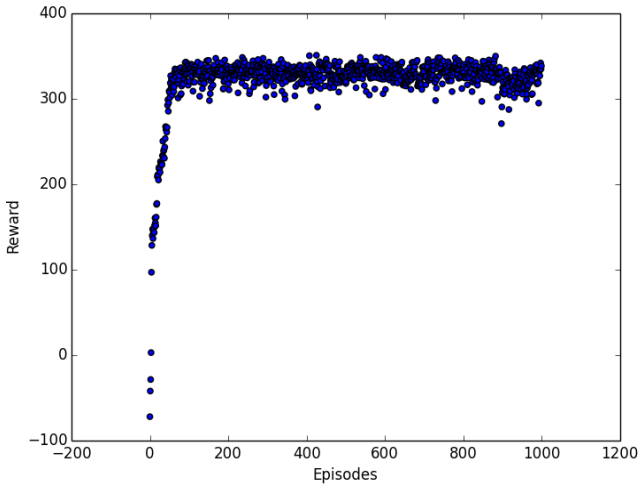
Danny Gaeta
CS 440
MP 2 Part 2

1. In WorldWithoutTheif
    a. I observe that for the first few episodes - my agent's reward is a negative value (which is approaching 0.0 for these first few episodes). However, after the first few negative values that are approaching 0.0, my agent attains a reward of 0.0 for all of it's remaining episodes (i.e. episodes 3-1000 are all 0.0 reward values). We can conclude that since epsilon is set to 0.0 this means that we have no exploration mechanism for choosing actions that are not simply choosing an action in which the agent believes to have the best long-term effect. Our Reinforcement learning requires clever exploration mechanisms and in this example we limit our exploration mechanisms by not including a random choice.
    b. After setting epsilon to 0.1, I observe my agent's reward score approaching 0.0 for the first 80 episodes. Where the most extreme negative score's occur at episode 0, and gradually become less negative as the episodes approach episode 80. Around the episode 80 mark (and a few episodes here after), the agent's score fluctuates between single digit positive and negative values. However, after about the 89th episode, the agent's score quickly progresses to a 3 digit positive reward value (about 130). It is here on after the agent's scores are all mostly positive 3 digit values that vary between (100 and 225). We can conclude that because we have small random exploration mechanism in our Q-Learner, the agent is now able to find more clever solutions. (It seems that a small value of epsilon is optimal. Randomly selecting actions, without reference to an estimated probability distribution, would give rise to very poor performance.)
    c. After setting epsilon to 0.5, all of my agent's reward value varies between negative reward scores of -50.0 and -20.0, while receiving a positive single digit reward value a few times. We are observing poor performance with epsilon equal to 0.5 because our agent will choose a random action and disregard the best long term effect every 1 of 2 actions. Thus, our random exploration mechanism is as dominant as our q-learning model. And selecting actions, without reference to our q-learning model is giving rise to a very poor performance.
2. WorldWithTheif
    a. With epsilon = 0.05 and when the agent DOES NOT know where the thief is, performance is very poor (Reward values vary between -25.0 and -10.0). We can conclude that this is due to the missing knowledge of the agent not knowing where the thief is. In other words, the agent needs a better representation of the world it operates in, in order to perform at a higher level. Furthermore, because the agent does not have a more complete representation of the world it operates in, the actions it chooses are not sufficiently accounting for the negative score adjustment associated with running into the thief, and thus the long term effect that the agent is optimizing for are not correctly computed.

b. With epsilon = 0.05 and when the agent DOES know where the thief is, performance is very strong. Reward values start at a negative value of about -70.0, but within 7 episodes the agent's reward value reaches a positive value of about 40.0 (After about the 10th episode, reward scores vary around a score of 280.0). We can attribute this strong performance to a more accurate representation for the agent of the world it operates in. Thus, the agent is able to accurately formulate long term effect q values.
c. For my investigations, I began with the assumptions that I wanted a very small degree of randomness (this was so that my system would choose actions based on the best long term effect most of the time, but still include a mechanism of randomness for exploring) -- for these reasons I started with an epsilon of 0.05. Next, I wanted a learning rate that was suitable for a stochastic environment. In which, the agent would overwrite old learning less than half of the time. So I began with a learning rate of 0.3. After a bit of experimentation, I found that I needed by agent to use new information at a lower factor. So I started decreasing my learning rate. Likewise, I found that my agent needed a decreased factor of randomness and needed to stick to focusing on long term effect more. Thus, after experimenting a bit, I found my best scores came at: rate = 0.22 and epsilon = 0.01

3. Graphs placed on next page
    a. For the single simulation, the scores begin negative then sharply rise to about 320. From here on after the scores fluctuates from about 300 - 320.
    b. When we run the simulation 9 more times and average the score received at each corresponding episode, we obtain a graph that looks a lot like the graph from a single simulation but now the scores have less variance. Once the scores reach about 305, they compactly oscillate between 300 and 310. In relation to the graph of a single simulation, this graph is much smoother with less outliers.

1 Simulation



10 Simulations Averaged