OIT 661: Lecture 6
Learning Treatment Policies
(for class use only, please do not distribute)

Recap. 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:

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

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 $\sigma^2(x) = \text{Var}\left[Y_i(w) \,\middle|\, X_i = x\right]$ does not depend on w.

In the first lectures our goal was estimate the average treatment effect, $\tau = \mathbb{E}\left[Y_i(1) - Y_i(0)\right]$, while last time we discussed estimation of the full conditional average treatment effect function $\tau(x)$. In this lecture, we move away from pure estimation problems, and instead seek learn good decision rules that exploit treatment effect heterogeneity. This problem is clearly linked to that of heterogeneous treatment effect estimation (e.g., if we knew $\tau(x)$, then treating those with $\tau(x) > 0$ would be Bayes-optimal); however, in the face of statistical uncertainty, these two tasks may not be identical.

Policy learning. For our purposes, a treatment assignment $\pi(x)$ is a mapping

$$\pi : \mathcal{X} \to \{0,\, 1\}\,, \tag{2}$$

such that individuals with features $X_i = x$ get treated if and only if $\pi(x) = 1$. Our goal is to find a policy that maximizes expected utility which, assuming SUTVA, can be written as (today, we'll always consider Y_i to be a utility to avoid discussions of risk preferences, etc.)

$$\text{Utility}(\pi) = \mathbb{E}\left[Y_i\left(\pi(X_i)\right)\right]. \tag{3}$$

In the following, it will be convenient to rescale utility in terms of an improvement over a randomized treatment allocation:

$$\frac{1}{2}Q(\pi) = \mathbb{E}\left[Y_i\left(\pi(X_i)\right)\right] - \mathbb{E}\left[\frac{Y_i(0) + Y_i(1)}{2}\right], \tag{4}$$

where Q stands for the "quality" of the policy.

For any class of policies Π, the optimal policy π^* (if it exists) is defined as

$$\pi^* = \text{argmax}\left\{Q(\pi') : \pi' \in \Pi\right\}, \tag{5}$$

while the regret of any other policy is

$$R(\pi) = \sup\left\{Q(\pi') : \pi' \in \Pi\right\} - Q(\pi). \tag{6}$$

Our goal is to learn a policy with guaranteed bounds of $R(\hat{\pi})$; this criterion is called the minimax regret criterion.

Policy learning via empirical maximization. If the optimal policy π^* is a maximizer of the true quality function $Q(\pi)$ over $\pi \in \Pi$, then it is natural to learn $\hat{\pi}$ by maximizing an estimated quality function:

$$\hat{\pi} = \text{argmax}\left\{\widehat{Q}\left(\pi\right) : \pi \in \Pi\right\}. \tag{7}$$

If we know the treatment propensities $e(x)$, then it turns out we have access to a simple, unbiased choice for $\widehat{Q}(\pi)$:

$$\widehat{Q}_{IPW}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\left(2\pi(X_i) - 1\right)\left(\frac{W_i Y_i}{e(X_i)} - \frac{(1-W_i)Y_i}{1-e(X_i)}\right). \tag{8}$$

We can verify unbiasedness as follows:

$$\begin{aligned}
\mathbb{E}\left[\widehat{Q}_{IPW}(\pi)\right] &= \mathbb{E}\left[\left(2\pi(X_i) - 1\right)\left(\frac{W_i Y_i}{e(X_i)} - \frac{(1-W_i)Y_i}{1-e(X_i)}\right)\right] \\
&= \mathbb{E}\left[\left(2\pi(X_i) - 1\right)\mathbb{E}\left[\frac{W_i Y_i}{e(X_i)} - \frac{(1-W_i)Y_i}{1-e(X_i)} \,\middle|\, X_i\right]\right] \\
&= \mathbb{E}\left[\left(2\pi(X_i) - 1\right)\left(\mu_{(1)}(X_i) - \mu_{(0)}(X_i)\right)\right] \\
&= \mathbb{E}\left[2\mu_{(\pi(X_i))}(X_i) - \left(\mu_{(0)}(X_i) + \mu_{(1)}(X_i)\right)\right] \\
&= Q(\pi).
\end{aligned}$$

This is like an IPW estimator for the ATE, except we "earn" the treatment effect for the i-th sample when $\pi(X_i) = 1$, and "pay" the treatment effect when $\pi(X_i) = 0$.

Now, optimizing $\widehat{Q}_{IPW}(\pi)$ may seem like a daunting task. However, we can re-write the objective as

$$\begin{aligned}
\widehat{Q}_{IPW}(\pi) &= \frac{1}{n}\sum_{i=1}^{n}\underbrace{\left(2\pi(X_i) - 1\right)\text{sign}\left(\Gamma_i\right)}_{\text{classification objective}}\underbrace{\left|\Gamma_i\right|}_{\text{sample weight}}, \\
\Gamma_i &= \frac{W_i Y_i}{e(X_i)} - \frac{(1-W_i)Y_i}{1-e(X_i)}.
\end{aligned} \tag{9}$$

In other words, maximizing $\widehat{Q}_{IPW}(\pi)$ is equivalent to optimizing a weighted classification objective. From this connection, we directly obtain regret bounds for the learned policy. If we assume that $|Y_i| \leq M$ and $\eta \leq e(X_i) \leq 1-\eta$, such as to make the weights Γ_i bounded, and assume that Π has a bounded Vapnik-Chervonenkis dimension, then the regret of the learned policy $\hat{\pi}$ is bounded as

$$R\left(\hat{\pi}_{IPW}\right) = \mathcal{O}_P\left(\frac{M}{\eta}\sqrt{\frac{\text{VC}(\pi)}{n}}\right),\ \hat{\pi}_{IPW} = \text{argmax}_{\pi\in\Pi}\left\{\widehat{Q}_{IPW}\left(\pi\right)\right\}, \quad (10)$$

where $\text{VC}(\pi)$ denotes the VC-dimension of Π.

The role of the policy class Π. This problem setup may appear unusual. We started with a non-parametric model (i.e., $\mu_{(w)}(x)$ and $e(x)$ can be generic), and with a well-defined Bayes treatment assignment rule, $\pi_{bayes}(x) = 1\left(\{\tau(x) > 0\}\right)$. However, from this point, our goal was not to find a way to approximate $\pi_{bayes}(x)$; rather, given another, pre-specified class of policies Π, we want to learn a nearly regret-optimal representative from Π. For example, Π could consist of linear decision rules, k-sparse decision rules, depth-ℓ decision trees, etc. Note, in particular, that we never assumed that $\pi_{bayes}(\cdot) \in \Pi$.

The reason for this tension is that the features X_i play two distinct roles here. First, the X_i may be needed to achieve unconfoundedness

$$\{Y_i(0),\, Y_i(1)\} \perp\!\!\!\perp W_i \,\big|\, X_i. \quad (11)$$

In general, the more pre-treatment variables we have access to, the more plausible unconfoundedness becomes. In order to have a credible model of nature, it's good to have flexible, non-parametric models for $e(x)$ and $\mu_{(w)}(x)$ using a wide variety of features.

On the other hand, when we want to deploy a policy $\pi(\cdot)$, we should be much more careful about what features we use to make decisions and the form of the policy $\pi(\cdot)$:

- We should not use certain features, e.g., features that are difficult to measure in a deployed system, features that are gameable by participants in the system, or features that correspond to legally protected classes.
- We may have budget constraints (e.g., at most 15% of people get treated), or marginal budget constraints (e.g., the total amount of funds allocated to each state stays fixed, but we may re-prioritize funds within states).
- We may have functional form constraints on $\pi(\cdot)$ (e.g., if the policy needs to be communicated to employees in a non-electronic format, or audited using non-quantitative methods).

Given any such constraints set by a practitioner, we can construct a class of allowable policies Π that respects these feature exclusion, budget, and functional form constraints.

Efficient policy evaluation and learning. Although the IPW policy learning method discussed above has some nice properties (e.g., $\sqrt{n}$ regret consistency), we may still ask whether it is the best possible such method. To get a better understanding of this issue, it is helpful to turn back to our discussions of ATE estimation.

In order to learn a good policy $\hat{\pi}$, it is intuitively helpful to start with a good method $\widehat{Q}(\pi)$ for evaluating the quality of individual policies π. And here, we can start by noting that

$$\begin{aligned} Q(\pi) &= 2\mathbb{E}\left[Y_i\left(\pi(X_i)\right)\right] - \mathbb{E}\left[Y_i(0) + Y_i(1)\right] \\ &= \mathbb{E}\left[Y_i\left(\pi(X_i)\right)\right] - \mathbb{E}\left[Y_i\left(1 - \pi(X_i)\right)\right]. \end{aligned} \tag{12}$$

In other words, $Q(\pi)$ is the ATE in an experiment where we compare deploying the policy $\pi(\cdot)$ to and experiment where we always deploy the *opposite* of $\pi(\cdot)$.

Now, given this formulation as an ATE estimation problem, we know that the oracle IPW estimator is OK, but not efficient. Moreover, we know that the oracle estimator $\widehat{Q}^*(\pi)$ that estimates $Q(\pi)$ by averaging an efficient score attains the semiparametric efficiency bound; and, in our case, $\widehat{Q}^*(\pi)$ is

$$\begin{aligned} &\widehat{Q}^*(\pi) = \frac{1}{n}\sum_{i=1}^{n}\left(2\pi(x) - 1\right)\Gamma_i^*, \\ &\Gamma_i^* := \mu_{(1)}(X_i) - \mu_{(0)}(X_i) + W_i\frac{Y_i - \mu_{(1)}(X_i)}{e(X_i)} - (1 - W_i)\frac{Y_i - \mu_{(0)}(X_i)}{1 - e(X_i)}. \end{aligned} \tag{13}$$

Moreover, assuming the existence of $o_P(n^{-1/4})$-consistent regression adjustments for $\hat{\mu}_{(w)}(x)$ and $\hat{e}(x)$ we can construct a double machine learning estimator that emulates the efficient oracle:

$$\begin{aligned} &\widehat{Q}_{DML}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\left(2\pi(x) - 1\right)\widehat{\Gamma}_i, \\ &\widehat{\Gamma}_i := \hat{\mu}_{(1)}^{(-k(i))}(X_i) - \hat{\mu}_{(0)}^{(-k(i))}(X_i) \\ &\qquad + W_i\frac{Y_i - \hat{\mu}_{(1)}^{(-k(i))}(X_i)}{\hat{e}^{(-k(i))}(X_i)} - (1 - W_i)\frac{Y_i - \hat{\mu}_{(0)}^{(-k(i))}(X_i)}{1 - \hat{e}^{(-k(i))}(X_i)}, \end{aligned} \tag{14}$$

and note that this is also a weighted classification objective.

We already know from lecture 4 that $\widehat{Q}_{DML}(\pi)$ is pointwise asymptotically equivalent to $\widehat{Q}^*(\pi)$, i.e., for any fixed policy π the difference between the two quantities decays faster than $1/\sqrt{n}$. However, more is true: If Π is a VC class, then

$$\sqrt{n}\sup\left\{\left|\widehat{Q}_{DML}(\pi) - \widehat{Q}^*(\pi)\right| : \pi \in \Pi\right\} \to_p 0. \tag{15}$$

This result, along with an empirical process theory argument (outside the scope of this class) then imply that the regret of double machine learning policy evaluation is bounded on the order of

$$\begin{aligned} &R\left(\hat{\pi}_{DML}\right) = \mathcal{O}_P\left(\sqrt{\frac{V^* \operatorname{VC}(\pi)}{n}}\right), \quad \hat{\pi}_{DML} = \operatorname{argmax}_{\pi\in\Pi}\left\{\widehat{Q}_{DML}\left(\pi\right)\right\}, \\ &V^* = \mathbb{E}\left[\tau^2(X_i)\right] + \mathbb{E}\left[\frac{\sigma^2(X_i)}{e(X_i)(1-e(X_i))}\right]. \end{aligned} \tag{16}$$

See Athey & W. (2017) for details, as well as lower bounds. Effectively, the above bound is optimal in a regime where treatment effects just barely peak out of the noise.

Closing thoughts. In relation to last lecture, the prescriptions from this lecture may seem almost paradoxical. One might think that a good way to learn policies is the following:

1. Get a good estimate $\hat{\tau}(x)$ of the heterogeneous treatment effect, and then
2. Treat those people with $\hat{\tau}(x) > 0$.

However, the bounds seen today look more like ATE-estimation: Effectively, we evaluate each policy individually (this is just an ATE problem), and then optimize the result. In other words, we have made the problem of policy learning look like many ATE problems, rather than a single HTE estimation problem. Two comments are in line.

First, this result is intricately tied to the fact that we want to learn regret-optimal policies over "simple" classes Π (essentially, finite-dimensional classes). If we wanted to learn non-parametric policies, the problem would start to look more like a HTE estimation problem (see Hirano & Porter for details).

Second, even though the approach for learning $\hat{\pi}_{DML}$ superficially has nothing to do with HTE estimation, we can still algorithmically introduce a connection. In the definition of $\widehat{Q}_{DML}(\pi)$, we had a term

$$\hat{\mu}_{(1)}^{(-k(i))}(X_i) - \hat{\mu}_{(0)}(X_i)^{(-k(i))},$$

which of course can be interpreted as a HTE estimate $\hat{\tau}^{(-k(i))}(X_i)$. Although our regret bounds don't depend on the accuracy of this HTE estimator, nothing stops us from making sure we choose the $\hat{\mu}_{(w)}(X_i)$ such as to make the resulting $\hat{\tau}(X_i)$ estimator correspond to one of the "good" methods seen in last lecture. In practice, this appears to be helpful.

Further reading. Athey & W. have results on the connections between efficient policy learning and optimal regret bounds. Hiarano & Porter develop asymptotic minimax theory in the case where there are no constraints on Π, so we're just trying to attain the quality of the Bayes decision rule. Swaminathan & Joachims generalize the IPW-learning results to a multi-class setting.

Athey, Susan, and Stefan Wager. **Efficient policy learning.** arXiv:1702.02896, 2017.

Hirano, Keisuke, and Jack R. Porter. **Asymptotics for statistical treatment rules.** *Econometrica*, 77(5), 2009.

Swaminathan, Adith, and Thorsten Joachims. **Batch learning from logged bandit feedback through counterfactual risk minimization.** *Journal of Machine Learning Research*, 16, 2015.