CS440: Artificial Intelligence
MP1 – Part 2

Aditi Mhapsekar (mhapsek2)

1.

a. The reward always stays at 0 through all the episodes when epsilon is set to 0 – the agent never crosses over to the region to the right of the column filled with slippery grids. This is because in the cells adjacent to the slippery column (and to the left), the q values for going in any other direction is 0 as opposed to a negative value when crossing over the slippery grids. Since there is no randomness here, it will always pick the grids with higher q values (in this case 0), thereby never exploring regions that might give the agent a greater positive reward.

b. Greater positive rewards are observed in this case, unlike the previous scenario, due to the randomness in taking an action by ignoring q values $1/10^{th}$ of the times, the agent ends up crossing the column of slippery grids and exploring areas that offer a higher reward by delivering the packages. Also, because of this there is a component of exploration associated with the actions of the agent – there is a greater variance in the rewards gained by agents policy through the episodes.

c. Negative rewards are observed through most of the episodes in this case. This is intuitive since ½ the time the actions of the agent are purely random and there is no logical basis for going in a particular direction i.e. the q values don't hold as much of a significance. The agent can keep moving back and forth over the slippery region resulting in large negative rewards.

2.

a. Negative reward values are seen through the episodes. Since the agent is not aware of the location of the thief, the packages end up getting stolen resulting in a negative reward through all the episodes despite the fact that the agent is designed to explore sufficiently well.

b. Positive reward values are observed in this case. Since the agent is aware of the location of the thief and can at the same time explore new areas it ends up delivering both the packages and gaining positive reward.

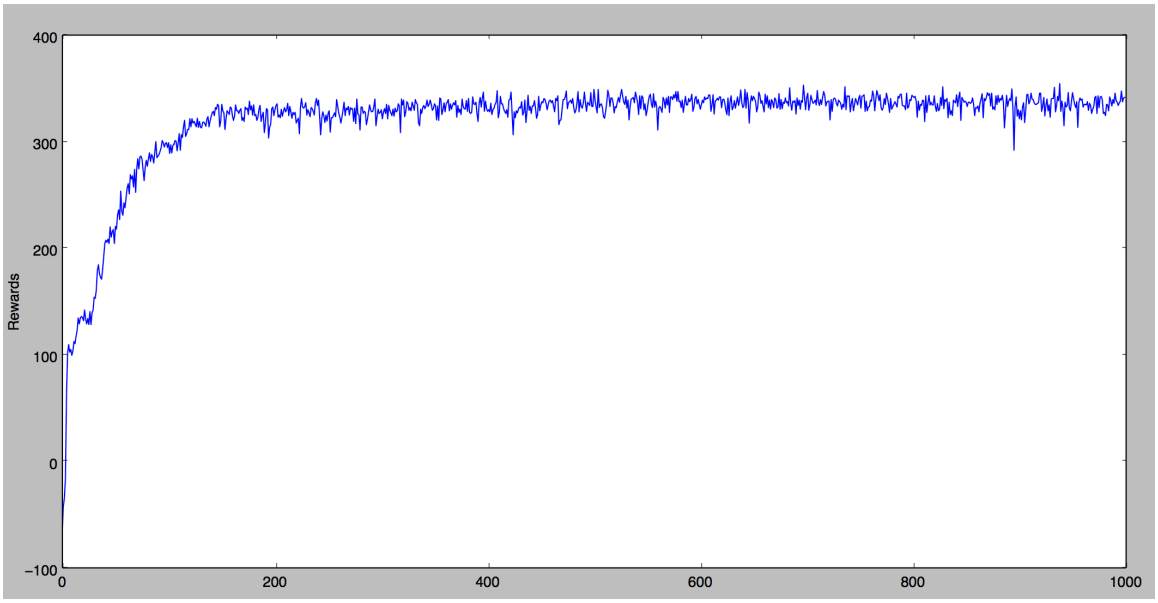
c. Tried out epsilon values in the range 0.001 - 0.5. The best results were obtained for the epsilon value of **0.007**. The objective is to introduce a little

randomness while at the same time respecting the q values to the extent possible. Little exploration is no good neither is exploring all the time – clearly very high values and very low values don't do as well.

Learning rate values between 0.01 and 0.5 were tried out. The best results (most consistently high) were obtained with the value of **0.15**.

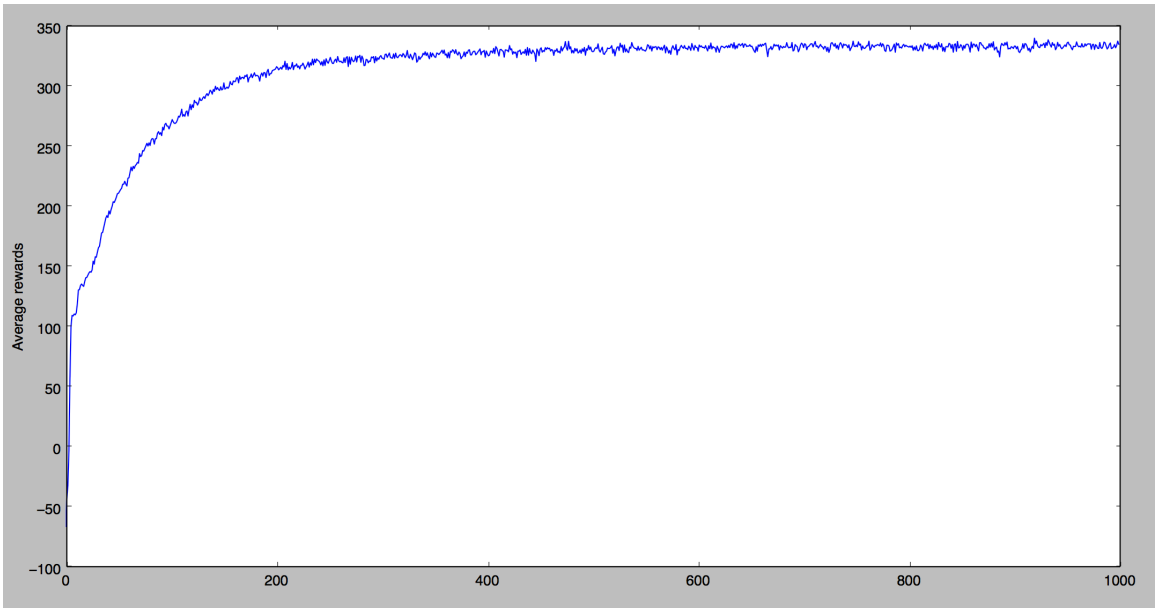
3.

Total rewards at the end of each of the 1000 episodes with parameters set to the values mentioned in 2c:



Initially there is the learning phase where the agent tries to learn the q values for each of the cells. With progressing episodes, this stabilizes as can be observed by the plot. We notice some spikes though, which can be attributed to the random exploration.

Total rewards averaged over 10 simulations:



The spikes observed in the previous chart smoothen out – this plot gets rid of the variance observed in any one run.