Data-efficient Policy Evaluation through Behavior Policy Search

Josiah Hanna\(^1\) Philip Thomas\(^2\) Peter Stone\(^1\)
Scott Niekum\(^1\)

\(^1\)University of Texas at Austin
\(^2\)University of Massachusetts, Amherst

August 8th, 2017
Policy Evaluation
1. Demonstrate that importance-sampling for policy evaluation can outperform on-policy policy evaluation.
Outline

1. Demonstrate that importance-sampling for policy evaluation can outperform on-policy policy evaluation.

2. Show how to improve the behavior policy for importance-sampling policy evaluation.
Outline

1. Demonstrate that importance-sampling for policy evaluation can outperform on-policy policy evaluation.

2. Show how to improve the behavior policy for importance-sampling policy evaluation.

3. Empirically evaluate (1) and (2).
Background

- Finite-horizon MDP.
- Agent selects actions with a stochastic policy, $\pi$.
- The policy and environment determine a distribution over trajectories, $H : S_0, A_0, R_0, S_1, A_1, R_1, \ldots, S_L, A_L, R_L$
Policy Evaluation

Policy performance:

$$\rho(\pi) := \mathbb{E} \left[ \sum_{t=0}^{L} \gamma^t R_t \Bigg| H \sim \pi \right]$$
Policy performance:

\[
\rho(\pi) := \mathbb{E} \left[ \sum_{t=0}^{L} \gamma^t R_t \mid H \sim \pi \right]
\]

Given a target policy, \( \pi_e \), estimate \( \rho(\pi_e) \).
Policy Evaluation

Policy performance:

\[ \rho(\pi) := \mathbb{E} \left[ \sum_{t=0}^{L} \gamma^t R_t \middle| H \sim \pi \right] \]

Given a target policy, \( \pi_e \), estimate \( \rho(\pi_e) \).

- Let \( \pi_e \equiv \pi_{\theta_e} \)
Monte Carlo Policy Evaluation

Given a dataset $\mathcal{D}$ of trajectories where $\forall H \in \mathcal{D}$, $H \sim \pi_e$:

$$MC(\mathcal{D}) := \frac{1}{|\mathcal{D}|} \sum_{H_i \in \mathcal{D}} \sum_{t=0}^{L} \gamma^t R_t^{(i)}$$
Target policy \( \pi_e \) samples the high-rewarding first action with probability 0.01.
Target policy $\pi_e$ samples the high-rewarding first action with probability 0.01.

Monte Carlo evaluation of $\pi_e$ has high variance.
Target policy $\pi_e$ samples the high-rewarding first action with probability 0.01.

Monte Carlo evaluation of $\pi_e$ has high variance.

Importance-sampling with a behavior policy that samples either action with equal probability gives a low variance evaluation.
Importance-Sampling Policy Evaluation

Given a dataset $D$ of trajectories where $\forall H_i \in D$, $H_i$ is sampled from a behavior policy $\pi_i$:

$$IS(D) := \frac{1}{|D|} \sum_{H_i \in D} \prod_{t=0}^{L} \frac{\pi_e(A_t|S_t)}{\pi_i(A_t|S_t)} \sum_{t=0}^{L} \gamma^t R^{(i)}_t$$

re-weighting factor

---

1Precup, Sutton, and Singh (2000)
Importance-Sampling Policy Evaluation

Given a dataset $D$ of trajectories where $\forall H_i \in D$, $H_i$ is sampled from a behavior policy $\pi_i$:

$$IS(D) := \frac{1}{|D|} \sum_{H_i \in D} \prod_{t=0}^{L} \frac{\pi_e(A_t|S_t)}{\pi_i(A_t|S_t)} \sum_{t=0}^{L} \gamma^t R_{t}^{(i)}$$

re-weighting factor

For convenience:

$$IS(H, \pi) := \prod_{t=0}^{L} \frac{\pi_e(A_t|S_t)}{\pi(A_t|S_t)} \sum_{t=0}^{L} \gamma^t R_{t}$$

$^{1}$Precup, Sutton, and Singh (2000)

Josiah Hanna, Philip Thomas, Peter Stone, Scott Niekum

Data-efficient Policy Evaluation through Behavior Policy Search
Importance-sampling can achieve zero mean-squared error policy evaluation with only a single trajectory!
The Optimal Behavior Policy

Importance-sampling can achieve zero mean-squared error policy evaluation with only a single trajectory!

We cannot analytically determine this policy.

- Requires $\rho(\pi_e)$ be known!
The Optimal Behavior Policy

Importance-sampling can achieve zero mean-squared error policy evaluation with only a single trajectory!

We cannot analytically determine this policy.

- Requires $\rho(\pi_e)$ be known!
- Requires the reward function be known.
Importance-sampling can achieve zero mean-squared error policy evaluation with only a single trajectory!

We cannot analytically determine this policy.

- Requires $\rho(\pi_e)$ be known!
- Requires the reward function be known.
- Requires deterministic transitions.
Adapt the behavior policy towards the optimal behavior policy.
Behavior Policy Search

Adapt the behavior policy towards the optimal behavior policy.

At each iteration, $i$:

1. Choose behavior policy parameters, $\theta_i$, based on all observed data $D$. 
Adapt the behavior policy towards the optimal behavior policy.

At each iteration, $i$:

1. Choose behavior policy parameters, $\theta_i$, based on all observed data $\mathcal{D}$.

2. Sample $m$ trajectories, $H \sim \theta_i$ and add to a data set $\mathcal{D}$. 
Behavior Policy Search

Adapt the behavior policy towards the optimal behavior policy.

At each iteration, $i$:

1. Choose behavior policy parameters, $\theta_i$, based on all observed data $\mathcal{D}$.

2. Sample $m$ trajectories, $H \sim \theta_i$ and add to a data set $\mathcal{D}$.

3. Estimate $\rho(\pi_e)$ with trajectories in $\mathcal{D}$. 
Behavior Policy Gradient

**Key Idea:** Adapt the behavior policy parameters, $\theta$, with gradient descent on the mean squared error of importance-sampling.

$$
\theta_{i+1} = \theta_i - \alpha \frac{\partial}{\partial \theta} \text{MSE}[\text{IS}(H_i, \theta)]
$$
**Key Idea:** Adapt the behavior policy parameters, $\theta$, with gradient descent on the mean squared error of importance-sampling.

$$
\theta_{i+1} = \theta_i - \alpha \frac{\partial}{\partial \theta} \text{MSE}[\text{IS}(H_i, \theta)]
$$

- $\text{MSE}[\text{IS}(H, \theta)]$ is **not** computable.
- $\frac{\partial}{\partial \theta} \text{MSE}[\text{IS}(H, \theta)]$ is computable.
Behavior Policy Gradient Theorem

Theorem

$$\frac{\partial}{\partial \theta} \text{MSE} (\text{IS}(H, \theta)) = \mathbb{E}_{\pi_{\theta}} \left[ - \text{IS}(H, \theta)^2 \sum_{t=0}^{L} \frac{\partial}{\partial \theta} \log \left( \pi_{\theta}(A_t | S_t) \right) \right]$$
Empirical Results

Data-efficient Policy Evaluation through Behavior Policy Search

Josiah Hanna, Philip Thomas, Peter Stone, Scott Niekum

UT Austin

Cartpole Swing-up

Acrobot
Empirical Results

![Cartpole Swing-up](image1)

![Acrobot](image2)

Josiah Hanna, Philip Thomas, Peter Stone, Scott Niekum

Data-efficient Policy Evaluation through Behavior Policy Search
GridWorld Results

High Variance Policy

Low Variance Policy

Josiah Hanna, Philip Thomas, Peter Stone, Scott Niekum

Data-efficient Policy Evaluation through Behavior Policy Search
GridWorld Results

High Variance Policy

Low Variance Policy
Variance Reduction

![Graph showing variance reduction over iterations]
Investigated an extension to the doubly-robust off-policy estimator.\(^2\)

Investigated where BPG is most effective empirically.

\(^2\)Jiang and Li(2016), Thomas and Brunskill(2016)
Behavior policy search makes off-policy evaluation more accurate than on-policy evaluation.

Behavior Policy Gradient is an effective behavior policy search method.
Open Questions

1. Can behavior policy search improve policy improvement?
Open Questions

1. Can behavior policy search improve policy improvement?

2. Are there better measures of a good behavior policy?
Open Questions

1. Can behavior policy search improve policy improvement?

2. Are there better measures of a good behavior policy?

3. Is the final behavior policy found by BPG applicable to other target policies?
Thanks for your attention!
Questions?
Nan Jiang and Lihong Li. Doubly robust off-policy evaluation for reinforcement learning.  
*arXiv preprint arXiv:1511.03722, 2016.*

P.S. Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning.  
*arXiv preprint arXiv:1604.00923, 2016.*
Prior Work: Importance Sampling
Prior Work: Importance Sampling