Thomas Lesniak (tlesnia2)
CS440 MP1 part 2

1a) The reward remains at 0 and the agent moves only up and down because there is 100% greediness, and the best action is never moving to an area where a negative reward could be had.

1b) The agent now attempts to deliver both packages but is not able to make it back to the company to pick up more. Now that we introduce the epsilon, there is a chance that the agent will move into risk areas, however once the agent delivers it's packages it does not know a better way to get a higher reward because epsilon is so low.

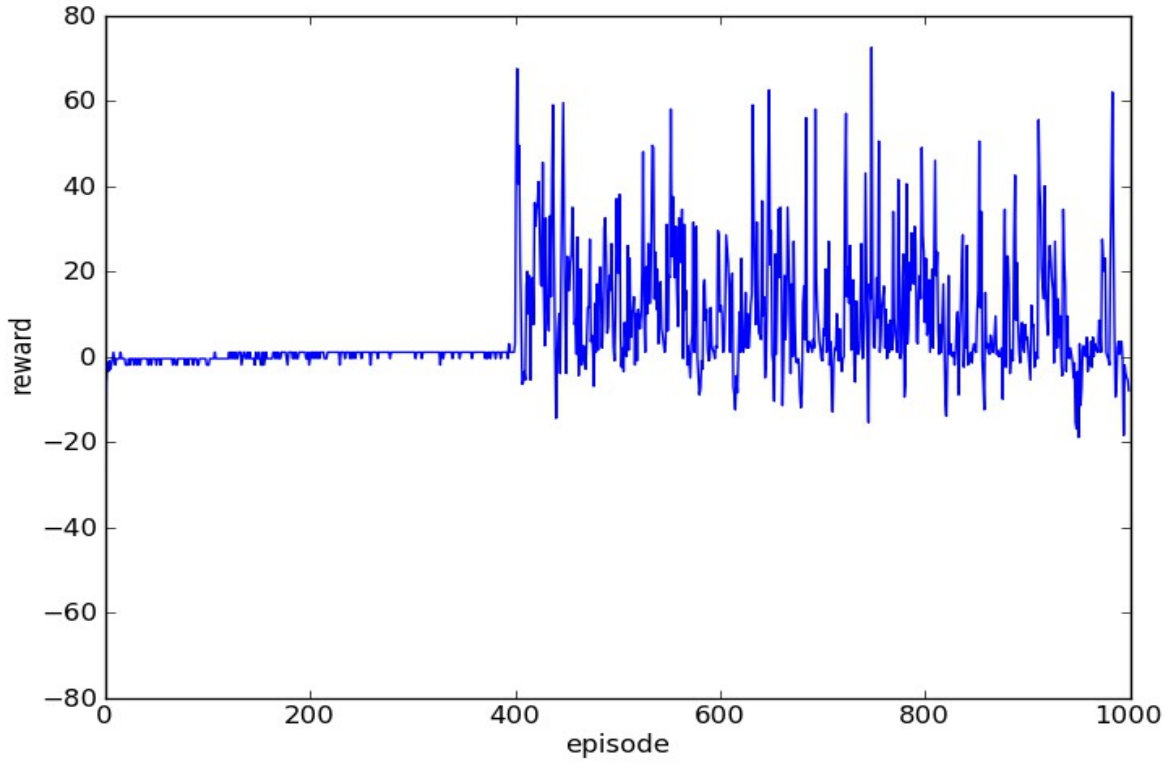
1c) Increasing epsilon above a certain point seems to generate more and more negative rewards. while the agent attempted to deliver the packages at low, non-zero epsilon, it often failed to with increased epsilon. This is because with higher epsilon, we choose more and more random moves over best moves.

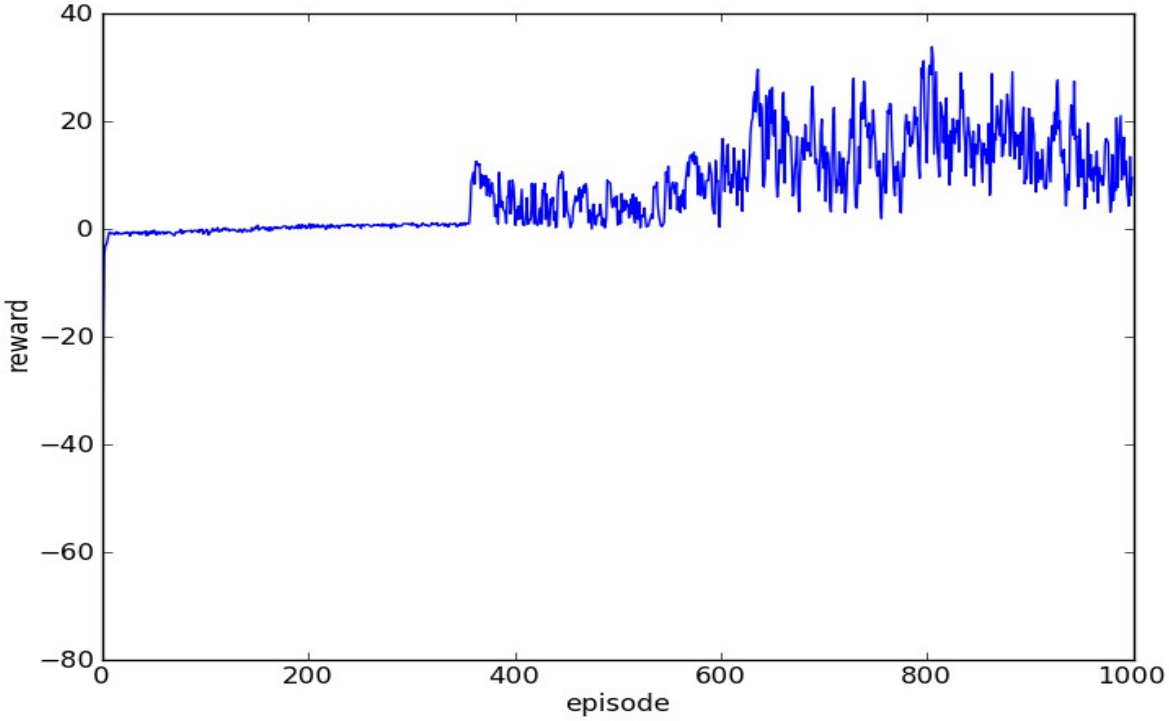
2a) The agent does not seem to be able to deliver the packages on a consistent basis. This is because the utility is being taken away by the unknown thief, but also because the epsilon value is so low that it is very easy for the bot to move in a cycle.

2b) The agent learns to avoid the thief and deliver the packages, but again, because the epsilon is so low he is unable to make it back to the company to pick up more packages due to the low epsilon value.

2c) Given that as epsilon approaches 1 (keeping other values constant), the agent performs worse, the ideal epsilon should be some relatively small, non-zero, positive decimal. Testing between .05 and .2. The most consistent and highest-reward values for epsilon and the learning rate are epsilon = 0.05 and rate = 0.2 according to my investigations. Extreme values (close to 1.0 and 0.0) for the learning rate yielded relatively low but stable results. rate = 0.2 seemed to yield the highest rewards.

The first image shows the first run's reward output, while the second image shows the average across 10 runs.


80
60
40
20
0
−20
−40
−60
−80
reward
0
200
400
600
800
1000
episode


40
20
0
−20
−40
−60
−80
reward
0
200
400
600
800
1000
episode

The graphs tend to start off around of 0 but suddenly jump around the halfway (400 steps or so). Once a high reward has been reached, it is easy to for the agent to climb back up to a high reward again, although due to the randomness it is not always going to yield higher and higher rewards. Looking at the average graph, the result I an upwards trend in the rewards gathered as time goes on.