# OFFLINE POLICY EVALUATION IN A CONTEXTUAL BANDIT PROBLEM

SANG HOON KIM

Abstract. Many real-world problems can be described as a contextual bandit problem, which is an important branch of machine learning. Since online evaluation in a contextual bandit problem is expensive, effective offline evaluation is necessary. This paper provides the formal definition of a contextual bandit problem and a way to do offline evaluation with an unbiased estimate with low variance.

## Contents

## 1. Introduction

A contextual bandit problem is a specific case of reinforcement learning, which plays an important role in machine learning. In a contextual bandit problem, an agent has a policy to choose an action based on a given context in order to maximize the reward from the action. Many real-world problems can be modeled as a contextual bandit problem. Internet advertising is a good example. A search engine (agent) should choose a proper advertisement (action) for a user (context) to increase the probability that the user clicks the advertisement. In this case, we can think of the reward as 1 if the user clicks and 0 if not. What makes contextual bandit problem distinguished from others is that the reward is partially observed; we only know the reward from the action chosen. The search engine only knows if the user clicked the advertisement it gave, but does not know if any other advertisement would have been clicked if given.

To maximize the reward, it is of utmost importance to have a good policy. However, it is very expensive to test a policy online; if the search engine tests a policy which gives only a terrible advertisement by actually using it online, users would no longer use the search engine. Therefore, we want to do offline policy evaluation: evaluating a new policy from the data logged by a deployed policy. Offline policy evaluation could sound counter-intuitive because if the new policy chooses an action that the deployed policy did not choose for a given context, there is no way to find the reward from the action chosen by the new policy. With several assumptions, this paper introduces a way to get an unbiased estimate with low variance of the expected reward of a policy we want to evaluate.

## 2. Problem definition

Given a space of contexts $\mathcal{X}$ and a finite space of actions $\mathcal{A}$, the contextual bandit setting is as follows: on each round,

1. The agent checks a vector of a context $x \in \mathcal{X}$.
2. An action (or arm) $a$ is chosen from $\mathcal{A}$.
3. A reward $r \in [0, 1]$ for the action $a$ is shown. However, the rewards of the other actions are not.

We assume that contexts are IID from an unknown distribution $D(x)$. Also, the distribution over rewards $D(r|a, x)$ does not change but is unknown. For each round $k$, a triple $(x_k, a_k, r_k)$ is generated.

A stationary policy is a function from $\mathcal{X}$ to the collection of probability distributions over $\mathcal{A}$; it chooses an action depending only on a given context. At the time step $n$, a non-stationary policy is a function from $\mathcal{X}$ and the historical information, which is $\{(x_1, a_1, r_1), \ldots, (x_{n-1}, a_{n-1}, r_{n-1})\}$, to the probability distribution over $\mathcal{A}$; if the reward of $a$ for a given context $x$ was bad, it can choose a different action for the same context later. Every policy we consider here is stationary. We assume that for a given stream of $n$ contexts, triples $(x_k, a_k, r_k)$ for $k = 1, \ldots, n$ are generated by a policy, which we call the *exploration policy* or *old policy.* The exploration history up to round $k$ is denoted by $z_k = (x_1, a_1, r_1, \ldots, x_k, a_k, r_k)$. The value $\mu(a|x)$ is defined to be the conditional probability of the old policy choosing action $a$ for context $x$. The old policy is assumed to be stochastic, rather than deterministic, i.e. $\mu(a|x) < 1$ for every $a$ and $x$.

Given the input data $z_n$, what we want to do is "evaluating" another policy, which we call the *target policy* or *new policy*; we want to estimate the expected reward of the new policy. Similarly, $\nu(a|x)$ is defined to be the conditional probability of the new policy choosing action $a$ for context $x$. The expected reward is as follows:

$$V(\nu) = \mathbf{E}_{x\sim D}\mathbf{E}_{a\sim\nu(\cdot|x)}\mathbf{E}_{r\sim D(\cdot|x,a)}[r].$$

For unbiased estimate of $V$, we assume that if $\nu(a|x) > 0$, then $\mu(a|x) > 0$; this is true if $\mu(a|x) > 0$ for every $a$ and $x$. The reason why this assumption is needed can be easily shown with an example; suppose the old policy chooses only action $A$ for any context whereas the new policy chooses only action $B$. Then, the policy evaluation is hopeless. Because $\nu$ is fixed in this study, we denote $V(\nu)$ by $V$.

## 3. Basic approaches

What makes the policy evaluation of a contextual bandit problem very difficult compared to other problems is that rewards are partially observed; the agent could only know the rewards of the actions chosen by the old policy. The rewards of the other actions are unknown. Thus, for a given context, we cannot know the reward of the new policy if it chooses an action different from the action chosen by the old policy for that context. There are two existing approaches to overcome this limitation.

### 3.1. DM (Direct Method).

The first one is called the *direct method* (DM). It makes a reward estimate $\hat{r}(x,a)$ conditioned on the context and action. Then, the policy value is estimated by

$$\hat{V}_{DM} = \frac{1}{n}\sum_{k=1}^{n}\sum_{a\in\mathcal{A}} \nu(a|x_k)\hat{r}(x_k,a).$$

This estimator obviously works well if the estimate $\hat{r}(x,a)$ is accurate. However, it is likely that $\hat{r}(x,a)$ is formed without the knowledge of $\nu$, so $\hat{r}(x,a)$ can be inaccurate in the fields significant for $\nu$. Also, it is often difficult to accurately estimate expected rewards; even if we know someone very much, we cannot always know if he likes what we would give to him.

### 3.2. IPS (Inverse Propensity Score).

The second one is the *inverse propensity score* (IPS); it forms an estimate $\hat{\mu}(a|x)$ of the conditional probability $\mu(a|x)$ of the old policy and thus accounts for the bias in the input data made by the old policy.

$$\hat{V}_{IPS} = \frac{1}{n}\sum_{k=1}^{n}\frac{\nu(a_k|x_k)}{\hat{\mu}(a_k|x_k)}\cdot r_k$$

If $\hat{\mu}(a|x) \approx \mu(a|x)$, the IPS is an, approximately, unbiased estimate of $V$, which is shown below.

**Lemma 3.1.** *(Unbiasedness of IPS) For any $x$ and $a$, if $\mu(a|x) \in (0,1)$ as $\nu(a|x) > 0$, IPS is an unbiased estimate of $V(\nu)$, which means the expected value of IPS is the same with $V(\nu)$.*

*Proof.* Let $r(x,a) := \mathbf{E}[r|x,a]$ and $\hat{r}_{IPS}(x_i,a) = r_i \cdot \frac{I(a_i=a)}{\mu(a_i|x_i)}$. Then,

$$\begin{aligned}\mathbf{E}[\hat{r}_{IPS}(x_i,a)|x_i,a] &= \sum_{a'\in\mathcal{A}} \mu(a'|x_i)\mathbf{E}[\hat{r}_{IPS}(x_i,a)|x_i,a,a_i=a'] \\ &= \sum_{a'\in\mathcal{A}} \mu(a'|x_i)r(x,a)\frac{I(a'=a)}{\mu(a'|x_i)} = r(x,a).\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbf{E}[\hat{V}_{IPS}] = \mathbf{E}[\frac{1}{n}\sum_{k=1}^{n}\frac{\nu(a_k|x_k)}{\hat{\mu}(a_k|x_k)}\cdot r_k] = \mathbf{E}[\frac{1}{n}\sum_{k=1}^{n}\nu(a_k|x_k)\cdot r(x,a)]& \\ = V(\nu).&\end{aligned}$$

□

As IPS is unbiased, it works better than any other biased estimator in general. However, IPS suffers a high variance if the random variable $\frac{\nu(a|x)}{\hat{\mu}(a|x)}$ in the denominator is very big as $\nu(a|x)$ is big while $\hat{\mu}(a|x)$ is small. Because we can estimate $V$ only once as we have one sample, the high variance makes our estimate far from the true value, which could make a little bit biased estimator with low variance better than IPS.

## 4. Doubly robust estimator

Doubly robust estimator $\hat{V}_{DR}$ (DR) exploits both estimators; as its name shows, if at least one of the estimators $\hat{r}$ and $\hat{\mu}$ is accurate, the doubly robust estimator is unbiased while having less variance than the other estimators.

$$\hat{V}_{DR} = \frac{1}{n}\sum_{k=1}^{n}[\hat{r}(x_k,\nu) + \frac{\nu(a_k|x_k)}{\hat{\mu}(a_k|x_k)} \cdot (r_k - \hat{r}(x_k,a_k))],$$

$$\text{where } \hat{r}(x,\nu) = \sum_{a\in\mathcal{A}} \nu(a|x)\hat{r}(x|a).$$

Both DM and IPS are special cases of DR with $\hat{\mu} = \infty$ and $\hat{r} = 0$, respectively. In reality, it is likely that neither $\hat{r}$ nor $\hat{\mu}$ is accurate (However, it should be noted that $\hat{\mu}$ is usually much easier to estimate than $\hat{r}$ because in many situations, the agent has almost full knowledge of the old policy that he deployed). Therefore, the most important question: when $\hat{r}$ and $\hat{\mu}$ deviate from truth values, how does $\hat{V}_{DR}$ perform?

## 5. Theoretical analysis

### 5.1. **Assumptions and first consequences.**

There are several assumptions for our theoretical analysis:

(1) $r(x,a) \in [0,1]$ and $\mu(a|x) \in (0,\infty]$.
(2) $r$ is independent of $z_n$.
(3) $\mu$ is conditionally independent of $\{(x_l,a_l,r_l)\}_{l\geq k}$ conditioned on $z_{k-1}$.

The second assumptions means that $r$ is fixed and thus any past history does not affect future rewards. The third assumptions means that $\mu$ does not depend on future.

Analogous to $\hat{r}(x,a)$ and $\hat{r}(x,\nu)$, we define

$$r^*(x,a) = \mathbf{E}_D[r|x,a],$$
$$r^*(x,\nu) = \mathbf{E}_\nu[r|x].$$

We can think those values represent the true rewards.

We define the error terms of $\hat{r}$ and $\hat{\mu}$ as follows:

$$\Delta(x,a) = \hat{r}(x,a) - r^*(x,a),$$
$$\varrho(x,a) = \mu(a|x)/\hat{\mu}(a|x).$$

We also assume that for some $M \geq 0$, with probability one under $\mu$:

$$\frac{\nu(a_k|x_k)}{\hat{\mu}(a_k|x_k)} \leq M.$$

This can be always satisfied with $\hat{\mu} \geq 1/M$, which is possible if $\mathcal{A}$ is finite and $|M| \geq |\mathcal{A}|$.

To bound the error of $\hat{V}_{DR}$, we analyze a single term first:

$$\hat{V}_k = \hat{r}(x_k, \nu) + \frac{\nu(a_k|x_k)}{\hat{\mu}(a_k|x_k)} \cdot (r_k - \hat{r}(x_k, a_k)).$$

We bound its range, bias, and conditional variance with the three following lemmas[1].

**Lemma 5.1.** *The range of* $\hat{V}_k$ *is bounded as*

$$|\hat{V}_k| \leq 1 + M.$$

*Proof.*

$$\begin{aligned} |\hat{V}_k| &= |\hat{r}(x_k, \nu) + \frac{\nu(a_k|x_k)}{\hat{\mu}(a_k|x_k)} \cdot (r_k - \hat{r}(x_k, a_k))| \\ &\leq |\hat{r}(x_k, \nu)| + \frac{\nu(a_k|x_k)}{\hat{\mu}(a_k|x_k)} \cdot |r_k - \hat{r}(x_k, a_k)| \\ &\leq 1 + M. \end{aligned}$$

The last inequality follows from the assumption that $\hat{r}, r_k \in [0, 1]$. □

**Lemma 5.2.** *The expectation of the term* $\hat{V}_k$ *is*

$$\mathbf{E}_\mu[\hat{V}_k] = \mathbf{E}_{(x,a)\sim\nu}[r^*(x,a) + (1 - \varrho(x,a))\Delta(x,a)].$$

*Proof.*

$$\begin{aligned} \mathbf{E}_\mu[\hat{V}_k] &= \mathbf{E}_{(x,a,r)\sim\mu}[\hat{r}(x_k, \nu) + \frac{\nu(a_k|x_k)}{\mu(a_k|x_k)} \cdot \varrho \cdot (r - \hat{r})] \\ &= \mathbf{E}_{x\sim D}[\hat{r}(x, \nu)] + \mathbf{E}_{x\sim D}[\sum_{a\in\mathcal{A}} \mu(a|x) \mathbf{E}_{r\sim D(\cdot|x,a)}[\frac{\nu(a|x)}{\mu(a|x)} \cdot \varrho \cdot (r - \hat{r})]] \\ &= \mathbf{E}_{x\sim D}[\hat{r}(x, \nu)] + \mathbf{E}_{x\sim D}[\sum_{a\in\mathcal{A}} \nu(a|x) \mathbf{E}_{r\sim D(\cdot|x,a)}[\varrho \cdot (r - \hat{r})]] \\ &= \mathbf{E}_{(x,a)\sim\nu}[\hat{r}] + \mathbf{E}_{(x,a,r)\sim\nu}[\varrho \cdot (r - \hat{r})] \\ &= \mathbf{E}_{(x,a)\sim\nu}[r^* + (\hat{r} - r^*) + \varrho \cdot (r^* - \hat{r})] \qquad (5.3) \\ &= \mathbf{E}_{(x,a)\sim\nu}[r^* + (1 - \varrho)\Delta] \end{aligned}$$

□

**Lemma 5.4.** *The variance of the term* $\hat{V}_k$ *can be decomposed and bounded as follows:*

*(1)*

$$\begin{aligned} &\mathbf{V}_\mu[\hat{V}_k] \\ &= \mathbf{V}_{x\sim D}[\mathbf{E}_{a\sim\nu(\cdot|x)}[r^*(x,a) + (1 - \varrho(x,a)) \cdot \Delta(x,a)]] \\ &\quad -\mathbf{E}_{x\sim D}[\mathbf{E}_{a\sim\nu(\cdot|x)}[\varrho(x,a)\Delta(x,a)]^2] \\ &\quad +\mathbf{E}_{(x,a)\sim\nu}[\frac{\nu(a|x)}{\hat{\mu}(a|x)} \cdot \varrho(x,a) \cdot \mathbf{V}_{r\sim D(\cdot|x,a)}[r]] \\ &\quad +\mathbf{E}_{(x,a)\sim\nu}[\frac{\nu(a|x)}{\hat{\mu}(a|x)} \cdot \varrho(x,a) \cdot \Delta(x,a)^2] \end{aligned}$$

*(2)*

$$\begin{aligned} &\mathbf{V}_{\mu}[\hat{V}_k] \\ \leq\ &\mathbf{V}_{x\sim D}[r^*(x,\nu)] + 2\mathbf{E}_{(x,a)\sim\nu}[|(1-\varrho(x,a))\Delta(x,a)|] \\ &+M\cdot\mathbf{E}_{(x,a)\sim\nu}[\varrho(x,a)\cdot\mathbf{E}_{r\sim D(\cdot|x,a)}[(r-\hat{r}(x,a))^2]] \end{aligned}$$

*Proof.* (1)

$$\begin{aligned} \mathbf{E}_{\mu}[\hat{V}_k^2] &= \mathbf{E}_{(x,a,r)\sim\mu}[\hat{r}(x,\nu) + \frac{\nu(a|x)}{\hat{\mu}(a|x)}\cdot\varrho\cdot(r-\hat{r})] \\ &= \mathbf{E}_{x\sim D}[\hat{r}(x,\nu)^2] + 2\cdot\mathbf{E}_{(x,a,r)\sim\mu}[\hat{r}(x,\nu)\cdot\frac{\nu(a|x)}{\hat{\mu}(a|x)}\cdot\varrho\cdot(r-\hat{r})] \\ &\quad +\mathbf{E}_{(x,a,r)\sim\mu}[\frac{\nu(a|x)}{\mu(a|x)}\cdot\frac{\nu(a|x)}{\hat{\mu}(a|x)}\cdot\varrho\cdot(r-\hat{r})^2] \end{aligned}$$

$$\begin{aligned} &= \mathbf{E}_{x\sim D}[\hat{r}(x,\nu)^2] + 2\cdot\mathbf{E}_{(x,a,r)\sim\nu}[\hat{r}(x,\nu)\cdot\varrho\cdot(r-\hat{r})] \\ &\quad +\mathbf{E}_{(x,a,r)\sim\nu}[\frac{\nu(a|x)}{\hat{\mu}(a|x)}\cdot\varrho\cdot(r-\hat{r})^2] \end{aligned} \tag{5.5}$$

$$= \mathbf{E}_{(x,a)\sim\nu}[(\hat{r}(x,\nu)-\varrho\Delta)^2] - \mathbf{E}_{(x,a)\sim\nu}[\varrho^2\Delta^2] + E \tag{5.6}$$

$$\text{where } E := \mathbf{E}_{(x,a,r)\sim\nu}[\frac{\nu(a|x)}{\hat{\mu}(a|x)}\cdot\varrho\cdot(r-\hat{r})^2].$$

We want to get an expression for the variance of $\hat{V}_k$. By (5.3),

$$\mathbf{E}_{\mu}[\hat{V}_k] = \mathbf{E}_{(x,a)\sim\nu}[\hat{r}(x,\nu) - \varrho\Delta]. \tag{5.7}$$

We combine this equation with (5.6) and get

$$\begin{aligned} \mathbf{V}_{\mu}[\hat{V}_k] &= \mathbf{V}_{(x,a)\sim\nu}[\hat{r}(x,\nu)-\varrho\Delta] - \mathbf{E}_{(x,a)\sim\nu}[\varrho^2\Delta^2] + E \\ &= \mathbf{V}_{x\sim D}[\mathbf{E}_{a\sim\nu(\cdot|x)}[\hat{r}(x,\nu)-\varrho\Delta]] + \mathbf{E}_{x\sim D}[\mathbf{V}_{a\sim\nu(\cdot|x)}[\hat{r}(x,\nu)-\varrho\Delta]] \\ &\quad -\mathbf{E}_{x\sim D}[\mathbf{V}_{a\sim\nu(\cdot|x)}[\varrho\Delta]] - \mathbf{E}_{x\sim D}[\mathbf{E}_{a\sim\nu(\cdot|x)}[\varrho\Delta]^2] + E \\ &= \mathbf{V}_{x\sim D}[\mathbf{E}_{a\sim\nu(\cdot|x)}[r^* - (1-\varrho)\Delta]] - \mathbf{E}_{x\sim D}[\mathbf{E}_{a\sim\nu(\cdot|x)}[\varrho\Delta]^2] + E. \end{aligned}$$

By decomposing $E$,

$$\begin{aligned} E &= \mathbf{E}_{(x,a,r)\sim\nu}[\frac{\nu(a|x)}{\hat{\mu}(a|x)}\cdot\varrho\cdot(r-r^*)^2] + \mathbf{E}_{(x,a)\sim\nu}[\frac{\nu(a|x)}{\hat{\mu}(a|x)}\cdot\varrho\cdot(r^*-\hat{r})^2] \\ &= \mathbf{E}_{(x,a)\sim\nu}[\frac{\nu(a|x)}{\hat{\mu}(a|x)}\cdot\varrho\cdot\mathbf{V}_{r\sim D(\cdot|x,a)}[r]] + \mathbf{E}_{(x,a)\sim\nu}[\frac{\nu(a|x)}{\hat{\mu}(a|x)}\cdot\varrho\cdot\Delta^2]. \end{aligned}$$

Thus, we get the first part of Lemma 5.4.

(2)

$$\begin{aligned} \hat{r}(x,\nu)^2 &= (r^*(x,\nu) + \mathbf{E}_{a\sim\nu(\cdot|x)}[\Delta(x,a)])^2 \\ &= r^*(x,\nu)^2 + 2r^*(x,\nu)\mathbf{E}_{a\sim\nu(\cdot|x)}[\Delta(x,a)] + \mathbf{E}_{a\sim\nu(\cdot|x)}[\Delta(x,a)]^2 \\ &= r^*(x,\nu)^2 + 2\hat{r}(x,\nu)\mathbf{E}_{a\sim\nu(\cdot|x)}[\Delta(x,a)] - \mathbf{E}_{a\sim\nu(\cdot|x)}[\Delta(x,a)]^2 \\ &\leq r^*(x,\nu)^2 + 2\hat{r}(x,\nu)\mathbf{E}_{a\sim\nu(\cdot|x)}[\Delta(x,a)]. \end{aligned}$$

We plug this inequality into (5.5):

$$\begin{aligned}
&\mathbf{E}_\mu[\hat{V}_k^2] \\
&\le \mathbf{E}_{x\sim D}[r^*(x,\nu)^2] + 2\mathbf{E}_{x\sim D}[\hat{r}(x,\nu)\mathbf{E}_{a\sim\nu(\cdot|x)}[\Delta]] \\
&\quad +2\mathbf{E}_{(x,a)\sim\nu}[\hat{r}(x,\nu)\cdot(1-\varrho)\cdot\Delta] + E \\
&= \mathbf{E}_{x\sim D}[r^*(x,\nu)^2] + 2\mathbf{E}_{(x,a)\sim\nu}[\hat{r}(x,\nu)\cdot(1-\varrho)\cdot\Delta] + E.
\end{aligned}$$

Combine the above inequality with (5.6) as

$$\mathbf{E}_\mu[\hat{V}_k] = \mathbf{E}_{(x,a)\sim\nu}[r^*(x,\nu) - (1-\varrho)\Delta],$$

we have

$$\begin{aligned}
&\mathbf{V}_\mu[\hat{V}_k] \\
&\le \mathbf{V}_{x\sim D}[r^*(x,\nu)] + 2\mathbf{E}_{(x,a)\sim\nu}[\hat{r}(x,\nu) - (1-\varrho)\Delta] \\
&\quad -2\mathbf{E}_{x\sim D}[r^*(x,\nu)]\mathbf{E}_{(x,a)\sim\nu}[(1-\varrho)\Delta] - \mathbf{E}_{(x,a)\sim\nu}[(1-\varrho)\Delta]^2 + E \\
&\le \mathbf{V}_{x\sim D}[r^*(x,\nu)] + 2\mathbf{E}_{(x,a)\sim\nu}[(\hat{r}(x,\nu) - 1/2)(1-\varrho)\Delta] \\
&\quad -2\mathbf{E}_{x\sim D}[r^*(x,\nu) - 1/2]\mathbf{E}_{(x,a)\sim\nu}[(1-\varrho)\Delta] + E \\
&\le \mathbf{V}_{x\sim D}[r^*(x,\nu)] + \mathbf{E}_{(x,a)\sim\nu}[|(1-\varrho)\Delta|] + |\mathbf{E}_{(x,a)\sim\nu}[(1-\varrho)\Delta]| + E,
\end{aligned}$$

where the last inequality follows from Hölder's inequality and the observations that $|\hat{r} - 1/2| \le 1/2$ and $|r^* - 1/2| \le 1/2$. The second part of Lemma 5.4 follows by the bound

$$E \le M \cdot \mathbf{E}_{(x,a)\sim\nu}[\varrho\mathbf{E}_{r\sim D(\cdot|x,a)}[(r-\hat{r})^2]].$$

□

Lemma 5.2 shows that the range of $\hat{V}_k$ is controlled by $\nu(a_k|x_k)/\hat{\mu}(a_k|x_k)$. Lemma 5.3 shows that as $\Delta$ and $\varrho$ get more accurate (the estimates $\hat{r}$ and $\hat{\mu}$ get more accurate), the bias of $\hat{V}_k$ gets smaller.

Lemma 5.4(1) shows that various factors control the variance of $\hat{V}_k$. The first term in Lemma 5.4(1) shows the variance due to the randomness over $x$. The second term can decrease the variance as it has a negative value. The third and fourth terms are due to IPS as both of them have $\nu(a|x)/\hat{\mu}(a|x)$.

Lemma 5.4(2) gives an upper bound to the variance. The first term is the variance of the estimated variable over $x$. The second term is due to the error in the estimates $\hat{r}$ and $\hat{\mu}$; if either of them is accurate, the second term is 0. The last term is due to IPS, so if we do not use IPS, $\varrho = 0$ and the term is 0. With non-zero $\varrho$, it decreases with better $\hat{r}$.

5.2. **Bias analysis.** Lemma 5.3 gives a bound to the bias of DR estimator, which is

$$|\mathbf{E}_\mu[\hat{V}_{DR}] - V| = \frac{1}{n}|\mathbf{E}_\mu[\sum_{k=1}^{n}\mathbf{E}_{(x,a)\sim\nu}[(1-\varrho(x,a))\Delta(x,a)]]|.$$

Because we are only dealing with stationary policies, the next theorem immediately follows.

**Theorem 5.8.** *If the exploration policy $\mu$ is stationary so that its estimator $\hat{\mu}$ is also stationary, the expression above is simplified to*

$$|\mathbf{E}_\mu[\hat{V}_{DR}]| = |\mathbf{E}_\nu[(1-\varrho(x,a))\Delta(x,a)]|$$

*Proof.*

$$\begin{aligned}|\mathbf{E}_\mu[\hat{V}_{DR}] - V| &= \frac{1}{n}|\mathbf{E}_\mu[\sum_{k=1}^{n}\mathbf{E}_{(x,a)\sim\nu}[(1-\varrho(x,a))\Delta(x,a)]]| \\ &= \frac{n}{n}|\mathbf{E}_{(x,a)\sim\nu}[(1-\varrho(x,a))\Delta(x,a)]| \\ &= |\mathbf{E}_\nu[(1-\varrho(x,a))\Delta(x,a)]|.\end{aligned}$$

□

Also,

$$|\mathbf{E}_\mu[\hat{V}_{DM}] - V| = |\mathbf{E}_\nu[\Delta(x,a)]|$$
$$|\mathbf{E}_\mu[\hat{V}_{IPS}] - V| = |\mathbf{E}_\nu[r^*(x,a)(1-\varrho(x,a))]|$$

as DM and IPS are special cases of DR with $\hat{\mu} = \infty$ and $\hat{r} = 0$, respectively.

In general, both $\hat{\mu}$ and $\hat{r}$ do not dominate each other. However, as we can see in the equations above, if either of the estimators is accurate, the expected value of DR is close to the truth value $V$. Moreover, if $|\varrho - 1|$ is much smaller than 1, DR will perform better than DM. Also, if $\varrho \approx 1$ and $|\Delta|$ is much smaller than $|r^*|$, DR will perform better than IPS. Therefore, DR performs better than the others while taking advantage of both of them.

5.3. **Variance analysis.** In the previous section, we saw that the expected value of DR is closer to the truth value $V$ than that of DM or IPS. Having lower variance means that as the sample size increases, the estimate converges fast to the truth value. Thus, it is worth investigating the variance of the estimators. For a simple comparison between the estimators, we assume that the new policy $\nu$ is deterministic, i.e. given a context, the new policy deterministically chooses an action.

**Theorem 5.9.** *If the old policy $\mu$ is stationary and the new policy $\nu$ is deterministic, the variance of DR estimator is*

$$\begin{aligned}V_\mu[\hat{V}_{DR}] = \frac{1}{n}(&\mathbf{V}_{(x,a)\sim\nu}[r^*(x,a) + (1-\varrho(x,a))\Delta(x,a)] \\ &+\mathbf{E}_{(x,a)\sim\nu}[\frac{1}{\hat{\mu}(a|x)}\cdot\varrho(x,a)\cdot\mathbf{V}_{r\sim D(\cdot|x,a)}[r]] \\ &+\mathbf{E}_{(x,a)\sim\nu}[\frac{1-\mu(a|x)}{\hat{\mu}(a|x)}\cdot\varrho(x,a)\Delta(x,a)^2])\end{aligned}$$

*Proof.* Taking $\nu(a|x) = 1$ for the chosen action $a$ and $\mu$ as stationary, the theorem immediately follows from Lemma 2.4(1). □

The theorem above shows that the variance can be viewed as being controlled by three terms. The first term is due to the randomness over $x$. The other two terms are due to the importance weighting in IPS, which means that they disappear in DM.

We also get the variance of IPS

$$\begin{aligned} V_\mu[\hat{V}_{IPS}] = \frac{1}{n}(\mathbf{V}_{(x,a)\sim\nu}[\varrho(x,a)r^*(x,a)] \\ +\mathbf{E}_{(x,a)\sim\nu}[\frac{1}{\hat{\mu}(a|x)} \cdot \varrho(x,a) \cdot \mathbf{V}_{r\sim D(\cdot|x,a)}[r]] \\ +\mathbf{E}_{(x,a)\sim\nu}[\frac{1-\mu(a|x)}{\hat{\mu}(a|x)} \cdot \varrho(x,a)\Delta(x,a)^2]). \end{aligned}$$

Assuming that $\varrho \approx 1$, the first term of the variance of IPS would be similar with that of DR. The second term is the same for both of them. However, the third term can get much bigger for IPS if $\mu(a|x)$ is much smaller than 1 and $|\Delta(x,a)|$ is smaller than $r^*(x,a)$ for the actions chosen by $\nu$.

Also, the variance of DM is

$$V_\mu[\hat{V}_{IPS}] = \frac{1}{n}\mathbf{V}_{(x,a)\sim\nu}[r^*(x,a) + \Delta(x,a)].$$

Since the variance of DM does not depend on the factors that control the variance of DR or IPS, it is usually significantly lower than that of DR or IPS. However, we know that it is likely that the agent has almost full knowledge of the old policy and have $\varrho \approx 1$ with a good estimate $\hat{\mu}$. In that case, the large bias of DM causes a larger error than relatively bigger variance of DR.

## 6. Conclusion

We showed that theoretically, doubly robust estimator is unbiased and has almost always lower variance than widely used inverse propensity score method. Because it gives a more accurate estimate than the basic approaches introduced, we expect that it can be a common method for offline policy evaluation for contextual bandit problem.

## Acknowledgments

It is a pleasure to thank my mentor, William Cooperman, for all his great help in everything including writing this paper. He has always given me profound advice, and nothing could be possible without the help of my amazing mentor. I would also like to express my very great appreciation to Professor Jon Peter May for providing such a wonderful opportunity to participate in the UChicago Mathematics Research Experience for Undergraduates program (REU).

## References

[1] Dudk, M., Erhan, D., Langford, J., and Li, L. (2014). Doubly robust policy evaluation and optimization. In *Statistical Science*

[2] Strehl, A., Langford, J., Li, L. and Kakade, S. (2010). Learning from logged implicit exploration data. In *Advances in Neural Information Processing Systems (NIPS)*