High-Dimensional Continuous Control Using Generalized Advantage Estimation

Presented by Jialun Lyu and Zhibo Zhang

Original Paper: Schulman, John & Moritz, Philipp & Levine, Sergey & Jordan, Michael & Abbeel, Pieter. (2015). High-Dimensional Continuous Control Using Generalized Advantage Estimation.

Motivation

In class, we saw the policy gradient for a discounted reword problem that has the following form:

$$\nabla J(\theta) = E[\sum_{t=0}^{\infty} \gamma^t \mathrm{r}_t \nabla_\theta log\pi_\theta(a_t|s_t)]$$

We would introduce to you a new form of policy gradient

$$\nabla J(\theta) = E[\sum_{t=0}^{\infty} A^{\pi,\gamma}(s_t, a_t) \nabla_\theta log\pi_\theta(a_t|s_t)]$$

where

$$A^{\pi,\gamma}(s_t, a_t) = Q^{\pi,\gamma}(s_t, a_t) - V^{\pi,\gamma}(s_t)$$

is called advantage function.

We would derive General Advantage Estimator for $A^{\pi,\gamma}(s_t, a_t)$,

$$\hat{A}^{GAE(\gamma,\lambda)}$$

Which is controlled by two parameters γ, λ

Problem Setting

Our usual setting in RL:

A trajectory $(s_0, a_0, s_1, a_1, \ldots)$ is generated by $a_t \sim \pi(a_t|s_t)$ and Transitional rule $s_{t+1} \sim P(s_{t+1}|s_t, a_t)$

Our goal is to maximize the expected total reward $\sum_{t=0}^{\infty} r_t$, where $r_t = r(a_t, s_t)$ is received at each timestamp.

In a discounted total reward version, we have $\sum_{t=0}^{\infty} \gamma^t r_t$, where γ is the discount factor.

We can avoid convergence problem due to infinite horizon and control the scale of the overall return through settings of γ.

Advantage Function

$$A^{\pi,\gamma}(s_t, a_t) = Q^{\pi,\gamma}(s_t, a_t) - V^{\pi,\gamma}(s_t)$$

Intuitively, $A^{\pi,\gamma}(s_t, a_t)$ is the "advantage" of taking a specific action a_t at state s_t, comparing to the "average" reward across all possible actions at s_t.

Recall that mathematically,

$$Q^{\pi,\gamma}(s_t, a_t) = E_{s_{t+1:\infty}, a_{t+1:\infty}}[\sum_{l=0}^{\infty} \gamma^l r_{t+l}]$$

and

$$V^{\pi,\gamma}(s_t) = E_{s_{t+1:\infty}, a_{t:\infty}}[\sum_{l=0}^{\infty} \gamma^l r_{t+l}]$$

The difference between $Q^{\pi,\gamma}(s_t, a_t)$ and $V^{\pi,\gamma}(s_t)$ is the range of expected value.

$A^{\pi,\gamma}(s_t, a_t)$ VS $\gamma^{\mathrm{t}}\mathrm{r}_{\mathrm{t}}$ in Policy Gradient

Why do we use advantage function in policy gradient?

1. A step in the policy gradient direction should increase the probability of better-than-average actions and decrease the probability of worse-than-average actions.

2. Choosing $A^{\pi,\gamma}(s_t, a_t)$ makes sure that for any step t, $A^{\pi,\gamma}(s_t, a_t)\nabla_\theta log\pi_\theta(a_t|s_t)$ points to the direction of increased $\pi_\theta(a_t|s_t)$ if and only if $A^{\pi,\gamma}(a_t, s_t) > 0$.

3. Comparing to using overall return, using advantage function results in a typically smaller variance for policy gradient.

Estimating $A^{\pi,\gamma}(s_t, a_t)$

We now want to find an estimator $\hat{A}$ such that the policy gradient has form

$$E[\sum_{t=0}^{\infty} \hat{A}_t \nabla_\theta log \pi_\theta(a_t|s_t)]$$

to estimate

$$E[\sum_{t=0}^{\infty} A^{\pi,\gamma}(s_t, a_t) \nabla_\theta log \pi_\theta(a_t|s_t)]$$

If the above two expectation is equivalent, we say $\hat{A}_t$ is “unbiased”.

Finding $\hat{A}_t$

There are several options for $\hat{A}_t$, such that the policy gradient is unbiased.

Suppose that we have some value function, denoted V.

Here we introduce TD (Temporal Difference) residual using V ,

$$\delta_t^V = r_t + \gamma V(s_{t+1}) - V(s_t)$$

In fact, if we know the true value function $V^{\pi,\gamma}$,

Taking expected value $E_{s_{t+1}}\left[\delta_t^{V^{\pi,\gamma}}\right] = E_{s_{t+1}}[Q^{\pi,\gamma}(s_t,a_t) - V^{\pi,\gamma}(s_t)] = A^{\pi,\gamma}(s_t,a_t)$

$\delta_t^{V^{\pi,\gamma}}$ is unbiased to estimate $A^{\pi,\gamma}(s_t,a_t)$!

But we don’t know $V^{\pi,\gamma}$, however, we can use δ_t^V as a starting point to construct an estimator with some approximation V .

Finding $\hat{A}_t$

Consider taking sum of k of these TD residuals, denoted by $\hat{A}_t^{(k)}$.

$$\hat{A}_t^{(1)} = \delta_t^V = -V(s_t) + r_t + \gamma V(s_{t+1})$$

$$\hat{A}_t^{(2)} = \delta_t^V + \gamma \delta_{t+1}^V = -V(s_t) + r_t + \gamma r_{t+1} + \gamma^2 V(s_{t+2})$$

$$\hat{A}_t^{(k)} = \sum_{l=0}^{k-1} \gamma^l \delta_{t+l}^V = -V(s_t) + r_t + \gamma r_{t+1} + \cdots + \gamma^{k-1} r_{t+k-1} + \gamma^k V(s_{t+k})$$

As k increases the “inaccuracy” in $V(s_{t+k})$ becomes smaller.

$$\hat{A}_t^{(\infty)} = \sum_{l=0}^{\infty} \gamma^l \delta_{t+l}^V = -V(s_t) + \sum_{l=0}^{\infty} \gamma^l r_{t+l}$$

We have a very compact form for $\hat{A}_t^{(\infty)}$, which is essentially the empirical return minus the value function baseline.

Finding $\hat{A}_t$

Now we have

$$\hat{A}_t^{(\infty)} = -V(s_t) + \sum_{l=0}^{\infty} \gamma^l r_{t+l}$$

What about the “inaccurate” baseline $V(s_t)$?

As it turns out, we don’t need to worry about the choice of V!

The intuition is that $V(s_t)$ is a function of s_t, when multiplied by $\nabla_\theta log\pi_\theta(a_t|s_t)$ and taking expected value over the future trajectory $(s_{t+1:\infty}, a_{t:\infty})$. The term

$$E_{s_{t+1:\infty},a_{t:\infty}}[\nabla_\theta log\pi_\theta(a_t|s_t)V(s_t)] = 0$$

because of $\mathrm{E}[\nabla_\theta log\pi_\theta(a_t|s_t)] = 0$, and $V(s_t)$ can be taken outside of the expectation!

Main takeaway: $\hat{A}_t^{(\infty)}$ is ”unbiased” regardless of the choice of V!

Final remark on $\hat{A}_t^{(k)}$

Now we have $\hat{A}_t^{(1)}$, …, $\hat{A}_t^{(k)}$, …, $\hat{A}_t^{(\infty)}$.

The bias decreases as k goes from 1 to ∞.

This is because $\gamma^k V(s_{t+k})$ becomes more and more discounted by γ as k increases.

The bias as the result of inaccurate V becomes smaller as well.

But the variance increases, as k increases! You may interpret that, as k increases, the number of random variables in the summation increases. Each of them contributes some variation, as the result $\hat{A}_t^{(\infty)}$ has the largest variation.

So far, $\hat{A}_t^{(k)}$ allows us to control bias-variance through k.

Final step! We want to ”build” a general expression using $\hat{A}_t^{(k)}$.

General Advantage Estimation (GAE)

Finally, we can build GAE, which is controlled by two parameters γ, λ, which is defined to be an exponentially weighted summation of $\hat{A}_t^{(1)}$, $\hat{A}_t^{(2)}$, …, $\hat{A}_t^{(\infty)}$

$$\hat{A}^{GAE(\gamma,\lambda)} = (1-\lambda)(\hat{A}_t^{(1)} + \lambda\hat{A}_t^{(2)} + \lambda^2\hat{A}_t^{(3)} + \cdots)$$

$\hat{A}_t^{(1)}$ has weight $(1-\lambda)$, $\hat{A}_t^{(k)}$ has weight $\lambda^k(1-\lambda)$

$$\lambda \in [0,1]$$

is the control of weight for $\hat{A}_t^{(k)}$

If we want GAE to be more unbiased,

we should assign more weight towards $\hat{A}_t^{(\infty)}$ or larger k.

If we want GAE have less variation,

we should assign more weight towards $\hat{A}_t^{(1)}$ or smaller k.

General Advantage Estimation (GAE)

After some cleanup,

$$\hat{A}^{GAE(\gamma,\lambda)} = \sum_{l=0}^{\infty}(\gamma\lambda)^{l}\delta_{t+l}^{V}$$

where δ_{t+l}^{V} is the TD residual with form $\delta_{t+l}^{V} = r_{t+l} + \gamma V(s_{t+l+1}) - V(s_{t+l})$.

GAE can be expressed by exponentially weighted summation from both $\hat{A}_t^k$ or TD residuals.

Finally, we can plug in $\hat{A}^{GAE(\gamma,\lambda)}$ into an empirical estimator for policy gradient to update our policy using N trajectories :

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{\infty}\hat{A}_t^{GAE(\gamma,\lambda)}\nabla_\theta log\pi_\theta(a_t^n|s_t^n)$$

How do we handle infinite time horizon (limitation due to computational cost)?

We could sample trajectories until the MDP terminates, after which the reward would be 0.

Or empirically, we can sample trajectories up to some very large timestamp, as large as our compute power allows.

We can set a “cut off” for λ, after some small number, we treat the weight λ as 0 when computing $\hat{A}^{GAE(\gamma,\lambda)}$.

Summary

We want to estimate $E[\sum_{t=1}^{\infty} A^{\pi,\gamma} \nabla_\theta log\pi_\theta(a_t|s_t)]$ as our policy gradient.

One of the parameters, γ , can be interpreted discount factor in a discounted return problem. We can also view γ as a control to the scale of the expected total return.

Then we aim to estimate $E[A^{\pi,\gamma} \nabla_\theta log\pi_\theta(a_t|s_t)]$ by using GAE,

$$\hat{A}^{GAE(\gamma,\lambda)} = \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l}^V$$

Using some approximate value function V.

Finally, we can control the bias and variance of the estimator in accordance with our needs in our problem.

$\lambda \to 0$ means estimated policy gradient is becoming more biased and has smaller variance.

$\lambda \to 1$ means estimated policy gradient is becoming less biased and has larger variance.

Schulman, John & Moritz, Philipp & Levine, Sergey & Jordan, Michael & Abbeel, Pieter. (2015). High-Dimensional Continuous Control Using Generalized Advantage Estimation.