State Relevance for Off-Policy Evaluation

Simon P. Shen [1] **Yecheng Jason Ma** [2] **Omer Gottesman** [3] **Finale Doshi-Velez** [1]

Abstract

Importance sampling-based estimators for off-policy evaluation (OPE) are valued for their simplicity, unbiasedness, and reliance on relatively few assumptions. However, the variance of these estimators is often high, especially when trajectories are of different lengths. In this work, we introduce *Omitting-States-Irrelevant-to-Return Importance Sampling* (OSIRIS), an estimator which reduces variance by strategically omitting likelihood ratios associated with certain states. We formalize the conditions under which OSIRIS is unbiased and has lower variance than ordinary importance sampling, and we demonstrate these properties empirically.

1. Introduction

In the context of reinforcement learning, our work focuses on the off-policy evaluation (OPE) problem, where the goal is to estimate the value of a given policy using historical data collected under a different policy (Sutton & Barto, 2018). OPE is often a necessary step in many real-world applications of reinforcement learning whenever running evaluation policies is costly or risky, for example, in healthcare and education settings (Murphy et al., 2001; Mandel et al., 2014). In particular, we focus on the OPE approaches based on importance sampling (IS), which correct the historical data to account for the difference between the policies (Precup et al., 2000). IS-based estimators are popular for their appealing statistical properties (Thomas & Brunskill, 2016; Jiang & Li, 2016; Farajtabar et al., 2018; Thomas, 2015).

However, current IS-based estimators struggle in scenarios involving trajectories of different lengths. In these settings, IS-based estimators have large variance that increases with trajectory length because IS weights are products of likelihood ratios (Doroudi et al., 2018). Resolving this issue is important as many domains have trajectories with different lengths: in health settings, patients' length of stays may vary drastically; in education settings, students may spend different amounts of time using various online tools.

Motivated by the observation that IS variance is driven by a large and varying number of likelihood ratios, we present a new estimator, ***O**mitting-**S**tates-**I**rrelevant-to-**R**eturn **I**mportance **S**ampling* (OSIRIS), which strategically omits likelihood ratios associated with certain states. The goal of the estimator is to reduce IS variance while introducing minimal bias. We first identify the variance contributed to ordinary IS by likelihood ratios that would be omitted by OSIRIS. This analysis motivates the method's idea to omit likelihood ratios corresponding to "irrelevant" states, where the action taken does not affect the trajectory return, and this omission criterion keeps OSIRIS unbiased. Based on this criterion, we describe a practical algorithm using a statistical test to estimate state relevance. Finally, we experimentally validate this implementation of OSIRIS on a suite of discrete- and continuous-state environments. Because the estimator's procedure is to set "irrelevant" likelihood ratios to 1, it can be easily used alongside other variants of importance sampling estimators.

2. Background

Markov Decision Process We consider a standard reinforcement learning framework in which an agent, characterized by a policy π, interacts with a finite Markov decision process (MDP), characterized by a tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$. $\mathcal{S}$ and $\mathcal{A}$ represent the state and action spaces, respectively; $P(s' \mid s, a)$ and $R(s, a)$ represent the transition distribution and the reward function, respectively; and $\gamma \in [0, 1]$ is the temporal discount factor. The agent starts at an initial state s_1 drawn from the initial state distribution $P(s_1)$. At each time step t, the agent performs an action $a_t \sim \pi(\cdot \mid s_t)$, observes reward $r_t = R(s_t, a_t)$, and transitions to state $s_{t+1} \sim P(\cdot \mid s_t, a_t)$. Once the agent reaches a terminal state (e.g. at time $T+1$), the trajectory is complete and is defined as $\tau = (s_1, a_1, r_1, \ldots, s_T, a_T, r_T, s_{T+1})$. The discounted rewards collected between any times t_1 and t_2 in trajectory τ is defined as $g_{t_1:t_2}(\tau) \equiv \sum_{t=t_1}^{t_2} \gamma^{t-t_1} r_t$, and the full trajectory return is simply denoted by $g(\tau) \equiv g_{1:T}(\tau)$.

[1] Harvard University, Cambridge, MA [2] University of Pennsylvania, Philadelphia, PA [3] Brown University, Providence, RI. Correspondence to: Simon P. Shen <simonshen@fas.harvard.edu>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

Policy Evaluation In the policy evaluation problem, we are given historical data as a batch of trajectories $\mathcal{D} \equiv \{\tau^{(1)}, ..., \tau^{(N)}\}$ collected under a behavior policy π_b, and we want to estimate the value $V^{\pi_e} \equiv \underset{\tau \sim \pi_e}{\mathbb{E}} [g(\tau)]$ of an evaluation policy π_e. We use the notation $\tau \sim \pi$ to indicate that trajectories τ are sampled from the joint probability distribution $P(\tau; \pi) = P(s_1) \prod_{t=1}^{T} \pi(a_t \mid s_t) P(s_{t+1} \mid s_t, a_t)$, and $\mathcal{D} \sim \pi$ indicates i.i.d. sampling of N such trajectories.

If $\pi_b = \pi_e$, we can perform *on-policy* evaluation with the unbiased Monte Carlo (MC) estimator $\hat{V}^{\pi_e}_{\text{MC}}(\mathcal{D}) \equiv \frac{1}{|\mathcal{D}|} \sum_{\tau \in \mathcal{D}} g(\tau)$.

Importance Sampling In many real-world scenarios, $\pi_b \neq \pi_e$, so we can only perform *off-policy* evaluation. Among OPE methods, estimators based on importance sampling (IS) are valued for their simplicity, unbiasedness, and reliance on relatively few assumptions (Precup et al., 2000). The ordinary IS estimator is given by the weighted average

$$\hat{V}^{\pi_e}_{\text{IS}}(\mathcal{D}) \equiv \frac{1}{|\mathcal{D}|} \sum_{\tau \in \mathcal{D}} g(\tau) \rho(\tau) \tag{1}$$

where the weight $\rho(\tau)$ is the product of likelihood ratios

$$\rho(\tau) \equiv \prod_{t=1}^{T} \frac{\pi_e(a_t \mid s_t)}{\pi_b(a_t \mid s_t)} \tag{2}$$

We also introduce shorthand notation: $\rho_t(\tau) \equiv \frac{\pi_e(a_t \mid s_t)}{\pi_b(a_t \mid s_t)}$ and $\rho_{t_1:t_2}(\tau) \equiv \prod_{t=t_1}^{t_2} \rho_t(\tau)$. The IS estimator is an unbiased estimator of V^{π_e} but typically has large variance, which is the subject of Section 4.1. To reduce variance, the commonly used weighted IS (WIS) estimator effectively scales the importance weights $\rho(\tau)$ to be between 0 and 1: $\hat{V}^{\pi_e}_{\text{WIS}}(\mathcal{D}) = \frac{\sum_{\tau \in \mathcal{D}} g(\tau)\rho(\tau)}{\sum_{\tau \in \mathcal{D}} \rho(\tau)}$ (Precup et al., 2000). This estimator becomes biased but is consistent.

3. Related Work

There has recently been significant interest in improving the accuracy of estimation based on importance sampling. Some approaches have included importance weight truncation (Ionides, 2008; Su et al., 2019), confidence bounds (Thomas et al., 2015a;b; Papini et al., 2019; Metelli et al., 2020), and doubly robust estimation (Jiang & Li, 2016; Thomas & Brunskill, 2016; Su et al., 2019). In this work, we focus on the unique challenges presented by the long-horizon setting (Doroudi et al., 2018).

Instead of weighting entire observed trajectories, a recent family of methods showed promising results by calculating weights using estimates of the steady-state visitation distribution (Liu et al., 2018; Xie et al., 2019). This approach was shown to improve asymptotic behavior with respect to the horizon. However, all IS estimators will suffer when trajectories are long, so our fundamentally different approach is to strategically *shorten* the horizon. Furthermore, there is still merit in trying to treat trajectories as a whole rather than breaking them apart into individual transitions, especially if the state space is believed to be partially observable.

Doroudi et al. (2018) propose a per-horizon estimator which first groups trajectories by their lengths, performs WIS on each group, and finally averages these sub-estimates using horizon-corrective weights. However, if each such group is small, the sub-estimates will have large bias because WIS is biased. Furthermore, it is difficult to estimate the distribution of trajectory lengths under π_e, which is necessary to compute the horizon-corrective weights.

Guo et al. (2017) are motivated by options-based policies and leverage temporal abstraction to modify the per-decision importance sampling (PDIS) estimator, which weights the individual rewards from each transition. For settings without access to well-defined options-based policies, the authors suggest dropping all but the k-most recent likelihood ratios from each PDIS weight, where k is chosen to minimize an estimate of the estimator's MSE. However, in addition to the difficulty of accurately estimating MSE, the estimator tends to overweight rewards near the end of a trajectory because they are multiplied by fewer likelihood ratios. Our perspective reveals using state relevance to flexibly omit likelihood ratios without favoring any part of the trajectory a priori.

Rowland et al. (2020) present a framework for conditional importance sampling, including the return-conditioned importance sampling (RCIS) estimator, which uses IS weights that are conditioned on trajectory return. This approach is designed to remove noise that is unrelated to determining trajectory return. However, RCIS uses a regressor to fit IS weights that are uncorrected and thus still vulnerable to the trajectory length issues discussed in Section 4.1. Under this conditional IS framework, OSIRIS weights are conditioned on richer relevance information rather than only the trajectory returns, and OSIRIS's likelihood ratio omission is designed to directly address trajectory length issues.

4. Variance Reduction by Omitting Likelihood Ratios

4.1. Sources of Trajectory Length Variance

We are motivated by the observation that a major source of variance in IS-based estimators is the *large* and *variable* number of likelihood ratios in the IS weight.

Because the IS weight is the product of likelihood ratios, each of which is a random variable, the IS weight tends to have larger variance when trajectories are *long*:

Proposition 1. *For any subsets $\mathcal{T}_1 \subsetneq \mathcal{T}_2 \subseteq \{1, \ldots, T\}$, if $\pi_e \neq \pi_b$, and $\rho_1(\tau), \ldots, \rho_T(\tau)$ are mutually independent, then*

$$\underset{\tau\sim\pi_b}{\mathrm{Var}}\Big[\prod_{t\in\mathcal{T}_1}\rho_t(\tau)\,|\,s_1,\ldots,s_T\Big] < \underset{\tau\sim\pi_b}{\mathrm{Var}}\Big[\prod_{t\in\mathcal{T}_2}\rho_t(\tau)\,|\,s_1,\ldots,s_T\Big] \quad (3)$$

The proof is in Appendix B.1. To address this issue, we propose to strategically omit likelihood ratios from the IS weight product by setting them to 1, which has zero variance.

Furthermore, when trajectory lengths are *variable*, they contribute variance that is not inherently meaningful to OPE. Intuitively, there are at least two kinds of information that can be represented in the IS weight: the individual actions taken during the trajectory and the trajectory length. First, each likelihood ratio in the IS weight measures how well a transition in the historical data follows the evaluation policy. This feature is intentional so that, in expectation, the overall IS weight corrects for the difference between the behavior and evaluation policies. However the distribution of the IS weight is also directly related to the number of likelihood ratios multiplied together. This is a result of the skew of the distribution of likelihood ratios: because the behavior probability π_b appears in the denominator of the likelihood ratio, it is very rare to observe large/exploding likelihood ratios for transitions in the finite historical data, which are sampled from π_b, but it can be very common to observe small/vanishing likelihood ratios. The effect of this skew is that longer trajectories tend to have smaller IS weights because they multiply more likelihood ratios. This idea is formalized in:

Proposition 2. *For any subsets $\mathcal{T}_1 \subsetneq \mathcal{T}_2 \subseteq \{1, \ldots, T\}$, if $\pi_e \neq \pi_b$, then*

$$\underset{\tau\sim\pi_b}{\mathbb{E}}\Big[\log\prod_{t\in\mathcal{T}_1}\rho_t(\tau)\Big] > \underset{\tau\sim\pi_b}{\mathbb{E}}\Big[\log\prod_{t\in\mathcal{T}_2}\rho_t(\tau)\Big] \quad (4)$$

The proof is in Appendix B.2. The log allows us to easily reveal the relationship between trajectory length and IS weight, and it is appropriate because the IS weight multiplies likelihood ratios together. Although the IS estimator is unbiased in expectation, this relationship can be problematic for finite data sizes and especially when trajectory lengths are long and highly variable. In these cases, the IS weights can become dominated by the information about the number of likelihood ratios rather than the meaningful information about how well the trajectory follows the evaluation policy.

4.2. Omission of Meaningless Likelihood Ratios

We have identified two problems: when trajectory lengths are *variable*, IS tends to overweight short trajectories in a way that is not inherently meaningful; and when trajectory lengths are *long*, the extra likelihood ratios contribute extra IS variance overall. Our goal is then to strategically omit likelihood ratios in a way that preserves/emphasizes meaningful variance related to the actions taken in the historical data while minimizing meaningless variance that is only related to trajectory length. We begin by decomposing the IS estimator variance, which will suggest such a method.

Assume we have a mapping $\theta' : \mathcal{S} \to \{0, 1\}$ that identifies which likelihood ratios should be kept vs omitted in the IS weight.[1] Omitting likelihood ratios is equivalent to setting them to 1. This procedure can easily be applied to any IS-based estimator (see Extensions in Section 5), but for now we formally define the procedure on the ordinary IS estimator:

Definition 1. *Given any mapping $\theta' : \mathcal{S} \to \{0, 1\}$, the OSIRIS weight is defined as*

$$\rho_{\theta'}(\tau) \equiv \prod_{t=1}^{T}\big[\rho_t(\tau)\big]^{\theta'(s_t)} \quad (5)$$

and accordingly, the OSIRIS estimator is defined as

$$\hat{V}^{\pi_e}_{\mathrm{OSIRIS}}(\mathcal{D};\theta') \equiv \frac{1}{|\mathcal{D}|}\sum_{\tau\in\mathcal{D}} g(\tau)\,\rho_{\theta'}(\tau) \quad (6)$$

We also introduce notation for the product of the omitted likelihood ratios: $\rho^{\complement}_{\theta'}(\tau) \equiv \prod_{t=1}^{T}\big[\rho_t(\tau)\big]^{1-\theta'(s_t)}$. This quantity directly relates the OSIRIS estimator to the ordinary IS estimator:

$$\hat{V}^{\pi_e}_{\mathrm{IS}}(\tau) = \hat{V}^{\pi_e}_{\mathrm{OSIRIS}}(\tau;\theta')\cdot\rho^{\complement}_{\theta'}(\tau) \quad (7)$$

where we have abused notation by writing $\hat{V}^{\pi_e}(\tau)$ to represent the single-trajectory estimator $\hat{V}^{\pi_e}(\{\tau\})$. We use this fact to decompose the IS estimator variance:

Theorem 1. *Given any mapping $\theta' : \mathcal{S} \to \{0, 1\}$, the variance of the OSIRIS estimator is:*

$$\underset{\mathcal{D}\sim\pi_b}{\mathrm{Var}}[\hat{V}^{\pi_e}_{\mathrm{OSIRIS}}(\mathcal{D};\theta')] = \underset{\mathcal{D}\sim\pi_b}{\mathrm{Var}}[\hat{V}^{\pi_e}_{\mathrm{IS}}(\mathcal{D})] + \frac{1}{|\mathcal{D}|}\underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}^{\pi_e}_{\mathrm{IS}}(\tau^{(1)})]^2 - \frac{1}{|\mathcal{D}|}\underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}^{\pi_e}_{\mathrm{OSIRIS}}(\tau^{(1)};\theta')]^2 \quad (8a)$$

$$-\frac{1}{|\mathcal{D}|}\underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}^{\pi_e}_{\mathrm{OSIRIS}}(\tau^{(1)};\theta')^2]\underset{\mathcal{D}\sim\pi_b}{\mathrm{Var}}[\rho^{\complement}_{\theta'}(\tau^{(1)})] \quad (8b)$$

$$-\frac{1}{|\mathcal{D}|}\underset{\mathcal{D}\sim\pi_b}{\mathrm{Cov}}[\hat{V}^{\pi_e}_{\mathrm{OSIRIS}}(\tau^{(1)};\theta')^2,\,\rho^{\complement}_{\theta'}(\tau^{(1)})^2] \quad (8c)$$

[1] The analysis here in Section 4.2 works for any general mapping θ'. But we write the input space as just $\mathcal{S}$ in order to smoothly introduce the idea of *state* relevance in Section 5.

The derivation is in Appendix C.1. Term (8a) adjusts for the locations of the respective estimator distributions. It is generally much smaller in magnitude than all other terms, which is plausible by the Cauchy-Schwarz inequality.[2] Term (8b) represents the variance of the omitted likelihood ratios. This term cannot be positive, so it can only act to decrease variance. The key conclusion is about Term (8c), which mirrors[3] the bias term:

Theorem 2. *Given any mapping $\theta' : \mathcal{S} \to \{0, 1\}$, the mean of the OSIRIS estimator is:*

$$\mathop{\mathbb{E}}_{\mathcal{D}\sim\pi_b}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\mathcal{D};\theta')] = \\ V^{\pi_e} - \mathop{\text{Cov}}_{\mathcal{D}\sim\pi_b}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\tau^{(1)};\theta'),\, \rho^{\mathsf{C}}_{\theta'}(\tau^{(1)})] \quad (9)$$

The derivation is in Appendix C.2. If the covariance between $\hat{V}^{\pi_e}_{\text{OSIRIS}}(\tau;\theta')$ and $\rho^{\mathsf{C}}_{\theta'}(\tau)$ is zero, then the estimator is unbiased by Theorem 2 and likely has reduced variance by Theorem 1. This observation suggests an algorithm that chooses θ' to minimize this covariance term.

However, accurate estimation of the covariance term is challenging: high-variance estimates of the covariance can introduce large variance to the OPE estimator by outputting drastically different θ' per sample, which would increase length variance. Accurate estimation of the covariance term is further complicated by independence requirements that restrict the usable data. First, Equations 8c and 9 say the covariance term should be estimated over i.i.d. data sets $\mathcal{D}$, each of which has its own θ'; but in practice, any calculation of θ' will probably incorporate all data in a single sample of $\mathcal{D}$. Furthermore, Theorems 1 and 2 use the fact that $\mathbb{E}_{\mathcal{D}\sim\pi_b}[\rho^{\mathsf{C}}_{\theta'}(\tau^{(1)})] = 1$, which is generally only true if θ' is calculated using $\mathcal{D} \setminus \{\tau^{(1)}\}$ (Appendix A).

Towards getting around these problems associated with picking θ', we further interpret the covariance term between $\hat{V}^{\pi_e}_{\text{OSIRIS}}(\tau;\theta')$ and $\rho^{\mathsf{C}}_{\theta'}(\tau)$ as a measure of the statistical dependence between the trajectory outcome under π_e and the product of omitted likelihood ratios, respectively. As discussed in Section 4.1, the product of likelihood ratios can represent both the number of transitions included and how well those transitions follow π_e. Assuming that trajectory length is not inherently meaningful to trajectory outcome, the information contained in $\rho^{\mathsf{C}}_{\theta'}(\tau)$ is then primarily about the actions taken during the trajectory. Indeed, the covariance term involving the *product* of likelihood ratios is the sum of covariances involving each *individual* likelihood ratio (Appendix C.3). This decomposition suggests omitting individual likelihood ratios that are independent of the kept terms in the IS estimator. We refine this idea towards a practical estimator algorithm in the next section.

5. State Relevance

We have motivated the idea to reduce IS variance without introducing bias by omitting individual likelihood ratios where there is no statistical dependence between the action taken and the trajectory outcome. Now we present another perspective of this idea, suggesting a practical algorithm that circumvents the challenges associated with directly estimating the covariance term.

We want to calculate θ to measure the dependence of the trajectory outcome on each individual action taken. Because the actions are sampled from policies that are conditioned on states, we assume that there are some states where the action taken does not matter to the trajectory outcome. We formalize this idea as the relevance of a state:

Definition 2. *A state $s \in \mathcal{S}$ is irrelevant if*

$$\mathop{\mathbb{E}}_{\tau\sim\pi_e}[g_{t:T}(\tau) \mid s_t = s,\, a_t = a] = \text{constant}, \quad \forall a \in \mathcal{A}. \quad (10)$$

Otherwise, s is relevant. Using this condition, we define the true relevance mapping $\theta : \mathcal{S} \to \{0, 1\}$ where $\theta(s) = 0$ if s is irrelevant, and $\theta(s) = 1$ if s is relevant.[4]

In other words, a state is irrelevant if the average return-to-go is not affected by the action taken in that state. For example, if state s^A always takes the agent to s^B and no reward is given for that transition, then s^A is considered irrelevant. Or if paths diverge out of s^A but later converge to s^B and no reward is given for those transitions, then s^A is still considered irrelevant. s^A can even be irrelevant if non-zero rewards are given, as long as the expected value of the return-to-go (aka state-action value function $Q^{\pi_e}(s, a) \equiv \mathbb{E}_{\tau\sim\pi_e}[g_{t:T}(\tau) \mid s_t = s,\, a_t = a]$) remains the same regardless of the action taken in s^A. Likelihood ratios corresponding to all transitions between s^A and s^B still get multiplied in the ordinary IS weight. But these likelihood ratios have no effect in correcting the difference between π_e and π_b, so they just add meaningless variance to the IS estimator.

Definition 2 tells us that in these irrelevant states, the trajectory outcome is unaffected by the action taken. As such, we can pretend that in irrelevant states, the evaluation policy will actually draw actions from π_b rather than π_e. This idea is formalized in:

Lemma 1. *Let π'_e be a composite policy:*

$$\pi'_e(a \mid s;\theta) \equiv \begin{cases} \pi_e(a \mid s) & \textit{if } \theta(s) = 1 \textit{ (relevant)} \\ \pi_b(a \mid s) & \textit{if } \theta(s) = 0 \textit{ (irrelevant)} \end{cases} \quad (11)$$

[2] Terms (8b) and (8c) are of the form $\mathbb{E}[x^2]$ while the terms in (8a) are of the form $\mathbb{E}[x]^2$.

[3] By definition, the variance involves *second* moments, so the variables in Term (8c) are squared but not in Equation 9.

[4] If Equation 10 is true for some t, then it must be true for all t by the MDP setup.

Then the policy values of π_e and π'_e are equal:

$$V^{\pi_e} = V^{\pi'_e} \tag{12}$$

The proof is in Appendix D.1. Performing importance sampling while treating π'_e as the evaluation policy is equivalent to setting the importance sampling likelihood ratios to $\frac{\pi_b(a\,|\,s)}{\pi_b(a\,|\,s)} = 1$ wherever s is irrelevant. This perspective is exactly the likelihood ratio omission strategy in OSIRIS. Thus, the OSIRIS estimator using the true state relevance mapping is unbiased:

Theorem 3. *Given the true relevance mapping θ, the mean of the OSIRIS estimator is*

$$\mathop{\mathbb{E}}_{\mathcal{D}\sim\pi_b}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\mathcal{D};\theta)] = V^{\pi_e} \tag{13}$$

The proof is in Appendix D.2, where we also show that the estimator remains unbiased if we keep irrelevant likelihood ratios, but it is biased if we omit relevant likelihood ratios.

Implementation A key assumption of our analysis so far is that we have access to the true relevance mapping θ, but practically, we will almost always need to estimate relevance from the historical data. Definition 2 naturally suggests a class of algorithms to do so by estimating the state-action value function $Q^{\pi_e}(s,a)$ and then comparing the estimate $\hat{Q}^{\pi_e}(s,a)$ for different actions a within state s. If $\hat{Q}^{\pi_e}(s,a)$ is (approximately[5]) constant for all actions a, then the algorithm should consider s irrelevant. Otherwise s should be relevant.

We will focus on presenting one possible implementation, summarized in Algorithm 1, that uses IS with a statistical test because it is simple yet effective and can guarantee estimator consistency. For each visit to state s in the historical data, we pre-calculate the associated return-to-go $g_{t:T}(\tau^{(n)})$ multiplied by the IS weight-to-go $\rho_{t:T}(\tau^{(n)})$. This is an unbiased estimate of the expected return-to-go under the evaluation policy π_e after taking action a in state s (aka $Q^{\pi_e}(s,a)$). Although these samples likely come from different times t, the MDP setup says they still come from the same distribution conditioned on s. Thus, for a given s, we collect these estimates together for all visits to s into either the list $\mathcal{G}_+$ or $\mathcal{G}_-$ depending on whether the likelihood ratio corresponding to the transition is > 1 or ≤ 1, respectively. This procedure effectively produces a binary action space, which allows us to use popular two-sample statistical tests[6] to compare $\mathcal{G}_+$ and $\mathcal{G}_-$. In particular, we use Welch's t-test (Welch, 1951) where the null hypothesis is Equation 10 with $|\mathcal{A}| = 2$. Because the statistical test assumes the individual samples in $\mathcal{G}_+$ and $\mathcal{G}_-$ are i.i.d., we need to assume that state s is visited at most once per trajectory, which is otherwise sampled i.i.d.

Algorithm 1 Estimating state relevance $\hat{\theta}(s;\mathcal{D})$

```
Input: state $s$, data $\mathcal{D}$, significance level $\alpha$
Initialize $\mathcal{G}_+ \leftarrow \emptyset$ and $\mathcal{G}_- \leftarrow \emptyset$.
for $n = 1$ to $N$ and $t = 1$ to $T$ do
    if $s_t^{(n)} = s$ then
        if $\rho_t(\tau^{(n)}) > 1$ then
            Append $\mathcal{G}_+ \leftarrow \mathcal{G}_+ \cup \{g_{t:T}(\tau^{(n)})\rho_{t:T}(\tau^{(n)})\}$
        else
            Append $\mathcal{G}_- \leftarrow \mathcal{G}_- \cup \{g_{t:T}(\tau^{(n)})\rho_{t:T}(\tau^{(n)})\}$
        end if
    end if
end for
Perform Welch's $t$-test comparing the samples $\mathcal{G}_+$ and $\mathcal{G}_-$ with significance level $\alpha$
if null hypothesis is rejected then
    Output: $\hat{\theta}(s;\mathcal{D}) \equiv 1$ (relevant)
else
    Output: $\hat{\theta}(s;\mathcal{D}) \equiv 0$ (irrelevant)
end if
```

By using Equation 10 as the null hypothesis, this state relevance estimation procedure assumes that states are irrelevant until "proven" relevant. Now, we will characterize the effect of this property on accuracy.

Incorrectly identifying a truly irrelevant state as relevant occurs with probability $\leq \alpha$, resulting from a type I error of the statistical test. This error will not affect the bias of the OSIRIS estimator because the likelihood ratios corresponding to truly irrelevant states can take any value without introducing bias (Appendix D.2). Although Term (8c) of the variance should also be unaffected by this error because it mirrors the bias term, Term (8b) will likely increase variance because fewer likelihood ratios are omitted (Proposition 1).

Meanwhile, the type II error, of incorrectly identifying a truly relevant state as irrelevant, introduces bias (Appendix D.2). But this bias goes to zero as data size increases because the statistical test is consistent:

Theorem 4. *If $|\mathcal{A}| = 2$ and $\alpha > 0$, then as $|\mathcal{D}| \to \infty$*

$$\mathop{\mathbb{E}}_{\mathcal{D}\sim\pi_b}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\mathcal{D};\hat{\theta}(\,\cdot\,;\mathcal{D}))] = V^{\pi_e} \tag{14}$$

The proof is in Appendix D.3. The purpose of the assumption $|\mathcal{A}| = 2$ is to account for the fact that our binary classification (for practical implementation of the statistical test) loses information about the actions. Generally, the action

[5] Inevitably, $\hat{Q}^{\pi_e}$ will be an imperfect estimator. But this error could be advantageous because it effectively enforces a softer version of Definition 2, which may allow the estimator to also ignore states that are "mostly" irrelevant.

[6] If Smirnov's non-parametric test (Hodges, 1958) is used instead, we observe the same qualitative trends as our reported results (Appendix F.1).

space will be larger than 2, but we expect the estimator to still be consistent if the action space is partitioned in a way that that preserves any dependence between actions and return-to-go. We proposed classifying actions based on their likelihood ratios, which is reasonable because following vs deviating from the evaluation policy could cause drastic differences in trajectory outcome. For example, in a policy improvement setting where the policies are ϵ-greedy, following the optimal action would lead to higher rewards than a random action. Furthermore, our partitioning strategy is based on the likelihood ratios, which are the entities that directly appear in the covariance term in Equations 8 and 9. In Appendix F.1, we explored another partition strategy, which gave the same qualitative trends as our reported results.

Altogether, because α trades off the probabilities of type I and II error, it can also be seen as a parameter that trades off OSIRIS bias and variance by mixing two OPE approaches: a naive average of behavior trajectory returns ($\alpha = 0$ i.e. all likelihood ratios are irrelevant) and an unbiased but high-variance IS estimator ($\alpha = 1$ i.e. all likelihood ratios are relevant).[7] This perspective also points out that for smaller data size, Algorithm 1 will more aggressively label states as irrelevant. Equivalently, it prioritizes variance reduction for smaller data size, which is reasonable because variance is inversely related to data size (Theorem 1).

Alternative Implementations Clearly, the estimation in Algorithm 1 will be more accurate if the lists $\mathcal{G}_+, \mathcal{G}_-$ contain more data, so it effectively requires a discrete state space. Nonetheless, the method can easily be extended to settings with continuous state spaces: in Section 6, we show the method still works if we use discretized states to perform the statistical test and use the original continuous states for all other calculations in the estimator.

In Appendix F.1, we present empirical results for other possible implementations. Notably, we also tried directly estimating $\hat{Q}^{\pi_e}$ with a neural network model, which can certainly handle comparisons within continuous states and across more than two actions. This approach produces the same qualitative trends as our reported results, which demonstrates the extent to which our analysis is robust to specific implementation choices. An important conclusion from these results is that simpler implementations are generally advantageous because they involve fewer degrees of freedom that can vary across datasets and thus introduce less estimator variance. At the same time, large variance in the estimate of Q^{π_e} should be limited in its effect on our overall OSIRIS estimator because $\hat{Q}^{\pi_e}$ is only used for comparisons. Furthermore, the binary output of the comparison $\hat{\theta}(s; \mathcal{D}) \in \{0, 1\}$ further weakens any statistical dependence between $\hat{\theta}(s; \mathcal{D})$ and each individual trajectory in the dataset, which could violate the independence assumptions discussed in Section 4.2.

Extensions From the perspective presented in Lemma 1, the OSIRIS procedure is equivalent to performing ordinary IS while treating the composite policy π_e' as the evaluation policy, and doing so does not introduce any new bias. Thus, it is natural to use variants of OSIRIS corresponding to any variant of IS (Thomas & Brunskill, 2016; Thomas, 2015; Jiang & Li, 2016). For example, analogous to WIS, we provide empirical results in Section 6 for OSIRWIS:

$$\hat{V}^{\pi_e}_{\text{OSIRWIS}}(\mathcal{D}; \theta') \equiv \frac{\sum_{\tau \in \mathcal{D}} g(\tau)\, \rho_{\theta'}(\tau; \theta')}{\sum_{\tau \in \mathcal{D}} \rho_{\theta'}(\tau)}. \tag{15}$$

This principle also directly extends to a step-wise IS framework (Precup et al., 2000; Jiang & Li, 2016). Alternatively, because step-wise IS is fundamentally estimating the expected *reward* collected at each time step, it presents an interesting opportunity to define and use the relevance of a state to the reward $\Delta t \in \mathbb{Z}$ time steps away (rather than to the overall trajectory outcome):

Theorem 5. *Let state $s \in \mathcal{S}$ be irrelevant to the reward Δt-steps away if*

$$\mathop{\mathbb{E}}_{\tau \sim \pi_e} \left[r_{t+\Delta t} \,|\, s_t = s, a_t = a \right] = \text{constant}, \quad \forall a \in \mathcal{A}. \tag{16}$$

Otherwise, s is relevant to the reward Δt-steps away. Using this condition, we define $\theta_{\Delta t} : \mathcal{S} \to \{0, 1\}$ where $\theta_{\Delta t}(s) = 0$ if s is irrelevant to the reward Δt-steps away, and otherwise $\theta_{\Delta t}(s) = 1$. Then

$$\hat{V}^{\pi_e}_{\substack{\text{step-wise}\\ \text{OSIRIS}}}(\mathcal{D}; \theta_{\Delta t}) \equiv \frac{1}{|\mathcal{D}|} \sum_{\tau \in \mathcal{D}} \sum_{t'=1}^{T} \left(\gamma^{t'-1} r_{t'} \prod_{t=1}^{T} \left[\rho_t(\tau) \right]^{\theta_{t'-t}(s_t)} \right) \tag{17}$$

is an unbiased estimator of V^{π_e}.

The proof is in Appendix D.4, which uses similar techniques as for Lemma 1 and Theorem 3. Notice that the per-decision importance sampling (PDIS) estimator (Precup et al., 2000) can be seen as a special case of this strategy if we use the relevance mapping $\theta^{\text{PDIS}}_{\Delta t}(s) \equiv \mathbf{1}\{\Delta t \geq 0\}$, and this observation is reflected in Equation 16 because the MDP setup assumes a_t depends only on s_t.

Using this step-wise OSIRIS framework, it is straightforward to extend OSIRIS to the doubly robust (DR) estimator (Jiang & Li, 2016) by replacing $\rho_{t'}(\tau) r_{t'}$ in each term of Equation 17 with the corresponding DR estimate $\rho_{t'}(\tau) \big[r_{t'} - \hat{Q}^{\pi_e}(s_{t'}, a_{t'}) \big] + \sum_{a \in \mathcal{A}} \pi_e(a \,|\, s_{t'}) \hat{Q}^{\pi_e}(s_{t'}, a)$ where $\hat{Q}^{\pi_e}$ is some estimate of Q^{π_e}. Alternatively, the principle behind ordinary OSIRIS can be applied.

[7] Interestingly, the jump in the estimator's behavior from $\alpha = 0$ to $\alpha > 0$ is non-continuous because it involves the first introduction of an IS likelihood ratio.

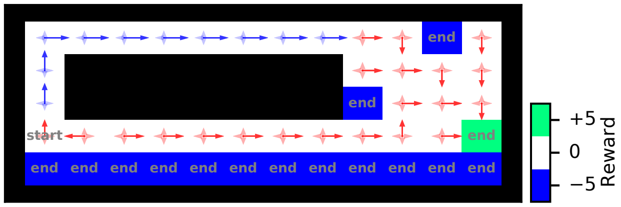


Figure 1: Environment and policies for Gridworld experiments. States with blue arrows comprise the "corridor."

6. Experiments

In this section, we experimentally validate the efficacy of the likelihood ratio omission procedure of OSIRIS and the relevance estimation procedure in Algorithm 1. We demonstrate that they improve estimator accuracy. We demonstrate that this occurs by reducing meaningless variance associated with trajectory length while strategically using $\hat{\theta}(s; \mathcal{D})$ to identify meaningful variance to be kept.

Environment Descriptions Detailed descriptions of the environments and methods are in Appendix E. In summary: All policies are ϵ-greedy where ϵ is smaller in the evaluation policy. We consider variants of the discrete gridworld shown in Figure 1. The agent spends a variable number of transitions dilly-dallying in the "corridor" before navigating to terminal states that award either $+5$ or -5. Compared to this *Dilly-Dallying Gridworld*, the only difference in the *Express Gridworld* variant is that the behavior policy uses an even smaller ϵ but only in the corridor, which reduces the spread of the number of dilly-dallying transitions and thus alleviates trajectory length issues in IS. We also demonstrate that our implementation in Algorithm 1 extends to continuous state spaces, specifically the popular benchmark environments *Cart Pole* and *Lunar Lander*. For only the calculation of state relevance, we discretized the state space by creating linearly spaced bins per state dimension. In Cart Pole, the agent receives $+1$ reward for each transition while it can keep an inverted pendulum upright. In Lunar Lander, the agent is rewarded for landing on target, penalized for firing its engines, and harshly penalized for crashing. Unless specified otherwise, results are aggregated from 200 trials where $|\mathcal{D}| = 25$ for the Gridworlds and $|\mathcal{D}| = 50$ for Cart Pole and Lunar Lander. All code and models used to generate these results are publicly accessible at `github.com/dtak/osiris`.

OSIRIS/OSIRWIS mean-squared errors are generally lower than their IS counterparts. The estimator means, standard deviations, and RMSEs for each environment are listed together in Table 1. OSIRIS/OSIRWIS generally outperforms IS, WIS, PHWIS (Doroudi et al., 2018), INCRIS (Guo et al., 2017), and MIS (Xie et al., 2019). The RMSE improvement is mostly driven by variance reduction. Express Gridworld is the exception (see discussion below), where OSIRIS/OSIRWIS is not expected to do better because we modified the environment to produce trajectories with less length variability. MIS also performs well in the Gridworlds where the state space is small.

OSIRWIS bias decreases as $|\mathcal{D}|$ or α increases. The distributions of OPE value estimates are plotted for different data sizes $|\mathcal{D}|$ and values of α in Figure 2. These results reflect the expected consistency behavior (Theorem 4). As the data size increases, the OSIRWIS estimator mean approaches the true policy value. As α increases, OSIRWIS becomes more similar to the WIS estimator.

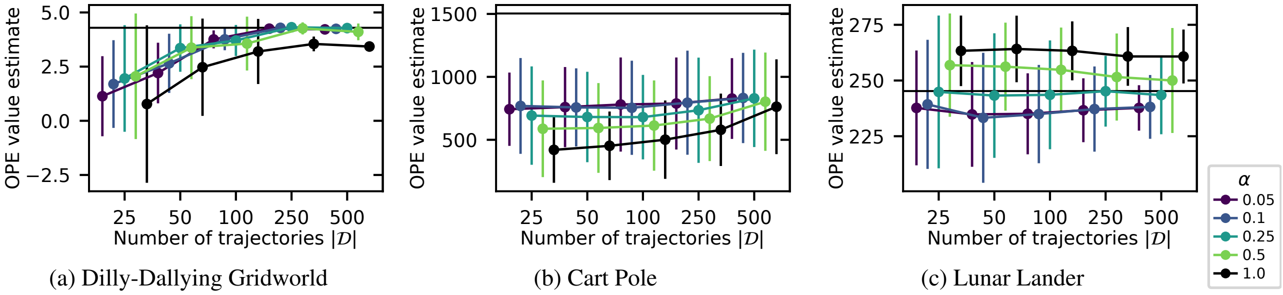
Dilly-dallying contributes high variance to IS/WIS but is ignored by OSIRIS/OSIRWIS. In Section 4.1, we argued that IS weights correlate with the trajectory length, so if trajectory length varies, then that can directly contribute meaningless variance to the IS estimate. This point is reflected in Figure 3.

Consider the Dilly-Dallying Gridworld environment, where the agent will dilly-dally in the corridor (blue arrows in Figure 1). While the random number of dilly-dally transitions affects trajectory length, it does not affect the trajectory return because the agent accumulates zero reward in the corridor and almost always exits the corridor from the same state. IS/WIS is distracted by dilly-dally transitions in the corridor: the likelihood ratios corresponding to these low-probability transitions are less than 1, so each dilly-dally transition makes the IS/WIS weight smaller. A large and variable number of dilly-dally likelihood ratios dominates the IS/WIS weights and masks the informative likelihood ratios corresponding to transitions outside the corridor. As such, the IS/WIS weights become more informative of the number of dilly-dally transitions (trajectory length) than whether a behavior trajectory is representative of the evaluation policy (scatter plot in Figure 3a). Because trajectory length is not directly meaningful in this setting, this information hurts IS/WIS accuracy by contributing meaningless variance. Meanwhile, the OSIRIS/OSIRWIS estimator significantly reduces this source of variance (boxplots in Figure 3a) by omitting likelihood ratios corresponding to these corridor states (Figure 4a).

In contrast, the Express Gridworld variant generates trajectories with less dilly-dallying in the corridor. Although this modification of the environment/policies does not affect policy value, it reduces the meaningless spread of trajectory lengths. Because there are fewer distracting likelihood ratios, the IS/WIS estimator becomes more accurate. Meanwhile, the OSIRIS/OSIRWIS estimator is robust to

Table 1: Comparison of mean squared errors for IS-based estimators. OSIRIS/OSIRWIS with $\alpha = 0.05$ generally outperforms its IS counterparts except in Express Gridworld (see text).

		IS	WIS	PHWIS	INCRIS	MIS	OSIRIS	OSIRWIS	ON-POLICY
DILLY-DALLYING GRIDWORLD	MEAN	**3.5**	1.1	−0.3	−0.1	5.8	1.3	1.1	4.3
	STD	6.8	3.5	**1.0**	4.8	1.6	2.0	1.8	0.6
	RMSE	6.9	4.7	4.7	6.5	**2.2**	3.6	3.7	0.6
EXPRESS GRIDWORLD	MEAN	3.0	2.7	0.3	**4.3**	5.1	0.7	0.8	4.3
	STD	3.5	2.2	**1.1**	4.0	1.7	1.2	1.3	0.6
	RMSE	3.7	2.7	4.1	4.0	**1.9**	3.8	3.7	0.6
CART POLE	MEAN	**1073.5**	452.2	608.2	1048.4	2498.6	1068.9	759.7	1503.6
	STD	10202.2	272.5	**97.6**	4437.7	583.1	3961.8	318.5	244.8
	RMSE	10211.3	1086.1	900.7	4461.0	1153.3	3985.5	**809.2**	244.8
LUNAR LANDER	MEAN	305.6	264.2	239.6	83.8	281.9	**244.6**	234.8	245.3
	STD	768.7	14.9	**9.1**	149.6	139.6	55.3	23.5	6.8
	RMSE	771.1	24.1	**10.7**	220.2	144.3	55.3	25.7	6.8



(a) Dilly-Dallying Gridworld

(b) Cart Pole

(c) Lunar Lander

Figure 2: Distributions of OSIRWIS estimates showing estimator consistency. Dots represent means, error bars represent standard deviations, and horizontal line represents the mean of the on-policy MC estimator. Colors indicate α values, where $\alpha = 1$ is equivalent to ordinary WIS.

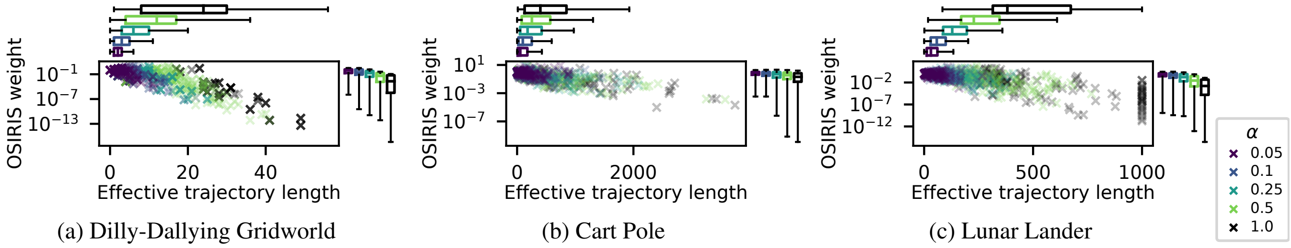
this modification: estimator accuracy remains constant (Table 1), and the other reported results still show the same qualitative trends (Appendix F.2).

While is difficult to precisely interpret dilly-dallying in the high-dimensional Cart Pole and Lunar Lander environments, we still observe the same variance reduction trend (Figures 3b and c).

Estimated state relevance $\hat{\theta}$ identifies key decision points where trajectory outcome is sensitive to the action taken. Figure 4 plots the proportion of trials in which the indicated state was estimated to be relevant (i.e. $\hat{\theta}(s;\mathcal{D}) = 1$). As designed, $\hat{\theta}(s;\mathcal{D}) = 1$ identifies key decision points where trajectory outcome is sensitive to the action taken, and often many different trajectory outcomes are accessible from the state.

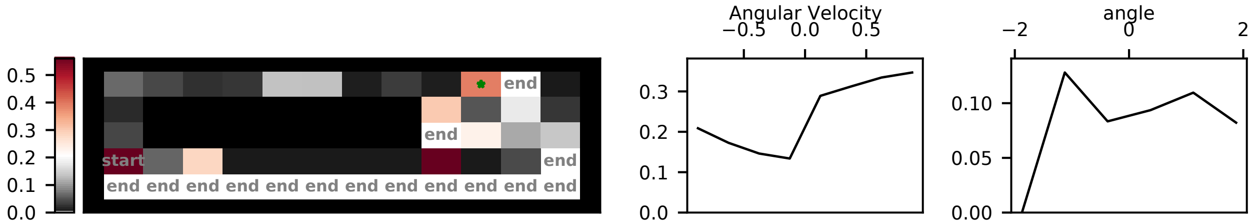
For example, in the Dilly-Dallying Gridworld, consider the state marked with a green star in Figure 4a. Here, if the agent moves east, then the corresponding likelihood ratio < 1, and the trajectory ends with -5 return. If the agent moves south, then the corresponding likelihood ratio > 1, and the agent will very likely end the trajectory with $+5$ return. Thus, there is a clear relationship between likelihood ratio and trajectory return in this state, which is reflected by OSIRIS's tendency to set $\hat{\theta}(s;\mathcal{D}) = 1$. Meanwhile, $\hat{\theta}(s;\mathcal{D})$ is often 0 in the corridor (blue arrows in Figure 1). In any of these corridor states, the likelihood ratio > 1 if the agent takes the main policy action or < 1 if it dilly-dallies. However, by the Markov assumption, this decision is independent of the agent's future actions to receive either $+5$ or -5 trajectory return. As such, the statistical test should not find any relationship between likelihood ratio and trajectory return, which is reflected by the tendency for $\hat{\theta}(s;\mathcal{D}) = 0$.

The Lunar Lander and Cart Pole environments also demonstrate this principle: $\hat{\theta}(s;D) = 1$ identifies relevant states where the agent can avoid crashing by following the optimal policy action but will likely crash if a random action is taken. This establishes a positive correlation between trajectory return and likelihood ratio that is detected by the statistical test. In Cart Pole, the detected relevant states tend to have large angular velocity that can be stabilized by taking the optimal policy action instead of a random action (Figure 4b).



(a) Dilly-Dallying Gridworld (b) Cart Pole (c) Lunar Lander

Figure 3: Scatter plots show correlation between OSIRIS weights and effective trajectory lengths $\sum_{t=1}^{T} \hat{\theta}(s_t)$. Boxplots show variance reduction of OSIRIS weights by shortening and evening of the effective trajectory lengths. Colors indicate α values, where $\alpha = 1$ is equivalent to ordinary IS.



(a) Dilly-Dallying Gridworld (b) Cart Pole (c) Lunar Lander

Figure 4: Mean of estimated state relevance $\hat{\theta}(s)$ from visits to the indicated states is represented by color (a) or on the y axis (b, c). States identified as relevant (i.e. $\hat{\theta}(s) = 1$) are key decision points where trajectory outcome is sensitive to action taken.

Similarly, in Lunar Lander, when the agent is not level, it is important whether the agent chooses stabilizing actions (Figure 4c). But interestingly, if it has rotated too far, the state tends to be considered irrelevant – at this point, there is no hope as the agent will likely crash no matter what action it takes. This principle can be observed in the other state dimensions too (Appendix F.3).

7. Discussion and Conclusion

We presented the OSIRIS estimator to reduce importance sampling variance in settings with long and varying trajectory lengths. The algorithm strategically identifies and omits irrelevant likelihood ratios in a way that introduces minimal bias. This procedure can technically be applied to any IS-based estimator or even for the purposes of interpretability (Gottesman et al., 2020).

We also describe when OSIRIS will provide the most benefit. In environments where trajectory length is not directly correlated with the trajectory return, OSIRIS will shine because the covariance term in Equations 8 and 9 is small: length is meaningless (assumed in these environments) and the individual omitted likelihood ratios are also meaningless (by state irrelevance). In contrast, in environments like Cart Pole, where trajectory length is directly related to trajectory return, likelihood ratio omission could disrupt any naturally occurring relationship between trajectory return and IS weight (aka trajectory length). Depending on the direction of the relationship, which determines the sign of the covariance term (aka the bias term), this could be favorable or harmful by shifting the estimator distribution closer to or further from the true value.

More broadly, OSIRIS's likelihood ratio omission directly addresses the variance of *long* trajectories, but depending on the environment, it may create more *varying* trajectory length, e.g. by omitting several likelihood ratios in some trajectories while leaving others untouched. While we did not see this as a dominating factor in our experiments, future extensions could address this issue of uneven lengths by keeping only the $T_{\max}$-most relevant likelihood ratios for some constant $T_{\max}$. Nonetheless, because we omit irrelevant likelihood ratios, we at least expect the OSIRIS weight to give more attention to the relevant likelihood ratios.

Acknowledgements

We thank the reviewers for their thoughtful comments. Part of this work was completed while S.P.S. was supported by NSF GRFP. S.P.S. and F.D.V. also acknowledge support from NSF RI-Small 2007076.

References

Doroudi, S., Thomas, P., and Brunskill, E. Importance sampling for fair policy selection. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 5239–5243, 2018.

Farajtabar, M., Chow, Y., and Ghavamzadeh, M. More robust doubly robust off-policy evaluation. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1446–1455, 2018.

Gottesman, O., Futoma, J., Liu, Y., Parbhoo, S., Celi, L., Brunskill, E., and Doshi-Velez, F. Interpretable off-policy evaluation in reinforcement learning by highlighting influential transitions. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 3658–3667, 2020.

Guo, Z. D., Thomas, P. S., and Brunskill, E. Using options and covariance testing for long horizon off-policy policy evaluation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 2493–2502, 2017.

Hodges, J. L. The significance probability of the smirnov two-sample test. *Arkiv för matematik*, 3(5):469–486, 1958.

Ionides, E. L. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.

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

Liu, Q., Li, L., Tang, Z., and Zhou, D. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 5356–5366, 2018.

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

Metelli, A. M., Papini, M., Montali, N., and Restelli, M. Importance sampling techniques for policy optimization. *Journal of Machine Learning Research*, 21:1–75, 2020.

Murphy, S. A., Van der Laan, M. J., Robins, J. M., Bierman, K. L., Coie, J. D., Greenberg, M. T., Lochman, J. E., McMahon, R. J., and Pinderhughes, E. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423, 2001.

Papini, M., Metelli, A. M., Lupo, L., and Restelli, M. Optimistic policy optimization via multiple importance sampling. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 4989–4999, 2019.

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

Rowland, M., Harutyunyan, A., van Hasselt, H., Borsa, D., Schaul, T., Munos, R., and Dabney, W. Conditional importance sampling for off-policy learning. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pp. 45–55, 2020.

Su, Y., Dimakopoulou, M., Krishnamurthy, A., and Dudík, M. Doubly robust off-policy evaluation with shrinkage. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 9167–9176, 2019.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

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

Thomas, P. S. *Safe reinforcement learning*. PhD thesis, University of Massachusetts Amherst, 2015.

Thomas, P. S., Theocharous, G., and Ghavamzadeh, M. High confidence off-policy evaluation. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pp. 3000–3006, 2015a.

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

Welch, B. L. On the comparison of several mean values: An alternative approach. *Biometrika*, 38(3/4):330–336, 1951.

Xie, T., Ma, Y., and Wang, Y. X. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 1–36, 2019.