To appear in *AAAI Spring 2018 Symposium on Data Efficient Reinforcement Learning*
Palo Alto, CA, USA March 2018

Towards a Data Efficient Off-Policy Policy Gradient

Josiah P. Hanna and **Peter Stone**
Dept. of Computer Science
The University of Texas at Austin
Austin, TX 78712 USA
{jphanna,pstone}@cs.utexas.edu

Abstract

The ability to learn from off-policy data – data generated from past interaction with the environment – is essential to data efficient reinforcement learning. Recent work has shown that the use of off-policy data not only allows the re-use of data but can even improve performance in comparison to on-policy reinforcement learning. In this work we investigate if a recently proposed method for learning a better data generation policy, commonly called a behavior policy, can also increase the data efficiency of policy gradient reinforcement learning. Empirical results demonstrate that with an appropriately selected behavior policy we can estimate the policy gradient more accurately. The results also motivate further work into developing methods for adapting the behavior policy as the policy we are learning changes.

Introduction

Off-policy RL is necessary for data efficient reinforcement learning. The standard way to incorporate off-policy data into reinforcement learning is to use importance sampling. Unfortunately, policy improvement with importance sampling may exhibit instability due to increased variance (Levine and Koltun 2013; Thomas, Theocharous, and Ghavamzadeh 2015). Recent work has shown that importance sampling can actually lead to more data efficient policy evaluation (Hanna et al. 2017). This work introduced a method called behavior policy gradient (BPG) and demonstrated it can find data generation policies that give low variance importance sampling evaluations. Here we investigate the problem of policy improvement with a data generation policy that has been learned with BPG. Specifically, we investigate whether a behavior policy that gives low variance evaluation of an initial policy can also be used to effectively estimate the direction of the policy gradient and if this same policy can be used for multiple policy gradient updates. Empirical results show that 1) off-policy policy gradient estimates with such a behavior policy lead to larger performance gains with a single update and 2) that this improvement can be realized for a limited number of policy improvement steps before off-policy gradient estimates lead to worse performance than on-policy gradient estimates.

Preliminaries

We assume the environment is represented as a finite horizon, episodic MDP. The agent interacts with the environment in a series of episodes by selecting actions from a policy π. Each episode can be described as a trajectory, H, that consists of a sequence of states, actions, and rewards: $H = S_0, A_0, R_0, ...S_L, A_L, R_L$. The return of a trajectory, denoted $g(h)$, is the sum of rewards along the trajectory: $g(H) = \sum_{t=0}^{L} R_t$. We assume π is a parameterized, stochastic policy with parameter vector $\boldsymbol{\theta}$ and write $H \sim \pi$ to denote sampling a trajectory by following policy π for one episode. The expected value of a policy, π, is $J(\pi) = \mathbf{E}[g(H)|H \sim \pi]$.

In reinforcement learning, policy improvement is the iterative process of updating a policy towards a policy with higher expected return. Denote the initial policy as $\pi_{\boldsymbol{\theta}_0}$. At step i a policy improvement method updates $\boldsymbol{\theta}_i$ to $\boldsymbol{\theta}_{i+1}$ such that $J(\pi_{\boldsymbol{\theta}_{i+1}}) > J(\pi_{\boldsymbol{\theta}_i})$. Policy improvement can continue for a fixed number of iterations or until there is no longer an increase in the expected return.

Naturally, policy improvement requires interaction with the environment. We will refer to the policy that generates the trajectories for a step of policy improvement as the *behavior policy*. The policy being updated is the *target policy*. Methods where the target policy is also the behavior policy are termed *on-policy*; otherwise, they are *off-policy*.

Policy Gradient Reinforcement Learning Policy gradient methods are a popular class of reinforcement learning algorithms used for policy improvement (Deisenroth et al. 2013). Policy gradient methods attempt to maximize the expected return of a policy $\pi_{\boldsymbol{\theta}}$ with respect to the policy parameters $\boldsymbol{\theta}$. This gradient can be derived as:

$$\frac{\partial}{\partial \boldsymbol{\theta}} J(\pi_{\boldsymbol{\theta}}) = \mathbf{E}\left[g(H) \sum_{t=0}^{L} \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_t | S_t) \right] \qquad (1)$$

where $H \sim \pi_{\boldsymbol{\theta}}$. The simplest policy gradient method is the REINFORCE algorithm which adapts the policy with unbiased estimates of Eq. 1 (Williams 1992). In this form, estimates of the policy gradient often suffer from high variance. Extensive work has gone in to reducing this variance in order to scale policy gradient methods to complex problems (e.g., (Peters, Mülling, and Altun 2010; Greensmith et al. 2001; Schulman et al. 2015; 2016; Gu et al. 2017)). As a result, policy gradient methods are a widely applied class of RL algorithms.

Note that policy gradient methods are typically on-policy methods in that we estimate the gradient at π with trajectories sampled from π. In practice this means that at step i of learning, policy π_i is used to collect a dataset of trajectories, $\mathcal{D}_i$, $\mathcal{D}_i$ is used to estimate (1), a gradient step is taken on $\boldsymbol{\theta}_i$, and then $\mathcal{D}_i$ *is discarded* and the process repeats with policy π_{i+1}.

Behavior Policy Search This section describes a recently proposed off-policy method for policy evaluation that uses *importance sampling* to lower the variance of policy evaluation. In the next section we will adapt this idea to the policy gradient setting.

Consider the policy evaluation setting where our goal is to evaluate a *target policy*, π. The simplest approach is to execute π for multiple episodes and average the resulting returns. Unfortunately, this Monte Carlo estimator may have high variance when the target policy rarely experiences trajectories with high-magnitude return.

Instead of running π, we can instead run a different behavior policy, π_b and weight the resulting returns according to the likelihood of seeing them under π instead of π_b. This approach allows us to over-sample these rare, high-magnitude returns and then weight them according to their true likelihood. Importance sampling is an *unbiased* method for computing the re-weighting. The importance sampled return of a trajectory H is:

$$\mathrm{IS}(H, \pi_b) = \prod_{t=0}^{L} \frac{\pi(A_t | S_t}{\pi_b(A_t | S_t)} \cdot g(H)$$

Given a dataset of trajectories, $\mathcal{D}$, generated by π_b the importance sampling estimator is the mean of $\mathrm{IS}(H, \pi_b)$ over all $H \in \mathcal{D}$.

Recent work by Hanna et al. demonstrated that it is possible to find a behavior policy that leads to lower variance policy evaluation compared to Monte Carlo policy evaluation (Hanna et al. 2017). Their *behavior policy gradient* (BPG) method used gradient descent on the variance of the importance sampling estimator to adapt a parameterized behavior policy towards a locally optimal behavior policy. The result of running BPG for a particular target policy π is a behavior policy, π_b, that generates data for low variance importance sampling evaluation of a π. This low variance evaluation is only guaranteed for a static target policy.

Off-Policy Policy Gradient

This section discusses how we can apply behavior policy search to policy gradient methods. While there have been many important contributions since Williams' original REINFORCE work, we will primarily discuss REINFORCE and note that other approaches (e.g., optimal baselines (Greensmith et al. 2001; Peters and Schaal 2008), trust-regions (Peters, Mülling, and Altun 2010; Schulman et al. 2015), etc.) could be combined with the presented approach in future work.

The REINFORCE method can be adapted to an off-policy variant by using unbiased estimates of an importance-sampled version of Equation 1

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

where $H \sim \pi_b$. As in policy evaluation, if π_b is chosen arbitrarily gradient estimates are likely to have high variance. On the other hand, if we can select π_b appropriately then our gradient estimate may have less variance than the on-policy version.

We will select π_b to be a behavior policy that minimizes the variance of an importance sampling evaluation of the current policy. This approach allows us to directly apply BPG to learn π_b. In contrast, previous work has considered the trace of the gradient covariance matrix as the measure of gradient variance (Peters and Schaal 2008; Gu et al. 2017; Ciosek and Whiteson 2017; Bouchard et al. 2016). Minimizing this variance measure is equivalent to minimizing the variance of each component of the gradient. This measure has been used in previous work on adaptive importance sampling for stochastic gradient descent (Bouchard et al. 2016; Ciosek and Whiteson 2017). One downside of this measure is that it may be sensitive to the scale of the policy parameterization. Minimizing the variance of policy evaluation is scale-invariant although it is *not* guaranteed to lower policy gradient variance.

The other challenge with developing an off-policy REINFORCE method is the need to track the current policy. If we start with π_b that gives low variance policy gradient estimates for the initial policy it may not give low variance estimates after the initial policy has changed. One of our experiments attempts to evaluate the scale of this problem.

Empirical Results

We present two experiments using the Cartpole domain implented in OpenAI gym (Brockman et al. 2016). The policy is a softmax distribution over actions where the logits come from a linear combination of state variables.

Method	Average Return (std.)
Random π_b	54.92 (8.27)
On-policy	55.081 (1.31)
Optimized π_b	**68.656 (15.7)**

Table 1: Comparison of one-step improvement in average return when estimating the policy gradient with off-policy and on-policy policy REINFORCE. For each behavior policy we sample 200 trajectories and estimate the policy gradient direction with (2). The gradient step size is computed with a line search. Results are averaged over 50 independent runs.

The initial behavior policy is trained with BPG to minimize the variance of an importance sampling evaluation of the initial policy.

We design experiments to answer the questions 1) does a behavior policy selected with BPG lead to better estimation of the policy gradient direction and 2) can a behavior policy selected with BPG be used for multiple policy gradient updates?

Policy Improvement Step Quality

Our first experiment compares the quality of the update direction computed with an off-policy REINFORCE method to the quality of the update direction computed with standard REINFORCE. In order to make this comparison, we sample a batch of trajectories with the initial policy and another batch with π_b. We estimate the on-policy REINFORCE gradient, the off-policy REINFORCE gradient estimated with a behavior policy trained with BPG to evaluate the initial policy, and the off-policy REINFORCE gradient estimated with a randomly initialized behavior policy. For each method we select the optimal step-size for each method with a line search on (π). We use a line search to avoid conflating gradient direction with gradient magnitude.

Table 1 shows that the average gradient direction computed with off-policy REINFORCE leads to a much larger increase in expected return. However, we also point out that the variance of the performance improvement is also higher. While in most cases expected performance increases above the increase obtained by on-policy REINFORCE or random policy off-policy REINFORCE, the fact that the variance of the improvement has increased may suggest that lowering the variance of policy evaluation does *not* necessarily lead to a lower variance policy gradient estimate.

Multi-step Policy Improvement

Our second experiment investigates if a behavior policy trained to evaluate the initial policy can be used to estimate the policy gradient at other policies. For this experiment, we collect a single set of 100 trajectories with the behavior policy and adapt the target policy with off-policy REINFORCE for 10 iterations.

Figure 1 demonstrate that an improved π_b for importance sampling evaluation can lead to faster learning compared to on-policy REINFORCE – even without re-sampling new trajectories. However, the improvement is relegated to the first few iterations of policy improvement before the target policy has changed significantly.

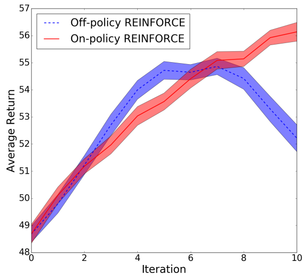

Figure 1: Comparison of multi-step improvement in average return when estimating the policy gradient with off-policy and on-policy REINFORCE.

Discussion and Open Questions

Our empirical results have shown that off-policy policy gradient estimates can give a more accurate estimate of the direction of the policy gradient better than on-policy policy gradient estimates. Our results also show that off-policy REINFORCE with a behavior policy trained with BPG can lead to faster initial learning but that performance degrades once the current policy has been adapted away from the initial policy. In order to develop a complete, low variance off-policy REINFORCE method it will be important to address the question of how to adapt the behavior policy so that it continues to lower variance as the current policy changes.

An alternative to adapting the behavior policy to track the current policy is to start with a behavior policy that generalizes to other policies along the trajectory of learning. One approach towards finding such a policy would be to regularize BPG so that it does not overfit to the policy it is trained to evaluate or to use meta-learning techniques to learn a behavior policy that can be quickly adapted to estimate the policy gradient for a new target policy (Finn, Abbeel, and Levine 2017).

Conclusion

We have presented preliminary steps towards a policy gradient algorithm that uses off-policy data for more efficient updates. We have described how a recently proposed behavior policy search method could be adapted to the policy improvement setting. We then presented experiments showing that a carefully selected behavior policy can improve the step direction of the REINFORCE method and that this same behavior policy can be used for multiple updates before it performs worse than an on-policy update. These results indicate that research into how to adapt the behavior policy as the policy being learned changes has the potential to further improve the data efficiency of policy gradient reinforcement learning.

Acknowledgements

We thank the anonymous reviewers for useful comments on the work and its presentation. This work has taken place in the Learning Agents Research Group (LARG) at the Artificial Intelligence Laboratory, The University of Texas at Austin. LARG research is supported in part by NSF (IIS-1637736, IIS-1651089, IIS-1724157), Intel, Raytheon, and Lockheed Martin. Peter Stone serves on the Board of Directors of Cogitai, Inc. The terms of this arrangement have been reviewed and approved by the University of Texas at Austin in accordance with its policy on objectivity in research.

References

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

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

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

Deisenroth, M. P.; Neumann, G.; Peters, J.; et al. 2013. A survey on policy search for robotics. *Foundations and Trends® in Robotics* 2(1–2):1–142.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*.

Greensmith, E.; Bartlett, P. L.; Baxter, J.; et al. 2001. Variance reduction techniques for gradient estimates in reinforcement learning. In *Proceedings of the 14th Conference on Neural Information Processing Systems (NIPS)*, 1507–1514.

Gu, S.; Lillicrap, T.; Ghahramani, Z.; Turner, R. E.; and Levine, S. 2017. Q-prop: Sample-efficient policy gradient with an off-policy critic.

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

Levine, S., and Koltun, V. 2013. Guided policy search. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 1–9.

Peters, J., and Schaal, S. 2008. Reinforcement learning of motor skills with policy gradients. *Neural networks* 21(4):682–697.

Peters, J.; Mülling, K.; and Altun, Y. 2010. Relative entropy policy search. In *AAAI*, 1607–1612. Atlanta.

Schulman, J.; Levine, S.; Abbeel, P.; Jordan, M.; and Moritz, P. 2015. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, 1889–1897.

Schulman, J.; Moritz, P.; Levine, S.; Jordan, M.; and Abbeel, P. 2016. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations*.

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

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