

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
Edmonton, Alberta, CA*

Peter Stone

PSTONE@CS.UTEXAS.EDU

*Department of Computer Science
The University of Texas at Austin, Sony AI
Austin, TX, USA*

Scott Niekum

SNIEKUM@CS.UTEXAS.EDU

*College of Information and Computer Sciences
University of Massachusetts
Amherst, MA, USA*

Editor: TBD

Abstract

1 We consider the task of evaluating a policy for a *Markov decision process* (MDP). The
2 standard unbiased technique for evaluating a policy is to deploy the policy and observe its
3 performance. We show that the data collected from deploying a different policy, commonly
4 called the *behavior policy*, can be used to produce unbiased estimates with lower mean
5 squared error than this standard technique. We derive an analytic expression for a *minimal*
6 *variance behavior policy* – a behavior policy that minimizes the mean squared error of the
7 resulting estimates. Because this expression depends on terms that are unknown in practice,
8 we propose a novel policy evaluation sub-problem, *behavior policy search*: searching for a
9 behavior policy that reduces mean squared error. We present two behavior policy search
10 algorithms and empirically demonstrate their effectiveness in lowering the mean squared
11 error of policy performance estimates.¹

12 **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) 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

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

59 2. Background

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

64 2.1 Notation

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

73 A policy, $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$, is a function mapping states to probability distributions over
 74 \mathcal{A} . Let Π be the set of all such policies. We use $\pi(a|s) = \pi(s, a)$ to denote the conditional
 75 probability of action a given state s . In this work, we consider parameterized policies, π_{θ} ,
 76 where the distribution over actions given a state is determined by a vector $\theta \in \Theta$, where
 77 $\Theta \subseteq \mathbf{R}^d$ for some dimension d . Furthermore, we require $\pi_{\theta}(a|s)$ to be twice-differentiable
 78 with respect to θ at every state-action pair and for $\frac{\partial}{\partial \theta} \pi_{\theta}(a|s)$ and $\frac{\partial^2}{\partial^2 \theta} \pi_{\theta}(a|s)$ to be bounded
 79 by a finite constant for all states, actions, and values of θ .

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

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

96 2.2 Batch Policy Evaluation

97 In the batch policy evaluation problem, we are given an *evaluation policy*, π_e , for which we
 98 would like to estimate $v(\pi_e)$. We assume there exists a policy parameter vector θ_e such
 99 that $\pi_e = \pi_{\theta_e}$ and that this vector is known. We consider an incremental setting where,
 100 at iteration i , we sample a single trajectory H_i with a policy π_{θ_i} and add (H_i, θ_i) to a set
 101 D . We use D_i to denote the set at iteration i (including (H_i, θ_i)) where $D_0 = \emptyset$. We use
 102 superscripts on states, actions, and rewards to denote the trajectory in which they occur:
 103 $H_i := (S_0^i, A_0^i, R_0^i, \dots, S_{l-1}^i, A_{l-1}^i, R_{l-1}^i)$.

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

$$\text{MSE}[\text{PE}] := \mathbf{E} \left[\left(\text{PE}(D_i) - v(\pi_e) \right)^2 \right],$$

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

112 In this work, we focus on unbiased estimators. An *unbiased* estimator is an estimator
 113 whose estimates have expected value equal to $v(\pi_e)$. For unbiased estimators, minimizing
 114 variance is equivalent to minimizing MSE. While biased estimators (like bootstrapping
 115 methods ([Sutton and Barto, 2018](#), Chapter 6) and approximate models ([Kearns and Singh,](#)
 116 [2002](#))) can sometimes produce lower MSE estimates, some applications may call for unbiased
 117 estimators.

118 The algorithms we introduce only consider the problem of selecting θ_i and estimating
 119 $v(\pi_e)$ to minimize the MSE at iteration i . That is, they do *not* consider how the selection of
 120 θ_i will impact our future ability to select an appropriate θ_j for $j > i$ and thus to produce
 121 more accurate estimates in the future.

122 2.3 Monte Carlo Batch Policy Evaluation

123 Perhaps the simplest batch policy evaluation method is the *on-policy Monte-Carlo* (MC)
 124 estimator. As an on-policy method, the Monte Carlo estimator requires $\theta_i = \theta_e$ for all
 125 iterations i . The estimate of $v(\pi_e)$ at iteration i is the mean return:

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

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

2. Being a strongly consistent estimator of $v(\pi_e)$ means that $\Pr \left(\lim_{i \rightarrow \infty} \overline{\text{MC}}(\pi_e, D_i) = v(\pi_e) \right) = 1$. 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 $\frac{\pi_e(a|s)}{\pi_\theta(a|s)}$ exists and is bounded above by (an unknown) $c < \infty$, $\forall s \in \mathcal{S}, \forall a \in \mathcal{A}, \forall \theta \in \Theta$.*

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 π_θ is bounded away from zero. Under this assumption, the re-weighted IS return of a trajectory, H , sampled from behavior policy π_b is:

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

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:

$$\overline{\text{IS}}(\pi_e, D_i) := \frac{1}{i} \sum_{j=1}^i \text{IS}(\pi_e, H_j, \pi_{\theta_j}).$$

Note that when π_{θ_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

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

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

174 [Frank et al. \(2008\)](#) consider adaptive importance sampling for TD-learning ([Sutton,](#)
 175 [1988](#)) in MDPs. They assume a known probability of a rare event taking place and assume
 176 learning occurs in a simulator where this probability can be changed. They propose two
 177 algorithms that adapt the probability of a rare event and use importance sampling to remove
 178 bias from the distribution shift. These algorithms lead to faster convergence of TD-learning
 179 algorithms. In contrast to this work, we only assume that we know the evaluation policy
 180 and adapt the behavior policy for low variance importance sampling estimates. We also
 181 only consider estimating $v(\pi_e)$ instead of the expected return from all states, i.e., the state
 182 value-function.

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

188 The one work we know of that adapts the behavior policy is the work of [Bouchard et al.](#)
 189 [\(2016\)](#) who adapt the behavior policy to lower the variance of batch policy gradient estimates.
 190 Their algorithm adapts the behavior policy to lower the variance of each component of the
 191 vector-valued off-policy policy gradient estimate for a different, target policy. This approach
 192 is shown to lead to faster learning on a Grid World domain compared to on-policy batch
 193 policy gradient learning. In contrast to this work, we study the problem of batch policy
 194 evaluation of a fixed policy.

195 3.2 Variance Reduction for Policy Evaluation

196 Aside from adaptive importance sampling, other methods exist for lowering the variance
 197 of on-policy estimates. Control variates ([Zinkevich et al., 2006](#); [White and Bowling, 2009](#);
 198 [Jiang and Li, 2016](#); [Thomas and Brunskill, 2016](#)) are a widely used technique for variance
 199 reduction in RL. As we show in Section 9.3, this technique can be used in conjunction with
 200 adaptive importance sampling.

201 [Veness et al. \(2011\)](#) use common random numbers and antithetic variates to lower
 202 the variance of policy evaluation in Monte Carlo tree search (MCTS). These techniques
 203 require the environment to be known and appear to be inapplicable to the general RL policy
 204 evaluation problem. We note that the algorithms we introduce could potentially be applied,

205 in combination with the methods of [Veness et al. \(2011\)](#), to lower the variance of value
 206 estimates in MCTS.

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

216 4. The Behavior Policy Search Problem

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

231 4.1 Motivating Off-Policy Sampling for Lower Variance Importance Sampling

232 To gain intuition for how importance sampling can lower the variance of Monte Carlo returns,
 233 we first examine why importance sampling often increases variance in RL. First, we make
 234 the straightforward observation that any particular behavior policy will induce a particular
 235 distribution over weighted returns and the weighted returns will have some variance under
 236 this distribution. In the case of on-policy sampling, this distribution is just the distribution
 237 of unweighted returns since $\pi_e(a|s) = \pi_b(a|s)$ and all importance weights are equal to one.
 238 Since choosing $\pi_b \neq \pi_e$ means the importance weights themselves have non-zero variance, it
 239 is natural to assume that the variance of the weighted returns can only increase when we
 240 multiply non-zero variance unweighted returns with non-zero variance weights. In fact, this
 241 case often does arise in RL when the behavior policy is dictated by circumstance (e.g., when
 242 using historical logged data) ([Thomas et al., 2015a](#)).

243 Looking closer at why the variance can be magnified under off-policy sampling, we can
 244 see that some importance weights are greater than 1 while others are less than 1. Weights
 245 greater than 1 will magnify the magnitude of the associated return while weights less than
 246 1 will lessen this magnitude. As a consequence, we can see that if we could select π_b such

247 that the largest magnitude unweighted returns receive weights less than 1 and the smallest
 248 magnitude returns received weights greater than 1 then the overall variance of the weighted
 249 returns would decrease relative to the variance of the unweighted returns. In effect, the
 250 spread of possible return values would decrease and hence variance would decrease as well.

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

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

255 Rearranging the terms of this expression yields:

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

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

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

280 The parameters τ_P , τ_R , τ_π allow us to vary the transition entropy, per-state reward
 281 variance, and evaluation policy entropy respectively of the randomly generated MDP- π_e
 282 pairs. Our objective is to see under what settings there exists a behavior policy that lowers
 283 the variance of importance sampling compared to using $\pi_b = \pi_e$. Figure 2 plots variance
 284 reduction as a function of the three task parameters that we vary. In all settings that we

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

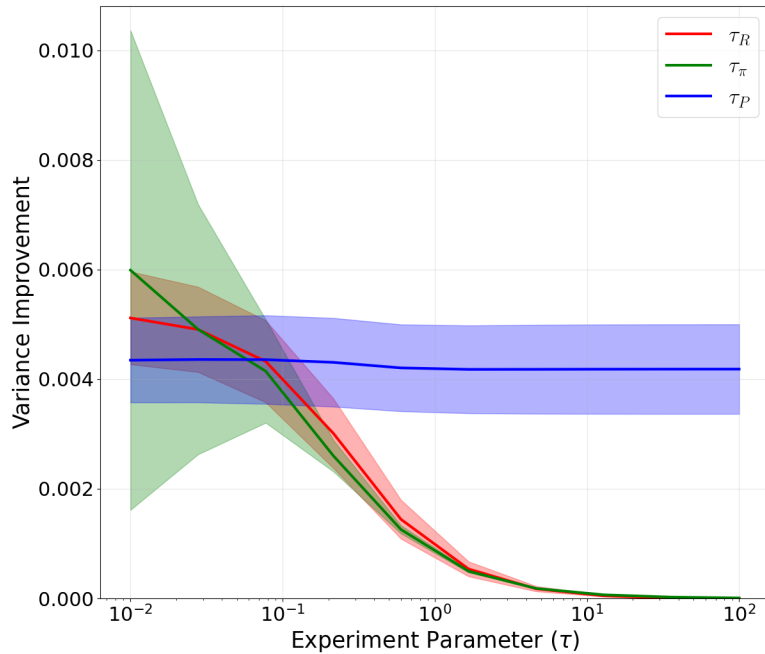


Figure 1: Variance Improvement

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.

285 consider we find that adapting the behavior policy leads to no worse variance than using
 286 π_e though the degree of possible variance reduction varies across settings. In particular,
 287 the three cases where adapting the behavior policy leads to minimal variance reduction
 288 are 1) when the reward function has low variance across actions (high τ_R), 2) when π_e is
 289 near uniform random (higher τ_π), and 3) when π_e is deterministic. The last case sometimes
 290 occurs for the smaller tested τ_π and explains the wider confidence interval see in Figure 2.
 291 While this experiment does not establish there will always be a $\pi_b \neq \pi_e$ that decreases the
 292 variance of importance sampling, it shows that it is in principle possible in some cases to
 293 lower variance by adapting the behavior policy. This finding motivates the behavior policy
 294 search problem which we introduce in the next subsection.

295 4.2 The Behavior Policy Search Problem

296 With the potential to lower the variance of importance sampling via off-policy sampling in
 297 mind, we now introduce the *behavior policy search* (BPS) problem for finding π_b that lowers

298 the MSE of estimates of $v(\pi_e)$. While the previous subsection focused on the IS-estimator,
 299 this subsection considers the more general class of unbiased off-policy value estimators.

300 A BPS problem is defined by the inputs:

- 301 1. An evaluation policy π_e with policy parameters θ_e .
- 302 2. An initial behavior policy, π_{θ_0} , with policy parameters θ_0 . We assume from here on
 303 that $\theta_0 = \theta_e$.
- 304 3. An off-policy policy evaluation estimator, $\text{OPE}(\pi_e, H, \pi_{\theta})$, that takes a trajectory,
 305 $H \sim \pi_{\theta}$ and returns an estimate of $v(\pi_e)$.

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

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

315 In a BPS problem, the i^{th} iteration proceeds as follows. First, given all of the past
 316 behavior policies, $\{\pi_{\theta_j}\}_{j=1}^{i-1}$, and the resulting trajectories, $\{H_j\}_{j=1}^{i-1}$, the BPS algorithm must
 317 select θ_i . The policy π_{θ_i} is run for one episode to generate the trajectory H_i . Then the BPS
 318 algorithm estimates $v(\pi_e)$ as the mean of OPE in the available data, D_i :

$$\overline{\text{OPE}}(\pi_e, D_i) := \frac{1}{i} \sum_{j=1}^i \text{OPE}(\pi_e, H_j, \pi_{\theta_j}).$$

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

325 It is worth noting that adapting the behavior policy increases the computational com-
 326 plexity of estimating $v(\pi_e)$. The exact increase will depend on the behavior policy search
 327 algorithm used and the dimension of θ , however, it seems unlikely that a behavior policy
 328 search algorithm will match the computational simplicity of simply running the evaluation
 329 policy. Thus practitioners must decide whether computational or statistical efficiency is
 330 more appropriate for a particular application.

331 4.3 Statistical Properties of Behavior Policy Search Estimates

332 To enable better statistical efficiency, so far we have focused on reducing variance while
 333 evaluating π_e by adequately adjusting π_b . Before we present concrete algorithms for behavior
 334 policy search, it is important to ensure that any such search procedure does not give up other

335 desired statistical properties like unbiasedness, consistency, and finite sample rates. These
 336 properties when using unbiased estimates are typically established under the assumption
 337 that the trajectories $\{H_j\}_{j=1}^i$ are independent (Thomas, 2015). However, notice that when
 338 using a behavior policy search algorithm the policy parameters $(\theta_j)_{j=1}^i$ will be iteratively
 339 obtained and hence need *not* be independent of each other, and thus even the trajectories
 340 $\{H_j\}_{i=j}^i$ in D_i need not be independent of each other either. Moreover, the distribution
 341 of the random variable $\text{OPE}(\pi_e, H_j, \pi_{\theta_j})$ can vary when θ_j is different for different values
 342 of $j \in \{1, \dots, i\}$ as well. These two factors combined violate *both* the independence and
 343 identical distribution assumptions that are often required to establish statistical guarantees
 on estimators. Figure 3 presents a graphical depiction of the concern.

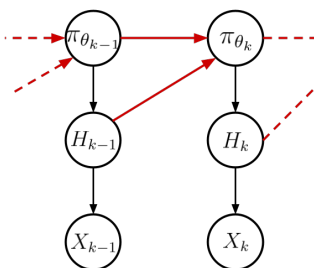


Figure 3: Graphical depiction of the dependencies between the variables. Here, X_k is the random variable corresponding to the estimate $\text{OPE}(\pi_e, H_k, \pi_{\theta_k})$ using the trajectory H_k generated using the behavior policy π_{θ_k} . Red arrows correspond to any learning algorithm that 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.

344

345 This problem occurs even in settings beyond the behavior policy search problem. For
 346 instance, many reinforcement learning methods leverage off-policy trajectories to update
 347 policy parameters, which are consequently used to generate new trajectories. Therefore,
 348 $\overline{\text{OPE}}$ estimates using these trajectories violate the i.i.d. assumption as well. This raises the
 349 question:

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

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

361 For our results to hold for estimates computed as the mean of a set of unbiased estimates,
 362 $\{\text{OPE}(\pi_e, H, \pi_{\theta_j})\}_{j=1}^i$, we require the following assumption.

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

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

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

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

368

Proof The proof is presented in Appendix A. ■

369 An important consequence of Proposition 1 is that, despite lacking independence from each
370 other, the $(\text{OPE}(\pi_e, H_j, \pi_{\theta_j}))_{j=1}^i$ estimates are *uncorrelated*. We formalize this statement
371 below and then use it to establish other properties of the $\overline{\text{OPE}}(\pi_e, D_i)$ estimate.
372

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,*

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

373

Proof The proof is presented in Appendix A. ■

374

Remark 1. *While Lemma 1 implies that the expected value (first moment) of $\text{OPE}(\pi_e, H_j, \pi_{\theta_j})$
376 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
377 on $\text{OPE}(\pi_e, H_k, \pi_{\theta_k})$.
378*

379 A desired property for any estimator is that it provides a more accurate estimate as
380 the amount of data increases. Typically, Kolmogorov's strong law (Sen and Singer, 1993,
381 Theorem 2.3.10) is used to show consistency of estimators, however, it requires random
382 variables to be independent. While the independence (and identical distribution) assumption
383 is violated in our setting, we show below that asymptotic consistency can still be established.
384

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

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

385

Proof The core idea of the proof relies upon results from Proposition 1 and Lemma 1 to
386 show that mean-squared-error of $\overline{\text{OPE}}(\pi_e, D_i)$ asymptotically converges to 0. The complete
387 proof is presented in Appendix A. ■
388

389 While asymptotic consistency is desirable, it is often also essential to quantify finite sample
390 rates to understand the dependency on the sample size, construct confidence intervals, etc.
391 Because i.i.d. assumptions are violated in our setup, it is not immediately clear if existing
392 methods that make the i.i.d. assumption can be leveraged as-is to provide finite sample rates.
393 To resolve this difficulty, we use a common technique based on Martingales to obtain finite
394 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 $j \in \{1, 2, \dots, i\}$, then $\forall \delta \in [0, 1]$,

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

395

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

399

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

403 **Remark 3.** The concentration bound given in Proposition 3 depends upon the range of the
 404 $\text{OPE}(\pi_e, H_j, \pi_{\theta_j})$ estimates. Taking importance sampling as an example and assuming the
 405 returns $g(h)$ are bounded, we can observe that for any $\pi_{\theta_j} \neq \pi_e$ the range of $\text{IS}(\pi_e, H_j, \pi_{\theta_j})$
 406 increases and so the bound becomes looser (see Thomas et al. (2015a) for additional dis-
 407 cussion). Thus, even if a behavior policy search algorithm lowers variance (and thus MSE)
 408 compared to on-policy sampling, Proposition 3 still assigns the estimate a looser finite-sample
 409 bound than the estimate from on-policy sampling. An alternative to Hoeffding-style bounds
 410 are Student’s t-Test bounds which depend on the sample variance. We would expect t-Test
 411 bounds to return a tighter error bound for behavior policy search algorithms that compute
 412 behavior policies that lower the variance of off-policy evaluation. However, t-Test bounds
 413 require the assumption that $\overline{\text{OPE}}(\pi_e, D_i)$ is normally distributed and this assumption is typi-
 414 cally false for small data sets. While this requirement invalidates the error bound, Thomas
 415 et al. (2015b) note that in certain cases t-Test bounds are overly conservative which makes
 416 them suitable for applications of high-confidence off-policy evaluation.

417 Propositions 1, 2, and 3 ensure that the statistical guarantees on unbiasedness, consistency,
 418 and finite sample rates can still be achieved even if any behavior policy search algorithm
 419 results in non i.i.d. returns. In the following sections, we now introduce concrete solution
 420 algorithms for the behavior policy search problem. We will first introduce an algorithm that
 421 optimizes the behavior policy to minimize the variance of an importance sampling estimate.
 422 We then introduce an algorithm that optimizes the behavior policy to minimize a measure of
 423 divergence between a minimal-variance behavior policy and the current behavior policy. We
 424 will also introduce behavior policy search algorithms for extensions to the basic importance
 425 sampling estimator.

426 5. Behavior Policy Gradient on the Variance

427 Our first behavior policy search algorithm is derived from the perspective of selecting
 428 the behavior policy that minimizes the MSE of the importance sampling estimator. As

429 importance sampling is unbiased, minimizing the MSE is equivalent to minimizing variance.
 430 We introduce an analytic expression for the gradient of the MSE of the importance sampling
 431 estimator and a stochastic gradient descent algorithm that adapts π_{θ} to minimize the MSE
 432 between the importance sampling estimate and $v(\pi_e)$. Our algorithm – *behavior policy*
 433 *gradient on the variance* (BPG-V) – begins with on-policy estimates (sets $\theta_0 = \theta_e$) and
 434 adapts the behavior policy with gradient descent on the MSE with respect to θ . The gradient
 435 of the MSE is given by the following theorem:

Theorem 1 (Behavior Policy Gradient of the Variance).

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

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

442

443 BPG-V uses stochastic gradient descent in place of exact gradient descent: replacing the
 444 expectation in Theorem 1 with an unbiased estimate. While in theory, the single trajectory
 445 H_i is sufficient for an unbiased estimate of this gradient, in practice, we can obtain a more
 446 accurate descent direction by sampling a batch, B_i , of k trajectories with π_{θ_i} . In the BPS
 447 setting, sampling a batch of trajectories is equivalent to holding θ fixed for k iterations and
 448 then updating θ with the k most recent trajectories used to compute the gradient estimate.⁴

449 Full details of BPG-V are given in Algorithm 1. At iteration i , BPG-V samples a batch,
 450 B_i , of k trajectories with π_{θ_i} and adds $\{(H_{i:k+j}, \pi_{\theta_i})_{j=1}^k\}$ to D_{i-1} to yield data set D_i (Lines
 451 4 – 5). Then BPG-V updates θ_i with an empirical estimate of the expectation in Theorem 1
 452 (Line 6). After n iterations, BPG-V returns an estimate of $v(\pi_e)$ (Line 8) given as:

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

453 As a behavior policy search algorithm, these BPG-V estimates inherit all the results shown
 454 in Section 4.2: unbiasedness, consistency, finite-sample rates, and independence between
 455 $\text{IS}(\pi_e, H_i, \pi_{\theta_i})$ and $\text{IS}(\pi_e, H_j, \pi_{\theta_j})$ for any two iterations i and j .

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

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.

Algorithm 1 Behavior Policy Gradient on the Variance

Input: Evaluation policy parameters, θ_e , batch size k , a step-size for each iteration, α_i , and number of iterations n .

Output: Final behavior policy parameters θ_n and the IS estimate of $v(\pi_e)$ using all sampled trajectories.

```

1:  $\theta_0 \leftarrow \theta_e$ 
2:  $D_0 = \{\}$ 
3: for all  $i \in 0..n$  do
4:    $B_i = \text{Sample } k \text{ trajectories } H \sim \pi_{\theta_i}$ 
5:    $D_{i+1} = D_i \cup B_i$ 
6:    $\theta_{i+1} = \theta_i + \frac{\alpha_i}{k} \sum_{j=1}^k \text{IS}(\pi_e, H_j, \pi_{\theta_i})^2 \sum_{t=0}^{l-1} \frac{\partial}{\partial \theta} \log \pi_{\theta_i}(A_t^j | S_t^j)$ 
7: end for
8: Return  $\theta_n, \overline{\text{IS}}(\pi_e, D_n)$ 

```

462 **Convergence of BPG-V**

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

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

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

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

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

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

472
473 With further assumptions on the policy class of π_{θ} we can derive stronger convergence
474 guarantees. In particular, if θ is the parameters of a linear-softmax policy than the MSE
475 objective is convex with respect to θ and local minima of the MSE are also global minima.
476 A linear-softmax policy is a policy over a finite set of actions where the probability of each
477 action is defined as a softmax distribution with logits from a linear combination of state
478 features. Formally, let $\phi : \mathcal{S} \rightarrow \mathbf{R}^q$ for integer q be a state feature function that maps
479 states to feature vectors. For each action, $a \in \mathcal{A}$, we have a vector $\theta_a \in \mathbf{R}^q$ and θ is the
480 concatenation of all θ_a . A linear-softmax policy defines the probability of action a in state s
481 as:

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

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

484 **Proof** See Appendix D. ■

485

486 **Remark 4.** The result that the MSE of the importance-sampled return is a convex function
 487 of θ is somewhat surprising given that the mean return is a non-convex function under the
 488 same assumption of linear-softmax policies (Agarwal et al., 2019).

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

495 6. Behavior Policy Gradient on the KL-Divergence

496 The preceding section derived an algorithm that searches for a lower variance behavior
 497 policy by incrementally decreasing the variance with stochastic gradient descent. In this
 498 section, we explore an alternative approach to finding a behavior policy that minimizes
 499 variance. Specifically, we first derive a sufficient condition for a behavior policy to minimize
 500 the variance of the importance sampling estimator. We then introduce an algorithm that
 501 searches for a behavior policy π_θ that comes closest to satisfying this condition.

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

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

507 **Proposition 5.** Let $w_\pi(h) := \prod_{t=0}^{l-1} \pi(a_t|s_t)$. Assume $\exists \tilde{h} \in \mathcal{H}$ such that $g(\tilde{h}) \cdot \Pr(H =$
 508 $\tilde{h}|\pi_e) \neq 0$, i.e., there is non-zero probability that π_e generates a trajectory with non-zero
 509 return. If $\exists \pi \in \Pi$ s.t.

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

510 then π is a minimal-variance behavior policy.

511 **Proof** See Appendix E for a full proof. ■

512

513 We now introduce a second algorithm that attempts to find π_θ that comes closest to
 514 satisfying the condition given in Proposition 5. Note that a policy, π_b^* , that satisfies this
 515 expression will induce the following distribution over trajectories:

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

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

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

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

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

529 **Proof** See Appendix F for full proof. ■

530
 531 Theorem 3 gives a similar gradient to that in Theorem 1, except it takes the absolute
 532 value of $\text{IS}(\pi_e, H, \pi_\theta)$ instead of squaring it. Like BPG-V, BPG-KL begins with on-policy
 533 estimates and adapts the behavior policy with gradient descent on the KL-divergence with
 534 respect to θ . Pseudo-code for the BPG-KL algorithm is given in Algorithm 2. The only
 535 difference between BPG-V and BPG-KL is the method of adapting the behavior policy (Line
 536 6); both algorithms still use importance sampling as the underlying off-policy estimator
 537 to return estimates of $v(\pi_e)$. As a behavior policy search algorithm, BPG-KL inherits the
 538 unbiasedness, consistency, and finite-sample rates given by Proposition 1, Proposition 2, and
 539 Proposition 3 respectively.

540 Convergence of BPG-KL

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

543 **Proposition 6.** *Under Assumption 1 and Assumption 3, BPG-KL converges. That is,*
 544 $D_{\text{KL}}(\Pr(H|\pi_b^*) || \Pr(H|\pi_\theta))$ *converges to a finite value and* $\lim_{i \rightarrow \infty} \frac{\partial}{\partial \theta} D_{\text{KL}}(\Pr(H|\pi_b^*) || \Pr(H|\pi_\theta)) =$
 545 0 .

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

549
 550 Additionally, we can show convexity of the KL-objective under an assumption of linear-
 551 softmax policies.

Algorithm 2 Behavior Policy Gradient on the KL-Divergence

Input: Evaluation policy parameters, θ_e , batch size k , a step-size for each iteration, α_i , and number of iterations n .

Output: Final behavior policy parameters θ_n and the IS estimate of $v(\pi_e)$ using all sampled trajectories.

```

1:  $\theta_0 \leftarrow \theta_e$ 
2:  $D_0 = \{\}$ 
3: for all  $i \in 0 \dots n$  do
4:    $B_i = \text{Sample } k \text{ trajectories } H \sim \pi_{\theta_i}$ 
5:    $D_{i+1} = D_i \cup B_i$ 
6:    $\theta_{i+1} = \theta_i + \frac{\alpha_i}{k} \sum_{j=1}^k |\text{IS}(\pi_e, H_j, \pi_{\theta_i})| \sum_{t=0}^{l-1} \frac{\partial}{\partial \theta} \log \pi_{\theta_i}(A_t^j | S_t^j)$ 
7: end for
8: Return  $\theta_n, \overline{\text{IS}}(\pi_e, D_n)$ 

```

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

554 **Proof** See Appendix H. ■

555

556 Proposition 6 and Theorem 4 jointly imply convergence to a global minimum (Zinkevich,
 557 2003). A counterintuitive observation is that global minimization of the KL-objective does
 558 not necessarily imply that BPG-KL converges to lower variance importance-sampled returns
 559 compared to on-policy sampling. First, observe that, since we minimize the KL between the
 560 minimal-variance behavior policy and a policy within a specific family of behavior policies,
 561 we may not converge to a minimal-variance behavior policy (which may be unrepresentable in
 562 the family of linear soft-max policies). While the policy at BPG-KL’s convergence would be
 563 closer *in terms of KL* to a minimal-variance behavior policy than any other linear soft-max
 564 policy, we have not ruled out the possibility that the policy would yield sub-optimal variance
 565 for the importance sampling returns. While this case may be theoretically possible, our
 566 experimental results (in Section 9) show that BPG-V and BPG-KL perform similarly in
 567 practice, suggesting that minimizing the KL also minimizes variance compared to on-policy
 568 sampling in practice.

569 7. Interpreting BPG-V and BPG-KL Updates

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

575 gradient ascent on $v(\pi_\theta)$ using unbiased estimates of the gradient of $v(\pi_\theta)$:

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

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

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

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

592 The main difference between the two algorithms is that BPG-V puts more emphasis on
 593 increasing the probability of high magnitude return trajectories. For example if one return
 594 has double the return of another then it has quadruple the emphasis under BPG-V whereas
 595 with BPG-KL doubling the return only doubles the emphasis. BPG-V is minimizing our
 596 target objective (low MSE) while BPG-KL attempts to find a policy that is close (in terms
 597 of KL-divergence) to the optimal solution to our target objective.

598 8. Behavior Policy Search for Importance Sampling Extensions

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

604 8.1 Baselined Importance Sampling

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

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

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

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

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

615 8.2 Doubly Robust and Per-Decision Importance Sampling

616 In cases where an approximate model of the environment is available, the *doubly robust*
617 (DR) estimator (Jiang and Li, 2016; Thomas and Brunskill, 2016) lowers the variance of
618 importance sampling using the control variate technique (Lemieux, 2014). In this section,
619 we introduce a behavior policy search algorithm that uses the DR estimator for estimates of
620 $v(\pi_e)$.

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

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

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

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

633 The MSE gradient for the DR estimator is given by the following corollary to Theorem 1:
634

Corollary 1.

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

635 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 $H \sim \pi_\theta$.

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

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

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

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

Corollary 2.

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

649 where the expectation is taken over $H \sim \pi_\theta$.

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

652

653 8.3 Weighted Importance Sampling

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

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

658 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.

659 For finite sample sizes, weighted importance sampling is a biased estimator, however,
 660 it lowers variance due to the importance weights themselves. Though often noted to lower
 661 variance compared to the basic importance sampling estimator (Thomas et al., 2015b;
 662 Mahmood et al., 2014), if the behavior policy is optimized for basic importance sampling,
 663 then it may *harm* the efficiency of policy evaluation. We illustrate this fact with an example.

664 Consider a two-armed bandit problem in which the policy selects arm 1 with probability θ
 665 and arm 2 with probability $1 - \theta$. Let the outcome of pulling arm 1 be a reward of 100 and
 666 the outcome of arm 2 be a reward of 1. The evaluation policy is defined as $\theta_e := 0.1$. The
 667 minimal-variance behavior policy for the basic importance sampling estimator (computed
 668 with (2)) is $\theta_b^* \approx 0.917$.

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

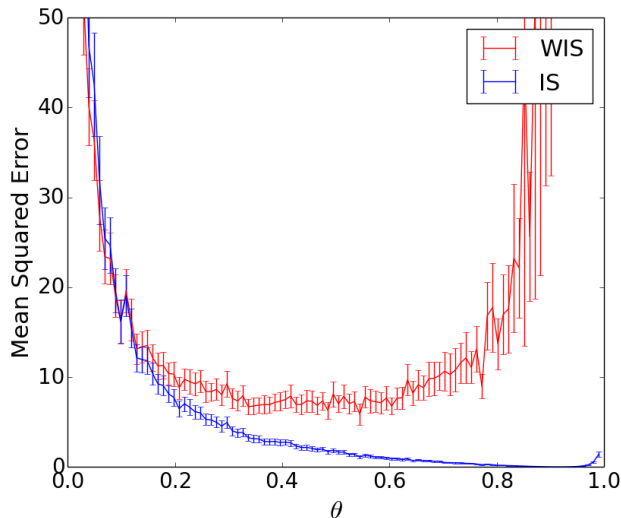


Figure 4: An example where optimizing the behavior policy for the MSE of the basic IS estimator increases the MSE of WIS. For 100 values of θ evenly spaced between 0.01 and 0.99, a data set of size 50 is collected and both the IS estimate and WIS estimate are computed and the squared error calculated. 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.

679 9. Empirical Study

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

- 682 • Can behavior policy search with BPG-V and BPG-KL reduce the MSE of batch policy
683 evaluation compared to on-policy estimates in both tabular and continuous domains?
- 684 • Does adapting the behavior policy of the doubly robust estimator with DR-BPG lower
685 MSE compared to the on-policy doubly robust estimator?
- 686 • Does the rareness of actions that cause high magnitude rewards affect the performance
687 gap between BPG-V and Monte Carlo estimates?

688 9.1 Empirical Set-up

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

691 **Grid World** The first domain is the Grid World
692 domain showed in Figure 5. All grid locations without
693 a reward shown have a reward of -1 . The action
694 set contains the four cardinal directions and actions
695 move the agent in its intended direction (except when
696 moving into a wall, which produces no movement).
697 The agent begins in $(0, 0)$, $\gamma = 1$, and $l = 100$. Each
698 state-action pair, (s, a) , has a parameter $\theta_{s,a}$ and the
699 probability of taking action a in state s is given by
700 the softmax distribution:

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

701 In this domain it is unnecessary to represent the policy with function approximation and
702 we can study BPG-V and BPG-KL without concern of whether our class of function
703 approximator includes a lower variance behavior policy. We obtain two evaluation policies
704 by applying a simple REINFORCE algorithm to maximize the expected return, starting
705 from a policy that selects actions uniformly at random. We then select one evaluation policy
706 from the early stages of learning – an improved policy but still far from converged –, π_1 ,
707 and one after learning has converged, π_2 . We run our set of experiments once with $\pi_e := \pi_1$
708 and a second time with $\pi_e := \pi_2$. The ground truth value of $v(\pi_e)$ is computed with value
709 iteration for both choices of π_e .

710 **Control Tasks** We also study BPG-V and BPG-KL on four tasks with real-valued state
711 variables. The first two of these are the continuous control Cart Pole Swing Up and Acrobot
712 tasks implemented within RLLAB (Duan et al., 2016), the third task is the Cart Pole task
713 from OpenAI Gym (Brockman et al., 2016), and the final task is the PyBullet (Coumans
714 and Bai, 2016–2019) variant of the Hopper domain from OpenAI gym (Brockman et al.,
715 2016). In contrast to the tabular Grid World domain, these domains require that BPG-V
716 and BPG-KL optimize the behavior policy within a given class of function approximator.
717 For Cart Pole Swing Up and Acrobot, π_e is a two layer neural network with 32 tanh units
718 per layer that maps the state to the mean of a Gaussian distribution over the continuous
719 action space. For Cart Pole Swing Up, π_e was learned using 10 iterations of the TRPO
720 algorithm (Schulman et al., 2015) applied to a randomly initialized policy. For Acrobot, π_e

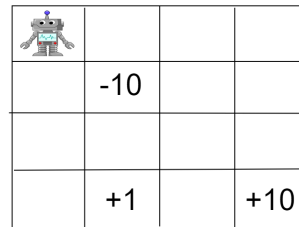


Figure 5: Grid World Domain

721 was learned using 60 iterations. For Cart Pole and Hopper, π_e is a neural network with two
 722 layers of 64 tanh hidden units in each layer and is trained using 200 iterations of proximal
 723 policy optimization (Schulman et al., 2017). For Cart Pole the network maps the state to a
 724 softmax distribution over actions while in Hopper the network maps the state to a Gaussian
 725 distribution over the continuous-valued actions. For Cart Pole Swing Up and Acrobot we use
 726 $l = 50$ and $\gamma = 1$; CartPole and Hopper use $l = 200$ (with early termination possible) and
 727 $\gamma = 1$. For step-size selection at each iteration BPG-V and BPG-KL use the largest possible
 728 step-size subject to a constraint on the KL-divergence between the old and new policy. This
 729 type of update has been shown to be more stable than constant step-size updates in the
 730 policy gradient RL literature (Kakade, 2001; Peters and Schaal, 2008; Schulman et al., 2015).
 731 The ground truth value of $v(\pi_e)$ in all domains is computed with 1,000,000 Monte Carlo
 732 roll-outs.

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

$$\mathbf{E} \left[-\text{IS}(\pi_e, H, \pi_{\theta_{i-1}})^2 \mid H \sim \pi_{\theta_{i-1}} \right]$$

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

$$\mathbf{E} \left[-|\text{IS}(\pi_e, H, \pi_{\theta_{i-1}})| \mid H \sim \pi_{\theta_{i-1}} \right].$$

736 The baseline b_i is estimated with trajectories from iteration $i - 1$ where for the first iteration
 737 $b_i = 0$. The gradient with baseline for BPG-V is an estimate of:

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

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

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

739 Adding or subtracting a constant leaves the gradient unchanged in expectation since
 740 $b_i \mathbf{E} \left[\sum_{t=0}^{l-1} \frac{\partial}{\partial \theta} \log \pi_{\theta}(A_t | S_t) \right] = 0$. However, the baseline variants of BPG-V and BPG-KL
 741 have lower variance gradient estimates so that the estimated gradient is closer in direction
 742 to the true gradient. Note that this baseline is for gradient estimation and is different than
 743 using a constant baseline for importance sampling.

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

749 9.2 Main Results

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

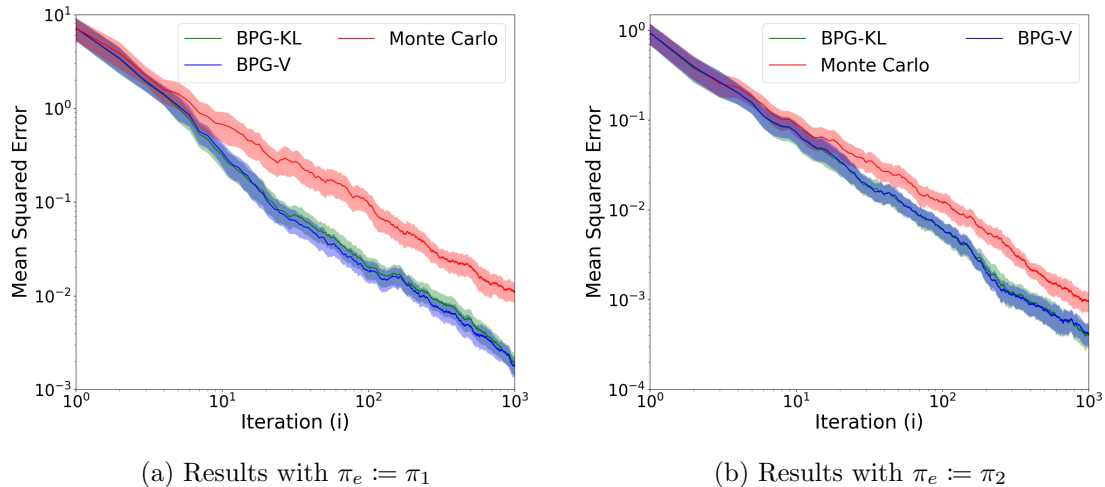


Figure 6: Grid World experiments when π_e is a partially optimized policy, π_1 , (6a) and a converged policy, π_2 , (6b). Results are averaged over 100 trials of 1000 iterations with a shaded region representing a 95% confidence interval. The vertical axis shows the mean squared error and the horizontal 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, $p < 0.05$).

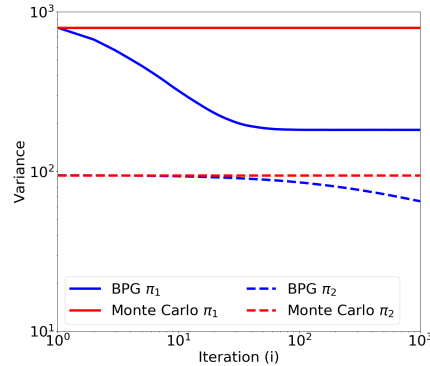
752 9.2.1 GRID WORLD

753 Figure 6 compares BPG-V, BPG-KL, and the on-policy Monte Carlo estimator for both Grid
 754 World policies, π_1 and π_2 . At each iteration, each method collects 100 additional trajectories.
 755 BPG-V gradient estimates will tend to have a different magnitude than BPG-KL gradient
 756 estimates because the importance-sampled return is squared instead of its absolute value
 757 taken. We normalize the gradient estimates to have magnitude one and use a step-size of
 758 0.1 for both methods in order to have similar magnitude behavior policy changes for each
 759 method.

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

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

770 To understand the disparity in performance when π_e changes, we plot the variance of the
 771 Monte Carlo return under π_e (Figures 7b and 7c). These plots show the variance of π_1 is
 772 much higher; it sometimes samples returns with twice the magnitude of any sampled by π_2 .



(a) Variance Reduction

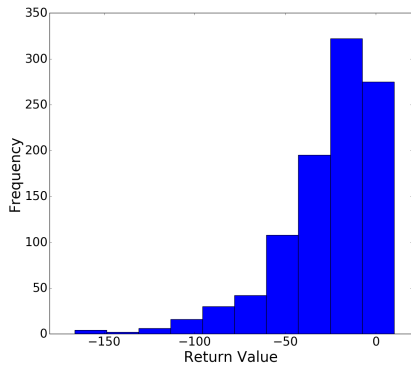
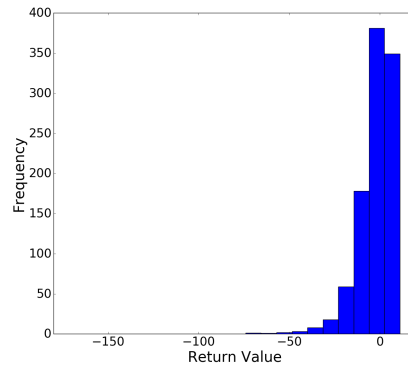
(b) Histogram of returns using π_1 (c) Histogram of returns using π_2

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 .

773 To quantify the decrease in variance from behavior policy search, we also measure and plot
 774 the variance of $\text{IS}(\pi_e, H, \pi_{\theta_i})$ for the BPG-V algorithm (Figure 7a). Figure 7a shows much
 775 higher initial variance for importance sampling evaluation of π_1 . The high initial variance
 776 means there is much more room for BPG-V and BPG-KL to improve the behavior policy
 777 when θ_e is the partially optimized policy, π_1 .

778 BPG-V and BPG-KL require setting two parameters for the stochastic gradient descent
 779 update: a step-size, α , and a batch-size, k . We ablate these parameters to test the sensitivity
 780 of performance to their values. Again, we use normalized gradient estimates to ensure
 781 comparability of the algorithms given the same step-size.

782 To ablate step-size, we run each algorithm for 1000 iterations with a batch-size of $k = 100$
 783 for different settings of α . Our point of comparison is the MSE of the estimate at the final
 784 iteration. Figure 8a shows that both BPG-V and BPG-KL perform as well as or better than
 785 Monte Carlo for a wide range of step-size values. However, for very high values ($\alpha = 5$ or
 786 $\alpha = 10$), the estimates may diverge.

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

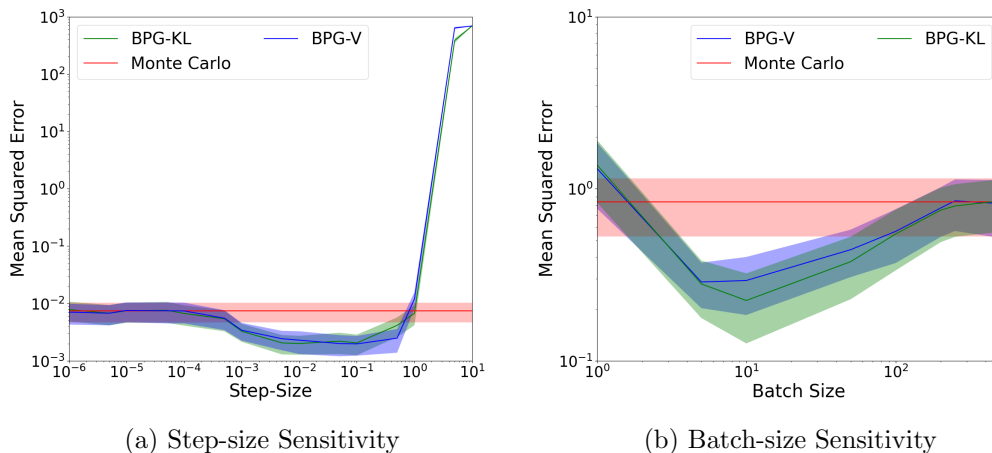
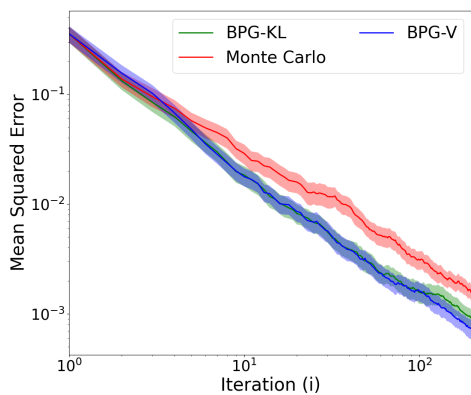


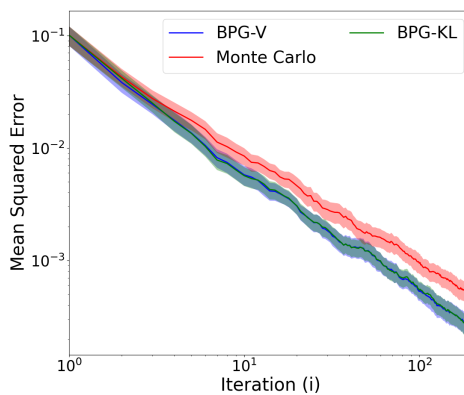
Figure 8: BPG-V and BPG-KL parameter sensitivity. Figure 8a shows performance as a function of the algorithm step-size, α , and Figure 8b shows performance as a function of the algorithm batch-size, k . In both figures the vertical axis is mean squared error of the importance sampling estimate. The horizontal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

794 9.2.2 CONTROL TASKS

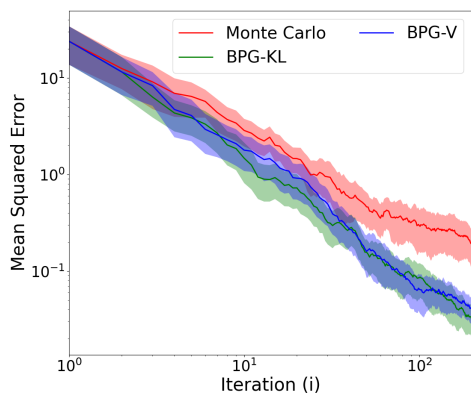
795 Figure 9 shows reduction of MSE on the Cart Pole Swing Up Acrobot, Cart Pole, and
 796 Hopper domains. Each method uses a step-size of 5×10^{-5} . Again we see that both BPG-V
 797 and BPG-KL reduce MSE faster than Monte Carlo value estimation and that both methods
 798 perform similarly to one another. In contrast to the discrete Grid World experiment, these
 799 experiment demonstrates the applicability of BPG-V and BPG-KL to both continuous states
 800 and actions. These results also demonstrates that BPG-V and BPG-KL (and more generally
 801 behavior policy search) can lower the variance of batch policy evaluation when the policy
 802 must generalize across different states and actions.



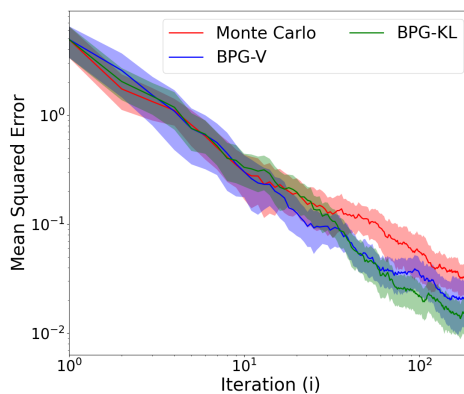
(a) Cart Pole Swing Up MSE.



(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 $\theta_i = \theta_e$ 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

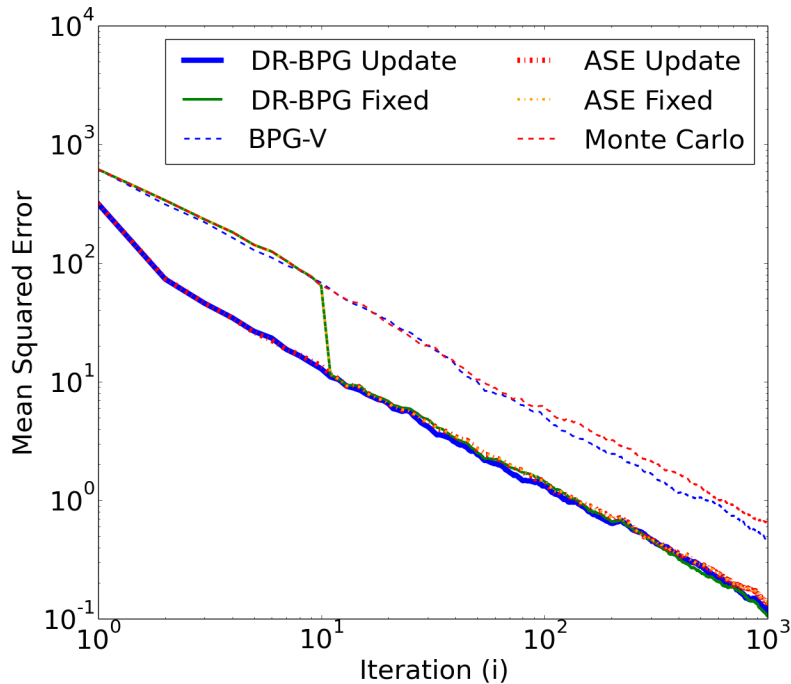


Figure 10: Comparison of DR-BPG and ASE (on-policy DR) on a larger stochastic Grid World. For the fixed model methods, the significant drop in MSE at iteration 10 is due to the introduction of the model control variate. For visual clarity we omit error bars. The mean difference between the final estimate 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.

846 between on- and off-policy policy evaluation. For this experiment, we use a variant of the
 847 deterministic Grid World where taking the UP action in the initial state (the upper left
 848 corner) causes a transition to the terminal state with a reward of +50. We use π_1 from
 849 our earlier Grid World experiments but we vary the probability, p , of choosing UP when in
 850 the initial state, i.e., with probability p the agent will receive a large reward and end the
 851 trajectory. We use BPG-V with a step-size of 10^{-5} and unnormalized gradient estimates as
 852 the behavior policy search algorithm for all values of p . We plot the relative decrease of the
 853 variance after 500 iterations as a function of p over 100 trials for each value of p . We use
 854 relative variance to normalize across problem instances. Note that under this measure, even
 855 when p is close to 1, the relative variance remains greater than zero because as p approaches
 856 1 the initial variance also goes to zero.

857 This experiment illustrates that as the initial variance increases, the amount of improve-
 858 ment BPG-V can achieve increases. As p becomes closer to 1, the rare high magnitude
 859 event becomes less rare and the initial variance becomes closer to zero. When this happens,
 860 BPG-V barely improves over the variance of Monte Carlo (in terms of absolute variance
 861 there is no improvement). When π_e rarely takes the high rewarding UP action (p close to 0),

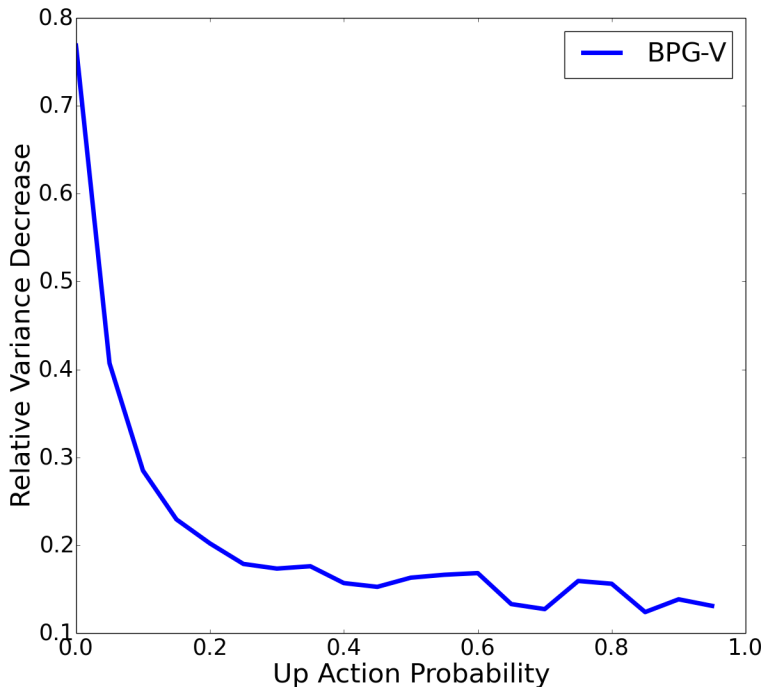


Figure 11: Varying the probability of a high rewarding terminal action in the Grid World domain. Each point on the horizontal axis is the probability of taking this action. The vertical axis gives the mean relative decrease in variance after adapting θ for 500 iterations. Denoting the initial variance 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.

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

865 10. Discussion

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

875 It is also necessary that the evaluation policy contributes to the variance of the returns.
 876 If all variance is due to the environment then it seems unlikely that BPG-V or BPG-KL
 877 will offer much improvement. For example, [Ciosek and Whiteson \(2017\)](#) consider a variant
 878 of the Mountain Car task ([Singh and Sutton, 1996](#)) where the dynamics can trigger a rare
 879 event – independent of the action – in which rewards are multiplied by 1000. No behavior
 880 policy adaptation can lower the variance due to this event.

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

887 When to Perform Behavior Policy Search?

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

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

905 11. Future Work

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

911 11.1 Evaluating Multiple Evaluation Policies

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

915 policy that provides low variance importance sampling evaluation of that evaluation policy.
 916 An important direction for future work is to develop behavior policy search algorithms that
 917 optimize the behavior policy for a *set* of evaluation policies. A straightforward way to adapt
 918 either BPG-KL or BPG-V to multiple evaluation policies is to use a linear combination
 919 of the objective they minimize for each evaluation policy. For example, BPG-V minimizes
 920 $\mathbf{E}[\text{IS}(\pi_e, H, \pi_\theta)^2 | H \sim \pi_\theta]$ for the single evaluation policy π_e . If instead, we wished to
 921 minimize the variance of evaluating a set of policies, $\{\pi_1, \dots, \pi_m\}$, a multi-policy variant of
 922 BPG-V could minimize $\sum_{j=1}^m \mu(\pi_j) \mathbf{E}[\text{IS}(\pi_j, H, \pi_\theta)^2 | H \sim \pi_\theta]$ where, we define $\mu(\pi_j)$ to be
 923 an emphasis factor that provides the relative importance of evaluating each policy in the
 924 set of evaluation policies. This approach would be straightforward, however, it might be
 925 the case that lowering the variance for one evaluation policy might increase the variance of
 926 evaluating another.

927 11.2 Behavior Policy Search for Value Function Learning

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

942 11.3 Behavior Policy Search for Policy Improvement

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

950 11.4 Theoretical Variance Reduction

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

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

961 12. Conclusion

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

974 Acknowledgments

975 We thank Daniel Brown and the anonymous reviewers of the earlier version of this work
 976 for insightful comments on the work and its presentation. This work has taken place in
 977 the Learning Agents Research Group (LARG) and the Personal Autonomous Robotics
 978 Lab (PeARL) at the Artificial Intelligence Laboratory, The University of Texas at Austin.
 979 LARG research is supported in part by grants from the National Science Foundation (CPS-
 980 1739964, IIS-1724157, NRI-1925082), the Office of Naval Research (N00014-18-2243), Future
 981 of Life Institute (RFP2-000), Army Research Office (W911NF-19-2-0333), DARPA, Lockheed
 982 Martin, General Motors, and Bosch. PeARL research is supported in part by the NSF
 983 (IIS-1724157, IIS-1638107, IIS-1749204, IIS-1925082) and ONR (N00014-18-2243). The
 984 views and conclusions contained in this document are those of the authors alone.

985 References

- 986 Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. Reinforcement learning: Theory and
 987 algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 2019.
- 988 Ahamed, T. I., Borkar, V. S., and Juneja, S. Adaptive importance sampling technique for
 989 Markov chains using stochastic approximation. *Operations Research*, 54(3):489–504, 2006.
- 990 Akyildiz, Ö. D. and Míguez, J. Convergence rates for optimised adaptive importance
 991 samplers. *Statistics and Computing*, 31:1–17, 2021.

- 992 Archibald, T., McKinnon, K., and Thomas, L. On the generation of markov decision
993 processes. *Journal of the Operational Research Society*, 46(3):354–361, 1995.
- 994 Asis, K. D., Hernandez-Garcia, J. F., Holland, G. Z., and Sutton, R. S. Multi-step
995 reinforcement learning: A unifying algorithm, 2017.
- 996 Azuma, K. Weighted sums of certain dependent random variables. *Tohoku Mathematical*
997 *Journal, Second Series*, 19(3):357–367, 1967.
- 998 Bastani, M. Model-free intelligent diabetes management using machine learning. Master’s
999 thesis, Department of Computing Science, University of Alberta, 2014.
- 1000 Bertsekas, D. P. and Tsitsiklis, J. N. Gradient convergence in gradient methods with errors.
1001 *SIAM Journal on Optimization*, 10:627–642, 2000.
- 1002 Bouchard, G., Trouillon, T., Perez, J., and Gaidon, A. Online learning to sample. *arXiv*
1003 *preprint arXiv:1506.09016*, 2016.
- 1004 Boyd, S., Boyd, S. P., and Vandenberghe, L. *Convex optimization*. Cambridge university
1005 press, 2004.
- 1006 Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and
1007 Zaremba, W. OpenAI gym. *arXiv preprint arXiv:1606.01540*, 2016.
- 1008 Ciosek, K. and Whiteson, S. OFFER: Off-environment reinforcement learning. In *Proceedings*
1009 *of the 31st AAAI Conference on Artificial Intelligence (AAAI)*, 2017.
- 1010 Coumans, E. and Bai, Y. Pybullet, a python module for physics simulation for games,
1011 robotics, and machine learning. <http://pybullet.org>, 2016–2019.
- 1012 Desai, P. Y. and Glynn, P. W. Simulation in optimization and optimization in simulation:
1013 A Markov chain perspective on adaptive Monte Carlo algorithms. In *Proceedings of the*
1014 *33rd conference on Winter simulation*, pages 379–384. IEEE Computer Society, 2001.
- 1015 Duan, Y., Chen, X., Houthoofd, R., Schulman, J., and Abbeel, P. Benchmarking deep
1016 reinforcement learning for continuous control. In *Proceedings of the 33rd International*
1017 *Conference on Machine Learning*, 2016.
- 1018 Frank, J., Mannor, S., and Precup, D. Reinforcement learning in the presence of rare events.
1019 In *Proceedings of the 25th International Conference on Machine Learning*, pages 336–343.
1020 ACM, 2008.
- 1021 Gelada, C. and Bellemare, M. G. Off-policy deep reinforcement learning by bootstrapping
1022 the covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*,
1023 volume 33, pages 3647–3655, 2019.
- 1024 Guo, Z. D., Thomas, P. S., and Brunskill, E. Using options and covariance testing for long
1025 horizon off-policy policy evaluation. In *Proceedings of the 31st Conference on Advances in*
1026 *Neural Information Processing Systems (NeurIPS)*, 2017.

- 1027 Hallak, A. and Mannor, S. Consistent on-line off-policy evaluation. In *Proceedings of the*
1028 *34th International Conference on Machine Learning*, pages 1372–1383, 2017.
- 1029 Hammersley, J. and Handscomb, D. Monte Carlo methods. *Ltd., London*, page 40, 1964.
- 1030 Hanna, J. P., Thomas, P., Stone, P., and Niekum, S. Data-efficient policy evaluation through
1031 behavior policy search. In *Proceedings of the 34th International Conference on Machine*
1032 *Learning (ICML)*, 2017.
- 1033 Hanna, J. P., Niekum, S., and Stone, P. Importance sampling in reinforcement learning with
1034 an estimated behavior policy. *Machine Learning*, pages 1–51, 2021.
- 1035 Jiang, N. and Li, L. Doubly robust off-policy evaluation for reinforcement learning. In
1036 *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016.
- 1037 Kakade, S. A natural policy gradient. In *Proceedings of the 14th Conference on Advances in*
1038 *Neural Information Processing Systems (NeurIPS)*, volume 14, pages 1531–1538, 2001.
- 1039 Kearns, M. and Singh, S. Near-optimal reinforcement learning in polynomial time. *Machine*
1040 *Learning*, 49(2-3):209–232, 2002.
- 1041 Lemieux, C. Control variates. *Wiley StatsRef: Statistics Reference Online*, pages 1–8, 2014.
- 1042 Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and
1043 Wierstra, D. Continuous control with deep reinforcement learning. *CoRR*, abs/1509.02971,
1044 2015.
- 1045 Liu, Q., Li, L., Tang, Z., and Zhou, D. Breaking the curse of horizon: Infinite-horizon
1046 off-policy estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*,
1047 volume 31, pages 5356–5366, 2018.
- 1048 Mahmood, A. R., Hasselt, H. P. van, and Sutton, R. S. Weighted importance sampling for
1049 off-policy learning with linear function approximation. In *Advances in Neural Information*
1050 *Processing Systems*, pages 3014–3022, 2014.
- 1051 Owen, A. B. *Monte Carlo theory, methods and examples*. 2013.
- 1052 Peters, J. and Schaal, S. Natural actor-critic. *Neurocomputing*, 71(7-9):1180–1190, 2008.
- 1053 Piot, B., Geist, M., and Pietquin, O. Difference of convex functions programming for
1054 reinforcement learning. *Advances in Neural Information Processing Systems*, 27, 2014.
- 1055 Popoviciu, T. Sur les équations algébriques ayant toutes leurs racines réelles. *Mathematica*,
1056 9:129–145, 1935.
- 1057 Precup, D., Sutton, R. S., and Singh, S. Eligibility traces for off-policy policy evaluation. In
1058 *Proceedings of the 17th International Conference on Machine Learning (ICML)*, pages
1059 759–766, 2000.
- 1060 Puterman, M. L. *Markov decision processes: Discrete stochastic dynamic programming*.
1061 John Wiley & Sons, 2014.

- 1062 Robbins, H. and Monro, S. A stochastic approximation method. *The annals of mathematical*
1063 *statistics*, pages 400–407, 1951.
- 1064 Rubinstein, R. Y. and Kroese, D. P. *Simulation and the Monte Carlo method*, volume 10.
1065 John Wiley & Sons, 2016.
- 1066 Schulman, J., Levine, S., Moritz, P., Jordan, M., and Abbeel, P. Trust region policy
1067 optimization. In *Proceedings of the 32nd International Conference on Machine Learning*
1068 *(ICML)*, 2015.
- 1069 Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy
1070 optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- 1071 Sen, P. K. and Singer, J. M. *Large Sample Methods in Statistics: An Introduction with*
1072 *Applications*. Chapman & Hall, 1993.
- 1073 Singh, S. P. and Sutton, R. S. Reinforcement learning with replacing eligibility traces.
1074 *Machine Learning*, 22(1-3):123–158, 1996.
- 1075 Strehl, A. L., Li, L., and Littman, M. L. Reinforcement learning in finite MDPs: PAC
1076 analysis. *Journal of Machine Learning Research*, 10:2413–2444, 2009.
- 1077 Sutton, R. S. Learning to predict by the methods of temporal differences. *Machine Learning*,
1078 3(1):9–44, 1988.
- 1079 Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. MIT Press, 2018.
- 1080 Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. Policy gradient methods for
1081 reinforcement learning with function approximation. In *Proceedings of the 13th Conference*
1082 *on Advances in Neural Information Processing Systems (NeurIPS)*, 2000.
- 1083 Sutton, R. S., Modayil, J., Delp, M., Degris, T., Pilarski, P. M., White, A., and Precup,
1084 D. Horde: A scalable real-time architecture for learning knowledge from unsupervised
1085 sensorimotor interaction. In *The 10th International Conference on Autonomous Agents*
1086 *and Multiagent Systems-Volume 2*, pages 761–768, 2011.
- 1087 Swaminathan, A. and Joachims, T. The self-normalized estimator for counterfactual learning.
1088 In *Proceedings of the 29th Conference on Advances in Neural Information Processing*
1089 *Systems (NeurIPS)*, pages 3231–3239, 2015.
- 1090 Theodorou, G., Thomas, P. S., and Ghavamzadeh, M. Personalized ad recommendation
1091 systems for life-time value optimization with guarantees. In *Proceedings of the 27th*
1092 *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1806–1812, 2015.
- 1093 Thomas, P. S. *Safe reinforcement learning*. PhD thesis, University of Massachusetts Libraries,
1094 2015.
- 1095 Thomas, P. S. and Brunskill, E. Data-efficient off-policy policy evaluation for reinforcement
1096 learning. In *Proceedings of the 33rd International Conference on Machine Learning*
1097 *(ICML)*, 2016.

- 1098 Thomas, P. S. and Brunskill, E. Importance sampling with unequal support. In *Thirty-first*
1099 *AAAI conference on artificial intelligence*, 2017.
- 1100 Thomas, P. S., Theodorou, G., and Ghavamzadeh, M. High confidence off-policy evaluation.
1101 In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2015a.
- 1102 Thomas, P. S., Theodorou, G., and Ghavamzadeh, M. High confidence policy improvement.
1103 In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015b.
- 1104 Veness, J., Lanctot, M., and Bowling, M. Variance reduction in Monte-Carlo tree search. In
1105 *Proceedings of the 24th Conference on Advances in Neural Information Processing Systems*
1106 *(NeurIPS)*, pages 1836–1844, 2011.
- 1107 White, M. and Bowling, M. Learning a value analysis tool for agent evaluation. In
1108 *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI)*,
1109 pages 1976–1981, 2009.
- 1110 Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforce-
1111 ment learning. *Machine Learning*, 8(3-4):229–256, 1992.
- 1112 Yang, M., Nachum, O., Dai, B., Li, L., and Schuurmans, D. Off-policy evaluation via the
1113 regularized lagrangian. In *Advances in Neural Information Processing Systems (NeurIPS)*,
1114 volume 33, 2020.
- 1115 Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent.
1116 In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages
1117 928–936, 2003.
- 1118 Zinkevich, M., Bowling, M., Bard, N., Kan, M., and Billings, D. Optimal unbiased estimators
1119 for evaluating agent performance. In *Proceedings of the 21st National Conference on*
1120 *Artificial Intelligence (AAAI)*, pages 573–578, 2006.

1121 **Appendix A. Statistical Properties of Behavior Policy Search Estimates**

1122 In this appendix, we prove that the estimates from behavior policy search algorithms that
 1123 are computed as the mean of an unbiased off-policy estimator, OPE, such as IS, are unbiased
 1124 and consistent estimates of $v(\pi_e)$ and we provide a theoretical finite-rate bound on the
 1125 estimate. Typically, such results rely on i.i.d. sampling of trajectories from a single π_{θ_j} or at
 1126 least independent sampling from a behavior policy that is independent of other behavior
 1127 policies. In our case, the difficulty is that the estimate at iteration i depends on all π_{θ_j} for
 1128 $i = 1 \dots i$ and each π_{θ_j} is *not* independent of the others. Further, as π_{θ_j} may be different
 1129 from π_{θ_k} when $j \neq k$, both the assumptions of independence and identical distribution do
 1130 not hold. Nevertheless, we prove here that behavior policy search algorithms still produce
 1131 unbiased and consistent estimates of $v(\pi_e)$ at each iteration and have finite-rate bounds
 1132 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 $n \in \mathbb{N}$,*

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

1133

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

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

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

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

1134 Observe that (5) factors out the probability of observing parameter θ_j (which depends on
 1135 past parameters and trajectories) and the expected value of OPE given the value of θ_j
 1136 (which is independent of past parameters and trajectories *given* the value of θ_j). In Figure
 1137 3, this idea can be observed from d-separation: *conditioned* on a specific instance of π_{θ_k} the
 1138 estimates X_k are independent of previous parameters and trajectories.

Therefore, as OPE is an unbiased estimator for any *fixed* policy π_{θ_j} under Assumption 2, (5) can be expressed as,

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

Therefore, combining (4) and (6),

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

1139

■

1140

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,*

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

1141

Proof We begin by first establishing conditional independence in expectation between $\text{OPE}(\pi_e, H_j, \pi_{\theta_j})$ and $\text{OPE}(\pi_e, H_k, \pi_{\theta_k})$ for any $j \neq k$. For brevity, let $Z_j := \text{OPE}(\pi_e, H_j, \pi_{\theta_j})$.

$$\begin{aligned} \mathbb{E}[\text{OPE}(\pi_e, H_j, \pi_{\theta_j}) | \text{OPE}(\pi_e, H_k, \pi_{\theta_k})] &= \mathbb{E}[Z_j | Z_k] \\ &= \int_{\Theta} p(\theta_j = \theta | Z_k) \sum_{h \in \mathcal{H}} p(H_j = h | \pi_{\theta_j}, Z_k) \text{OPE}(\pi_e, h, \pi_{\theta_j}) \, d\theta \\ &\stackrel{(a)}{=} \int_{\Theta} p(\theta_j = \theta | Z_k) \left(\sum_{h \in \mathcal{H}} p(H_j = h | \pi_{\theta_j}) \text{OPE}(\pi_e, h, \pi_{\theta_j}) \right) \, d\theta \\ &\stackrel{(b)}{=} \int_{\Theta} p(\theta_j = \theta | Z_k) v(\pi_e) \, d\theta \\ &= v(\pi_e), \end{aligned} \tag{7}$$

where (a) follows from the fact that *given* the policy π_{θ_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 $\text{OPE}(\pi_e, H_j, \pi_{\theta_j})$ and $\text{OPE}(\pi_e, H_k, \pi_{\theta_k})$ can now be expressed as,

$$\begin{aligned} \text{Cov}(\text{OPE}(\pi_e, H_j, \pi_{\theta_j}), \text{OPE}(\pi_e, H_k, \pi_{\theta_k})) &= \text{Cov}(Z_j, Z_k) \\ &= \mathbb{E}[Z_j Z_k] - \mathbb{E}[Z_j] \mathbb{E}[Z_k] \\ &\stackrel{(b)}{=} \mathbb{E}[\mathbb{E}[Z_j | Z_k] Z_k] - \mathbb{E}[Z_j] \mathbb{E}[Z_k] \\ &\stackrel{(c)}{=} v(\pi_e) \mathbb{E}[Z_k] - \mathbb{E}[Z_j] \mathbb{E}[Z_k] \\ &\stackrel{(d)}{=} v(\pi_e)^2 - v(\pi_e)^2 \\ &= 0, \end{aligned}$$

1144

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

1145

1146

■

1147

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

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

1148

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

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

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 $\text{OPE}(\pi_e, H_j, \theta_j)$ has variance bounded above by some finite constant \tilde{c} . Therefore, as $\forall j, \text{Var}[\text{OPE}(\pi_e, H_j, \theta_j)] < \tilde{c}$, it follows from (8) that $\text{Var}[\overline{\text{OPE}}(\pi_e, D_i)] \rightarrow 0$. 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,

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

Now from Markov's inequality,

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

1149 Combining (9) and (10),

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

1150 ■

1151

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

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

1152

1153 **Proof** We begin by observing that since the $(\text{OPE}(\pi_e, H_j, \theta_j))_{j=1}^i$ are sequentially dependent,
 1154 if we can convert them into a Martingale sequence then we can use concentration inequalities
 1155 for Martingales to obtain convergence rates for $\text{OPE}(\pi_e, D_i)$.

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

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

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

$$\begin{aligned} \mathbb{E}[Y_j | Y_{j-1}] &= \mathbb{E}[\text{OPE}(\pi_e, H_j, \pi_{\theta_j}) | Y_{j-1}] - \mathbb{E}[v(\pi_e) | Y_{j-1}] + \mathbb{E}[Y_{j-1} | Y_{j-1}] \\ &= \mathbb{E}[\text{OPE}(\pi_e, H_j, \pi_{\theta_j}) | Y_{j-1}] - v(\pi_e) + Y_{j-1}. \end{aligned} \quad (12)$$

To simplify (12) further, notice that,

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

where (a) follows from the fact that Y_{j-1} only contains information from iterates till $j-1$ (inclusive) and $\text{OPE}(\pi_e, H_j, \pi_{\theta_j})$ is independent of the past *conditioned* on the value of θ_j . Step (b) follows from the fact that $\text{OPE}(\pi_e, H_j, \pi_{\theta_j})$ (the OPE estimate) is an unbiased estimator of $v(\pi_e)$ for any fixed behavior policy π_{θ} 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 $\text{OPE}(\pi_e, H_j, \pi_{\theta_j})$ 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 $(Y_j)_{j=0}^i$,

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

First, considering $|Y_i - Y_0|$:

$$|Y_i - Y_0| \stackrel{(a)}{=} \left| \sum_{j=1}^i (Y_j - Y_{j-1}) \right| \stackrel{(b)}{=} \left| \sum_{j=1}^i (\text{OPE}(\pi_e, H_j, \pi_{\theta_j}) - v(\pi_e)) \right|, \quad (15)$$

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

$$\begin{aligned}
 & \Pr \left(\left| \sum_{j=1}^i (\text{OPE}(\pi_e, H_j, \pi_{\theta_j}) - v(\pi_e)) \right| > \epsilon \right) \leq 2 \exp \left(\frac{-2\epsilon^2}{|D_i| \tilde{c}^2} \right) \\
 & \Pr \left(\left| \sum_{j=1}^i \text{OPE}(\pi_e, H_j, \pi_{\theta_j}) - |D_i| v(\pi_e) \right| > \epsilon \right) \leq 2 \exp \left(\frac{-2\epsilon^2}{|D_i| \tilde{c}^2} \right) \\
 & \Pr \left(\left| \frac{1}{|D_i|} \sum_{j=1}^i \text{OPE}(\pi_e, H_j, \pi_{\theta_j}) - v(\pi_e) \right| > \frac{\epsilon}{|D_i|} \right) \stackrel{(c)}{\leq} 2 \exp \left(\frac{-2\epsilon^2}{|D_i| \tilde{c}^2} \right) \\
 & \Pr \left(|\overline{\text{OPE}}(\pi_e, D_i) - v(\pi_e)| > \frac{\epsilon}{|D_i|} \right) \stackrel{(d)}{\leq} 2 \exp \left(\frac{-2\epsilon^2}{|D_i| \tilde{c}^2} \right) \\
 & \Pr \left(|\overline{\text{OPE}}(\pi_e, D_i) - v(\pi_e)| > \epsilon \right) \stackrel{(e)}{\leq} 2 \exp \left(\frac{-2|D_i| \epsilon^2}{\tilde{c}^2} \right), \quad (16)
 \end{aligned}$$

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

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

1156 thereby giving the desired error rate of $O \left(\frac{1}{\sqrt{|D_i|}} \right)$.

1157

1158

1159 Appendix B. Behavior Policy Gradient of the Variance

1160 In this section, we derive the gradient of the variance of importance sampling with respect
 1161 to the behavior policy parameters. We first derive an analytic expression for the gradient of
 1162 the variance of an arbitrary, unbiased off-policy policy evaluation estimator, $\text{OPE}(\pi_e, H, \pi_{\theta})$.
 1163 From our general derivation we derive the gradient of the variance of the basic importance
 1164 sampling estimator and then extend to the doubly robust and per-decision estimators.

1165 B.1 MSE Gradient for an Unbiased Off-Policy Policy Evaluation Method

1166 Lemma 2 gives the gradient of the MSE for any unbiased off-policy policy evaluation method.

1167

Lemma 2.

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

1168 **Proof** We begin by decomposing $\Pr(H = h|\pi)$ into two components – one that depends on
1169 π and the other that does not. Recall that we defined:

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

1170 and define

$$p(h) := \Pr(H = h|\pi)/w_{\pi}(h),$$

1171 for any π such that h is in the support of π (any such π will result in the same value of
1172 $p(h)$). These two definitions mean that $\Pr(H = h|\pi) = p(h)w_{\pi}(h)$.

The MSE of the OPE estimator is given by:

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

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

$$\text{MSE}(\text{OPE}(\pi_e, H, \pi_{\boldsymbol{\theta}})) = \text{Var}[\text{OPE}(\pi_e, H, \pi_{\boldsymbol{\theta}})] \tag{17}$$

$$= \mathbf{E} [\text{OPE}(\pi_e, H, \pi_{\boldsymbol{\theta}})^2 | H \sim \pi_{\boldsymbol{\theta}}] - \mathbf{E}[\text{OPE}(\pi_e, H, \pi_{\boldsymbol{\theta}}) | H \sim \pi_{\boldsymbol{\theta}}]^2 \tag{18}$$

$$= \mathbf{E} [\text{OPE}(\pi_e, H, \pi_{\boldsymbol{\theta}})^2 | H \sim \pi_{\boldsymbol{\theta}}] - v(\pi_e)^2 \tag{19}$$

1173 (18) follows from (17) by the definition of the variance and (19) follows from (18) because
1174 the expectation of an unbiased estimator of $v(\pi_e)$ is $v(\pi_e)$.

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

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

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

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

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

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

1175

1176

1177 B.2 Behavior Policy Gradient of the Variance

1178 We now use Lemma 2 to prove the Behavior Policy Gradient of the Variance Theorem.

Theorem 1 (Behavior Policy Gradient of the Variance).

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

1179 **Proof** We first derive $\frac{\partial}{\partial \boldsymbol{\theta}} \text{IS}(\pi_e, H, \pi_{\boldsymbol{\theta}})^2$. Theorem 1 then follows directly from using
 1180 $\frac{\partial}{\partial \boldsymbol{\theta}} \text{IS}(\pi_e, H, \pi_{\boldsymbol{\theta}})^2$ as $\frac{\partial}{\partial \boldsymbol{\theta}} \text{OPE}(\pi_e, H, \pi_{\boldsymbol{\theta}})^2$ in Lemma 2.

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

1181 where **(a)** uses (21) to differentiate $w_{\pi_{\theta}}(H)$.

1182 Substituting this expression and $\text{IS}(\pi_e, H, \pi_{\theta})$ for $\text{OPE}(\pi_e, H, \pi_{\theta})$ into Lemma 2 completes
 1183 the proof of Theorem 1:

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

1184 ■

1185

1186 B.3 MSE Gradient for the Doubly Robust Estimator

1187 We also present an extension of the IS MSE gradient to the Doubly Robust (DR) estimator.
 1188 Recall that for a single trajectory, H , DR is defined as:

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

1189 where \hat{v}^{π_e} is an approximation of the state-value function of π_e , \hat{q}^{π_e} is an approximation of
 1190 the action-value function of π_e , and $w_{\pi, t} := \prod_{j=0}^t \pi(A_j | S_j)$.

1191 The gradient of the MSE of the DR estimator is given by the following corollary to
 1192 Theorem 1:

Corollary 1.

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

1193 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 $H \sim \pi_{\theta}$.

1194 Proof

1195 As with Theorem 1, we first derive $\frac{\partial}{\partial \theta} \text{DR}(\pi_e, H, \pi_{\theta})^2$. Corollary 1 then follows directly
 1196 from using $\frac{\partial}{\partial \theta} \text{DR}(\pi_e, H, \pi_{\theta})^2$ as $\frac{\partial}{\partial \theta} \text{OPE}(\pi_e, H, \pi_{\theta})^2$ in Lemma 2.

1197 Let $\delta_t := R_t - \hat{q}^{\pi_e}(S_t, A_t) + \hat{v}^{\pi_e}(S_{t+1})$.

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

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

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

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

1198

■

1199

1200

1201

1202

1203

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.

1204

Appendix C. Convergence of BPG-V

1205

1206

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

1207

1208

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 $\lim_{i \rightarrow \infty} \frac{\partial}{\partial \boldsymbol{\theta}} \text{MSE}[\text{IS}(\pi_e, H_i, \pi_{\boldsymbol{\theta}_i})] = 0$.*

1209

1210

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:

1211

1212

1213

1214

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

1215

1216

1217

1218

1219

Other conditions of Proposition 3 in (Bertsekas and Tsitsiklis, 2000) are satisfied by the unbiasedness of the gradient estimates used by BPG-V. We 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 $\frac{\partial}{\partial \boldsymbol{\theta}} \text{MSE}[\text{IS}(\pi_e, H, \pi_{\boldsymbol{\theta}})]$ which can be seen to be continuously differentiable under

1220 our assumption that π_θ is continuously differentiable and Assumption 1 which implies that
 1221 $\text{IS}(\pi_e, H, \pi_\theta)$ always exists.

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

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

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):

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

1222 which is bounded because Assumption 1 implies (c.1) is bounded and (c.2) is the same as
 1223 (d) which we next show is bounded.

For (d):

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

1224 which is bounded because each $\frac{\partial}{\partial \theta} \pi_\theta(a_t | s_t)$ is bounded by construct, and $\prod_{t'=0, t' \neq t}^{l-1} \pi_\theta(a_{t'} | s_{t'}) \leq$
 1225 1.

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

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

1226 which is bounded under the construct that π_{θ} is twice differentiable with bounded first and
 1227 second derivatives. Thus we conclude that the MSE objective is continuously differentiable
 1228 with a Lipschitz derivative.

1229 Finally, we have to show that the variance of the gradient estimate used by BPG-V is
 1230 bounded. To do so, we show that the gradient estimate with any single trajectory is bounded
 1231 which implies the variance of the estimates used by BPG-V is bounded because the variance
 1232 of a bounded random variable is itself bounded.

For any trajectory h , collected by following π_{θ} , an unbiased estimate of the MSE estimate is given as:

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

1233 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
 1234 of Equation (23), $\frac{w_{\pi_e}(h)^2 g(h)^2}{w_{\pi_{\theta}}(h)^3}$ is bounded under Assumption 1 and $\frac{\partial}{\partial \theta} w_{\pi_{\theta}}(h)$ was shown to
 1235 be bounded in Equation (22). Thus we conclude that the variance of the gradient estimate
 1236 used by BPG-V is bounded. Proposition 4 now follows from Proposition 3 of Bertsekas and
 1237 Tsitsiklis (2000). ■

1238

1239

1240 Appendix D. Convexity of Variance Objective

1241 In this appendix, we prove that, when π_{θ} is a linear-softmax policy, then the variance
 1242 objective minimized by BPG-V is convex in the policy parameters. Thus, BPG-V is
 1243 guaranteed to converge to the parameter vector that minimizes the variance of the IS-return
 1244 under standard stochastic gradient descent step-size conditions (Bertsekas and Tsitsiklis,
 1245 2000).

1246 A linear-softmax policy is a policy over a finite set of actions where the probability of
 1247 each action is defined as a softmax distribution with logits from a linear combination of
 1248 state features. Formally, let $\phi : \mathcal{S} \rightarrow \mathbf{R}^q$ for integer q be a state feature function that maps
 1249 states to feature vectors. For each action, $a \in \mathcal{A}$, we have a vector $\theta_a \in \mathbf{R}^q$ and θ is the
 1250 concatenation of all θ_a . A linear-softmax policy defines the probability of action a in state s
 1251 as:

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

1252 **Theorem 2.** *Assume π_{θ} is a linear-softmax policy. Then, $\text{MSE}[\text{IS}(\pi_e, H, \theta)]$ is a convex*
 1253 *function w.r.t. θ .*

1254 **Proof**

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

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: $\Pr(H = h|\pi) = p(h) * w_{\pi}(h)$.

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

1255 where c_1 and c_2 are functions of h that are constant w.r.t. θ . Furthermore, $c_2(h) = e^{c_1(h)}$
 1256 and therefore must be positive. We next show that $e^{-\ln w_{\pi_{\theta}}(h)}$ is convex in θ . We then
 1257 have a linear combination of convex functions with positive weights which is itself a convex
 1258 function. Note that we do not have to worry about taking the log of a non-positive value.
 1259 For any h such that $g(h)$, $w_{\pi_e}(h)$, or $w_{\pi_{\theta}}(h)$ is zero, then $\Pr(H = h|\pi_{\theta}) * \text{IS}(\pi_e, h, \theta)^2$ is zero
 1260 and can be ignored in the summation. The only potential negative value is $g(h)$ but it is
 1261 squared within the logarithm and can thus be replaced with its absolute value.

1262 We next introduce the following lemma that shows that $-\ln w_{\pi_{\theta}}(h)$ is a convex function
 1263 with respect to θ .

1264 **Lemma 2.** Assume π_θ is a linear-softmax policy. Then for $w_{\pi_\theta}(h) := \prod_{t=0}^{l-1} \pi_\theta(a_t|s_t)$,
 1265 $-\ln w_{\pi_\theta}(h)$ is a convex function w.r.t. θ for any trajectory $h = (s_0, a_0, \dots, s_{l-1}, a_{l-1})$.

Proof

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

1266 Next, we show that each $-\ln \pi_\theta(a_t|s_t)$ is convex under the linear-softmax policy param-
 1267 eterization:

$$-\ln \pi_\theta(a|s) = \ln \left(\sum_{b \in \mathcal{A}} e^{\theta_b^T \phi(s)} \right) - \theta_a^T \phi(s)$$

1268 The log-sum-exp function is convex (Boyd et al., 2004, Chapter 3, Example 3.13) and
 1269 subtracting a linear function does not change convexity. Thus, (25) is a sum of convex
 1270 functions which is convex. ■

1271
 1272 Continuing with the proof of Theorem 4, Lemma 2 implies that $e^{-\ln w_{\pi_\theta}(h)}$ is the
 1273 exponential of a convex function. The exponential of a convex function is convex (Boyd et al.,
 1274 2004, Chapter 3, Eq 3.11) and thus $e^{-\ln w_{\pi_\theta}(h)}$ is convex in θ . Finally, we have that (24) is a
 1275 linear combination of convex functions with positive weights. Thus, $\mathbf{E}[\text{IS}(\pi_e, H, \theta)^2 | H \sim \pi_\theta]$
 1276 is a convex function which concludes the proof. ■

1277

1278

1279 Appendix E. Minimal-Variance Behavior Policy

1280 In this appendix we prove Proposition 5 that gives a sufficient condition for a minimal-
 1281 variance behavior policy:

1282 **Proposition 5.** Let $w_\pi(h) := \prod_{t=0}^{l-1} \pi(a_t|s_t)$. Assume $\exists \tilde{h} \in \mathcal{H}$ such that $g(\tilde{h}) \cdot \Pr(H =$
 1283 $\tilde{h} | \pi_e) \neq 0$, i.e., there is non-zero probability that π_e generates a trajectory with non-zero
 1284 return. If $\exists \pi \in \Pi$ s.t.

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

1285 then π is a minimal-variance behavior policy.

1286 **Proof** Recall that we defined $w_\pi(h) := \prod_{t=0}^{l-1} \pi(a_t|s_t)$ and define $p(h) := d_0(s_0) \prod_{t=1}^{l-1} P(s_t|s_{t-1}, a_{t-1})$.
 1287 From these definitions, note that $\Pr(H = h | \pi) = w_\pi(h)p(h)$.

The variance of the importance sampling estimator is:

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

$$(27)$$

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:

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

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:

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

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

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

Observe that Equation (29) holds $\forall \tilde{h} \in \mathcal{H}$ and thus λ^* 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 λ^* to obtain the optimal choice of $w_{\pi_b}(\tilde{h})$:

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

The constant λ^* can be determined by ensuring the equality constraint is satisfied giving $\lambda^* = \sum_{h \in \mathcal{H}} p(h) w_{\pi_e}(h) |g(h)| = \mathbf{E}[|g(H)| | H \sim \pi_e]$. 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

1304 the case that $g(h) = 0$ for all $h \in \mathcal{H}$ that Proposition 5 gives an undefined value for $w_{\pi_b^*}$.
 1305 However we can ignore this case as, if $g(h) = 0$ for all $h \in \mathcal{H}$, then the variance of the
 1306 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 $w_{\pi_b^*}$.

$$\text{Var}[\text{IS}(\pi_e, H, \pi_b^*)] = \mathbf{E} \left[\left(\frac{w_{\pi_e}(H)}{w_{\pi_b^*}(H)} g(H) \right)^2 \middle| H \sim \pi_b^* \right] - v(\pi_e)^2 \quad (31)$$

$$= \mathbf{E} \left[|g(H)| \middle| H \sim \pi_e \right]^2 \underbrace{\mathbf{E} \left[\left(\frac{g(H)}{|g(H)|} \right)^2 \middle| H \sim \pi_b^* \right]}_{=1} - v(\pi_e)^2 \quad (32)$$

$$= \mathbf{E} \left[|g(H)| \middle| H \sim \pi_e \right]^2 - v(\pi_e)^2 \quad (33)$$

$$= \mathbf{E} \left[\frac{w_{\pi_e}(H)}{w_{\pi_b}(H)} |g(H)| \middle| H \sim \pi_b \right]^2 - v(\pi_e)^2 \quad (34)$$

$$\leq \mathbf{E} \left[\left(\frac{w_{\pi_e}(H)}{w_{\pi_b}(H)} |g(H)| \right)^2 \middle| H \sim \pi_b \right] - v(\pi_e)^2 \quad (35)$$

$$= \text{Var}[\text{IS}(\pi_e, H, \pi_b)]$$

1307 where (32) follows (31) by plugging in the solution for $w_{\pi_b^*}$ given by (30) and factoring
 1308 out the constant λ , (33) follows from (32) because the expected value of 1 is 1 under any
 1309 distribution, (34) follows (33) by using importance sampling to change the expectation to
 1310 be in terms of trajectories from any behavior policy π_b instead of π_e , and the inequality
 1311 follows from Jensen's inequality. Finally, we can drop the absolute value in (35) because
 1312 it is squared. Thus we can conclude that $\text{Var}[\text{IS}(\pi_e, H, \pi_b^*)] \leq \text{Var}[\text{IS}(\pi_e, H, \pi_b)]$ for any
 1313 behavior policy π_b . \blacksquare

1314

1315 Appendix F. Behavior Policy Gradient of the KL

1316 In this appendix we derive Theorem 3, which gives the gradient, with respect to the policy
 1317 parameters, of the KL-divergence between the distribution of trajectories under a minimal-
 1318 variance behavior policy, $\Pr(H|\pi_b^*)$, and the distribution of trajectories under π_θ , $\Pr(H|\pi_\theta)$.
 1319 This gradient is:

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

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

1320 **Proof** From Proposition 5 we know that a minimal-variance behavior policy is any policy,
 1321 π_b^* , that satisfies the condition:

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

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

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

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

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

1322

Differentiating with respect to θ , we obtain:

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

1323 where the second step uses the multi-factor product rule and the final step uses importance
1324 sampling to convert from an expectation under π_e to one under π_{θ} . \blacksquare

1325

1326 Appendix G. Convergence of BPG-KL

1327 In this section, we prove that BPG-KL (Algorithm 2) converges under an appropriately
1328 chosen step-size.

1329 **Proposition 6.** *Under Assumption 1 and Assumption 3, BPG-KL converges. That is,*
1330 $D_{\text{KL}}(\Pr(H|\pi_b^*) || \Pr(H|\pi_{\theta}))$ *converges to a finite value and* $\lim_{i \rightarrow \infty} \frac{\partial}{\partial \theta} D_{\text{KL}}(\Pr(H|\pi_b^*) || \Pr(H|\pi_{\theta})) =$
1331 0 .

1332 **Proof** Similar to Proposition 4, the proof follows from an application of Proposition 3 in
 1333 Bertsekas and Tsitsiklis (2000). A minor nuance for Proposition 6 is that BPG-KL does *not*
 1334 use unbiased estimates of the true KL-gradient but uses unbiased estimates of an expression
 1335 that is just proportional to the true KL-gradient. However, the proportionality constant is
 1336 fixed with respect to θ and so we can ignore it when showing convergence.

1337 To apply Proposition 3 in Bertsekas and Tsitsiklis (2000), we must show the following
 1338 conditions:

- 1339 1. $D_{\text{KL}}(\Pr(H|\pi_b^*)||\Pr(H|\pi_\theta))$ is continuously differentiable w.r.t. θ .
- 1340 2. The gradient of the KL objective, $\frac{\partial}{\partial\theta} D_{\text{KL}}(\Pr(H|\pi_b^*)||\Pr(H|\pi_\theta))$, is Lipschitz continuous
 1341 w.r.t. θ .
- 1342 3. The gradient estimate used by BPG-KL has bounded variance.

1343 Theorem 3 gives us an expression that is proportional to $\frac{\partial}{\partial\theta} D_{\text{KL}}(\Pr(H|\pi_b^*)||\Pr(H|\pi_\theta))$ which
 1344 can be seen to be continuously differentiable under our assumption that π_θ is continuously
 1345 differentiable and Assumption 1 which implies that $\text{IS}(\pi_e, H, \pi_\theta)$ is bounded.

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

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

1346 The denominator in (36) cannot be zero as otherwise $w_\theta(h)$ would be zero and the cor-
 1347 responding trajectory could be ignored in the expectation. Furthermore, by construct,
 1348 $\frac{\partial}{\partial\theta} \pi_\theta(a_t|s_t)$ and $\frac{\partial^2}{\partial\theta^2} \pi_\theta(a_t|s_t)$ exist and are bounded. Thus we conclude that the first and
 1349 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 π_θ , an unbiased estimate of the KL gradient estimate is given as:

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

1350 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
1351 of Equation (37), $\frac{w_{\pi_e}(h)|g(h)|}{w_{\pi_\theta}(h)^2}$ is bounded under Assumption 1 and $\frac{\partial}{\partial \theta} w_{\pi_\theta}(h)$ was shown to
1352 be bounded in Equation (22). Thus we conclude that the variance of the gradient estimate
1353 used by BPG-KL is bounded. Proposition 6 now follows from Proposition 3 of (Bertsekas
1354 and Tsitsiklis, 2000).

1355

■

1356

1357 Appendix H. Convexity of KL-Divergence Objective

1358 In this appendix, we prove that, when π_θ is a linear-softmax policy, then the objective
1359 minimized by BPG-KL is convex in the policy parameters. Thus, BPG-KL is guaranteed
1360 to converge to the parameter vector that minimizes the KL divergence with the minimal-
1361 variance behavior policy under standard stochastic gradient descent step-size conditions
1362 (Bertsekas and Tsitsiklis, 2000).

1363 **Theorem 4.** *Assume π_θ is a linear-softmax policy. Then, $D_{\text{KL}}(\text{Pr}(H|\pi_{b^*})||\text{Pr}(H|\pi_\theta))$ is a*
1364 *convex function w.r.t. θ .*

1365 Proof

1366 Recall that for the minimal-variance behavior policy we have:

$$w_{\pi_{b^*}}(h) = \frac{|g(h)|w_{\pi_e}(h)}{\lambda},$$

1367 where $\lambda = \mathbf{E}[|g(H)||H \sim \pi_\theta]$.

The KL-divergence between the minimal-variance behavior policy and policy π_{θ} is given as:

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

1368 Functions c_1 , c_2 , and c_3 are positive for any h and constant with respect to θ . Lemma 2
 1369 says that $-\log w_{\pi_{\theta}}(h)$ is convex w.r.t θ . Thus, (38) is a weighted sum of convex functions
 1370 with positive weights which is itself convex. Thus the KL-divergence objective optimized
 1371 by BPG-KL is convex with respect to θ . An interesting observation from the proof is that
 1372 the KL-divergence between the trajectory distribution of π_{θ} and any trajectory distribution
 1373 that does not depend on θ is also a convex function (under the assumption that π_{θ} is a
 1374 linear-softmax policy).

1375

■

1376