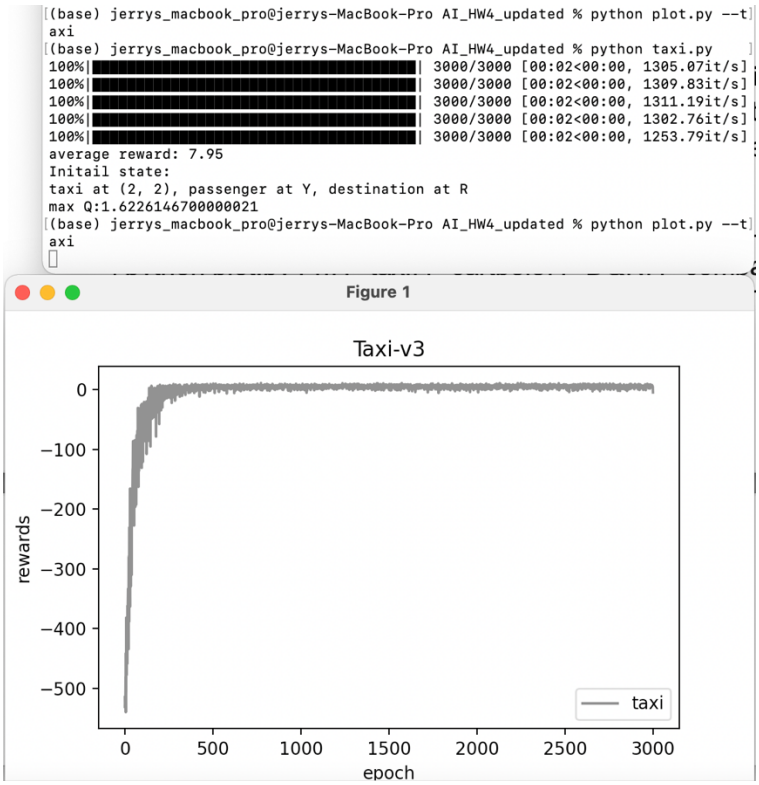
Homework 4: Reinforcement Learning Report

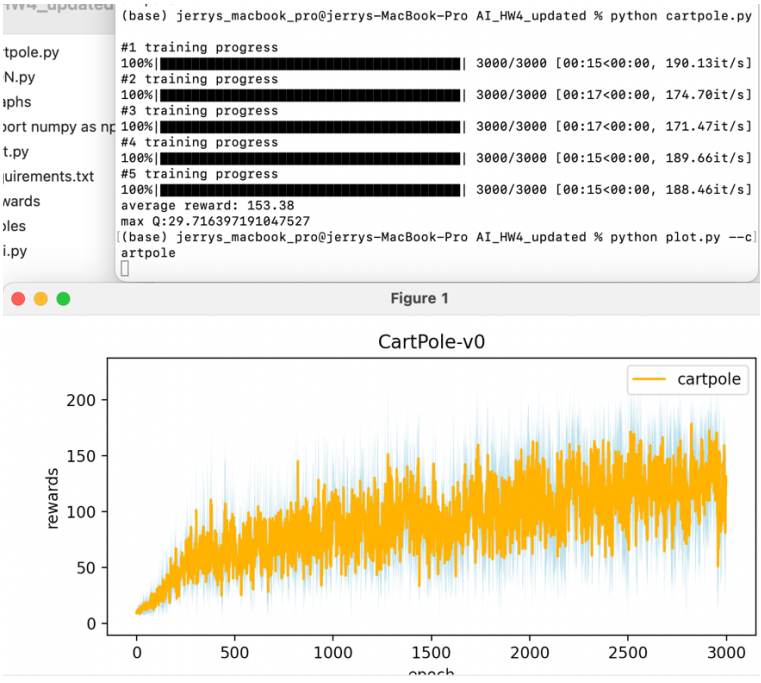
109550116 楊傑宇

Part I. Experiment Results :

Taxi:

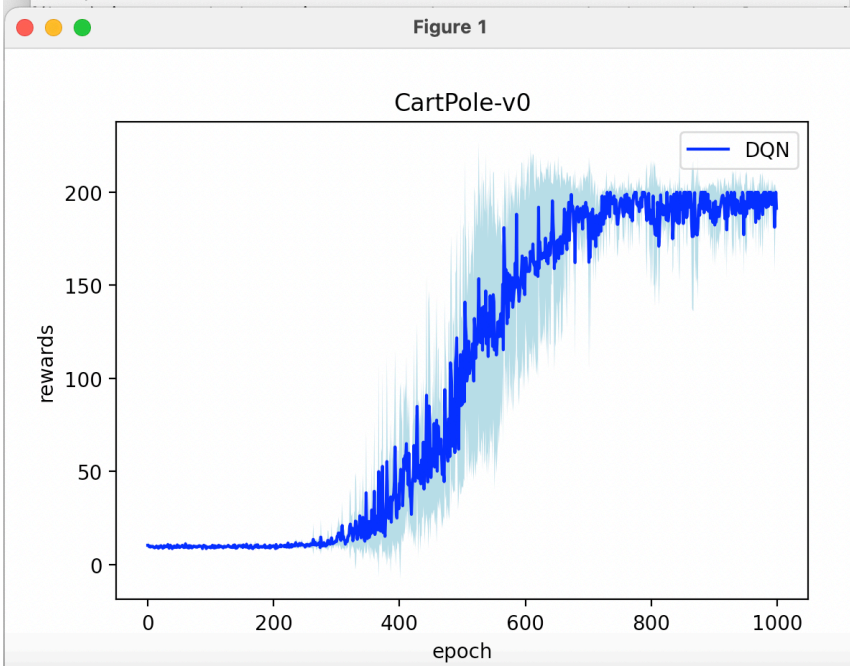


Cartpole:

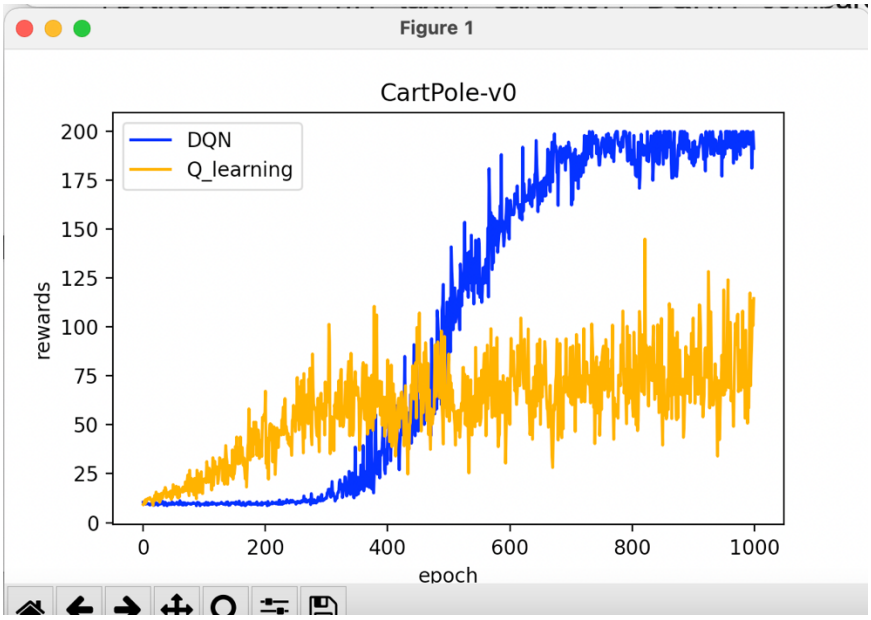


DQN:

```
umpy.array() before converting to a tensor. (Triggered internally at  /Users/dis
tiller/project/pytorch/torch/csrc/utils/tensor_new.cpp:210.)
  q_eval = self.evaluate_net(torch.FloatTensor(b_state)).gather(1, action)
100%|████████████████████████████████████| 1000/1000 [02:45<00:00,  6.04it/s]
#2 training progress
100%|████████████████████████████████████| 1000/1000 [02:58<00:00,  5.62it/s]
#3 training progress
100%|████████████████████████████████████| 1000/1000 [02:52<00:00,  5.81it/s]
#4 training progress
100%|████████████████████████████████████| 1000/1000 [02:58<00:00,  5.59it/s]
#5 training progress
100%|████████████████████████████████████| 1000/1000 [02:09<00:00,  7.71it/s]
reward: 193.5
max Q:34.068077087402344
```

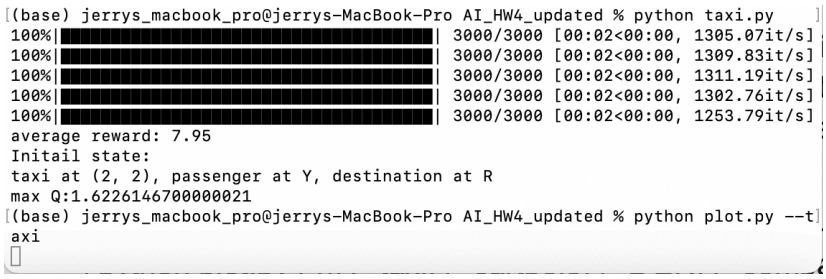


Compare:



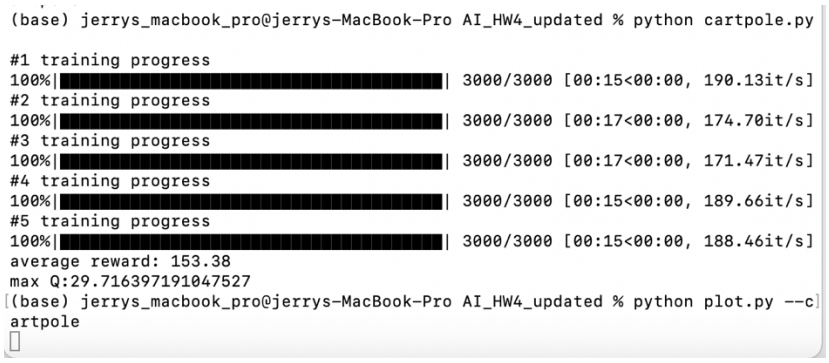
Part II. Question Answering (50%):

1. Calculate the optimal Q-value of a given state in Taxi-v3 (the state is assigned in google sheet), and compare with the Q-value you learned (Please screenshot the result of the "check_max_Q" function to show the Q-value you learned). (4%)



My optimal Q-value is `-0.5856821172999993`

2. Calculate the max Q-value of the initial state in CartPole-v0, and compare with the Q-value you learned. (Please screenshot the result of the "check_max_Q" function to show the Q-value you learned) (4%)



My optimal Q-value is `4.816074831364288`

3.
   a. Why do we need to discretize the observation in Part 2? (2%)
      Because we get a continuous interval, which is not easy to judge cartpole position.
   b. How do you expect the performance will be if we increase "num_bins"? (2%)
      I think performance will be better because it can calculate more accurately.
   c. Is there any concern if we increase "num_bins"? (2%)
      It may need more time to run the code and more memory to save qtable.

4. Which model (DQN, discretized Q learning) performs better in Cartpole-v0, and what are the reasons? (3%)
   Because DQN use memory buffer save the record until 1000 times, then it starts learning.

5.
    a. What is the purpose of using the epsilon greedy algorithm while choosing an action? (2%)
       The epsilon-greedy approach selects the action with the highest estimated reward most of the time
    b. What will happen, if we don't use the epsilon greedy algorithm in the CartPole-v0 environment? (3%)
       We can't balance between exploration and exploitation. it can't perform truly situation.
    c. Is it possible to achieve the same performance without the epsilon greedy algorithm in the CartPole-v0 environment? Why or Why not? (3%)
       I think it is possible if we have another method to determine the factor between exploration and exploitation.
    d. Why don't we need the epsilon greedy algorithm during the testing section? (2%)
       Because when we are in testing section, qtable have been done, we just get the max q-value between each action.
6. Why is there "with torch.no_grad():" in the "choose_action" function in DQN? (3%)
   It lets tensor with gradient currently attached with the current computational graph is now detached from the current graph
7.
    a. Is it necessary to have two networks when implementing DQN? (1%)
       No.
    b. What are the advantages of having two networks? (3%)
       It helps keep runaway bias from bootstrapping from dominating the system numerically, causing the estimated Q values to diverge.
    c. What are the disadvantages? (2%)
       training two networks separately will double the convergence time.
8.
    a. What is a replay buffer(memory)? Is it necessary to implement a replay buffer? What are the advantages of implementing a replay buffer? (5%)
       - Replay buffers is to store trajectories of experience when executing a policy in an environment.
       - Yes
       - More efficient use of previous experience, by learning with it multiple times.

b. Why do we need batch size? (3%)
   Batch size controls the accuracy of the estimate of the error gradient when training neural networks.
c. Is there any effect if we adjust the size of the replay buffer(memory) or batch size? Please list some advantages and disadvantages. (2%)
   The more batch size is, the more training examples used in the estimate and the more accurate this estimate will be. However it may slow our code.

9.

a. What is the condition that you save your neural network? (1%)
   I don't use condition to save, however I think I can save every 100 times
b. What are the reasons? (2%)
   It can accelerate my code.

10. What have you learned in the homework? (2%)
    In taxi, I learned the epsilon greedy algorithm and how to calculate q-value. In cartpole, I learned how to discretize. In DQN, Beside the operation of DQN, I learned how to convert dimension between variables more.