DOUBLY ROBUST BIAS REDUCTION IN INFINITE HORIZON OFF-POLICY ESTIMATION

Ziyang Tang [*]
The University of Texas at Austin
ztang@cs.utexas.edu

Yihao Feng [*]
The University of Texas at Austin
yihao@cs.utexas.edu

Lihong Li
Google Research
lihong@google.com

Dengyong Zhou
Google Research
dennyzhou@google.com

Qiang Liu
The University of Texas at Austin
lqiang@cs.utexas.edu

ABSTRACT

Infinite horizon off-policy policy evaluation is a highly challenging task due to the excessively large variance of typical importance sampling (IS) estimators. Recently, Liu et al. (2018a) proposed an approach that substantially reduces the variance of infinite horizon off-policy evaluation by estimating the stationary density ratio, but at the cost of introducing biases due to errors in density ratio estimation. In this paper, we develop a bias-reduced augmentation of their method, which can take advantage of a learned value function to improve accuracy. Our method is *doubly robust* in that the bias vanishes when either the density ratio or value function estimation is perfect. In general, when either of them is accurate, the bias can also be reduced. Both theoretical and empirical results show that our method yields significant advantages over previous methods.

1 INTRODUCTION

A key problem in reinforcement learning (RL) (Sutton & Barto, 1998) is off-policy policy evaluation: given a fixed *target policy* of interest, estimating the average reward garnered by an agent that follows the policy, by only using data collected from different *behavior policies*. This problem is widely encountered in many real-life applications (e.g., Murphy et al., 2001; Li et al., 2011; Bottou et al., 2013; Thomas et al., 2017), where online experiments are expensive and high-quality simulators are difficult to build. It also serves as a key algorithmic component of off-policy policy optimization (e.g., Dudík et al., 2011; Jiang & Li, 2016; Thomas & Brunskill, 2016; Liu et al., 2019b).

There are two major families of approaches to policy evaluation. The first approach is to build a simulator that mimics the reward and next-state transitions of the real environment (e.g., Fonteneau et al., 2013). While straightforward, this approach strongly relies on the model assumptions in building the simulator, which may invalidate evaluation results. The second approach is to use importance sampling to correct the sampling bias in off-policy data, so that an (almost) unbiased estimator can be obtained (Liu, 2001; Strehl et al., 2010; Bottou et al., 2013). A major limitation, however, is that importance sampling can become inaccurate due to high variance. In particular, most existing IS-based estimators compute the weight as the product of the importance ratios of many steps in the trajectory, causing excessively high variance for problems with long or infinite horizon, yielding a *curse of horizon* (Liu et al., 2018a).

Recently, Liu et al. (2018a) proposes a new estimator for infinite-horizon off-policy evaluation, which presents significant advantages to standard importance sampling methods. Their method directly estimates the density ratio between the state stationary distributions of the target and behavior policies, instead of the trajectories, thus avoiding exponential blowup of variance in the horizon. While Liu et al.'s method shows much promise by significantly *reducing the variance*, in practice, it may suffer from *high bias* due to the error or model misspecficiation when estimating the density ratio function.

[*] The first two authors contributed equally to this work.

In this paper, we develop a *doubly robust* estimator for infinite horizon off-policy estimation, by integrating Liu et al.'s method with information from an additional value function estimation. This significantly reduces the bias of Liu et al.'s method once either the density ratio, or the value function estimation is accurate (hence doubly robust). Since Liu et al.'s method already promises low variance, our additional bias reduction allows us to achieve significantly better accuracy for practical problems.

Technically, our new *bias reduction* method provides a new angle of double robustness for off-policy evaluation, orthogonal to the existing literature of doubly robust policy evaluation that solely devotes to *variance reduction* (Jiang & Li, 2016; Thomas & Brunskill, 2016; Farajtabar et al., 2018), mostly based on the idea of control variates (e.g. Asmussen & Glynn, 2007). Our double robustness for bias reduction is significantly different, and instead yields an intriguing connection with the fundamental primal-dual relations between the density (ratio) functions and value functions (e.g., Bertsekas, 2000; Puterman, 2014). This new perspective can inspire more efficient algorithms for policy evaluation, and lead to unified frameworks for these types of double robustness in future work.

2 Background

Infinite Horizon Off-Policy Estimation Let $M = \langle \mathcal{S}, \mathcal{A}, r, \boldsymbol{T}, \mu_0 \rangle$ be a Markov decision process (MDP) with state space $\mathcal{S}$, action space $\mathcal{A}$, reward function r, transition probability function $\boldsymbol{T}$, and initial-state distribution μ_0. A policy π maps states to distributions over $\mathcal{A}$, with $\pi(a|s)$ being the probability of selecting a given s. The average discounted reward for π, with a given discount $\gamma \in (0,1)$ [1], is defined as

$$R^\pi := \lim_{T\to\infty} \mathbb{E}_{\tau\sim\pi}\left[\frac{\sum_{t=0}^T \gamma^t r_t}{\sum_{t=0}^T \gamma^t}\right],$$

where $\tau = \{s_t, a_t, r_t\}_{0\le t\le T}$ is a trajectory with states, actions, and rewards collected by following policy π in the MDP, starting from $s_0 \sim \mu_0$. Given a set of n trajectories, $\mathcal{D} = \{s_t^{(i)}, a_t^{(i)}, r_t^{(i)}\}_{1\le i\le n, 0\le t\le T}$, collected under a behavior policy $\pi_0(a|s)$, the off-policy evaluation problem aims to estimate the average discounted reward R^π for another target policy $\pi(a|s)$.

Estimation via Value Function The value function for policy π is defined as the expected accumulated discounted future rewards started from a certain state: $V^\pi(s) = \mathbb{E}_{\tau\sim\pi}[\sum_{t=0}^\infty \gamma^t r_t | s_0 = s]$. We use $r^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}[r(s,a)]$ to denote the average reward for state s given policy π. Under the definition, the value function can be seen as a fixed point of the Bellman equation:

$$V^\pi(s) = r^\pi(s) + \gamma \mathcal{P}^\pi V^\pi(s), \qquad \mathcal{P}^\pi V^\pi(s) := \mathbb{E}_{a\sim\pi(\cdot|s), s'\sim \boldsymbol{T}(\cdot|s,a)}[V^\pi(s')], \quad \forall s \in \mathcal{S}, \tag{1}$$

where $\mathcal{P}^\pi V(s)$ is the average of the next value function given the current state s and policy π; see Appendix A.1 for details.

The value function and the expected reward R^π is related in the following straightforward way

$$R^\pi = (1-\gamma)\mathbb{E}_{s\sim\mu_0}[V^\pi(s)], \tag{2}$$

where the expectation is with respect to the distribution $\mu_0(s)$ of the initial states s_0 at time t. Therefore, given an approximation $\widehat{V}$ of V^π, and samples $\mathcal{D}_0 := \{s_0^{(i)}\}_{1\le i\le n_0}$ drawn from $\mu_0(s)$, we can estimate R^π by

$$\widehat{R}^\pi_{\text{VAL}}[\widehat{V}] = \frac{(1-\gamma)}{n_0}\sum_{i=1}^{n_0} \widehat{V}(s_0^{(i)}).$$

Note that this estimator is off-policy in nature, since it requires no samples from the target policy π.

Estimation via State Density Function Denote $d_{\pi,t}(\cdot)$ as average visitation of s_t in time step t. The state density function, or the discounted average visitation, is defined as:

$$d_\pi(s) := \lim_{T\to\infty} \frac{\sum_{t=0}^T \gamma^t d_{\pi,t}(s)}{\sum_{t=0}^T \gamma^t} = (1-\gamma)\sum_{t=0}^\infty \gamma^t d_{\pi,t}(s),$$

[1] For average case when $\gamma = 1$, the definition of R^π is the same. However, the definition of value function is different. We will assume $\gamma < 1$ throughout our main paper for simplicity; for average case check appendix B for more details.

where $(1-\gamma)$ can be viewed as the normalization factor introduced by $\sum_{t=0}^{\infty}\gamma^t$.

Similar to Bellman equation for value function, the state density function can also be viewed as a fixed point to the following recursive equation (Liu et al., 2018a, Lemma 3):

$$d_\pi(s') = (1-\gamma)\mu_0(s') + \gamma\mathcal{T}^\pi d_\pi(s'), \quad \text{where} \quad \mathcal{T}^\pi d_\pi(s') := \sum_{s,a} \boldsymbol{T}(s'|s,a)\pi(a|s)d_\pi(s). \tag{3}$$

The operator $\mathcal{T}^\pi$ is an adjoint operator of $\mathcal{P}^\pi$ used in (1); see Appendix A.1 for a discussion.

If the density function d_π is known, it provides an alternative way for estimating the expected reward R^π, by noting that

$$R^\pi = \mathbb{E}_{s\sim d_\pi, a\sim\pi(\cdot|s)}[r(s,a)]. \tag{4}$$

We can see that both density function d_π and value function V^π can be used to estimate the expected reward R^π. We clarify the connection in detail in Appendix A.1.

Off-Policy State Visitation Importance Sampling Equation (4) can not be directly used for off-policy estimation, since it involves expectation under the behavior policy π. Liu et al. (2018a) addressed this problem by introducing a change of measures via importance sampling:

$$R^\pi = \mathbb{E}_{s\sim d_{\pi_0}, a\sim\pi_0(\cdot|s)}\left[w_{\pi/\pi_0}(s)\frac{\pi(a|s)}{\pi_0(a|s)}r(s,a)\right], \qquad \text{with} \qquad w_{\pi/\pi_0}(s) = \frac{d_\pi(s)}{d_{\pi_0}(s)}, \tag{5}$$

where $w_{\pi/\pi_0}(s)$ is the density ratio function of d_π and d_{π_0}.

Given an approximation $\widehat{w}$ of w_{π/π_0}, and samples $\mathcal{D} = \{s_t^{(i)}, a_t^{(i)}, r_t^{(i)}\}_{1\le i\le n, 0\le t\le T}$ collected from policy π_0, we can estimate R^π as:

$$\widehat{R}^\pi_{\text{SIS}}[\widehat{w}] = \frac{1}{Z}\sum_{i=1}^{n}\sum_{t=0}^{T}\gamma^t\widehat{w}(s_t^{(i)})\frac{\pi(a_i|s_i)}{\pi_0(a_i|s_i)}r_i, \qquad Z = \sum_{i=1}^{n}\sum_{t=0}^{T}\gamma^t\widehat{w}(s_t^{(i)})\frac{\pi(a_t^{(i)}|s_t^{(i)})}{\pi_0(a_t^{(i)}|s_t^{(i)})}, \tag{6}$$

where Z is the normalized constant of the importance weights.

3 Doubly Robust Estimator

Doubly robust estimator is first proposed into reinforcement learning community to solve contextual bandit problem by Dudík et al. (2011) as an estimator combining *inverse propensity score* (IPS) estimator and *direct method* (DM) estimator.

Jiang & Li (2016) introduce the idea of doubly robust estimator into off-policy evaluation in reinforcement learning. It incorporates an approximate value function as a control variate to reduce the variance of importance sampling estimator. Inspired by previous works, we propose a new doubly robust estimator based on our infinite horizon off-policy estimator $\widehat{R}^\pi_{\text{SIS}}$.

3.1 Doubly Robust Estimator for Infinite Horizon MDP

The value-based estimator $\widehat{R}^\pi_{\text{VAL}}[\widehat{V}]$ and density-ratio-based estimator $\widehat{R}^\pi_{\text{SIS}}[\widehat{w}]$ are expected to be accurate when $\widehat{V}$ and $\widehat{w}$ are accurate, respectively. Our goal is to combine their advantages, obtaining a doubly robust estimator that is accurate once either $\widehat{V}$ or $\widehat{w}$ or is accurate.

To simplify the problem, it is useful to examine the limit of infinite samples, with which $\widehat{R}^\pi_{\text{VAL}}[\widehat{V}]$ and $\widehat{R}^\pi_{\text{SIS}}[\widehat{w}]$ converge to the following limit of expectations:

$$R^\pi_{\text{SIS}}[\widehat{w}] := \lim_{n,T\to\infty}\widehat{R}^\pi_{\text{SIS}}[\widehat{w}] = \sum_s r^\pi(s)d_{\pi_0}(s)\widehat{w}(s), \tag{7}$$

$$R^\pi_{\text{VAL}}[\widehat{V}] := \lim_{n_0\to\infty}\widehat{R}^\pi_{\text{VAL}}[\widehat{V}] = (1-\gamma)\sum_s \widehat{V}(s)\mu_0(s). \tag{8}$$

Here and throughout this work, we assume $\widehat{V}$ and $\widehat{w}$ are fixed pre-defined approximations, and only consider the randomness from the data $\mathcal{D}$.

A first observation is that we expect to have $r^\pi \approx \widehat{V} - \gamma \mathcal{P}^\pi \widehat{V}$ by Bellman equation (1), when $\hat{V}$ approximates the true value V^π. Plugging this into $R^\pi_{\text{SIS}}[\widehat{w}]$ in Equation (7), we obtain the following "bridge estimator", which incorporates information from both $\widehat{w}$ and $\widehat{V}$:

$$R^\pi_{\text{bridge}}[\widehat{V}, \widehat{w}] = \sum_s \left(\widehat{V}(s) - \gamma \mathcal{P}^\pi \widehat{V}(s)\right) d_{\pi_0}(s) \widehat{w}(s), \tag{9}$$

where operator $\mathcal{P}^\pi$ is defined in Bellman equation (1). The corresponding empirical estimator is defined by

$$\widehat{R}^\pi_{\text{bridge}}[\widehat{V}, \widehat{w}] = \sum_{i=1}^{n} \sum_{t=0}^{T-1} \left(\frac{1}{Z_1} \gamma^t \widehat{w}(s_t^{(i)}) \widehat{V}(s_t^{(i)}) - \frac{1}{Z_2} \gamma^{t+1} \widehat{w}(s_t^{(i)}) \frac{\pi(a_i|s_i)}{\pi_0(a_i|s_i)} \widehat{V}(s_{t+1}^{(i)}) \right), \tag{10}$$

where $Z_1 = \sum_{i=1}^n \sum_{t=0}^{T-1} \gamma^t \widehat{w}(s_t^{(i)})$ and $Z_2 = \sum_{i=1}^n \sum_{t=0}^{T-1} \gamma^{t+1} \widehat{w}(s_t^{(i)}) \beta_{\pi/\pi_0}(a_t^{(i)}|s_t^{(i)})$ are self-normalized constant of important weights each empirical estimation.

However, directly estimating R^π using the bridge estimator $\widehat{R}^\pi_{\text{bridge}}[\widehat{V}, \widehat{w}]$ yields a poor estimation, because it includes the errors from both $\widehat{w}$ and $\widehat{V}$ and can be "*doubly worse*". However, we can construct our "*doubly robust*" estimator by canceling $R^\pi_{\text{bridge}}[\widehat{V}, \widehat{w}]$ out from $R^\pi_{\text{SIS}}[\widehat{w}]$ and $R^\pi_{\text{VAL}}[\widehat{V}]$:

$$R^\pi_{\text{DR}}[\widehat{V}, \widehat{w}] = \underbrace{\sum_s r^\pi(s) d_{\pi_0}(s) \widehat{w}(s)}_{R^\pi_{\text{SIS}}[\widehat{w}]} + \underbrace{(1-\gamma) \sum_s \widehat{V}(s) \mu_0(s)}_{R^\pi_{\text{VAL}}[\widehat{V}]} - \underbrace{\sum_s \left(\widehat{V}(s) - \gamma \mathcal{P}^\pi \widehat{V}(s)\right) d_{\pi_0}(s) \widehat{w}(s)}_{R^\pi_{\text{bridge}}[\widehat{V}, \widehat{w}]}. \tag{11}$$

And its corresponding empirical estimator can be written as:

$$\widehat{R}^\pi_{\text{DR}}[\widehat{V}, \widehat{w}] = \widehat{R}^\pi_{\text{SIS}}[\widehat{w}] + \widehat{R}^\pi_{\text{VAL}}[\widehat{V}] - \widehat{R}^\pi_{\text{bridge}}[\widehat{V}, \widehat{w}] \,.$$

Doubly Robust Bias Reduction The double robustness of $\widehat{R}^\pi_{\text{DR}}[\widehat{V}, \widehat{w}]$ is reflected in the following key theorem, which shows that it is accurate once either $\widehat{V}$ or $\widehat{w}$ is accurate.

Theorem 3.1 (Doubly Robustness). *Let $R^\pi_{DR}[\widehat{V}, \widehat{w}] := \lim_{n_0, n, T \to \infty} \widehat{R}^\pi_{DR}[\widehat{V}, \widehat{w}]$ be the limit of $\widehat{R}^\pi_{DR}$ when it has infinite samples. Following the definition above, we have*

$$R^\pi_{DR}[\widehat{V}, \widehat{w}] - R^\pi = \mathbb{E}_{s \sim d_{\pi_0}} \left[\varepsilon_{\widehat{w}}(s) \varepsilon_{\widehat{V}}(s) \right], \tag{12}$$

where $\varepsilon_{\widehat{V}}$ and $\varepsilon_{\widehat{w}}$ are errors of $\widehat{V}$ and $\widehat{w}$, respective, defined as follows

$$\varepsilon_{\widehat{w}} = \frac{d_\pi(s)}{d_{\pi_0}(s)} - \widehat{w}(s), \qquad \varepsilon_{\widehat{V}}(s) = \widehat{V}(s) - r^\pi(s) - \gamma \mathcal{P}^\pi \widehat{V}(s).$$

The error $\varepsilon_{\widehat{w}}$ of $\widehat{w}$ is measured by the difference with the true density ratio $d_\pi(s)/d_{\pi_0}(s)$, and the error $\varepsilon_{\widehat{V}}$ of $\widehat{V}$ is measured using the Bellman residual.

From the theorem we can see that, if $\widehat{V}$ is exact ($\widehat{V} \equiv V^\pi$), we have $\varepsilon_{\widehat{V}} \equiv 0$; if $\widehat{w}$ is exact ($\widehat{w} \equiv d_\pi/d_{\pi_0}$), we have $\varepsilon_{\widehat{w}} \equiv 0$. Therefore, our estimator is consistent (i.e., $\lim_{n, n_0 \to \infty} \widehat{R}^\pi_{\text{DR}}[\widehat{V}, \widehat{w}] = R^\pi$) if either $\widehat{V}$ or $\widehat{w}$ are exact. The estimator is thus doubly robust in this sense. In contrast, $\widehat{R}^\pi_{\text{SIS}}[\widehat{w}]$ and $\widehat{R}^\pi_{\text{VAL}}[\widehat{V}]$ can be more sensitive to the error of $\widehat{w}$ and $\widehat{V}$, respectively:

$$R^\pi_{\text{SIS}}[\widehat{w}] - R^\pi = \mathbb{E}_{s \sim d_{\pi_0}} \left[\varepsilon_{\widehat{w}}(s) r^\pi(s) \right], \qquad R^\pi_{\text{VAL}}[\widehat{V}] - R^\pi = \mathbb{E}_{s \sim d_{\pi_0}} \left[w_{\pi/\pi_0}(s) \varepsilon_{\widehat{V}}(s) \right].$$

Variance Analysis Different from the bias reduction, the doubly robust estimator does not guarantee to the reduce the variance over $\widehat{R}^\pi_{\text{SIS}}[\widehat{w}]$ or $\widehat{R}^\pi_{\text{VAL}}[\widehat{V}]$ in general. However, as we show in the following result, we can break the variance of $\widehat{R}^\pi_{\text{DR}}[\widehat{V}, \widehat{w}]$ into two parts. The first is the variance of $\widehat{R}^\pi_{\text{VAL}}[\widehat{V}]$ which is often relatively small. The second is generally no greater than the variance of $\widehat{R}^\pi_{\text{SIS}}[\widehat{w}]$, and can be much smaller when $\widehat{V} \approx V^\pi$. Moreover, $\widehat{R}^\pi_{\text{VAL}}[\widehat{V}]$ and $\widehat{R}^\pi_{\text{SIS}}[\widehat{w}]$ avoid the curse of horizon by design, so their variances tend to be much smaller than the corresponding estimators that apply IPS correction on the trajectories.

Algorithm 1 Infinite Horizon Doubly Robust Estimator

```
Input: Transition data $\mathcal{D}_{\pi_0} = \{s_t^{(i)}, a_t^{(i)}, r_t^{(i)}\}_{1\le i\le n, 0\le t\le T}$ from policy $\pi_0$;
       $\mathcal{D}_0 = \{s_0^{(j)}\}_{1\le j\le n_0}$ be samples from initial distribution $\mu_0$
       target policy $\pi$, value function estimate $\widehat{V}$; density ratio estimate $\widehat{w}$.
Estimation: Use $\widehat{R}^{\pi}_{\mathrm{DR}}$ in (11) to estimate $R^\pi$ using sample from $\mathcal{D}$ and $\mathcal{D}_0$.
```

Proposition 3.1 (Variance Analysis)**.** *Assume $\widehat{R}^{\pi}_{DR}[\widehat{V}, \widehat{w}]$ is estimated based on sample $\mathcal{D}_0 \sim \mu_0$ and $\mathcal{D}_{\pi_0} \sim d_{\pi_0}$, which we assume to be independent of each other. We have*

$$Var_{\mathcal{D}_0, \mathcal{D}_{\pi_0}}\left[\widehat{R}^{\pi}_{DR}[\widehat{V}, \widehat{w}]\right] = Var_{\mathcal{D}_0}\left[\widehat{R}^{\pi}_{VAL}[\widehat{V}]\right] \; + \; Var_{\mathcal{D}_{\pi_0}}\left[\widehat{R}^{\pi}_{res}[\widehat{V}, \widehat{w}]\right], \tag{13}$$

with $\widehat{R}^{\pi}_{res}[\widehat{V}, \widehat{w}] := \widehat{R}^{\pi}_{SIS}[\widehat{w}] - \widehat{R}^{\pi}_{bridge}[\widehat{V}, \widehat{w}] = \widehat{\mathbb{E}}_{\mathcal{D}_{\pi_0}}\left[\widehat{\varepsilon}_{\widehat{V}}(s)\widehat{w}(s)\right]$ and $\widehat{\varepsilon}_{\widehat{V}}(s)$ the TD error of $\widehat{V}$,

$$\widehat{\varepsilon}_{\widehat{V}}(s) = \widehat{r^{\pi}}(s) - \widehat{V}(s) + \gamma\widehat{\mathcal{P}^{\pi}}\widehat{V}(s),$$

with $\widehat{r^{\pi}}(s) = r(s,a)\pi(a|s)/\pi_0(a|s)$, $\widehat{\mathcal{P}^{\pi}}\widehat{V}(s) = \pi(a|s)/\pi_0(a|s)\widehat{V}(s')$. Therefore, $Var_{\mathcal{D}_{\pi_0}}(\widehat{R}^{\pi}_{res}[\widehat{V}, \widehat{w}])$ can be small when $\widehat{V}$ is close to the true value V^π, or $\widehat{\varepsilon}_{\widehat{V}}(s) \approx 0$.

From the proposition we can see that when $\widehat{V}$ is close to the true value V^π, the variance of the residual $\widehat{\varepsilon}_{\widehat{V}}$ may be negligibly small compared to the variance of $\widehat{R}^{\pi}_{\mathrm{SIS}}[\widehat{w}]$. A further comparison of the variances between $\widehat{R}^{\pi}_{\mathrm{res}}[\widehat{V}, \widehat{w}]$ and $\widehat{R}^{\pi}_{\mathrm{SIS}}[\widehat{w}]$ is provided in Appendix A.3. In the case when the TD error $\widehat{\varepsilon}_{\widehat{V}}$ is negligible, we have $\mathrm{Var}_{\mathcal{D}_0, \mathcal{D}_{\pi_0}}\left[\widehat{R}^{\pi}_{\mathrm{DR}}[\widehat{V}, \widehat{w}]\right] \approx \mathrm{Var}_{\mathcal{D}_0}\left[\widehat{R}^{\pi}_{\mathrm{VAL}}[\widehat{V}]\right]$. Typically, the variance of $\widehat{R}^{\pi}_{\mathrm{VAL}}[\widehat{V}]$ is smaller than $\widehat{R}^{\pi}_{\mathrm{SIS}}[\widehat{w}]$ since the variance of importance sampling methods heavily depends on the effective sample size, which is less controllable compared to $\widehat{R}^{\pi}_{\mathrm{VAL}}[\widehat{V}]$. Therefore, the variance of our doubly robust estimator may even be smaller than that of $\widehat{R}^{\pi}_{\mathrm{SIS}}[\widehat{w}]$ in practice.

The variance in (13) is a sum of two terms because of the assumption that samples from μ_0 and d_{π_0} are independent. In practice, they have dependency but it is possible to couple the samples from μ_0 and d_{π_0} in a certain way to further decrease the variance, which we leave as future work.

Proposed Algorithm for Off-Policy Evaluation Suppose we have already obtained $\widehat{V}$ and $\widehat{w}$, as estimations of V^π and w_{π/π_0}, respectively, we can directly use equation (11) to estimate R^π. A detail procedure is described in Algorithm 1.

4 Double Robustness and Lagrangian Duality

In this section, we reveal a surprising connection between our double robustness and Lagrangian duality. We show that our doubly robust estimator is equivalent to the Lagrangian function of a primal-dual formulation of policy evaluation. This connection is of interest in itself, and may provide a foundation for deriving more effective algorithms.

We start with the following classic, optimization formulation of policy evaluation (Puterman, 2014):

$$R^\pi = \min_V \left\{ (1-\gamma)\sum_s \mu_0(s)V(s) \qquad \text{s.t.} \qquad V(s) \ge r^\pi(s) + \gamma\mathcal{P}^\pi V(s), \quad \forall s \right\}, \tag{14}$$

where we find V to maximize its average value, subject to an inequality constraint on the Bellman equation. It can be shown that the solution of (14) is achieved by the true value function V^π, hence yielding a true expected reward R^π.

Introducing a Lagrangian multiplier $\rho \ge 0$, we can derive the Lagrangian function $L(V, \rho)$ of (14),

$$L(V, \rho) = (1-\gamma)\sum_s \mu_0(s)V(s) - \sum_s \rho(s)(V(s) - r^\pi(s) + \gamma\mathcal{P}^\pi V(s)). \tag{15}$$

Comparing $L(V, \rho)$ with our estimator $\widehat{R}^{\pi}_{\mathrm{DR}}[\widehat{V}, \widehat{w}]$ in (11), we can see that they are in fact equivalent in expectation.

Theorem 4.1 (Primal Dual). *I) Define* $w_{\rho/\pi_0}(s) = \frac{\rho(s)}{d_{\pi_0}(s)}$. *We have*

$$L(V, \rho) = R^{\pi}_{DR}[V, w_{\rho/\pi_0}], \qquad \textit{and hence} \qquad L(V, d_\pi) = L(V^\pi, \rho) = R^\pi, \ \forall\, V, \rho,$$

which suggests that $L(V, \rho)$ *is "doubly robust" in that it equals* R^π *if either* $V = V^\pi$ *or* $\rho = d_\pi$.

II) The primal problem (14) *forms a strong duality with the following dual problem,*

$$R^\pi = \max_{\rho \geq 0} \left\{ \sum_s \rho(s) r^\pi(s) \qquad \textit{s.t.} \qquad \rho(s') = (1-\gamma)\mu_0(s') + \gamma \mathcal{T}^\pi \rho(s'), \qquad \forall s' \right\}, \tag{16}$$

where $\mathcal{T}^\pi$ *is defined in* (3).

This shows that the dual problem is equivalent to constraint ρ using the fixed point equation (3) and maximize the average reward given distribution ρ. Since the unique fixed point of (3) is $d_\pi(s)$, the solution of (16) naturally yields the true reward R^π, hence forming a zero duality gap with (14).

It is natural to intuit the double robustness of the Lagrangian function. From (15), $L(V, \rho)$ can be viewed as estimating the reward R^π using value function with a correction of Bellman residual $(V - r^\pi - \gamma \mathcal{P}^\pi V)$. If $V = V^\pi$, the estimation equals the true reward and the correction equals zero. From the dual problem (16), $L(V, \rho)$ can be viewed as estimating R^π using density function ρ, corrected by the residual $(\rho - (1-\gamma)\mu_0 - \gamma \mathcal{T}^\pi \rho)$. We again get the true reward if $\rho = d_\pi$.

It turns out that we can use the primal-dual formula when $\gamma = 1$ to obtain the double robust estimator for the average reward case. We clarify it in appendix B.

Remark The fact that the density function d_π forms a dual variable of the value function V^π is widely known in the optimal control and reinforcement learning literature (e.g., Bertsekas, 2000; Puterman, 2014; de Farias & Van Roy, 2003), and has been leveraged in various works for policy optimization. However, it does not seem to be well exploited in the literature of off-policy policy evaluation.

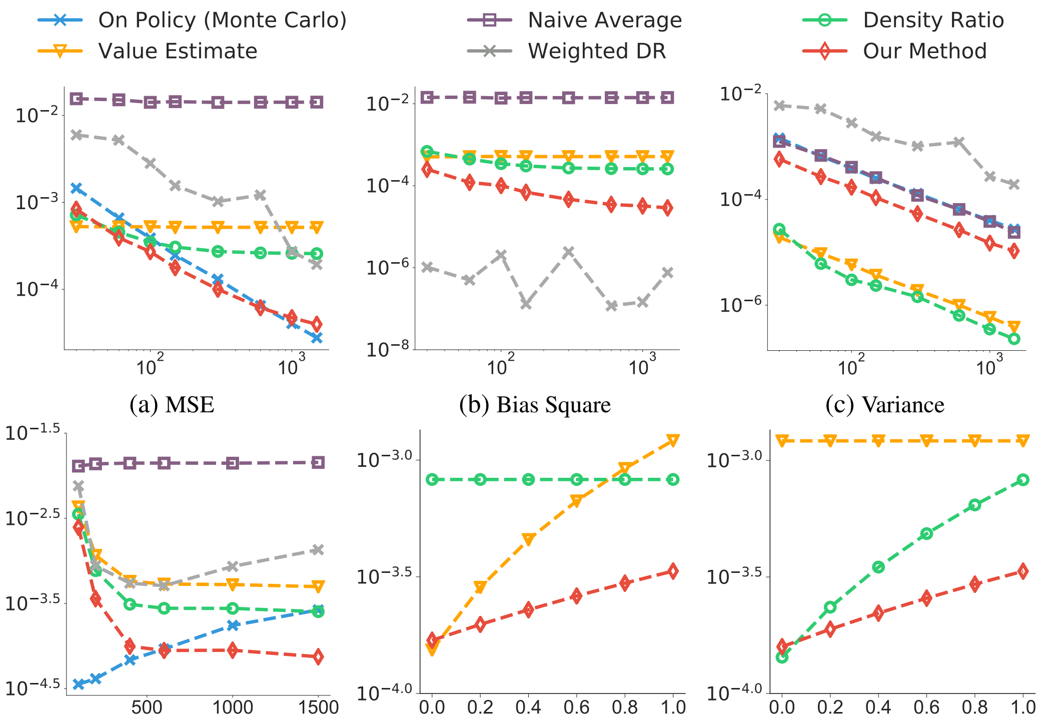
5 Related Work

Off-Policy Value Evaluation The problem of off-policy value evaluation has been studied extensively in contextual bandits and MDPs (Fonteneau et al., 2013; Li et al., 2015; Jiang & Li, 2016; Thomas & Brunskill, 2016; Liu et al., 2018b; Farajtabar et al., 2018; Hanna et al., 2019; Xie et al., 2019; Rowland et al., 2019). However, most of the existing works are based on importance sampling (IS) to correct the mismatch between the distribution of the whole trajectories induced by the behavior and target policies, which faces the "curse of horizon" (Liu et al., 2018a) when extended to long-horizon (or infinite-horizon) problems.

Several other works (Guo et al., 2017; Hallak & Mannor, 2017; Liu et al., 2018a; Gelada & Bellemare, 2019; Nachum et al., 2019; Liu et al., 2019a) have been proposed to address the high variance issue in the long-horizon problems. Liu et al. (2018a) apply importance sampling on the average visitation distribution of state-action pairs, instead of the distribution of the whole trajectories, providing an effective approach to break "the curse of horizon". However, they require to learn a density ratio function over the whole state-action pairs, which may induce large bias. Our work incorporates the density ratio and value function estimation, which significantly reduces the induced bias of two estimators, resulting a doubly robust estimator.

Our work is also closely related to DR techniques used in finite horizon problems (Murphy et al., 2001; Dudík et al., 2011; Jiang & Li, 2016; Thomas & Brunskill, 2016; Farajtabar et al., 2018), which incorporate an approximate value function as control variates to IS estimators. Different from existing DR approaches, our work is related to the well known duality between the density and the value function, which reveals the relationship between density (ratio) learning (Liu et al., 2018a) and value function learning. Based on this interesting observation, we further obtain the doubly robust estimator for estimating average reward in infinite-horizon problems.

A recent work (Xie et al., 2019) has also explored the idea of tracking marginal state distribution shifts at every single time step in an episode, instead of the stationary state distribution across all time steps (Liu et al., 2018a). It provides additional benefits for understanding and analyzing problems



(a) MSE

(b) Bias Square

(c) Variance

(d) MSE with H changes

(e) Bias square with α changes

(f) Bias square with β changes

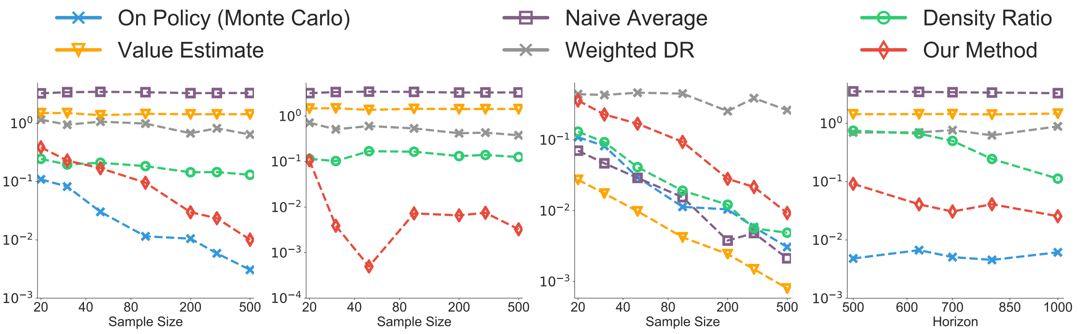
Figure 1: Off Policy Evaluation Results on Taxi. Default parameter, discounted factor $\gamma = 0.99$, mixed ratio $\alpha = \beta = 1$, horizon length $H = 600$. For (a)-(c) the x-axis is the number of trajectories and y-axis corresponds to MSE, Bias Square and Variance, respectively. For (d) we fix the total number of samples (number of trajectories times horizon length) and change the horizon length as x-axis and observe the MSE. (e) and (f) show the change the mixed ratio of α, β with the change of bias. We repeat each experiment for 1000 runs.

such as short horizon and time-variant MDPs. They call this approach *marginalized importance sampling* (MIS). Later on, Kallus & Uehara (2019a) incorporate the DR technique to improve the MIS estimator. Independent of this work, Kallus & Uehara (2019b); Uehara et al. (2019) extend previous works to propose estimators can be viewed as a variant of ours.

Primal-Dual Value Learning Primal-dual optimization techniques have been widely used for off-policy value function learning and policy optimization (Liu et al., 2015; Chen & Wang, 2016; Dai et al., 2017; 2018; Feng et al., 2019). Nevertheless, the duality between density and value function has not been well explored in the literature of off policy value estimation. Our work proposes a new doubly robustness technique for off-policy value estimation, which can be naturally viewed as the Lagrangian function of the primal-dual formulation of policy evaluation, providing an alternative unified view for off policy value evaluation.

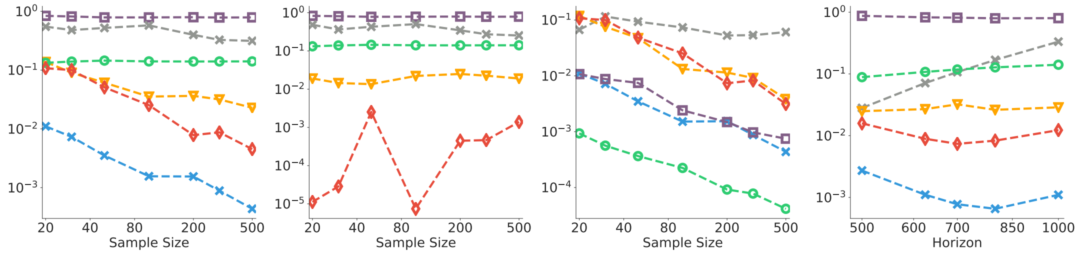
6 Experiment

In this section, we conduct simulation experiments on different environmental settings to compare our new doubly robust estimator with existing methods. We mainly compare with infinite horizon based estimator including state importance sampling estimator (Liu et al. (2018a)) and value function estimator. We do not report results on the vanilla trajectory-based importance sampling estimators because of their significant higher variance, but we do compare with the doubly robust version induced by Thomas & Brunskill (2016) (self-normalized variant of Jiang & Li (2016)). In all experiments we compare with Monte Carlo and naive average as Liu et al. (2018a) suggested. The ground truth for each environment is calculated by averaging Monte Carlo estimation with a very large sample size.



(a) MSE (b) Bias Square (c) Variance (d) MSE with H changes

Figure 2: Off Policy Evaluation Results on Puck-Mountain. We set discounted factor $\gamma = 0.995$ as default. For (a)-(c) we set the horizon $H = 1000$ and the x-axis is the number of trajectories for used for evaluation. For (d) we fix the total number of samples and change the horizon length.



(a) MSE (b) Bias Square (c) Variance (d) MSE with H changes

Figure 3: Off Policy Evaluation Results on InvertedPendulum-v2. We set discounted factor $\gamma = 0.995$ as default. For (a)-(c) we set the horizon $H = 1000$ and the x-axis is the number of trajectories for used for evaluation. For (d) we fix the total number of samples and change the horizon length.

Taxi Environment We follow Liu et al. (2018a)'s tabular environment *Taxi*, which has 2000 states and 6 actions in total. For more experimental details, please check appendix C.1.

We pre-train two different $\widehat{V}$ and $\tilde{V}$ trained with a small and fairly large size of samples, respectively, where $\tilde{V}$ is very close to true value function V^π but $\widehat{V}$ is relatively further from it. Similarly we pre-train $\widehat{\rho}$ and $\tilde{\rho} \approx d_\pi$. For estimation we use a mixed ratio α, β to control the bias of the input V, ρ, where $V = \alpha\widehat{V} + (1-\alpha)\tilde{V}$ and $\rho = \beta\widehat{\rho} + (1-\beta)\tilde{\rho}$.

Figure 1(a)-(c) show results of comparison for different methods as we changing the number of trajectories. We can see that the MSE performance of value function($\widehat{R}^\pi_{\text{VAL}}$) and state visitation importance sampling($\widehat{R}^\pi_{\text{SIS}}$) estimators are mainly impeded by their large biases, while our method has much less bias thus it can keep decreasing as sample size increase and achieves same performance as on policy estimator. Figure 1(d) shows results if we change the horizon length. Notice that here we keep the number of samples to be the same, so if we increase our horizon length we will decrease the number of trajectories in the same time. We can see that our method alongside with all infinite horizon methods will get better result as horizon length increase. Figure 1(e)-(f) indicate the "double robustness" of our method, where our method benefits from either a better V or a better ρ.

Puck-Mountain Puck-Mountain is an environment similar to Mountain-Car, except that the goal of Puck-Mountain is to push the puck as high as possible in a local valley, which has a continuous state space of $\mathbb{R}^2$ and a discrete action space similar to Mountain-Car. We use the softmax functions of an optimal Q-function as both target policy and behavior policy, where the temperature of the behavior policy is higher (encouraging exploration). For more details of constructing policies and training algorithms for density ratio and value functions, please check appendix C.2.

Figure 2(a)-(c) show results of comparison for different methods as we changing the number of trajectories. Similar to taxi, we find our method has much lower bias than density ratio and value function estimation, which yields a better MSE. In Figure 2(d) the performance for all infinite horizon estimator will not degenerate as horizon increases, while finite horizon method such as finite weighted horizon doubly robust will suffer from larger variance as horizon increases.

InvertedPendulum InvertedPendulum is a pendulum that has its center of mass above its pivot point. We use the implementation of InvertedPendulum from OpenAI gym (Brockman et al., 2016), which is a continuous control task with state space in $\mathcal{R}^4$ and we discrete the action space to be $\{-1, -0.3, -0.2, 0, 0.2, 0.3, 1\}$. More experiment details can be found in appendix C.2.

In Figure 3(a)-(c) our method again significantly reduces the bias, which yields a better MSE comparing with value and density estimation. Figure 3(d) also shows that our method consistently outperforms all other methods as the horizon increases with a fixed total timesteps.

7 Conclusion

In this paper, we develop a new doubly robust estimator based on the infinite horizon density ratio and off policy value estimation. Our new proposed doubly robust estimator can be accurate as long as one of the estimators are accurate, which yields a significant advantage comparing to previous estimators. Future directions include deriving more novel optimization algorithms to learn value function and density(ratio) function by using the primal dual framework.

Acknowledgement

This work is supported in part by NSF CRII 1830161 and NSF CAREER 1846421. We would like to acknowledge Google Cloud for their support.

References

Søren Asmussen and Peter W Glynn. *Stochastic simulation: algorithms and analysis*, volume 57. Springer Science & Business Media, 2007.

Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 2nd edition, 2000. ISBN 1886529094.

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

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Yichen Chen and Mengdi Wang. Stochastic primal-dual methods and sample complexity of reinforcement learning. *arXiv preprint arXiv:1612.02516*, 2016.

Bo Dai, Niao He, Yunpeng Pan, Byron Boots, and Le Song. Learning from conditional distributions via dual kernel embeddings. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1458–1467, 2017. CoRR abs/1607.04579.

Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. SBEED: Convergent reinforcement learning with nonlinear function approximation. In *Proceedings of the Thirty-Fifth International Conference on Machine Learning (ICML)*, pp. 1133–1142, 2018.

Daniela Pucci de Farias and Benjamin Van Roy. The linear programming approach to approximate dynamic programming. *Operations Research*, 51(6):850–865, 2003.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pp. 1097–1104, 2011.

Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 1446–1455, 2018.

Yihao Feng, Lihong Li, and Qiang Liu. A kernel loss for solving the bellman equation. *Neural Information Processing Systems (NeurIPS)*, 2019.

Raphael Fonteneau, Susan A. Murphy, Louis Wehenkel, and Damien Ernst. Batch mode reinforcement learning based on the synthesis of artificial trajectories. *Annals of Operations Research*, 208(1): 383–416, 2013.

Carles Gelada and Marc G Bellemare. Off-policy deep reinforcement learning by bootstrapping the covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3647–3655, 2019.

Zhaohan Guo, Philip S. Thomas, and Emma Brunskill. Using options and covariance testing for long horizon off-policy policy evaluation. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pp. 2489–2498, 2017.

Assaf Hallak and Shie Mannor. Consistent on-line off-policy evaluation. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 1372–1383, 2017.

Josiah Hanna, Scott Niekum, and Peter Stone. Importance sampling policy evaluation with an estimated behavior policy. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 2605–2613, 2019.

Nan Jiang and Lihong Li. Doubly robust off-policy evaluation for reinforcement learning. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pp. 652–661, 2016.

Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *arXiv preprint arXiv:1908.08526*, 2019a.

Nathan Kallus and Masatoshi Uehara. Efficiently breaking the curse of horizon: Double reinforcement learning in infinite-horizon processes. *arXiv preprint arXiv:1909.05850*, 2019b.

Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the 4th International Conference on Web Search and Data Mining (WSDM)*, pp. 297–306, 2011.

Lihong Li, Rémi Munos, and Csaba Szepesvári. Toward minimax off-policy value estimation. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 608–616, 2015.

Bo Liu, Ji Liu, Mohammad Ghavamzadeh, Sridhar Mahadevan, and Marek Petrik. Finite-sample analysis of proximal gradient td algorithms. In *UAI*, pp. 504–513. Citeseer, 2015.

Jun S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer Series in Statistics. Springer-Verlag, 2001. ISBN 0387763694.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pp. 5361–5371, 2018a.

Yao Liu, Omer Gottesman, Aniruddh Raghu, Matthieu Komorowski, Aldo A Faisal, Finale Doshi-Velez, and Emma Brunskill. Representation balancing mdps for off-policy policy evaluation. In *Advances in Neural Information Processing Systems*, pp. 2644–2653, 2018b.

Yao Liu, Pierre-Luc Bacon, and Emma Brunskill. Understanding the curse of horizon in off-policy evaluation via conditional importance sampling. *arXiv preprint arXiv:1910.06508*, 2019a.

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Off-policy policy gradient with state distribution correction. *arXiv preprint arXiv:1904.08473*, 2019b.

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

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *Neural Information Processing Systems (NeurIPS)*, 2019.

Martin L Puterman. *Markov Decision Processes.: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

Mark Rowland, Anna Harutyunyan, Hado van Hasselt, Diana Borsa, Tom Schaul, Rémi Munos, and Will Dabney. Conditional importance sampling for off-policy learning. *arXiv preprint arXiv:1910.07479*, 2019.

Alexander L. Strehl, John Langford, Lihong Li, and Sham M. Kakade. Learning from logged implicit exploration data. In *Advances in Neural Information Processing Systems 23 (NIPS-10)*, pp. 2217–2225, 2010.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, March 1998. ISBN 0-262-19398-1.

Philip S. Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pp. 2139–2148, 2016.

Philip S. Thomas, Georgios Theocharous, Mohammad Ghavamzadeh, Ishan Durugkar, and Emma Brunskill. Predictive off-policy policy evaluation for nonstationary decision problems, with applications to digital marketing. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI)*, pp. 4740–4745, 2017.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and Q-function learning for off-policy evaluation, 2019. arXiv:1910.12809.

Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. Optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. *Neural Information Processing Systems (NeurIPS)*, 2019.

A Proof

A.1 Transition Operator for Bellman Equation

For simplicity, we define the following two operators thorough our proofs to simplify our notations.

Definition A.1. *Given a policy π and the unknown environment transition $\boldsymbol{T}$, we define $\mathcal{T}^\pi$ and $\mathcal{P}^\pi$ over any function $f : \mathcal{S} \to \mathcal{R}$ as*

$$\begin{aligned}
\left(\mathcal{T}^\pi f\right)(s') &= \sum_{s,a} \boldsymbol{T}(s'|s,a)\pi(a|s)f(s) \\
\left(\mathcal{P}^\pi f\right)(s) &= \sum_{s',a} \boldsymbol{T}(s'|s,a)\pi(a|s)f(s')
\end{aligned}$$

Using these operator notations, we can rewrite the above two recursive equations as:

$$\begin{aligned}
V^\pi &= r^\pi + \gamma \mathcal{P}^\pi V^\pi, \\
d_\pi &= (1-\gamma)\mu_0 + \gamma \mathcal{T}^\pi d_\pi,
\end{aligned}$$

where $r^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}[r(s,a)]$.

These transition operators have the following nice adjoint property.

Lemma A.2. *For two function f and g, if the following summation is finite, we will have*

$$\sum_s \left(\mathcal{P}^\pi f\right)(s) g(s) = \sum_s f(s) \left(\mathcal{T}^\pi g\right)(s). \tag{17}$$

Proof.

$$\begin{aligned}
\sum_s \left(\mathcal{P}^\pi f\right)(s) g(s) =& \sum_s \left(\sum_{s',a} \boldsymbol{T}(s'|s,a)\pi(a|s)f(s') \right) g(s) \\
=& \sum_{s'} f(s') \left(\sum_{s,a} \boldsymbol{T}(s'|s,a)\pi(a|s)g(s) \right) \\
=& \sum_{s'} f(s') \left(\mathcal{T}^\pi g\right)(s')
\end{aligned}$$

□

Using this property we can actually using Bellman Equations to re-derive the two different way to get R^π.

$$\begin{aligned}
R^\pi =& \lim_{T\to\infty} \mathbb{E}_{\tau^{(i)}\sim\pi}[\frac{\sum_{t=0}^T \gamma^t r_t}{\sum_{t=0}^T \gamma^t}] \\
=& \sum_s V^\pi(s)(1-\gamma)\mu_0(s) \\
=& \sum_s V^\pi(s) \left(\mathcal{I} - \gamma\mathcal{T}^\pi\right) d_\pi(s) \\
=& \sum_s \left(\mathcal{I} - \gamma\mathcal{P}^\pi\right) V^\pi(s) d_\pi(s) \\
=& \sum_s r^\pi(s) d_\pi(s).
\end{aligned}$$

A.2 Proof of Theorem 3.1

Theorem 3.1 (Doubly Robustness)**.** *Let $R^{\pi}_{DR}[\widehat{V}, \widehat{w}] := \lim_{n_0, n, T \to \infty} \widehat{R}^{\pi}_{DR}[\widehat{V}, \widehat{w}]$ be the limit of $\widehat{R}^{\pi}_{DR}$ when it has infinite samples. Following the definition above, we have*

$$R^{\pi}_{DR}[\widehat{V}, \widehat{w}] - R^{\pi} = \mathbb{E}_{s \sim d_{\pi_0}} \left[\varepsilon_{\widehat{w}}(s) \varepsilon_{\widehat{V}}(s) \right], \tag{12}$$

where $\varepsilon_{\widehat{V}}$ and $\varepsilon_{\widehat{w}}$ are errors of $\widehat{V}$ and $\widehat{w}$, respective, defined as follows

$$\varepsilon_{\widehat{w}} = \frac{d_{\pi}(s)}{d_{\pi_0}(s)} - \widehat{w}(s), \qquad\qquad \varepsilon_{\widehat{V}}(s) = \widehat{V}(s) - r^{\pi}(s) - \gamma \mathcal{P}^{\pi} \widehat{V}(s).$$

The error $\varepsilon_{\widehat{w}}$ of $\widehat{w}$ is measured by the difference with the true density ratio $d_{\pi}(s)/d_{\pi_0}(s)$, and the error $\varepsilon_{\widehat{V}}$ of $\widehat{V}$ is measured using the Bellman residual.

Proof. Using the property of the operator, we can rewrite $(1-\gamma)\mu_0(s)$ using Bellman equation as $d_{\pi} - \gamma \mathcal{T}^{\pi} d_{\pi}$, thus we have

$$\begin{aligned}
R^{\pi}_{\text{VAL}}[\widehat{V}] =& (1-\gamma) \sum_{s} \widehat{V}(s) \mu_0(s) \\
=& \sum_{s} \widehat{V}(s) \left(d_{\pi} - \gamma \mathcal{T}^{\pi} d_{\pi} \right)(s) \\
=& \sum_{s} \left(I - \gamma \mathcal{P}^{\pi} \right) \widehat{V}(s) d_{\pi}(s).
\end{aligned}$$

and similarly if we break r^{π} as $(I - \gamma \mathcal{P}^{\pi}) V^{\pi}$, for $R^{\pi}_{\text{SIS}}[\widehat{w}]$ we have:

$$R^{\pi}_{\text{SIS}}[\widehat{w}] = \sum_{s} \left(I - \gamma \mathcal{P}^{\pi} \right) V^{\pi}(s) d_{\widehat{w}}(s),$$

where $d_{\widehat{w}} = d_{\pi_0} \widehat{w}$ for short.

Compare with $R^{\pi} = \sum_{s} \left(I - \gamma \mathcal{P}^{\pi} \right) V^{\pi}(s) d_{\pi}(s)$, we can see the main difference between R^{π}_{SIS} and R^{π}_{VAL} with R^{π} are they replace d_{π} and $d_{\widehat{w}}$ and V^{π} with $\widehat{V}$ respectively. If we add them together and minus the connection estimator, we have we will have:

$$\begin{aligned}
R^{\pi}_{\text{DR}}[\widehat{V}, \widehat{w}] - R^{\pi} =& R^{\pi}_{\text{SIS}}[\widehat{w}] + R^{\pi}_{\text{VAL}}[\widehat{V}] - R^{\pi}_{\text{bridge}}[\widehat{V}, \widehat{w}] - R^{\pi} \\
=& \sum_{s} \Big(\left(I - \gamma \mathcal{P}^{\pi} \right) V^{\pi}(s) d_{\widehat{w}}(s) + \left(I - \gamma \mathcal{P}^{\pi} \right) \widehat{V}(s) d_{\pi}(s) \\
& \qquad - \left(I - \gamma \mathcal{P}^{\pi} \right) \widehat{V}(s) d_{\widehat{w}}(s) - \left(I - \gamma \mathcal{P}^{\pi} \right) V^{\pi}(s) d_{\pi}(s) \Big) \\
=& \sum_{s} \left(I - \gamma \mathcal{P}^{\pi} \right) (V^{\pi} - \widehat{V})(s) (d_{\widehat{w}} - d_{\pi})(s) \\
=& \sum_{s} d_{\pi_0}(s) \varepsilon_{\widehat{V}}(s) \varepsilon_{\widehat{w}}(s) \\
=& \mathbb{E}_{s \sim d_{\pi_0}} \left[\varepsilon_{\widehat{V}}(s) \varepsilon_{\widehat{w}}(s) \right],
\end{aligned}$$

where the third equation is because $\left(I - \gamma \mathcal{P}^{\pi} \right) (V^{\pi} - \widehat{V}) = \left(I - \gamma \mathcal{P}^{\pi} \right) V^{\pi} - \left(I - \gamma \mathcal{P}^{\pi} \right) \widehat{V} = r^{\pi} - \left(I - \gamma \mathcal{P}^{\pi} \right) \widehat{V}$. Similarly, the bias for R^{π}_{DR} and R^{π}_{SIS} can be calculated as:

$$\begin{aligned}
R^{\pi}_{\text{SIS}}[\widehat{w}] - R^{\pi} =& \sum_{s} \left(I - \gamma \mathcal{P}^{\pi} \right) V^{\pi}(s) \left(d_{\widehat{w}}(s) - d_{\pi}(s) \right) \\
=& \sum_{s} r^{\pi}(s) d_{\pi_0}(s) \varepsilon_{\widehat{w}}(s) \\
=& \mathbb{E}_{s \sim d_{\pi_0}} [\varepsilon_{\widehat{w}}(s) r^{\pi}(s)],
\end{aligned}$$

and

$$\begin{aligned}
R^{\pi}_{\text{VAL}}[\widehat{V}] - R^{\pi} =& \sum_{s} \left(I - \gamma \mathcal{P}^{\pi} \right) (\widehat{V}(s) - V^{\pi}(s)) d_{\pi}(s) \\
=& \sum_{s} \varepsilon_{\widehat{V}}(s) w_{\pi/\pi_0}(s) d_{\pi_0}(s) \\
=& \mathbb{E}_{s \sim d_{\pi_0}} [w_{\pi/\pi_0}(s) \varepsilon_{\widehat{V}}(s)].
\end{aligned}$$

□

A.3 More discussions on the Variance in Proposition 3.1

Recall the proposition 3.1.

Proposition 3.1 (Variance Analysis)**.** *Assume $\widehat{R}^{\pi}_{DR}[\widehat{V},\widehat{w}]$ is estimated based on sample $\mathcal{D}_0 \sim \mu_0$ and $\mathcal{D}_{\pi_0} \sim d_{\pi_0}$, which we assume to be independent of each other. We have*

$$\mathit{Var}_{\mathcal{D}_0,\mathcal{D}_{\pi_0}}\left[\widehat{R}^{\pi}_{\mathit{DR}}[\widehat{V},\widehat{w}]\right] = \mathit{Var}_{\mathcal{D}_0}\left[\widehat{R}^{\pi}_{\mathit{VAL}}[\widehat{V}]\right] \; + \; \mathit{Var}_{\mathcal{D}_{\pi_0}}\left[\widehat{R}^{\pi}_{\mathit{res}}[\widehat{V},\widehat{w}]\right], \tag{13}$$

with $\widehat{R}^{\pi}_{\mathit{res}}[\widehat{V},\widehat{w}] := \widehat{R}^{\pi}_{\mathit{SIS}}[\widehat{w}] - \widehat{R}^{\pi}_{\mathit{bridge}}[\widehat{V},\widehat{w}] = \widehat{\mathbb{E}}_{\mathcal{D}_{\pi_0}}\left[\widehat{\varepsilon}_{\widehat{V}}(s)\widehat{w}(s)\right]$ and $\widehat{\varepsilon}_{\widehat{V}}(s)$ the TD error of $\widehat{V}$,

$$\widehat{\varepsilon}_{\widehat{V}}(s) = \widehat{r^{\pi}}(s) - \widehat{V}(s) + \gamma\widehat{\mathcal{P}^{\pi}}\widehat{V}(s),$$

with $\widehat{r^{\pi}}(s) = r(s,a)\pi(a|s)/\pi_0(a|s)$, $\widehat{\mathcal{P}^{\pi}}\widehat{V}(s) = \pi(a|s)/\pi_0(a|s)\widehat{V}(s')$. Therefore, $\mathit{Var}_{\mathcal{D}_{\pi_0}}(\widehat{R}^{\pi}_{\mathit{res}}[\widehat{V},\widehat{w}])$ can be small when $\widehat{V}$ is close to the true value V^{π}, or $\widehat{\varepsilon}_{\widehat{V}}(s) \approx 0$.

Proof. Since $\widehat{R}^{\pi}_{\text{DR}}[\widehat{V},\widehat{w}] = \widehat{R}^{\pi}_{\text{VAL}}[\widehat{V}] + \widehat{R}^{\pi}_{\text{res}}[\widehat{V},\widehat{w}]$ and we have the assumption that the samples from $\mathcal{D}_0$ and $\mathcal{D}_{\pi_0}$ are independent, the proposition comes immediately. □

Intuitively, the variance of $\widehat{R}^{\pi}_{\text{res}}[\widehat{V},\widehat{w}]$ is less than $\widehat{R}^{\pi}_{\text{SIS}}[\widehat{w}]$. We want to quantitatively study every pieces of orthogonal randomness of $\widehat{R}^{\pi}_{\text{res}}[\widehat{V},\widehat{w}]$ and $\widehat{R}^{\pi}_{\text{SIS}}[\widehat{w}]$ in details, and we state the analysis as the following theorem.

Theorem A.3. *Let $\mathit{Var}_{\mathcal{D}_{\pi_0}}\left[\widehat{R}^{\pi}_{\mathit{res}}[\widehat{V},\widehat{w}]\right]$ be defined in Proposition 3.1, and suppose the normalization is constant (and hence an ordinary importance sampling which is unbiased). Then we can further break it into two terms*

$$\mathit{Var}_{\mathcal{D}_{\pi_0}}\left[\widehat{R}^{\pi}_{\mathit{res}}[\widehat{V},\widehat{w}]\right] = \frac{1}{n}\left(\mathit{Var}\left[\widehat{w}(s)\varepsilon_{\widehat{V}}(s)\right] + \mathbb{E}\left[\widehat{w}(s)^2\left(\delta_1(s,a) + \gamma\delta_2(s,a,s')\right)^2\right]\right), \tag{18}$$

where $\varepsilon_{\widehat{V}}(s) = \widehat{V}(s) - r^{\pi}(s) - \gamma\mathcal{P}^{\pi}\widehat{V}(s')$ is the Bellman residual(not the empirical one), $\delta_1(s,a) = \frac{\pi(a|s)}{\pi_0(a|s)}r(s,a) - r^{\pi}(s)$ is the randomness for action and $\delta_2(s,a,s') = \frac{\pi(a|s)}{\pi_0(a|s)}\widehat{V}(s') - \mathcal{P}^{\pi}\widehat{V}(s)$ is the randomness for transition operator over function $\widehat{V}$. Both δ_1 and δ_2 is zero mean if we condition over s.

Compared with $\mathit{Var}\left[\widehat{R}^{\pi}_{\mathit{SIS}}[\widehat{w}]\right]$ we have:

$$\mathit{Var}\left[\widehat{R}^{\pi}_{\mathit{SIS}}[\widehat{w}]\right] = \frac{1}{n}(\mathit{Var}\left[\widehat{w}(s)r^{\pi}(s)\right] + \mathbb{E}\left[\widehat{w}(s)^2\delta_1(s,a)^2\right]) \tag{19}$$

Proof. $\widehat{R}^{\pi}_{\text{res}}[\widehat{V},\widehat{w}]$ can be written as

$$\widehat{R}^{\pi}_{\text{res}}[\widehat{V},\widehat{w}] = \frac{1}{n}\sum\widehat{w}(s)\left(\beta_{\pi/\pi_0}(a|s)(r + \gamma\widehat{V}(s')) - \widehat{V}(s)\right),$$

where $\beta_{\pi/\pi_0}(a|s)$ is short for $\frac{\pi(a|s)}{\pi_0(a|s)}$. We can break $\beta_{\pi/\pi_0}(a|s)(r + \gamma\widehat{V}(s')) - \widehat{V}(s)$ into

$$\begin{aligned}
&\beta_{\pi/\pi_0}(a|s)(r(s,a) + \gamma\widehat{V}(s')) - \widehat{V}(s)\\
=&(-\widehat{V}(s) + r^{\pi}(s) + \gamma\mathcal{P}^{\pi}\widehat{V}(s)) + \underbrace{(\beta_{\pi/\pi_0}(a|s)r(s,a) - r^{\pi}(s))}_{\delta_1} + \gamma\underbrace{(\beta_{\pi/\pi_0}(a|s)\widehat{V}(s') - \mathcal{P}^{\pi}\widehat{V}(s))}_{\delta_2}\\
=& -\varepsilon_{\widehat{V}}(s) + \delta_1(s,a) + \gamma\delta_2(s,a,s').
\end{aligned}$$

where $\varepsilon_{\widehat{V}} = \widehat{V} - r^\pi - \mathcal{P}^\pi \widehat{V}$ is the Bellman residual and the if we condition over s we have the expectations for δ_1 and δ_2 are 0. Also notice that if we condition over s then $\varepsilon_{\widehat{V}}(s)$ become a constant thus it is independent to δ_1 and δ_2. Thus we have:

$$\begin{aligned}&\mathrm{Var}\left[\widehat{w}(s)\left(\beta_{\pi/\pi_0}(a|s)(r+\gamma\widehat{V}(s'))-\widehat{V}(s)\right)\right]\\ =&\mathrm{Var}\left[\widehat{w}(s)\left(-\varepsilon_{\widehat{V}}(s)+\delta_1(s,a)+\gamma\delta_2(s,a,s')\right)\right]\\ =&\mathrm{Var}\left[\widehat{w}(s)\varepsilon_{\widehat{V}}(s)\right]+\mathbb{E}\left[\widehat{w}(s)^2\left(\delta_1(s,a)+\gamma\delta_2(s,a,s')\right)^2\right]\end{aligned}$$

Therefore we have:

$$\mathrm{Var}\left[\widehat{R}^\pi_{\text{DR}}[\widehat{V},\widehat{w}]\right] = \frac{(1-\gamma)^2}{n_0}\mathrm{Var}[\widehat{V}(s_0)]+\frac{1}{n}(\mathrm{Var}\left[\widehat{w}(s)\varepsilon_{\widehat{V}}(s)\right]+\mathbb{E}\left[\widehat{w}(s)^2\left(\delta_1(s,a)+\gamma\delta_2(s,a,s')\right)^2\right]).$$

For $\mathrm{Var}\left[\widehat{R}^\pi_{\text{SIS}}[\widehat{w}]\right]$ we have:

$$\begin{aligned}\mathrm{Var}\left[\widehat{R}^\pi_{\text{SIS}}[\widehat{w}]\right] =&\frac{1}{n}\mathrm{Var}\left[\widehat{w}(s)\beta_{\pi/\pi_0}(a|s)r(s,a)\right]\\ =&\frac{1}{n}\mathrm{Var}\left[\widehat{w}(s)r^\pi(s)+\widehat{w}(s)\delta_1(s,a)\right]\\ =&\frac{1}{n}(\mathrm{Var}\left[\widehat{w}(s)r^\pi(s)\right]+\mathbb{E}\left[\widehat{w}^2(s)\delta_1(s,a)^2\right]).\end{aligned}$$

□

From the theorem we can see that the variance of residual comes from two parts, the majority part relies on the variance of $|\varepsilon_{\widehat{V}}(s)|$ is usually much smaller than r^π as the majority variance of state visitation importance sampling. When $\widehat{V} \approx V^\pi$, $\varepsilon_{\widehat{V}}(s) \approx 0$, the variance of residual only comes from $\delta_1(s,a)$ and $\delta_2(s,a,s')$.

In practice when we get trajectories data from π_0, to get samples uniformly from d_{π_0}, we can either draw sample s_t depends on its discounted factor γ^t or add an importance weight γ^t as we did in our empirical estimator in equation (6) and equation (10).

A.4 Proof of Theorem 4.1

Theorem 4.1 (Primal Dual). *I) Define $w_{\rho/\pi_0}(s) = \frac{\rho(s)}{d_{\pi_0}(s)}$. We have*

$$L(V,\rho) = R^\pi_{DR}[V, w_{\rho/\pi_0}], \qquad \textit{and hence} \qquad L(V,d_\pi) = L(V^\pi,\rho) = R^\pi,\ \forall\, V,\rho,$$

which suggests that $L(V,\rho)$ is "doubly robust" in that it equals R^π if either $V = V^\pi$ or $\rho = d_\pi$.

II) The primal problem (14) *forms a strong duality with the following dual problem,*

$$R^\pi = \max_{\rho\geq 0}\left\{\sum_s \rho(s)r^\pi(s) \qquad s.t. \qquad \rho(s') = (1-\gamma)\mu_0(s') + \gamma\mathcal{T}^\pi\rho(s'), \qquad \forall s'\right\}, \tag{16}$$

where $\mathcal{T}^\pi$ is defined in (3).

Proof. The Lagrangian can be written as:

$$\begin{aligned}L(V,\rho) =&(1-\gamma)\sum_s\mu_0(s)V(s) - \sum_s\rho(s)\left(V(s)-r^\pi(s)-\gamma\mathcal{P}^\pi V(s)\right)\\ =&\underbrace{\sum_s(1-\gamma)\mu_0(s)V(s)}_{=R^\pi_{\text{VAL}}[V]} - \underbrace{\sum_s\rho(s)\left(I-\gamma\mathcal{P}^\pi\right)V(s)}_{=R^\pi_{\text{bridge}}[V,w_{\rho/\pi_0}]} + \underbrace{\sum_s\rho(s)r^\pi(s)}_{=R^\pi_{\text{SIS}}[w_{\rho/\pi_0}]}\\ =&\sum_s(1-\gamma)\mu_0(s)V(s) - \sum_s\left(I-\gamma\mathcal{T}^\pi\right)\rho(s)V(s) + \sum_s\rho(s)r^\pi(s)\\ =&\sum_s\left((1-\gamma)\mu_0(s) - (I-\gamma\mathcal{T}^\pi)\rho(s)\right)V(s) + \sum_s\rho(s)r^\pi(s).\end{aligned}$$

We can see that the Lagrangian $L(V, \rho)$ is actually our doubly robust estimator $R^{\pi}_{\text{DR}}[V, w_{\rho/\pi_0}]$.

From the last equation we can derive our dual as:

$$\begin{aligned} \max_{\rho \geq 0} \quad & \sum_s \rho(s) r^{\pi}(s) \\ \text{s.t.} \quad & \rho(s) = (1-\gamma)\mu_0(s) + \gamma \mathcal{T}^{\pi}\rho(s), \ \forall s\,. \end{aligned}$$

□

B Doubly Robust Estimator for Average Case

B.1 Primal Dual Framework

We start from primal dual framework to get our doubly robust estimator similar to section 4. To estimate the average reward of a given policy π, we can consider solve the following linear programming:

$$\begin{aligned} \max_{\rho \geq 0} & \sum_s \rho(s) r^{\pi}(s) \\ \text{s.t.} & \sum_s \rho(s) = 1, \quad \rho(s) = \mathcal{T}^{\pi}\rho(s), \ \forall s, \end{aligned} \tag{20}$$

where $\rho(s)$ is the stationary distribution of states under $\mathcal{P}^{\pi}$, and the objective is the average reward given π.

Consider the Lagrangian of above linear programming:

$$\begin{aligned} L(V, \rho, \bar{v}) &= \sum_s \rho(s) r^{\pi}(s) - \sum_s V(s)(\rho(s) - \mathcal{T}^{\pi}\rho(s)) - \bar{v}(\sum_s \rho(s) - 1) \\ &= \bar{v} - \sum_s \rho(s)(V(s) - r^{\pi}(s) - \mathcal{P}^{\pi}V(s) + \bar{v}). \end{aligned} \tag{21}$$

From Equation (21) we can get the dual formula as:

$$\begin{aligned} \min_{V, \bar{v}} \quad & \bar{v} \\ \text{s.t.} \quad & \bar{v} + V(s) - \mathcal{P}^{\pi}V(s) - r^{\pi}(s) \geq 0, \ \ \forall s, \end{aligned} \tag{22}$$

where $V(s)$ is the value function and $\bar{v}$ is the average reward we want to optimize.

Notice that in average case, $V^{\pi}(s)$ can be viewed as the fixed-point solution to the following Bellman equation:

$$V^{\pi}(s) - \mathbb{E}_{s', a|s \sim d_{\pi}}[V^{\pi}(s')] = \mathbb{E}_{a|s \sim \pi}[r(s, a) - \bar{v}].$$

Note that this explains the constraint and only if we pick $\bar{v} = R^{\pi}$, we can find a V to guarantee the constraint $\bar{v} + V(s) - \mathcal{P}^{\pi}V(s) - r^{\pi}(s) \geq 0$ holds true.

B.2 Doubly Robust Estimator

We want to build the doubly robust estimator via the lagrangian. However, the Lagrangian consist of three term ρ, V and $\bar{v}$. It is counter-intuitive if we already given an estimator of $\bar{v} \approx R^{\pi}$ into our estimator.

A better way to solve this problem is to remove the constraint $\sum \rho(s) = 1$, but we divide it as an self-normalization. Then our Lagrangian becomes

$$L(V, \rho) = \frac{\sum_s \rho(s) r^{\pi}(s) - \sum_s V(s)(\rho(s) - \mathcal{T}^{\pi}\rho(s))}{\sum \rho(s)}.$$

which can be utilized to define the doubly robust estimator for average reward.

Definition B.1. *Given a learned value function $\hat{V}(s)$ for policy π and an estimated density ratio $\hat{w}(s)$ for $w_{\pi/\pi_0}(s)$, we define*

$$\widehat{R}^{\pi}_{DR}[\widehat{V}, \widehat{w}] := \frac{\sum_{s,a,r,s' \in \mathcal{D}} \widehat{w}(s) \left(\beta_{\pi/\pi_0}(a|s)(r + \widehat{V}(s')) - \widehat{V}(s) \right)}{\sum_{s \in \mathcal{D}} \widehat{w}(s)},$$

where $\beta_{\pi/\pi_0}(a|s) = \frac{\pi(a|s)}{\pi_0(a|s)}$.

Similarly to Theorem 3.1 we will have our double robustness for our average doubly robust estimator:

Theorem B.2. *Suppose we have infinite samples and we can get*

$$R^{\pi}_{DR}[\widehat{V}, \widehat{w}] = \frac{\mathbb{E}_{s \sim d_{\pi_0}} \left[\widehat{w}(s) \left(r_\pi(s) - \widehat{V}(s) + \mathcal{P}^\pi \widehat{V}(s) \right) \right]}{\mathbb{E}_{s \sim d_{\pi_0}} \left[\widehat{w}(s) \right]}.$$

Then we have

$$R^{\pi}_{DR}[\widehat{V}, \widehat{w}] - R^\pi = \mathbb{E}_{s \sim d_{\pi_0}} \left[\varepsilon_{\widehat{w}}(s) \varepsilon_{\widehat{V}}(s) \right], \tag{23}$$

where $\varepsilon_{\widehat{V}}$ and $\varepsilon_{\widehat{w}}$ are errors of $\widehat{V}$ and $\widehat{w}$, respective, defined as follows

$$\varepsilon_{\widehat{w}} = \frac{\widehat{w}(s)}{\mathbb{E}_{s \sim d_{\pi_0}} \left[\widehat{w}(s) \right]} - \frac{d_\pi(s)}{d_{\pi_0}(s)}, \qquad \varepsilon_{\widehat{V}}(s) = r_\pi - \widehat{V} + \mathcal{P}^\pi \widehat{V} - R^\pi.$$

Proof. A key observation is that

$$\begin{aligned} \mathbb{E}_{s \sim d_{\pi_0}}[w_{\pi/\pi_0}(s) \varepsilon_{\widehat{V}}(s)] =& \mathbb{E}_{s \sim d_\pi}[r^\pi(s) - \widehat{V}(s) + \mathcal{P}^\pi \widehat{w}(s) - R^\pi] \\ =& \left(\mathbb{E}_{s \sim d_\pi}[r^\pi(s)] - R^\pi \right) + \mathbb{E}_{s \sim d_\pi}[-\widehat{V}(s) + \mathcal{P}^\pi \widehat{w}(s)] \\ =& 0 \end{aligned}$$

Thus we have

$$\begin{aligned} R^\pi_{\text{DR}}[\widehat{V}, \widehat{w}] - R^\pi =& \mathbb{E}_{s \sim d_{\pi_0}} \left[\frac{\widehat{w}(s)}{\mathbb{E}_{s \sim d_{\pi_0}} \left[\widehat{w}(s) \right]} \left(R^\pi + \varepsilon_{\widehat{V}}(s) \right) \right] - R^\pi \\ =& \mathbb{E}_{s \sim d_{\pi_0}} \left[\frac{\widehat{w}(s)}{\mathbb{E}_{s \sim d_{\pi_0}} \left[\widehat{w}(s) \right]} \varepsilon_{\widehat{V}}(s) \right] \\ =& \mathbb{E}_{s \sim d_{\pi_0}} \left[\frac{\widehat{w}(s)}{\mathbb{E}_{s \sim d_{\pi_0}} \left[\widehat{w}(s) \right]} \varepsilon_{\widehat{V}}(s) \right] - \mathbb{E}_{s \sim d_{\pi_0}}[w_{\pi/\pi_0}(s) \varepsilon_{\widehat{V}}(s)] \\ =& \mathbb{E}_{s \sim d_{\pi_0}} \left[\varepsilon_{\widehat{w}}(s) \varepsilon_{\widehat{V}}(s) \right]. \end{aligned}$$

□

Similar to discounted case we have $R^\pi_{\text{DR}}[\widehat{V}, \widehat{w}] = R^\pi$ if either $\widehat{w}$ or $\widehat{V}$ is accurate.

C Experimental Details

C.1 Tabular Case: Taxi

Behavior and Target Policies Choosing We use an on-policy Q-learning to get a sequence of policy $\pi_0, \pi_1, ..., \pi_{19}$ as data size increases. We pick the last policy π_{19} (almost optimum) as our target policy and π_{18} as our behavior policy to guarantee that those policies are not far away. We set our discounted factor $\gamma = 0.99$.

Train $\widehat{V}$ and $\widehat{\rho}$ Separate from testing, we use a set of independent sample to first train a value function $\widehat{V}$ and a density function $\widehat{\rho}$. Both $\widehat{V}$ and $\widehat{\rho}$ have bias due to finite sample approximation.

For training $\widehat{V}$ and $\widehat{\rho}$, we choose to use Monte Carlo method to estimate $\widehat{V}$ and $\widehat{\rho}$. We first use the finite samples to get an estimated model $\widehat{T}(s'|s,a)$ and rewards function $\widehat{r}(s,a)$ and $\widehat{d_0}$. Then we solve the following linear equation (by iteration like power method, which is actually Monte Carlo):

$$\begin{aligned}
\widehat{V}(s) &= \sum_a \pi(a|s)\widehat{Q}^\pi(s,a), \\
\widehat{Q}^\pi(s,a) &= \widehat{r}(s,a) + \gamma \sum_{s'} \widehat{T}(s'|s,a)\widehat{V}(s'), \\
\widehat{\nu}(s,a) &= \widehat{\rho}(s)\pi(a|s), \\
\widehat{\rho}(s) &= (1-\gamma)\widehat{d_0}(s) + \gamma \sum_{s,a} \widehat{T}(s'|s,a)\widehat{\nu}(s,a).
\end{aligned}$$

Estimate R^π Using $\widehat{V}$ and $\widehat{\rho}$ We put $\widehat{V}$ and $\widehat{\rho}$ into the Lagrangian as equation (15) as our doubly robust estimator. For those states we haven't visited, we set $\widehat{V}(s)$ and $\widehat{\rho}(s)$ as 0 and we self-normalized the $\widehat{\rho}$ to get a fair estimation.

C.2 Continuous States Off-Policy Evaluation

Evaluation Environments We evaluate our method on two infinite horizon environments: Puck-Mountain and InvertedPendulum.

Puck-Mountain is an environment similar to Mountain-car, except that the goal of the task is to push the puck as high as possible in a local valley whose initial position is at the bottom of the valley. If the ball reaches the top sides of the valley, it will hit a roof and changes the speed to its opposite direction with half of its original speeds. The rewards was determined by the current velocity and height of the puck.

InvertedPendulum is a pendulum that has its center of mass above its pivot point. It is unstable and without additional help will fall over. We train a near optimal policy that can make the pendulum balance for a long horizon. For both behavior and target policies, we assume they are good enough to keep the pendulum balance and will never fall down until they reach the maximum timesteps. We use the implementation from OpenAI Gym (Brockman et al., 2016) and changing the dynamic by adding some additional zero mean Gaussian noise to the transition dynamic.

Behavior and Target Policies Learning We use the open source implementation[2] of deep Q-learning to train a 32 × 32 MLP parameterized Q-function to converge. We then use the softmax policy of learned the Q-function as the target policy π, which has a default temperature $\tau = 1$. For the behavior policy π_0, we set a relative large temperature which encourages exploration. We set the temperature of the behavior policy $\tau_0 = 1.88$ for Puck-Mountain and $\tau_0 = 1.50$ for InvertedPendulum respectively.

Training of density ratio $\hat{w}(s)$ and value function $\hat{V}(s)$ We use a seperate training dataset with 200 trajectories whose horizon length is 1000 to learn the density ratio $\hat{w}(s)$ and the value function

[2] https://github.com/openai/baselines

```
Algorithm 2 Optimization of density ratio $\widehat{w}$
  Input: Transition data $\mathcal{D} = \{s_i, a_i, s_i', r_i\}_{i=1}^n$ from the behavior policy $\pi_0$; a target policy $\pi$ for which we want to estimate the expected reward. Discount factor $\gamma \in (0,1)$, starting state $\mathcal{D}_0 = \{s_j^{(0)}\}_{j=1}^m$ from initial distribution.
  Initial the density ratio $w(s) = w_\theta(s)$ to be a neural network parameterized by $\theta$, $f(s) = f_\beta(s)$ to be a neural network parameterized by $\beta$. We need to ensure that the final layer of $\theta$ is a softmax layer.
  for iteration = 1,2,...,T do
    Randomly choose a batch $\mathcal{M} \subseteq \{1, \ldots, n\}$ uniformly from the transition data $\mathcal{D}$ and a batch $\mathcal{M}_0 \subseteq \{1, \ldots, m\}$ uniformly from start states $\mathcal{D}_0$.
    for iteration = 1,2,..., K do
      Update the parameter $\beta$ by $\beta \leftarrow \beta + \epsilon_\beta \nabla_\beta \hat{L}(w_\theta, \phi_\beta)$, where
        $\hat{L}(w_\theta, f_\beta) = \frac{1}{|\mathcal{M}|}\sum_{i\in\mathcal{M}}\left((w(s_i') - \gamma w(s_i)\frac{\pi(a_i|s_i)}{\pi_0(a_i|s_i)} - \frac{1}{2}f(s_i'))f(s_i')\right) - (1-\gamma)\frac{1}{|\mathcal{M}_0|}\sum_{j\in\mathcal{M}_0} f(s_j)$
    end for
    Update the parameter $\theta$ by $\theta \leftarrow \theta - \epsilon_\theta \nabla_\theta \hat{L}(w_\theta, f_\beta)$.
  end for
  Output: the density ratio $\widehat{w} = w_\theta$.
```

```
Algorithm 3 Optimization of value function $\widehat{V}$
  Input: Transition data $\mathcal{D} = \{s_i, a_i, s_i', r_i\}_{i=1}^n$ from the behavior policy $\pi_0$; a target policy $\pi$ for which we want to estimate the expected reward. Discount factor $\gamma \in (0,1)$.
  Initial the value function $V(s) = V_\phi(s)$ to be a neural network parameterized by $\phi$, $f(s) = f_\beta(s)$ to be a neural network parameterized by $\beta$.

  for iteration = 1,2,...,T do
    Randomly choose a batch $\mathcal{M} \subseteq \{1, \ldots, n\}$ uniformly from the transition data $\mathcal{D}$.
    for iteration = 1,2,..., K do
      Update the parameter $\beta$ by $\beta \leftarrow \beta + \epsilon_\beta \nabla_\beta \hat{L}(V_\phi, \phi_\beta)$, where
        $\hat{L}(V_\phi, f_\beta) = \frac{1}{|\mathcal{M}|}\sum_{i\in\mathcal{M}}\left(\left(V_\phi(s_i) - \frac{\pi(a_i|s_i)}{\pi_0(a_i|s_i)}(r_i + \gamma V_\phi(s_i'))\right) f_\beta(s_i) - \frac{1}{2} f_\beta(s_i)^2\right)$
    end for
    Update the parameter $\phi$ by $\phi \leftarrow \phi - \epsilon_\phi \nabla_\phi \hat{L}(V_\phi, f_\beta)$.
  end for
  Output: the density ratio $\widehat{V} = V_\phi$.
```

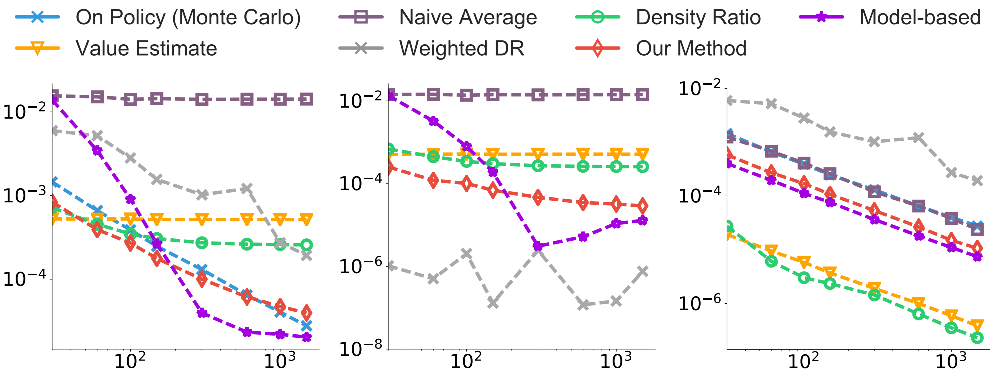
$\hat{V}(s)$. For the training of density ratio, we adapt the algorithm 2 in Liu et al. (2018a) to train a neural network parameterized $w_\theta(s)$. Instead of taking the test function $f(s)$ into an RKHS $\mathcal{H}_\mathcal{K}$, we parameterize the test function $f(s) = f_\beta(s)$ to be a neural network with parameter β, and perform minimax optimization over parametr θ and β. A detail description can be found in Algorithm 2.

Similarly, for the training of value function, we use primal-dual based optimization methods (Dai et al., 2018; Feng et al., 2019) to minimize the bellman residual:

$$\min_{\phi} \max_{f_\beta \in \mathcal{F}} \frac{1}{|\mathcal{M}|}\sum_{i\in\mathcal{M}}\left(\left(V_\phi(s_i) - \frac{\pi(a_i|s_i)}{\pi_0(a_i|s_i)}(r_i + \gamma V_\phi(s_i'))\right) f_\beta(s_i) - \frac{1}{2} f_\beta(s_i)^2\right),$$

where $V_\phi(s)$ is the parameterized value function and $f_\beta(s)$ is the test function. We also perform minimax over parameter ϕ and β. A detail description can be found in Algorithm 3.

For the network structures, we use 32×32 feed forward neural networks to parameterize value function V_ϕ and density ratio $w_\theta(s)$, and we use one hidden neural network with 10 units to parameterize the test function $f_\beta(s)$. We use Adam Optimizer for all our experiments.



(a) MSE (b) Bias Square (c) Variance

Figure 4: Additional results on Taxi, with an extra curve for model based method.

Estimate R^π using $\widehat{V}$ and $\widehat{w}$ Given data samples from the policy π_0, We can directly use $\widehat{R}^{\pi}_{\mathrm{DR}}$ in equation (11) to estimate R^π.

D Additional Experimental Results

We add a standard model-based baseline in Figure 4 following the same procedure as Liu et al. (2018a).[3] Figure 4 shows that the performance of the model-based approach largely depends of the size of the data, which is consistent with that in Liu et al. (2018a).

[3] https://github.com/zt95/infinite-horizon-off-policy-estimation.