Reinforcement Learning and Bandits **Spring 2019**

Off-policy Evaluation and Learning

Instructors: Alekh Agarwal, Sham Kakade *Lecture 7*

In the last lecture, we saw the basic setup of the contextual bandit problem, as well as the really simple greedy strategy which can be used in an i.i.d. setting, but only if we observe the reward of every action at each round. In this lecture, we will introduce the relevant ideas to extend the computation of such greedy strategies to the partial feedback setting.

7.1 Formal problem setting

For readers familiar with supervised machine learning, off-policy evaluation and learning questions are probably the most natural ones in the contextual bandits. For the off-policy setting, the most natural description consists of a fixed and unchaging distribution $D_t \equiv D$ over contexts and rewards. The learner has access to a dataset $(x_i, a_i, r_i)_{i=1}^n$, where $x_i \sim D$, $r_i \sim D(\cdot|x_i, a_i)$ for all i and actions a_i are chosen according to some fixed *logging policy* μ. In practice, μ might be the behavior resulting from following some existing system, or from an experiment designed in order to collect data for subsequent learning.

In *off-policy evaluation*, the learner has a *target policy* π of interest, and seeks to estimate the value of π, which we recap from the previous lecture as:

$$V(\pi) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{x\sim D_t}\left[\mathbb{E}[r_t|x, \pi(x)]\right]. \tag{7.1}$$

Note that this is a *counterfactual question*. That is, we seek to ask *what would have happened* if we had use π at the time of data collection. This might not be a causal statement about the future if the data distribution is non-stationary. In stationary settings, we can however interpret the estimates we obtain as accurate predictions about the future.

It is clear that such an estimation is not possible in general. If μ never chooses actions which π would pick, then we cannot plausibly estimate $V(\pi)$. More generally, if π and μ systematically disagree on some fraction ϵ of the contexts, then a bias of $\mathcal{O}(\epsilon)$ in any estimate of $V(\pi)$ is unavoidable in general. If π is known at data collection time, then the simplest alternative is to choose the actions according to π with some positive probability (say $1/2$) on each context. However, π is often not known at data collection time, or we wish to evaluate not just one but a number of different policies. In such cases, the general solution relies on collecting data using some randomized policy μ. Estimates of $V(\pi)$ can then be constructed, whose quality scales with the amount of similarity between π and μ.

In *off-policy learning*, the learner has access to a policy class π, and wishes to find a policy $\widehat{\pi}_n$ from the dataset collected with μ such that

$$V(\widehat{\pi}_n) \geq \max_{\pi\in\Pi} V(\pi) - \epsilon_n,$$

for some suitable slack ϵ_n. In principle, there is an elementary way of doing learning given the evaluation ability. We could simply evaluate all the different policies in Π on our data, and pick the best. This is quite inefficient computationally, of course. The focus of this part will be computational, and we will see how the off-policy learning questions can be turned into specific types of classification problems, such as mutliclass classification.

We conclude the setting by laying down a couple of formal assumptions which will be necessary throughout this lecture. In order to be concise, it would be helpful to extend the distributions μ and π to be viewed as distributions over the (x, a, r) triples, where we have contexts x drawn according to D, actions a drawing according to μ or π given

x and rewards according to $D(\cdot|a,x)$. Throughout this lecture, we assume that the number of actions is finite, although the techniques largely apply to continuous action spaces as well with slight generalizations.

Assumption 1 (Bounded rewards)**.** For any x, a, the rewards are bounded in $[0,1]$. For a fixed (x,a), we denote $\mathrm{Var}[r|x,a] = \sigma^2(x,a)$.

Assumption 2 (Coverage in logging)**.** For any x, a, if $\pi(a|x) > 0$, then we also have $\mu(a|x) > 0$.

7.2 Off-policy evaluation

For evaluating a policy π using data collected according to μ, it suffices to ask a simpler question. For a fixed context x and action a, can we estimate $v(x,a) := \mathbb{E}[r|x,a]$ using data collected according to μ? It turns out that we have already seen an instance of such an estimator in defining the reduction from adversarial bandits to full information in Lecture 5. We begin with this estimator before presenting improvements.

7.2.1 Inverse Propensity Scoring

Recalling that we have n i.i.d. samples (x_i, a_i, r_i) according to μ, we translate the notation to we present the basic estimator again as

$$\widehat{v}_{\texttt{IPS}}(x_i, a) = r_i \frac{\mathbf{1}(a_i = a)}{\mu(a_i|x_i)}. \tag{7.2}$$

Here the name `IPS` stands for *Inverse Propensity Scoring*, where the probability $\mu(a_i|x_i)$ is referred to as the propensity score. This estimator goes back at least to the work of Horvitz and Thompson [1952]. The estimator is unbiased under Assumption 2.

Theorem 3 (Unbiasedness of `IPS`)**.** *For any x and a, if $\mu(a|x) > 0$, then $\mathbb{E}[\widehat{v}_{\texttt{IPS}}(x,a) \mid x,a] = v(x,a)$.*

Proof: The proof essentially follows from definitions. We have

$$\begin{aligned}
\mathbb{E}[\widehat{v}_{\texttt{IPS}}(x_i,a) \mid x_i, a] &= \sum_{a' \in \mathcal{A}} \mu(a'|x_i) \mathbb{E}[\widehat{v}_{\texttt{IPS}}(x_i,a) \mid x_i, a, a_i = a'] \\
&\overset{(a)}{=} \sum_{a' \in \mathcal{A}} \mu(a'|x_i) v(x_i, a') \frac{\mathbf{1}(a' = a)}{\mu(a'|x_i)} \\
&\overset{(b)}{=} v(x_i, a).
\end{aligned}$$

Here the equality (a) uses the fact that the only randomness present after fixing a, x and $a_i = a'$ is that in the rewards being drawn according to D, which has the expected value of $v(x_i, a')$ as per our earlier definition. Equality (b) uses the fact that the indicator only preserves the term corresponding to $a' = a$ in the sum, as long as $\mu(a|x_i) \neq 0$, as per our assumption. ■

The result tells us that we can estimate the reward of any action in an unbiased manner, given logged exploration data. A natural unbiased estimator for any policy π is then immediately obtained as well.

Corollary 4. *Given logged exploration data with a policy μ, and an evaluation policy π such that Assumption 2 holds, the following `IPS` estimator for value of π is unbiased:*

$$\widehat{V}_{\texttt{IPS}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \sum_{a \in \mathcal{A}} \pi(a|x_i) \widehat{v}_{\texttt{IPS}}(x_i, a) = \frac{1}{n} \sum_{i=1}^{n} \frac{\pi(a_i|x_i)}{\mu(a_i|x_i)} r_i. \tag{7.3}$$

This basic result tells us that given sufficiently large log of exploration data, we can aim to estimate any policies whose actions are chosen with non-zero probabilities in the logs. It is natural to ask how large is sufficiently large? A first step towards answering this is by understanding what controls the variance of the IPS estimator. In order to state the subsequent results concisely, we introduce the notation $w(x,a) = \pi(x,a)/\mu(x,a)$ to denote the propensity scores. With this notation, we have

$$\widehat{V}_{\texttt{IPS}}(\pi) = \frac{1}{n} w(x_i, a_i) r_i.$$

We have the following result.

Theorem 5 (Variance of IPS)**.** *Under Assumption 2, we have*

$$\mathrm{Var}[\widehat{V}_{\texttt{IPS}}(\pi)] = \frac{1}{n}\left[\mathbb{E}_{(x,a,r)\sim\mu}[\sigma^2(x,a)w^2(x,a)] + \mathrm{Var}_{(x,a,r)\sim\mu}[v(x,a)w(x,a)]\right].$$

Proof: Since the (x_i, a_i, r_i) are jointly i.i.d., we know that

$$\mathrm{Var}[\widehat{V}_{\texttt{IPS}}(\pi)] = \frac{1}{n}\mathrm{Var}_{(x,a,r)\sim\mu}[w(x,a)r(x,a)],$$

where we recall our notational convention of extending μ to a distribution over triples.

By the law of total variance (i.e. $\mathrm{Var}[Y] = \mathbb{E}[\mathrm{Var}[Y|X]] + \mathrm{Var}[E[Y|X]]$), we have

$$\begin{aligned}
\mathrm{Var}_{(x,a,r)\sim\mu}[w(x,a)r(x,a)] &= \mathbb{E}_\mu\left[\mathrm{Var}[r(x,a)w(x,a)|x,a]\right] + \mathrm{Var}_\mu\left[\mathbb{E}[w(x,a)r(x,a)|x,a]\right] \\
&= \mathbb{E}_\mu\left[w^2(x,a)\,\mathrm{Var}[r(x,a)|x,a]\right] + \mathrm{Var}_\mu\left[w(x,a)v(x,a)\right] \\
&= \mathbb{E}_\mu\left[w^2(x,a)\sigma^2(x,a)\right] + \mathrm{Var}_\mu\left[w(x,a)v(x,a)\right].
\end{aligned}$$

Dividing by n completes the proof. ■

The two terms in the variance upper bound are quite intuitive. The first term captures the joint variance due to the distributions of $(x,a,r) \sim \mu$. This is unavoidable in off-policy evaluation, even if the context distribution is degenerate and consists of just one context. It scales quadratically with both the variance in rewards, as well as the propensity score w. In off-policy evaluation, we are generally trying to keep the dependence on w as mild as possible, so as to allow the policy π to be as dissimilar from μ as possible while still having good evaluation. From the first term, we see that the dependence on w is poor if the rewards have a sizeable variance, but becomes much smaller when the rewards are nearly deterministic.

The second term has already taken variance over the randomness in rewards, and is capturing the joint variance due to the actions a and the contexts x. This term generally scales quadratically in v and w, and can be large even if the rewards are completely deterministic. The dependence stays this way, even if the context distribution is degenerate and consists of just a single x.

Remark: Owing to the second term, we will generally think of the variance of IPS as scaling quadratically in the propensity score w. Thus, we have smaller errors in estimating the value of π when the propensity scores are small, meaning that μ and π are very similar in the induced distributions over actions.

Remark: Since IPS is unbiased, its variance is also equal to its Mean Squared Error (MSE) in estimating $V(\pi)$. That is, Theorem 5 also provides an upper bound $\mathbb{E}\left[\left(\widehat{V}_{\texttt{IPS}}(\pi) - V(\pi)\right)^2\right]$.

As the remark indicates, we can further interpret Theorem 5 as giving a bound on how far the IPS estimate might be from the true value of the policy π, as a function of the context and reward distributions, as well as the mismatch

between μ and π. However, the MSE is an error bound only in expectation. A more refined question is to further examine how fast the errors in our estimation start behaving like this expectation. Alternatively, given a finite number n of samples in our log, can we obtain an upper bound on $\widehat{V}_{\texttt{IPS}}(\pi) - V(\pi)$ which holds with overwhelming probability? This is the issue we tackle with our next result.

Theorem 6 (Deviation bound for `IPS`)**.** *Under conditions of Theorem 5, let us further assume that the propensity scores satisfy $w(x,a) \leq \hat{w}_{\max}$ almost surely. Then with probability at least $1-\delta$, we have*

$$\left|\widehat{V}_{\texttt{IPS}}(\pi) - V(\pi)\right| \leq \hat{w}_{\max}\sqrt{\frac{1}{2n}\log\frac{2}{\delta}}.$$

Proof: The proof follows from an application of Hoeffding's inequality (Lemma 6.1 in the previous lecture). By Theorem 3, we know that $\widehat{V}_{\texttt{IPS}}(\pi)$ is a sum of n i.i.d. random variables, each with mean $V(\pi)$. Since the rewards are in $[0,1]$, and the propensity scores are non-negative and at most $\hat{w}_{\max}$ almost surely, each random variable takes values in $[0, \hat{w}_{\max}]$. The result now directly follows from Lemma 6.1. ■

Remark: This theorem seems remarkably crude in comparison with our previous result in Theorem 5. While Theorem 5 carefully accounts for the mismatch between π and μ as well as the context and reward distributions, Theorem 6 only looks at the largest propensity score. Even if π and μ mostly agree on their actions, but choose very different things on occassion, largest score can be large meaning the theorem will allow the error in `IPS` to be rather large.

This discrepancy between the results of Theorem 5 and 6 is because Hoeffding's inequality does not account for any properties of a random variable apart from the range in which they take values. But a random variable might have a small variance despite the range being large. Should this lead to better concetration around the mean? We will formalize this notation by using Bernstein's inequality, which is an improvement upon Hoeffding's in taking the variance of the random variables into account.

Lemma 7 (Bernstein's inequality)**.** *Suppose $X_1, \ldots, X_n$ are i.i.d. with 0 mean, variance σ^2 and $|X_i| \leq M$ almost surely. Then with probability at least $1-\delta$, we have*

$$\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \leq \sqrt{\frac{2\sigma^2}{n}\log\frac{2}{\delta}} + \frac{2M}{3n}\log\frac{2}{\delta}.$$

Noting that we have bounds on both the range and variance of the `IPS` estimator from Theorems 5 and 6, we immediately obtain a different upper bound by the application of Bernstein's inequality.

Theorem 8 (Sharper deviation bound for `IPS`)**.** *Under conditions of Theorem 6, with probability at least $1-\delta$, we have*

$$\left|\widehat{V}_{\texttt{IPS}}(\pi) - V(\pi)\right| \leq \sqrt{2\log\frac{2}{\delta}\,\frac{\mathrm{Var}_{(x,a,r)\sim\mu}[\widehat{V}_{\texttt{IPS}}(\pi)]}{n}} + \frac{2\hat{w}_{\max}}{3n}\log\frac{2}{\delta}.$$

Remark: We notice that the first term in the upper bound scales roughly as square root of the variance bound in Theorem 5, which is to be expected since we are bounding the error in absolute value rather than the MSE. Here, we see that the additional $\log 2/\delta$ term is present in order to account for the failure probability, while the MSE bound only considers expectation.

Remark: We also observe that the second term still depends on the largest propensity score $\hat{w}_{\max}$. However, this term now goes to zero as $1/n$ rather than $1/\sqrt{n}$ in Theorem 6. This is a major improvement when n is large, since $1/n$ is substantially smaller than $1/\sqrt{n}$. As a result, we will say that the asymptotically dominant term when n gets large is governed by the variance of the `IPS` estimator, like in the MSE bound. At the same time, given a desired error threshold, Theorem 8 allows us to obtain an estimate of the number of samples before $\widehat{V}_{\texttt{IPS}}(\pi)$ can approximate $V(\pi)$ to that error.

Computation of `IPS` estimates: So far we have only discussed the statistical aspects of $\widehat{V}_{\texttt{IPS}}$. However, its relatively simple structure implies that it is computationally cheap to obtain. Given our log, it suffices to just stream over the data in it. For each sample, we check whether the recorded action matches that of π, and if so, accumulate the propensity-weighted reward. Consequently, $\widehat{V}_{\texttt{IPS}}(\pi)$ for a fixed policy π is computed in time linear in the size of the logs, and requires constant storage due to the streaming nature of the computation.

7.2.2 Improved Estimators beyond `IPS`

As we have seen so far, `IPS` fulfils the simplest goal of providing unbiased off-policy evaluation. However, it suffers from a large variance when the propensity scores are large. There is a whole body of literature in machine learning, statistics and econometrics that has subsequently worked on developing better estimators to address this shortcoming of `IPS`. We will present one such estimator in some detail and provide pointers to other improvements.

7.2.2.1 Doubly Robust Estimators

A common approach to off-policy evaluation comes from viewing it as a supervised learning problem. Given data (x_i, a_i, r_i), we can treat this as a supervised dataset to learn a mapping $\widehat{v} \ : \ \mathcal{X} \times \mathcal{A} \mapsto \mathbb{R}$. Given such a mapping, a straightforward estimator for $V(\pi)$ is the use $\widehat{v}$ as the true reward function for evaluating π, that is

$$\widehat{V}_{\texttt{DM}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \widehat{v}_{\texttt{DM}}(\pi(x_i)). \tag{7.4}$$

Here `DM` stands for *Model-Based*, to reflect that this reward estimator is based on first fitting a model $\widehat{v}_{\texttt{DM}}$ of the reward function. Note that there is no importance weighting here, since given $\widehat{v}$, we can evaluate the reward of π's action on each round, whether we chose that action in our dataset or not.

The quality of `DM` estimators heavily relies upon the quality of $\widehat{v}$ in estimating v. If we happen to find an estimator $\widehat{v}$ which has a small MSE in estimating v, then it is easy to show that $\widehat{V}_{\texttt{DM}}(\pi)$ is also a good estimator for $V(\pi)$. However, in general $\widehat{v}$ might be a biased estimator for v, and the bias might not even go down with the number of samples n in our dataset. For instance, if we estimate $\widehat{v}$ as the solution to a linear regression for predicting the rewards, but the true reward function is highly non-linear, then any estimator $\widehat{v}$ we come up with is going to be extremely inaccurate in many cases. This is in contrast with `IPS`, where the error goes down provably as the size of our exploration data increases.

At the same time, when $\widehat{v}$ is a good estimator for v, then the resulting estimator $\widehat{V}_{\texttt{DM}}(\pi)$ can have significantly lower variance than `IPS`. This arises since there is no division by potentially small probabilites in $\widehat{V}_{\texttt{DM}}(\pi)$. For instance, if $\widehat{v}$ was estimated by a linear regression, and if v is indeed a linear function of the contexts for each action, then we might hope to quickly obtain a small MSE in predicting the rewards and obtain a high quality `DM` estimator.

This motivates a natural quest for estimators which retain the worst-case robustness of `IPS`-style unbiased approaches, while benefitting from good reward models when possible. One such line of inquiry leads to the so-called doubly robust estimators [Bang and Robins, 2005, Rotnitzky et al., 2012, Dudík et al., 2011]. To formally define the doubly robust (DR) estimator, we first create a different reward estimator given a context and action, and assuming access to a model-based reward estimator $\widehat{v}_{\texttt{DM}}(x, a)$:

$$\widehat{v}_{\texttt{DR}}(x_i, a) = \widehat{v}_{\texttt{DM}}(x, a) + (r_i - \widehat{v}_{\texttt{DM}}(x, a)) \frac{\mathbf{1}(a_i = a)}{\mu(a_i|x_i)}. \tag{7.5}$$

Remark: In order to avoid carrying subscripts througout, we will now use $\widetilde{v}$ to denote the arbitrary reward function which goes into the doubly robust estimator, rather than the more cumbersome $\widehat{v}_{\texttt{DM}}$ notation.

That is, we follow the prediction of the reward estimator $\widetilde{v}(x,a)$ to evaluate any (x,a) pair's reward. But this estimator might be arbitrarily biased. So on the actions a_i in our dataset, we add an importance weighted reward taking the $r_i - \widetilde{v}(x,a)$. This additional term can be seen as a correction for the bias in $\widetilde{v}(x,a)$. When $\widetilde{v}(x,a)$ is close to r_i, the residual is small meaning that we have a small penalty in variance from the importance weights. But if the estimator is completely off, we still correct the bias as opposed to blindly following $\widetilde{v}(x,a)$. We now define the doubly robust estimator for policies in the natural manner, and study its bias and variance properties formally.

$$\widehat{V}_{\texttt{DR}}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \widehat{v}_{\texttt{DR}}(x_i, \pi(x_i)) = \frac{1}{n}\sum_{i=1}^{n}\left[\widetilde{v}(x,a) + (r_i - \widetilde{v}(x,a))\,\frac{\mathbf{1}(a_i = a)}{\mu(a_i|x_i)}\right]. \tag{7.6}$$

Theorem 9 (Unbiasedness of DR)**.** *Given logged exploration data with a policy μ, an evaluation policy π such that Assumption 2 holds and an arbitrary reward estimator $\widetilde{v}$, $\mathbb{E}[\widehat{V}_{\texttt{DR}}(\pi)] = V(\pi)$.*[1]

Proof: The proof follows by showing that for every x_i, a such that $\mu(a|x_i) > 0$, we have $\mathbb{E}[\widehat{v}_{\texttt{DR}}(x_i,a)|x_i,a] = v(x_i,a)$. This follows essentially from the unbiasedness of IPS, since

$$\begin{aligned}\mathbb{E}[\widehat{v}_{\texttt{DR}}(x_i,a)|x_i,a] &= \widetilde{v}(x_i,a) + \mathbb{E}\left[(r_i - \widetilde{v}(x,a))\,\frac{\mathbf{1}(a_i=a)}{\mu(a_i|x_i)}\right] \\ &\overset{(a)}{=} \widetilde{v}(x_i,a) + v(x,a) - \widetilde{v}(x,a) = \widehat{v}(x,a).\end{aligned}$$

Here (a) follows since the second term is effectively IPS, but using a reward of $r_i - \widetilde{v}(x,a)$ instead of just r_i. Consequently, the unbiasedness of IPS tells that the conditional expectation is equal to $v(x,a) - \widetilde{v}(x,a)$. Noting that under the conditions of the theorem, π only takes actions with non-zero probabilities under μ completes the proof. ■

The unbiaedness result tells us that DR is indeed robust to the potential poor quality of $\widetilde{v}$, resulting in the first type of robustness in this estimator. We will now see the second type of robustness, which is to large importance weights when the reward predictor $\widetilde{v}$ is good.

Theorem 10 (Variance of DR)**.** *Under Assumption 2, we have*

$$\begin{aligned}\mathrm{Var}[\widehat{V}_{\texttt{DR}}(\pi)] = \frac{1}{n}\Big[&\mathbb{E}_{(x,a,r)\sim\mu}[\sigma^2(x,a)w^2(x,a)] + \mathbb{E}_{x\sim D}\left[\mathrm{Var}_{a\sim\mu(\cdot|x)}\, w(x,a)(v(x,a) - \widetilde{v}(x,a))\right] \\ &+ \mathrm{Var}_{x\sim D}\left[\mathbb{E}_{a\sim\mu(\cdot|x)}\left[v(x,a)w(x,a)|x\right]\right]\Big].\end{aligned}$$

Remark: The first term is same as in the variance of IPS, and is still quadratic in w. The main saving comes in the second variance term of IPS replaced by the second and third terms in Theorem 10. First note that the law of total variance

$$\mathrm{Var}_{(x,a,r)\sim\mu}[w(x,a)\widehat{v}(x,a)] = E_{x\sim D}\left[\mathrm{Var}_{a\sim\mu(\cdot|x)}\, w(x,a)v(x,a)\right] + \mathrm{Var}_{x\sim D}\left[\mathbb{E}_{a\sim\mu(\cdot|x)}\left[v(x,a)w(x,a)|x\right]\right].$$

So the only change is the residual relative to $\widetilde{v}$ in the second term. When $\widetilde{v}$ is a good predictor for $\widehat{v}$, this term can be quite small. Intuitively, this term is capturing the expected variance due to the action selection based on μ, and observed actions and rewards in our dataset become largely irrelevant to the quality of estimation, if someone gives us a great reward predictor for all contexts and actions.

At the same time, access to $\widetilde{v}$ does not completely avoid the quadratic dependence on w if one of the reward or context distributions is non-degenerate, as shown by the first and third terms. The reward variance arises as the

[1] The expectation does not need to include any randomness in $\widetilde{v}$.

importance weighted term in the DR estimator takes the difference between an observed reward and $\widetilde{v}$. Even if $\widetilde{v}$ was deterministically equal to $\widehat{v}$, this term will be on the order of reward variance on average. The last term in variance bound of DR comes from the richness of the context distribution. If we see only n samples, the sample average reward over these contexts is only a proxy for the actual expectation over contexts, with the gap governed by the variance in the context distribution.

Proof: We will be terse since the proof large mirrors that of IPS. We note that

$$\begin{aligned}\mathrm{Var}[\widehat{V}_{\mathtt{DR}}(\pi)] &= \mathbb{E}[\mathrm{Var}[\widehat{V}_{\mathtt{DR}}(\pi)|x,a]] + \mathrm{Var}[\mathbb{E}[\widehat{V}_{\mathtt{DR}}(\pi)|x,a]] \\ &= \mathbb{E}[w^2(x,a)\sigma^2(x,a)] + \underbrace{\mathbb{E}_x[\mathrm{Var}_{a\sim\mu(\cdot|x)}\,\mathbb{E}_r[\widehat{V}_{\mathtt{DR}}(\pi)|x,a]]}_{\mathcal{T}} \\ &\quad + \mathrm{Var}_x[\mathbb{E}_{a\sim\mu(\cdot|x)}\mathbb{E}_r[\widehat{V}_{\mathtt{DR}}(\pi)|x,a]].\end{aligned}$$

Now the third term directly follows due to the unbiasedness of the doubly robust estimator. For the middle term, note that each individual summand in $\widehat{V}_{\mathtt{DR}}(\pi)$ can be written as

$$\widehat{v}(x_i,\pi(x_i)) = \widetilde{v}(x_i,\pi(x_i)) + (r_i - \widetilde{v}(x_i,\pi(x_i))\frac{\mathbf{1}(a_i=\pi(x_i))}{\mu(a_i|x_i)}.$$

Since $\widetilde{v}(x_i,\pi(x_i))$ does not depend on the action a_i chosen in our dataset, this is a constant offset which does not count towards the variance in the term $\mathcal{T}$, which is only with respect to the conditional distribution over actions. Taking variance just of the remaining estimator gives the result. ■

By taking the variance upper bound, and using it in Bernstein's inequality, we also obtain a finite sample error bound on the quality of $\widehat{V}_{\mathtt{DR}}(\pi)$. This bound also improves upon that for IPS in the same ways as the variance bound.

Where does $\widetilde{v}$ come from? A natural next question to consider is how to get reward estimators $\widetilde{v}$ in practice. A common practice is to in turn take the logged data, and perform regression with it to obtain $\widetilde{v}$. From the variance upper bound in Theorem 9, we see that the second term is effectively the MSE of $\widetilde{v}$ in predicting $\widehat{v}$, although on a sample reweighted according to the propensity scores w. Consequently, a natural estimation procedure for $\widetilde{v}$ is to take some class of regression functions and minimize propensity-weighted squared loss over that class. That is,

$$\widetilde{v} = \arg\min_{f\in\mathcal{F}} \sum_{(x,a,r)\in S} w(x,a)(f(x,a)-r)^2,$$

where f is our class of regression functions and S is the dataset we use for regression. Common choices for $\mathcal{F}$ are linear functions using some joint feature representations of contexts and actions, regression trees or random forests and neural networks. In statistics and econometrics, where the features are often relatively low-dimensional, more non-parametric regression approaches are also often used.

One concern with this approach might be that the regression function $\widetilde{v}$ is itself dependent on the samples now, and this dependence was not accounted for in our bias and variance calculations above. To avoid dependence issues, the common approach is to take the entire logged dataset D and break it up into two equal halves D_1 and D_2. We first train $\widetilde{v}_1$ on D_1 and use it for evaluating π on D_2. We similarly train $\widetilde{v}_2$ on D_2 and use it for D_1. The two estimates for $V(\pi)$ are finally averaged to lower the variance. This is also easily generalized to a k-fold or leave-one-out methodology.

7.2.3 Other improvements and lower bounds

Note that we have only considered unbiased estimators for $V(\pi)$ so far. Noting that the MSE of an estimator is sum of its variance and squared bias, it is natural to consider estimators with a small bias that might have a lower variance.

Biased estimators using propensity score surrogates $\hat{w}$ Since the biggest source of variance in our estimators is large propensity scores, several works have investigated the use of lower variance surrogates. Two common approaches are to pick a threshold τ, and use $\hat{w}(x,a) = \max(w(x,a), \tau)$ (see e.g. [Bembom and van der Laan, 2008]) or $\hat{w}(x,a) = w(x,a)\mathbf{1}(w(x,a) < \tau)$ (e.g. [Bottou et al., 2013]). Alternatively, some works also use the observed propensities $w(x,a)$ to create a smoother surrogate $\hat{w}(x,a)$, which is then used in evaluation.

For the first two proposals, controlling the bias is easy. We are effectively allowing bias whenever $w(x,a) > \tau$, so the probability of this event yields a direct estimate on the bias.

When a reward estimator $\widetilde{v}$ is present, we can take these approaches a bit further. In the work of Bottou et al. [2013], the reward is estimated by 0 whenever $w(x,a)$ is larger than τ. A natural alternative is to instead replace it with $\widetilde{v}(x,a)$ on this event. This is the approach motivated in Wang et al. [2016] under the name of SWITCH estimators, and they also prescribe an empirical procedure for setting a good threshold τ. As the empirical results in these works show, such modifications can result in substantial improvements in our value estimates.

Lower bounds on error: There are a number of estimators for off-policy evaluation, each with its corresponding MSE upper bound. It is natural to ask how far we can lower the error, and what are fundamental lower bounds for the hardness of the problem of off-policy evaluation. This question has been studied in the literature from at least two perspectives.

Wang et al. [2016] take a finite sample minimax approach to the problem. That is, given n samples according to μ, a fixed evaluation policy π and context distribution D, *what is the smallest possible error of any potential estimator of $V(\pi)$, when we take worst case over all possible reward distributions?* Under certain moment conditions, they show that the simplest `IPS` estimator is already optimal up to constant factors according to this criterion. However, they also observe that this is only true in the worst-case over arbitrary reward structures, and we have already seen possible improvements when we have additional knowledge of the function $\widehat{v}(x,a)$, allowing us to create a good predictor $\widetilde{v}(x,a)$ for `DR` or SWITCH estimators.

A second viewpoint comes from the asymptotic theory, where a similar question is asked as above, but in the limit of infinite sample size, and the emphasis is on identifying the precise constants in the error bound. Under certain conditions (see e.g. [Rothe, 2016]), effectively implying that a non-trivial estimator $\widetilde{v}$ can be obtained, `DR` is shown to be better than `IPS` in such settings.

7.3 Off-policy learning

Off-policy evaluation is a useful tool, if we have a collection of data and a policy in mind that we wish to evaluate. But more commonly, we only have our dataset, and would like to find a policy attains a high value. Often, we may further have a parametrized policy class in mind, such as by the weights in a linear function or neural network. In this case, we would really like to find a set of weights such that the policy corresponding to them has a high estimate for $V(\pi)$ based on the data we have.

7.3.1 Statistical Analysis

Given the power to do off-policy evaluation, this question is not too hard statistically. Indeed, let us suppose that we have access to a finite policy class π, with $|\Pi| = N$. Then we can easily prove the following corollary of Theorem 6.

Corollary 11 (Deviation bounds for policy optimization)**.** *Given a finite policy class Π with $|\Pi| = N$, assume that*

the conditions of Theorem 6 are satisfied for each $\pi \in \Pi$. Then with probability at least $1-\delta$, we have

$$\sup_{\pi\in\Pi} |\widehat{V}_{\texttt{IPS}}(\pi) - V(\pi)| \leq \hat{w}_{\max}\sqrt{\frac{1}{2n}\log\frac{2N}{\delta}}.$$

In particular, we can define $\widehat{\pi}_n = \arg\max_{\pi\in\Pi} \widehat{V}_{\texttt{IPS}}(\pi)$. Applying the corollary twice, once with $\widehat{\pi}_n$ and once with $\pi^\star$, we obtain

$$V_{\texttt{IPS}}(\pi) \geq V(\pi^\star) - \hat{w}_{\max}\sqrt{\frac{2}{n}\log\frac{2N}{\delta}}. \tag{7.7}$$

The corollary follow from Theorem 6 using a union bound over policies in Π. Effectively, this is the same calculation we saw for analyzing the GREEDY and τ-GREEDY algorithms in the previous lecture, except we do not need full information or uniform randomization anymore! We can also use the same idea with better deviation bounds, such as in Theorem 8, but we have to instead use a uniform upper bound on the variance for all policies in Π now.

7.3.2 Computational Algorithms

While the statistical analysis of off-policy learning follows relatively straightforwardly given that for off-policy evaluation, the computational story is now significantly more complex. A naïve algorithm for off-policy learning would evaluate our favorite off-policy evaluator for every $\pi \in \Pi$, and then return the policy with the best estimated value. This would require $\mathcal{O}(N)$ computation, prohibitive if Π is continuously parametrized or exponentially large.

However, it is difficult to do much better. Consider a special case of two actions, with r_i taking values in $\{0,1\}$ and $\mu(a|x_i) = 0.5$ for each action a. Assume further exactly one action gets a reward of one on each context. Suppose we can efficiently compute the best policy $\widehat{\pi}_n$ for a sample generated from this distribution, for our policy class Π. It means that we can compute

$$\begin{aligned}
\widehat{\pi}_n &= \arg\max_{\pi\in\Pi} \widehat{V}_{\texttt{IPS}}(\pi) = \arg\max_{\pi\in\Pi} \sum_{i=1}^{n} r_i \frac{\mathbf{1}(a_i = \pi(x_i))}{1/K} \\
&= \arg\max_{\pi\in\Pi} \sum_{i=1}^{n} \mathbf{1}(a_i = \pi(x_i), r_i = 1) \\
&= \arg\max_{\pi\in\Pi} \sum_{i\,:\,r_i=1} \mathbf{1}(a_i = \pi(x_i)) \\
&= \arg\min_{\pi\in\Pi} \sum_{i\,:\,r_i=1} \mathbf{1}(\bar{a_i} \neq \pi(x_i)),
\end{aligned}$$

where $\bar{a_i}$ is the action not in the logs (for this two action case). Note that the last computation is a $0-1$ loss minimization problem for binary classification. For most classes Π of interest, this computation is known to be NP-hard (examples include linear as well as non-linear classes such as decision trees and neural networks). Consequently, it is hopeless to expect that we can have computationally efficient algorithms for off-policy learning under all data distributions, which we have just shown to be a more general case of an NP-hard computation.

However, the important thing to remember in this conclusion is that it is something which holds in worst-case over all distributions. On the other hand, there are a number of distributional assumptions under which $0-1$ loss minimization can be done efficiently in theory, or where we have empirical evidence of the ability to learn good classifiers. In order

to leverage the existence of several efficient classification algorithms, tailored to representations of common interest, we will build our policy learning algorithms which assume access to a certain kind of classification algorithm over the policies $\pi \in \Pi$.

Formally, we first define a cost-sensitive classification problem.

Definition 12 (Cost-sensitive classification)**.** Suppose we are given an input space $\mathcal{X}$ and let D be any distribution over $\mathcal{X} \times [0,1]^K$. Then the cost-sensitive classification error of a function $h : \mathcal{X} \mapsto \{1, 2, \ldots, K\}$ is given by $\mathbb{E}_{(x,c)\sim D}[c(h(x))]$.

An algorithm for cost-sensitive classification then takes a set $\mathcal{H}$ of cost-sensitive classifiers, and promises to return

$$\arg\min_{h\in\mathcal{H}} \mathbb{E}_{(x,c)\sim D}[c(h(x))].$$

For instance, a common structure for designing cost-sensitive classifiers is to start with a base class of regressors f, where $f(x,a) \in [0,1]$ for any $a \in \{1,2,\ldots,K\}$. Then we can associate a classifier with f that predicts $h_f(x) = \arg\min_{a\in\{1,2,\ldots,K\}} f(x,a)$. That is, the classifier h_f treats f as a proxy for the expected cost of a given x and predicts accordingly. For such classifiers, it is know that if there exists $f^\star \in \mathcal{F}$ such that $f^\star(x,a) = \mathbb{E}_{(x,c)\sim D}[c(a)|x]$, then a good classifier can be found by first regressing on the observed costs and then taking the classifier induced by the resulting function. This strategy works whenever we have efficient regression algorithms such as with linear representations, regression tree ensembles etc. Alternatively, reductions from cost-sensitive classification to multiclass classification also exist, so that we can use existing classification algorithms to find a good cost-sensitive classifier.

In summary, cost-sensitive classification problems admit practical algorithms for most representations of interest. This observation matters as we will now show that computations required for off-policy learning can be efficiently performed assuming access to a cost-sensitive classification algorithm. We begin by noting that our off-policy evaluation estimates so far take the form

$$\widehat{V}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \widehat{v}(x_i, \pi(x_i)).$$

Defining $c(x,a) = -\widehat{v}(x,a)$, we further observe that

$$\arg\max_{\pi\in\Pi} \widehat{V}(\pi) = \arg\max_{\pi\in\Pi} \frac{1}{n}\sum_{i=1}^{n} \widehat{v}(x_i, \pi(x_i)) = \arg\min_{\pi\in\Pi} \frac{1}{n}\sum_{i=1}^{n} c(x_i, \pi(x_i)),$$

where we have used the definition of costs as negative rewards to switch from minimization to maximization in the last equality. That is, *off-policy learning is equivalent to cost-sensitive classification for appropriate cost definition.*

As a final remark, we will also show that the problem simplifies even further in the case of the `IPS` estimator. Note that

$$\begin{aligned}
\arg\max_{\pi\in\Pi} \sum_{i=1}^{n} \widehat{v}(x_i, \pi(x_i)) &= \arg\max_{\pi\in\Pi} \sum_{i=1}^{n} r_i \frac{\mathbf{1}(a_i = \pi(x_i))}{\mu(a_i|x_i)} \\
&= \arg\max_{\pi\in\Pi} \sum_{i=1}^{n} \frac{r_i}{\mu(a_i|x_i)} \mathbf{1}(a_i = \pi(x_i)) \\
&= \arg\min_{\pi\in\Pi} \sum_{i=1}^{n} \frac{r_i}{\mu(a_i|x_i)} \mathbf{1}(a_i \neq \pi(x_i)).
\end{aligned}$$

Now the last problem is simply a multiclass classification problem, where we treat each example with a different importance weight, rather than the standard formulation of equal importance weight 1 on every example. Since most existing multiclass classification techniques allow this additional importance weights, there is indeed a large number of algorithms at our disposal to find the best policy according to the `IPS` estimator's value.

Given all this discussion, we finally also observe that the computation of the GREEDY estimator in the previous lecture is precisely a cost-sensitive classification problem, and is efficiently implementable assuming access to such an algorithm.

7.4 Evaluating contextual bandit algorithms

So far, we have seen how to estimate the value of a given policy, and how to search for the best policy from a class. This might be a good idea if we have done a one-off experiment to do some randomized data collection, and want to use the best policy we find from the data to drive our application for the future. In many real world applications though, the data distribution as well as actions available change over time, meaning that a policy which was once good might no longer stay good. This also means that the randomized data we collected at some point in the past may no longer point us towards sound conclusions in the future. This is effectively the difference between causal and counterfactually accurate predictions we discussed before. Our techniques give sound counterfactual predictions, but they might not be causally correct if the past and present have diverged.

As a result, a preferred mode in such applications is to not do one-off randomized experiments, but actually use *explore-exploit* algorithms to constantly obtain some amount of randomization over plausibly good actions. In the previous lecture, we saw two examples, EXP4 and τ-GREEDY of such explore-exploit algorithms.[2] These explore-exploit algorithms for contextual bandits have tunable parameters as well, such as the amount of uniform exploration in EXP4 and the parameter τ in τ-GREEDY. *Given some exploration data, how do we evaluate the performance of an explore-exploit algorithm for different settings of its parameters?* If we could answer this, then we can instantiate the algorithm with the best setting we find when running it.

It turns out, however, that this question cannot be answered for arbitrary bandit algorithms. This is shown below through a pathological example which was descibed in Li et al. [2011].

Example 13 (Impossibility of evaluating contextual bandit algorithms from logged data)**.** Consider a CB problem with $K = 2$ actions, $x \in \{0, 1\}$ and $r(1) = 1$, $r(2) = 0$ independent of which context is chosen. Consider an algorithm `A` which chooses $a_t = 1$ if the first context $x_1 = 1$, otherwise always chooses $a_t = 2$. Suppose x_t are chosen uniformly at random each time. Now with probability 0.5, `A` always plays action 1 and with probability 0.5, always plays action 2. Hence, it has an expected reward of 0.5. However, given any logged dataset, we will either have recorded $x_1 = 1$ on the first example, or $x_1 = 0$. In the first case, we will estimate the average reward of `A` to be 1, while it will be 0 in the second case.

The example points out a pathology. Since exploration algorithms are stateful objects, one large logged dataset is not enough to get a good estimate of the algorithms average reward. Of course, one could take many such datasets, and for the example above we will see a reward of 1 on roughly half the datasets and we can conclude that the average reward is 0.5. But this is extremely data inefficient.

A saving grace is that the above example is truly pathological. For most contextual bandit algorithms of interest, we can indeed estimate their expected reward using logged data. While we will not go into the precise conditions here, we intuitively think of good contextual bandits as reasonably stable, so that any indiosyncracies in the logged dataset do not hurt these algorithms significantly.

[2] Of course τ-GREEDY also explores just once and is not a good fit to such changing environments. We will see better algorithms in the next lecture.

One evaluator for exploration algorithms is shown in Algorithm 1. The algorithm goes over the dataset, and for each x queries the action of our exploration algorithm $\texttt{A}$ one x. If the chosen action is the same as the one in our dataset, we supply its reward to $\texttt{A}$, or we do not provide any feedback. Throughout, given a set of examples denoted by h (for history), we will use $\texttt{A}(h, x)$ to denote the action chosen by $\texttt{A}$ on x after observing rewards for its chosen actions on the samples in h.

Algorithm 1 Off-policy evaluation of exploration algorithms

Require: Exploration algorithm $\texttt{A}$ and dataset D
  Set $h = \emptyset$ ▷ An initially empty history
  Set $G_A = 0$ ▷ Initialize net reward of the algorithm
  Set $T = 0$ ▷ Initialize number of updates done
  for $(x, a, p, r) \in D$ **do**
    if $\texttt{A}(h, x) = a$ **then**
      Update $h \leftarrow h \cup \{(x, a, r)\}$
      Update $G_A \leftarrow G_A + r$
      Update $T \leftarrow T + 1$
    end if
  end for
  return G_A/T

To get intuition about the procedure, let us consider the case when the logged data has uniformly chosen actions, that os $p = 1/K$ in all the records. Then on each example, $\texttt{A}$'s action matches the recorded one with probability $1/K$. That is, we would expect to end up with $T \approx |D|/K$, and the estimate we get is for the expected reward which $\texttt{A}$ obtains after running on $|D|/K$ contextual bandit examples.

An alternative evaluator we can use is closer to $\texttt{IPS}$. For each (x, a, p, r) in the logged data, we first query $a' = \texttt{A}(h, x)$. We then update $h = h \cup \{(x, a, p, r')\}$, where $r' = r\mathbf{1}(a = a')/p$. That is, we update the algorithm using an importance weighted reward, which is unbiased, but injects additional variance as usual.

In general, the evaluation of exploration algorithms in this manner is often prone to significant variance though, and is not quite as robust to its counterpart for fixed policies. For this reason, exploration algorithms often end up being eventually evaluated using online testing.

References

Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

Oliver Bembom and Mark J van der Laan. Data-adaptive selection of the truncation level for inverse-probability-of-treatment-weighted estimators. 2008.

Léon Bottou, Jonas Peters, Joaquin Quinonero Candela, Denis Xavier Charles, Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Y Simard, and Ed Snelson. Counterfactual reasoning and learning systems: the example of computational advertising. *Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *ICML*, 2011.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 1952.

Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 297–306. ACM, 2011.

Christoph Rothe. The value of knowing the propensity score for estimating average treatment effects. *IZA Discussion Paper Series*, 2016.

Andrea Rotnitzky, Quanhong Lei, Mariela Sued, and James M Robins. Improved double-robust estimation in missing data and causal inference models. *Biometrika*, 99(2):439–456, 2012.

Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudík. Optimal and adaptive off-policy evaluation in contextual bandits. *arXiv preprint arXiv:1612.01205*, 2016.