ESE 650, SPRING 2024

HOMEWORK 4

YUNPENG WANG [YUNPENGW@SEAS],

Solution 1 (Time spent: 1 hour)**.**

The links to the recorded videos are shown below:

(1) Cartpole: https://youtu.be/WMVQ041gj-g
(2) Acrobot: https://youtu.be/HQYywdKw8MY
(3) Walker: https://youtu.be/0TDStJpuSfQ
(4) Humanoid Stand: https://youtu.be/iiiLDbqobk4

Solution 2 (Time spent: 20 hours)**.**

Following the classical way of implementing PPO, I'm able to build a model to control the humanoid robot to "walk". The video of the walker in the simulation is https://youtu.be/LuSwg0o5Alw.

0.1. **Mini-batch update.** Mini-batch is a good way to balance optimization and memory consumption. After collecting some trajectories consisting of many steps, I randomly select some steps, use them to compute the loss, and update the network.

0.2. **Entropy Regularization.** One problem I saw in the training is that the reward drops in the middle of training as shown in figure 1. I believe that the entropy regularization causes it.

```
entropy_loss = entropy.mean()
actor_loss = actor_loss - self.ent_coef * entropy_loss
```

LISTING 1. Entropy Regularization

0.3. **Reward vs Timestep.** After setting the entropy coefficient to zero, the reward increases monotonically that can be seen in figure 2.

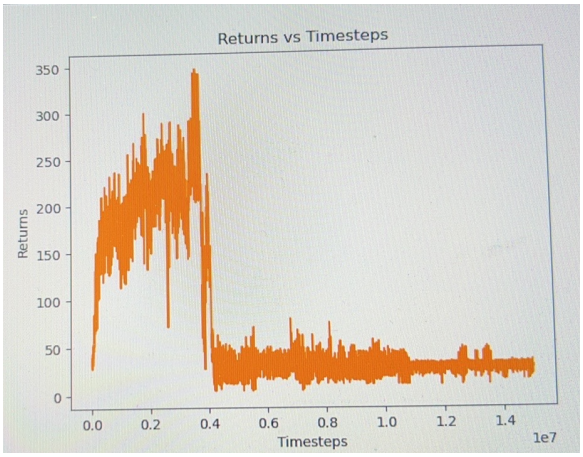

FIGURE 1. Reward drops in the middle of training

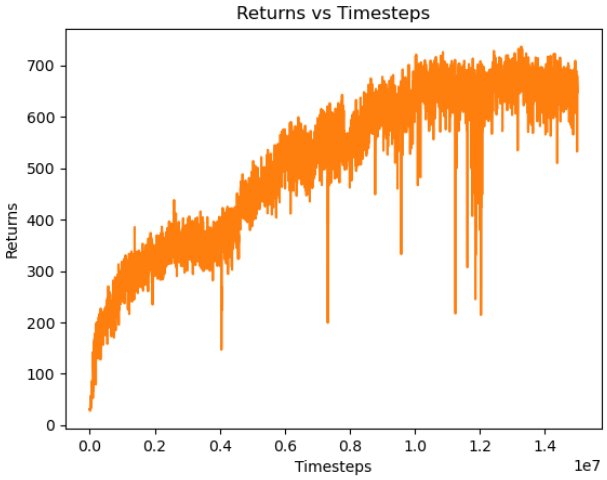

FIGURE 2. Return vs Timestep