Doubly Robust Off-policy Value Evaluation for Reinforcement Learning

Nan Jiang
Computer Science & Engineering, University of Michigan

NANJIANG@UMICH.EDU

Lihong Li
Microsoft Research

LIHONGLI@MICROSOFT.COM

Abstract

We study the problem of *off-policy value evaluation* in reinforcement learning (RL), where one aims to estimate the value of a new policy based on data collected by a different policy. This problem is often a critical step when applying RL to real-world problems. Despite its importance, existing general methods either have uncontrolled bias or suffer high variance. In this work, we extend the *doubly robust* estimator for bandits to sequential decision-making problems, which gets the best of both worlds: it is guaranteed to be unbiased and can have a much lower variance than the popular importance sampling estimators. We demonstrate the estimator's accuracy in several benchmark problems, and illustrate its use as a subroutine in safe policy improvement. We also provide theoretical results on the inherent hardness of the problem, and show that our estimator can match the lower bound in certain scenarios.

1. Introduction

We study the *off-policy value evaluation* problem, where one aims to estimate the value of a policy with data collected by another policy (Sutton & Barto, 1998). This problem is critical in many real-world applications of reinforcement learning (RL), 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), it is often hard to run a controlled experiment to estimate the treatment effect, and off-policy

value evaluation is a form of counterfactual reasoning that infers the causal effect of a new intervention from historical data (Holland, 1986; Pearl, 2009).

There are roughly two classes of approaches to off-policy value evaluation. The first is to fit an MDP model from data via regression, and evaluate the policy against the model. Such a regression based approach has a relatively low variance and works well when the model can be learned to satisfactory accuracy. However, for complex real-world problems, it is often hard to specify a function class in regression that is efficiently 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 function class, resulting in a bias that cannot be easily quantified. The second class of approaches are based on the idea of importance sampling (IS), which corrects 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 size of the problem's state space, but its variance can be too large for the method to be useful when the horizon is long (Mandel et al., 2014).

In this work, we propose a new off-policy value evaluation estimator that can achieve the best of regression based approaches (low variance) and importance sampling based approaches (no bias). Our contributions are three-fold:

1. A simple doubly robust (DR) estimator is proposed for RL that extends and subsumes a previous off-policy estimator for contextual bandits.

2. 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 off-policy value evaluation.

3. On benchmark problems, the new estimator is much more accurate than importance sampling baselines, while remaining unbiased in contrast to regression-based approaches. As an application, we show how such a better estimator can benefit *safe* policy iteration with a more effective policy improvement step.

2. Related Work

This paper focuses on off-policy value evaluation in finite-horizon problems, which are often a natural way to model real-world problems like dialogue systems. The goal is to estimate the expected return of start states drawn randomly from a distribution. This differs from (and is somewhat easier than) the setting considered in some previous work, often known as off-policy *policy* evaluation, which aims to estimate the whole value function (Precup et al., 2000; 2001; Sutton et al., 2015). Both settings find important yet different uses in practice, and share the same core difficulty of dealing with distribution mismatch.

The DR technique was first studied in statistics (Rotnitzky & Robins, 1995) to improve the robustness of estimation against model misspecification, and a DR estimator has been developed for dynamic treatment regime (Murphy et al., 2001). DR was later applied to policy learning in contextual bandits (Dudík et al., 2011), and its finite-time variance is shown to be typically lower than IS. The DR estimator in this work extends the work of Dudík et al. (2011) to sequential decision-making problems. In addition, we show that in certain scenarios the variance of DR matches the statistical lower bound of the estimation problem.

An important application of off-policy value evaluation is to ensure that a new policy to be deployed does not have degenerate performance in policy iteration; example algorithms for this purpose include conservative policy iteration (Kakade & Langford, 2002) and safe policy iteration (Pirotta et al., 2013). More recently, Thomas et al. (2015a) incorporate lower confidence bounds with IS in approximate policy iteration to ensure that the computed policy meets a minimum value guarantee. Our work compliments their interesting use of confidence intervals by providing DR as a drop-in replacement of IS. We show that after such a replacement, an agent can accept 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 with $P(s'|s, a)$ being the probability of seeing state s' after taking action a at state s, $R : S \times A \to \mathbb{R}$ is the mean reward function with $R(s, a)$ being the immediate goodness of (s, a), and $\gamma \in [0, 1]$ is the discount factor. Let μ be the initial state distribution. A (stationary) policy $\pi : S \times A \to [0, 1]$ assigns each state $s \in S$ a distribution over actions, where $a \in A$ has probability $\pi(a|s)$. The distribution of a H-step trajectory $\tau = (s_1, a_1, r_1, \ldots, s_H, a_H, r_H, s_{H+1})$ is specified

by μ and π as follows: $s_1 \sim \mu$, and for $t = 1, \ldots, H$, $a_t \sim \pi_0(s_t)$, $s_{t+1} \sim P(\cdot|s_t, a_t)$, and r_t has mean $R(s_t, a_t)$ conditioned on (s_t, a_t). We refer to such a distribution as $\tau \sim (\mu, \pi)$ in short. The H-step discounted value of π is

$$v^{\pi, H} := \mathbb{E}_{\tau \sim (\mu, \pi)} \left[\sum_{t=1}^{H} \gamma^{t-1} r_t \right]. \tag{1}$$

When the value of π 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 value evaluation we can still use a finite H (usually set to be on the order of $O(1/(1 - \gamma))$) so that $v^{\pi, H}$ approximates $v^{\pi, \infty}$ with a bounded error that diminishes as H increases.

3.2. Off-policy Value Evaluation

For simplicity, we assume that the data (a set of length-H trajectories) is sampled using a fixed stochastic policy[1] π_0, known as the *behavior policy*. Our goal is to estimate $v^{\pi_1, H}$, the value of a given *target policy* π_1 from data trajectories. Below we review two popular families of estimators for off-policy value evaluation.

Notation Since we are only interested in the value of π_1, the dependence of value functions on policy is omitted. In terms like $V^{\pi_1, H-t+1}(s_t)$, we also omit the dependence on horizon and abbreviate as $V(s_t)$, assuming there are $H + 1 - t$ remaining steps. Also, all (conditional) expectations are taken with respect to the distribution induced by (μ, π_0), unless stated otherwise. Finally, we use the shorthand: $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid s_1, a_1, \ldots, s_{t-1}, a_{t-1}]$ for conditional expectations, and $\mathbb{V}_t[\cdot]$ for variances similarly.

3.2.1. REGRESSION ESTIMATORS

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

$$Q^h(s, a) := \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[R(s, a) + \gamma V^{h-1}(s') \right], \tag{2}$$
$$V^h(s) := \mathbb{E}_{a \sim \pi_1(\cdot|s)} \left[Q^h(s, a) \right]. \tag{3}$$

This suggests a two-step, *regression based* procedure for off-policy value evaluation: first, fit an MDP model $\widehat{M}$ from data; second, compute the value function from Eqn.(3) using the estimated parameters $\widehat{P}$ and $\widehat{R}$. Evaluating the resulting value function, $\widehat{V}^H(s)$, on a sample of initial states and the average will be an estimate of $v^{\pi_1, H}$. (Alternatively, one could generate artificial trajectories for evaluation without explicitly referring to a model (Fonteneau et al., 2013).) When an exact state representation is used and each state-action pair appears sufficiently often in the data, such regres-

---

[1] Analyses in this paper can be easily extended to handle data trajectories that are associated with different behavior policies.

sion estimators have provably low variances and negligible biases (Mannor et al., 2007), and often outperform alternatives in practice (Paduraru, 2013).

However, real-world problems usually have a large or even infinite state space, and many state-action pairs will not be observed even once in the data, rendering the necessity of generalization in model fitting. To generalize, one can either apply function approximation to fitting $\widehat{M}$ (Jong & Stone, 2007; Grünewälder et al., 2012), or to fitting the value function directly (Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998; Dann et al., 2014). While the use of function approximation makes 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, thus breaks the credibility of estimations given by regression based approaches (Farahmand & Szepesvári, 2011; Marivate, 2015; Jiang et al., 2015).

3.2.2. IMPORTANCE SAMPLING ESTIMATORS

The IS estimator provides an unbiased estimate of π_1's value by averaging the following function of each trajectory $(s_1, a_1, r_1, \ldots, s_{H+1})$ in the data: define the per-step importance ratio as $\rho_t := \pi_1(a_t|s_t)/\pi_0(a_t|s_t)$, and the cumulative importance ratio $\rho_{1:t} := \prod_{t'=1}^{t} \rho_{t'}$; the basic (trajectory-wise) IS estimator, and an improved step-wise version are given as follows:

$$V_{\text{IS}} := \rho_{1:H} \cdot \left(\sum_{t=1}^{H} \gamma^{t-1} r_t \right), \tag{4}$$

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

Given a dataset D, the IS estimator is simply the average estimate over the trajectories, namely $\frac{1}{|D|} \sum_{i=1} V_{\text{IS}}^{(i)}$, where $|D|$ is the number of trajectories in D and $V_{\text{IS}}^{(i)}$ is IS applied to the i-th trajectory. (This averaging step will be omitted for the other estimators in the rest of this paper, and we will only specify the estimate for a single trajectory). Typically, IS (even the step-wise version) suffers from very high variance, which easily grows exponentially in horizon.

A variant of IS, weighted importance sampling (WIS), is a biased but consistent estimator, given as follows together with its step-wise version: 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 from each trajectory in D, the estimates given by trajectory-wise and step-wise WIS are respectively

$$V_{\text{WIS}} = \frac{\rho_{1:H}}{w_H} \left(\sum_{t=1}^{H} \gamma^{t-1} r_t \right), \tag{6}$$

$$V_{\text{step-WIS}} = \sum_{t=1}^{H} \gamma^{t-1} \frac{\rho_{1:t}}{w_t} r_t. \tag{7}$$

WIS has lower variance than IS, and its step-wise version is considered as the most practical point estimator in the IS family (Precup, 2000; Thomas, 2015). We will compare to the step-wise IS/WIS baselines in the experiments.

3.3. Doubly Robust Estimator for Contextual Bandits

Contextual bandits may be considered as MDPs with horizon 1, and the sample trajectories take 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 contextual bandits (Dudík et al., 2011) is defined as:

$$V_{\text{DR}} := \widehat{V}(s) + \rho \left(r - \widehat{R}(s, a) \right), \tag{8}$$

where $\rho := \frac{\pi_1(a|s)}{\pi_0(a|s)}$ and $\widehat{V}(s) := \sum_a \pi_1(a|s)\widehat{R}(s, a)$. It is easy to verify that $\widehat{V}(s) = \mathbb{E}_{a \sim \pi_0}\left[\rho \widehat{R}(s, a) \right]$, as long as $\widehat{R}$ and ρ are independent, which implies the unbiasedness of the estimator. Furthermore, if $\widehat{R}(s, a)$ is a good estimate of r, the magnitude of $r - \widehat{R}(s, a)$ can be much smaller than that of r. Consequently, the variance of $\rho(r - \widehat{R}(s, a))$ *tends to* be smaller than that of ρr, implying that DR often has a lower variance than IS (Dudík et al., 2011).

In the case where the importance ratio ρ is unknown, DR estimates both ρ 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 π_1 is often computed from data, and for DR to stay unbiased, π_1 should not depend on the samples used in Eqn.(8); the same requirement applies to IS. While $\widehat{R}$ should be independent of such samples as well, it is not required that π_1 and $\widehat{R}$ be independent of each other. For example, we can use the same dataset to compute π_1 and $\widehat{R}$, although an independent dataset is still needed to run the DR estimator in Eqn.(8). In other situations where π_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 Eqn.(8). The same requirements and procedures apply to the sequential case (discussed below). In Section 6, we will empirically validate our extension of DR in both kinds of situations.

4. DR Estimator for the Sequential Setting

4.1. The Estimator

We now extend the DR estimator for bandits to the sequential case. A key observation is that Eqn.(5) can be written in a recursive form. Define $V_{\text{step-IS}}^0 := 0$, and for $t = 1, \ldots, H$,

$$V_{\text{step-IS}}^{H+1-t} := \rho_t \left(r_t + \gamma V_{\text{step-IS}}^{H-t} \right). \tag{9}$$

It can be shown that $V_{\text{step-IS}}^H$ is equivalent to $V_{\text{step-IS}}$ given in Eqn.(5). While the rewriting is straight-forward, the recursive form provides a novel and interesting insight that

is key to the extension of the DR estimator: that is, 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 $r_t + \gamma V_{\text{step-IS}}^{H-t}$, whose expected value is $Q(s_t, a_t)$. Then, if we are supplied with $\widehat{Q}$, an estimate of Q (possibly via regression on a separate dataset), we can apply the bandit DR estimator at each horizon, and obtain the following unbiased estimator: define $V_{\text{DR}}^0 := 0$, and

$$V_{\text{DR}}^{H+1-t} := \widehat{V}(s_t) + \rho_t \left(r_t + \gamma V_{\text{DR}}^{H-t} - \widehat{Q}(s_t, a_t) \right). \quad (10)$$

The DR estimate of the policy value is then $V_{\text{DR}} := V_{\text{DR}}^H$.

4.2. Variance Analysis

In this section, we analyze the variance of DR in Theorem 1 and show that DR is preferable than step-wise IS when a good value function $\widehat{Q}$ is available. The analysis is given in the form of the variance of the estimate for a *single* trajectory, and the variance of the estimate averaged over a dataset D will be that divided by $|D|$ due to the *i.i.d.* nature of D. Due to space limit, the proof is deferred to Appendix A.

Theorem 1. *V_{DR} is an unbiased estimator of $v^{\pi_1, H}$, whose variance is given recursively as follows: $\forall t = 1, \ldots, H$,*

$$\mathbb{V}_t \left[V_{DR}^{H+1-t} \right] = \mathbb{V}_t \left[V(s_t) \right] + \mathbb{E}_t \left[\mathbb{V}_t \left[\rho_t \Delta(s_t, a_t) \mid s_t \right] \right]$$
$$+ \mathbb{E}_t \left[\rho_t^2 \, \mathbb{V}_{t+1} \left[r_t \right] \right] + \mathbb{E}_t \left[\gamma^2 \rho_t^2 \, \mathbb{V}_{t+1} \left[V_{DR}^{H-t} \right] \right], \quad (11)$$

where $\Delta(s_t, a_t) := \widehat{Q}(s_t, a_t) - Q(s_t, a_t)$ and $\mathbb{V}_{H+1} \left[V_{DR}^0 \mid s_H, a_H \right] = 0$.

On the RHS of Eqn.(11), the first 3 terms are variances due to different sources of randomness at time step t: state transition randomness, action stochasticity in π_0, and reward randomness, respectively; the 4th term contains the variance from future steps. The key conclusion is that DR's variance depends on $\widehat{Q}$ via the error function $\Delta = \widehat{Q} - Q$ in the 2nd term, hence DR with a good $\widehat{Q}$ will enjoy reduced variance, and in general outperform step-wise IS as the latter is simply DR's special case with a trivial value function $\widehat{Q} \equiv 0$.

4.3. Confidence Intervals

As mentioned in the introduction, an important motivation for off-policy value 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 a confidence interval (CI). The calculation of CIs for DR is straight-forward, since DR is an unbiased estimator applied to *i.i.d.* trajectories and standard concentration bounds 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, δ the chosen confidence level, and b the range of the estimate, which is a function of the maximal magnitudes of r_t, $\widehat{Q}(s_t, a_t)$, ρ_t and γ. The application of more sophisticated bounds for off-policy value 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 with normally approximated CIs can lead to more effective and reliable policy improvement than IS.

4.4. An Extension

From Theorem 1, it is clear that DR only reduces the variance due to action stochasticity, and may suffer a large variance even with a perfect Q-value function $\widehat{Q} = Q$, as long as the MDP has substantial stochasiticity in rewards and/or state transitions. It is, however, possible to address such a limitation. For example, one modification of DR that further reduces the variance in state transitions is:

$$V_{\text{DR-v2}}^{H+1-t} = \widehat{V}(s_t) + \rho_t \Big(r_t + \gamma V_{\text{DR-v2}}^{H-t}$$
$$- \widehat{R}(s_t, a_t) - \gamma \widehat{V}(s_{t+1}) \frac{\widehat{P}(s_{t+1}|s_t, a_t)}{P(s_{t+1}|s_t, a_t)} \Big), \quad (12)$$

where $\widehat{P}$ is the transition probability of the MDP model that we use to compute $\widehat{Q}$. While we can show that this estimator is unbiased and reduces the state-transition-induced variance with a good reward & transition functions $\widehat{R}$ and $\widehat{P}$ (we omit proof), it 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 reward function may be poorly estimated), we can assume that $P(\cdot) = \widehat{P}(\cdot)$, and the last term in Eqn.(12) becomes simply $\gamma \widehat{V}(s_{t+1})$. This generally reduces more variance than the original DR at the cost of introducing a small bias. The bias is bounded in Proposition 1, whose proof is deferred to Appendix B. In Section 6.1.3 we will demonstrate the use of such an estimator by an experiment.

Proposition 1. *Define $\epsilon = \max_{s,a} \|\widehat{P}(\cdot|s, a) - P(\cdot|s, a)\|_1$. Then, the bias of DR-v2, computed by Eqn.(12) with the approximation $\widehat{P}/P \equiv 1$, is bounded by $\epsilon V_{\max} \sum_{t=1}^{H} \gamma^t$, where $V_{\max}$ is a bound on the magnitude of $\widehat{V}$.*

5. Hardness of Off-policy Value Evaluation

In Section 4.4, we showed the possibility of reducing variance due to state transition stochasticity in a special scenario. A natural question is whether there exists an estimator that can reduce such variance without relying on strong assumptions like $\widehat{P} \approx P$. In this section, we answer this question by providing hardness results on off-policy value evaluation via the *Cramer-Rao lower bound* (or *C-R bound* for short), and comparing the C-R bound to the variance of DR.

Before stating the results, we emphasize that, as in other estimation problems, the C-R bound depends crucially on how the MDP is parameterized, because the parameterization captures our prior knowledge about the problem. In general, the more structural knowledge is encoded in parameterization, the easier it is to recover the true parameters from data, and the lower the C-R bound will be. While strong assumptions (*e.g.*, parametric form of value function) are often made in the *training* phase to make RL problems tractable, one may not want to count them as prior knowledge in *evaluation*, as every assumption made in evaluation decreases the credibility of the value estimate. (This is why regression-based methods are not trustworthy; see Section 3.2.1.) Therefore, we first present the result for the hardest case when no assumptions (other than discrete decisions & outcomes) – especially the Markov assumption that the last observation is a state – are made, to ensure the most credible estimate. A relaxed case is discussed afterwards.

Definition 1. An MDP is a *discrete tree MDP* if

- State is represented by history: that is, $s_t = h_t$, where $h_t := o_1 a_1 \cdots o_{t-1} a_{t-1} o_t$. The o_t's are called *observations*. 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 $s' = hao$, with probability $P(o|h, a)$.
- As a simplification, we assume $\gamma = 1$, and non-zero rewards only occur at the end of each trajectory. An additional observation o_{H+1} encodes the reward randomness so that reward function $R(h_{H+1})$ is deterministic, and an MDP is solely parameterized by transition probabilities.

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

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

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

Proof of Obs.1. The result follows directly by unfolding the recursion in Eqn.(11) and noticing that $\Delta \equiv 0$, $\mathbb{V}_{t+1}[r_t] \equiv 0$ for $t < H$, and $\mathbb{V}_{H+1}\big[V(s_{H+1})\big] = \mathbb{V}_{H+1}\big[r_H\big]$. $\square$

Implication When minimal prior knowledge is available, the lower bound in Theorem 2 equals the variance of DR with a perfect Q-value function, hence the part of variance due to state transition stochasticity (which DR fails to improve even with a good Q-value function) is intrinsic to the problem and cannot be eliminated without extra knowledge. Moreover, the more accurate $\widehat{Q}$ is, the lower the variance DR tends to have. A related hardness result is given by Li et al. (2015a) for MDPs with *known* transition probabilities.

Relaxed Case In Appendix C, we discuss a relaxed case where the MDP has a Directed Acyclic Graph (DAG) structure, allowing different histories of the same length to be identified as the same state, making the problem easier than

the tree case. The two cases share almost identical proofs, and below we give a concise proof of Theorem 2; readers can consult Appendix C for a fully expanded version.

Proof of Theorem 2. In the proof, it will be convenient to index rows and columns of a matrix (or vector) by histories, so that $A_{h,h'}$ denotes the (h, h') entry of matrix A. Furthermore, given a real-valued function f, $[f(h, h')]_{h,h'}$ denotes a matrix whose (h, h') entry is given by $f(h, h')$.

We parameterize a discrete tree MDP by $\mu(o)$ and $P(o|h, a)$, for h of length $1, \ldots, H$. For convenience, we treat $\mu(o)$ as $P(o|\emptyset)$, and the model parameters can be encoded as a vector θ with $\theta_{hao} = P(o|h, a)$, where ha contains $|ha| = 0, \ldots, H$ alternating observations & actions.

These parameters are subject to the normalization constraints that have to be taken into consideration in the C-R bound, namely $\forall h, a, \sum_{o \in O} P(o|h, a) = 1$. In matrix form, we have $F\theta = \mathbf{1}$, where F is a block-diagonal matrix with each block being a row vector of 1's; specifically, $F_{ha,h'a'o} = \mathbf{1}\{ha = h'a'\}$. Note that F is the Jacobian of the constraints. Let U be a matrix whose column vectors consist of an orthonormal basis for the null space of F. From Moore Jr (2010, Eqn. (3.3) and Corollary 3.10), we obtain a Constrained Cramer-Rao Bound (CCRB):

$$KU(U^\top I U)^{-1} U^\top K^\top, \tag{14}$$

where I is the Fisher Information Matrix (FIM) without taking the constraints into consideration, and K the Jacobian of the quantity $v^{\pi_1, H}$ that we want to estimate. Our calculation of the CCRB consists of four main steps.

1) Calculation of I: By definition, the FIM I is computed as $\mathbb{E}\left[\left(\frac{\partial \log P_0(h_{H+1})}{\partial \theta}\right)\left(\frac{\partial \log P_0(h_{H+1})}{\partial \theta}\right)^\top\right]$, with $P_0(h_{H+1}) := \mu(o_1)\pi_0(a_1|o_1)P(o_2|o_1, a_1)\ldots P(o_{H+1}|h_H, a_H)$ being the probability of observing h_{H+1} under policy π_0.

Define a new notation $g(h_{H+1})$ as a vector of indicator functions, such that $g(h_{H+1})_{hao} = 1$ whenever hao is a prefix of h_{H+1}. Using this notation, we have $\frac{\partial \log P_0(h_{H+1})}{\partial \theta} = \theta^{\circ-1} \circ g(h_{H+1})$, where $\circ$ denotes element-wise power/multiplication. We rewrite the FIM as $I = \mathbb{E}\left[[\theta_h^{-1}\theta_{h'}^{-1}]_{h,h'} \circ \left(g(h_{H+1})g(h_{H+1})^\top\right)\right] = [\theta_h^{-1}\theta_{h'}^{-1}]_{h,h'} \circ \mathbb{E}\left[g(h_{H+1})g(h_{H+1})^\top\right]$. Now we compute $\mathbb{E}\left[g(h_{H+1})g(h_{H+1})^\top\right]$. This matrix takes 0 in all the entries indexed by hao and $h'a'o'$ when neither of the two strings is a prefix of the other. For the other entries, without loss of generality, assume $h'a'o'$ is a prefix of hao; the other case is similar as I is symmetric. 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 $\mathbb{E}\left[g(h_{H+1})_{(hao)} \cdot g(h_{H+1})_{(h'a'o')}\right] = P_0(hao)$, and consequently $I_{(hao),(h'a'o')} = \frac{P_0(hao)}{P(o|h,a)P(o'|h',a')} = \frac{P_0(ha)}{P(o'|h',a')}$.

2) Calculation of $(U^\top I U)^{-1}$: Since I is quite dense, it

is hard to compute the inverse of $U^\top I U$ directly. Note, however, that for any matrix X with matching dimensions, $U^\top I U = U^\top (F^\top X^\top + I + XF)U$, because by definition U is orthogonal to F. Observing this, we design X to make $D = F^\top X^\top + I + XF$ diagonal so that $U^\top D U$ is easy to invert. This is achieved by letting $X_{(h'a'o'),(ha)} = 0$ 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 D is diagonal with $D_{(hao),(hao)} = I_{(hao),(hao)} = \frac{P_0(ha)}{P(o|h,a)}$.

With the above trick, we have $(U^\top I U)^{-1} = (U^\top D U)^{-1}$. Since CCRB is invariant to the choice of U, we choose U to be $\mathrm{diag}(\{U_{(ha)}\})$, where $U_{(ha)}$ is a diagonal block with columns forming an orthonormal basis of the null space of the none-zero part of $F_{(ha),(\cdot)}$ (an all-1 row vector). It is easy to verify that such U exists and is column orthonormal, with $FU = [0]_{(ha),(ha)}$. We also rewrite $D = \mathrm{diag}(\{D_{(ha)}\})$ where $D_{(ha)}$ is a diagonal matrix with $(D_{(ha)})_{o,o} = \frac{P_0(ha)}{P(o|h,a)}$, and we have $U(U^\top I U)^{-1}U^\top = \mathrm{diag}(\{U_{(ha)}(U_{(ha)}^\top D_{(ha)} U_{(ha)})^{-1} U_{(ha)}^\top\})$.

The final step is to notice that each block in the expression above is simply $\frac{1}{P_0(ha)}$ times the CCRB of a multinomial distribution $p = P(\cdot|h,a)$, which is $\mathrm{diag}(p) - pp^\top$ (Moore Jr, 2010, Eqn. (3.12)).

3) Calculation of K: Recall that we want to estimate
$$v = v^{\pi_1, H} = \sum_{o_1} \mu(o_1) \sum_{a_1} \pi_1(a_1|o_1) \cdots$$
$$\sum_{o_{H+1}} P(o_{H+1}|h_H, a_H) R(h_{H+1}).$$

Its Jacobian, $K = \partial v/\partial \theta$, can be computed by $K_{(hao)} = P_1(ha)V(hao)$, where $P_1(o_1 a_1 \cdots o_t a_t) := \mu(o_1)\pi_1(a_1)\cdots P(o_t|h_{t-1},a_{t-1})\pi_1(a_t|h_t)$ is the probability of observing a sequence under policy π_1.

4) The C-R bound: Putting all the pieces together, Eqn.(14) is equal to
$$\sum_{ha} \frac{P_1(ha)^2}{P_0(ha)} \Big(\sum_o P(o|h,a)V(hao)^2$$
$$- \Big(\sum_o P(o|h,a)V(hao) \Big)^2 \Big)$$
$$= \sum_{t=0}^{H} \sum_{|ha|=t} P_0(ha) \frac{P_1(ha)^2}{P_0(ha)^2} \mathbb{V}\big[V(hao) \mid h,a\big].$$
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, the lower bound is equal to
$$\sum_{t=0}^{H} \mathbb{E}\big[\rho_{1:t}^2 \mathbb{V}_{t+1}\big[V(s_{t+1})\big]\big] = \sum_{t=1}^{H+1} \mathbb{E}\big[\rho_{1:(t-1)}^2 \mathbb{V}_t\big[V(s_t)\big]\big].$$

6. Experiments

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

1. (IS) Step-wise IS of Eqn.(5);
2. (WIS) Step-wise WIS of Eqn.(7);
3. (REG) Regression estimator (details to be specified for each domain in the "model fitting" paragraphs);
4. (DR) Doubly robust estimator of Eqn.(10);
5. (DR-bsl) DR with a state-action independent $\widehat{Q}$.

6.1. Comparison of Mean Squared Errors

In these experiments, we compare the accuracy of the point estimate given by each estimator. For each domain, a policy π_{train} is computed as the optimal policy of the MDP model estimated from a training dataset D_{train} (generated using π_0), and the target policy π_1 is set to be $(1-\alpha)\pi_{\text{train}} + \alpha\pi_0$ for $\alpha \in \{0, 0.25, 0.5, 0.75\}$. The parameter α controls similarity between π_0 and π_1. A larger α tends to make off-policy evaluation easier, at the cost of yielding a more conservative policy when π_{train} is potentially of high quality.

We then apply the five estimators on a separate dataset D_{eval} to estimate the value of π_1, compare the estimates to the groundtruth value, and take the average estimation errors across multiple draws of D_{eval}. Note that for the DR estimator, the supplied Q-value function $\widehat{Q}$ should be independent of the data used in Eqn.(10) to ensure unbiasedness. We therefore split D_{eval} further into two subsets D_{reg} and D_{test}, estimate $\widehat{Q}$ from D_{reg} and apply DR on D_{test}.

In the above procedure, DR does not make full use of data, as the data in D_{reg} do not go into the sample average in Eqn.(10). To address this issue, we propose a more data-efficient way of applying DR in the situation when $\widehat{Q}$ has to be estimated from (a subset of) D_{eval}, and we call it k-fold DR, inspired by k-fold cross validation in supervised learning: we partition D_{eval} into k subsets, apply Eqn.(8) to each subset with $\widehat{Q}$ estimated from the remaining data, and finally average the estimate over all subsets. Since the estimate from each subset is unbiased, the overall average remains unbiased, and has lower variance since all trajectories go into the sample average. We only show the results of 2-fold DR as model fitting is time-consuming.

6.1.1. MOUNTAIN CAR

Domain description Mountain car is a popular RL benchmark problem with a 2-dimensional continuous state space, 3 discrete actions, and deterministic dynamics (Singh & Sutton, 1996). We set the horizon to be 100 time steps, and the initial state distribution to uniformly random. The behavior policy is uniformly random over the 3 actions.

Model fitting We use state aggregations: the two state variables are multiplied by 2^6 and 2^8 respectively, and the rounded integers are treated as the abstract state. We then estimate an MDP model from data using a tabular approach.

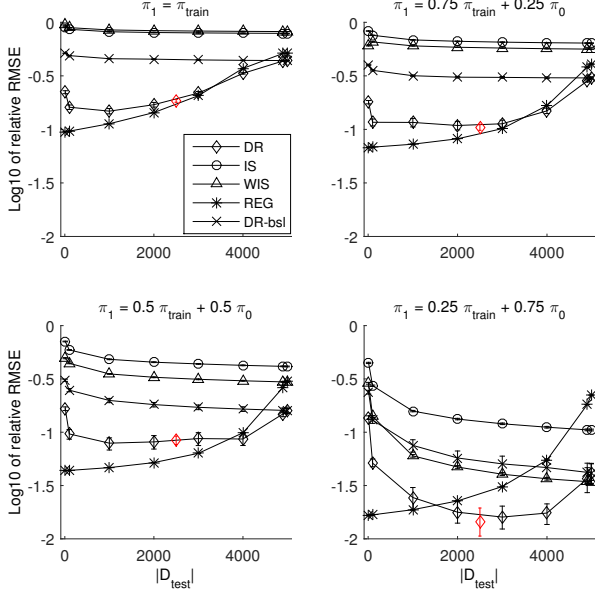**Data sizes & other details** We choose $|D_{\text{train}}| = 2000$ and

Figure 1. Comparison of the methods as point estimators on Mountain Car. 5000 trajectories are generated for off-policy evaluation, and all the results are from over 4000 runs. The subgraphs correspond to the target policies produced by mixing π_{train} and π_0 with different portions. X-axes show the size of D_{test}, the part of the data used for IS/WIS/DR/DR-bsl. The remaining data are used by the regression estimator (REG; DR uses it as $\widehat{Q}$). Y-axes show the RMSE of the estimates divided by the true value in logarithmic scale. We also show the error of 2-fold DR as an isolated point ($\diamond$).

$|D_{\text{eval}}| = 5000$. DR-bsl uses the step-dependent constant $\widehat{Q}(s_t, a_t) = R_{\min}(1 - \gamma^{H-t+1})/(1 - \gamma)$.

Results See Fig.1 for the errors of IS/WIS/DR-bsl/DR on D_{test}, and REG on D_{reg}. As $|D_{\text{test}}|$ increases, IS/WIS gets increasingly better, while REG gets worse as D_{reg} contains less data. Since DR depends on both halves of the data, it achieves the best error at some intermediate $|D_{\text{test}}|$, and beats using all the data for IS/WIS in all the 4 graphs. DR-bsl shows the accuracy of DR with $\widehat{Q}$ being a constant guess, and it 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. A state contains 4 integer variables, each representing location or direction. There are 8 actions, and $R_{\min} = -3 - 4\sqrt{2}$ and $R_{\max} = 0$.

Model fitting We apply Kernel-based Reinforcement Learning (Ormoneit & Sen, 2002) with smoothing kernel $\exp(-\|\cdot\|/0.25)$, where $\|\cdot\|$ is the ℓ_2-distance in $S \times A$.

Data sizes & other details The data sizes are $|D_{\text{train}}| = 1000$ and $|D_{\text{eval}}| = 2500$. DR-bsl uses the step-dependent
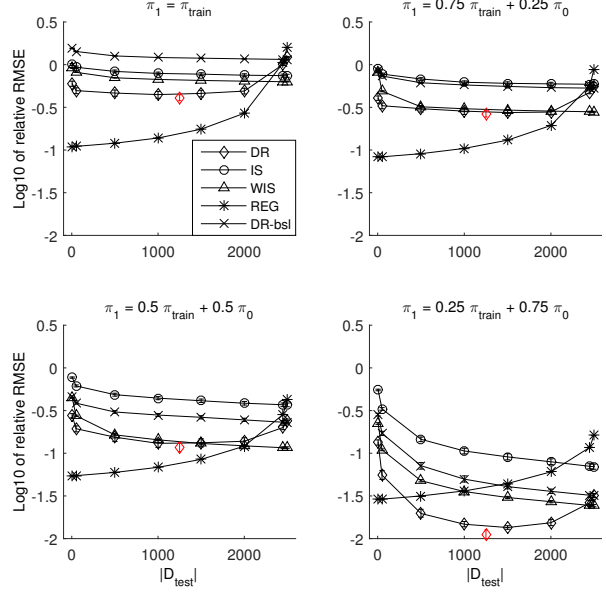

Figure 2. Comparison of the methods as point estimators on Sailing (4000 runs). 2500 trajectories are used in off-policy evaluation.

constant $\widehat{Q}(s_t, a_t) = R_{\min}/2 \cdot (1 - \gamma^{H-t+1})/(1 - \gamma)$.

Results See Fig.2. The results are qualitatively similar to Mountain Car results in Fig.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 regression estimator) with a significant margin, and a further improvement is achieved by 2-fold DR.

6.1.3. KDD CUP 1998 DONATION DATASET

In the last domain, we use the donation dataset from KDD Cup 1998 (Hettich & Bay, 1999), which records the email interactions between an agent and potential donators. A state contains 5 integer features, and there are 12 discrete actions. All trajectories are 22-steps long and there is no discount. Since no groundtruth values are available for the target policies, we fit a simulator from the true data (see appendix for details), and use it as groundtruth for everything henceforward: the true value of a target policy is computed by Monte-Carlo roll-outs in the simulator, and the off-policy evaluation methods also use data generated from the simulator (under a uniformly random policy). Among the compared estimators, we replace DR with DR-v2 (Section 4.4; reason explained below), and use the 2-fold trick.

The MDP model used to compute $\widehat{Q}$ is estimated as follows: each state variable is assumed to evolve independently (a reasonable assumption for this dataset), 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 fit by linear
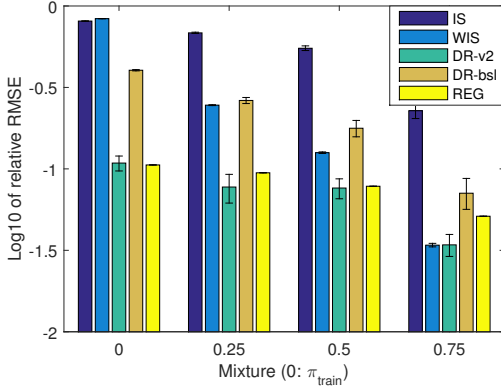
Figure 3. Results on the donation dataset, averaged over about 5000 runs. DR-v2 is the estimator given in Eqn.(12) with the 2-fold trick. The whole dataset is applied to other estimators. X-axis shows the portion of which π_0 is mixed into π_{train}.

regression using the first 3 state features (on the contrast, all the 5 features are used when fitting the simulator's reward function). Consequently, we get a model with an almost perfect transition function and a relatively inaccurate reward function, and DR-v2 is supposed to work well in such a situation. The results are shown in Fig.3, where DR-v2 is the best estimator in all situations: it beats WIS when π_1 is far from π_0, and beats REG when π_1 and π_0 are close.

6.2. Application to Safe Policy Improvement

In this experiment, we apply the off-policy value evaluation methods in safe policy improvement. Given a batch dataset D, the agent uses part of it (D_{train}) to find candidate policies, which may be poor due to data insufficiency and/or inaccurate approximation. The agent then evaluates these candidates on the remaining data (D_{test}) and chooses a policy based on the evaluation. In this common scenario, DR has an extra advantage: D_{train} can be reused to estimate $\widehat{Q}$, and it is not necessary to hold out part of D_{test} for regression.

Due to the high variance of IS and its variants, acting greedily w.r.t. the point estimate is not enough to promote safety. A typical approach is to select the policy that has the highest lower confidence bound (Thomas et al., 2015b), and hold on to the current behavior policy if none of the bounds is better than the behavior policy's value. More specifically, the bound is $V_\dagger - C\sigma_\dagger$, where V is the point estimate, σ is the empirical standard error, and $C \geq 0$ controls confidence level. $\dagger$ is a placeholder for any method that works by averaging a function of sample trajectories; examples considered in this paper are the IS and the DR estimators.

The experiment is conducted in Mountain Car, and most of the setting is the same as Section 6.1.1. Since we do not address the exploration-exploitation problem, we keep the behavior policy fixed as uniformly random, and evaluate the recommended policy once in a while as the
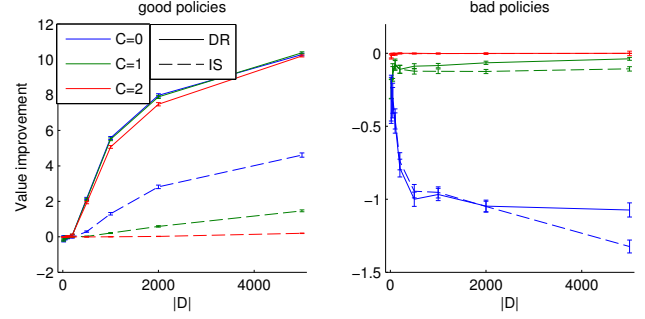

Figure 4. Safe policy improvement in Mountain Car. X-axis shows the size of data and y-axis shows the true value of the recommended policy subtracted by the value of the behavior policy.

agent gets more and more data. The candidate policies are generated as follows: we split $|D|$ so that $|D_{\text{train}}|/|D| \in \{0.2, 0.4, 0.6, 0.8\}$; for each split, we compute optimal π_{train} from the model estimated on D_{train}, mix π_{train} and π_0 with rate $\alpha \in \{0, 0.1, \ldots, 0.9\}$, compute their confidence bounds by applying IS/DR on $D \setminus D_{\text{train}}$, and finally pick the policy with the highest score over all splits and α's.

The results are shown on the left panel of Fig.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 π_0. However, here π_{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 evidence, both algorithms achieve the best value with $C = 0$, raising the concern that DR might make unsafe recommendations when π_{train} is poor.

To falsify this hypothesis, we conduct another experiment in parallel, where we have π_{train} minimize the value instead of maximizing it, resulting in policies worse than the behavior policy, and the results are shown on the right panel. Clearly, as C becomes smaller, the algorithms become 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.

7. Conclusions

We proposed a doubly robust (DR) estimator for off-policy value evaluation, and showed its high accuracy as a point estimator and usefulness in safe policy improvement. Hardness results on the problem are also provided, and the variance of DR can match the lower bound in certain scenarios. Building on a preliminary version of this work, Thomas & Brunskill (2016) showed that our DR estimator can be viewed as an application of control variates for variance reduction, and designed more advanced DR-based estimators. Future work includes applying such DR techniques to real-world problems to access its effectiveness in practice.

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.

Farahmand, Amir-massoud and Szepesvári, Csaba. Model selection in reinforcement learning. *Machine learning*, 85(3):299–332, 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.

Jiang, Nan, Kulesza, Alex, and Singh, Satinder. Abstraction Selection in Model-based Reinforcement Learning. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 179–188, 2015.

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.

Li, Lihong, Munos, Remi, and Szepesvári, Csaba. Toward minimax off-policy value estimation. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics (AISTATS-15)*, pp. 608–616, 2015a.

Li, Xiujun, Li, Lihong, Gao, Jianfeng, He, Xiaodong, Chen, Jianshu, Deng, Li, and He, Ji. Recurrent reinforcement learning: A hybrid approach, 2015b. arXiv:1509.03044.

Mandel, Travis, Liu, Yun-En, Levine, Sergey, Brunskill, Emma, and Popovic, Zoran. Offline policy evaluation across representations with applications to educational games. In *Proceedings of the 13th International Conference on Autonomous Agents and Multi-Agent Systems*, pp. 1077–1084, 2014.

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.

Marivate, Vukosi N. *Improved Empirical Methods in Reinforcement-Learning Evaluation*. PhD thesis, Rutgers University, New Brunswick, NJ, 2015.

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.

Pirotta, Matteo, Restelli, Marcello, Pecorino, Alessio, and Calandriello, Daniele. Safe policy iteration. In *Proceedings of the 30th International Conference on Machine Learning*, 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., 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 and Brunskill, Emma. Data-Efficient Off-Policy Policy Evaluation for Reinforcement Learning. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

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.

A. Proof of Theorem 1

Proof. For the base case $t = H + 1$, since $V_{\text{DR}}^0 = V(s_{H+1}) = 0$, it is obvious that at the $(H + 1)$-th step the estimator is unbiased with 0 variance, and the theorem holds. For the inductive step, suppose the theorem holds for step $t + 1$. At time step t, we have:

$$\mathbb{V}_t\left[V_{\text{DR}}^{H+1-t}\right]$$
$$= \mathbb{E}_t\left[\left(V_{\text{DR}}^{H+1-t}\right)^2\right] - \left(\mathbb{E}_t\left[V(s_t)\right]\right)^2$$
$$= \mathbb{E}_t\left[\left(\widehat{V}(s_t) + \rho_t\left(r_t + \gamma V_{\text{DR}}^{H-t} - \widehat{Q}(s_t, a_t)\right)\right)^2 - V(s_t)^2\right] + \mathbb{V}_t\left[V(s_t)\right]$$
$$= \mathbb{E}_t\left[\left(\rho_t Q(s_t, a_t) - \rho_t \widehat{Q}(s_t, a_t) + \widehat{V}(s_t)\right.\right.$$
$$\left.\left. + \rho_t\left(r_t + \gamma V_{\text{DR}}^{H-t} - Q(s_t, a_t)\right)\right)^2 - V(s_t)^2\right]$$
$$+ \mathbb{V}_t\left[V(s_t)\right]$$
$$= \mathbb{E}_t\left[\left(-\rho_t\Delta(s_t, a_t) + \widehat{V}(s_t) + \rho_t(r_t - R(s_t, a_t))\right.\right.$$
$$\left.\left. + \rho_t\gamma\left(V_{\text{DR}}^{H-t} - \mathbb{E}_{t+1}\left[V(s_{t+1})\right]\right)\right)^2$$
$$- V(s_t)^2\right] + \mathbb{V}_t\left[V(s_t)\right] \qquad (15)$$
$$= \mathbb{E}_t\left[\mathbb{E}_t\left[\left(-\rho_t\Delta(s_t, a_t) + \widehat{V}(s_t)\right)^2 - V(s_t)^2 \,\Big|\, s_t\right]\right]$$
$$+ \mathbb{E}_t\left[\mathbb{E}_{t+1}\left[\rho_t^2(r_t - R(s_t, a_t))^2\right]\right] + \mathbb{V}_t\left[V(s_t)\right]$$
$$+ \mathbb{E}_t\left[\mathbb{E}_{t+1}\left[\left(\rho_t\gamma\left(V_{\text{DR}}^{H-t} - \mathbb{E}_{t+1}\left[V(s_{t+1})\right]\right)\right)^2\right]\right]$$
$$= \mathbb{E}_t\left[\mathbb{V}_t\left[-\rho_t\Delta(s_t, a_t) + \widehat{V}(s_t) \mid s_t\right]\right] + \mathbb{E}_t\left[\rho_t^2 \,\mathbb{V}_{t+1}\left[r_t\right]\right]$$
$$+ \mathbb{E}_t\left[\rho_t^2\gamma^2 \,\mathbb{V}\left[V_{\text{DR}}^{H-t} \mid s_t, a_t\right]\right] + \mathbb{V}_t\left[V(s_t)\right]$$
$$= \mathbb{E}_t\left[\mathbb{V}_t\left[\rho_t\Delta(s_t, a_t) \mid s_t\right]\right] + \mathbb{E}_t\left[\rho_t^2 \,\mathbb{V}_{t+1}\left[r_t\right]\right]$$
$$+ \mathbb{E}_t\left[\rho_t^2\gamma^2 \,\mathbb{V}_{t+1}\left[V_{\text{DR}}^{H-t}\right]\right] + \mathbb{V}_t\left[V(s_t)\right].$$

This completes the proof. Note that from Eqn.(15) to the next step, we have used the fact that conditioned on s_t and a_t, $r_t - R(s_t, a_t)$ and $V_{\text{DR}}^{H-t} - \mathbb{E}_{t+1}\left[V(s_{t+1})\right]$ 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. □

B. Bias of DR-v2

Proof of Proposition 1. Let $V_{\text{DR-v2'}}$ denote Eqn.(12) with approximation $\widehat{P} = P$. Since $V_{\text{DR-v2}}$ is unbiased, the bias of $V_{\text{DR-v2'}}$ is then the expectation of $V_{\text{DR-v2'}} - V_{\text{DR-v2}}$. Define

$$\beta_t = \mathbb{E}_t\left[V_{\text{DR-v2'}}^{H+1-t} - V_{\text{DR-v2}}^{H+1-t}\right].$$

Then, β_1 is the bias we try to quantify, and is a constant. In general, β_t is a random variable that depends

on $s_1, a_1, \ldots, s_{t-1}, a_{t-1}$. Now we have

$$\beta_t = \mathbb{E}_t\left[\rho_t\gamma\left(V_{\text{DR-v2'}}^{H-t} - V_{\text{DR-v2}}^{H-t}\right)\right.$$
$$\left. - \rho_t\gamma\widehat{V}(s_{t+1})\left(\frac{\widehat{P}(s_{t+1}|s_t, a_t)}{P(s_{t+1}|s_t, a_t)} - 1\right)\right]$$
$$= \mathbb{E}_t\left[\rho_t\gamma\beta_{t+1}\right] - \mathbb{E}_t\left[\rho_t\gamma\widehat{V}(s_{t+1})\left(\frac{\widehat{P}(s_{t+1}|s_t, a_t)}{P(s_{t+1}|s_t, a_t)} - 1\right)\right].$$

In the second term of the last expression, the expectation is taken over the randomness of a_t and s_{t+1}; we keep a_t as a random variable and integrate out s_{t+1}, and get

$$\mathbb{E}_t\left[\rho_t\gamma\widehat{V}(s_{t+1})\left(\frac{\widehat{P}(s_{t+1}|s_t, a_t)}{P(s_{t+1}|s_t, a_t)} - 1\right)\right]$$
$$= \mathbb{E}_t\left[\mathbb{E}_{t+1}\left[\rho_t\gamma\widehat{V}(s_{t+1})\left(\frac{\widehat{P}(s_{t+1}|s_t, a_t)}{P(s_{t+1}|s_t, a_t)} - 1\right)\right]\right]$$
$$= \mathbb{E}_t\left[\rho_t\gamma\sum_{s'} P(s'|s_t, a_t)\widehat{V}(s')\left(\frac{\widehat{P}(s'|s_t, a_t)}{P(s'|s_t, a_t)} - 1\right)\right]$$
$$= \mathbb{E}_t\left[\rho_t\gamma\sum_{s'}\widehat{V}(s')\left(\widehat{P}(s'|s_t, a_t) - P(s'|s_t, a_t)\right)\right].$$

Recall that the expectation of the importance ratio is always 1, hence

$$\beta_t \leq \mathbb{E}_t\left[\rho_t\gamma\left(\beta_{t+1} + \epsilon V_{\max}\right)\right]$$
$$= \mathbb{E}_t\left[\rho_t\gamma\beta_{t+1}\right] + \gamma\epsilon V_{\max}.$$

With an abuse of notation, we reuse β_t as its maximal absolute magnitude over all sample paths $s_1, a_1, \ldots, s_{t-1}, a_{t-1}$. Clearly we have $\beta_{H+1} = 0$, and

$$\beta_t \leq \gamma(\beta_{t+1} + \epsilon V_{\max}).$$

Hence, $\beta_1 \leq \epsilon V_{\max}\sum_{t=1}^{H}\gamma^t.$ □

C. Cramer-Rao bound for discrete DAG MDPs

Here, we prove a lower bound for the relaxed setting where the MDP is a layered Directed Acyclic Graph instead of a tree. In such MDPs, the regions of the state space reachable in different time steps are disjoint (just as tree MDPs), but trajectories that separate in early steps can reunion at a same state later.

Definition 2 (Discrete DAG MDP). An MDP is a *discrete Directed Acyclic Graph (DAG) MDP* if:
- The state space and the action space are finite.
- For any $s \in S$, there exists a unique $t \in \mathbb{N}$ such that, $\max_{\pi:S \to A} P(s_t = s \mid \pi) > 0$. In other words, a state only occurs at a particular time step.
- As a simplification, we assume $\gamma = 1$, and non-zero rewards only occur at the end of each H-step long tra-

jectory. We use an additional state s_{H+1} to encode the reward randomness so that reward function $R(s_{H+1})$ is deterministic and the domain can be solely parameterized by transition probabilities.

Theorem 3. *For discrete DAG MDPs, the variance of any unbiased estimator is lower bounded by*

$$\sum_{t=1}^{H+1} \mathbb{E}\left[\frac{P_1(s_{t-1}, a_{t-1})^2}{P_0(s_{t-1}, a_{t-1})^2} \mathbb{V}_t\left[V(s_t)\right]\right],$$

where for trajectory τ,
$P_0(\tau) = \mu(s_1)\pi_0(a_1|s_1)P(s_2|s_1, a_1)\ldots P(s_{H+1}|s_H, a_H)$, *and $P_0(s_t, a_t)$ is its marginal probability; $P_1(\cdot)$ is similarly defined for π_1.*

Remark Compared to Theorem 2, the cumulative importance ratio $\rho_{1:t-1}$ is replaced by the state-action occupancy ratio $P_1(s_{t-1}, a_{t-1})/P_0(s_{t-1}, a_{t-1})$ in Theorem 3. The two ratios are equal when each state can only be reached by a unique sample path. In general, however, $\mathbb{E}\left[P_1(s_{t-1}, a_{t-1})^2/P_0(s_{t-1}, a_{t-1})^2 \mathbb{V}_t\left[V(s_t)\right]\right] \leq \mathbb{E}\left[\rho_{1:t-1}^2 \mathbb{V}_t\left[V(s_t)\right]\right]$, hence DAG MDPs are easier than tree MDPs for off-policy value evaluation.

Below we give the proof of Theorem 3, which is almost identical to the proof of Theorem 2.

Proof of Theorem 3. We parameterize the MDP by $\mu(s_1)$ and $P(s_{t+1}|s_t, a_t)$ for $t = 1, \ldots, H$. For convenience we will treat $\mu(s_1)$ as $P(s_1|\emptyset)$, so all the parameters can be represented as $P(s_{t+1}|s_t, a_t)$ (for $t = 0$ there is a single s_0 and a). These parameters are subject to the normalization constraints that have to be taken into consideration in the Cramer-Rao bound, namely $\forall t, s_t, a_t, \sum_{s_{t+1}} P(s_{t+1}|s_t, a_t) = 1$.

$$\begin{bmatrix} 1\cdots 1 & & & \\ & 1\cdots 1 & & \\ & & \ddots & \\ & & & 1\cdots 1 \end{bmatrix}\theta = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \quad (16)$$

where $\theta_{s_t, a_t, s_{t+1}} = P(o|s_t, a_t)$. The matrix on the left is effectively the Jacobian of the constraints, which we denote as F. We index its rows by (s_t, a_t), so $F_{(s_t, a_t), (s_t, a_t, s_{t+1})} = 1$ and other entries are 0. Let U be a matrix whose column vectors consist an orthonormal basis for the null space of F. From Moore Jr (2010, Eqn. (3.3) and Corollary 3.10), we have the Constrained Cramer-Rao Bound (CCRB) being[2]

---

[2]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.

(the dependence on θ in all terms are omitted):

$$KU(U^\top IU)^{-1}U^\top K^\top, \quad (17)$$

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

$$I = \mathbb{E}\left[\left(\frac{\partial \log P_0(\tau)}{\partial \theta}\right)\left(\frac{\partial \log P_0(\tau)}{\partial \theta}\right)^\top\right]. \quad (18)$$

To calculate I, we define a new notation $g(\tau)$, which is a vector of indicator functions and $g(\tau)_{s_t, a_t, s_{t+1}} = 1$ when (s_t, a_t, s_{t+1}) appears in trajectory τ. Using this notation, we have

$$\frac{\partial \log P_0(\tau)}{\partial \theta} = \theta^{\circ -1} \circ g(\tau), \quad (19)$$

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

$$I = \mathbb{E}\left[[\theta_i^{-1}\theta_j^{-1}]_{ij} \circ (g(\tau)g(\tau)^\top)\right]$$
$$= [\theta_i^{-1}\theta_j^{-1}]_{ij} \circ \mathbb{E}\left[(g(\tau)g(\tau)^\top)\right], \quad (20)$$

where $[\theta_i^{-1}\theta_j^{-1}]_{ij}$ is a matrix expressed by its (i, j)-th element. Now we compute $\mathbb{E}\left[g(\tau)g(\tau)^\top\right]$. On the diagonal, it is $P_0(s_t, a_t, s_{t+1})$, so the diagonal of I is $\frac{P_0(s_t, a_t)}{P(s_{t+1}|s_t, a_t)}$; for non-diagonal entries whose row indexing and column indexing tuples are at the same time step, the value is 0; in other cases, suppose row is (s_t, a_t, s_{t+1}) and column is $s_{t'}, a_{t'}, s_{t'+1}$, and without loss of generality assume $t' < t$, then the entry is $P_0(s_{t'}, a_{t'}, s_{t'+1}, s_t, a_t, s_{t+1})$, with the corresponding entries in I being $\frac{P_0(s_{t'}, a_{t'}, s_{t'+1}, s_t, a_t, s_{t+1})}{P(s_{t'+1}|s_{t'}, a_{t'})P(s_{t+1}|s_t, a_t)} = P_0(s_{t'}, a_{t'})P_0(s_t, a_t|s_{t'+1})$.

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

$$U^\top IU = U^\top(F^\top X^\top + I + XF)U, \quad (21)$$

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 $D_{(s_t, a_t, s_{t+1}),(s_t, a_t, s_{t+1})} = I_{(s_t, a_t, s_{t+1}),(s_t, a_t, s_{t+1})} = \frac{P_0(s_t, a_t)}{P(s_{t+1}|s_t, a_t)}$. 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 particular X we take is $X_{(s_{t'}, a_{t'}, s_{t'+1}),(s_t, a_t)} = -P_0(s_{t'}, a_{t'})P_0(s_t, a_t|s_{t'+1})\mathbb{I}(t' < t)$, and 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 $n_{(s_t, a_t)}$ be the number of 1's in $F_{(s_t, a_t), (\cdot)}$, and $U_{(s_t, a_t)}$ be the $n_{(s_t, a_t)} \times (n_{(s_t, a_t)} - 1)$ matrix

with orthonormal columns in the null space of $\begin{bmatrix} 1 & \dots & 1 \end{bmatrix}$ ($n_{(s_t,a_t)}$ 1's); finally, we choose U to be a block diagonal matrix $U = \text{diag}(\{U_{(s_t,a_t)}\})$, where $U_{(s_t,a_t)}$'s are the diagonal blocks, and it is easy to verify that U is column orthonormal and $FU = 0$. Similarly, we write $D = \text{diag}(\{D_{(s_t,a_t)}\})$ where $D_{(s_t,a_t)}$ is a diagonal matrix with $(D_{(s_t,a_t)})_{s_{t+1},s_{t+1}} = P_0(s_t,a_t)/P(s_{t+1}|s_t,a_t)$, and

$$U(U^\top IU)^{-1}U^\top = U(U^\top DU)^{-1}U^\top$$
$$= U(\text{diag}(\{U_{(s_t,a_t)}^\top\})\text{diag}(\{D_{(s_t,a_t)}\})\text{diag}(\{U_{(s_t,a_t)}\}))^{-1}U$$
$$= U\text{diag}(\{(U_{(s_t,a_t)}^\top D_{(s_t,a_t)}U_{(s_t,a_t)})^{-1}\})U$$
$$= \text{diag}(\{U_{(s_t,a_t)}(U_{(s_t,a_t)}^\top D_{(s_t,a_t)}U_{(s_t,a_t)})^{-1}U_{(s_t,a_t)}^\top\}). \tag{22}$$

Notice that each block in Eqn.(22) is simply $1/P_0(s_t,a_t)$ times the CCRB of a multinomial distribution $P(\cdot|s_t,a_t)$. The CCRB of a multinomial distribution p can be easily computed by an alternative formula (Moore Jr, 2010, Eqn. (3.12))), which gives $\text{diag}(p) - pp^\top$, so we have,

$$U_{(s_t,a_t)}(U_{(s_t,a_t)}^\top D_{(s_t,a_t)}U_{(s_t,a_t)})^{-1}U_{(s_t,a_t)}^\top$$
$$= \frac{\text{diag}(P(\cdot|s_t,a_t)) - P(\cdot|s_t,a_t)P(\cdot|s_t,a_t)^\top}{P_0(s_t,a_t)}. \tag{23}$$

We then calculate K. Recall that we want to estimate

$$v = v^{\pi_1,H} = \sum_{s_1}\mu(s_1)\sum_{a_1}\pi_1(a_1|s_1)\dots$$
$$\sum_{s_{H+1}}P(s_{H+1}|s_H,a_H)R(s_{H+1}), \tag{24}$$

and its Jacobian is $K = (\partial v/\partial\theta_t)^\top$, with $K_{(s_t,a_t,s_{t+1})} = P_1(s_t,a_t)V(s_{t+1})$, where $P_1(\tau) = \mu(s_1)\pi_1(a_1)\dots P(s_{H+1}|s_H,a_H)$ and $P_1(s_t,a_t)$ is the marginal probability.

Finally, putting all the pieces together, we have Eqn.(17) equal to

$$\sum_{s_t,a}\frac{P_1(s_t,a_t)^2}{P_0(s_t,a_t)}\Big(\sum_{s_{t+1}}P(s_{t+1}|s_t,a_t)V(s_{t+1})^2$$
$$- \Big(\sum_{s_{t+1}}P(s_{t+1}|s_t,a_t)V(s_{t+1})\Big)^2\Big)$$
$$= \sum_{t=0}^{H}\sum_{s_t}P_0(s_t,a_t)\frac{P_1(s_t,a_t)^2}{P_0(s_t,a_t)^2}\mathbb{V}\big[V(s_{t+1})\mid s_t,a\big]$$
$$= \sum_{t=0}^{H}\mathbb{E}\Big[\frac{P_1(s_t,a_t)^2}{P_0(s_t,a_t)^2}\mathbb{V}_{t+1}\big[V(s_{t+1})\big]\Big]$$
$$= \sum_{t=1}^{H+1}\mathbb{E}\Big[\frac{P_1(s_{t-1},a_{t-1})^2}{P_0(s_{t-1},a_{t-1})^2}\mathbb{V}_t\big[V(s_t)\big]\Big]. \qquad \square$$

D. Experiment Details

Here, we provide full details on the experiments that are omitted in the main paper due to space limit.

D.1. Mountain Car

Domain Description   Mountain car is a widely used benchmark problem for RL with a 2-dimensional continuous state space (position and velocity) and deterministic dynamics (Singh & Sutton, 1996). 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 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 such that given (s,a), the next state s' is obtained by calling the original transition function 4 times holding a fixed, 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 state variables 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 aggregated state-action pairs are assumed to have reward $R_{\min} = -1$ and a self-loop transition. Both the models that produces π_{train} and that used for off-policy evaluation are constructed in the same way.

Data sizes & other details   The dataset sizes are $|D_{\text{train}}| = 2000$ and $|D_{\text{eval}}| = 5000$. We split D_{eval} such that $D_{\text{test}} \in \{10, 100, 1000, 2000, 3000, 4000, 4900, 4990\}$. DR-bsl uses the step-dependent constant function

$$\widehat{Q}(s_t,a_t) = \frac{R_{\min}(1 - \gamma^{H-t+1})}{1 - \gamma}.$$

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. The same treatment is also applied to the experiment on Sailing.

D.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 10×10) 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 that depends on moving direction, wind direction, and other factors, ranging from $R_{\min} = -3 - 4\sqrt{2}$ to $R_{\max} = 0$ (absorbing). The problem is non-discounting, and we use $\gamma = 0.99$ for easy convergence when computing π_{train}.

Model Construction We apply Kernel-based Reinforcement Learning (Ormoneit & Sen, 2002) and supply 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 ℓ_2-distance in $S \times A$,[3] and b is the kernel bandwidth, set to 0.25.

Data sizes & other details The data sizes are $|D_{\text{train}}| = 1000$ and $|D_{\text{eval}}| = 2500$, and we split D_{eval} such that $D_{\text{test}} \in \{5, 50, 500, 1000, 1500, 2000, 2450, 2495\}$. DR-bsl uses the step-dependent constant function

$$\widehat{Q}(s_t, a_t) = \frac{R_{\min}}{2} \frac{1 - \gamma^{H-t+1}}{1 - \gamma},$$

for the reason that in Sail $R_{\min}$ is rarely reached hence too pessimistic as a rough estimate of the magnitude of reward obtained per step.

D.3. KDD Cup 1998 Donation Dataset

Here are further details for experiments with the KDD donation dataset:

1. The size of dataset generated from the simulator for off-policy evaluation is equal to that of the true dataset (the one we use to fit the simulator at the very beginning; there are 3754 trajectories in that dataset).

2. The policy π_{train} is generated by training a recurrent neural network on the original data to fit a Q-value function (Li et al., 2015b).

3. 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.

---

[3]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.