EE675A : Introduction to Reinforcement Learning

Lecture 22: Actor-critic and Baseline concept in Policy Gradients

12-04-2023

Lecturer: Subrahmanya Swamy Peruru *Scribe: Harsh Saroha*

In the previous lecture, we learned about policy gradient methods and there used cases and also re caped the policy gradient method for bandit setting. In this lecture, we will be focusing on basic policy gradient method and explore actor-critic method and look into its advantages.

1 Bandit Setting

In a bandit setting, the objective to find an optimal policy is:

$$\max_{\theta} \eta(\theta) = \sum_{a} \pi(a;\theta)\mu(a) \tag{1}$$

We have derived the gradient previously as:

$$\nabla \eta(\theta) = E_{\pi}[\mu(a)\nabla_{\theta} \log(\pi(a;\theta)] \tag{2}$$

The $Vanilla$ version of the policy gradient is referred to as **"REINFORCE"** and approximates the gradient as

$$
\begin{aligned}
&= E_{\pi}[E[R_t|A_t]\nabla_{\theta} \log(\pi(A_t;\theta)] \\
&= E_{\pi}[R_t \nabla_{\theta} \log(\pi(A_t;\theta)] \\
&= R_t \nabla_{\theta} \log(\pi(A_t;\theta) \qquad (3)
\end{aligned}
$$

2 Actor-Critic Version

Instead of using R_t as a sample estimate of $\mu(A_t)$ we could use $\hat{\mu}(A_t)$ which is the sample average of the rewards obtained for arm At so far. $\hat{\mu}(A_t)$ is a better estimate of $\mu(A_t)$ than R_t as it has lower variance. Therefore the policy gradient method convergence improves. So, the actor-critic version of policy gradient approximates the gradient.

$$\nabla \eta_\theta \approx \hat{\mu}(A_t) \nabla_\theta \log(\pi(A_t; \theta)) \tag{4}$$

In the Actor-Critic method, the policy is referred to as the $actor$ that proposes a set of possible actions given a state that is tells us how to act or behave, and the estimated value function is referred to as the $critic$, which evaluates actions taken by the $actor$ based on the given policy.

3 Baseline

The baseline can be any function, even a random variable, as long as it does not vary with a; the equation remains valid because the subtracted quantity is zero.
If we subtract a baseline b as $E_\pi[(\mu(A_t) - b)\nabla_\theta \log(\pi(A_t; \theta)]$, still the gradient remains unchanged if b is not a function of A_t
In other words,

$$E_\pi[(\mu(A_t) - b)\nabla_\theta \log(\pi(A_t; \theta)] = E_\pi[\mu(A_t)\nabla_\theta \log(\pi(A_t; \theta)] \tag{5}$$

Proof:
It suffices to show that $E_\pi[b\nabla_\theta \log(\pi(A_t; \theta)] = 0$
Consider

$$= E_\pi[b\nabla_\theta \log(\pi(A_t; \theta)]$$

$$= \sum_a \pi(a)[b\nabla_\theta \log(\pi(A_t; \theta)]$$

If b is not a functions of A_t = a then,

$$= b \sum_a \pi(a) \frac{1}{\pi(a)} \nabla_\theta \pi(a)$$

$$= b\nabla_\theta \sum_a \pi(a)$$

$$= b\nabla_{theta}(1)$$

$$= 0$$

Typical Baseline used: Theoretically, any baseline "b" which is not a function of A_t works. Depending on the applications and domain knowledge, "b" can be cleverly chosen to obtain better convergence. A typical baseline used is average rewards. $\bar{R}$ **= average rewards obtained till time "t-1" across all the arms.**

$$\nabla_\theta \approx (R_t - \bar{R}) \nabla_\theta \log \pi(A_t; \theta) \tag{6}$$

4 Full RL/MDP setting

Both the concepts of Actor-Critic and baseline can be generalized to the MDP setting.

First, let us use the concept of baseline to simplify the gradient of

$$J(\theta) = E_{S_0}[v_\pi(S_0)]$$

(where S_0 is the starting distance) that we have previously obtained as

$$\nabla J(\theta) = E[G(\upsilon) \sum_{t=0}^{T} \nabla_\theta \log(\pi(A_t|S_t)]$$

where $\upsilon = S_0, A_0, S_1, A_1, ...$ trajectory

$$= E[\sum_{t=0}^{T} \sum_{k=0}^{T} \gamma^k R_{k+1} \nabla_\theta \log \pi(A_t|S_t)]$$

since $\{R_k\}_{k<t}$ doesn't depend on A_t

$$E[R_k \nabla_\theta \log \pi(A_t|S_t)] = 0, \forall k < t$$

$$\nabla_\theta J(\theta) = E_\pi[\sum_{t=0}^{T} \sum_{k=t}^{T} \gamma^k R_{k+1} \nabla_\theta \log \pi(A_t|S_t)]$$

$$\nabla_\theta J(\theta) = E_\pi[\sum_{t=0}^{T} \gamma^t \sum_{k=t}^{T} \gamma^{k-t} R_{k+1} \nabla_\theta \log \pi(A_t|S_t)]$$

$$\nabla_\theta J(\theta) = E_\pi[\sum_{t=0}^{T} \gamma^t G_t \nabla_\theta \log \pi(A_t|S_t)]$$

Further, since

$$E_\pi[G_t|S_t, A_t] = Q_\pi(S_t, A_t)$$

we can write

$$\nabla_\theta J(\theta) = E_\pi[\sum_{t=0}^{T} \gamma^t Q_\pi(S_t, A_t) \nabla_\theta \log \pi(A_t|S_t)] \quad (7)$$

5 Different variants of policy gradient

$$Q_\pi(S_t, A_t) \nabla_\theta \log \pi(A_t|S_t) \quad (8)$$

5.1 REINFORCE (Monte-Carlo policy gradient)

$$G_t \nabla_\theta \log \pi(A_t|S_t) \quad (9)$$

5.2 Actor-Critic

$$\hat{Q}_\pi(S_t, A_t; W) \nabla_\theta \log \pi(A_t|S_t) \quad (10)$$

where $\hat{Q_\pi}$ is approximate estimate of Q_π Here we have two set of parameters θ which take care of policy parameters and w for approximate values for parameters.

6 Advantages of Actor-Critic

$$A_\pi(S_t, A_t) \nabla_\theta \log \pi(A_t|S_t)$$

where $A_\pi(S_t, A_t) = Q_\pi(S_t, A_t) - v_\pi(S_t)$ is the advantage function
NOTE:

- Here we used the baseline trick
- Baseline 'b' can be anything which doesn't depend on A_t.
- In particular, 'b' can depend on s_t
- A typical baseline used is $v_\pi(s_t)$ (Just like $\bar{R}$ the average reward in the bandit problem)

Average reward concept: For continuous MDP setting, instead of using discounted return, average reward is generally used as the metric for optimal policy.

Average reward per time step

$$J(\pi) = E_\pi[R_{t+1}]$$

$$= \mu_\pi(s) E_\pi[R_{t+1}|S]$$

where $\mu_\pi(s)$ is probability of being in state s while following π.

For more details on these topics refer to [1] and [2].

References

[1] D. Silvers. Course on reinforcement learning.

[2] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.