Doubly robust inference and cross-fitting

1 Consistency does not yet give a standard error

Prerequisite: Doubly robust estimation. The previous note asked whether the estimate approaches the right effect as the sample grows. This note asks a further question: **how much does the estimate vary across samples, and how can we put a confidence interval around it?** We use n for the number of observations throughout.

1.1 Getting close to the truth is only the first requirement

Write τ for the true average treatment effect on the treated (ATT), and $\hat{\tau}$ for our estimate. **Consistency** means that the estimation error $\hat{\tau} - \tau$ tends to zero in probability: with more data, the chance of an error larger than any fixed tolerance tends to zero.

A confidence interval asks for more. Its width also shrinks with the sample size. An error can therefore tend to zero and still be large compared with the interval we report. To justify an interval, we must understand the estimation error **relative to the shrinking standard error**, not just in absolute terms.

For example, suppose ordinary sampling variation gives a standard error of $1/\sqrt{n}$. An additional error of $1/\sqrt{n}$ goes to zero, but it remains one full standard error at every sample size. An additional error of $1/n$ is different: measured in standard-error units, it is $1/\sqrt{n}$, which disappears. The first extra error can change confidence-interval coverage; the second becomes negligible for that purpose.

1.2 Why is $1/\sqrt{n}$ the scale we compare with?

Start with an ordinary average of independent observations. If each observation has variance σ^2, the average of n observations has variance σ^2/n and standard deviation $\sigma/\sqrt{n}$. This is why we compare errors with $n^{-1/2} = 1/\sqrt{n}$. Under the usual finite-variance assumptions, the central limit theorem (CLT) also supplies a normal approximation for this average.[1]

Our treatment-effect estimator is more complicated than a simple average. Section 2 will show that, if the propensity and untreated-outcome regression were known, its leading error would nevertheless be an average of mean-zero contributions. These two ingredients are the nuisance functions we otherwise have to estimate. The known-function estimator has this same $n^{-1/2}$ scale under the stated assumptions.

[1] Let $W_1, \ldots, W_n$ be independent observations with the same distribution, and let f be a fixed function with $m = \mathbb{E}[f(W)]$ and $0 < \sigma^2 = \mathrm{Var}(f(W)) < \infty$. Write $\bar{f}_n = n^{-1} \sum_{i=1}^{n} f(W_i)$. Independence gives $\mathrm{Var}(\bar{f}_n - m) = n^{-2} \sum_{i=1}^{n} \sigma^2 = \sigma^2/n$. The **central limit theorem** states that $\sqrt{n}(\bar{f}_n - m) \xrightarrow{d} N(0, \sigma^2)$, where $\xrightarrow{d}$ means convergence in distribution. Multiplication by $\sqrt{n}$ magnifies the shrinking error until it has a stable limiting distribution. These conclusions require the stated sampling and variance assumptions; they are not claims about every possible average.

1.3 What must the error from fitting the functions satisfy?

To separate the two sources of error, call the hypothetical estimate using the true nuisance functions $\tilde{\tau}$. We can write the following exact identity before working out any formulas for the estimator, as shown in Equation 1:

$$\hat{\tau} - \tau = \underbrace{\tilde{\tau} - \tau}_{\text{sampling error with true functions}} + \underbrace{\hat{\tau} - \tilde{\tau}}_{\text{extra error from fitted functions}}. \tag{1}$$

Let $\Delta_n = \hat{\tau} - \tilde{\tau}$ denote the extra error. Our aim is to show that it becomes negligible compared with $1/\sqrt{n}$. Equivalently,

$$\frac{\Delta_n}{n^{-1/2}} = \sqrt{n}\,\Delta_n \xrightarrow{p} 0. \tag{2}$$

Read this as: **measure the extra error in units of ordinary sampling error; the resulting ratio must go to zero.** This is stronger than $\Delta_n \xrightarrow{p} 0$. In probability it means that, for every tolerance $\eta > 0$,

$$\mathbb{P}\left(|\Delta_n| > \frac{\eta}{\sqrt{n}}\right) \longrightarrow 0. \tag{3}$$

Notice that the threshold on the right inside the probability is shrinking. We need the extra error to fall below any fixed fraction η of the $n^{-1/2}$ scale with probability tending to one.

Here are three deterministic examples to show the distinction. The same comparisons apply in probability when the error depends on the training data.

Extra error Δ_n	Does it tend to zero?	In sampling-error units: $\sqrt{n}\Delta_n$	Negligible on this scale?
$n^{-1/3}$	Yes	$n^{1/6} \to \infty$	No
$cn^{-1/2}$, fixed $c \neq 0$	Yes	c	No
n^{-1}	Yes	$n^{-1/2} \to 0$	Yes

The reason this comparison gives us the desired inference is visible after multiplying the error decomposition by $\sqrt{n}$:

$$\sqrt{n}(\hat{\tau} - \tau) = \sqrt{n}(\tilde{\tau} - \tau) + \sqrt{n}\Delta_n. \tag{4}$$

If the first term approaches a centered normal distribution and the second goes to zero, their sum has the same limiting distribution. If instead $\sqrt{n}\Delta_n \to c \neq 0$, the normal limit is shifted by c. A centered normal confidence interval would then use the wrong center. A random extra term that does not disappear can also change the limiting variability.

1.4 The shorthand we will use for these comparisons

The statement $\Delta_n = o_p(n^{-1/2})$, read "little-o in probability of $n^{-1/2}$," means exactly $\sqrt{n}\Delta_n \xrightarrow{p} 0$. The subscript p allows for random errors. For a nonrandom sequence, $o(n^{-1/2})$ means the corresponding ordinary numerical limit is zero.

More generally, for a positive scale c_n, $R_n = o_p(c_n)$ means $R_n/c_n \xrightarrow{p} 0$. Thus $o_p(1)$ means simply "goes to zero in probability." We use $O_p(c_n)$, with a capital O, when R_n/c_n stays bounded in probability. That is, we can make the chance that $|R_n/c_n|$ exceeds a sufficiently large fixed bound as small as desired, for all sufficiently large n. The ratio need not approach zero. In particular, $O_p(n^{-1/2})$ permits an error on the same scale as the standard error; $o_p(n^{-1/2})$ requires an error negligible relative to that scale.

The target is not to make the estimator's entire error smaller than $n^{-1/2}$. Ordinary sampling variation remains. We want the *additional* error from fitting the nuisance functions to be smaller. Nor will we require each nuisance function itself to be estimated at that speed: their errors enter the population bias as a product.

1.5 Recall the estimator's ingredients

Keep $W = (Y, D, X)$ for an observation's outcome, treatment indicator, and covariates. Write $\pi = \mathbb{E}[D] > 0$ for the population treated share, $p_0(X) = \mathbb{P}(D = 1 \mid X)$ for the true propensity, and $\mu_0(X) = \mathbb{E}[Y \mid D = 0, X]$ for the true untreated-outcome regression. These last two are the **nuisance functions**: ingredients needed to estimate τ, rather than the effect we want to report. The causal assumptions and overlap from the previous note remain in force. Define

$$A_p(W) = D - (1 - D)\frac{p(X)}{1 - p(X)}, \qquad B(W; \mu, p) = A_p(W)(Y - \mu(X)). \tag{5}$$

The previous note showed $\mathbb{E}B(W; \mu_0, p_0) = \pi\tau$. Its Section 5, "When both functions are wrong: derive the bias", also derived the **product identity**. Since its population effect formula is $T(\mu, p) = \mathbb{E}B(W; \mu, p)/\pi$, multiplying its bias formula by π gives

$$\mathbb{E}B(W; \mu, p) - \pi\tau = \mathbb{E}\left[\frac{p_0 - p}{1 - p}(\mu_0 - \mu)\right]. \tag{6}$$

The right side of Equation 6 multiplies the propensity error by the regression error, with weight $1/(1 - p)$. If either function is correct, this population bias is zero.

For example, the rule $f(W) = D$ returns the treatment indicator. It is a **fixed function**: it is not fitted to the sample and does not change with sample size. Its sample average, the treated share, is still random because the observations are random. Likewise, $W \mapsto B(W; \mu_0, p_0)$ is fixed: its ingredients are population functions, even though we do not know them. The CLT can be applied to an appropriate centered version of this function.

1.6 The two pieces of extra error we need to control

There are two separate tasks:

1. Bound the change in the **population average** when the fitted functions replace the truth.
2. Bound the extra **sample fluctuation** caused by evaluating those fitted functions on data.

For the first task, use the product identity recalled above from Section 5 of the previous note: the population change is $\mathbb{E}[(p_0 - p)(\mu_0 - \mu)/(1 - p)]$. Section 4 bounds this product when the fitted functions replace μ, p. Cross-fitting, which evaluates each observation using functions trained on other observations, helps with the second task. Neither step removes the need for overlap, moment assumptions, or adequate fits.

The conclusions in Section 4–Section 8 assume **both** fitted functions converge to their truths. Consistency when just one model is correct is a different result. Section 9 explains why its standard error needs additional care.

2 Use an estimating equation that includes the treated count

Our aim in this section is to find the ordinary sampling error that remains **even if we know both nuisance functions**. We will express that error as an average of contributions from individual observations. This will tell us which quantity to put into the CLT and, eventually, how to calculate a standard error. We first rewrite the estimator so that fluctuations in the treated count are included. Later sections ask when fitting the nuisance functions leaves this leading error unchanged.

2.1 Start with the notation for an average

We observe n independent draws $W_i = (Y_i, D_i, X_i)$ from the same population. The symbol $\mathbb{P}_n$ is an **averaging operator**: it tells us to evaluate a function at every observation and average the resulting numbers. For any function f,

$$\mathbb{P}_n f := \frac{1}{n}\sum_{i=1}^{n} f(W_i). \tag{7}$$

Thus $\mathbb{P}_n f$ is a sample average, whereas $\mathbb{E}[f(W)]$ is a population average. The letter f is a placeholder for the rule being averaged. Two examples are

$$\mathbb{P}_n D = \frac{1}{n}\sum_{i=1}^{n} D_i = \hat{\pi}, \qquad \mathbb{P}_n B(W;\mu,p) = \frac{1}{n}\sum_{i=1}^{n} B(W_i;\mu,p). \tag{8}$$

In the second expression, μ, p specify which function of W we average; the semicolon in $B(W;\mu,p)$ separates the observation from those choices.

The estimator from Section 7 of the previous note is

$$\hat{\tau} = \frac{\sum_{i=1}^{n} B(W_i;\hat{\mu},\hat{p})}{\sum_{i=1}^{n} D_i} = \frac{\mathbb{P}_n B(W;\hat{\mu},\hat{p})}{\hat{\pi}}. \tag{9}$$

Dividing numerator and denominator by n gives the last expression. We work with samples containing at least one treated observation, so $\hat{\pi} > 0$. Both numerator and denominator vary from sample to sample.

2.2 Rewrite the ratio as an equation

Let t denote a candidate value for the treatment effect; τ still denotes the true effect. Define a new function

$$h(W;t,\mu,p) := B(W;\mu,p) - Dt. \tag{10}$$

For one observation, h takes its numerator contribution B and subtracts t if that observation is treated. Averaging gives

$$\mathbb{P}_n h(W;t,\mu,p) = \mathbb{P}_n B(W;\mu,p) - t\hat{\pi}. \tag{11}$$

At the fitted functions, setting this average equal to zero and solving for t recovers exactly $\hat{\tau}$, as Equation 12 shows:

$$\mathbb{P}_n h(W;\hat{\tau},\hat{\mu},\hat{p}) = 0 \quad \Longleftrightarrow \quad \mathbb{P}_n B(W;\hat{\mu},\hat{p}) = \hat{\tau}\hat{\pi}. \tag{12}$$

This is an **estimating equation**: an equation whose solution is the estimate. Its population counterpart is zero at the truth because

$$\mathbb{E}[h(W;\tau,\mu_0,p_0)] = \mathbb{E}[B(W;\mu_0,p_0)] - \mathbb{E}[D]\tau = \pi\tau - \pi\tau = 0. \tag{13}$$

The benefit of this rewriting is that B and the treated indicator D now appear together in a single observation's contribution.

2.3 First imagine the nuisance functions are known

We now do a thought experiment: suppose someone gives us the exact propensity and untreated-outcome functions. We would still get a different effect estimate from each random sample, because the observed people, outcomes, and treated count vary. We will work out that remaining sampling error first. It gives us a benchmark: later we ask how accurately we must learn the two functions to retain the same leading error and standard error.

For this calculation, use the true μ_0, p_0 while still estimating the treated share from the sample. Write

$$B_0(W) := B(W; \mu_0, p_0), \qquad B_{0i} := B_0(W_i), \qquad \tilde{\tau} := \frac{\mathbb{P}_n B_0}{\hat{\pi}}. \tag{14}$$

The tilde distinguishes this hypothetical estimator from the feasible $\hat{\tau}$, which uses fitted functions. We cannot compute $\tilde{\tau}$ in practice; it provides the benchmark sampling error we want to understand.

Subtract τ using a common denominator and substitute $\hat{\pi} = \mathbb{P}_n D$:

$$\tilde{\tau} - \tau = \frac{\mathbb{P}_n B_0 - \hat{\pi}\tau}{\hat{\pi}} = \frac{\mathbb{P}_n (B_0 - D\tau)}{\hat{\pi}}. \tag{15}$$

This is an exact identity. To apply a CLT to its numerator, we must explain what is being averaged and why its population mean is zero.

2.4 What are the two contributions, and why do they have mean zero?

Define the true treatment odds as $o_0(X) = p_0(X)/(1 - p_0(X))$. Expanding B_0 splits each observation's numerator contribution into

$$B_0(W) - D\tau = \underbrace{D\{Y - \mu_0(X) - \tau\}}_{\text{treated contribution}} - \underbrace{(1-D)o_0(X)\{Y - \mu_0(X)\}}_{\text{control correction}}. \tag{16}$$

For a **treated observation**, $Y - \mu_0(X)$ compares the observed outcome with the predicted untreated outcome. Its average among treated people equals the ATT under our identifying assumptions; it is not necessarily that person's individual treatment effect. Subtracting τ centers this comparison around its treated-population mean. Therefore

$$\begin{aligned} \mathbb{E}[D\{Y - \mu_0(X) - \tau\}] &= \pi\{\mathbb{E}[Y - \mu_0(X) \mid D = 1] - \tau\} \\ &= \pi(\tau - \tau) = 0. \end{aligned} \tag{17}$$

For a **control observation**, $Y - \mu_0(X)$ is the untreated-outcome prediction error. By the definition of μ_0, its mean conditional on $X, D = 0$ is zero. Multiplying by odds that depend only on X preserves this zero conditional mean:

$$\begin{aligned} \mathbb{E}[(1-D)o_0(X)\{Y - \mu_0(X)\} \mid X] &= (1 - p_0(X))o_0(X)\mathbb{E}[Y - \mu_0(X) \mid X, D = 0] \\ &= 0. \end{aligned} \tag{18}$$

Averaging over X gives zero unconditionally too. These two centered contributions have mean zero separately. In the shorter expression $B_0 - D\tau$, however, B_0 and $D\tau$ each have mean $\pi\tau$; it is their difference that has mean zero. None of these population statements requires a realized sample average to be exactly zero.

2.5 Which part of the error survives on the CLT scale?

For readability, call this mean-zero function $h_0(W) := h(W; \tau, \mu_0, p_0) = B_0(W) - D\tau$. Assume $0 < \mathbb{E}[h_0(W)^2] < \infty$. The variance calculation and CLT in the footnote give $\mathbb{P}_n h_0 = O_p(n^{-1/2})$. Similarly, because D is a bounded indicator, $\hat{\pi} - \pi = O_p(n^{-1/2})$ and $\hat{\pi} \xrightarrow{p} \pi > 0$.

Now add and subtract the expression with the population denominator π:

$$\begin{aligned} \tilde{\tau} - \tau &= \frac{\mathbb{P}_n h_0}{\hat{\pi}} \\ &= \underbrace{\frac{\mathbb{P}_n h_0}{\pi}}_{\text{leading average}} + \underbrace{\left(\frac{1}{\hat{\pi}} - \frac{1}{\pi}\right)\mathbb{P}_n h_0}_{\text{remainder}}. \end{aligned} \tag{19}$$

The leading average is of order $n^{-1/2}$. For the remainder, note that

$$\frac{1}{\hat{\pi}} - \frac{1}{\pi} = -\frac{\hat{\pi} - \pi}{\pi\hat{\pi}} = O_p(n^{-1/2}), \tag{20}$$

since the denominator converges to $\pi^2 > 0$. The remainder is consequently a product of two errors of order $n^{-1/2}$, hence is $O_p(n^{-1})$. Multiplying it by $\sqrt{n}$ gives $O_p(n^{-1/2})$, which converges to zero. No independence between these two errors is needed for this rate calculation.

This explains the terminology: the **first-order error** is the leading $n^{-1/2}$ average, which remains visible after multiplication by $\sqrt{n}$. The **second-order remainder** here is a product of two sampling errors and disappears on that scale. Define

$$\phi_0(W) = \frac{h_0(W)}{\pi} = \frac{B_0(W) - D\tau}{\pi} = \frac{D(Y - \mu_0(X) - \tau) - (1 - D)o_0(X)(Y - \mu_0(X))}{\pi}. \tag{21}$$

Then the leading-error expansion in Equation 22 reads

$$\tilde{\tau} - \tau = \frac{1}{n}\sum_{i=1}^{n} \phi_0(W_i) + O_p(n^{-1}), \qquad \sqrt{n}(\tilde{\tau} - \tau) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \phi_0(W_i) + o_p(1). \tag{22}$$

This is why ϕ_0 is called the **influence function**: observation i contributes $\phi_0(W_i)/n$ to the leading approximation of the estimation error. The CLT applies to this average of a fixed, mean-zero function. Its variance $\mathbb{E}[\phi_0(W)^2]/n$ is the basis for the standard error.

2.6 Does this mean estimating the treated share is only second order?

No. The treated share already contributes to the first-order error through $-D\tau$ inside h_0. To make this visible, rewrite the leading average as

$$\mathbb{P}_n\phi_0 = \frac{\mathbb{P}_n B_0 - \pi\tau}{\pi} - \frac{\tau}{\pi}(\hat{\pi} - \pi). \tag{23}$$

The first term is the centered numerator fluctuation. The second is the correction for the fluctuating treated share; it generally has the same $n^{-1/2}$ order. Only the further change from dividing the **already centered** $\mathbb{P}_n h_0$ by $\hat{\pi}$ instead of π is second order.

For example, if every observation's numerator were exactly $B_0 = D\tau$, then $\tilde{\tau} = \tau\hat{\pi}/\hat{\pi} = \tau$ in every sample with treated observations. The two fluctuations above would cancel exactly, and $\phi_0 = 0$. This limiting illustration shows why we must keep numerator and treated-count fluctuations together.

Using $B_0/\pi - \tau$ as the influence function would instead describe an estimator that divides by the known population share. Our estimator divides by the sample share, which is why its influence function contains $-D\tau/\pi$. The equation $h = B - Dt$ keeps that distinction explicit. The next sections study what changes when we also have to estimate μ_0, p_0.

3 Orthogonality: why small function errors enter at second order

Section 2 found the sampling error when both functions are known. In practice we insert estimates, so we now ask: **if the functions we insert are slightly wrong, how much does the population estimating equation change?** At the true effect τ, its population average should be zero. A change away from zero means that the equation would point us toward a different effect even with an unlimited sample.

Here we study that population change, before considering random sample fluctuations. Keep τ and the population distribution of (Y, D, X) unchanged; vary only the functions inserted into h. We will see that the two function errors have to multiply before this population average changes.

3.1 Describe a particular pattern of errors, then shrink it

A function can be wrong by different amounts at different values of X. To describe one possible pattern, choose two functions $a(x)$ and $b(x)$: $a(x)$ specifies the propensity error at covariate value x, and $b(x)$ specifies the regression error there. Introduce a number r that scales both errors up or down:

$$p_r(x) = p_0(x) + ra(x), \qquad \mu_r(x) = \mu_0(x) + rb(x). \tag{24}$$

At $r = 0$ we use the truth; at $r = 1$ we add the chosen errors; at $r = 1/2$ we add half as much error at every x. The functions a, b are called **directions** because they specify how we move away from the true functions. **Fixed** means we choose this pattern once and hold it unchanged as r varies. It does not mean the errors must be constant across people, or that the functions are fitted to a sample.

For example, use the previous note's three covariate values $x = 0, 1, 2$, with $p_0 = (0.25, 0.50, 0.75)$ and $\mu_0 = (0, 1, 4)$. Choose the following pattern:

x	Propensity direction $a(x)$	Regression direction $b(x)$	$p_r(x)$ at $r = 0.1$	$\mu_r(x)$ at $r = 0.1$
0	0.1	1	0.26	0.1
1	0	−1	0.50	0.9
2	−0.1	2	0.74	4.2

At $r = 0.05$, each error in the last two columns would be halved. The actual treatment probabilities and outcome means in the population stay at p_0 and μ_0; only the values we put into our estimating equation change. This is a sensitivity calculation, not a recipe for fitting either function.

3.2 Why is there no term proportional to the error size?

Recall that $h = B - D\tau$, so its population mean at the true effect is exactly the numerator bias in the product identity. Along our chosen pattern, $p_0 - p_r = -ra$ and $\mu_0 - \mu_r = -rb$. Substituting gives

$$\begin{aligned} \mathbb{E}h(W; \tau, \mu_r, p_r) &= \mathbb{E}\left[\frac{p_0 - p_r}{1 - p_r}(\mu_0 - \mu_r)\right] \\ &= \mathbb{E}\left[\frac{(-ra)(-rb)}{1 - p_r}\right] \\ &= r^2\mathbb{E}\left[\frac{ab}{1 - p_r}\right]. \end{aligned} \tag{25}$$

The two minus signs cancel, and each error supplies one factor of r. That is where r^2 comes from. Provided the denominator stays away from zero, the change is bounded by a constant times r^2, rather than $|r|$. Halving small errors in both functions therefore makes their population effect roughly one quarter as large; the denominator also changes slightly.

To make this precise, consider any fixed directions for which p_r stays a valid probability with $1 - p_r \geq \epsilon > 0$ near $r = 0$, and $\mathbb{E}[|a(X)b(X)|] < \infty$. A derivative at zero asks for the change in the population average **per unit of** r as r tends to zero. Since the average at $r = 0$ is zero, we can bound that ratio by

$$\left| \frac{\mathbb{E}h(W; \tau, \mu_r, p_r) - \mathbb{E}h(W; \tau, \mu_0, p_0)}{r} \right| \leq \frac{|r|}{\epsilon} \mathbb{E}|ab| \longrightarrow 0. \tag{26}$$

This zero derivative in Equation 26 is **Neyman orthogonality**. Starting at the true functions, the population equation has no first-order response to any such pattern of small function errors. Here "second order" refers to the squared error scale r^2; Section 4 will connect that scale to sample size.

3.3 What would happen without the control correction?

For comparison, using only the outcome regression gives the equation $h_{\text{OR}} = D(Y - \mu - \tau)$. Replacing μ_0 by $\mu_0 + rb$ subtracts rDb from each contribution, so

$$\mathbb{E}h_{\text{OR}}(W; \tau, \mu_0 + rb) = -r\mathbb{E}[Db] = -r\mathbb{E}[p_0 b]. \tag{27}$$

For instance, if $b(x) = 1$ everywhere, every untreated-outcome prediction is too high by r, and the population equation changes by $-r\pi$. This is proportional to r: the regression error passes through directly. In AIPW, the control correction cancels this change when the propensity is correct. If the propensity is also wrong, the remaining population change involves the product of the two errors, as we just calculated.

This calculation explains the link with double robustness. Setting either error identically to zero gives the exact global identity from the previous note. Moving both functions a small distance gives a second-order local change. Orthogonality by itself is not a claim that either function can be estimated arbitrarily badly, nor does it control fitting and evaluating on the same data.

4 How accurate must the fitted functions be?

The previous section shrank a chosen pattern of errors using one number r. Actual fitted functions can make different errors at every covariate value, and the pattern can change with the training sample. We now need a way to summarize each fitted function's overall error by a single number. We can then ask how quickly those two numbers must shrink as we collect more data for their contribution to be negligible relative to $n^{-1/2}$ sampling error.

4.1 Turn errors across covariate values into one measure of accuracy

Imagine we have finished training $\hat{\mu}$ and keep that fitted function fixed. For a new person with covariates x, its error in estimating the true untreated-outcome mean is $\hat{\mu}(x) - \mu_0(x)$. Square that error so positive and negative mistakes cannot cancel. Average the squared errors across the population, weighting covariate values by how common they are. Finally take the square root to return to the units of the outcome. This gives

$$\|\hat{\mu} - \mu_0\|_2 = \left\{\mathbb{E}_X[(\hat{\mu}(X) - \mu_0(X))^2]\right\}^{1/2}. \tag{28}$$

This is a **root mean squared error in the fitted function**. It compares $\hat{\mu}$ with the true conditional mean μ_0, rather than with a noisy individual outcome Y. Even a perfect estimate of μ_0 need not predict every person's outcome exactly.

For example, suppose $X = 0, 1, 2$ each occurs with probability $1/3$, and the regression errors at these values are $0.1, -0.1, 0.2$, as in Section 3. Then

$$\|\hat{\mu} - \mu_0\|_2 = \sqrt{\tfrac{1}{3}(0.1)^2 + \tfrac{1}{3}(-0.1)^2 + \tfrac{1}{3}(0.2)^2} = \sqrt{0.02} \approx 0.141. \tag{29}$$

The same construction gives $\|\hat{p} - p_0\|_2$ for errors in the estimated treatment probabilities. The weights in both averages come from the whole population's distribution of X, which is the distribution in our bias identity.

The notation $\|f\|_2$ is called the L_2 **norm**. For any function f of the covariates, it abbreviates

$$\|f\|_2 = \{\mathbb{E}[f(X)^2]\}^{1/2}. \tag{30}$$

Here we use it with an error function, such as $f = \hat{\mu} - \mu_0$. A "fresh draw of X" means covariates from a new, independent person in the same population. The expectation averages over that person's possible covariates while holding the trained function fixed. A different training sample gives a different fitted function and hence a different norm; that is why the norm itself is random in the rate statements below.

This is a theoretical measure of accuracy: we cannot directly compute these errors from data because the true functions are unknown. Its purpose here is to state conditions under which the standard-error calculation is justified.

4.2 Use the two accuracy measures to bound the population bias

Why use this particular measure? The bias identity contains the **average product** of the two function errors. The Cauchy–Schwarz inequality says $\mathbb{E}|fg| \le \|f\|_2\|g\|_2$: that average product cannot exceed the product of their root mean squared sizes. No independence between the two errors is required.

First assume the fitted propensity stays below $1-\epsilon$, so the weight $1/(1-\hat{p})$ is at most $1/\epsilon$. Then take absolute values in the product identity and apply the inequality:

$$\begin{aligned} |\mathbb{E}B(W; \hat{\mu}, \hat{p}) - \pi\tau| &\le \frac{1}{\epsilon}\mathbb{E}|(p_0 - \hat{p})(\mu_0 - \hat{\mu})| \\ &\le \frac{1}{\epsilon}\|\hat{p} - p_0\|_2\|\hat{\mu} - \mu_0\|_2. \end{aligned} \tag{31}$$

Thus the product-rate condition for negligible population bias is Equation 32:

$$\|\hat{p} - p_0\|_2 \|\hat{\mu} - \mu_0\|_2 = o_p(n^{-1/2}). \tag{32}$$

Dividing by the fixed positive π gives the same requirement for ATT bias.

4.3 Read the rate condition carefully

If the two errors are $O_p(n^{-a})$ and $O_p(n^{-b})$, their product is $O_p(n^{-(a+b)})$. It is $o_p(n^{-1/2})$ when $a + b > 1/2$. For example, $a = 1/3$ and $b = 1/5$ give $a + b = 8/15 > 1/2$. Both errors being $o_p(n^{-1/4})$ also suffices.

Two errors merely of order $O_p(n^{-1/4})$ have product $O_p(n^{-1/2})$, which is not enough. Likewise the boundary pair $O_p(n^{-1/3})$, $O_p(n^{-1/6})$ needs a further little-o improvement. An estimator's label, such as forest or lasso, does not establish any of these rates; they depend on assumptions about the functions and covariates.

4.4 Conditions for the sample-fluctuation argument

We have bounded the population bias. To control the extra sample fluctuation as well, the remaining sections use the following sufficient conditions. Each has a particular job in the proof:

- **Independent observations from the same population.** This lets us calculate the variance of an average by adding individual variances. The number of folds K is fixed, and each fold contains a positive fraction of the sample.
- **A positive treated share and bounded weights.** We require $\pi > 0$ and both true and fitted propensities in $[0, 1 - \epsilon]$, for a fixed $\epsilon > 0$. These keep the denominators used in the argument away from zero.
- **Controlled outcome variation.** We assume $\mathbb{E}[D(Y - \mu_0)^2] < \infty$ for the treated contribution and, for controls,

  $$\mathbb{E}[(Y - \mu_0(X))^2 \mid X, D = 0] \leq C \tag{33}$$

  for a finite constant C. The latter condition prevents very large outcome variation at some covariate values from amplifying small propensity errors.
- **Fits that improve with more data.** In every fold, both L_2 errors tend to zero in probability, and their product is $o_p(n^{-1/2})$ as above. The first requirement is what we mean by L_2 **consistency**.

These conditions are sufficient, rather than the weakest possible assumptions. The next section explains what folds are and how to construct their fits.

5 Cross-fitting: fit elsewhere, evaluate here

Section 4 controlled the population bias of fitted functions. There is still a sampling problem: the same observations may influence the fitted functions and then be used to evaluate them. The resulting contributions are no longer evaluations of a fixed rule on independent new observations.

For example, a very flexible outcome regression can make the residual $Y_i - \hat{\mu}(X_i)$ tiny for the controls it was trained on. This need not mean it predicts equally well for other controls. If we use those training residuals in the control correction, the fitting procedure has already adapted to the outcomes entering the estimating equation.

Cross-fitting separates fitting from evaluation for each observation. We first describe how to do it, then identify exactly which part of the error this separation lets us control.

5.1 Start with a two-fold example

Randomly divide the observation indices into two roughly equal groups, called folds. Train on one group and evaluate on the other, then swap their roles:

Observations to evaluate	Data used to fit the propensity	Data used to fit the outcome regression
Fold 1	All observations in fold 2	Controls in fold 2
Fold 2	All observations in fold 1	Controls in fold 1

Each observation is evaluated once, using functions whose training did not include it. Across the two rounds, every observation also helps train the fit used for the other fold. We therefore obtain contributions for the whole sample, despite using separate observations for training and evaluation in each round. Any preprocessing or tuning of a fold's fit must also use only that fold's training data.

5.2 Write the same procedure for K folds

Split observation indices into K disjoint groups $I_1, \ldots, I_K$ using a fixed split or random assignment independent of the observations (Y, D, X). Write $n_k = |I_k|$ for the number in fold k, and $\alpha_k = n_k/n$ for its fraction of the full sample. For each fold k:

1. Fit $\hat{p}_k$ on all observations outside I_k.
2. Fit $\hat{\mu}_k$ on controls outside I_k.
3. For each $i \in I_k$, compute $\widehat{B}_i = B(W_i; \hat{\mu}_k, \hat{p}_k)$.

The subscript k labels the fold where the functions will be **evaluated**; they were trained outside that fold. The resulting $\widehat{B}_i$ is the fitted numerator contribution for observation i. Pool these contributions and divide by the full sample treated share:

$$\hat{\tau} = \frac{n^{-1} \sum_i \widehat{B}_i}{\hat{\pi}}, \qquad \hat{\pi} = \mathbb{P}_n D. \tag{34}$$

This is one ratio for the whole sample. Averaging separate treatment-effect ratios from the folds would generally give a different estimate, since their treated shares can differ.

5.3 Measure what fitting changed for one observation

Recall $B_0(W) = B(W; \mu_0, p_0)$, the contribution with true functions. Define

$$\delta_k(W) = B(W; \hat{\mu}_k, \hat{p}_k) - B_0(W). \tag{35}$$

The symbol δ_k means **the change in an observation's contribution caused by using the fitted functions**. This is sometimes called a score difference. It is zero if the fitted functions equal the true ones.

We need to distinguish two averages of this change:

$$P\delta_k := \mathbb{E}[\delta_k(W) \mid \text{training data for fold } k], \qquad P_k\delta_k := \frac{1}{n_k}\sum_{i\in I_k}\delta_k(W_i). \tag{36}$$

In $P\delta_k$, W is an independent new observation from the population: hold the trained functions fixed and average over new people. In $P_k\delta_k$, evaluate those same functions on the actual held-out people in fold k. We will use P and P_k for these averaging operations on other functions too. The quantity $P\delta_k$ can vary across training samples even though it is a population average for a given fit.

Add and subtract the population average:

$$\underbrace{P_k\delta_k}_{\text{observed average change}} = \underbrace{P\delta_k}_{\text{population change}} + \underbrace{(P_k - P)\delta_k}_{\text{fluctuation around that change}}. \tag{37}$$

Here $(P_k - P)\delta_k$ is shorthand for $P_k\delta_k - P\delta_k$. For example, the average change on a particular held-out fold could be 0.03 when the population change for that fit is 0.02. The remaining 0.01 comes from which people happened to enter the held-out fold.

5.4 Put these changes back into the estimator

The hypothetical estimator $\tilde{\tau}$ and the feasible $\hat{\tau}$ use the same treated share $\hat{\pi}$. Subtracting them therefore gives

$$\begin{aligned}\Delta_n := \hat{\tau} - \tilde{\tau} &= \frac{1}{\hat{\pi}}\frac{1}{n}\sum_k\sum_{i\in I_k}\delta_k(W_i)\\ &= \frac{1}{\hat{\pi}}\sum_k \alpha_k P_k\delta_k\\ &= \frac{1}{\hat{\pi}}\left\{\sum_k \alpha_k P\delta_k + \sum_k \alpha_k (P_k - P)\delta_k\right\}.\end{aligned} \tag{38}$$

The weights $\alpha_k = n_k/n$ turn averages within folds into an average over all observations. This is the promised split of the extra error into a population change and an additional sample fluctuation.

Recall the goal from Section 1: $\sqrt{n}\Delta_n \xrightarrow{p} 0$. Multiply the last identity by $\sqrt{n}$ and combine it with Section 2's exact expression for $\tilde{\tau} - \tau$:

$$\begin{aligned}\sqrt{n}(\hat{\tau} - \tau) = \frac{1}{\hat{\pi}}\Bigg\{&\frac{1}{\sqrt{n}}\sum_i (B_{0i} - D_i\tau)\\ &+ \underbrace{\sqrt{n}\sum_k \alpha_k (P_k - P)\delta_k}_{R_{\text{sample}}}\\ &+ \underbrace{\sqrt{n}\sum_k \alpha_k P\delta_k}_{R_{\text{bias}}}\Bigg\}.\end{aligned} \tag{39}$$

The first term is the ordinary sampling error we want to keep. The two terms labelled R_{sample} and R_{bias} are the errors we want to remove. **They already include the factor $\sqrt{n}$.** Showing that they are $o_p(1)$ is therefore the same as showing that their unscaled contributions are $o_p(n^{-1/2})$. Since $\hat{\pi} \xrightarrow{p} \pi > 0$, division by $\hat{\pi}$ does not change whether they vanish.

5.5 One remainder is already controlled; one remains

Section 4 gives $P\delta_k = o_p(n^{-1/2})$ for each fold, so $\sqrt{n}P\delta_k = o_p(1)$. There are only a fixed number of folds, and their weights sum to one. Their weighted sum, R_{bias}, therefore vanishes too. This is where the product rate for the two function errors enters.

What remains is R_{sample}. Section 6 uses the separation of training and evaluation to control it. Fits for different folds use overlapping data when $K > 2$, and even with two folds the fits and evaluations reuse the same observations in different roles. We will consequently work one fold at a time, without treating the fitted contributions across folds as independent.

6 Why the extra sample fluctuation disappears

Our task is to show that the held-out average of δ_k gets close enough to its population average, even after multiplication by $\sqrt{n}$. Two facts work together: averaging independent held-out observations reduces variation, and better nuisance fits make the differences δ_k themselves smaller.

We will first see why a small population mean square $P\delta_k^2$ is enough, then show that consistency of both nuisance functions makes it small. Here $P\delta_k^2$ means $P[\delta_k(W)^2]$, the average of the squared difference; it is not $(P\delta_k)^2$, the square of its average.

6.1 Hold the training data fixed to recover an ordinary average

Focus on one fold k. **Conditioning on its training data** means treating the observed training sample, and hence the functions fitted to it, as given. The observations in the held-out fold are still independent draws from the population. From this perspective, δ_k is just a fixed rule applied to new observations, exactly the situation where an average is easy to study. For a random split, also hold the fold assignment fixed throughout this argument.

Let v_k be the variance of $\delta_k(W)$ for a new observation, keeping the training data fixed. Then

$$\mathbb{E}[(P_k - P)\delta_k \mid \text{training}] = 0, \qquad \text{Var}((P_k - P)\delta_k \mid \text{training}) = \frac{v_k}{n_k}. \tag{40}$$

The centering by $P\delta_k$ is essential: we are studying fluctuation around the population change, whose size Section 4 has already handled.

The contribution of this fold to the remainder in Section 5 is

$$R_k := \sqrt{n}\alpha_k(P_k - P)\delta_k = \frac{1}{\sqrt{n}}\sum_{i \in I_k}\{\delta_k(W_i) - P\delta_k\}. \tag{41}$$

Multiplying a random variable by a constant multiplies its variance by the square of that constant. Thus

$$\begin{aligned} \text{Var}(R_k \mid \text{training}) &= n\alpha_k^2\frac{v_k}{n_k} = \alpha_k v_k \\ &\leq \alpha_k P\delta_k^2. \end{aligned} \tag{42}$$

The last inequality uses "variance = mean square minus squared mean." As the conditional variance bound in Equation 42 shows, the averaging has offset the $\sqrt{n}$ scaling; now we only need $P\delta_k^2 \xrightarrow{p} 0$. For equal-sized folds, for example, the variance bound is $P\delta_k^2/K$.

This explains an important difference between the two remainders. The population bias needs a rate fast enough to beat $n^{-1/2}$. For the **centered** sample fluctuation, it is enough here that the mean square of the contribution difference tends to zero at all.

6.2 Why do accurate functions make $P\delta_k^2$ small?

For this calculation only, omit the fold subscript k and write $e_\mu = \hat{\mu} - \mu_0$. Starting from $B = A_p(Y - \mu)$, add and subtract $A_{\hat{p}}(Y - \mu_0)$:

$$\begin{aligned} \delta &= A_{\hat{p}}(Y - \hat{\mu}) - A_{p_0}(Y - \mu_0) \\ &= (A_{\hat{p}} - A_{p_0})(Y - \mu_0) - A_{\hat{p}} e_\mu. \end{aligned} \tag{43}$$

The first term is the change in weights times the true outcome residual. The second is the change in the predicted outcome times the fitted weight. We bound their mean squares separately.

For treated observations, $A_p = 1$ for any p, so changing p does not change their weight. For controls, subtract the two odds using a common denominator:

$$\begin{aligned} A_{\hat{p}} - A_{p_0} &= -(1 - D)\left\{ \frac{\hat{p}}{1 - \hat{p}} - \frac{p_0}{1 - p_0} \right\} \\ &= -(1 - D)\frac{\hat{p} - p_0}{(1 - \hat{p})(1 - p_0)}. \end{aligned} \tag{44}$$

Both denominators are at least ϵ, so the absolute weight change is at most $(1 - D)|\hat{p} - p_0|/\epsilon^2$. Also $|A_{\hat{p}}| \le M$, where we can take $M = \max\{1, (1 - \epsilon)/\epsilon\}$.

The inequality $(u + v)^2 \le 2u^2 + 2v^2$ now lets us square the two parts without having to control their cross-product:

$$\begin{aligned} P\delta^2 &\le \frac{2}{\epsilon^4}\mathbb{E}[(1 - D)(\hat{p} - p_0)^2(Y - \mu_0)^2] + 2M^2\|e_\mu\|_2^2 \\ &\le \frac{2C}{\epsilon^4}\|\hat{p} - p_0\|_2^2 + 2M^2\|\hat{\mu} - \mu_0\|_2^2 = o_p(1). \end{aligned} \tag{45}$$

For the second inequality, first average the squared residual among controls at a given X. Section 4 bounds that average by C, so, with the fitted functions still held fixed,

$$\begin{aligned} &\mathbb{E}[(1 - D)(\hat{p} - p_0)^2(Y - \mu_0)^2] \\ &\quad = \mathbb{E}[(\hat{p} - p_0)^2(1 - p_0)\mathbb{E}[(Y - \mu_0)^2 \mid X, D = 0]] \\ &\quad \le C\mathbb{E}[(\hat{p} - p_0)^2] = C\|\hat{p} - p_0\|_2^2. \end{aligned} \tag{46}$$

Both function errors tend to zero in L_2, so the upper bound on $P\delta^2$ tends to zero. We have proved the condition needed for the conditional variance of R_k to vanish.

6.3 Turn the variance bound into convergence in probability

There is one detail left: the variance bound depends on the random training sample. We know it tends to zero in probability, rather than having a fixed small value for every possible training sample.

Choose an error tolerance $a > 0$ and a variance threshold $v > 0$. Split training samples into those for which $\alpha_k P\delta_k^2 > v$ and the rest. The first group has probability tending to zero. In the second group, R_k has conditional mean zero and variance at most v. Chebyshev's inequality bounds the conditional chance of $|R_k| > a$ by v/a^2. Together these observations give

$$\mathbb{P}(|R_k| > a) \le \mathbb{P}(\alpha_k P\delta_k^2 > v) + \frac{v}{a^2}. \tag{47}$$

First let n grow: the first term vanishes. Then choose v as small as desired: the second term can be made arbitrarily small. Thus $\mathbb{P}(|R_k| > a) \to 0$ for every $a > 0$, which is exactly $R_k = o_p(1)$.

Finally, $R_{\text{sample}} = \sum_{k=1}^{K} R_k$. If the absolute value of this sum exceeds a, at least one $|R_k|$ must exceed a/K. Hence

$$\mathbb{P}\left(\left|\sum_{k=1}^{K} R_k\right| > a\right) \leq \sum_{k=1}^{K} \mathbb{P}(|R_k| > a/K) \longrightarrow 0. \tag{48}$$

This step uses a fixed number of folds, but no independence between folds.

Cross-fitting's role is now explicit: it allowed us to treat the fitted function as fixed while averaging over independent evaluation observations. Without that separation, holding the fit's training data fixed would also hold the evaluation observations fixed, so this conditional-variance argument would not apply. Other restrictions can sometimes justify reusing the same data, but they require a different argument.

7 The limiting distribution and a usable standard error

We have now reached the goal set out in Section 1: $\sqrt{n}(\hat{\tau} - \tilde{\tau}) \xrightarrow{p} 0$. Fitting the nuisance functions adds an error negligible relative to ordinary sampling variation. We can therefore use the known-function calculation to describe the leading error of the estimator we actually compute.

7.1 Keep the leading average and let the extra errors disappear

With both remainders vanishing, Section 5's decomposition becomes

$$\sqrt{n}(\hat{\tau} - \tau) = \frac{1}{\hat{\pi}} \frac{1}{\sqrt{n}} \sum_{i} (B_{0i} - D_i \tau) + o_p(1). \tag{49}$$

The remaining sum uses the true functions: its terms are independent and have mean zero. Section 2 defined $\phi_0(W) = (B_0(W) - D\tau)/\pi$. Since $\hat{\pi} \xrightarrow{p} \pi > 0$, we can replace $\hat{\pi}$ by π here with another $o_p(1)$ change, giving

$$\sqrt{n}(\hat{\tau} - \tau) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_0(W_i) + o_p(1). \tag{50}$$

Assume $0 < V := \mathbb{E}[\phi_0(W)^2] < \infty$. This ensures a finite, positive variance for the limiting normal distribution. The CLT applies to the sum; adding a term that tends to zero does not change its limit. Thus Equation 51 gives

$$\sqrt{n}(\hat{\tau} - \tau) \xrightarrow{d} N(0, V), \qquad V = \mathbb{E}[\phi_0(W)^2]. \tag{51}$$

The expression $N(0, V)$ describes the **rescaled** error. On the original scale, the large-sample approximation is

$$\hat{\tau} - \tau \text{ approximately } N\left(0, \frac{V}{n}\right). \tag{52}$$

Consequently, the standard error we need to estimate is $\sqrt{V/n}$. Here V/n is the variance in this approximation, not a claim about the estimator's exact finite-sample variance.

7.2 Replace the unknown influence contributions by fitted ones

The formula for ϕ_0 contains unknown quantities. Replace each by its sample counterpart, using the held-out $\hat{B}_i$ already computed in Section 5:

$$\hat{\phi}_i = \frac{\hat{B}_i - D_i \hat{\tau}}{\hat{\pi}}. \tag{53}$$

For a treated observation, subtract $\hat{\tau}$ from $\hat{B}_i$; for a control, subtract zero. Divide both by $\hat{\pi}$. The resulting $\hat{\phi}_i$ estimates observation i's contribution to the leading error.

These fitted contributions average to exactly zero:

$$\frac{1}{n}\sum_i \hat{\phi}_i = \frac{n^{-1}\sum_i \hat{B}_i - \hat{\pi}\hat{\tau}}{\hat{\pi}} = 0, \tag{54}$$

because $\hat{\tau}$ was defined as $n^{-1}\sum_i \hat{B}_i/\hat{\pi}$. Their mean square therefore measures their spread around their sample mean. Use it to estimate V, then divide by n and take a square root as in Equation 55:

$$\hat{V} = \frac{1}{n}\sum_i \hat{\phi}_i^2, \qquad \widehat{\mathrm{se}}(\hat{\tau}) = \sqrt{\frac{\hat{V}}{n}}. \tag{55}$$

This calculation needs just one fitted contribution per observation. We do not compute a variance separately within each fold and treat the fold estimates as independent. The justification instead comes from showing that these contributions approximate the independent true $\phi_0(W_i)$.

7.3 Why does the fitted mean square estimate the right variance?

We need $\hat{V} \xrightarrow{p} V$ before substituting it into a confidence interval. There are two steps: fitted contributions must get close to true ones, and the true contributions' sample mean square must approach its population mean. In the expressions below, $\mathbb{P}_n$ simply averages the indicated n values, including when the fitted value uses a different function in each fold.

First, Section 6 showed that, for each fold, $P\delta_k^2 \xrightarrow{p} 0$. The held-out mean square $P_k\delta_k^2$ has conditional expectation $P\delta_k^2$. Because it is nonnegative, Markov's inequality makes it small in probability too.[2] Pooling the folds gives

$$\mathbb{P}_n(\hat{B} - B_0)^2 := \frac{1}{n}\sum_i (\hat{B}_i - B_{0i})^2 = \sum_k \alpha_k P_k\delta_k^2 \xrightarrow{p} 0. \tag{56}$$

We also have $\hat{\tau} \xrightarrow{p} \tau$ from the leading-error expansion and $\hat{\pi} \xrightarrow{p} \pi > 0$ from the law of large numbers. To see how all three facts enter, subtract the two influence contributions explicitly:

$$\hat{\phi}_i - \phi_0(W_i) = \frac{\hat{B}_i - B_{0i} - D_i(\hat{\tau} - \tau)}{\hat{\pi}} + \left(\frac{1}{\hat{\pi}} - \frac{1}{\pi}\right)(B_{0i} - D_i\tau). \tag{57}$$

The average squared numerator error in the first term tends to zero; the denominator tends to a positive constant. In the second term, the multiplier tends to zero and the average of $(B_{0i} - D_i\tau)^2$ stays bounded by the law of large numbers. Squaring and averaging therefore gives $\mathbb{P}_n(\hat{\phi} - \phi_0)^2 \xrightarrow{p} 0$.

Second, closeness of the contributions implies closeness of their mean squares. Factor $\hat{\phi}_i^2 - \phi_0(W_i)^2$ as a difference times a sum, and apply the same Cauchy–Schwarz inequality used in Section 4:

$$|\mathbb{P}_n\hat{\phi}^2 - \mathbb{P}_n\phi_0^2| \le \{\mathbb{P}_n(\hat{\phi} - \phi_0)^2\}^{1/2}\{\mathbb{P}_n(\hat{\phi} + \phi_0)^2\}^{1/2} \xrightarrow{p} 0. \tag{58}$$

The first factor tends to zero. The second stays bounded because $\mathbb{P}_n(\hat{\phi} + \phi_0)^2 \le 2\mathbb{P}_n(\hat{\phi} - \phi_0)^2 + 8\mathbb{P}_n\phi_0^2$ and the law of large numbers gives $\mathbb{P}_n\phi_0^2 \xrightarrow{p} V$. Therefore $\hat{V} = \mathbb{P}_n\hat{\phi}^2 \xrightarrow{p} V$.

[2] For any $a, v > 0$, split training samples according to whether $P\delta_k^2 > v$. On the remaining samples, conditional Markov gives $\mathbb{P}(P_k\delta_k^2 > a \mid \text{training}) \le v/a$. Hence $\mathbb{P}(P_k\delta_k^2 > a) \le \mathbb{P}(P\delta_k^2 > v) + v/a$. Let n grow, then v tend to zero, as in Section 6. We do not need the unconditional expectation of $P\delta_k^2$ to converge to zero.

7.4 Form the confidence interval

Since $\hat{V}$ approaches the positive V, dividing the estimation error by the estimated standard error gives

$$\frac{\hat{\tau} - \tau}{\sqrt{\hat{V}/n}} \xrightarrow{d} N(0,1). \tag{59}$$

A standard normal variable lies between -1.96 and 1.96 with probability approximately 0.95. Rearranging that comparison gives the interval in Equation 60:

$$\hat{\tau} \pm 1.96\sqrt{\hat{V}/n}. \tag{60}$$

Across repeated large samples satisfying our assumptions, these intervals cover the fixed true effect approximately 95% of the time. The argument is asymptotic, so it does not promise exact coverage at any particular sample size.

8 A numerical influence-function check

We now put numbers into the influence function to see what its variance means. This calculation uses the true functions, so it illustrates the limiting variance V that Section 7 estimates; it is not a simulation of fitting them.

Recall the three-value example: $X = 0, 1, 2$ each occurs with probability $1/3$, the treatment probabilities are $(1/4, 1/2, 3/4)$, and $\mu_0(X) = X^2$. For this calculation set the outcome noise $U = 0$, so controls have outcome X^2 and treated people have outcome $X^2 + 2X$. The true effect is $\tau = 8/3$ and the treated share is $\pi = 1/2$.

Even though outcomes are now determined by (D, X), a new sample still contains a different mix of covariate values and treatment indicators. We will calculate the sampling uncertainty due to that changing mix.

8.1 Work out each possible contribution

For controls, $Y - \mu_0(X) = 0$, so $\phi_0 = 0$. For treated observations, $Y - \mu_0(X) = 2X$, giving

$$\phi_0 = 2D(2X - 8/3). \tag{61}$$

For instance, a treated person with $X = 0$ has $\phi_0 = -16/3$: their comparison $2X = 0$ is below the ATT and pulls the estimate downward. A treated person with $X = 2$ has $\phi_0 = 8/3$ and pulls it upward. Each person's contribution to the leading estimation error is ϕ_0/n.

To average these contributions, we need their probabilities in the **whole population**, not just among treated people. For example, $\mathbb{P}(X = 0, D = 1) = \mathbb{P}(X = 0)p_0(0) = (1/3)(1/4) = 1/12$. The corresponding probabilities for $X = 1, 2$ are $1/6, 1/4$.

Treated cell x	Population probability	ϕ_0	Probability times ϕ_0^2
0	1/12	$-16/3$	64/27
1	1/6	$-4/3$	8/27
2	1/4	8/3	16/9

The listed probabilities sum to $1/2$. The remaining half of the population consists of controls, all with $\phi_0 = 0$ in this example. They therefore add zero to both the mean and mean square. The mean is

$$\frac{1}{12}\left(-\frac{16}{3}\right) + \frac{1}{6}\left(-\frac{4}{3}\right) + \frac{1}{4}\frac{8}{3} = -\frac{4}{9} - \frac{2}{9} + \frac{6}{9} = 0, \tag{62}$$

Since this mean is zero, the variance is the probability-weighted mean square:

$$V = \frac{64}{27} + \frac{8}{27} + \frac{48}{27} = \frac{40}{9}. \tag{63}$$

Translate V into a standard error

The large-sample standard error is $\sqrt{40/(9n)}$. At $n = 100$ it is about 0.211, giving a 95% interval half-width of about 0.413; at $n = 400$ both numbers halve. These are illustrations using the known population V. In an application we substitute $\widehat{V}$, and finite-sample coverage remains approximate.

This positive variance comes entirely from changes in the composition of the treated sample. It also illustrates why Section 2 kept the random treated count in the calculation. Using $B_0/\pi - \tau$ as the contribution would give $V = 104/9$ instead, describing an estimator that divides by the known population share rather than the sample share.

8.2 What efficiency adds

So far we have shown how to calculate and estimate this estimator's leading variance. A separate question is whether another estimator could have a smaller leading variance under the same model assumptions.

Under the nonparametric ATT model with unknown propensity, ϕ_0 is the efficient influence function: its variance is the lower bound for regular, asymptotically linear estimation in that model. "Asymptotically linear" means the leading error can be written as an average of individual contributions, as ours can. "Regular" requires suitably stable limiting behavior under small changes to the data-generating distribution; it rules out apparent improvements that work only at an isolated distribution. This bound is an external theorem, not established by checking that ϕ_0 has mean zero. See Hahn (1998), On the role of the propensity score.

Our derivation shows that the cross-fitted estimator has this influence function under the stated moment, consistency, and rate conditions. Consequently it attains the bound under those conditions. Consistency alone does not imply efficiency.

9 What changes if only one nuisance model is correct?

The previous note established consistency when either nuisance function is correct. The inference proof here used more: **both function errors tend to zero**, and their product is small relative to $n^{-1/2}$. We now examine why the weaker consistency result does not automatically justify the same standard error.

9.1 A zero limiting bias need not mean a negligible estimation error

Suppose the fitted functions converge to limiting functions μ^* and p^*. The star denotes what the fitting method approaches with unlimited data; this need not be the truth. Consider $p^* = p_0$ but $\mu^* \neq \mu_0$: the propensity fit becomes correct, while the regression retains a systematic error.

At the limit, the product identity still gives $\mathbb{E}h(W;\tau,\mu^*,p_0) = 0$. This is the double-robust consistency result. In a finite sample, however, the fitted propensity differs from p_0. That small error now multiplies a regression error that does **not** shrink. There is only one small factor left.

For a concrete example, suppose $p_0(x) = 1/2$ everywhere, the limiting regression is too high by one unit, $\mu^*(x) = \mu_0(x) + 1$, and we insert $p_r(x) = 1/2 + r$ for a small r. Then

$$\mathbb{E}h(W;\tau,\mu^*,p_r) = \mathbb{E}\left[\frac{(-r)(-1)}{1/2 - r}\right] = \frac{r}{1/2 - r} \approx 2r. \tag{64}$$

The change is proportional to r, rather than r^2. If a propensity estimation error is of order $n^{-1/2}$, this population change can also be of order $n^{-1/2}$ and survive the rescaling in Section 1. It still goes to zero, which explains how consistency and a missing first-order uncertainty term can coexist.

9.2 Repeat the direction calculation around the wrong limit

The general calculation is the same as Section 3, but starts at (μ^*, p_0) rather than (μ_0, p_0). Hold μ^* fixed and set $p_r = p_0 + ra$ for a fixed admissible direction a. The product identity gives

$$\mathbb{E}h(W;\tau,\mu^*,p_r) = -r\mathbb{E}\left[\frac{a(\mu_0 - \mu^*)}{1 - p_r}\right]. \tag{65}$$

Since the mean at $r = 0$ is zero, dividing by r and taking the limit gives

$$\frac{\mathbb{E}h(W;\tau,\mu^*,p_r)}{r} = -\mathbb{E}\left[\frac{a(\mu_0 - \mu^*)}{1 - p_r}\right] \longrightarrow -\mathbb{E}\left[\frac{a(\mu_0 - \mu^*)}{1 - p_0}\right] \tag{66}$$

under an integrable bound. There is no reason for this derivative to be zero: the nonvanishing regression error remains in it. Thus estimating the correct propensity can contribute to the leading error. The analogous issue arises with correct regression and wrong propensity.

There is another change too: $B(W; \mu^*, p_0)$ generally differs from $B_0(W)$, even though they have the same population mean. Its sampling variation can therefore differ. The proof in Section 6 that $P\delta_k^2 \xrightarrow{p} 0$ relative to B_0 no longer follows. Both the sampling contributions at the limiting functions and the effects of fitting those functions must be reconsidered.

9.3 What would a valid standard error have to include?

This is why **double robustness of consistency does not automatically imply double robustness of inference**. For parametric fits under appropriate regularity conditions, one route is to analyze the propensity fit, regression fit, and effect estimate together. Their **joint estimating equations** keep track of all three sets of estimation errors; a **sandwich variance** combines their variation and the ways those errors pass into the effect estimate. Simply substituting the limiting, possibly wrong functions into the influence formula from Section 2 does not generally account for these extra terms.

The extension note develops a different route: choose propensity and regression fitting equations that cancel these first-order terms even at their possibly wrong limiting functions. That result uses specific fitting methods and regularity assumptions, not arbitrary machine learning.

9.4 Interpretation of the result

To connect the proof to an application, keep the roles of its assumptions separate:

- **The sampling design determines the independent units.** For a panel, keep all periods of a unit in one fold. With clustered observations, splitting and the variance calculation must respect independent clusters; the observation-level proof above is not enough.
- **Overlap controls how much individual controls are weighted.** Inspect treatment probabilities and the amount of comparable control information. Capping fitted propensities below one prevents extreme weights, but matches this proof only if the true propensity satisfies that bound and the fits remain consistent.
- **Function accuracy controls the population error.** Cross-fitting supplies separation of fitting and evaluation. It does not, by itself, establish the product rate required in Section 4.
- **Causal assumptions determine what the estimator targets.** A small standard error describes little sampling variation; it does not establish that the limit is the intended causal effect.

For the general cross-fitting framework see Chernozhukov et al. (2018), Double/debiased machine learning. The proof here specializes the argument to the ATT ratio estimator.