[CSCI-GA 3033-090]
Special Topics: Deep Reinforcement Learning
Homework - 3 Policy Gradient Algorithms

Xu Cao
xc2057@nyu.edu

25 oktober 2021

Question 1

Finished, please see Fig. 1, Fig. 2 and Fig. 3. I use seed 3, 5, 9 to run the code.

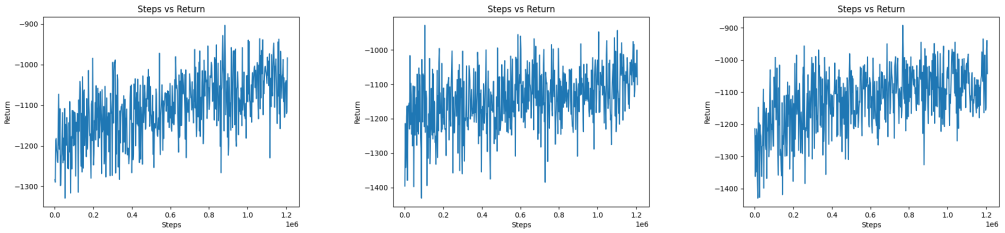
REINFORCE performs not good in these environments. For LunarLanderContinuous-v2, it shows an increasing trend, but it finally did not achieve the request threshold. For Pendulum-v0 and ipedalWalker-v3, its return change very very slow during training.

To explain this phenomenon, we should analyze the pseudo code of REINFORCE. REINFORCE has large variance and unstable training process. The REINFORCE computes the return R_t by $R_t = \sum_{t'=t}^{T-1} \gamma r_{t'}$, which is also related to the step size for parameter update. However, if the step is too far, we will continually stay in a bad policy region. To solve this problem, a good choice is using Actor-Critic.

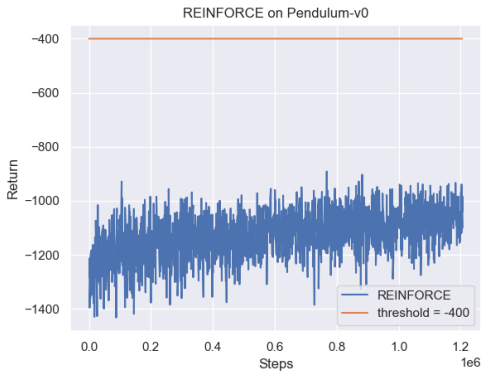
Question 2

Finished, please see Fig. 4, Fig. 5 and Fig. 6. I use seed 3, 5, 9 to run the code. The comparison is also shown in the figure. The figures include the homework expected mean performance (Pendulum: -400, BipedalWalker: 125, LunarLanderContinuous: 100).

The results show that PPO is better than REINFORCE on these three environments, and PPO also achieves requested expected mean performance on all three environments.

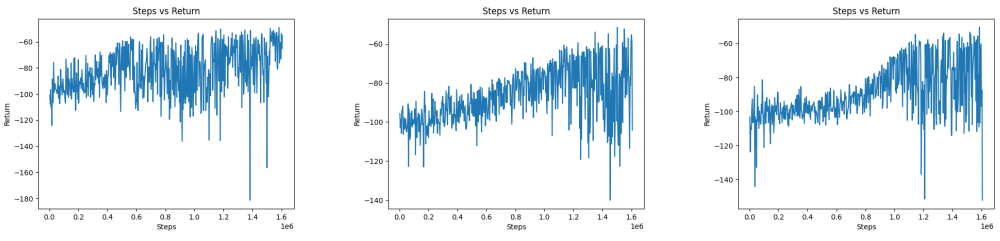


(a) (b) (c)

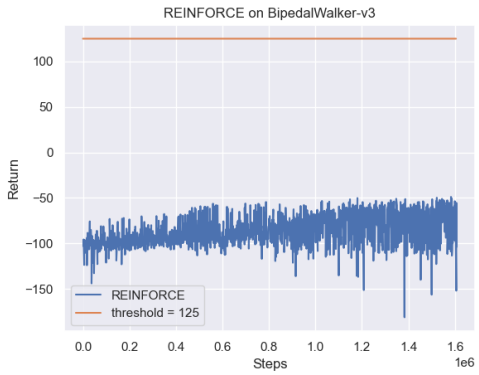


(d)

Figur 1: REINFORCE on Pendulum-v0, from (a) to (c) is seed 3, 5, 9; (d) is the average performance.

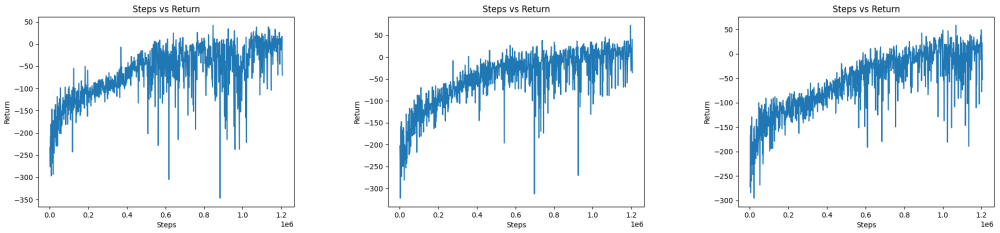


(a) (b) (c)

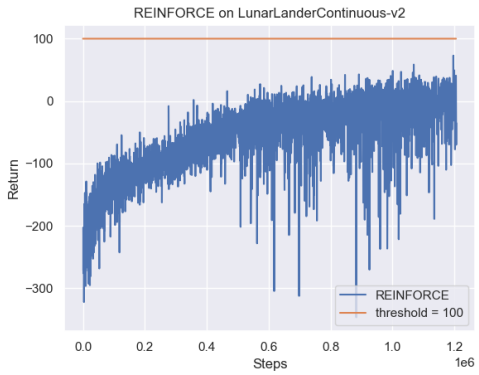


(d)

Figur 2: REINFORCE on BipedalWalker-v3, from (a) to (c) is seed 3, 5, 9; (d) is the average performance.

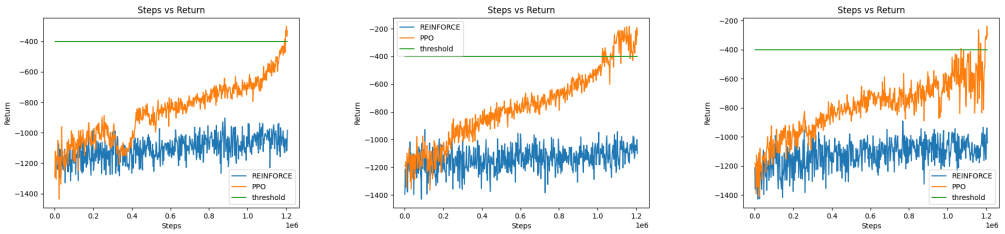


(a) (b) (c)

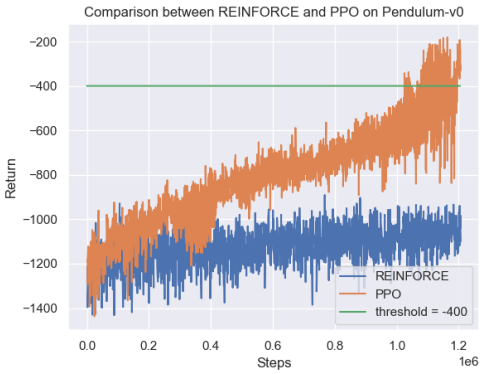


(d)

Figur 3: REINFORCE on LunarLanderContinuous-v2, from (a) to (c) is seed 3, 5, 9; (d) is the average performance.

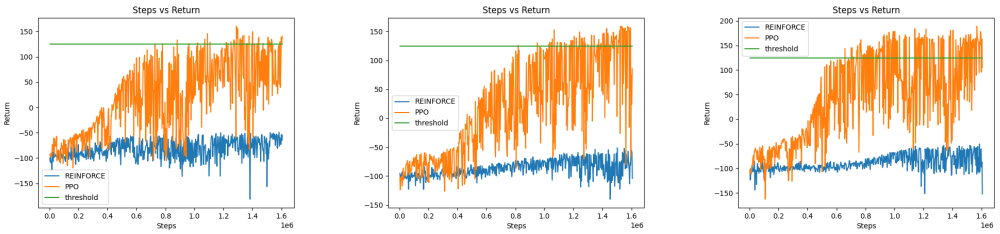


(a) (b) (c)

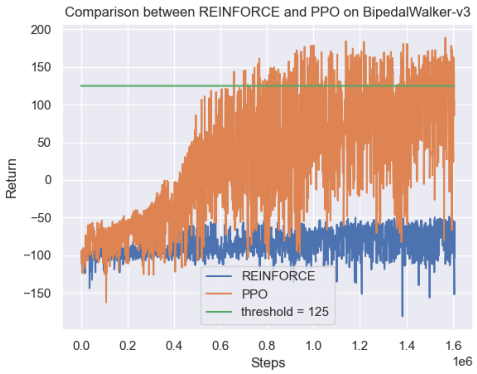


(d)

Figur 4: Comparison of REINFORCE and PPO on Pendulum-v0, from (a) to (c) is seed 3, 5, 9; (d) is the average performance. The orange line is PPO. The blue line is REINFORCE. The green line is the threshold.

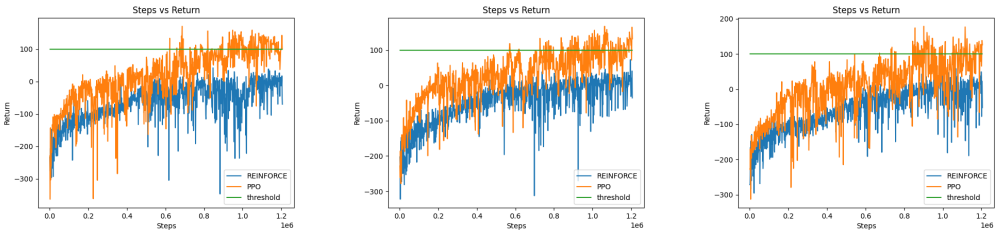


(a) (b) (c)

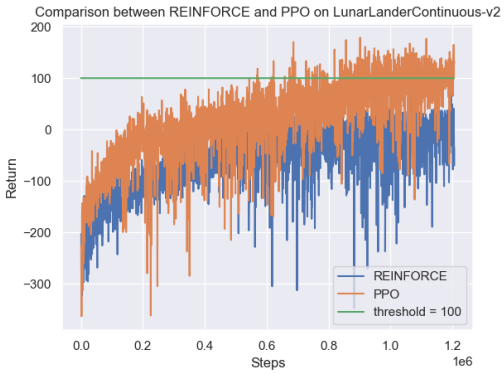


(d)

Figur 5: Comparison of REINFORCE and PPO on BipedalWalker-v3, from (a) to (c) is seed 3, 5, 9; (d) is the average performance. The orange line is PPO. The blue line is REINFORCE. The green line is the threshold.



(a) (b) (c)



(d)

Figur 6: Comparison of REINFORCE and PPO on LunarLanderContinuous-v2, from (a) to (c) is seed 3, 5, 9; (d) is the average performance. The orange line is PPO. The blue line is REINFORCE. The green line is the threshold.