Control Variates

David S. Rosenberg

1 Introduction

Suppose we observe $X \in \mathcal{X}$ and $Y \in \mathbb{R}$, where (X, Y) has some unknown joint distribution. Our goal is to estimate $\mathbb{E}Y$. Perhaps the simplest estimator of $\mathbb{E}Y$ is just Y. It's unbiased with variance $\text{Var}(Y)$. Is there anything we can do with X to reduce the variance?

Suppose we have a function $f : \mathcal{X} \to \mathbb{R}$, and we think that $f(X)$ gives a reasonable prediction for Y. Perhaps f is based on some prior research or a common-sense guess. We also assume that we know $\mathbb{E}f(X)$. For example, we can compute $\mathbb{E}f(X)$ if know the marginal distribution of X, which is sometimes a reasonable assumption. Is there some way to use $f(X)$ to get a better estimate of $\mathbb{E}Y$?

Consider the following decomposition of Y:

$$Y \;\; = \;\; Y - f(X) + f(X) - \mathbb{E}f(X) + \mathbb{E}f(X)$$

What's interesting to note is that the $f(X) - \mathbb{E}f(X)$ in the middle of this decomposition has expectation 0. That means we can remove those terms without changing the expectation. So if we let

$$\hat{\mu} = \hat{\mu}(X, Y) = Y - f(X) + \mathbb{E}f(X),$$

then $\mathbb{E}\left[\hat{\mu}\right] = \mathbb{E}Y$ and so $\hat{\mu}$ is a new unbiased estimator for $\mathbb{E}Y$. The variance of this estimator is

$$\text{Var}\left(\hat{\mu}\right) = \text{Var}\left(Y - f(X)\right).$$

So does $\hat{\mu}$ have lower variance than Y? Sometimes yes, sometimes no.

Consider the extreme case that f predicts Y perfectly, i.e. $Y = f(X)$. Then our new estimator has $\text{Var}\left(\hat{\mu}\right) = 0$, which is great. On the other hand, if $f(X) = -Y$, then

$$\text{Var}\left(\hat{\mu}\right) = \text{Var}\left(Y - f(X)\right) = \text{Var}\left(2Y\right) = 4\text{Var}\left(Y\right),$$

which has much larger variance than Y.

In the sense of MSE, we know that $x \mapsto \mathbb{E}[Y \mid X = x]$ is the best function for approximating Y given $X = x$. Suppose we let $f(x) = \mathbb{E}[Y \mid X = x]$. Then we get

$$\text{Var}(\hat{\mu}) \quad = \quad \text{Var}(Y - \mathbb{E}[Y \mid X]).$$

Recall that the projection-residual decomposition of the variance of Y is

$$\text{Var}Y = \text{Var}(Y - \mathbb{E}[Y \mid X]) + \text{Var}(\mathbb{E}[Y \mid X]).$$

So we see that if we can take $f(x) = \mathbb{E}[Y \mid X = x]$, then we reduce the variance of our estimator by $\text{Var}(\mathbb{E}[Y \mid X])$. This is the amount of variation in Y that we can account for with X.

In the setting described above, $f(X)$ is called a **control variate**. A control variate is a random variable with **known expectation** [Owe13, Sec 8.9]. To be effective in creating lower-variance estimators of $\mathbb{E}Y$, we need $f(X) \approx Y$.

Exercise 1. Suppose $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. from an unknown distribution. Suppose we have a function $f(x)$ with known $\mathbb{E}f(X)$. Using f as a control variate, give an estimator for $\mathbb{E}Y$ in terms of our sample of size n, and derive an expression for the variance. Compare this to the variance of the naive estimator $\hat{\mu}_{\text{mean}} = (Y_1 + \cdots + Y_n)/n$.

Solution 1. Building on the calculations above, let

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i) + \mathbb{E}f(X)).$$

Then

$$\text{Var}(\hat{\mu}) = \frac{1}{n}\text{Var}(Y_i - f(X_i))$$

compared to

$$\text{Var}(\hat{\mu}_{\text{mean}}) = \frac{1}{n}\text{Var}(Y_i).$$

Exercise 2. Show the estimator $\hat{\mu} = Y - f(X) + \mathbb{E}f(X)$ with control variate $f(X)$ is unchanged if we replace $f(x)$ by $f'(x) = f(x) + c$. How do we interpret this?

Solution 2. This is trivial: $\hat{\mu}' = Y - f'(X) + \mathbb{E}f'(X) = Y - f(X) - c + \mathbb{E}[f(x) + c] = \hat{\mu}$. The interpretation is simply that, since we're always subtracting off the mean of $f(X)$, any additive shifts are irrelevant.

2 Empirical Example

To illustrate how a control variate works in practice, let's consider the following joint distribution of (X, Y):

$$\begin{aligned} X &\sim \text{Unif}[0, 6] \\ Y \mid X &\sim \mathcal{N}\left(6 + 1.3\sin(X), \left[.3 + \frac{1}{4}|3 - X|\right]^2\right) \end{aligned}$$

For simplicity, we're going to consider the scenario in which we get a sample of size $n = 1$ from this distribution, call it (X, Y). We want to use this sample to estimate $\mathbb{E}Y$. To get our control variate $f(x)$, we'll take a preliminary sample of size $n = 100$ from the distribution and fit a simple regression tree model. Using this same data, we estimate $\mathbb{E}f(X)$, which we'll write as $\hat{\mathbb{E}}f(X)$.

Figure 1 has a plot of a sample of size 1000 from the distribution. For each sampled point (x, y), we also plot $(x, \mathbb{E}[Y \mid X = x])$, which is the best prediction of Y given just $X = x$. We also plot $f(x)$, which is the prediction we learned from the preliminary sample.

For each sample (X, Y), we form several estimators of $\mathbb{E}Y$. These include using Y directly (which ignores X), using the preliminary regression function $f(X)$ (which ignores Y), and using the estimator with control variate $Y - f(X) + \hat{\mathbb{E}}f(X)$ (which uses both X and Y). We also included an idealized version of this control variate estimator, where we assume that $\mathbb{E}f(X)$ is known (though this would be reasonable if we had knew the marginal distribution of X or had a large sample from that distribution). Finally, we used the even more idealized estimator in which we assume we know $\mathbb{E}[Y \mid X = x]$, and use that as a control variate. In Figure 2 we plot histograms for the estimates produces by these 5 estimators over the 1000 repetitions of samples of size $n = 1$. Note the relative variances of these distributions.

To make a more objective comparison of these 5 estimators, we increased the repetitions to 100000 and tabulated the mean, SD, bias, and RMSE of each estimator across the repetitions (1). There are several things to note in the table. First, as predicted by the theory, Y, $y - f(x) + \mathbb{E}[f(X)]$, and $y - \mathbb{E}[Y \mid X = x] + \mathbb{E}Y$ all have bias very close to 0. Using the estimate $\hat{\mathbb{E}}[f(X)]$ increases the bias significantly, but it's still so much smaller than the SD that the change in RMSE is not appreciable. The benefit of using X and Y in combination here is clear in the reduction of RMSE. Even though $f(x)$ is not a particularly good approximation of $\mathbb{E}[Y \mid X = x]$, it still gives an appreciable reduction in RMSE over using Y

Fig. 1: This plot shows the sampled (x, y) pairs, along with the conditional expectation and our regression tree predictions.

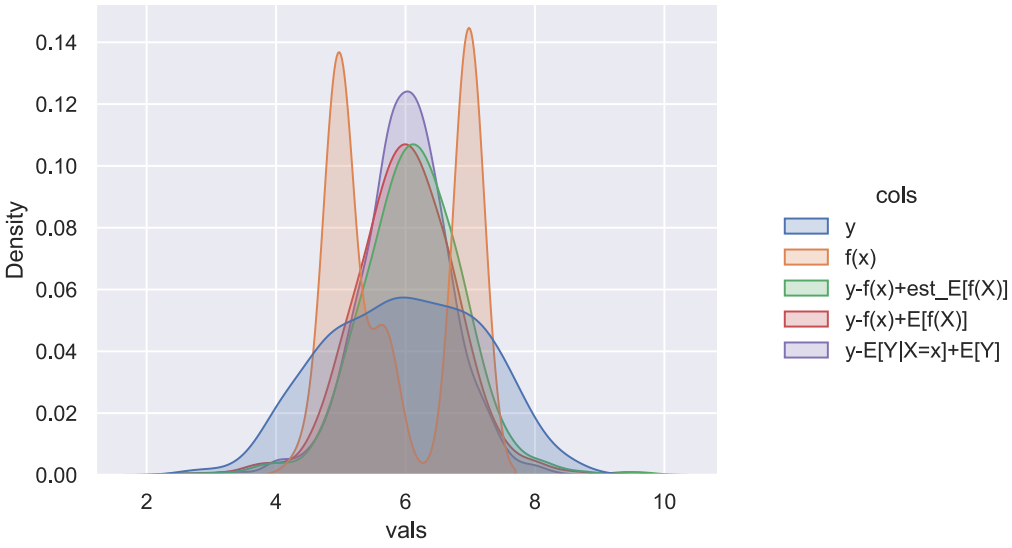

Fig. 2: This plot show histograms of various estimators of $\mathbb{E}Y \approx 6.008$.

estimator	mean	SD	bias	RMSE
y	6.009909	1.176118	0.001862	1.176119
$f(x)$	5.956776	0.951484	-0.051271	0.952864
$y - f(x) + \hat{\mathbb{E}}[f(X)]$	6.129088	0.808296	0.121041	0.817309
$y - f(x) + \mathbb{E}[f(X)]$	6.007692	0.808296	-0.000355	0.808296
$y - \mathbb{E}[Y \mid X = x] + \mathbb{E}Y$	6.009719	0.712861	0.001671	0.712863

Tab. 1: Performance results for various estimators of $\mathbb{E}Y = 5.96$ across $n = 300$ trials.

alone, and also improves significantly over $f(X)$.

Exercise 3. In Table 1, can you see a relationship between the the SD's of y, $f(x)$, and $y - f(x) + \hat{\mathbb{E}}[f(X)]$? (Hint: consider the corresponding variances.)

3 Control variate for IPW estimates in MAR setting

Again we assume that $X \in \mathcal{X}$ and $Y \in \mathbb{R}$ have some unknown joint distribution, and our objective is to estimate $\mathbb{E}Y$. However, this time Y is sometimes "missing". That is, sometimes we only observe X rather than the pair (X, Y). The probability that Y is observed depends on the value of the covariate X. We assume that the probability of observing Y is $\pi(X)$, for some function $\pi : \mathcal{X} \to (0, 1)$. The function $\pi(x)$ is called the **propensity score function**. We traditionally introduce an indicator variable $R = \mathbb{1}[Y \text{ is observed}]$, and we make the "missing at random" (MAR) assumption that Y and R are conditionally independent given X. This assumption is written as $Y \perp\!\!\!\perp R \mid X$. In this context, we can think of our observation as a triple (X, R, RY).

The **inverse propensity weighted** estimator of $\mathbb{E}Y$ is defined as

$$\hat{\mu}_{\text{ipw}} = \frac{RY}{\pi(X)}.$$

Note that we can always evaluate this, even when Y is missing, since in that case

$R = 0$ (and our estimate is 0 as well). This estimator is unbiased, since

$$\begin{aligned}
\mathbb{E}\hat{\mu}_{\text{ipw}} &= \mathbb{E}\left[\mathbb{E}\left[\frac{RY}{\pi(X)} \mid X\right]\right] \quad \text{Adam's Law} \\
&= \mathbb{E}\left[\frac{1}{\pi(X)}\mathbb{E}\left[RY \mid X\right]\right] \quad \text{Taking out what is known} \\
&= \mathbb{E}\left[\frac{1}{\pi(X)}\mathbb{E}\left[R \mid X\right]\mathbb{E}\left[Y \mid X\right]\right] \quad \text{By MAR assumption} \\
&= \mathbb{E}\left[\mathbb{E}\left[Y \mid X\right]\right] \quad \text{since } \mathbb{E}\left[R \mid X\right] = \mathbb{P}\left(R = 1 \mid X\right) = \pi(X) \\
&= \mathbb{E}Y \quad \text{Adam's Law}
\end{aligned}$$

As before, suppose we have a function $f : \mathcal{X} \to \mathbb{R}$, and we think that $f(X)$ gives a reasonable prediction for Y. As before, we assume we know (or can get a very good estimate of) $\mathbb{E}f(X)$. Can we again use f to get an estimator with lower variance?

Note that $Rf(X)/\pi(X)$ is a natural plug-in estimator for $RY/\pi(X)$. Let's try to use that as a control variate. We need to compute the expectation of the control variate:

$$\begin{aligned}
\mathbb{E}\left[\frac{Rf(X)}{\pi(X)}\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{Rf(X)}{\pi(X)} \mid X\right]\right] \quad \text{Adam's Law} \\
&= \mathbb{E}\left[\frac{f(X)}{\pi(X)}\mathbb{E}\left[R \mid X\right]\right] \quad \text{Taking out what is known} \\
&= \mathbb{E}\left[f(X)\right] \quad \text{Same argument as for } \mathbb{E}\hat{\mu}_{\text{ipw}}
\end{aligned}$$

So we propose the following estimator:

$$\hat{\mu}_{\text{ipw-cv}} = \frac{RY}{\pi(X)} - \frac{Rf(X)}{\pi(X)} + \mathbb{E}\left[f(X)\right],$$

where "cv" in the subscript refers to "control variate". Note that this estimator is unbiased, no matter how well or badly $f(x)$ predicts Y. We can also write this as

$$\hat{\mu}_{\text{ipw-cv}} = \begin{cases} \mathbb{E}\left[f(X)\right] & R = 0 \\ \frac{Y - f(X)}{\pi(X)} + \mathbb{E}\left[f(X)\right] & R = 1. \end{cases}$$

Here we see that if Y is missing, then we just give $\mathbb{E}\left[f(X)\right]$ as our estimate for $\mathbb{E}Y$, which seems reasonable. If Y is not missing, we "correct" $\mathbb{E}\left[f(X)\right]$ by

an amount related to the residual $Y - f(X)$. Why do we scale the residual by $1/\pi(X)$? Well... in the fully observed case where $\pi(X) = 1$, we can think of $\hat{\mu} = Y - f(X) + \mathbb{E}[f(X)]$ as correcting a prior estimate $\mathbb{E}[f(X)]$ by the residual $Y - f(X)$. In our missing data case, we can only make our correction when we observe Y. When we don't observe Y, we can only produce our baseline $\mathbb{E}[f(X)]$. In scaling by $1/\pi(X)$, it's like overcompensating the correction to make up for all the instances where we are not able to make the correction.

Exercise 4. Suppose $(X, R, RY), (X_1, R_1, R_1Y_1), \ldots, (X_n, R_n, R_nY_n)$ are i.i.d. in the MAR setting, for which $Y_i \perp\!\!\!\perp R_i \mid X_i$, and $\mathbb{P}(R = 1 \mid X) = \pi(X)$, for some known propensity function $\pi : \mathcal{X} \to (0, 1)$. Suppose we have a function $f(x)$ with known $\mathbb{E}f(X)$. Using f as a control variate, give an unbiased estimator for $\mathbb{E}Y$ in terms of our sample of size n.

Solution 3. Building on the calculations above, let

$$\begin{aligned}\hat{\mu}_{\text{ipw-cv}} &= \mathbb{E}[f(X)] + \frac{1}{n}\sum_{i=1}^{n}\left(\frac{R_iY_i}{\pi(X_i)} - \frac{R_if(X_i)}{\pi(X_i)}\right) \\ &= \mathbb{E}[f(X)] + \frac{1}{n}\sum_{i:R_i=1}\frac{Y_i - f(X_i)}{\pi(X_i)}.\end{aligned}$$

3.1 Practical issues and "doubly robust estimators"

In practice, we rarely have a good $f(x)$ before looking at the data. Standard practice is to estimate $f(x)$ from the dataset we have. Let's consider a parameterized class of functions $f(x; \theta) : \mathcal{X} \to \mathbb{R}$ for $\theta \in \mathbb{R}^d$. We can fit θ using least squares:

$$\hat{\theta} = \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} \left(f(X_i; \theta) - Y_i\right)^2.$$

We can then plug in $f(x, \hat{\theta})$ wherever we had $f(x)$.

Similarly, we generally don't know the propensity score function $\pi(x)$. This too can be estimated from the dataset we have. For example, we can use a logistic regression model $\pi(x; \gamma) : \mathcal{X} \to (0, 1)$ for $\gamma \in \mathbb{R}^d$ and use maximum likelihood to fit it:

$$\hat{\gamma} = \operatorname*{arg\,max}_{\gamma \in \mathbb{R}^d} \prod_{i=1}^{n} \left[\pi(X_i; \gamma)\right]^{R_i} \left[1 - \pi\left(X_i; \gamma\right)\right]^{1-R_i}.$$

Then we'd plug in $\pi(x;\hat{\gamma})$ for $\pi(x)$. Plugging in parametric estimates for $f(x)$ and $\pi(x)$ and replacing $\mathbb{E}f(X)$ by its plug-in estimate $\frac{1}{n}\sum_{i=1}^{n} f(X_i;\hat{\theta})$, we get

$$\begin{aligned}\hat{\mu}_{\text{ipw-cv}} &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{R_iY_i}{\pi(X_i;\hat{\gamma})} - \frac{R_if(X_i;\hat{\theta})}{\pi(X_i;\hat{\gamma})} + f(X_i;\hat{\theta})\right)\\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{R_iY_i}{\pi(X_i;\hat{\gamma})} - \left[\frac{R_i-\pi(X_i;\hat{\gamma})}{\pi(X_i;\hat{\gamma})}\right] f(X_i;\hat{\theta})\right)\end{aligned}$$

The approach we presented above to estimate $\hat{\theta}$ and $\hat{\gamma}$ is the one that seems to be used most commonly in practice. The disadvantage of this approach is that it's difficult to derive expressions for the bias and the variance of the resulting estimator because we are reusing the same dataset multiple times. I'm only aware of asymptotic results. For example, if as $n \to \infty$ we have $\hat{\theta} \xrightarrow{P} \theta^*$ and $\hat{\gamma} \xrightarrow{P} \gamma^*$, then one can show[1] that

$$\begin{aligned}\hat{\mu}_{\text{ipw-cv}} &\xrightarrow{P} \mathbb{E}\left[\frac{RY}{\pi(X;\gamma^*)} - \left[\frac{R-\pi(X;\gamma^*)}{\pi(X;\gamma^*)}\right] f(X;\theta^*)\right]\\ &= \mathbb{E}Y + \mathbb{E}\left[\left(\frac{R-\pi(X;\gamma^*)}{\pi(X;\gamma^*)}\right)[Y - f(X;\theta^*)]\right]\\ &= \mathbb{E}Y + \mathbb{E}\left[[Y-f(X;\theta^*)]\underbrace{\left(\frac{1}{\pi(X;\gamma^*)}\left(\pi(X)-\pi(X;\gamma^*)\right)\right)}_{h(X)}\right] \quad \text{(see below)}\end{aligned}$$

If the regression model $f(x;\theta)$ is "correct", or "well specified", in the sense that $f(x;\theta^*) = \mathbb{E}[Y \mid X = x]$, then

$$\begin{aligned}\mathbb{E}\left[[Y-f(X;\theta^*)]\,h(X)\right] &= \mathbb{E}\left[(Y-\mathbb{E}[Y \mid X])\,h(X)\right]\\ &= 0 \quad \text{By projection interpretation}\end{aligned}$$

and so we have $\hat{\mu}_{\text{ipw-cv}} \xrightarrow{P} \mathbb{E}Y$ regardless of how good or bad $\pi(x;\gamma^*)$ is as an estimate for $\pi(x)$. On the other hand, if the propensity model is "correct" (i.e. well specified), in the sense that $\pi(x,\gamma^*) = \pi(x)$, then $h(X) = 0$ and we again get $\hat{\mu}_{\text{ipw-cv}} \xrightarrow{P} \mathbb{E}Y$, regardless of how good or bad is $f(x;\theta^*)$ as an estimate for $f(x)$. Because we only need one of $\pi(x,\gamma^*)$ and $f(x,\theta^*)$ to be correct for the estimator to be consistent, we call this estimator **doubly robust**.

[1] Proving this is well beyond the scope of this course, but see [SV18].

Exercise 5. Complete the derivation given above by showing that

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{R-\pi(X;\gamma^*)}{\pi(X;\gamma^*)}\right)[Y-f(X;\theta^*)]\right] \\
&= \mathbb{E}\left[[Y-f(X;\theta^*)]\underbrace{\left(\frac{1}{\pi(X;\gamma^*)}\left(\pi(X)-\pi(X;\gamma^*)\right)\right)}_{h(X)}\right]
\end{aligned}$$

Solution 4. These calculations should be getting pretty familiar by now. (If not, you probably haven't read the note on Conditional Expectations.)

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{R-\pi(X;\gamma^*)}{\pi(X;\gamma^*)}\right)[Y-f(X;\theta^*)]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(\frac{R-\pi(X;\gamma^*)}{\pi(X;\gamma^*)}\right)[Y-f(X;\theta^*)] \mid X,Y\right]\right] \\
&= \mathbb{E}\left[[Y-f(X;\theta^*)]\left(\frac{1}{\pi(X;\gamma^*)}\left(\mathbb{E}\left[R \mid X,Y\right]-\pi(X;\gamma^*)\right)\right)\right] \\
&= \mathbb{E}\left[[Y-f(X;\theta^*)]\underbrace{\left(\frac{1}{\pi(X;\gamma^*)}\left(\pi(X)-\pi(X;\gamma^*)\right)\right)}_{h(X)}\right].
\end{aligned}$$

3.2 Another approach to fitting $\hat{\theta}$ and $\hat{\gamma}$

Rather than reusing the same dataset to fit $\hat{\theta}$ and $\hat{\gamma}$, another approach is to split the dataset and use one part to fit $\hat{\theta}, \hat{\gamma}$ and then use the other part to evaluate the estimator. For this case, we can derive[2] closed form expressions for the bias and variance of $\hat{\mu}_{\text{ipw-cv}}$ in terms of how good the estimates are for $f(x)$ and $\pi(x)$.

References

[DLL11] Miroslav Dudík, John Langford, and Lihong Li, *Doubly robust policy evaluation and learning*, Proceedings of the 28th International Confer-

[2] See [DLL11, Thms 1 and 2] for the slightly more general scenario of offline policy evaluation, which we'll discuss later in the course. These results are a bit of work, but not beyond the scope of the class.

ence on International Conference on Machine Learning (Madison, WI, USA), ICML'11, Omnipress, 2011, pp. 1097–1104.

[Owe13] Art B. Owen, *Monte carlo theory, methods and examples*, 2013.

[SV18] Shaun R. Seaman and Stijn Vansteelandt, *Introduction to double robust methods for incomplete data*, Statistical Science **33** (2018), no. 2, 184–197.