MP1

Part 2: Experimentation

1. a) The robot would never go past the first line of slippery patches. It was not willing to investigate past the negative reward inducing slippery patches to find the goals.

b) The results increased dramatically over the first twenty or so episodes, but did not really converge. The resulting policy would sometimes have the agent get stuck in one state, however. The epsilon value was high enough to perform fairly well during all the tests, but there still wasn't enough randomness to find the optimal policy.

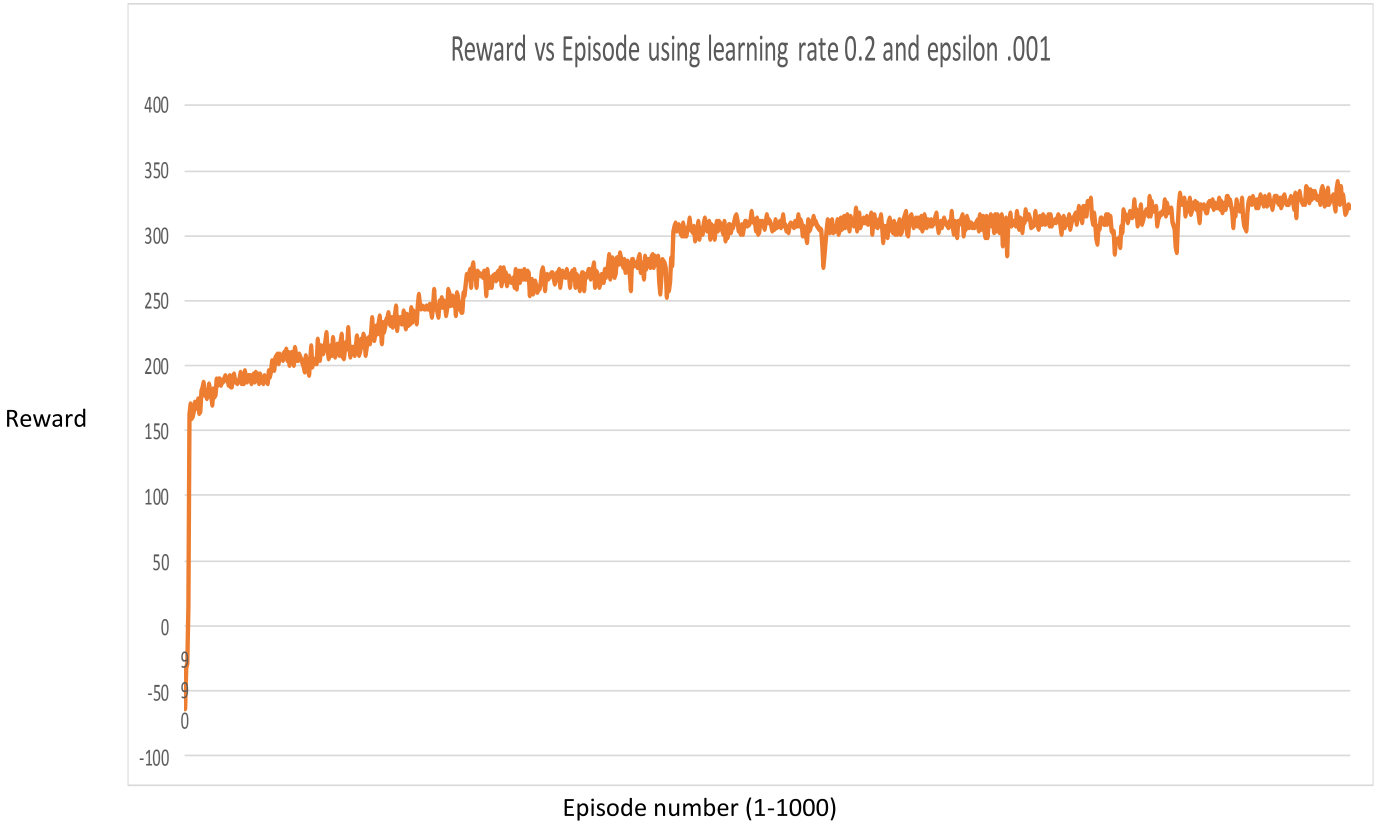
c) The results of the episodes never got very high due to all the randomness in the robot's actions. However, the policy created was very reasonable and did not get stuck like the other epsilon values. In this case, the robot was doing too much exploration and not putting its results to use as much as it could have.

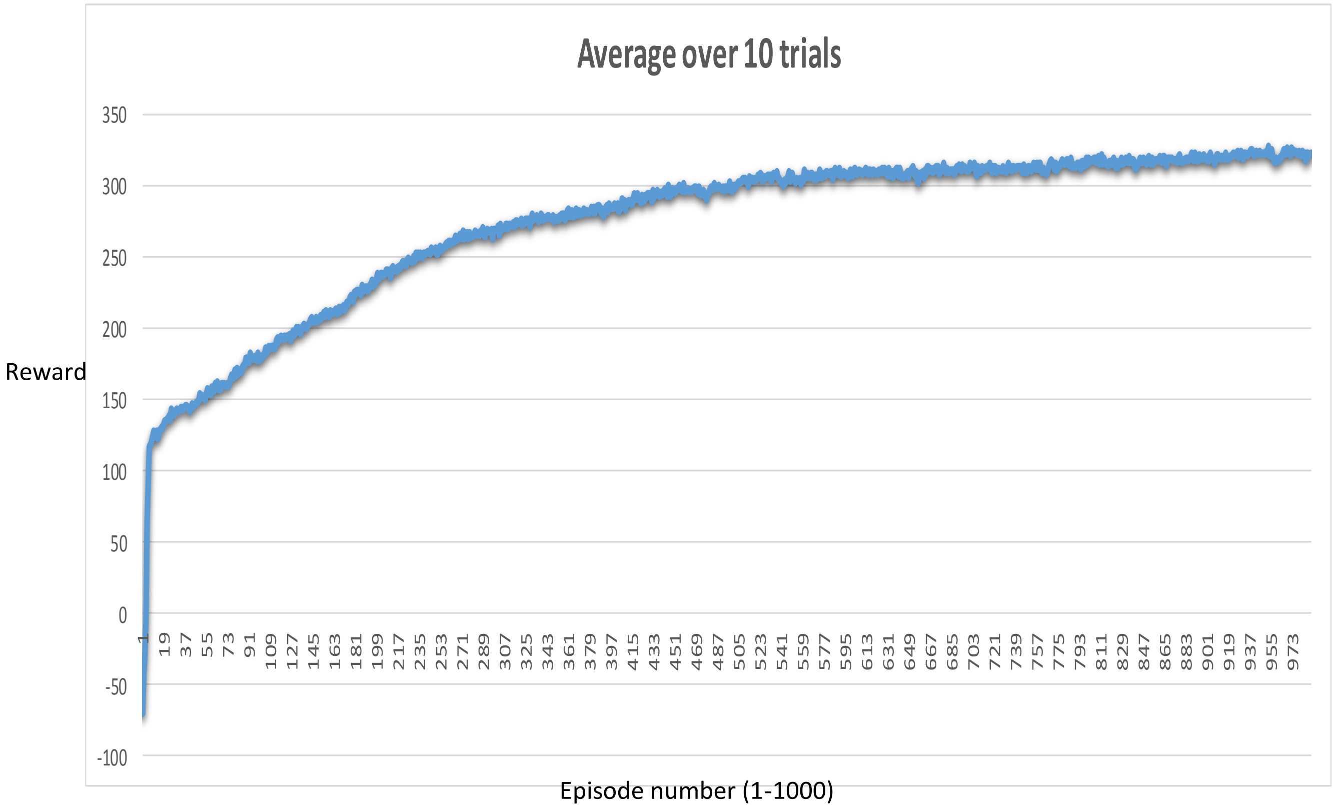
2. a) The robot never really incurs positive reward, and the resulting policy keeps in hiding in a corner. This result is reasonable, considering how difficult it is to deliver the packages without knowing the thief's location.

b) The robot performs much better now that it has knowledge of the thief's position. It is able to maneuver around it and incur large rewards, and acts reasonably.

c) The rates I came up with were .2 and .001. As you increased the learning rate, the rewards would converge faster, but end up with more variation at the end. The opposite happened with the epsilon values: as they increased, the values converged faster and the end result had more variance. The values I came up with had the simulation converge near the end, getting results pretty close to optimal but also converging to a fairly small range.

3. (See graphs on next page) The vast amount of improvement is done at the start of the trials, and then the policy is fine tuned in the latter episodes. This occurs because the agent is able to easily distinguish between obviously bad actions and better ones. However, the agent must also try actions many times to find replicate the reward function, which takes more trials and episodes to accomplish. On average, the episodes resemble logarithmic growth of the reward as a function of episode number.


Reward vs Episode using learning rate 0.2 and epsilon .001
400
350
300
250
200
150
100
50
0
-50
-100
9
9
0
Reward
Episode number (1-1000)


Average over 10 trials
350
300
250
200
150
100
50
0
-50
-100
1
19
37
55
73
91
109
127
145
163
181
199
217
235
253
271
289
307
325
343
361
379
397
415
433
451
469
487
505
523
541
559
577
595
613
631
649
667
685
703
721
739
757
775
793
811
829
847
865
883
901
919
937
955
973
Reward
Episode number (1-1000)