CS 440 MP1 Part 2
Avdhesh Garodia

1. World Without Thief

a) The reward went to 0.0 and stayed there. This happened since we have an epsilon value of 0.0, which means our agent does no random exploration. We are stuck in a local optima where the agent finds the best solution is the move up and stay there never crossing any slippery paths at all

b) In this case, epsilon equals 0.1, which means that 10% of the agent's actions are random. This allows the agent to explore beyond the slippery region and see that it can gain a positive reward if it does so. We have gotten out of the local optima since we are able to explore randomly. We now get a positive reward, however there is still some fluctuation in the results since there is a still a lot of randomness involved. The reward was anywhere from positive 50 to positive 230.

c) In this case, epsilon equals 0.5, which means that 50% of the agent's actions are random. This gave negative results as a result anywhere from negative .5 to negative 70. This happens since half of the actions are random which is very extreme.

2. World With Thief

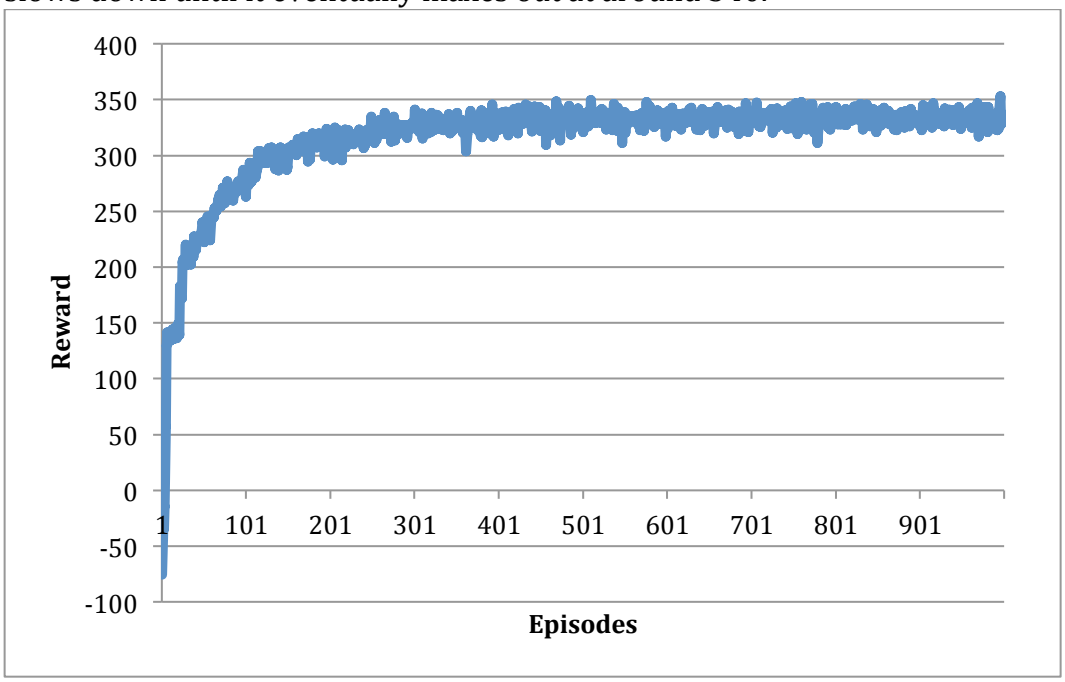
a) Setting epsilon to 0.05 and setting knows thief to N. The agent performs poorly getting negative rewards for each episode. Without knowing the location of the thief the agent will mostly likely run into the thief while delivering the packages, it only learns to avoid the slippery paths.

b) Keeping epsilon at 0.05 but setting knows thief to Y. In this case the agent performs well getting high positive rewards approaching 300. Setting knows thief to Y adds more states depending on the location of the thief. This allows the agent to avoid the thief.

c) Searching for the best learning rate I found that an epsilon of .1 is too high and that it was better to lower the epsilon to .01 so that only 1% of our moves are random. For the learning rate the reward did not change to much and a rate anywhere between .08-.13 would give be very similar results. I picked .09. With these values I was able to get rewards around 320-350. A small improvement over the previous settings

3. Graphs

Plotting the simulation from 2c we get a graph that resembles the graph of a logistic function. As the episodes increase, reward increases really fast at first and then slows down until it eventually maxes out at around 340.



Taking the average over 10 simulations shows us the logistic function more clearly.