# Doubly Robust Off-policy Evaluation for Reinforcement Learning


**Nan Jiang** NANJIANG@UMICH.EDU
Computer Science & Engineering, University of Michigan

**Lihong Li** LIHONGLI@MICROSOFT.COM
Microsoft Research



## Abstract

We study the problem of evaluating a policy that is different from the one that generates data. Such a problem, known as *off-policy evaluation* in reinforcement learning (RL), is encountered whenever one wants to estimate the value of a new solution, based on historical data, before actually deploying it in the real system, which is a critical step of applying RL in most real-world applications. Despite the fundamental importance of the problem, existing general methods either have uncontrolled bias or suffer high variance. In this work, we extend the so-called *doubly robust* estimator for bandits to sequential decision-making problems, which gets the best of both worlds: it is guaranteed to be unbiased and has low variance, and as a point estimator, it outperforms the most popular importance-sampling estimator and its variants in most occasions. We also provide theoretical results on the hardness of the problem, and show that our estimator can match the asymptotic lower bound in certain scenarios.


## 1. Introduction

We study the *off-policy evaluation* problem, in which one aims to estimate the value of a policy with data collected by another policy (Sutton & Barto, 1998). This problem is of fundamental importance in AI, and reinforcement learning (RL) in particular. For example, it is critical in intra-option learning when an agent tries to optimize the policies of different options (that is, temporally abstracted actions, sometimes referred to as skills) using the *same* stream of experience generated by some behavior policy (Sutton et al., 1998). Off-policy evaluation may also be viewed as the statistical problem of estimating the causal effect of an intervention from historical data, without actually running an experiment (Holland, 1986; Pearl, 2009).

In practice, off-policy evaluation is critical in many real-world applications of reinforcement learning, whenever it is infeasible to estimate policy value by running the policy, because doing so is expensive, risky, or unethical/illegal. In robotics and business/marketing applications, for instance, it is often risky (thus expensive) to run a policy without an estimate of the policy's quality (Li et al., 2011; Bottou et al., 2013; Thomas et al., 2015a). In medical and public-policy domains (Murphy et al., 2001; Hirano et al., 2003), often it is hard to run a controlled experiment to estimate the treatment effect. Off-policy evaluation is therefore critical in a wide range of important applications.

There are roughly two classes of approaches to off-policy evaluation. The first is to build an outcome model that predicts, given a current state and action, the reward and possibly the next states. Such a model-based approach has low variance and works well when the outcome model can be learned to satisfactory accuracy. However, for complex real-world problems, it is often hard to specify a model class which is learnable with limited data while at the same time has a small approximation error. Furthermore, it is in general impossible to estimate the approximation error of a model class, resulting in a bias in model-based estimates that cannot be easily quantified. The second class of approaches are based on the idea of importance sampling, which correct the mismatch between the distributions induced by the target policy and by the behavior policy (Precup et al., 2000). Such approaches have the salient properties of being unbiased and independent of the problem's state space, but its variance can be too large for the method to be useful, especially when the decision horizon is large.

In this work, we propose a new off-policy evaluation algorithm that can achieve the best of model-based approaches (low variance) and importance-sampling-based approaches (no bias). Our contributions include:

- A simple doubly-robust estimator is proposed for RL that subsumes a previous one for contextual bandits.

- The estimator's statistical properties are analyzed, which suggests its superiority over previous approaches. Furthermore, in certain scenarios we prove that the estimator's variance matches the Cramer-Rao lower bound for the off-policy evaluation problem.
- On benchmark problems, the new estimator is shown to be much more accurate than importance-sampling baselines, while remaining unbiased in contrast to model-based approaches. As an application, we also show how such a better estimator can benefit policy iteration to obtain improved policies with higher value.

## 2. Related Works

This paper focuses on off-policy evaluation in finite-horizon problems, which are often a more natural way to model real-world problems. The goal is to estimate expected return of start states drawn randomly from a distribution. It differs from previous work that aims to obtain the whole value function (Precup et al., 2000; 2001; Sutton et al., 2015), but the two settings share the same core difficulty of predicting long-term reward in an off-policy way.

There are two main classes of approaches to off-policy evaluation. The first is to estimate a model from data, and then evaluate a new policy against this model. Examples are Pednault et al. (2002) and variants that generates approximate models by artificial trajectories (Fonteneau et al., 2013). The second is to use importance sampling (IS) to obtain unbiased estimates of policy values (Precup et al., 2000; 2001). The main drawback of IS is its high variance, especially when the horizon is large. In this work, we use the doubly-robust (DR) technique to combine these two classes of approaches, resulting in a new off-policy estimator that is unbiased and has lower variance.

DR was first studied in statistics (Rotnitzky & Robins, 1995) to improve robustness of estimations when there can be model misspecification, and a DR estimator is developed for dynamic treatment regime (Murphy et al., 2001), to estimate the value as a function of some static side information. DR was later applied to policy learning in contextual bandits (Dudík et al., 2011), whose finite-time variance is shown to be typically better than IS. The DR estimator in this work generalizes the work of Dudík et al. (2011) to sequential decision-making problems. In addition, we show that in certain scenarios DR with a perfect model is optimal as its variances matches the lower bound.

One of the motivations to study off-policy learning is to ensure that a new policy has a good enough quality before it gets deployed. Such a safety guarantee has been studied in the literature. In conservative policy iteration (Kakade & Langford, 2002) and safe policy iteration (Pirotta et al., 2013), local adjustments to policies are made to ensure monotonic improvements over iterations. More recently, Thomas et al. (2015a) incorporates lower confidence bounds in policy iteration to ensure the computed policies in the iterations have a minimum performance guarantee. In this paper, we show that by replacing IS with DR, an agent can often deploy good policies more aggressively hence obtain higher reward, while maintaining the same level of safety against bad policies.

## 3. Background

### 3.1. Markov Decision Processes

An MDP is defined by $M = \langle S, A, P, R, \gamma \rangle$, where $S$ is the state space, $A$ is the action space, $P : S \times A \times S \to \mathbb{R}$ is the transition function where $P(s'|s,a)$ specifies the probability of seeing state $s'$ after taking action $a$ at state $s$, and $R : S \times A \to \mathbb{R}$ is the mean reward function that determines the short-term goodness of a state-action pair $(s,a)$ by $R(s,a)$. Let $\mu$ be the initial state distribution. A (stationary) policy $\pi : S \times A \to \mathbb{R}$ maps each state $s \in S$ to a distribution over actions, where $a \in A$ is assigned probability $\pi(a|s)$. Given an initial state distribution $\mu$, the discounted value of $\pi$ w.r.t. a finite horizon $H$ is defined as

$$v^{\pi,H} := \mathbb{E}_{\substack{s_1 \sim \mu, a_t \sim \pi(\cdot|s_t) \\ s_{t+1} \sim P(\cdot|s_t,a_t)}} \Big[ \sum_{t=1}^{H} \gamma^{t-1} r_t \Big],$$

where $r_t$ is the observed reward with mean $R(s_t, a_t)$. When the value of $\pi$ is conditioned on $s_1 = s$ (and $a_1 = a$), we define it as the state (and action) value function $V^{\pi,H}(s)$ (and $Q^{\pi,H}(s,a)$). In some discounted problems the true horizon is infinite, but for the purpose of policy evaluation we can still use a finite $H$ (usually set to be $O(1/(1-\gamma))$) so that $v^{\pi,H}$ approximates $v^{\pi,\infty}$ with a bounded error that diminishes as $H$ increases.

### 3.2. Off-policy Evaluation

For simplicity, we assume that the data (a set of length-$H$ trajectories) is sampled using a fixed policy[1] $\pi_0$, which we call the *behavior policy*. Our goal is to estimate $v^{\pi_1,H}$, the value of a given *target policy* $\pi_1$ from data trajectories $D = \{(s_1^{(i)}, a_1^{(i)}, r_1^{(i)}, \ldots, s_{H+1}^{(i)}), i = 1, \ldots, |D|\}$. Below we review two popular families of estimators for this off-policy evaluation problem.

**Notation Simplification** Since we will only be interested in the value of $\pi_1$ in this paper, we will omit the dependence of value functions on policy unless specified otherwise. Also, whenever the argument of a value function is a state that appears at the $t$-th step of a trajectory,

[1] All the analyses in this paper can be straight-forwardly extended to handle datasets comprising of trajectories generated using different behavior policies.

we omit the dependence on horizon and always assume that there are $H+1-t$ remaining steps. For example, $V(s_t) = V^{\pi_1, H+1-t}(s_t)$. Furthermore, all (conditional) expectations are taken with respect to the distribution induced by $(\mu, \pi_0)$, unless stated otherwise. Finally, given a trajectory $\tau = (s_1, a_1, r_1, \ldots, s_{H+1})$, we use the shorthand: $\mathbb{E}_t\big[\,\cdot\,\big] := \mathbb{E}\big[\,\cdot\,\big|\, s_1, a_1, \ldots, s_{t-1}, a_{t-1}\big]$, and similar for $\mathbb{V}_t\big[\,\cdot\,\big]$.

#### 3.2.1. Model-based Estimator

If the true parameters of the MDP are known, the value of the target policy can be computed recursively from the Bellman equations: let $V^0(s) \equiv 0$, and for $h = 1, 2, \ldots, H$,

$$\begin{aligned} Q^h(s,a) &:= \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[R(s,a) + \gamma V^{h-1}(s')\right], \\ V^h(s) &:= \mathbb{E}_{a \sim \pi_1(\cdot|s)}\left[Q^h(s,a)\right]. \end{aligned} \tag{1}$$

This suggests a two-step procedure for off-policy evaluation, which is called *model-based off-policy evaluation*: first, fit an MDP model $\widehat{M}$ from data; second, compute the value function from Equation 1 using the estimated parameters $\widehat{P}$ and $\widehat{R}$. When an exact state representation is used and each state-action pair appears sufficient number of times in the data, the model-based estimator has provably low variance and negligible bias (Mannor et al., 2007), and often outperforms other alternatives in practice (Paduraru, 2013).

However, real-world problems usually have huge or even infinitely large state spaces, and many state-action pairs will not be observed even once in the data, rendering the necessity of generalization in model fitting. To generalize over the state (and action) space, one can either choose to apply function approximators to fitting $\widehat{M}$ (Jong & Stone, 2007; Grünewälder et al., 2012), or to fitting the value function directly[2] (Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998; Dann et al., 2014). While the use of function approximators make the problem tractable, it can introduce bias to the estimated value when the MDP parameters or the value function cannot be represented in the corresponding function class. Such a bias is in general hard to quantify from data, hence breaks the credibility of estimations given by the model-based approach.

#### 3.2.2. Importance Sampling Estimators

The basic importance sampling (IS) estimator provides an unbiased estimate of $\pi_1$'s value by averaging the following function of each trajectory $(s_1, a_1, r_1, \ldots, s_{H+1})$. Let us define the per-step importance sampling ratio as $\rho_t := \pi_1(a_t|s_t)/\pi_0(a_t|s_t)$, and the cumulative importance sampling ratio $\rho_{1:t} := \prod_{t'=1}^{t} \rho_{t'}$. The basic (trajectory-wise) importance sampling estimator is

$$V_{\text{IS}} := \rho_{1:H} \cdot \Big( \sum_{t=1}^{H} \gamma^{t-1} r_t \Big). \tag{2}$$

Given a dataset $D$ containing multiple trajectories, the IS estimator is simply the average estimate over the trajectories, namely $\frac{1}{|D|}\sum_{i=1} V_{\text{IS}}^{(i)}$, where $V_{\text{IS}}^{(i)}$ is IS applied to the $i$-th trajectory. (This averaging step will be ommited for the other estimators in the rest of this paper, and we will only specify the estimate for a single trajectory).

An improved step-wise version of IS multiplies the reward at each horizon only by the cumulative importance sampling ratio preceding it, and enjoys slightly smaller variance:

$$V_{\text{step-IS}} := \sum_{t=1}^{H} \gamma^{t-1} \rho_{1:t}\, r_t. \tag{3}$$

Typically, IS (even the step-wise version) suffers from very high variance, which in general grows exponentially with horizon.

One variant of IS, weighted importance sampling (WIS), is a biased but consistent estimator, given as follows: define $w_t = \sum_{i=1}^{|D|} \rho_{1:t}^{(i)}/|D|$ as the average cumulative important ratio at horizon $t$ in a dataset $D$, then WIS estimates the return from each trajectory in $D$ as

$$V_{\text{WIS}} = \frac{\rho_{1:H}}{w_H} \Big( \sum_{t=1}^{H} \gamma^{t-1} r_t \Big). \tag{4}$$

In general, WIS has smaller variance than IS, at the cost of being biased. There is also a step-wise version for WIS, which is considered to be the most practial point estimator in the IS family (Precup, 2000; Thomas, 2015). It is obtained by applying WIS to each horizon separately:

$$V_{\text{step-WIS}} = \sum_{t=1}^{H} \gamma^{t-1} \frac{\rho_{1:t}}{w_t} r_t. \tag{5}$$

We will compare our proposed method to both step-wise IS and step-wise WIS in the experiment section.

### 3.3. Doubly Robust Estimator for Contextual Bandits

Contextual bandits may be considered as MDPs with horizon 1, with initial states $s$ drawn *i.i.d.* from some distribution $\mu$. As in the general MDP case, a sampling trajectory is in the form of $(s, a, r)$. Suppose now we are given an estimated reward function $\widehat{R}$, possibly from performing regression over a separate dataset, then the doubly robust estimator for

[2]Conventionally, fitting value functions directly without fitting an MDP model is known as value-based (or direct) RL and is distinguished from model-based (or indirect) RL. Such a difference, however, is not important for most part of this paper, therefore we use the term "model" to refer to the value function obtained by either of these two methods.

contextual bandits (Dudík et al., 2011) is defined as follows: with $\rho := \frac{\pi_1(a|s)}{\pi_0(a|s)}$ and $\widehat{V}(s) := \mathbb{E}_{a\sim\pi_1(\cdot|s)}\big[\widehat{R}(s,a)\big]$,

$$V_{\text{DR}} := \widehat{V}(s) + \rho\left(r - \widehat{R}(s,a)\right). \tag{6}$$

This estimator is known to be unbiased when $\widehat{R}$ does not depend on the sample $(s,a,r)$, and tends to have much lower variance than IS when $\widehat{R}$ approximates $R$ reasonably well (Dudík et al., 2011).

In the case where the importance ratio $\rho$ is unknown, DR estimates both $\rho$ and the reward function from data using some parametric function classes. The name "doubly robust" refers to fact that if *either* function class is properly specified, the DR estimator is asymptotically unbiased, offering two chances to ensure consistency. In this paper, however, we are only interested in DR's variance reduction benefit.

**Requirement of independence** In practice, the target policy $\pi_1$ is often computed from data, and for DR to stay unbiased, it is required that $\pi_1$ does not depend on the samples used in Equation 6; the same requirement applies to IS. While $\widehat{R}$ should be independent of such samples as well, it is not required that $\pi_1$ and $\widehat{R}$ are independent of each other. Therefore, we can use the same dataset to generate $\pi_1$ and fit $\widehat{R}$. In other situations where $\pi_1$ is given directly, to apply DR we can randomly split the data into two parts, one for fitting $\widehat{R}$ and the other for applying Equation 6. The same requirements and procedures apply to the sequential case, discussed in the next section. In the experiments, we will show that our extension of DR outperforms other estimators regardless of which dataset the model-based estimation is obtained from.

# 4. Doubly Robust Estimator for the Seqential Setting

## 4.1. Algorithm

Here, we will extend the DR estimator for bandits to the sequential case. A critical step is to rewrite Equation 3 in a recursive form. Define $V^0_{\text{step-IS}} := 0$, and for $t = 1,\ldots,H$,

$$V^{H+1-t}_{\text{step-IS}} := \rho_t\left(r_t + \gamma V^{H-t}_{\text{step-IS}}\right). \tag{7}$$

It can be shown that $V^H_{\text{step-IS}}$ is equivalent to $V_{\text{step-IS}}$ given in Equation 3. In this form, we can view the step-wise importance sampling estimator as dealing with a bandit problem at each horizon $t = 1,\ldots,H$, where $s_t$ is the context, $a_t$ is the action taken, and the observed stochastic return is $r_t + \gamma V^{H-t}_{\text{step-IS}}$, whose expected value is $Q(s_t,a_t)$. Then, if we are supplied $\widehat{Q}$, an estimation of $Q$, we can apply the bandit doubly robust estimator at each horizon, and obtain the following unbiased estimator: let $V^0_{\text{DR}} := 0$,

$$V^{H+1-t}_{\text{DR}} := \widehat{V}(s_t) + \rho_t\Big(r_t + \gamma V^{H-t}_{\text{DR}} - \widehat{Q}(s_t,a_t)\Big). \tag{8}$$

**Implementation Note** Recall that the dependence of $\widehat{V}$ and $\widehat{Q}$ on the remaining number of steps is omitted (see Section 3.2). When computed from an estimated MDP model, the value functions for different number of remaining steps may be obtained by applying Bellman update operator iteratively $H$ times starting from $\widehat{V}^0(s) \equiv 0$.

## 4.2. Variance Analysis

Below we compare the variance of step-wise IS and DR and show how DR reduces variance with a good model $\widehat{Q}$ from a theoretical perspective. Note that we only give the variance of the estimate for a *single* trajectory, and the variance for the estimate based on a dataset $D$ will be divided by $|D|$, the number of trajectories, due to the *i.i.d.* nature of $D$.

**Theorem 1.** *Step-wise importance sampling (Equation 7) is an unbiased estimator of $v^{\pi,H}$, whose variance is given in the following recursive form: for $t = 1,2,\ldots,H$,*

$$\begin{aligned}\mathbb{V}_t\big[V^{H+1-t}_{\textit{step-IS}}\big] &= \mathbb{V}_t\big[V(s_t)\big] + \mathbb{E}_t\Big[\mathbb{V}_t\big[\rho_t Q(s_t,a_t) \bigm| s_t\big]\Big] \\ &+ \mathbb{E}_t\Big[\rho_t^2\,\mathbb{V}_{t+1}\big[r_t\big]\Big] + \mathbb{E}_t\Big[\gamma^2\rho_t^2\,\mathbb{V}_{t+1}\big[V^{H-t}_{\textit{step-IS}}\big]\Big],\end{aligned} \tag{9}$$

*and $\mathbb{V}\big[V^0_{\textit{step-IS}} \bigm| s_H, a_H\big] = 0$. In comparison, Equation 8 is also an unbiased estimator of $v^{\pi,H}$ with variance given recursively by*

$$\begin{aligned}\mathbb{V}_t\big[V^{H+1-t}_{\textit{DR}}\big] &= \mathbb{V}_t\big[V(s_t)\big] + \mathbb{E}_t\Big[\mathbb{V}_t\big[\rho_t \Delta(s_t,a_t) \bigm| s_t\big]\Big] \\ &+ \mathbb{E}_t\Big[\rho_t^2\,\mathbb{V}_{t+1}\big[r_t\big]\Big] + \mathbb{E}_t\Big[\gamma^2\rho_t^2\,\mathbb{V}_{t+1}\big[V^{H-t}_{\textit{DR}}\big]\Big],\end{aligned} \tag{10}$$

*where $\Delta(s_t,a_t) = \widehat{Q}(s_t,a_t) - Q(s_t,a_t)$ and $\mathbb{V}_{H+1}\big[V^0_{\textit{DR}} \bigm| s_H, a_H\big] = 0$.*

*Proof.* We only prove using mathematical induction for the case of DR as step-wise IS can be viewed as DR's special case with $\widehat{Q}(s_t,a_t) \equiv 0$.

For the base case $t = H+1$, since $V^0_{\text{DR}} = V(s_{H+1}) = 0$, it is obvious that at the $(H+1)$-th step the estimator is unbiased with 0 variance. The theorem holds for $t = H+1$.

For the inductive step, suppose the theorem holds for step $t+1$. At time step $t$, we have:

$$\begin{aligned}&\mathbb{V}_t\big[V^{H+1-t}_{\text{DR}}\big] \\ &= \mathbb{E}_t\Big[\big(V^{H+1-t}_{\text{DR}}\big)^2\Big] - \Big(\mathbb{E}_t\big[V(s_t)\big]\Big)^2 \\ &= \mathbb{E}_t\Big[\,\Big(\widehat{V}(s_t) + \rho_t\big(r_t + \gamma V^{H-t}_{\text{DR}} - \widehat{Q}(s_t,a_t)\big)\Big)^2\end{aligned}$$

$$
\begin{aligned}
&\quad - V(s_t)^2\Big] + \mathbb{V}_t\big[V(s_t)\big] \\
&= \mathbb{E}_t\Big[\Big(\rho_t Q(s_t,a_t) - \rho_t \widehat{Q}(s_t,a_t) + \widehat{V}(s_t) \\
&\quad + \rho_t\big(r_t + \gamma V_{\text{DR}}^{H-t} - Q(s_t,a_t)\big)\Big)^2 - V(s_t)^2\Big] \\
&\quad + \mathbb{V}_t\big[V(s_t)\big] \\
&= \mathbb{E}_t\Big[\Big( - \rho_t \Delta(s_t,a_t) + \widehat{V}(s_t) + \rho_t(r_t - R(s_t,a_t)) \\
&\quad + \rho_t\gamma\big(V_{\text{DR}}^{H-t} - \mathbb{E}_{t+1}\big[V(s_{t+1})\big]\big)\Big)^2 \\
&\quad - V(s_t)^2\Big] + \mathbb{V}_t\big[V(s_t)\big] \qquad (11) \\
&= \mathbb{E}_t\Big[\mathbb{E}_t\Big[\big( - \rho_t\Delta(s_t,a_t) + \widehat{V}(s_t)\big)^2 - V(s_t)^2 \;\Big|\; s_t\Big]\Big] \\
&\quad + \mathbb{E}_t\Big[\mathbb{E}_{t+1}\big[\rho_t^2(r_t - R(s_t,a_t))^2\big]\Big] + \mathbb{V}_t\big[V(s_t)\big] \\
&\quad + \mathbb{E}_t\Big[\mathbb{E}_{t+1}\Big[\Big(\rho_t\gamma\big(V_{\text{DR}}^{H-t} - \mathbb{E}_{t+1}\big[V(s_{t+1})\big]\big)\Big)^2\Big]\Big] \\
&= \mathbb{E}_t\Big[\mathbb{V}_t\big[ - \rho_t\Delta(s_t,a_t) + \widehat{V}(s_t) \mid s_t\big]\Big] + \mathbb{E}_t\Big[\rho_t^2\, \mathbb{V}_{t+1}\big[r_t\big]\Big] \\
&\quad + \mathbb{E}_t\Big[\rho_t^2\gamma^2\, \mathbb{V}\big[V_{\text{DR}}^{H-t} \mid s_t, a_t\big]\Big] + \mathbb{V}_t\big[V(s_t)\big] \\
&= \mathbb{E}_t\Big[\mathbb{V}_t\big[\rho_t\Delta(s_t,a_t) \mid s_t\big]\Big] + \mathbb{E}_t\Big[\rho_t^2\, \mathbb{V}_{t+1}\big[r_t\big]\Big] \\
&\quad + \mathbb{E}_t\Big[\rho_t^2\gamma^2\, \mathbb{V}_{t+1}\big[V_{\text{DR}}^{H-t}\big]\Big] + \mathbb{V}_t\big[V(s_t)\big].
\end{aligned}
$$

This completes the proof. Note that from Equation 11 to the next step, we used the fact that conditioned on $s_t$ and $a_t$, $r_t - R(s_t,a_t)$ and $V_{\text{DR}}^{H-t} - \mathbb{E}_{t+1}\big[V(s_{t+1})\big]$ are independent and have zero means, and all the other terms are constants. Therefore, the square of the sum equals the sum of squares in expectation. □

### 4.3. Confidence Intervals

As mentioned in the introduction, an important motivation for off-policy evaluation is to guarantee safety before deploying a policy. For this purpose, we have to characterize the uncertainty in our estimates, usually in terms of confidence intervals (CIs). The calculation of CIs for DR is straight-forward, since DR is an unbiased estimator of *i.i.d.* trajectories and standard concentration results apply. For example, Hoeffding's inequality states that for random variables with bounded range $b$, the deviation of the average from $n$ independent samples from the expected value is at most $b\sqrt{\frac{1}{2n}\log\frac{2}{\delta}}$ with probability at least $1-\delta$. In the case of DR, $n = |D|$ is the number of trajectories, $\delta$ the chosen confidence level, and $b$ the range of the estimate, which is a function of the maximal magnitudes of $r_t$, $\widehat{Q}(s_t,a_t)$, $\rho_t$ and $\gamma$, respectively. The application of more sophisticated bounds for off-policy evaluation in RL can be found in Thomas et al. (2015a). In practice, however, strict CIs are usually too pessimistic, and normal approximations are used instead (Thomas et al., 2015b). In the experiments, we will see how DR combined with normally approximated CIs can lead to more significant safe policy improvement compared to IS.

### 4.4. Extension

From the analysis in Section 4.2, it is clear that DR only reduces the variance due to action stochasticity, and cannot reduce the variance due to the inherent randomness in reward and state transitions, therefore DR can still suffer a large variance even with a perfect model $\widehat{Q} = Q$, when the MDP has substantial stochasticity. One way to further reduce the variance in state transitions is to use the following modified version of DR:

$$
\begin{aligned}
V_{\text{DR-v2}}^{H+1-t} = \widehat{V}(s_t) + \rho_t\Big(& r_t + \gamma V_{\text{DR-v2}}^{H-t} \\
& - \widehat{R}(s_t,a_t) - \gamma\widehat{V}(s_t,a_t)\tfrac{\widehat{P}(s_{t+1}|s_t,a_t)}{P(s_{t+1}|s_t,a_t)}\Big), \qquad (12)
\end{aligned}
$$

where $\widehat{P}$ is the transition probability of the model that we use to compute $\widehat{Q}$. While it is clear this is an unbiased estimator and reduces the state transition variance with good reward & transition model $\widehat{R}$ and $\widehat{P}$ (we omit proof), the estimator is impractical as the true transition function $P$ is unknown. However, in problems where we are confident that the transition dynamics can be estimated accurately (but the estimated reward function is only approximate), we can assume that $P(\cdot|\cdot,\cdot) = \widehat{P}(\cdot|\cdot,\cdot)$, and the last term in Equation 12 becomes simply $\gamma\widehat{V}(s_t,a_t)$, which is now computable. This introduces a small bias in the estimate, but reduces more variance than the original DR. In Section 6.1.3 we will use an experiment to demonstrate the use of such an estimator.

## 5. Lower Bound for Off-policy Evaluation

In Section 4.4, we discussed the possibility of reducing variance due to state transition stochasticity in a special scenario where transition function can be estimated accurately with high confidence, which does not happen very often in practice. A natural question to ask is whether there exists some estimator that can reduce such variance, without relying on extra strong assumptions.

In this section, we provide a hardness result on the off-policy evaluation problem, and show that when no prior knowledge of the domain is available, the lower bound of variance of any estimator equals the variance of DR with a perfect model. Furthermore, DR's variance gracefully degrades when the model error increases; in other words, DR smoothly bridges the gap between simple IS, with a trivial zero value function, and an optimal estimator, with a perfect value function.

Before stating the theorem, we first clarify what we mean by "no prior knowledge is known for the domain". In particular, it means that whether Markovian assumption holds, i.e., whether the last observation is automatically a sufficient statistics of history (namely state), is unknown, and the only valid state representation is the history itself. We capture such a most general class of domains by the following definition of discrete tree MDPs, with some simplifications.

**Definition 1.** An MDP is a *discrete tree MDP* if it satisfies the following conditions:

- State is represented by a sequence of alternating *observations* and actions, namely $s = o_1 a_1 \dots o_{t-1} a_{t-1} o_t$, or abbreviated as $s = h_t$. We assume discrete observations and actions.
- Initial states take the form of $s = o_1$. Upon taking action $a$, a state $s = h$ can only transition to a next state in the form of $s' = hao$ (*i.e.*, $h$ appended with $a$ and $o$), and the transition probability is given by the parameter $P(o|h, a)$.
- As a simplification, we assume that $\gamma = 1$, and non-zero rewards only occur at the end of each $H$-step long trajectory.[3] We use an additional observation $o_{H+1}$ to encode the reward randomness so that reward function $R(h_{H+1})$ is deterministic and the domain can be solely parameterized by transition probabilities.

**Theorem 2.** *For discrete tree MDPs, the variance of any unbiased estimator is lower bounded by*

$$\sum_{i=1}^{H+1} \mathbb{E}\Big[\rho_{1:(t-1)}^2 \mathbb{V}_t\big[V(s_t)\big]\Big]. \tag{13}$$

*The bound also applies to the asymptotic mean square error of biased estimators.*

**Observation 1.** *The variance of DR applied to a discrete tree MDP with $\widehat{Q} = Q$ is equal to Equation 13.*

The theorem follows from Cramer-Rao bound (CRB) for the off-policy evaluation problem, and the claim follows directly by unfolding the recursive form of Equation 10 and noticing that $\Delta \equiv 0$, $\mathbb{V}_{t+1}\big[r_t\big] \equiv 0$ for $t = 1, \dots, H-1$, and $\mathbb{V}_{H+1}\big[V(s_{H+1})\big]$ is just a re-writing of $\mathbb{V}_{H+1}\big[r_H\big]$.

*Proof of Theorem 2.* We parameterize the discrete tree MDPs by $\mu(o)$ and $P(o|h, a)$ for $h$ of length $1, \dots, H$. For convenience we will treat $\mu(o)$ as $P(o|\emptyset)$, so all the parameters can be represented as $P(o|h, a)$ where $ha$ can contain $0, \dots, H$ alternating observations & actions (and we call this number the length of $ha$, or $|ha|$). These parameters are subject to the normalization constraints that have to be taken into consideration in the Cramer-Rao bound, namely $\forall h, a, \sum_{o \in O} P(o|h, a) = 1$. If we sort the parameters $P(o|h, a)$ in the alphabetical order of $ha$ and then $o$, then the constraints can be written as

$$\begin{bmatrix} 1 \cdots 1 & & & \\ & 1 \cdots 1 & & \\ & & \ddots & \\ & & & 1 \cdots 1 \end{bmatrix} \theta = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \tag{14}$$

where $\theta_{hao} = P(o|h, a)$. The matrix on the left is effectively the Jacobian of the constraints, which we denote as $F$. We index its rows by $ha$, so $F_{(ha),(hao)} = 1$ and other entries are 0. Let $U$ be a matrix whose column vectors consist an orthonormal basis for the null space of $F$. From Equation (3.3) and Corollary 3.10 in Moore Jr (2010), we have the Constrained Cramer-Rao Bound (CCRB) being[4] (the dependence on $\theta$ in all terms are omitted):

$$KU(U^\top IU)^{-1} U^\top K^\top, \tag{15}$$

where $I$ is the Fisher Information Matrix (FIM), and $K$ is the Jacobian of the quantity we want to estimate; they are computed below. We start with $I$, which is

$$I = \mathbb{E}\Big[ \left(\frac{\partial \log P_0(h_{H+1})}{\partial \theta}\right) \left(\frac{\partial \log P_0(h_{H+1})}{\partial \theta}\right)^\top \Big], \tag{16}$$

where

$$P_0(h_{H+1}) = \mu(o_1)\pi_0(a_1|o_1)P(o_2|o_1, a_1) \dots P(o_{H+1}|h_H, a_H).$$

To calculate $I$, we define a new notation $g(h_{H+1})$, which is a vector of indicator functions and $g(h_{H+1})_{hao} = 1$ when $hao$ is a prefix of $h_{H+1}$. Using this notation, we have

$$\frac{\partial \log P_0(h_{H+1})}{\partial \theta} = \theta^{\circ -1} \circ g(h_{H+1}), \tag{17}$$

where $\circ$ denotes element-wise power/multiplication. Then we can rewrite the FIM as

$$\begin{aligned} I &= \mathbb{E}\Big[[\theta_i^{-1}\theta_j^{-1}]_{ij} \circ (g(h_{H+1})g(h_{H+1})^\top)\Big] \\ &= [\theta_i^{-1}\theta_j^{-1}]_{ij} \circ \mathbb{E}\big[(g(h_{H+1})g(h_{H+1})^\top)\big], \end{aligned} \tag{18}$$

where $[\theta_i^{-1}\theta_j^{-1}]_{ij}$ is a matrix expressed by its $(i, j)$-th element. Now we compute $\mathbb{E}\big[(g(h_{H+1})g(h_{H+1})^\top)\big]$. This matrix takes 0 in all the entries indexed by $hao$

[3] We suspect that this simplification does not affect the generality of our theoretical results: for problems with random rewards at each step, we can encode the reward randomness as part of observation as long as the reward distribution is discrete, and delay the sum of discounted rewards to the end of trajectory. Both the Cramer-Rao bound and the variance of DR with a perfect model seem to be invariant to these transformations.

[4] In fact, existing literature on Contrained Cramer-Rao Bound does not deal with the situation where the unconstrained parameters break the normalization constraints (which we are facing). However, this can be easily tackled by changing the model slightly to $P(o|h, a) = \theta_{hao} / \sum_{o'} \theta_{hao'}$, which resolves the issue and gives the same result.

and $h'a'o'$ when neither of the two strings is a prefix of the other (because, for any observed trajectory, either $g(h_{H+1})_{hao}$ or $g(h_{H+1})_{h'a'o'}$ is 0). For the rest entries, without loss of generality assume $h'a'o'$ is a prefix of $hao$. Since $g(h_{H+1})_{hao}g(h_{H+1})_{h'a'o'} = 1$ if and only if $hao$ is a prefix of $h_{H+1}$, we have $\mathbb{E}\big[g(h_{H+1})_{hao}g(h_{H+1})_{h'a'o'}\big] = P_0(hao)$, and consequently $I_{hao,h'a'o'} = \frac{P_0(hao)}{P(o|h,a)P(o'|h',a')} = \frac{P_0(ha)}{P(o'|h',a')}$ (note the independence on $o$, which is crucial for $I$ to be diagonalizable; see below).

Then, we calculate $(U^\top IU)^{-1}$. To avoid the difficulty of taking inverse, we apply the following trick to diagonalize $I$: note that for any matrix $X$ with matching dimensions,

$$U^\top IU = U^\top(F^\top X^\top + I + XF)U, \tag{19}$$

because by definition $U$ is orthogonal to $F$. We can design $X$ so that $D = F^\top X^\top + I + XF$ is a diagonal matrix, and $D_{(hao),(hao)} = I_{(hao),(hao)} = \frac{P_0(ha)}{P(o|h,a)}$. This is achieved by having $XF$ eliminate all the non-diagonal entries of $I$ in the upper triangle without touching anything on the diagonal or below, and by symmetry $F^\top X^\top$ will deal with the lower triangle. The construction is as follows: let $X_{(h'a'o'),(ha)} = 0$ except when $h'a'o'$ is a prefix of $ha$, in which case we set $X_{(h'a'o'),(ha)} = -\frac{P_0(ha)}{P(o'|h',a')}$. It is not hard to verify that this construction diagonalizes $I$.

With the diagonalization trick, we have $(U^\top IU)^{-1} = (U^\top DU)^{-1}$. Since CCRB is invariant to the choice of $U$, and we observe that the rows of $F$ are orthogonal, we choose $U$ as follows: let $n_{(ha)}$ be the number of 1's in $F_{(ha),(\cdot)}$, and $U_{(ha)}$ be the $n_{(ha)}\times(n_{(ha)}-1)$ matrix with orthonormal columns in the null space of $\begin{bmatrix}1 & \dots & 1\end{bmatrix}$ ($n_{(ha)}$ 1's); finally, we choose $U$ to be a block diagonal matrix $U = \mathrm{diag}(\{U_{(ha)}\})$, where $U_{(ha)}$'s are the diagonal blocks, and it is easy to verify that $U$ is column orthonormal and $FU = 0$. Similarly, we write $D = \mathrm{diag}(\{D_{(ha)}\})$ where $D_{(ha)}$ is a diagonal matrix with $(D_{(ha)})_{oo} = P_0(ha)/P(o|h,a)$, and

$$\begin{aligned} &U(U^\top IU)^{-1}U^\top = U(U^\top DU)^{-1}U^\top \\ &= U(\mathrm{diag}(\{U_{(ha)}^\top\})\mathrm{diag}(\{D_{(ha)}\})\mathrm{diag}(\{U_{(ha)}\}))^{-1}U \\ &= U\mathrm{diag}(\{\big(U_{(ha)}^\top D_{(ha)}U_{(ha)}\big)^{-1}\})U \\ &= \mathrm{diag}(\{U_{(ha)}\big(U_{(ha)}^\top D_{(ha)}U_{(ha)}\big)^{-1}U_{(ha)}^\top\}). \end{aligned} \tag{20}$$

Notice that each block in Equation 20 is simply $1/P_0(ha)$ times the CCRB of a multinomial distribution $P(\cdot|h,a)$. The CCRB of a multinomial distribution $p$ can be easily computed by an alternative formula (Equation (3.12) in (Moore Jr, 2010)), which gives $\mathrm{diag}(p) - pp^\top$, so we have,

$$\begin{aligned} &U_{(ha)}\big(U_{(ha)}^\top D_{(ha)}U_{(ha)}\big)^{-1}U_{(ha)}^\top \\ &= \frac{\mathrm{diag}(P(\cdot|h,a)) - P(\cdot|h,a)P(\cdot|h,a)^\top}{P_0(ha)}. \end{aligned} \tag{21}$$

We then turn to the calculation of $K$. Recall that we want to estimate

$$\begin{aligned} v = v^{\pi_1,H} = \sum_{o_1}\mu(o_1)\sum_{a_1}\pi_1(a_1|o_1)\dots \\ \sum_{o_{H+1}} P(o_{H+1}|h_H,a_H)R(h_{H+1}), \end{aligned} \tag{22}$$

and its Jacobian is $K = (\partial v/\partial\theta)^\top$, with $K_{(hao)} = P_1(ha)V(hao)$, where $P_1(o_1a_1\dots o_ta_t) = \mu(o_1)\pi_1(a_1)\dots P(o_t|h_{t-1},a_{t-1})\pi_1(a_t|h_t)$.

Finally, putting all the pieces together, we have Equation 15 equal to

$$\begin{aligned} &\sum_{ha}\frac{P_1(ha)^2}{P_0(ha)}\Big(\sum_o P(o|h,a)V(hao)^2 \\ &\qquad - \big(\sum_o P(o|h,a)V(hao)\big)^2\Big) \\ &= \sum_{t=0}^{H}\sum_{|ha|=t}P_0(ha)\frac{P_1(ha)^2}{P_0(ha)^2}\mathbb{V}\big[V(hao)\bigm| h,a\big]. \end{aligned}$$

Noticing that $P_1(ha)/P_0(ha)$ is the cumulative importance ratio, and $\sum_{|ha|=t}P_0(ha)(\cdot)$ is taking expectation over sample trajectories, we have the above expression equal to

$$\sum_{t=0}^{H}\mathbb{E}\Big[\rho_{1:t}^2\mathbb{V}_{t+1}\big[V(s_{t+1})\big]\Big] = \sum_{t=1}^{H+1}\mathbb{E}\Big[\rho_{1:(t-1)}^2\mathbb{V}_t\big[V(s_t)\big]\Big].$$

And this completes the proof. □

## 6. Experiments

Throughout this section, we will be concerned with the comparison among the following estimators. For compactness, in this section, we drop the prefix "step-wise" from step-wise IS & WIS.

1. (IS) Step-wise importance sampling of Equation 3;
2. (WIS) Step-wise weighted importance sampling of Equation 5;
3. (model) MDP Model estimated from data (specified later for each domain)
4. (DR) Doubly robust estimator of Equation 8
5. (DR-bsl) Doubly robust using a constant guess.

### 6.1. Comparison of Mean Squared Errors

In this experiment, we evaluate the quality of each method mentioned above as point estimates by looking at the mean squared errors of their estimates. For each domain, a policy $\pi_{\text{train}}$ is computed as the optimal policy of the MDP model estimated from a training dataset $D_{\text{train}}$ (generated using $\pi_0$), and the target policy $\pi_1$ is set to be $(1-\alpha)\pi_{\text{train}} + \alpha\pi_0$ for

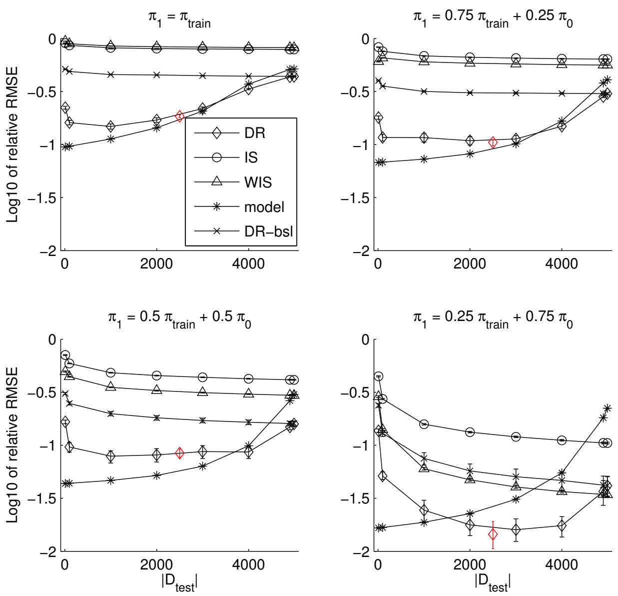


*Figure 1.* Comparison of point estimators on Mountain Car. 5000 trajectories are generated for off-policy evaluation, and all the results are from over 4000 re-draws. The subgraphs correspond to the target policies produced by mixing $\pi_{\text{train}}$ and $\pi_0$ with different portions. X-axis shows the size of $D_{\text{test}}$, the part of the data used for IS/WIS/DR. The remaining part of the data is used to construct the model, which DR uses to calculate $\widehat{Q}$. The y-axis shows the root mean square error of the estimators divided by the true value of the target policy in logarithmic scale. We also show the error of 2-fold DR as an isolated point (◇). Since the estimators in the IS family typically has a highly skewed distribution, the estimates can occasionally go largely out of range, and we crop such outliers in $[V_{\min}, V_{\max}]$ to ensure that we can get statistically significant experiment results within a reasonable number of simulations.

$\alpha \in \{0, 0.25, 0.5, 0.75\}$ respectively. Such a *mixture* policy is commonly used to yield an improved policy from an existing one; *e.g.*, Kakade & Langford (2002). The parameter $\alpha$ controls similarity between $\pi_0$ and $\pi_1$. We then apply the estimators mentioned above to a separate dataset $D_{\text{eval}}$ to estimate the value of $\pi_1$, and average the estimation errors across multiple draws of $D_{\text{eval}}$. Note that for the doubly robust estimator, the estimated model should be independent of data used in Equation 8 to ensure unbiasedness. We therefore split $D_{\text{eval}}$ into two two subsets $D_{\text{model}}$ and $D_{\text{test}}$, build model on $D_{\text{model}}$, and apply the estimator to $D_{\text{test}}$. We pick several split points and report the error on each of them.

The DR estimator described above does not make full use of data, since only part of the data ($D_{\text{test}}$) is used to compute the Monte Carlo sample average in Equation 8. We also propose a more data-efficient version of DR when $\widehat{Q}$ has to be estimated from $D_{\text{eval}}$, namely $k$-fold DR, which is inspired by $k$-fold cross validation in supervised learning: we partition $D_{\text{eval}}$ into $k$ subsets, apply Equation 6 to each subset with $\widehat{Q}$ estimated from the union of the rest $k-1$ subsets, and finally average the estimate. Since the estimate from each subset is unbiased, the overall average is unbiased and potentially has lower variance since the final estimate is averaged over all trajectories (though the estimate for each trajectory is not i.i.d.). Since model estimation is expensive, we only show the result of 2-fold DR in our experiments.

The detailed experiment settings on each domain is below.

#### 6.1.1. Mountain Car

**Domain Description** The mountain car problem (Singh & Sutton, 1996) is a continuous control problem with deterministic dynamics. The state space is $[-1.2, 0.6] \times [-0.07, 0.07]$, and there are 3 discrete actions. The agent receives $-1$ reward every time step with a discount factor $0.99$, and an episode terminates when the first dimension of state reaches the right boundary. The initial state distribution is set to be uniformly random, and behavior policy is uniformly random over the 3 actions. The typical horizon for this problem is $400$, which can be too large for IS and its variants, therefore we accelerate the dynamics in such a way that given $(s, a)$, the next state $s'$ is obtained by calling the original transition function $4$ times, and we set the horizon to $100$. A similar modification was taken by Thomas (2015), where every $20$ steps are compressed as one step.

**Model Construction** The model we construct for this domain uses a simple discretization (state aggregation): the two dimensions are multiplied by $2^6$ and $2^8$ respectively and the rounded integers are treated as the abstract state. We then estimate the model parameters from data using a tabular approach. Unseen abstract state-action pairs are assumed to have reward $R_{\min}$ and a self-loop transition. Both the model that produces $\pi_{\text{train}}$ and that used on $D$ are constructed in this way. On the other hand, DR-bsl uses the step-dependent constant $\widehat{Q}(s_t, a_t) = R_{\min}(1-\gamma^t)/(1-\gamma)$.

**Other** The dataset sizes are $|D_{\text{train}}| = 2000$ and $|D_{\text{eval}}| = 5000$. We split $D_{\text{eval}}$ such that $D_{\text{test}} \in \{10, 100, 1000, 2000, 3000, 4000, 4900, 4990\}$.

**Results** See Figure 1. As $|D_{\text{test}}|$ increases, IS/WIS gets monotonically better, while the model gets worse as less data is put into model estimation ($D_{\text{model}}$). Since DR's performance depends on both halves of the data, it achieves the best error at some intermediate value of $|D_{\text{test}}|$, and outperforms using all the data for IS/WIS in all the $4$ graphs. DR-bsl shows the performance of DR with a constant guess as $\widehat{Q}$, which already outperforms IS/WIS most of the time.

### 6.1.2. SAILING

**Domain Description** The sailing domain (Kocsis & Szepesvári, 2006) is a stochastic shortest-path problem, where the agent sails on a grid (in our experiment, a map of size $10 \times 10$) with wind blowing in random directions, aiming at the terminal location on the top-right corner. The state is represented by 4 integer variables, representing either location or direction. At each step, the agent chooses to move in one of the 8 directions (moving against the wind or running off the grid is prohibited), and receives a negative reward (cost) that depends on moving direction, wind direction, and other factors, ranging between $R_{\min} = -3 - 4\sqrt{2}$ and $R_{\max} = 0$ (absorbing). The problem is non-discounting, and we use $\gamma = 0.99$ in planning for easy convergence when computing $\pi_{\text{train}}$.

**Model Construction** Unlike Mountain Car, valid actions vary from state to state in Sailing, so it is difficult to apply state aggregations. Instead, we apply Kernel-based Reinforcement Learning (Ormoneit & Sen, 2002) to generalize by supplying a smoothing kernel in the joint space of states and actions. The kernel we use takes the form $\exp(\|\cdot\|/b)$, where $\|\cdot\|$ is the $\ell_2$-distance in $S \times A$,[5] and $b$ is the kernel bandwidth, set to 0.25.

**Other** The data sizes are $|D_{\text{train}}| = 1000$ and $|D_{\text{eval}}| = 2500$, and we split $D_{\text{eval}}$ such that $D_{\text{test}} \in \{5, 50, 500, 1000, 1500, 2000, 2450, 2495\}$. DR-bsl uses the step-dependent constant $\widehat{Q}(s_t, a_t) = R_{\min}/2 \cdot (1 - \gamma^t)/(1-\gamma)$, for the reason that in Sail $R_{\min}$ is rarely reached hence too pessimistic as a rough estimate of the value.

**Results** See Figure 2. The results are qualitatively similar to Mountain Car results in Figure 1, except that: (1) WIS is as good as DR in the 2nd and 3rd graph; (2) in the 4th graph, DR with a 3:2 split outperforms all the other estimators, including the model, with a significant margin, and a further improvement is achieved by using 2-fold DR.

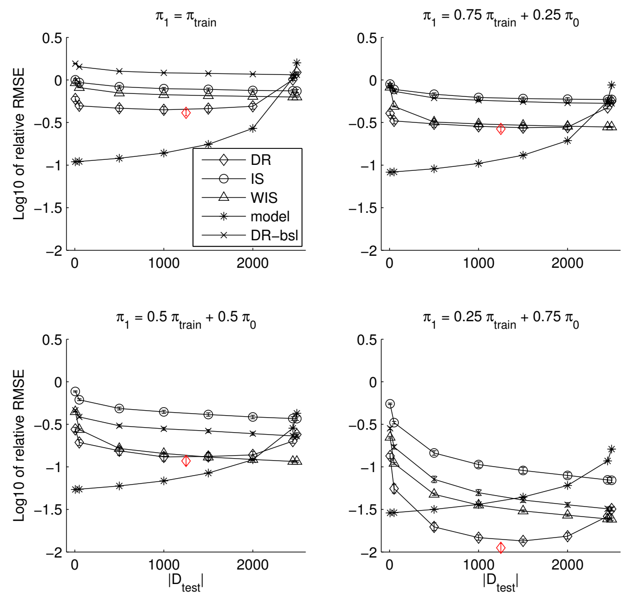


*Figure 2.* Comparison of point estimators on Sailing. 2500 trajectories are generated for off-policy evaluation, and all the results are from over 4000 re-draws. Other details are the same as Figure 1.

### 6.1.3. KDD CUP 1998 DONATION DATASET

In our last domain, we use the donation dataset from KDD Cup 1998 (Hettich & Bay, 1999), which summarizes the sequential interactions between the agent and the potential donators via emails, with the reward being donated money. The state is represented using 5 integer features, and there are 12 discrete actions representing sending different types of emails at a time period, or not sending emails at all. All trajectories are 22-steps long ($H = 22$), and there is no discount. Since this problem only provides data and no groundtruth of the target policy's value is provided, we fit a simulator from the true data (see details in the next paragraph), and use the simulator as groundtruth for everything henceforward: the true value of the target policy is computed by rolling out Monte-carlo trajectories in the simulator, and the trajectories used for off-policy evaluation are also generated from the simulator with uniformly random behavior policy, with the number of trajectories equal to the size of the original dataset (3754 trajectories). The other settings are the same as the experiments in Section 6.1, except that we replace DR with DR-v2 (Equation 12; assuming $\widehat{P} = P$) and use the 2-fold trick. The policy $\pi_{\text{train}}$ is generated by training a recurrent neural network on the original data to fit the Q-function.

The MDP model used to compute $\widehat{Q}$ is estimtated as follows: for the transition dynamics, each state variable is assumed to involve independently, and the marginal transition probabilities are estimated using a tabular approach, which is exactly how the simulator is fit from real data. Reward function, on the other hand, is estimated by linear regression for each action separately, with the first 3 dimensions of state as features (on the contrast, all the 5 dimensions are used when fitting the simulator's reward function from real data). Consequently, we get a model with an almost perfect transition function and a relatively inaccurate reward function. The results are shown in Figure 3, where DR-v2 is the best estimator in all situations: it beats WIS when $\pi_1$ is far from $\pi_0$, and beats the model-based $\widehat{Q}$ when $\pi_1$ and $\pi_0$ are close.

[5]The difference of two directions is defined as the angle between them (in degrees) divided by 45°. For computational efficiency, the kernel function is cropped to 0 whenever two state-action pairs deviate more than 1 in any of the dimensions.

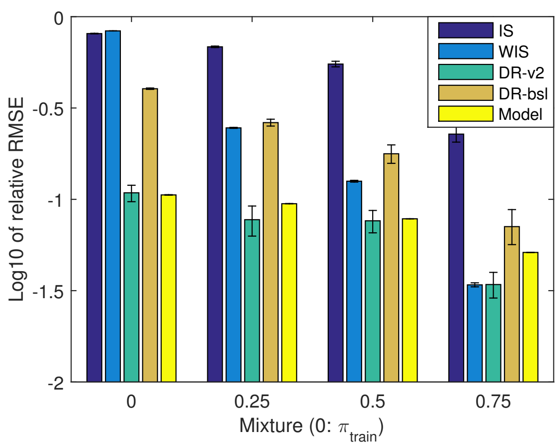


*Figure 3.* Results of donation data, averaged over about 5000 runs. The same legend as Figures 1&2 is used, except for the new estimator DR-v2 given in Equation 12. DR-v2 uses the 2-fold trick; the whole dataest is applied to other estimators. Since there are many possible next-states for each state-action pair, for computational efficiency we use a sparse-sample approach when estimating $\widehat{Q}$ using the fitted model $\widehat{M}$: for each $(s, a)$, we randomly sample several next-states from $\widehat{P}(\cdot|s, a)$, and cache them as a particle representation for the next-state distribution. The number of particles is set to 5 which is enough to ensure high accuracy.

### 6.2. Application to Safe Policy Improvement

In this experiment, we apply the off-policy evaluation methods in safe policy improvement. Given a batch dataset $D$, the agent uses part of it ($D_{\text{train}}$) to find some candidate policies, some or all of which could be bad due to insufficient sample and/or approximation in learning, and recommending such bad policies are considered undesired in many scenarios. Hence, the agent estimate the value of the candidate policies on the rest data ($D_{\text{test}}$) via off-policy evaluation, and recommend the policy with the highest estimated value. In this problem, we have an additional reason to favor DR: $D_{\text{train}}$ can be reused to estimate $\widehat{Q}$, hence it is no longer necessary to hold out part of $D_{\text{test}}$ for model construction.

Due to the high variance of the importance sampling estimators and its variants, acting greedily w.r.t. the point estimate is not enough to promote safety, and we also want to take the estimation uncertainty into consideration. Therefore, we select the policy that has the highest lower confidence bound using Student's t-test approximation (Thomas et al., 2015b),[6] and hold on to the current behavior policy if none of the lower confidence bounds is better than the threshold. More specifically, the score is $V_\dagger(D_{\text{test}}; \pi) - C\sigma_\dagger(D_{\text{test}}; \pi)$, where $\sigma$ is the empirical standard error, $C$ is a hyperparameter that controls confidence level, and $\dagger$ is a placeholder for any method that can provide confidence intervals. Among the methods considered in this paper, importance sampling and doubly robust estimators satisfy this criterion, while the model-based and WIS estimators have unknown bias.

[6] Following their paper, we normalize all the rewards into $[0, 1]$ by the transformation $(r - R_{\min})/(R_{\max} - R_{\min})$ when computing the confidence interval.

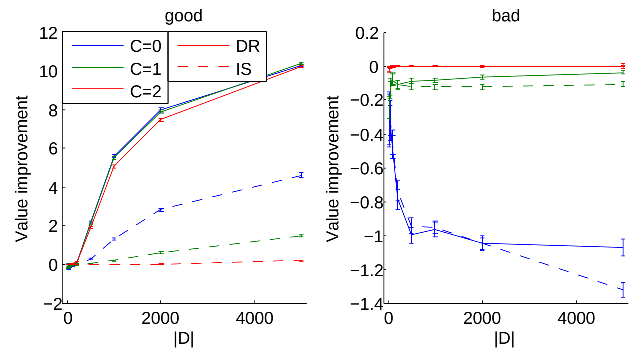


*Figure 4.* Safe policy improvement in Mountain Car. X-axis shows the size of data and y-axis shows the true value of the recommended policy subtracted by the value of the behavior policy.

The experiment is conducted in Mountain Car, and most of the setting is the same as in Section 6.1.1. Since we do not address the exploration and exploitation problem, we keep the behavior policy fixed as uniformly random, and evaluate the recommended policy once a while as the agent gets more and more data. In particular, the datasizes are $|D| \in \{20, 50, 100, 200, 500, 1000, 2000, 5000\}$. The candidate policies are generated as follows: first, we consider different ways of splitting $|D|$ so that $|D_{\text{train}}|/|D| \in \{0.2, 0.4, 0.6, 0.8\}$; for each split, we compute optimal $\pi_{\text{train}}$ from the model estimated from $D_{\text{train}}$, and mix it with $\pi_0$ with rate $\alpha \in \{0, 0.1, \ldots, 0.9\}$, and altogether we get 40 candidate policies.

The results are shown in the left panel of Figure 4. From the figure, it is clear that DR's value improvement largely outperforms IS, primarily because IS is not able to accept a target policy that is too different from $\pi_0$. However, in this situation $\pi_{\text{train}}$ is mostly a good policy (except when $|D|$ is very small) hence the more aggressive an algorithm is, the more value it gets. As an evidence, both algorithms achieve the best value with $C = 0$ (except when $|D|$ is very small), raising the suspicion that DR might make unsafe recommendations when $\pi_{\text{train}}$ is bad. To eliminate such a concern, we conduct another experiment in parallel, where we have $\pi_{\text{train}}$ minimizing the value instead of maximizing it, resulting in policies worse than the target policy, and the results are shown in the right panel. Here, we see that as $C$ becomes smaller, the algorithms become less and less safe, and with the same $C$ DR is as safe as IS if not better at $|D| = 5000$. Overall, we conclude that DR can be a drop-in replacement for IS in safe policy improvement.

## References


Bertsekas, Dimitri P and Tsitsiklis, John N. *Neuro-Dynamic Programming (Optimization and Neural Computation Series, 3)*. Athena Scientific, 1996.

Bottou, Léon, Peters, Jonas, Quiñonero-Candela, Joaquin, Charles, Denis Xavier, Chickering, D. Max, Portugaly, Elon, Ray, Dipankar, Simard, Patrice, and Snelson, Ed. Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising. *Journal of Machine Learning Research*, 14:3207–3260, 2013.

Dann, Christoph, Neumann, Gerhard, and Peters, Jan. Policy evaluation with temporal differences: A survey and comparison. *The Journal of Machine Learning Research*, 15(1):809–883, 2014.

Dudík, Miroslav, Langford, John, and Li, Lihong. Doubly Robust Policy Evaluation and Learning. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 1097–1104, 2011.

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

Grünewälder, S, Lever, G, Baldassarre, L, Pontil, M, and Gretton, A. Modelling transition dynamics in MDPs with RKHS embeddings. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 535–542, 2012.

Hettich, S. and Bay, S. D. The UCI KDD Archive. http://kdd.ics.uci.edu, 1999.

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

Holland, Paul W. Statistics and causal inference. *Journal of the American Statistical Association*, 81(6):945–960, 1986.

Jong, Nicholas K and Stone, Peter. Model-based function approximation in reinforcement learning. In *Proceedings of the 6th International Conference on Autonomous Agents and Multiagent Systems*, pp. 95, 2007.

Kakade, Sham and Langford, John. Approximately Optimal Approximate Reinforcement Learning. In *Proceedings of the 19th International Conference on Machine Learning*, volume 2, pp. 267–274, 2002.

Kocsis, Levente and Szepesvári, Csaba. Bandit based monte-carlo planning. In *Machine Learning: ECML 2006*, pp. 282–293. Springer Berlin Heidelberg, 2006.

Li, Lihong, Chu, Wei, Langford, John, and Wang, Xuanhui. Unbiased Offline Evaluation of Contextual-bandit-based News Article Recommendation Algorithms. In *Proceedings of the 4th International Conference on Web Search and Data Mining*, pp. 297–306, 2011.

Mannor, Shie, Simester, Duncan, Sun, Peng, and Tsitsiklis, John N. Bias and variance approximation in value function estimates. *Management Science*, 53(2):308–322, 2007.

Moore Jr, Terrence Joseph. *A theory of Cramer-Rao bounds for constrained parametric models*. PhD thesis, University of Maryland, College Park, 2010.

Murphy, Susan A., van der Laan, Mark, and Robins, James M. Marginal Mean Models for Dynamic Regimes. *Journal of the American Statistical Association*, 96(456):1410–1423, 2001.

Ormoneit, Dirk and Sen, Śaunak. Kernel-based reinforcement learning. *Machine learning*, 49(2-3):161–178, 2002.

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

Pearl, Judea. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009. ISBN 052189560X.

Pednault, Edwin P. D., Abe, Naoki, and Zadrozny, Bianca. Sequential cost-sensitive decision-making with reinforcement learning. In *KDD*, pp. 259–268, 2002.

Pirotta, Matteo, Restelli, Marcello, Pecorino, Alessio, and Calandriello, Daniele. Safe policy iteration. In *Proceedings of the Thirtieth International Conference on Machine Learning (ICML-13)*, number 3, pp. 307–317, 2013.

Precup, Doina. *Temporal abstraction in reinforcement learning*. PhD thesis, University of Massachusetts Amherst, 2000.

Precup, Doina, Sutton, Richard S., and Singh, Satinder P. Eligibility Traces for Off-Policy Policy Evaluation. In *Proceedings of the 17th International Conference on Machine Learning*, pp. 759–766, 2000.

Precup, Doina, Sutton, Richard S., and Dasgupta, Sanjoy. Off-Policy Temporal-Difference Learning with Funtion Approximation. In *Proceedings of the 18th International Conference on Machine Learning*, pp. 417–424, 2001.

Rotnitzky, Andrea and Robins, James M. Semiparametric regression estimation in the presence of dependent censoring. *Biometrika*, 82(4):805–820, 1995.

Singh, Satinder P and Sutton, Richard S. Reinforcement learning with replacing eligibility traces. *Machine learning*, 22(1-3):123–158, 1996.

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

Sutton, Richard S., Precup, Doina, and Singh, Satinder P. Intra-option Learning about Temporally Abstract Actions. In *Proceedings of the 15th International Conference on Machine Learning*, pp. 556–564, 1998.

Sutton, Richard S., Mahmood, Ashique Rupam, and White, Martha. An emphatic approach to the problem of off-policy temporal-difference learning. CoRR abs/1503.04269, 2015.

Thomas, Philip. *Safe Reinforcement Learning*. PhD thesis, University of Massachusetts Amherst, 2015.

Thomas, Philip, Theocharous, Georgios, and Ghavamzadeh, Mohammad. High Confidence Off-Policy Evaluation. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 2015a.

Thomas, Philip, Theocharous, Georgios, and Ghavamzadeh, Mohammad. High Confidence Policy Improvement. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 2380–2388, 2015b.