HW3 of ECEN 689 RL *

Tianqi Li

April 25, 2022

1 Point-v0

1.1 Analytical gradient

Given the policy distribution $\pi(s|a) \sim \mathcal{N}(\mu, I)$ with mean $\mu = \theta^T s$, the analytical solution is

$$\nabla \log \pi_\theta(s|a) = \nabla_\theta \log \frac{1}{2\pi} \exp(-\frac{1}{2}(a-\mu)^T I(a-\mu)) \tag{1a}$$

$$= -\frac{1}{2}\nabla_\theta (a-\mu)^T I(a-\mu) \tag{1b}$$

$$= -\frac{1}{2}\nabla_\theta (a-\theta^T s)^T I(a-\theta^T s) \tag{1c}$$

$$= -(a-\theta^T s)s^T \tag{1d}$$

1.2 Implement Policy Gradient

Plot of training curves (discounted sum of rewards per iteration) is shown in Fig 1 and 2. As the plot shows, both equation helps the learning in the task Point-v0. The average return of both approaches is -17.5 after 100 iterations.

The rendering of final policy is in assets/Particle-v0_plot_0.mp4

2 CartPole-v0

3 plots un-discounted cumulative reward per episode is shown in Fig 3 - 5. From the observation, only equation 3 (A2C) method managed to solve the problem, which achieves 200 average reward around 15 iterations. The other two approaches are not able to solve this problem.

References

*The code is maintained in `https://github.com/TianqiLi7398/RL-hw/tree/main/hw4`

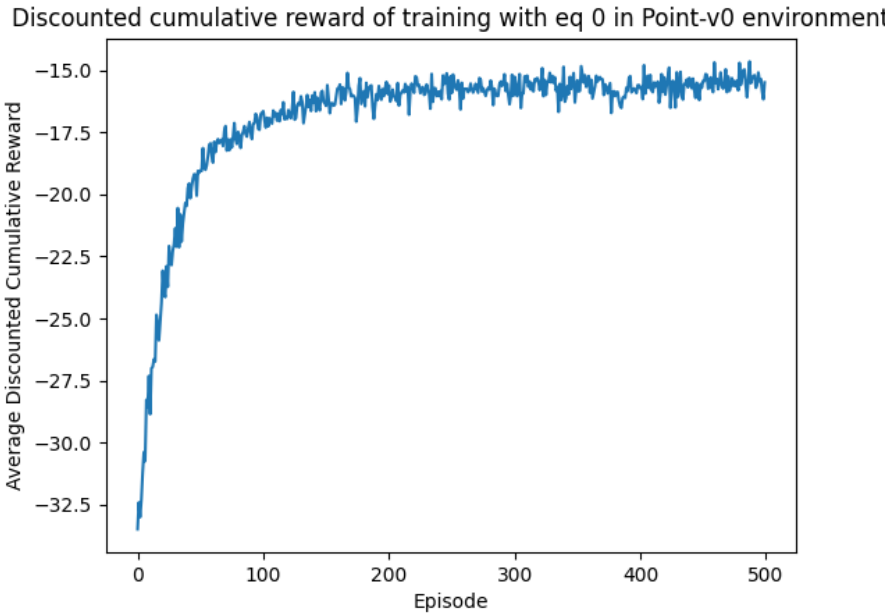

Figure 1: Point-v0 training discounted reward by equation 1

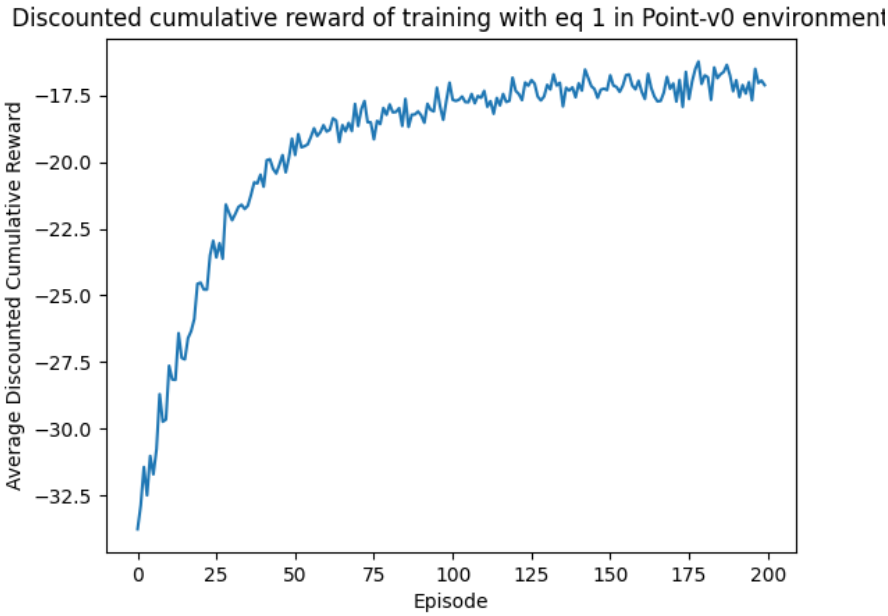

Figure 2: Point-v0 training discounted reward by equation 2

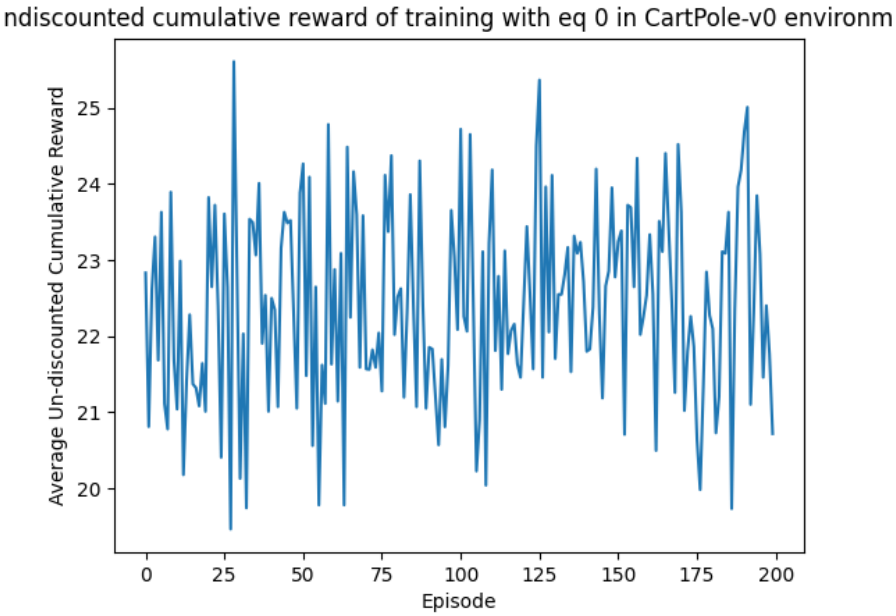

Figure 3: Training discounted reward by equation 1

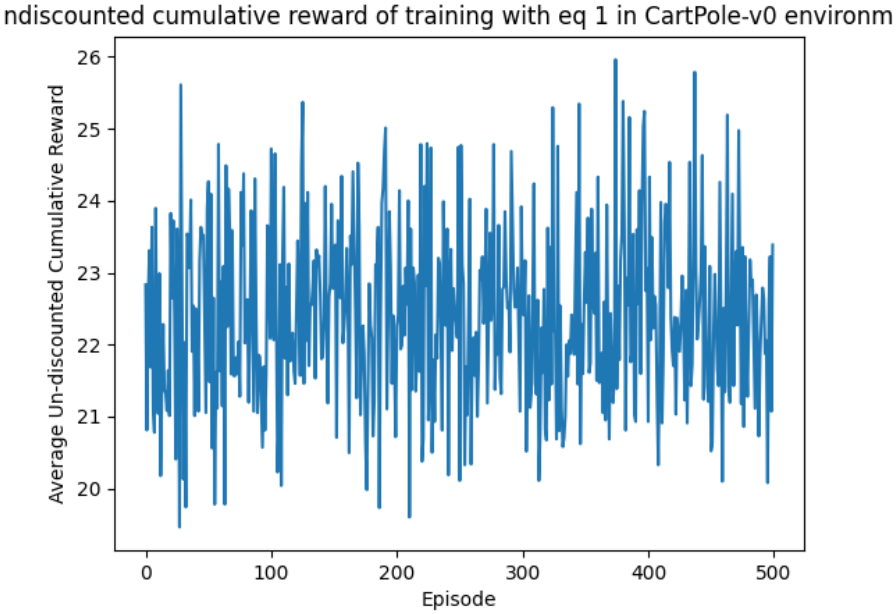

Figure 4: Training discounted reward by equation 2

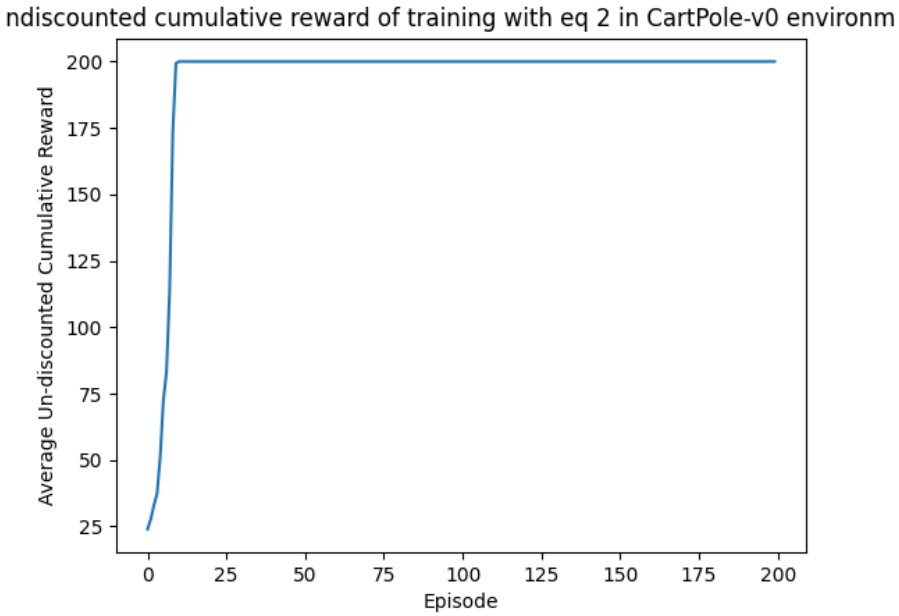

Figure 5: Training discounted reward by equation 3