In *Proceedings of the 36th International Conference on Machine Learning* (ICML 2019),
Long Beach, CA, USA June 2019

Importance Sampling Policy Evaluation with an Estimated Behavior Policy

Josiah P. Hanna [1] **Scott Niekum** [1] **Peter Stone** [1]

Abstract

We consider the problem of off-policy evaluation in Markov decision processes. Off-policy evaluation is the task of evaluating the expected return of one policy with data generated by a different, *behavior* policy. Importance sampling is a technique for off-policy evaluation that re-weights off-policy returns to account for differences in the likelihood of the returns between the two policies. In this paper, we study importance sampling with an estimated behavior policy where the behavior policy estimate comes from the same set of data used to compute the importance sampling estimate. We find that this estimator often lowers the mean squared error of off-policy evaluation compared to importance sampling with the true behavior policy or using a behavior policy that is estimated from a separate data set. Intuitively, estimating the behavior policy in this way corrects for error due to sampling in the action-space. Our empirical results also extend to other popular variants of importance sampling and show that estimating a non-Markovian behavior policy can further lower large-sample mean squared error even when the true behavior policy is Markovian.

1. Introduction

Sequential decision-making tasks, such as a robot manipulating objects or an autonomous vehicle deciding when to change lanes, are ubiquitous in artificial intelligence. For these tasks, *reinforcement learning* (RL) algorithms provide a promising alternative to hand-coded skills, allowing sequential decision-making agents to acquire policies autonomously given only a reward function measuring task performance (Sutton & Barto, 1998). When applying RL to real world problems, an important problem that often comes up is *policy evaluation*. In policy evaluation, the goal is to determine the expected return – sum of rewards – that an *evaluation policy*, π_e, will obtain when deployed on the task of interest.

In *off-policy* policy evaluation, we are given data (in the form of state-action-reward trajectories) generated by a second *behavior policy*, π_b. We then use these trajectories to evaluate π_e. Accurate off-policy policy evaluation is especially important when we want to know the value of a policy before it is deployed in the real world or have many policies to evaluate and want to avoid running each one individually. *Importance sampling* addresses this problem by re-weighting returns generated by π_b such that they are unbiased estimates of π_e (Precup et al., 2000). While the basic importance sampling estimator is often noted in the literature to suffer from high variance, more recent importance sampling estimators have lowered this variance (Thomas & Brunskill, 2016; Jiang & Li, 2016). Regardless of additional variance reduction techniques, all importance sampling variants compute the likelihood ratio $\frac{\pi_e(a|s)}{\pi_b(a|s)}$ for all state-action pairs in the off-policy data.

In this paper, we propose to replace $\pi_b(a|s)$ with its empirical estimate – that is, we replace the probability of sampling an action in a particular state with the frequency at which that action actually occurred in that state in the data. It is natural to assume that such an estimator will yield worse performance since it replaces a known quantity with an estimated quantity. However, research in the multi-armed bandit (Li et al., 2015; Narita et al., 2019), causal inference (Hirano et al., 2003; Rosenbaum, 1987), and Monte Carlo integration (Henmi et al., 2007; Delyon & Portier, 2016) literature has demonstrated that estimating the behavior policy can *improve* the mean squared error of importance sampling policy evaluation. Motivated by these results, we study the performance of such methods for policy evaluation in full Markov decision processes.

Specifically, we study a family of estimators that, given a dataset, $\mathcal{D}$, of trajectories, use $\mathcal{D}$ both to estimate the behavior policy and then to compute the importance sampling estimate. Though related to methods in the statistics literature, the so-called regression importance sampling methods are specific to Markov decision processes where actions taken at one time-step influence the states and rewards at future time-steps. We show empirically that regression importance

[1] The University of Texas at Austin, Austin, Texas, USA. Correspondence to: Josiah P. Hanna <jphanna@cs.utexas.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

sampling *lowers* the mean squared error of importance sampling off-policy evaluation in both discrete and continuous action spaces. Though our study is primarily empirical, we present theoretical results that, when the policy class of the estimated behavior policy is specified correctly, regression importance sampling is consistent and has asymptotically lower variance than using the true behavior policy for importance sampling. To the best of our knowledge, we are the first to study this method for policy evaluation in Markov decision processes.

2. Preliminaries

This section formalizes our problem and introduces importance sampling off-policy evaluation.

2.1. Notation

We assume the environment is a finite horizon, episodic *Markov decision process* with state space $\mathcal{S}$, action space $\mathcal{A}$, transition probabilities, P, reward function R, horizon L, discount factor γ, and initial state distribution d_0 (Puterman, 2014). A *Markovian* policy, π, is a function mapping the current state to a probability distribution over actions; a policy is *non-Markovian* if its action distribution is conditioned on past states or actions. For simplicity, we assume that $\mathcal{S}$ and $\mathcal{A}$, are finite and that probability distributions are probability mass functions.[1] Let $H := (S_0, A_0, R_0, S_1, \dots, S_{L-1}, A_{L-1}, R_{L-1})$ be a *trajectory*, $g(H) := \sum_{t=0}^{L-1} \gamma^t R_t$ be the *discounted return* of trajectory H, and $v(\pi) := \mathbf{E}[g(H)|H \sim \pi]$ be the expected discounted return when the policy π is used starting from state S_0 sampled from the initial state distribution. We assume that the transition and reward functions are unknown and that the episode length, L, is a finite constant.

In off-policy policy evaluation, we are given a fixed *evaluation policy*, π_e, and a data set of m trajectories and the policies that generated them: $\mathcal{D} := \{H_i, {\pi_b}^{(i)}\}_{i=1}^m$ where $H_i \sim {\pi_b}^{(i)}$. We assume that $\forall \{H_i, {\pi_b}^{(i)}\} \in \mathcal{D}$, ${\pi_b}^{(i)}$ is Markovian i.e., actions in $\mathcal{D}$ are independent of past states and actions given the immediate preceding state. Our goal is to design an off-policy estimator, OPE, that takes $\mathcal{D}$ and estimates $v(\pi_e)$ with minimal mean squared error (MSE). Formally, we wish to minimize $\mathbf{E}_\mathcal{D}[(\text{OPE}(\pi_e, \mathcal{D}) - v(\pi_e))^2]$.

2.2. Importance Sampling

Importance Sampling (IS) is a method for reweighting returns generated by a *behavior* policy, π_b, such that they are unbiased returns from the *evaluation* policy. Given a set of m trajectories and the policy that generated each trajectory, the IS off-policy estimate of $v(\pi_e)$ is:

$$\text{IS}(\pi_e, \mathcal{D}) := \frac{1}{m} \sum_{i=1}^m g(H^{(i)}) \prod_{t=0}^{L-1} \frac{\pi_e(A_t^{(i)}|S_t^{(i)})}{{\pi_b}^{(i)}(A_t^{(i)}|S_t^{(i)})}. \quad (1)$$

We refer to (1) – that uses the true behavior policy – as the ordinary importance sampling (OIS) estimator and refer to $\frac{\pi_e(A|S)}{\pi_b(A|S)}$ as the OIS weight for action A in state S.

The importance sampling estimator with OIS weights can be understood as a Monte Carlo estimate of $v(\pi_e)$ with a correction for the distribution shift caused by sampling trajectories from π_b instead of π_e. As more data is obtained, the empirical frequency of any trajectory approaches the expected frequency under π_b and then the OIS weight corrects the weighting of each trajectory to reflect the expected frequency under π_e.

3. Sampling Error in Importance Sampling

The ordinary importance sampling estimator (1) is known to have high variance. A number of importance sampling variants have been proposed to address this problem, however, all such variants use the OIS weight. The common reliance on OIS weights suggest that an implicit assumption in the RL community is that OIS weights lead to the most accurate estimate. Hence, when an application requires estimating an unknown π_b in order to compute importance weights, the application is implicitly assumed to only be approximating the desired weights.

However, OIS weights themselves are sub-optimal in at least one respect: the weight of each trajectory in the OIS estimate is inaccurate unless we happen to observe each trajectory according to its true probability. When the empirical frequency of any trajectory is unequal to its expected frequency under π_b, the OIS estimator puts either too much or too little weight on the trajectory. We refer to error due to some trajectories being either over- or under-represented in $\mathcal{D}$ as *sampling error*. Sampling error may be unavoidable when we desire an unbiased estimate of $v(\pi_e)$. However, correcting for it by properly weighting trajectories will, in principle, give us a lower mean squared error estimate.

The problem of sampling error is related to a Bayesian objection to Monte Carlo integration techniques: OIS ignores information about the closeness of trajectories in $\mathcal{D}$ (O'Hagan, 1987; Ghahramani & Rasmussen, 2003). This objection is easiest to understand in deterministic and discrete environments though it also holds for stochastic and continuous environments. In a deterministic environment, additional samples of any trajectory, h, provide no new information about $v(\pi_e)$ since only a single sample of h is required to know $g(h)$. However, the more times a particular trajectory appears, the more weight it receives in an OIS estimate even though the correct weighting of $g(h)$, $\Pr(h|\pi_e)$,

[1] Unless otherwise noted, all results and discussion apply equally to the discrete and continuous setting.

is known since π_e is known. In stochastic environments, it is reasonable to give more weight to recurring trajectories since the recurrence provides additional information about the unknown state-transition and reward probabilities. However, ordinary importance sampling also relies on sampling to approximate the known policy probabilities.

Finally, we note that the problem of sampling error applies to any variant of importance sampling using OIS weights, e.g., weighted importance sampling (Precup et al., 2000), per-decision importance sampling (Precup et al., 2000), the doubly robust estimator (Jiang & Li, 2016; Thomas & Brunskill, 2016), and the MAGIC estimator (Thomas & Brunskill, 2016). Sampling error is also a problem for on-policy Monte Carlo policy evaluation since Monte Carlo is the special case of OIS when the behavior policy is the same as the evaluation policy.

4. Regression Importance Sampling

In this section we introduce the primary focus of our work: a family of estimators called regression importance sampling (RIS) estimators that correct for sampling error in $\mathcal{D}$ by importance sampling with an estimated behavior policy. The motivation for this approach is that, though $\mathcal{D}$ was sampled with π_b, the trajectories in $\mathcal{D}$ may appear as if they had been generated by a different policy, $\pi_\mathcal{D}$. For example, if π_b would choose between two actions with equal probability in a particular state, the data might show that one action was selected more often than the other in that state. Thus instead of using OIS to correct from π_b to π_e, we introduce RIS that corrects from $\pi_\mathcal{D}$ to π_e.

We assume that, in addition to $\mathcal{D}$, we are given a policy class – a set of policies – Π^n where each $\pi \in \Pi^n$ is a distribution over actions conditioned on an n-step state-action history: $\pi : \mathcal{S}^{n+1} \times \mathcal{A}^n \rightarrow [0, 1]$. Let $H_{t-n:t}$ be the trajectory segment: $S_{t-n}, A_{t-n}, ...S_{t-1}, A_{t-1}, S_t$ where if $t - n < 0$ then $H_{t-n:t}$ denotes the beginning of the trajectory until step t. The RIS(n) estimator first estimates the maximum likelihood behavior policy in Π^n given $\mathcal{D}$:

$$\pi_\mathcal{D}{}^{(n)} := \operatorname*{argmax}_{\pi \in \Pi^n} \sum_{H \in \mathcal{D}} \sum_{t=0}^{L-1} \log \pi(a|H_{t-n:t}). \qquad (2)$$

The RIS(n) estimate is then the importance sampling estimate with $\pi_\mathcal{D}{}^{(n)}$ replacing π_b:

$$\text{RIS}(n)(\pi_e, \mathcal{D}) := \frac{1}{m} \sum_{i=1}^{m} g(H_i) \prod_{t=0}^{L-1} \frac{\pi_e(A_t|S_t)}{\pi_\mathcal{D}{}^{(n)}(A_t|H_{t-n:t})}$$

Analogously to OIS, we refer to $\frac{\pi_e(A_t|S_t)}{\pi_\mathcal{D}{}^{(n)}(S_t|H_{t-n:t})}$ as the RIS(n) weight for action A_t, state S_t, and trajectory segment $H_{t-n:t}$. Note that the RIS(n) weights are always well-defined since $\pi_\mathcal{D}{}^{(n)}$ never places zero probability mass on any action that occurred in $\mathcal{D}$.

4.1. Correcting Importance Sampling Sampling Error

We now present an example illustrating how RIS corrects for sampling error in off-policy data.

Consider a deterministic MDP with finite $|\mathcal{S}|$ and $|\mathcal{A}|$. Let $\mathcal{H}$ be the (finite) set of possible trajectories under π_b and suppose that our observed data, $\mathcal{D}$, contains at least one of each $h \in \mathcal{H}$. In this setting, the maximum likelihood behavior policy can be computed with count-based estimates. We define $c(h_{i:j})$ as the number of times that trajectory segment $h_{i:j}$ appears during any trajectory in $\mathcal{D}$. Similarly, we define $c(h_{i:j}, a)$ as the number of times that action a is observed following trajectory segment $h_{i:j}$ during any trajectory in $\mathcal{D}$. RIS(n) estimates the behavior policy as:

$$\pi_\mathcal{D}(a|h_{i-n:i}) := \frac{c(h_{i-n:i}, a)}{c(h_{i-n:i})}.$$

Observe that both OIS and all variants of RIS can be written in one of two forms:

$$\underbrace{\frac{1}{m} \sum_{i=1}^{m} \frac{w_{\pi_e}(h_i)}{w_\pi(h_i)} g(h_i)}_{(i)} = \underbrace{\sum_{h \in \mathcal{H}} \frac{c(h)}{m} \frac{w_{\pi_e}(h)}{w_\pi(h)} g(h)}_{(ii)}$$

where $w_\pi(h) = \prod_{t=0}^{L-1} \pi(a_t|s_t)$ and for OIS $\pi := \pi_b$ and for RIS(n) $\pi := \pi_\mathcal{D}{}^{(n)}$ as defined in Equation (2).

If we had sampled trajectories using $\pi_\mathcal{D}{}^{(L-1)}$ instead of π_b, in our deterministic environment, the probability of each trajectory would be $\Pr(H|\pi_\mathcal{D}{}^{(L-1)}) = \frac{c(H)}{m}$. Thus Form (ii) can be written as:

$$\mathbf{E}\left[\frac{w_{\pi_e}(H)}{w_\pi(H)} g(H) | H \sim \pi_\mathcal{D}{}^{(L-1)}\right].$$

To emphasize what we have shown so far: OIS and RIS are both sample-average estimators whose estimates can be written as exact expectations. However, this exact expectation is under the distribution that trajectories were observed and *not* the distribution of trajectories under π_b.

Consider choosing $w_\pi := w_{\pi_\mathcal{D}}^{(L-1)}$ as RIS$(L-1)$ does. This choice results in (ii) being exactly equal to $v(\pi_e)$[2] On the other hand, choosing $w_\pi := w_{\pi_b}$ will *not* return $v(\pi_e)$ unless we happen to observe each trajectory at its expected frequency (i.e., $\pi_\mathcal{D}{}^{(L-1)} = \pi_b$).

Choosing w_π to be $w_{\pi_\mathcal{D}{}^{(n)}}$ for $n < L - 1$ also does *not* result in $v(\pi_e)$ being returned in this example. This observation is surprising because even though we know that the true $\Pr(h|\pi_b) = \prod_{t=0}^{L-1} \pi_b(a_t|s_t)$, it does not follow

[2] This statement follows from the importance sampling identity: $\mathbf{E}[\frac{\Pr(H|\pi_e)}{\Pr(H|\pi)} g(h)|H \sim \pi] = \mathbf{E}[g(H)|H \sim \pi_e] = v(\pi_e)$ and the fact that we have assumed a deterministic environment.

that the estimated probability of a trajectory is equal to the product of the estimated Markovian action probabilities, i.e., that $\frac{c(h)}{m} = \prod_{t=0}^{L-1} \pi_{\mathcal{D}}{}^{(0)}(a_t|s_t)$. With a finite number of samples, the data may have higher likelihood under a non-Markovian behavior policy – possibly even a policy that conditions on all past states and actions. Thus, to fully correct for sampling error, we must importance sample with an estimated non-Markovian behavior policy. However, $w_{\pi_{\mathcal{D}}{}^{(n)}}$ with $n < L-1$ still provides a better sampling error correction than w_{π_b} since any $\pi_{\mathcal{D}}{}^{(n)}$ will reflect the statistics of $\mathcal{D}$ while π_b does not. This statement is supported by our empirical results comparing $\text{RIS}(0)$ to OIS and a theoretical result we present in the following section that states that $\text{RIS}(n)$ has lower asymptotic variance than OIS for all n.

Before concluding this section, we discuss two limitations of the presented example – these limitations are *not* present in our theoretical or empirical results. First, the example lacks stochasticity in the rewards and transitions. In stochastic environments, sampling error arises from sampling states, actions, and rewards while in deterministic environments, sampling error only arises from sampling actions. Neither RIS nor OIS can correct for state and reward sampling error since such a correction requires knowledge of what the true state and reward frequencies are and these quantities are typically unknown in the MDP policy evaluation setting.

Second, we assumed that $\mathcal{D}$ contains at least one of each trajectory possible under π_b. If a trajectory is absent from $\mathcal{D}$ then $\text{RIS}(L-1)$ has non-zero bias. Theoretical analysis of this bias for both $\text{RIS}(L-1)$ and other RIS variants is an open question for future analysis.

4.2. Theoretical Properties of RIS

Here, we briefly summarize new theoretical results (full proofs appear in the appendices) as well as a connection to prior work from the multi-armed bandit literature:

- **Proposition 1:** For all n, $\text{RIS}(n)$ is a biased estimator, however, it is consistent provided $\pi_b \in \Pi^n$ (see Appendix A for a full proof).
- **Corollary 1:** For all n, if $\pi_b \in \Pi^n$ then RIS has asymptotic variance at most that of OIS. This result is a corollary to a result by Henmi et al. (2007) for general Monte Carlo integration (see Appendix B for a full proof). We highlight that the derivation of this result includes some $o(n)$ and $o_p(1)$ terms that may be large for small sample sizes; the lower variance is asymptotic and we leave analysis of the finite-sample variance of RIS to future work.
- **Connection to REG:** For finite MDPs, Li et al. (2015) introduce the *regression* (REG) estimator and show it has asymptotic lower minimax MSE than OIS provided the estimator has full knowledge of the environment's transition probabilities. With this knowledge REG can correct for sampling error in both the actions and state transitions. $\text{RIS}(L-1)$ is an approximation to REG that only corrects for sampling error in the actions. The derivation of the connection between REG and $\text{RIS}(L-1)$ is given in Appendix C.

We also note that prior theoretical analysis of importance sampling with an estimated behavior policy has made the assumption that $\pi_{\mathcal{D}}$ is estimated independently of $\mathcal{D}$ (Dudík et al., 2011; Farajtabar et al., 2018). This assumption simplifies the theoretical analysis but makes it inapplicable to regression importance sampling.

4.3. RIS with Function Approximation

The example in Section 4.1 presented RIS with count-based estimation of $\pi_{\mathcal{D}}$. In many practical settings, count-based estimation of $\pi_{\mathcal{D}}$ is intractable and we must rely on function approximation. For example, in our final experiments we learn $\pi_{\mathcal{D}}$ as a Gaussian distribution over actions with the mean given by a neural network. Two practical concerns arise when using function approximation for RIS: avoiding over-fitting and selecting the function approximator.

RIS uses all of the data available for off-policy evaluation to both estimate $\pi_{\mathcal{D}}$ and compute the off-policy estimate of $v(\pi_e)$. Unfortunately, the RIS estimate may suffer from high variance if the function approximator is too expressive and $\pi_{\mathcal{D}}$ is over-fit to our data. Additionally, if the policy class of π_b is unknown, it may be unclear what is the right function approximation representation for $\pi_{\mathcal{D}}$. A practical solution is to use a validation set – distinct from $\mathcal{D}$ – to select an appropriate policy class and appropriate regularization criteria for RIS. This solution is a small departure from the previous definition of RIS as selecting $\pi_{\mathcal{D}}$ to maximize the log likelihood on $\mathcal{D}$. Rather, we select $\pi_{\mathcal{D}}$ to maximize the log likelihood on $\mathcal{D}$ while avoiding over-fitting. This approach represents a trade-off between robust empirical performance and potentially better but more sensitive estimation with RIS.

5. Empirical Results

We present an empirical study of the RIS estimator across several policy evaluation tasks. Our experiments are designed to answer the following questions:

1. What is the empirical effect of replacing OIS weights with RIS weights in sequential decision making tasks?
2. How important is using $\mathcal{D}$ to both estimate the behavior policy and compute the importance sampling estimate?
3. How does the choice of n affect the MSE of $\text{RIS}(n)$?

With non-linear function approximation, our results suggest that the standard supervised learning approach of model

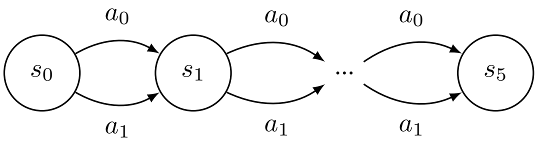


Figure 1: The SinglePath MDP. This environment has 5 states, 2 actions, and $L = 5$. The agent begins in state 0 and both actions either take the agent from state n to state $n + 1$ or cause the agent to remain in state n. **Not shown:** If the agent takes action a_1 it remains in its current state with probability 0.5.

selection using hold-out validation loss may be sub-optimal for the regression importance sampling estimator. Thus, we also investigate the question:

4. Does minimizing hold-out validation loss set yield the minimal MSE regression importance sampling estimator when estimating $\pi_\mathcal{D}$ with gradient descent and neural network function approximation?

5.1. Empirical Set-up

We run policy evaluation experiments in several domains. We provide a short description of each domain here; a complete description and additional experimental details are given in Appendix E.[3]

- **Gridworld:** This domain is a 4×4 Gridworld used in prior off-policy evaluation research (Thomas & Brunskill, 2016; Hanna et al., 2017). RIS uses count-based estimation of π_b. This domain allows us to study RIS separately from questions of function approximation.
- **SinglePath:** See Figure 1 for a description. This domain is small enough to allow implementations of RIS$(L-1)$ and the REG method from Li et al. (2015). All RIS methods use count-based estimation of π_b.
- **Linear Dynamical System:** This domain is a point-mass agent moving towards a goal in a two dimensional world by setting x and y acceleration. Policies are linear in a second order polynomial transform of the state features. We estimate $\pi_\mathcal{D}$ with least squares.
- **Simulated Robotics:** We also use two continuous control tasks from the OpenAI gym: Hopper and HalfCheetah.[4] In each task, we use neural network policies with 2 layers of 64 $\tanh$ hidden units each for π_e and π_b.

5.2. Empirical Results

We now present our empirical results. Except where specified otherwise, RIS refers to RIS(0).

Finite MDP Policy Evaluation Our first experiment compares several importance sampling variants implemented with both RIS weights and OIS weights. Specifically, we use the basic IS method described in Section 2, the *weighted* IS estimator (Precup et al., 2000), and the *weighted doubly robust* estimator (Thomas & Brunskill, 2016).

Figure 2(a) shows the MSE of the evaluated methods averaged over 100 trials. The results show that using RIS weights improves all IS variants relative to OIS weights.[5]

We also evaluate alternative data sources for estimating $\pi_\mathcal{D}$ in order to establish the importance of using $\mathcal{D}$ to both estimate $\pi_\mathcal{D}$ and compute the value estimate. Specifically, we consider:

1. **Independent Estimate**: In addition to $\mathcal{D}$, this method has access to an additional set, $\mathcal{D}_\texttt{train}$. The behavior policy is estimated with $\mathcal{D}_\texttt{train}$ and the policy value estimate is computed with $\mathcal{D}$. Since (s, a) pairs in $\mathcal{D}$ may be absent from $\mathcal{D}_\texttt{train}$ we use Laplace smoothing to ensure that the importance weights are well-defined.
2. **Extra-data Estimate**: This baseline is the same as **Independent Estimate** except it uses both $\mathcal{D}_\texttt{train}$ and $\mathcal{D}$ to estimate π_b. Only $\mathcal{D}$ is used to compute the policy value estimate.

Figure 2(b) shows that these alternative data sources for estimating π_b decrease accuracy compared to RIS and OIS. **Independent Estimate** has high MSE when the sample size is small but its MSE approaches that of OIS as the sample size grows. We understand this result as showing that this baseline cannot correct for sampling error in the off-policy data since the behavior policy estimate is unrelated to the data used in the off-policy evaluation. **Extra-data Estimate** initially has high MSE but its MSE decreases faster than that of OIS. Since this baseline estimates π_b with data that includes $\mathcal{D}$, it can partially correct for sampling error – though the extra data harms its ability to do so. Only estimating $\pi_\mathcal{D}$ with $\mathcal{D}$ and $\mathcal{D}$ alone improves performance over OIS for all sample sizes.

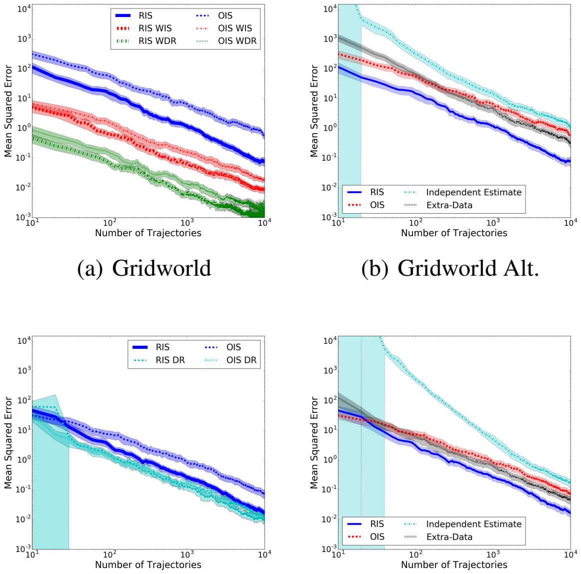
We also repeat these experiments for the on-policy setting and present results in Figure 2(c) and Figure 2(d). We observe similar trends as in the off-policy experiments suggesting that RIS can lower variance in Monte Carlo sampling methods even when OIS weights are otherwise unnecessary.

RIS(n) In the Gridworld domain it is difficult to observe the performance of RIS(n) for various n because of the long horizon: smaller n perform similarly and larger n scale poorly with L. To see the effects of different n more clearly, we use the SinglePath domain. Figure 3 gives the mean

[3] Code is provided at https://github.com/LARG/regression-importance-sampling.

[4] For these tasks we use the Roboschool versions: https://github.com/openai/roboschool

[5] We also implemented and evaluated *per-decision* importance sampling and the ordinary *doubly robust* estimator and saw similar results. However we defer these results to Appendix F for clarity.



(a) Gridworld

(b) Gridworld Alt.

(c) Gridworld On-Policy

(d) Gridworld On-Policy Alt.

Figure 2: Gridworld policy evaluation results. In all subfigures, the x-axis is the number of trajectories collected and the y-axis is mean squared error. Axes are log-scaled. The shaded region gives a 95% confidence interval. (a) Gridworld Off-policy Evaluation: The main point of comparison is the RIS variant of each method to the OIS variant of each method. (b) Gridworld $\pi_{\mathcal{D}}$ Estimation Alternatives: This plot compares RIS and OIS to two methods that replace the true behavior policy with estimates from data sources other than $\mathcal{D}$. Subfigures (c) and (d) repeat experiments (a) and (b) with the behavior policy from (c) and (d) as the evaluation policy.

squared error for OIS, RIS, and the REG estimator of Li et al. (2015) that has full access to the environment's transition probabilities. For RIS, we use $n = 0, 3, 4$ and each method is ran for 200 trials.

Figure 3 shows that higher values of n and REG tend to give inaccurate estimates when the sample size is small. However, as data increases, these methods give increasingly accurate value estimates. In particular, REG and RIS(4) produce estimates with MSE more than 20 orders of magnitude below that of RIS(3) (Figure 3 is cut off at the bottom for clarity of the rest of the results). REG eventually passes the performance of RIS(4) since its knowledge of the transition probabilities allows it to eliminate sampling error in both the actions and the environment. In the low-to-medium data regime, only RIS(0) outperforms OIS. However, as data increases, the MSE of all RIS methods and REG decreases faster than that of OIS. The similar performance of $RIS(L-1)$ and REG supports the connection between these methods that we discuss in Section 4.2.

RIS with Linear Function Approximation Our next set of experiments consider continuous state and action spaces in the Linear Dynamical System domain. RIS represents $\pi_{\mathcal{D}}$ as a Gaussian policy with mean given as a linear function of the state features. Similar to in Gridworld, we compare three variants of IS, each implemented with RIS and OIS weights: the ordinary IS estimator, weighted IS (WIS), and per-decison IS (PDIS). Each method is averaged over 200 trials and results are shown in Figure 4(a).

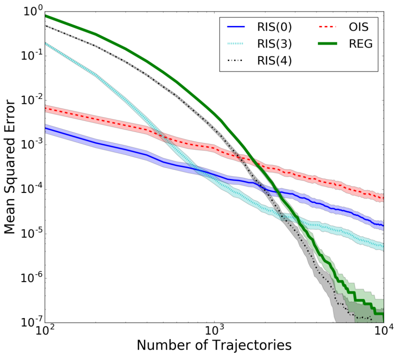
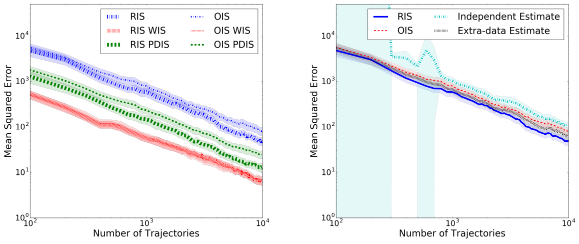


Figure 3: Off-policy evaluation in the SinglePath MDP for various n. The curves for REG and RIS(4) have been cut-off to more clearly show all methods. These methods converge to an MSE value of approximately 1×10^{-31}.



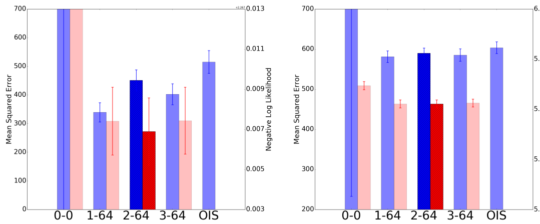
(a) LDS

(b) LDS Alt. Weights

Figure 4: Linear dynamical system results. Figure 4(a) shows the mean squared error (MSE) for three IS variants with and without RIS weights. Figure 4(b) shows the MSE for different methods of estimating the behavior policy compared to RIS and OIS. Axes and scaling are the same as in Figure 2(a).

We see that RIS weights improve both IS and PDIS, while both WIS variants have similar MSE. This result suggests that the MSE improvement from using RIS weights depends, at least partially, on the variant of IS being used.

Similar to Gridworld, we also consider estimating $\pi_{\mathcal{D}}$ with either an independent data-set or with extra data and see a similar ordering of methods. **Independent Estimate** gives high variance estimates for small sample sizes but then approaches OIS as the sample size grows. **Extra-Data Estimate** corrects for some sampling error and has lower MSE than OIS. RIS lowers MSE compared to all baselines.



(a) Hopper

(b) HalfCheetah

Figure 5: Figures 5(a) and 5(b) compare different neural network architectures (specified as #-layers-#-units) for regression importance sampling on the Hopper and HalfCheetah domain. The darker, blue bars give the MSE for each architecture and OIS. Lighter, red bars give the negative log likelihood of a hold-out data set. Our main point of comparison is the MSE of the architecture with the lowest hold-out negative log likelihood (given by the darker pair of bars) compared to the MSE of IS.

RIS with Neural Networks Our remaining experiments use the Hopper and HalfCheetah domains. RIS represents $\pi_\mathcal{D}$ as a neural network that maps the state to the mean of a Gaussian distribution over actions. The standard deviation of the Gaussian is given by state-independent parameters. In these experiments, we sample a single batch of 400 trajectories and compare the MSE of RIS and IS on this batch. We repeat this experiment 200 times for each method.

Figure 5 compares the MSE of RIS for different neural network architectures. Our main point of comparison is RIS using the architecture that achieves the lowest validation error during training (the darker bars in Figure 5). Under this comparison, the MSE of RIS with a two hidden layer network is lower than that of OIS in both Hopper and HalfCheetah though, in HalfCheetah, the difference is statistically insignificant. We also observe that the policy class with the best validation error does *not* always give the lowest MSE (e.g., in Hopper, the two hidden layer network gives the lowest validation loss but the network with a single layer of hidden units has $\approx 25\%$ less MSE than the two hidden layer network). This last observation motivates our final experiment.

RIS Model Selection Our final experiment aims to better understand how hold-out validation error relates to the MSE of the RIS estimator when using gradient descent to estimate neural network approximations of $\pi_\mathcal{D}$. This experiment duplicates our previous experiment, except every 25 steps of gradient descent we stop optimizing $\pi_\mathcal{D}$ and compute the RIS estimate with the current $\pi_\mathcal{D}$ and its MSE. We also compute the training and hold-out validation negative log-likelihood. Plotting these values gives a picture of how the MSE of RIS changes as our estimate of $\pi_\mathcal{D}$ changes. Figure 6 shows this plot for the Hopper domain.

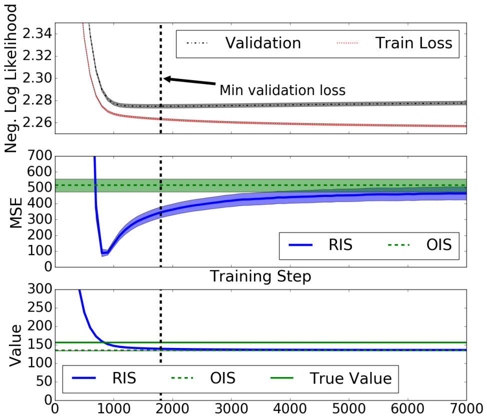


Figure 6: Mean squared error and estimate of the importance sampling estimator during training of $\pi_\mathcal{D}$. The x-axis is the number of gradient descent steps. The top plot shows the training and validation loss curves. The y-axis of the top plot is the average negative log-likelihood. The y-axis of the middle plot is mean squared error (MSE). The y-axis of the bottom plot is the value of the estimate. MSE is minimized close to, but slightly before, the point where the validation and training loss curves indicate that overfitting is beginning. This point corresponds to where the RIS estimate transitions from over-estimating to under-estimating.

We see that the policy with minimal MSE and the policy that minimizes validation loss are misaligned. If training is stopped when the validation loss is minimized, the MSE of RIS is lower than that of OIS (the intersection of the RIS curve and the vertical dashed line in Figure 6. However, the $\pi_\mathcal{D}$ that minimizes the validation loss curve is *not* identical to the $\pi_\mathcal{D}$ that minimizes MSE.

To understand this result, we also plot the average RIS estimate throughout behavior policy learning (bottom of Figure 6). We can see that at the beginning of training, RIS tends to *over-estimate* $v(\pi_e)$ because the probabilities given by $\pi_\mathcal{D}$ to the observed data will be small (and thus the RIS weights are large). As the likelihood of $\mathcal{D}$ under $\pi_\mathcal{D}$ increases (negative log likelihood decreases), the RIS weights become smaller and the estimates tend to *under-estimate* $v(\pi_e)$. The implication of these observations, for RIS, is that during behavior policy estimation the RIS estimate will likely have zero MSE at some point. Thus, there may be an early stopping criterion – besides minimal validation loss – that would lead to lower MSE with RIS, however, to date we have not found one. Note that OIS also tends to under-estimate policy value in MDPs as has been previously analyzed by Doroudi et al. (2017). Appendix F shows the same observations in the HalfCheetah domain.

6. Related Work

In this section we survey work related to behavior policy estimation for importance sampling. Methods related to RIS have been studied for Monte Carlo integration (Henmi et al., 2007; Delyon & Portier, 2016) and causal inference (Hirano et al., 2003; Rosenbaum, 1987). The REG method (discussed below) can be seen as the direct extension of these methods to MDPs. In contrast to these works, we study policy evaluation in Markov decision processes which introduces sequential structure into the samples and unknown stochasticity in the state transitions. These methods have also, to the best of our knowledge, *not* been studied in Markov decision processes or for sequential data.

Li et al. (2015) study the *regression* (REG) estimator for off-policy evaluation and show that its minimax MSE is asymptotically optimal though it might perform poorly for small sample sizes. Though REG and RIS are equivalent for multi-armed bandit problems, for MDPs, the definition of REG and any RIS method diverge. Figure 3 shows that all tested RIS methods improve over REG for small sample sizes though REG has lower asymptotic MSE. Intuitively, REG corrects for sampling error in both the action selection and state transitions through knowledge of the true state-transition function. However, such knowledge is usually unavailable and, in these cases, REG is inapplicable.

Narita et al. (2019) study behavior policy estimation for policy evaluation and improvement in multi-armed bandit problems. They also show lower asymptotic variance (as we do), however, their results are only for the bandit setting.

In the contextual bandit literature, Dudik et al. (2011) present finite sample bias and variance results for importance sampling that is applicable when the behavior policy probabilities are different than the true behavior policy. Farajtabar et al. (2018) extended these results to full MDPs. These works make the assumption that $\pi_{\mathcal{D}}$ is estimated independently from the data used in the final IS evaluation. In contrast, RIS uses the same set of data to both estimate π_b and compute the IS evaluation. This choice allows RIS to correct for sampling error and improve upon the OIS estimate (as shown in Figure 2(b), 2(d), and 4(b)).

A large body of work exists on lowering the variance of importance sampling for off-policy evaluation. Such approaches include control variates (Jiang & Li, 2016; Thomas & Brunskill, 2016), normalized importance weights (Precup et al., 2000; Swaminathan & Joachims, 2015), and importance ratio clipping (Bottou et al., 2013). These variance reduction strategies are complementary to regression importance sampling; any of these methods can be combined with RIS for further variance reduction.

7. Discussion and Future Work

Our experiments demonstrate that regression importance sampling can obtain lower mean squared error than ordinary importance sampling for off-policy evaluation in Markov decision process environments. The main practical conclusion of our paper is the importance of estimating $\pi_{\mathcal{D}}$ with the same data used to compute the importance sampling estimate. We also demonstrate that estimating a behavior policy that conditions on trajectory segments – instead of only the preceding state – improves performance in the large sample setting.

For all n, $\text{RIS}(n)$ is consistent and has lower asymptotic variance than OIS. There remain theoretical questions concerning the finite-sample setting and relaxing the assumption that we estimate $\pi_{\mathcal{D}}$ from a policy class that includes the true behavior policy. The connection to the REG estimator and our empirical results suggest that RIS with n close to L may suffer from high bias. Future work that quantifies or bounds this bias will give us a better understanding of RIS methods. Relaxing the assumption that $\pi_b \in \Pi$ or analyzing the case when $\pi_b \notin \Pi$ is also an important next step for bridging the gap between our presented theory and the use of RIS in settings where the policy class of π_b is unknown.

In this paper we focused on *batch* policy evaluation where $\mathcal{D}$ is given and fixed. Studying RIS for *online* policy evaluation setting is an interesting direction for future work. Finally, incorporating RIS into policy improvement methods is an interesting direction for future work. In work parallel to our own, two of the authors (Hanna & Stone, 2019) explored using an estimated behavior policy to lower sampling error in on-policy policy gradient learning. However, our approach in that paper only focuses on reducing variance in the one-step action selection while RIS could lower variance in the full return estimation.

8. Conclusion

We have studied a class of off-policy evaluation importance sampling methods, called regression importance sampling methods, that apply importance sampling after first estimating the behavior policy that generated the data. Notably, RIS estimates the behavior policy from the same set of data that is also used for the IS estimate. Computing the behavior policy estimate and IS estimate from the same set of data allows RIS to correct for the sampling error inherent to importance sampling with the true behavior policy. We evaluated RIS across several policy evaluation tasks and show that it improves over ordinary importance sampling – that uses the true behavior policy – in several off-policy policy evaluation tasks. Finally, we showed that, as the sample size grows, it can be beneficial to ignore knowledge that the true behavior policy is Markovian.

Acknowledgments

We would like to thank Garrett Warnell, Ishan Durugkar, Philip Thomas, Qiang Liu, Faraz Torabi, Leno Felipe da Silva, Marc Bellemare, Finale Doshi-Velez and the anonymous reviewers for insightful comments that suggested new directions to study and improved the final presentation of the work. This work has taken place in the Learning Agents Research Group (LARG) and the Personal Autonomous Robots Lab (PEARL) at the Artificial Intelligence Laboratory, The University of Texas at Austin. LARG research is supported in part by NSF (IIS-1637736, IIS-1651089, IIS-1724157), ONR (N00014-18-2243), FLI (RFP2-000), ARL, DARPA, Intel, Raytheon, and Lockheed Martin. PeARL research is supported in part by the NSF (IIS-1724157, IIS-1638107, IIS-1617639, IIS-1749204) and ONR(N00014-18-2243). Josiah Hanna is supported by an IBM PhD Fellowship. Peter Stone serves on the Board of Directors of Cogitai, Inc. The terms of this arrangement have been reviewed and approved by the University of Texas at Austin in accordance with its policy on objectivity in research.

References

Bottou, L., Peters, J., Quiñonero-Candela, J., Charles, D. X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P., and Snelson, E. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

Delyon, B. and Portier, F. Integral approximation by kernel smoothing. *Bernoulli*, 22(4):2177–2208, 2016.

Doroudi, S., Thomas, P. S., and Brunskill, E. Importance sampling for fair policy selection. In *Uncertainty in Artificial Intelligence (UAI)*, 2017.

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

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

Ghahramani, Z. and Rasmussen, C. E. Bayesian monte carlo. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 505–512, 2003.

Hanna, J. and Stone, P. Reducing sampling error in the monte carlo policy gradient estimator. In *Proceedings of the 19th International Conference on Autonomous Agents and Multi-agent Systems (AAMAS)*, 2019.

Hanna, J., Thomas, P. S., Stone, P., and Niekum, S. Data-efficient policy evaluation through behavior policy search. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Henmi, M., Yoshida, R., and Eguchi, S. Importance sampling via the estimated sampler. *Biometrika*, 94(4):985–991, 2007.

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

Jiang, N. and Li, L. Doubly robust off-policy evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016.

Li, L., Munos, R., and Szepesvári, C. Toward minimax off-policy value estimation. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015.

Narita, Y., Yasui, S., and Yata, K. Efficient counterfactual learning from bandit feedback. In *The 35th AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

O'Hagan, A. Monte carlo is fundamentally unsound. *The Statistician*, pp. 247–249, 1987.

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

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Rosenbaum, P. R. Model-based direct adjustment. *Journal of the American Statistical Association*, 82(398):387–394, 1987.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

Swaminathan, A. and Joachims, T. The self-normalized estimator for counterfactual learning. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 3231–3239, 2015.

Thomas, P. S. and Brunskill, E. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016.

A. Regression Importance Sampling is Consistent

In this appendix we show that the regression importance sampling (RIS) estimator is a consistent estimator of $v(\pi_e)$ under two assumptions. The main intuition for this proof is that RIS is performing policy search on an estimate of the log-likelihood, $\widehat{\mathcal{L}}(\pi|\mathcal{D})$, as a surrogate objective for the true log-likelihood, $\mathcal{L}(\pi)$. Since π_b has generated our data, π_b is the optimal solution to this policy search. As long as, for all π, $\widehat{\mathcal{L}}(\pi|\mathcal{D})$ is a consistent estimator of $\mathcal{L}(\pi)$ then selecting $\pi_\mathcal{D} = \operatorname*{argmax}_\pi \widehat{\mathcal{L}}(\pi|\mathcal{D})$ will converge probabilistically to π_b and the RIS estimator will be the same as the OIS estimator which is a consistent estimator of $v(\pi_e)$. If the set of policies we search over, Π, is countable then this argument is almost enough to show RIS to be consistent. The difficulty (as we explain below) arises when Π is *not* countable.

Our proof takes inspiration from Thomas and Brunskill who show that their Magical Policy Search algorithm converges to the optimal policy by maximizing a surrogate estimate of policy value (**?**). They show that performing policy search on a policy value estimate, $\hat{v}(\pi)$, will almost surely return the policy that maximizes $v(\pi)$ if $\hat{v}(\pi)$ is a consistent estimator of $v(\pi)$. The proof is almost identical; the notable difference is substituting the log-likelihood, $\mathcal{L}(\pi)$, and a consistent estimator of the log-likelihood, $\widehat{\mathcal{L}}(\pi|\mathcal{D})$, in place of $v(\pi)$ and $\hat{v}(\pi)$.

A.1. Definitions and Assumptions

Let $(\Omega, \mathcal{F}, \mu)$ be a probability space and $D_m : \Omega \rightarrow \mathcal{D}$ be a random variable. $D_m(\omega)$ is a sample of m trajectories with $\omega \in \Omega$. Let d_{π_b} be the distribution of states under π_b. Define the expected log-likelihood:

$$\mathcal{L}(\pi) = \mathbf{E}\left[\log \pi(A|S) | S \sim d_{\pi_b}, A \sim \pi_b\right]$$

and its sample estimate from samples in $D_m(\omega)$:

$$\widehat{\mathcal{L}}(\pi|D_m(\omega)) = \frac{1}{mL} \sum_{H \in D_m(\omega)} \sum_{t=0}^{L-1} \log \pi(A_t^H|S_t^H).$$

where S_t^H and A_t^H are the random variables representing the state and action that occur at time-step t of trajectory H.

Assuming for all s, a the variance of $\log \pi(a|s)$ is bounded, $\widehat{\mathcal{L}}(\pi|D_m(\omega))$ is a consistent estimator of $\mathcal{L}(\pi)$. We make this assumption explicit:

Assumption 1. *(Consistent Estimation of Log likelihood). For all $\pi \in \Pi$, $\widehat{\mathcal{L}}(\pi|D_m(\omega)) \xrightarrow{a.s.} \mathcal{L}(\pi)$.*

This assumption will hold when the support of π_b is a subset of the support of π for all $\pi \in \Pi$, i.e., no $\pi \in \Pi$ places zero probability measure on an action that π_b might take. We can ensure this assumption is satisfied by only considering $\pi \in \Pi$ that place non-zero probability on any action that π_b has taken.

We also make an additional assumption about the piece-wise continuity of the log-likelihood, $\mathcal{L}$, and the estimate of the log-likelihood, $\widehat{\mathcal{L}}$. First we present two necessary definitions as given by Thomas and Brunskill (**?**):

Definition 1. *(Piecewise Lipschitz continuity). We say that a function $f : M \rightarrow \mathbb{R}$ on a metric space (M, d) is piecewise Lipschitz continuous with respect to Lipschitz constant K and with respect to a countable partition, $\{M_1, M_2, ...\}$ if f is Lipschitz continuous with Lipschitz constant K on all metric spaces in $\{(M_i, d_i)\}_{i=1}^\infty$.*

Definition 2. *(δ-covering). If (M, d) is a metric space, a set $X \subset M$ is a δ-covering of (M, d) if and only if $\max_{y \in M} \min_{x \in X} d(x, y) \leq \delta$.*

We now present our final assumption:

Assumption 2. *(Piecewise Lipschitz objectives). Our policy class, Π, is equipped with a metric, d_Π, such that for all $D_m(\omega)$ there exist countable partition of Π, $\Pi^\mathcal{L} := \{\Pi_1^\mathcal{L}, \Pi_2^\mathcal{L}, ...\}$ and $\Pi^{\widehat{\mathcal{L}}} := \{\Pi_1^{\widehat{\mathcal{L}}}, \Pi_2^{\widehat{\mathcal{L}}}, ...\}$, where $\mathcal{L}$ and $\widehat{\mathcal{L}}(\cdot|D_m(\omega))$ are piecewise Lipschitz continuous with respect to $\Pi^\mathcal{L}$ and $\Pi^{\widehat{\mathcal{L}}}$ with Lipschitz constants K and $\widehat{K}$ respectively. Furthermore, for all $i \in \mathbb{N}_{>0}$ and all $\delta > 0$ there exist countable δ-covers of $\Pi_i^\mathcal{L}$ and $\Pi_i^{\widehat{\mathcal{L}}}$.*

As pointed out by Thomas and Brunskill, this assumption holds for the most commonly considered policy classes but is also general enough to hold for other settings (see Thomas and Brunskill (**?**) for further discussion of Assumptions 1 and 2 and the related definitions).

A.2. Consistency Proof

Note that:

$$\pi_b = \operatorname*{argmax}_{\pi \in \Pi} \mathcal{L}(\pi)$$

$$\pi_{\mathcal{D}} = \operatorname*{argmax}_{\pi \in \Pi} \widehat{\mathcal{L}}(\pi | D_m(\omega)).$$

Define the KL-divergence ($D_{\texttt{KL}}$)) between π_b and $\pi_{\mathcal{D}}$ in state s as: $\delta_{\texttt{KL}}(s) = D_{\texttt{KL}}(\pi_b(\cdot|s), \pi_{\mathcal{D}}(\cdot|s))$.

Lemma 1. *If Assumptions 1 and 2 hold then* $\mathbf{E}_{d_{\pi_b}}[\delta_{\texttt{KL}}(s)] \xrightarrow{a.s.} 0.$

Proof. Define $\Delta(\pi, \omega) = |\widehat{\mathcal{L}}(\pi | D_m(\omega)) - \mathcal{L}(\pi)|$. From Assumption 1 and one definition of almost sure convergence, for all $\pi \in \Pi$ and for all $\epsilon > 0$:

$$\Pr\left(\liminf_{m \to \infty} \{\omega \in \Omega : \Delta(\pi, \omega) < \epsilon\} \right) = 1. \tag{3}$$

Thomas and Brunskill point out that because Π may not be countable, (3) may not hold at the same time for all $\pi \in \Pi$. More precisely, it does *not* immediately follow that for all $\epsilon > 0$:

$$\Pr\left(\liminf_{m \to \infty} \{\omega \in \Omega : \forall \pi \in \Pi, \Delta(\pi, \omega) < \epsilon\} \right) = 1. \tag{4}$$

Let $C(\delta)$ denote the union of all of the policies in the δ-covers of the countable partitions of Π assumed to exist by Assumption 2. Since the partitions are countable and the δ-covers for each region are assumed to be countable, we have that $C(\delta)$ is countable for all δ. Thus, for all $\pi \in C(\delta)$, (3) holds simulatenously. More precisely, for all $\delta > 0$ and for all $\epsilon > 0$:

$$\Pr\left(\liminf_{m \to \infty} \{\omega \in \Omega : \forall \pi \in C(\delta), \Delta(\pi, \omega) < \epsilon\} \right) = 1. \tag{5}$$

Consider a $\pi \notin C(\delta)$. By the definition of a δ-cover and Assumption 2, we have that $\exists \pi' \in \Pi_i^{\mathcal{L}}, d(\pi, \pi') \leq \delta$. Since Assumption 2 requires $\mathcal{L}$ to be Lipschitz continuous on $\Pi_i^{\mathcal{L}}$, we have that $|\mathcal{L}(\pi) - \mathcal{L}(\pi')| \leq K\delta$. Similarly $|\widehat{\mathcal{L}}(\pi|D_m(\omega)) - \widehat{\mathcal{L}}(\pi'|D_m(\omega))| \leq \widehat{K}\delta$. So, $|\widehat{\mathcal{L}}(\pi|D_m(\omega)) - \mathcal{L}(\pi)| \leq |\widehat{\mathcal{L}}(\pi|D_m(\omega)) - \mathcal{L}(\pi')| + K\delta \leq |\widehat{\mathcal{L}}(\pi'|D_m(\omega)) - \mathcal{L}(\pi')| + (\widehat{K} + K)\delta$. Then it follows that for all $\delta > 0$:

$$\begin{gathered}(\forall \pi \in C(\delta), \Delta(\pi, \omega) \leq \epsilon) \to \\ \left(\forall \pi \in \Pi, \Delta(\pi, \omega) < \epsilon + (K + \widehat{K})\delta \right).\end{gathered}$$

Substituting this into (5) we have that for all $\delta > 0$ and for all $\epsilon > 0$:

$$\Pr\left(\liminf_{m \to \infty} \{\omega \in \Omega : \forall \pi \in \Pi, \Delta(\pi, \omega) < \epsilon + (K + \widehat{K})\delta\} \right) = 1$$

The next part of the proof massages (5) into a statement of the same form as (4). Consider the choice of $\delta \coloneqq \epsilon/(K + \widehat{K})$. Define $\epsilon' = 2\epsilon$. Then for all $\epsilon' > 0$:

$$\Pr\left(\liminf_{m \to \infty} \{\omega \in \Omega : \forall \pi \in \Pi, \Delta(\pi, \omega) < \epsilon'\} \right) = 1 \tag{6}$$

Since $\forall \pi \in \Pi, \Delta(\pi, \omega) < \epsilon'$, we obtain:

$$\Delta(\pi_b, \omega) < \epsilon' \tag{7}$$

$$\Delta(\pi_{\mathcal{D}}, \omega) < \epsilon' \tag{8}$$

and then applying the definition of Δ

$$\begin{aligned}
\mathcal{L}(\pi_{\mathcal{D}}) &\overset{(a)}{\leq} \mathcal{L}(\pi_b) && (9)\\
&\overset{(b)}{<} \widehat{\mathcal{L}}(\pi_b|D_m(\omega)) + \epsilon' && (10)\\
&\overset{(c)}{\leq} \widehat{\mathcal{L}}(\pi_{\mathcal{D}}|D_m(\omega)) + \epsilon' && (11)\\
&\overset{(d)}{\leq} \mathcal{L}(\pi_{\mathcal{D}}) + 2\epsilon' && (12)
\end{aligned}$$

where (a) comes from the fact that π_b maximizes $\mathcal{L}$, (b) comes from (7), (c) comes from the fact that $\pi_{\mathcal{D}}$ maximizes $\widehat{\mathcal{L}}(\cdot|D_m(\omega))$, and (d) comes from (8). Considering (9) and (12), it follows that $|\mathcal{L}(\pi_{\mathcal{D}}) - \mathcal{L}(\pi_b)| < 2\epsilon'$. Thus, (6) implies that:

$$\forall \epsilon' > 0, \Pr\left(\liminf_{m\to\infty}\{\omega \in \Omega : |\mathcal{L}(\pi_{\mathcal{D}}) - \mathcal{L}(\pi_b)| < 2\epsilon'\}\right) = 1$$

Using $\epsilon'' := 2\epsilon'$ we obtain:

$$\forall \epsilon'' > 0, \Pr\left(\liminf_{m\to\infty}\{\omega \in \Omega : |\mathcal{L}(\pi_{\mathcal{D}}) - \mathcal{L}(\pi_b)| < \epsilon''\}\right) = 1$$

From the definition of the KL-Divergence,

$$\mathcal{L}(\pi_{\mathcal{D}}) - \mathcal{L}(\pi_b) = \mathbf{E}_{d_{\pi_b}}[\delta_{\texttt{KL}}(s)]$$

and we obtain that:

$$\forall \epsilon > 0, \Pr\left(\liminf_{n\to\infty}\{\omega \in \Omega : |-\mathbf{E}_{d_{\pi_b}}[\delta_{\texttt{KL}}(s)]| < \epsilon\}\right) = 1$$

And finally, since the KL-Divergence is non-negative:

$$\forall \epsilon > 0, \Pr\left(\liminf_{m\to\infty}\{\omega \in \Omega : \mathbf{E}_{d_{\pi_b}}[\delta_{\texttt{KL}}(s)]| < \epsilon\}\right) = 1,$$

which, by the definition of almost sure convergence, means that $\mathbf{E}_{d_{\pi_b}}[\delta_{\texttt{KL}}(s)] \xrightarrow{a.s.} 0$. □

Proposition 1. *If Assumptions 1 and 2 hold, then* $\operatorname{RIS}(n)$ *is a consistent estimator of* $v(\pi_e)$*:* $\operatorname{RIS}(n)(\pi_e, \mathcal{D}) \xrightarrow{a.s.} v(\pi_e)$.

Proof. Lemma 1 shows that as the amount of data increases, the behavior policy estimated by RIS will almost surely converge to the true behavior policy. Almost sure convergence to the true behavior policy means that RIS almost surely converges to the ordinary OIS estimate. Since OIS is a consistent estimator of $v(\pi_e)$, RIS is also a consistent estimator of $v(\pi_e)$. □

B. Asymptotic Variance Proof

In this appendix we prove that, $\forall n$, $\operatorname{RIS}(n)$ has asymptotic variance at most that of OIS. We give this result as a corollary to Theorem 1 of Henmi et al. (2007) that holds for general Monte Carlo integration. Note that while we define distributions as probability mass functions, this result can be applied to continuous-valued state and action spaces by replacing probability mass functions with density functions.

Corollary 1. *Let* $\Pi^n_{\boldsymbol{\theta}}$ *be a class of twice differentiable policies,* $\pi_{\boldsymbol{\theta}}(\cdot|s_{t-n}, a_{t-n}, \dots, s_t)$*. If* $\exists \tilde{\boldsymbol{\theta}}$ *such that* $\pi_{\tilde{\boldsymbol{\theta}}} \in \Pi^n_{\boldsymbol{\theta}}$ *and* $\pi_{\tilde{\boldsymbol{\theta}}} = \pi_b$ *then*

$$\operatorname{Var}_A(\operatorname{RIS}(n)(\pi_e, \mathcal{D})) \leq \operatorname{Var}_A(\operatorname{IS}(\pi_e, \mathcal{D}, \pi_b))$$

where Var_A *denotes the asymptotic variance.*

Corollary 1 states that the asymptotic variance of RIS(n) must be at least as low as that of OIS.

We first present Theorem 1 from Henmi et al. (2007) and adopt their notation for its presentation. Consider estimating $v = \mathbf{E}_p\left[f(x)\right]$ for probability mass function p and real-valued function f. Given parameterized and twice differentiable probability mass function $q(\cdot|\tilde{\boldsymbol{\theta}})$, we define the ordinary importance sampling estimator of v as $\tilde{v} = \frac{1}{m}\sum_{i=1}^{m}\frac{p(x_i)}{q(x_i, \tilde{\boldsymbol{\theta}})}f(x_i)$. Similarly, define $\hat{v} = \frac{1}{m}\sum_{i=1}^{m}\frac{p(x_i)}{q(x_i, \hat{\boldsymbol{\theta}})}f(x_i)$ where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\tilde{\boldsymbol{\theta}}$ given the m samples from $q(\cdot|\tilde{\boldsymbol{\theta}})$. The following theorem relates the asymptotic variance of $\hat{v}$ to that of $\tilde{v}$.

Theorem 1.

$$\mathrm{Var}_A(\hat{v}) \leq \mathrm{Var}_A(\tilde{v})$$

where Var_A *denotes the asymptotic variance.*

Proof. See Theorem 1 of Henmi et al. (2007). □

Theorem 1 shows that the maximum likelihood estimated parameters of the sampling distribution yield an asymptotically lower variance estimate than using the true parameters, $\tilde{\boldsymbol{\theta}}$. To specialize this theorem to our setting, we show that the maximum likelihood behavior policy parameters are also the maximum likelihood parameters for the trajectory distribution of the behavior policy. First specify the class of sampling distribution: $\Pr(h; \boldsymbol{\theta}) = p(h)w_{\boldsymbol{\theta}}(h)$ where $p(h) = d_0(s_0)\prod_{t=1}^{L-1}P(s_t|s_{t-1}, a_{t-1})$ and $w_{\boldsymbol{\theta}}(h) = \prod_{t=0}^{L-1}\pi_{\boldsymbol{\theta}}(a_t|s_{t-n}, a_{t-n}, \ldots, s_t)$. We now present the following lemma:

Lemma 2.

$$\operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{h\in\mathcal{D}}\sum_{t=0}^{L-1}\log\pi_{\boldsymbol{\theta}}(a_t|s_{t-n}, a_{t-n}, \ldots, s_t) = \operatorname*{argmax}_{\boldsymbol{\theta}}\sum_{h\in\mathcal{D}}\log\Pr(h;\boldsymbol{\theta})$$

Proof.

$$\begin{aligned}
&\operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{h\in\mathcal{D}}\sum_{t=0}^{L-1}\log\pi_{\boldsymbol{\theta}}(a_t|s_{t-n}, a_{t-n}, \ldots, s_t)\\
&= \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{h\in\mathcal{D}}\sum_{t=0}^{L-1}\log\pi_{\boldsymbol{\theta}}(a_t|s_{t-n}, a_{t-n}, \ldots, s_t) + \underbrace{\log d(s_0) + \sum_{t=1}^{L-1}\log P(s_t|s_{t-1}, a_{t-1})}_{\text{const w.r.t. }\boldsymbol{\theta}}\\
&= \operatorname*{argmax}_{\boldsymbol{\theta}}\sum_{h\in\mathcal{D}}\log w_{\boldsymbol{\theta}}(h) + \log p(h)\\
\boldsymbol{\theta} &= \operatorname*{argmax}_{\boldsymbol{\theta}}\sum_{h\in\mathcal{D}}\log\Pr(h;\theta)
\end{aligned}$$

□

Finally, we combine Lemma 2 with Theorem 1 to prove Corollary 1:

Corollary 1. *Let* $\Pi_{\boldsymbol{\theta}}^n$ *be a class of policies,* $\pi_{\boldsymbol{\theta}}(\cdot|s_{t-n}, a_{t-n}, \ldots, s_t)$ *that are twice differentiable with respect to* $\boldsymbol{\theta}$*. If* $\exists\boldsymbol{\theta}\in\Pi_{\boldsymbol{\theta}}^n$ *such that* $\pi_{\boldsymbol{\theta}} = \pi_b$ *then*

$$\mathrm{Var}_A(\mathrm{RIS}(n)(\pi_e, \mathcal{D})) \leq \mathrm{Var}_A(\mathrm{IS}(\pi_e, \mathcal{D}, \pi_b))$$

where Var_A *denotes the asymptotic variance.*

Proof. Define $f(h) = g(h)$, $p(h) = \Pr(h|\pi_e)$ and $q(h|\boldsymbol{\theta}) = \Pr(h|\pi_{\boldsymbol{\theta}})$. Lemma 2 implies that:

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}\in\Pi_{\boldsymbol{\theta}}}\sum_{h\in\mathcal{D}}\sum_{t=0}^{L}\log\pi_{\boldsymbol{\theta}}(a_t|s_t)$$

is the maximum likelihood estimate of $\tilde{\boldsymbol{\theta}}$ (where $\pi_{\tilde{\boldsymbol{\theta}}} = \pi_b$ and $\Pr(h|\tilde{\boldsymbol{\theta}})$ is the probability of h under π_b) and then Corollary 1 follows directly from Theorem 1. □

Note that for RIS(n) with $n > 0$, the condition that $\pi_{\tilde{\boldsymbol{\theta}}} \in \Pi^n$ can hold even if the distribution of $A_t \sim \pi_{\tilde{\boldsymbol{\theta}}}$ (i.e., $A_t \sim \pi_b$) is only conditioned on s_t. This condition holds when $\exists \pi_{\boldsymbol{\theta}} \in \Pi^n$ such that $\forall s_{t-n}, a_{t-n}, \dots a_{t-1}$:

$$\pi_{\tilde{\boldsymbol{\theta}}}(a_t|s_t) = \pi_{\boldsymbol{\theta}}(a_t|s_{t-n}, a_{t-n}, \dots, s_t),$$

i.e., the action probabilities only vary with respect to s_t.

C. Connection to the REG estimator

In this appendix we show that $\text{RIS}(L-1)$ is an approximation of the REG estimator studied by Li et al. (2015). This connection is notable because Li et al. showed REG is asymptotically minimax optimal, however, in MDPs, REG requires knowledge of the environment's transition and initial state distribution probabilities while $RIS(L-1)$ does not. For this discussion, we recall the definition of the probability of a trajectory for a given MDP and policy:

$$\Pr(h|\pi) = d_0(s_0)\pi(a_0|s_0)P(s_1|s_0,a_0)\cdots P(s_{L-1}|s_{L-2},a_{L-2})\pi(a_{L-1}|s_{L-1}).$$

We also define $\mathcal{H}$ to be the set of all state-action trajectories possible under π_b of length L: $s_0, a_0, ...s_{L-1}, a_{L-1}$.

Li et al. introduce the regression estimator (REG) for multi-armed bandit problems (2015). This method estimates the mean reward for each action as $\hat{r}(a, \mathcal{D})$ and then computes the REG estimate as:

$$\texttt{REG}(\pi_e, \mathcal{D}) = \sum_{a \in \mathcal{A}} \pi_e(a)\hat{r}(a, \mathcal{D}).$$

This estimator is identical to RIS(0) in multi-armed bandit problems (Li et al., 2015). The extension of REG to finite horizon MDPs estimates the mean return for each trajectory as $\hat{g}(h, \mathcal{D})$ and then computes the estimate:

$$\texttt{REG}(\pi_e, \mathcal{D}) = \sum_{h \in \mathcal{H}} \Pr(h|\pi_e)\hat{g}(h, \mathcal{D}).$$

Since this estimate uses $\Pr(h|\pi_e)$ it requires knowledge of the initial state distribution, d_0, and transition probabilities, P.

We now elucidate a relationship between $\text{RIS}(L-1)$ and REG even though they are different estimators. Let $c(h)$ denote the number of times that trajectory h appears in $\mathcal{D}$. We can rewrite REG as an importance sampling method with a count-based estimate of the probability of a trajectory in the denominator:

$$\begin{aligned}
\texttt{REG}(\pi_e, \mathcal{D}) =& \sum_{h \in \mathcal{H}} \Pr(h|\pi_e)\hat{g}(h, \mathcal{D}) && (13)\\
=& \frac{1}{m}\sum_{h \in \mathcal{H}} c(h)\frac{\Pr(h|\pi_e)}{c(h)/m}\hat{g}(h, \mathcal{D}) && (14)\\
=& \frac{1}{m}\sum_{i=1}^{m} \frac{\Pr(h_i|\pi_e)}{c(h_i)/m} g(h_i) && (15)
\end{aligned}$$

The denominator in (15) can be re-written as a telescoping product to obtain an estimator that is similar to RIS($L-1$):

$$\begin{aligned}
\texttt{REG}(\pi_e, \mathcal{D}) =& \frac{1}{m}\sum_{i=1}^{m} \frac{\Pr(h_i|\pi_e)}{c(h_i)/m} g(h_i)\\
=& \frac{1}{m}\sum_{i=1}^{m} \frac{\Pr(h_i|\pi_e)}{\frac{c(s_0)}{m}\frac{c(s_0,a_0)}{c(s_0)}\cdots\frac{c(h_i)}{c(h_i/a_{L-1})}} g(h_i)\\
=& \frac{1}{m}\sum_{i=1}^{m} \frac{d_0(s_0)\pi_e(a_0|s_0)P(s_1|s_0,a_0)\cdots P(s_{L-1}|s_{L-2},a_{L-2})\pi_e(a_{L-1}|s_{L-1})}{\hat{d}(s_0)\pi_{\mathcal{D}}(a_0|s_0)\hat{P}(s_1|s_0,a_0)\cdots\hat{P}(s_{L-1}|h_{0:L-1})\pi_{\mathcal{D}}(a_{L-1}|h_{i:j})} g(h_i).
\end{aligned}$$

This expression differs from RIS($L-1$) in two ways:

1. The numerator includes the initial state distribution and transition probabilities of the environment.

2. The denominator includes count-based estimates of the initial state distribution and transition probabilities of the environment where the transition probabilities are conditioned on all past states and actions.

If we assume that the empirical estimates of the environment probabilities in the denominator are equal to the true environment probabilities then these factors cancel and we obtain the RIS($L-1$) estimate. This assumption will almost always be false except in deterministic environments. However, showing that RIS($L-1$) is approximating REG suggests that RIS($L-1$) may have similar theoretical properties to those elucidated for REG by Li et al. (2015). Our SinglePath experiment (See Figure 2 in the main text) supports this conjecture: RIS($L-1$) has high bias in the low to medium sample size but have asymptotically lower MSE compared to other methods. REG has even higher bias in the low to medium sample size range but has asymptotically lower MSE compared to RIS($L-1$). RIS with smaller n appear to decrease the initial bias but have larger MSE as the sample size grows. The asymptotic benefit of RIS for all n is also corroborated by Corollary 1 in Appendix B though Corollary 1 does *not* tell us anything about how different RIS methods compare asymptotically. The asymptotic benefit of REG compared to RIS methods can be understood as REG correcting for sampling error in both the action selection and state transitions.

D. Sampling Error with Continuous Actions

In Section 3 of the main text we discussed how ordinary importance sampling can suffer from sampling error. Then, in Section 4, we presented an example showing how RIS corrects for sampling error in $\mathcal{D}$ in deterministic and finite MDPs. Most of this discussion assumed that the state and action spaces of the MDP were finite. Here, we discuss sampling error in continuous action spaces. The primary purpose of this discussion is intuition and we limit discussion to a setting that can be easily visualized. We consider a deterministic MDP with scalar, real-valued actions, reward $R : \mathcal{A} \rightarrow \mathbb{R}$, and $L = 1$.

We assume the support of π_b and π_e is bounded and for simplicity assume the support to be $[0, 1]$. Policy evaluation is equivalent to estimating the integral:

$$v(\pi_e) = \int_0^1 R(a)\pi_e(a)da \tag{16}$$

and the ordinary importance sampling estimate of this quantity with m samples from π_b is:

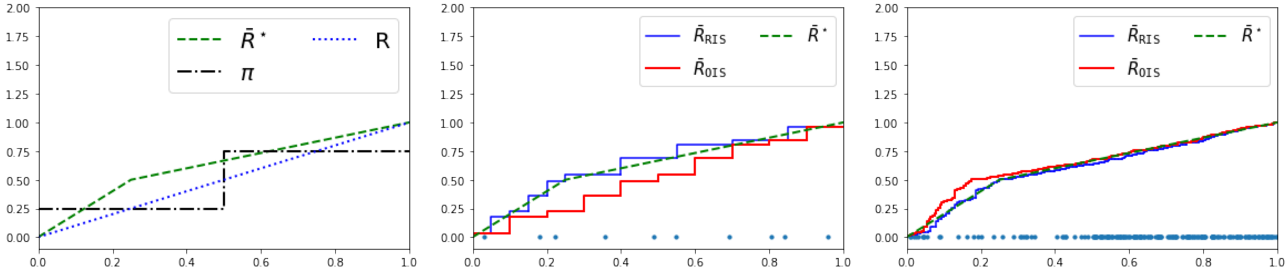
$$\frac{1}{m}\sum_{i=1}^{m}\frac{\pi_e(a_i)}{\pi_b(a_i)}R(a_i). \tag{17}$$

Even though the OIS estimate is a sum over a finite number of samples, we show it is exactly equal to an integral over a particular piece-wise function. We assume (w.l.o.g) that the a_i's are in non-decreasing order, ($a_0 <= a_i <= a_m$). Imagine that we place the $R(a_i)$ values uniformly across the interval $[0, 1]$ so that they divide the range $[0, 1]$ into m equal bins. In other words, we maintain the relative ordering of the action samples but ignore the spatial relationship between samples. We now define piece-wise constant function $\bar{R}_{\text{OIS}}$ where $\bar{R}_{\text{OIS}}(a) = R(a_i)$ if a is in the i^{th} bin. The ordinary importance sampling estimate is exactly equal to the integral $\int_0^1 \bar{R}_{\text{OIS}}(a)da$.

It would be reasonable to assume that $\bar{R}_{\text{OIS}}(a)$ is approximating $R(a)\pi_e(a)$ since the ordinary importance sampling estimate (17) is approximating (16), i.e., $\lim_{m\rightarrow\infty} \bar{R}_{\text{OIS}}(a) = R(a)\pi_e(a)$. In reality, $\bar{R}_{\text{OIS}}$ approaches a *stretched* version of R where areas with high density under π_e are stretched and areas with low density are contracted. We call this stretched version of R, $\bar{R}^{\star}$. The integral of $\int_0^1 \bar{R}^{\star}(a)da$ is $v(\pi_e)$.

Figure 7(a) gives a visualization of an example $\bar{R}^{\star}$ using on-policy Monte Carlo sampling from an example π_e and linear R. In contrast to the true $\bar{R}^{\star}$, the OIS approximation to $\bar{R}$, $\bar{R}_{\text{OIS}}$ stretches ranges of R according to the number of samples in that range: ranges with many samples are stretched and ranges without many samples are contracted. As the sample size grows, any range of R will be stretched in proportion to the probability of getting a sample in that range. For example, if the probability of drawing a sample from $[a, b]$ is 0.5 then $\bar{R}^{\star}$ stretches R on $[a, b]$ to cover half the range $[0, 1]$. Figure 7 visualizes $\bar{R}_{\text{OIS}}$ the OIS approximation to $\bar{R}^{\star}$ for sample sizes of 10 and 200.

In this analysis, sampling error corresponds to over-stretching or under-stretching R in any given range. The limitation of ordinary importance sampling can then be expressed as follows: given π_e, we know the correct amount of stretching for any range and yet OIS ignores this information and stretches based on the empirical proportion of samples in a particular range. On the other hand, RIS first divides by the empirical pdf (approximately undoing the stretching from sampling) and then



(a) Policy and Reward

(b) 10 Sample Approximation

(c) 200 Sample Approximation

Figure 7: Policy evaluation in a continuous armed bandit task. Figure 7(a) shows a reward function, R, and the pdf of a policy, π, with support on the range $[0, 1]$. With probability 0.25, π selects an action less than 0.5 with uniform probability; otherwise π selects an action greater than 0.5. The reward is equal to the action chosen. All figures show $\bar{R}^\star$: a version of R that is stretched according to the density of π; since the range $[0.5, 1]$ has probability 0.75, R on this interval is stretched over $[0.25, 1]$. Figure 7(b) and 7(c) show $\bar{R}^\star$ and the piece-wise $\bar{R}_{\text{OIS}}$ and $\bar{R}_{\text{RIS}}$ approximations to $\bar{R}^\star$ after 10 and 200 samples respectively.

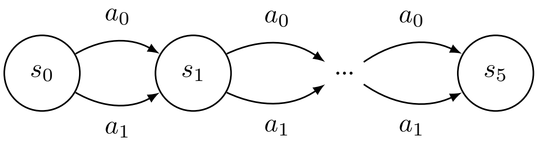


Figure 8: The SinglePath MDP referenced in Section 4 of the main text. **Not shown:** If the agent takes action a_1 it remains in its current state with probability 0.5.

multiplies by the true pdf to stretch R a more accurate amount. Figure 7 also visualizes the $\bar{R}_{\text{RIS}}$ approximation to $\bar{R}^\star$ for sample sizes of 10 and 200. In this figure, we can see that $\bar{R}_{\text{RIS}}$ is a closer approximation to $\bar{R}^\star$ than $\bar{R}_{\text{OIS}}$ for both sample sizes. In both instances, the mean squared error of the RIS estimate is less than that of the OIS estimate.

Since R may be unknown until sampled, we will still have non-zero MSE. However the standard OIS estimate has error due to *both* sampling error and unknown R values.

E. Extended Empirical Description

In this appendix we provide additional details for our experimental domains. Code is provided at `https://github.com/LARG/regression-importance-sampling`.

SinglePath: This environment is shown in Figure **??** with horizon $L = 5$. In each state, π_b selects action, a_0, with probability $p = 0.6$ and π_e selects action, a_0, with probability $1 - p = 0.4$. Action a_0 causes a deterministic transition to the next state. Action a_1 causes a transition to the next state with probability 0.5, otherwise, the agent remains in its current state. The agent receives a reward of 1 for action a_0 and 0 otherwise. RIS uses count-based estimation of π_b and REG uses count-based estimation of trajectories. REG is also given the environment's transition matrix, P.

Gridworld: This domain is a 4×4 Gridworld with a terminal state with reward 100 at $(3, 3)$, a state with reward -10 at $(1, 1)$, a state with reward 1 at $(1, 3)$, and all other states having reward -1. The domain has been used in prior off-policy policy evaluation work (**?**Thomas & Brunskill, 2016; Hanna et al., 2017; Farajtabar et al., 2018). The action set contains the four cardinal directions and actions move the agent in its intended direction (except when moving into a wall which produces no movement). The agent begins in $(0, 0)$, $\gamma = 1$, and $L = 100$. All policies use a softmax action selection distribution with temperature 1 and a separate parameter, θ_{sa}, for each state, s, and action a. The probability of taking action a in state s is given by:

$$\pi(a|s) = \frac{e^{\theta_{sa}}}{\sum_{a' \in \mathcal{A}} e^{\theta_{sa'}}}$$

The first set of experiments uses a behavior policy, π_b, that can reach the high reward terminal state and an evaluation policy, π_e, that is the same policy with lower entropy action selection. The second set of experiments uses the same behavior policy as both behavior and evaluation policy. RIS estimates the behavior policy with the empirical frequency of actions in each state. This domain allows us to study RIS separately from questions of function approximation.

Linear Dynamical System This domain is a point-mass agent moving towards a goal in a two dimensional world by setting x and y acceleration. The state-space is the agent's x and y position and velocity. The agent acts for $L = 20$ time-steps under linear-gaussian dynamics and receives a reward that is proportional to its distance from the goal. Specifically, if $\mathbf{s}_t$ is the agent's state vector and it takes action $\mathbf{a}_t$, then the resulting next state is:

$$\mathbf{s}_{t+1} = A \cdot \mathbf{s}_t + B \cdot \mathbf{a}_t + \epsilon_t$$

where $\epsilon_t \sim \mathcal{N}(0, I)$, A is the identity matrix, and

$$B = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The agent's policy is a linear map from state features to the mean of a Gaussian distribution over actions. For the state features, we use second order polynomial basis functions so that policies are non-linear in the state features but we can still estimate $\pi_{\mathcal{D}}$ efficiently with ordinary least squares. We obtain a basic policy by optimizing the linear weights of this policy for 10 iterations of the Cross-Entropy method (**?**). The evaluation policy uses a standard deviation of 0.5 and the true π_b uses a standard deviation of 0.6.

Continuous Control We also use two continuous control tasks from the OpenAI gym: Hopper and HalfCheetah.[6] The state and action dimensions of each task are shown in Table 1. In each task, we use neural network policies with 2 layers of 64 hidden units each for π_e and π_b. Each policy maps the state to the mean of a Gaussian distribution with state-independent standard deviation. We obtain π_e and π_b by running the OpenAI Baselines (**?**) version of proximal policy optimization (PPO) (**?**) and then selecting two policies along the learning curve. For both environments, we use the policy after 30 updates for π_e and after 20 updates for π_b. These policies use $\tanh$ activations on their hidden units since these are the default in the OpenAI Baselines PPO implementation.

Environment	State Dimension	Action Dimension
Hopper	15	3
Half Cheetah	26	6

Table 1: State and action dimension for each OpenAI Roboschool environment.

RIS estimates the behavior policy with gradient descent on the negative log-likelihood of the neural network. Specifically, we interpret the neural network outputs, $\mu(s)$, as the mean of a multi-variate Gaussian distribution with diagonal covariance matrix. We use a state-independent parameter vector, σ, to represent the log-standard deviation of the Gaussian distribution. Given m, state-action pairs, RIS uses the loss function:

$$\mathcal{L} = \sum_{i=1}^{m} 0.5((a_i - \mu(s_i))/e^{\sigma})^2 + \sigma$$

Minimizing $\mathcal{L}$ is equivalent to minimizing a squared-error loss function with regards to estimating μ.

In our experiments we use a learning rate of 1×10^{-3} and L2-regularization with a weight of 0.02. The multi-layer behavior policies learned by RIS use relu activations. The specific architectures considered for $\pi_{\mathcal{D}}$ have either 0, 1, 2, or 3 hidden layers with 64 units in each hidden layer.

In these domains we only consider a batch size of 400 trajectories for estimating $\pi_{\mathcal{D}}$ and computing the policy value estimate. For determining early stopping and measuring validation error we use a separate batch of 80 trajectories (20% of the policy evaluation data).

[6] For these tasks we use the Roboschool versions: https://github.com/openai/roboschool

F. Extended Empirical Results

This appendix includes two additional plots that space constraints limited from the main text.

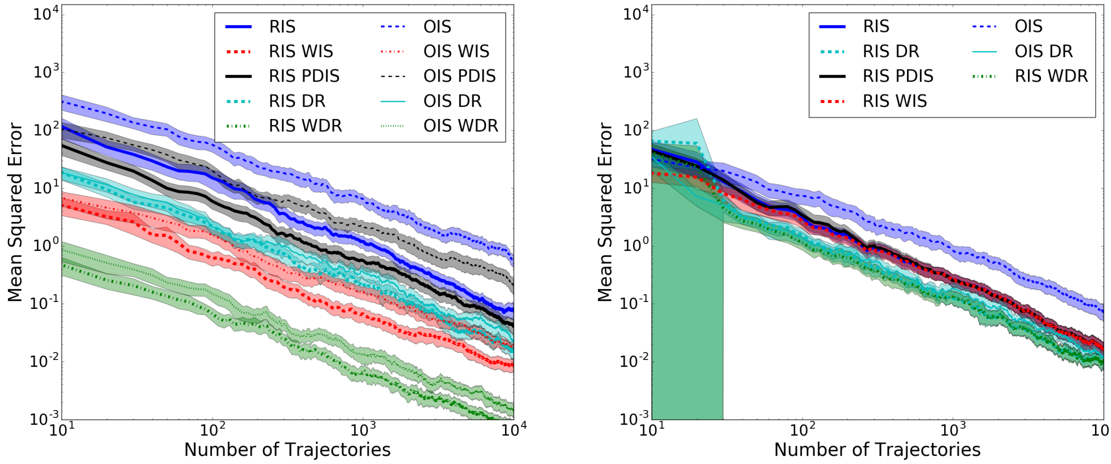
F.1. Importance Sampling Variants

This appendix presents additional importance sampling methods that are implemented with both OIS weights and RIS weights. Specifically, we implement the following:

- The ordinary importance sampling estimator described in Section 2.
- The weighted importance sampling estimator (WIS) (Precup et al., 2000) that normalizes the importance weights with their sum.
- Per-decision importance sampling (PDIS) (Precup et al., 2000) that importance samples the individual rewards.
- The doubly-robust (DR) estimator (Jiang & Li, 2016; Thomas & Brunskill, 2016) that uses a model of P and r to lower the variance of PDIS.
- The weighted doubly robust (WDR) estimator (Thomas & Brunskill, 2016) that uses weighted importance sampling to lower the variance of the doubly robust estimator.

Since DR and WDR require a model of the environment, we estimate a count-based model with half of the available data in $\mathcal{D}$.

Figure 9(a) gives results for all 5 of these IS variants implemented with both RIS weights and OIS weights. Figure 9(b) gives the same results except for the on-policy setting. Note that in the on-policy setting, PDIS and WIS are identical to IS and WDR is identical to DR when implemented with OIS weights. Thus we only present the RIS versions of these methods. In addition to the results for ordinary IS, WIS, and WDR that are also in the main text, Figure 9 shows RIS weights improve DR and PDIS.



(a) Gridworld Off-Policy

(b) Gridworld On-Policy

Figure 9: Policy evaluation results for Gridworld. In all subfigures, the x-axis is the number of trajectories collected and the y-axis is mean squared error. Axes are log-scaled. The shaded region gives a 95% confidence interval. The main point of comparison is the RIS variant of each method to the OIS variant of each method, e.g., RIS WIS compared to OIS WIS. Results are averaged over 100 trials.

F.2. Gradient Descent Policy Estimation

This appendix shows how the MSE of RIS changes during estimation of $\pi_{\mathcal{D}}$ in the HalfCheetah domain. Figure 10 gives the results. As in the Hopper domain, we see that the minimal validation loss policy and the minimal MSE policy are misaligned. The RIS estimate initially over-estimates the policy value and then begins under-estimating. Further discussion of these observations are given in Section 6 of the main text.

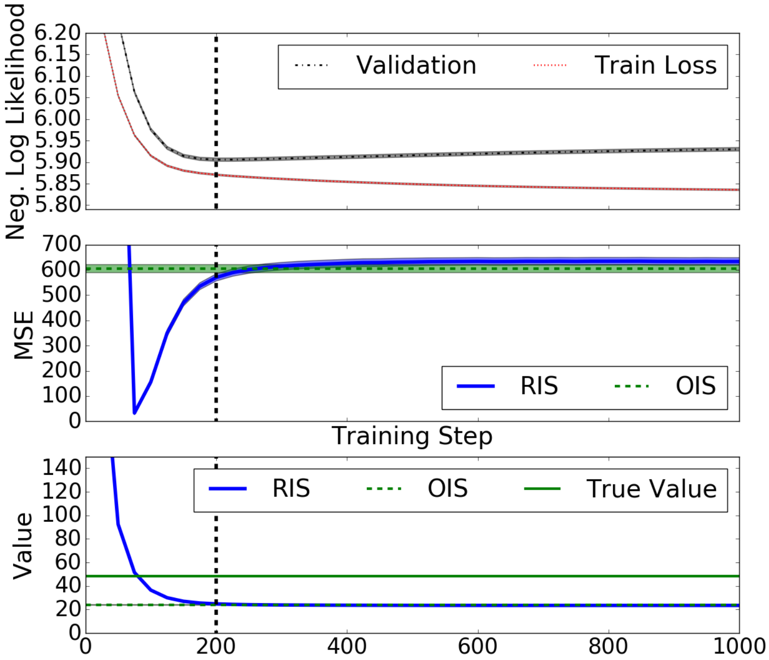


Figure 10: Mean squared error and estimate of the importance sampling estimator during training of $\pi_{\mathcal{D}}$. The x-axis is the number of gradient ascent steps. The top plot shows the training and validation loss curves. The y-axis of the top plot is the average negative log-likelihood. The y-axis of the middle plot is mean squared error (MSE). The y-axis of the bottom plot is the value of the estimate. MSE is minimized close to, but slightly before, the point where the validation and training loss curves indicate that overfitting is beginning. This point corresponds to where the RIS estimate transitions from over-estimating to under-estimating the policy value. Results are averaged over 200 trials and the shaded region represents a 95% confidence interval around the mean result.