Homework 2 CS 440

Problem 1, in WorldWithoutThief

(a) With $\epsilon =$, the robot will never make an action that might be costly because it will never take any random actions, so, it will never walk through a puddle, and will therefore never leave the left side of the map.

(b) The robot can now take actions that might be costly, and will take them at random as it learns about the map. In this case, the robot figures out that it has to walk through puddles. Now the robot can move through puddles and deliver the packages. The reward increases slowly, but it does increase.

(c) In this case, the robot can get a very large reward, but does so with less consistency. This is because the robot takes random actions equally as much as it follows the policy it is creating, so it is much more difficult to generate a good policy.

Problem 2, in WorldWithThief

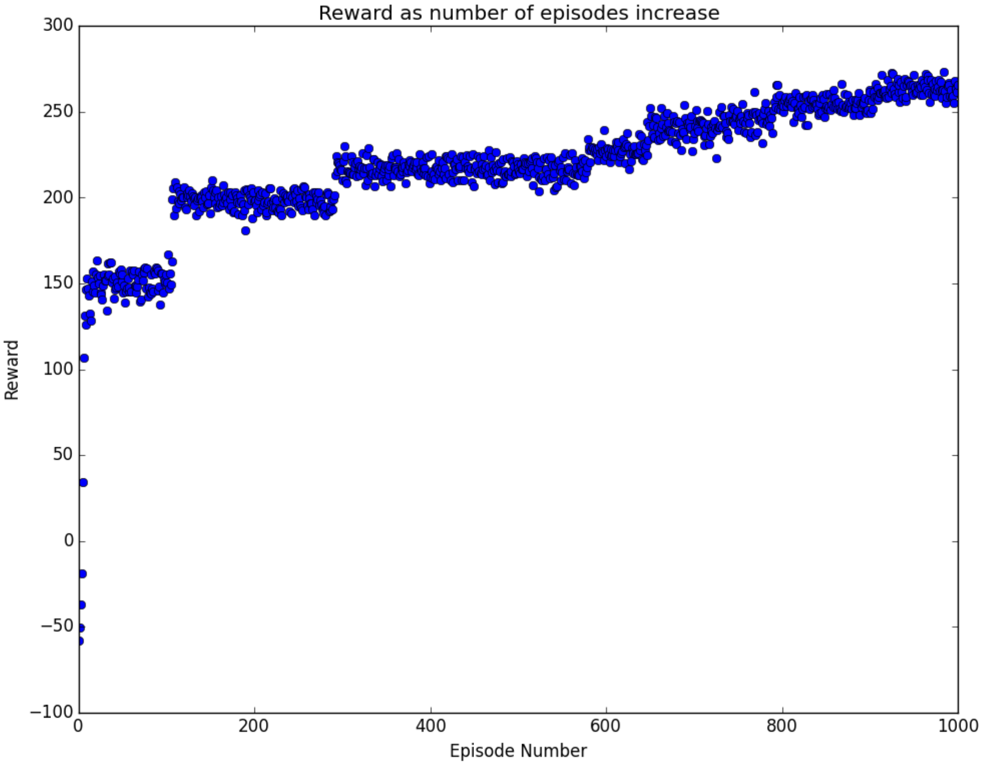
(a) We see that the robot almost always gets a negative reward, because it keeps running into the thief. Because the robot doesn't know where the thief is, it cannot do anything to avoid it.

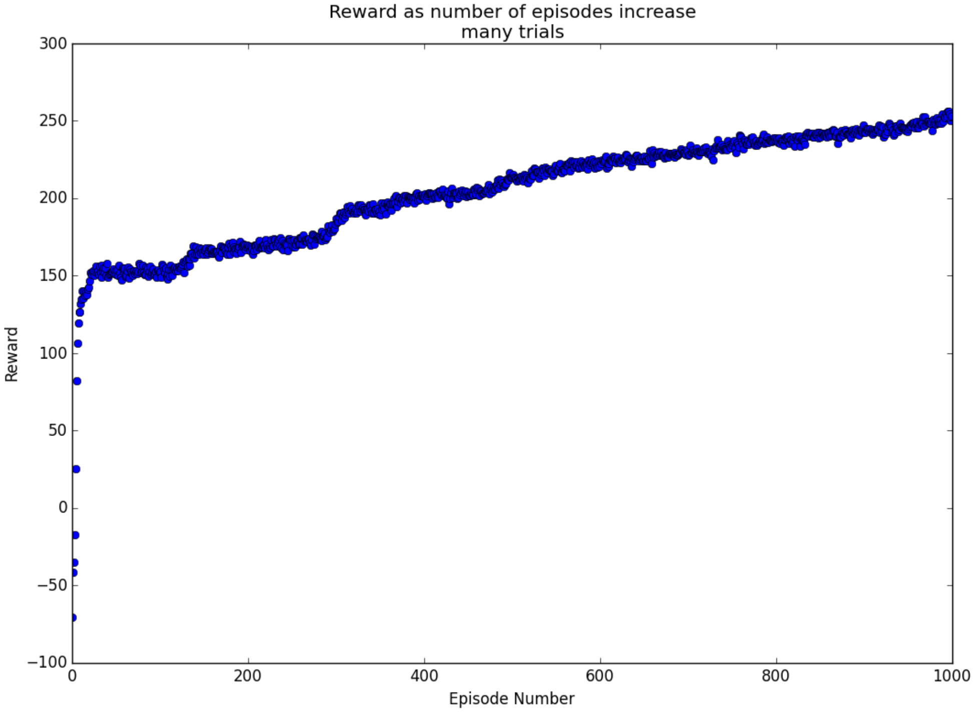
(b) With knowledge of the thief's location, the robot can learn to avoid the thief most of the time, and can learn to walk through puddles when it is not holding packages, thus, we see very high rewards.

(c) I got the best results with $\epsilon = 0.008$ and $r = 0.01$. The reward was 303.500000. To find this, I wrote a bash script to vary the parameters in the QLearningAgent file, then rerun the simulation.

Problem 3

I notice that initially the robot performs very poorly, and that, as the number of episodes increases, the change in benefit decreases. In the single trial case, there seem to be points at which the robot gets stuck, but this staged growth does not appear in the averaged plot.




Reward as number of episodes increase
many trials
Reward
Episode Number
300
250
200
150
100
50
0
−50
−100
0
200
400
600
800
1000