HIGH-DIMENSIONAL CONTINUOUS CONTROL USING GENERALIZED ADVANTAGE ESTIMATION

John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan and Pieter Abbeel
Department of Electrical Engineering and Computer Science
University of California, Berkeley
fjoschu,pcmoritz,levine,jordan,pabbeelg@eecs.berkeley.edu

1 Introduction

For a long time,scholars make effort on better estimation for reinforcement learning,for example: estimator of value function or advantage function.REINFORCE(Monte Carlo) provided a unbiased estimator but with large cost of variance.Temporal difference offered another estimator with some bias but less variance. Policy gradient methods directly optimize the cumulative reward and can be applied to neural network simply. Two challenges conquered by this paper.

- **Large number of samples typically required.**
  We know that reinforcement learning surpasses human but with less efficiency.That's why we need tremendous amount of data.This problem can be solved by estimator with lower variance at cost of tolerable bias.
- **Stable and steady improvement.**
  In practice,line search (gradient descent) might fail somehow in reinforcement learning.For this problem trust region optimization procedure can deal with it for both the policy and the value function.

2 Problem Formulation

Consider undiscounted formulation of policy optimization first:

- Initial State $S_0 \sim P_0$
- Trajectory (S_0, a_0, S_1, a_1) for $a_t \sim \pi(a_t, S_t)$,$S_{t+1} \sim P(S_{t+1}, |S_t, a_t)$
- Return $r_t = r(S_t, a_t, S_{t+1})$

Note that we assume that $\sum_{n=0}^{+\infty} r_t < +\infty$.
Also,for the discounted reward, the discount factor γ will be used in bias-variance trade-off.
We can also rewrite the following formula.

$$\sum_{n=-\infty}^{+\infty} \gamma^t r_t = \sum_{n=-\infty}^{+\infty} r_t^* \quad , for \quad \gamma^t r_t = r_t^*$$

Preprint. Under review.

3 Theoretical Analysis

- **Policy Gradient**

  repeatedly estimating the gradient g:=$\mathbb{E}\left[\sum_{t=0}^{+\infty}\Psi_t\nabla_\theta log\pi_\theta(a_t|S_t)\right]$

  where Ψ_t may be one of the following:
  1. $\sum_{t=0}^{\infty} r_t$
  2. $\sum_{t=0}^{\infty} r_{t^*}$ (reward after following action a_t)
  3. $\sum_{t^*=t}^{\infty} r_{t^*} - b(S_{t^*})$ (baseline version of 2)
  4. $Q^\pi(S_t, a_t)$
  5. $A^\pi(S_t, a_t)$
  6. $r_t + V(S_{t+1}) - V(S_t)$

  For $\Psi_t = A^\pi(S_t, a_t)$(actually we do not know what it actually is,but we have to estimate it) We know $A^\pi(S,a)$ = $Q^\pi(S,a)$ $-V^\pi(S,a)$,measures whether or not the action is better or worse than the policy's default behavior.Then,with the gradient term $\Psi_t\nabla log\pi_\theta(a_t|S_t)$ points in the direction of increased $\pi_\theta(a_t, S_t)$ if and only if $A^\pi(a_t, S_t) > 0$.

- **Discount factor** γ

  The discount factor γ we see in $\sum_{t=0}^{\infty}\gamma^t r_t$ is actually a control term to reduce variance.Here we have to define the discounted approximation to the policy gradient first:

  $$g^\gamma := \mathbb{E}\left[\sum_{t=0}^{\infty} A^{\pi,\gamma}(S_t, a_t)\nabla log\pi_\theta(s_t, a_t)\right]$$

  Before we continue, we have to introduce the definition of γ-just.

  Def. The estimator $\hat{A}_t$ is γ-just for all t ,if

  $\mathbb{E}\left[\sum_{t=0}^{\infty}\hat{A}_t(s_{0:\infty}, a_{0:\infty})\nabla_\theta log\pi_\theta(a_t|s_t)\right]$ = $\mathbb{E}\left[\sum_{t=0}^{\infty} A^{\pi,\gamma}(S_t, a_t)\nabla log\pi_\theta(s_t, a_t)\right]$

  Also,is follows

  $$\mathbb{E}\left[\sum_{t=0}^{\infty}\hat{A}_t(s_{0:\infty}, a_{0:\infty})\nabla_\theta log\pi_\theta(a_t|s_t)\right] = g^\gamma$$

- **Advantage Function Estimation**

  Let V be an approximate value fimctoion.

  Define δ_t^V=$r_t + \gamma V(S_{t+1}) - V(S_t)$a.k.a,the TD residual.

  If we have $V = V^{\pi,\gamma}$,then it is a γ-just an unbiased advantage estimator of $A^{\pi,\gamma}$:

  $$\begin{aligned}\mathbb{E}\left[\delta^{V_t^{\pi,\gamma}}\right] &= \mathbb{E}\left[r_t + \gamma V^{\pi,\gamma}(S_{t+1}) - V^{\pi,\gamma}(S_t)\right] \\ &= \mathbb{E}\left[Q^{\pi,\gamma}(S_t, a_t) - V^{\pi,\gamma}(S_t)\right] = A^{\pi,\gamma}(S_t, a_t).\end{aligned}$$

  But ideally,this hold for only γ-just V,otherwise, this yield biased policy gradient estimator.

  Now define $\hat{A}_t^{(k)} := \sum_{l=0}^{k-1}\gamma^l\delta_{t+l}^V$a.k.a k-step estimate of the returns but

minus a baseline term $-Vs_t$
Note the bias generally becomes smaller as k $\to \infty$.Taking k $\to$,we get
$$\hat{A}_t^{\infty} = -V(s_t) + \sum_{l=0}^{\infty} \gamma^l r_{t+l}$$
Then we can define the generalized advantage estimator GAE(γ, λ),defined as the exponentially -weighted average of these k-step estimators:
$$\hat{A}_t^{GAE(\gamma,\lambda)} := \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l}^V$$
we can see that
$GAE(\gamma, 0): \quad \hat{A}_t := \delta_t \qquad = r_t + \gamma V(s_{t+1}) - V(st)$
$GAE(\gamma, 0): \quad \hat{A}_t := \sum_{l=0}^{\infty} \gamma^l \delta_{t+l} = r_t + \gamma V(s_{t+1}) - V(st)$
We can see for $\lambda = 0$,the TD residual form GAE we know introduce lower bias but with some bias.
For $\lambda = 1$,is actually MC return minus baseline.

- **Interpretation as reward**
Reward shaping refers to the following transformation of the reward function of an MDP:let $\Phi : \mathrm{S} \to \mathbb{R}$ be an arbitrary scalar-valued function on state space.
Define the transformed reward function $\hat{r}$ to be
$$\hat{r}(s, s, s') = r(s, a, s') + \gamma\Phi(s') - \Phi(s)$$
Then we can see that the discounted sum of rewards of a trajectory starting with state S_t :
$\sum_{l=0}^{\infty} \gamma^l \hat{r}(s_{t+l}, a_t, s_{t+l+1}) = \sum_{l=0}^{\infty} \gamma^l r s_{t+l}, a_t, s_{t+l+1} - \Phi(s_t)$
Then with the construction,let
$\tilde{Q}^{\pi,\gamma}(s, a) = Q^{\pi,\gamma}(s, a) - \Phi(s)$
$\tilde{V}^{\pi,\gamma}(s, a) = V^{\pi,\gamma}(s, a) - \Phi(s)$
$\tilde{A}^{\pi,\gamma}(s, a) = (Q^{\pi,\gamma}(s, a) - \Phi(s)) - (V^{\pi,\gamma}(s, a) - \Phi(s)) = A^{\pi,\gamma}(s, a)$
Now we consider discount $\gamma\lambda$,this is actually a steeper discount factor than γ since $0 \leq \lambda \leq 1$.Now,let $\Phi = V$ We see that
$\sum_{l=0}^{\infty} (\gamma\lambda)^l \hat{r}(s_{t+l}, a_t, s_{t+l+1}) = \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l}^V = \hat{A}_t^{GAE(\gamma,\lambda)}$
It's useful to introduce the notion of a response function $\mathcal{X}$ It's defined as follows:
$$\mathcal{X}(l; s_t, a_t) = \mathbb{E}\big[r_{t+l} | s_t, a_t\big] - \mathbb{E}\big[r_{t+l} | s_t\big]$$
Note that $A^{\pi,\gamma}(s, a) = \sum_{l=0}^{\infty} \gamma^l \mathcal{X}(l; s, a)$
We can see that
$\nabla_\theta log\pi_\theta(a_t|s_t) A^{\pi,\gamma}(s_t, a_t)$ = $\nabla_\theta log\pi_\theta(a_t|s_t) \sum_{l=0}^{\infty} \mathcal{X}(l; s_t, a_t)$
Using a discount $\gamma < 1$ corresponding to dropping the terms with $l \gg 1/(1-\gamma)$
Thus the error introduced by the approximation will be small if $\mathcal{X}$ rapidly decays as l increases,i.e,if the effect of an action on rewards is 'forgotten' after $\approx 1/(1-\gamma)$ timesteps.

4 Conclusion

With the proof above,we can see this paper provide an estimator with steeper discount factor $\gamma and \lambda$ which can interpreted from MDP and discounted factor. With a better estimator, we can get better estimation.This paper not only provides a better way to control the bias-variance trade-off but also

combined the traditional MDP to neural method.GAE is such an important way for variance reduction that the baseline TRPO is implemented with GAE method.

References

High-Dimensional Continuous Control Using Generalized Advantage Estimation