Hanzhao Deng
UIN:661663177
NetID:hdeng7

CS440 MP1 Part 2

1.

a) The agent does get better over time, but after 1000 episodes the total reward is still not ever big. Its because the agent spends too much time exploiting what he has learned and never explore.

b) The agent gets better over time and the total reward after 1000 episodes is noticeably higher than the agent with a epsilon of zero. There is a balance between exploring and exploiting, resulting in good policy.

c) After 1000 episodes most of the time the total reward is still negative. The agent explores too much and behave too randomly since have of the time the agent just chooses a random direction to go.

2

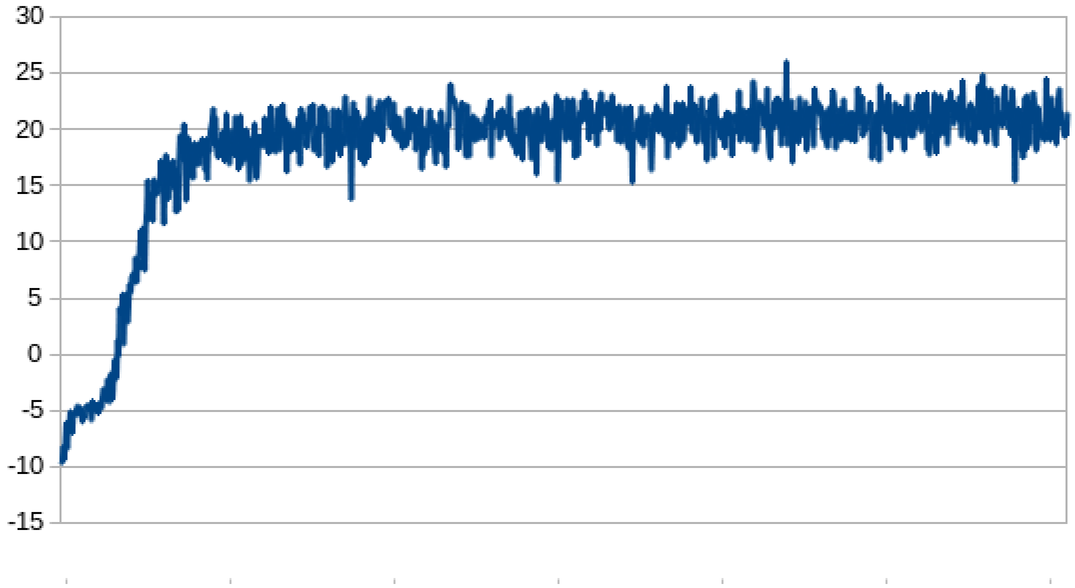
a) My agent get stuck in the right hand side corner. Because he doesn't know the position of the thief the best policy he has is to not decrease the reward by going blindly into the thief column so the agent just get stuck in the corner.

b) My agent can successfully deliver packages and collect rewards because through Q-learning he has learned how to react to the theifs in different locations.

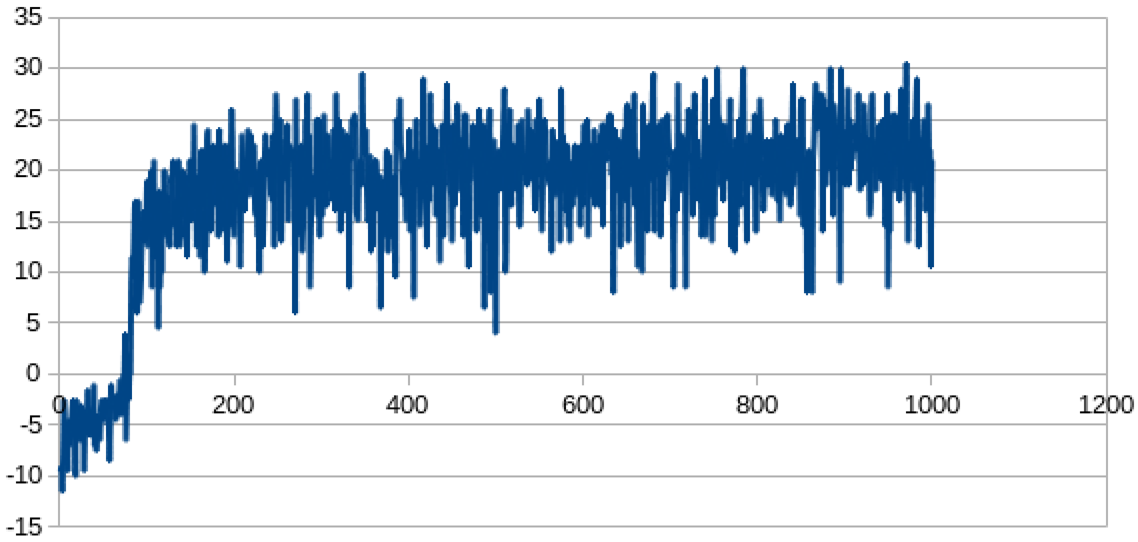
c) I modified the simulator so that it takes two more arguments, epsilon and learning rate for the agent. I then wrote a bash script to run the program 900 times, changing the epsilon from 0.00 to 0.99 and changing the learning rate from 0.1 to 0.9. I also modified the episodes output of the simulator so that they write the total number of rewards they get in the 1000 episodes in a file with the filename of its epsilon and learning rate. After that I wrote a simple script to find the file that has the contains the largest number.The best epsilon and learning rate i found was epsilon = 0.10 and learning rate = 0.2.

3.

This is the graph generated by the average of 10 simulation with epsilon = 0.10 and learning rate = 0.2:



This is the graph generated by the average of one simulation with epsilon = 0.10 and learning rate = 0.2:



The one with average value of 10 simulations have much more smooth line and more stable.