Behavior Policy Gradient Supplemental Material

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

A. Proof of Theorem 1

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

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

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

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

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

Lemma 1.

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

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

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

and

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

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

The MSE of the OPE estimator is given by:

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

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

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

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

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

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

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

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

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

□

A.2. Behavior Policy Gradient Theorem

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

Theorem 1.

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

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

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

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

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

Substituting this expression into Lemma 1 completes the proof:

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

□

A.3. Doubly Robust Estimator

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

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

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

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

Corollary 1.

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

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

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

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

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

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

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

□

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

B. BPG's Off-Policy Estimates are Unbiased

This appendix proves that the estimate of BPG is an unbiased estimate of $\rho(\pi_e)$. If only trajectories from a single $\boldsymbol{\theta}_i$ were used then clearly $\operatorname{IS}(\cdot,\boldsymbol{\theta}_i)$ is an unbiased estimate of $\rho(\pi_e)$. The difficulty is that the BPG's estimate at iteration n depends on all $\boldsymbol{\theta}_i$ for $i = 1 \ldots n$ and each $\boldsymbol{\theta}_i$ is *not* independent of the others. Nevertheless, we prove here that BPG produces an unbiased estimate of $\rho(\pi_e)$ at each iteration. Specifically, we will show that $\mathbf{E}\left[\operatorname{IS}(H_n,\boldsymbol{\theta}_n|\boldsymbol{\theta}_0 = \boldsymbol{\theta}_e)\right]$ is an unbiased estimate of $\rho(\pi_e)$, where the IS estimate is conditioned on $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_e$. To make the dependence of $\boldsymbol{\theta}_i$ on $\boldsymbol{\theta}_{i-1}$ explicit, we will write $f(H_{i-1}) \coloneqq \boldsymbol{\theta}_i$ where $H_{i-1} \sim \pi_{\boldsymbol{\theta}_{i-1}}$. Notice that, even though BPG's off-policy estimates are unbiased, they are *not* statistically independent. This means that concentration inequalities, like Hoeffding's inequality, cannot be applied directly. We conjecture that the conditional independence properties of BPG (specifically that H_i is independent of H_{i-1} given θ_i), are sufficient for Hoeffding's inequality to be applicable.

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

C. Supplemental Experiment Description

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

Gridworld: This domain is a 4x4 Gridworld with a terminal state with reward 10 at $(3,3)$, a state with reward -10 at $(1,1)$, a state with reward 1 at $(1,3)$, and all other states having reward -1. The action set contains the four cardinal directions and actions move the agent in its intended direction (except when moving into a wall which produces no movement). The agent begins in (0,0), $\gamma = 1$, and $L = 100$. All policies use softmax action selection with temperature 1 where the probability of taking an action a in a state s is given by:

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

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

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

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

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

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

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

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

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

References

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

Schulman, John, Levine, Sergey, Moritz, Philipp, Jordan, Michael, and Abbeel, Pieter. Trust region policy optimization. In *International Conference on Machine Learning, ICML*, 2015.