CS440 MP1 Part 2

Zhuojun Yao zyao10

1.(a) In my observation, the reward of all episodes are equal to 0. Because if the agent does not perform random actions, he won't step into the slipperiness because the reward of doing so is negative, which is lower than the utility of not stepping into it (reward = 0). He will choose to just move around in grid without slipperiness in all steps. Therefore, the total reward will be unchanged and equal to zero.

(b)When epsilon is equal to 0.1, the reward will be positive, and the simulator shows that the average reward will be between 120 and 140. This is because right now the epsilon of QLearningAgent is no longer equal to 0 but have a small value. Therefore, the agent will mostly perform greedily but have a small probability to act randomly, so it may step into the slipperiness. After stepping into it, it may not drop the package and will find that continue to deliver the goods leads to a better reward and therefore doing so greedily, based on the WorldWithThief. Consequently, the total rewards of each episode will be positive.

(c)When epsilon is larger and equal to 0.5, the agent will act very irrationally. Therefore, the average reward of 1000 episode is approximately -28. It's because even though high epsilon will cause the agent to step into the slipperiness, it will also lead it to somewhere unrelated to the goal state. For example, it may step into the slipperiness while standing in another slipperiness. It will raise the possibility that the package being dropped. Thus, even though some of the episode may have a positive reward, most of it will show negative results.

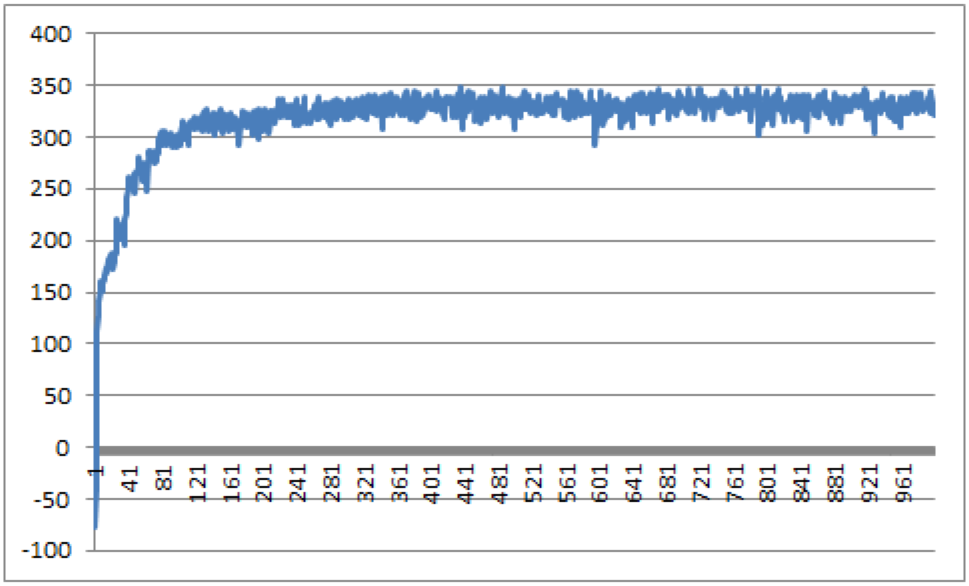
2.(a) When the agent doesn't know the position of the thief, the reward after 1000 episode is approximately -12. First of all, because the epsilon is very small, the agent will mostly act greedily. However, since the agent doesn't know the exact position of the thief, its actio won't base on the current position of thief and whether or not it will be caught by the thief. Thus, it will have a high possibility to be caught by the thief and get negative reward. As a result, the final rewards of the agent will be negative.

(b) When the agent realizes the position of the thief, the reward is approximately 270. It's because when the agent know the position of the thief. It can be smart enough to choose the next step to avoid caught by the thief. Thus, it's very likely for it to get a high positive reward

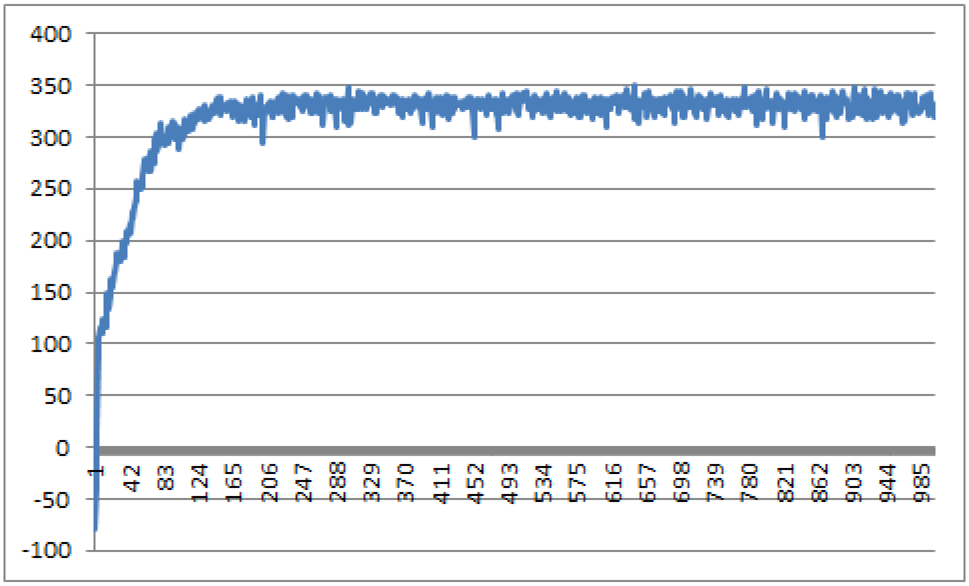
(c) When the learning rate is 0.15 and the epsilon is 0.01, the agent receives a highest reward. I use simulator to print out the average reward of each procedure. Then, I change the learning rate and epsilon independently in the change of 0.01 at the QLearningAgent to get the highest average award. Since epsilon should not be too large due to past experience and it cannot be zero, I try it from a very small amount, like 0.01, and increase it to see whether there is a increase of rewards. Finally it shows up that with everything else unchanged, reward gets its largest value when epsilon equals to 0.01. Same story about learning rate. Since the problem is stochastic, the learning rate should be low. Then I start testing it from 0.4, keep decreasing, and finally find the result of 0.15.

3. The starting episode rewards of the agent is negative, and it keeps raising at the starting episodes. This is because the agent has studied from the past episode. After about 200 episodes, the reward of each episode become stable and converge to a horizontal line.

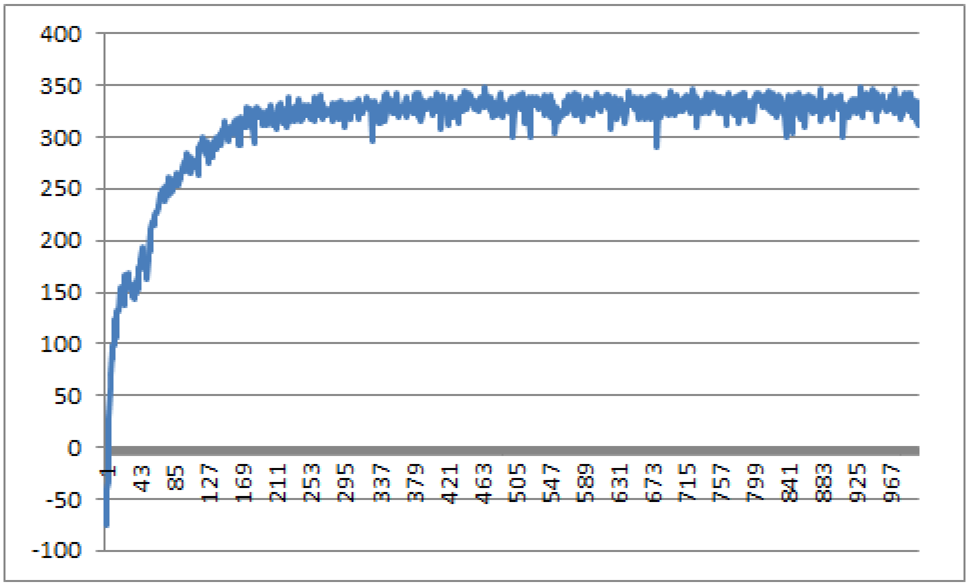
After testing ten times of the agent, it got the average reward of 316.7 per episode. (Totally 10 * 1000 episodes).This average reward is approximately the same as the convergent lines.



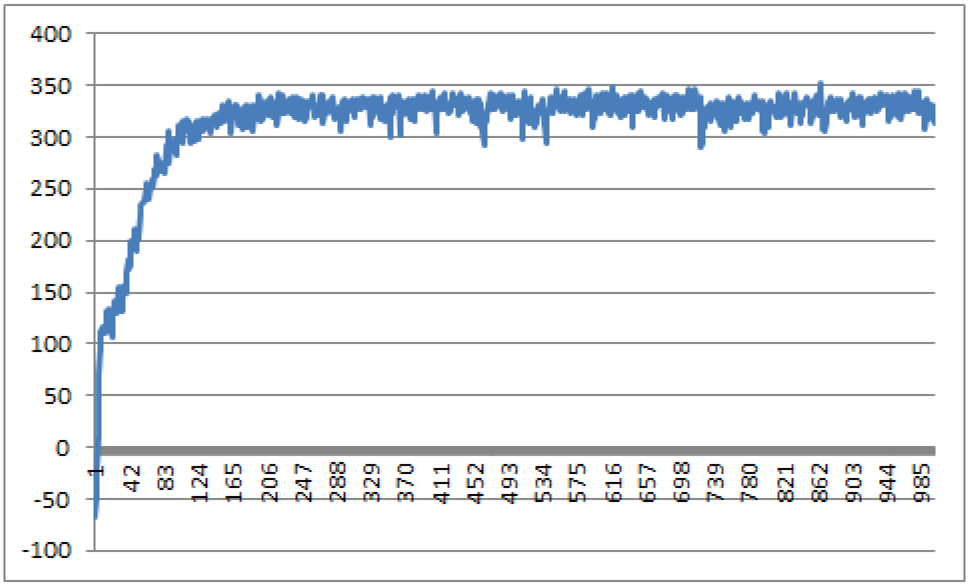
First stimulation



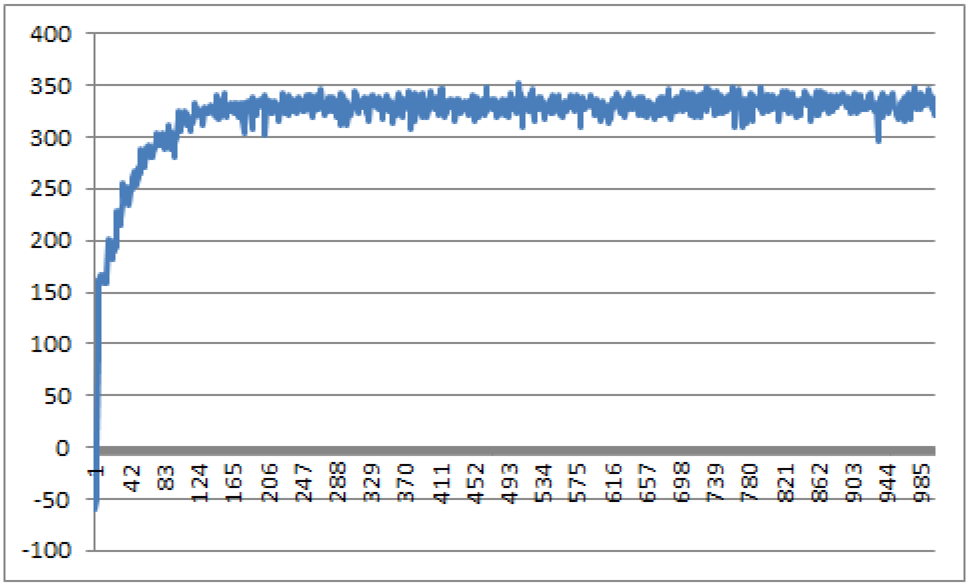
Second stimulation



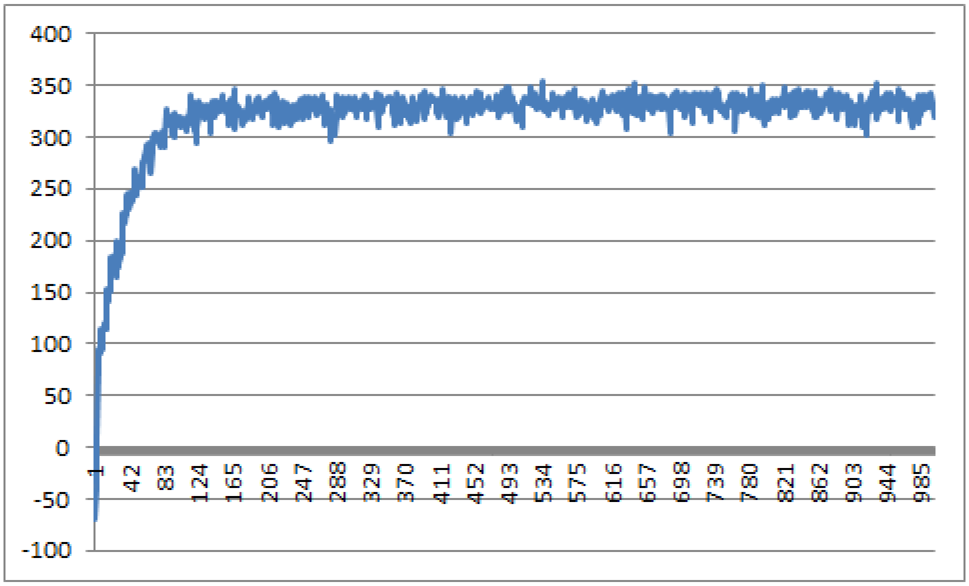
Third stimulation



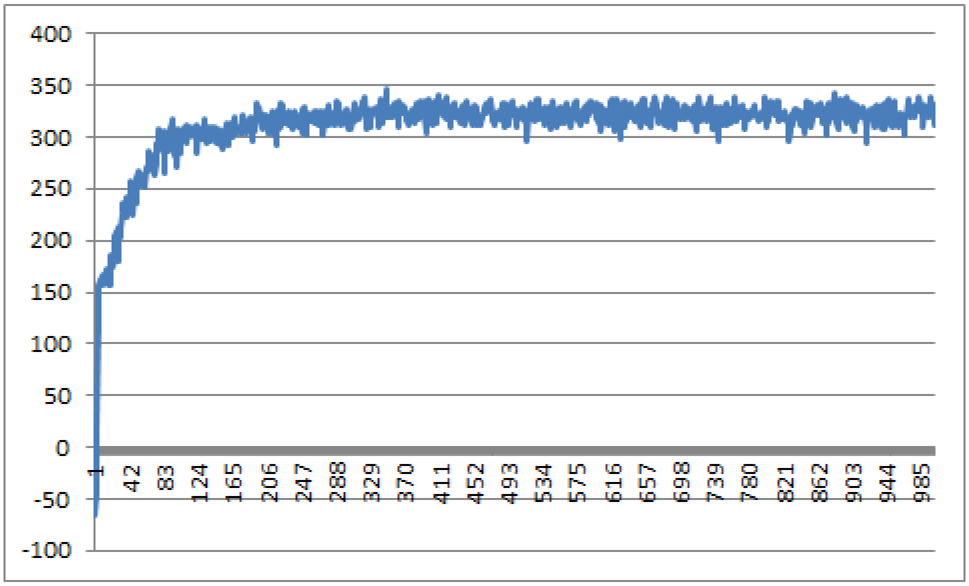
Four stimulation



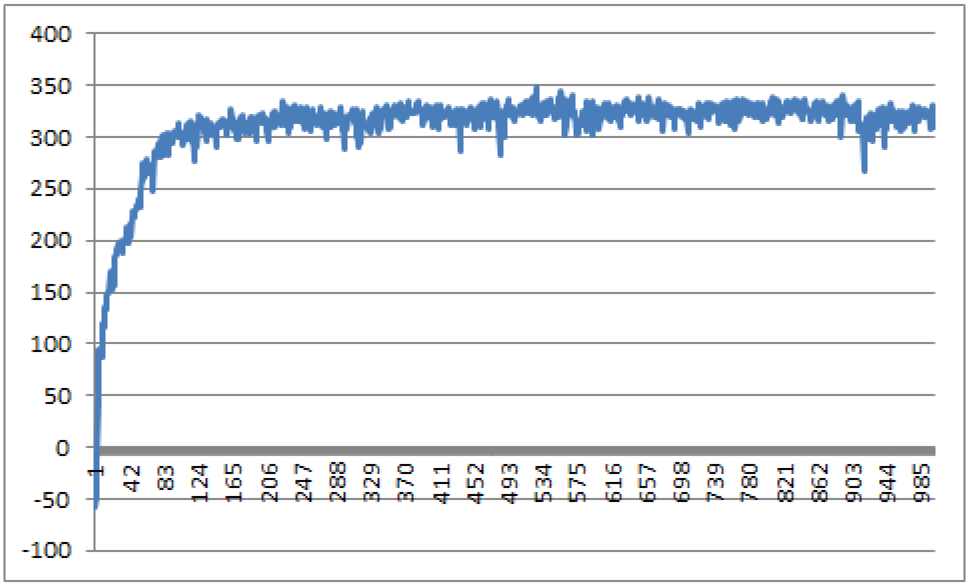
Fifth stimulation



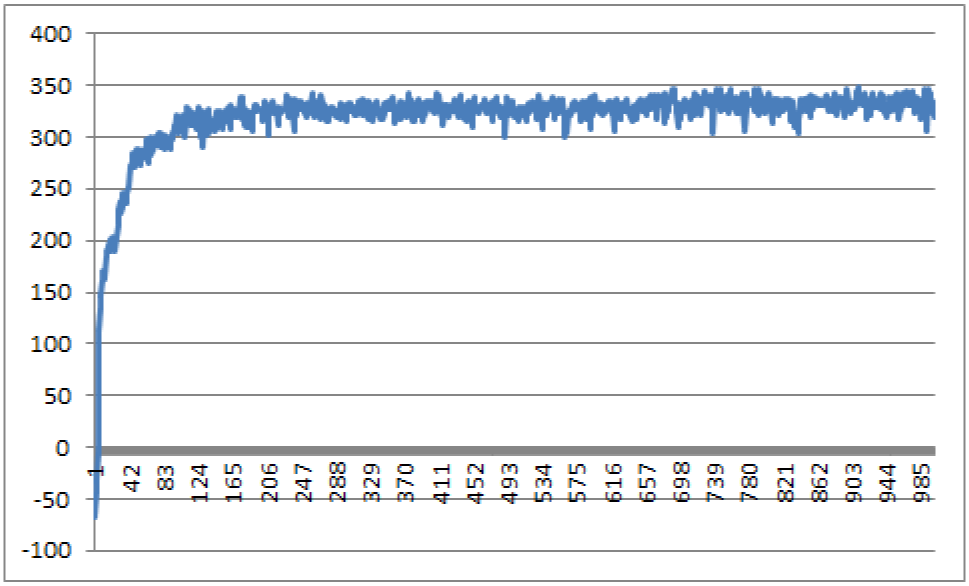
Sixth stimulation



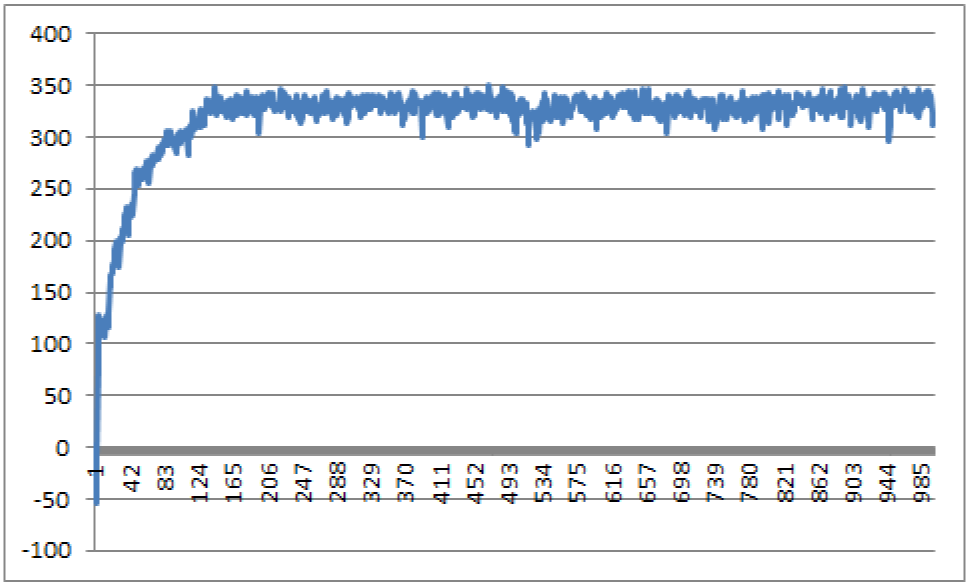
Seventh stimulation



Eighth stimulation



Ninth stimulation



Tenth stimulation

Ten rounds' average rewards and total average rewards

318.8	312.3	314.1	319.6	321.5	320.0	311.5	310.2	319.2	319.1	**316.6**
99	585	875	405	765	76	46	675	765	425	**971**