Parker Brown
9/27/14
ECE 448

MP 2.1 Write Up

1a.
The agent receives a negative reward for the first episode, and then receives a 0 reward for every subsequent episode. The policy that is generated has the robot move to the upper left corner of the map and just stay there, sometimes attempting to move up another square but remaining in a fixed spot for the entire series of steps.

b.
The agent receives a negative reward for roughly the first 30 episodes. After that the agent receives all positive rewards ranging from 50-200. The policy generated actually sees the agent delivering packages correctly.

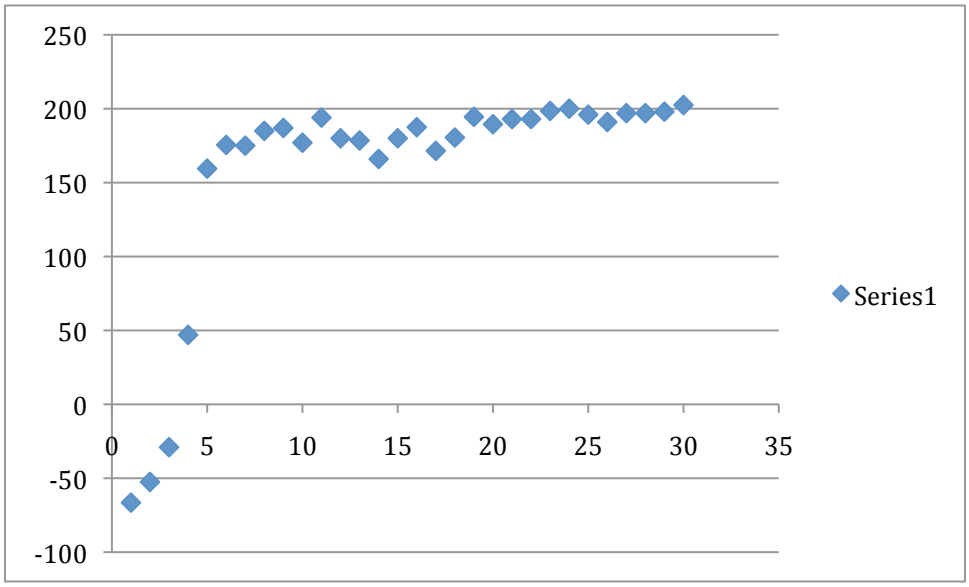
c.
For an epsilon of 0.5, the agent receives a negative reward between -10 and -50 for every episode. For an epsilon of 0.9, the agent receives even more negative rewards for every episode, in the range of -60 to -70. I believe with the epsilon that high, the agent makes too many random moves. This causes it harm because it can't implement its learned policy.

2a.
The agent receives mostly negative rewards in the -10 to -20 range, with the occasionally barely positive reward. There is no general increasing or decreasing trend between episodes. I believe this is because the agent does not know of the thief, it can't learn to avoid it.

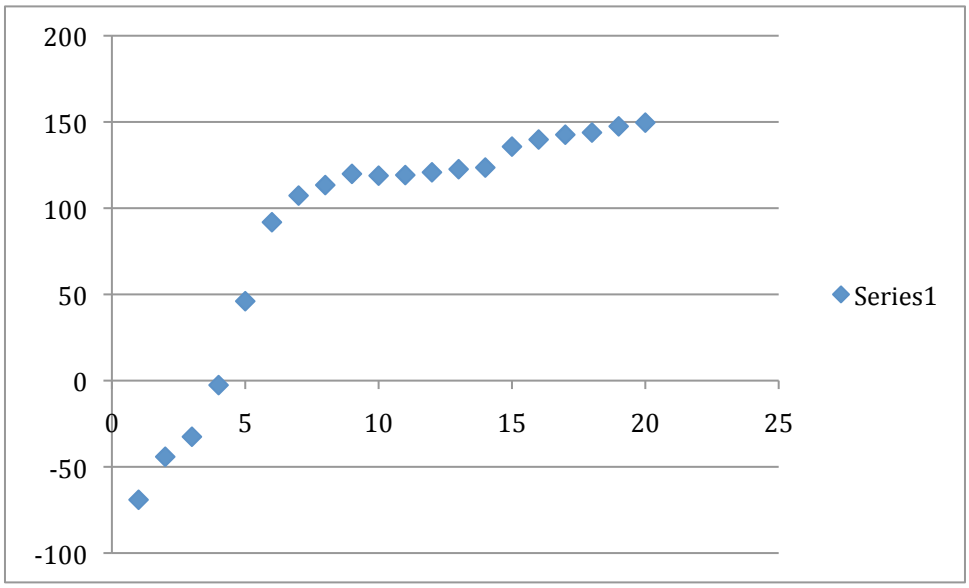
b.
The agent receives a negative reward for the first few episodes, but then receives a positive reward in the 200-300 range for the remainder of the episodes. Because the agent now knows about the thief, he can learn to avoid it, which greatly increases the reward as seen.

c.
The best epsilon I found was anything in the .005 to .01 range. Anything greater than that is causes too many random steps, which can cause the agent to get negative rewards. The best learning rate I found was a number in the range of 0.05 to 0.1. Anything greater or less than thought saw a large decrease in the reward the agent was getting every episode.

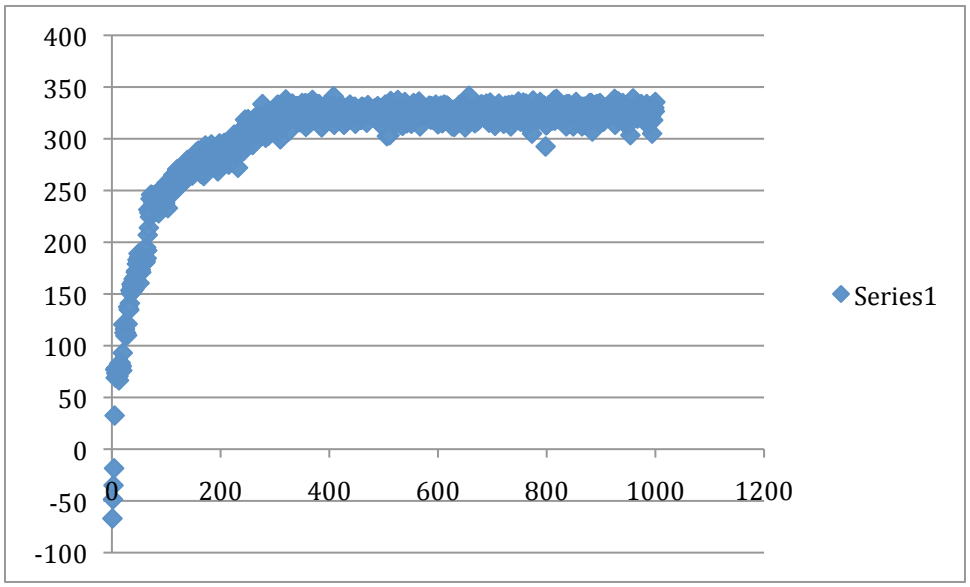
3.



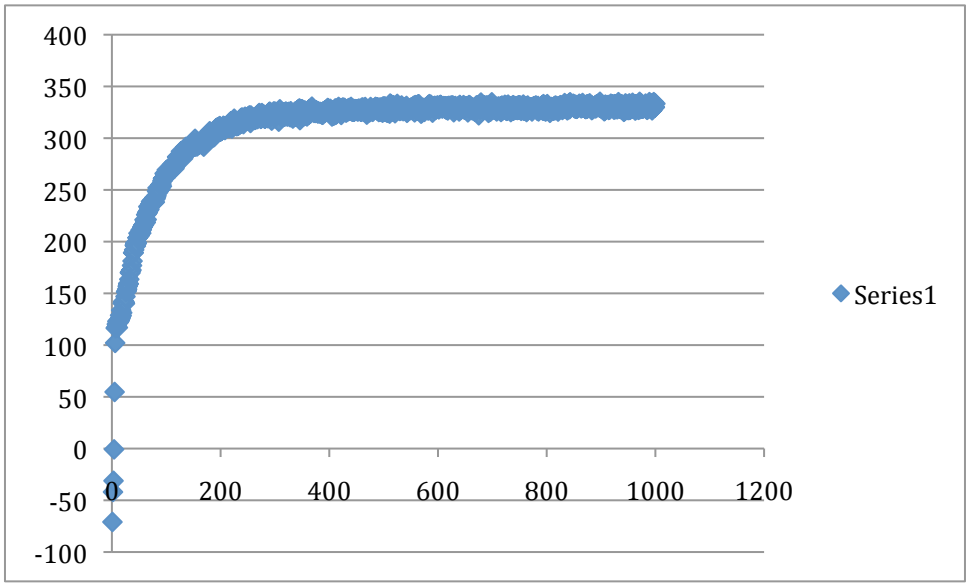
This chart shows the reward at the end of each episode for the first 30 episodes. As you can see, for the first few episodes the agent does poorly but quickly learns and reaches a steady reward value for each episode by the 20th episode.



This chart shows the reward at the end of each episode averaged out over 10 runs. It again shows a quick learning rate followed by a steady state reward value.



For my first 2 graphs, I only included 30 episodes to show the learning rate early on. Here is the graph for all 1000 episodes.



Here is the graph of the reward for 1000 episodes, averaged over 10 runs.