RL HW2 DDPG, TRPO, and PPO

110704054 謝弘逵

Problem 1.

(i)

代入 $\pi_\theta = \pi_{\theta_1}$

$$L_{\pi_{\theta_1}}(\pi_{\theta_1}) = \eta(\pi_{\theta_1}) + \sum_s d_\mu^{\pi_{\theta_1}}(s) \sum_a \pi_{\theta_1}(a|s) A^{\pi_{\theta_1}}(s,a)$$

For Any State s,

$$\sum_a \pi_{\theta_1}(a|s) A^{\pi_{\theta_1}}(s,a) = \sum_a \pi_{\theta_1}(a|s)\left[Q^{\pi_{\theta_1}}(s,a) - V^{\pi_{\theta_1}}(s)\right]$$

$$= V^{\pi_{\theta_1}}(s) - V^{\pi_{\theta_1}}(s) = 0$$

$$\Rightarrow L_{\pi_{\theta_1}}(\pi_{\theta_1}) = \eta(\pi_{\theta_1}) + 0 = \eta(\pi_{\theta_1}) \#$$

(ii)

$$\nabla_\theta L_{\pi_{\theta_1}}(\pi_\theta)\Big|_{\theta=\theta_1} = \nabla_\theta \eta(\pi_\theta) + \sum_s d_\mu^{\pi_{\theta_1}}(s) \sum_a \nabla_\theta \left[\pi_\theta(a|s)\right] A^{\pi_{\theta_1}}(s,a)\Big|_{\theta=\theta_1}$$

$A = Q - V$ 代入：

$$\nabla_\theta L_{\pi_{\theta_1}}(\pi_\theta)\Big|_{\theta=\theta_1} = \nabla_\theta \eta(\pi_{\theta_1}) + \sum_s d_\mu^{\pi_{\theta_1}}(s) \nabla_\theta \left(V^{\pi_{\theta_1}}(s) - V^{\pi_{\theta_1}}(s)\right)$$

$$= \nabla_\theta \eta(\pi_{\theta_1}) \#$$

Problem 2.

Lagrangian :

$$\mathcal{L}(\theta,\lambda) = -g^T(\theta-\theta_k) + \lambda\left[\frac{1}{2}(\theta-\theta_k)^T H(\theta-\theta_k) - \delta\right]$$

$$\nabla_\theta \mathcal{L}(\theta,\lambda) = -g + \lambda H(\theta-\theta_k)^T \overset{!}{=} 0$$

$$\Rightarrow \theta(\lambda) = \theta_k + \frac{1}{\lambda} H^{-1} g$$

$\theta(\lambda)$代入$\mathcal{L}$:

$$D(\lambda) = \min_\theta \mathcal{L}(\theta,\lambda) = \mathcal{L}(\theta(\lambda),\lambda)$$

$$= -g^T\left(\frac{1}{\lambda}H^{-1}g\right) + \lambda\left[\frac{1}{2}\left(\frac{1}{\lambda}H^{-1}g\right)^T H\left(\frac{1}{\lambda}H^{-1}g\right) - \delta\right]$$

$$= -\frac{1}{2\lambda} g^T H^{-1} g - \lambda\delta$$

$$\frac{\partial D}{\partial \lambda} = 0 \Rightarrow \lambda^* = \sqrt{\frac{g^T H^{-1} g}{2\delta}}$$

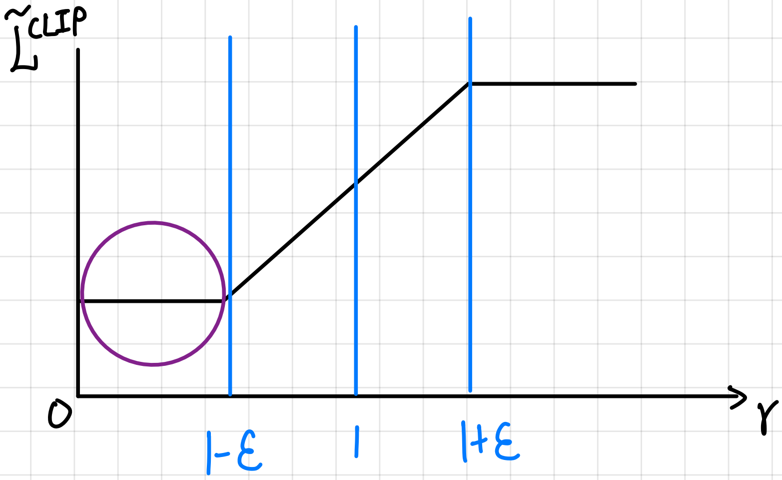
(b)

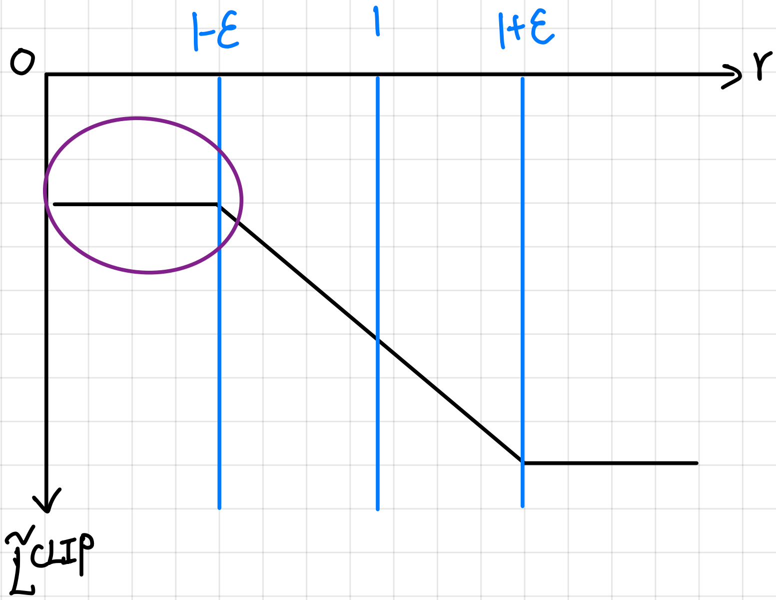
代回求θ^*

$\Rightarrow$ 步長$\alpha = \frac{1}{\lambda^*}$, 則

$$\theta^* = \theta_k + \alpha H^{-1} g, \quad \alpha = \frac{1}{\lambda^*} = \sqrt{\frac{2\delta}{g^T H^{-1} g}}$$

Problem 3.





原始 L^{CLIP}: 根據優勢的正負動態調整裁剪策略,優勢正→保守更新

優勢負→懲罰加重

變体 $\tilde{L}^{CLIP}$: 無論優勢正負,都在 $r<1-\varepsilon$ 或 $r>1+\varepsilon$ 雙測鎖死

Homework 2 Technical Report

Name: Hing-Ting Hsieh

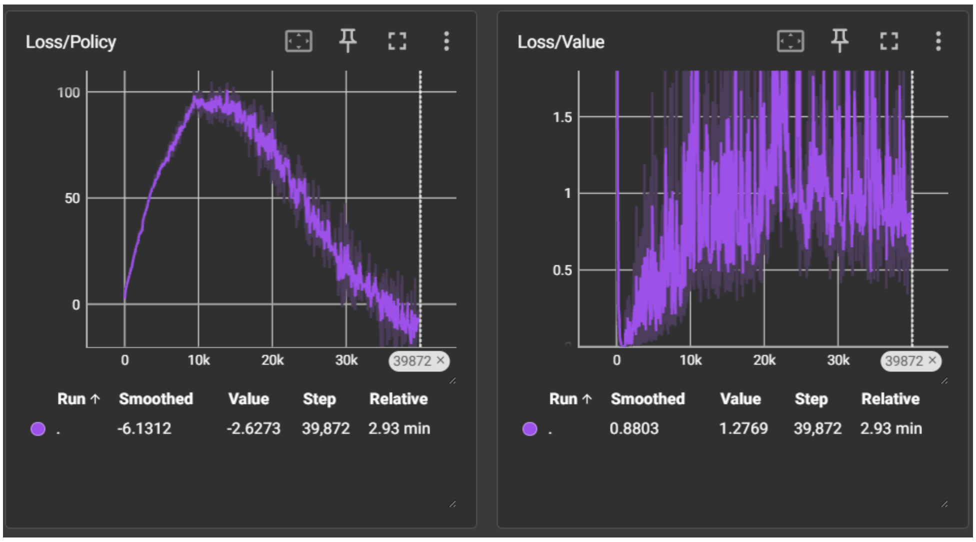
Student ID: 110704054

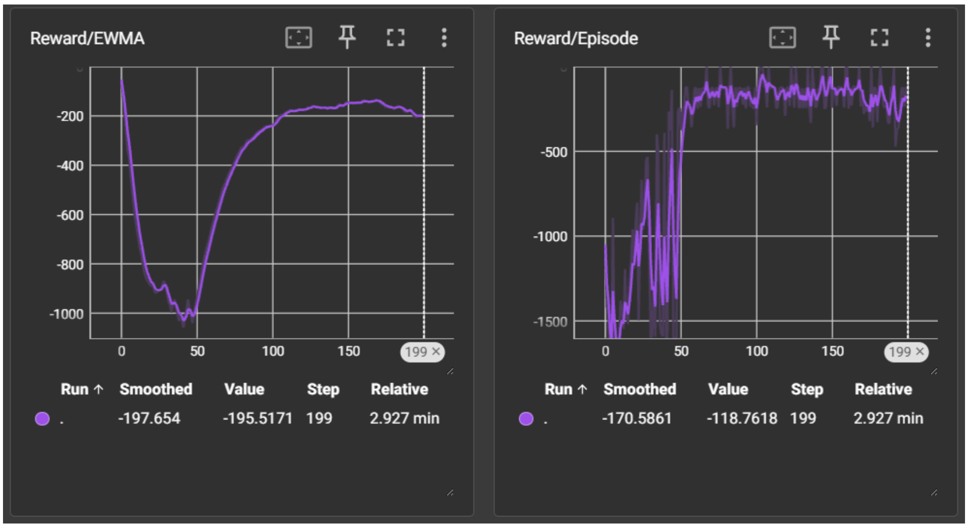
I. Experiment of Simple DDPG

1. Pendulum-v1

Actor Learning Rate	0.0001
Critic Learning Rate	0.001
Batch Size	128
Hidden Size	128
Episodes	200

Result: Reach well policy within 200 episodes.



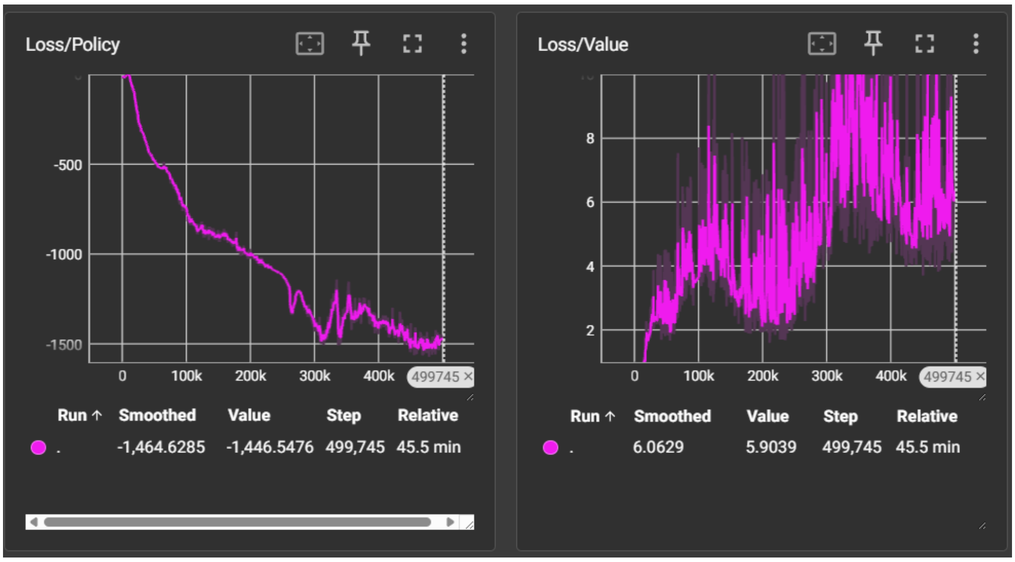


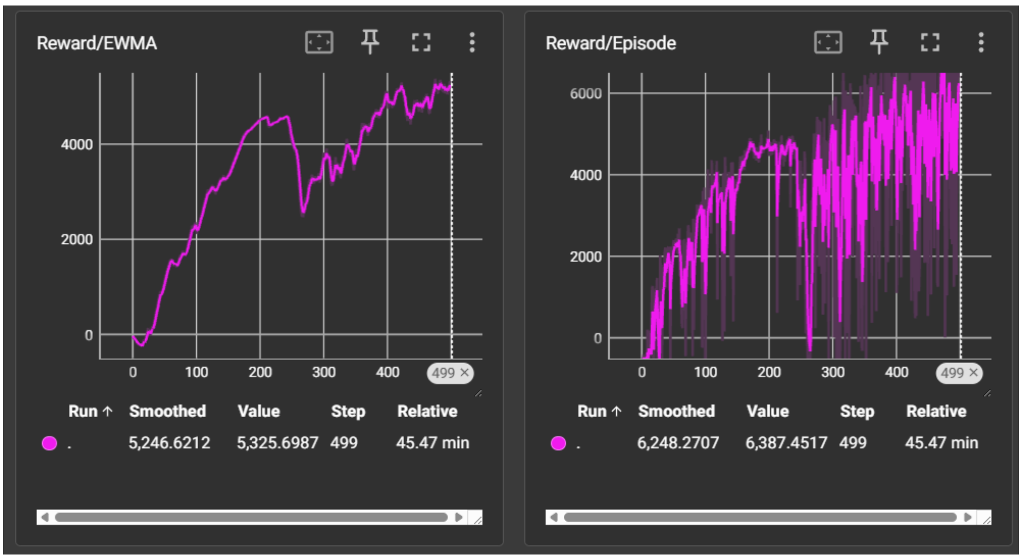
2. Halfcheetah-v2

I implemented the same architecture of the previous task but more batch size and increased tau value.

Actor Learning Rate	0.0001
Critic Learning Rate	0.001
Batch Size	256
Hidden Size	128
Tau	0.02
Episodes	500
Noise Scale	0.1

Result: It went quite well at 500 episodes, but still need more episodes to reach great performance, but the result is still fine.





II. Experiment of DDPG with CDQ

1. Halgcheetah-v2

   After implementing CDQ method, I found that Q-values are conservative and more realistic, furthermore, we got more stable training curves with same hyper parameters.

   Result: It reached well policy within 350 episodes.