IEMS 469 Project 2

Senmiao Wang

November 14, 2021

1 Overall

For both questions, I try to apply the advantage actor-critic algorithm. In order to achieve this goal, I first construct policy network and value network. In cartpole problem, I use the multi-layer perceptron with one hidden layer for both networks, and in pong problem, I switch to CNN for this problem is more related to image recognition.

When we define the temporal difference error to be

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

By policy gradient theorem with baseline,

$$\begin{aligned}\nabla_\theta J(\pi_\theta) &= \mathbb{E}\left[\sum_{t=0}^{\infty} \log \pi_\theta(a_t|s_t)\Big(Q^{\pi_\theta}(s_t,a_t) - V^{\pi_\theta}(s_t)\Big)\right] \\ &= \mathbb{E}\left[\sum_{t=0}^{\infty} \log \pi_\theta(a_t|s_t)\Big(r_t + \gamma V^{\pi_\theta}(s_{t+1}) - V^{\pi_\theta}(s_t)\Big)\right] \quad = \mathbb{E}\left[\sum_{t=0}^{\infty} \log \pi_\theta(a_t|s_t)\delta_t\right]\end{aligned} \tag{1.1}$$ {?}

Therefore, when we get a trajectory $\{s_0, a_0, s_1, \ldots, s_H\}$, for value network, the critic loss is defined to be

$$l_c(\omega) = \frac{1}{2H}\sum_{t=0}^{H-1} r_t + \gamma V_{\omega-}(s_{t+1}) - V_\omega(s_t)$$

where $\omega-$ is the target network introduced for stationarity.

The actor loss is defined to be

$$l_a(\theta,\omega) = -\frac{1}{H}\sum_{t=0}^{H-1}\Big(r_t + \gamma V_{\omega-}(s_{t+1}) - V_\omega(s_t)\Big)\log \pi_\theta(a_t|s_t)$$

And update the target network $\omega- \leftarrow \omega$ at each step.

2 Cartpole-v0

In the .ipynb file.

3 Pong-v0

I tried my best, but fail to get the convergence of the algorithm. I post the code on the .ipynb file.