Importance Sampling Policy Evaluation with an Estimated Behavior Policy

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

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 (2016b). 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 $D_m : \Omega \to \mathcal{D}$ 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 (2016b):

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

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

Definition 1. *(Piecewise Lipschitz continuity). We say that a function $f : M \to \mathbb{R}$ 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 (2016b) 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_{KL})) between π_b and $\pi_{\mathcal{D}}$ in state s as: $\delta_{\text{KL}}(s) = D_{\text{KL}}(\pi_b(\cdot|s), \pi_{\mathcal{D}}(\cdot|s))$.

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

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 $\epsilon > 0$:

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

Thomas and Brunskill point out that because Π may not be countable, (1) may not hold at the same time for all $\pi \in \Pi$. More precisely, it does *not* immediately follow that for all $\epsilon > 0$:

$$\Pr\left(\liminf_{m \to \infty} \{\omega \in \Omega : \forall \pi \in \Pi, \Delta(\pi, \omega) < \epsilon\}\right) = 1. \tag{2}$$

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 $\pi \in C(\delta)$, (1) holds simulatenously. More precisely, for all $\delta > 0$ and for all $\epsilon > 0$:

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

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 $\delta > 0$:

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

Substituting this into (3) 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 (3) into a statement of the same form as (2). Consider the choice of $\delta \coloneqq \epsilon/(K + \widehat{K})$. Define $\epsilon' = 2\epsilon$. Then for all $\epsilon' > 0$:

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

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

$$\Delta(\pi_b,\omega) < \epsilon' \tag{5}$$

$$\Delta(\pi_{\mathcal{D}},\omega) < \epsilon' \tag{6}$$

and then applying the definition of Δ

$$\begin{aligned}
\mathcal{L}(\pi_{\mathcal{D}}) &\overset{(a)}{\leq} \mathcal{L}(\pi_b) && (7)\\
&\overset{(b)}{<} \widehat{\mathcal{L}}(\pi_b|D_m(\omega)) + \epsilon' && (8)\\
&\overset{(c)}{\leq} \widehat{\mathcal{L}}(\pi_{\mathcal{D}}|D_m(\omega)) + \epsilon' && (9)\\
&\overset{(d)}{\leq} \mathcal{L}(\pi_{\mathcal{D}}) + 2\epsilon' && (10)
\end{aligned}$$

where (a) comes from the fact that π_b maximizes $\mathcal{L}$, (b) comes from (5), (c) comes from the fact that $\pi_{\mathcal{D}}$ maximizes $\widehat{\mathcal{L}}(\cdot|D_m(\omega))$, and (d) comes from (6). Considering (7) and (10), it follows that $|\mathcal{L}(\pi_{\mathcal{D}}) - \mathcal{L}(\pi_b)| < 2\epsilon'$. Thus, (4) 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 $\epsilon'' \coloneqq 2\epsilon'$ 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* $\text{RIS}(n)$ *is a consistent estimator of* $v(\pi_e)$*:* $\text{RIS}(n)(\pi_e, \mathcal{D}) \xrightarrow{a.s.} v(\pi_e)$.

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, $\forall n$, $\text{RIS}(n)$ 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,* $\pi_{\boldsymbol{\theta}}(\cdot|s_{t-n}, a_{t-n}, \ldots, s_t)$*. If* $\exists \tilde{\boldsymbol{\theta}}$ *such that* $\pi_{\tilde{\boldsymbol{\theta}}} \in \Pi^n_{\boldsymbol{\theta}}$ *and* $\pi_{\tilde{\boldsymbol{\theta}}} = \pi_b$ *then*

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

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.

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

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.

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

Proof.

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

□

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

Corollary 1. *Let $\Pi^n_{\boldsymbol{\theta}}$ be a class of policies, $\pi_{\boldsymbol{\theta}}(\cdot|s_{t-n}, a_{t-n}, \dots, s_t)$ that are twice differentiable with respect to $\boldsymbol{\theta}$. If $\exists \boldsymbol{\theta} \in \Pi^n_{\boldsymbol{\theta}}$ 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 $\mathrm{RIS}(L-1)$ 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 $\mathrm{RIS}(L-1)$ 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:

$$\begin{aligned}
\texttt{REG}(\pi_e, \mathcal{D}) &= \sum_{h \in \mathcal{H}} \Pr(h|\pi_e)\hat{g}(h, \mathcal{D}) && (11)\\
&= \frac{1}{m} \sum_{h \in \mathcal{H}} c(h) \frac{\Pr(h|\pi_e)}{c(h)/m} \hat{g}(h, \mathcal{D}) && (12)\\
&= \frac{1}{m} \sum_{i=1}^{m} \frac{\Pr(h_i|\pi_e)}{c(h_i)/m} g(h_i) && (13)
\end{aligned}$$

The denominator in (13) 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:

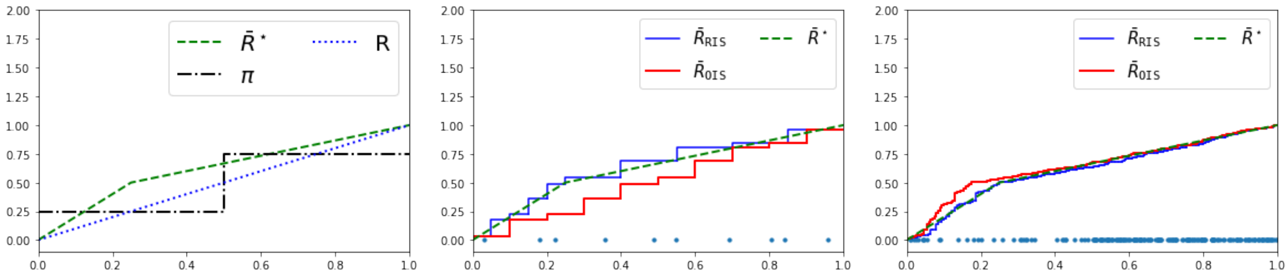
$$v(\pi_e) = \int_0^1 R(a)\pi_e(a)da \tag{14}$$

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

$$\frac{1}{m}\sum_{i=1}^{m} \frac{\pi_e(a_i)}{\pi_b(a_i)} R(a_i). \tag{15}$$

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 (15) is approximating (14), 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



(a) Policy and Reward

(b) 10 Sample Approximation

(c) 200 Sample Approximation

Figure 1. Policy evaluation in a continuous armed bandit task. Figure 1(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 1(b) and 1(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.

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 1(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 1 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 multiplies by the true pdf to stretch R a more accurate amount. Figure 1 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 2 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, 2015; Thomas & Brunskill, 2016a; 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

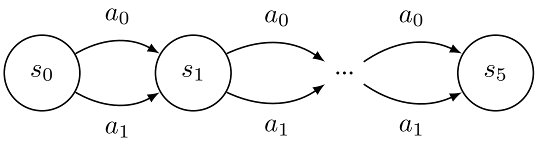


Figure 2. The SinglePath MDP referenced in Section 4 of the main text. **Not shown:** If the agent takes action a_1 it remains in its current state with probability 0.5.

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 (Rubinstein & Kroese, 2013). 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.[1] 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 (Dhariwal et al., 2017) version of proximal policy optimization (PPO) (Schulman et al., 2017) 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

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

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).

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, 2016a) that uses a model of P and r to lower the variance of PDIS.
- The weighted doubly robust (WDR) estimator (Thomas & Brunskill, 2016a) 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 3(a) gives results for all 5 of these IS variants implemented with both RIS weights and OIS weights. Figure 3(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 3 shows RIS weights improve DR and PDIS.

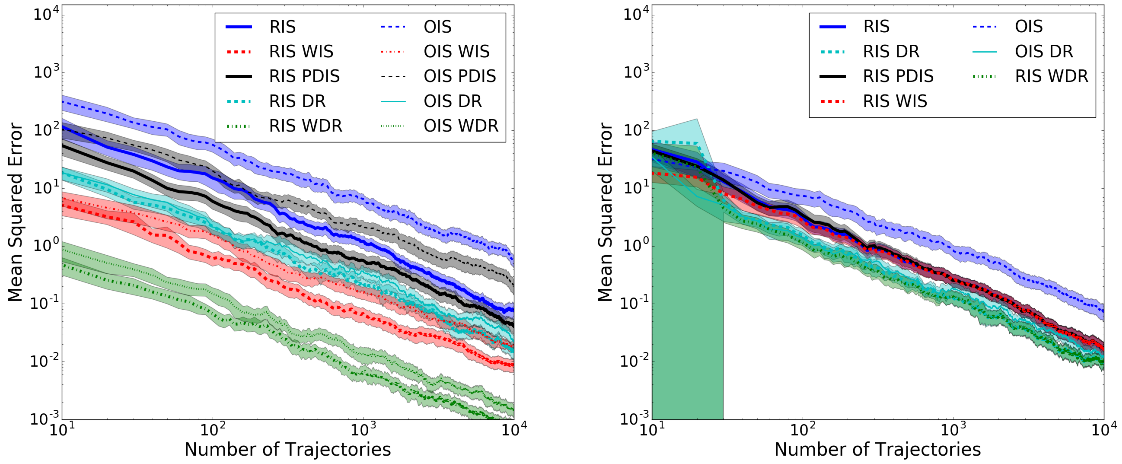
F.2. Gradient Descent Policy Estimation

This appendix shows how the MSE of RIS changes during estimation of $\pi_{\mathcal{D}}$ in the HalfCheetah domain. Figure 4 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.

References

Dhariwal, Prafulla, Hesse, Christopher, Klimov, Oleg, Nichol, Alex, Plappert, Matthias, Radford, Alec, Schulman, John, Sidor, Szymon, and Wu, Yuhuai. Openai baselines. https://github.com/openai/baselines, 2017.

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



(a) Gridworld Off-Policy

(b) Gridworld On-Policy

Figure 3. 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.

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

Henmi, Masayuki, Yoshida, Ryo, and Eguchi, Shinto. Importance sampling via the estimated sampler. *Biometrika*, 94(4): 985–991, 2007.

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

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

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

Rubinstein, Reuven Y and Kroese, Dirk P. *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning*. Springer Science & Business Media, 2013.

Schulman, John, Wolski, Filip, Dhariwal, Prafulla, Radford, Alec, and Klimov, Oleg. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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

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

Thomas, Philip S and Brunskill, Emma. Magical policy search: Data efficient reinforcement learning with guarantees of global optimality. *European Workshop On Reinforcement Learning*, 2016b.

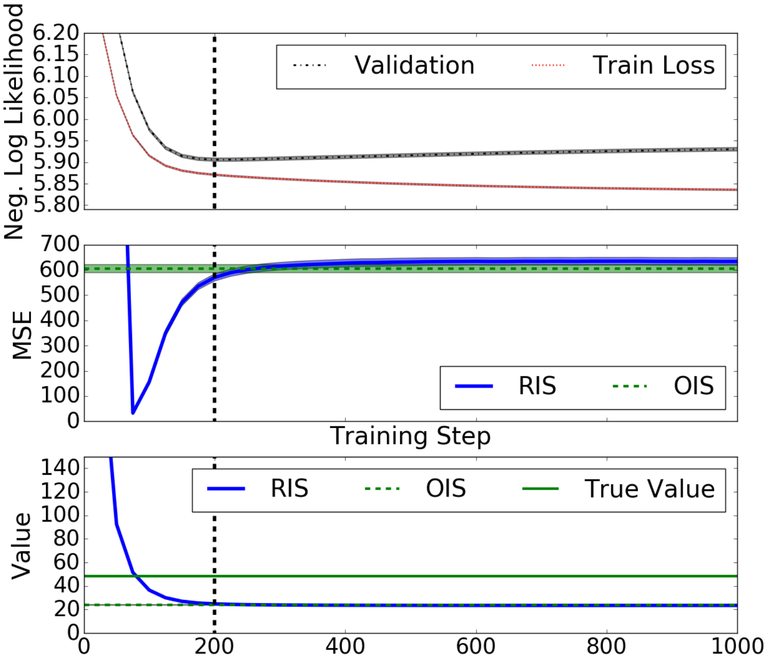


Figure 4. Mean squared error and estimate of the importance sampling estimator during training of $\pi_{\mathcal{D}}$. The x-axis is the number of gradient ascent steps. The top plot shows the training and validation loss curves. The y-axis of the top plot is the average negative log-likelihood. The 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.