Exercise 1

Xuebing Li - 678500
ELEC-E8125 - Reinforcement Learning

September 9, 2022

1 Task 1

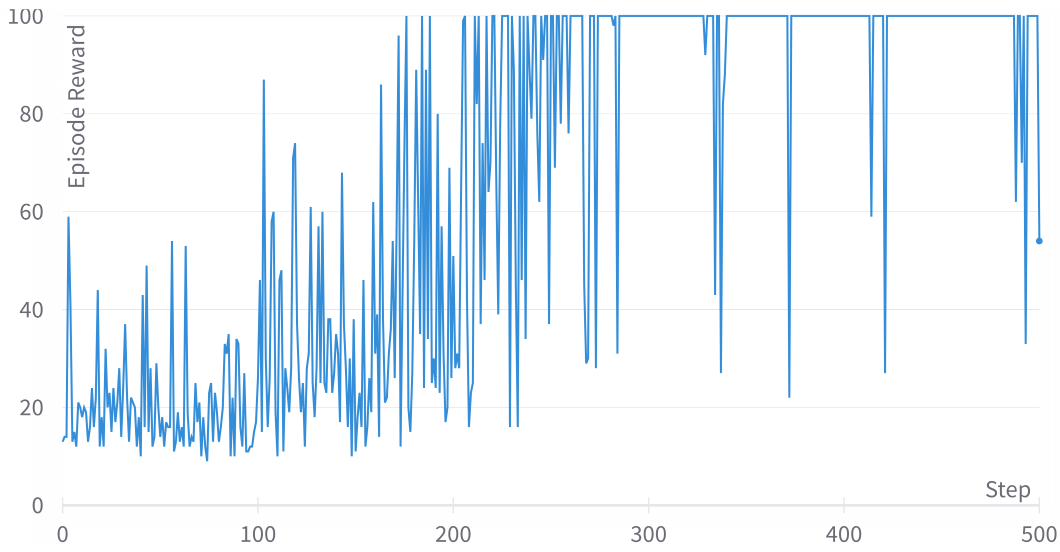

Figure 1: Episode reward during training.

I have trained a model with 100 timestamps. The reward of each episode during the training phase is illustrated in Fig. 1. When testing the model with 1000 timestamps, the average reward is 927.8.

1.1 Question 1.1

The model trained over 100 timestamps does not work for a 1000-timestamp evaluation scenario all the time. The root cause is that the trained model is not robust enough for the CartPole problem. There may be some reasons that may causes the performance issue of the trained model. E.g, the training timestamps is too small, the network is not deep enough, the DRL algorithm in use is not good enough, etc. I test the trained model under a 100-timestamp evaluation scenario for 5 times and the reward is always 100. So, I believe

Experiment No.	Seed	Averaged Reward
#1	1	927.8
#2	2	1000
#3	3	287.1
#4	4	378.8
#5	5	574.6

Table 1: The average test reward over repeated experiments.

that the insufficient training timestamps is an important cause of the performance issue of the trained model.

2 Task 2

I have trained a model under an 100-timestamp scenario and have evaluated the model under a 1000-timestamp scenario for 5 times. The average test reward for each is shown in Table 1.

2.1 Question 2.1

The performance of the trained model is not the same every time. The key cause is the use of seed. Firstly, random variables are not actually random in gym. Whenever the seed is defined, the sequence of random variables generated by the seed is also defined. Secondly, the randomness of the CartPole environment relies purely on top of the randomness of the random variables. It means that a fixed value of seed makes CartPole behave always the same. Thirdly, the training program does not have any stochastic behavior at the time of training.

So, putting all of them together. For two independent runs, if the seed is the same, the CartPole environment, the training result, and the evaluation result are all the same. Otherwise, both training and evaluation results will be different.

2.2 Question 2.2

When evaluating a single model, we expect the environment to be as random as possible to prove that the model is robust. However, when comparing two models, we need to ensure that the environment in use is the same for both models to be fair.

3 Task 3

3.1 Behavior 1

```python
def get_reward(self, prev_state, action, next_state):
    return 1 if action == 1 else 0
```

The training reward is illustrated in Fig. 2

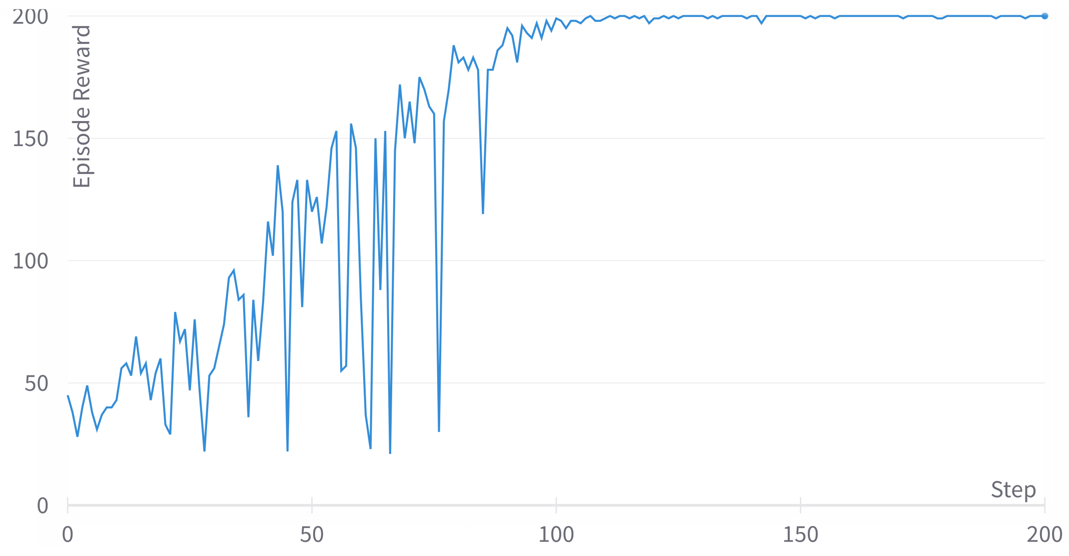

Figure 2: Training reward in behavior 1, task 3

3.2 Behavior 2

```python
def get_reward(self, prev_state, action, next_state):
    return 1 if self.get_terminal_state() else 0
```

The training reward is illustrated in Fig. 3

4 Task 4

The two plots generate by plot_rew.ipynb are shown in Fig. 4 and Fig. 5, respectively.

4.1 Question 4.1

The highest rewards are achieved when the end point of the arm is close to the target (1, 1), i.e., within the distance of 0.25. The corresponding states are marked as white in Fig. 4.

4.2 Question 4.2

No, the policy has not learnt to reach the goal in an optimal way. From Fig. 5, it shows that the model only takes two actions: J1+ and J2+. It is obvious that, in some states, taking J1- or J2- will achieve the goal faster. More specifically, the training program did not motivate the model to solve the problem with less steps.

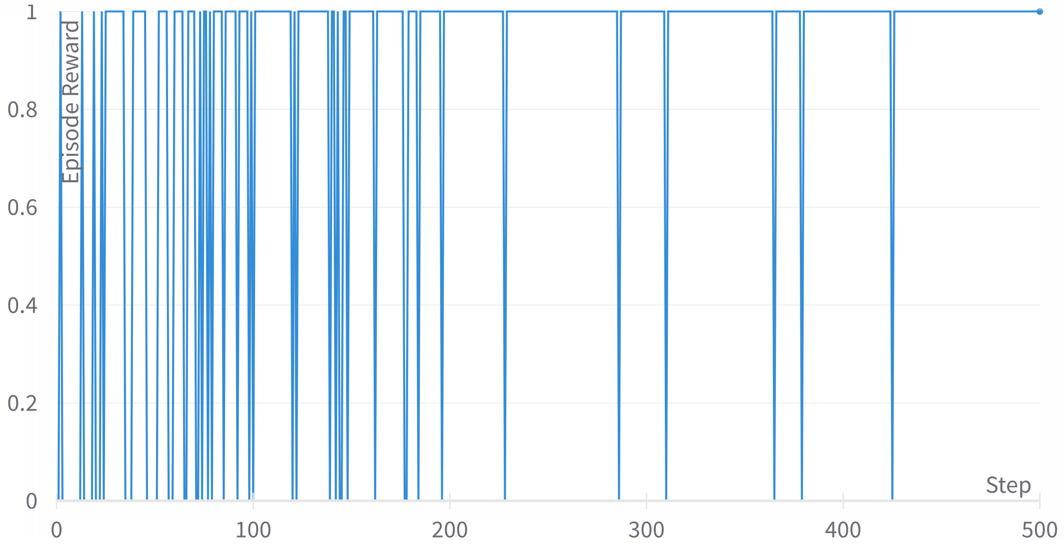

Figure 3: Training reward in behavior 2, task 3

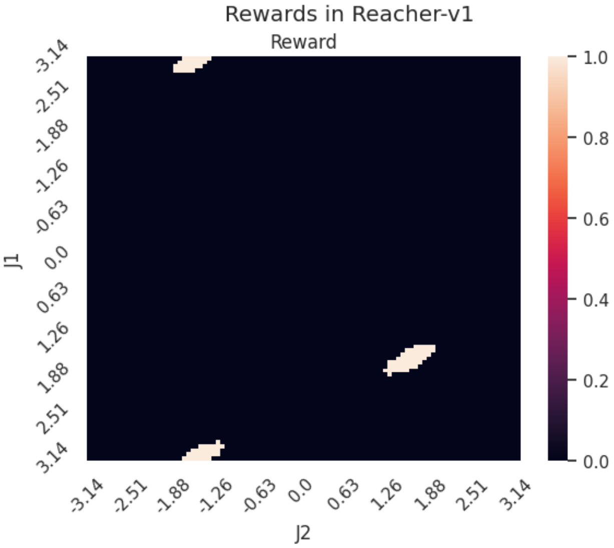

Figure 4: The reward w.r.t. the states

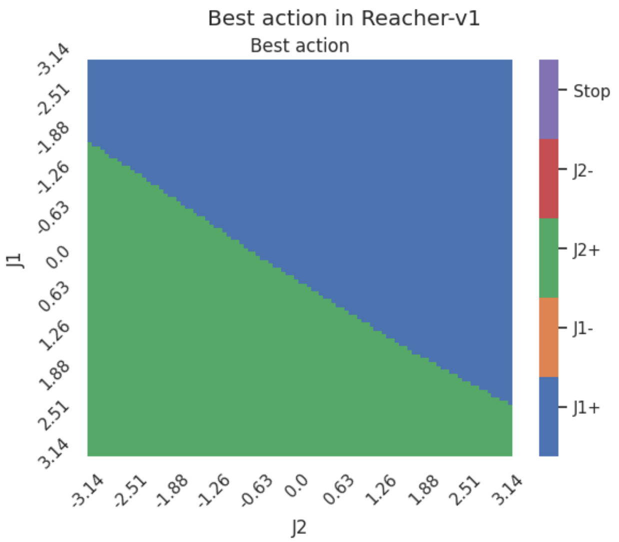

Figure 5: The best action w.r.t. the states

5 Feedback

How much time did you spend? Approximate 2 hours to solve the questions. But it takes much more time to setup everything. I understand that latex is good for academic writing. But I did not see any beneficial of using latex when writing assignment reports, especially on short reports. It will take less time if I can use Microsoft Word. The same question applies to the use of wandb. Using pyplot to export diagrams shall be much easier than downloading from the website. Wandb may be very useful in the future, but I cannot see its importance in this assignment.
Any hard task? Nope.

References