James Wegner

jewegne2

1.

a. The robot kept going north and then stopped once it hit the boundry. Since there was no epsilon value the agent takes the first lowest q value action, which is to go north. This resulted in the reward for each episode to be 0.

b. The robot was able to successfully deliver the packages. It also avoids the slippery areas as much as possible even when not carrying packages. Since the epsilon value was very small it rarely experimented with the slippery areas. The rewards for each episode were also very high.

c. Since the epsilon value was high the robot took more chances, which resulted in mostly negative rewards for each episode since the robot would travel more on the slippery areas. The robot avoided the slippery areas when carrying packages but when returning from customer one the robot would travel along the slippery path to the company since it was the shortest path. Since the epsilon value was high the robot was able to discover the short path along the slippery areas from customer one to the company.

2.

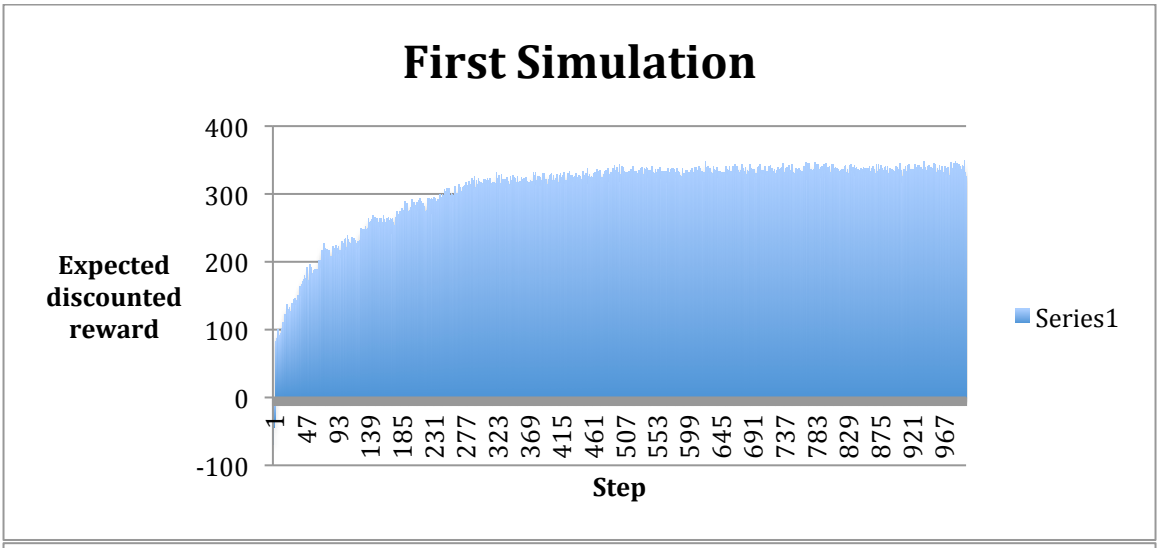
a. The robot kept traveling north and got stuck at the boundary, which led to negative rewards in the episodes. The epsilon value seems to be too low so the robot was not able to find a path to the customers due to running into the thief.

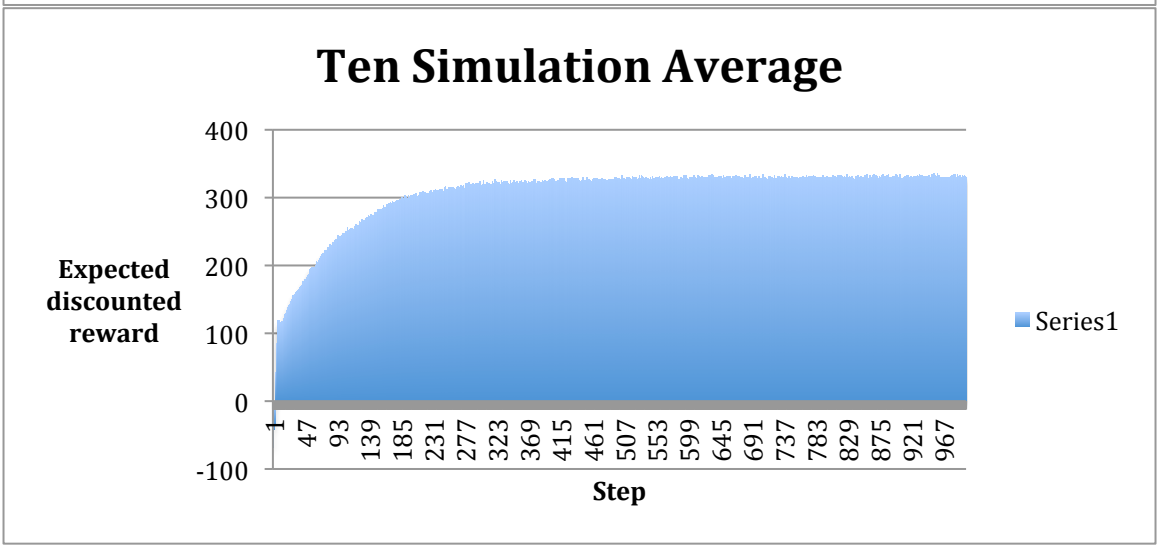
b. The robot was able to effectively avoid the thieves and deliver the packages, which led to positive rewards in the episodes. Since the robot knew where the thief was and we still used an epsilon value the robot was able to take chances and find a good policy for avoiding the thief while delivering packages.

c. I found that setting the rate to 0.2 and epsilon to 0.01 gave me the best results. If I tried moving the epsilon higher the rewards sharply dropped. I also tried higher rates with a higher epsilon, which led to negative rewards. If I kept the epsilon low and raised the rate the rewards started to drop. It looks like to find a good solution there must be a little randomness, but not too much.

3. (charts on next page)

The expected discounted reward for both charts start out negative for the first few episodes then grow logarithmically. They also stop increasing around 250 episodes. The first chart does continue to have some highs and lows after 250 episodes but averages at about 329. The second chart is more flat after 250 episodes and averages at about 327.


First Simulation
Expected discounted reward
400
300
200
100
0
-100
1
47
93
139
185
231
277
323
369
415
461
507
553
599
645
691
737
783
829
875
921
967
Step
Series1


Ten Simulation Average
Expected discounted reward
400
300
200
100
0
-100
1
47
93
139
185
231
277
323
369
415
461
507
553
599
645
691
737
783
829
875
921
967
Step
Series1