MS&E 226: Fundamentals of Data Science

Lecture 15: Causal inference from observational data

Ramesh Johari

From experiments to observational data

"Observational data" refers to settings where the treatment was not necessarily assigned at random; as we learned, this is a setting where we are vulnerable to selection bias (or confounding).

Typically, our data consists of:

- the outcome of interest Y;
- the treatment assignment indicator $W \in \{0, 1\}$, and
- features or covariates $\vec{X}$.

Informally, observational causal inference *relies* on $\vec{X}$ to help us mitigate or eliminate selection bias, and recover accurate estimates of the causal effect of treatment.

A motivating example

Example: Treatment and disease severity

Let's consider a stylized example:

- Y: Disease severity; larger Y means worse outcome
- $W \in \{0, 1\}$: Indicator for whether treatment is taken to reduce disease risk
- X_1: Age; uniformly distributed on $[20, 50]$
- $X_2 \in \{0, 1\}$: High blood pressure (HBP) indicator; $\mathbb{P}(X_2 = 1) = 0.1$

Data generating process ($n = 2000$):

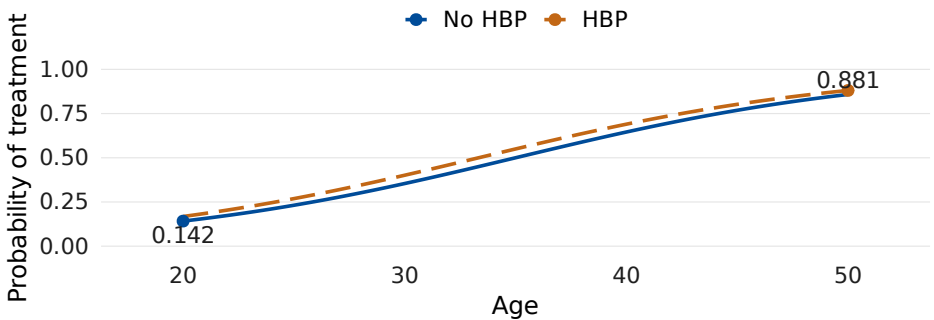
$$Y = 70 - 5W + X_1 + 10X_2 + \epsilon, \quad \epsilon \sim N(0, 4)$$

$$e(X) = \mathbb{P}(W = 1 \mid X_1, X_2) = \frac{\exp(0.12X_1 + 0.2X_2 - 4.2)}{1 + \exp(0.12X_1 + 0.2X_2 - 4.2)} \quad \text{(logistic)}$$

Older, HBP $\implies$ higher severity, and more likely to get treatment.
Note that ATE = β_W – the treatment effect is the same for all individuals, regardless of X_1, X_2.

The propensity score

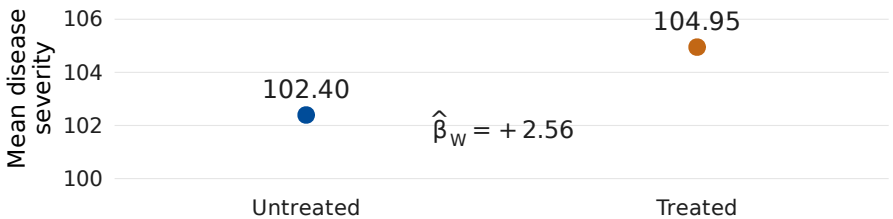
The probability of treatment given $\vec{X}$, $e(x) = \mathbb{P}(W = 1|\vec{X})$, is called the *propensity score*.



It varies from 0.142 (when age $X_1 = 20$ and HBP absent, $X_2 = 0$) to 0.881 (when age $X_1 = 50$ and HBP present, $X_2 = 1$).

Confounding

Suppose we ignore X_1, X_2, and just use the basic regression $Y \sim \hat{\beta}_0 + \hat{\beta}_W W$.



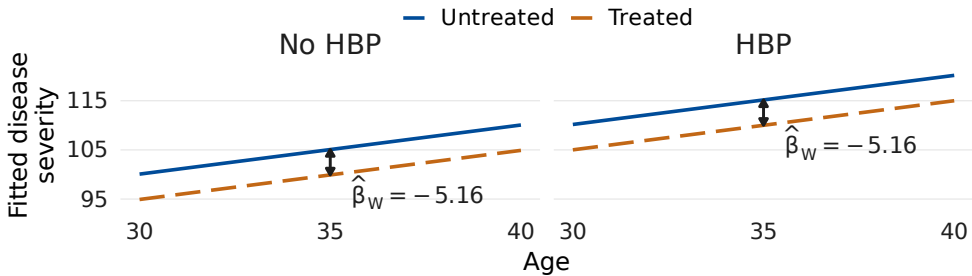
HC3 standard error for $\hat{\beta}_W$: 0.3805.

The naive estimate is +2.56, which has the opposite sign from the true ATE = −5.

Age and HBP both are confounders; in particular, older patients and those with HBP are more likely to be treated, but these patients have worse outcomes at baseline, masking the true benefit of treatment.

Controlling for observables

Now estimate: $Y \sim \hat{\beta}_0 + \hat{\beta}_W W + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$. HC3 SE: 0.102606.



By controlling for observables, we recover an (approximately) unbiased estimate (95% confidence interval $[-5.36, -4.96]$).

Why? Among patients of the *same age* with *same HBP* status, treatment assignment is *as if* it were randomly assigned.

The unconfoundedness assumption

This is a key assumption in all causal inference from observational data, referred to as *unconfoundedness*:

Conditional on $\vec{X}$, the treatment assignment W is independent of the potential outcomes $Y(0), Y(1)$.

(Also called: *ignorability*, *selection on observables*, *no hidden confounders*.)

The idea is that to the extent that there is any selection bias, i.e., any correlation between treatment and the potential outcomes, it is *only* because of the observed features.

The main limitation of observational data

Unconfoundedness makes observational causal inference possible, because it allows us to "pretend" assignment was random (given the features we observe).

But in general, unconfoundedness is an unverifiable claim!

In other words, because by definition hidden confounders are not observed, "you don't know what you don't know".

The rest of this lecture should be consumed with this warning in mind:
In observational causal inference, there is always a chance that you were misled by hidden confounding.

The goal is to reasonably argue that you have ruled out major sources of confounding through the features you observed.

Unconfoundedness and the "simple" regression

Returning to the "simple" OLS regression $Y \sim \hat{\beta}_0 + \hat{\beta}_W W + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$, when will $\hat{\beta}_W$ be a consistent estimate of the true ATE?

Unconfoundedness ensures the available (observable) control covariates completely resolve any potential confounding.

Under unconfoundedness, it can be shown that $\hat{\beta}_W$ consistently estimates the ATE as long as the model itself is correct: the true treatment effect is constant (regardless of covariates), and the true outcome model is linear in the covariates.

In our synthetic example, these assumptions held, which is why we obtain an accurate estimate of the true ATE.

Propensity scores

More covariates

Since we want unconfoundedness to hold, it's natural to want *as many covariates as possible* to help us remove selection bias.

What should we do if the set of features is very large, i.e., $\vec{X}$ has high dimension?

Last lecture we suggested the T-learner (T for "two"):

- Train separate machine learning models $\hat{\mu}_1(\vec{X})$ to predict Y for treated units ($W = 1$), and $\hat{\mu}_0(\vec{X})$ to predict Y for control units ($W = 0$).
- Compute $\widehat{\text{CATE}}(\mathbf{X}_i) = \hat{\mu}_1(\mathbf{X}_i) - \hat{\mu}_0(\mathbf{X}_i)$ for each unit i in the sample.
- Compute $\widehat{\text{ATE}} = (1/n) \sum_{i=1}^{n} \widehat{\text{CATE}}(\mathbf{X}_i)$.

But the T-learner can be problematic...

The T-learner is trying to estimate the potential outcomes *separately*, instead of the CATE *directly*.

ML models can have *bias* – e.g., models like lasso introduce bias to reduce variance when there are many features. We can't be sure the bias of the two models $\hat{\mu}_1$ and $\hat{\mu}_0$ will cancel out when we compute $\widehat{\text{CATE}}$, which means our eventual estimate $\widehat{\text{ATE}}$ may be biased.

Further, the resulting $\widehat{\text{ATE}}$ estimate can have high *variance* because we are learning two separate models (with many covariates each). These might individually have high variance, even if CATE($\vec{X}$) itself does not vary much across $\vec{X}$.

Finally, (in part because of potentially high variance), we don't have reliable ways to quantify uncertainty (standard errors and confidence intervals).

A key result: The propensity score suffices

Hypothetically, consider a collection of individuals with different features $\vec{X}$, with the *same* propensity score $p = e(\vec{X})$ (e.g., some are older without HBP, others are younger with HBP).

Some of these people with propensity score p will be treated, and some will not. But if *all* we know about them is their propensity score, we learn nothing about their features — regardless of their treatment assignment!

In other words: Among individuals with the same *propensity score*, treatment is independent of features.

A key result: The propensity score suffices

Under unconfoundedness, selection bias is *only* due to observable features.

Therefore, among individuals with the same propensity score, there is no further selection bias:

Under unconfoundedness, treatment is independent *of the potential outcomes, among individuals with the same propensity score* (i.e., W is independent of $Y(0), Y(1)$, given $e(\vec{X}) = p$).

In other words: It is *as if* for all individuals with the same propensity score p, we have run a randomized experiment with probability p of treatment assignment.

The IPW estimator

Using the propensity score

It is *as if* for all individuals with the same propensity score p, we have run a randomized experiment with probability p of treatment assignment.

Imagine we had a group of n individuals, *all* with known propensity score p; some are treated ($W_i = 1$), and some are not ($W_i = 0$).

How do we estimate the average treatment effect for this group?

Estimating ATE within a propensity score group

For a group with the same propensity score p, treatment is "as if" random.

If n total individuals in this group, then:

- For approximately np individuals, $W_i = 1$, and we observe $Y_i = Y_i(1)$.
- For approximately $n(1-p)$ individuals, $W_i = 0$, and we observe $Y_i = Y_i(0)$.

So to estimate $\mathbb{E}[Y(1)]$ and $\mathbb{E}[Y(0)]$:

$$\widehat{\mathbb{E}[Y(1)]} \approx \frac{1}{n} \sum_{i:W_i=1} \frac{Y_i}{p}$$

$$\widehat{\mathbb{E}[Y(0)]} \approx \frac{1}{n} \sum_{i:W_i=0} \frac{Y_i}{1-p}$$

In other words, we weight observations by *inverse propensity*.

The IPW estimator

In reality, individuals have *different* propensity scores $e(\vec{X}_i)$ based on their features.

The preceding discussion leads to the *inverse propensity weighting* (IPW) estimator, which involves two steps:

1. *Estimate propensity scores.* Fit a model $\hat{e}(\vec{X}_i)$ to *predict treatment from features* (e.g., logistic regression of W on $\vec{X}$.)

2. *Weight by inverse propensity.* Compute:

$$\widehat{\text{IPW}} = \frac{1}{n} \sum_{i:W_i=1} \frac{Y_i}{\hat{e}(\vec{X}_i)} - \frac{1}{n} \sum_{i:W_i=0} \frac{Y_i}{1-\hat{e}(\vec{X}_i)}$$

In our example, this yields $\widehat{\text{IPW}} = -4.717$ (recall that the true ATE $= -5$).

The overlap assumption

When the propensity score at $\vec{X}$ is too low or too high, we see *very few* individuals with features $\vec{X}$ in treatment or control, respectively.

Further, since we divide by $\hat{e}(\mathbf{X}_i)$ and $1 - \hat{e}(\mathbf{X}_i)$, the IPW estimator will be highly sensitive to observations that have extremely high or very low $\hat{e}(\mathbf{X}_i)$, leading to *high variance*.

Thus we impose the *overlap assumption*: propensities are *neither* too small or too large, i.e., in the population of interest, for all $\vec{X}$, there is a bound $\delta > 0$ such that:

$$\delta < e(\vec{X}) < 1 - \delta.$$

Diagnostic criteria [$*$]

In practice, since we only have access to *estimated* propensity scores, and we can't verify unconfoundedness, the following two diagnostic criteria are used to evaluate whether overlap is reasonable:

1. *Histogram overlap*: Plot histograms of $\hat{e}(\vec{X}_i)$ for treated and control groups. The distributions should overlap substantially. If they don't, we have poor overlap and IPW will be unreliable.
2. *Covariate balance*: After applying IPW weights, check that the *weighted* distributions of covariates are similar between treated and control groups. This verifies the reweighting is working — if successful, weighted groups should look similar in terms of features.

See R Markdown notebook for these diagnostics in our example.

Double robustness and the AIPW estimator

Estimators and challenges

We have seen two approaches to estimation of ATE in the presence of many covariates:

- The T-learner directly tries to model treatment and control outcomes respectively as $\hat{\mu}_1(\vec{X})$ and $\hat{\mu}_0(\vec{X})$, but the resulting estimate of ATE will be biased if the models are misspecified.
- The IPW estimator "reconstructs" treatment and control groups through inverse propensity weighting, via estimated propensity scores $\hat{e}(\vec{X})$; of course, it will also be biased if $\hat{e}(\vec{X})$ is misspecified.
- The T-learner can have high variance because it is learning two separate models; while the IPW estimator can have high variance if propensities are small.

Amazingly, we can *combine* them and end up with an estimator that has *low bias* and *low variance*!

The AIPW estimator

The *augmented inverse propensity weighting* (AIPW) estimator is:

$$\widehat{\text{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\Bigg[\hat{\mu}_1(\mathbf{X}_i) - \hat{\mu}_0(\mathbf{X}_i) + W_i\frac{Y_i - \hat{\mu}_1(\mathbf{X}_i)}{\hat{e}(\mathbf{X}_i)} - (1 - W_i)\frac{Y_i - \hat{\mu}_0(\mathbf{X}_i)}{1 - \hat{e}(\mathbf{X}_i)}\Bigg].$$

- *First term*: Outcome regression estimate $\hat{\mu}_1(\mathbf{X}_i) - \hat{\mu}_0(\mathbf{X}_i)$ (like T-learner)
- *Second term*: IPW-weighted *residuals* that correct for outcome model misspecification

The "augmentation" term uses IPW to fix errors in the outcome models.

The double robustness property

Why does the AIPW estimator perform so well?

It has a remarkable property known as *double robustness*:

- Suppose the outcome models $\hat{\mu}_0(\mathbf{X}_i)$ and $\hat{\mu}_1(\mathbf{X}_i)$ are nearly unbiased. Then the residuals will have *conditional mean near zero* $\implies$ *propensity scores* don't need to be unbiased.
- Suppose the propensity scores are nearly unbiased. Then the IPW terms will accurately "correct" any bias in the corresponding outcome model difference $\hat{\mu}_1(\mathbf{X}_i) - \hat{\mu}_0(\mathbf{X}_i)$ to obtain a nearly unbiased CATE at $\mathbf{X}_i$.

In other words, the AIPW estimator successfully *combines* both approaches to yield an estimator that is *robust* to misspecification of either of the two models (hence "double" robustness).

Overfitting

When using flexible ML methods for $\hat{e}(\mathbf{X})$, $\hat{\mu}_0(\mathbf{X})$, and $\hat{\mu}_1(\mathbf{X})$, we face a fundamental problem we have seen before:

Using the same data to both fit the models *AND* evaluate the final causal estimate can lead to overfitting bias.

Why is this a problem?

- When we use predictions from fitted models on the *same data* they were trained on, they may be *overoptimistic*.
- I.e., the propensity scores $\hat{e}(\mathbf{X}_i)$ and outcome predictions $\hat{\mu}_0(\mathbf{X}_i)$, $\hat{\mu}_1(\mathbf{X}_i)$ will be systematically biased.
- This bias propagates directly into our $\widehat{\text{ATE}}$ estimates.
- Further, any standard error estimates will be too low, and confidence intervals will be too narrow, due to overoptimism.

Cross-fitting via sample splitting

Cross-fitting via sample splitting implements the ML principle of separating training and evaluation data:

1. *Split data* into K folds (e.g., $K = 2$ or $K = 5$).
2. *For each fold* $k = 1, \ldots, K$:
   (i) Fit propensity model $\hat{e}(\mathbf{X})$ and outcome models $\hat{\mu}_0(\mathbf{X}), \hat{\mu}_1(\mathbf{X})$ on *all other folds* (folds $\neq k$)
   (ii) Use the resulting trained models to make predictions *on fold* k *only*
   (iii) Compute AIPW summand $\hat{\Gamma}(i)$ for each unit i in fold k, using the predictions from step (ii)
3. *Average* the fold-specific estimates: $\widehat{\text{AIPW}}_{\text{CF}} = \frac{1}{n} \sum_{i=1}^{n} \hat{\Gamma}(i)$

Cross-fitting via sample splitting: Notes

- When *fitting* models in step (i), *all* steps needed to yield a fitted model must be carried out, including parameter tuning (e.g., using cross-validation). During this process, *only* data from folds $\neq k$ should be used.
- In step (iii), $\hat{\Gamma}(i)$ is the summand from the definition of $\widehat{\text{AIPW}}$:

$$\hat{\Gamma}(i) = \hat{\mu}_1(\mathbf{X}_i) - \hat{\mu}_0(\mathbf{X}_i) + \frac{W_i(Y_i - \hat{\mu}_1(\mathbf{X}_i))}{\hat{e}(\mathbf{X}_i)} - \frac{(1 - W_i)(Y_i - \hat{\mu}_0(\mathbf{X}_i))}{1 - \hat{e}(\mathbf{X}_i)},$$

where the models in the previous equation are computed using all the data *except* the fold k that contains unit i.

Standard errors and confidence intervals

AIPW also allows for a very simple variance estimator:

$$\widehat{SE}^2_{\text{AIPW}} = \frac{1}{n^2} \sum_{i=1}^{n} (\hat{\Gamma}(i) - \widehat{\text{AIPW}})^2.$$

It can be shown that as long as the outcome model and propensity model are "reasonable", then a central limit theorem holds, i.e., for large n, $\widehat{\text{AIPW}}$:

- is asymptotically normal;
- centered at a mean that is the true ATE; and
- has standard error approximately $\widehat{\text{SE}}_{\text{AIPW}}$ estimated as above.

(See appendix.)

These facts mean we can build confidence intervals and run hypothesis tests on $\widehat{\text{AIPW}}$, exactly as discussed earlier in the course.

AIPW for randomized experiments

Note that in a randomized experiment, we know the treatment probability *by design*. This fact makes it easy to use AIPW for experiments:

If treatment was assigned with probability q to all units, we simply use $\hat{e}(\vec{X}_i) = q$ for all i in the AIPW estimator.

Why use AIPW for experiments?

- We can leverage flexible modern ML models for outcome prediction $\hat{\mu}_0(\vec{X})$ and $\hat{\mu}_1(\vec{X})$ to capture complex relationships.
- *Double robustness* ensures *consistent* estimation (vanishing bias) of ATE even if outcome models are misspecified (since propensities are known exactly).
- Further, AIPW addresses the drawbacks of the T-learner to yield *low variance* estimation; and allows us to estimate standard errors and confidence intervals for inference.

Simulation example

Comparing IPW and T-learner to AIPW

We simulate a high-dimensional confounded setting to compare finite-sample performance:

Data generating process:

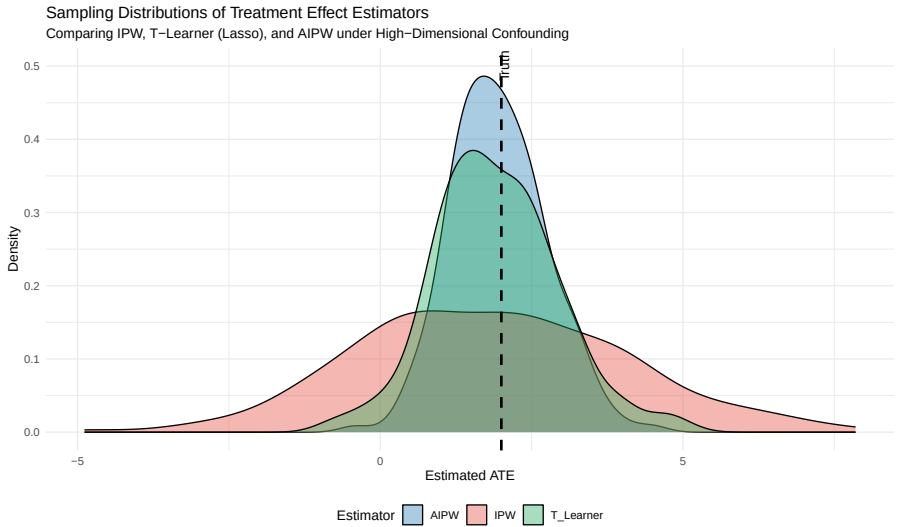
- $n = 300$ observations, $p = 300$ covariates, $s = 20$ active features
- First s covariates affect *both* propensity and outcome (confounding)
- Outcome coefficients: $\beta_y[1:s] \sim \text{Unif}(1,3) \times \{\pm 1\}$
- Propensity coefficients: $\beta_w[1:s] \sim \text{Unif}(0.5,1.5) \times \{\pm 1\}$
- True ATE $= 2.0$, noise $\sigma = 1.5$

Comparing IPW and T-learner to AIPW

Estimation: For each of 200 simulated datasets, we estimate ATE using:

- T-Learner with separate Lasso outcome models
- IPW with sparse logistic regression propensity scores (lasso)
- AIPW combining both, with 5-fold cross-fitting

Simulation results: AIPW dominates



Summary

Key points

Observational causal inference requires *unconfoundedness* given your features – an assumption you can't verify, except through qualitative argument and domain knowledge.

But if unconfoundedness and overlap hold, AIPW is a remarkably powerful tool to estimate causal effects:

- Estimate outcomes and propensities using modern ML techniques.
- Combine them into a *low bias, low variance* estimator of ATE.
- Obtain valid standard errors and confidence intervals, to enable inference.

Appendix: Asymptotic normality of $\widehat{\text{AIPW}}$ [$*$]

Preliminaries [$*$]

Let $\widehat{\text{AIPW}}_{\text{CF}}$ denote the cross-fitted AIPW estimator, and let $\hat{e}(\vec{X})$, $\hat{\mu}_0(\vec{X})$, and $\hat{\mu}_1(\vec{X})$ denote the estimated propensity and outcome models.

We will need some notation:

- For a function $h(\vec{X})$, we write $\|h\|_{P,2} = \sqrt{\mathbb{E}[h(\vec{X})^2]}$, where the expectation is over the distribution of covariates. (This is called the L_2 *norm* of the function, with respect to the covariate distribution.)
- A sequence of random variables Z_n is $o_p(1/\sqrt{n})$ if $Z_n\sqrt{n} \to 0$ in probability as $n \to \infty$. In other words, $Z_n \to 0$ "faster" than $1/\sqrt{n}$.

The central limit theorem [$*$]

Theorem. Suppose observations are i.i.d., unconfoundedness and overlap hold. In addition, suppose the following conditions hold:

1. *Bounded moments*: $\mathbb{E}[Y^2(0)]$ and $\mathbb{E}[Y^2(1)]$ are finite.
2. *Estimated propensities*: $\delta < \hat{e}(\vec{X}) < 1 - \delta$ for a fixed $\delta > 0$.
3. *Convergence rates*: All three models are consistent in L_2. For both $w = 0$ and $w = 1$, the *product* of the RMSEs of the outcome models and the propensity model is $o_p(1/\sqrt{n})$:

$$\|\hat{\mu}_w - \mu_w\|_{P,2} \times \|\hat{e} - e\|_{P,2} = o_p(n^{-1/2})$$

Then $\widehat{\text{AIPW}}_{\text{CF}}$ is asymptotically normal: $\sqrt{n}(\widehat{\text{AIPW}}_{\text{CF}} - \text{ATE}) \xrightarrow{d} N(0, \sigma^2_{\text{AIPW}})$.

Remarks on the theorem [$*$]

- The models $\hat{\mu}_0$, $\hat{\mu}_1$, and $\hat{e}$ are random, because they depend on the realization of the training data; the convergence in probability is over this randomness.
- Consistency (i.e., convergence to the true ATE) only requires *one* of the two models to be correct (double robustness); the preceding CLT requires *both* to converge fast enough.
- However, the key gain in this result is that *neither* outcome nor propensity models need to converge at the $1/\sqrt{n}$ parametric rate, to still obtain a valid CLT.
- The theorem assumes cross-fitting is used. This ensures the models $\hat{\mu}_0, \hat{\mu}_1, \hat{e}$ are independent of the data used to estimate $\hat{\Gamma}(i)$, avoiding "overfitting bias."
- *Variance estimation*: The sample variance estimator is consistent:

$$\frac{1}{n}\sum^{n}(\hat{\Gamma}(i) - \widehat{\mathsf{AIPW}}_{\mathsf{CF}})^2 \xrightarrow{p} \sigma^2_{\mathsf{AIPW}}.$$