OIT 661: Lecture 4
Double Machine Learning
(for class use only, please do not distribute)

Recap: Estimating average treatment effects under unconfoundedness. We have been discussing the following estimation problem. We observe data $(X_i,\, Y_i,\, W_i,) \in \mathcal{X} \times \mathbb{R} \times \{0,\, 1\}$ according to the potential outcomes model, such that there are potential outcomes $\{Y_i(0),\, Y_i(1)\}$ for which $Y_i = Y_i(W_i)$. We are not necessarily in a randomized controlled trial; however, we assume unconfoundedness, i.e., that treatment assignment is as good as random conditionally on the features X_i:

$$\{Y_i(0),\, Y_i(1)\} \perp\!\!\!\perp W_i \,\big|\, X_i. \tag{1}$$

We seek to estimate the average treatment effect $\tau = \mathbb{E}\left[Y_i(1) - Y_i(0)\right]$.

Throughout, we write $e(x) = \mathbb{P}\left[W_i = 1 \,\middle|\, X_i = x\right]$ for the propensity score, $\mu_{(w)}(x) = \mathbb{E}\left[Y_i(w) \,\middle|\, X_i = x\right]$ for the conditional response surfaces of the potential outcomes, $\tau(x) = \mu_{(1)}(x) - \mu_{(0)}(x)$ for the conditional average treatment effect given $X_i = x$, and assume that $\sigma^2(x) = \mathrm{Var}\left[Y_i(w) \,\middle|\, X_i = x\right]$ for not depend on w.

In the last class, we observed something of a surprising phenomenon. We considered 3 seemingly different estimators $\hat{\tau}$ for τ, namely bucketing when $\mathcal{X}$ is finite, IPW with a covariate balancing propensity score, and oracle doubly robust estimation, and found that they all have the same asymptotic distribution:

$$\sqrt{n}\left(\hat{\tau} - \tau\right) \Rightarrow \mathcal{N}\left(0,\, V^*\right), \quad V^* = \mathrm{Var}\left[\tau(X)\right] + \mathbb{E}\left[\frac{\sigma^2(X)}{e(X)(1 - e(X))}\right]. \tag{2}$$

The goal of this lecture is to further build on this phenomenon.

Semiparametric efficiency for ATE estimation. The starting point for our discussion is the following fact: Given the unconfoundedness assumption (1) but and no further parametric assumptions on $\mu_{(w)}(x)$ and $e(x)$, then no "regular" estimator of τ can improve on the behavior in (2). Moreover, any estimator attaining this bound is asymptotically equivalent to

$$\hat{\tau}^* = \frac{1}{n}\sum_{i=1}^{n}\left(\mu_{(1)}(X_i) - \mu_{(0)}(X_i) + W_i\frac{Y_i - \mu_{(1)}(X_i)}{e(X_i)} - (1 - W_i)\frac{Y_i - \mu_{(0)}(X_i)}{1 - e(X_i)}\right), \tag{3}$$

in the sense that $\sqrt{n}(\hat{\tau} - \hat{\tau}^*) \to_p 0$. We call V^* the *semiparametric efficient variance* for ATE estimation, and $\hat{\tau}^*$ the efficient score estimator. This result is a Cramer-Rao type bound for non-parametric average treatment effect estimation; for more details, see the paper by Hahn (1998).[1]

In the previous lectures, we discussed two concrete estimators that attain the semiparametric efficiency bound—the bucketing estimator for discrete $\mathcal{X}$ and the CBPS in the linear-logistic specification—and a close look at the derivations will establish that they are in fact asymptotically equivalent to (2). However, both of these estimators are rather restrictive: Bucketing needs $\mathcal{X}$ to be discrete, and CBPS is parametric. In this lecture, we develop a much more flexible strategy attaining the semiparametric efficiency bound.

Double machine learning. Unlike in the previous two lectures, we no longer want to make any parametric assumptions about $\mu_{(w)}(x)$ and $e(x)$; rather, much like in the machine learning literature, we simply assume that we can "learn" them. We assume that there exist some "good enough" estimators $\hat{\mu}_{(w)}(x)$ and $\hat{e}(x)$ (provided by, e.g., a sparse wavelet regression or a deep net), and want to use them to build an efficient treatment effect estimator.

Given this setup, a natural thing to try is to simply plug our learned predictions $\hat{\mu}_{(w)}(x)$ and $\hat{e}(x)$ into the efficient score function:

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i) + W_i\frac{Y_i - \hat{\mu}_{(1)}(X_i)}{\hat{e}(X_i)} - (1-W_i)\frac{Y_i - \hat{\mu}_{(0)}(X_i)}{1-\hat{e}(X_i)}\right), \tag{4}$$

However, this estimator is extremely unwieldy to work with. In particular, it is very hard to describe what might happen if $\hat{\mu}_{(w)}(x)$ and $\hat{e}(x)$ fit to the data too aggressively.

But we can overcome this challenge using cross-fitting. We first split the data (at random) into two halves $\mathcal{I}_1$ and $\mathcal{I}_2$, and then use an estimator

$$\begin{aligned}\hat{\tau}_{CF} =& \frac{|\mathcal{I}_1|}{n}\,\hat{\tau}^{\mathcal{I}_1} + \frac{|\mathcal{I}_2|}{n}\,\hat{\tau}^{\mathcal{I}_2}, \quad \hat{\tau}^{\mathcal{I}_1} = \frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i) - \hat{\mu}^{\mathcal{I}_2}_{(0)}(X_i)\right. \\ &\left. + W_i\frac{Y_i - \hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i)}{\hat{e}^{\mathcal{I}_2}(X_i)} - (1-W_i)\frac{Y_i - \hat{\mu}^{\mathcal{I}_2}_{(0)}(X_i)}{1-\hat{e}^{\mathcal{I}_2}(X_i)}\right),\end{aligned} \tag{5}$$

[1]To see why $\hat{\tau}^*$ has variance V^*/n, recall that we can decompose it into 3 uncorrelated parts: $\mu_{(1)}(X_i) - \mu_{(0)}(X_i)$ with variance $\mathrm{Var}\left[\tau(X)\right]$, $W_i\left(Y_i - \mu_{(1)}(X_i)\right)/e(X_i)$ with variance $\mathbb{E}\left[\sigma^2(x)/e(x)\right]$, and $(1-W_i)\left(Y_i - \mu_{(1)}(X_i)\right)/\left(1-e(X_i)\right)$ with variance $\mathbb{E}\left[\sigma^2(x)/(1-e(x))\right]$.

where the $\hat{\mu}_{(w)}^{\mathcal{I}_2}(\cdot)$ and $\hat{e}^{\mathcal{I}_2}(\cdot)$ are estimates of $\mu_{(w)}(\cdot)$ and $e(\cdot)$ obtained using only the half-sample $\mathcal{I}_2$, and $\hat{\tau}^{\mathcal{I}_2}$ is defined analogously (with the roles of $\mathcal{I}_1$ and $\mathcal{I}_2$ swapped). In other words, $\hat{\tau}^{\mathcal{I}_1}$ is a treatment effect estimator on $\mathcal{I}_1$ that uses $\mathcal{I}_2$ to estimate its nuisance components, and vice-versa.

This cross-estimation construction allows us to, asymptotically, ignore the idiosyncrasies of the specific machine learning adjustment we chose to use, and to simply rely on the following high-level conditions:

1. Overlap: The true propensity score is bounded away from 0 and 1, such that $\eta < e(x) < 1 - \eta$ for all $x \in \mathcal{X}$.

2. Consistency: All machine learning adjustments are sup-norm consistent,

$$\sup_{x \in \mathcal{X}} \left| \hat{\mu}_{(w)}^{\mathcal{I}_2}(x) - \mu_{(w)}(x) \right|, \quad \sup_{x \in \mathcal{X}} \left| \hat{e}^{\mathcal{I}_2}(x) - e(x) \right| \to_p 0.$$

3. Risk decay: The product of the errors for the outcome and propensity models decays as

$$\mathbb{E}\left[\left(\hat{\mu}_{(w)}^{\mathcal{I}_2}(X) - \mu_{(w)}(X) \right)^2 \right] \mathbb{E}\left[\left(\hat{e}^{\mathcal{I}_2}(X) - e(X) \right)^2 \right] = o_P\left(\frac{1}{n} \right),$$

   where the randomness above is taken over both the training of $\hat{\mu}_{(w)}$ and $\hat{e}$ and the test example X. Note that if $\hat{\mu}_{(w)}$ and $\hat{e}$ were both attained the parametric "$\sqrt{n}$-consistent" rate, then the error product would be bounded as $\mathcal{O}(1/n^2)$. A simple say to satisfy this condition is to have all regression adjustments be $o_P(n^{-1/4})$ consistent in RMSE.

Note that none of these conditions depend on the internal structure of the machine learning method used. Moreover, (3) depends on the mean-squared error of the risk adjustments, and so justifies tuning the $\hat{\mu}_{(w)}$ and $\hat{e}$ estimates via cross-validation.

Given these assumptions, we characterize the cross-fitting estimator (5) by coupling it with the oracle efficient score estimator (3), i.e.,

$$\sqrt{n}\left(\hat{\tau}_{CF} - \hat{\tau}^* \right) \to_p 0. \tag{6}$$

To do so, we first note that we can write

$$\hat{\tau}^* = \frac{|\mathcal{I}_1|}{n} \hat{\tau}^{\mathcal{I}_1,*} + \frac{|\mathcal{I}_2|}{n} \hat{\tau}^{\mathcal{I}_2,*}$$

analogously to (5) (because $\hat{\tau}^*$ uses oracle nuisance components, the cross-fitting construction doesn't change anything for it). Moreover, we can decompose $\hat{\tau}^{\mathcal{I}_1}$ itself as

$$\begin{aligned}
\hat{\tau}^{\mathcal{I}_1} &= \hat{m}_{(1)}^{\mathcal{I}_1} - \hat{m}_{(0)}^{\mathcal{I}_1}, \\
\hat{m}_{(1)}^{\mathcal{I}_1} &= \frac{1}{|\mathcal{I}_1|} \sum_{i \in \mathcal{I}_1} \left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i) + W_i \frac{Y_i - \hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i)}{\hat{e}^{\mathcal{I}_2}(X_i)} \right),
\end{aligned} \tag{7}$$

etc., and define $\hat{m}_{(0)}^{\mathcal{I}_1,*}$ and $\hat{m}_{(1)}^{\mathcal{I}_1,*}$ analogously. Given this buildup, in order to verify (6), it suffices to show that

$$\sqrt{n} \left(\hat{m}_{(1)}^{\mathcal{I}_1} - \hat{m}_{(1)}^{\mathcal{I}_1,*} \right) \to_p 0, \tag{8}$$

etc., across folds and treatment statuses.

We now study the term in (8) by decomposing it as follows:

$$\begin{aligned}
&\hat{m}_{(1)}^{\mathcal{I}_1} - \hat{m}_{(1)}^{\mathcal{I}_1,*} \\
&= \frac{1}{|\mathcal{I}_1|} \sum_{i \in \mathcal{I}_1} \left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i) + W_i \frac{Y_i - \hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i)}{\hat{e}^{\mathcal{I}_2}(X_i)} - \mu_{(1)}(X_i) - W_i \frac{Y_i - \mu_{(1)}(X_i)}{e(X_i)} \right) \\
&= \frac{1}{|\mathcal{I}_1|} \sum_{i \in \mathcal{I}_1} \left(\left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i) - \mu_{(1)}(X_i) \right) \left(1 - \frac{W_i}{e(X_i)} \right) \right) \\
&\quad + \frac{1}{|\mathcal{I}_1|} \sum_{i \in \mathcal{I}_1} W_i \left(\left(Y_i - \mu_{(1)}(X_i) \right) \left(\frac{1}{\hat{e}^{\mathcal{I}_2}(X_i)} - \frac{1}{e(X_i)} \right) \right) \\
&\quad - \frac{1}{|\mathcal{I}_1|} \sum_{i \in \mathcal{I}_1} W_i \left(\left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i) - \mu_{(1)}(X_i) \right) \left(\frac{1}{\hat{e}^{\mathcal{I}_2}(X_i)} - \frac{1}{e(X_i)} \right) \right)
\end{aligned}$$

Now, we can verify that these are small for different reasons. For the first term, we intricately use the fact that, thanks to our double machine learning construction, $\hat{\mu}_{(w)}^{\mathcal{I}_2}$ can effectively be treated as deterministic. Thus after conditioning on $\mathcal{I}_2$, the summands used to build this term become mean-zero and

independent (2nd and 3rd equalities below)

$$
\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(\left(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i)-\mu_{(1)}(X_i)\right)\left(1-\frac{W_i}{e(X_i)}\right)\right)\right)^2\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(\left(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i)-\mu_{(1)}(X_i)\right)\left(1-\frac{W_i}{e(X_i)}\right)\right)\right)^2 \,\middle|\, \mathcal{I}_2\right]\right] \\
&= \mathbb{E}\left[\mathrm{Var}\left[\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(\left(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i)-\mu_{(1)}(X_i)\right)\left(1-\frac{W_i}{e(X_i)}\right)\right) \,\middle|\, \mathcal{I}_2\right]\right] \\
&= \frac{1}{|\mathcal{I}_1|}\mathbb{E}\left[\mathrm{Var}\left[\left(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i)-\mu_{(1)}(X_i)\right)\left(1-\frac{W_i}{e(X_i)}\right) \,\middle|\, \mathcal{I}_2\right]\right] \\
&= \frac{1}{|\mathcal{I}_1|}\mathbb{E}\left[\mathbb{E}\left[\left(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i)-\mu_{(1)}(X_i)\right)^2\left(\frac{1}{e(X_i)}-1\right) \,\middle|\, \mathcal{I}_2\right]\right] \\
&= \frac{1}{\eta\,|\mathcal{I}_1|}\mathbb{E}\left[\left(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i)-\mu_{(1)}(X_i)\right)^2\right] = \frac{o_P(1)}{n}
\end{aligned}
$$

by consistency (2), because $\mathcal{I}_1 \sim n/2$. The key step in this argument was the 3rd equality: Because the summands become independent and mean-zero after conditioning, we "earn" a factor $1/|\mathcal{I}_1|$ due to concentration of iid sums. The second summand in our decomposition here can also be bounded similarly (thanks to overlap). Finally, for the last summand, we simply use Cauchy-Schwarz:

$$
\begin{aligned}
&\frac{1}{|\mathcal{I}_1|}\sum_{\{i:i\in\mathcal{I}_1,\,W_i=1\}}\left(\left(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i)-\mu_{(1)}(X_i)\right)\left(\frac{1}{\hat{e}^{\mathcal{I}_2}(X_i)}-\frac{1}{e(X_i)}\right)\right) \\
&\quad\leq \sqrt{\frac{1}{|\mathcal{I}_1|}\sum_{\{i:i\in\mathcal{I}_1,\,W_i=1\}}\left(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i)-\mu_{(1)}(X_i)\right)^2} \\
&\qquad\times\sqrt{\frac{1}{|\mathcal{I}_1|}\sum_{\{i:i\in\mathcal{I}_1,\,W_i=1\}}\left(\frac{1}{\hat{e}^{\mathcal{I}_2}(X_i)}-\frac{1}{e(X_i)}\right)^2} = o_P\left(\frac{1}{\sqrt{n}}\right)
\end{aligned}
$$

by risk decay (3). (To establish this fact, also note that by consistency (2), the estimated propensities will all eventually also be uniformly bounded away from 0, $\eta/2 \leq \hat{e}^{\mathcal{I}_2}(X_i) \leq 1-\eta/2$, and so the MSE for the inverse weights decays at the same rate as the MSE for the propensities themselves.)

The upshot is that by using cross-fitting, we can transform any $o_P(n^{-1/4})$-consistent machine learning method into an efficient ATE estimator. Also,

the proof was remarkably short (at least compared to a typical proof in the semiparametric efficiency literature).

Condensed notation. We will be encountering cross-fit estimators frequently in this class. From now on, we'll use the following notation: We define the data into K folds (above, $K = 2$), and compute estimators $\hat{\mu}_{(w)}^{(-k)}(x)$, etc., excluding the k-th fold. Then, writing $k(i)$ as the mapping that takes an observation and puts it into one of the k folds, we can write

$$\hat{\tau}_{CF} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_{(1)}^{(-k(i))}(X_i) - \hat{\mu}_{(0)}^{(-k(i))}(X_i) + W_i\frac{Y_i - \hat{\mu}_{(1)}^{(-k(i))}(X_i)}{\hat{e}^{(-k(i))}(X_i)} - (1-W_i)\frac{Y_i - \hat{\mu}_{(0)}^{(-k(i))}(X_i)}{1-\hat{e}^{(-k(i))}(X_i)}\right), \tag{9}$$

which (almost) fits on one line.

Confidence intervals. It is also important to be able to quantify uncertainty of treatment effect estimates. Cross-fitting also makes this easy. Recall from last class that the empirical variance of the efficient score converges to the efficient variance V_*:

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(\mu_{(1)}(X_i) - \mu_{(0)}(X_i) + W_i\frac{Y_i - \mu_{(1)}(X_i)}{e(X_i)} - (1-W_i)\frac{Y_i - \mu_{(0)}(X_i)}{1-e(X_i)} - \hat{\tau}^*\right)^2 \to_p V^*, \tag{10}$$

where $\hat{\tau}^*$ is as in (3). Our previous derivation then establishes that the same holds for cross-fitting:

$$\widehat{V}_{CF} := \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{\mu}_{(1)}^{(-k(i))}(X_i) - \hat{\mu}_{(0)}^{(-k(i))}(X_i) + W_i\frac{Y_i - \hat{\mu}_{(1)}^{(-k(i))}(X_i)}{\hat{e}^{(-k(i))}(X_i)} - (1-W_i)\frac{Y_i - \hat{\mu}_{(0)}^{(-k(i))}(X_i)}{1-\hat{e}^{(-k(i))}(X_i)} - \hat{\tau}_{CF}\right) \to_p V^*. \tag{11}$$

We can thus produce level-α confidence intervals for τ as

$$\tau \in \left(\hat{\tau}_{CF} \pm \frac{1}{\sqrt{n}}\Phi^{-1}(1-\frac{\alpha}{2})\sqrt{\widehat{V}_{CF}}\right),$$

where $\Phi(\cdot)$ is the standard Gaussian CDF, and these will achieve coverage with probability $1-\alpha$ in large samples.

Estimating the average treatment effect of the treated. Recall that another relevant estimand is the average treatment effect on the treated (ATT), $\tau_{ATT} = \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, W_i = 1\right]$. This problem looks slightly different than the ATE estimation problem, because now $Y_i(1)$ is observed for all the observations of interest, whereas $Y_i(0)$ is always missing. However, an analogous theory still holds. The efficient score estimator is

$$\begin{aligned}\hat{\tau}^*_{ATT} = &\frac{1}{|\{W_i = 1\}|} \sum_{\{i:W_i=1\}} \big(Y_i - \mu_{(0)}(X_i)\big) \\ &- \frac{1}{|\{W_i = 0\}|} \frac{\mathbb{P}\left[W_i = 0\right]}{\mathbb{P}\left[W_i = 1\right]} \sum_{\{i:W_i=0\}} \big(Y_i - \mu_{(0)}(X_i)\big) \frac{e(X_i)}{(1-e(X_i))},\end{aligned} \tag{12}$$

and we can turn this into a feasible cross-fit estimator as above,

$$\begin{aligned}\hat{\tau}_{ATT} = &\frac{1}{|\{W_i = 1\}|} \sum_{\{i:W_i=1\}} \left(Y_i - \mu^{(-k(i))}_{(0)}(X_i)\right) \\ &- \sum_{\{i:W_i=0\}} \left(Y_i - \hat{\mu}^{(-k(i))}_{(0)}(X_i)\right) \frac{\hat{e}^{(-k(i))}(X_i)}{(1-\hat{e}^{(-k(i))}(X_i))} \Bigg/ \sum_{\{i:W_i=0\}} \frac{\hat{e}^{(-k(i))}(X_i)}{(1-\hat{e}^{(-k(i))}(X_i))}.\end{aligned} \tag{13}$$

Closing thoughts. People often ask whether using machine learning methods for causal inference necessarily means that our analysis becomes "uninterpretable." However, from a certain perspective, the results shown here may provide some counter evidence. We used "heavy" machine learning to obtain our estimates for $\hat{\mu}_{(w)}(x)$ and $\hat{e}(x)$—these methods were treated as pure black boxes, and we never looked inside—and yet the scientific questions we are trying to answer remain just as crisp as before (i.e., we want the ATE or the ATT). Perhaps our results even got *more* interpretable (or, at least, credible), because we did not need to rely on linear or logistic specification to build our estimators for τ.

For further reading, see the following:

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. **Double/debiased machine learning for treatment and structural parameters.** *The Econometrics Journal*, 2017.

Imbens, Guido. **Nonparametric estimation of average treatment effects under exogeneity: A review.** *The Review of Economics and Statistics*, 86(1), 2004.

Hahn, Jinyong. **On the role of the propensity score in efficient semiparametric estimation of average treatment effects.** *Econometrica*, 1998.

van der Laan, Mark, and Sherri Rose. *Targeted learning: Causal inference for observational and experimental data.* **Springer Science & Business Media**, 2011.