In Proceedings of the 34th International Conference on Machine Learning (ICML 2017)
Sydney, Australia August 2017

Data-Efficient Policy Evaluation Through Behavior Policy Search

Josiah P. Hanna [1] **Philip S. Thomas** [2,3] **Peter Stone** [1] **Scott Niekum** [1]

Abstract

We consider the task of evaluating a policy for a *Markov decision process* (MDP). The standard unbiased technique for evaluating a policy is to deploy the policy and observe its performance. We show that the data collected from deploying a different policy, commonly called the *behavior policy*, can be used to produce unbiased estimates with lower mean squared error than this standard technique. We derive an analytic expression for the *optimal behavior policy*—the behavior policy that minimizes the mean squared error of the resulting estimates. Because this expression depends on terms that are unknown in practice, we propose a novel policy evaluation sub-problem, *behavior policy search*: searching for a behavior policy that reduces mean squared error. We present a behavior policy search algorithm and empirically demonstrate its effectiveness in lowering the mean squared error of policy performance estimates.

1. Introduction

Many sequential decision problems, including diabetes treatment (Bastani, 2014), digital marketing (Theocharous et al., 2015), and robot control (Lillicrap et al., 2015), are modeled as *Markov decision processes* (MDPs) and solved using *reinforcement learning* (RL) algorithms. One important problem when applying RL to real problems is *policy evaluation*. The goal in policy evaluation is to estimate the expected *return* (sum of rewards) produced by a policy. We refer to this policy as the *evaluation policy*, π_e. The standard policy evaluation approach is to repeatedly deploy π_e and average the resulting returns. While this naïve Monte Carlo estimator is unbiased, it may have high variance.

[1]The University of Texas at Austin, Austin, Texas, USA [2]The University of Massachusetts, Amherst, Massachusetts, USA [3]Carnegie Mellon University, Pittsburgh, Pennsylvania, USA. Correspondence to: Josiah P. Hanna <jphanna@cs.utexas.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

Methods that evaluate π_e while selecting actions according to π_e are termed *on-policy*. Previous work has addressed variance reduction for on-policy returns (Zinkevich et al., 2006; White & Bowling, 2009; Veness et al., 2011). An alternative approach is to estimate the performance of π_e while following a different, *behavior policy*, π_b. Methods that evaluate π_e with data generated from π_b are termed *off-policy*. *Importance sampling* (IS) is one standard approach for using off-policy data in RL. IS reweights returns observed while executing π_b such that they are unbiased estimates of the performance of π_e.

Presently, IS is usually used when off-policy data is already available or when executing π_e is impractical. If π_b is not chosen carefully, IS often has high variance (Thomas et al., 2015). For this reason, an implicit assumption in the RL community has generally been that on-policy evaluation is more accurate when it is feasible. However, IS can also be used for variance reduction when done with an appropriately selected distribution of returns (Hammersley & Handscomb, 1964). While IS-based variance reduction has been explored in RL, this prior work has required knowledge of the environment's transition probabilities and remains on-policy (Desai & Glynn, 2001; Frank et al., 2008; Ciosek & Whiteson, 2017). In contrast to this earlier work, we show how careful selection of the behavior policy can lead to lower variance policy evaluation than using the evaluation policy and do not require knowledge of the environment's transition probabilities.

In this paper, we formalize the selection of π_b as the *behavior policy search* problem. We introduce a method for this problem that adapts the policy parameters of π_b with gradient descent on the variance of importance-sampling. Empirically we demonstrate behavior policy search with our method lowers the mean squared error of estimates compared to on-policy estimates. To the best of our knowledge, this work is the first to propose adapting the behavior policy to obtain better policy evaluation in RL. Furthermore we present the first method to address this problem.

2. Preliminaries

This section details the policy evaluation problem setting, the Monte Carlo and Advantage Sum on-policy methods, and importance-sampling for off-policy evaluation.

2.1. Background

We use notational standard MDPNv1 (Thomas, 2015), and for simplicity, we assume that $\mathcal{S}, \mathcal{A}$, and R are finite.[1] Let $H := (S_0, A_0, R_0, S_1, \ldots, S_L, A_L, R_L)$ be a *trajectory* and $g(H) := \sum_{t=0}^{L} \gamma^t R_t$ be the *discounted return* of trajectory H. Let $\rho(\pi) := \mathbf{E}[g(H)|H \sim \pi]$ be the expected discounted return when the stochastic policy π is used from S_0 sampled from the initial state distribution. In this work, we consider parameterized policies, $\pi_{\boldsymbol{\theta}}$, where the distribution over actions is determined by the vector $\boldsymbol{\theta}$. We assume that the transitions and reward function are unknown and that L is finite.

We are given an *evaluation policy*, π_e, for which we would like to estimate $\rho(\pi_e)$. We assume there exists a policy parameter vector $\boldsymbol{\theta}_e$ such that $\pi_e = \pi_{\boldsymbol{\theta}_e}$ and that this vector is known. We consider an incremental setting where, at iteration i, we sample a single trajectory H_i with a policy $\pi_{\boldsymbol{\theta}_i}$ and add $\{H_i, \boldsymbol{\theta}_i\}$ to a set $\mathcal{D}$. We use $\mathcal{D}_i$ to denote the set at iteration i. Methods that always (i.e., $\forall i$) choose $\boldsymbol{\theta}_i = \boldsymbol{\theta}_e$ are on-policy; otherwise, the method is off-policy. A policy evaluation method, PE, uses all trajectories in $\mathcal{D}_i$ to estimate $\rho(\pi_e)$. Our goal is to design a policy evaluation algorithm that produces estimates of $\rho(\pi_e)$ that have low *mean squared error* (MSE). Formally, the goal of policy evaluation with PE is to minimize $(\text{PE}(\mathcal{D}_i) - \rho(\pi_e))^2$. While other measures of policy evaluation accuracy could be considered, we follow earlier work in using MSE (e.g., (Thomas & Brunskill, 2016; Precup et al., 2000)).

We focus on unbiased estimators of $\rho(\pi_e)$. While biased estimators (e.g., bootstrapping methods (Sutton & Barto, 1998), approximate models (Kearns & Singh, 2002), etc.) can sometimes produce lower MSE estimates they are problematic for high risk applications requiring confidence intervals. For unbiased estimators, minimizing variance is equivalent to minimizing MSE.

2.2. Monte-Carlo Estimates

Perhaps the most commonly used policy evaluation method is the *on-policy Monte-Carlo* (MC) estimator. The estimate of $\rho(\pi_e)$ at iteration i is the average return:

$$\text{MC}(\mathcal{D}_i) := \frac{1}{i+1} \sum_{j=0}^{i} \sum_{t=0}^{L} \gamma^t R_t = \frac{1}{i+1} \sum_{j=0}^{i} g(H_j).$$

This estimator is unbiased and strongly consistent given mild assumptions.[2] However, this method can have high variance.

2.3. Advantage Sum Estimates

Like the Monte-Carlo estimator, the *advantage sum* (ASE) estimator selects $\boldsymbol{\theta}_i = \boldsymbol{\theta}_e$ for all i. However, it introduces a control variate to reduce the variance without introducing bias. This control variate requires an approximate model of the MDP to be provided. Let the reward function of this model be given as $\hat{r}(s,a)$. Let $\hat{q}^{\pi_e}(s_t, a_t) = \mathbf{E}[\sum_{t'=t}^{L} \gamma^{t'} \hat{r}(s_{t'}, a_{t'})]$ and $\hat{v}^{\pi_e}(s_t) = \mathbf{E}[\hat{q}^{\pi_e}(s_t, a_t)|a_t \sim \pi_e]$, i.e., the action-value function and state-value function of π_e in this approximate model. Then, the advantage sum estimator is given by:

$$\text{AS}(D_i) := \frac{1}{i+1} \sum_{j=0}^{i} \sum_{t=0}^{L} \gamma^t (R_t - \hat{q}^{\pi_e}(S_t, A_t) + \hat{v}^{\pi_e}(S_t)).$$

Intuitively, ASE is replacing part of the randomness of the Monte Carlo return with the known expected return under the approximate model. Provided $q^{\pi_e}(S_t, A_t) - \hat{v}^{\pi_e}(S_t)$ is sufficiently correlated with R_t, the variance of ASE is less than that of MC.

Notice that, like the MC estimator, the ASE estimator is *on-policy*, in that the behavior policy is always the policy that we wish to evaluate. Intuitively it may seems like this choice should be optimal. However, we will show that it is not—selecting behavior policies that are different from the evaluation policy can result in estimates of $\rho(\pi_e)$ that have lower variance.

2.4. Importance Sampling

Importance Sampling is a method for reweighting returns from a *behavior* policy, $\boldsymbol{\theta}$, such that they are unbiased returns from the *evaluation* policy. In RL, the re-weighted IS return of a trajectory, H, sampled from $\pi_{\boldsymbol{\theta}}$ is:

$$\text{IS}(H, \boldsymbol{\theta}) := g(H) \prod_{t=0}^{L} \frac{\pi_e(S_t|A_t)}{\pi_{\boldsymbol{\theta}}(S_t|A_t)}.$$

The IS off-policy estimator is then a Monte Carlo estimate of $\mathbf{E}\left[IS(H, \boldsymbol{\theta})|H \sim \pi_{\boldsymbol{\theta}}\right]$:

$$\text{IS}(\mathcal{D}_i) := \frac{1}{i+1} \sum_{j=0}^{i} \text{IS}(H_j, \boldsymbol{\theta}_j).$$

In RL, importance sampling allows off-policy data to be used as if it were on-policy. In this case the variance of the IS estimate is often much worse than the variance of on-policy MC estimates because the behavior policy is not

[1]The methods, and theoretical results discussed in this paper are applicable to both finite and infinite $\mathcal{S}, \mathcal{A}$ and R as well as *partially-observable* Markov decision processes.

[2]Being a strongly consistent estimator of $\rho(\pi_e)$ means that $\Pr\left(\lim_{i\to\infty} \text{MC}(\mathcal{D}_i) = \rho(\pi_e)\right) = 1$. If $\rho(\pi_e)$ exists, MC is strongly consistent by the Khintchine Strong law of large numbers (Sen & Singer, 1993).

chosen to minimize variance, but is a policy that is dictated by circumstance.

3. Behavior Policy Search

Importance sampling was originally intended as a variance reduction technique for Monte Carlo evaluation (Hammersley & Handscomb, 1964). When an evaluation policy rarely samples trajectories with high magnitude returns a Monte Carlo evaluation will have high variance. If a behavior policy can increase the probability of observing such trajectories then the off-policy IS estimate will have lower variance than an on-policy Monte Carlo estimate. In this section we first describe the theoretical potential for variance reduction with an appropriately selected behavior policy. In general this policy will be unknown. Thus, we propose a policy evaluation subproblem — the behavior policy search problem — solutions to which will adapt the behavior policy to provide lower mean squared error policy performance estimates. To the best of our knowledge, we are the first to propose behavior policy adaptation for policy evaluation.

3.1. The Optimal Behavior Policy

An appropriately selected behavior policy can lower variance to zero. While this fact is generally known for importance-sampling, we show here that this policy exists for any MDP and evaluation policy under two restrictive assumptions: all returns are positive and the domain is deterministic. In the following section we describe how an initial policy can be adapted towards the optimal behavior policy even when these conditions fail to hold.

Let $w_\pi(H) := \prod_{t=0}^{L} \pi(A_t|S_t)$. Consider a behavior policy $\pi_b^\star$ such that for any trajectory, H:

$$\rho(\pi_e) = \mathrm{IS}(H, \pi_b^\star) = g(H)\frac{w_{\pi_e}(H)}{w_{\pi_b^\star}(H)}.$$

Rearranging the terms of this expressions yields:

$$w_{\pi_b^\star}(H) = g(H)\frac{w_{\pi_e}(H)}{\rho(\pi_e)}.$$

Thus, if we can select $\pi_b^\star$ such that the probability of observing any $H \sim \pi_b^\star$ is $\frac{g(H)}{\rho(\pi_e)}$ times the likelihood of observing $H \sim \pi_e$ then the IS estimate has zero MSE with only a single sampled trajectory. Regardless of $g(H)$, the importance-sampled return will equal $\rho(\pi_e)$.

Furthermore, the policy $\pi_b^\star$ exists within the space of all feasible stochastic policies. Consider that a stochastic policy can be viewed as a mixture policy over time-dependent (i.e., action selection depends on the current time-step) deterministic policies. For example, in an MDP with one state, two actions and a horizon of L there are 2^L possible time-dependent deterministic policies, each of which defines a unique sequence of actions. We can express any evaluation policy as a mixture of these deterministic policies. The optimal behavior policy $\pi_b^\star$ can be expressed similarly and thus the optimal behavior policy exists.

Unfortunately, the optimal behavior policy depends on the unknown value $\rho(\pi_e)$ as well as the unknown reward function R (via $g(H)$). Thus, while there exists an optimal behavior policy for IS – which is not π_e – in practice we cannot analytically determine $\pi_b^\star$. Additionally, $\pi_b^\star$ may not be representable by any $\boldsymbol{\theta}$ in our policy class.

3.2. The Behavior Policy Search Problem

Since the optimal behavior policy cannot be analytically determined, we instead propose the behavior policy search (BPS) problem for finding π_b that lowers the MSE of estimates of $\rho(\pi_e)$. A BPS problem is defined by the inputs:

1. An evaluation policy π_e with policy parameters $\boldsymbol{\theta}_e$.
2. An off-policy policy evaluation algorithm, $\mathrm{OPE}(H, \boldsymbol{\theta})$, that takes a trajectory, $H \sim \pi_{\boldsymbol{\theta}}$, or, alternatively, a set of trajectories, and returns an estimate of $\rho(\pi_e)$.

A BPS solution is a policy, $\pi_{\boldsymbol{\theta}_b}$ such that off-policy estimates with OPE have lower MSE than on-policy estimates. Methods for this problem are BPS algorithms.

Recall we have formalized policy evaluation within an incremental setting where one trajectory for policy evaluation is generated each iteration. At the i^{th} iteration, a BPS algorithm selects a behavior policy that will be used to generate a trajectory, H_i. The policy evaluation algorithm, OPE, then estimates $\rho(\pi_e)$ using trajectories in $\mathcal{D}_i$. Naturally, the selection of the behavior policy depends on how OPE estimates $\rho(\pi_e)$.

In a BPS problem, the i^{th} iteration proceeds as follows. First, given all of the past behavior policies, $\{\boldsymbol{\theta}_i\}_{i=0}^{i-1}$, and the resulting trajectories, $\{H_i\}_{i=0}^{i-1}$, the BPS algorithm must select $\boldsymbol{\theta}_i$. The policy $\pi_{\boldsymbol{\theta}_i}$ is then run for one episode to create the trajectory H_i. Then the BPS algorithm uses OPE to estimate $\rho(\pi_e)$ given the available data, $D_i := \{(\boldsymbol{\theta}_i, H_i)\}_{i=0}^{i}$. In this paper, we consider the one-step problem of selecting $\boldsymbol{\theta}_i$ and estimating $\rho(\pi_e)$ at iteration i in a way that minimizes MSE. That is, we do not consider how our selection of $\boldsymbol{\theta}_i$ will impact our future ability to select an appropriate $\boldsymbol{\theta}_j$ for $j > i$ and thus to produce more accurate estimates in the future.

One natural question is: if we are given a limit on the number of trajectories that can be sampled, is it better to "spend" some of our limited trajectories on BPS instead of using on-policy estimates? Since each $OPE(H_i, \boldsymbol{\theta}_i)$ is an unbiased estimator of $\rho(\pi_e)$, we can use all sampled trajectories to compute $\mathrm{OPE}(\mathcal{D}_i)$. Provided for all iterations, $\mathrm{Var}[\mathrm{OPE}(H, \boldsymbol{\theta}_i)] \leq Var[MC]$ then, in expectation, a BPS algorithm will always achieve lower MSE than MC, showing that it is, in fact, worthwhile to do so. This claim is supported by our empirical study.

4. Behavior Policy Gradient Theorem

We now introduce our primary contributions: an analytic expression for the gradient of the mean squared error of the IS estimator and a stochastic gradient descent algorithm that adapts $\boldsymbol{\theta}$ to minimize the MSE between the IS estimate and $\rho(\pi_e)$. Our algorithm — **B**ehavior **P**olicy **G**radient (BPG) — begins with on-policy estimates and adapts the behavior policy with gradient descent on the MSE with respect to $\boldsymbol{\theta}$. The gradient of the MSE with respect to the policy parameters is given by the following theorem:

Theorem 1.

$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{MSE}[\mathrm{IS}(H, \boldsymbol{\theta})] = \mathbf{E}\left[- \mathrm{IS}(H, \boldsymbol{\theta})^2 \sum_{t=0}^{L} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_t|S_t) \right]$$

where the expectation is taken over $H \sim \pi_{\boldsymbol{\theta}}$.

Proof. Proofs for all theoretical results are included in Appendix A.

BPG uses stochastic gradient descent in place of exact gradient descent: replacing the intractable expectation in Theorem 1 with an unbiased estimate of the true gradient. In our experiments, we sample a batch, $\mathcal{B}_i$, of k trajectories with $\pi_{\boldsymbol{\theta}_i}$ to lower the variance of the gradient estimate at iteration i. In the BPS setting, sampling a batch of trajectories is equivalent to holding $\boldsymbol{\theta}$ fixed for k iterations and then updating $\boldsymbol{\theta}$ with the k most recent trajectories used to compute the gradient estimate.

Full details of BPG are given in Algorithm 1. At iteration i, BPG samples a batch, $\mathcal{B}_i$, of k trajectories and adds $\{(\boldsymbol{\theta}_i, H_i)_{i=0}^{k}\}$ to a data set $\mathcal{D}$ (Lines 4-5). Then BPG updates $\boldsymbol{\theta}$ with an empirical estimate of Theorem 1 (Line 6). After n iterations, the BPG estimate of $\rho(\pi_e)$ is $\mathrm{IS}(\mathcal{D}_n)$ as defined in Section 2.4.

Given that the step-size, α_i, is consistent with standard gradient descent convergence conditions, BPG will converge to a behavior policy that locally minimizes the variance (Bertsekas & Tsitsiklis, 2000). At best, BPG converges to the globally optimal behavior policy *within the parameterization of* π_e. Since the parameterization of π_e determines the class of representable distributions it is possible that the theoretically optimal behavior policy is unrepresentable under this parameterization. Nevertheless, a suboptimal behavior policy still yields better estimates of $\rho(\pi_e)$, provided it decreases variance compared to on-policy returns.

Algorithm 1 Behavior Policy Gradient

Input: Evaluation policy parameters, $\boldsymbol{\theta}_e$, batch size k, a step-size for each iteration, α_i, and number of iterations n.

Output: Final behavior policy parameters $\boldsymbol{\theta}_n$ and the IS estimate of $\rho(\pi_e)$ using all sampled trajectories.

1: $\boldsymbol{\theta}_0 \leftarrow \boldsymbol{\theta}_e$
2: $\mathcal{D}_0 = \{\}$
3: **for all** $i \in 0...n$ **do**
4: $\quad \mathcal{B}_i =$ Sample k trajectories $H \sim \pi_{\boldsymbol{\theta}_i}$
5: $\quad \mathcal{D}_{i+1} = \mathcal{D}_i \cup \mathcal{B}_i$
6: $\quad \boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i + \frac{\alpha_i}{k} \sum_{H \in \mathcal{B}} \mathrm{IS}(H, \boldsymbol{\theta})^2 \sum_{t=0}^{L} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}_i}(A_t|S_t)$
7: **end for**
8: **Return** $\boldsymbol{\theta}_n$, $\mathrm{IS}(\mathcal{D}_n)$

4.1. Control Variate Extension

In cases where an approximate model is available, we can further lower variance adapting the behavior policy of the *doubly robust* estimator (Jiang & Li, 2016; Thomas & Brunskill, 2016). Based on a similar intuition as the Advantage Sum estimator (Section 2.3), the Doubly Robust (DR) estimator uses the value functions of an approximate model as a control variate to lower the variance of importance-sampling.[3] We show here that we can adapt the behavior policy to lower the mean squared error of DR estimates. We denote this new method DR-BPG for **D**oubly **R**obust **B**ehavior **P**olicy **G**radient.

Let $w_{\pi,t}(H) = \prod_{i=0}^{t} \pi(A_t|S_t)$ and recall that $\hat{v}^{\pi_e}$ and $\hat{q}^{\pi_e}$ are the state and action value functions of π_e in the approximate model. The DR estimator is:

$$\mathrm{DR}(H, \boldsymbol{\theta}) := \hat{v}(S_0) + \sum_{t=0}^{L} \frac{w_{\pi_e,t}}{w_{\pi_{\boldsymbol{\theta}},t}} (R_t - \hat{q}^{\pi_e}(S_t, A_t) + \hat{v}^{\pi_e}(S_{t+1})).$$

We can reduce the mean squared error of DR with gradient descent using unbiased estimates of the following corollary to Theorem 1:

Corollary 1.

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{MSE}\left[DR(H, \boldsymbol{\theta})\right] = \mathbf{E}[(\mathrm{DR}(H, \boldsymbol{\theta})^2 \sum_{t=0}^{L} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_t|S_t) \\ - 2\,\mathrm{DR}(H, \boldsymbol{\theta}) (\sum_{t=0}^{L} \gamma^t \delta_t \frac{w_{\pi_e,t}}{w_{\boldsymbol{\theta},t}} \sum_{i=0}^{t} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_i|S_i))] \end{aligned}$$

where $\delta_t = R_t - \hat{q}(S_t, A_t) + \hat{v}(S_{t+1})$ and the expectation is taken over $H \sim \pi_{\boldsymbol{\theta}}$.

The first term of $\frac{\partial}{\partial \boldsymbol{\theta}} MSE$ is analogous to the gradient of the importance-sampling estimate with $\mathrm{IS}(H, \boldsymbol{\theta})$ replaced

[3] DR lowers the variance of *per-decision* importance-sampling which importance samples the per time-step reward.

by $\mathrm{DR}(H, \boldsymbol{\theta})$. The second term accounts for the covariance of the DR terms.

AS and DR both assume access to a model, however, they make no assumption about where the model comes from except that it must be independent of the trajectories used to compute the final estimate. In practice, AS and DR perform best when all trajectories are used to estimate the model and then used to estimate $\rho(\pi_e)$ (Thomas & Brunskill, 2016). However, for DR-BPG, changes to the model change the surface of the MSE objective we seek to minimize and thus DR-BPG will only converge once the model stops changing. In our experiments, we consider both a changing and a fixed model.

4.2. Connection to REINFORCE

BPG is closely related to existing work in policy gradient RL (c.f., (Sutton et al., 2000)) and we draw connections between one such method and BPG to illustrate how BPG changes the distribution of trajectories. REINFORCE (Williams, 1992) attempts to maximize $\rho(\pi_{\boldsymbol{\theta}})$ through gradient ascent on $\rho(\pi_{\boldsymbol{\theta}})$ using the following unbiased gradient of $\rho(\pi_{\boldsymbol{\theta}})$:

$$\frac{\partial}{\partial \boldsymbol{\theta}} \rho(\pi_{\boldsymbol{\theta}}) = \mathbf{E}\left[g(H) \sum_{t=0}^{L} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_t|S_t) \middle| H \sim \pi_{\boldsymbol{\theta}} \right].$$

Intuitively, REINFORCE increases the probability of all actions taken during H as a function of $g(H)$. This update increases the probability of actions that lead to high return trajectories. BPG can be interpreted as REINFORCE where the return of a trajectory is the square of its importance-sampled return. Thus BPG increases the probability of all actions taken along H as a function of $IS(H, \boldsymbol{\theta})^2$. The magnitude of $IS(H, \boldsymbol{\theta})^2$ depends on two qualities of H:

1. $g(H)^2$ is large (i.e., a high magnitude event).
2. H is rare relative to its probability under the evaluation policy (i.e., $\prod_{t=0}^{L} \frac{\pi_e(A_t|S_t)}{\pi_{\boldsymbol{\theta}}(A_t|S_t)}$ is large).

These two qualities demonstrate a balance in how BPG changes trajectory probabilities. Increasing the probability of a trajectory under $\pi_{\boldsymbol{\theta}}$ will decrease $IS(H, \boldsymbol{\theta})^2$ and so BPG increases the probability of a trajectory when $g(H)^2$ is large enough to offset the decrease in $IS(H, \boldsymbol{\theta})^2$ caused by decreasing the importance weight.

5. Empirical Study

This section presents an empirical study of variance reduction through behavior policy search. We design our experiments to answer the following questions:

- Can behavior policy search with BPG reduce policy evaluation MSE compared to on-policy estimates in both tabular and continuous domains?
- Does adapting the behavior policy of the Doubly Robust estimator with DR-BPG lower the MSE of the Advantage Sum estimator?
- Does the rarety of actions that cause high magnitude rewards affect the performance gap between BPG and Monte Carlo estimates?

5.1. Experimental Set-up

We address our first experimental question by evaluating BPG in three domains. We briefly describe each domain here; full details are available in appendix C.

The first domain is a 4x4 Gridworld. We obtain two evaluation policies by applying REINFORCE to this task, starting from a policy that selects actions uniformly at random. We then select one evaluation policy, π_1, from the early stages of learning – an improved policy but still far from converged – and one after learning has converged, π_2. We run all experiments once with $\pi_e := \pi_1$ and a second time with $\pi_e := \pi_2$.

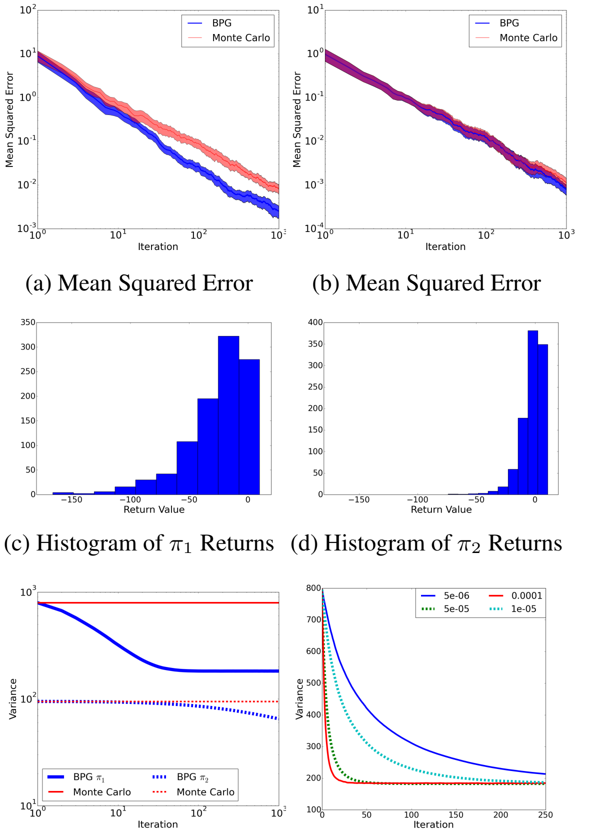
Our second and third tasks are the continuous control Cart-pole Swing Up and Acrobot tasks implemented within RL-LAB (Duan et al., 2016). The evaluation policy in each domain is a neural network that maps the state to the mean of a Gaussian distribution. Policies are partially optimized with trust-region policy optimization (Schulman et al., 2015) applied to a randomly initialized policy.

5.2. Main Results

Gridworld Experiments Figure 1 compares BPG to Monte Carlo for both Gridworld policies, π_1 and π_2. Our main point of comparison is the mean squared error (MSE) of both estimates at iteration i over 100 trials. For π_1, BPG significantly reduces the MSE of on-policy estimates (Figure 1a). For π_2, BPG also reduces MSE, however, it is only a marginal improvement.

At the end of each trial we used the final behavior policy to collect 100 more trajectories and estimate $\rho(\pi_e)$. In comparison to a Monte Carlo estimate with 100 trajectories from π_1, MSE is 85.48 % lower with this improved behavior policy. For π_2, the MSE is 31.02 % lower. This result demonstrates that BPG can find behavior policies that substantially lower MSE.

To understand the disparity in performance between these two instances of policy evaluation, we plot the distribution of $g(H)$ under π_e (Figures 1c and 1d). These plots show the variance of π_1 to be much higher; it sometimes samples returns with twice the magnitude of any sampled by π_2. To quantify this difference, we also measure the variance of $\mathrm{IS}(H, \boldsymbol{\theta}_i)$ as $\mathbf{E}\left[\mathrm{IS}(H)^2 \middle| H \sim \pi_{\boldsymbol{\theta}_i}\right] - \mathbf{E}\left[\mathrm{IS}(H) | H \sim \pi_{\boldsymbol{\theta}_i}\right]^2$ where the expectations are estimated with 10,000 trajecto-



(a) Mean Squared Error (b) Mean Squared Error

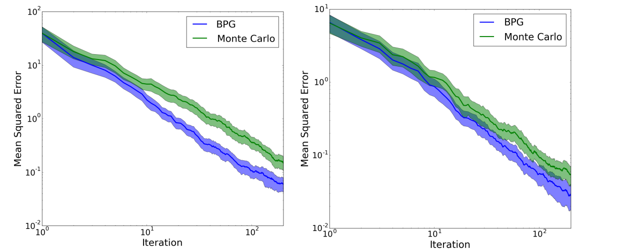
(c) Histogram of π_1 Returns (d) Histogram of π_2 Returns

(e) Variance Reduction (f) Learning Rate Sensitivity

Figure 1: Gridworld experiments when π_e is a partially optimized policy, π_1, (1a) and a converged policy, π_2, (1b). The first and second rows give results for π_1 on the left and π_2 on the right. Results are averaged over 100 trials of 1000 iterations with error bars given for 95 % confidence intervals. In both instances, BPG lowers MSE more than on-policy Monte Carlo returns (statistically significant, $p < 0.05$). The second row shows the distribution of returns under the two different π_e. Figure 1e shows a substantial decrease in variance when evaluating π_1 with BPG and a lesser decrease when evaluating π_2 with BPG. Figure 1f shows the effect of varying the step-size parameter for representative α (BPG diverged for high settings of α). We ran BPG for 250 iterations per value of α and averaged over 5 trials. Axes in 1a, 1b, and 1e are log-scaled.

ries. This evaluation is repeated 5 times per iteration and the reported variance is the mean over these evaluations. The decrease in variance for each policy is shown in Figure 1e. The high initial variance means there is much more room for BPG to improve the behavior policy when $\boldsymbol{\theta}_e$ is the partially optimized policy.

We also test the sensitivity of BPG to the learning rate parameter. A critical issue in the use of BPG is selecting the step size parameter α. If α is set too high we risk making too large of an update to $\boldsymbol{\theta}$ — potentially stepping to a worse behavior policy. If we are too conservative then it will take many iterations for a noticeable improvement over Monte Carlo estimation. Figure 1f shows variance reduction for a number of different α values in the GridWorld domain. We found BPG in this domain was robust to a variety of step size values. We do not claim this result is representative for all problem domains; stepsize selection in the behavior policy search problem is an important area for future work.



(a) Cart-pole Swing Up MSE (b) Acrobot MSE

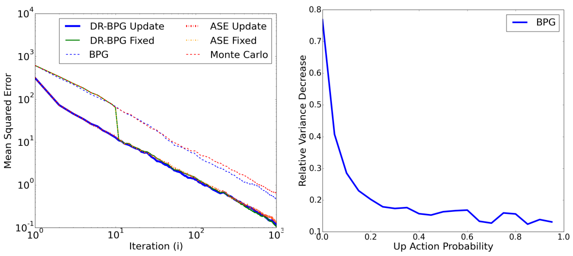
Figure 2: Mean squared error reduction on the Cart-pole Swing Up and Acrobot domains. We adapt the behavior policy for 200 iterations and average results over 100 trials. Error bars are for 95% confidence intervals.

Continuous Control Figure 2 shows reduction of MSE on the Cartpole Swing-up and Acrobot domains. Again we see that BPG reduces MSE faster than Monte Carlo evaluation. In contrast to the discrete Gridworld experiment, this experiment demonstrates the applicability of BPG to the continuous control setting. While BPG significantly outperforms Monte Carlo evaluation in Cart-pole Swing-up, the gap is much smaller in Acrobot. This result also demonstrates BPG (and behavior policy search) when the policy must generalize across different states.

5.3. Control Variate Extensions

In this section, we evaluate the combination of model-based control variates with behavior policy search. Specifically, we compare the AS estimator with Doubly Robust BPG (DR-BPG). In these experiments we use a 10x10 stochastic gridworld. The added stochasticity increases the difficulty of building an accurate model from trajectories.

Since these methods require a model we construct this model in one of two ways. The first method uses all trajectories in $\mathcal{D}$ to build the model and then uses the same set to estimate $\rho(\pi_e)$ with ASE or DR. The second method uses trajectories from the first 10 iterations to build the model and then fixes the model for the remaining iterations. For DR-BPG, behavior policy search starts at iteration 10 un-



(a) Control Variate MSE (b) Rare Event Improvement

Figure 3: **3a:** Comparison of DR and ASE on a larger stochastic Gridworld. For the fixed model methods, the significant drop in MSE at iteration 10 is due to the introduction of the model control variate. For clarity we do not show error bars. The difference between the final estimate of DR-BPG and ASE with the fixed model is statistically significant ($p < 0.05$); the difference between the same methods with a constantly improving model is not. **3b:** Varying the probability of a high rewarding terminal action in the GridWorld domain. Each point on the horizontal axis is the probability of taking this action. The vertical axis gives the relative decrease in variance after adapting $\boldsymbol{\theta}$ for 500 iterations. Denoting the initial variance as v_i and the final variance as v_f, the relative decrease is computed as $\frac{v_i - v_f}{v_i}$. Error bars for 95% confidence intervals are given but are small.

der this second condition. We call the first method "update" and the second method "fixed." The update method invalidates the theoretical guarantees of these methods but learns a more accurate model. In both instances, we build maximum likelihood tabular models.

Figure 3 demonstrates that combining BPG with a model-based control variate (DR-BPG) can lead to further reduction of MSE compared to the control variate alone (ASE). Specifically, with the fixed model, DR-BPG outperformed all other methods. DR-BPG using the update method for building the model performed competitively with ASE although not statistically significantly better. We also evaluate the final learned behavior policy of the fixed model variant of DR-BPG. For a batch size of 100 trajectories, the DR estimator with this behavior policy improves upon the ASE estimator with the same model by 56.9 %.

For DR-BPG, estimating the model with all data still allowed steady progress towards lower variance. This result is interesting since a changing model changes the surface of our variance objective and thus gradient descent on the variance has no theoretical guarantees of convergence. Empirically, we observe that setting the learning rate for DR-BPG was more challenging for either model type. Thus while we have shown BPG can be combined with control variates, more work is needed to produce a robust method.

5.4. Rareness of Event

Our final experiment aims to understand how the gap between on- and off-policy variance is affected by the probability of rare events. The intuition for why behavior policy search can lower the variance of on-policy estimates is that a well selected behavior policy can cause rare and high magnitude events to occur. We test this intuition by varying the probability of a rare, high magnitude event and observing how this change affects the performance gap between on- and off-policy evaluation. For this experiment, we use a variant of the deterministic Gridworld where taking the UP action in the initial state (the upper left corner) causes a transition to the terminal state with a reward of $+50$. We use π_1 from our earlier Gridworld experiments but we vary the probability of choosing UP when in the initial state. So with probability p the agent will receive a large reward and end the trajectory. We use a constant learning rate of 10^{-5} for all values of p and run BPG for 500 iterations. We plot the relative decrease of the variance as a function of p over 100 trials for each value of p. We use relative variance to normalize across problem instances. Note that under this measure, even when p is close to 1, the relative variance is not equal to zero because as p approaches 1 the initial variance also goes to zero.

This experiment illustrates that as the initial variance increases, the amount of improvement BPG can achieve increases. As p becomes closer to 1, the initial variance becomes closer to zero and BPG barely improves over the variance of Monte Carlo (in terms of absolute variance there is no improvement). When the π_e rarely takes the high rewarding UP action (p close to 0), BPG improves policy evaluation by increasing the probability of this action. This experiment supports our intuition for why off-policy evaluation can outperform on-policy evaluation.

6. Related Work

Behavior policy search and BPG are closely related to existing work on adaptive importance-sampling. While adaptive importance-sampling has been studied in the estimation literature, we focus here on adaptive importance-sampling for MDPs and Markov Reward Processes (i.e., an MDP with a fixed policy). Existing work on adaptive IS in RL has considered changing the transition probabilities to lower the variance of policy evaluation (Desai & Glynn, 2001; Frank et al., 2008) or lower the variance of policy gradient estimates (Ciosek & Whiteson, 2017). Since the transition probabilities are typically unknown in RL, adapting the behavior policy is a more general approach to adaptive IS. Ciosek and Whiteson also adapt the distribution of trajectories with gradient descent on the variance (Ciosek & Whiteson, 2017) with respect to parameters of the transition probabilities. The main focus of this work is increasing the probability of simulated rare events so that policy improvement can learn an appropriate response. In contrast, we address the problem of policy evaluation and differentiate with respect to the (known) policy parameters.

The cross-entropy method (CEM) is a general method for adaptive importance-sampling. CEM attempts to minimize the Kullback-Leibler divergence between the current sampling distribution and the optimal sampling distribution. As discussed in Section 3.1, this optimal behavior policy only exists under a set of restrictive conditions. In contrast we adapt the behavior policy by minimizing variance.

Other methods exist for lowering the variance of on-policy estimates. In addition to the control variate technique used by the Advantage Sum estimator (Zinkevich et al., 2006; White & Bowling, 2009), Veness et al. consider using common random numbers and antithetic variates to reduce the variance of roll-outs in Monte Carlo Tree Search (MCTS) (2011). These techniques require a model of the environment (as is typical for MCTS) and do not appear to be applicable to the general RL policy evaluation problem. BPG could potentially be applied to find a lower variance rollout policy for MCTS.

In this work we have focused on unbiased policy evaluation. When the goal is to minimize MSE it is often permissible to use biased methods such as temporal difference learning (van Seijen & Sutton, 2014), model-based policy evaluation (Kearns & Singh, 2002; Strehl et al., 2009), or variants of weighted importance sampling (Precup et al., 2000). It may be possible to use similar ideas to BPG to reduce bias and variance although this appears to be difficult since the bias contribution to the mean squared error is squared and thus any gradient involving bias requires knowledge of the estimator's bias. We leave behavior policy search with biased off-policy methods to future work.

7. Discussion and Future Work

Our experiments demonstrate that behavior policy search with BPG can lower the variance of policy evaluation. One open question is characterizing the settings where adapting the behavior policy substantially improves over on-policy estimates. Towards answering this question, our Gridworld experiment showed that when π_e has little variance, BPG can only offer marginal improvement. BPG increases the probability of observing rare events with a high magnitude. If the evaluation policy never sees such events then there is little benefit to using BPG. However, in expectation and with an appropriately selected step-size, BPG will never lower the data-efficiency of policy evaluation.

It is also necessary that the evaluation policy contributes to the variance of the returns. If all variance is due to the environment then it seems unlikely that BPG will offer much improvement. For example, Ciosek and Whiteson (2017) consider a variant of the Mountain Car task where the dynamics can trigger a rare event — independent of the action — in which rewards are multiplied by 1000. No behavior policy adaptation can lower the variance due to this event.

One limitation of gradient-based BPS methods is the necessity of good step-size selection. In theory, BPG can never lead to worse policy evaluation compared to on-policy estimates. In practice, a poorly selected step-size may cause a step to a worse behavior policy at step i which may increase the variance of the gradient estimate at step $i + 1$. Future work could consider methods for adaptive step-sizes, second order methods, or natural behavior policy gradients.

One interesting direction for future work is incorporating behavior policy search into policy improvement. A similar idea was explored by Ciosek and Whiteson who explored off-environment learning to improve the performance of policy gradient methods (2017). The method presented in that work is limited to simulated environments with differential dynamics. Adapting the behavior policy is a potentially much more general approach.

8. Conclusion

We have introduced the behavior policy search problem in order to improve estimation of $\rho(\pi_e)$ for an evaluation policy π_e. We present a solution — Behavior Policy Gradient — for this problem which adapts the behavior policy with stochastic gradient descent on the variance of the importance-sampling estimator. Experiments demonstrate BPG lowers the mean squared error of estimates of $\rho(\pi_e)$ compared to on-policy estimates. We also demonstrate BPG can further decrease the MSE of estimates in conjunction with a model-based control variate method.

9. Acknowledgements

We thank Daniel Brown and the anonymous reviewers for useful comments on the work and its presentation. This work has taken place in the Personal Autonomous Robotics Lab (PeARL) and Learning Agents Research Group (LARG) at the Artificial Intelligence Laboratory, The University of Texas at Austin. PeARL research is supported in part by NSF (IIS-1638107, IIS-1617639). LARG research is supported in part by NSF (CNS-1330072, CNS-1305287, IIS-1637736, IIS-1651089), ONR (21C184-01), AFOSR (FA9550-14-1-0087), Raytheon, Toyota, AT&T, and Lockheed Martin. Josiah Hanna is supported by an NSF Graduate Research 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

Bastani, Meysam. *Model-free intelligent diabetes management using machine learning*. PhD thesis, Masters thesis, Department of Computing Science, University of Alberta, 2014.

Bertsekas, Dimitri P. and Tsitsiklis, John N. Gradient convergence in gradient methods with erros. 10:627–642, 2000.

Ciosek, Kamil and Whiteson, Shimon. OFFER: Off-environment reinforcement learning. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

Desai, Paritosh Y and Glynn, Peter W. Simulation in optimization and optimization in simulation: A Markov chain perspective on adaptive Monte Carlo algorithms. In *Proceedings of the 33rd conference on Winter simulation*, pp. 379–384. IEEE Computer Society, 2001.

Duan, Yan, Chen, Xi, Houthooft, Rein, Schulman, John, and Abbeel, Pieter. Benchmarking deep reinforcement learning for continuous control. In *In Proceedings of the 33rd International Conference on Machine Learning*, 2016.

Frank, Jordan, Mannor, Shie, and Precup, Doina. Reinforcement learning in the presence of rare events. In *Proceedings of the 25th International Conference on Machine learning*, pp. 336–343. ACM, 2008.

Hammersley, JM and Handscomb, DC. Monte Carlo methods, methuen & co. *Ltd., London*, pp. 40, 1964.

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.

Kearns, Michael and Singh, Satinder. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2-3):209–232, 2002.

Lillicrap, Timothy P., Hunt, Jonathan J., Pritzel, Alexander, Heess, Nicolas, Erez, Tom, Tassa, Yuval, Silver, David, and Wierstra, Daan. Continuous control with deep reinforcement learning. *CoRR*, abs/1509.02971, 2015.

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*, pp. 759–766, 2000.

Schulman, John, Levine, Sergey, Moritz, Philipp, Jordan, Michael, and Abbeel, Pieter. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

Sen, P.K. and Singer, J.M. *Large Sample Methods in Statistics: An Introduction with Applications*. Chapman & Hall, 1993.

Strehl, Alexander L, Li, Lihong, and Littman, Michael L. Reinforcement learning in finite mdps: PAC analysis. *Journal of Machine Learning Research*, 10:2413–2444, 2009.

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

Sutton, Richard S., McAllester, David, Singh, Satinder, and Mansour, Yishay. Policy gradient methods for reinforcement learning with function approximation. In *Proceedings of the 13th Conference on Neural Information Processing Systems (NIPS)*, 2000.

Theocharous, Georgios, Thomas, Philip S., and Ghavamzadeh, Mohammad. Personalized ad recommendation systems for life-time value optimization with guarantees. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1806–1812, 2015.

Thomas, Philip S. A notation for Markov decision processes. *ArXiv*, arXiv:1512.09075v1, 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)*, 2016.

Thomas, Philip S., Theocharous, Georgios, and Ghavamzadeh, Mohammad. High confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2015.

van Seijen, Harm and Sutton, Richard S. True online TD (λ). In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, volume 14, pp. 692–700, 2014.

Veness, J., Lanctot, M., and Bowling, M. Variance reduction in Monte-Carlo tree search. In *Proceedings of the 24th Conference on Neural Information Processing Systems*, pp. 1836–1844, 2011.

White, M. and Bowling, M. Learning a value analysis tool for agent evaluation. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1976–1981, 2009.

Williams, Ronald J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Zinkevich, M., Bowling, M., Bard, N., Kan, M., and Billings, D. Optimal unbiased estimators for evaluating agent performance. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, pp. 573–578, 2006.

A. Proof of Theorem 1

In Appendix A, we give the full derivation of our primary theoretical contribution — the importance-sampling (IS) variance gradient. We also present the variance gradient for the doubly-robust (DR) estimator.

We first derive an analytic expression for the gradient of the variance of an arbitrary, unbiased off-policy policy evaluation estimator, $\mathrm{OPE}(H, \boldsymbol{\theta})$. Importance-sampling is one such off-policy policy evaluation estimator. From our general derivation we derive the gradient of the variance of the IS estimator and then extend to the DR estimator.

A.1. Variance Gradient of an Unbiased Off-Policy Policy Evaluation Method

We first present a lemma from which $\frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{MSE}[IS(H, \boldsymbol{\theta})]$ and $\frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{MSE}[DR(H, \boldsymbol{\theta})]$ can both be derived.

Lemma 1 gives the gradient of the mean squared error (MSE) of an unbiased off-policy policy evaluation method.

Lemma 1.

$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{MSE}[\mathrm{OPE}(H, \boldsymbol{\theta})] = \mathbf{E}\left[\mathrm{OPE}(H, \boldsymbol{\theta})^2 (\sum_{t=0}^{L} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_t|S_t)) + \frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{OPE}(H, \boldsymbol{\theta})^2 \middle| H \sim \pi_{\boldsymbol{\theta}} \right]$$

Proof. We begin by decomposing $\Pr(H|\pi)$ into two components—one that depends on π and the other that does not. Let

$$w_\pi(H) \coloneqq \prod_{t=0}^{L} \pi(A_t|S_t),$$

and

$$p(H) \coloneqq \Pr(H|\pi)/w_\pi(H),$$

for any π such that $H \in \mathrm{supp}(\pi)$ (any such π will result in the same value of $p(H)$). These two definitions mean that $\Pr(H|\pi) = p(H) w_\pi(H)$.

The MSE of the OPE estimator is given by:

$$\mathrm{MSE}[\mathrm{OPE}(H, \boldsymbol{\theta})] = \mathrm{Var}[\mathrm{OPE}(H, \boldsymbol{\theta})] + \underbrace{(\mathbf{E}[\mathrm{OPE}(H, \boldsymbol{\theta})] - \rho(\pi_e))^2}_{\text{bias}^2}.$$

Since the OPE estimator is unbiased, i.e., $\mathbf{E}[\mathrm{OPE}(H, \boldsymbol{\theta})] = \rho(\pi_e)$, the second term is zero and so:

$$\begin{aligned}
\mathrm{MSE}(\mathrm{OPE}(H, \boldsymbol{\theta})) &= \mathrm{Var}(\mathrm{OPE}(H, \boldsymbol{\theta})) \\
&= \mathbf{E}\left[\mathrm{OPE}(H, \boldsymbol{\theta})^2 \middle| H \sim \pi_{\boldsymbol{\theta}}\right] - \mathbf{E}[\mathrm{OPE}(H, \boldsymbol{\theta}) | H \sim \pi_{\boldsymbol{\theta}}]^2 \\
&= \mathbf{E}\left[\mathrm{OPE}(H, \boldsymbol{\theta})^2 \middle| H \sim \pi_{\boldsymbol{\theta}}\right] - \rho(\pi_e)^2
\end{aligned}$$

To obtain the MSE gradient, we differentiate $\mathrm{MSE}(\mathrm{OPE}(H, \boldsymbol{\theta}))$ with respect to $\boldsymbol{\theta}$:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{MSE}[\mathrm{OPE}(H, \boldsymbol{\theta})] &= \frac{\partial}{\partial \boldsymbol{\theta}} \left[\mathbf{E}\left[\mathrm{OPE}(H, \boldsymbol{\theta})^2 \middle| H \sim \pi_{\boldsymbol{\theta}}\right] - \rho(\pi_e)^2 \right] \\
&= \frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{E}_{H \sim \pi_{\boldsymbol{\theta}}}\left[\mathrm{OPE}(H, \boldsymbol{\theta})^2\right] \\
&= \frac{\partial}{\partial \boldsymbol{\theta}} \sum_H \Pr(H|\boldsymbol{\theta}) \mathrm{OPE}(H, \boldsymbol{\theta})^2 \\
&= \sum_H \Pr(H|\boldsymbol{\theta}) \frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{OPE}(H, \boldsymbol{\theta})^2 + \mathrm{OPE}(H, \boldsymbol{\theta})^2 \frac{\partial}{\partial \boldsymbol{\theta}} \Pr(H|\boldsymbol{\theta}) \\
&= \sum_H \Pr(H|\boldsymbol{\theta}) \frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{OPE}(H, \boldsymbol{\theta})^2 + \mathrm{OPE}(H, \boldsymbol{\theta})^2 p(H) \frac{\partial}{\partial \boldsymbol{\theta}} w_{\pi_{\boldsymbol{\theta}}}(H) \qquad (1)
\end{aligned}$$

Consider the last factor of the last term in more detail:

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}} w_{\pi_{\boldsymbol{\theta}}}(H) =& \frac{\partial}{\partial \boldsymbol{\theta}} \prod_{t=0}^{L} \pi_{\boldsymbol{\theta}}(A_t|S_t) \\
\overset{\text{(a)}}{=}& \left(\prod_{t=0}^{L} \pi_{\boldsymbol{\theta}}(A_t|S_t) \right) \left(\sum_{t=0}^{L} \frac{\frac{\partial}{\partial \boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(A_t|S_t)}{\pi_{\boldsymbol{\theta}}(A_t|S_t)} \right) \\
=& w_{\pi_{\boldsymbol{\theta}}}(H) \sum_{t=0}^{L} \frac{\partial}{\partial \boldsymbol{\theta}} \log\left(\pi_{\boldsymbol{\theta}}(A_t|S_t)\right),
\end{aligned}
$$

where **(a)** comes from the multi-factor product rule. Continuing from (A.1) we have that:

$$
\frac{\partial}{\partial \boldsymbol{\theta}} \operatorname{MSE}(\operatorname{OPE}(H, \boldsymbol{\theta})) = \mathbf{E}\left[\operatorname{OPE}(H, \boldsymbol{\theta})^2 \sum_{t=0}^{L} \frac{\partial}{\partial \boldsymbol{\theta}} \log\left(\pi_{\boldsymbol{\theta}}(A_t|S_t)\right) + \frac{\partial}{\partial \boldsymbol{\theta}} \operatorname{OPE}(H, \boldsymbol{\theta})^2 \middle| H \sim \pi_{\boldsymbol{\theta}} \right].
$$

□

A.2. Behavior Policy Gradient Theorem

We now use Lemma 1 to prove the Behavior Policy Gradient Theorem which is our main theoretical contribution.

Theorem 2.

$$
\frac{\partial}{\partial \boldsymbol{\theta}} \operatorname{MSE}[\operatorname{IS}(H, \boldsymbol{\theta})] = \mathbf{E}\left[-\operatorname{IS}(H, \boldsymbol{\theta})^2 \sum_{t=0}^{L} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_t|S_t) \middle| H \sim \pi_{\boldsymbol{\theta}} \right]
$$

where the expectation is taken over $H \sim \pi_{\boldsymbol{\theta}}$.

Proof. We first derive $\frac{\partial}{\partial \boldsymbol{\theta}} \operatorname{IS}(H, \boldsymbol{\theta})^2$. Theorem 1 then follows directly from using $\frac{\partial}{\partial \boldsymbol{\theta}} \operatorname{IS}(H, \boldsymbol{\theta})^2$ as $\frac{\partial}{\partial \boldsymbol{\theta}} \operatorname{OPE}(H, \boldsymbol{\theta})^2$ in Lemma 1.

$$
\begin{aligned}
\operatorname{IS}(H, \boldsymbol{\theta})^2 =& \left(\frac{w_{\pi_e}}{w_{\boldsymbol{\theta}}} g(H) \right)^2 \\
\frac{\partial}{\partial \boldsymbol{\theta}} \operatorname{IS}(H, \boldsymbol{\theta})^2 =& \frac{\partial}{\partial \boldsymbol{\theta}} \left(\frac{w_{\pi_e}(H)}{w_{\boldsymbol{\theta}}(H)} g(H) \right)^2 \\
=& 2 \cdot g(H) \frac{w_{\pi_e}(H)}{w_{\boldsymbol{\theta}}(H)} \frac{\partial}{\partial \boldsymbol{\theta}} \left(g(H) \frac{w_{\pi_e}(H)}{w_{\boldsymbol{\theta}}(H)} \right) \\
\overset{\text{(a)}}{=}& -2 \cdot g(H) \frac{w_{\pi_e}(H)}{w_{\boldsymbol{\theta}}(H)} \left(g(H) \frac{w_{\pi_e}(H)}{w_{\boldsymbol{\theta}}(H)} \right) \sum_{t=0}^{L} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_t|S_t) \\
=& -2 \operatorname{IS}(H, \boldsymbol{\theta})^2 \sum_{t=0}^{L} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_t|S_t)
\end{aligned}
$$

where **(a)** comes from the multi-factor product rule and using the likelihood-ratio trick (i.e., $\frac{\frac{\partial}{\partial \boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(A|S)}{\pi_{\boldsymbol{\theta}}(A|S)} = \log \pi_{\boldsymbol{\theta}}(A|S)$)

Substituting this expression into Lemma 1 completes the proof:

$$
\frac{\partial}{\partial \boldsymbol{\theta}} \operatorname{MSE}[\operatorname{IS}(H, \boldsymbol{\theta})] = \mathbf{E}\left[-\operatorname{IS}(H, \boldsymbol{\theta})^2 \sum_{t=0}^{L} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_t|S_t) \middle| H \sim \pi_{\boldsymbol{\theta}} \right]
$$

□

A.3. Doubly Robust Estimator

Our final theoretical result is a corollary to the Behavior Policy Gradient Theorem: an extension of the IS variance gradient to the Doubly Robust (DR) estimator. Recall that for a single trajectory DR is given as:

$$\mathrm{DR}(H, \boldsymbol{\theta}) \coloneqq \hat{v}^{\pi_e}(S_0) + \sum_{t=0}^{L} \gamma^t \frac{w_{\pi_e,t}}{w_{\boldsymbol{\theta},t}} \left(R_t - \hat{q}^{\pi_e}(S_t, A_t) + \hat{v}^{\pi_e}(S_{t+1})\right)$$

where $\hat{v}^{\pi_e}$ is the state-value function of π_e under an approximate model, $\hat{q}^{\pi_e}$ is the action-value function of π_e under the model, and $w_{\pi,t} \coloneqq \prod_{j=0}^{t} \pi(A_j|S_j)$.

The gradient of the mean squared error of the DR estimator is given by the following corollary to the Behavior Policy Gradient Theorem:

Corollary 2.

$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{MSE}\left[DR(H, \boldsymbol{\theta})\right] \quad = \quad \mathbf{E}[(\mathrm{DR}(H, \boldsymbol{\theta})^2 \sum_{t=0}^{L} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_t|S_t) \quad - \quad 2\,\mathrm{DR}(H, \boldsymbol{\theta})(\sum_{t=0}^{L} \gamma^t \delta_t \frac{w_{\pi_e,t}}{w_{\boldsymbol{\theta},t}} \sum_{i=0}^{t} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_i|S_i))]$$

where $\delta_t = R_t - \hat{q}(S_t, A_t) + \hat{v}(S_{t+1})$ and the expectation is taken over $H \sim \pi_{\boldsymbol{\theta}}$.

Proof. As with Theorem 1, we first derive $\frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{DR}(H, \boldsymbol{\theta})^2$. Corollary 1 then follows directly from using $\frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{DR}(H, \boldsymbol{\theta})^2$ as $\frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{OPE}(H, \boldsymbol{\theta})^2$ in Lemma 1.

$$\mathrm{DR}(H, \boldsymbol{\theta})^2 = \left(\hat{v}^{\pi_e}(S_0) + \sum_{t=0}^{L} \gamma^t \frac{w_{\pi_e,t}}{w_{\boldsymbol{\theta},t}} \left(R_t - \hat{q}^{\pi_e}(S_t, A_t) + \hat{v}^{\pi_e}(S_{t+1})\right)\right)^2$$

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{DR}(H, \boldsymbol{\theta})^2 =& \frac{\partial}{\partial \boldsymbol{\theta}} \left(\hat{v}^{\pi_e}(S_0) + \sum_{t=0}^{L} \gamma^t \frac{w_{\pi_e,t}}{w_{\boldsymbol{\theta},t}} \left(R_t - \hat{q}^{\pi_e}(S_t, A_t) + \hat{v}^{\pi_e}(S_{t+1})\right)\right)^2 \\
=& 2\,\mathrm{DR}(H, \boldsymbol{\theta}) \frac{\partial}{\partial \boldsymbol{\theta}} \left(\hat{v}^{\pi_e}(S_0) + \sum_{t=0}^{L} \gamma^t \frac{w_{\pi_e,t}}{w_{\boldsymbol{\theta},t}} (R_t - \hat{q}^{\pi_e}(S_t, A_t) + \hat{v}^{\pi_e}(S_{t+1}))\right) \\
=& -2\,\mathrm{DR}(H, \boldsymbol{\theta})(\sum_{t=0}^{L} \gamma^t \frac{w_{\pi_e,t}}{w_{\boldsymbol{\theta},t}} (R_t - \hat{q}^{\pi_e}(S_t, A_t) + \hat{v}^{\pi_e}(S_{t+1})) \sum_{i=0}^{t} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_i|S_i))
\end{aligned}$$

Thus the $\mathrm{DR}(H, \boldsymbol{\theta})$ gradient is:

$$= \mathbf{E}\left[\mathrm{DR}(H, \boldsymbol{\theta})^2 \sum_{t=0}^{L} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_t|S_t) - 2\,\mathrm{DR}(H, \boldsymbol{\theta})(\sum_{t=0}^{L} \gamma^t \frac{w_{\pi_e,t}}{w_{\boldsymbol{\theta},t}} (R_t - \hat{q}^{\pi_e}(S_t, A_t) + \hat{v}^{\pi_e}(S_{t+1})) \sum_{i=0}^{t} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_i|S_i)) \middle| H \sim \pi_{\boldsymbol{\theta}}\right]$$

□

The expression for the DR behavior policy gradient is more complex than the expression for the IS behavior policy gradient. Lowering the variance of DR involves accounting for the covariance of the sum of terms. Intuitively, accounting for the covariance increases the complexity of the expression for the gradient.

B. BPG's Off-Policy Estimates are Unbiased

This appendix proves that BPG's estimate is an unbiased estimate of $\rho(\pi_e)$. If only trajectories from a single $\boldsymbol{\theta}_i$ were used then clearly $\mathrm{IS}(\cdot, \boldsymbol{\theta}_i)$ is an unbiased estimate of $\rho(\pi_e)$. The difficulty is that the BPG's estimate at iteration n depends on all $\boldsymbol{\theta}_i$ for $i = 1 \dots n$ and each $\boldsymbol{\theta}_i$ is *not* independent of the others. Nevertheless, we prove here that BPG produces an unbiased estimate of $\rho(\pi_e)$ at each iteration. Specifically, we will show that $\mathbf{E}\left[\mathrm{IS}(H_n, \boldsymbol{\theta}_n)|\boldsymbol{\theta}_0 = \boldsymbol{\theta}_e)\right]$ is an unbiased estimate of $\rho(\pi_e)$, where the IS estimate is conditioned on $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_e$. To make the dependence of $\boldsymbol{\theta}_i$ on $\boldsymbol{\theta}_{i-1}$ explicit, we will write $f(H_{i-1}) := \boldsymbol{\theta}_i$ where $H_{i-1} \sim \pi_{\boldsymbol{\theta}_{i-1}}$. We use $\Pr(h|\boldsymbol{\theta})$ as shorthand for $\Pr(H = h|\boldsymbol{\theta})$.

$$
\begin{aligned}
\mathbf{E}\left[\mathrm{IS}(H_n, \boldsymbol{\theta}_n)|\boldsymbol{\theta} = \boldsymbol{\theta}_e)\right] =& \sum_{h_0} \Pr(h_0|\boldsymbol{\theta}_0) \sum_{h_1} \Pr(h_1|f(h_0)) \cdots \underbrace{\sum_{h_n} \Pr(h_n|f(h_{n-1}))\,\mathrm{IS}(h_n)}_{\rho(\pi_e)} \\
=& \rho(\pi_e) \sum_{h_0} \Pr(h_0|\boldsymbol{\theta}_0) \sum_{h_1} \Pr(h_1|f(h_0)) \cdots \\
=& \rho(\pi_e)
\end{aligned}
$$

Notice that, even though BPG's off-policy estimates at each iteration are unbiased, they are *not* statistically independent. This means that concentration inequalities, like Hoeffding's inequality, cannot be applied directly. We conjecture that the conditional independence properties of BPG (specifically that H_i is independent of H_{i-1} given θ_i), are sufficient for Hoeffding's inequality to be applicable.

C. Supplemental Experiment Description

This appendix contains experimental details in addition to the details contained in Section 5 of the paper.

Gridworld: This domain is a 4x4 Gridworld with a terminal state with reward 10 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 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 softmax action selection with temperature 1 where the probability of taking an action a in a state s is given by:

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

We obtain two evaluation policies by applying REINFORCE to this task, starting from a policy that selects actions uniformly at random. We then select one evaluation policy from the early stages of learning – an improved policy but still far from converged –, π_1, and one after learning has converged, π_2. We run our set of experiments once with $\pi_e := \pi_1$ and a second time with $\pi_e := \pi_2$. The ground truth value of $\rho(\pi_e)$ is computed with value iteration for both π_e.

Stochastic Gridworld: The layout of this Gridworld is identical to the deterministic Gridworld except the terminal state is at $(9,9)$ and the $+1$ reward state is at $(1,9)$. When the agent moves, it moves in its intended direction with probability 0.9, otherwise it goes left or right with equal probability. Noise in the environment increases the difficulty of building an accurate model from trajectories.

Continuous Control: We evaluate BPG on two continuous control tasks: Cart-pole Swing Up and Acrobot. Both tasks are implemented within RLLAB (Duan et al., 2016) (full details of the tasks are given in Appendix 1.1). The single task modification we make is that in Cart-pole Swing Up, when a trajectory terminates due to moving out of bounds we give a penalty of -1000. This modification increases the variance of π_e. We use $\gamma = 1$ and $L = 50$. Policies are represented as conditional Gaussians with mean determined by a neural network with two hidden layers of 32 tanh units each and a state-independent diagonal covariance matrix. In Cart-pole Swing Up, π_e was learned with 10 iterations of the TRPO algorithm (Schulman et al., 2015) applied to a randomly initialized policy. In Acrobot, π_e was learned with 60 iterations. The ground truth value of $\rho(\pi_e)$ in both domains is computed with 1,000,000 Monte Carlo roll-outs.

Domain Independent Details In all experiments we subtract a constant control variate (or baseline) in the gradient estimate from Theorem 1. The baseline is $b_i = \mathbf{E}\left[-\mathrm{IS}(H)^2 \middle| H \sim \boldsymbol{\theta}_{i-1}\right]$ and our new gradient estimate is:

$$
\mathbf{E}\left[(-\mathrm{IS}^2 - b_i) \sum_{t=0}^{L} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_t|S_t) \middle| H \sim \pi_{\boldsymbol{\theta}}\right]
$$

Adding or subtracting a constant does not change the gradient in expectation since $b_i \cdot \mathbf{E}\left[\sum_{t=0}^{L} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_t|S_t)\right] = 0$. BPG with a baseline has lower variance so that the estimated gradient is closer in direction to the true gradient.

We use batch sizes of 100 trajectories per iteration for Gridworld experiments and size 500 for the continuous control tasks. The step-size parameter was determined by a sweep over $[10^{-2}, 10^{-6}]$

Early Stopping Criterion In all experiments we run BPG for a fixed number of iterations. In general, BPS can continue for a fixed number of iterations or until the variance of the IS estimator stops decreasing. The true variance is unknown but can be estimated by sampling a set of k trajectories with $\boldsymbol{\theta}_i$ and computing the *uncentered* variance: $\frac{1}{k}\sum_{j=0}^{k} \mathrm{OPE}(H_j, \boldsymbol{\theta}_j)^2$. This measure can be used to empirically evaluate the quality of each $\boldsymbol{\theta}$ or determine when a BPS algorithm should terminate behavior policy improvement.