Reinforcement Learning Homework 8

Deadline 10th December 2024
Students Sahiti Chebolu, Surabhi S Nath, Xin Sui

Question 1. Importance weighted policy gradient

We assume one step MDPs here.

- We want to maximise the importance weighted rewards: $J_\theta(\theta) = \mathbb{E}_{\pi_{\theta old}}\left[\frac{\pi_\theta(s,a)}{\pi_{\theta old}(s,a)} r\right]$

- We expand the expectation by summing over the state-action distribution induced by $\pi_{\theta^{old}}$:
$\nabla_\theta J_\theta(\theta) = \nabla_\theta \sum_s \mu(s) \sum_a \pi_{\theta^{old}}(s,a) \frac{\pi_\theta(s,a)}{\pi_{\theta old}(s,a)} r$

- Assuming that $\mu(s)$ doesn't change with π_θ, the only term that is dependant on θ is $\pi_\theta(a,s)$: $\nabla_\theta J_\theta(\theta) = \sum_s \mu(s) \sum_a \pi_{\theta^{old}}(s,a) \frac{\nabla_\theta \pi_\theta(s,a)}{\pi_{\theta old}(s,a)} r$

- This is nothing but: $\nabla_\theta J_\theta(\theta) = \mathbb{E}_{\pi_{\theta old}}\left[\frac{\nabla_\theta \pi_\theta(s,a)}{\pi_{\theta old}(s,a)} r\right]$

Question 2. PPO

(a)-(c)

Check code and plots in `PPO.py` and `Gym-PPO-plot.ipynb`. Training reward does increase through training for all values of ϵ except 0.01. For some (such as $\epsilon = 0.2$), training reward decreases after reaching a peak.

(d)

Results for $\epsilon = 0.5$ look promising. Here, the training rewards increased and stayed stable. Plots for re-runs with different seeds are in `Gym-PPO-plot.ipynb`. On re-runs, while rewards did increase through training, they were not as promising as the initial run.

(e)

To test, we used the weights stored in the best run with $\epsilon = 0.5$ (this was the very first run for us). We got the following test rewards in 10 test episodes:

Episode	Reward
1	66
2	77
3	126
4	69
5	120
6	31
7	93
8	86
9	112
10	31

The lander was able to touch the surface of the moon and stay there in most cases.