[Lecture 21]

0.1 Doubly-Robust Estimators

Recall that in the previous section we defined the inverse propensity weighted estimator

$$\mathbb{E}_{\tilde{p}}[Y(1) - Y(0)] = \mathbb{E}_{\tilde{p}}\Big[\Big(\frac{\mathbb{I}[T=1]}{\tilde{p}(T \mid X)} - \frac{\mathbb{I}[T=0]}{\tilde{p}(T \mid X)}\Big)Y(T)\Big]. \tag{1}$$

To actually estimate the left-hand-side, we take the empirical average over n samples.

There are a couple of downsides of this estimator. One is that the variance of this estimator can be large. Specifically, we can compute it as

$$\frac{1}{n}\Big(\mathbb{E}_{\tilde{p}}\Big[\frac{1}{\tilde{p}(T=1 \mid X)}Y(1)^2 + \frac{1}{\tilde{p}(T=0 \mid X)}Y(0)^2\Big] - \mathbb{E}_{\tilde{p}}[Y(1) - Y(0)]^2\Big). \tag{2}$$

If the propensity weights are near zero then the variance explodes (similarly to the issue with χ^2-divergence that we saw earlier).

Another issue is that estimating the propensity weights themselves is non-trivial, and if we use the wrong propensity weights, then the estimate could be arbitrarily wrong.

We will explore an idea that partially mitigates both issues; it reduces the variance when the propensity weights are correct (although doesn't avoid the exploding variance issue), and in some cases it produces a correct estimate even if the propensity weights are wrong.

The basic idea is as follows: suppose that we have some prediction $\bar{Y}(1, X)$, $\bar{Y}(0, X)$ of what will happen under $T = 1$, $T = 0$ conditioned on X. Since these predictions only require knowing X and not T, an alternate estimate of the treatment effect can be obtained by adding and subtracting $\bar{Y}$:

$$\mathbb{E}_{\tilde{p}}[Y(1) - Y(0)] = \mathbb{E}_{\tilde{p}}[\bar{Y}(1, X) - \bar{Y}(0, X)] + \mathbb{E}_{\tilde{p}}[(Y(1) - \bar{Y}(1, X)) - (Y(0) - \bar{Y}(0, X))] \tag{3}$$

$$= \mathbb{E}_{\tilde{p}}[\bar{Y}(1, X) - \bar{Y}(0, X)] + \mathbb{E}_{\tilde{p}}\Big[\Big(\frac{\mathbb{I}[T=1]}{\tilde{p}(T \mid X)} - \frac{\mathbb{I}[T=0]}{\tilde{p}(T \mid X)}\Big)(Y(T) - \bar{Y}(T, X))\Big]. \tag{4}$$

In other words, we first use our prediction $\bar{Y}$ to form a guess of the average treatment effect, then use inverse propensity weighting to correct the guess so as to obtain an unbiased estimate. This can yield substantial improvements when $Y(T) - \bar{Y}(T, X)$ is much smaller in magnitude than $Y(T)$. For instance, a patient's cholesterol after taking a cholesterol-reducing drug is still highly-correlated with their initial cholesterol, so in that case we can take $\bar{Y}(T, X)$ to be the pre-treatment cholesterol level. Even though this is independent of T it can substantially reduce the variance of the estimate! (We will formally bound the variance below.)

Bias of the estimate. Call the prediction $\bar{Y}$ unbiased if $\mathbb{E}[Y \mid X, T] = \bar{Y}(T, X)$. The first key property of (4) is that it is unbiased as long as *either* $\bar{Y}$ is unbiased, or the propensity weights are correct. Indeed, if the prediction is unbiased then the first term is the average treatment effect while the second term is zero. Conversely, if the propensity weights are correct then the second term exactly estimates the difference between the predicted and true treatment effect. Correspondingly, (4) is called a *doubly-robust estimator*.

We can actually say more about the bias. Suppose that instead of the true propensity weights, we have an incorrect guess $\hat{p}(t \mid x)$. Then the bias of the estimate is the difference between $\mathbb{E}_{\tilde{p}}[Y(1) - Y(0)]$ and (4), which is

$$\mathbb{E}_{\tilde{p}}[Y(1) - Y(0)] - \mathbb{E}[\bar{Y}(1, X) - \bar{Y}(0, X)] - \mathbb{E}_{\tilde{p}}\Big[\Big(\frac{\mathbb{I}[T=1]}{\hat{p}(T \mid X)} - \frac{\mathbb{I}[T=0]}{\hat{p}(T \mid X)}\Big)(Y(T) - \bar{Y}(T, X))\Big] \tag{5}$$

$$= \mathbb{E}_{\tilde{p}}\Big[(Y(1) - \bar{Y}(1, X))\Big(1 - \frac{\mathbb{I}[T=1]}{\hat{p}(t=1 \mid X)}\Big) + (Y(0) - \bar{Y}(0, X))\Big(1 - \frac{\mathbb{I}[T=0]}{\hat{p}(t=0 \mid X)}\Big)\Big]. \tag{6}$$

Focusing on the first term, and using the independence of T and Y conditioned on X, we have

$$\mathbb{E}_{\tilde{p}}\Big[(Y(1) - \bar{Y}(1,X))\Big(1 - \frac{\mathbb{I}[T=1]}{\hat{p}(t=1 \mid X)}\Big)\Big] \tag{7}$$

$$= \mathbb{E}_{\tilde{p}}\Big[(\mathbb{E}[Y(1) \mid X] - \bar{Y}(1,X))\Big(1 - \frac{\tilde{p}(t=1 \mid X)}{\hat{p}(t=1 \mid X)}\Big) \mid X\Big] \tag{8}$$

$$\leq \mathbb{E}_{\tilde{p}}[(\mathbb{E}[Y(1) \mid X] - \bar{Y}(1,X))^2]^{1/2}\mathbb{E}_{\tilde{p}}\Big[\Big(1 - \frac{\tilde{p}(t=1 \mid X)}{\hat{p}(t=1 \mid X)}\Big)^2\Big]^{1/2}, \tag{9}$$

meaning that the bias of the estimator is the *product* of the biases of $\bar{Y}$ and $\hat{p}$ (measured as the expected squared errors in (9)).

Variance of the estimate. We can obtain a somewhat similar relation for the variance. Usually the variance of $\bar{Y}(1,X) - \bar{Y}(0,X)$ is small compared to the propensity-weighted term, so again focusing on the $T=1$ case we have

$$\mathsf{Var}\Big[(Y(1) - \bar{Y}(1,X))\frac{\mathbb{I}[T=1]}{\hat{p}(t=1 \mid X)}\Big] \leq \mathbb{E}\Big[\mathbb{E}_Y[(Y(1) - \bar{Y}(1,X))^2 \mid X]\frac{\tilde{p}(t=1 \mid X)}{\hat{p}(t=1 \mid X)^2}\Big]. \tag{10}$$

The variance is substantially reduced when $Y(1)$ is close to $\bar{Y}(1,X)$. We cannot always hope for this, e.g. if Y has a large amount of intrinsic variance even conditioned on X. But in many cases even trivial $\bar{Y}$ can predict most of the variance in Y—for instance, for any chronic disease the patient's post-treatment status is well-predicted by their pre-treatment status. And the value of a stock tomorrow is well-predicted by its value today.

Semi-parametric estimation. It may seem difficult to estimate $\bar{Y}(\cdot, X)$ and $\tilde{p}(t=1 \mid X)$, since any parametric model could be mis-specified and lead to biased estimates. One idea is to estimate these both non-parametrically, and then apply the doubly-robust estimator above. This is an instance of *semi-parametric estimation*, because while we estimate $\bar{Y}$ and $\tilde{p}(t \mid X)$ non-parametrically, the doubly-robust estimator itself is parametric (i.e a simple sample estimate of the mean), and in some cases we obtain non-parametric rates. This is explored in detail in **?** for estimating conditional average treatment effects; we describe the basic idea here. Since the squared error in an estimate is $\mathsf{Bias}^2 + \mathsf{Variance}/n$, the bias will dominate the error in the doubly-robust estimator as long as the variance doesn't explode (of course, the variance can often explode if the propensity weights are too close to 0 or 1, and the following idea won't help in that case).

We saw above that the bias of the doubly-robust estimator is the product of the biases in $\bar{Y}$ and $\hat{p}$, which are both given as expected squared errors between the true and estimated value. In non-parametric estimation, we typically get convergence rates of $\mathcal{O}(n^{-\alpha})$ for some $\alpha < 1/2$ (note that $\alpha = 1/2$ is what we typically get for parametric estimation). The parameter α typically depends on the dimension of the problem and the smoothness of the function class we wish to estimate. Since the doubly-robust bias is the product of the biases, we end up with a bias of $\mathcal{O}(n^{-2\alpha})$ as long as $\bar{Y}$ and $\hat{p}$ each converge at a $n^{-\alpha}$ rate. As long as $\alpha > 1/4$, this yields a parametric rate (the variance term will then asymptotically dominate as it only converges at $1/\sqrt{n}$).