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.

[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.

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 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 $H := (S_0, A_0, R_0, S_1, \ldots, S_{L-1}, A_{L-1}, R_{L-1})$ 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 $\text{RIS}(n)$ 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 $\text{RIS}(n)$ 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 $\text{RIS}(n)$ weight for action A_t, state S_t, and trajectory segment $H_{t-n:t}$. Note that the $\text{RIS}(n)$ 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 $\mathcal{D}$. $\text{RIS}(n)$ 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 $\text{RIS}(n)$ $\pi := \pi_\mathcal{D}{}^{(n)}$ 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 $\text{RIS}(L-1)$ 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

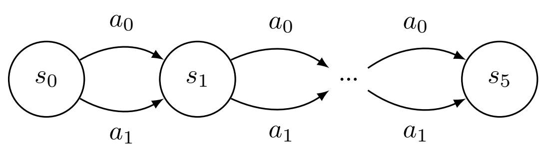


Figure 1: The SinglePath MDP. This environment has 5 states, 2 actions, and $L = 5$. The agent begins in state 0 and both actions either take the agent from state n to state $n + 1$ or cause the agent to remain in state n. **Not shown:** If the agent takes action a_1 it remains in its current state with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

[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`

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:

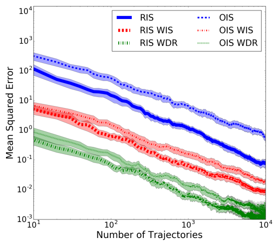
1. **Independent Estimate**: In addition to $\mathcal{D}$, this method has access to an additional set, $\mathcal{D}_{\texttt{train}}$. The behavior policy is estimated with $\mathcal{D}_{\texttt{train}}$ and the policy value estimate is computed with $\mathcal{D}$. Since (s, a) pairs in $\mathcal{D}$ 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.

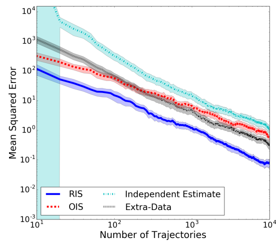
We also repeat these experiments for the on-policy setting and present results in Figure 2(c) and Figure 2(d). We observe similar trends as in the off-policy experiments suggesting that RIS can lower variance in Monte Carlo sampling methods even when OIS weights are otherwise unnecessary.

RIS(n) In the Gridworld domain it is difficult to observe the performance of $\text{RIS}(n)$ 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

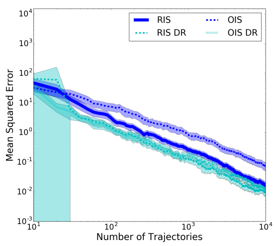
[5]We also implemented and evaluated *per-decision* importance sampling and the ordinary *doubly robust* estimator and saw similar results. However we defer these results to Appendix F for clarity.



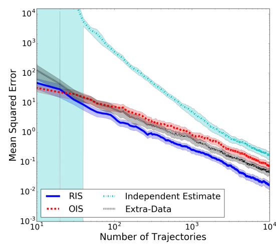
(a) Gridworld



(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 $\text{RIS}(4)$ produce estimates with MSE more than 20 orders of magnitude below that of $\text{RIS}(3)$ (Figure 3 is cut off at the bottom for clarity of the rest of the results). REG eventually passes the performance of $\text{RIS}(4)$ 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 $\text{RIS}(0)$ 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

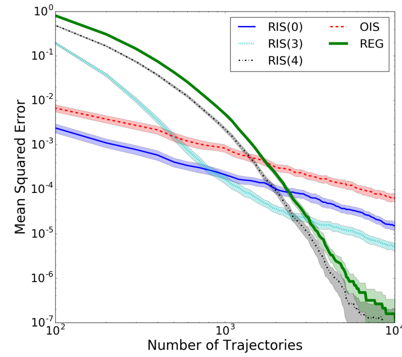
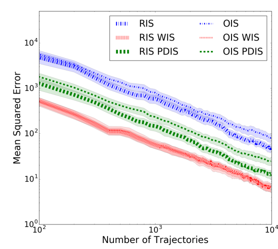
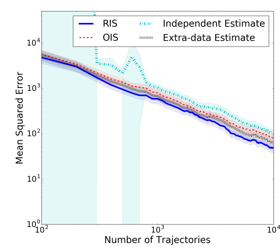


Figure 3: Off-policy evaluation in the SinglePath MDP for various n. The curves for REG and $\text{RIS}(4)$ have been cut-off to more clearly show all methods. These methods converge to an MSE value of approximately 1×10^{-31}.



(a) LDS



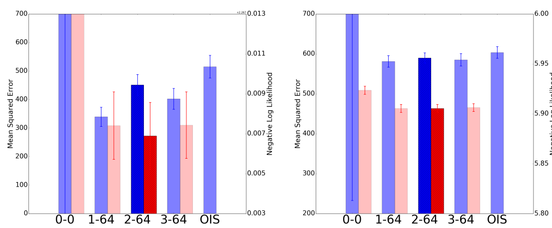
(b) LDS Alt. Weights

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).

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).

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.

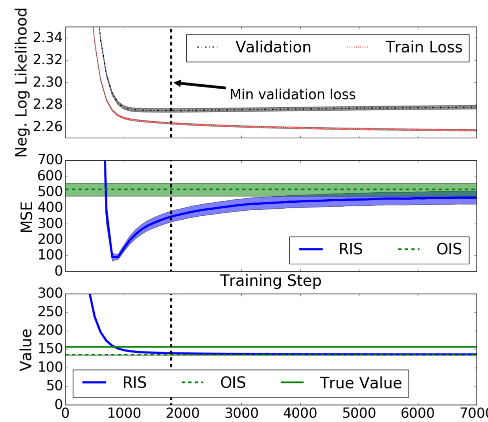


Figure 6: Mean squared error and estimate of the importance sampling estimator during training of $\pi_\mathcal{D}$. The x-axis is the number of gradient descent steps. The top plot shows the training and validation loss curves. The y-axis of the top plot is the average negative log-likelihood. The y-axis 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.