Jessica Chu
CS440 - MP1 Part 2

1a) The first episode had a negative reward, which was usually -8.0 over the course of several runs. The rest of the episodes had 0 reward. Because the agent never makes random decisions, it only knows learns about negative awards and learns to do nothing.

1b) The episodes at the start have mostly negative rewards, but become closer to 0 and are sometimes positive from around episode 100, and around episode 150, rewards became consistently positive. Because there was some random choices involved, the agent learned further despite getting negative rewards at first.

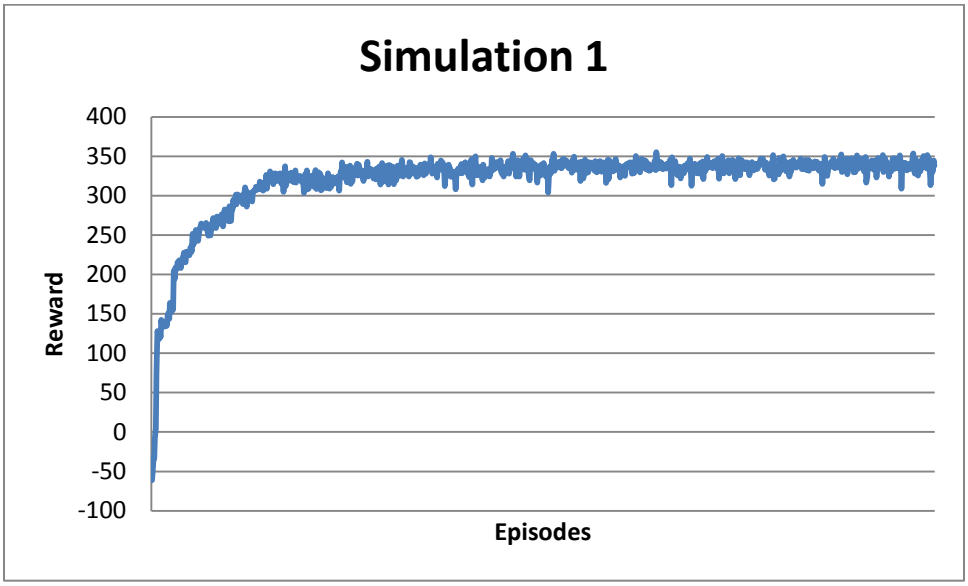
1c) Almost all episodes result in negative rewards. The agent did not make optimal choices half the time, and so the random choices overwhelmed the optimal choices and the agent received mostly negative rewards.

2a) Almost all episodes result in negative rewards, between -5 and -20. The performance of the agent was not good, because the agent was unaware of the thief's policy and made choices optimal to a world without a thief.

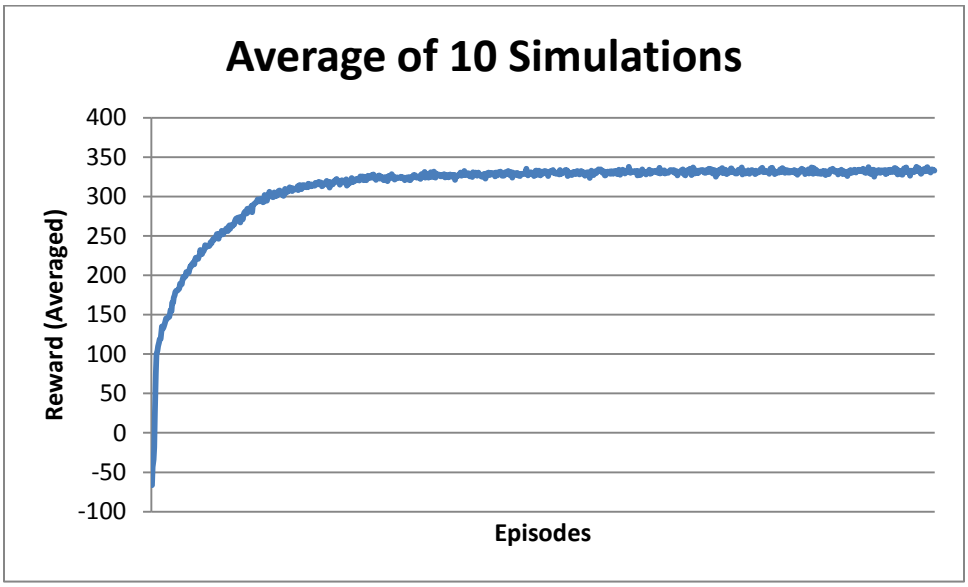
2b) The first few episodes result in negative rewards, but they increase over episodes, then quickly ramp up. Around the 60th or 70th episode, the rewards begin to level, hovering around 250 to 300 reward. This happened because the agent learned to avoid the thief, because running into the thief would lead to negative reward.

2c) First, I changed the learning rate. With epsilon set to 0.05, I tried learning rate values of 0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5, 0.6, and 0.7. Of these, 0.15 gave the highest average reward value (273.278). Values began decreasing for values smaller and larger than 0.15. Using a learning rate of 0.15, I changed the epsilon value. I tried epsilon values of 0.005, 0.01, 0.015, 0.02, and 0.05. Of these, 0.01 gave the highest average reward value (318.913). Values began decreasing for values smaller and larger than 0.01. Therefore, I conclude that the agent will get the highest average rewards from a learning rate value of 0.15 and an epsilon value of 0.01.

3)



The reward increases over episodes and converges between 300 and 350.



The reward increases over episodes and converges around 325 or 330. This is a much tighter bound than the single simulation. This is because the average reward of each episode in each simulation averages out to be closer to this convergence value.