Name: Jiatong Shen UID: 668298212 Date: 29/09/14

Problem 1

(1) When $\epsilon = 0.0$, what I observe is that total reward is 0.0 and robot tends to stay in some certain place, for example in may in the left up corner of the network.The reason is when $\epsilon = 0.0$, the robot carries out a greedy plan. It will perform a random walk only to break the ties in the Q value table. As a result, when there is a minus Q value in the table, the robot will never try to go across that squares. So the robot will tend to stay in the left two columns to avoid reducing its Q value.

(2) When $\epsilon = 0.1$, at the beginning the reward of each episode is changing around 0, since it does not have an image of the world and needs to develop the world which will lead to some possible errors. After the robot has got enough knowledge, the reward increases dramatically. Then the maximum value is oscillating around the average, because there are unexpected cases happening on each episode.
The result is better than $\epsilon = 0.0$, because the agent is more adventurous and will try to go across the squares whose slipperiness is not 0.

(3) When *epsilon* = 0.5, the result is worse than the first two choices. The reason is that the robot is too adventurous and becomes more careless towards the Q value which contains information about the past moves. The above situations show that the ϵ value should be calibrated really carefully to avoid potential cases where the robot becomes very randomly and cannot find a reasonable way.

Problem 2

(1) During this situation, the performance of the agent is not good, it still stuck at a local optimization without fully discovering the world. The reason is that staying at the this point will keep a maximum Q value. When the robot does not know where the thief is, it does not predict where the thief is and will not try to avoid meeting with it.

(2) When the position of the thief is known, it is like an extra dimension of state. After several trials, a robot will tend to move away from the thief. I think that will explain why there is a great improvement when the position of the thief is known.

(3) Since random values are used, so every time the situation is different, it is a little difficult to say which one is better. In order to deal with such dilemma, I calculate the mean of reward for each learning rate and set the epsilon to 0.05. I assume that for different *epsilon*, the largest reward has the same learning rate. And I calculate mean for 3 times, meaning the I will run the document for 3 times. Then the one with the maximum mean will be the best.

learningrate	Mean One	Mean Two	Mean Three
0.2	270.2835	270.0290	269.2105
0.16	272.2130	270.9815	273.0795
0.15	274.3420	270.9180	271.2055
0.14	270	273.3990	273.1750
0.13	271.3260	272.2525	271.6505
0.12	273.7280	275.9785	273.1665
0.05	270.1710	267.3710	271.5690

$\epsilon = 0.05$

The standard to find the optimal epsilon is $\arg\max_i\{MeanOne, MeanTwo, MeanThree\}$. So the optimal leaning rate is 0.12. Since it has the largest rewards 275.9875.

epsilon, LR=0.12	1	2	3
0.05	272.0370	271.6990	272.9835
0.04	285.4450	286.6395	283.7460
0.03	299.7555	298.3695	292.9415
0.02	310.0290	304.1700	307.1905
0.01	307.0540	310.2375	307.5319
0.007	311.7160	308.2540	304.7065
0.005	301.9990	308.4555	298.4645

According to the above table, $\epsilon = 0.07$ is an optimal choice.
I think when the learning rate is large, like 0.2, the rewards are smaller because it changes very dramatically when there is something unexpected happens meaning that it will underline the potential reward or penalty, but the penalty is big and more easily to meet with. Do a high learning rate will more or less makes the agent more careful.
For a small learning rate value, this has been reversed. So in order to get a larger value, the agent should combine the characteristics of both side, meaning that it is both willing to be adventurous and careful.
All in all the optimal value for learning rate is 0.12 and for ϵ is 0.07.

Problem 3

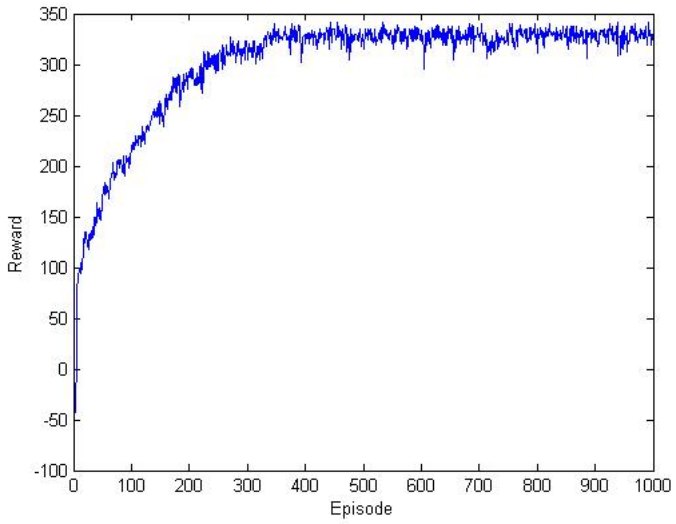

Figure 1: Rewards After Each Episode

From the graph, reward increases dramatically from 1^{st} to round 100^{th} sample, and then the curve oscillates around 270 and does not make any dramatical change.

We can find that the overall trend of this graph is similar to Figure 1. The

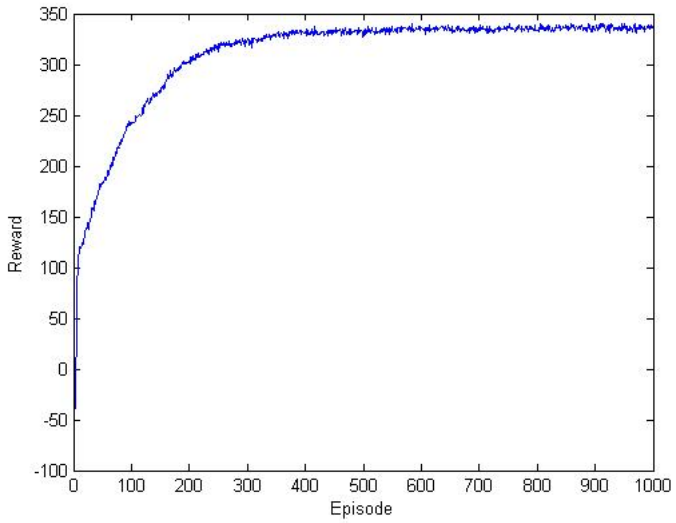

Figure 2: Rewards From Average 10 Times

mainly difference is that the last figure is much more steady than the first image since it is the average of several data. That shows that if we exclude the influence of random numbers with averaging rewards on different time periods, the results will be more steady. Besides, the overall reward does not change very much after 100 samples which show that a local minimum value has been found and the mean does not change much.