Journal of Machine Learning Research 25 (2024) 1-58 Submitted 4/21; Revised 7/23; Published 10/24

Data-Efficient Policy Evaluation Through Behavior Policy Search

Josiah P. Hanna JPHANNA@CS.WISC.EDU
Computer Sciences Department
University of Wisconsin – Madison
Madison, WI, USA

Yash Chandak, Philip S. Thomas {YCHANDAK,PTHOMAS}@CS.UMASS.EDU
College of Information and Computer Sciences
University of Massachusetts
Amherst, MA, USA

Martha White WHITEM@UALBERTA.CA
Department of Computing Science
University of Alberta and the Alberta Machine Intelligence Institute (Amii)
Edmonton, Alberta, CA

Peter Stone PSTONE@CS.UTEXAS.EDU
Department of Computer Science
The University of Texas at Austin and Sony AI
Austin, TX, USA

Scott Niekum SNIEKUM@CS.UMASS.EDU
College of Information and Computer Sciences
University of Massachusetts
Amherst, MA, USA

Editor: Pradeep Ravikumar

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 a *minimal variance behavior policy* – a 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 two behavior policy search algorithms and empirically demonstrate their effectiveness in lowering the mean squared error of policy performance estimates.[1]

Keywords: Off-policy reinforcement learning; policy evaluation; importance sampling

1. A shorter version of this work first appeared at the International Conference on Machine Learning (ICML) (Hanna et al., 2017).

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* 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 (Hammersley and Handscomb, 1964), it may have high variance.

Methods that evaluate π_e while selecting actions according to π_e are termed *on-policy*. Previous work has addressed variance reduction for methods that collect data on-policy (e.g., Zinkevich et al. (2006); White and Bowling (2009); Veness et al. (2011); Hanna et al. (2021)). 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 re-weights returns observed while executing π_b such that they are unbiased estimates of the performance of π_e (Thomas, 2015).

Presently, IS is usually used when off-policy data is already available or when executing π_e is impractical. In such circumstances, IS often has high variance (Thomas et al., 2015a; Jiang and Li, 2016; Guo et al., 2017). 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 and 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 and Glynn, 2001; Frank et al., 2008; Ciosek and Whiteson, 2017). In contrast to this earlier work, we show how careful selection of the behavior policy can lead to lower variance batch policy evaluation than using the evaluation policy without requiring knowledge of the environment's transition probabilities.

In this work, we formalize the selection of π_b as the *behavior policy search* problem. After formalizing this problem, we introduce two algorithms for this problem that adapt the policy parameters of π_b to find a behavior policy that provides lower variance importance sampling estimates. The first method directly minimizes the variance of the importance sampling estimator using gradient descent on the parameters of π_b. The second method uses gradient descent to minimize the KL-divergence between the behavior policy and a derived *minimal-variance* behavior policy. Empirically we demonstrate that behavior policy search with both of our methods 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 lower mean squared error policy evaluation in RL. Furthermore we present the first methods to address this problem.

This article builds upon and includes work first presented at the 34th International Conference on Machine Learning (ICML) (Hanna et al., 2017). Going beyond this earlier work, we formally derive a condition that a minimal-variance behavior policy must satisfy, we introduce a second behavior policy search algorithm, derive formal convergence and convexity results, prove statistical properties of our algorithms, and we extend the empirical

study contained in the original work. Taken together, these contributions and the earlier work comprise a complete study of behavior policy search for data-efficient policy evaluation.

2. Background

We first present the notation used throughout this work. We then formalize the *batch* policy evaluation problem for Markov decision processes and discuss two common approaches to this problem. Finally we survey literature related to batch policy evaluation and the use of *adaptive importance sampling* in reinforcement learning.

2.1 Notation

We assume the environment is a finite-horizon, episodic *Markov decision process* (MDP) with state set $\mathcal{S}$, action set $\mathcal{A}$, transition function, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$, bounded reward function $r : \mathcal{S} \times \mathcal{A} \to [r_{\texttt{min}}, r_{\texttt{max}}]$, horizon l, discount factor $\gamma \in [0, 1]$, and initial state distribution $d_0 : \mathcal{S} \to [0, 1]$ (Puterman, 2014). We use $P(s'|s, a) = P(s, a, s')$ to denote the conditional probability of transitioning to state s' after taking action a in state s. We assume that $\mathcal{S}$ and $\mathcal{A}$ are finite though our empirical analysis is conducted in both finite and infinite $\mathcal{S}$ and $\mathcal{A}$ MDPs. We assume that the transition and reward functions are unknown and that the maximum episode length, l, is a finite constant.

A policy, $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, is a function mapping states to probability distributions over $\mathcal{A}$. Let Π be the set of all such policies. We use $\pi(a|s) = \pi(s, a)$ to denote the conditional probability of action a given state s In this work, we consider parameterized policies, $\pi_{\boldsymbol{\theta}}$, where the distribution over actions given a state is determined by a vector $\boldsymbol{\theta} \in \Theta$, where $\Theta \subseteq \mathbf{R}^d$ for some dimension d. Furthermore, we require $\pi_{\boldsymbol{\theta}}(a|s)$ to be twice-differentiable with respect to $\boldsymbol{\theta}$ at every state-action pair and for $\frac{\partial}{\partial \boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(a|s)$ and $\frac{\partial^2}{\partial^2 \boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(a|s)$ to be bounded by a finite constant for all states, actions, and values of $\boldsymbol{\theta}$.

The agent interacts with the environment MDP as follows: The agent begins in initial state $S_0 \sim d_0$. At discrete time-step t the agents takes action $A_t \sim \pi(A|S_t)$. The environment responds with $R_t \coloneqq r(S_t, A_t)$ and $S_{t+1} \sim P(\cdot|S_t, A_t)$ according to the reward function and transition function. The agent's interaction with the environment terminates after l steps regardless of the agent's current state or action. We allow the possibility of termination before l steps by including a special terminal state, s_∞. If the agent enters the terminal state, s_∞, it remains there and receives zero reward until step l is reached. Note that the finite-horizon assumption implies that the current time-step of interaction must be included as part of the current state.

Let $h \coloneqq (s_0, a_0, r_0, s_1, \ldots, s_{l-1}, a_{l-1}, r_{l-1})$ be a *trajectory* and $g(h) \coloneqq \sum_{t=0}^{l-1} \gamma^t r_t$ be the *discounted return* of h. Note that $g(h)$ is bounded since the per-time-step reward is bounded. Any policy defines a distribution over trajectories, $\Pr(H = h|\pi)$, where H is a random variable denoting a trajectory. We will write $H \sim \pi$ to denote sampling a trajectory by following π as described in the preceding paragraph and $\mathcal{H} \coloneqq \mathcal{S}^l \times \mathcal{A}^l \times \mathcal{R}^l$ to denote the set of all possible trajectories. Finally, we define the *value* of a policy, $v(\pi) \coloneqq \mathbf{E}[g(H)|H \sim \pi]$, as the expected discounted return when sampling a trajectory with policy π.

2.2 Batch Policy Evaluation

In the batch policy evaluation problem, we are given an *evaluation policy*, π_e, for which we would like to estimate $v(\pi_e)$. 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 D. We use D_i to denote the set at iteration i (including $(H_i, \boldsymbol{\theta}_i)$) where $D_0 = \emptyset$. We use superscripts on states, actions, and rewards to denote the trajectory in which they occur: $H_i \coloneqq (S_0^i, A_0^i, R_0^i, ..., S_{l-1}^i, A_{l-1}^i, R_{l-1}^i)$.

A batch policy evaluation method, PE, uses all trajectories in D_i to estimate $v(\pi_e)$. Methods that always (i.e., $\forall i$) choose $\boldsymbol{\theta}_i = \boldsymbol{\theta}_e$ are on-policy; otherwise, the method is off-policy. Our goal is to design a batch policy evaluation algorithm that produces estimates of $v(\pi_e)$ that have low *mean squared error* (MSE). Formally, we express this goal as selecting PE to minimize:

$$\text{MSE}\Big[\text{PE}\Big] \coloneqq \mathbf{E}\Big[\Big(\text{PE}(D_i) - v(\pi_e)\Big)^2\Big],$$

where D_i is a random variable representing the data set at iteration i. While other measures of policy evaluation accuracy could be considered, we follow earlier work in using MSE (e.g., Thomas and Brunskill (2016); Precup et al. (2000)).

In this work, we focus on unbiased estimators. An *unbiased* estimator is an estimator whose estimates have expected value equal to $v(\pi_e)$. For unbiased estimators, minimizing variance is equivalent to minimizing MSE. While biased estimators (like bootstrapping methods (Sutton and Barto, 2018, Chapter 6) and approximate models (Kearns and Singh, 2002)) can sometimes produce lower MSE estimates, some applications may call for unbiased estimators.

The algorithms we introduce only consider the problem of selecting $\boldsymbol{\theta}_i$ and estimating $v(\pi_e)$ to minimize the MSE at iteration i. That is, they do *not* consider how the 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.

2.3 Monte Carlo Batch Policy Evaluation

Perhaps the simplest batch policy evaluation method is the *on-policy Monte-Carlo* (MC) estimator. As an on-policy method, the Monte Carlo estimator requires $\boldsymbol{\theta}_i = \boldsymbol{\theta}_e$ for all iterations i. The estimate of $v(\pi_e)$ at iteration i is the mean return:

$$\overline{\text{MC}}(\pi_e, D_i) \coloneqq \frac{1}{i}\sum_{j=1}^{i}\sum_{t=0}^{l-1}\gamma^t R_t^j = \frac{1}{i}\sum_{j=1}^{i} g(H_j).$$

This estimator is unbiased and strongly consistent given mild assumptions.[2] However, this method can have high variance (Sutton and Barto, 2018, Chapter 5).

2. Being a strongly consistent estimator of $v(\pi_e)$ means that $\Pr\left(\lim_{i\to\infty}\overline{\text{MC}}(\pi_e, D_i) = v(\pi_e)\right) = 1$. If $v(\pi_e)$ exists, the Monte Carlo estimator is strongly consistent (Sen and Singer, 1993).

2.4 Importance Sampling Policy Evaluation

The Monte Carlo estimator requires that all trajectories are collected on-policy by running π_e. It can be generalized to the *off-policy* setting by re-weighting returns from any *behavior policy*, π_b, such that they are unbiased estimates of the expected return of the *evaluation* policy (Sutton and Barto, 2018, Chapter 5). The off-policy Monte Carlo estimator is known in the RL literature as the *importance sampling* (IS) estimator. Notice that if trajectories under the behavior policy π_b are not informative for evaluating π_e, then this re-weighting procedure may not be feasible. Therefore, to avoid such problems we make a standard assumption that is needed for importance sampling.

Assumption 1 *The quotient* $\frac{\pi_e(a|s)}{\pi_{\boldsymbol{\theta}}(a|s)}$ *exists and is bounded above by (an unknown)* $c < \infty$, $\forall s \in \mathcal{S}, \forall a \in \mathcal{A}, \forall \boldsymbol{\theta} \in \Theta$.

Intuitively, Assumption 1 says that any outcome that is possible under the evaluation policy π_e is also possible under any of the behavior policies. Assumption 1 can be trivially satisfied by ensuring $\pi_{\boldsymbol{\theta}}$ is bounded away from zero. Under this assumption, the re-weighted IS return of a trajectory, H, sampled from behavior policy π_b is:

$$\mathrm{IS}(\pi_e, H, \pi_b) \coloneqq g(H) \prod_{t=0}^{l-1} \frac{\pi_e(A_t|S_t)}{\pi_b(A_t|S_t)}. \tag{1}$$

Intuitively, the IS return up-weights returns that were more likely under π_e than π_b and down-weights returns that were less likely under π_e compared to π_b. The IS estimator at iteration i is then:

$$\overline{\mathrm{IS}}(\pi_e, D_i) \coloneqq \frac{1}{i} \sum_{j=1}^{i} \mathrm{IS}(\pi_e, H_j, \pi_{\boldsymbol{\theta}_i}).$$

Note that when $\pi_{\boldsymbol{\theta}_i}$ and π_e are the same for all i, the IS estimator is identical to the Monte Carlo estimator.

In RL, importance sampling allows off-policy data to be used as if it were on-policy. Importance sampling is both unbiased and consistent, however, like the Monte Carlo estimator, it may suffer from high variance (Thomas, 2015). The variance of IS may in fact be worse than that of on-policy Monte Carlo because the importance weights themselves can contribute to the variance (Sutton and Barto, 2018, Chapter 5). In many uses of IS in reinforcement learning, the variance of the IS estimate is often much worse than the variance of on-policy MC estimates because the behavior policy is not chosen to minimize variance, but is a policy that is dictated by circumstance.

3. Related Work

The methods we will introduce can be classified as *adaptive importance sampling* methods. This section surveys the related literature of adaptive importance sampling for reinforcement learning. We also discuss additional literature on lowering variance for policy evaluation.

3.1 Adaptive Importance Sampling

In this work we introduce algorithms that lower the variance of batch policy evaluation by adapting the behavior policy and then importance sampling to correct for the distribution shift. Such algorithms are closely related to existing work on adaptive importance-sampling. While adaptive IS has been studied in the Monte Carlo simulation literature, we focus here on adaptive IS for MDPs and Markov reward processes (MRPs), i.e., Markov chains with rewards at each state. Existing work on adaptive IS in RL has considered changing the transition probabilities of the MDP to lower the variance of policy evaluation. Since the transition probabilities are typically uncontrollable in RL, adapting the behavior policy is a more general approach to adaptive IS in RL.

Desai and Glynn (2001) and Ahamed et al. (2006) consider adaptive importance sampling for estimating the expected cost until termination in an MRP. They introduce algorithms that perform adaptive importance sampling by modifying the state transition matrix of the Markov chain. In contrast to these works, we focus on policy evaluation in MDPs.

Frank et al. (2008) consider adaptive importance sampling for TD-learning (Sutton, 1988) in MDPs. They assume a known probability of a rare event taking place and assume learning occurs in a simulator where this probability can be changed. They propose two algorithms that adapt the probability of a rare event and use importance sampling to remove bias from the distribution shift. These algorithms lead to faster convergence of TD-learning algorithms. In contrast to this work, we only assume that we know the evaluation policy and adapt the behavior policy for low variance importance sampling estimates. We also only consider estimating $v(\pi_e)$ instead of the expected return from all states, i.e., the state value-function.

Ciosek and Whiteson (2017) adapt the environment transition probabilities to minimize the variance of each component of an *on-policy* policy gradient estimate. This work assumes a known environment transition function and that learning is done in a simulator where the transition function can be modified. In contrast, we focus on the problem of batch policy evaluation in an unknown environment and lower variance through off-policy data collection.

The one prior work we know of that adapts the behavior policy is the work of Bouchard et al. (2016) who adapt the behavior policy to lower the variance of batch policy gradient estimates. Their algorithm adapts the behavior policy to lower the variance of each component of the vector-valued off-policy policy gradient estimate for a different, target policy. This approach is shown to lead to faster learning on a Grid World domain compared to on-policy batch policy gradient learning. In contrast to this work, we study the problem of batch policy evaluation of a fixed policy. While this work was in submission, additional related work has considered the use of offline data for behavior policy search (Liu and Zhang, 2024), safety constraints in behavior policy search (Mukherjee et al., 2024a; Wan et al., 2022), and data collection for biased estimators of policy value (Zhong et al., 2022; Mukherjee et al., 2022, 2024b; Corrado and Hanna, 2023; Mutný et al., 2023; Arnold et al., 2022).

3.2 Variance Reduction for Policy Evaluation

Aside from adaptive importance sampling, other methods exist for lowering the variance of on-policy estimates. Control variates (Zinkevich et al., 2006; White and Bowling, 2009; Jiang and Li, 2016; Thomas and Brunskill, 2016) are a widely used technique for variance

reduction in RL. As we show in Section 9.3, this technique can be used in conjunction with adaptive importance sampling.

Veness et al. (2011) use common random numbers and antithetic variates to lower the variance of policy evaluation in Monte Carlo tree search (MCTS). These techniques require the environment to be known and appear to be inapplicable to the general RL policy evaluation problem. We note that the algorithms we introduce could potentially be applied, in combination with the methods of Veness et al. (2011), to lower the variance of value estimates in MCTS.

In this work we focus on unbiased batch policy evaluation. When the goal is to minimize MSE it is often permissible to use biased methods such as temporal difference learning (Sutton, 1988), model-based policy evaluation (Kearns and Singh, 2002; Strehl et al., 2009), variants of weighted importance sampling (Precup et al., 2000), stationary distribution corrections (Hallak and Mannor, 2017; Liu et al., 2018; Gelada and Bellemare, 2019; Yang et al., 2020), or tree back-ups (Precup et al., 2000; Asis et al., 2017). It may be possible to use adaptive importance sampling to reduce bias and variance although the methods we introduce are *not* directly extensible to accomplish bias and variance reduction. We leave behavior policy search with biased off-policy methods to future work.

4. The Behavior Policy Search Problem

The importance sampling estimator (1) is often viewed as a high variance technique for using off-policy data – in fact the standard RL textbook states, in reference to methods using importance sampling, that "off-policy learning is inherently of greater variance than on-policy learning" (Sutton and Barto, 2018, Chapter 5). However, outside of RL, importance sampling was originally intended as a variance reduction technique for Monte Carlo evaluation (Hammersley and Handscomb, 1964). In this section we first provide intuition for how importance sampling with a behavior policy different than π_e can reduce the variance of importance sampling. This intuition motivates us to propose a policy evaluation sub-problem – the behavior policy search problem – solutions to which are policies that provide lower MSE off-policy batch policy evaluation than on-policy estimators. We then prove statistical properties on the off-policy estimates that are produced as we adapt the behavior policy, showing that such estimates are unbiased and consistent and that we can construct confidence intervals on the estimates. To the best of our knowledge, we are the first to propose behavior policy adaptation for lower variance policy evaluation.

4.1 Motivating Off-Policy Sampling for Lower Variance Importance Sampling

To gain intuition for how importance sampling can lower the variance of Monte Carlo returns, we first examine why importance sampling often increases variance in RL. First, we make the straightforward observation that any particular behavior policy will induce a particular distribution over weighted returns and the weighted returns will have some variance under this distribution. In the case of on-policy sampling, this distribution is just the distribution of unweighted returns since $\pi_e(a|s) = \pi_b(a|s)$ and all importance weights are equal to one. Since choosing $\pi_b \neq \pi_e$ means the importance weights themselves have non-zero variance, it is natural to assume that the variance of the weighted returns can only increase when we multiply non-zero variance unweighted returns with non-zero variance weights. In fact, this

case often does arise in RL when the behavior policy is dictated by circumstance (e.g., when using historical logged data) (Thomas et al., 2015a).

Looking closer at why the variance can be magnified under off-policy sampling, we can see that some importance weights are greater than 1 while others are less than 1. Weights greater than 1 will magnify the magnitude of the associated return while weights less than 1 will lessen this magnitude. As a consequence, we can see that if we could select π_b such that the largest magnitude unweighted returns receive weights less than 1 and the smallest magnitude returns received weights greater than 1 then the overall variance of the weighted returns would decrease relative to the variance of the unweighted returns. In effect, the spread of possible return values would decrease and hence the variance would decrease as well.

In fact, there is even a special case in which a well-chosen behavior policy could decrease the variance of an importance sampling estimate to zero. Consider the case when d_0 and P are deterministic, all rewards are positive and imagine we have a behavior policy ${\pi_b}^\star$ such that for all $h \in \mathcal{H}$:

$$v(\pi_e) = \text{IS}(\pi_e, h, {\pi_b}^\star) = g(h)\frac{\Pr(H = h|\pi_e)}{\Pr(H = h|{\pi_b}^\star)}.$$

Rearranging the terms of this expression yields:

$$\Pr(H = h|{\pi_b}^\star) = g(h)\frac{\Pr(H = h|\pi_e)}{v(\pi_e)}. \tag{2}$$

Thus, if we could select ${\pi_b}^\star$ such that the probability of observing any $H \sim {\pi_b}^\star$ is $\frac{g(H)}{v(\pi_e)}$ times the likelihood of observing $H \sim \pi_e$, then the IS estimate has zero variance with only a single sampled trajectory! Regardless of the value of $g(H)$, the importance weight under ${\pi_b}^\star$ will scale $g(H)$ exactly to $v(\pi_e)$ for all possible realizations of H and the importance-sampled return will equal $v(\pi_e)$.

While in principle importance weights can be used to decrease the variance of the unweighted returns under π_e, we have yet to show that one should expect there to exist a behavior policy that yields the necessary importance weights for any MDP and evaluation policy pair. We consider this question with a small scale empirical study on randomly generated MDP-π_e pairs. Specifically, we randomly generate MDPs from the class of Garnet MDPs (Archibald et al., 1995; Piot et al., 2014) with 10 states, 2 actions, a branching factor of 2 (each state-action pair leads to at most 2 possible next states), and a maximum horizon of 3. The transition probabilities are given by a softmax distribution with temperature τ_P. Both rewards and π_e's action probabilities are given by a softmax distribution over actions in each state. These distributions use temperature parameter τ_R and τ_π respectively. Logits for all softmax distributions are sampled uniformly from $[0, 1]$. The small size of these randomly generated MDPs allows us to analytically compute the variance of an importance sampling estimate with a particular behavior policy. Furthermore, we can analytically compute the gradient of the variance with respect to the softmax parameters of the policy.[3] For a randomly generated MDP-π_e pair, we first compute the variance with $\pi_b \leftarrow \pi_e$. We then compute the gradient, $\mathbf{g}$, of the variance and create a new behavior policy with a single step of gradient descent, $\boldsymbol{\theta}_b \leftarrow \boldsymbol{\theta}_e - \alpha\frac{\mathbf{g}}{||\mathbf{g}||_2}$ where $\alpha = 0.001$ is a scalar step-size parameter.

3. We will elaborate on the derivation of this gradient in Section 5.

Finally, we compute the variance with $\pi_b \leftarrow \pi_{\boldsymbol{\theta}_b}$ and measure the difference between the initial variance and new variance.

The parameters τ_P, τ_R, τ_π allow us to vary the transition entropy, per-state reward variance, and evaluation policy entropy respectively of the randomly generated MDP-π_e pairs. Our objective is to see under what settings there exists a behavior policy that lowers the variance of importance sampling compared to using $\pi_b = \pi_e$. Figure 2 plots variance reduction as a function of the three task parameters that we vary. In all settings that we consider we find that adapting the behavior policy leads to no worse variance than using π_e though the degree of possible variance reduction varies across settings. In particular, the three cases where adapting the behavior policy leads to minimal variance reduction are 1) when the reward function has low variance across actions (high τ_R), 2) when π_e is near uniform random (higher τ_π), and 3) when π_e is deterministic. The last case sometimes occurs for the smaller tested τ_π and explains the wider confidence interval see in Figure 2. While this experiment does not establish there will always be a $\pi_b \neq \pi_e$ that decreases the variance of importance sampling, it shows that it is in principle possible in some cases to lower variance by adapting the behavior policy. This finding motivates the behavior policy search problem which we introduce in the next subsection.

4.2 The Behavior Policy Search Problem

With the potential to lower the variance of importance sampling via off-policy sampling in mind, we now introduce the *behavior policy search* (BPS) problem for finding π_b that lowers the MSE of estimates of $v(\pi_e)$. While the previous subsection focused on the IS-estimator, this subsection considers the more general class of unbiased off-policy value estimators.

A BPS problem is defined by the inputs:

1. An evaluation policy π_e with policy parameters $\boldsymbol{\theta}_e$.

2. An initial behavior policy, $\pi_{\boldsymbol{\theta}_0}$, with policy parameters $\boldsymbol{\theta}_0$. We assume from here on that $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_e$.

3. An off-policy policy evaluation estimator, $\text{OPE}(\pi_e, H, \pi_{\boldsymbol{\theta}})$, that takes a trajectory, $H \sim \pi_{\boldsymbol{\theta}}$ and returns an estimate of $v(\pi_e)$.

A BPS solution is a policy, $\pi_{\boldsymbol{\theta}_b}$, that generates trajectories, H, such that $\text{OPE}(\pi_e, H, \pi_{\boldsymbol{\theta}_b})$ has lower MSE than $\text{OPE}(\pi_e, H, \pi_e)$. Algorithms for this problem are BPS algorithms.

Recall that we consider an incremental batch policy evaluation setting where at each iteration i we can select a behavior policy to collect a trajectory and add this trajectory to a dataset containing trajectories collected at earlier iterations. At the i^{th} iteration, a BPS algorithm selects a behavior policy that will be used to generate a trajectory, H_i. We then add trajectory H_i to dataset D_{i-1} to form dataset D_i. Finally, we estimate $v(\pi_e)$ as the mean value of OPE across all trajectories in D. Naturally, the selection of the behavior policy depends on how the estimator estimates $v(\pi_e)$.

In a BPS problem, the i^{th} iteration proceeds as follows. First, given all of the past behavior policies, $\{\pi_{\boldsymbol{\theta}_j}\}_{j=1}^{i-1}$, and the resulting trajectories, $\{H_j\}_{j=1}^{i-1}$, the BPS algorithm must select $\boldsymbol{\theta}_i$. The policy $\pi_{\boldsymbol{\theta}_i}$ is run for one episode to generate the trajectory H_i. Then the BPS

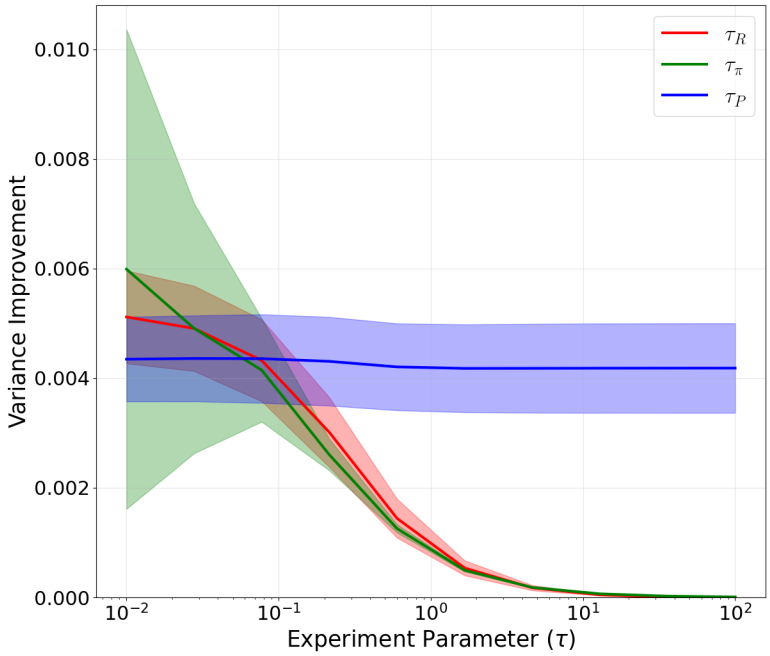


Figure 1: Variance Improvement

Figure 2: Reduction of variance on random MDPs with varying properties. The vertical axis shows change in the variance of importance sampling after adapting the behavior policy's parameters with a single step of gradient descent on the variance. The horizontal axis is the MDP parameter that is varied. Higher indicates a larger reduction in variance and the shaded region indicates a 95% confidence interval.

algorithm estimates $v(\pi_e)$ as the mean of OPE in the available data, D_i:

$$\overline{\text{OPE}}(\pi_e, D_i) \coloneqq \frac{1}{i}\sum_{j=1}^{i} \text{OPE}(\pi_e, H_j, \pi_{\boldsymbol{\theta}_j}).$$

At the final iteration, the algorithm returns the final policy parameters and the estimate of $v(\pi_e)$ using all trajectories collected while running the algorithm. If for all iterations, the variance of OPE with $H \sim \pi_{\boldsymbol{\theta}_i}$ is less than that of OPE with $H \sim \pi_e$ (i.e., on-policy policy evaluation) then a BPS algorithm will have lower variance than an on-policy policy evaluation. Thus adapting the behavior policy is statistically more efficient than simply collecting all trajectories with π_e.

It is worth noting that adapting the behavior policy increases the computational complexity of estimating $v(\pi_e)$. The exact increase will depend on the behavior policy search algorithm used and the dimension of $\boldsymbol{\theta}$, however, it seems unlikely that a behavior policy search algorithm will match the computational simplicity of simply running the evaluation

policy. Thus practitioners must decide whether computational or statistical efficiency is more appropriate for a particular application.

4.3 Statistical Properties of Behavior Policy Search Estimates

To enable better statistical efficiency, so far we have focused on reducing variance while evaluating π_e by adequately adjusting π_b. Before we present concrete algorithms for behavior policy search, it is important to ensure that any such search procedure does not give up other desired statistical properties like unbiasedness, consistency, and finite sample rates. These properties when using unbiased estimates are typically established under the assumption that the trajectories $\{H_j\}_{j=1}^i$ are independent (Thomas, 2015). However, notice that when using a behavior policy search algorithm the policy parameters $(\theta_j)_{j=1}^i$ will be iteratively obtained and hence need *not* be independent of each other, and thus even the trajectories $\{H_j\}_{i=j}^i$ in D_i need not be independent of each other either. Moreover, the distribution of the random variable $\text{OPE}(\pi_e, H_j, \pi_{\theta_j})$ can vary when θ_j is different for different values of $j \in \{1, ..., i\}$ as well. These two factors combined violate *both* the independence and identical distribution assumptions that are often required to establish statistical guarantees on estimators. Figure 3 presents a graphical depiction of the concern.

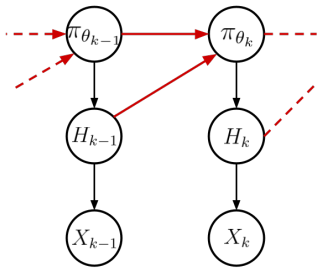


Figure 3: Graphical depiction of the dependencies between the variables. Here, X_k is the random variable corresponding to the estimate $\text{OPE}(\pi_e, H_k, \pi_{\theta_k})$ using the trajectory H_k generated using the behavior policy π_{θ_k}. Red arrows correspond to any learning algorithm that looks at the sampled trajectories and updates policy $\pi_{\theta_{k-1}}$ to π_{θ_k}, thereby introducing the dependency between X's. Further, the distribution of X_{k-1} and X_k can vary as θ_{k-1} and θ_k can be different. These factors combined violate *both* the independence and identical distribution assumptions.

This problem occurs even in settings beyond the behavior policy search problem. For instance, many reinforcement learning methods leverage off-policy trajectories to update policy parameters, which are consequently used to generate new trajectories. Therefore, $\overline{\text{OPE}}$ estimates using these trajectories violate the i.i.d. assumption as well. This raises the question:

Can we obtain statistical properties for the $\overline{\text{OPE}}$ estimate, similar to what is possible under the i.i.d. setting, in the above settings where the i.i.d. assumptions are violated?

In what follows, we answer this question positively and show that despite the violation of the i.i.d. assumption, strong guarantees on unbiasedness, consistency, and concentration rates can still be obtained. First, we present these results in a generic form that applies regardless of how the behavior policy is updated. As trajectories may not be i.i.d. in other applications of off-policy evaluation, these results are of independent interest apart from behavior policy search. In the context of this article, these results establish unbiasedness, consistency, and concentration rates for the specific behavior policy search algorithms that we introduce. For simplicity, we will only consider the case that any algorithm (stochastically) selects $\pi_{\theta_{j+1}}$ given only the previous parameter θ_j and the corresponding trajectory H_j.

For our results to hold for estimates computed as the mean of a set of unbiased estimates, $\{\text{OPE}(\pi_e, H, \pi_{\boldsymbol{\theta}_j})\}_{j=1}^{i}$, we require the following assumption.

Assumption 2 *The unbiased, off-policy policy evaluation estimator* $\text{OPE}(\pi_e, H, \pi_{\boldsymbol{\theta}})$ *is bounded in the range* $[\texttt{min}, \texttt{max}]$ *for finite constants* `min` *and* `max` *for all trajectories and choices of* $\pi_{\boldsymbol{\theta}}$.

For the IS-estimator, Assumption 1 and bounded rewards imply that Assumption 2 is satisfied.

Proposition 1 *Under Assumption 2,* $\overline{\text{OPE}}(\pi_e, D_n)$ *is an unbiased estimator of* $v(\pi_e)$ *for any* $n \in \mathbb{N}$,

$$\mathbb{E}\left[\overline{\text{OPE}}(\pi_e, D_n)\right] = v(\pi_e).$$

Proof The proof is presented in Appendix A. ■

An important consequence of Proposition 1 is that, despite lacking independence from each other, the $(\text{OPE}(\pi_e, H_j, \pi_{\boldsymbol{\theta}_j}))_{j=1}^{i}$ estimates are *uncorrelated*. We formalize this statement below and then use it to establish other properties of the $\overline{\text{OPE}}(\pi_e, D_i)$ estimate.

Lemma 1 *Under Assumption 2,* $\forall j \in \mathbb{N}$, *and* $\forall k \in \mathbb{N}$, *where* $j \neq k$, $\text{OPE}(\pi_e, H_j, \pi_{\theta_j})$ *and* $\text{OPE}(\pi_e, H_k, \pi_{\theta_k})$ *are uncorrelated. That is,*

$$\forall j \neq k, \quad \text{Cov}\left(\text{OPE}(\pi_e, H_j, \pi_{\theta_j}), \text{OPE}(\pi_e, H_k, \pi_{\theta_k})\right) = 0.$$

Proof The proof is presented in Appendix A. ■

Remark 1 *While Lemma* 1 *implies that the expected value (first moment) of* $\text{OPE}(\pi_e, H_j, \pi_{\theta_j})$ *is independent of* $\text{OPE}(\pi_e, H_k, \pi_{\theta_k})$, *the higher moments of* $\text{OPE}(\pi_e, H_j, \pi_{\theta_j})$ *may still depend on* $\text{OPE}(\pi_e, H_k, \pi_{\theta_k})$.

A desired property for any estimator is that it provides a more accurate estimate as the amount of data increases. Typically, Kolmogorov's strong law (Sen and Singer, 1993, Theorem 2.3.10) is used to show consistency of estimators, however, it requires random variables to be independent. While the independence (and identical distribution) assumption is violated in our setting, we show below that asymptotic consistency can still be established.

Proposition 2 *Under Assumption 2,* $\overline{\text{OPE}}(\pi_e, D_i)$ *converges to* $v(\pi_e)$ *in probability. That is, for* $\epsilon > 0$,

$$\lim_{i \to \infty} \Pr\left(|\overline{\text{OPE}}(\pi_e, D_i) - v(\pi_e)| > \epsilon\right) = 0.$$

Proof The core idea of the proof relies upon results from Proposition 1 and Lemma 1 to show that mean-squared-error of $\overline{\text{OPE}}(\pi_e, D_i)$ asymtotically converges to 0. The complete proof is presented in Appendix A. ■

While asymptotic consistency is desirable, it is often also essential to quantify finite sample rates to understand the dependency on the sample size, construct confidence intervals, etc. Because i.i.d. assumptions are violated in our setup, it is not immediately clear if existing methods that make the i.i.d. assumption can be leveraged as-is to provide finite sample rates. To resolve this difficulty, we use a common technique based on Martingales to obtain finite sample rates for the specific setting of our interest.

Proposition 3 *Under Assumption* 2*, let* $\tilde{c}$ *be the range of* $\text{OPE}(\pi_e, H_j, \pi_{\theta_j})$ *for any* $j \in \{1, 2, ..., i\}$*, then* $\forall \delta \in [0, 1]$*,*

$$\Pr\left(\left|\overline{\text{OPE}}(\pi_e, D_i) - v(\pi_e)\right| > \tilde{c}\sqrt{\frac{\ln(2/\delta)}{2|D_i|}}\right) \leq \delta,$$

Proof The core idea of the proof relies upon modeling the sequence $\left(\left(\text{OPE}(\pi_e, D_j, \pi_{\theta_j})\right)_{j=1}^{i}\right.$ as Martingales and then using concentration inequalities for the Martingales. The complete proof is provided in Appendix A. ■

Remark 2 *Note that Proposition* 3 *reduces to naively applying Hoeffding's inequality on the OPE estimates* $(\text{OPE}(\pi_e, H_j, \pi_{\theta_j}))_{j=1}^{i}$*, even though neither independence nor the identical distribution assumption holds.*

Remark 3 *The concentration bound given in Proposition* 3 *depends upon the range of the* $\text{OPE}(\pi_e, H_j, \pi_{\boldsymbol{\theta}_j})$ *estimates. Taking importance sampling as an example and assuming the returns* $g(h)$ *are bounded, we can observe that for any* $\pi_{\boldsymbol{\theta}_j} \neq \pi_e$ *the range of* $\text{IS}(\pi_e, H_j, \pi_{\boldsymbol{\theta}_j})$ *increases and so the bound becomes looser (see Thomas et al. (2015a) for additional discussion). Thus, even if a behavior policy search algorithm lowers variance (and thus MSE) compared to on-policy sampling, Proposition* 3 *still assigns the estimate a looser finite-sample bound than the estimate from on-policy sampling. An alternative to Hoeffding-style bounds are Student's t-Test bounds which depend on the sample variance. We would expect t-Test bounds to return a tighter error bound for behavior policy search algorithms that compute behavior policies that lower the variance of off-policy evaluation. However, t-Test bounds require the assumption that* $\overline{\text{OPE}}(\pi_e, D_i)$ *is normally distributed and this assumption is typically false for small data sets. While this requirement invalidates the error bound, Thomas et al. (2015b) note that in certain cases t-Test bounds are overly conservative which makes them suitable for applications of high-confidence off-policy evaluation.*

Propositions 1, 2, and 3 ensure that the statistical guarantees on unbiasedness, consistency, and finite sample rates can still be achieved even if any behavior policy search algorithm results in non i.i.d. returns. In the following sections, we now introduce concrete solution

algorithms for the behavior policy search problem. We will first introduce an algorithm that optimizes the behavior policy to minimize the variance of an importance sampling estimate. We then introduce an algorithm that optimizes the behavior policy to minimize a measure of divergence between a minimal-variance behavior policy and the current behavior policy. We will also introduce behavior policy search algorithms for extensions to the basic importance sampling estimator.

5. Behavior Policy Gradient on the Variance

Our first behavior policy search algorithm is derived from the perspective of selecting the behavior policy that minimizes the MSE of the importance sampling estimator. As importance sampling is unbiased, minimizing the MSE is equivalent to minimizing variance. We introduce an analytic expression for the gradient of the MSE of the importance sampling estimator and a stochastic gradient descent algorithm that adapts $\pi_{\boldsymbol{\theta}}$ to minimize the MSE between the importance sampling estimate and $v(\pi_e)$. Our algorithm – *behavior policy gradient on the variance* (BPG-V) – begins with on-policy estimates (sets $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_e$) and adapts the behavior policy with gradient descent on the MSE with respect to $\boldsymbol{\theta}$. The gradient of the MSE is given by the following theorem:

Theorem 1 (Behavior Policy Gradient of the Variance)

$$\frac{\partial}{\partial \boldsymbol{\theta}} \operatorname{MSE}\Big[\operatorname{IS}(\pi_e, H, \pi_{\boldsymbol{\theta}})\Big] = \mathbf{E}\left[-\operatorname{IS}(\pi_e, H, \pi_{\boldsymbol{\theta}})^2 \sum_{t=0}^{l-1} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_t|S_t) \middle| H \sim \pi_{\boldsymbol{\theta}}\right]$$

Proof See Appendix B for full proof. The proof of Theorem 1 relies on the fact that the MSE of an estimator is the sum of its variance and the square of its bias. Since importance sampling is unbiased, its MSE is equal to its variance. Thus, the gradient of the MSE given by Theorem 1 is also the gradient of the variance which can be estimated without knowledge of $v(\pi_e)$. Importantly, this gradient can be estimated with trajectories sampled from $\pi_{\boldsymbol{\theta}}$, even though the MSE is defined using $v(\pi_e)$. ■

BPG-V uses stochastic gradient descent in place of exact gradient descent: replacing the expectation in Theorem 1 with an unbiased estimate. While in theory, the single trajectory H_i is sufficient for an unbiased estimate of this gradient, in practice, we can obtain a more accurate descent direction by sampling a batch, B_i, of k trajectories with $\pi_{\boldsymbol{\theta}_i}$. 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.[4]

Full details of BPG-V are given in Algorithm 1. At iteration i, BPG-V samples a batch, B_i, of k trajectories with $\pi_{\boldsymbol{\theta}_i}$ and adds $\{(H_{i\cdot k+j}, \pi_{\boldsymbol{\theta}_i})_{j=1}^{k}\}$ to D_{i-1} to yield data set D_i (Lines 4 – 5). Then BPG-V updates $\boldsymbol{\theta}_i$ with an empirical estimate of the expectation in Theorem 1

4. In principle, we could also re-use trajectories from earlier iterations in our gradient estimate after applying a second importance sampling correction. Informal experiments on a Gridworld domain showed some benefit (i.e., faster variance reduction) from including trajectories from recent batches but an increase in variance when including trajectories from older batches.

(Line 6). After n iterations, BPG-V returns an estimate of $v(\pi_e)$ (Line 8) given as:

$$\overline{\text{IS}}(\pi_e, D_n) = \frac{1}{nk} \sum_{i=1}^{n} \sum_{j=1}^{k} \text{IS}(\pi_e, H_{i \cdot k+j}, \pi_{\boldsymbol{\theta}_i}).$$

As a behavior policy search algorithm, these BPG-V estimates inherit all the results shown in Section 4.2: unbiasedness, consistency, finite-sample rates, and independence between $\text{IS}(\pi_e, H_i, \pi_{\boldsymbol{\theta}_i})$ and $\text{IS}(\pi_e, H_j, \pi_{\boldsymbol{\theta}_j})$ for any two iterations i and j.

Algorithm 1 Behavior Policy Gradient on the Variance
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 $v(\pi_e)$ using all sampled trajectories.

1: $\boldsymbol{\theta}_0 \leftarrow \boldsymbol{\theta}_e$
2: $D_0 = \{\}$
3: **for all** $i \in 0...n$ **do**
4: $\quad B_i =$ Sample k trajectories $H \sim \pi_{\boldsymbol{\theta}_i}$
5: $\quad D_{i+1} = D_i \cup B_i$
6: $\quad \boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i + \frac{\alpha_i}{k} \sum_{j=1}^{k} \text{IS}(\pi_e, H_j, \pi_{\boldsymbol{\theta}_i})^2 \sum_{t=0}^{l-1} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}_i}(A_t^j | S_t^j)$
7: **end for**
8: **Return** $\boldsymbol{\theta}_n$, $\overline{\text{IS}}(\pi_e, D_n)$

Since BPG-V requires collecting trajectories to estimate the variance-gradient, a natural question is whether this gradient can be estimated more efficiently than $v(\pi_e)$. The key insight is that we do *not* require perfect gradient estimation; the gradient only must be estimated well enough to provide a reliable descent direction. Thus we can improve the behavior policy with lower accuracy gradient estimates to obtain a more accurate policy value estimate.

Convergence of BPG-V

We now discuss the theoretical convergence of the BPG-V algorithm. We make the following assumption on the step-size parameter, α_i, at each iteration:

Assumption 3 *The step-size α_i is chosen such that:*

$$\sum_{i=0}^{\infty} \alpha_i = \infty \qquad\qquad \sum_{i=0}^{\infty} \alpha_i^2 < \infty.$$

This assumption is also known as the Robbins and Monroe condition (Robbins and Monro, 1951) and is widely used in convergence results in stochastic approximation.

Proposition 4 *Under Assumption* 1 *and Assumption* 3*, BPG-V converges. That is,* $\text{MSE}[\text{IS}(\pi_e, H_i, \pi_{\boldsymbol{\theta}_i})]$ *converges to a finite value and* $\lim_{i \to \infty} \frac{\partial}{\partial \boldsymbol{\theta}} \text{MSE}[\text{IS}(\pi_e, H_i, \pi_{\boldsymbol{\theta}_i})] = 0$.

Proof See Appendix C for a full proof. The result is an application of Proposition 3 in (Bertsekas and Tsitsiklis, 2000). In Appendix C we show that the MSE objective satisfies the assumptions needed to apply this result. ■

With further assumptions on the policy class of $\pi_{\boldsymbol{\theta}}$ we can derive stronger convergence guarantees. In particular, if $\boldsymbol{\theta}$ is the parameters of a linear-softmax policy than the MSE objective is convex with respect to $\boldsymbol{\theta}$ and local minima of the MSE are also global minima. A linear-softmax policy is a policy over a finite set of actions where the probability of each action is defined as a softmax distribution with logits from a linear combination of state features. Formally, let $\phi : \mathcal{S} \to \mathbf{R}^q$ for integer q be a state feature function that maps states to feature vectors. For each action, $a \in \mathcal{A}$, we have a vector $\boldsymbol{\theta}_a \in \mathbf{R}^q$ and $\boldsymbol{\theta}$ is the concatenation of all $\boldsymbol{\theta}_a$. A linear-softmax policy defines the probability of action a in state s as:

$$\pi_{\boldsymbol{\theta}}(a|s) = \frac{e^{\boldsymbol{\theta}_a^T \phi(s)}}{\sum_{b \in \mathcal{A}} e^{\boldsymbol{\theta}_b^T \phi(s)}}.$$

Theorem 2 *Assume $\pi_{\boldsymbol{\theta}}$ is a linear-softmax policy. Then,* $\text{MSE}[\text{IS}(\pi_e, H, \boldsymbol{\theta})]$ *is a convex function w.r.t. $\boldsymbol{\theta}$.*

Proof See Appendix D. ■

Remark 4 *The result that the MSE of the importance-sampled return is a convex function of $\boldsymbol{\theta}$ is somewhat surprising given that the mean return is a non-convex function under the same assumption of linear-softmax policies (Agarwal et al., 2019).*

Proposition 4 and Theorem 2 imply that BPG-V converges to the globally minimal variance behavior policy in the family of linear-softmax policies Zinkevich (2003). Since we have assumed that π_e belongs to the same parameterized family of policies that we optimize over, BPG-V converges to a behavior policy that will have no higher variance than π_e. In addition to having lower variance, the estimate remains unbiased by Proposition 1, consistent by Proposition 2, and has finite-sample error given by Proposition 3.

6. Behavior Policy Gradient on the KL-Divergence

The preceding section derived an algorithm that searches for a lower variance behavior policy by incrementally decreasing the variance with stochastic gradient descent. In this section, we explore an alternative approach to finding a behavior policy that minimizes variance. Specifically, we first derive a sufficient condition for a behavior policy to minimize the variance of the importance sampling estimator. We then introduce an algorithm that searches for a behavior policy $\pi_{\boldsymbol{\theta}}$ that comes closest to satisfying this condition.

We first define a *minimal variance behavior policy* and then provide a condition that is sufficient for a behavior policy to be a minimal variance behavior policy.

Definition 1 (Minimal-Variance Behavior Policy) *A minimal-variance behavior policy is a policy, $\pi_b{}^\star$, such that* $\text{Var}[\text{IS}(\pi_e, H, \pi_b{}^\star)] \leq \text{Var}[\text{IS}(\pi_e, H, \pi)]$, $\forall \pi \in \Pi$. *Since the variance is lower bounded by zero, such a policy trivially exists.*

Proposition 5 *Let* $w_\pi(h) \coloneqq \prod_{t=0}^{l-1} \pi(a_t|s_t)$. *Assume* $\exists \tilde{h} \in \mathcal{H}$ *such that* $g(\tilde{h}) \cdot \Pr(H = \tilde{h}|\pi_e) \neq 0$, *i.e., there is non-zero probability that* π_e *generates a trajectory with non-zero return. If* $\exists \pi \in \Pi$ *s.t.*

$$\forall h \in \mathcal{H}, w_\pi(h) = |g(h)| \frac{w_{\pi_e}(h)}{\mathbf{E}\left[|g(H)| \;\middle|\; H \sim \pi_e \right]}.$$

then π *is a minimal-variance behavior policy.*

Proof See Appendix E for a full proof. ■

We now introduce a second algorithm that attempts to find $\pi_{\boldsymbol{\theta}}$ that comes closest to satisfying the condition given in Proposition 5. Note that a policy, $\pi_b{}^\star$, that satisfies this expression will induce the following distribution over trajectories:

$$\Pr(H = h|\pi_b{}^\star) \propto \Pr(H = h|\pi_e) \cdot |g(h)|.$$

Though a Markovian policy $\pi_b{}^\star$ that induces this distribution may *not* necessarily exists within a given parameterized policy class, we can still attempt to find $\pi_{\boldsymbol{\theta}}$ that induces a similar trajectory distribution. Thus, our second algorithm attempts to minimize the Kullback-Leibler (KL) divergence between $\Pr(H = h|\pi_b{}^\star)$ and $\Pr(H = h|\pi_{\boldsymbol{\theta}})$. To do so, we first introduce an analytic expression for the gradient of the KL divergence between these trajectory distributions and then use unbiased estimates of this gradient to perform stochastic gradient descent on the behavior policy parameters. We call this second algorithm *behavior policy gradient on the KL-Divergence* (BPG-KL). While BPG-V minimizes our ultimate objective (MSE), BPG-KL minimizes divergence from a minimal-variance solution, given by Proposition 5. We note that this objective has been used before for adaptive IS outside of RL (Rubinstein and Kroese, 2016).

The gradient of the KL-divergence with respect to the policy parameters is proportional to the expression given by the following theorem:

Theorem 3 (Behavior Policy Gradient of the KL-Divergence)

$$\frac{\partial}{\partial \boldsymbol{\theta}} D_{\text{KL}}(\Pr(H|\pi_b{}^\star) || \Pr(H|\pi_{\boldsymbol{\theta}})) \propto \mathbf{E}\left[-\left| \text{IS}(\pi_e, H, \pi_{\boldsymbol{\theta}}) \right| \sum_{t=0}^{l-1} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_t|S_t) \middle| H \sim \pi_{\boldsymbol{\theta}} \right].$$

Proof See Appendix F for full proof. ■

Theorem 3 gives a similar gradient to that in Theorem 1, except it takes the absolute value of $\text{IS}(\pi_e, H, \pi_{\boldsymbol{\theta}})$ instead of squaring it. Like BPG-V, BPG-KL begins with on-policy estimates and adapts the behavior policy with gradient descent on the KL-divergence with

respect to $\boldsymbol{\theta}$. Pseudo-code for the BPG-KL algorithm is given in Algorithm 2. The only difference between BPG-V and BPG-KL is the method of adapting the behavior policy (Line 6); both algorithms still use importance sampling as the underlying off-policy estimator to return estimates of $v(\pi_e)$. As a behavior policy search algorithm, BPG-KL inherits the unbiasedness, consistency, and finite-sample rates given by Proposition 1, Proposition 2, and Proposition 3 respectively.

Algorithm 2 Behavior Policy Gradient on the KL-Divergence
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 $v(\pi_e)$ using all sampled trajectories.

1: $\boldsymbol{\theta}_0 \leftarrow \boldsymbol{\theta}_e$
2: $D_0 = \{\}$
3: **for all** $i \in 0...n$ **do**
4: $\quad B_i =$ Sample k trajectories $H \sim \pi_{\boldsymbol{\theta}_i}$
5: $\quad D_{i+1} = D_i \cup B_i$
6: $\quad \boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i + \frac{\alpha_i}{k} \sum_{j=1}^{k} |\operatorname{IS}(\pi_e, H_j, \pi_{\boldsymbol{\theta}i})| \sum_{t=0}^{l-1} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}_i}(A_t^j | S_t^j)$
7: **end for**
8: **Return** $\boldsymbol{\theta}_n$, $\overline{\operatorname{IS}}(\pi_e, D_n)$

Convergence of BPG-KL

Like BPG-V, we can show that BPG-KL converges and that, under a linear-softmax policy assumption, the objective optimized by BPG-KL is convex.

Proposition 6 *Under Assumption* 1 *and Assumption* 3*, BPG-KL converges. That is,* $D_{\mathtt{KL}}(\Pr(H|\pi_b{}^\star) || \Pr(H|\pi_{\boldsymbol{\theta}}))$ *converges to a finite value and* $\lim_{i\to\infty} \frac{\partial}{\partial \boldsymbol{\theta}} D_{\mathtt{KL}}(\Pr(H|\pi_b{}^\star) || \Pr(H|\pi_{\boldsymbol{\theta}})) = 0$.

Proof See Appendix G for a full proof. The result is an application of Proposition 3 in (Bertsekas and Tsitsiklis, 2000). In Appendix G we show that the KL-objective satisfies the assumptions needed to apply this result. ■

Additionally, we can show convexity of the KL-objective under an assumption of linear-softmax policies.

Theorem 4 *Assume* $\pi_{\boldsymbol{\theta}}$ *is a linear-softmax policy. Then,* $D_{\mathtt{KL}}(\Pr(H|\pi_b{}^\star) || \Pr(H|\pi_{\boldsymbol{\theta}}))$ *is a convex function w.r.t.* $\boldsymbol{\theta}$*.*

Proof See Appendix H. ■

Proposition 6 and Theorem 4 jointly imply convergence to a global minimum (Zinkevich, 2003). A counterintuitive observation is that global minimization of the KL-objective does

not necessarily imply that BPG-KL converges to lower variance importance-sampled returns compared to on-policy sampling. First, observe that, since we minimize the KL between the minimal-variance behavior policy and a policy within a specific family of behavior policies, we may not converge to a minimal-variance behavior policy (which may be unrepresentable in the family of linear soft-max policies). While the policy at BPG-KL's convergence would be closer *in terms of KL* to a minimal-variance behavior policy than any other linear soft-max policy, we have not ruled out the possibility that the policy would yield sub-optimal variance for the importance sampling returns. While this case may be theoretically possible, our experimental results (in Section 9) show that BPG-V and BPG-KL perform similarly in practice, suggesting that minimizing the KL also minimizes variance compared to on-policy sampling in practice.

7. Interpreting BPG-V and BPG-KL Updates

We can gain intuition for how BPG-V and BPG-KL update the behavior policy by comparing their updates to existing algorithms in policy gradient RL (cf. Sutton et al. (2000)). Here, we draw a connection between one such family of algorithms and our new behavior policy search methods to illustrate how these methods change the distribution of trajectories. The REINFORCE family of algorithms (Williams, 1992) attempts to maximize $v(\pi_{\boldsymbol{\theta}})$ through gradient ascent on $v(\pi_{\boldsymbol{\theta}})$ using unbiased estimates of the gradient of $v(\pi_{\boldsymbol{\theta}})$:

$$\frac{\partial}{\partial \boldsymbol{\theta}} v(\pi_{\boldsymbol{\theta}}) = \mathbf{E}\left[g(H) \sum_{t=0}^{l-1} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_t|S_t) \middle| H \sim \pi_{\boldsymbol{\theta}} \right].$$

Intuitively, REINFORCE methods increase 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-V can be interpreted as a REINFORCE method where the return of a trajectory is the square of its importance-sampled return. Thus BPG-V increases the probability of all actions taken along H as a function of $\text{IS}(\pi_e, H, \boldsymbol{\theta})^2$. BPG-KL can be interpreted as a REINFORCE method where the return of a trajectory is the absolute value of its importance-sampled return. Thus BPG-KL increases the probability of all actions taken along H as a function of $|\text{IS}(\pi_e, H, \boldsymbol{\theta})|$. Recall that $\text{IS}(\pi_e, H, \boldsymbol{\theta}) = g(H) \prod_{t=0}^{l-1} \frac{\pi_e(A_t|S_t)}{\pi_{\boldsymbol{\theta}}(A_t|S_t)}$. Thus, the magnitude of both $\text{IS}(\pi_e, H, \boldsymbol{\theta})^2$ and $|\text{IS}(\pi_e, H, \boldsymbol{\theta})|$ depends on two qualities of H:

1. The magnitude of $g(H)$ (whether positive or negative).
2. The relative likelihood of H under π_e compared to $\pi_{\boldsymbol{\theta}}$ (i.e., $\prod_{t=0}^{l-1} \frac{\pi_e(A_t|S_t)}{\pi_{\boldsymbol{\theta}}(A_t|S_t)}$).

These two qualities demonstrate a balance in how BPG-V and BPG-KL change trajectory probabilities. Increasing the probability of a trajectory under $\pi_{\boldsymbol{\theta}}$ will decrease the magnitude of $\text{IS}(\pi_e, H, \boldsymbol{\theta})$ and so BPG-V and BPG-KL increase the probability of a trajectory when the magnitude of $g(H)$ is large enough to offset the decrease in the magnitude of $\text{IS}(\pi_e, H, \boldsymbol{\theta})$ caused by decreasing the importance weight.

The main difference between the two algorithms is that BPG-V puts more emphasis on increasing the probability of high magnitude return trajectories. For example if one return has double the return of another then it has quadruple the emphasis under BPG-V whereas

with BPG-KL doubling the return only doubles the emphasis. BPG-V is minimizing our target objective (low MSE) while BPG-KL attempts to find a policy that is close (in terms of KL-divergence) to the optimal solution to our target objective.

8. Behavior Policy Search for Importance Sampling Extensions

The behavior policy search algorithms introduced in Sections 5 and 6 both use the basic importance sampling estimator for estimating $v(\pi_e)$. In this section we introduce behavior policy search algorithms that use other unbiased off-policy estimators: doubly robust and per-decision estimators. We also discuss behavior policy search for *weighted* importance sampling.

8.1 Baselined Importance Sampling

Instead of using importance sampling to evaluate $v(\pi_e)$, we can instead estimate

$$\mathbf{E}\left[g(H) - b \;\middle|\; H \sim \pi_e\right] + b \tag{3}$$

for some constant b. With a constant baseline, the baselined importance sampling estimate of $v(\pi_e)$ after n iterations becomes:

$$\overline{\text{IS}}(\pi_e, D_i, b) \coloneqq b + \frac{1}{n}\sum_{j=1}^{n}\prod_{t=0}^{l-1}\frac{\pi_e(A_t^j|S_t^j)}{\pi_{\boldsymbol{\theta}_i}(A_t^j|S_t^j)}(g(H_j) - b).$$

While the on-policy Monte Carlo estimate of (3) is identical to the Monte Carlo estimate of $v(\pi_e)$, an off-policy importance sampling estimate benefits from a baseline if b is closer to $v(\pi_e)$ than $v(\pi_e)$ is to 0. The lower variance is due to $b \cdot \prod_{t=0}^{l-1}\frac{\pi_e(A_t^j|S_t^j)}{\pi_{\boldsymbol{\theta}_i}(A_t^j|S_t^j)}$ serving as a control variate for the importance sampled $g(H)$ (Thomas and Brunskill, 2017). BPG-V and BPG-KL only require a small modification to use a constant baseline: we replace all occurrences of $g(H)$ with $g(H) - b$ in the algorithms and then add b to the final estimate returned.

8.2 Doubly Robust and Per-Decision Importance Sampling

In cases where an approximate model of the environment is available, the *doubly robust* (DR) estimator (Jiang and Li, 2016; Thomas and Brunskill, 2016) lowers the variance of importance sampling using the control variate technique (Lemieux, 2014). In this section, we introduce a behavior policy search algorithm that uses the DR estimator for estimates of $v(\pi_e)$.

The DR estimator computes the average difference between the observed importance-sampled rewards and the predicted expected reward under a model of the environment's transition and reward function. Provided the expected reward predictions are correlated with the true rewards, DR has lower variance than using the importance-sampled rewards alone. The DR estimate for a single trajectory, H, is given by:

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

where $w_{\pi,t}(H) = \prod_{i=0}^{t} \pi(A_t|S_t)$ and $\hat{v}^{\pi_e}$ and $\hat{q}^{\pi_e}$ be the state and action value functions of π_e in the approximate model.

We show here that we can adapt the behavior policy to lower the MSE of DR estimates. As of this writing, it is an open problem whether there exists a form for a minimal-variance behavior policy for DR. Therefore we only introduce a method that adapts the behavior policy from the perspective of minimizing variance. We denote this new method DR-BPG for *doubly robust behavior policy gradient*.

The MSE gradient for the DR estimator is given by the following corollary to Theorem 1:

Corollary 1

$$\frac{\partial}{\partial \boldsymbol{\theta}} \operatorname{MSE}\Big[DR(\pi_e, H, \pi_{\boldsymbol{\theta}}, \hat{q}^{\pi_e}, \hat{v}^{\pi_e})\Big] = \mathbf{E}\Bigg[\operatorname{DR}(\pi_e, H, \boldsymbol{\theta}, \hat{q}^{\pi_e}, \hat{v}^{\pi_e})^2 \sum_{t=0}^{l-1} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_t|S_t) \\ - 2 \operatorname{DR}(\pi_e, H, \pi_{\boldsymbol{\theta}}, \hat{q}^{\pi_e}, \hat{v}^{\pi_e}) \Bigg(\sum_{t=0}^{l-1} \gamma^t \delta_t \frac{w_{\pi_e,t}}{w_{\boldsymbol{\theta},t}} \sum_{i=0}^{t} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_i|S_i) \Bigg) \Bigg]$$

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

Proof See Appendix B.3 for the full proof. ■

The first term of $\frac{\partial}{\partial \boldsymbol{\theta}} \operatorname{MSE}$ is analogous to the gradient of the importance-sampling estimate with $\operatorname{IS}(\pi_e, H, \boldsymbol{\theta})$ replaced by $\operatorname{DR}(\pi_e, H, \boldsymbol{\theta}, \hat{q}^{\pi_e}, \hat{v}^{\pi_e})$. The second term accounts for the covariance of the DR terms over time.

In practice, DR has been noted to perform best when all available trajectories are used to estimate the approximate model and then also used to estimate $v(\pi_e)$ (Thomas and Brunskill, 2016). However, for DR-BPG, updating the model as $\pi_{\boldsymbol{\theta}}$ is learned will change the the surface of the MSE objective we seek to minimize and thus DR-BPG will only converge once the model stops changing. Computing the model from the same data used in the DR estimate also violates assumptions made for the theoretical analysis of DR (Thomas and Brunskill, 2016). In our experiments, we consider both a changing and a fixed model.

Finally, as a special case of Corollary 1, we obtain the variance gradient for the per-decision importance sampling estimator (Precup et al., 2000).

Corollary 2

$$\frac{\partial}{\partial \boldsymbol{\theta}} \operatorname{MSE}\Big[\operatorname{PDIS}(\pi_e, H, \pi_{\boldsymbol{\theta}})\Big] = \mathbf{E}\Bigg[\operatorname{PDIS}(\pi_e, H, \boldsymbol{\theta})^2 \sum_{t=0}^{l-1} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_t|S_t) \\ - 2 \operatorname{PDIS}(\pi_e, H, \pi_{\boldsymbol{\theta}}) \Bigg(\sum_{t=0}^{l-1} \gamma^t R_t \frac{w_{\pi_e,t}}{w_{\boldsymbol{\theta},t}} \sum_{i=0}^{t} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_i|S_i) \Bigg) \Bigg]$$

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

Proof Set $\hat{q}^{\pi_e}$ and v^{π_e} to 0 for all states, actions, and time-steps and the DR estimator reduces to the per-decision estimator and then Corollary 2 follows from 1. ■

8.3 Weighted Importance Sampling

Another common variance reduction technique for importance sampling is to use weighted (also known as self-normalized) importance sampling (Precup et al., 2000; Swaminathan and Joachims, 2015). The weighted importance sampling estimator for a set of m trajectory-behavior-policy pairs is defined as:

$$\overline{\text{WIS}}(\pi_e, D) \coloneqq \frac{1}{Z} \sum_{j=1}^{m} \text{IS}(\pi_e, H_j, \pi_j),$$

where the normalization factor, $Z = \sum_{j=1}^{m} \frac{w_{\pi_e}(H_j)}{w_{\pi_j}(H_j)}$, is the sum of all importance weights.

For finite sample sizes, weighted importance sampling is a biased estimator, however, it lowers variance due to the importance weights themselves. Though often noted to lower variance compared to the basic importance sampling estimator (Thomas et al., 2015b; Mahmood et al., 2014), if the behavior policy is optimized for basic importance sampling, then it may *harm* the efficiency of policy evaluation. We illustrate this fact with an example. Consider a two-armed bandit problem in which the policy selects arm 1 with probability θ and arm 2 with probability $1 - \theta$. Let the outcome of pulling arm 1 be a reward of 100 and the outcome of arm 2 be a reward of 1. The evaluation policy is defined as $\theta_e \coloneqq 0.1$. The minimal-variance behavior policy for the basic importance sampling estimator (computed with Equation (2)) is $\theta_b^\star \approx 0.917$.

Figure 4 shows the MSE of weighted importance sampling compared to the basic importance sampling estimator for different values of θ. Estimates are computed with data sets of size 50 and the squared error is averaged over 500 different data sets. For values of θ greater than 0.5, the MSE of weighted importance sampling increases even while the MSE of the basic importance sampling estimator continues to decrease. This example illustrates that weighted importance sampling can harm the accuracy of policy evaluation estimates when using a behavior policy chosen to lower the variance of the basic importance sampling estimator. Since we focus on unbiased policy evaluation estimators, we leave how to best determine the behavior policy for a weighted importance sampling estimate as an open question.

9. 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-V and BPG-KL reduce the MSE of batch policy evaluation 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 MSE compared to the on-policy doubly robust estimator?
- Does the rareness of actions that cause high magnitude rewards affect the performance gap between BPG-V and Monte Carlo estimates?

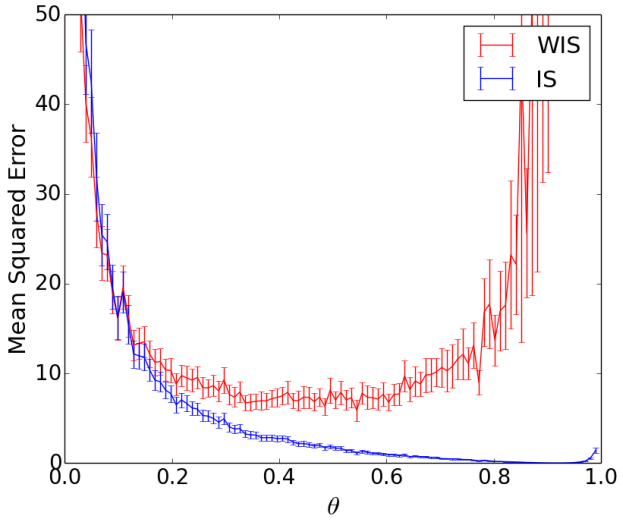


Figure 4: An example where optimizing the behavior policy for the MSE of the basic IS estimator increases the MSE of WIS. For 100 values of θ evenly spaced between 0.01 and 0.99, a data set of size 50 is collected and both the IS estimate and WIS estimate are computed and the squared error calculated. The process is repeated 500 times and the mean squared error reported with 95% confidence intervals shown. The horizontal axis gives the parameter value and the vertical axis gives mean squared error.

9.1 Empirical Set-up

We address our first experimental question by evaluating BPG-V and BPG-KL on several policy evaluation tasks.

Grid World The first domain is the Grid World domain showed in Figure 5. All grid locations without a reward shown have a reward of -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 $l = 100$. Each state-action pair, (s, a), has a parameter $\theta_{s,a}$ and the probability of taking action a in state s is given by the softmax distribution:

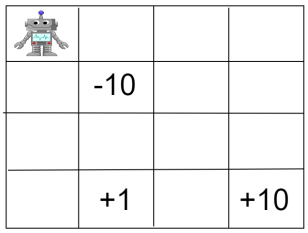


Figure 5: Grid World Domain

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

In this domain it is unnecessary to represent the policy with function approximation and we can study BPG-V and BPG-KL without concern of whether our class of function approximator includes a lower variance behavior policy. We obtain two evaluation policies by applying a simple REINFORCE algorithm to maximize the expected return, 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 $\pi_e \coloneqq \pi_1$ and a second time with $\pi_e \coloneqq \pi_2$. The ground truth value of $v(\pi_e)$ is computed with value iteration for both choices of π_e.

Control Tasks We also study BPG-V and BPG-KL on four tasks with real-valued state variables. The first two of these are the continuous control Cart Pole Swing Up and Acrobot tasks implemented within RLLAB (Duan et al., 2016), the third task is the Cart Pole task from OpenAI Gym (Brockman et al., 2016), and the final task is the PyBullet (Coumans and Bai, 2016–2019) variant of the Hopper domain from OpenAI gym (Brockman et al., 2016). In contrast to the tabular Grid World domain, these domains require that BPG-V and BPG-KL optimize the behavior policy within a given class of function approximator. For Cart Pole Swing Up and Acrobot, π_e is a two layer neural network with 32 tanh units per layer that maps the state to the mean of a Gaussian distribution over the continuous action space. For Cart Pole Swing Up, π_e was learned using 10 iterations of the TRPO algorithm (Schulman et al., 2015) applied to a randomly initialized policy. For Acrobot, π_e was learned using 60 iterations. For Cart Pole and Hopper, π_e is a neural network with two layers of 64 tanh hidden units in each layer and is trained using 200 iterations of proximal policy optimization (Schulman et al., 2017). For Cart Pole the network maps the state to a softmax distribution over actions while in Hopper the network maps the state to a Gaussian distribution over the continuous-valued actions. For Cart Pole Swing Up and Acrobot we use $l = 50$ and $\gamma = 1$; CartPole and Hopper use $l = 200$ (with early termination possible) and $\gamma = 1$. For step-size selection at each iteration BPG-V and BPG-KL use the largest possible step-size subject to a constraint on the KL-divergence between the old and new policy. This type of update has been shown to be more stable than constant step-size updates in the policy gradient RL literature (Kakade, 2001; Peters and Schaal, 2008; Schulman et al., 2015). The ground truth value of $v(\pi_e)$ in all domains is computed with 1,000,000 Monte Carlo roll-outs.

In all experiments, for both BPG-V and BPG-KL, we use a constant control variate (or baseline) when estimating the gradient. For BPG-V, the baseline, b_i, is an estimate of:

$$\mathbf{E}\left[-\operatorname{IS}(\pi_e, H, \pi_{\boldsymbol{\theta}_{i-1}})^2 \;\middle|\; H \sim \pi_{\boldsymbol{\theta}_{i-1}}\right]$$

and for BPG-KL, the baseline, b_i, is an estimate of

$$\mathbf{E}\left[-|\operatorname{IS}(\pi_e, H, \pi_{\boldsymbol{\theta}_{i-1}})| \;\middle|\; H \sim \pi_{\boldsymbol{\theta}_{i-1}}\right].$$

The baseline b_i is estimated with trajectories from iteration $i-1$ where for the first iteration $b_i = 0$. The gradient with baseline for BPG-V is an estimate of:

$$\mathbf{E}\left[\left(-\operatorname{IS}(\pi_e, H, \pi_{\boldsymbol{\theta}})^2 - b_i\right)\sum_{t=0}^{l-1}\frac{\partial}{\partial \boldsymbol{\theta}}\log \pi_{\boldsymbol{\theta}}(A_t|S_t)\middle| H \sim \pi_{\boldsymbol{\theta}}\right]$$

and the gradient with baseline for BPG-KL is an estimate of:

$$\mathbf{E}\left[\left(-|\operatorname{IS}(\pi_e, H, \pi_{\boldsymbol{\theta}})| - b_i\right)\sum_{t=0}^{l-1}\frac{\partial}{\partial \boldsymbol{\theta}}\log \pi_{\boldsymbol{\theta}}(A_t|S_t)\middle| H \sim \pi_{\boldsymbol{\theta}}\right].$$

Adding or subtracting a constant leaves the gradient unchanged in expectation since $b_i \mathbf{E}\left[\sum_{t=0}^{l-1} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_t|S_t)\right] = 0$. However, the baseline variants of BPG-V and BPG-KL have lower variance gradient estimates so that the estimated gradient is closer in direction to the true gradient. Note that this baseline is for gradient estimation and is different than using a constant baseline for importance sampling.

In all domains we run multiple trials where each trial consists of a fixed number of iterations. At each iteration, each algorithm collects a batch of trajectories and computes a new estimate of $v(\pi_e)$. We use batch sizes of 100 trajectories per iteration for Grid World experiments and size 500 for the continuous control tasks. All algorithms have access to the same number of trajectories at the same iteration across trials.

9.2 Main Results

In this section we present our empirical results to address the questions outlined at the beginning of Section 9.

9.2.1 Grid World

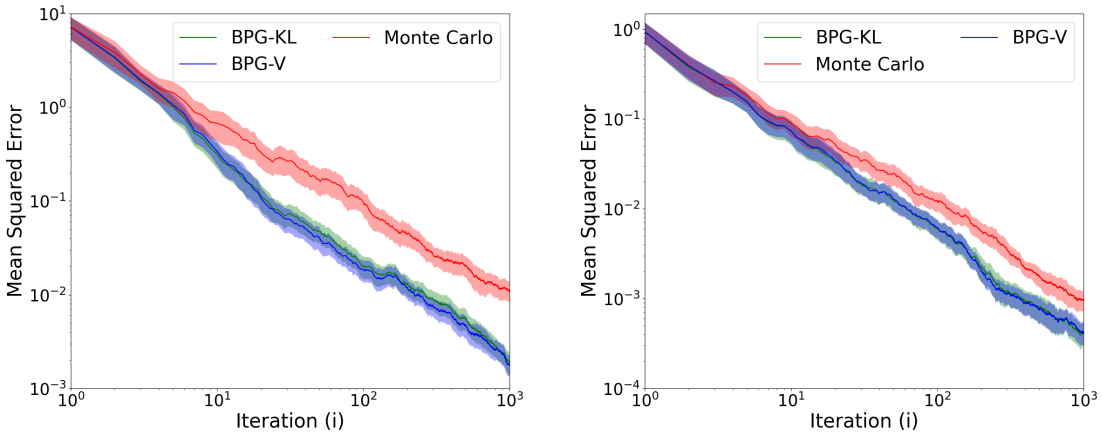
Figure 6 compares BPG-V, BPG-KL, and the on-policy Monte Carlo estimator for both Grid World policies, π_1 and π_2. At each iteration, each method collects 100 additional trajectories. BPG-V gradient estimates will tend to have a different magnitude than BPG-KL gradient estimates because the importance-sampled return is squared instead of its absolute value taken. We normalize the gradient estimates to have magnitude one and use a step-size of 0.1 for both methods in order to have similar magnitude behavior policy changes for each method.

Our main point of comparison is the MSE of both estimates at iteration i over 100 trials. For π_1, BPG-V and BPG-KL reduce the MSE of on-policy estimates (Figure 6a) by up to an order of magnitude. For π_2, BPG-V and BPG-KL also reduce MSE, however, it is a more marginal improvement. In both cases, BPG-V and BPG-KL perform almost identically.

At the end of each trial we used the final behavior policy to collect 100 more trajectories and estimate $v(\pi_e)$. For BPG-V, in comparison to a Monte Carlo estimate with 100 trajectories from π_1, MSE is 73.52% lower with this improved behavior policy; for π_2, the MSE is 64.6% lower. For BPG-KL and π_1, the MSE is 77.78% lower with the final behavior policy; for π_2, the MSE is 46.28% lower. This result demonstrates that BPG-V and BPG-KL can find behavior policies that substantially lower MSE.

To understand the disparity in performance when π_e changes, we plot the variance of the Monte Carlo return under π_e (Figures 7b and 7c). These plots show the variance of π_1 is much higher; it sometimes samples returns with twice the magnitude of any sampled by π_2. To quantify the decrease in variance from behavior policy search, we also measure and plot the variance of $\text{IS}(\pi_e, H, \pi_{\boldsymbol{\theta}_i})$ for the BPG-V algorithm (Figure 7a). Figure 7a shows much higher initial variance for importance sampling evaluation of π_1. The high initial variance means there is much more room for BPG-V and BPG-KL to improve the behavior policy when $\boldsymbol{\theta}_e$ is the partially optimized policy, π_1.

BPG-V and BPG-KL require setting two parameters for the stochastic gradient descent update: a step-size, α, and a batch-size, k. We ablate these parameters to test the sensitivity



(a) Results with $\pi_e \coloneqq \pi_1$

(b) Results with $\pi_e \coloneqq \pi_2$

Figure 6: Grid World experiments when π_e is a partially optimized policy, π_1, (6a) and a converged policy, π_2, (6b). Results are averaged over 100 trials of 1000 iterations with a shaded region representing a 95% confidence interval. The vertical axis shows the mean squared error and the horizontal axis shows the iteration number. Axes are log-scaled. In both instances, BPG-V and BPG-KL lower MSE more than on-policy Monte Carlo returns (statistically significant, $p < 0.05$).

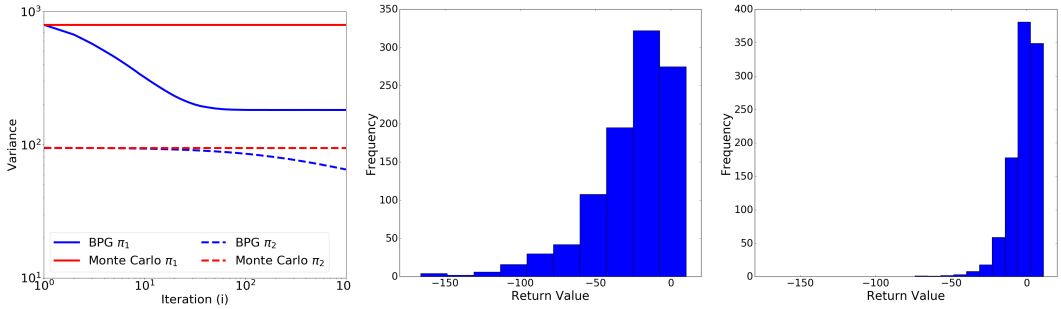
of performance to their values. Again, we use normalized gradient estimates to ensure comparability of the algorithms given the same step-size.

To ablate step-size, we run each algorithm for 1000 iterations with a batch-size of $k = 100$ for different settings of α. Our point of comparison is the MSE of the estimate at the final iteration. Figure 8a shows that both BPG-V and BPG-KL perform as well as or better than Monte Carlo for a wide range of step-size values. However, for very high values ($\alpha = 5$ or $\alpha = 10$), the estimates may diverge.

To ablate batch-size, we run each algorithm until it has collected 1000 trajectories with different settings of k. So a trial using $k = 500$ will collect 500 trajectories, adapt the behavior policy once, and then collect 500 more trajectories to compute the final estimate. Both algorithms use a step-size of 0.1. As with step-size, we see that both algorithms perform as well as or better than Monte Carlo evaluation for most batch-size settings. With the smallest tested batch ($k = 1$), BPG-V and BPG-KL perform worse, presumably because the gradient estimates are poor and so the algorithms fail to improve the behavior policy.

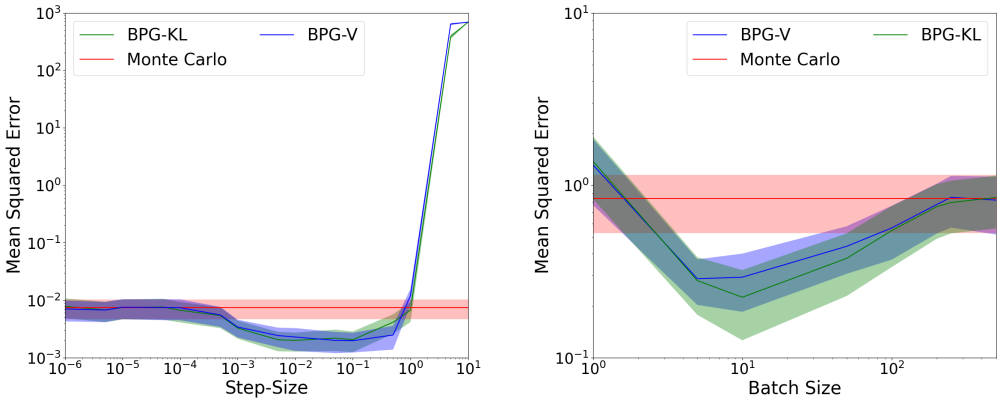
9.2.2 Control Tasks

Figure 9 shows reduction of MSE on the Cart Pole Swing Up Acrobot, Cart Pole, and Hopper domains. Each method uses a step-size of 5×10^{-5}. Again we see that both BPG-V and BPG-KL reduce MSE faster than Monte Carlo value estimation and that both methods perform similarly to one another. In contrast to the discrete Grid World experiment, these experiment demonstrates the applicability of BPG-V and BPG-KL to both continuous states



(a) Variance Reduction (b) π_1 return distribution (c) π_2 return distribution

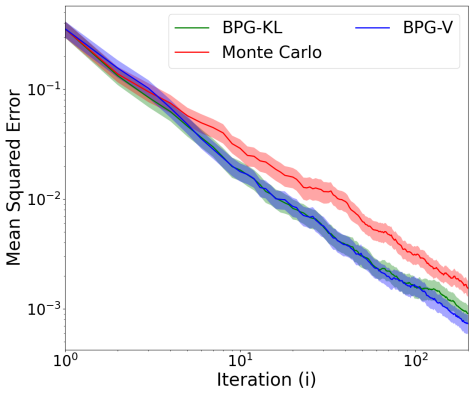
Figure 7: Comparison of variance reduction between π_1 and π_2 in Grid World domain. Figure 7a shows variance on the vertical axis and iteration number on the horizontal axis. These axes are log-scaled. Results are plotted for Monte Carlo value estimation with π_1 and π_2 and for BPG-V evaluations of π_1 and π_2. Results are averaged over 100 trials of 1000 iterations. Figures 7b and 7c give the distribution of returns under the two different π_e. Taken together these plots show that the variance of a Monte Carlo evaluation of π_1 is much higher than a Monte Carlo evaluation of π_2. Thus a behavior policy search algorithm has more room for variance reduction when evaluating π_1.



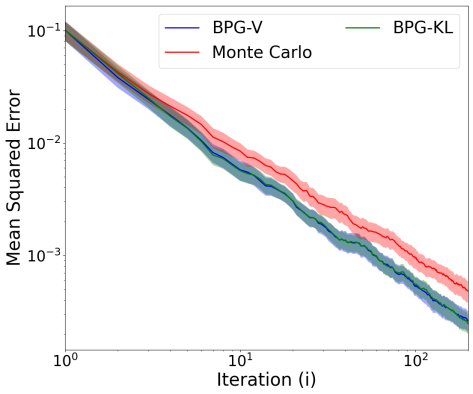
(a) Step-size Sensitivity (b) Batch-size Sensitivity

Figure 8: BPG-V and BPG-KL parameter sensitivity. Figure 8a shows performance as a function of the algorithm step-size, α, and Figure 8b shows performance as a function of the algorithm batch-size, k. In both figures the vertical axis is mean squared error of the importance sampling estimate. The horizontal axis is the parameter being ablated. Axes are log-scaled.

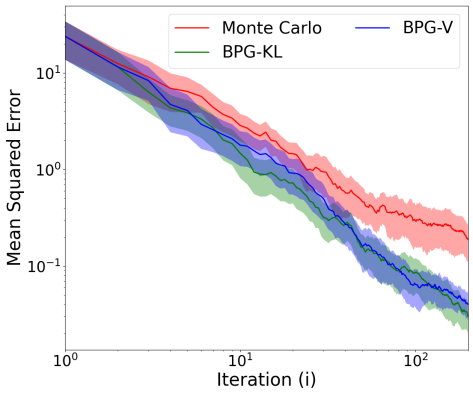
and actions. These results also demonstrates that BPG-V and BPG-KL (and more generally behavior policy search) can lower the variance of batch policy evaluation when the policy must generalize across different states and actions.



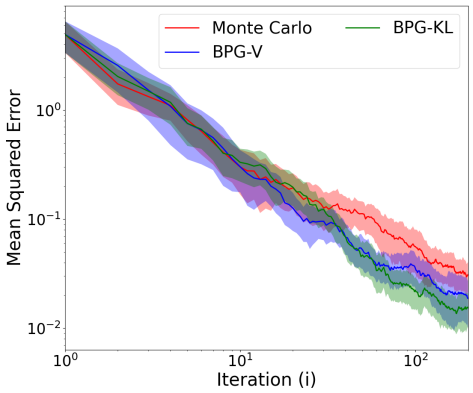
(a) Cart Pole Swing Up MSE.



(b) Acrobot MSE.



(c) Cart Pole MSE.



(d) Hopper MSE.

Figure 9: Mean squared error reduction on the Cart Pole Swing Up Acrobot, Cart Pole, and Hopper domains. The vertical axis gives MSE and the horizontal axis is the iteration number. Axes are log-scaled. We adapt the behavior policy for 200 iterations and average results over 100 trials. Error bars are for 95% confidence intervals.

9.3 Control Variate Extension Results

In this section, we evaluate the combination of model-based control variates with behavior policy search. Specifically, we compare doubly robust BPG-V (DR-BPG) with an on-policy doubly robust estimator that uses $\boldsymbol{\theta}_i = \boldsymbol{\theta}_e$ for all i. We refer to the on-policy doubly robust estimator as the *advantage-sum* estimator (ASE) as it has appeared previously in the literature under this name (Zinkevich et al., 2006; White and Bowling, 2009; Veness et al., 2011).

In these experiments we use a 10x10 stochastic Grid World where the added stochasticity and increased size increase the difficulty of building an accurate model from data. The layout of this Grid World is identical to the deterministic Grid World 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. Stochasticity in the environment 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 D to build the model and then uses the same set to estimate $v(\pi_e)$ 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 under this second condition. We call the first method "Update" and the second method "Fixed." The update method invalidates consistency guarantees of these methods but learns a more accurate model. In both instances, we build the models with count-based estimates of the transition probabilities.

Figure 10 demonstrates that combining BPG-V with a model-based control variate (DR-BPG) can lead to further reduction of MSE compared to either the control variate(ASE) or behavior policy search (BPG) alone. 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%. BPG-V outperforms Monte Carlo but both methods do significantly worse than the methods using a model-based control-variate.

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. Informally, we observed that setting the step-size, α, for DR-BPG was more challenging for either model type. Thus while we have shown BPG-V can be combined with control variates, more work is needed to produce a robust method.

9.4 Rareness of Event Study

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

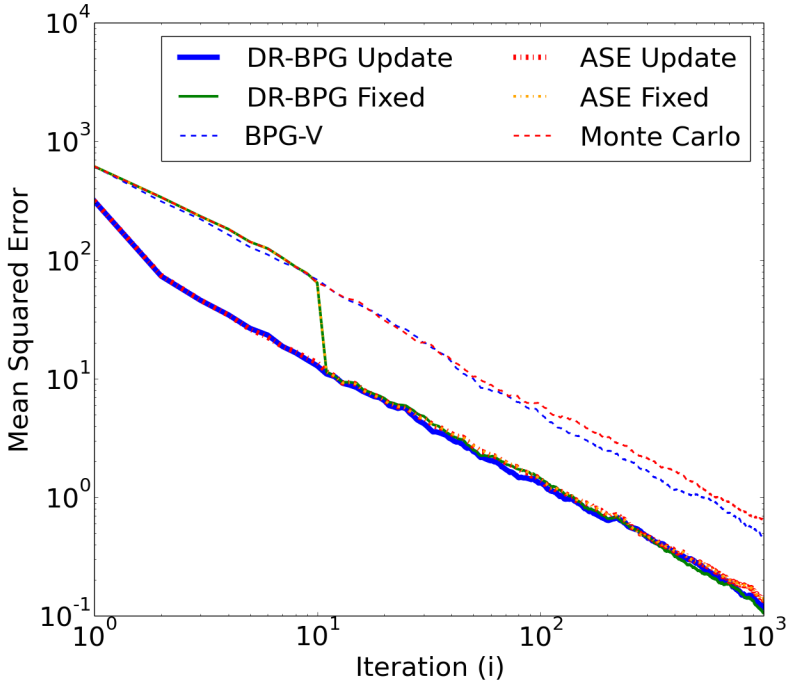


Figure 10: Comparison of DR-BPG and ASE (on-policy DR) on a larger stochastic Grid World. For the fixed model methods, the significant drop in MSE at iteration 10 is due to the introduction of the model control variate. For visual clarity we omit error bars. The mean difference between the final estimate of DR-BPG and ASE with the fixed model averaged over 300 trials is statistically significant ($p < 0.05$); the difference between the same methods with a constantly improving model is not.

between on- and off-policy policy evaluation. For this experiment, we use a variant of the deterministic Grid World 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 Grid World experiments but we vary the probability, p, of choosing UP when in the initial state, i.e., with probability p the agent will receive a large reward and end the trajectory. We use BPG-V with a step-size of 10^{-5} and unnormalized gradient estimates as the behavior policy search algorithm for all values of p. We plot the relative decrease of the variance after 500 iterations 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 remains greater than 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-V can achieve increases. As p becomes closer to 1, the rare high magnitude event becomes less rare and the initial variance becomes closer to zero. When this happens, BPG-V barely improves over the variance of Monte Carlo (in terms of absolute variance there is no improvement). When π_e rarely takes the high rewarding UP action (p close to 0),

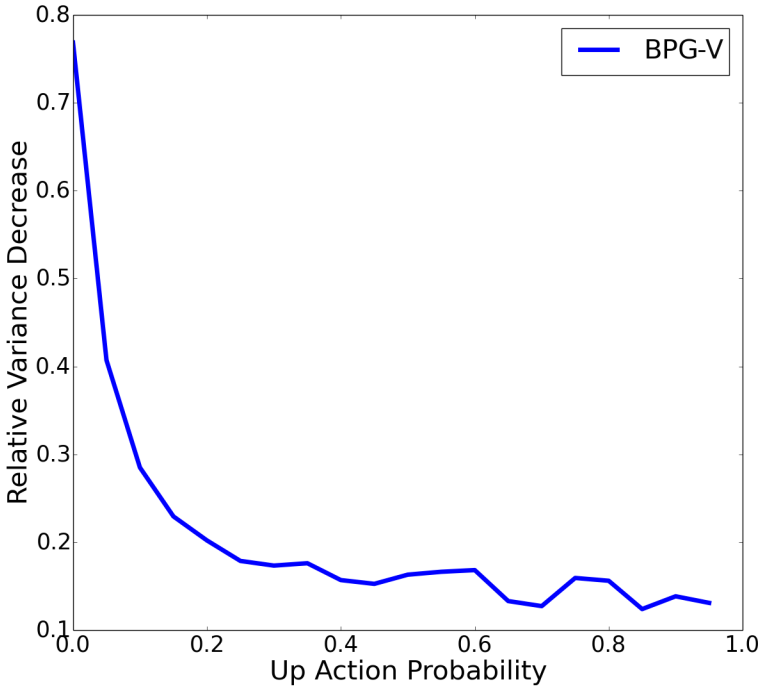


Figure 11: Varying the probability of a high rewarding terminal action in the Grid World domain. Each point on the horizontal axis is the probability of taking this action. The vertical axis gives the mean relative decrease in variance after adapting $\boldsymbol{\theta}$ for 500 iterations. Denoting the initial variance as V_i and the final variance as V_f, the relative decrease is computed as $\frac{V_i - V_f}{V_i}$. Results are averaged over 100 trials. A 95% confidence interval region is shaded around the mean but is small.

BPG-V lowers the variance of policy evaluation by increasing the probability of this action. This experiment supports our intuition for why off-policy data collection can be preferable to on-policy data collection.

10. Discussion

Our experiments demonstrate that behavior policy search with either BPG-V or BPG-KL can lower the variance of batch 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-V or BPG-KL can only offer marginal improvement. BPG-V and BPG-KL increase the probability of observing rare events with a high magnitude. If the evaluation policy never sees such events then there is less benefit to using a behavior policy search algorithm. However, with an appropriately selected step-size, BPG-V and BPG-KL will never, in expectation, 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-V or BPG-KL will offer much improvement. For example, Ciosek and Whiteson (2017) consider a variant of the Mountain Car task (Singh and Sutton, 1996) 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 behavior policy search methods is the necessity of good step-size selection. In expectation, BPG-V and BPG-KL 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 gradients.

When to Perform Behavior Policy Search?

We conclude with a discussion of the question of when should one prefer behavior policy search to just choosing the evaluation policy as the behavior policy. From our experiments with random MDPs in Section 4.1, we find that the most potential improvement is when the evaluation policy is stochastic (but not uniform random) and there is variation in the reward across the action space. This observation dovetails with the intuition that BPS is most useful when there are rare trajectories with high magnitude return under the evaluation policy because such settings are where the variance of on-policy Monte Carlo is highest. This intuition was demonstrated experimentally in Section 9.4. On the other hand, when π_e is deterministic or uniform random there may be little or no room for improvement.

In settings where π_e is already a near optimal behavior policy for itself, the need to set hyper-parameters for BPG-V and BPG-KL may not be worth any additional variance reduction that could be gained through behavior policy search. Both methods lack guarantees that the behavior policy improves at every iteration and if intermediate behavior policies *increase* variance (e.g., due to variance in the behavior policy gradient estimate) then the final estimate may have higher squared error than if π_e had just been ran to collect all trajectories. Thus, we recommend behavior policy search for settings where the variance of the return under π_e is anticipated to be high.

11. Future Work

In this section, we outline directions for future work to further develop the utility of behavior policy search for reinforcement learning. As an overarching direction, we note that this work assumed a finite-horizon, episodic, and fully observable environment. Future work should consider what is the minimal-variance behavior policy and how to perform behavior policy search in infinite-horizon, continuing, or partially observable environments.

11.1 Evaluating Multiple Evaluation Policies

A common motivation for collecting data in an off-policy fashion is when we want to learn about multiple evaluation policies with the same stream of data (e.g., (Sutton et al., 2011)). In this work, we have assumed a single evaluation policy and considered finding a behavior

policy that provides low variance importance sampling evaluation of that evaluation policy. An important direction for future work is to develop behavior policy search algorithms that optimize the behavior policy for a *set* of evaluation policies. A straightforward way to adapt either BPG-KL or BPG-V to multiple evaluation policies is to use a linear combination of the objective they minimize for each evaluation policy. For example, BPG-V minimizes $\mathbf{E}[\text{IS}(\pi_e, H, \pi_{\boldsymbol{\theta}})^2 | H \sim \pi_{\boldsymbol{\theta}}]$ for the single evaluation policy π_e. If instead, we wished to minimize the variance of evaluating a set of policies, $\{\pi_1, ..., \pi_m\}$, a multi-policy variant of BPG-V could minimize $\sum_{j=1}^{m} \mu(\pi_j) \mathbf{E}[\text{IS}(\pi_j, H, \pi_{\boldsymbol{\theta}})^2 | H \sim \pi_{\boldsymbol{\theta}}]$ where, we define $\mu(\pi_j)$ to be an emphasis factor that provides the relative importance of evaluating each policy in the set of evaluation policies. This approach would be straightforward, however, it might be the case that lowering the variance for one evaluation policy might increase the variance of evaluating another.

11.2 Behavior Policy Search for Value Function Learning

This work has focused on batch policy evaluation in which we collect a set of trajectories and estimate $v(\pi_e)$. A more general policy evaluation problem is to estimate the value function: the function that gives the expected return of a policy from any state in the MDP. A first question for extending behavior policy search to value function learning is, "what is the minimal-variance behavior policy when learning a value function for a fixed policy?" The answer to this question may give insight into how to best adapt the behavior policy for low variance evaluation. One facet of this question is whether the minimal-variance behavior policy for estimating the expected return from one state is the same as the minimal-variance behavior policy for another. As with lowering variance for multiple evaluation policies, it may be necessary to assume a measure of the relative importance of states. Another facet of the minimal-variance behavior policy question concerns the use of intermediate value estimates or *bootstrapping*. The variance of a return estimate that uses an intermediate value estimate may change as the intermediate value estimate changes. Thus the minimal-variance behavior policy may be non-stationary as the value function is learned.

11.3 Behavior Policy Search for Policy Improvement

The primary goal of reinforcement learning is policy improvement: learning a policy that maximizes the expected sum of discounted rewards. A final direction for future work is to apply behavior policy search to policy improvement. Behavior policy search could aid policy improvement by lowering the variance of policy gradient estimation or improving value function learning for value-based methods. Regardless of the underlying approach, one fundamental difficulty will be balancing finding a behavior policy that lowers variance while maintaining sufficient exploration to find an optimal policy.

11.4 Theoretical Variance Reduction

We have shown empirically that behavior policy search methods can produce lower variance importance sampling estimates than on-policy data collection. Future work should establish in theory that variance is reduced and at what rate the variance decreases. It is known that the importance sampling estimator has variance $\frac{\sigma^2}{n}$ where n is the number of trajectories

and σ is the variance of the importance sampled return under a fixed sampling distribution (Owen, 2013). Prior work on adaptive IS outside of RL suggests that the rate of $\frac{1}{n}$ cannot be improved (Akyildiz and Míguez, 2021). Thus future work should focus on analysis of how σ decreases as the behavior policy changes. Such analysis could provide further guidance on identifying the settings where behavior policy search is preferable to simply running the evaluation policy for policy evaluation.

12. Conclusion

In this work we have shown that off-policy importance sampling policy evaluation can have lower variance than on-policy Monte Carlo policy evaluation. We derived a condition for the minimal-variance behavior policy. We then introduced the behavior policy search (BPS) problem in order to improve estimation of $v(\pi_e)$ for an evaluation policy π_e. We present two solution algorithms for this problem: the Behavior Policy Gradient on the Variance algorithm and the Behavior Policy Gradient on the KL-Divergence algorithm. BPG-V adapts the behavior policy with stochastic gradient descent on the variance of the importance-sampling estimator. BPG-KL adapts the behavior policy with stochastic gradient descent on the KL-divergence between the current behavior policy and the minimal-variance behavior policy. Experiments demonstrate that both algorithms lower the MSE of estimates of $v(\pi_e)$ compared to on-policy estimates. We also demonstrate BPS can further decrease the MSE of estimates in conjunction with a model-based control variate method.

Acknowledgments

We thank Daniel Brown and the anonymous reviewers of the earlier version of this work for insightful comments on the work and its presentation. Josiah Hanna acknowledges support from NSF (IIS-2410981), American Family Insurance through a research partnership with the University of Wisconsin—Madison's Data Science Institute, the Wisconsin Alumni Research Foundation, and Sandia National Labs through a University Partnership Award. A portion of this work has taken place in the Learning Agents Research Group (LARG) at UT Austin and the Safe, Correct, and Aligned Learning and Robotics Lab (SCALAR) at The University of Massachusetts Amherst. LARG research is supported in part by NSF (FAIN-2019844, NRT-2125858), ONR (N00014-18-2243), ARO (W911NF-23-2-0004, W911NF-17-2-0181), Lockheed Martin, and UT Austin's Good Systems grand challenge. Peter Stone serves as the Executive Director of Sony AI America and receives financial compensation for this work. 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. SCALAR research is supported in part by the NSF (IIS-2323384) and AFOSR (FA9550-20-1-0077).

References

A. Agarwal, N. Jiang, S. M. Kakade, and W. Sun. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 2019.

T. I. Ahamed, V. S. Borkar, and S. Juneja. Adaptive importance sampling technique for Markov chains using stochastic approximation. *Operations Research*, 54(3):489–504, 2006.

Ö. D. Akyildiz and J. Míguez. Convergence rates for optimised adaptive importance samplers. *Statistics and Computing*, 31:1–17, 2021.

T. Archibald, K. McKinnon, and L. Thomas. On the generation of markov decision processes. *Journal of the Operational Research Society*, 46(3):354–361, 1995.

S. M. R. Arnold, P. L'Ecuyer, L. Chen, Y.-f. Chen, and F. Sha. Policy Learning and Evaluation with Randomized Quasi-Monte Carlo. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

K. D. Asis, J. F. Hernandez-Garcia, G. Z. Holland, and R. S. Sutton. Multi-step reinforcement learning: A unifying algorithm, 2017.

K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.

M. Bastani. Model-free intelligent diabetes management using machine learning. Master's thesis, Department of Computing Science, University of Alberta, 2014.

D. P. Bertsekas and J. N. Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10:627–642, 2000.

G. Bouchard, T. Trouillon, J. Perez, and A. Gaidon. Online learning to sample. *arXiv preprint arXiv:1506.09016*, 2016.

S. Boyd, S. P. Boyd, and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. OpenAI gym. *arXiv preprint arXiv:1606.01540*, 2016.

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

N. Corrado and J. P. Hanna. On-Policy Policy Gradient Reinforcement Learning Without On-Policy Sampling. *Arxiv Pre-Print*, 2023.

E. Coumans and Y. Bai. Pybullet, a python module for physics simulation for games, robotics, and machine learning. `http://pybullet.org`, 2016–2019.

P. Y. Desai and P. W. Glynn. 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*, pages 379–384. IEEE Computer Society, 2001.

Y. Duan, X. Chen, R. Houthooft, J. Schulman, and P. Abbeel. Benchmarking deep reinforcement learning for continuous control. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

J. Frank, S. Mannor, and D. Precup. Reinforcement learning in the presence of rare events. In *Proceedings of the 25th International Conference on Machine Learning*, pages 336–343. ACM, 2008.

C. Gelada and M. G. Bellemare. Off-policy deep reinforcement learning by bootstrapping the covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3647–3655, 2019.

Z. D. Guo, P. S. Thomas, and E. Brunskill. Using options and covariance testing for long horizon off-policy policy evaluation. In *Proceedings of the 31st Conference on Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

A. Hallak and S. Mannor. Consistent on-line off-policy evaluation. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1372–1383, 2017.

J. Hammersley and D. Handscomb. Monte Carlo methods. *Ltd., London*, page 40, 1964.

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

J. P. Hanna, S. Niekum, and P. Stone. Importance sampling in reinforcement learning with an estimated behavior policy. *Machine Learning*, pages 1–51, 2021.

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

S. Kakade. A natural policy gradient. In *Proceedings of the 14th Conference on Advances in Neural Information Processing Systems (NeurIPS)*, volume 14, pages 1531–1538, 2001.

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

C. Lemieux. Control variates. *Wiley StatsRef: Statistics Reference Online*, pages 1–8, 2014.

T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. *CoRR*, abs/1509.02971, 2015.

Q. Liu, L. Li, Z. Tang, and D. Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, pages 5356–5366, 2018.

S. Liu and S. Zhang. Improving Monte Carlo Evaluation with Offline Data. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

A. R. Mahmood, H. P. van Hasselt, and R. S. Sutton. Weighted importance sampling for off-policy learning with linear function approximation. In *Advances in Neural Information Processing Systems*, pages 3014–3022, 2014.

S. Mukherjee, J. P. Hanna, and R. Nowak. ReVar: Strengthening Policy Evaluation via Reduced Variance Sampling. In *Proceedings of the 38th International Conference on Uncertainty in Artificial Intelligence (UAI)*, 2022.

S. Mukherjee, J. P. Hanna, and R. Nowak. SaVeR: Optimal Data Collection Strategy for Safe Policy Evaluation in Tabular MDP. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024a.

S. Mukherjee, Q. Xie, J. P. Hanna, and R. Nowak. SPEED: Experimental Design for Policy Evaluation in Linear Heteroscedastic Bandits. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024b.

M. Mutný, T. Janik, and A. Krause. Active Exploration via Experiment Design in Markov Chains. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

A. B. Owen. *Monte Carlo theory, methods and examples.* 2013.

J. Peters and S. Schaal. Natural actor-critic. *Neurocomputing*, 71(7-9):1180–1190, 2008.

B. Piot, M. Geist, and O. Pietquin. Difference of convex functions programming for reinforcement learning. *Advances in Neural Information Processing Systems*, 27, 2014.

T. Popoviciu. Sur les équations algébriques ayant toutes leurs racines réelles. *Mathematica*, 9:129–145, 1935.

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

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

H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

R. Y. Rubinstein and D. P. Kroese. *Simulation and the Monte Carlo method*, volume 10. John Wiley & Sons, 2016.

J. Schulman, S. Levine, P. Moritz, M. Jordan, and P. Abbeel. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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

S. P. Singh and R. S. Sutton. Reinforcement learning with replacing eligibility traces. *Machine Learning*, 22(1-3):123–158, 1996.

A. L. Strehl, L. Li, and M. L. Littman. Reinforcement learning in finite MDPs: PAC analysis. *Journal of Machine Learning Research*, 10:2413–2444, 2009.

R. S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44, 1988.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Proceedings of the 13th Conference on Advances in Neural Information Processing Systems (NeurIPS)*, 2000.

R. S. Sutton, J. Modayil, M. Delp, T. Degris, P. M. Pilarski, A. White, and D. Precup. Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 761–768, 2011.

A. Swaminathan and T. Joachims. The self-normalized estimator for counterfactual learning. In *Proceedings of the 29th Conference on Advances in Neural Information Processing Systems (NeurIPS)*, pages 3231–3239, 2015.

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

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

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

P. S. Thomas and E. Brunskill. Importance sampling with unequal support. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

P. S. Thomas, G. Theocharous, and M. Ghavamzadeh. High confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2015a.

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

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

R. Wan, B. Kveton, and R. Song. Safe Exploration for Efficient Policy Evaluation and Comparison. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2022.

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

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

M. Yang, O. Nachum, B. Dai, L. Li, and D. Schuurmans. Off-policy evaluation via the regularized lagrangian. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

R. Zhong, D. Zhang, L. Schäfer, S. V. Albrecht, and J. P. Hanna. Robust On-Policy Sampling for Data-Efficient Policy Evaluation in Reinforcement Learning. In *Proceedings of Neural and Information Processing Systems (NeurIPS)*, 2022.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 928–936, 2003.

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

Appendix A. Statistical Properties of Behavior Policy Search Estimates

In this appendix, we prove that the estimates from behavior policy search algorithms that are computed as the mean of an unbiased off-policy estimator, OPE, such as IS, are unbiased and consistent estimates of $v(\pi_e)$ and we provide a theoretical finite-rate bound on the estimate. Typically, such results rely on i.i.d. sampling of trajectories from a single $\pi_{\boldsymbol{\theta}_j}$ or at least independent sampling from a behavior policy that is independent of other behavior policies. In our case, the difficulty is that the estimate at iteration i depends on all $\pi_{\boldsymbol{\theta}_j}$ for $i = 1 \dots i$ and each $\pi_{\boldsymbol{\theta}_j}$ is *not* independent of the others. Further, as π_{θ_j} may be different from π_{θ_k} when $j \neq k$, both the assumptions of independence and identical distribution do not hold. Nevertheless, we prove here that behavior policy search algorithms still produce unbiased and consistent estimates of $v(\pi_e)$ at each iteration and have finite-rate bounds similar to Hoeffding's bounds.

Proposition 1 *Under Assumption 2,* $\overline{\text{OPE}}(\pi_e, D_n)$ *is an unbiased estimator of* $v(\pi_e)$ *for any* $n \in \mathbb{N}$*,*

$$\mathbb{E}\left[\overline{\text{OPE}}(\pi_e, D_n)\right] = v(\pi_e).$$

Proof We begin by expanding $\mathbb{E}\left[\overline{\text{OPE}}(\pi_e, D_n)\right]$,

$$\mathbb{E}\left[\overline{\text{OPE}}(\pi_e, D_n)\right] = \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n} \text{OPE}(\pi_e, H_j, \pi_{\boldsymbol{\theta}_j})\right] = \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}\left[\text{OPE}(\pi_e, H_j, \pi_{\boldsymbol{\theta}_j})\right]. \quad (4)$$

Recall that in $\text{OPE}(\pi_e, H_j, \pi_{\boldsymbol{\theta}_j})$ the random variables are the parameters $\boldsymbol{\theta}_j$ under the (stochastic) algorithm and the trajectory H_j generated using π_{θ_j}. Therefore,

$$\mathbb{E}\left[\text{OPE}(\pi_e, H_j, \pi_{\boldsymbol{\theta}_j})\right] = \int_{\Theta} p(\boldsymbol{\theta}_j = \boldsymbol{\theta}) \left(\sum_{h \in \mathcal{H}} p(H_j = h | \pi_{\boldsymbol{\theta}_j}) \, \text{OPE}(\pi_e, h, \pi_{\boldsymbol{\theta}_j})\right) \mathrm{d}\boldsymbol{\theta}. \quad (5)$$

Observe that (5) factors out the probability of observing parameter $\boldsymbol{\theta}_j$ (which depends on past parameters and trajectories) and the expected value of OPE given the value of $\boldsymbol{\theta}_j$ (which is independent of past parameters and trajectories *given* the value of $\boldsymbol{\theta}_j$). In Figure 3, this idea can be observed from d-separation: *conditioned* on a specific instance of π_{θ_k} the estimates X_k are independent of previous parameters and trajectories.

Therefore, as OPE is an unbiased estimator for any *fixed* policy $\pi_{\boldsymbol{\theta}_j}$ under Assumption 2, (5) can be expressed as,

$$\begin{aligned}\mathbb{E}[\text{OPE}(\pi_e, H_j, \pi_{\boldsymbol{\theta}_j})] &= \int_{\Theta} p(\boldsymbol{\theta}_j = \boldsymbol{\theta}) v(\pi_e) \mathrm{d}\boldsymbol{\theta} \\ &= v(\pi_e). \end{aligned} \quad (6)$$

Therefore, combining (4) and (6),

$$\mathbb{E}\left[\overline{\text{OPE}}(\pi_e, D_n)\right] = \frac{1}{n}\sum_{j=1}^{n} v(\pi_e) = v(\pi_e).$$

■

Lemma 1 *Under Assumption 2, $\forall j \in \mathbb{N}$, and $\forall k \in \mathbb{N}$, where $j \neq k$, $\mathrm{OPE}(\pi_e, H_j, \pi_{\theta_j})$ and $\mathrm{OPE}(\pi_e, H_k, \pi_{\theta_k})$ are uncorrelated. That is,*

$$\forall j \neq k, \quad \mathrm{Cov}\left(\mathrm{OPE}(\pi_e, H_j, \pi_{\theta_j}), \mathrm{OPE}(\pi_e, H_k, \pi_{\theta_k})\right) = 0.$$

Proof We begin by first establishing conditional independence in expectation between $\mathrm{OPE}(\pi_e, H_j, \pi_{\boldsymbol{\theta}_j})$ and $\mathrm{OPE}(\pi_e, H_k, \pi_{\boldsymbol{\theta}_k})$ for any $j \neq k$. For brevity, let $Z_j \coloneqq \mathrm{OPE}(\pi_e, H_j, \pi_{\boldsymbol{\theta}_j})$.

$$\begin{aligned}
\mathbb{E}\left[\mathrm{OPE}(\pi_e, H_j, \pi_{\boldsymbol{\theta}_j}) \middle| \mathrm{OPE}(\pi_e, H_k, \pi_{\boldsymbol{\theta}_k})\right] &= \mathbb{E}\left[Z_j | Z_k\right] \\
&= \int_\Theta p(\boldsymbol{\theta}_j = \boldsymbol{\theta} | Z_k) \sum_{h \in \mathcal{H}} p(H_j = h | \pi_{\boldsymbol{\theta}_j}, Z_k) \,\mathrm{OPE}(\pi_e, h, \pi_{\boldsymbol{\theta}_j}) \,\mathrm{d}\boldsymbol{\theta} \\
&\overset{(a)}{=} \int_\Theta p(\boldsymbol{\theta}_j = \boldsymbol{\theta} | Z_k) \left(\sum_{h \in \mathcal{H}} p(H_j = h | \pi_{\boldsymbol{\theta}_j}) \,\mathrm{OPE}(\pi_e, h, \pi_{\boldsymbol{\theta}_j}) \right) \mathrm{d}\boldsymbol{\theta} \\
&\overset{(b)}{=} \int_\Theta p(\boldsymbol{\theta}_j = \boldsymbol{\theta} | Z_k) v(\pi_e) \,\mathrm{d}\boldsymbol{\theta} \\
&= v(\pi_e), \qquad (7)
\end{aligned}$$

where (a) follows from the fact that *given* the policy $\pi_{\boldsymbol{\theta}_j}$, H_j is independent of the Z_k (see Fig 3), and (b) follows from arguments similar to those used in the proof of Proposition 1. The co-variance between $\mathrm{OPE}(\pi_e, H_j, \pi_{\boldsymbol{\theta}_j})$ and $\mathrm{OPE}(\pi_e, H_k, \pi_{\boldsymbol{\theta}_k})$ can now be expressed as,

$$\begin{aligned}
\mathrm{Cov}\left(\mathrm{OPE}(\pi_e, H_j, \pi_{\boldsymbol{\theta}_j}), \mathrm{OPE}(\pi_e, H_k, \pi_{\boldsymbol{\theta}_k})\right) &= \mathrm{Cov}\left(Z_j, Z_k\right) \\
&= \mathbb{E}\left[Z_j Z_k\right] - \mathbb{E}\left[Z_j\right]\mathbb{E}\left[Z_k\right] \\
&\overset{(b)}{=} \mathbb{E}\left[\mathbb{E}[Z_j | Z_k] Z_k\right] - \mathbb{E}\left[Z_j\right]\mathbb{E}\left[Z_k\right] \\
&\overset{(c)}{=} v(\pi_e)\mathbb{E}\left[Z_k\right] - \mathbb{E}\left[Z_j\right]\mathbb{E}\left[Z_k\right] \\
&\overset{(d)}{=} v(\pi_e)^2 - v(\pi_e)^2 \\
&= 0,
\end{aligned}$$

where (b) follows from the law of total expectation, (c) follows from (7), and (d) follows from (6).

■

Proposition 2 *Under Assumption 2, $\overline{\text{OPE}}(\pi_e, D_i)$ converges to $v(\pi_e)$ in probability. That is, for $\epsilon > 0$,*

$$\lim_{i\to\infty} \Pr\left(|\overline{\text{OPE}}(\pi_e, D_i) - v(\pi_e)| > \epsilon\right) = 0.$$

Proof We begin by expanding the variance of $\overline{\text{OPE}}(\pi_e, D_i)$,

$$\begin{aligned}
\text{Var}\left[\overline{\text{OPE}}(\pi_e, D_i)\right] &= \text{Var}\left[\frac{1}{i}\sum_{j=1}^{i} \text{OPE}(\pi_e, H_j, \theta_j)\right] \\
&= \frac{1}{i^2}\left[\sum_{j=1}^{i} \text{Var}\left[\text{OPE}(\pi_e, H_j, \theta_j)\right] + 2\sum_{j=1}^{i}\sum_{k=1}^{i} \text{Cov}\left(\text{OPE}(\pi_e, H_j, \theta_j), \text{OPE}(\pi_e, H_k, \theta_k)\right)\right] \\
&\overset{(a)}{=} \frac{1}{i^2}\left[\sum_{j=1}^{i} \text{Var}\left[\text{OPE}(\pi_e, H_j, \theta_j)\right]\right], \qquad (8)
\end{aligned}$$

where (a) follows using uncorrelatedness established in Lemma 1. Further, from Assumption 2, $\text{OPE}(\pi_e, H_j, \theta_j)$ is a bounded random variable for all j and thus it follows from Popoviciu's inequality (Popoviciu, 1935) that $\text{OPE}(\pi_e, H_j, \pi_{\boldsymbol{\theta}_j})$ has variance bounded above by some finite constant $\tilde{c}$. Therefore, as $\forall j, \text{Var}[\text{OPE}(\pi_e, H_j, \pi_{\boldsymbol{\theta}_j})] < \hat{c}$, it follows from (8) that $\text{Var}[\text{OPE}(\pi_e, D_i)] \to 0$. As $\overline{\text{OPE}}(\pi_e, D_i)$ is unbiased (Proposition 1) and has no variance in the limit it follows from the bias-variance decomposition of mean-squared error that,

$$\begin{aligned}
\lim_{i\to\infty} \quad \mathbb{E}\left[\left(\overline{\text{OPE}}(\pi_e, D_i) - v(\pi_e)\right)^2\right] &= \lim_{i\to\infty} \left(\mathbb{E}\left[\overline{\text{OPE}}(\pi_e, D_i)\right] - v(\pi_e)\right)^2 \\
&\quad + \text{Var}\left[\overline{\text{OPE}}(\pi_e, D_i)\right] = 0. \qquad (9)
\end{aligned}$$

Now from Markov's inequality,

$$\Pr(|\overline{\text{OPE}}(\pi_e, D_i) - v(\pi_e)| > \epsilon) \leq \frac{\mathbb{E}\left[\left(\overline{\text{OPE}}(\pi_e, D_i) - v(\pi_e)\right)^2\right]}{\epsilon^2} \qquad (10)$$

Combining (9) and (10),

$$\lim_{i\to\infty} \Pr(|\overline{\text{OPE}}(\pi_e, D_i) - v(\pi_e)| > \epsilon) = 0.$$

■

Proposition 3 *Under Assumption 2, let $\tilde{c}$ be the range of $\text{OPE}(\pi_e, H_j, \pi_{\theta_j})$ for any $j \in \{1, 2, ..., i\}$, then $\forall \delta \in [0, 1]$,*

$$\Pr\left(\left|\overline{\text{OPE}}(\pi_e, D_i) - v(\pi_e)\right| > \tilde{c}\sqrt{\frac{\ln(2/\delta)}{2|D_i|}}\right) \leq \delta,$$

Proof We begin by observing that since the $(\text{OPE}(\pi_e, H_j, \theta_j))_{j=1}^i$ are sequentially dependent, if we can convert them into a Martingale sequence then we can use concentration inequalities for Martingales to obtain convergence rates for $\overline{\text{OPE}}(\pi_e, D_i)$.

Let $Y_0, Y_1, ..., Y_i$ denote the desired Martingale sequence constructed using the OPE estimates $\text{OPE}(\pi_e, H_1, \theta_1), ..., \text{OPE}(\pi_e, H_i, \theta_n)$, where

$$\begin{aligned} Y_0 &= 0, \\ Y_j &= \text{OPE}(\pi_e, H_j, \pi_{\boldsymbol{\theta}_j}) - v(\pi_e) + Y_{j-1}. \end{aligned} \tag{11}$$

From (11) notice that $\forall j \geq 1$,

$$\begin{aligned} \mathbb{E}\left[Y_j | Y_{j-1}\right] &= \mathbb{E}[\text{OPE}(\pi_e, H_j, \pi_{\boldsymbol{\theta}_j})|Y_{j-1}] - \mathbb{E}[v(\pi_e)|Y_{j-1}] + \mathbb{E}[Y_{j-1}|Y_{j-1}] \\ &= \mathbb{E}[\text{OPE}(\pi_e, H_j, \pi_{\boldsymbol{\theta}_j})|Y_{j-1}] - v(\pi_e) + Y_{j-1}. \end{aligned} \tag{12}$$

To simplify (12) further, notice that,

$$\begin{aligned} \mathbb{E}[\text{OPE}(\pi_e, H_j, \pi_{\boldsymbol{\theta}_j})|Y_{j-1}] &= \int_{\mathbb{R}} p(\text{OPE}(\pi_e, H_j, \pi_{\boldsymbol{\theta}_j}) = x|Y_{j-1}) x \, \mathrm{d}x \\ &\overset{(a)}{=} \int_{\Theta} p(\boldsymbol{\theta}_j = \boldsymbol{\theta}|Y_{j-1}) \left(\int_{\mathbb{R}} p(\text{OPE}(\pi_e, H_j, \pi_{\boldsymbol{\theta}_j}) = x|\pi_{\boldsymbol{\theta}_j}) x \, \mathrm{d}x \right) \mathrm{d}\boldsymbol{\theta} \\ &\overset{(b)}{=} \int_{\Theta} p(\boldsymbol{\theta}_j = \boldsymbol{\theta}|Y_{j-1}) v(\pi_e) \mathrm{d}\boldsymbol{\theta} \\ &= v(\pi_e), \end{aligned} \tag{13}$$

where (a) follows from the fact that Y_{j-1} only contains information from iterates till $j-1$ (inclusive) and $\text{OPE}(\pi_e, H_j, \pi_{\boldsymbol{\theta}_j})$ is independent of the past *conditioned* on the value of θ_j. Step (b) follows from the fact that $\text{OPE}(\pi_e, H_j, \pi_{\boldsymbol{\theta}_j})$ (the OPE estimate) is an unbiased estimator of $v(\pi_e)$ for any fixed behavior policy $\pi_{\boldsymbol{\theta}}$ under Assumption 1. Combining (13) and (12), it can be observed that $(Y_j)_{j=1}^i$ is a Martingale sequence as

$$\mathbb{E}[Y_j | Y_{j-1}] = Y_{j-1}.$$

Since $\text{OPE}(\pi_e, H_j, \pi_{\boldsymbol{\theta}_j})$ is bounded (under Assumption 1), Y_j is also bounded. Consequently, the differences between Y_j and Y_{j-1} are also bounded. Applying Azuma's inequality (Azuma, 1967) for Martingales to the sequence $(Y_j)_{j=0}^i$,

$$\Pr\left(|Y_i - Y_0| > \epsilon\right) \leq 2 \exp\left(\frac{-2\epsilon^2}{|D_i| \tilde{c}^2}\right). \tag{14}$$

First, considering $|Y_i - Y_0|$:

$$|Y_i - Y_0| \overset{(a)}{=} \left| \sum_{j=1}^{i} (Y_j - Y_{j-1}) \right| \overset{(b)}{=} \left| \sum_{j=1}^{i} (\text{OPE}(\pi_e, H_j, \pi_{\boldsymbol{\theta}_j}) - v(\pi_e)) \right|, \tag{15}$$

where (a) follows by telescoping the summation and (b) follows from (11). Combining (15) and (14),

$$\Pr\left(\left|\sum_{j=1}^{i}(\mathrm{OPE}(\pi_e, H_j, \pi_{\boldsymbol{\theta}_j}) - v(\pi_e))\right| > \epsilon\right) \leq 2\exp\left(\frac{-2\epsilon^2}{|D_i|\tilde{c}^2}\right)$$

$$\Pr\left(\left|\sum_{j=1}^{i}\mathrm{OPE}(\pi_e, H_j, \pi_{\boldsymbol{\theta}_j}) - |D_i|v(\pi_e)\right| > \epsilon\right) \leq 2\exp\left(\frac{-2\epsilon^2}{|D_i|\tilde{c}^2}\right)$$

$$\Pr\left(\left|\frac{1}{|D_i|}\sum_{j=1}^{i}\mathrm{OPE}(\pi_e, H_j, \pi_{\boldsymbol{\theta}_j}) - v(\pi_e)\right| > \frac{\epsilon}{|D_i|}\right) \overset{(c)}{\leq} 2\exp\left(\frac{-2\epsilon^2}{|D_i|\tilde{c}^2}\right)$$

$$\Pr\left(\left|\overline{\mathrm{OPE}}(\pi_e, D_i) - v(\pi_e)\right| > \frac{\epsilon}{|D_i|}\right) \overset{(d)}{\leq} 2\exp\left(\frac{-2\epsilon^2}{|D_i|\tilde{c}^2}\right)$$

$$\Pr\left(\left|\overline{\mathrm{OPE}}(\pi_e, D_i) - v(\pi_e)\right| > \varepsilon\right) \overset{(e)}{\leq} 2\exp\left(\frac{-2|D_i|\varepsilon^2}{\tilde{c}^2}\right), \tag{16}$$

where (c) follows from dividing both sides within the LHS by $|D_i|$, (d) follows from definition of $\overline{\mathrm{OPE}}(\pi_e, D_i)$, and (e) follows from relabeling $\varepsilon \coloneqq \frac{\epsilon}{|D_i|}$. Finally, relabeling the RHS in (16) to δ, one can obtain,

$$\Pr\left(\left|\overline{\mathrm{OPE}}(\pi_e, D_i) - v(\pi_e)\right| > \tilde{c}\sqrt{\frac{\ln(2/\delta)}{2|D_i|}}\right) \leq \delta,$$

thereby giving the desired error rate of $O\left(\frac{1}{\sqrt{|D_i|}}\right)$.

■

Appendix B. Behavior Policy Gradient of the Variance

In this section, we derive the gradient of the variance of importance sampling with respect to the behavior policy parameters. We first derive an analytic expression for the gradient of the variance of an arbitrary, unbiased off-policy policy evaluation estimator, $\mathrm{OPE}(\pi_e, H, \pi_{\boldsymbol{\theta}})$. From our general derivation we derive the gradient of the variance of the basic importance sampling estimator and then extend to the doubly robust and per-decision estimators.

B.1 MSE Gradient for an Unbiased Off-Policy Policy Evaluation Method

Lemma 2 gives the gradient of the MSE for any unbiased off-policy policy evaluation method.

Lemma 2

$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{MSE}\Big[\mathrm{OPE}(\pi_e, H, \pi_{\boldsymbol{\theta}})\Big] = \mathbf{E}\Big[\,\mathrm{OPE}(\pi_e, H, \pi_{\boldsymbol{\theta}})^2 \Big(\sum_{t=0}^{l-1} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_t|S_t)\Big) + \frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{OPE}(\pi_e, H, \pi_{\boldsymbol{\theta}})^2 \;\Big|\; H \sim \pi_{\boldsymbol{\theta}}\Big].$$

Proof We begin by decomposing $\Pr(H = h|\pi)$ into two components – one that depends on π and the other that does not. Recall that we defined:

$$w_\pi(h) \coloneqq \prod_{t=0}^{l-1} \pi(a_t|s_t),$$

and define

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

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

The MSE of the OPE estimator is given by:

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

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

$$\begin{aligned}
\mathrm{MSE}(\mathrm{OPE}(\pi_e, H, \pi_{\boldsymbol{\theta}})) &= \mathrm{Var}[\mathrm{OPE}(\pi_e, H, \pi_{\boldsymbol{\theta}})] && (17)\\
&= \mathbf{E}\left[\mathrm{OPE}(\pi_e, H, \pi_{\boldsymbol{\theta}})^2 \middle| H \sim \pi_{\boldsymbol{\theta}}\right] - \mathbf{E}[\mathrm{OPE}(\pi_e, H, \pi_{\boldsymbol{\theta}})|H \sim \pi_{\boldsymbol{\theta}}]^2 && (18)\\
&= \mathbf{E}\left[\mathrm{OPE}(\pi_e, H, \pi_{\boldsymbol{\theta}})^2 \middle| H \sim \pi_{\boldsymbol{\theta}}\right] - v(\pi_e)^2 && (19)
\end{aligned}$$

(18) follows from (17) by the definition of the variance and (19) follows from (18) because the expectation of an unbiased estimator of $v(\pi_e)$ is $v(\pi_e)$.

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

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{MSE}[\mathrm{OPE}(\pi_e, H, \pi_{\boldsymbol{\theta}})] &= \frac{\partial}{\partial \boldsymbol{\theta}} \left(\mathbf{E}\left[\mathrm{OPE}(\pi_e, H, \pi_{\boldsymbol{\theta}})^2 \middle| H \sim \pi_{\boldsymbol{\theta}}\right] - \underbrace{v(\pi_e)^2}_{\text{const}}\right)\\
&= \frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{E}\left[\mathrm{OPE}(\pi_e, H, \pi_{\boldsymbol{\theta}})^2 \middle| H \sim \pi_{\boldsymbol{\theta}}\right]\\
&= \frac{\partial}{\partial \boldsymbol{\theta}} \sum_{h \in \mathcal{H}} \Pr(H = h|\pi_{\boldsymbol{\theta}})\, \mathrm{OPE}(\pi_e, h, \pi_{\boldsymbol{\theta}})^2\\
&= \sum_{h \in \mathcal{H}} \Pr(H = h|\pi_{\boldsymbol{\theta}}) \frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{OPE}(\pi_e, h, \pi_{\boldsymbol{\theta}})^2 +\\
&\quad \mathrm{OPE}(\pi_e, h, \pi_{\boldsymbol{\theta}})^2 \frac{\partial}{\partial \boldsymbol{\theta}} \Pr(H = h|\pi_{\boldsymbol{\theta}})\\
&= \sum_{h \in \mathcal{H}} \Pr(H = h|\pi_{\boldsymbol{\theta}}) \frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{OPE}(\pi_e, h, \pi_{\boldsymbol{\theta}})^2 +\\
&\quad \mathrm{OPE}(\pi_e, h, \pi_{\boldsymbol{\theta}})^2 p(h) \frac{\partial}{\partial \boldsymbol{\theta}} w_{\pi_{\boldsymbol{\theta}}}(h) && (20)
\end{aligned}$$

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

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\theta}} w_{\pi_{\boldsymbol{\theta}}}(h) =& \frac{\partial}{\partial \boldsymbol{\theta}} \prod_{t=0}^{l-1} \pi_{\boldsymbol{\theta}}(a_t|s_t) \\ \overset{\text{(a)}}{=}& \left(\prod_{t=0}^{l-1} \pi_{\boldsymbol{\theta}}(a_t|s_t)\right) \left(\sum_{t=0}^{l-1} \frac{\frac{\partial}{\partial \boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(a_t|s_t)}{\pi_{\boldsymbol{\theta}}(a_t|s_t)}\right) \\ \overset{\text{(b)}}{=}& w_{\pi_{\boldsymbol{\theta}}}(h) \sum_{t=0}^{l-1} \frac{\partial}{\partial \boldsymbol{\theta}} \log\left(\pi_{\boldsymbol{\theta}}(a_t|s_t)\right), \end{aligned} \tag{21}$$

where **(a)** comes from the product rule of differentiation and **(b)** comes from the likelihood-ratio trick (i.e., $\frac{\frac{\partial}{\partial \boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(A|S)}{\pi_{\boldsymbol{\theta}}(A|S)} = \log(\pi_{\boldsymbol{\theta}}(A|S))$) and the definition of $w_{\pi_{\boldsymbol{\theta}}}(h)$. Continuing from (20) we have that:

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\theta}} \operatorname{MSE}(\operatorname{OPE}(\pi_e, H, \pi_{\boldsymbol{\theta}})) = \mathbf{E}\Bigg[\operatorname{OPE}(\pi_e, H, \pi_{\boldsymbol{\theta}})^2 \sum_{t=0}^{l-1} \frac{\partial}{\partial \boldsymbol{\theta}} \log\left(\pi_{\boldsymbol{\theta}}(A_t|S_t)\right) + \\ \frac{\partial}{\partial \boldsymbol{\theta}} \operatorname{OPE}(\pi_e, H, \pi_{\boldsymbol{\theta}})^2 \;\Bigg|\; H \sim \pi_{\boldsymbol{\theta}} \Bigg].\end{aligned}$$

■

B.2 Behavior Policy Gradient of the Variance

We now use Lemma 2 to prove the Behavior Policy Gradient of the Variance Theorem.

Theorem 1 (Behavior Policy Gradient of the Variance)

$$\frac{\partial}{\partial \boldsymbol{\theta}} \operatorname{MSE}\Big[\operatorname{IS}(\pi_e, H, \pi_{\boldsymbol{\theta}})\Big] = \mathbf{E}\left[-\operatorname{IS}(\pi_e, H, \pi_{\boldsymbol{\theta}})^2 \sum_{t=0}^{l-1} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_t|S_t) \middle| H \sim \pi_{\boldsymbol{\theta}} \right]$$

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

$$\begin{aligned}\operatorname{IS}(\pi_e, H, \pi_{\boldsymbol{\theta}})^2 &= \left(\frac{w_{\pi_e}}{w_{\pi_{\boldsymbol{\theta}}}} g(H)\right)^2 \\ \frac{\partial}{\partial \boldsymbol{\theta}} \operatorname{IS}(\pi_e, H, \pi_{\boldsymbol{\theta}})^2 &= \frac{\partial}{\partial \boldsymbol{\theta}} \left(\frac{w_{\pi_e}(H)}{w_{\pi_{\boldsymbol{\theta}}}(H)} g(H)\right)^2 \\ &= 2g(H) \frac{w_{\pi_e}(H)}{w_{\pi_{\boldsymbol{\theta}}}(H)} \frac{\partial}{\partial \boldsymbol{\theta}} \left(g(H) \frac{w_{\pi_e}(H)}{w_{\pi_{\boldsymbol{\theta}}}(H)}\right) \\ &\overset{\text{(a)}}{=} -2g(H) \frac{w_{\pi_e}(H)}{w_{\pi_{\boldsymbol{\theta}}}(H)} \left(g(H) \frac{w_{\pi_e}(H)}{w_{\pi_{\boldsymbol{\theta}}}(H)}\right) \sum_{t=0}^{l-1} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_t|S_t) \\ &= -2 \operatorname{IS}(\pi_e, H, \pi_{\boldsymbol{\theta}})^2 \sum_{t=0}^{l-1} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_t|S_t),\end{aligned}$$

where **(a)** uses (21) to differentiate $w_{\pi_{\boldsymbol{\theta}}}(H)$.

Substituting this expression and $\text{IS}(\pi_e, H, \pi_{\boldsymbol{\theta}})$ for $\text{OPE}(\pi_e, H, \pi_{\boldsymbol{\theta}})$ into Lemma 2 completes the proof of Theorem 1:

$$\frac{\partial}{\partial \boldsymbol{\theta}} \text{MSE}[\text{IS}(\pi_e, H, \pi_{\boldsymbol{\theta}})] = \mathbf{E}\left[-\text{IS}(\pi_e, H, \pi_{\boldsymbol{\theta}})^2 \sum_{t=0}^{l-1} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_t|S_t) \middle| H \sim \pi_{\boldsymbol{\theta}} \right].$$

■

B.3 MSE Gradient for the Doubly Robust Estimator

We also present an extension of the IS MSE gradient to the Doubly Robust (DR) estimator. Recall that for a single trajectory, H, DR is defined as:

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

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

The gradient of the MSE of the DR estimator is given by the following corollary to Theorem 1:

Corollary 1

$$\frac{\partial}{\partial \boldsymbol{\theta}} \text{MSE}\left[DR(\pi_e, H, \pi_{\boldsymbol{\theta}}, \hat{q}^{\pi_e}, \hat{v}^{\pi_e})\right] = \mathbf{E}\Bigg[\text{DR}(\pi_e, H, \boldsymbol{\theta}, \hat{q}^{\pi_e}, \hat{v}^{\pi_e})^2 \sum_{t=0}^{l-1} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_t|S_t) - 2\,\text{DR}(\pi_e, H, \pi_{\boldsymbol{\theta}}, \hat{q}^{\pi_e}, \hat{v}^{\pi_e}) \left(\sum_{t=0}^{l-1} \gamma^t \delta_t \frac{w_{\pi_e,t}}{w_{\boldsymbol{\theta},t}} \sum_{i=0}^{t} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_i|S_i)\right)\Bigg]$$

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

Proof

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

Let $\delta_t \coloneqq R_t - \hat{q}^{\pi_e}(S_t, A_t) + \hat{v}^{\pi_e}(S_{t+1})$.

$$\text{DR}(\pi_e, H, \pi_{\boldsymbol{\theta}})^2 = \left(\hat{v}^{\pi_e}(S_0) + \sum_{t=0}^{l-1} \gamma^t \frac{w_{\pi_e,t}}{w_{\pi_{\boldsymbol{\theta}},t}} \delta_t\right)^2$$

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

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

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\theta}} \operatorname{MSE}\left[\operatorname{DR}(\pi_e, H, \pi_{\boldsymbol{\theta}})\right] = \mathbf{E}[&\operatorname{DR}(\pi_e, H, \pi_{\boldsymbol{\theta}})^2 \sum_{t=0}^{l-1} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_t|S_t) \\ &- 2 \operatorname{DR}(\pi_e, H, \pi_{\boldsymbol{\theta}}) (\sum_{t=0}^{l-1} \gamma^t \delta_t \frac{w_{\pi_e,t}}{w_{\pi_{\boldsymbol{\theta}},t}} \sum_{i=0}^{t} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_i|S_i)) | H \sim \pi_{\boldsymbol{\theta}}]\end{aligned}$$

■

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.

Appendix C. Convergence of BPG-V

In this section, we prove that BPG-V (Algorithm 1) converges under an appropriately chosen step-size.

Proposition 4 *Under Assumption* 1 *and Assumption* 3*, BPG-V converges. That is,* $\operatorname{MSE}[\operatorname{IS}(\pi_e, H_i, \pi_{\boldsymbol{\theta}_i})]$ *converges to a finite value and* $\lim_{i\to\infty} \frac{\partial}{\partial \boldsymbol{\theta}} \operatorname{MSE}[\operatorname{IS}(\pi_e, H_i, \pi_{\boldsymbol{\theta}_i})] = 0$*.*

Proof The proof follows from an application of Proposition 3 in (Bertsekas and Tsitsiklis, 2000). To apply this result, we must show that BPG-V satisfies the following conditions:

1. $\operatorname{MSE}[\operatorname{IS}(\pi_e, H, \pi_{\boldsymbol{\theta}})]$ is continuously differentiable w.r.t. $\boldsymbol{\theta}$.
2. The gradient of the MSE objectives, $\frac{\partial}{\partial \boldsymbol{\theta}} \operatorname{MSE}[\operatorname{IS}(\pi_e, H, \pi_{\boldsymbol{\theta}})]$, is Lipschitz continuous w.r.t. $\boldsymbol{\theta}$.
3. The variance of the gradient estimate used by BPG-V is bounded.

Other conditions of Proposition 3 in (Bertsekas and Tsitsiklis, 2000) are satisfied by the unbiasedness of the gradient estimates used by BPG-V$\dot{\text{W}}$e also note that the MSE objective is bounded below by zero which rules out the case of BPG-V converging to an MSE of $-\infty$ which is technically allowed by Proposition 3 of (Bertsekas and Tsitsiklis, 2000). Theorem 1 gives us $\frac{\partial}{\partial \boldsymbol{\theta}} \operatorname{MSE}[\operatorname{IS}(\pi_e, H, \pi_{\boldsymbol{\theta}})]$ which can be seen to be continuously differentiable under

our assumption that $\pi_{\boldsymbol{\theta}}$ is continuously differentiable and Assumption 1 which implies that $\mathrm{IS}(\pi_e, H, \pi_{\boldsymbol{\theta}})$ always exists.

We next show that the second derivative of the MSE objective is bounded which implies the Lipschitz continuity of $\frac{\partial}{\partial \boldsymbol{\theta}}\mathrm{MSE}[\mathrm{IS}(\pi_e, H, \pi_{\boldsymbol{\theta}})]$.

$$
\begin{aligned}
\frac{\partial^2}{\partial^2 \boldsymbol{\theta}}\mathrm{MSE}[\mathrm{IS}(\pi_e, H, \pi_{\boldsymbol{\theta}})] =& \frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{E}\left[-\mathrm{IS}(\pi_e, H, \pi_{\boldsymbol{\theta}})^2 \underbrace{\sum_{t=0}^{l-1} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_t|S_t)}_{\frac{\partial}{\partial \boldsymbol{\theta}} \log w_{\boldsymbol{\theta}}(H)} \middle| H \sim \pi_{\boldsymbol{\theta}} \right] \\
=& \frac{\partial}{\partial \boldsymbol{\theta}} \sum_{h \in \mathcal{H}} p(h) w_{\boldsymbol{\theta}}(h) \left(-\mathrm{IS}(\pi_e, h, \pi_{\boldsymbol{\theta}})^2 \frac{\partial}{\partial \boldsymbol{\theta}} \log w_{\boldsymbol{\theta}}(h) \right) \\
\overset{(a)}{=}& \frac{\partial}{\partial \boldsymbol{\theta}} \sum_{h \in \mathcal{H}} p(h) w_{\boldsymbol{\theta}}(h) \left(-\mathrm{IS}(\pi_e, h, \pi_{\boldsymbol{\theta}})^2 \frac{\frac{\partial}{\partial \boldsymbol{\theta}} w_{\boldsymbol{\theta}}(h)}{w_{\boldsymbol{\theta}}(h)} \right) \\
=& \frac{\partial}{\partial \boldsymbol{\theta}} \sum_{h \in \mathcal{H}} -p(h) \mathrm{IS}(\pi_e, h, \pi_{\boldsymbol{\theta}})^2 \frac{\partial}{\partial \boldsymbol{\theta}} w_{\boldsymbol{\theta}}(h) \\
\overset{(b)}{=}& \sum_{h \in \mathcal{H}} -p(h) \left[\underbrace{\frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{IS}(\pi_e, h, \pi_{\boldsymbol{\theta}})^2}_{\text{(c)}} \underbrace{\frac{\partial}{\partial \boldsymbol{\theta}} w_{\boldsymbol{\theta}}(h)}_{\text{(d)}} + \underbrace{\mathrm{IS}(\pi_e, h, \pi_{\boldsymbol{\theta}})^2 \frac{\partial^2}{\partial^2 \boldsymbol{\theta}} w_{\boldsymbol{\theta}}(h)}_{\text{(e)}} \right]
\end{aligned}
$$

where (**a**) comes from the chain rule of calculus and (**b**) comes from the product rule of calculus. We can now show that each term (c, d, and e) is bounded. First, for (c):

$$
\frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{IS}(\pi_e, h, \pi_{\boldsymbol{\theta}})^2 = \underbrace{\frac{-2 g(h)^2 w_{\pi_e}(h)^2}{w_{\boldsymbol{\theta}}(h)^3}}_{\text{(c.1)}} \underbrace{\frac{\partial}{\partial \boldsymbol{\theta}} w_{\boldsymbol{\theta}}(h)}_{\text{(c.2)}},
$$

which is bounded because Assumption 1 implies (c.1) is bounded and (c.2) is the same as (d) which we next show is bounded.

For (d):

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}} w_{\boldsymbol{\theta}}(h) =& \frac{\partial}{\partial \boldsymbol{\theta}} \prod_{t=0}^{l-1} \pi_{\boldsymbol{\theta}}(a_t|s_t) \\
=& \sum_{t=0}^{l-1} \frac{\partial}{\partial \boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(a_t|s_t) \prod_{t'=0, t' \neq t}^{l-1} \pi_{\boldsymbol{\theta}}(a_{t'}|s_{t'}),
\end{aligned} \tag{22}
$$

which is bounded because each $\frac{\partial}{\partial \boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(a_t|s_t)$ is bounded by construct, and $\prod_{t'=0, t' \neq t}^{l-1} \pi_{\boldsymbol{\theta}}(a_{t'}|s_{t'}) \leq 1$.

Finally, for (e), we just consider $\frac{\partial^2}{\partial^2 \boldsymbol{\theta}} w_{\boldsymbol{\theta}}(h)$ since Assumption 1 implies that $\text{IS}(\pi_e, h, \pi_{\boldsymbol{\theta}})^2$ exists and is bounded.

$$\begin{aligned}\frac{\partial^2}{\partial^2 \boldsymbol{\theta}} w_{\boldsymbol{\theta}}(h) =& \frac{\partial}{\partial \boldsymbol{\theta}} \sum_{t=0}^{l-1} \frac{\partial}{\partial \boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(a_t|s_t) \prod_{t'=0, t' \neq t}^{l-1} \pi_{\boldsymbol{\theta}}(a_t|s_t) \\ =& \sum_{t=0}^{l-1} \frac{\partial^2}{\partial^2 \boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(a_t|s_t) \prod_{t' \neq t} \pi_{\boldsymbol{\theta}}(a_{t'}|s_{t'}) + \frac{\partial}{\partial \boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(a_t|s_t) \sum_{t' \neq t} \frac{\partial}{\partial \boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(a_{t'}|s_{t'}) \prod_{t'' \neq t, t'} \pi_{\boldsymbol{\theta}}(a_{t''}|s_{t''}),\end{aligned}$$

which is bounded under the construct that $\pi_{\boldsymbol{\theta}}$ is twice differentiable with bounded first and second derivatives. Thus we conclude that the MSE objective is continuosly differentiable with a Lipschitz derivative.

Finally, we have to show that the variance of the gradient estimate used by BPG-V is bounded. To do so, we show that the gradient estimate with any single trajectory is bounded which implies the variance of the estimates used by BPG-V is bounded because the variance of a bounded random variable is itself bounded.

For any trajectory h, collected by following $\pi_{\boldsymbol{\theta}}$, an unbiased estimate of the MSE estimate is given as:

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\theta}} \text{MSE}[\text{IS}(\pi_e, H, \pi_{\boldsymbol{\theta}}] \approx& -\text{IS}(\pi_e, h, \pi_{\boldsymbol{\theta}})^2 \sum_{t=0}^{l-1} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a_t|s_t) \\ \overset{(a)}{=}& -\frac{w_{\pi_e}(h)^2 g(h)^2}{w_{\pi_{\boldsymbol{\theta}}}(h)^2} \frac{\partial}{\partial \boldsymbol{\theta}} \log w_{\pi_{\boldsymbol{\theta}}}(h) \\ \overset{(b)}{=}& -\frac{w_{\pi_e}(h)^2 g(h)^2}{w_{\pi_{\boldsymbol{\theta}}}(h)^2} \frac{\frac{\partial}{\partial \boldsymbol{\theta}} w_{\pi_{\boldsymbol{\theta}}}(h)}{w_{\pi_{\boldsymbol{\theta}}}(h)} \\ =& -\frac{w_{\pi_e}(h)^2 g(h)^2}{w_{\pi_{\boldsymbol{\theta}}}(h)^3} \frac{\partial}{\partial \boldsymbol{\theta}} w_{\pi_{\boldsymbol{\theta}}}(h) \qquad (23)\end{aligned}$$

where (a) uses $w_\pi(h) = \prod_{t=0}^{l-1} \pi(a_t|s_t)$ and (b) uses the likelihood-ratio trick. On the RHS of Equation (23), $\frac{w_{\pi_e}(h)^2 g(h)^2}{w_{\pi_{\boldsymbol{\theta}}}(h)^3}$ is bounded under Assumption 1 and $\frac{\partial}{\partial \boldsymbol{\theta}} w_{\pi_{\boldsymbol{\theta}}}(h)$ was shown to be bounded in Equation (22). Thus we conclude that the variance of the gradient estimate used by BPG-V is bounded. Proposition 4 now follows from Proposition 3 of Bertsekas and Tsitsiklis (2000).

■

Appendix D. Convexity of Variance Objective

In this appendix, we prove that, when $\pi_{\boldsymbol{\theta}}$ is a linear-softmax policy, then the variance objective minimized by BPG-V is convex in the policy parameters. Thus, BPG-V is guaranteed to converge to the parameter vector that minimizes the variance of the IS-return under standard stochastic gradient descent step-size conditions (Bertsekas and Tsitsiklis, 2000).

A linear-softmax policy is a policy over a finite set of actions where the probability of each action is defined as a softmax distribution with logits from a linear combination of state features. Formally, let $\phi : \mathcal{S} \to \mathbf{R}^q$ for integer q be a state feature function that maps states to feature vectors. For each action, $a \in \mathcal{A}$, we have a vector $\boldsymbol{\theta}_a \in \mathbf{R}^q$ and $\boldsymbol{\theta}$ is the concatenation of all $\boldsymbol{\theta}_a$. A linear-softmax policy defines the probability of action a in state s as:

$$\pi_{\boldsymbol{\theta}}(a|s) = \frac{e^{\boldsymbol{\theta}_a^T \phi(s)}}{\sum_{b \in \mathcal{A}} e^{\boldsymbol{\theta}_b^T \phi(s)}}.$$

Theorem 2 *Assume $\pi_{\boldsymbol{\theta}}$ is a linear-softmax policy. Then,* $\mathrm{MSE}[\mathrm{IS}(\pi_e, H, \boldsymbol{\theta})]$ *is a convex function w.r.t. $\boldsymbol{\theta}$.*

Proof

$$\mathrm{Var}[\mathrm{IS}(\pi_e, H, \boldsymbol{\theta})] = \mathbf{E}[\mathrm{IS}(\pi_e, H, \boldsymbol{\theta})^2 | H \sim \pi_{\boldsymbol{\theta}}] - v(\pi_e)^2$$

We can ignore $v(\pi_e)^2$ since it is a constant and only shifts the objective. Recall from Appendix E, that we can factor trajectory probabilities, $\Pr(H = h|\pi)$, into factors that depend on π and factors that do not: $\Pr(H = h|\pi) = p(h) * w_\pi(h)$.

$$\begin{aligned}
\mathbf{E}[\mathrm{IS}(\pi_e, H, \boldsymbol{\theta})^2 | H \sim \pi_{\boldsymbol{\theta}}] &= \sum_{h \in \mathcal{H}} \Pr(H = h|\pi_{\boldsymbol{\theta}}) \mathrm{IS}(\pi_e, h, \boldsymbol{\theta})^2 \\
&= \sum_{h \in \mathcal{H}} e^{\ln(\Pr(H=h|\pi_{\boldsymbol{\theta}}) \mathrm{IS}(\pi_e, h, \boldsymbol{\theta})^2)} \\
&= \sum_{h \in \mathcal{H}} e^{\ln(w_{\pi_{\boldsymbol{\theta}}}(h) p(h) \frac{w_{\pi_e}(h)^2}{w_{\pi_{\boldsymbol{\theta}}}(h)^2} g(h)^2)} \\
&= \sum_{h \in \mathcal{H}} e^{\ln w_{\pi_{\boldsymbol{\theta}}}(h) + \ln p(h) + \ln w_{\pi_e}(h)^2 + \ln g(h)^2 - \ln w_{\pi_{\boldsymbol{\theta}}}(h)^2} \\
&= \sum_{h \in \mathcal{H}} e^{-\ln w_{\pi_{\boldsymbol{\theta}}}(h) + \underbrace{\ln p(h) + \ln w_{\pi_e}(h)^2 + \ln g(h)^2}_{\text{const w.r.t.} \boldsymbol{\theta}}(h)} \\
&= \sum_{h \in \mathcal{H}} e^{-\ln w_{\pi_{\boldsymbol{\theta}}}(h) + c_1(h)} \\
&= \sum_{h \in \mathcal{H}} c_2(h) e^{-\ln w_{\pi_{\boldsymbol{\theta}}}(h)} \qquad (24)
\end{aligned}$$

where c_1 and c_2 are functions of h that are constant w.r.t. $\boldsymbol{\theta}$. Furthermore, $c_2(h) = e^{c_1(h)}$ and therefore must be positive. We next show that $e^{-\ln w_{\pi_{\boldsymbol{\theta}}}(h)}$ is convex in $\boldsymbol{\theta}$. We then have a linear combination of convex functions with positive weights, which is itself a convex function. Note that we do not have to worry about taking the log of a non-positive value for the following reasons. For any h such that $g(h)$, $w_{\pi_e}(h)$, or $w_{\pi_{\boldsymbol{\theta}}}(h)$ is zero, then $\Pr(H = h|\pi_{\boldsymbol{\theta}}) * \mathrm{IS}(\pi_e, h, \boldsymbol{\theta})^2$ is zero and can be ignored in the summation. The only potential negative value is $g(h)$ but it is squared within the logarithm and can thus be replaced with its absolute value.

We next introduce the following lemma that shows that $-\ln w_{\pi_{\boldsymbol{\theta}}}(h)$ is a convex function with respect to $\boldsymbol{\theta}$.

Lemma 2 *Assume $\pi_{\boldsymbol{\theta}}$ is a linear-softmax policy. Then for $w_{\pi_{\boldsymbol{\theta}}}(h) \coloneqq \prod_{t=0}^{l-1} \pi_{\boldsymbol{\theta}}(a_t|s_t)$, $-\ln w_{\pi_{\boldsymbol{\theta}}}(h)$ is a convex function w.r.t. $\boldsymbol{\theta}$ for any trajectory $h = (s_0, a_0, ..., s_{l-1}, a_{l-1})$.*

Proof

$$\begin{aligned} -\ln w_{\pi_{\boldsymbol{\theta}}}(h) &= -\ln \prod_{t=0}^{l-1} \pi_{\boldsymbol{\theta}}(a_t|s_t) \\ &= \sum_{t=0}^{l-1} -\ln \pi_{\boldsymbol{\theta}}(a_t|s_t). \end{aligned} \tag{25}$$

Next, we show that each $-\ln \pi_{\boldsymbol{\theta}}(a_t|s_t)$ is convex under the linear-softmax policy parameterization:

$$-\ln \pi_{\boldsymbol{\theta}}(a|s) = \ln(\sum_{b \in \mathcal{A}} e^{\boldsymbol{\theta}_b^T \phi(s)}) - \boldsymbol{\theta}_a^T \phi(s)$$

The log-sum-exp function is convex (Boyd et al., 2004, Chapter 3, Example 3.13) and subtracting a linear function does not change convexity. Thus, (25) is a sum of convex functions which is convex. ■

Continuing with the proof of Theorem 4, Lemma 2 implies that $e^{-\ln w_{\pi_{\boldsymbol{\theta}}}(h)}$ is the exponential of a convex function. The exponential of a convex function is convex (Boyd et al., 2004, Chapter 3, Eq 3.11) and thus $e^{-\ln w_{\pi_{\boldsymbol{\theta}}}(h)}$ is convex in $\boldsymbol{\theta}$. Finally, we have that (24) is a linear combination of convex functions with positive weights. Thus, $\mathbf{E}[\mathrm{IS}(\pi_e, H, \boldsymbol{\theta})^2 | H \sim \pi_{\boldsymbol{\theta}}]$ is a convex function which concludes the proof.

■

Appendix E. Minimal-Variance Behavior Policy

In this appendix we prove Proposition 5 that gives a sufficient condition for a minimal-variance behavior policy:

Proposition 5 *Let $w_\pi(h) \coloneqq \prod_{t=0}^{l-1} \pi(a_t|s_t)$. Assume $\exists \tilde{h} \in \mathcal{H}$ such that $g(\tilde{h}) \cdot \Pr(H = \tilde{h}|\pi_e) \neq 0$, i.e., there is non-zero probability that π_e generates a trajectory with non-zero return. If $\exists \pi \in \Pi$ s.t.*

$$\forall h \in \mathcal{H}, w_\pi(h) = |g(h)| \frac{w_{\pi_e}(h)}{\mathbf{E}\left[|g(H)| \,\middle|\, H \sim \pi_e \right]}.$$

then π is a minimal-variance behavior policy.

Proof Recall that we defined $w_\pi(h) \coloneqq \prod_{t=0}^{l-1} \pi(a_t|s_t)$ and define $p(h) \coloneqq d_0(s_0) \prod_{t=1}^{l-1} P(s_t|s_{t-1}, a_{t-1})$. From these definitions, note that $\Pr(H = h|\pi) = w_\pi(h)p(h)$.

The variance of the importance sampling estimator is:

$$\begin{aligned}\mathrm{Var}\left[\mathrm{IS}(\pi_e, H, \pi_b)\right] =& \mathbf{E}\left[\left(\frac{w_{\pi_e}(H)}{w_{\pi_b}(H)}g(H)\right)^2 \;\middle|\; H \sim \pi_b\right] - \mathbf{E}\left[\left(\frac{w_{\pi_e}(H)}{w_{\pi_b}(H)}g(H)\right) \;\middle|\; H \sim \pi_b\right]^2 \quad (26)\\ =& \mathbf{E}\left[\left(\frac{w_{\pi_e}(H)}{w_{\pi_b}(H)}g(H)\right)^2 \;\middle|\; H \sim \pi_b\right] - v(\pi_e)^2, \quad (27)\end{aligned}$$

where (27) follows from (26) since the IS return is unbiased (Thomas, 2015). To prove Proposition 5 we need to find $w_\pi(h)$ for each trajectory, $h \in \mathcal{H}$, such that (27) is minimized subject to the constraints that $\sum_{h\in\mathcal{H}} p(h)w_{\pi_b}(h) = 1$ and $\forall h \in \mathcal{H}$, $w_{\pi_b}(h) > 0$. These constraints enforce that the choices for $w_{\pi_b}(h)$ lead to a valid probability distribution over trajectories.

We ignore $v(\pi_e)^2$ because it is a constant that does not affect the critical points of the variance and arrive at the constrained minimization problem:

$$\begin{aligned}&\min_{w_{\pi_b}} \sum_{h\in\mathcal{H}} \Pr(H = h|\pi_b)\left(\frac{g(h)w_{\pi_e}(h)}{w_{\pi_b}(h)}\right)^2\\ &s.t. \sum_{h\in\mathcal{H}} p(h)w_{\pi_b}(h) = 1\\ &\quad \forall h \in \mathcal{H}, w_{\pi_b}(h) \geq 0\end{aligned}$$

We will consider a relaxed version of this minimization problem that ignores the inequality constraints; as we show, doing so still leads to a feasible solution to the original problem. The Lagrangian for the relaxed constrained minimization problem is:

$$\mathcal{L}(w_{\pi_b}, \lambda) = \sum_{h\in\mathcal{H}} \Pr(h|\pi_b)\left(\frac{g(h)w_{\pi_e}(h)}{w_{\pi_b}(h)}\right)^2 + \lambda\left(\sum_{h\in\mathcal{H}} p(h)w_{\pi_b}(h) - 1\right). \quad (28)$$

Differentiating with respect to $w_{\pi_b}(\tilde{h})$ for any trajectory $\tilde{h}$, we obtain:

$$\frac{\partial}{\partial w_{\pi_b}(\tilde{h})}\mathcal{L}(w_{\pi_b}, \lambda) = -p(\tilde{h})(\frac{g(\tilde{h})w_{\pi_e}(\tilde{h})}{w_{\pi_b}(\tilde{h})})^2 + \lambda p(\tilde{h}).$$

Setting $\frac{\partial}{\partial w_{\pi_b}(\tilde{h})}\mathcal{L}(w_{\pi_b}, \lambda) = 0$, we obtain:

$$\lambda^\star w_{\pi_b{}^\star}(\tilde{h}) = |g(\tilde{h})|w_{\pi_e}(\tilde{h}). \quad (29)$$

Observe that Equation (29) holds $\forall \tilde{h} \in \mathcal{H}$ and thus $\lambda^\star$ must be non-zero since $|g(\tilde{h})|w_{\pi_e}(\tilde{h}) > 0$ for at least one $\tilde{h} \in \mathcal{H}$ by assumption. Thus, we can divide both sides by $\lambda^\star$ to obtain the optimal choice of $w_{\pi_b}(\tilde{h})$:

$$w_{\pi_b{}^\star}(\tilde{h}) = \frac{|g(\tilde{h})|w_{\pi_e}(\tilde{h})}{\lambda^\star}. \quad (30)$$

The constant $\lambda^\star$ can be determined by ensuring the equality constraint is satisfied giving $\lambda^\star = \sum_{h\in\mathcal{H}} p(h)w_{\pi_e}(h)|g(h)| = \mathbf{E}[|g(H)||H \sim \pi_e]$. Furthermore, this form clearly makes $w_{\pi_b}(h)$ positive $\forall h \in \mathcal{H}$, satisfying the constraint that $\forall h \in \mathcal{H}$, $w_{\pi_b}(h) > 0$. Note that in

the case that $g(h) = 0$ for all $h \in \mathcal{H}$ that Proposition 5 gives an undefined value for $w_{\pi_b{}^\star}$. However we can ignore this case as, if $g(h) = 0$ for all $h \in \mathcal{H}$, then the variance of the IS-estimator is trivially zero for any choice of behavior policy.

So far we have found a critical point for the Lagrangian given by (28). In order to establish that this critical point is indeed a global minimum we show that no other choice for w_{π_b} has lower variance than $w_{\pi_b{}^\star}$.

$$\begin{aligned}
\text{Var}[\text{IS}(\pi_e, H, \pi_b{}^\star)] &= \mathbf{E}\left[\left(\frac{w_{\pi_e}(H)}{w_{\pi_b{}^\star}(H)} g(H)\right)^2 \,\middle|\, H \sim \pi_b{}^\star\right] - v(\pi_e)^2 && (31)\\
&= \mathbf{E}\left[|g(H)| \,\middle|\, H \sim \pi_e\right]^2 \mathbf{E}\Big[\underbrace{\left(\frac{g(H)}{|g(H)|}\right)^2}_{=1} \,\Big|\, H \sim \pi_b{}^\star\Big] - v(\pi_e)^2 && (32)\\
&= \mathbf{E}\left[|g(H)| \,\middle|\, H \sim \pi_e\right]^2 - v(\pi_e)^2 && (33)\\
&= \mathbf{E}\left[\frac{w_{\pi_e}(H)}{w_{\pi_b}(H)}|g(H)| \,\middle|\, H \sim \pi_b\right]^2 - v(\pi_e)^2 && (34)\\
&\leq \mathbf{E}\left[\left(\frac{w_{\pi_e}(H)}{w_{\pi_b}(H)}|g(H)|\right)^2 \,\middle|\, H \sim \pi_b\right] - v(\pi_e)^2 \\
&= \text{Var}[\text{IS}(\pi_e, H, \pi_b)] && (35)
\end{aligned}$$

where (32) follows (31) by plugging in the solution for $w_{\pi_b{}^\star}$ given by (30) and factoring out the constant λ, (33) follows from (32) because the expected value of 1 is 1 under any distribution, (34) follows (33) by using importance sampling to change the expectation to be in terms of trajectories from any behavior policy π_b instead of π_e, and the inequality follows from Jensen's inequality. Finally, we can drop the absolute value in (35) because it is squared. Thus we can conclude that $\text{Var}[\text{IS}(\pi_e, H, \pi_b{}^\star)] \leq \text{Var}[\text{IS}(\pi_e, H, \pi_b)]$ for any behavior policy π_b. ■

Appendix F. Behavior Policy Gradient of the KL

In this appendix we derive Theorem 3, which gives the gradient, with respect to the policy parameters, of the KL-divergence between the distribution of trajectories under a minimal-variance behavior policy, $\Pr(H|\pi_b{}^\star)$, and the distribution of trajectories under $\pi_{\boldsymbol{\theta}}$, $\Pr(H|\pi_{\boldsymbol{\theta}})$. This gradient is:

Theorem 3 (Behavior Policy Gradient of the KL-Divergence)

$$\frac{\partial}{\partial \boldsymbol{\theta}} D_{\text{KL}}(\Pr(H|\pi_b{}^\star) || \Pr(H|\pi_{\boldsymbol{\theta}})) \propto \mathbf{E}\left[-\left|\text{IS}(\pi_e, H, \pi_{\boldsymbol{\theta}})\right| \sum_{t=0}^{l-1} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_t|S_t) \,\middle|\, H \sim \pi_{\boldsymbol{\theta}}\right].$$

Proof From Proposition 5 we know that a minimal-variance behavior policy is any policy, $\pi_b{}^\star$, that satisfies the condition:

$$\forall h \in \mathcal{H}, w_{\pi_b{}^\star}(h) = |g(h)| \frac{w_{\pi_e}(h)}{\mathbf{E}[|g(H)||H \sim \pi_e]}.$$

The KL-divergence between two probability distributions p and q with shared support is defined to be $D_{\mathtt{KL}}(p,q) \coloneqq \mathbf{E}[\log(\frac{p(X)}{q(X)})|X \sim p]$. Thus, the KL-divergence between the trajectory distribution of any minimal-variance behavior policy and that of the current behavior policy $\pi_{\boldsymbol{\theta}}$ is given by:

$$\begin{aligned} D_{\mathtt{KL}}(\Pr(H|{\pi_b}^\star)||\Pr(H|\pi_{\boldsymbol{\theta}})) &= \mathbf{E}\left[\log\frac{Pr(H|{\pi_b}^\star)}{\Pr(H|\pi_{\boldsymbol{\theta}})}\middle| H \sim {\pi_b}^\star\right] \\ &= \mathbf{E}\left[\log\frac{w_{{\pi_b}^\star}(H)}{w_{\boldsymbol{\theta}}(H)}\middle| H \sim {\pi_b}^\star\right]. \end{aligned}$$

Using Proposition 5 and defining $\lambda \coloneqq \mathbf{E}[|g(H)||H \sim \pi_e]$, we can expand the $w^\star_{\pi_b}(H)$ terms:

$$\begin{aligned} D_{\mathtt{KL}}(\Pr(H|{\pi_b}^\star)||\Pr(H|\pi_{\boldsymbol{\theta}})) &= \mathbf{E}\Big[\log w_{\pi_e}(H) - \log w_{\pi_{\boldsymbol{\theta}}}(H) + \log|g(H)| - \log\lambda \;\Big|\; H \sim {\pi_b}^\star\Big] \\ &= \sum_{h\in\mathcal{H}} \Pr(h|\pi_e)\frac{|g(h)|}{\lambda}\Big(\log w_{\pi_e}(H) - \log w_{\pi_{\boldsymbol{\theta}}}(H) + \log|g(H)| - \log\lambda\Big) \\ &= \frac{1}{\lambda}\mathbf{E}\Big[|g(H)|\Big(\log w_{\pi_e}(H) - \log w_{\pi_{\boldsymbol{\theta}}}(H) + \log|g(H)| - \log\lambda\Big) \;\Big|\; H \sim \pi_e\Big] \\ &= \frac{1}{\lambda}\mathbf{E}\Big[-|g(H)|\log w_{\pi_{\boldsymbol{\theta}}}(H) \;\Big|\; H \sim \pi_e\Big] \\ &+ \frac{1}{\lambda}\underbrace{\mathbf{E}\Big[|g(H)|(\log w_{\pi_e}(H) + \log|g(H)| - \log\lambda) \;\Big|\; H \sim \pi_e\Big]}_{\text{const w.r.t. } \boldsymbol{\theta}}. \end{aligned}$$

Differentiating with respect to $\boldsymbol{\theta}$, we obtain:

$$\begin{aligned} \frac{\partial}{\partial\boldsymbol{\theta}} D_{\mathtt{KL}}(\Pr(H|{\pi_b}^\star)||\Pr(H|\pi_{\boldsymbol{\theta}})) &\propto \mathbf{E}\left[-|g(H)|\frac{\partial}{\partial\boldsymbol{\theta}}\log w_{\pi_{\boldsymbol{\theta}}}(H)\middle| H \sim \pi_e\right] \\ &= \mathbf{E}\left[-|g(H)|\sum_{t=0}^{l-1}\frac{\partial}{\partial\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(A_t|S_t)\middle| H \sim \pi_e\right] \\ &= \mathbf{E}\left[-|\mathrm{IS}(\pi_e,H,\pi_{\boldsymbol{\theta}})|\sum_{t=0}^{l-1}\frac{\partial}{\partial\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(A_t|S_t)\middle| H \sim \pi_{\boldsymbol{\theta}}\right] \end{aligned}$$

where the second step uses the multi-factor product rule and the final step uses importance sampling to convert from an expectation under π_e to one under $\pi_{\boldsymbol{\theta}}$. ■

Appendix G. Convergence of BPG-KL

In this section, we prove that BPG-KL (Algorithm 2) converges under an appropriately chosen step-size.

Proposition 6 *Under Assumption* 1 *and Assumption* 3*, BPG-KL converges. That is,* $D_{\mathtt{KL}}(\Pr(H|{\pi_b}^\star)||\Pr(H|\pi_{\boldsymbol{\theta}}))$ *converges to a finite value and* $\lim_{i\to\infty}\frac{\partial}{\partial\boldsymbol{\theta}} D_{\mathtt{KL}}(\Pr(H|{\pi_b}^\star)||\Pr(H|\pi_{\boldsymbol{\theta}})) = 0$.

Proof Similar to Proposition 4, the proof follows from an application of Proposition 3 in Bertsekas and Tsitsiklis (2000). A minor nuance for Proposition 6 is that BPG-KL does *not* use unbiased estimates of the true KL-gradient but uses unbiased estimates of an expression that is just proportional to the true KL-gradient. However, the proportionality constant is fixed with respect to $\boldsymbol{\theta}$ and so we can ignore it when showing convergence.

To apply Proposition 3 in Bertsekas and Tsitsiklis (2000), we must show the following conditions:

1. $D_{\mathtt{KL}}(\Pr(H|\pi_b{}^\star)||\Pr(H|\pi_{\boldsymbol{\theta}}))$ is continuously differentiable w.r.t. $\boldsymbol{\theta}$.
2. The gradient of the KL objective, $\frac{\partial}{\partial \boldsymbol{\theta}} D_{\mathtt{KL}}(\Pr(H|\pi_b{}^\star)||\Pr(H|\pi_{\boldsymbol{\theta}}))$, is Lipschitz continuous w.r.t. $\boldsymbol{\theta}$.
3. The gradient estimate used by BPG-KL has bounded variance.

Theorem 3 gives us an expression that is proportional to $\frac{\partial}{\partial \boldsymbol{\theta}} D_{\mathtt{KL}}(\Pr(H|\pi_b{}^\star)||\Pr(H|\pi_{\boldsymbol{\theta}}))$ which can be seen to be continuously differentiable under our assumption that $\pi_{\boldsymbol{\theta}}$ is continuously differentiable and Assumption 1 which implies that $\mathrm{IS}(\pi_e, H, \pi_{\boldsymbol{\theta}})$ is bounded.

We next show that the KL objective has bounded second derivative which implies the Lipschitz continuity of $\frac{\partial}{\partial \boldsymbol{\theta}} D_{\mathtt{KL}}(\Pr(H|\pi_b{}^\star)||\Pr(H|\pi_{\boldsymbol{\theta}}))$.

$$
\begin{aligned}
\frac{\partial^2}{\partial^2 \boldsymbol{\theta}} D_{\mathtt{KL}}(\Pr(H|\pi_b{}^\star)||\Pr(H|\pi_{\boldsymbol{\theta}})) =& \frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{E}\left[-|\mathrm{IS}(\pi_e, H, \pi_{\boldsymbol{\theta}})| \underbrace{\sum_{t=0}^{l-1} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_t|S_t)}_{\frac{\partial}{\partial \boldsymbol{\theta}} \log w_{\boldsymbol{\theta}}(H)} \middle| H \sim \pi_{\boldsymbol{\theta}} \right] \\
=& \frac{\partial}{\partial \boldsymbol{\theta}} \sum_{h \in \mathcal{H}} -p(h) w_{\pi_e}(h)|g(h)| \frac{\partial}{\partial \boldsymbol{\theta}} \log w_{\boldsymbol{\theta}}(h) \\
=& \sum_{h \in \mathcal{H}} -p(h) w_{\pi_e}(h)|g(h)| \frac{\partial^2}{\partial^2 \boldsymbol{\theta}} \log w_{\boldsymbol{\theta}}(h) \\
=& \sum_{h \in \mathcal{H}} -p(h) w_{\pi_e}(h)|g(h)| \sum_{t=0}^{l-1} \frac{\partial}{\partial \boldsymbol{\theta}} \frac{\frac{\partial}{\partial \boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(a_t|s_t)}{\pi_{\boldsymbol{\theta}}(a_t|s_t)} \\
=& \sum_{h \in \mathcal{H}} -p(h) w_{\pi_e}(h)|g(h)| \cdot \\
& \cdot \sum_{t=0}^{l-1} \frac{\pi_{\boldsymbol{\theta}}(a_t|s_t) \frac{\partial^2}{\partial^2 \boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(a_t|s_t) - (\frac{\partial}{\partial \boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(a_t|s_t))^2}{\pi_{\boldsymbol{\theta}}(a_t|s_t)^2}
\end{aligned} \tag{36}
$$

The denominator in (36) cannot be zero as otherwise $w_{\boldsymbol{\theta}}(h)$ would be zero and the corresponding trajectory could be ignored in the expectation. Furthermore, by construct, $\frac{\partial}{\partial \boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(a_t|s_t)$ and $\frac{\partial^2}{\partial^2 \boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(a_t|s_t)$ exist and are bounded. Thus we conclude that the first and second derivative of the KL objective exist and the first derivative is Lipschitz.

Finally, we have to show that the variance of the gradient estimate used by BPG-KL is bounded. To do so, we show that the gradient estimate with any single trajectory is bounded

which implies the variance of the estimates used by BPG-KL is bounded. For any trajectory h, collected by following $\pi_{\boldsymbol{\theta}}$, an unbiased estimate of the KL gradient estimate is given as:

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\theta}} D_{\mathtt{KL}}(\Pr(H|\pi_b{}^\star)||\Pr(H|\pi_{\boldsymbol{\theta}})) &\approx -|\mathrm{IS}(\pi_e, h, \pi_{\boldsymbol{\theta}})| \sum_{t=0}^{l-1} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a_t|s_t) \\ &\overset{(a)}{=} -\frac{w_{\pi_e}(h)|g(h)|}{w_{\pi_{\boldsymbol{\theta}}}(h)} \frac{\partial}{\partial \boldsymbol{\theta}} \log w_{\pi_{\boldsymbol{\theta}}}(h) \\ &\overset{(b)}{=} -\frac{w_{\pi_e}(h)|g(h)|}{w_{\pi_{\boldsymbol{\theta}}}(h)} \frac{\frac{\partial}{\partial \boldsymbol{\theta}} w_{\pi_{\boldsymbol{\theta}}}(h)}{w_{\pi_{\boldsymbol{\theta}}}(h)} \\ &= -\frac{w_{\pi_e}(h)|g(h)|}{w_{\pi_{\boldsymbol{\theta}}}(h)^2} \frac{\partial}{\partial \boldsymbol{\theta}} w_{\pi_{\boldsymbol{\theta}}}(h) \end{aligned} \tag{37}$$

where (a) uses $w_\pi(h) = \prod_{t=0}^{l-1} \pi(a_t|s_t)$ and (b) uses the likelihood-ratio trick. On the RHS of Equation (37), $\frac{w_{\pi_e}(h)|g(h)|}{w_{\pi_{\boldsymbol{\theta}}}(h)^2}$ is bounded under Assumption 1 and $\frac{\partial}{\partial \boldsymbol{\theta}} w_{\pi_{\boldsymbol{\theta}}}(h)$ was shown to be bounded in Equation (22). Thus we conclude that the variance of the gradient estimate used by BPG-KL is bounded. Proposition 6 now follows from Proposition 3 of (Bertsekas and Tsitsiklis, 2000).

■

Appendix H. Convexity of KL-Divergence Objective

In this appendix, we prove that, when $\pi_{\boldsymbol{\theta}}$ is a linear-softmax policy, then the objective minimized by BPG-KL is convex in the policy parameters. Thus, BPG-KL is guaranteed to converge to the parameter vector that minimizes the KL divergence with the minimal-variance behavior policy under standard stochastic gradient descent step-size conditions (Bertsekas and Tsitsiklis, 2000).

Theorem 4 *Assume $\pi_{\boldsymbol{\theta}}$ is a linear-softmax policy. Then, $D_{\mathtt{KL}}(\Pr(H|\pi_b{}^\star)||\Pr(H|\pi_{\boldsymbol{\theta}}))$ is a convex function w.r.t. $\boldsymbol{\theta}$.*

Proof

Recall that for the minimal-variance behavior policy we have:

$$w_{\pi_b{}^\star}(h) = \frac{|g(h)| w_{\pi_e}(h)}{\lambda},$$

where $\lambda = \mathbf{E}[|g(H)||H \sim \pi_{\boldsymbol{\theta}}]$.

The KL-divergence between the minimal-variance behavior policy and policy $\pi_{\boldsymbol{\theta}}$ is given as:

$$\begin{aligned}
D_{\mathtt{KL}}(\Pr(H|\pi_b{}^\star)||\Pr(H|\pi_{\boldsymbol{\theta}})) &= \mathbf{E}\left[\log\frac{Pr(H|\pi_b{}^\star)}{\Pr(H|\pi_{\boldsymbol{\theta}})}\middle| H\sim\pi_b{}^\star\right] \\
&= \mathbf{E}\left[\log\frac{w_{\pi_b{}^\star}(H)}{w_{\pi_{\boldsymbol{\theta}}}(H)}\middle| H\sim\pi_b{}^\star\right]. \\
&= \sum_{h\in\mathcal{H}}\Pr(H=h|\pi_b{}^\star)\log\frac{|g(h)|w_{\pi_e}(h)}{w_{\pi_{\boldsymbol{\theta}}}(h)\lambda} \\
&= \sum_{h\in\mathcal{H}}\underbrace{p(h)\frac{|g(h)|}{\lambda}w_{\pi_e}(h)}_{\text{const w.r.t. }\boldsymbol{\theta}}\log\frac{|g(h)|w_{\pi_e}(h)}{w_{\pi_{\boldsymbol{\theta}}}(h)\lambda} \\
&= \sum_{h\in\mathcal{H}}c_1(h)(\log(\frac{|g(h)|}{\lambda}w_{\pi_e}(h))-\log w_{\pi_{\boldsymbol{\theta}}}(h)) \\
&= \sum_{h\in\mathcal{H}}-c_3(h)\log w_{\pi_{\boldsymbol{\theta}}}(h)+c_2(h) \qquad (38)
\end{aligned}$$

Functions c_1, c_2, and c_3 are positive for any h and constant with respect to $\boldsymbol{\theta}$. Lemma 2 says that $-\log w_{\pi_{\boldsymbol{\theta}}}(h)$ is convex w.r.t $\boldsymbol{\theta}$. Thus, (38) is a weighted sum of convex functions with positive weights which is itself convex. Thus the KL-divergence objective optimized by BPG-KL is convex with respect to $\boldsymbol{\theta}$. An interesting observation from the proof is that the KL-divergence between the trajectory distribution of $\pi_{\boldsymbol{\theta}}$ and any trajectory distribution that does not depend on $\boldsymbol{\theta}$ is also a convex function (under the assumption that $\pi_{\boldsymbol{\theta}}$ is a linear-softmax policy).

■