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.
More Robust Doubly Robust Off-policy Evaluation

(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 (Dudik 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 defineMDPs 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 $\langle \mathcal{X}, \mathcal{A}, P, \rho, P_0, \gamma \rangle$, where $\mathcal{X}$ and $\mathcal{A}$ are the state and action spaces, $P(x, a)$ is the distribution of the bounded random variable $r(x, a) \in [0, R_{\text{max}}]$ of the immediate reward of taking action $a$ in state $x$, $P_0$ 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 $P \pi$ the state transition of the Markov chain induced by policy $\pi$, 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 $\pi$, and by $R_{0:T-1}(\xi) = \sum_{t=0}^{T-1} \gamma^t r_t$ the return of trajectory $\xi$. Note that in $\xi$, $x_0 \sim P_0$, and $\forall t \in \{1, \ldots, T - 1\}$, $a_t \sim \pi(x_t|x_{t-1}, a_{t-1}, r_{t-1}, x_{T-1}, x_T)$, and $r_t \sim P_r(x_t|x_{t-1}, a_{t-1})$. These distributions together define $P_\xi$, i.e., the distribution of trajectory $\xi$. We evaluate a policy $\pi$ by the expectation of the return of the T-step trajectories it generates, i.e., $\rho_\pi^T = \mathbb{E}_{\xi \sim P_\xi}[R_{0:T-1}(\xi)]$. If we set $T$ to be of order $O(1/(1 - \gamma))$, then $\rho_\pi^T$ would be a good approximation of the infinite-horizon performance $\rho_\infty^\pi$. Throughout the paper, we assume that $T$ has been selected such that $\rho_\pi^T \approx \rho_\infty^\pi$, and thus, we refer to $\rho_\pi^T$ as the performance of policy $\pi$.

We further define the value (action-value) function of a policy $\pi$ 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 $\pi$. Note that $\rho_\pi^T = \mathbb{E}_{x \sim P_0}[V_\pi(x)]$.

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 $D = \{(\xi_1^{(1)}, \ldots, \xi_n^{(1)})\}_{i=1}^n$ independently generated by the behavior policy $\pi_b$.

We define the cumulative importance ratio as follows. For each $T$-step trajectory $\xi \in D$, 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 $\omega_{t_1,t_2} = \prod_{t=t_1}^{t_2} \frac{P_\pi(x_{t+1}|x_t, a_t)}{P_\pi_b(x_{t+1}|x_t, a_t)}$, otherwise. In case the behavior policy $\pi_b$ is unknown, we define $\hat{\omega}_{t_1,t_2}$ exactly as $\omega_{t_1,t_2}$, with $\pi_b$ replaced by its approximation $\hat{\pi}_b$.

Under Assumption 1, it is easy to see that $\rho_\pi^T = \mathbb{E}_{\hat{\pi}}[\sum_{t=0}^{T-1} \gamma^t r_t] = \mathbb{E}_{\hat{\pi}}[\sum_{t=0}^{T-1} \gamma^t \omega_{0:t} r_t]$. Similar equalities hold for the value and action-value functions of $\pi$, i.e., $V_\pi(x) = \mathbb{E}_{\hat{\pi}}[\sum_{t=0}^{T-1} \gamma^t \omega_{0:t} r_t | x_0 = x] = \mathbb{E}_{\hat{\pi}}[\sum_{t=0}^{T-1} \gamma^t \omega_{0:t} r_t | x_0 = x]$ and $Q_\pi(x, a) = \mathbb{E}_{\hat{\pi}}[\sum_{t=0}^{T-1} \gamma^t \omega_{0:t} r_t | x_0 = x, a_0 = a] = \mathbb{E}_{\hat{\pi}}[\sum_{t=0}^{T-1} \gamma^t \omega_{0:t} r_t | x_0 = x, a_0 = a]$.

1The results of this paper can be easily extended to the case that the trajectories are generated by multiple behavior policies.
More Robust Doubly Robust Off-policy Evaluation

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 \( \pi_e \). 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 contains sufficient number of the states and actions relevant to the evaluation of \( \pi_e \), 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.\(^1\)

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 \( \pi_e \) 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).

\(^1\)Note that we shall use separate datasets for learning the model in DM and evaluating the policy.

\(^2\)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 \( Q^\pi_e \), the action-value function of the evaluation policy \( \pi_e \), and then use it to evaluate its performance as

\[
\hat{\rho}^\pi_e = \frac{1}{n} \sum_{i=1}^{n} \sum_{x \in X, a \in A} \pi_e(a|x^i_0)Q^\pi_e(x^i_0, a; \beta^*_n).
\]

We model \( Q^\pi_e \) using a parameterized class of functions with parameter \( \beta \in \mathbb{R}^n \) and learn \( \beta \) by solving the following weighted MSE problem

\[
\beta^*_n \in \arg \min_{\beta \in \mathbb{R}^n} \mathbb{E}_{\pi_e \sim \mu_{\pi_e}} \left[ \left( Q^e(x, a) - \hat{Q}^e(x, a; \beta) \right)^2 \right], \tag{2}
\]

where \( \mu_{\pi_e} \) is the \( \gamma \)-discounted horizon-\( T \) state-action occupancy of \( \pi_e \), i.e., \( \mu_{\pi_e}(x, a) = \frac{1}{n} \sum_{t=0}^{T-1} \gamma^t \mathbb{E}_{\pi_e} \left[ 1 \{ x_t = x, a_t = a \} \right] \) and \( 1 \{ \cdot \} \) is the indicator function. Since the actions in the data set \( D \) are generated by \( \pi_0 \), we rewrite the objective function of the optimization problem (2) as

\[
\beta^*_n \in \arg \min_{\beta \in \mathbb{R}^n} \mathbb{E}_{\pi_0 \sim \mu_{\pi_0}} \left[ \omega_0 T \left( \hat{R}_{\pi_0:1:T-1}(\xi) - \hat{Q}^e(x_t, a_t; \beta) \right)^2 \right], \tag{3}
\]

where \( \hat{R}_{\pi_0:1:T-1}(\xi) = \sum_{t=0}^{T-1} \gamma^t \omega_{t+1} r(x_t, a_t) \) is the Monte Carlo estimate of \( Q^e(x_t, a_t) \). The proof of the equivalence of the objective functions (2) and (3) can be found in Appendix A. We obtain \( \beta^*_n \) by solving the sample average approximation (SAA) of (3), i.e.,

\[
\beta^*_n \in \arg \min_{\beta \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \sum_{t=0}^{T-1} \gamma^t \mathbb{E}_{\pi_0 \sim \mu_{\pi_0}} \left[ \hat{R}_{\pi_0:1:T-1}(\xi^i) - \hat{Q}^e(x_t^i, a_t^i; \beta) \right]^2, \tag{4}
\]

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

\[
\beta^*_n \in \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\pi_0(a_i|x_i)} \left( r(x_i, a_i) - \hat{Q}(x_i, a_i; \beta) \right)^2, \tag{5}
\]

with weights \( 1/\pi_0(a_i|x_i) \) for the actions consistent with \( \pi_e \).

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.,

\[
\hat{\rho}_{IS} = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=0}^{T-1} \gamma^t r_t^i = \frac{1}{n} \sum_{i=1}^{n} \omega_0^{i:T-1} R_{0:T-1}^{i}, \tag{6}
\]
where \( \omega_0^{(i)} \) and \( r_t^{(i)} \) are the cumulative importance ratio and reward at step \( t \) of trajectory \( \xi^{(i)} \in \mathcal{D} \), respectively, and \( p_0^{(i)} = R_0^{(i)} \). 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.,

\[
\hat{\pi}_{\text{step-IS}} = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=0}^{T-1} \gamma^t \omega_0^{(i)} r_t^{(i)}.
\]

If the behavior policy \( \pi_b \) is unknown, which is the case in many applications, then either \( \pi_b \) 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

\[
E_{\pi_e} [b_{\text{IS}}(\xi)R_0^{(i)}(\xi)] = \sum_{t=0}^{T-1} \gamma^t E_{\pi_e} [b_{\text{IS}}(\xi) r_t] = \sum_{t=0}^{T-1} \gamma^t E_{\pi_e} [b_{\text{IS}}(\xi) r_t],
\]

respectively, where \( b_{\text{IS}}(\xi) = 1 - \lambda_0(\xi) = 1 - \prod_{t=0}^{T-1} \frac{\pi_b(x_t, a_t)}{\pi_e(x_t, a_t)} \), with \( \pi_b \) being our approximation of \( \pi_b \) (see the proofs in Appendix B). Note that when \( \pi_e \) is known, i.e., \( \pi_b = \pi_0 \), we have \( \delta_t = 0 \), 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.,

\[
\hat{\pi}_{\text{WIS}} = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=0}^{T-1} \gamma^t \omega_0^{(i)} r_t^{(i)},
\]

\[
\hat{\pi}_{\text{step-WIS}} = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=0}^{T-1} \gamma^t \omega_0^{(i)} r_t^{(i)},
\]

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

\[
\hat{\pi}_{\text{DR}}(\beta) = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=0}^{T-1} \left[ \gamma^t \omega_0^{(i)} Q^{\pi_e}(x_t^{(i)}, a_t^{(i)}; \beta) - \gamma^t \omega_0^{(i)} \hat{V}^{\pi_e}(x_t^{(i)}; \beta) \right].
\]

Eq. (7) clearly shows that a DR estimator contains both the cumulative importance ratio \( \omega \) (IS part) and the model estimates \( \hat{V}^{\pi_e} \) and \( Q^{\pi_e} \) (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 \( E_{\pi_e} [\sum_{t=0}^{T-1} \gamma^t \lambda_0(\xi) \delta_t(\xi) \Delta(x_t, a_t)] \) (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 \( \pi_b \) is known, \( \delta_t = 0 \) and the DR estimator (7) is unbiased. Throughout this paper, we assume that \( \pi_e \) 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 \( \pi_b \) 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, \( \beta \in \mathbb{R}^n \), 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 \( \pi_b \) 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

\[
\hat{\pi}_{\text{DR}}(\beta) = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=0}^{T-1} \left[ \gamma^t \omega_0^{(i)} Q^{\pi_e}(x_t^{(i)}, a_t^{(i)}; \beta) - \gamma^t \omega_0^{(i)} \hat{V}^{\pi_e}(x_t^{(i)}; \beta) \right] + \hat{V}^{\pi_e}(x_t^{(i)}; \beta),
\]

where \( \hat{Q}(x, a; \beta) \approx Q(x, a) = E_{\pi_b} [r(x, a)] \) and \( \hat{V}^{\pi_e}(x; \beta) = E_{\pi_b[s]} [\hat{Q}(x, a; \beta)] \). We further define the DM bias \( \Delta(x, a) = \hat{Q}(x, a; \beta) - Q(x, a) \), and error in learning the behavior policy \( \delta(x, a) = \lambda(x, a) - \pi_b(a|x) \). Proposition 1 proves the bias and variance of DR for stochastic evaluation policy \( \pi_e \).

**Proposition 1.** The bias and variance of the DR estimator (8) for stochastic \( \pi_e \) may be written as
The variance of the variance formulations of Theorem 1 is 1) the variance of the DR estimator has no dependence on the unknown term \( \Delta \), and thus, its derivative w.r.t. \( \beta \) is computable, 2) the expectation in (11) is w.r.t. \( P^\pi_{\beta} \), which makes it possible to replace \( J(\beta) \) with its unbiased SAA

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

where \( \mathcal{D} = \{(x_i, a_i, r_i)\}_{i=1}^{n} \) is the data set generated by the behavior policy \( \pi_b \), 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 \( q_\beta \), which in case that \( \tilde{Q}(x, a; \beta) \) is smooth, makes it possible to efficiently optimize \( J_n(\beta) \) with stochastic gradient descent. Moreover, when \( \nabla_{\beta} \tilde{Q}(x, a; \beta) \) can be explicitly written, we can obtain \( \beta^*_n \in \arg\min_{\beta} J_n(\beta) \), by solving the first order optimality condition \( \sum_{i=1}^{n} \omega(x_i, a_i)q_\beta(x_i, a_i, r_i) \cdot \Omega_{\pi_b}(x_i)D_{\pi_b}(x_i)\nabla_{\beta} \tilde{Q}(x_i; \beta) = 0 \).

In case the evaluation policy is deterministic, the variance \( nV_{\pi e}^2(\hat{\rho}_{\text{DR}}^\pi) \) in (10) becomes

\[
\begin{align*}
\mathbb{E}_{\pi e}^n \left[ \frac{1}{n} \sum_{i=1}^{n} \left( \frac{1 - \pi_b(a|x_i)}{\pi_b(a|x_i)} (r(x_i, a_i) - \tilde{Q}(x_i, a_i; \beta))^2 \right) \right] \\
+ \mathbb{E}_{\pi e}^n \left[ \frac{1}{n} \sum_{i=1}^{n} \left( \frac{1 - \pi_b(a|x_i)}{\pi_b(a|x_i)} (r(x_i, a_i) - \tilde{Q}(x_i, a_i; \beta))^2 \right) \right]
\end{align*}
\]

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

\[
\beta^*_n \in \arg\min_{\beta} J_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{1 - \pi_b(a|x_i)}{\pi_b(a|x_i)} (r(x_i, a_i) - \tilde{Q}(x_i, a_i; \beta))^2 \right] = 0.
\]

Comparing this WLS with that in the DM approach in Section 5, we note that MRDR changes the weights from \( 1/\pi_b \) to \( (1 - \pi_b)/\pi^2_b \), and this way increases the penalty of the samples whose actions are the same as those suggested by \( \pi_e \), but have low probability under \( \pi_b \), 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
As opposed to the DR variance reported in Jiang & Li (2016), ours in (13) has no dependence on the DM bias \( \text{Q} \).

**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 \( \text{Q} \). This allows us to easily compute the MRDR model parameter from the gradient of (13).

Let’s define \( \beta^* \in \arg\min_{\beta \in \mathbb{R}^n} \mathbb{E}_{\beta} \left( \hat{\beta}_{\text{MRDR}} (\beta) \right) \) as the minimizer of the DR variance. We may write \( \beta^* \) using the variance formulation of Theorem 2, and after dropping the \( \beta \)-independent terms, as \( \beta^* \in \arg\min_{\beta \in \mathbb{R}^n} \sum_{t=0}^{T-1} \mathbb{E}_{\beta} \left[ \gamma^{2t} \mathbb{E}_{\beta} \left( \omega_i \beta \right) \right] \). Similar to the derivation of (11) for bandits, we can show that

\[
\beta^* \in \arg\min_{\beta \in \mathbb{R}^n} J(\beta) = \sum_{t=0}^{T-1} \gamma^{2t} \mathbb{E}_{\beta} \left( \omega_i \beta \right).
\]

As shown in Proposition 6, \( J(\beta) \) is a quadratic convex function of \( \beta \), which means that if the approximation \( \hat{\beta}_{\text{MRDR}} (\cdot, \beta) \) is smooth in \( \beta \), then this problem can be effectively solved by gradient descent. Since the expectation in (14) is w.r.t. \( \mathbb{E}_{\beta} \), we may use the trajectories in \( \mathcal{D} \) (generated by \( \pi_b \)), replace \( J(\beta) \) with its unbiased SAA, \( J_n(\beta) \), and solve it for \( \beta^* \), i.e.,

\[
\beta^*_n \in \arg\min_{\beta \in \mathbb{R}^n} J_n(\beta) = \sum_{t=0}^{T-1} \gamma^{2t} \mathbb{E}_{\beta} \left( \omega_i \hat{\beta}_{\text{MRDR}} (\cdot, \beta) \right) \Omega_{\pi_b}(x_i) q_\beta(x_i, a_i, \hat{\beta}_{\text{MRDR}}(x_i) ; \beta).
\]

Since \( J_n(\beta) \) is strongly consistent, \( \beta^*_n \rightarrow \beta^* \) almost surely. If we can explicitly write \( \nabla_{\beta} Q(x, a; \beta) \), then \( \beta^*_n \) is the solution of equation 0 \( \sum_{i=1}^{n} \sum_{t=0}^{T-1} \gamma^{2t} \mathbb{E}_{\beta} \left( \omega_i \hat{\beta}_{\text{MRDR}} (\cdot, \beta) \right) \Omega_{\pi_b}(x_i) q_\beta(x_i, a_i, \hat{\beta}_{\text{MRDR}}(x_i) ; \beta) \nabla_{\beta} Q(x_i, a_i; \beta) \). 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:

\[
J_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} \gamma^{2t} \mathbb{E}_{\beta} \left( \omega_i \hat{\beta}_{\text{MRDR}} (\cdot, \beta) \right) 1 \left\{ \pi_e(x_i) = a_i \right\}
\]

\[
1 - \frac{\pi_b(a_i|x_i; \beta)}{\pi_e(a_i|x_i; \beta)} \left( \hat{\beta}_{\text{MRDR}} - \hat{\beta}_{\text{MRDR}} (x_i, a_i; \beta) \right)^2.
\]

The intuition behind the weights in WLS (16) is 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 \( 1 \left\{ \pi_e(x_t) = a_t \right\} \).

### 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., \( \lim_{n \rightarrow \infty} \hat{\beta}_{\text{MRDR}}(\cdot, \beta) = \beta^* \) 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 \( \pi \), 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 \( \alpha \) and \( \beta \), 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 \( \pi \) 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 \( \frac{1-(\alpha + \beta \times u)}{l-1} \). 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.

In the contextual bandit experiments, it is clear that in most cases the proposed MRDR estimator is superior to all alternative 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 for 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 DRO. This also illustrates the benefits of optimizing the model parameter based on the knowledge of trajectory distribution \( P_{\xi}^* \), which is generated by the evaluation policy.

<table>
<thead>
<tr>
<th>Evaluation Policy</th>
<th>( \alpha )</th>
<th>( \beta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friendly I</td>
<td>0.9</td>
<td>0</td>
</tr>
<tr>
<td>Friendly II</td>
<td>0.7</td>
<td>0.2</td>
</tr>
<tr>
<td>Neutral</td>
<td>0.5</td>
<td>0.2</td>
</tr>
<tr>
<td>Adversary I</td>
<td>0.3</td>
<td>0.2</td>
</tr>
<tr>
<td>Adversary II</td>
<td>0.5</td>
<td>0.2</td>
</tr>
</tbody>
</table>

**Table 2. Vehicle**

<table>
<thead>
<tr>
<th>Behavior Policy</th>
<th>DM</th>
<th>IS</th>
<th>DR</th>
<th>MRDR</th>
<th>DRO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friendly I</td>
<td>0.3273</td>
<td>0.0347</td>
<td>0.0217</td>
<td>0.0202</td>
<td>0.0224</td>
</tr>
<tr>
<td>Friendly II</td>
<td>0.3499</td>
<td>0.0517</td>
<td>0.0331</td>
<td>0.0318</td>
<td>0.0356</td>
</tr>
<tr>
<td>Neutral</td>
<td>0.4384</td>
<td>0.087</td>
<td>0.0604</td>
<td>0.0549</td>
<td>0.0722</td>
</tr>
<tr>
<td>Adversary I</td>
<td>0.405</td>
<td>0.0937</td>
<td>0.0616</td>
<td>0.0516</td>
<td>0.0769</td>
</tr>
<tr>
<td>Adversary II</td>
<td>0.405</td>
<td>0.1131</td>
<td>0.0712</td>
<td>0.0602</td>
<td>0.0952</td>
</tr>
</tbody>
</table>

**Table 3. SatImage**

<table>
<thead>
<tr>
<th>Behavior Policy</th>
<th>DM</th>
<th>IS</th>
<th>DR</th>
<th>MRDR</th>
<th>DRO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friendly I</td>
<td>0.2884</td>
<td>0.0128</td>
<td>0.0071</td>
<td>0.0063</td>
<td>0.0073</td>
</tr>
<tr>
<td>Friendly II</td>
<td>0.3328</td>
<td>0.0191</td>
<td>0.0107</td>
<td>0.0087</td>
<td>0.0119</td>
</tr>
<tr>
<td>Neutral</td>
<td>0.3848</td>
<td>0.0413</td>
<td>0.0246</td>
<td>0.0186</td>
<td>0.0335</td>
</tr>
<tr>
<td>Adversary I</td>
<td>0.3963</td>
<td>0.0459</td>
<td>0.027</td>
<td>0.0195</td>
<td>0.0383</td>
</tr>
<tr>
<td>Adversary II</td>
<td>0.4093</td>
<td>0.0591</td>
<td>0.0364</td>
<td>0.0262</td>
<td>0.0521</td>
</tr>
</tbody>
</table>

**Table 4. PenDigits**

<table>
<thead>
<tr>
<th>Behavior Policy</th>
<th>DM</th>
<th>IS</th>
<th>DR</th>
<th>MRDR</th>
<th>DRO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friendly I</td>
<td>0.4014</td>
<td>0.0103</td>
<td>0.0056</td>
<td>0.0037</td>
<td>0.0059</td>
</tr>
<tr>
<td>Friendly II</td>
<td>0.4628</td>
<td>0.0159</td>
<td>0.0092</td>
<td>0.0056</td>
<td>0.0194</td>
</tr>
<tr>
<td>Neutral</td>
<td>0.564</td>
<td>0.0450</td>
<td>0.0314</td>
<td>0.0138</td>
<td>0.0412</td>
</tr>
<tr>
<td>Adversary I</td>
<td>0.5861</td>
<td>0.0503</td>
<td>0.0366</td>
<td>0.0172</td>
<td>0.0472</td>
</tr>
<tr>
<td>Adversary II</td>
<td>0.5641</td>
<td>0.0646</td>
<td>0.0444</td>
<td>0.0188</td>
<td>0.0611</td>
</tr>
</tbody>
</table>

**Table 5. Letter**

<table>
<thead>
<tr>
<th>Behavior Policy</th>
<th>DM</th>
<th>IS</th>
<th>DR</th>
<th>MRDR</th>
<th>DRO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friendly I</td>
<td>0.392</td>
<td>0.0074</td>
<td>0.0056</td>
<td>0.0044</td>
<td>0.0057</td>
</tr>
<tr>
<td>Friendly II</td>
<td>0.4146</td>
<td>0.0102</td>
<td>0.0077</td>
<td>0.0054</td>
<td>0.0083</td>
</tr>
<tr>
<td>Neutral</td>
<td>0.4713</td>
<td>0.0467</td>
<td>0.0363</td>
<td>0.0315</td>
<td>0.0545</td>
</tr>
<tr>
<td>Adversary I</td>
<td>0.46</td>
<td>0.0587</td>
<td>0.0455</td>
<td>0.0385</td>
<td>0.0575</td>
</tr>
<tr>
<td>Adversary II</td>
<td>0.4728</td>
<td>0.0714</td>
<td>0.055</td>
<td>0.0481</td>
<td>0.0703</td>
</tr>
</tbody>
</table>
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 $\alpha, \beta$. 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.

For the model-based methods, we set the number of trajectories for sampling-based part of estimators to 512. The sample size for DR is set to 256 in all experiments. The corresponding MSE values are shown in Table 7. For the model-based methods, we set the number of trajectories for sampling-based part of estimators to 512. The sample size for DR is set to 256 in all experiments. The corresponding MSE values are shown in Table 7. For the model-based methods, we set the number of trajectories for sampling-based part of estimators to 512. The sample size for DR is set to 256 in all experiments. The corresponding MSE values are shown in Table 7.

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 DR 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.

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

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 $\alpha, \beta$. 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.

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 DR 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.

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


More Robust Doubly Robust Off-policy Evaluation


A. Proofs of Section 3.1

Proposition 2. Solving the weighted MSE problem
\[
\beta^* \in \arg \min_{\beta \in \mathbb{R}^n} \mathbb{E}_{(x,a) \sim \mu_{\pi_e}} \left[ (Q^{\pi_e}(x,a) - \hat{Q}^{\pi_e}(x,a;\beta))^2 \right],
\]
where \( \mu_{\pi_e} \) is the \( \gamma \)-discounted horizon-\( T \) state-action occupancy of \( \pi_e \) and \( \mathbb{1}\{\cdot\} \) is the indicator function is equivalent to solving
\[
\beta^* \in \arg \min_{\beta \in \mathbb{R}^n} \sum_{t=0}^{T-1} \gamma^t \mathbb{E}_{\rho_{\pi_e}^t} \left[ \omega_0 t \left( \hat{R}_{t:T-1}(\xi) - \hat{Q}^{\pi_e}(x_t,a_t;\beta) \right)^2 \right],
\]
where \( \hat{R}_{t:T-1}(\xi) = \sum_{\tau=t}^{T-1} \gamma^{\tau-t} \omega_{t+1:\tau} r(x_\tau,a_\tau) \).

Proof. We first expand the occupation measure \( \mu_{\pi_e} \) and use the change of measures by importance sampling as follows:
\[
\arg \min_{\beta \in \mathbb{R}^n} \mathbb{E}_{(x,a) \sim \mu_{\pi_e}} \left[ (Q^{\pi_e}(x,a) - \hat{Q}^{\pi_e}(x,a;\beta))^2 \right] = \arg \min_{\beta \in \mathbb{R}^n} \sum_{t=0}^{T-1} \gamma^t \cdot \mathbb{E}_{\rho_{\pi_e}^t} \left[ (Q^{\pi_e}(x_t,a_t) - \hat{Q}^{\pi_e}(x_t,a_t;\beta))^2 \right] =
\]
\[
\arg \min_{\beta \in \mathbb{R}^n} \sum_{t=0}^{T-1} \gamma^t \cdot \mathbb{E}_{\rho_{\pi_e}^t} \left[ \left( \mathbb{E}_{P_{\pi_e}} \left[ \sum_{\tau=t}^{T-1} \gamma^{\tau-t} r(x_\tau,a_\tau) \mid x_t,a_t \right] - \hat{Q}^{\pi_e}(x_t,a_t;\beta) \right) \right]^2 \]
\[
= \arg \min_{\beta \in \mathbb{R}^n} \sum_{t=0}^{T-1} \gamma^t \cdot \mathbb{E}_{\rho_{\pi_e}^t} \left[ \left( \mathbb{E}_{P_{\pi_e}} \left[ \sum_{\tau=t}^{T-1} \gamma^{\tau-t} \omega_{t+1:\tau} r(x_\tau,a_\tau) \mid x_t,a_t \right] - \hat{Q}^{\pi_e}(x_t,a_t;\beta) \right) \right]^2 \]
\[
(a) \quad \text{To shorten the notations, let’s define } f(X,Y) = \sum_{\tau=t}^{T-1} \gamma^{\tau-t} \omega_{t+1:\tau} r(x_\tau,a_\tau) \text{ and } g(X;\beta) = \hat{Q}^{\pi_e}(x_t,a_t;\beta), \text{ where } X = (x_t,a_t) \text{ and } Y = \{x_{t+1},a_{t+1},r_{t+1},\ldots,x_{T-1},a_{T-1},r_{T-1}\} \text{ are random variables. Using these notations, we may write}
\]
\[
\arg \min_{\beta \in \mathbb{R}^n} \mathbb{E}_{X,Y} \left[ (f(X,Y) - g(X;\beta))^2 \right] = \arg \min_{\beta \in \mathbb{R}^n} \mathbb{E}_{X,Y} \left[ (f(X,Y) - \mathbb{E}_Y[f(X,Y)\mid X] + \mathbb{E}_Y[f(X,Y)\mid X] - g(X;\beta))^2 \right] =
\]
\[
\arg \min_{\beta \in \mathbb{R}^n} \mathbb{E}_{X,Y} \left[ (f(X,Y) - \mathbb{E}_Y[f(X,Y)\mid X])^2 + (\mathbb{E}_Y[f(X,Y)\mid X] - g(X;\beta))^2 \right]
\]
\[
+ 2(f(X,Y) - \mathbb{E}_Y[f(X,Y)\mid X]) \cdot (\mathbb{E}_Y[f(X,Y)\mid X] - g(X;\beta)) \quad (b)
\]
\[
\arg \min_{\beta \in \mathbb{R}^n} \mathbb{E}_{X,Y} \left[ (\mathbb{E}_Y[f(X,Y)\mid X] - g(X;\beta))^2 - 2(f(X,Y) - \mathbb{E}_Y[f(X,Y)\mid X]) \cdot g(X;\beta) \right] \quad (c)
\]
\[
= \arg \min_{\beta \in \mathbb{R}^n} \mathbb{E}_{X,Y} \left[ (\mathbb{E}_Y[f(X,Y)\mid X] - g(X;\beta))^2 \right],
\]
(b) Here we drop the terms that are independent of \( \beta \).

(c) This passage is due to the fact that \( \mathbb{E}_Y[f(X,Y) - \mathbb{E}_Y[f(X,Y)\mid X] \mid X] = 0. \)
B. Proofs of Section 3.2

B.1. Bias of the IS Estimator ($\pi_b$ is unknown)

In case the behavior policy $\pi_b$ is unknown, the IS estimator is written as

$$\hat{\rho}_{\text{IS}} = \frac{1}{n} \sum_{i=1}^{n} \omega_{0:T-1}^{(i)} \sum_{t=0}^{T-1} \gamma^t \tau_t^{(i)} = \frac{1}{n} \sum_{i=1}^{n} \omega_{0:T-1}^{(i)} R_{0:T-1}^{(i)}, \quad (17)$$

where $\omega_{0:T-1}^{(i)} = \prod_{t=0}^{T-1} \frac{\pi_b(a_t|x_t)}{\pi_b(a_t|x_t^*)}$ is the approximate cumulative importance ratio of trajectory $\xi^{(i)} \in \mathcal{D}$ and $\hat{\pi}_b$ is our approximation of the behavior policy $\pi_b$. Note that $\hat{\pi}_b$ should be computed from a data set different than $\mathcal{D} = \{\xi^{(i)}\}_{i=1}^{n}$ that we use for our estimator.

**Proposition 3.** The bias of the IS estimator (17) is

$$\text{Bias}(\hat{\rho}_{\text{IS}}) = \left| \rho_{\xi}^{\pi_e} - \mathbb{E}_{\rho_{\xi}^{\pi_b}}[\hat{\rho}_{\text{IS}}^{\xi}] \right| = \left| \mathbb{E}_{\rho_{\xi}^{\pi_b}}[\delta_{0:T-1}(\xi) R_{0:T-1}(\xi)] \right|, \quad (18)$$

where $\xi = (x_0, a_0, r_0, \ldots, x_T, a_T)$ is a trajectory and

$$\delta_{0:T-1}(\xi) = 1 - \prod_{t=0}^{T-1} \frac{\pi_b(a_t|x_t)}{\pi_b(a_t|x_t^*)}.$$

**Proof.** To prove (18), we first develop the term $\mathbb{E}_{\rho_{\xi}^{\pi_b}}[\hat{\rho}_{\text{IS}}^{\xi}]$ as follows:

$$\mathbb{E}_{\rho_{\xi}^{\pi_b}}[\hat{\rho}_{\text{IS}}^{\xi}] = \mathbb{E}_{\rho_{\xi}^{\pi_b}} \left[ \frac{1}{n} \sum_{i=1}^{n} \omega_{0:T-1}^{(i)} R_{0:T-1}^{(i)} \right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\rho_{\xi}^{\pi_b}} \left[ \omega_{0:T-1}^{(i)} R_{0:T-1}^{(i)} \right] \overset{(b)}{=} \mathbb{E}_{\rho_{\xi}^{\pi_b}} [\delta_{0:T-1} R_{0:T-1}] \quad (19)$$

(a) This is because the trajectories $\xi^{(i)} \in \mathcal{D}$ are i.i.d.

(b) We define $\lambda_{0:T-1} = \prod_{t=0}^{T-1} \frac{\pi_b(a_t|x_t)}{\pi_b(a_t|x_t^*)}$.

Given (19), we may write the bias as

$$\text{Bias}(\hat{\rho}_{\text{IS}}) = \left| \rho_{\xi}^{\pi_e} - \mathbb{E}_{\rho_{\xi}^{\pi_b}}[\hat{\rho}_{\text{IS}}^{\xi}] \right| = \left| \mathbb{E}_{\rho_{\xi}^{\pi_b}}[R_{0:T-1}] - \mathbb{E}_{\rho_{\xi}^{\pi_e}}[\lambda_{0:T-1} R_{0:T-1}] \right| = \left| \mathbb{E}_{\rho_{\xi}^{\pi_e}}[\delta_{0:T-1} R_{0:T-1}] \right|,$$

which concludes the proof. \qed

B.2. Bias of the step-IS Estimator ($\pi_b$ is unknown)

In case the behavior policy $\pi_b$ is unknown, the step-IS estimator is written as

$$\hat{\rho}_{\text{step-IS}} = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=0}^{T-1} \gamma^t \omega_{0:t}^{(i)} \tau_t^{(i)}, \quad (20)$$

where $\omega_{0:t}^{(i)} = \prod_{t=0}^{T-1} \frac{\pi_b(a_t|x_t)}{\pi_b(a_t|x_t^*)}$ is the approximate cumulative importance ratio of trajectory $\xi^{(i)} \in \mathcal{D}$ and $\hat{\pi}_b$ is our approximation of the behavior policy $\pi_b$. Note that $\hat{\pi}_b$ should be computed from a data set different than $\mathcal{D} = \{\xi^{(i)}\}_{i=1}^{n}$ that we use for our estimator.

**Proposition 4.** The bias of the step-IS estimator (20) is

$$\text{Bias}(\hat{\rho}_{\text{step-IS}}) = \left| \rho_{\xi}^{\pi_e} - \mathbb{E}_{\rho_{\xi}^{\pi_b}}[\hat{\rho}_{\text{step-IS}}^{\xi}] \right| = \left| \sum_{t=0}^{T-1} \gamma^t \mathbb{E}_{\rho_{\xi}^{\pi_b}}[\delta_{0:t}(\xi) r_t] \right|, \quad (21)$$
where $\xi = (x_0, a_0, r_0, \ldots, x_{T-1}, a_{T-1}, r_{T-1}, x_T)$ is a trajectory and

$$\delta_{0:t}(\xi) = 1 - \prod_{t=0}^{T} \frac{\pi_b(a_t | x_t)}{\hat{\pi}_b(a_t | x_t)}.$$ 

**Proof.** To prove (21), we first develop the term $\mathbb{E}_{\pi^b_{\xi}}[\hat{\rho}_{\text{step-IS}}]$ as follows:

$$\mathbb{E}_{\pi^b_{\xi}}[\hat{\rho}_{\text{step-IS}}] = \mathbb{E}_{\pi^b_{\xi}} \left[ \frac{1}{n} \sum_{i=1}^{n} \sum_{t=0}^{T-1} \gamma^t \hat{\omega}_{0:t} r_t \right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\pi^b_{\xi}} \left[ \sum_{t=0}^{T-1} \gamma^t \hat{\omega}_{0:t} r_t \right] = \mathbb{E}_{\pi^b_{\xi}} \left[ \sum_{t=0}^{T-1} \gamma^t \hat{\omega}_{0:t} r_t \right].$$ 

(a) This is because the trajectories $\xi^{(i)} \in \mathcal{D}$ are i.i.d.
(b) We define $\lambda_{0:t} = \prod_{t=0}^{T} \frac{\pi_b(a_t | x_t)}{\hat{\pi}_b(a_t | x_t)}$. Since step-IS is unbiased when $\pi_b$ is known, similar to (22), we may write

$$\rho^{\pi_e} = \mathbb{E}_{\pi^b_{\xi}} \left[ \frac{1}{n} \sum_{i=1}^{n} \sum_{t=0}^{T-1} \gamma^t \omega_{0:t} r_t \right] = \sum_{t=0}^{T-1} \gamma^t \mathbb{E}_{\pi^b_{\xi}} [\omega_{0:t} r_t] = \sum_{t=0}^{T-1} \gamma^t \mathbb{E}_{\pi^e_{\xi}} [r_t].$$ 

(23)

Given (22) and (23), we may write the bias as

$$\text{Bias}(\hat{\rho}_{\text{step-IS}}) = \left| \rho^{\pi_e} - \mathbb{E}_{\pi^b_{\xi}}[\hat{\rho}_{\text{step-IS}}] \right| = \left| \sum_{t=0}^{T-1} \gamma^t \mathbb{E}_{\pi^e_{\xi}} [r_t] - \sum_{t=0}^{T-1} \gamma^t \mathbb{E}_{\pi^b_{\xi}} [\lambda_{0:t} r_t] \right| = \left| \sum_{t=0}^{T-1} \gamma^t \mathbb{E}_{\pi^e_{\xi}} [\delta_{0:t} r_t] \right|,$$

which concludes the proof.
C. Proofs of Section 3.3

C.1. Bias of the DR Estimator (π_b is unknown)

When the behavior policy π_b is unknown, the DR estimator (7) is written as

\[
\hat{\rho}_{\text{DR}}^{\pi_b} = \frac{1}{n} \sum_{i=1}^{n} \gamma_t \left( \omega_{0:t}^{(i)} \pi_t^{(i)} - \omega_{0:t}^{(i)} \hat{Q}_t^{\pi_b}(x_t^{(i)}, a_t^{(i)}) + \omega_{0:t-1}^{(i)} \hat{\nu}^{\pi_b}(x_t^{(i)}) \right),
\]

where \( \omega_{0:T-1} = \prod_{\tau=0}^{T-1} \frac{\pi_b(a_t|x_\tau)}{\pi_{\hat{b}}(a_t|x_\tau)} \) is the approximate cumulative importance ratio and \( \pi_{\hat{b}} \) is our approximation of the behavior policy π_b.

**Proposition 5.** The bias of the DR estimator (24) is

\[
\text{Bias}(\hat{\rho}_{\text{DR}}^{\pi_b}) = E_\pi E_{P^{\pi_b}} \left[ \rho^{\pi_b} - E_{P^{\hat{\pi}_b}} \left[ \hat{\rho}_{\text{DR}}^{\pi_b} \right] \right] = E_\pi E_{P^{\pi_b}} \left[ \sum_{t=0}^{T-1} \gamma_t \left( 1 - \delta_{0:t-1}(\xi) \right) \delta_t(\xi) \Delta^{\pi_b}(x_t, a_t) \right],
\]

where \( \xi = (x_0, a_0, r_0, \ldots, x_{T-1}, a_{T-1}, r_{T-1}, x_T) \) is a trajectory and

\[
\delta_{0:t-1}(\xi) = 1 - \prod_{\tau=0}^{t-1} \frac{\pi_b(a_t|x_\tau)}{\pi_{\hat{b}}(a_t|x_\tau)}, \quad \delta_t(\xi) = 1 - \frac{\pi_b(a_t|x_t)}{\pi_{\hat{b}}(a_t|x_t)}.
\]

**Proof.** We first develop the term \( E_{P^{\pi_b}}[\hat{\rho}_{\text{DR}}^{\pi_b}] \) as follows:

\[
E_{P^{\pi_b}}[\hat{\rho}_{\text{DR}}^{\pi_b}] = E_{P^{\pi_b}} \left[ \frac{1}{n} \sum_{i=1}^{n} \sum_{t=0}^{T-1} \gamma_t \left( \omega_{0:t}^{(i)} \pi_t^{(i)} - \omega_{0:t}^{(i)} \hat{Q}_t^{\pi_b}(x_t^{(i)}, a_t^{(i)}) + \omega_{0:t-1}^{(i)} \hat{\nu}^{\pi_b}(x_t^{(i)}) \right) \right]
\]

\[
= E_{P^{\pi_b}} \left[ \sum_{t=0}^{T-1} \gamma_t \left( \omega_{0:t} \pi_t - \omega_{0:t} \hat{Q}_t^{\pi_b}(x_t, a_t) + \omega_{0:t-1} \hat{\nu}^{\pi_b}(x_t) \right) \right]
\]

\[\overset{(a)}{=} E_{P^{\pi_b}} \left[ \sum_{t=0}^{T-1} \gamma_t \left( \omega_{0:t} \pi_t - \omega_{0:t} \left( \hat{Q}_t^{\pi_b}(x_t, a_t) + \Delta^{\pi_b}(x_t, a_t) + \hat{\nu}^{\pi_b}(x_t) \right) \right) \right]
\]

\[\overset{(b)}{=} E_{P^{\pi_b}} \left[ \gamma^{T-1} \omega_{0:T-1} (r_{T-1} - Q^{\pi_b}(x_{T-1}, a_{T-1})) - E_{P^{\pi_b}} \left[ \sum_{t=0}^{T-2} \gamma_t \omega_{0:t} \pi_t (r_t + \gamma V^{\pi_b}(x_{t+1}) - Q^{\pi_b}(x_t, a_t)) \right] \right]
\]

\[\overset{(c)}{=} \rho^{\pi_b} - E_{P^{\pi_b}} \left[ \sum_{t=0}^{T-1} \gamma_t \left( \omega_{0:t} \Delta^{\pi_b}(x_t, a_t) - \omega_{0:t-1} \Delta^{\pi_b}(x_t) \right) \right]
\]

\[\overset{(d)}{=} \rho^{\pi_b} - E_{P^{\pi_b}} \left[ \sum_{t=0}^{T-1} \gamma_t \left( \lambda_{0:t} \Delta^{\pi_b}(x_t, a_t) - \lambda_{0:t-1} E_{a_t \sim \pi_b \cdot (x_t)} \left[ \Delta^{\pi_b}(x_t, a_t) \right] \right) \right]
\]

\[\overset{(e)}{=} \rho^{\pi_b} - E_{P^{\pi_b}} \left[ \sum_{t=0}^{T-1} \gamma_t \left( \lambda_{0:t} - \lambda_{0:t-1} \right) \Delta^{\pi_b}(x_t, a_t) \right] \overset{(f)}{=} \rho^{\pi_b} + E_{P^{\pi_b}} \left[ \sum_{t=0}^{T-1} \gamma_t \lambda_{0:t-1} \delta_t \Delta^{\pi_b}(x_t, a_t) \right]. \tag{25}
\]

(a) We define the bias of the DM estimator as \( \Delta^{\pi_b}(x, a) = \hat{Q}^{\pi_b}(x, a) - Q^{\pi_b}(x, a) \) and \( \Delta^{\pi_b}(x) = \hat{\nu}^{\pi_b}(x) - V^{\pi_b}(x). \)

(b) Application of the telescopic sum.

\[\overset{4}{\text{Note that with abuse of notation, here we used } \Delta^{\pi_b} \text{ as both a function of } \mathcal{X} \times \mathcal{A} \text{ and a function of } \mathcal{X}.}\]
(e) It comes from the fact that (i) $\rho_{\pi_e} = \mathbb{E}_{x_0 \sim p_0} \left[ V_{\pi_e}(x_0) \right]$, (ii) $\mathbb{E}_{p_{\pi_b}^\xi} \left[ \sum_{t=0}^{T-2} \gamma^t \omega_{0:t} \left( r_t + \gamma V_{\pi_e}(x_{t+1}) - Q_{\pi_e}(x_t, a_t) \right) \right] = 0$, and (iii) $r_{T-1} - Q_{\pi_e}(x_{T-1}, a_{T-1}) = 0$, because $V_{\pi_e}(x_T) = 0$.

(d) It comes from the definition of $\lambda_{0:t-1} = \prod_{\tau=0}^{t-1} \pi_b(a_\tau|x_\tau)$ and the fact that $\Delta_{\pi_e}(x_t) = \mathbb{E}_{a_t \sim \pi_e(\cdot|x_t)} \left[ \Delta_{\pi_e}(x_t, a_t) \right]$.

(e) It comes from the fact that $\mathbb{E}_{p_{\pi_e}^\xi} \left[ \lambda_{0:t-1} \mathbb{E}_{a_t \sim \pi_e(\cdot|x_t)} \left[ \Delta_{\pi_e}(x_t, a_t) \right] \right] = \mathbb{E}_{p_{\pi_e}^\xi} \left[ \lambda_{0:t-1} \Delta_{\pi_e}(x_t, a_t) \right]$.

(f) It comes from the definition of $\delta_t$.

Finally, 25 concludes the proof.
More Robust Doubly Robust Off-policy Evaluation

D. Proofs of Section 4.1

D.1. Proof of Proposition 1

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

\[
\text{Bias}(\hat{\rho}_{\text{DR}}^{\pi_x}) = \left| \rho^{\pi_x} - E_{\hat{\rho}^{\pi_x}}[\hat{\rho}_{\text{DR}}^{\pi_x}] \right| = \left| E_{\hat{\rho}^{\pi_x}}[\delta(x, a) \Delta(x, a)] \right|,
\]

\[
n\forall \hat{\rho}^{\pi_x} (\hat{\rho}_{\text{DR}}^{\pi_x}) = E_{\hat{\rho}^{\pi_x}} \left[ \omega(x, a) (r(x, a) - Q(x, a))^2 \right] + \forall_{\rho_0} \left( E_{\pi_x} [Q(x, a) + \delta(x, a) \Delta(x, a)] \right)
\]

\[+ E_{\rho_0, \pi_x} \left[ \omega(x, a) \left( (1 - \delta(x, a)) \Delta(x, a) \right)^2 - E_{\pi_x} \left[ (1 - \delta(x, a)) \Delta(x, a) \right]^2 \right].\]

Proof. Bias: For the bias of the estimator, note that

\[
\text{Bias}(\hat{\rho}_{\text{DR}}^{\pi_x}) = \left| \rho^{\pi_x} - E_{\hat{\rho}^{\pi_x}}[\hat{\rho}_{\text{DR}}^{\pi_x}] \right| = \left| \rho^{\pi_x} - E_{\hat{\rho}^{\pi_x}} \left[ \frac{\pi_x(a|x)}{\hat{\pi}_b(a|x)} (r(x, a) - \hat{Q}(x, a; \beta)) + \hat{V}^{\pi_x}(x; \beta) \right] \right|
\]

\[
= \left| \rho^{\pi_x} - E_{\hat{\rho}^{\pi_x}} \left[ \frac{\pi_x(a|x)}{\hat{\pi}_b(a|x)} \left( Q(x, a) - \hat{Q}(x, a; \beta) \right) \right] \right| - E_{\hat{\rho}^{\pi_x}} \left[ \hat{V}^{\pi_x}(x; \beta) \right]
\]

\[
= \left| \rho^{\pi_x} - E_{\hat{\rho}^{\pi_x}} \left[ (1 - \delta(x, a)) \left( Q(x, a) - \hat{Q}(x, a; \beta) \right) \right] \right| - E_{\hat{\rho}^{\pi_x}} \left[ \hat{V}^{\pi_x}(x; \beta) \right]
\]

\[
= \left| \rho^{\pi_x} - E_{\hat{\rho}^{\pi_x}} \left[ \delta(x, a) \Delta(x, a) \right] - E_{\hat{\rho}^{\pi_x}} \left[ \frac{\pi_x(a|x)}{\hat{\pi}_b(a|x)} \left( Q(x, a) - \hat{Q}(x, a; \beta) \right) \right] \right| + E_{\hat{\rho}^{\pi_x}} \left[ \hat{Q}(x, a; \beta) \right] - E_{\hat{\rho}^{\pi_x}} \left[ \hat{V}^{\pi_x}(x; \beta) \right]
\]

\[
= \left| E_{\rho^{\pi_x}} \left[ \delta(x, a) \Delta(x, a) \right] \right|.
\]

Variance: Before proving the variance, from the proof of bias, we notice that

\[
E_{\rho^{\pi_x}} \left[ \hat{\rho}_{\text{DR}}^{\pi_x} \right] = E_{\rho^{\pi_x}} \left[ Q(x, a) + \delta(x, a) \Delta(x, a) \right].
\]

For the variance of the estimator, note that

\[
n\forall E_{\rho^{\pi_x}} \left[ \hat{\rho}_{\text{DR}}^{\pi_x} \right] = \forall_{\hat{\rho}^{\pi_x}} \left[ \frac{\pi_x(a|x)}{\hat{\pi}_b(a|x)} (r(x, a) - \hat{Q}(x, a; \beta)) + \hat{V}^{\pi_x}(x; \beta) \right]
\]

\[
= \forall_{\hat{\rho}^{\pi_x}} \left[ \frac{\pi_x(a|x)}{\hat{\pi}_b(a|x)} \left( r(x, a) - \hat{Q}(x, a; \beta) \right) \right] + \left( E_{\rho^{\pi_x}} \left[ Q(x, a) + \delta(x, a) \Delta(x, a) \right] \right)^2
\]

\[
= \forall_{\hat{\rho}^{\pi_x}} \left[ \frac{\pi_x(a|x)}{\hat{\pi}_b(a|x)} \left( r(x, a) - Q(x, a) - \Delta(x, a) \right) \right] + \left( E_{\rho^{\pi_x}} \left[ Q(x, a) + \Delta(x, a) \right] \right)^2
\]

\[
= \left( E_{\rho^{\pi_x}} \left[ Q(x, a) + \delta(x, a) \Delta(x, a) \right] \right)^2
\]

\[
\overset{(a)}{=} \left( E_{\rho^{\pi_x}} \left[ \frac{\pi_x(a|x)}{\hat{\pi}_b(a|x)} \left( r(x, a) - Q(x, a) \right) \right] + \left( E_{\rho^{\pi_x}} \left[ \frac{\pi_x(a|x)}{\hat{\pi}_b(a|x)} \Delta(x, a) \right] \right)^2 \right)^2
\]

\[
- 2 E_{\rho^{\pi_x}} \left[ \frac{\pi_x(a|x)}{\hat{\pi}_b(a|x)} (x) E_{\pi_x} \left[ Q(x, a) + \Delta(x, a) \right] \right] - \left( E_{\rho^{\pi_x}} \left[ Q(x, a) + \delta(x, a) \Delta(x, a) \right] \right)^2
\]

\[
- 2 E_{\rho^{\pi_x}} \left[ \frac{\pi_x(a|x)}{\hat{\pi}_b(a|x)} \left( r(x, a) - Q(x, a) \right) \right] + 2 E_{\rho^{\pi_x}} \left[ \frac{\pi_x(a|x)}{\hat{\pi}_b(a|x)} \left( r(x, a) - Q(x, a) \right) \right] E_{\pi_x} \left[ Q(x, a) + \Delta(x, a) \right]
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]
More Robust Doubly Robust Off-policy Evaluation

\[
\begin{align*}
&= \mathbb{E}_{p_{\pi\beta}} \left[ \left( \frac{\pi_e(a|x)}{\pi_{\hat{b}}(a|x)} (r(x, a) - Q(x, a)) \right)^2 \right] + \mathbb{E}_{p_{\gamma\pi}} \left[ \frac{\pi_e(a|x)}{\pi_{\hat{b}}(a|x)} \left( \frac{\pi_e(a|x)}{\pi_{\hat{b}}(a|x)} \Delta(x, a) \right)^2 \right] - \mathbb{E}_{p_{\gamma\pi}} \left[ (Q(x, a) + \delta(x, a) \Delta(x, a))^2 \right]
+ \mathbb{E}_{p_{\beta}} \left[ \left( \mathbb{E}_{\pi_e} (Q(x, a) + \Delta(x, a))^2 \right] - 2\mathbb{E}_{p_{\gamma\pi}} \left[ (1 - \delta(x, a)) \Delta(x, a) \mathbb{E}_{\pi_e} (Q(x, a) + \Delta(x, a)) \right]
\end{align*}
\]

\[
\begin{align*}
&= \mathbb{E}_{p_{\pi\beta}} \left[ \left( \frac{\pi_e(a|x)}{\pi_{\hat{b}}(a|x)} (r(x, a) - Q(x, a)) \right)^2 \right] + \mathbb{E}_{p_{\gamma\pi}} \left[ \frac{\pi_e(a|x)}{\pi_{\hat{b}}(a|x)} \left( \frac{\pi_e(a|x)}{\pi_{\hat{b}}(a|x)} \Delta(x, a) \right)^2 \right] - \mathbb{E}_{p_{\gamma\pi}} \left[ (Q(x, a) + \delta(x, a) \Delta(x, a))^2 \right]
+ \mathbb{E}_{p_{\gamma\pi}} \left[ (Q(x, a) + \delta(x, a) \Delta(x, a))^2 \right] + \mathbb{E}_{p_{\gamma\pi}} \left[ (Q(x, a) + \delta(x, a) \Delta(x, a))^2 \right]
\end{align*}
\]

which concludes the proof.

(a) Here we used the fact that

\[
\mathbb{E}_{p_{\pi\beta}} \left[ \frac{\pi_e(a|x)}{\pi_{\hat{b}}(a|x)} \Delta(x, a) (r(x, a) - Q(x, a)) \right] = \mathbb{E}_{p_{\gamma\pi}} \left[ \frac{\pi_e(a|x)}{\pi_{\hat{b}}(a|x)} \Delta(x, a) \mathbb{E}_{\pi_e} (Q(x, a) + \Delta(x, a)) \right] = 0
\]

and

\[
\mathbb{E}_{p_{\gamma\pi}} \left[ \frac{\pi_e(a|x)}{\pi_{\hat{b}}(a|x)} (r(x, a) - Q(x, a)) \mathbb{E}_{\pi_e} (Q(x, a) + \Delta(x, a)) \right] = \mathbb{E}_{p_{\gamma\pi}} \left[ \frac{\pi_e(a|x)}{\pi_{\hat{b}}(a|x)} \mathbb{E}_{\pi_e} (Q(x, a) + \Delta(x, a)) \mathbb{E}_{\pi_e} (r(x, a) - Q(x, a)) \right] = 0
\]

(b) Developing the last two terms, we obtain

\[
\begin{align*}
&= \mathbb{E}_{p_{\gamma\pi}} \left[ \left( \mathbb{E}_{\pi_e} (Q(x, a) + \Delta(x, a))^2 \right] - 2\mathbb{E}_{p_{\gamma\pi}} \left[ (1 - \delta(x, a)) \Delta(x, a) \mathbb{E}_{\pi_e} (Q(x, a) + \Delta(x, a)) \right]
+ \mathbb{E}_{p_{\gamma\pi}} \left[ \frac{\pi_e(a|x)}{\pi_{\hat{b}}(a|x)} \Delta(x, a) \mathbb{E}_{\pi_e} (Q(x, a) + \Delta(x, a)) \right]
\end{align*}
\]

Note that when \(\pi_e\) is deterministic, the variance proved in Proposition 1 may be written as

\[
n\mathbb{E}_{p_{\pi\beta}} \left( \frac{\pi_e(a|x)}{\pi_{\hat{b}}(a|x)} (r(x, a) - Q(x, a)) \right)^2 + \mathbb{E}_{p_{\gamma\pi}} \left[ V_{\pi\gamma} (x) + \delta(x, \pi_e(x)) \Delta(x, \pi_e(x)) \right]
\]

which is equal to the variance reported in *Dudik et al.* (2011).
D.2. Proof of Theorem 1

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

(i)  
\[
n \frac{\pi_e}{\pi_b} (\frac{\hat{\pi}_e}{\hat{\pi}_b}) = E_{\pi_e} \left[ \omega(x,a) \left( \frac{\pi_e(a|x)}{\pi_b(a|x)} \left( r(x,a) - Q(x,a) \right) \right)^2 \right] + \frac{1}{\pi_b(a|x)} \omega(x,a) r(x,a) ^ 2
\]

(ii)  
\[
n \frac{\pi_e}{\pi_b} (\frac{\hat{\pi}_e}{\hat{\pi}_b}) = E_{\pi_e} \left[ \omega(x,a) q_b(x,a,r) \right] \Omega_e \left( \frac{\hat{\pi}_e}{\hat{\pi}_b} \right) + C
\]

where $\Omega_e(x) = \text{diag}[1/\pi_e(a|x)]_{a \in A} - ee^T$ with $e = [1, ..., 1]^T$; $q_b(x,a,r) = D_{\pi_e}(x) \hat{Q}(x,b) - 1(a)r$ is a row vector with $D_{\pi_e}(x) = \text{diag}[\pi_e(a|x)]_{a \in A}$, row vector $\hat{Q}(x,b) = [\hat{Q}(x,a,b)]_{a \in A}$, and the row vector of indicator functions $l(a) = [1 \{ a' = a \}]_{a' \in A}$; and finally

\[
C = \left[ E_{\pi_e} \left[ \frac{\pi_e(a|x)}{\pi_b(a|x)} \left( r(x,a) \right) \right]^2 \right] + \frac{1}{\pi_b(a|x)} \omega(x,a) r(x,a) ^ 2.
\]

Proof. In this proof, we first derive the expression in (i) using the variance formulation from Lemma 1 and later deduce the expression in (ii) using the completion of squares.

For the derivation of (i), recall from Proposition 1 that when the behavior policy $\pi_b$ is known, then the policy approximation error $\delta$ vanishes uniformly, and the variance formulation is given by

\[
n \frac{\pi_e}{\pi_b} (\frac{\hat{\pi}_e}{\hat{\pi}_b}) = E_{\pi_e} \left[ \left( \frac{\pi_e(a|x)}{\pi_b(a|x)} \left( r(x,a) - Q(x,a) \right) \right)^2 \right] + \frac{1}{\pi_b(a|x)} \omega(x,a) r(x,a) ^ 2.
\]

Using the following total variance decomposition rule:

\[
\nu_{P_0,P_r} [E_{\pi_e} [r(x,a)]] = E_{P_0} \left[ \nu_{P_r} [E_{\pi_e} [r(x,a)]] \right] + E_{P_0} \left[ \nu_{P_r} [E_{\pi_e} [r(x,a)]] \right],
\]

we can show that

\[
\nu_{P_0} [V^{\pi_e} (x)] = \nu_{P_0} \left[ E_{P_r} \left[ E_{\pi_e} [r(x,a)] \right] \right]
\]

\[
= \nu_{P_0} \left[ E_{\pi_e} [r(x,a)] \right] - \nu_{P_0} \left[ E_{P_r} \left[ E_{\pi_e} [r(x,a)] \right] \right]
\]

\[
= \nu_{P_0} \left[ E_{\pi_e} [r(x,a)] \right] - \nu_{P_0} \left[ \left( E_{\pi_e} [r(x,a)] - V^{\pi_e} (x) \right)^2 \right],
\]

where the last equality follows from the fact that $\nu_{P_r} [E_{\pi_e} [r(x,a)] = E_{P_r} \left[ (E_{\pi_e} [r(x,a)] - V^{\pi_e} (x))^2 \right]$. Utilizing this equality, the variance $\nu_{P_0} (\frac{\hat{\pi}_e}{\hat{\pi}_b})$ can be re-written as

\[
n \frac{\pi_e}{\pi_b} (\frac{\hat{\pi}_e}{\hat{\pi}_b}) = E_{\pi_e} \left[ \left( \frac{\pi_e(a|x)}{\pi_b(a|x)} \left( r(x,a) - Q(x,a) \right) \right)^2 \right] + \nu_{P_0} \left[ E_{\pi_e} [r(x,a)] \right] - \nu_{P_0} \left[ \left( E_{\pi_e} [r(x,a)] - Q(x,a) \right)^2 \right]
\]

\[
+ \nu_{P_0,\pi_b} \left[ \left( \frac{\pi_e(a|x)}{\pi_b(a|x)} \left( r(x,a) \right) \right)^2 \right] - \nu_{P_0} \left[ \left( \frac{\pi_e(a|x)}{\pi_b(a|x)} \left( r(x,a) \right) \right)^2 \right],
\]

which is due to the following importance sampling properties:

\[
E_{\pi_e} \left[ \frac{\pi_e(a|x)}{\pi_b(a|x)} \left( \Delta(x,a) \right) \right] = E_{\pi_b} \left[ \left( \frac{\pi_e(a|x)}{\pi_b(a|x)} \right)^2 \left( \Delta(x,a) \right) \right],
\]
and

\[ 
E_{π_0}[Δ(x, a)] = E_{π_0} \left[ \frac{π_e(a|x)}{π_0(a|x)} Δ(x, a) \right].
\]

Now recall the definition of the model error \( Δ(x, a) = \hat{Q}(x, a) - Q(x, a) \). Since \( E_{π_0}[r(x, a)] = Q(x, a) \) for every state \( x ∈ X \) and action \( a ∈ A \), one can easily show that

\[
E_{π_0, P_0} \left[ \left( \frac{π_e(a|x)}{π_0(a|x)} (r(x, a) - Q(x, a)) - Δ(x, a) \right)^2 \right] = E_{π_0, P_0} \left[ \left( \frac{π_e(a|x)}{π_0(a|x)} (r(x, a) - Q(x, a) - Δ(x, a)) \right)^2 \right] = E_{π_0, P_0} \left[ \left( \frac{π_e(a|x)}{π_0(a|x)} (r(x, a) - \hat{Q}(x, a)) \right)^2 \right],
\]

because the corresponding cross-term cancels out. By using the same argument, one can also show that

\[
E_{P_0, P_0} \left[ (E_{π_0}[r(x, a) - Q(x, a)])^2 + (E_{π_0}[Δ(x, a)])^2 \right] = E_{P_0, P_0} \left[ (E_{π_0}[r(x, a) - Q(x, a)) - Δ(x, a)])^2 \right] = E_{P_0, P_0} \left[ (E_{π_0}[r(x, a) - \hat{Q}(x, a; β)])^2 \right]
\]

Collectively, the variance expression \( nV_{P_0}^0(\hat{r}^0_{DR}) \) can be further simplified by the following chain of equalities:

\[
\begin{align*}
\text{case } nV_{P_0}^0(\hat{r}^0_{DR}) & = V_{P_0, P_0} \left[ E_{π_0, P_0} \left[ \left( \frac{π_e(a|x)}{π_0(a|x)} (r(x, a) - Q(x, a) - Δ(x, a)) \right)^2 \right] \right] \\
& - E_{P_0, P_0} \left[ \left( \frac{π_e(a|x)}{π_0(a|x)} (r(x, a) - Q(x, a) - Δ(x, a)) \right)^2 \right] \\
& = V_{P_0, P_0} \left[ E_{π_0}[r(x, a)] \right] + V_{P_0, P_0} \left[ \left( \frac{π_e(a|x)}{π_0(a|x)} (r(x, a) - \hat{Q}(x, a; β)) \right)^2 \right] - E_{P_0, P_0} \left[ \left( \frac{π_e(a|x)}{π_0(a|x)} (r(x, a) - \hat{Q}(x, a; β)) \right)^2 \right] \\
& = V_{P_0, P_0} \left[ E_{π_0}[r(x, a)] \right] + V_{P_0, P_0} \left[ \left( \frac{π_e(a|x)}{π_0(a|x)} (r(x, a) - \hat{Q}(x, a; β)) \right)^2 \right] - E_{P_0, P_0} \left[ \left( \frac{π_e(a|x)}{π_0(a|x)} (r(x, a) - \hat{Q}(x, a; β)) \right)^2 \right] \\
& = V_{P_0, P_0} \left[ \left( \frac{π_e(a|x)}{π_0(a|x)} (r(x, a) - \hat{Q}(x, a; β)) \right)^2 \right] - E_{P_0, P_0} \left[ \left( \frac{π_e(a|x)}{π_0(a|x)} (r(x, a) - \hat{Q}(x, a; β)) \right)^2 \right] \\
& = V_{P_0, P_0} \left[ \left( \frac{π_e(a|x)}{π_0(a|x)} (r(x, a) - \hat{Q}(x, a; β)) \right)^2 \right]
\end{align*}
\]

To simplify the above variance formulation, one also notices that the conditional variance term can be re-written as

\[
\begin{align*}
V_{π_0} \left[ \frac{π_e(a|x)}{π_0(a|x)} (r(x, a) - \hat{Q}(x, a; β)) \right] & = E_{π_0} \left[ \left( \frac{π_e(a|x)}{π_0(a|x)} (r(x, a) - \hat{Q}(x, a; β)) \right)^2 \right] - \left( E_{π_0} \left[ r(x, a) - \hat{Q}(x, a; β) \right) \right)^2 \\
& = E_{π_0} \left[ \left( \frac{π_e(a|x)}{π_0(a|x)} (r(x, a) - \hat{Q}(x, a; β)) \right)^2 + 2 \left( \frac{π_e(a|x)}{π_0(a|x)} (r(x, a) - \hat{Q}(x, a; β)) \right) \left( \hat{V}^π_e(x; β) \right) - \left( r_{π_0(x)} \right)^2 \right] - (\hat{V}^π_e(x; β))^2 \\
& = E_{π_0} \left[ \left( \frac{π_e^2(a|x)}{π_0^2(a|x)} \hat{Q}(x, a); β \right)^2 \right] - (\hat{V}^π_e(x; β))^2 + E_{π_0} \left[ \left( \frac{π_e^2(a|x)}{π_0^2(a|x)} (r(x, a)) \right)^2 \right] - r^2_{π_e(x)} \\
& = -2E_{π_0} \left[ \left( \frac{π_e(a|x)}{π_0(a|x)} (r(x, a) \left( \frac{π_e(a|x)}{π_0(a|x)} \hat{Q}(x, a; β) - \hat{V}^π_e(x; β) \right) \right),
\end{align*}
\]

(27)
where \( r_{\pi_e(x)} = \mathbb{E}_{\pi_e}[r(x, a)] \) is the reward random variable induced by policy \( \pi_e \). The second equality is based on expanding the term \( (\mathbb{E}_{\pi_e}[r(x, a)] - \hat{Q}(x, a; \beta))^2 \) and recalling that \( \hat{V}^{\pi_e}(x; \beta) = \mathbb{E}_{\pi_e}[Q(x, a; \beta)] \), and the third equality is based on expanding the term \( \left( \frac{\pi_e(a|x)}{\pi_{\pi_e}(a|x)} r(x, a) - \hat{Q}(x, a; \beta) \right)^2 \) and collecting the cross-terms.

Therefore, by combining both the arguments in (26) and (27), and recalling that \( \omega(x, a) = \pi_e(a|x)/\pi_0(a|x) \), the variance expression \( n\mathbb{V}_{\rho_{\pi_e}}(\hat{\rho}_{\text{DR}}) \) becomes

\[
\mathbb{V}_{\rho_{\pi_e}}(\hat{\rho}_{\text{DR}}) = \mathbb{V}_{\rho_0}(\mathbb{E}_{\pi_e}[r(x, a)]) + \mathbb{E}_{\rho_0} \mathbb{E}_{\pi_e} \left[ \frac{\pi^2(a'|x)}{\pi^2_0(a'|x)} \hat{Q}(x, a'; \beta)^2 \right] - (\hat{V}^{\pi_e}(x; \beta))^2 + \mathbb{E}_{\rho_0} \left[ \frac{\pi^2_e(a|x)}{\pi^2_0(a|x)} r(x, a)^2 \right] - r_{\pi_e(x)}^2
\]

(28)

For the derivation of (ii), consider the parts of the variance objective function that is dependent of \( \beta \), i.e.,

\[
\mathbb{E}_{\rho_{\pi_e}} \left[ \omega(x, a) \left( \frac{\omega(x, a') \hat{Q}(x, a'; \beta)^2}{\pi_{\pi_e}(a|x)} - \hat{V}^{\pi_e}(x; \beta)^2 - 2r(x, a)(\omega(x, a) \hat{Q}(x, a; \beta) - \hat{V}^{\pi_e}(x; \beta)) \right) \right]
\]

From the definition of \(|A| \times |A|\) symmetric positive-semidefinite matrix \( \Omega_{\pi_e}(x) = \text{diag}[1/\pi_0(a|x)]_{a \in A} - ee^T \) and \(|A| \times 1\) vector \( q_{\beta}(x, a, r) = D_{\pi_e}(x) \hat{Q}(x; \beta) - I(a)r \), one can easily sees that

\[
q_{\beta}(x, a, r) \Omega_{\pi_e}(x) q_{\beta}(x, a, r) = q_{\beta}(x, a, r)^T \cdot \text{diag}[1/\pi_0(a|x)]_{a \in A} \cdot q_{\beta}(x, a, r) - (q_{\beta}(x, a, r)^T e)^2,
\]

for which

\[
(q_{\beta}(x, a, r)^T e)^2 = \hat{V}^{\pi_e}(x; \beta)^2 + r(x, a)^2 - 2r(x, a) \cdot \hat{V}^{\pi_e}(x; \beta)
\]

(29)

and

\[
q_{\beta}(x, a, r)^T \cdot \text{diag}[1/\pi_0(a|x)]_{a \in A} \cdot q_{\beta}(x, a, r) = \mathbb{E}_{\pi_e} \left[ \omega(x, a') \hat{Q}(x, a'; \beta)^2 \right] - 2r(x, a) \omega(x, a) \hat{Q}(x, a; \beta) + \frac{r(x, a)^2}{\pi_0^2(a|x)}.
\]

(30)

By collecting the above terms in (28), (29), and (30), one thus can show that

\[
n\mathbb{V}_{\rho_{\pi_e}}(\hat{\rho}_{\text{DR}}) = \mathbb{E}_{\rho_{\pi_e}} \left[ \omega(x, a) \cdot q_{\beta}(x, a, r)^T \Omega_{\pi_e}(x) q_{\beta}(x, a, r) \right] + \mathbb{V}_{\rho_0}(\mathbb{E}_{\pi_e}[r(x, a)]) + \mathbb{E}_{\rho_{\pi_e}} \left[ \mathbb{E}_{\pi_0}(r(x, a)^2) \right] - \mathbb{E}_{\rho_{\pi_e}} \left[ \left( 1 + \omega(x, a) - \frac{1}{\pi_0^2(a|x)} \right) \omega(x, a) r(x, a)^2 \right].
\]

Therefore by recalling the definition of \( C \), the derivation of ii) immediately follows.

The proof of this theorem is completed by combining both parts of the above derivations. \( \square \)

**D.3. Proof of Proposition 6**

**Proposition 6.** For any state \( x \in \mathcal{X} \), the symmetric matrix \( \Omega_{\pi_0}(x) \) is positive semi-definite (PSD).

**Proof.** For any non-zero vector \( v \in \mathbb{R}^{|\mathcal{A}|} \), consider the quadratic form at a given arbitrary state \( x \in \mathcal{X} \):

\[
v^T (\text{diag}(1/\pi_0(a|x))_{a' \in \mathcal{A}}) - ee^T) v = \sum_{a \in \mathcal{A}} v_a^2 / \pi_0(a|x) - (\sum_{a \in \mathcal{A}} v_a)^2.
\]
Since $\sum_{a \in A} \pi_a(a|s) = 1$, notice that
\[ \sum_{a \in A} v_a^2/\pi_b(a|x) = \sum_{a \in A} \sqrt{\pi_b(a|x)}^2 \sum_{a \in A} (v_a/\sqrt{\pi_b(a|x)})^2. \]

Moreover, one deduces that
\[ \sum_{a \in A} \sqrt{\pi_b(a|x)}^2 \sum_{a \in A} (v_a/\sqrt{\pi_b(a|x)})^2 \geq (\sum_{a \in A} v_a)^2 \]
from Cauchy-Schwarz inequality. This implies that $\text{diag}(\{1/\pi_b(a'|x)\}_{a' \in A}) - ee^\top$ is a positive semi-definite matrix, which completes the proof. \qed
We utilize the variance formulation of the DR estimator in the contextual bandit setting and construct the variance of the DR estimator for RL using mathematical induction.

E.1.1. BASE CASE, WHEN $T = 2$

Consider the case when $T = 2$, where the variance of the DR estimator has the form:

$$ n \cdot \mathbb{V}_{P_t^x} [R_{0:1}] := \mathbb{V}_{P_t^x} \left[ \omega_0 \left[ r_0 + \gamma \omega_1 \left( r_1 - \hat{Q}^{x_1} (x_1, a_1; \beta) \right) + \hat{V}^{x_1} (x_1; \beta) \right] - \hat{Q}^{x_1} (x_0, a_0; \beta) \right] + \hat{V}^{x_1} (x_0; \beta) \right]. $$

Using the total variance decomposition rule

$$ \mathbb{V}_{P_t^x} [\cdot] = \mathbb{E}_{F_0} \left[ \mathbb{V}_{F_1} [\cdot | F_0] \right] + \mathbb{V}_{F_0} \left[ \mathbb{E}_{F_1} [\cdot | F_0] \right], $$

this formulation can be immediately re-written as

$$ \mathbb{V}_{P_t^x} [R_{0:1}] = \mathbb{E}_{F_0} \left[ \mathbb{E}_{F_1} \left[ \omega_0 \left[ r_0 + \gamma \omega_1 \left( r_1 - \hat{Q}^{x_1} (x_1, a_1; \beta) \right) + \hat{V}^{x_1} (x_1; \beta) \right] - \hat{Q}^{x_1} (x_0, a_0; \beta) \right] + \hat{V}^{x_1} (x_0; \beta) | F_0 \right] + \mathbb{V}_{F_0} \left[ \mathbb{E}_{F_1} \left[ \omega_0 \left[ r_0 + \gamma \omega_1 \left( r_1 - \hat{Q}^{x_1} (x_1, a_1; \beta) \right) + \hat{V}^{x_1} (x_1; \beta) \right] - \hat{Q}^{x_1} (x_0, a_0; \beta) \right] + \hat{V}^{x_1} (x_0; \beta) | F_0 \right]. $$

Now consider the second term in (31). By the linearity of expectation operator $\mathbb{E}_{F_1}$ and by the following property:

$$ \mathbb{E}_{F_1} \left[ \hat{Q}^{x_1} (x_1, a_1; \beta) \right] = \mathbb{E}_{x_1} \left[ \mathbb{E}_{a_1 \sim \pi_0} \left[ \omega_1 \hat{Q}^{x_1} (x_1, a_1; \beta) \right] \right] = \mathbb{E}_{x_1} \left[ \hat{V}^{x_1} (x_1; \beta) \right], $$

this term can be re-written as follows:

$$ \mathbb{V}_{F_0} \left[ \omega_0 \left[ r_0 + \gamma \omega_1 \left( r_1 - \hat{Q}^{x_1} (x_1, a_1; \beta) \right) + \hat{V}^{x_1} (x_1; \beta) \right] - \hat{Q}^{x_1} (x_0, a_0; \beta) \right] + \hat{V}^{x_1} (x_0; \beta) \right] = \mathbb{E}_{F_0} \left[ \omega_0 \left[ r_0 + \gamma \mathbb{E}_{F_1} \left[ \omega_1 \left( r_1 - \hat{Q}^{x_1} (x_1, a_1; \beta) \right) + \hat{V}^{x_1} (x_1; \beta) \right] - \hat{Q}^{x_1} (x_0, a_0; \beta) \right] + \hat{V}^{x_1} (x_0; \beta) \right]. \quad (32) $$

For the first term in (31), by the translational invariance of the variance operator $\mathbb{V}_{F_1}$, one has the following simplification:

$$ \mathbb{E}_{F_0} \left[ \omega_0 \left[ r_0 + \gamma \omega_1 \left( r_1 - \hat{Q}^{x_1} (x_1, a_1; \beta) \right) + \hat{V}^{x_1} (x_1; \beta) \right] - \hat{Q}^{x_1} (x_0, a_0; \beta) \right] + \hat{V}^{x_1} (x_0; \beta) \right] = \mathbb{E}_{F_0} \left[ \gamma^2 \omega_0^2 \cdot \mathbb{E}_{F_1} \left[ \omega_1 \left( r_1 - \hat{Q}^{x_1} (x_1, a_1; \beta) \right) + \hat{V}^{x_1} (x_1; \beta) \right] + \hat{V}^{x_1} (x_0; \beta) \right]. \quad (33) $$

Therefore, by combining the above results, one shows the variance formulation in Theorem 2 for the base case when $T = 2$.

E.1.2. GENERAL CASE BY MATHEMATICAL INDUCTION

By the principle of mathematical induction, at step $k = j$ for any arbitrary time step $j$, the variance of the $T - j$ step truncated return variable is assumed to have the following form:

$$ n \cdot \mathbb{V}_{P_t^x} [R_{j:T-1}] = \sum_{i=j}^{T-1} \mathbb{E}_{F_{j+1:i}} \left[ \gamma^{2(t-j)} \omega_{j:i}^2 \mathbb{V}_{F_0} \left[ \omega_i \left[ \mathbb{E}_{F_{i:T-1}} [R_{i:T-1}] - \hat{Q}^{x_1} (x_1, a_1; \beta) \right] + \hat{V}^{x_1} (x_1; \beta) \right] \right], \quad (34) $$
Again, by the total variance decomposition rule

$$\nu_{\epsilon_k}^{\pi^*} \left[ R_{j-1:T-1} \right] = \nu_{\epsilon_k}^{\pi^*} \left[ \sum_{t=j-1}^{T-1} \gamma^{t-j+1} \omega_j - \sum_{t=j-1}^{T-1} \gamma^{t-j+1} \left( \omega_j \cdot \hat{Q}^\pi (x_t, a_t; \beta) - \bar{\omega}_{j-1} \cdot \hat{V}^\pi (x_t; \beta) \right) \right]$$

$$= \nu_{\epsilon_k}^{\pi^*} \left[ \omega_j \cdot r_{j-1} + \gamma \hat{R}_{j:T-1} - \hat{Q}^\pi (x_{j-1}, a_{j-1}; \beta) \right] + \hat{V}^\pi (x_{j-1}; \beta) .$$

Again, by the total variance decomposition rule

$$\nu_{\epsilon_k}^{\pi^*} \left[ \cdot \right] = \mathbb{E}_{X_j \sim 1} \left[ \mathbb{E}_{X_j \sim 1} \left[ \cdot \mid X_j \sim 1 \right] \right] + \mathbb{E}_{X_j \sim 1} \left[ \mathbb{E}_{X_j \sim 1} \left[ \cdot \mid X_j \sim 1 \right] \right] ,$$

this variance formulation can be re-written using the following chain of equalities:

$$\nu_{\epsilon_k}^{\pi^*} \left[ R_{j-1:T-1} \right]$$

$$= \mathbb{E}_{X_j \sim 1} \left[ \mathbb{E}_{X_j \sim 1} \left[ r_{j-1} + \gamma R_{j:T-1} - \hat{Q}^\pi (x_{j-1}, a_{j-1}; \beta) \right] + \hat{V}^\pi (x_{j-1}; \beta) \mid X_j \sim 1 \right]$$

$$+ \mathbb{E}_{X_j \sim 1} \left[ \mathbb{E}_{X_j \sim 1} \left[ r_{j-1} + \gamma R_{j:T-1} - \hat{Q}^\pi (x_{j-1}, a_{j-1}; \beta) \right] + \hat{V}^\pi (x_{j-1}; \beta) \mid X_j \sim 1 \right]$$

$$= \mathbb{E}_{X_j \sim 1} \left[ \omega_j \cdot r_{j-1} + \gamma \mathbb{E}_{X_j \sim 1} \left[ R_{j:T-1} \mid X_j \sim 1 \right] - \hat{Q}^\pi (x_{j-1}, a_{j-1}; \beta) \right] + \hat{V}^\pi (x_{j-1}; \beta)$$

$$+ \mathbb{E}_{X_j \sim 1} \left[ \omega_j \cdot r_{j-1} \left[ \omega_j \cdot R_{j:T-1} \mid X_j \sim 1 \right] - \hat{Q}^\pi (x_{j-1}, a_{j-1}; \beta) \right] + \hat{V}^\pi (x_{j-1}; \beta) + \mathbb{E}_{X_j \sim 1} \left[ \gamma^2 \omega_j \cdot \mathbb{E}_{X_j \sim 1} \left[ R_{j:T-1} \mid X_j \sim 1 \right] \right] .$$

The second equality is due to the linearity of the expectation operator $\mathbb{E}_{X_j \sim 1}$, and the square-scaling and the translational invariance property of the variance operator, i.e.,

$$\mathbb{E}_{X_j \sim 1} \left[ r_{j-1} + \gamma R_{j:T-1} - \hat{Q}^\pi (x_{j-1}, a_{j-1}; \beta) \right] + \hat{V}^\pi (x_{j-1}; \beta) = \gamma^2 \omega_j \cdot \mathbb{E}_{X_j \sim 1} \left[ R_{j:T-1} \mid X_j \sim 1 \right] .$$

The third equality is due to the fact that

$$\mathbb{E}_{X_j \sim 1} \left[ R_{j-1:T-1} \right] = r_{j-1} + \gamma \mathbb{E}_{X_j \sim 1} \left[ R_{j:T-1} \mid X_j \sim 1 \right]$$

The fourth equality is based on the translational invariance property of $\nu_{\epsilon_k}^{\pi^*}$, and the assumption of mathematical induction, and finally the fifth equality is merely based on re-formulating the summation and noticing that $\omega_{t_1:t_2} = 1$ if $t_1 > t_2$. Therefore, the variance of the truncated return random variable holds in the case when $k = j - 1$.

Based on the argument of induction, one concludes that for any $k \in \{0, \ldots, T-1\}$, the variance of the $T-k$ truncated return in (34) holds.

To complete the proof of Theorem 2, one needs to show that for any $t \geq 0$,

$$\nu_{\epsilon_t} \left[ \omega_t \left[ \mathbb{E}_{X_{t+1:T-1} \sim 1} \left[ \hat{R}_{t:T-1} - \hat{Q}^\pi (x_t, a_t; \beta) \right] \right] \right] = \nu_{\epsilon_t} \left[ \omega_t \left[ \hat{R}_{t:T-1} - \hat{Q}^\pi (x_t, a_t; \beta) \right] + C_t .$$
By expanding the variance expression on the left side, by utilizing the completion of squares, and by noticing that \( \bar{R}_{t:T-1} \) is an unbiased sampling-based estimate of \( \mathcal{F}_{t:T-1} [\bar{R}_{t:T-1}] \), one has the following chain of equalities:

\[
\forall F, \quad \mathbb{E}_F \left[ \omega_t \left[ \mathbb{E}_{F_{t+1:T-1}} [\bar{R}_{t:T-1}] - \hat{Q}^\pi_e (x_t, a_t; \beta) \right] \right] \\
= \mathbb{E}_F \left[ \left( \omega_t \left[ \mathbb{E}_{F_{t+1:T-1}} [\bar{R}_{t:T-1}] - \hat{Q}^\pi_e (x_t, a_t; \beta) \right] \right)^2 \right] - \left( \mathbb{E}_F \left[ \omega_t \left[ \mathbb{E}_{F_{t+1:T-1}} [\bar{R}_{t:T-1}] - \hat{Q}^\pi_e (x_t, a_t; \beta) \right] \right) \right)^2 \\
= \mathbb{E}_F \left[ \left( \omega_t \left[ \mathbb{E}_{F_{t+1:T-1}} [\bar{R}_{t:T-1}] - \hat{Q}^\pi_e (x_t, a_t; \beta) \right] \right)^2 \right] - \left( \mathbb{E}_F \left[ \omega_t \left[ \bar{R}_{t:T-1} - \hat{Q}^\pi_e (x_t, a_t; \beta) \right] \right) \right)^2 \\
= \mathbb{E}_F \left[ \left( \omega_t \left[ \bar{R}_{t:T-1} - \hat{Q}^\pi_e (x_t, a_t; \beta) \right] \right)^2 \right] - 2\omega_t^2 \left( \bar{R}_{t:T-1} - \hat{Q}^\pi_e (x_t, a_t; \beta) \right) \left( \bar{R}_{t:T-1} - \mathbb{E}_{F_{t+1:T-1}} [\bar{R}_{t:T-1}] \right) \\
+ \mathbb{E}_F \left[ \omega_t^2 \left( \bar{R}_{t:T-1} - \mathbb{E}_{F_{t+1:T-1}} [\bar{R}_{t:T-1}] \right)^2 \right] - \left( \mathbb{E}_F \left[ \omega_t \left[ \bar{R}_{t:T-1} - \hat{Q}^\pi_e (x_t, a_t; \beta) \right] \right) \right)^2 \\
= \mathbb{E}_F \left[ \left( \omega_t \left[ \bar{R}_{t:T-1} - \hat{Q}^\pi_e (x_t, a_t; \beta) \right] \right)^2 \right] - \left( \mathbb{E}_F \left[ \omega_t \left[ \bar{R}_{t:T-1} - \hat{Q}^\pi_e (x_t, a_t; \beta) \right] \right) \right)^2 \\
+ \mathbb{E}_F \left[ \omega_t^2 \left( \bar{R}_{t:T-1} - \mathbb{E}_{F_{t+1:T-1}} [\bar{R}_{t:T-1}] \right)^2 \right] - 2\omega_t^2 \bar{R}_{t:T-1} \left( \bar{R}_{t:T-1} - \mathbb{E}_{F_{t+1:T-1}} [\bar{R}_{t:T-1}] \right) \\
+ \underbrace{\mathbb{E}_F \left[ \omega_t^2 \left( \bar{R}_{t:T-1} - \mathbb{E}_{F_{t+1:T-1}} [\bar{R}_{t:T-1}] \right)^2 \right]}_{C_t} - 2\omega_t^2 \bar{R}_{t:T-1} \left( \bar{R}_{t:T-1} - \mathbb{E}_{F_{t+1:T-1}} [\bar{R}_{t:T-1}] \right) \\
= \mathbb{E}_F \left[ \omega_t \left[ \bar{R}_{t:T-1} - \hat{Q}^\pi_e (x_t, a_t; \beta) \right] \right] + C_t.
\]

Therefore, by combining all the previous arguments, one proves the claim of Theorem 2.
F. Proofs of Section 4.3

Proposition 7. Suppose the importance weight \( \omega_{t_1, t_2} \) is bounded at any time interval \((t_1, t_2)\), such that \( 0 \leq t_1 \leq t_2 \leq T - 1 \). Then the MRDR estimator is strongly consistent, i.e., \( \lim_{n \to \infty} \rho_{n, n, \beta}^{MRDR} (\beta^*) = \rho^{\pi_e} \) almost surely.

**Proof.** Consider the general DR estimator in (7) for any given model parameter \( \beta \). In this proof we use the notation \( \hat{\rho}_{n, n}^{\pi_e} (\beta) \) to emphasize that the sample-size of constructing the DR estimator is \( n \). By Lemma 12 of (Thomas & Brunskill, 2016), one first can show that

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \sum_{t=0}^{T-1} \gamma^t \omega_{0,t}^i \pi_e (x_t^i, a_t^i; \beta) = \rho^{\pi_e} \text{ almost surely.}
\]

Now consider the two terms that constitute to the control variate:

\[
X_n = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=0}^{T-1} \gamma^t \omega_{0,t}^i \hat{Q}^{\pi_e} (x_t^i, a_t^i; \beta), \quad Y_n = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=0}^{T-1} \gamma^t \omega_{0,t-1}^i \hat{V}^{\pi_e} (x_t^i; \beta).
\]

Again by Lemma 12 in (Thomas & Brunskill, 2016), one can show that

\[
\lim_{n \to \infty} X_n = \mathbb{E} \left[ \sum_{t=0}^{T-1} \gamma^t \hat{Q}^{\pi_e} (x_t, a_t; \beta) \mid x_0, \pi_e \right] \text{ almost surely.}
\]

On the other hand, by Lemma 13 in (Thomas & Brunskill, 2016), one can also show that

\[
\lim_{n \to \infty} Y_n = \mathbb{E} \left[ \sum_{t=0}^{T-1} \gamma^t \hat{V}^{\pi_e} (x_t; \beta) \mid x_0, \pi_e \right] \text{ almost surely.}
\]

This further implies that with probability 1, the following chain of equalities hold:

\[
\lim_{n \to \infty} Y_n = \mathbb{E} \left[ \sum_{t=0}^{T-1} \gamma^t \mathbb{E}_{\pi_e} \left[ \hat{Q}^{\pi_e} (x_{t+1}, a_t; \beta) \mid x_0, \pi_e \right] \right]
\]

\[
\mathbb{E} \left[ \sum_{t=0}^{T-1} \gamma^t \hat{Q}^{\pi_e} (x_t; a_t; \beta) \mid x_0, \pi_e \right] .
\]

Therefore, by combining the above results, for any DR estimators we have that \( \lim_{n \to \infty} \hat{\rho}_{n, \beta}^{\pi_e} = \rho^{\pi_e} \) almost surely. This convergence property is point-wise and it holds on any arbitrary choices of model parameters \( \beta \).

Utilizing the above result, and using both the triangular inequality and the Slutsky’s theorem, one can further show that

\[
\lim_{n \to \infty} |\hat{\rho}_{n, \beta}^{MRDR, (\beta^*)} - \rho^{\pi_e}| \leq \lim_{n \to \infty} |\hat{\rho}_{n, \beta}^{MRDR, (\beta^*)} - \rho_{n, \beta}^{MRDR, (\beta^*)}| + \lim_{n \to \infty} |\rho_{n, \beta}^{MRDR, (\beta^*)} - \rho^{\pi_e}| = 0 + 0,
\]

the convergence result is by using Slutsky’s theorem and the fact that \( \beta_n \to \beta^* \), the second convergence result is based on the above consistency result. Therefore, the proof of the main claim in this lemma is completed. \(\square\)
G. Details of Experimental Setup

G.1. Contextual Bandit

Following the setting in Dudık et al. (2011), we evaluate the policy evaluation algorithms in a classification setting. The data comes from standard datasets from UCI repositories. Table 11 shows the statistics of these data-sets. The standard procedure of turning a classification problem into a contextual bandit problem is described as follows.

Classification to Contextual Bandits. Consider a multi-label classification dataset \((x_i, y_i)_{i=1}^{m}\) where \(x_i \in \mathbb{R}^d\) is the data, \(y_i \in \{1, \ldots, l\}\) is its class, and \(m\) and \(l\) are the number of data points and classes respectively. A classification algorithm will assign a class to each data point \(f : \mathbb{R}^d \rightarrow \{1, \ldots, l\}\). If we treat each data point as a context then a deterministic (stochastic) classification algorithm can be viewed as a deterministic (stochastic) policy that maps contexts to actions.

Classification Policy. Assume we have a classification model given by \(f : \mathcal{X} \rightarrow \{1, \ldots, l\}\). In the contextual bandit setting, a policy \(\pi\) can be constructed from this model when one views the feature \(x \in \mathcal{X}\) as the context and the predicted label \(\hat{y}\) as the action chosen by this policy. During inference, one obtains label prediction \(\{\hat{y}_i\}_{i=1}^{N_{\text{test}}}\) from the evaluation data-set \(\{(x_i, y_i)\}_{i=1}^{N_{\text{test}}}\), and one can construct the set of actions that is executed by the evaluation policy as \(a_i = \hat{y}_i\) for \(i \in \{1, \ldots, N_{\text{test}}\}\). Correspondingly, at context \(x_i\) the agent receives a unit reward, i.e., \(r_i = 1\), if it accurately predicts the label (that is, \(a_i = \hat{y}_i = \hat{y}_i\)), and it receives no reward, i.e., \(r_i = 0\), otherwise. In this case, one can generate triplets of \(\{(x_i, a_i, r_i)\}_{i=1}^{N_{\text{test}}}\) using the evaluation results of the classification model, where the reward of a policy would the accuracy of the corresponding classification algorithm. (Dudık et al., 2011) used the same setup in their evaluation procedure. Note that applying the evaluation policy to a classification dataset can lead to many bandit datasets because the policy is stochastic and can choose different actions in accordance to the same context. This way, we generate \(N = 500\) different bandit test datasets from a given classification dataset. The accuracy and the significance tests are reported after taking average over these \(N\) runs.

Given the above transformation our experiment is set up as follows:

- The dataset is randomly split into training (70\%) for the models involved in estimators and test set (30\%) for generating samples for the importance sampling part of the estimators.
- On the training set we run logistic regression with ordinary least squares fitting to build a classifier.
- The accuracy of the above classifier on the whole data is measured and serves as ground truth value (\(\rho^{\pi_e}\)).
- We soften the baseline deterministic policy (a.k.a. learned classifier) to get a stochastic policy regarded as the evaluation policy (\(\pi_e\)).
- We further soften the base policy to form the behavior policy (\(\pi_b\)).
- The behavior policy is applied to training dataset and the output bandit data is used to train the model part of estimators.
- The behavior policy is applied to the test data to get behavior samples used in the importance sampling part of the estimators.

Results. Table 12 to Table 20 show all the results of the contextual bandit experiment. Besides all the OPE algorithms mentioned in the main paper, we also include the results from the MRDR0 estimator, which finds the model parameters by naively minimizing the empirical second order moment of the DR estimator. Theoretically the methodology of MRDR0 is identical to that of MRDR. However in practice, it turns out MRDR outperforms MRDR0 significantly. Theoretical comparisons of these methods are remained as future work. In most experiments, the proposed MRDR estimator outperforms its alternatives (statistically) significantly. Especially for medium to large datasets the performance improvement is more significant. We believe that data limitation is the main cause of impeding the performance.
### Table 12. Glass

<table>
<thead>
<tr>
<th>Behavior Policy</th>
<th>DM0</th>
<th>DM</th>
<th>IS</th>
<th>DR</th>
<th>MRDR</th>
<th>DR0</th>
<th>MRDR0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friendly I</td>
<td>0.4303</td>
<td>0.4153</td>
<td>0.0643</td>
<td>0.0557</td>
<td><strong>0.0485</strong></td>
<td>0.0562</td>
<td>0.1085</td>
</tr>
<tr>
<td>Friendly II</td>
<td>0.4709</td>
<td>0.4415</td>
<td>0.0937</td>
<td>0.083</td>
<td><strong>0.0793</strong></td>
<td>0.0848</td>
<td>0.1584</td>
</tr>
<tr>
<td>Neutral</td>
<td>0.5607</td>
<td>0.5037</td>
<td>0.2018</td>
<td>0.1928</td>
<td><strong>0.1739</strong></td>
<td>0.1987</td>
<td>0.3039</td>
</tr>
<tr>
<td>Adversary I</td>
<td>0.5533</td>
<td>0.4782</td>
<td>0.2238</td>
<td>0.199</td>
<td>0.2082</td>
<td>0.2125</td>
<td>0.384</td>
</tr>
<tr>
<td>Adversary II</td>
<td>0.5624</td>
<td>0.4947</td>
<td>0.2808</td>
<td><strong>0.2579</strong></td>
<td>0.2845</td>
<td>0.2711</td>
<td>0.4664</td>
</tr>
</tbody>
</table>

### Table 13. Ecoli

<table>
<thead>
<tr>
<th>Behavior Policy</th>
<th>DM0</th>
<th>DM</th>
<th>IS</th>
<th>DR</th>
<th>MRDR</th>
<th>DR0</th>
<th>MRDR0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friendly I</td>
<td>0.4275</td>
<td>0.4007</td>
<td>0.0828</td>
<td>0.0705</td>
<td><strong>0.0683</strong></td>
<td>0.071</td>
<td>0.13</td>
</tr>
<tr>
<td>Friendly II</td>
<td>0.4958</td>
<td>0.4367</td>
<td>0.1114</td>
<td>0.0872</td>
<td><strong>0.0869</strong></td>
<td>0.0902</td>
<td>0.1891</td>
</tr>
<tr>
<td>Neutral</td>
<td>0.7195</td>
<td>0.5601</td>
<td>0.2258</td>
<td><strong>0.1768</strong></td>
<td>0.2112</td>
<td>0.208</td>
<td>0.3938</td>
</tr>
<tr>
<td>Adversary I</td>
<td>0.7335</td>
<td>0.5819</td>
<td>0.247</td>
<td><strong>0.1975</strong></td>
<td>0.4165</td>
<td>0.2317</td>
<td>0.5386</td>
</tr>
<tr>
<td>Adversary II</td>
<td>0.7663</td>
<td>0.6383</td>
<td>0.3099</td>
<td>0.2556</td>
<td><strong>0.2387</strong></td>
<td>0.2983</td>
<td>0.5055</td>
</tr>
</tbody>
</table>

### Table 14. Vehicle

<table>
<thead>
<tr>
<th>Behavior Policy</th>
<th>DM0</th>
<th>DM</th>
<th>IS</th>
<th>DR</th>
<th>MRDR</th>
<th>DR0</th>
<th>MRDR0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friendly I</td>
<td>0.3593</td>
<td>0.3273</td>
<td>0.0347</td>
<td>0.0217</td>
<td><strong>0.0202</strong></td>
<td>0.0224</td>
<td>0.0706</td>
</tr>
<tr>
<td>Friendly II</td>
<td>0.4238</td>
<td>0.3499</td>
<td>0.0517</td>
<td>0.0331</td>
<td><strong>0.0318</strong></td>
<td>0.0356</td>
<td>0.1013</td>
</tr>
<tr>
<td>Neutral</td>
<td>0.5771</td>
<td>0.4384</td>
<td>0.087</td>
<td>0.0604</td>
<td><strong>0.0549</strong></td>
<td>0.0722</td>
<td>0.1534</td>
</tr>
<tr>
<td>Adversary I</td>
<td>0.5703</td>
<td>0.405</td>
<td>0.0937</td>
<td>0.0616</td>
<td><strong>0.0516</strong></td>
<td>0.0769</td>
<td>0.1709</td>
</tr>
<tr>
<td>Adversary II</td>
<td>0.6023</td>
<td>0.405</td>
<td>0.1131</td>
<td>0.0712</td>
<td><strong>0.0602</strong></td>
<td>0.0952</td>
<td>0.2142</td>
</tr>
</tbody>
</table>

### Table 15. Yeast

<table>
<thead>
<tr>
<th>Behavior Policy</th>
<th>DM0</th>
<th>DM</th>
<th>IS</th>
<th>DR</th>
<th>MRDR</th>
<th>DR0</th>
<th>MRDR0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friendly I</td>
<td>0.2989</td>
<td>0.2772</td>
<td>0.022</td>
<td>0.0165</td>
<td><strong>0.0148</strong></td>
<td>0.0169</td>
<td>0.041</td>
</tr>
<tr>
<td>Friendly II</td>
<td>0.3563</td>
<td>0.3083</td>
<td>0.033</td>
<td>0.0256</td>
<td><strong>0.0231</strong></td>
<td>0.0268</td>
<td>0.057</td>
</tr>
<tr>
<td>Neutral</td>
<td>0.4996</td>
<td>0.3712</td>
<td>0.0907</td>
<td>0.077</td>
<td><strong>0.0727</strong></td>
<td>0.0867</td>
<td>0.1531</td>
</tr>
<tr>
<td>Adversary I</td>
<td>0.4948</td>
<td>0.3544</td>
<td>0.1165</td>
<td>0.0932</td>
<td><strong>0.0853</strong></td>
<td>0.1114</td>
<td>0.1917</td>
</tr>
<tr>
<td>Adversary II</td>
<td>0.5019</td>
<td>0.3235</td>
<td>0.1424</td>
<td><strong>0.1098</strong></td>
<td>0.1299</td>
<td>0.1365</td>
<td>0.2398</td>
</tr>
</tbody>
</table>

### Table 16. PageBlock

<table>
<thead>
<tr>
<th>Behavior Policy</th>
<th>DM0</th>
<th>DM</th>
<th>IS</th>
<th>DR</th>
<th>MRDR</th>
<th>DR0</th>
<th>MRDR0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friendly I</td>
<td>0.7918</td>
<td>0.7793</td>
<td>0.0151</td>
<td><strong>0.0139</strong></td>
<td>0.0152</td>
<td>0.014</td>
<td>0.0306</td>
</tr>
<tr>
<td>Friendly II</td>
<td>0.8188</td>
<td>0.7923</td>
<td>0.0224</td>
<td>0.0205</td>
<td><strong>0.0182</strong></td>
<td>0.0211</td>
<td>0.0358</td>
</tr>
<tr>
<td>Neutral</td>
<td>0.8311</td>
<td>0.795</td>
<td>0.0314</td>
<td>0.029</td>
<td><strong>0.0232</strong></td>
<td>0.03</td>
<td>0.0544</td>
</tr>
<tr>
<td>Adversary I</td>
<td>0.8272</td>
<td>0.7931</td>
<td>0.0316</td>
<td>0.0292</td>
<td><strong>0.0246</strong></td>
<td>0.0302</td>
<td>0.049</td>
</tr>
<tr>
<td>Adversary II</td>
<td>0.8465</td>
<td>0.8039</td>
<td>0.0378</td>
<td>0.0349</td>
<td><strong>0.031</strong></td>
<td>0.0368</td>
<td>0.0666</td>
</tr>
</tbody>
</table>
G.2. Reinforcement Learning

We hereby describe the reinforcement learning domains that are used in the experiments and provide all the results of the OPE experiments. Similar to the Bandit case, the accuracy and significance results are averaged over $N = 100$ runs with different randomly generated test behavioral trajectories.

The **ModelFail Domain**. The purpose of this domain is to simulate environments that are not known perfectly. The MDP consists of 4 states, however, the agent cannot distinguish between the states. The agent starts from the left most node (node 1); with action $a_1$ it goes to upper middle state, and with action $a_2$ it goes to lower middle state. From these two states with any action it transits to the terminating state (rightmost). If the transition occurs from the upper state, it receives reward 1 otherwise reward of $-1$ is granted. The horizon is $T = 2$, and the problem has a partially observable state (i.e., the agent only sees one state). The evaluation policy selects action $a_1$ and $a_2$ with probabilities 0.88 and 0.12 respectively everywhere. The behavior policy has the opposite behavior.

The **ModelWin Domain**. This environment is built to simulate the cases where the domain is known perfectly. As Fig. 1 shows, it consists of 3 states. The agent begins at state $s_1$ (the middle one). Using action $a_1$, it transits to $s_2$ with probability

---

### Table 17. OptDigits

<table>
<thead>
<tr>
<th>Behavior Policy</th>
<th>DM0</th>
<th>DM</th>
<th>IS</th>
<th>DR</th>
<th>MRDR</th>
<th>DR0</th>
<th>MRDR0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friendly I</td>
<td>0.387</td>
<td>0.3596</td>
<td>0.0142</td>
<td>0.0069</td>
<td><strong>0.0056</strong></td>
<td>0.0073</td>
<td>0.0337</td>
</tr>
<tr>
<td>Friendly II</td>
<td>0.4966</td>
<td>0.4263</td>
<td>0.0218</td>
<td>0.0115</td>
<td><strong>0.0087</strong></td>
<td>0.013</td>
<td>0.0436</td>
</tr>
<tr>
<td>Neutral</td>
<td>0.7787</td>
<td>0.5571</td>
<td>0.0632</td>
<td>0.0412</td>
<td><strong>0.0351</strong></td>
<td>0.0562</td>
<td>0.113</td>
</tr>
<tr>
<td>Adversary I</td>
<td>0.798</td>
<td>0.5698</td>
<td>0.0738</td>
<td>0.0494</td>
<td><strong>0.0473</strong></td>
<td>0.0672</td>
<td>0.1283</td>
</tr>
<tr>
<td>Adversary II</td>
<td>0.8232</td>
<td>0.5949</td>
<td>0.0862</td>
<td><strong>0.061</strong></td>
<td>0.0697</td>
<td>0.0812</td>
<td>0.137</td>
</tr>
</tbody>
</table>

### Table 18. SatImage

<table>
<thead>
<tr>
<th>Behavior Policy</th>
<th>DM0</th>
<th>DM</th>
<th>IS</th>
<th>DR</th>
<th>MRDR</th>
<th>DR0</th>
<th>MRDR0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friendly I</td>
<td>0.3212</td>
<td>0.2884</td>
<td>0.0128</td>
<td>0.0071</td>
<td><strong>0.0063</strong></td>
<td>0.0073</td>
<td>0.0248</td>
</tr>
<tr>
<td>Friendly II</td>
<td>0.4116</td>
<td>0.3328</td>
<td>0.0191</td>
<td>0.0107</td>
<td><strong>0.0087</strong></td>
<td>0.0119</td>
<td>0.0356</td>
</tr>
<tr>
<td>Neutral</td>
<td>0.5984</td>
<td>0.3848</td>
<td>0.0413</td>
<td>0.0246</td>
<td><strong>0.0186</strong></td>
<td>0.0335</td>
<td>0.0778</td>
</tr>
<tr>
<td>Adversary I</td>
<td>0.6243</td>
<td>0.3963</td>
<td>0.0459</td>
<td>0.027</td>
<td><strong>0.0195</strong></td>
<td>0.0383</td>
<td>0.0871</td>
</tr>
<tr>
<td>Adversary II</td>
<td>0.6585</td>
<td>0.4093</td>
<td>0.0591</td>
<td>0.0364</td>
<td><strong>0.0262</strong></td>
<td>0.0521</td>
<td>0.1094</td>
</tr>
</tbody>
</table>

### Table 19. PenDigits

<table>
<thead>
<tr>
<th>Behavior Policy</th>
<th>DM0</th>
<th>DM</th>
<th>IS</th>
<th>DR</th>
<th>MRDR</th>
<th>DR0</th>
<th>MRDR0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friendly I</td>
<td>0.43</td>
<td>0.4014</td>
<td>0.0103</td>
<td>0.0056</td>
<td><strong>0.0037</strong></td>
<td>0.0059</td>
<td>0.0202</td>
</tr>
<tr>
<td>Friendly II</td>
<td>0.5315</td>
<td>0.4628</td>
<td>0.0159</td>
<td>0.0092</td>
<td><strong>0.0056</strong></td>
<td>0.0194</td>
<td>0.0309</td>
</tr>
<tr>
<td>Neutral</td>
<td>0.7632</td>
<td>0.564</td>
<td>0.0450</td>
<td>0.0314</td>
<td><strong>0.0138</strong></td>
<td>0.0412</td>
<td>0.0882</td>
</tr>
<tr>
<td>Adversary I</td>
<td>0.7828</td>
<td>0.5861</td>
<td>0.0503</td>
<td>0.0366</td>
<td><strong>0.0172</strong></td>
<td>0.0472</td>
<td>0.0976</td>
</tr>
<tr>
<td>Adversary II</td>
<td>0.7915</td>
<td>0.5641</td>
<td>0.0646</td>
<td>0.0444</td>
<td><strong>0.0188</strong></td>
<td>0.0611</td>
<td>0.01281</td>
</tr>
</tbody>
</table>

### Table 20. Letter

<table>
<thead>
<tr>
<th>Behavior Policy</th>
<th>DM0</th>
<th>DM</th>
<th>IS</th>
<th>DR</th>
<th>MRDR</th>
<th>DR0</th>
<th>MRDR0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friendly I</td>
<td>0.4177</td>
<td>0.392</td>
<td>0.0074</td>
<td>0.0056</td>
<td><strong>0.0044</strong></td>
<td>0.0057</td>
<td>0.0132</td>
</tr>
<tr>
<td>Friendly II</td>
<td>0.4729</td>
<td>0.4146</td>
<td>0.0102</td>
<td>0.0077</td>
<td><strong>0.0054</strong></td>
<td>0.0083</td>
<td>0.0186</td>
</tr>
<tr>
<td>Neutral</td>
<td>0.6347</td>
<td>0.4713</td>
<td>0.0467</td>
<td>0.0363</td>
<td><strong>0.0315</strong></td>
<td>0.0456</td>
<td>0.0855</td>
</tr>
<tr>
<td>Adversary I</td>
<td>0.6379</td>
<td>0.46</td>
<td>0.0587</td>
<td>0.0455</td>
<td><strong>0.0385</strong></td>
<td>0.0575</td>
<td>0.1082</td>
</tr>
<tr>
<td>Adversary II</td>
<td>0.6421</td>
<td>0.4728</td>
<td>0.0714</td>
<td>0.055</td>
<td>0.0481</td>
<td>0.0703</td>
<td>0.1318</td>
</tr>
</tbody>
</table>
with deterministic transitions. The horizon is set to $T = 20$. The evaluation policy at $s_1$ will take action $a_1$ with probability $0.73$ and $a_2$ with probability $0.27$. The behavior policy has the opposite behavior. In both states $s_2$ and $s_3$, the agent takes actions uniformly at random.

The Maze Domain. Depicted by Fig. 1, the maze is a $4 \times 4$ gridworld domain with 4 actions (up-right-down-left), and with deterministic transitions. The horizon is set to $T = 100$. The reward is always 0 except when the agent visits the red and green states. The behavior policy is a hand coded policy that moves quickly to position $(2, 4)$. From there by taking the downward action, it has a low probability of transitioning to right until it hits position $(4, 4)$. The evaluation policy is near-deterministic version of the behavior policy. Details can be found in Thomas & Brunskill (2016).

The Mountain Car Domain. A car is stuck between two hills in interval $[-0.7, 0.5]$ and the agent should move back and forth to gain enough momentum to reach the top of the right hill. The state space consists of position and velocity and the 3 discrete actions are accelerating forward and backward and stay-still. The position for the initial state is randomly chosen in $[-0.6, -0.4]$ with velocity equal to 0. The horizon is set to $T = 250$ with a reward of $-1$ per step. The optimal (deterministic) policy is learned in an on-policy fashion via SARSA (Sutton & Barto, 1998). Different from the standard linear state-action features used in other experiments, in order to increase feature resolution, the linear feature is discretized values of horizon.

The Cart Pole Domain. This environment is extracted from OpenAI Gym (Brockman et al., 2016) and has been used in some OPE applications (Hanna et al., 2016). The state consists of position, angle, speed, and angular speed, in which the position and speed are limited to $[-2.4, 2.4]$ and $[-41.8, 41.8]$, respectively. The possible actions are to move the cart to either left or right, and the reward is 1 for each step before the pendulum falls over. The horizon is set to $T = 250$. For the initial state vector, each element is chosen uniformly at random from interval $[-0.05, 0.05]$. The (deterministic) policy is learned in an off-policy fashion using Double Deep Q-Learning (Hasselt et al., 2016). It is a Double DQN Agent for that uses Neural Network to approximate $Q$ function and replay memory and target $Q$ networks. The $Q$ network is 3-layer neural network with the first 2 layers fully-connected and size of 24 with relu activation function plus a final fully-connected layer of size 2 with linear activation function. The loss is set to MSE and optimized via ADAM optimizer. The memory size for double DQN is set to 2000 and learning starts after step 1000. The $\epsilon$-greedy scheme for exploration starts with $\epsilon = 1$ and multiply it with 0.999 at each step until 0.01. The training episodes is bounded by 150.

The full simulation results of OPE in reinforcement learning are given below. Similar to the contextual bandit setting, we also include the results for the MRDR0 method, whose performance is much worse when compared to that of MRDR.

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>DM0</th>
<th>DM</th>
<th>IS</th>
<th>DR</th>
<th>MRDR</th>
<th>DR0</th>
<th>MRDR0</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>0.75394</td>
<td>0.07152</td>
<td>1.37601</td>
<td>0.18461</td>
<td>0.1698</td>
<td>1.16084</td>
<td>2.71942</td>
</tr>
<tr>
<td>64</td>
<td>0.75394</td>
<td>0.07152</td>
<td>1.07213</td>
<td>0.1314</td>
<td>0.11405</td>
<td>0.9046</td>
<td>2.11678</td>
</tr>
<tr>
<td>128</td>
<td>0.75394</td>
<td>0.07152</td>
<td>0.752</td>
<td>0.09901</td>
<td>0.08188</td>
<td>0.63571</td>
<td>1.47735</td>
</tr>
<tr>
<td>256</td>
<td>0.75394</td>
<td>0.07152</td>
<td>0.55955</td>
<td>0.06565</td>
<td>0.05527</td>
<td>0.47211</td>
<td>1.10451</td>
</tr>
<tr>
<td>512</td>
<td>0.75394</td>
<td>0.07152</td>
<td>0.39533</td>
<td>0.04756</td>
<td>0.03819</td>
<td>0.33391</td>
<td>0.77813</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>DM0</th>
<th>DM</th>
<th>IS</th>
<th>DR</th>
<th>MRDR</th>
<th>DR0</th>
<th>MRDR0</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>0.06839</td>
<td>0.06182</td>
<td>0.78452</td>
<td>1.55244</td>
<td>1.46778</td>
<td>1.51858</td>
<td>1.48273</td>
</tr>
<tr>
<td>64</td>
<td>0.06839</td>
<td>0.06182</td>
<td>1.03207</td>
<td>1.14699</td>
<td>0.98433</td>
<td>1.40758</td>
<td>1.46993</td>
</tr>
<tr>
<td>128</td>
<td>0.06839</td>
<td>0.06182</td>
<td>0.78452</td>
<td>1.55244</td>
<td>0.98433</td>
<td>1.40758</td>
<td>1.46993</td>
</tr>
<tr>
<td>256</td>
<td>0.06839</td>
<td>0.06182</td>
<td>0.55955</td>
<td>0.06565</td>
<td>0.05527</td>
<td>0.47211</td>
<td>1.10451</td>
</tr>
<tr>
<td>512</td>
<td>0.06839</td>
<td>0.06182</td>
<td>0.39533</td>
<td>0.04756</td>
<td>0.03819</td>
<td>0.33391</td>
<td>0.77813</td>
</tr>
</tbody>
</table>
### Table 23. 4 × 4 Maze

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>DM0</th>
<th>DM</th>
<th>IS</th>
<th>DR</th>
<th>MRDR</th>
<th>DR0</th>
<th>MRDR0</th>
</tr>
</thead>
<tbody>
<tr>
<td>128</td>
<td>1.77598</td>
<td>1.77598</td>
<td>6.68579</td>
<td>0.70465</td>
<td><strong>0.57042</strong></td>
<td>0.70969</td>
<td>3.19432</td>
</tr>
<tr>
<td>256</td>
<td>1.77598</td>
<td>1.77598</td>
<td>3.50346</td>
<td>0.69886</td>
<td><strong>0.58871</strong></td>
<td>0.70211</td>
<td>1.20821</td>
</tr>
<tr>
<td>512</td>
<td>1.77598</td>
<td>1.77598</td>
<td>2.64257</td>
<td>0.60124</td>
<td>0.58879</td>
<td>0.60338</td>
<td>1.33286</td>
</tr>
<tr>
<td>1024</td>
<td>1.77598</td>
<td>1.77598</td>
<td>1.45434</td>
<td>0.5201</td>
<td><strong>0.4666</strong></td>
<td>0.52148</td>
<td>0.84007</td>
</tr>
<tr>
<td>2048</td>
<td>1.77598</td>
<td>1.77598</td>
<td>0.89668</td>
<td>0.3932</td>
<td><strong>0.31274</strong></td>
<td>0.39425</td>
<td>0.52836</td>
</tr>
</tbody>
</table>

### Table 24. Mountain Car

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>DM0</th>
<th>DM</th>
<th>IS</th>
<th>DR</th>
<th>MRDR</th>
<th>DR0</th>
<th>MRDR0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1024</td>
<td>15.60588</td>
<td>10.24318</td>
<td>3.26843</td>
<td>3.03239</td>
<td>3.1359</td>
<td>9.16269</td>
<td>5.94688</td>
</tr>
<tr>
<td>2048</td>
<td>13.34371</td>
<td>10.91577</td>
<td>2.50591</td>
<td>2.75933</td>
<td><strong>2.17138</strong></td>
<td>8.25527</td>
<td>5.52493</td>
</tr>
</tbody>
</table>

### Table 25. Cart Pole

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>DM0</th>
<th>DM</th>
<th>IS</th>
<th>DR</th>
<th>MRDR</th>
<th>DR0</th>
<th>MRDR0</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>87.74511</td>
<td>86.81935</td>
<td>70.58151</td>
<td>12.13028</td>
<td>16.45905</td>
<td>10.84913</td>
<td>33.04329</td>
</tr>
<tr>
<td>64</td>
<td>87.59229</td>
<td>87.00547</td>
<td>75.86198</td>
<td>14.82026</td>
<td>14.16847</td>
<td>15.69192</td>
<td>36.13142</td>
</tr>
<tr>
<td>128</td>
<td>84.92029</td>
<td>84.40824</td>
<td>77.38233</td>
<td>18.55218</td>
<td><strong>15.38549</strong></td>
<td>19.15905</td>
<td>30.95533</td>
</tr>
<tr>
<td>256</td>
<td>83.74754</td>
<td>83.31824</td>
<td>64.75034</td>
<td>9.96921</td>
<td>8.36612</td>
<td>9.36373</td>
<td>5.86685</td>
</tr>
<tr>
<td>512</td>
<td>84.34155</td>
<td>84.09259</td>
<td>65.72996</td>
<td>6.88534</td>
<td><strong>4.60712</strong></td>
<td>6.9962</td>
<td>12.19454</td>
</tr>
</tbody>
</table>

More Robust Doubly Robust Off-policy Evaluation