

Debiased machine learning estimation for causal effects

Karla Diaz-Ordaz,

University College London

APTS Week 4, September 2026

Recap from last lecture

- ▶ g-computation, IPW/Hajek, and AIPW all rely on at least one of the outcome or exposure models being correctly specified
- ▶ naive ML-based plug-in estimators are problematic
- ▶ today: how to construct estimators that can use machine learning and still have good statistical properties

Understanding the bias

- ▶ Let $\theta(P)$ be a model-free estimand (statistical functional of P) , e.g.

$$\theta = E\{Y(1)\} = E\{Q(X)\},$$

to declutter notation I have dropped the subindex 1; here $Q(X)$ denotes the true treated outcome regression

- ▶ We say that $\hat{\theta} = \theta(\hat{P})$ is a *plug-in* estimator when it is constructed by plugging in an estimator(s) of P (or aspects of P , such as the nuisance parameters) into the estimand formula, e.g.

$$\hat{\theta} = \hat{E}(Y(1)) = \frac{1}{n} \sum_{i=1}^n \hat{Q}_n(X_i),$$

recall: g-computation, last lecture, was already an example of this

- ▶ We are interested in understanding the error, i.e. the difference

$$\hat{\theta} - \theta.$$

Notation

- ▶ recall $O = (X, A, Y) \sim P$ the observed data
- ▶ an expectation for a new X is denoted $P\{f(X)\} = \int f(X) dP(X)$
- ▶ the sample average is $P_n\{f(X)\} = \frac{1}{n} \sum_{i=1}^n f(X_i)$
- ▶ We use **two further symbols** for a distribution:
 - ▶ $\hat{P}_n$: the **actual, specific** estimator we compute from the data (e.g. built from $\hat{Q}_n, \hat{\pi}_n$).
 - ▶ $\tilde{P}$: a **generic, arbitrary** distribution “close” to P (the true distribution). This is used *only* to derive the theory It stands for “some working distribution,” not any specific estimator
- ▶ The theory is written in terms of $\tilde{P}$ **holds for any estimator**
- ▶ Once the theory is derived, we **substitute** $\tilde{P}$ by our specific $\hat{P}_n$ obtaining practical results

What drives this bias?

- ▶ We wish to understand $\theta(\tilde{P}) - \theta(P)$, for $\tilde{P}$ some estimator of P .
- ▶ This depends on
 - ▶ the **quality of the nuisance parameter estimators** $\tilde{P}$;
 - ▶ **how sensitive θ is to changes in the law P**
- ▶ Both are difficult to examine in general
- ▶ We will therefore restrict ourselves to
 - ▶ specific, estimators $\hat{P}_n$
 - ▶ **estimands that are 'smooth'** in the data-generating law.
so that small differences between $\hat{P}_n$ and P give rise to small differences $\hat{\theta} - \theta$.

The von Mises expansion

- ▶ The sensitivity of $\theta(P)$ to changes in the DGP is characterised by the pathwise derivative:

$$\left. \frac{d\theta(P_t)}{dt} \right|_{t=0}$$

where $P_t = tP + (1 - t)\tilde{P}$ is a **one-dimensional parametric submodel** with $t \in [0, 1]$, so that $P_0 = \tilde{P}$ (and $P_1 = P$).

- ▶ not a practically useful model: it captures the notion of perturbing $\tilde{P}$ in the direction of P .
- ▶ This forms the basis of the *von Mises expansion*, a Taylor-expansion, but for functionals of a distribution

$$\theta(P) = \theta(\tilde{P}) + \left. \frac{d\theta(P_t)}{dt} \right|_{t=0}(1 - 0) + R(P, \tilde{P})$$

where $R(P, \tilde{P})$ is a remainder term.

The efficient influence function

- ▶ The *efficient influence function* $\phi(O, \tilde{P})$ (at $\tilde{P}$) is s.t.

$$\left. \frac{d}{dt} \theta(P_t) \right|_{t=0} = \tilde{P}\{\phi(O, \tilde{P}) s(O)\}$$

for any submodel through $\tilde{P}$ with score $s(O)$ at $t = 0$

- ▶ **Result 1:** $\phi(O, \tilde{P})$ is unique and has mean zero **under $\tilde{P}$** :
 $\tilde{P}\{\phi(O, \tilde{P})\} = 0$

(since $\int p_t(O) dO = 1$ for every t , differentiating at $t = 0$ gives $\int s(O) d\tilde{P}(O) = 0$; $\phi(O, \tilde{P})$ is built from exactly such scores, so it inherits this — true at *any* point in the model, not just the truth)

- ▶ **Result 2:**

$$\left. \frac{d}{dt} \theta(P_t) \right|_{t=0} = P\{\phi(O, \tilde{P})\} - \tilde{P}\{\phi(O, \tilde{P})\}$$

since, for the path $P_t = tP + (1 - t)\tilde{P}$, the score at $t = 0$ is $s(O) = \{p(O) - \tilde{p}(O)\} / \tilde{p}(O)$

Back to the von Mises expansion

- ▶ **Result 3**

$$\left. \frac{d}{dt} \theta(P_t) \right|_{t=0} = P\{\phi(O, \tilde{P})\}$$

using Results 1 and 2

- ▶ now recall the von Mises expansion:

$$\theta(P) = \theta(\tilde{P}) + \underbrace{\left. \frac{d\theta(P_t)}{dt} \right|_{t=0}}_{\text{pathwise derivative}} (1 - 0) + R(P, \tilde{P})$$

- ▶ substituting **Result 3** for the pathwise derivative term:

$$\theta(P) = \theta(\tilde{P}) + P\{\phi(O, \tilde{P})\} + R(P, \tilde{P})$$

- ▶ Rearranging to isolate the estimand difference:

$$\theta(\tilde{P}) - \theta(P) = -P\{\phi(O, \tilde{P})\} - R(P, \tilde{P})$$

Back to the von Mises expansion (contd.)

- ▶ Substituting $\tilde{P}$ by $\hat{P}_n$, and scaling by $\sqrt{n}$

$$\sqrt{n}\{\theta(\hat{P}_n) - \theta(P)\} = -\sqrt{n}P\{\phi(O, \hat{P}_n)\} - \sqrt{n}R(P, \hat{P}_n)$$

- ▶ The term $P\{\phi(O, \hat{P}_n)\}$ cannot be calculated from the data.
- ▶ Let's focus on $-\sqrt{n}P\{\phi(O, \hat{P}_n)\}$.

Adding and subtracting terms,

$$\begin{aligned} & -\sqrt{n}P\{\phi(O, \hat{P}_n)\} \\ &= \sqrt{n}(P_n - P)\{\phi(O, \hat{P}_n)\} - \sqrt{n}P_n\{\phi(O, \hat{P}_n)\} \quad (\text{add \& subtract } \sqrt{n}P_n\{\phi(O, \hat{P}_n)\}) \\ &= \sqrt{n}(P_n - P)\{\phi(O, P)\} - \sqrt{n}P_n\{\phi(O, \hat{P}_n)\} \\ &\quad + \sqrt{n}(P_n - P)\{\phi(O, \hat{P}_n) - \phi(O, P)\} \quad (\text{add \& subtract } \sqrt{n}(P_n - P)\{\phi(O, P)\}) \end{aligned}$$

Why is the estimand's efficient influence function key?

- ▶ Thus we obtain

$$\begin{aligned}\sqrt{n}\{\theta(\hat{P}_n) - \theta(P)\} &= \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^n \phi(O_i, P)}_{\text{linear}} - \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^n \phi(O_i, \hat{P}_n)}_{\text{bias}} \\ &\quad + \underbrace{\sqrt{n}(P_n - P)\{\phi(O, \hat{P}_n) - \phi(O, P)\}}_{\text{empirical process}} \\ &\quad + \underbrace{\sqrt{n}R(P, \hat{P}_n)}_{\text{remainder}}\end{aligned}$$

- ▶ The first term in the expansion is well-understood
- ▶ It is normally distributed in large samples, with mean 0, and easy-to-calculate variance.

Example: efficient influence function of $E\{Y(1)\}$

- ▶ Consider $\theta(P) = E\{E(Y|A=1, X)\}$
- ▶ The EIF is

$$\begin{aligned}\phi(O, P) = & \frac{A}{\pi(X)} \{Y - E(Y|A=1, X=x)\} \\ & + E(Y|A=1, X=x) - \theta(P)\end{aligned}$$

- ▶ **recall:** the AIPW from last lecture

$$\mu_1^{AIPW}(O) = Q_1(X) + \frac{A}{\pi(X)} \{Y - Q_1(X)\}$$

- ▶ $\mu_1^{AIPW}(O)$ is (up to centering by $\theta(P)$) the EIF

Zooming in on the second term...

- ▶ The second term

$$-\frac{1}{\sqrt{n}} \sum_{i=1}^n \phi_{\hat{P}_n}(O_i)$$

is usually not well understood.

- ▶ Its randomness originates from the randomness of the data O_i , but also the uncertainty in the data-adaptive estimators $\hat{P}_n$, which is ill understood.
- ▶ Moreover, the complex distribution of such estimators $\hat{P}_n$, may propagate into this term, thereby rendering $\hat{\theta}$ biased and non-normal.
- ▶ This is the root cause of **plug-in bias**.

Example: Plug-in bias of g-computation of $E\{Y(1)\}$

- For g-computation estimators, this second term is

$$\begin{aligned}& -\frac{1}{n} \sum_{i=1}^n \frac{A_i}{\pi(X_i)} \{Y_i - \hat{Q}_n(X_i)\} + \hat{Q}_n(X_i) - \frac{1}{n} \sum_{i=1}^n \hat{Q}_n(X_i) \\&= -\frac{1}{n} \sum_{i=1}^n \frac{A_i}{\pi(X_i)} \{Y_i - Q(X_i) + Q(X_i) - \hat{Q}_n(X_i)\} \\&= -\frac{1}{n} \sum_{i=1}^n \frac{A_i(Y_i - Q(X_i))}{\pi(X_i)} - \frac{1}{n} \sum_{i=1}^n \frac{A_i}{\pi(X_i)} \{Q(X_i) - \hat{Q}_n(X_i)\}\end{aligned}$$

- This is determined by the difference

$$Q(X_i) - \hat{Q}_n(X_i)$$

which is poorly understood.

Eliminating plug-in bias

- ▶ Since this bias term

$$-\frac{1}{n} \sum_{i=1}^n \phi_{\hat{P}_n}(O_i)$$

is known when the efficient influence function is given, there is also hope that we can remove it.

- ▶ We discuss 3 strategies to achieve this:
 - ▶ one-step plug-in estimators;
 - ▶ estimating equations estimators
 - ▶ targeted minimum loss estimation

The one-step plug-in estimator

- The identity

$$\hat{\theta} - \theta \approx \frac{1}{n} \sum_{i=1}^n \phi_P(O_i) - \frac{1}{n} \sum_{i=1}^n \phi_{\hat{P}_n}(O_i),$$

suggests that we can remove plug-in bias
by adding the bias term to the plug-in estimator:

$$\hat{\theta}_1 = \hat{\theta} + \frac{1}{n} \sum_{i=1}^n \phi_{\hat{P}_n}(O_i).$$

- For this one-step plug-in estimator $\hat{\theta}_1$, we then have

$$\hat{\theta}_1 - \theta \approx \frac{1}{n} \sum_{i=1}^n \phi_P(O_i).$$

Example: A one-step estimator of $E\{Y(1)\}$

- ▶ Starting from a plug-in estimator

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n \hat{Q}_n(X_i),$$

we obtain a one-step plug-in estimator as

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \hat{Q}_n(X_i) + \frac{1}{n} \sum_{i=1}^n \frac{A_i}{\hat{\pi}_n(X_i)} \left\{ Y_i - \hat{Q}_n(X_i) \right\} + \hat{Q}_n(X_i) - \hat{\theta} \\ &= \frac{1}{n} \sum_{i=1}^n \frac{A_i}{\hat{\pi}_n(X_i)} \left\{ Y_i - \hat{Q}_n(X_i) \right\} + \hat{Q}_n(X_i). \end{aligned}$$

- ▶ This is the augmented IPW estimator.

The estimating equations estimator

- ▶ A second strategy to remove plug-in bias is to use an estimator $\hat{\theta}_2$ that **forces the bias term to be zero**

$$0 = \frac{1}{n} \sum_{i=1}^n \phi_{\hat{P}_n}(O_i).$$

- ▶ $\hat{\theta}_2$ is the solution to the above estimating equation.
- ▶ It is therefore called the **estimating equations estimator**.
- ▶ This forms the basis of the **debiased/double ML** literature.

Example: An estimating equations estimator of $E\{Y(1)\}$

- Solving

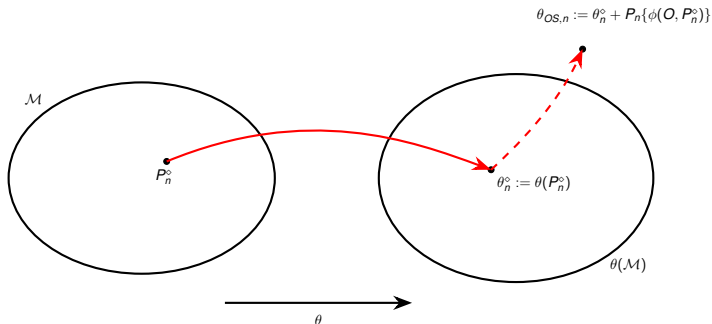
$$0 = \frac{1}{n} \sum_{i=1}^n \frac{A_i}{\hat{\pi}_n(X_i)} \left\{ Y_i - \hat{Q}_n(X_i) \right\} + \hat{Q}_n(X_i) - \hat{\theta},$$

happens to deliver the one-step plug-in estimator.

- This is not generally the case.

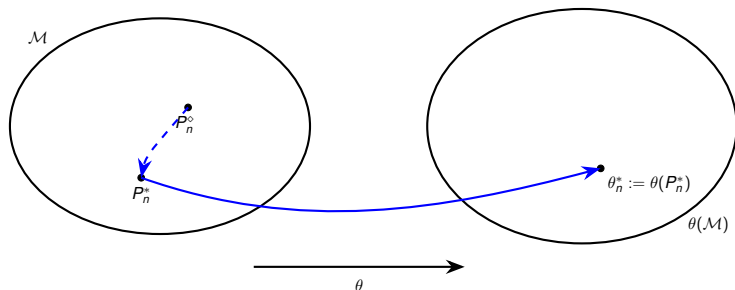
Downside of one-step correction

- ▶ Model space $\mathcal{M}$ (all distributions consistent with our assumptions) on the left, parameter space on the right
- ▶ Let $P_n^\diamond$ be the initial estimator (e.g., from some machine learning fit), therefore when mapped $\theta_n^\diamond = \theta(P_n^\diamond)$ is biased
- ▶ The one-step estimator performs an (additive) correction **in the parameter space** not a step to a new distribution in $\mathcal{M}$ (can land outside parameter space)



TMLE: targeting the distribution before mapping

- ▶ We can avoid this issues by performing the correction in the model space
- ▶ TMLE goes from $P_n^\diamond$ to a targeted P_n^* (in $\mathcal{M}$ therefore a valid distribution), **then maps** through θ
- ▶ Because θ_n^* comes from an actual $P_n^* \in \mathcal{M}$, it is guaranteed to land inside $\theta(\mathcal{M})$.



The Targeted Minimum loss estimator

1. TMLE tunes initial data-adaptive estimates $\hat{P}_n^{(0)}$
2. by building a parametric model around it with parameters estimated to force the bias term to be zero:

$$0 = \frac{1}{n} \sum_{i=1}^n \phi_{\hat{P}_n}(O_i).$$

3. this ensures that the solution equals the maximum likelihood estimator under a specific parametric model (the so-called least favourable submodel)
4. This results in a substitution estimator

Example: TMLE for $E\{Y(1)\}$

Estimand: $\theta(P) = P\{Q(X)\}$, with $Q(x) := E[Y \mid A = 1, X = x]$

- ▶ For $Y \in [0, 1]$ (rescale otherwise), tune initial predictions $\hat{Q}_n^{(0)}(X)$ by fitting (fluctuation model)

$$\text{logit}(E(Y|A, X)) = \text{logit}\hat{Q}_n^{(0)}(X) + \omega H(A, X)$$

- ▶ the **clever covariate** is $H(A, X) := \frac{I(A=1)}{\hat{\pi}_n(X)}$
- ▶ The “targeted” estimate $Q_n^*(X)$ is the fitted values based on the MLE for ω , which solves

$$0 = \frac{1}{n} \sum_{i=1}^n \frac{A_i}{\hat{\pi}_n(X_i)} \left\{ Y_i - \hat{Q}_n^{(1)}(X_i) \right\}.$$

- ▶ This is exactly the first term of ϕ (the EIF): H is “clever” precisely because it was reverse-engineered to make this happen.

Where does the rest of ϕ go?

The score only reproduces part of ϕ . The $E[Y \mid A = 1, X = x] - \theta(P)$ piece never appears in the fluctuation model

at convergence, TMLE defines the plug-in estimate as

$$\theta_n^* := P_n\{Q_n^*(X)\}.$$

Now check the full empirical mean of ϕ at the final fit:

$$P_n\{\phi(O, P_n^*)\} = \underbrace{P_n\{H(Y - Q_n^*)\}}_{= 0 \text{ by the MLE score equation}} + \underbrace{P_n\{Q_n^*(X)\} - \theta_n^*}_{= 0 \text{ by definition of } \theta_n^*}$$

Both terms vanish. The second term needs no separate fluctuation direction — it is absorbed automatically by how the plug-in parameter is computed.

Why?

- ▶ The clever covariate is derived directly from the EIF of the target estimands
- ▶ therefore different parameters (survival probabilities, quantiles) have clever covariates and fluctuation models
- ▶ For the ATE, we don't need to repeat the targeting step, but for other target estimands, we might
- ▶ The **logistic** fluctuation is why this TMLE automatically respects the bounds, $Y \in [0, 1]$
- ▶ Double robustness: the score $H(X)(Y - Q)$ vanishes in expectation if *either* H (i.e. π) or $\hat{Q}$ is correctly specified

How to achieve asymptotic normality?

- ▶ Recall our four-term decomposition (linear + bias + **empirical process** + **remainder**): we looked at three strategies to remove the bias term
- ▶ To show that one of those estimators, e.g.

$$\theta(\hat{P}_n) + \frac{1}{n} \sum_{i=1}^n \phi(O_i, \hat{P}_n)$$

is asymptotically normal:

- ▶ The empirical process term and remainder $\sqrt{n}R(P, \hat{P}_n)$ should **both** converge to zero.
- ▶ In that case, the variance of the debiased plug-in estimator can be estimated by the sample variance of the EIF

How to control empirical process terms?

- ▶ The **empirical process term**

$$\sqrt{n}(P_n - P)\{\phi(O, \hat{P}_n) - \phi(O, P)\}$$

arises because inference is based on

$$\sqrt{n}(P_n - P)\{\phi(O, P)\}$$

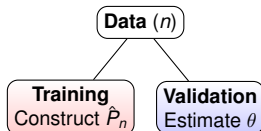
but we only have access to $\phi(O, \hat{P}_n)$.

- ▶ To control this term,
one can attempt to use empirical process theory
(Donsker conditions).
- ▶ This is often complex and not satisfactory for general
machine learning methods.

Sample-splitting and cross-fitting

Sample splitting

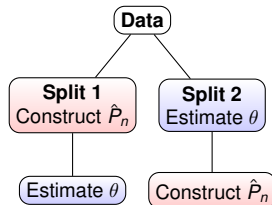
(Bickel, 1982; Schick, 1986; van der Vaart, 1998)



- ▶ This merely requires that $\hat{P}_n$ is consistent.
- ▶ But it is inefficient: half the data is 'wasted' on each of the two tasks, and this is sometimes associated with 'overfitting' or 'own observation' bias.

Cross-fitting solves this

(Zheng and van der Laan, 2011; Chernozhukov et al., 2018)



- ▶ swap the roles of the two splits, and take the **average** of the two estimates of θ : asymptotically, this regains full efficiency.

Why does sample splitting remove the empirical process term?

- ▶ If $\hat{P}_n$ is obtained from an external sample $O^N = (O_{n+1}, \dots, O_N)$, then **conditional on O^N** , the only randomness in $\phi(O, \hat{P}_n)$ comes from O , so

$$\sqrt{n}(P_n - P)\{\phi(O, \hat{P}_n) - \phi(O, P)\}$$

has mean zero and variance bounded by

$$P\left\{\{\phi(O, \hat{P}_n) - \phi(O, P)\}^2\right\}.$$

- ▶ By **Chebyshev's inequality**, this is enough: if that bound shrinks to zero, so does the empirical process term.

see e.g. Lemma 2 of Kennedy et al. (2020)

How to control the remainder term?

- ▶ The remainder $\sqrt{n}R(P, \hat{P}_n)$ is a **key ingredient** in causal machine learning.
- ▶ It is what justifies data adaptive estimation of nuisance parameters.
- ▶ However, it will need to be handled on a case-by-case basis.

What does the remainder look like?

- ▶ It follows from

$$\sqrt{n}\{\theta(\hat{P}_n) - \theta(P)\} = -\sqrt{n}P\{\phi(O, \hat{P}_n)\} - \sqrt{n}R(P, \hat{P}_n)$$

that the remainder term equals

$$\sqrt{n}R(P, \hat{P}_n) = -\sqrt{n}P\{\phi(O, \hat{P}_n)\} - \sqrt{n}\{\theta(\hat{P}_n) - \theta(P)\}.$$

- ▶ Since we know the influence function, we can calculate and bound the remainder.

The remainder term for $E\{Y(1)\}$

The remainder term for the one-step estimator equals

$$\begin{aligned} & \sqrt{n}R(P, \hat{P}_n) \\ &= -\sqrt{n}P \left\{ \frac{A\{Y - \hat{Q}_n(X)\}}{\hat{\pi}_n(X)} + \hat{Q}_n(X) - \theta(\hat{P}_n) \right\} - \sqrt{n}\{\theta(\hat{P}_n) - \theta(P)\} \\ &= -\sqrt{n}P \left\{ \frac{A\{Y - \hat{Q}_n(X)\}}{\hat{\pi}_n(X)} + \hat{Q}_n(X) - \theta(P) \right\} \\ &\quad - \sqrt{n}P \left\{ \left\{ \frac{\pi(X)}{\hat{\pi}_n(X)} - 1 \right\} \{Q(X) - \hat{Q}_n(X)\} \right\} \end{aligned}$$

This **product structure** is very common.

Bounding the remainder

- ▶ The Cauchy-Schwarz inequality: for any two random variables X and Y ,

$$|E(XY)| \leq \sqrt{E(X^2)E(Y^2)}$$

- ▶ Then $\sqrt{n}R(P, \hat{P}_n)$ can be upper bounded by

$$\sqrt{n}P \left\{ \left\{ \frac{\pi(X)}{\hat{\pi}_n(X)} - 1 \right\}^2 \right\}^{1/2} P \left\{ \{Q(X) - \hat{Q}_n(X)\}^2 \right\}^{1/2}.$$

Implications



$$\sqrt{n}P\left\{\left\{\frac{\pi(X)}{\hat{\pi}_n(X)} - 1\right\}^2\right\}^{1/2} P\left\{\{Q(X) - \hat{Q}_n(X)\}^2\right\}^{1/2}$$

is **asymptotically negligible** if **both** $\hat{\pi}_n(X)$ and $\hat{Q}_n(X)$ converge at faster than $n^{1/4}$ rate, but also more broadly.

- ▶ The remainder remains sufficiently small if e.g. $\hat{Q}_n(X)$ converges more slowly if $\hat{\pi}_n(X)$ has faster convergence.
- ▶ The remainder is zero when the propensity score is known.

Putting it together: asymptotic normality

- ▶ Let $\hat{\theta}_n$ denote **any** of the three estimators we have built from the EIF
- ▶ By construction, it satisfies $P_n\{\phi(O, \hat{P}_n)\} = 0$
- ▶ Substituting into our decomposition, the **bias term therefore vanishes** for all three:

$$\sqrt{n}\{\hat{\theta}_n - \theta(P)\} = \underbrace{\sqrt{n}P_n\{\phi(O, P)\}}_{\text{linear}} + \underbrace{\sqrt{n}(P_n - P)\{\phi(O, \hat{P}_n) - \phi(O, P)\}}_{\text{empirical process}} - \underbrace{\sqrt{n}R(P, \hat{P}_n)}_{\text{remainder}}$$

- ▶ We have shown **both** the empirical process term (via cross-fitting) and the remainder (via rate/double-robustness conditions) are $o_p(1)$

Theorem: asymptotic normality

Under these conditions, for any of the three constructions,

$$\sqrt{n}\{\hat{\theta}_n - \theta(P)\} \xrightarrow{d} N(0, \text{Var}_P[\phi(O, P)])$$

- ▶ $\text{Var}_P[\phi(O, P)]$ is the **semiparametric efficiency bound** (the smallest possible asymptotic variance among all regular estimators of $\theta(P)$)

So how do we do inference?

- ▶ Recall **Challenge 2**: plug-in ML estimators have **no valid uncertainty quantification**
- ▶ We now have exactly what we need: for any of the three estimators $\hat{\theta}_n$, a consistent estimator of the asymptotic variance is

$$\widehat{SE}(\hat{\theta}_n) = \sqrt{\frac{P_n\{\phi(O, \hat{P}_n)^2\} - \left(P_n\{\phi(O, \hat{P}_n)\}\right)^2}{n}}$$

- ▶ giving Wald-type confidence intervals

$$\hat{\theta}_n \pm 1.96 \times \widehat{SE}(\hat{\theta}_n)$$

and hypothesis tests, in the usual way

- ▶ **No bootstrap needed** and this is valid **even though** $\hat{P}_n$ was constructed using black-box machine learning

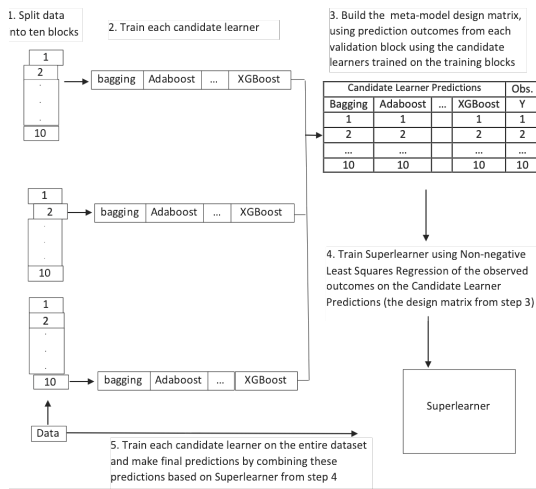
Implementation: Using AIPW R package

We will AIPW package to get the one-step debiased ML estimator for the ATE, using an ensemble of several models

```
SL.library=c("SL.glm", "SL.glm.interaction", "SL.glmnet", "SL.randomForest")
aipw_sl <-AIPW$new(Y=Y, A=A, W=W,
                  Q.SL.library=SL.library,
                  g.SL.library=SL.library,
                  k_split = 5)
aipw_sl$fit()
aipw_sl$summary()
```

Super Learner (SL)

- ▶ Combines multiple ML algorithms into a **heterogeneous ensemble**, to attenuate model misspecification
- ▶ Uses **cross-validation** to estimate each algorithm's risk (e.g. MSE)
- ▶ **Discrete SL**: picks algorithm with smallest CV-risk
- ▶ **Ensemble SL**: stacked regression
- ▶ **Oracle property**: asymptotically as accurate as the best learner in the library



Inspired by a figure in the "Targeted Learning" book (2011)

Summary

- ▶ Using data-adaptive (ML) methods to estimate nuisance models attenuates model misspecification concerns
- ▶ But plugging the resulting estimates directly into the estimand formula leads to **plug-in bias**
- ▶ The sample average of the EIF, evaluated at the plug-in fit, characterizes this bias
- ▶ Plug-in bias can therefore be eliminated by
 - ▶ adding this average to the initial plug-in (**one-step**);
 - ▶ using it as the basis of an **estimating equation**;
 - ▶ **targeting** the initial fit so that this average is zero (**TMLE**)
- ▶ When **all data-adaptive estimators $\hat{P}_n$ converge to the truth 'sufficiently' fast**, and **sample splitting** is used, then
 - ▶ the resulting estimators are asymptotically equivalent
 - ▶ inference can be based on the bootstrap,
 - ▶ or on Wald-type CIs, with $SE = n^{-1/2} \times \text{sample SD(EIF)}$

Selected references

- ▶ Van der Laan, M. J., & Rose, S. (2011). Targeted learning: causal inference for observational and experimental data. Springer Science & Business Media.
- ▶ Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, X., Newey, X., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.
- ▶ Hines, O., Dukes, O., Diaz-Ordaz, K., & Vansteelandt, S. (2022). Demystifying Statistical Learning Based on Efficient Influence Functions. *The American Statistician*, 76(3), 292–304.
<https://doi.org/10.1080/00031305.2021.2021984>

What is the efficient influence function of $E\{Y(1)\}$?

- Consider

$$\theta(P) = E\{E(Y \mid A = 1, X)\} = E\{Q(1, X)\}.$$

- Perturbing P in the direction of a point mass at $(\tilde{y}, \tilde{a}, \tilde{x})$, i.e.

$$f_t(y, a, x) = t l_{(\tilde{y}, \tilde{a}, \tilde{x})}(y, a, x) + (1 - t)f(y, a, x)$$

we obtain

$$\begin{aligned}\theta(P_t) &= \int y f_t(y \mid a = 1, x) f_t(x) dy dx \\ &= \int y \frac{f_t(y, a = 1, x) f_t(x)}{f_t(a = 1, x)} dy dx.\end{aligned}$$

example: pathwise differentiability continued

By the chain rule,

$$\begin{aligned}\frac{d\theta(P_t)}{dt}\Big|_{t=0} &= \int y \left\{ \frac{f(x)}{f(a=1, x)} \frac{d}{dt} f_t(y, a=1, x) \Big|_{t=0} \right. \\ &\quad - \frac{f(y, a=1, x)f(x)}{f(a=1, x)^2} \frac{d}{dt} f_t(a=1, x) \Big|_{t=0} \\ &\quad \left. + \frac{f(y, a=1, x)}{f(a=1, x)} \frac{d}{dt} f_t(x) \Big|_{t=0} \right\} dy dx \\ &= \int y \frac{f(y, a=1, x)f(x)}{f(a=1, x)} \left\{ \frac{l_{(y, a, x)}(\tilde{y}, \tilde{a}, \tilde{x})}{f(y, a=1, x)} \right. \\ &\quad \left. - \frac{l_{(a, x)}(\tilde{a}, \tilde{x})}{f(a=1, x)} + \frac{l_{(x)}(\tilde{x})}{f(x)} - 1 \right\} dy dx\end{aligned}$$

An example of pathwise differentiability

- ▶ Evaluating the integral gives

$$\left. \frac{d\theta(P_t)}{dt} \right|_{t=0} = \frac{l(\tilde{a}=1)}{\pi(\tilde{x})} \{ \tilde{y} - Q(1, \tilde{x}) \} + Q(1, \tilde{x}) - \theta(P)$$

where $\pi(x) := f(A=1 \mid X=x)$ is the propensity score, and $Q(1, x) := E(Y \mid A=1, X=x)$ as before.

- ▶ Because this has finite variance, we conclude that $\theta(P)$ is pathwise differentiable with efficient influence function

$$\phi(O, P) = \frac{l(A=1)}{\pi(X)} \{ Y - Q(1, X) \} + Q(1, X) - \theta(P)$$

- ▶ **recall:** this is exactly the (centered) AIPW correction term from Lecture 3 — $\mu_1^{AIPW}(O) - \theta(P)$

general TMLE

For given $\tilde{P} \in \mathcal{M}$, let $\mathcal{M}(\tilde{P}) = \{P(\omega) : \|\omega\| < \delta\} \subset \mathcal{M}$ be a fluctuation submodel with

$$(i) \ P(0) = \tilde{P} \quad \text{and}$$

$$(ii) \ \phi(O, \tilde{P}) \in \text{span} \left(\left. \frac{\partial}{\partial \omega} \log P(\omega) \right|_{\omega=0} \right),$$

where span refers to the closure of the linear span.

Suppose $P_n^{(k)}$ is a current estimate of P .

1. Set $P_n^{(k+1)}$ as the maximum likelihood estimator in $\mathcal{M}(P_n^{(k)})$.
2. Repeat until convergence, say to P_n^* .

Under conditions (on the empirical process and remainder terms, next) the targeted estimator $\theta_n^* := \theta(P_n^*)$ of θ is asymptotically linear with influence function $\phi(O, P)$.