with parameters n and $\Pr(X = x)$.

- The second term in (1.17) arises because the oracle propensity score does not guarantee any *finite-sample* balance. Due to the randomness of missingness within each group, the observed subsample may over- or under-represent a group, as captured by the discrepancy $\widetilde{n}_x - n_x$. This imbalance introduces additional variation into the estimator.

- The third term in (1.17) contains the idiosyncratic noises.

we can compute the variance of the estimator,

$$\begin{aligned} n \operatorname{Var}\left(\widetilde{\theta}\right) &= n \left[\mathbb{E}\left(\operatorname{Var}\left(\widetilde{\theta} \mid X \right) \right) + \operatorname{Var}\left(\mathbb{E}\left(\widetilde{\theta} \mid X \right) \right) \right] \\ &= \operatorname{Var}\left(\mu(X) \right) + \mathbb{E}\left(\frac{\operatorname{Var}(Y \mid X)}{p(X)} \right) + \mathbb{E}\left(\frac{1 - p(X)}{p(X)} \mu(X)^2 \right) \end{aligned}$$

- The term $\operatorname{Var}(\mu(X))$ is the variance contribution of the mean of the outcome.
- The term $\mathbb{E}\left(\frac{\operatorname{Var}(Y|X)}{p(X)}\right)$ represents the irreducible idiosyncratic noise ε.
- The term $\mathbb{E}\left(\frac{1-p(X)}{p(X)} \mu(X)^2\right)$ reflects the variance contribution of the imbalance term.

Note that in (1.17), replacing $p(x)$ by the estimated propensity score $\widehat{p}(x) = \frac{n_{x1}}{n_x}$ leads to

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \frac{D_i Y_i}{\widehat{p}(X_i)}.$$

Now $\widehat{n}_x = \frac{n_{x1}}{\widehat{p}(x)} = n_x$, the imputed weights exactly balance the realized subsamples, so we can follow the calculation in (1.17) and get Therefore,

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \mu(X_i) + \frac{1}{n}\sum_{i=1}^{n} \frac{D_i \varepsilon_i}{\widehat{p}(X_i)},$$

in which the imbalance term cancels out. By the standard CLT argument, we can show[1]

$$\sqrt{n}\widehat{\theta} \xrightarrow{d} = \mathcal{N}\left(0, \operatorname{Var}\left(\mu\left(X\right)\right) + \mathbb{E}\left(\frac{\operatorname{Var}(Y \mid X)}{p\left(X\right)}\right)\right).$$

Thus the efficiency gain is exactly the removal of the $\mathbb{E}\left(\frac{1-p(X)}{p(X)}\mu\left(X\right)^2\right)$ component.

1.1.3 Doubly Robust Estimator

Characterization of ATE via Outcome Regression. The CATE is made operational via

$$\begin{aligned} \operatorname{ATE}(x) &= \mathbb{E}\left[Y(1)|X = x\right] - \mathbb{E}\left[Y(0)|X = x\right] \\ &= \mathbb{E}\left[Y(1)|D(x) = 1, x\right] - \mathbb{E}\left[Y(0)|D(x) = 0, X_i = x\right] \\ &= \mathbb{E}\left[Y|D = 1, x\right] - \mathbb{E}\left[Y|D = 0, x\right]. \end{aligned} \tag{1.18}$$

Inside each age group, we just need to compute the difference between the averages of those treated and those untreated, respectively. Now, suppose the researcher has CATE at hand, and she wants to aggregate the CATE across subgroups into an overall ATE. Then by the law of iterated expectations:

$$\operatorname{ATE} = \mathbb{E}\left[\operatorname{ATE}(X)\right] = \mathbb{E}\left\{\mathbb{E}\left[Y|D = 1, X\right] - \mathbb{E}\left[Y|D = 0, X\right]\right\} = \mathbb{E}\left[\mu_1\left(X\right) - \mu_0\left(X\right)\right], \tag{1.19}$$

[1] Observe that $\widehat{\theta} = \sum_{x\in\{0,1\}} \frac{n_x}{n_{x1}} \left(\frac{1}{n}\sum_{i=1}^{n} D_i \mathbf{1}\{X_i = x\}\varepsilon_i\right)$. To analyze the asymptotic distribution, apply the Central Limit Theorem (CLT) separately to each term within the parentheses for $x \in \{0, 1\}$, and then invoke Slutsky's theorem to conclude.

where $\mu_d(x) = \mathbb{E}[Y|D=d, X=x]$. This leads to a simple and consistent (*but not necessarily optimal*) nonparametric regression estimator for ATE:

$$\widehat{\tau}^{\text{REG}} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i)\right), \tag{1.20}$$

where $\widehat{\mu}_j(\cdot)$ is a consistent nonparametric estimators for $\mu_j(\cdot)$.

Augmented Inverse Propensity Weighting (AIPW). Given the two characterizations, ATE can be estimated by either the IPW estimator (1.16) or the outcome regression estimator (1.20). Both estimators require the estimation of nuisance components ($p(\cdot)$ in $\widehat{\tau}^{\text{IPW}}$ and $\mu_d(\cdot)$ in $\widehat{\tau}^{\text{REG}}$). Consequently, the validity of these estimators relies heavily on the accuracy with which these nuisance components are estimated. Though the original motivation was mainly theoretical relying the *semiparametric efficiency theory*[2], it is natural to ask whether it is possible to combine both strategies to mitigate the bias and improve efficiency; see brief discussion in Ding (2024, Section 12.2) and references therein. This leads to the augmented inverse propensity weighting (AIPW) estimator (Robins et al., 1994),

$$\widehat{\tau}^{\text{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i) + \frac{D_i\left(Y_i - \widehat{\mu}_1(X_i)\right)}{\widehat{p}(X_i)} - \frac{(1-D_i)\left(Y_i - \widehat{\mu}_0(X_i)\right)}{1-\widehat{p}(X_i)}\right] \tag{1.21}$$

The AIPW estimator can be interpreted as first making a best effort at estimating τ by modeling the conditional means $\mu_0(x)$ and $\mu_1(x)$. Then, it adjusts for any remaining bias in the estimated conditional means $\widehat{\mu}_d(x)$ by applying IPW to the regression residuals. Statistically, AIPW not only inherits the robustness properties of both the regression and IPW estimators, but it can also improve upon both.

Weak Double Robustness. $\widehat{\tau}^{\text{AIPW}}$ is consistent if *either* $\widehat{\mu}_d(x)$ are consistent *or* $\widehat{p}(x)$ is consistent. To see this, first consider the case where $\widehat{\mu}_d(x)$ is consistent, i.e., $\widehat{\mu}_d(x) \approx \mu_d(x)$. Then,

$$\begin{aligned}\widehat{\tau}^{\text{AIPW}} = &\underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i)\right)}_{=\widehat{\tau}^{\text{REG}}} \\ &+ \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i}{\widehat{p}(X_i)}\left(Y_i - \widehat{\mu}_1(X_i)\right) - \frac{1-D_i}{1-\widehat{p}(X_i)}\left(Y_i - \widehat{\mu}_0(X_i)\right)\right)}_{\approx \text{ mean-zero noise}},\end{aligned}$$

because $\mathbb{E}\left(Y - \widehat{\mu}_d(X) \mid X, D\right) \approx 0$ under unconfoundedness, which zeros out the estimation

[2] See Bodhisattva Sen's note for a accessible introduction to semiparametric efficiency theory, and refer to Bickel et al. (1993) for the rigorous treatment.

error in the propensity score. Conversely, suppose $\widehat{p}(x)$ is consistent, i.e. $\widehat{p}(x) \approx p(x)$, then

$$\widehat{\tau}^{\text{AIPW}} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i Y_i}{\widehat{p}(X_i)} - \frac{(1-D_i)Y_i}{1-\widehat{p}(X_i)}\right)}_{=\widehat{\tau}^{\text{IPW}}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\mu}_1(X_i)\left(1-\frac{D_i}{\widehat{p}(X_i)}\right) - \widehat{\mu}_0(X_i)\left(1-\frac{1-D_i}{1-\widehat{p}(X_i)}\right)\right)}_{\approx \text{ mean-zero noise}},$$

since $\mathbb{E}\left(1 - D/\widehat{p}(X) \mid x\right) \approx 0$. Again, this can zero out the estimation error in the outcome regressions even if $\widehat{\mu}_d(x)$ is inconsistent.

Remark 1. Weak robustness ensures only the consistency of $\widehat{\tau}^{\text{AIPW}}$. If both $\widehat{\mu}_d(x)$ and $\widehat{p}(x)$ are consistently estimated using appropriate nonparametric or machine learning methods, then weak double robustness does not, by itself, provide further advantages. In practice, however, we are often interested not just in consistency, but also in the rate of convergence and in uncertainty quantification. Modern machine learning algorithms used to estimate $\mu_d(x)$ and $p(x)$ often cannot achieve the parametric $\sqrt{n}$ rate but rather slower rate like $n^{1/4}$ unless strong assumptions are imposed. This natural question is how robust is the performance of the AIPW estimator to the quality of the nuisance estimates $\mu_d(x)$ and $p(x)$, compared to the oracle AIPW estimator in which both nuisance components are known.

Oracle AIPW. The *oracle* AIPW estimator is constructed with true outcome means $\mu_1(x)$ and $\mu_0(x)$ and propensity score $p(x)$:

$$\widetilde{\tau}^{\text{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\left[\mu_1(X_i) - \mu_0(X_i) + \frac{D_i\left(Y_i - \mu_1(X_i)\right)}{p(X_i)} - \frac{1-D_i\left(Y_i - \mu_0(X_i)\right)}{1-p(X_i)}\right]. \tag{1.22}$$

Theorem 4 (Wager (2025) Proposition 3.1)**.** *Under the assumptions of Theorem 3,*

$$\sqrt{n}\left(\widetilde{\tau}^{\text{AIPW}} - \tau\right) = \mathcal{N}\left(0, V^*\right),$$

where

$$V^* = \text{Var}\left(\tau\left(X_i\right)\right) + \mathbb{E}\left[\frac{\sigma_1^2\left(X_i\right)}{p\left(X_i\right)} + \frac{\sigma_0^2\left(X_i\right)}{1-p\left(X_i\right)}\right]. \tag{1.23}$$

It is easy to show the result by applying central limit theorem; see Wager (2025, Proposition 3.1) for a proof.

Remark 2. Note that $V^* = V^{\text{S}}$, but under general conditions beyond stratified design with discrete covariates. In fact, V^* is the *efficiency bound* for nonparametric ATE estimation

under unconfoundedness. Wager (2025, Theorem 3.4) provides a proof sketch using the argument of Chamberlain (1992). Given $\widetilde{\tau}^{\mathrm{AIPW}}$ achieves the *optimal* asymptotic behavior, for any other operational estimator, such as $\widehat{\tau}^{\mathrm{AIPW}}$, we aim to establish the asymptotic equivalence that $\sqrt{n}\left(\widehat{\tau}-\widetilde{\tau}^{AIPW}\right)=o_p\left(1\right)$.

Population bias decomposition of $\widehat{\tau}^{\mathrm{AIPW}}$. Let $\eta := (\mu_0, \mu_1, p)$ collect nuisance functions and let $\eta_0 := (\mu_{0,0}, \mu_{1,0}, p_0)$ denote the truth, where

$$p_0(x) := \Pr(D = 1 \mid X = x),\ \mu_{d,0}(x) := \mathbb{E}(Y \mid D = d, X = x),\ d \in \{0, 1\}.$$

Let $W := (Y, D, X)$. Define the AIPW score

$$\psi(W;\eta) := \mu_1(X) - \mu_0(X) + \frac{D}{p(X)}\big(Y - \mu_1(X)\big) - \frac{1-D}{1-p(X)}\big(Y - \mu_0(X)\big), \tag{1.24}$$

and the associated moment function

$$m(W;\tau,\eta) := \psi(W;\eta) - \tau. \tag{1.25}$$

Under unconfoundedness and overlap, the target $\tau_0 := \mathbb{E}(Y(1)-Y(0))$ satisfies the *population moment condition*

$$\mathbb{E}\big[m(W;\tau_0,\eta_0)\big] = 0. \tag{1.26}$$

Remark 3. Note that $\widehat{\tau}^{\mathrm{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\psi\left(W_i;\widehat{\eta}\right)$ and $\widetilde{\tau}^{\mathrm{AIPW}} = \frac{1}{n}\sum_{i}^{n}\psi\left(W_i;\eta_0\right)$.

Remark 4. In the following discussion, we first establish identities for *fixed* (non-random) functions $\widetilde{\eta}$. When we later substitute an estimated nuisance $\widehat{\eta} = \widehat{\eta}(\mathcal{D}_n)$ constructed from a sample $\mathcal{D}_n := \{W_i\}_{i=1}^n$, the same identities are used in the following precise sense: let $W^\star = (Y^\star, D^\star, X^\star)$ be an independent draw from the population, $W^\star \perp \mathcal{D}_n$; then all "population" expectations involving $\widehat{\eta}$ should be read as conditional expectations over $W^\star$ given $\mathcal{D}_n$, e.g.

$$\mathbb{E}\big[m(W;\tau_0,\widehat{\eta})\big] \quad \text{means} \quad \mathbb{E}\big[m(W^\star;\tau_0,\widehat{\eta}(\mathcal{D}_n)) \mid \mathcal{D}_n\big].$$

This distinction matters because if $\widehat{\eta}$ is trained on the full sample, then $\widehat{\eta}(X_i)$ may depend on (Y_i, D_i) through the fitting procedure; hence one generally cannot treat $\widehat{\eta}(X_i)$ as measurable w.r.t. $\sigma(X_i)$ in *in-sample* conditional expectations. Cross-fitting introduced later restores the needed conditional independence fold-by-fold.

Fix $\widetilde{\eta} := (\widetilde{\mu}_0, \widetilde{\mu}_1, \widetilde{p})$ with $\varepsilon \le \widetilde{p}(X) \le 1-\varepsilon$ a.s. Define $\delta_d(X) := \widetilde{\mu}_d(X) - \mu_{d,0}(X)$. Because $\widetilde{\mu}_d(X)$ and $\widetilde{p}(X)$ are $\sigma(X)$-measurable (they are *fixed* functions of X), we may condition on

X and pull them out of conditional expectations.

$$\begin{aligned}\mathbb{E}\left[\frac{D}{\widetilde{p}(X)}\left(Y-\widetilde{\mu}_1(X)\right)\middle| X\right] &= \frac{1}{\widetilde{p}(X)}\mathbb{E}\left[D\mathbb{E}\left(Y-\widetilde{\mu}_1(X)\mid D,X\right)\mid X\right]\\ &= \frac{1}{\widetilde{p}(X)}\mathbb{E}\left[D\left(\mu_1(X)-\widetilde{\mu}_1(X)\right)\mid X\right]\\ &= \frac{p_0(X)}{\widetilde{p}(X)}\left(\mu_1(X)-\widetilde{\mu}_1(X)\right) = -\frac{p_0(X)}{\widetilde{p}(X)}\delta_1(X),\end{aligned}$$

and similarly

$$\mathbb{E}\left[\frac{1-D}{1-\widetilde{p}(X)}\left(Y-\widetilde{\mu}_0(X)\right)\middle| X\right] = -\frac{1-p_0(X)}{1-\widetilde{p}(X)}\delta_0(X).$$

Substituting these into (1.24) gives

$$\begin{aligned}\mathbb{E}\left[\psi\left(W;\widetilde{\eta}\right)\mid X\right] &= \widetilde{\mu}_1(X)-\widetilde{\mu}_0(X)-\frac{p_0(X)}{\widetilde{p}(X)}\delta_1(X)+\frac{1-p_0(X)}{1-\widetilde{p}(X)}\delta_0(X)\\ &= \mu_1(X)-\mu_0(X)+\left(1-\frac{p_0(X)}{\widetilde{p}(X)}\right)\delta_1(X)+\left(\frac{1-p_0(X)}{1-\widetilde{p}(X)}-1\right)\delta_0(X).\end{aligned}$$

Simplying the above, we obtain the exact conditional representation

$$\mathbb{E}\left[\psi\left(W;\widetilde{\eta}\right)\mid X\right] = \mu_{1,0}(X)-\mu_{0,0}(X)+\left(\widetilde{p}(X)-p_0(X)\right)\left\{\frac{\widetilde{\mu}_1(X)-\mu_{1,0}(X)}{\widetilde{p}(X)}+\frac{\widetilde{\mu}_0(X)-\mu_{0,0}(X)}{1-\widetilde{p}(X)}\right\}.$$

Taking expectations over X yields the population bias identity

$$\mathbb{E}\left[\psi\left(W;\widetilde{\eta}\right)\right]-\tau_0 = \mathbb{E}\left[\left(\widetilde{p}(X)-p_0(X)\right)\left\{\frac{\widetilde{\mu}_1(X)-\mu_{1,0}(X)}{\widetilde{p}(X)}+\frac{\widetilde{\mu}_0(X)-\mu_{0,0}(X)}{1-\widetilde{p}(X)}\right\}\right]. \quad (1.27)$$

Then,

$$\mathbb{E}\big[m(W;\tau_0,\widetilde{\eta})\big] = \mathbb{E}\left[\left(\widetilde{p}(X)-p_0(X)\right)\left\{\frac{\delta_1(X)}{\widetilde{p}(X)}+\frac{\delta_0(X)}{1-\widetilde{p}(X)}\right\}\right]. \quad (1.28)$$

Equation (1.28) immediately yields:

$$\mathbb{E}[m(W;\tau_0,\widetilde{\eta})] = 0 \quad \text{if either} \quad \widetilde{\mu}_d = \mu_{d,0} \text{ a.s. for } d=0,1, \quad \text{or} \quad \widetilde{p} = p_0 \text{ a.s.}$$

This is the population version of *double robustness*. Moreover, (1.28) shows that, in general, the population bias is a *product of nuisance errors*.

Remark 5. Even if $\widehat{\eta} = \widehat{\eta}(\mathcal{D}_n)$ is estimated using the full sample (no sample splitting), the algebra behind (1.28) remains valid *conditionally* for an independent draw $W^\star$. Specifically,

with $W^\star \perp \mathcal{D}_n$,

$$\mathbb{E}\big[m(W^\star;\tau_0,\widehat{\eta}) \mid \mathcal{D}_n\big] = \mathbb{E}\left[\big(\widehat{p}(X^\star) - p_0(X^\star)\big)\left\{\frac{\widehat{\mu}_1(X^\star) - \mu_{1,0}(X^\star)}{\widehat{p}(X^\star)} + \frac{\widehat{\mu}_0(X^\star) - \mu_{0,0}(X^\star)}{1-\widehat{p}(X^\star)}\right\}\middle| \mathcal{D}_n\right]. \tag{1.29}$$

Equation (1.29) is the precise sense in which we apply the "exact bias decomposition" when $\widetilde{\eta} = \widehat{\eta}$. However, (1.29) does *not* justify in-sample manipulations such as $\mathbb{E}[D_i/\widehat{p}(X_i) \mid X_i] = p_0(X_i)/\widehat{p}(X_i)$, because $\widehat{p}(X_i)$ need not be $\sigma(X_i)$-measurable when trained on the same sample.

Neyman orthogonality. For a perturbation direction $h := (h_0, h_1, h_p)$, define a path $\eta_t := \eta_0 + th$. We say the moment $m(W;\tau,\eta)$ is *Neyman orthogonal at* (τ_0,η_0) if

$$\frac{d}{dt}\mathbb{E}\big[m(W;\tau_0,\eta_t)\big]\Big|_{t=0} = 0 \qquad \text{for all admissible directions } h. \tag{1.30}$$

Equivalently, the (Gateaux) derivative of $\eta \mapsto \mathbb{E}[m(W;\tau_0,\eta)]$ vanishes at η_0[3].

Step 3: compute derivatives for AIPW. It is convenient to rewrite ψ as

$$\psi(W;\eta) = \frac{DY}{p(X)} - \frac{(1-D)Y}{1-p(X)} + \mu_1(X)\left(1 - \frac{D}{p(X)}\right) - \mu_0(X)\left(1 - \frac{1-D}{1-p(X)}\right). \tag{1.31}$$

From (1.25) and (1.31), note that $\partial_\eta \mathbb{E}[m] = \partial_\eta \mathbb{E}[\psi]$ since τ_0 is fixed.

Derivative w.r.t. μ_1. Let $\mu_{1,t} := \mu_{1,0} + th_1$ and hold $(\mu_0, p) = (\mu_{0,0}, p_0)$ fixed. Then by (1.31),

$$\psi(W;\mu_{0,0},\mu_{1,t},p_0) - \psi(W;\eta_0) = t\,h_1(X)\left(1 - \frac{D}{p_0(X)}\right).$$

Hence

$$\begin{aligned}
\frac{d}{dt}\mathbb{E}\big[\psi(W;\mu_{0,0},\mu_{1,t},p_0)\big]\Big|_{t=0} &= \mathbb{E}\left[h_1(X)\left(1 - \frac{D}{p_0(X)}\right)\right] \\
&= \mathbb{E}\left[h_1(X)\,\mathbb{E}\left(1 - \frac{D}{p_0(X)}\middle| X\right)\right] \\
&= \mathbb{E}\left[h_1(X)\left(1 - \frac{\mathbb{E}(D \mid X)}{p_0(X)}\right)\right] = 0.
\end{aligned} \tag{1.32}$$

[3]Refer to Luenberger (1997, Section 7.2) for formal definitions, and other chapters for textbook treatment of foundations of functional spaces.

Derivative w.r.t. μ_0. Similarly, letting $\mu_{0,t} := \mu_{0,0} + t h_0$ and holding $(\mu_1, p) = (\mu_{1,0}, p_0)$ fixed,

$$\psi(W; \mu_{0,t}, \mu_{1,0}, p_0) - \psi(W; \eta_0) = -\,t\, h_0(X) \left(1 - \frac{1-D}{1-p_0(X)}\right),$$

so

$$\begin{aligned}\frac{d}{dt}\mathbb{E}\big[\psi(W; \mu_{0,t}, \mu_{1,0}, p_0)\big]\Big|_{t=0} &= -\mathbb{E}\left[h_0(X)\left(1 - \frac{1-D}{1-p_0(X)}\right)\right] \\ &= -\mathbb{E}\left[h_0(X)\,\mathbb{E}\left(1 - \frac{1-D}{1-p_0(X)}\,\middle|\, X\right)\right] = 0. \end{aligned} \tag{1.33}$$

Derivative w.r.t. p. Now let $p_t := p_0 + t h_p$ and hold $(\mu_0, \mu_1) = (\mu_{0,0}, \mu_{1,0})$ fixed. Using the original form of ψ,

$$\psi(W; \mu_{0,0}, \mu_{1,0}, p_t) = \mu_{1,0}(X) - \mu_{0,0}(X) + \frac{D}{p_t(X)}\big(Y - \mu_{1,0}(X)\big) - \frac{1-D}{1-p_t(X)}\big(Y - \mu_{0,0}(X)\big),$$

and differentiating at $t = 0$ yields

$$\frac{d}{dt}\psi(W; \mu_{0,0}, \mu_{1,0}, p_t)\Big|_{t=0} = -\frac{D\, h_p(X)}{p_0(X)^2}\big(Y - \mu_{1,0}(X)\big) - \frac{(1-D)\, h_p(X)}{(1-p_0(X))^2}\big(Y - \mu_{0,0}(X)\big).$$

Taking expectations and conditioning on X gives

$$\begin{aligned}\frac{d}{dt}\mathbb{E}\big[\psi(W; \mu_{0,0}, \mu_{1,0}, p_t)\big]\Big|_{t=0} &= -\mathbb{E}\left[\frac{h_p(X)}{p_0(X)^2}\mathbb{E}\big(D\big(Y - \mu_{1,0}(X)\big)\big|X\big)\right] \\ &\quad - \mathbb{E}\left[\frac{h_p(X)}{(1-p_0(X))^2}\mathbb{E}\big((1-D)\big(Y - \mu_{0,0}(X)\big)\big|X\big)\right]. \end{aligned} \tag{1.34}$$

Now, by the law of iterated expectations,

$$\mathbb{E}\big(D\big(Y - \mu_{1,0}(X)\big)\big|X\big) = \mathbb{E}(D\,\mathbb{E}(Y - \mu_{1,0}(X) \mid D, X)|X) = \mathbb{E}(D \cdot 0|X) = 0,$$

and similarly $\mathbb{E}((1-D)(Y - \mu_{0,0}(X)) \mid X) = 0$. Substituting into (1.34) yields

$$\frac{d}{dt}\mathbb{E}\big[\psi(W; \mu_{0,0}, \mu_{1,0}, p_t)\big]\Big|_{t=0} = 0. \tag{1.35}$$

Combining (1.32), (1.33), and (1.35), we obtain

$$\frac{d}{dt}\mathbb{E}\big[m(W; \tau_0, \eta_0 + th)\big]\Big|_{t=0} = \frac{d}{dt}\mathbb{E}\big[\psi(W; \eta_0 + th)\big]\Big|_{t=0} = 0 \quad \text{for all } h,$$

which verifies (1.30). Hence, the AIPW moment is Neyman orthogonal at (τ_0, η_0).

The Taylor expansion around η_0 has the form

$$\mathbb{E}[m(W;\tau_0,\widehat{\eta})] = \mathbb{E}[m(W;\tau_0,\eta_0)] + \partial_\eta \mathbb{E}[m(W;\tau_0,\eta)]\big|_{\eta=\eta_0}[\widehat{\eta}-\eta_0] + R_2(\widehat{\eta},\eta_0), \tag{1.36}$$

where $R_2(\widehat{\eta},\eta_0)$ is a second-order remainder. By (1.26), the first term is zero. By Neyman orthogonality (1.30), the linear term is also zero. Thus,

$$\mathbb{E}[m(W;\tau_0,\widehat{\eta})] = R_2(\widehat{\eta},\eta_0), \tag{1.37}$$

so any population bias from plugging in $\widehat{\eta}$ is *second order*.[4] Indeed, the exact bias identity applied conditionally (cf. (1.29)) yields

$$\mathbb{E}[m(W;\tau_0,\widehat{\eta})] = \mathbb{E}\left[(\widehat{p}(X)-p_0(X))\left\{\frac{\widehat{\mu}_1(X)-\mu_{1,0}(X)}{\widehat{p}(X)} + \frac{\widehat{\mu}_0(X)-\mu_{0,0}(X)}{1-\widehat{p}(X)}\right\}\right], \tag{1.38}$$

again with the convention that this is a conditional population expectation when $\widehat{\eta}$ is random. In particular, if $\widehat{p}(X) \in [\varepsilon, 1-\varepsilon]$ a.s., then

$$|\mathbb{E}[m(W;\tau_0,\widehat{\eta})]| \le \frac{1}{\varepsilon}\,\mathbb{E}\left[|\widehat{p}(X)-p_0(X)|\,|\widehat{\mu}_1(X)-\mu_{1,0}(X)+\widehat{\mu}_0(X)-\mu_{0,0}(X)|\right], \tag{1.39}$$

and by Cauchy–Schwarz,

$$|\mathbb{E}[m(W;\tau_0,\widehat{\eta})]| \le \frac{1}{\varepsilon}\,\|\widehat{p}-p_0\|_{L^2(P_X)}\left(\|\widehat{\mu}_1-\mu_{1,0}\|_{L^2(P_X)} + \|\widehat{\mu}_0-\mu_{0,0}\|_{L^2(P_X)}\right), \tag{1.40}$$

where $\|f\|_{L^2(P_X)} := (\mathbb{E}[f(X)^2])^{1/2}$ denotes the $L^2(P_X)$ norm.

The AIPW estimator can be written as $\widehat{\tau}^{\mathrm{AIPW}} = \mathbb{P}_n[\psi(W;\widehat{\eta})]$. A basic decomposition is

$$\widehat{\tau}^{\mathrm{AIPW}} - \tau_0 = (\mathbb{P}_n-\mathbb{P})\psi(W;\eta_0) + \underbrace{\mathbb{P}\big[\psi(W;\widehat{\eta})-\psi(W;\eta_0)\big]}_{\text{population bias}} + \underbrace{(\mathbb{P}_n-\mathbb{P})\big(\psi(W;\widehat{\eta})-\psi(W;\eta_0)\big)}_{\text{overfitting}}. \tag{1.41}$$

The first term is the usual sampling fluctuation at the truth. By (1.38), orthogonality (and the exact bias identity) control the *second term*: it is second order and vanishes if either nuisance is correct (weak double robustness in terms of consistency), and it is $o_p(n^{-1/2})$ under a product-rate condition (for root-n inference).

The subtlety is the *third term* in (1.41). Even if the population Neyman condition holds,

[4] If $\widehat{\eta}$ is data-dependent, read $\mathbb{E}[m(W;\tau_0,\widehat{\eta})]$ as $\mathbb{E}[m(W^\star;\tau_0,\widehat{\eta}) \mid \mathcal{D}_n]$

this third term can fail to be small if $\widehat{\eta}$ is estimated on the same sample used in $\mathbb{P}_n$. The reason is that $\psi(\cdot;\widehat{\eta})$ is then a *data-dependent* function evaluated on the same data, so controlling $(\mathbb{P}_n - \mathbb{P})f_n$ with $f_n = \psi(\cdot;\widehat{\eta}) - \psi(\cdot;\eta_0)$ requires strong uniform complexity conditions (e.g. Donsker/entropy conditions) that typically fail for modern ML.

Sample Spliting and Cross-fitting. To deal with the third term in (1.41), we rely on sample splitting and cross-fittingn to restore conditional i.i.d. structure and make the third term negligible under mild conditions Partition $\{1,\dots,n\}$ into K folds $I_1,\dots,I_K$. For each fold k, estimate nuisances $\widehat{\eta}^{(-k)}$ using only the training sample I_k^c, and evaluate on the held-out fold I_k. Define the cross-fitted AIPW estimator

$$\widehat{\tau}^{\mathrm{cf}} := \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k}\psi(W_i;\widehat{\eta}^{(-k)}). \tag{1.42}$$

Let $\mathbb{P}_{n,k} := |I_k|^{-1}\sum_{i\in I_k}$ be the fold-specific empirical measure and let $\Delta_k(W) := \psi(W;\widehat{\eta}^{(-k)}) - \psi(W;\eta_0)$. Then

$$\widehat{\tau}^{\mathrm{cf}} - \tau_0 = \sum_{k=1}^{K}\frac{|I_k|}{n}(\mathbb{P}_{n,k}-\mathbb{P})\psi(W;\eta_0) + \sum_{k=1}^{K}\frac{|I_k|}{n}\,\mathbb{P}[\Delta_k(W)] + \sum_{k=1}^{K}\frac{|I_k|}{n}(\mathbb{P}_{n,k}-\mathbb{P})\Delta_k(W). \tag{1.43}$$

The middle term is the population bias term and is handled by the same orthogonality/product-of-errors argument as before. The crucial point is the last term: conditional on the training data I_k^c, the function $\Delta_k(\cdot)$ is fixed, and the held-out observations $\{W_i : i \in I_k\}$ are i.i.d. and independent of $\widehat{\eta}^{(-k)}$. Therefore, conditional on I_k^c,

$$\mathbb{E}\big[(\mathbb{P}_{n,k}-\mathbb{P})\Delta_k(W) \mid I_k^c\big] = 0,$$
$$\mathrm{Var}\big((\mathbb{P}_{n,k}-\mathbb{P})\Delta_k(W) \mid I_k^c\big) = \frac{\mathrm{Var}(\Delta_k(W)\mid I_k^c)}{|I_k|}.$$

In particular, using $\mathrm{Var}(Z) \le \mathbb{E}[Z^2]$,

$$\mathbb{E}\big[((\mathbb{P}_{n,k}-\mathbb{P})\Delta_k(W))^2 \big| I_k^c\big] \le \frac{1}{|I_k|}\,\mathbb{E}\big[\Delta_k(W)^2 \mid I_k^c\big]. \tag{1.44}$$

Thus,

$$(\mathbb{P}_{n,k}-\mathbb{P})\Delta_k(W) = O_p\left(\frac{\|\Delta_k\|_{L^2(P)}}{\sqrt{|I_k|}}\right). \tag{1.45}$$

Consequently, if $\|\Delta_k\|_{L^2(P)} = o_p(1)$ (a mild stability/consistency requirement), then the last

term in (1.43) is $o_p(n^{-1/2})$ when $|I_k| \asymp n$. This is the key technical role of cross-fitting: it turns the difficult empirical-process term into an ordinary conditional LLN/CLT problem.

Quick summary of the key ingredients.

- *Neyman orthogonality.* The Neyman orthogonality condition (1.30) ensures that the population moment has no first-order sensitivity to nuisance errors, so the population bias is second order. For AIPW, this second-order term is exactly the product-of-errors expression (1.38), interpreted conditionally when $\widehat{\eta}$ is random (cf. Step 0 and (1.29)).

- *Cross-fitting.* Orthogonality alone does *not* control the empirical-process/overfitting term in (1.41) when $\widehat{\eta}$ is estimated on the same sample. Cross-fitting enforces foldwise independence between the fitted nuisances and the evaluation observations, yielding the variance bound (1.44) and the rate control (1.45) without imposing strong uniform complexity (Donsker) conditions on the first-stage estimators.

Bibliography

Abadie, A. and M. D. Cattaneo (2018). Econometric methods for program evaluation. *Annual Review of Economics 10*(1), 465–503.

Bickel, P. J., C. A. Klaassen, P. J. Bickel, Y. Ritov, J. Klaassen, J. A. Wellner, and Y. Ritov (1993). *Efficient and adaptive estimation for semiparametric models*, Volume 4. Springer.

Chamberlain, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica: Journal of the Econometric Society*, 567–596.

Ding, P. (2024). *A first course in causal inference*. Chapman and Hall/CRC.

Hirano, K., G. W. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*(4), 1161–1189.

Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association 47*(260), 663–685.

Leung, M. P. (2022). Causal inference under approximate neighborhood interference. *Econometrica 90*(1), 267–293.

Luenberger, D. G. (1997). *Optimization by vector space methods*. John Wiley & Sons.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association 89*(427), 846–866.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Wager, S. (2025). *Causal inference: A statistical learning approach*.