More Robust Doubly Robust Off-policy Evaluation

Mehrdad Farajtabar [*1] **Yinlam Chow** [*2] **Mohammad Ghavamzadeh** [2]

Abstract

We study the problem of off-policy evaluation (OPE) in reinforcement learning (RL), where the goal is to estimate the performance of a policy from the data generated by another policy(ies). In particular, we focus on the doubly robust (DR) estimators that consist of an importance sampling (IS) component and a performance model, and utilize the low (or zero) bias of IS and low variance of the model at the same time. Although the accuracy of the model has a huge impact on the overall performance of DR, most of the work on using the DR estimators in OPE has been focused on improving the IS part, and not much on how to learn the model. In this paper, we propose alternative DR estimators, called *more robust doubly robust* (MRDR), that learn the model parameter by minimizing the variance of the DR estimator. We first present a formulation for learning the DR model in RL. We then derive formulas for the variance of the DR estimator in both contextual bandits and RL, such that their gradients w.r.t. the model parameters can be estimated from the samples, and propose methods to efficiently minimize the variance. We prove that the MRDR estimators are strongly consistent and asymptotically optimal. Finally, we evaluate MRDR in bandits and RL benchmark problems, and compare its performance with the existing methods.

1. Introduction

In many real-world decision-making problems, in areas such as marketing, finance, robotics, and healthcare, deploying a policy without having an accurate estimate of its performance could be costly, unethical, or even illegal. This is why the problem of *off-policy evaluation* (OPE) has been heavily studied in contextual bandits (e.g., Dudík et al. 2011; Swaminathan et al. 2017) and reinforcement learning (RL) (e.g., Precup et al. 2000a; 2001; Paduraru 2013; Mahmood et al. 2014; Thomas et al. 2015a; Li et al. 2015; Jiang & Li 2016; Thomas & Brunskill 2016), and some of the results have been applied to problems in marketing (e.g., Li et al. 2011; Theocharous et al. 2015), healthcare (e.g., Murphy et al. 2001; Hirano et al. 2003), and education (e.g., Mandel et al. 2014; 2016). The goal in OPE is to estimate the performance of an *evaluation* policy, given a log of data generated by the *behavior* policy(ies). The OPE problem can be viewed as a form of counterfactual reasoning to infer causal effects of a new treatment from historical data (e.g., Bottou et al. 2013; Shalit et al. 2017; Louizos et al. 2017).

Three different approaches to OPE in RL can be identified in the literature. **1) Direct Method (DM)** which learns a model of the system and then uses it to estimate the performance of the evaluation policy. This approach often has low variance but its bias depends on how well the selected function class represents the system and on whether the number of samples is sufficient to accurately learn this function class. There are two major problems with this approach: *(a)* Its bias cannot be easily quantified, since in general it is difficult to quantify the approximation error of a function class; *(b)* It is not clear how to choose the loss function for model learning without the knowledge of the evaluation policy (or the distribution of the evaluation policies). Without this knowledge, we may select a loss function that focuses on learning the areas that are irrelevant for the evaluation policy(ies). **2) Importance Sampling (IS)** that uses the IS term to correct the mismatch between the distributions of the system trajectory induced by the evaluation and behavior policies. Although this approach is unbiased (under mild assumptions) in case the behavior policy is known, its variance can be very large when there is a big difference between the distributions of the evaluation and behavior policies, and grows exponentially with the horizon of the RL problem. **3) Doubly Robust (DR)** which is a combination of DM and IS, and can achieve the low variance of DM and no (or low) bias of IS. The DR estimator was first developed in statistics (e.g., Cassel et al. 1976; Robins et al. 1994; Robins & Rotnitzky 1995; Bang & Robins 2005) to estimate from incomplete data with the property that is unbiased when either of its DM or IS estimators is correct. It was brought to our community, first in contextual bandits by Dudík et al.

[*]Equal contribution [1]Georgia Tech [2]DeepMind. Correspondence to: Yinlam Chow <yinlamchow@google.com>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

(2011) and then in RL by Jiang & Li (2016). Thomas & Brunskill (2016) proposed two methods to reduce the variance of DR, with the cost of introducing a bias, one to select a low variance IS estimator, namely weighted IS (WIS), and one to blend DM and IS together (instead of simply combining them as in the standard DR approach) in a way to minimize the mean squared error (MSE).

In this paper, we propose to reduce the variance of DR in bandits and RL by designing the loss function used to learn the model in the DM part of the estimator. The main idea of our estimator, called *more robust doubly robust* (MRDR), is to learn the parameters of the DM model by minimizing the variance of the DR estimator. This idea has been investigated in statistics in the context of regression when the labels of a subset of samples are randomly missing (Cao et al., 2009). We first present a novel formulation for the DM part of the DR estimator in RL. We then derive formulas for the variance of the DR estimator in both bandits and RL in a way that its gradient w.r.t. the model parameters can be estimated from the samples. Note that the DR variances reported for bandits (Dudík et al., 2011) and RL (Jiang & Li, 2016) contain the bias of the DM component, which is unknown. We then propose methods to efficiently minimize the variance in both bandits and RL. Furthermore, we prove that the MRDR estimator is strongly consistent and asymptotically optimal. Finally, we evaluate the MRDR estimator in bandits and RL benchmark problems, and compare its performance with DM, IS, and DR approaches.

2. Preliminaries

In this paper, we consider the reinforcement learning (RL) problem in which the agent's interaction with the system is modeled as a Markov decision process (MDP). Note that the contextual bandit problem is a special case with horizon-1 decision-making. In this section, we first define MDPs and the relevant quantities that we are going to use throughout the paper, and then define the off-policy evaluation problem in RL, which is the main topic of this work.

2.1. Markov Decision Processes

A MDP is a tuple $\langle \mathcal{X}, \mathcal{A}, P_r, P, P_0, \gamma \rangle$, where $\mathcal{X}$ and $\mathcal{A}$ are the state and action spaces, $P_r(x,a)$ is the distribution of the bounded random variable $r(x,a) \in [0, R_{\max}]$ of the immediate reward of taking action a in state x, $P(\cdot|x,a)$ is the transition probability distribution, $P_0 : \mathcal{X} \to [0,1]$ is the initial state distribution, and $\gamma \in [0,1)$ is the discounting factor. A (stationary) policy $\pi : \mathcal{X} \times \mathcal{A} \to [0,1]$ is a stochastic mapping from states to actions, with $\pi(a|x)$ being the probability of taking action a in state x. We denote by P^π the state transition of the Markov chain induced by policy π, i.e., $P^\pi(x_{t+1}|x_t) = \sum_{a\in\mathcal{A}} \pi(a|x_t) P(x_{t+1}|x_t, a)$.

We denote by $\xi = (x_0, a_0, r_0, \ldots, x_{T-1}, a_{T-1}, r_{T-1}, x_T)$ a T-step trajectory generated by policy π, and by $R_{0:T-1}(\xi) = \sum_{t=0}^{T-1} \gamma^t r_t$ the return of trajectory ξ. Note that in ξ, $x_0 \sim P_0$, and $\forall t \in \{1, \ldots, T-1\}$, $a_t \sim \pi(\cdot|x_t)$, $x_{t+1} \sim P(\cdot|x_t, a_t)$, and $r_t \sim P_r(\cdot|x_t, a_t)$. These distributions together define P^π_ξ, i.e., the distribution of trajectory ξ. We evaluate a policy π by the expectation of the return of the T-step trajectories it generates, i.e., $\rho^\pi_T = \mathbb{E}_{\xi\sim P^\pi_\xi}\big[R_{0:T-1}(\xi)\big]$. If we set T to be of order $O\big(1/(1-\gamma)\big)$, then ρ^π_T would be a good approximation of the infinite-horizon performance ρ^π_∞. Throughout the paper, we assume that T has been selected such that $\rho^\pi_T \approx \rho^\pi_\infty$, and thus, we refer to $\rho^\pi = \rho^\pi_T$ as the performance of policy π. We further define the value (action-value) function of a policy π at each state x (state-action pair (x,a)), denoted by $V^\pi(x)$ ($Q^\pi(x,a)$), as the expectation of the return of a T-step trajectory generated by starting at state x (state-action pair (x,a)), and then following policy π. Note that $\rho^\pi = \mathbb{E}_{x\sim P_0}\big[V^\pi(x)\big]$.

Note that the contextual bandit setting is a special case of the setting described above, where $T = 1$, and as a result, the context is sampled from P_0 and there is no dynamic P.

2.2. Off-policy Evaluation Problem

The off-policy evaluation (OPE) problem is when we are given a set of T-step trajectories $\mathcal{D} = \{\xi^{(i)}\}_{i=1}^n$ *independently* generated by the *behavior* policy π_b,[1] and the goal is to have a good estimate of the performance of the *evaluation* policy π_e. We consider the estimator $\hat\rho^{\pi_e}$ good if it has low mean square error (MSE), i.e.,

$$\mathrm{MSE}(\rho^{\pi_e}, \hat\rho^{\pi_e}) \stackrel{\triangle}{=} \mathbb{E}_{P^{\pi_b}_\xi}\big[(\rho^{\pi_e} - \hat\rho^{\pi_e})^2\big]. \tag{1}$$

We make the following standard regularity assumption:

Assumption 1 (Absolute Continuity)**.** *For all state-action pairs $(x,a) \in \mathcal{X}\times\mathcal{A}$, if $\pi_b(a|x) = 0$ then $\pi_e(a|x) = 0$.*

In order to quantify the mismatch between the behavior and evaluation policies in generating a trajectory, we define *cumulative importance ratio* as follows. For each T-step trajectory $\xi \in \mathcal{D}$, the *cumulative importance ratio* from time step t_1 to time step t_2, where both t_1 and t_2 are in $\{0, \ldots, T\}$, is $\omega_{t_1,t_2} = 1$ if $t_1 > t_2$, and is $\omega_{t_1,t_2} = \prod_{\tau=t_1}^{t_2} \frac{\pi_e(a_\tau|x_\tau)}{\pi_b(a_\tau|x_\tau)}$, otherwise. In case the behavior policy π_b is *unknown*, we define $\widehat{\omega}_{t_1,t_2}$ exactly as ω_{t_1,t_2}, with π_b replaced by its approximation $\widehat{\pi}_b$. Under Assumption 1, it is easy to see that $\rho^{\pi_e} = \mathbb{E}_{P^{\pi_e}_\xi}[\sum_{t=0}^{T-1}\gamma^t r_t] = \mathbb{E}_{P^{\pi_b}_\xi}[\sum_{t=0}^{T-1}\gamma^t \omega_{0:t} r_t]$. Similar equalities hold for the value and action-value functions of π_e, i.e., $V^{\pi_e}(x) = \mathbb{E}_{P^{\pi_e}_\xi}[\sum_{t=0}^{T-1}\gamma^t r_t|x_0 = x] = \mathbb{E}_{P^{\pi_b}_\xi}[\sum_{t=0}^{T-1}\gamma^t \omega_{0:t} r_t|x_0 = x]$ and $Q^{\pi_e}(x,a) = \mathbb{E}_{P^{\pi_e}_\xi}[\sum_{t=0}^{T-1}\gamma^t r_t|x_0 = x, a_0 = a] = \mathbb{E}_{P^{\pi_b}_\xi}[\sum_{t=0}^{T-1}\gamma^t \omega_{0:t} r_t|x_0 = x, a_0 = a]$.

[1] The results of this paper can be easily extended to the case that the trajectories are generated by multiple behavior policies.

3. Existing Approaches to OPE

The objective of MRDR is to learn the model part of a DR estimator by minimizing its variance. MRDR is a variation of DR with a DM loss function derived from minimizing the DR's variance and is built on top of IS and DM. Therefore, before stating our main results in Section 4, we first provide a brief overview of these popular approaches.

3.1. Direct Estimators

The idea of the direct method (DM) is to first learn a model of the system and then use it to estimate the performance of the evaluation policy π_e. In the case of bandits, this model is the mean reward of each pair of context and arm, and in RL it is either the mean reward $r(x,a)$ and state transition $P(\cdot|x,a)$, or the value (action-value) $V(x)$ ($Q(x,a)$) function. In either case, if we select a good representation for the quantities that need to be learned, and our dataset[2] contains sufficient number of the states and actions relevant to the evaluation of π_e, then the DM estimator has low variance and small bias, and thus, has the potential to outperform the estimators resulted from other approaches.

As mentioned in Section 1, an important issue that has been neglected in the previous work on off-policy evaluation in RL is the loss function used in estimating the model in DM. As pointed out by Dudík et al. (2011), the direct approach has a problem if the model is estimated without the knowledge of the evaluation policy. This is because the distribution of the states and actions that are visited under the evaluation policy should be included in the loss function of the direct approach. In other words, if upon learning a model, we have no information about the evaluation policy (or the distribution of the evaluation policies), then it is not clear how to design the DM's loss function (perhaps a uniform distribution over the states and actions would be the most reasonable). Therefore, in this paper, we assume that the evaluation policy is known prior to learning the model.[3]

In their DM and DR experiments, both Jiang & Li (2016) and Thomas & Brunskill (2016) learn the MDP model, $r(x,a)$ and $P(\cdot|x,a)$, although all the model learning discussion in Thomas & Brunskill (2016) is about the reward of the evaluation policy π_e at every step t along the T-step trajectory, i.e., $r^{\pi_e}(x,t)$. More generally, in off-policy actor-critic algorithms (such as the Reactor algorithm proposed in Gruslys et al. 2017), where one can view the gradient estimation part as an off-policy evaluation problem, the DM state-action value function model is learned by minimizing the Bellman residual in an off-policy setting (Precup et al., 2000b; Munos et al., 2016; Geist & Scherrer, 2014).

[2]Note that we shall use separate datasets for learning the model in DM and evaluating the policy.

[3]Our results can be extended to the case that the distribution of the evaluation policies is known prior to learning the model.

However, neither of these three approaches incorporate the design of the DM loss function into the primary objective, perhaps because they consider the setting in which the model is learned independently.

Our approach to DM in RL: In this paper, we propose to learn Q^{π_e}, the action-value function of the evaluation policy π_e, and then use it to evaluate its performance as

$$\hat{\rho}^{\pi_e}_{\text{DM}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in\mathcal{A}} \pi_e(a|x_0^{(i)})\widehat{Q}^{\pi_e}(x_0^{(i)},a;\beta_n^*).$$

We model Q^{π_e} using a parameterized class of functions with parameter $\beta \in \mathbb{R}^\kappa$ and learn β by solving the following weighted MSE problem

$$\beta^* \in \arg\min_{\beta\in\mathbb{R}^\kappa} \mathbb{E}_{(x,a)\sim\mu_{\pi_e}}\Big[\big(Q^{\pi_e}(x,a) - \widehat{Q}^{\pi_e}(x,a;\beta)\big)^2\Big], \quad (2)$$

where μ_{π_e} is the γ-discounted horizon-T state-action occupancy of π_e, i.e., $\mu_{\pi_e}(x,a) = \frac{1-\gamma}{1-\gamma^T}\sum_{t=0}^{T-1}\gamma^t \mathbb{E}_{P_\xi^{\pi_e}}\big[\mathbf{1}\{x_t = x, a_t = a\}\big]$ and $\mathbf{1}\{\cdot\}$ is the indicator function. Since the actions in the data set $\mathcal{D}$ are generated by π_b, we rewrite the objective function of the optimization problem (2) as

$$\sum_{t=0}^{T-1}\gamma^t \mathbb{E}_{P_\xi^{\pi_b}}\Big[\omega_{0:t}\big(\bar{R}_{t:T-1}(\xi) - \widehat{Q}^{\pi_e}(x_t,a_t;\beta)\big)^2\Big], \quad (3)$$

where $\bar{R}_{t:T-1}(\xi) = \sum_{\tau=t}^{T-1}\gamma^{\tau-t}\omega_{t+1:\tau}r(x_\tau,a_\tau)$ is the Monte Carlo estimate of $Q^{\pi_e}(x_t,a_t)$. The proof of the equivalence of the objective functions (2) and (3) can be found in Appendix A. We obtain β_n^* by solving the sample average approximation (SAA) of (3), i.e.,

$$\beta_n^* \in \arg\min_{\beta\in\mathbb{R}^\kappa} \sum_{t=0}^{T-1}\gamma^t \cdot \frac{1}{n}\sum_{i=1}^{n}\omega_{0:t}^{(i)}\big[\bar{R}_{t:T-1}(\xi^{(i)}) - \widehat{Q}^{\pi_e}(x_t^{(i)},a_t^{(i)};\beta)\big]^2. \quad (4)$$

Since the SAA estimator (4) is unbiased, for large enough n, $\beta_n^* \to \beta^*$ almost surely. We define the bias of our DM estimator at each state-action pair as $\Delta(x,a) = \widehat{Q}^{\pi_e}(x,a;\beta) - Q^{\pi_e}(x,a)$. Note that in contextual bandits with deterministic evaluation policy, the SAA (4) may be written as the weighted least square (WLS) problem

$$\beta_n^* \in \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbf{1}\{\pi_e(x_i)=a_i\}}{\pi_b(a_i|x_i)}\big(r(x_i,a_i) - \widehat{Q}(x_i,a_i;\beta)\big)^2, \quad (5)$$

with weights $1/\pi_b(a_i|x_i)$ for the actions consistent with π_e.

3.2. Importance Sampling Estimators

Another common approach to off-policy evaluation in RL is to use importance sampling (IS) to estimate the performance of the evaluation policy, i.e.,

$$\hat{\rho}^{\pi_e}_{\text{IS}} = \frac{1}{n}\sum_{i=1}^{n}\omega_{0:T-1}^{(i)}\sum_{t=0}^{T-1}\gamma^t r_t^{(i)} = \frac{1}{n}\sum_{i=1}^{n}\omega_{0:T-1}^{(i)}R_{0:T-1}^{(i)}, \quad (6)$$

where $\omega^{(i)}_{0:T-1}$ and $r^{(i)}_t$ are the cumulative importance ratio and reward at step t of trajectory $\xi^{(i)} \in \mathcal{D}$, respectively, and $R^{(i)}_{0:T-1} = R_{0:T-1}(\xi^{(i)})$. Under Assumption 1, the IS estimator (6) is *unbiased*.

A variant of IS that often has less variance, while still unbiased, is *step-wise importance sampling* (step-IS), i.e.,

$$\hat{\rho}^{\pi_e}_{\text{step-IS}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{T-1}\gamma^t \omega^{(i)}_{0:t}\, r^{(i)}_t.$$

If the behavior policy π_b is *unknown*, which is the case in many applications, then either π_b or the importance ratio $\omega = \pi_e/\pi_b$ needs to be estimated, and thus, IS may no longer be unbiased. In this case, the bias of IS and step-IS are $\left|\mathbb{E}_{P^{\pi_e}_\xi}\left[\delta_{0:T-1}(\xi)R_{0:T-1}(\xi)\right]\right|$ and $\left|\sum_{t=0}^{T-1}\gamma^t \mathbb{E}_{P^{\pi_e}_\xi}\left[\delta_{0:t}(\xi)r_t\right]\right|$, respectively, where $\delta_{0:t}(\xi) = 1 - \lambda_{0:t}(\xi) = 1 - \prod_{\tau=0}^{t}\frac{\pi_b(a_\tau|x_\tau)}{\widehat{\pi}_b(a_\tau|x_\tau)}$, with $\widehat{\pi}_b$ being our approximation of π_b (see the proofs in Appendix B). Note that when π_b is *known*, i.e., $\widehat{\pi}_b = \pi_b$, we have $\delta_{0:t} = 0$, and the bias of both IS and step-IS would be zero.

Although the unbiasedness of IS estimators is desirable for certain applications such as safety (Thomas et al., 2015b), their high variance (even in step-wise case), which grows exponentially with horizon T, restricts their applications. This is why another variant of IS, called *weighted importance sampling* (WIS), and particularly its step-wise version, i.e.,

$$\hat{\rho}^{\pi_e}_{\text{WIS}} = \sum_{i=1}^{n}\frac{\omega^{(i)}_{0:T-1}}{\sum_{i=1}^{n}\omega^{(i)}_{0:T-1}}\sum_{t=0}^{T-1}\gamma^t r^{(i)}_t = \sum_{i=1}^{n}\frac{\omega^{(i)}_{0:T-1}R^{(i)}_{0:T-1}}{\sum_{i=1}^{n}\omega^{(i)}_{0:T-1}},$$

$$\hat{\rho}^{\pi_e}_{\text{step-WIS}} = \sum_{i=1}^{n}\sum_{t=0}^{T-1}\gamma^t \frac{\omega^{(i)}_{0:t}\, r^{(i)}_t}{\sum_{i=1}^{n}\omega^{(i)}_{0:t}},$$

is considered more practical, especially where being biased is not crucial. The WIS estimators are biased but consistent and have lower variance than their IS counterparts.

3.3. Doubly Robust Estimators

Doubly robust (DR) estimators that combine DM and IS were first developed for regression (e.g., Cassel et al. 1976), brought to contextual bandits by Dudík et al. (2011), and to RL by Jiang & Li (2016) and Thomas & Brunskill (2016). The DR estimator for RL is defined as

$$\hat{\rho}^{\pi_e}_{\text{DR}}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{T-1}\Big[\gamma^t\omega^{(i)}_{0:t}r^{(i)}_t - \gamma^t\big(\omega^{(i)}_{0:t}\widehat{Q}^{\pi_e}(x^{(i)}_t,a^{(i)}_t;\beta) - \omega^{(i)}_{0:t-1}\widehat{V}^{\pi_e}(x^{(i)}_t;\beta)\big)\Big]. \tag{7}$$

Eq. 7 clearly shows that a DR estimator contains both the cumulative importance ratio ω (IS part) and the model estimates $\widehat{V}^{\pi_e}$ and $\widehat{Q}^{\pi_e}$ (DM part). Note that the IS part of the DR estimator (7) is based on step-wise IS. Thomas & Brunskill (2016) derived a DR estimator whose IS part is based on step-wise WIS. In this paper, we use step-wise IS for the IS part of our DR-based estimators, but our results can be easily extended to other IS estimators.

The bias of a DR estimator is the product of that of DM and IS, and thus, DR is unbiased whenever either IS or DM is unbiased. This is what the term "doubly robust" refers to. The bias of the DR estimator (7) is $\left|\mathbb{E}_{P^{\pi_e}_\xi}\left[\sum_{t=0}^{T-1}\gamma^t\lambda_{0:t-1}(\xi)\delta_t(\xi)\Delta(x_t,a_t)\right]\right|$ (see the proofs in Appendix C), and thus, it would be zero if either $\Delta(x_t,a_t)$ or $\delta_t(\xi)$ is zero. As discussed in Section 3.2, if π_b is known, $\delta_t = 0$ and the DR estimator (7) is unbiased. Throughout this paper, we assume that π_b is known, and thus, DR is unbiased as long as it uses unbiased variants of IS. However, our proposed estimator described in Section 4 can be extended to the case that π_b is unknown.

4. More Robust Doubly Robust Estimators

In this section, we present our class of more robust doubly robust (MRDR) estimators. The main idea of MRDR is to learn the DM parameter of a DR estimator, $\beta \in \mathbb{R}^\kappa$, by minimizing its variance. In other words, MRDR is a variation of DR with a DM loss function derived from minimizing the DR's variance. As mentioned earlier, we assume that the behavior policy π_b is known, and thus, both IS (step-IS) and DR estimators are unbiased. This means that our MRDR estimator is also unbiased, and since it is the result of minimizing the DR's variance, it has the lowest MSE among all the DR estimators.

4.1. MRDR Estimators for Contextual Bandits

Before presenting MRDR for RL, we first formulate it in the contextual bandit setting. We follow the setting of Dudík et al. (2011) and define the DR estimator as

$$\hat{\rho}^{\pi_e}_{\text{DR}}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\frac{\pi_e(a_i|x_i)}{\widehat{\pi}_b(a_i|x_i)}\big(r(x_i,a_i) - \widehat{Q}(x_i,a_i;\beta)\big) + \widehat{V}^{\pi_e}(x_i;\beta), \tag{8}$$

where $\widehat{Q}(x,a;\beta) \approx Q(x,a) = \mathbb{E}_{P_r}[r(x,a)]$ and $\widehat{V}^{\pi_e}(x;\beta) = \mathbb{E}_{a\sim\pi_e}[\widehat{Q}(x,a;\beta)]$. We further define the DM bias $\Delta(x,a) = \widehat{Q}(x,a;\beta) - Q(x,a)$, and error in learning the behavior policy $\delta(x,a) = 1 - \lambda(x,a) = 1 - \frac{\pi_b(a|x)}{\widehat{\pi}_b(a|x)}$. Proposition 1 proves the bias and variance of DR for stochastic evaluation policy π_e. Note that the results stated in Theorems 1 and 2 in Dudík et al. (2011) are only for deterministic π_e.

Proposition 1. *The bias and variance of the DR estimator* (8) *for stochastic* π_e *may be written as*

$$\mathit{Bias}(\hat{\rho}_{DR}^{\pi_e}) = \left|\rho^{\pi_e} - \mathbb{E}_{P_\xi^{\pi_b}}[\hat{\rho}_{DR}^{\pi_e}]\right| = \left|\mathbb{E}_{P_\xi^{\pi_e}}\left[\delta(x,a)\Delta(x,a)\right]\right|,$$

$$\begin{aligned} n\mathbb{V}_{P_\xi^{\pi_b}}(\hat{\rho}_{DR}^{\pi_e}) &= \mathbb{E}_{P_\xi^{\pi_b}}\left[\widehat{\omega}(x,a)^2\big(r(x,a)-Q(x,a)\big)^2\right] \\ &+ \mathbb{V}_{P_0}\left(\mathbb{E}_{\pi_e}\left[Q(x,a)+\delta(x,a)\Delta(x,a)\right]\right) \\ &+ \mathbb{E}_{P_0,\pi_e}\Big[\omega(x,a)\big(1-\delta(x,a)\big)^2\Delta(x,a)^2 \\ &- \mathbb{E}_{\pi_e}\left[\big(1-\delta(x,a)\big)\Delta(x,a)\right]^2\Big]. \end{aligned}$$

Proof. See Appendix D. □

As expected from a DR estimator, Proposition 1 shows that (8) is unbiased if either its DM part is unbiased, $\Delta = 0$, or its IS part is unbiased, $\delta = 0$. When the behavior policy π_b is known, and thus, $\delta(x,a) = 0$ for all x and a, the variance of (8) in Proposition 1 may be written as

$$\begin{aligned} n\mathbb{V}_{P_\xi^{\pi_b}}(\hat{\rho}_{\mathrm{DR}}^{\pi_e}) &= \mathbb{E}_{P_\xi^{\pi_b}}\left[\omega(x,a)^2\big(r(x,a)-Q(x,a)\big)^2\right] \\ &+ \mathbb{V}_{P_0}\left[V^{\pi_e}(x)\right] + \mathbb{E}_{P_0,\pi_e}\Big[\omega(x,a)\Delta(x,a)^2 - \mathbb{E}_{\pi_e}\left[\Delta(x,a)\right]^2\Big]. \end{aligned} \tag{9}$$

Unfortunately, the variance formulation (9) is not suitable for our MRDR method, because its derivative w.r.t. β contains a term $\Delta(x,a) = \widehat{Q}(x,a) - Q(x,a)$ that cannot be estimated from samples as the true expected reward Q is unknown. To address this issue, we derive a new formulation of the variance in Theorem 1, whose derivative does not contain such terms.

Theorem 1. *The variance of the DR estimator* (8) *for stochastic π_e may be written as the following two forms:*

$$\begin{aligned} n\mathbb{V}_{P_\xi^{\pi_b}}(\hat{\rho}_{DR}^{\pi_e}) &= \mathbb{E}_{P_\xi^{\pi_b}}\Big[\omega(x,a)\Big(\mathbb{E}_{\pi_e}\left[\omega(x,a')\widehat{Q}(x,a';\beta)^2\right] \\ &- \widehat{V}^{\pi_e}(x;\beta)^2 - 2r(x,a)\big(\omega(x,a)\widehat{Q}(x,a;\beta) - \widehat{V}^{\pi_e}(x;\beta)\big)\Big) \\ &+ \omega(x,a)^2 r(x,a)^2 - \mathbb{E}_{\pi_e}[r(x,a)]^2\Big] + \mathbb{V}_{P_0}\big(\mathbb{E}_{\pi_e}[r(x,a)]\big), \end{aligned} \tag{10}$$

$$\begin{aligned} n\mathbb{V}_{P_\xi^{\pi_b}}(\hat{\rho}_{DR}^{\pi_e}) &= \overbrace{\mathbb{E}_{P_\xi^{\pi_b}}\left[\omega(x,a)q_\beta(x,a,r)^\top \Omega_{\pi_b}(x) q_\beta(x,a,r)\right]}^{J(\beta)} \\ &+ C, \end{aligned} \tag{11}$$

where $\Omega_{\pi_b}(x) = \mathrm{diag}\big[1/\pi_b(a|x)\big]_{a\in\mathcal{A}} - ee^\top$ is a positive semi-definite matrix (see Proposition 6 in Appendix D fot the proof) with $e = [1,\dots,1]^\top$; $q_\beta(x,a,r) = D_{\pi_e}(x)\bar{Q}(x;\beta) - \mathbb{I}(a)r$ a row vector with $D_{\pi_e}(x) = \mathrm{diag}\big[\pi_e(a|x)\big]_{a\in\mathcal{A}}$, row vector $\bar{Q}(x;\beta) = \big[\widehat{Q}(x,a;\beta)\big]_{a\in\mathcal{A}}$, and the row vector of indicator functions $\mathbb{I}(a) = \big[\mathbf{1}\{a' = a\}\big]_{a'\in\mathcal{A}}$; and finally $C = \mathbb{V}_{P_0}\big(\mathbb{E}_{\pi_e}[r(x,a)]\big) - \mathbb{E}_{P_\xi^{\pi_b}}\big[\mathbb{E}_{\pi_e}[r(x,a)]^2\big] + \mathbb{E}_{P_\xi^{\pi_b}}\Big[\big(1 + \omega(x,a) - \frac{1}{\pi_b^2(a|x)}\big)\omega(x,a)r(x,a)^2\Big]$.

Proof. See Appendix D. □

The significance of the variance formulations of Theorem 1 is **1)** the variance of the DR estimator has no dependence on the unknown term Δ, and thus, its derivative w.r.t. β is computable, **2)** the expectation in (11) is w.r.t. $P_\xi^{\pi_b}$, which makes it possible to replace $J(\beta)$ with its unbiased SAA

$$J_n(\beta) = \frac{1}{n}\sum_{i=1}^n \omega(x_i,a_i) q_\beta(x_i,a_i,r_i)^\top \Omega_{\pi_b}(x_i) q_\beta(x_i,a_i,r_i),$$

where $\mathcal{D} = \{(x_i,a_i,r_i)\}_{i=1}^n$ is the data set generated by the behavior policy π_b, such that the optimizer of $J_n(\beta)$ converges to that of $J(\beta)$ almost surely, and **3)** $J(\beta)$ in (11) is a convex quadratic function of q_β, which in case that $\widehat{Q}(x,a;\beta)$ is smooth, makes it possible to efficiently optimize $J_n(\beta)$ with stochastic gradient descent. Moreover, when $\nabla_\beta \widehat{Q}(x,a;\beta)$ can be explicitly written, we can obtain $\beta_n^* \in \arg\min_\beta J_n(\beta)$, by solving the first order optimality condition $\sum_{i=1}^n \omega(x_i,a_i)$ $q_\beta(x_i,a_i,r_i)^\top \Omega_{\pi_b}(x_i) D_{\pi_e}(x_i) \nabla_\beta \bar{Q}(x_i;\beta) = 0$.

In case the evaluation policy is deterministic, the variance $n\mathbb{V}_{P_\xi^{\pi_b}}(\hat{\rho}_{\mathrm{DR}}^{\pi_e})$ in (10) becomes

$$\overbrace{\mathbb{E}_{P_\xi^{\pi_b}}\Big[\frac{\mathbf{1}\{\pi_e(x)=a\}}{\pi_b(a|x)}\cdot\frac{1-\pi_b(a|x)}{\pi_b(a|x)}\big(r(x,a)-\widehat{Q}(x,a;\beta)\big)^2\Big]}^{J(\beta)} + \mathbb{V}_{P_0}\big(\mathbb{E}_{\pi_e}[r(x,a)]\big).$$

This form of $J(\beta)$ allows us to find the model parameter of MRDR by solving the WLS

$$\begin{aligned} \beta_n^* \in \arg\min_\beta J_n(\beta) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{\pi_e(x_i)=a_i\}\times \\ \frac{1-\pi_b(a_i|x_i)}{\pi_b(a_i|x_i)^2}\big(r(x_i,a_i)-\widehat{Q}(x_i,a_i;\beta)\big)^2. \end{aligned} \tag{12}$$

Comparing this WLS with that in the DM approach in Section 5, we note that MRDR changes the weights from $1/\pi_b$ to $(1-\pi_b)/\pi_b^2$, and this way increases the penalty of the samples whose actions are the same as those suggested by π_e, but have low probability under π_b, and decreases the penalty of the rest of the samples.

4.2. MRDR Estimators for Reinforcement Learning

We now present our MRDR estimator for RL. We begin with the DR estimator for RL given by (7). Similar to the bandits case reported in Section 4.1, we first derive a formula for the variance of the estimator (7), whose derivative can be easily estimated from trajectories generated by the behavior policy. We then use this variance formulation as the objective function to find the MRDR model parameter.

Theorem 2. *The variance of the DR estimator in* (7) *can be written as*

$$n\mathbb{V}_{P_\xi^{\pi_b}}(\hat{\rho}_{DR}^{\pi_e}) = \sum_{t=0}^{T-1} \mathbb{E}_{\mathcal{F}_{0:t-1}}\Big[\gamma^{2t}\omega_{0:t-1}^2 \mathbb{V}_{\mathcal{F}_{t:T-1}}\Big(\omega_t\big(\bar{R}_{t:T-1} - \widehat{Q}^{\pi_e}(x_t,a_t;\beta)\big)\Big) + \widehat{V}^{\pi_e}(x_t;\beta) + C_t\Big] \quad (13)$$
$$+ \mathbb{E}_{\mathcal{F}_{0:t}}\Big[\gamma^{2t-2}\omega_{0:t-1}^2 \mathbb{V}_{\mathcal{F}_{t+1:T-1}}(\bar{R}_{t:T-1} \mid \mathcal{F}_t)\Big],$$

where $\mathcal{F}_{t_1:t_2}$ *is the filtration induced by the sequence* $\{x_{t_1}, a_{t_1}, r_{t_1}, \ldots, x_{t_2}, a_{t_2}, r_{t_2}\} \sim P_\xi^{\pi_b}$, $\bar{R}_{t:T-1} = r(x_t,a_t) + \gamma\sum_{\tau=t+1}^{T-1}\gamma^{\tau-(t+1)}\omega_{t+1:j}r(x_\tau,a_\tau)$, *and* $C_t = E_{\mathcal{F}_{t:T-1}}\Big[\omega_t^2\big(\bar{R}_{t:T-1} - \mathbb{E}_{\mathcal{F}_{t+1:T-1}}[\bar{R}_{t:T-1}]\big)^2 - 2\omega_t^2\bar{R}_{t:T-1}\big(\bar{R}_{t:T-1} - \mathbb{E}_{\mathcal{F}_{t+1:T-1}}[\bar{R}_{t:T-1}]\big)\Big]$ *is a* β*-independent term.*

Proof. The proof is by mathematical induction and is reported in Appendix E. □

As opposed to the DR variance reported in Jiang & Li (2016), ours in (13) has no dependence on the DM bias Δ, which contains the unknown term Q^{π_e}, and plus, all its expectations are over $P_\xi^{\pi_b}$. This allows us to easily compute the MRDR model parameter from the gradient of (13).

Let's define $\beta^* \in \arg\min_{\beta\in\mathbb{R}^\kappa} \mathbb{V}_{P_\xi^{\pi_b}}\big(\hat{\rho}_{\text{DR}}^{\pi_e}(\beta)\big)$ as the minimizer of the DR variance. We may write β^* using the variance formulation of Theorem 2, and after dropping the β-independent terms, as $\beta^* \in \arg\min_{\beta\in\mathbb{R}^\kappa}\sum_{t=0}^{T-1}\mathbb{E}_{\mathcal{F}_{0:t-1}}\Big[\gamma^{2t}\omega_{0:t-1}^2\mathbb{V}_{\mathcal{F}_t}\Big(\omega_t\big(\bar{R}_{t:T-1} - \widehat{Q}^{\pi_e}(x_t,a_t;\beta)\big) + \widehat{V}^{\pi_e}(x_t;\beta)\Big)\Big]$. Similar to the derivation of (11) for bandits, we can show that

$$\beta^* \in \arg\min_{\beta\in\mathbb{R}^\kappa} J(\beta) = \sum_{t=0}^{T-1}\gamma^{2t}\mathbb{E}_{\mathcal{F}_{0:t-1}}\big[\omega_{0:t-1}^2\cdot\omega_t\cdot \quad (14)$$
$$q_\beta(x_t,a_t,\bar{R}_{t:T-1})^\top\Omega_{\pi_b}(x_t)q_\beta(x_t,a_t,\bar{R}_{t:T-1})\big].$$

As shown in Proposition 6, $J(\beta)$ is a quadratic convex function of q_β, which means that if the approximation $\widehat{Q}^{\pi_e}(\cdot,\cdot;\beta)$ is smooth in β, then this problem can be effectively solved by gradient descent. Since the expectation in (14) is w.r.t. $P_\xi^{\pi_b}$, we may use the trajectories in $\mathcal{D}$ (generated by π_b), replace $J(\beta)$ with its unbiased SAA, $J_n(\beta)$, and solve it for β, i.e.,

$$\beta_n^* \in \arg\min_{\beta\in\mathbb{R}^\kappa} J_n(\beta) = \sum_{i=1}^{n}\sum_{t=0}^{T-1}\gamma^{2t}(\omega_{0:t-1}^{(i)})^2\cdot\omega_t^{(i)}\cdot \quad (15)$$
$$q_\beta(x_t^{(i)},a_t^{(i)},\bar{R}_{t:T-1}^{(i)})^\top\Omega_{\pi_b}(x_t^{(i)})q_\beta(x_t^{(i)},a_t^{(i)},\bar{R}_{t:T-1}^{(i)}).$$

Since $J_n(\beta)$ is strongly consistent, $\beta_n^* \to \beta^*$ almost surely. If we can explicitly write $\nabla_\beta\widehat{Q}(x,a;\beta)$, then β_n^* is the solution of equation $0 = \sum_{i=1}^n\sum_{t=0}^{T-1}\gamma^{2t}(\omega_{0:t-1}^{(i)})^2\omega_t^{(i)}$ $q_\beta(x_t^{(i)},a_t^{(i)},\bar{R}_{t:T-1}^{(i)})^\top\Omega_{\pi_b}(x_t^{(i)})D_{\pi_e}(x_t^{(i)})\nabla_\beta\bar{Q}(x_t^{(i)};\beta)$.

In case the evaluation policy is deterministic, we can further simplify $J_n(\beta)$ and derive the model parameter for MRDR by solving the following WLS problem:

$$J_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{T-1}\gamma^{2t}(\omega_{0:t-1}^{(i)})^2\omega_t^{(i)}\mathbf{1}\{\pi_e(x_t^{(i)}) = a_t^{(i)}\}$$
$$\frac{1-\pi_b(a_t^{(i)}|x_t^{(i)})}{\pi_b(a_t^{(i)}|x_t^{(i)})^2}\big(\bar{R}_{t:T-1}^{(i)} - \widehat{Q}^{\pi_e}(x_t^{(i)},a_t^{(i)};\beta)\big)^2. \quad (16)$$

The intuition behind the weights in WLS (16) is **1)** to adjust the difference between the occupancy measures of the behavior and evaluation policies, and **2)** to increase the penalty of the policy discrepancy term $\mathbf{1}\{\pi_e(x_t) = a_t\}$.

4.3. Other Properties of the MRDR Estimators

Strong Consistency Similar to the analysis in Thomas & Brunskill (2016) for weighted DR, we prove (in Appendix F) that the MRDR estimators are strongly consistent, i.e., $\lim_{n\to\infty}\hat{\rho}_{\text{MRDR},n}^{\pi_e}(\beta_n^*) = \rho^{\pi_e}$ almost surely. This implies that MRDR is a *well-posed* OPE estimator.

Asymptotic Optimality The MRDR estimator, by construction, has the lowest variance among the DR estimators of the form (7). On the other hand, the semi-parametric theory in multivariate regression (Robins et al., 1994) states that without extra assumption on the data distribution, the class of *unbiased*, *consistent*, and *asymptotically normal* OPE estimators is asymptotically equivalent to the DR estimators in (7). Utilizing this result, we can show that the MRDR estimators are asymptotically optimal (i.e., have minimum variance) in this class of estimators.

MRDR Extensions Similar to Thomas & Brunskill (2016), we can derive the *weighted* MRDR estimator by replacing the IS part of the MRDR estimator in (7) with (per-step) weighted importance sampling. This introduces bias, but potentially reduces its variance, and thus, its MSE.

Throughout the paper, we assumed that the data has been generated by a single behavior policy. We can extend our MRDR results to the case that there are more than one behavior policy by replacing the IS part of our estimator with *fused importance sampling* (Peshkin & Shelton, 2002).

5. Experiments

In this section, we demonstrate the effectiveness of the proposed MRDR estimation by comparing it with other state-of-the art methods from Section 3 on both contextual bandit and RL benchmark problems.

5.1. Contextual Bandit

Using the 9 benchmark experiments described in Dudík et al. (2011), we evaluate the OPE algorithms using the standard classification data-set from the UCI repository. Here we

follow the same procedure of transforming a classification data-set into a contextual bandit dataset. For the sake of brevity, detailed descriptions of the experimental setup will be deferred to the appendix.

Given a deterministic policy π, which is a logistic regression model trained by the classification data set, we discuss three methods of transforming it into stochastic policies. The first one, which is known as *friendly softening*, constructs a stochastic policy with the following smoothing procedure: Given two constants α and β, and a uniform (continuous) random variable $u \in [-0.5, 0.5]$. For each $a \in \{1, \ldots, l\}$, whenever $\pi(x) = a$, the stochastic policy $\pi_{\alpha,\beta}(x)$ returns a with probability $\alpha + \beta \times u$, and it returns k, which is a realization of the uniform (discrete) random variable in $\{1, \ldots, l\} \setminus \{a\}$ with probability $\frac{1-(\alpha+\beta\times u)}{l-1}$. The second one, which is known as *adversarial softening*, constructs a stochastic policy $\pi_{\alpha,\beta}(x)$ from policy π in a similar fashion. Whenever $\pi(x) = a$, $\pi_{\alpha,\beta}(x)$ returns $k \neq a$ with probability $\alpha + \beta \times u$, and it returns $\tilde{k}$, which is a realization of the uniform (discrete) random variable in $\{1, \ldots, l\}$ with probability $\frac{1-(\alpha+\beta\times u)}{l}$. The third one, which is the *neutral policy*, is a uniformly random policy. We will use these methods to construct behavior and evaluation policies. Table 1 summarizes their specifications.

Here we compare the MRDR method with the direct method (DM), the importance sampling (IS) method and two doubly robust (DR) estimators. The model parameter of the DM estimator is obtained by solving the SAA of the following problem: $\beta_{\text{DM}} \in \arg\min_{\beta\in\mathbb{R}^\kappa} \mathbb{E}_{(x,a)\sim P_\xi^{\pi_b}}[(Q^{\pi_e}(x,a) - \widehat{Q}^{\pi_e}(x,a;\beta))^2]$, which means all samples are weighted according to data, without consideration of the visiting distribution induced by the evaluation policy. The model parameters of the DR estimator is optimized based on the DM methodologies described in (2). Besides the standard DR estimator we also include another alternative that is known as DR0, which heuristically uses the model parameter from the vanilla DM method (which is called DM0 and assigns uniform weights over samples).

Below are results over the five behavior policies and five algorithms on the benchmark datasets. Due to the page limit, only the results of Vehicle, SatImage, PenDigits and Letter are included in the main paper, see Appendix G for the remaining results. We evaluate the accuracy of the estimation via root mean squares error (RMSE): $\sqrt{\sum_{j=1}^N (\hat{\rho}_j^{\pi_e} - \rho^{\pi_e})^2/N}$, where $\widehat{\rho}_j^{\pi_e}$ is the estimated value from the j-th dataset. Furthermore, we perform a 95% significance test *only* on MRDR and DR, with bold numbers indicating the corresponding method outperforms its counterpart significantly.

In the contextual bandit experiments, it is clear that in most cases the proposed MRDR estimator is superior to all alternative estimators (statistical) significantly. Similar to the results reported in Dudík et al. (2011), the DM method incurs much higher MSE than other methods in all of the experiments. This is potentially due to the issue of high bias in model estimation when the sample-size is small. In general the estimation error is increasing across rows from top to bottom. This is expected due to the increasing difficulties in the OPE tasks that is accounted by the increasing mis-matches between behavior and evaluation policies. Although there are no theoretical justifications, in most cases the performance of DR estimators (with the DM method described in Section 3.1) is better than that of DR0. This also illustrates the benefits of optimizing the model parameter based on the knowledge of trajectory distribution $P_\xi^{\pi_e}$, which is generated by the evaluation policy.

Table 1. Behavior and Evaluation Policies

	Policy	α	β
Evaluation Policy		0.9	0
Behavior Policies	Friendly I	0.7	0.2
	Friendly II	0.5	0.2
	Neutral	-	-
	Adversary I	0.3	0.2
	Adversary II	0.5	0.2

Table 2. Vehicle

Behavior Policy	DM	IS	DR	MRDR	DR0
Friendly I	0.3273	0.0347	0.0217	**0.0202**	0.0224
Friendly II	0.3499	0.0517	0.0331	**0.0318**	0.0356
Neutral	0.4384	0.087	0.0604	**0.0549**	0.0722
Adversary I	0.405	0.0937	0.0616	**0.0516**	0.0769
Adversary II	0.405	0.1131	0.0712	**0.0602**	0.0952

Table 3. SatImage

Behavior Policy	DM	IS	DR	MRDR	DR0
Friendly I	0.2884	0.0128	0.0071	**0.0063**	0.0073
Friendly II	0.3328	0.0191	0.0107	**0.0087**	0.0119
Neutral	0.3848	0.0413	0.0246	**0.0186**	0.0335
Adversary I	0.3963	0.0459	0.027	**0.0195**	0.0383
Adversary II	0.4093	0.0591	0.0364	**0.0262**	0.0521

Table 4. PenDigits

Behavior Policy	DM	IS	DR	MRDR	DR0
Friendly I	0.4014	0.0103	0.0056	**0.0037**	0.0059
Friendly II	0.4628	0.0159	0.0092	**0.0056**	0.0194
Neutral	0.564	0.0450	0.0314	**0.0138**	0.0412
Adversary I	0.5861	0.0503	0.0366	**0.0172**	0.0472
Adversary II	0.5641	0.0646	0.0444	**0.0188**	0.0611

Table 5. Letter

Behavior Policy	DM	IS	DR	MRDR	DR0
Friendly I	0.392	0.0074	0.0056	**0.0044**	0.0057
Friendly II	0.4146	0.0102	0.0077	**0.0054**	0.0083
Neutral	0.4713	0.0467	0.0363	**0.0315**	0.0456
Adversary I	0.46	0.0587	0.0455	**0.0385**	0.0575
Adversary II	0.4728	0.0714	0.055	0.0481	0.0703

5.2. Reinforcement Learning

In this section we present the experimental results of OPE in reinforcement learning. We first test the OPE algorithms on the standard domains ModelWin, ModelFail, and 4×4 Maze, with behavior and evaluation policies used in Thomas & Brunskill (2016). The schematic diagram of the domains is shown in Figure 1. To demonstrate the scalability of the proposed OPE methods, we also test the OPE algorithms on the following two domains with continuous state space: Mountain Car and Cart Pole. To construct the stochastic behavior and evaluation policies, we first compute the optimal policy using standard RL algorithms such as SARSA and Q-learning. Then these policies are constructed by applying friendly softening to the optimal policy with specific values of (α, β). For both domains, the evaluation policy is constructed using $(\alpha, \beta) = (0.9, 0.05)$, and the behavior policy is constructed analogously using $(\alpha, \beta) = (0.8, 0.05)$. Detailed explanations of the experimental setups can be found in Appendix G. In the following experiments we set the discounting factor to be $\gamma = 1$.

For both ModelFail and ModelWin domains, the number of training trajectories is set to 64, for Maze, Mountain Car, and Cart Pole domains this number is set to 1024. The number of trajectories for sampling-based part of estimators varies from 32 to 512 for the ModelWin, ModelFail, and Cart Pole domains, and varies from 128 to 2048 for the Maze and Mountain Car domains.

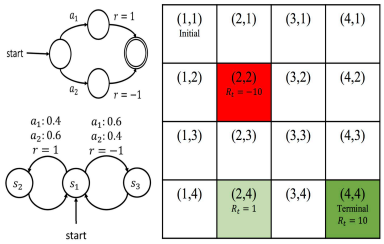

Figure 1. Environments from Thomas & Brunskill (2016). Top left: ModelFail; Bottom left: ModelWin; Right: Maze

In all of the above experiments, we compare results of MRDR with DM, IS, DR, and DR0 estimations by their corresponding MSE values. Similarly, the bold numbers represent cases when the performance of the MRDR estimator is statistically significantly better than that of the DR estimator. Similar to the contextual bandit setting, except for the ModelWin domain that is known to be in favor of the DM estimator (Thomas & Brunskill, 2016), in most cases MRDR estimator has significantly lower MSE than the other methods. Furthermore, when we increase the number of the evaluation trajectories, the accuracy of all the estimators in all the experiments is improved. Similar to the contextual bandit setting, significant performance improvement can be observed when one switches from DR0 to DR in the RL experiments.

Table 6. ModelFail

Sample Size	DM	IS	DR	MRDR	DR0
32	0.07152	1.37601	0.18461	**0.1698**	1.16084
64	0.07152	1.07213	0.1314	**0.11405**	0.9046
128	0.07152	0.752	0.09901	**0.08188**	0.63571
256	0.07152	0.55955	0.06565	**0.05527**	0.47211
512	0.07152	0.39533	0.04756	**0.03819**	0.33391

Table 7. Modelwin

Sample Size	DM	IS	DR	MRDR	DR0
32	0.06182	0.78452	1.55244	**1.46778**	1.51858
64	0.06182	1.03207	1.13856	**0.98433**	1.40758
128	0.06182	0.90166	1.4195	**1.27891**	1.52634
256	0.06182	0.78507	1.03575	**0.79849**	1.10332
512	0.06182	0.55647	0.89655	0.66791	0.97128

Table 8. 4×4 Maze

Sample Size	DM	IS	DR	MRDR	DR0
128	1.77598	6.68579	0.70465	**0.57042**	0.70969
256	1.77598	3.50346	0.69886	**0.58871**	0.70211
512	1.77598	2.64257	0.60124	0.58879	0.60338
1024	1.77598	1.45434	0.5201	**0.4666**	0.52148
2048	1.77598	0.89668	0.3932	**0.31274**	0.39425

Table 9. Mountain Car

Sample Size	DM	IS	DR	MRDR	DR0
128	17.80368	23.11318	16.14661	**14.96227**	19.46953
256	14.62359	14.82684	13.89212	**12.48327**	22.80573
512	13.22012	8.26484	8.01421	7.89474	7.96849
1024	10.24318	3.26843	3.03239	3.1359	9.16269
2048	10.91577	2.50591	2.75933	**2.17138**	8.25527

Table 10. Cart Pole

Sample Size	DM	IS	DR	MRDR	DR0
32	86.81935	70.58151	12.13028	16.45905	10.84913
64	87.00547	75.86198	14.82026	14.16847	15.69192
128	84.40824	77.38233	18.55218	**15.38549**	19.15905
256	83.31824	64.75034	9.96921	8.36612	9.36373
512	84.09259	65.72996	6.88534	**4.60712**	6.9962

6. Conclusions

In this paper, we proposed the class of more-robust doubly-robust (MRDR) estimators for off-policy evaluation in RL. In particular, we proposed a principled method to calculate the model in DR estimator, which aims at minimizing its variance. Furthermore, we showed that our estimator is consistent and asymptotically optimal in the class of unbiased, consistent and asymptotically normal estimators. Finally, we demonstrated the effectiveness of our MRDR estimator in bandits and RL benchmark problems.

Future work includes extending the MRDR estimator to the cases **1)** when there are multiple behavior policies, **2)** when the action set has a combinatorial structure, e.g., actions are in the form of slates (Swaminathan et al., 2017), and **3)** when the behavior policy is unknown.

References

Bang, H. and Robins, J. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–972, 2005.

Bottou, L., Peters, J., Quiñonero-Candela, J., Charles, D., Chickering, D. Max, Portugaly, E., Ray, D., Simard, P., and Snelson, E. Counterfactual reasoning and learning systems: The example of computational advertising. *JMLR*, 14:3207–3260–620, 2013.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. OpenAI Gym. arXiv:1606.01540, 2016.

Cao, W., Tsiatis, A., and Davidian, M. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96: 723–734, 2009.

Cassel, C., Särndal, C., and Wretman, J. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63: 615–620, 1976.

Dudík, M., Langford, J., and Li, L. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 1097–1104, 2011.

Geist, M. and Scherrer, B. Off-policy learning with eligibility traces: A survey. *The Journal of Machine Learning Research*, 15(1):289–333, 2014.

Gruslys, A., Azar, M., Bellemare, M., and Munos, R. The reactor: A sample-efficient actor-critic architecture. *arXiv preprint arXiv:1704.04651*, 2017.

Hanna, J., Stone, P., and Niekum, S. High confidence off-policy evaluation with models. *arXiv preprint arXiv:1606.06126*, 2016.

Hasselt, H. Van, Guez, A., and Silver, D. Deep reinforcement learning with double Q-learning. In *AAAI*, volume 16, pp. 2094–2100, 2016.

Hirano, K., Imbens, G., and Ridder, W. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Jiang, N. and Li, L. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 652–661, 2016.

Li, L., an d J. Langford, W. Chu, and Wang, X. Unbiased offline evaluation of contextual bandit-based news article recommendation algorithms. In *Proceedings of the 4th International Conference on Web Search and Data Mining*, pp. 297–306, 2011.

Li, L., Munos, R., and Szepesvàri, Cs. Toward minimax off-policy value estimation. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pp. 608–616, 2015.

Louizos, C., Shalit, U., Mooij, J., Sontag, D., Zemel, R., and Welling, M. Causal effect inference with deep latent-variable models. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 3076–3085, 2017.

Mahmood, A., van Hasselt, H., and Sutton, R. Weighted importance sampling for off-policy learning with linear function approximation. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, 2014.

Mandel, T., Liu, Y., Levine, S., Brunskill, E., and Popovic, Z. Off-policy evaluation across representations with applications to educational games. In *Proceedings of the 13th International Conference on Autonomous Agents and Multi-agent Systems*, pp. 1077–1084, 2014.

Mandel, T., Liu, Y., Brunskill, E., and Popovic, Z. Offline evaluation of online reinforcement learning algorithms. In *Proceedings of the 30th Conference on Artificial Intelligence*, 2016.

Munos, R., Stepleton, T., Harutyunyan, A., and Bellemare, M. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 1054–1062, 2016.

Murphy, S., van der Laan, M., and Robins, J. Marginal mean models for dynamic regimes. *Journal of American Statistical Association*, 96(456):1410–1423, 2001.

Paduraru, C. *Off-policy Evaluation in Markov Decision Processes*. PhD thesis, McGill University, 2013.

Peshkin, L. and Shelton, C. Learning from scarce experience. In *Proceedings of the 19th International Conference on Machine Learning*, pp. 498–505, 2002.

Precup, D., Sutton, R., and Singh, S. Eligibility traces for off-policy policy evaluation. In *Proceedings of the 17th International Conference on Machine Learning*, pp. 759–766, 2000a.

Precup, D., Sutton, R., and Singh, S. Eligibility traces for off-policy policy evaluation. In *ICML*, pp. 759–766. Citeseer, 2000b.

Precup, D., Sutton, R., and Dasgupta, S. Off-policy temporal difference learning with function approximation.

In *Proceedings of the 18th International Conference on Machine Learning*, pp. 417–424, 2001.

Robins, J. and Rotnitzky, A. Semi-parametric efficiency in multivariate regression models with missing data. *Journal of American Statistical Association*, 90:122–129, 1995.

Robins, J., Rotnitzky, A., and Zhao, L. Estimation of regression coefficients when some regressors are not always observed. *Journal of American Statistical Association*, 89(427):846–866, 1994.

Shalit, U., Johansson, F., and Sontag, D. Estimating individual treatment effect: Generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3076–3085, 2017.

Sutton, R. and Barto, A. *Reinforcement learning: An introduction*. MIT press Cambridge, 1998.

Swaminathan, A., Krishnamurthy, A., Agarwal, A., Dudík, M., Langford, J., Jose, D., and Zitouni, I. Off-policy evaluation for slate recommendation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 3635–3645, 2017.

Theocharous, G., Thomas, P., and Ghavamzadeh, M. Personalized ad recommendation systems for life-time value optimization with guarantees. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, pp. 1806–1812, 2015.

Thomas, P. and Brunskill, E. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 2139–2148, 2016.

Thomas, P., Theocharous, G., and Ghavamzadeh, M. High confidence off-policy evaluation. In *Proceedings of the 29th Conference on Artificial Intelligence*, 2015a.

Thomas, P., Theocharous, G., and Ghavamzadeh, M. High confidence policy improvement. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 2380–2388, 2015b.