Optimal Mixture Weights for Off-Policy Evaluation with Multiple Behavior Policies

Jinlin Lai Lixin Zou Jiaxing Song
Department of Computer Science and Technology
Tsinghua University
jinlinlai@cs.umass.edu,zoulx15@mails.tsinghua.edu.cn,jxsong@tsinghua.edu.cn

Abstract

Off-policy evaluation is a key component of reinforcement learning which evaluates a target policy with offline data collected from behavior policies. It is a crucial step towards safe reinforcement learning and has been used in advertisement, recommender systems and many other applications. In these applications, sometimes the offline data is collected from multiple behavior policies. Previous works regard data from different behavior policies equally. Nevertheless, some behavior policies are better at producing good estimators while others are not. This paper starts with discussing how to correctly mix estimators produced by different behavior policies. We propose three ways to reduce the variance of the mixture estimator when all sub-estimators are unbiased or asymptotically unbiased. Furthermore, experiments on simulated recommender systems show that our methods are effective in reducing the Mean-Square Error of estimation.

1 Introduction

In applications of reinforcement learning [24], it is usually unsafe or risky to use a policy without evaluating it. For example, in reinforcement learning based recommender systems, if a defective policy is deployed, it can cause irreversible loss like losing customers. To tackle the problem, Off-Policy Evaluation (OPE) algorithms are developed to evaluate a target policy with data collected from online behavior policies in an offline manner. OPE has been used to evaluate reinforcement learning applications in advertisements, recommender systems and many other areas [13, 4, 11, 25, 8]. The most influential algorithm in OPE is Doubly Robust estimation [6, 10]. Based on it, some recent works [27, 7, 2, 14] in OPE explore different ways to reduce the Mean-Square Error (MSE) of estimation. However, few of them discuss deeply about how to evaluate with multiple behavior policies. In such cases, most current methods will directly go through after regarding data from different behavior policies as a whole. However, in later sections of this paper, we show that better results can be reached if we split data by behavior policy, construct split estimators for the split data and consider the mixture estimator of the split estimators. The root cause of this issue is that some behavior policies are "adding" high variance to the result. Therefore, if we can assign high weights to "good" behavior policies and low weights to "bad" ones, better estimation can be obtained.

In this paper, we optimize the mixture weights for the mixture estimator by minimizing variance. This idea has been discussed in Agarwal et al. [1] for Importance Sampling estimators of contextual bandits. We generalize this idea to finite-horizon Markov Decision Process and derive naive mixture estimators for most OPE algorithms, including Importance Sampling, Weighted Importance Sampling, Doubly Robust estimation and Weighted Doubly Robust estimation. After exploiting the structure of reinforcement learning, we further propose mixture estimators and $\alpha\beta$ mixture estimators, which have theoretically lower variances than naive mixture estimators. To compute the optimal weights, we introduce Delta Method [18] from asymptotic statistics to estimate the variances and covariances

Offline Reinforcement Learning Workshop at Neural Information Processing Systems, 2020.

of the components of weighted estimators. In our experiments on simulated recommender systems, we show that mixture estimators are effective in reducing the MSE of all the estimators.

2 Preliminaries

2.1 Markov Decision Process

Markov Decision Process (MDP) [24] is represented by $< \mathcal{S}, \mathcal{A}, R, P, P_0, \gamma >$, where $\mathcal{S}$ and $\mathcal{A}$ are state space and action space, $R(s,a)$ is a random variable indicating the immediate reward of taking action a at state s, $P(\cdot|s,a)$ is the state transition distribution, $P_0(\cdot)$ is the distribution of initial state, and γ is the discount factor. To interact in such environment, a policy π is given and $\pi(a|s)$ is the probability of taking a in state s.

2.2 Off-policy Evaluation with Single Behavior Policy

In the literature of reinforcement learning, there are many algorithms to evaluate a new policy π with data collected from one behavioral policy π_0. In this section, we assume there are N data trajectories and the i-th data is $(s_{i,0}, a_{i,0}, r_{i,0}, s_{i,1}, a_{i,1}, r_{i,1}, ...)$.

Direct Method (DM) fits $\mathbb{E}[R(s,a)]$ and $P(\cdot|s,a)$ by regression [10]. With the approximated functions $\hat{R}(s,a)$ and $\hat{P}(\cdot|s,a)$, the value functions are recursively updated ($\hat{V}_0(s) = 0$):

$$\hat{Q}_t(s,a) = \hat{R}(s,a) + \gamma \mathbb{E}_{s' \sim \hat{P}(s'|s,a)}[\hat{V}_{t-1}(s')] \quad \hat{V}_t(s) = \mathbb{E}_{a \sim \pi(a|s)}[\hat{Q}_t(s,a)], \tag{1}$$

$$\hat{Q}(s,a) = \lim_{t \to \infty} \hat{Q}_t(s,a) \quad \hat{V}(s) = \lim_{t \to \infty} \hat{V}_t(s). \tag{2}$$

The value of the new policy would be estimated by $\hat{V}_{DM} = \mathbb{E}_{s \sim P_0(s)}[\hat{V}(s)]$.

Besides DM, another family of OPE technique is Importance Sampling(IS). IS estimates the value by $\hat{V}_{IS} = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{T} \gamma^t \rho_{i,t} r_{i,t}$, where $\rho_{i,t} = \prod_{\tau=0}^{t} \frac{\pi(a_{i,\tau}|s_{i,\tau})}{\pi_0(a_{i,\tau}|s_{i,\tau})}$ and T is the maximal horizon of data.

DM typically has low variance and high bias. IS is proved to be unbiased but suffers from high variance. Doubly Robust estimation(DR) [10] combines DM and IS. It can be regarded as IS with control variates so it has lower variance than IS. We follow Thomas and Brunskill [27] and formulate DR as $\hat{V}_{DR} = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{T} \gamma^t \left(\rho_{i,t-1} \hat{V}(s_{i,t}) + \rho_{i,t}(r_{i,t} - \hat{Q}(s_{i,t}, a_{i,t})) \right)$.

Weighted Importance Sampling(WIS) [20] is also a variance reduction technique for IS. It is derived by replacing $\frac{\rho_{i,t}}{N}$ in IS with $w_{i,t} = \frac{\rho_{i,t}}{\sum_{i=1}^{N} \rho_{i,t}}$. WIS is asymptotically unbiased and has lower variance than IS. Similarly, Weighted Doubly Robust estimation(WDR) [27] has lower variance than DR. It also replaces $\frac{\rho_{i,t}}{N}$ in DR with $w_{i,t}$.

2.3 Problem Setting

In applications of reinforcement learning, there might be multiple behavior policies in the same environment. Suppose we have M behavior policies $\pi_1, \pi_2, ..., \pi_M$. The i-th behavior policy π_i collects n_i data. The j-th data from π_i is $(s_{i,j,0}, a_{i,j,0}, r_{i,j,0}, s_{i,j,1}, a_{i,j,1}, r_{i,j,1}, ...)$. With the data from π_i, we can build an asymptotically unbiased estimator $\hat{V}_i$ to evaluate the target policy π. This paper begins with constructing the mixture estimator of the M estimators. The goal is to minimize the MSE of estimation. If we estimate θ with $\hat{\theta}$, the MSE is formulated as $\mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{E}^2[\hat{\theta} - \theta] + \mathbb{V}[\hat{\theta}]$. For IS and DR, MSE reduces to variance so we directly minimize the variance. For WIS and WDR, by Delta Method [18], the bias squared is $O(n^{-2})$ while the variance is $O(n^{-1})$. When n is large, the bias squared is dominated by variance so we neglect the bias of weighted estimators and also directly minimize the variance.

To make our idea clear, we define our symbol system of the following sections here. Bold letters are random variables (like $\boldsymbol{X}$). Letters with a subscript j are samples of the corresponding random variable (for example, $X_{i,j}$ is a sample of $\boldsymbol{X}_i$). Letters with a hat are estimators (like $\hat{X}$). Letters with an arrow are vectors (like $\overrightarrow{X}$).

Method	Estimator
IS	$\hat{V}_{IS} = \frac{1}{\sum_{i=1}^{M} n_i} \sum_{i=1}^{M} \sum_{j=1}^{n_i} \sum_{t=0}^{T} \gamma^t \rho_{i,j,t} r_{i,j,t}$
WIS	$\hat{V}_{WIS} = \sum_{i=1}^{M} \sum_{j=1}^{n_i} \sum_{t=0}^{T} \gamma^t w_{i,j,t} r_{i,j,t}$
DR	$\hat{V}_{DR} = \frac{1}{\sum_{i=1}^{M} n_i} \sum_{i=1}^{M} \sum_{j=1}^{n_i} \sum_{t=0}^{T} \gamma^t \left(\rho_{i,j,t-1} \hat{V}(s_{i,j,t}) + \rho_{i,j,t} (r_{i,j,t} - \hat{Q}(s_{i,j,t}, a_{i,j,t})) \right)$
WDR	$\hat{V}_{WDR} = \sum_{i=1}^{M} \sum_{j=1}^{n_i} \sum_{t=0}^{T} \gamma^t \left(w_{i,j,t-1} \hat{V}(s_{i,j,t}) + w_{i,j,t} (r_{i,j,t} - \hat{Q}(s_{i,j,t}, a_{i,j,t})) \right)$
SWIS	$\hat{V}_{SWIS} = \sum_{i=1}^{M} \frac{n_i}{\sum_{i'=1}^{M} n'_i} \sum_{j=1}^{n_i} \sum_{t=0}^{T} \gamma^t u_{i,j,t} r_{i,j,t}$
SWDR	$\hat{V}_{SWDR} = \sum_{i=1}^{M} \frac{n_i}{\sum_{i'=1}^{M} n'_i} \sum_{j=1}^{n_i} \sum_{t=0}^{T} \gamma^t \left(u_{i,j,t-1} \hat{V}(s_{i,j,t}) + u_{i,j,t} (r_{i,j,t} - \hat{Q}(s_{i,j,t}, a_{i,j,t})) \right)$

Table 1: Formulas for the vanilla estimators.

To guarantee the theoretical results in this paper, we give the following assumptions.

Assumption 1 *$\forall s, a, i$, if $\pi(a|s) > 0$, then $\pi_i(a|s) > 0$. Furthermore, there exists $\beta > 0$, such that $\forall i, t$, $\rho_{i,t} \leq \beta$.*

Assumption 2 *There exists $\zeta > 0$, such that $\forall s, a$, $\forall r \sim R(s, a)$, $|r| \leq \zeta$.*

Assumption 3 *For any estimator $\hat{\theta}$, we ignore its bias.*

Assumption 4 *Any two different data trajectories are independent.*

3 Optimal Mixture Weights with Multiple Behavior Policies

3.1 Value Estimators with Multiple Behavior Policies

With multiple behavior policies, we can construct value estimators based on Section 2.2. Define

$$\rho_{i,j,t} = \prod_{\tau=0}^{t} \frac{\pi(a_{i,j,\tau}|s_{i,j,\tau})}{\pi_i(a_{i,j,\tau}|s_{i,j,\tau})}, \tag{3}$$

$$w_{i,j,t} = \frac{\rho_{i,j,t}}{\sum_{i'=1}^{M} \sum_{j'=1}^{n_{i'}} \rho_{i',j',t}}, \tag{4}$$

then the IS, WIS, DR and WDR estimators are listed in Table 1.

Note that WIS and WDR normalize the importance weights across all data. We can also normalize inside each behavior policy by

$$u_{i,j,t} = \frac{\rho_{i,j,t}}{\sum_{j'=1}^{n_{i'}} \rho_{i,j',t}} \tag{5}$$

and construct split weighted estimators. We call them Split WIS (SWIS) and Split WDR (SWDR). See Table 1 for their formulas.

In our experiments, we show that there is little difference between the performances of weighted estimators and split weighted estimators. Nevertheless, the advantage of split weighted estimators is that they can be divided into sub-estimators. This makes optimizing the mixture weights of weighted estimators possible.

3.2 Naive Mixture Estimators

The central idea of this paper is to split each estimator into sub-estimators, assign weights to the sub-estimators and optimize the weights. For IS, SWIS, DR and SWDR, the first idea is to split them according to behavior policy. Taking IS for example, it can be rewritten as

$$\hat{V}_{IS} = \sum_{i=1}^{M} \frac{n_i}{\sum_{i'=1}^{M} n'_i} \hat{V}_{IS,i}, \tag{6}$$

where $\hat{V}_{IS,i} = \frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{t=0}^{T}\gamma^t\rho_{i,j,t}r_{i,j,t}$ is estimator for the target value V. We then replace $\frac{n_i}{\sum_{i'=1}^{M}n_i'}$ with mixture weights α_i, form $\hat{V}_{NMIS} = \sum_{i=1}^{M}\alpha_i\hat{V}_{IS,i}$ and optimize the weights. The weights should satisfy $\sum_{i=1}^{M}\alpha_i = 1$. The following theorem gives the optimal mixture weights of this problem:

Theorem 1 *Given M unbiased and independent estimators $\hat{V}_1, \hat{V}_2, ..., \hat{V}_M$ of a value V, the mixture estimator of them with the minimal variance is $\hat{V}_{MIX} = \sum_{i=1}^{M}\alpha_i^*\hat{V}_i$, where $\alpha_i^* = \frac{1}{\mathbb{V}[\hat{V}_i]\sum_{i'=1}^{M}\frac{1}{\mathbb{V}[\hat{V}_{i'}]}}$. The minimal variance is $\frac{1}{\sum_{i=1}^{M}\frac{1}{\mathbb{V}[\hat{V}_i]}}$.*

See Appendix A.1 for proof. The condition of independence comes from Assumption 4. Theorem 1 is the general case for Section 6 of Agarwal et al. [1] as well as the basis of Fixed Effect Model in Meta Analysis [3]. It can also be applied to SWIS, DR and SWDR. We call this estimator naive mixture estimator because it does not consider the properties of reinforcement learning.

3.3 Mixture Estimators for Off-Policy Evaluation

We can further split the estimators by t. Taking IS for example, it can be formulated as

$$\hat{V}_{IS} = \sum_{i=1}^{M}\sum_{t=0}^{T}\frac{n_i}{\sum_{i'=1}^{M}n_i'}\hat{V}_{IS,i,t}, \tag{7}$$

where $\hat{V}_{IS,i,t} = \frac{1}{n_i}\sum_{j=1}^{n_i}\gamma^t\rho_{i,j,t}r_{i,j,t}$ is estimator for the value at t denoted by V_t. We can replace $\frac{n_i}{\sum_{i'=1}^{M}n_i'}$ with $\alpha_{i,t}$ and form $\hat{V}_{MIS} = \sum_{i=1}^{M}\sum_{t=0}^{T}\alpha_{i,t}\hat{V}_{IS,i,t}$. The weights should satisfy $\forall t,\ \sum_{i=1}^{M}\alpha_{i,t} = 1$. The following proposition shows how to optimize such mixture weights.

Proposition 1 *Denote the covariance matrix of $[\hat{V}_{i,0}, \hat{V}_{i,1}, ..., \hat{V}_{i,T}]$ by Σ_i. If*

- *$\forall i_1, i_2, t_1, t_2$, if $i_1 \neq i_2$, then $\hat{V}_{i_1,t_1}$ and $\hat{V}_{i_2,t_2}$ are independent;*
- *$\forall i \forall t$, $\hat{V}_{i,t}$ is unbiased for V_t;*
- *$\forall i$ Σ_i is positive definite;*

then $\hat{V}_{MIXT} = \sum_{i=1}^{M}\sum_{t=0}^{T}\alpha_{i,t}\hat{V}_{i,t}$ is unbiased for V and the mixture weights that minimize variance of $\hat{V}_{MIXT}$ are $\overrightarrow{\alpha}_i^* = \Sigma_i^{-1}(\sum_{i'=1}^{M}\Sigma_{i'}^{-1})^{-1}\overrightarrow{e}$, where $\overrightarrow{\alpha}_i^* = [\alpha_{i,0}^*, \alpha_{i,1}^*, ..., \alpha_{i,T}^*]^T$ and $\overrightarrow{e}$ is $[1, 1, ..., 1]^T$. Moreover, $\mathbb{V}[\hat{V}_{MIXT}] \leq \mathbb{V}[\hat{V}_{MIX}]$ if $\forall i,\ \hat{V}_i = \sum_{t=0}^{T}\hat{V}_{i,t}$.

See appendix A.2 for proof. The assumption of positive definite is not hard to reach in real world problems. If the reward is constant at some horizon, we can simply remove it from our formula and still get a positive definite covariance matrix. The same results also hold for SWIS, DR and SWDR.

3.4 $\alpha\beta$ Mixture Estimators for Off-Policy Evaluation

Compared with IS and SWIS, DR and SWDR both have control variate terms. These terms reduce the variance of estimation [27]. Taking DR for example, we can divide the estimator to IS plus the control variates and formulate as

$$\hat{V}_{DR} = \sum_{i=1}^{M}\sum_{t=0}^{T}\frac{n_i}{\sum_{i'=1}^{M}n_i'}(\hat{V}_{IS,i,t} + \hat{W}_{DR,i,t}), \tag{8}$$

where $\hat{W}_{DR,i,t} = \frac{1}{n_i}\sum_{j=1}^{n_i}\gamma^t\left(\rho_{i,j,t-1}\hat{V}(s_{i,j,t}) - \rho_{i,j,t}\hat{Q}(s_{i,j,t}, a_{i,j,t})\right)$ are estimators for 0. Similar to the previous sections, we assign $\alpha_{i,t}$ to $\hat{V}_{IS,i,t}$ and $\beta_{i,t}$ to $\hat{W}_{DR,i,t}$ and form $\hat{V}_{\alpha\beta MDR} = \sum_{i=1}^{M}\sum_{t=0}^{T}(\alpha_{i,t}\hat{V}_{IS,i,t} + \beta_{i,t}\hat{W}_{DR,i,t})$. The weights should satisfy $\forall t,\ \sum_{i=1}^{M}\alpha_{i,t} = 1$. The following proposition derives optimal mixture weights for mixture estimator with control variates.

Proposition 2 *Given $M*(T+1)$ estimators $\hat{V}_{i,t}$ and $M*(T+1)$ estimators $\hat{W}_{i,t}$, if*

- $\forall i_1, i_2, t_1, t_2$, *if* $i_1 \neq i_2$, *then* $\hat{V}_{i_1,t_1}$ *and* $\hat{V}_{i_2,t_2}$ *are independent,* $\hat{W}_{i_1,t_1}$ *and* $\hat{W}_{i_2,t_2}$ *are independent,* $\hat{V}_{i_1,t_1}$ *and* $\hat{W}_{i_2,t_2}$ *are independent;*
- $\forall i, t$, $\mathbb{E}[\hat{V}_{i,t}] = V_t$ *and* $\mathbb{E}[\hat{W}_{i,t}] = 0$;

then the mixture weights that minimize the variance of mixture estimator with control variates for the estimators are $\overrightarrow{\alpha}_i^* = H_{i,11}(\sum_{i'=1}^{M} H_{i',11})^{-1}\overrightarrow{e}$, $\overrightarrow{\beta}_i^* = H_{i,21}(\sum_{i'=1}^{M} H_{i',11})^{-1}\overrightarrow{e}$, *where* $\begin{pmatrix} H_{i,11} & H_{i,12} \\ H_{i,21} & H_{i,22} \end{pmatrix}$ *is the precision matrix of* $[\hat{V}_{i,0}, \hat{V}_{i,1}, ..., \hat{V}_{i,t}, \hat{W}_{i,0}, \hat{W}_{i,1}, ..., \hat{W}_{i,t}]^T$.

See Appendix A.3 for proof. We call this estimator $\alpha\beta$ mixture estimator. The formulations for all three types of mixture estimators can be found in Appendix C.1.

4 Variance estimators

The mixture estimators in Section 3 rely on the estimation of variances and covariance matrixes. By Assumption 1 and 2, we can get strongly consistent variance estimators. Taking $\hat{V}_{IS,i}$ for example, it is formulated as

$$\hat{V}_{IS,i} = \frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{t=0}^{T}\gamma^t \rho_{i,j,t} r_{i,j,t} = \frac{1}{n_i}\sum_{j=1}^{n_i}\hat{V}_{IS,i,j}, \tag{9}$$

which can be interpreted by average of n_i samples of a random variable $\boldsymbol{V}_{IS,i}$. So $\mathbb{V}[\hat{V}_{IS,i}] = \frac{1}{n_i}\mathbb{V}[\boldsymbol{V}_{IS,i}]$. In this paper, we use half of the data to estimate $\mathbb{V}[\boldsymbol{V}_{IS,i}]$, plug the estimated variances into the formulas and estimate the value by the other half of the data. This strategy can be applied to components of IS and DR in all three types of mixture estimators.

However, for SWIS and SWDR, we can not regard each component as average of samples. Rather, we should regard them as function of average of samples. For example,

$$\begin{aligned}
\hat{V}_{SWDR,i} &= \sum_{j=1}^{n_i}\sum_{t=0}^{T}\gamma^t\left(u_{i,j,t-1}\hat{V}(s_{i,j,t}) + u_{i,j,t}(r_{i,j,t} - \hat{Q}(s_{i,j,t}, a_{i,j,t}))\right) \\
&= \sum_{t=0}^{T}\gamma^t\left(\frac{\sum_{j=1}^{n_i}\rho_{i,j,t-1}\hat{V}(s_{i,j,t})}{\sum_{j=1}^{n_i}\rho_{i,j,t-1}} + \frac{\sum_{j=1}^{n_i}\rho_{i,j,t}(r_{i,j,t} - \hat{Q}(s_{i,j,t}, a_{i,j,t}))}{\sum_{j=1}^{n_i}\rho_{i,j,t}}\right) \\
&= \sum_{t=0}^{T}\gamma^t\left(\frac{\frac{1}{n_i}\sum_{j=1}^{n_i}\rho_{i,j,t-1}\hat{V}(s_{i,j,t})}{\frac{1}{n_i}\sum_{j=1}^{n_i}\rho_{i,j,t-1}} + \frac{\frac{1}{n_i}\sum_{j=1}^{n_i}\rho_{i,j,t}(r_{i,j,t} - \hat{Q}(s_{i,j,t}, a_{i,j,t}))}{\frac{1}{n_i}\sum_{j=1}^{n_i}\rho_{i,j,t}}\right) \\
&\triangleq \sum_{t=0}^{T}\left(\frac{\frac{1}{n_i}\sum_{j=1}^{n_i}\hat{X}_{i,j,t}}{\frac{1}{n_i}\sum_{j=1}^{n_i}\hat{W}_{i,j,t}} + \frac{\frac{1}{n_i}\sum_{j=1}^{n_i}\hat{Z}_{i,j,t}}{\frac{1}{n_i}\sum_{j=1}^{n_i}\hat{Y}_{i,j,t}}\right)
\end{aligned} \tag{10}$$

We approximate $\mathbb{V}[\hat{V}_{SWDR,i,t}]$ by Delta Method [18]. See appendix B for introduction.

The variance and covariance estimators for components of all the estimators are in Appendix C.3.

5 Related Work

Recent advances in OPE can be split into two categories: Importance Sampling based OPE and stationary distribution based OPE. For Importance Sampling based OPE, previous works about mixture estimators mainly focus on mixing different kinds of estimators. Thomas and Brunskill [27] optimize a mixture weight between WDR and Direct Method to reduce the MSE of estimation. Following it, recent works [29, 21, 22, 23] propose more strategies to blend different off-policy evaluation algorithms. The mixture estimators in this paper are different from theirs. They mix different kinds of estimators, while we mix estimators of different behavior policies. Agarwal et al. [1]

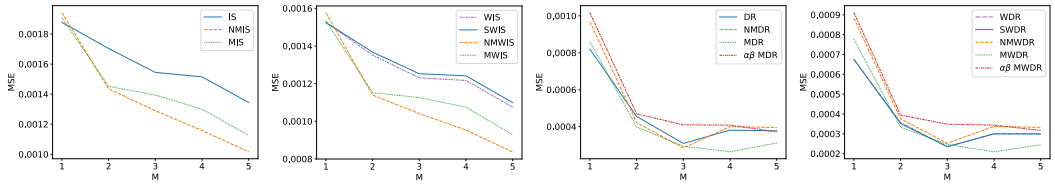

Figure 1: MSE of all four types of estimators.

derive Weighted IPS estimators, which is actually NMIS estimators for contextual bandits. Compared with them, we not only generalize the idea to finite-horizon MDP, but also apply our techniques to more OPE algorithms. Additionally, Balanced IPS estimator [1] or Multiple Importance Sampling [18] is another way to reduce variance for OPE of contextual bandits with multiple behavior policies. However, as far as we know, no previous work has generalized it to finite-horizon OPE so we do not compare with it. Stationary distribution based OPE are built on the estimation of stationary distributions or ratio of stationary distributions [15, 30, 26, 28, 16, 31]. To evaluate with multiple behavior policies, Nachum et al. [17] and Chen et al. [5] build the mixture policy of the behavior policies and estimate the ratio of stationary distributions. Different from them, we optimize the mixture weights of estimators while they regard data from different policies equally. It would be interesting to see how our methods help improve these methods.

6 Experiments

6.1 Experimental Settings

We construct a simulated recommender platform based on RecSim [9]. The interactions in this platform form a Partially Observable Markov Decision Process(POMDP) [24]. Detailed configuration of this environment can be found in Appendix D.1. We implemented IS, WIS, SWIS, DR, WDR and SWDR as baselines. We call naive mixture estimators for the methods NMIS, NMWIS, NMDR and NMWDR. Mixture estimators for the methods are called MIS, MWIS, MDR, MWDR. $\alpha\beta$ mixture estimators are called $\alpha\beta$ MDR and $\alpha\beta$ MWDR. See Appendix D.3 for implementation of the OPE algorithms. For mixture estimators and $\alpha\beta$ mixture estimators, we choose a hyper-parameter T, mix the values from 0 to T and simply add up the remains. See Appendix D.4 for discussion about it.

6.2 Results

With the chosen T, the MSE of the estimators with different M on test set are plotted in Figure 1. Numerical results can be found in Appendix E.2. In each figure, when M=1, the baselines have the lowest MSE. This is because only half of the samples in mixture estimators are used to estimate values. As M increases, the MSE of all estimators decrease. For IS and WIS, when M=5, both naive mixture estimators and mixture estimators are better than baselines. However, naive mixture estimators have the best results. There are two possible reasons for it. First, mixture estimators only mix the first several values while naive mixture estimators mix values of the whole horizon. Second, mixture estimators require estimation of covariance matrix, which may amplify the error of estimation. For DR and WDR, when M=5, mixture estimators produce the best results while naive mixture estimators and $\alpha\beta$ mixture estimators produce comparable results with baselines. This indicates that $\alpha\beta$ mixture estimators are not as effective as theory. We compute the average condition number for the estimated covariance matrixes of all the mixture estimators. See Appendix E.3. With relatively large condition numbers, the error of $\alpha\beta$ mixture estimators is amplified in matrix inversion.

7 Conclusion

We derive naive mixture estimators, mixture estimators and $\alpha\beta$ mixture estimators for OPE with multiple behavior policies. To estimate the mixture weights for weighted estimators, we introduce Delta Method to estimate the variances and covariances of weighted estimators. In our experiments on simulated recommender systems, we show that naive mixture estimators and mixture estimators are effective in reducing MSE while $\alpha\beta$ mixture estimators suffer from ill covariance matrixes. Future work can focus on mitigating this problem.

References

[1] Aman Agarwal, Soumya Basu, Tobias Schnabel, and Thorsten Joachims. Effective evaluation using logged bandit feedback from multiple loggers. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, pages 687–696. ACM, 2017. doi: 10.1145/3097983.3098155.

[2] Aurélien Bibaut, Ivana Malenica, Nikos Vlassis, and Mark J. van der Laan. More efficient off-policy evaluation through regularized targeted learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 654–663. PMLR, 2019.

[3] Michael Dale Borenstein, Larry V. Hedges, Julian P T Higgins, and Hannah R. Rothstein. Introduction to meta-analysis. 2009.

[4] Léon Bottou, Jonas Peters, Joaquin Quiñonero 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. *J. Mach. Learn. Res.*, 14(1): 3207–3260, 2013.

[5] Xinyun Chen, Lu Wang, Yizhe Hang, Heng Ge, and Hongyuan Zha. Infinite-horizon off-policy policy evaluation with multiple behavior policies. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[6] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 1097–1104. Omnipress, 2011.

[7] Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1446–1455. PMLR, 2018.

[8] Alexandre Gilotte, Clément Calauzènes, Thomas Nedelec, Alexandre Abraham, and Simon Dollé. Offline A/B testing for recommender systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*, pages 198–206. ACM, 2018. doi: 10.1145/3159652.3159687.

[9] Eugene Ie, Chih-Wei Hsu, Martin Mladenov, Vihan Jain, Sanmit Narvekar, Jing Wang, Rui Wu, and Craig Boutilier. Recsim: A configurable simulation platform for recommender systems. *CoRR*, abs/1909.04847, 2019.

[10] Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 652–661. JMLR.org, 2016.

[11] Thorsten Joachims and Adith Swaminathan. Counterfactual evaluation and learning for search, recommendation and ad placement. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, pages 1199–1201. ACM, 2016. doi: 10.1145/2911451.2914803.

[12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[13] Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February 9-12, 2011*, pages 297–306. ACM, 2011. doi: 10.1145/1935826.1935878.

[14] Anqi Liu, Hao Liu, Anima Anandkumar, and Yisong Yue. Triply robust off-policy evaluation. *CoRR*, abs/1911.05811, 2019.

[15] Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 5361–5371, 2018.

[16] Ali Mousavi, Lihong Li, Qiang Liu, and Denny Zhou. Black-box off-policy estimation for infinite-horizon reinforcement learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

[17] Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 2315–2325, 2019.

[18] Art B. Owen. *Monte Carlo theory, methods and examples*. 2013.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[20] M. J. D. POWELL and J. SWANN. Weighted Uniform Sampling — a Monte Carlo Technique for Reducing Variance. *IMA Journal of Applied Mathematics*, pages 228–236, 1966.

[21] Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudík. Doubly robust off-policy evaluation with shrinkage. *CoRR*, abs/1907.09623, 2019.

[22] Yi Su, Lequn Wang, Michele Santacatterina, and Thorsten Joachims. CAB: continuous adaptive blending for policy evaluation and learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6005–6014. PMLR, 2019.

[23] Yi Su, Pavithra Srinath, and Akshay Krishnamurthy. Adaptive estimator selection for off-policy evaluation. *CoRR*, abs/2002.07729, 2020.

[24] Richard S. Sutton and Andrew G. Barto. Reinforcement learning: An introduction. *IEEE Trans. Neural Networks*, 9(5):1054–1054, 1998. doi: 10.1109/TNN.1998.712192.

[25] Adith Swaminathan, Akshay Krishnamurthy, Alekh Agarwal, Miroslav Dudík, John Langford, Damien Jose, and Imed Zitouni. Off-policy evaluation for slate recommendation. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 3632–3642, 2017.

[26] Ziyang Tang, Yihao Feng, Lihong Li, Dengyong Zhou, and Qiang Liu. Doubly robust bias reduction in infinite horizon off-policy estimation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[27] Philip S. Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 2139–2148. JMLR.org, 2016.

[28] Masatoshi Uehara and Nan Jiang. Minimax weight and q-function learning for off-policy evaluation. *CoRR*, abs/1910.12809, 2019.

[29] Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudík. Optimal and adaptive off-policy evaluation in contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3589–3597. PMLR, 2017.

[30] Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 9665–9675, 2019.

[31] Ruiyi Zhang, Bo Dai, Lihong Li, and Dale Schuurmans. Gendice: Generalized offline estimation of stationary values. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.