CS 440 MP1 Report

In world without thief

(a) The result for 1000 episodes contains a lot of “reward=0.0”. Since epsilon equals to 0.0. Thus every time when the agent wants to choose an action, the next action must have greatest expected reward. At some point the agent reaches at the corner surrounding by traps, thus next movement should be out of bounder, making agent stay at the same place. Since the expected reward won’t be changed during this process. The agent will stay at the corner forever. Therefore, it gains 0.0 reward even it takes 10000 steps.

(b) At the end of the whole test, the rewards stay positive and some of them reach at the greatest reward points. The reason that it is different from original setting is that the positive epsilon gives agent opportunity to randomly pick a action, thus it can jump out of the corner. Therefore, it can take more possibilities.

(c) The reward points are random. Most of them are negative. The reason is that every state has 50% chance to randomly choose next action. Thus it won’t obey the greatest expected reward policy at a great chance. Then the whole system won’t be smart at all. It won’t learn lessons from previous experiment.

In world with Thief

(a) What I observed is that most of the rewards is negative. I guess the reason is that the agent is not sensitive to thief. That is every time the agent to choose next action, it won’t take thief penalty to consideration. Thus it will have a great chance to lose package, thus receive penalty.

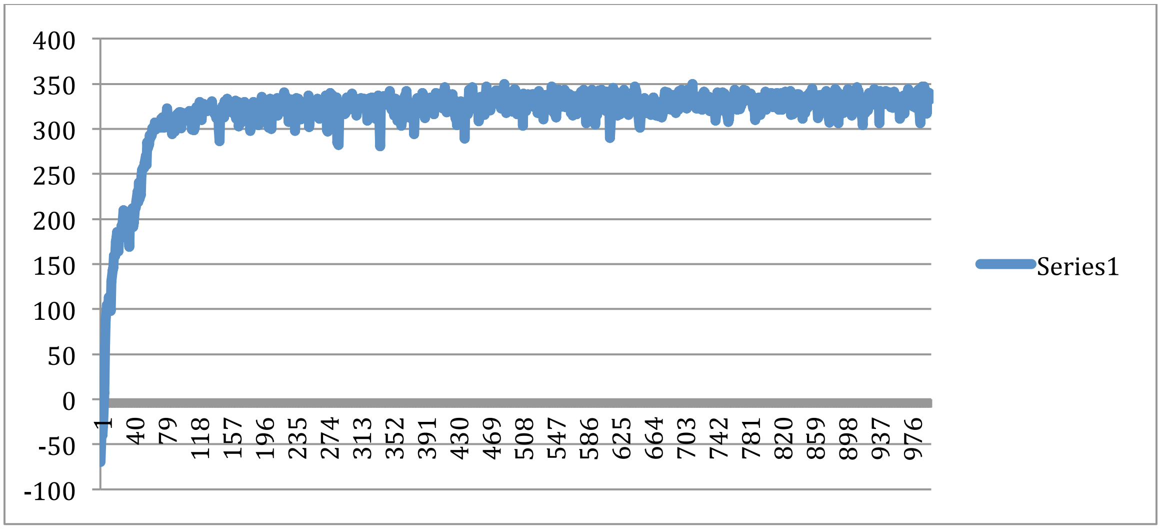
(b) The rewards are all positive and pretty high now. Since we consider thief now during action decision, thus the agent will take history record as reference and avoid thief. Thus it has a great opportunity to get higher reward.

(c) The optimal value for learning rate is 0.2 and epsilon is 0.01. I tested several combination of learning rate and epsilon. And calculate last 200 episodes’ average reward. And get following table:

	0	0.2	0.4	0.6	0.8	1
0.2	87	91	-106	-230	-238	-102
0.4	70	78	-112	-234	-239	-103
0.6	93	54	-128	-238	-239	-102
0.8	81	16	-147	-249	-244	-104

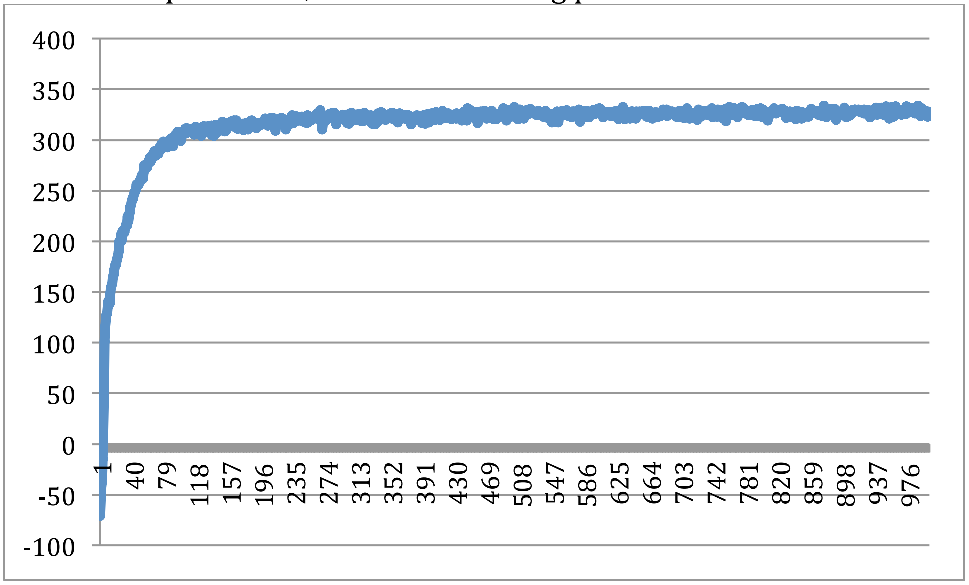
Thus we can see that the optimal choice may aroud learning rate = 0.2 and epsilon = 0.2. Then I test more around this setting. It turns out that 0.2/0.01 is the best one.

Optimal test



This graph depict the 1000 episodes' expected rewards. Notice that after 100 episodes, the expected reward are stale around 300-350.

After nine more experiments, we have following plot.



This graph shows the same pattern, but the plot is more smooth. That is the training for this agent can bring larger rewards for sure.