Machine Problem 1

Jayant Ahalawat

September 29, 2014

Answer 1

(a)

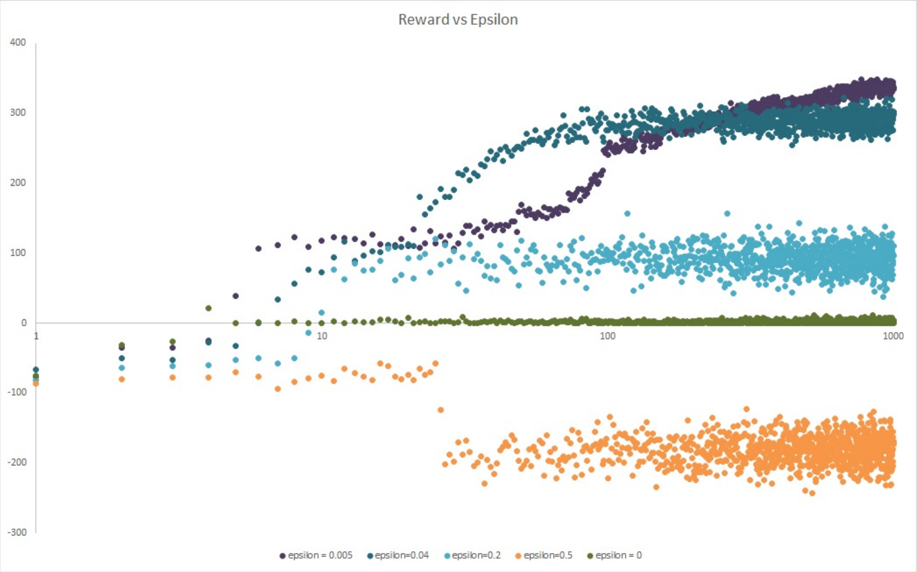
The agent gets a negative reward in the first episode and a reward of zero in all the subsequent episodes. This is because $\epsilon = 0$ in this case. Thus the agent is not performing any random actions. This is because of the purely greedy policy that we are following. This is preventing us from from exploring potentially better alternatives. Since in the first episode, the agent is getting negative rewards, subsequently it prefers getting no reward than a negative reward. That is there is no **exploration** in this case.

(b)

In this case, we have set $\epsilon = 0.1$. This implies that the agent will perform some random actions. Hence we observe that reward becomes positive after a few episodes since the agent subsequently explores better alternatives. However due to random action being undertaken by the robot, some negative rewards are observed later also.

(c)

In this case we are forcing the agent to act randomly 50% of the times. It is observed that rewards during simulation were mostly negative. However, if we run the policy simulator in this case, the performance of the agent is satisfactory. However the "regret" in this case would be higher. That is we are not learning the optimal policy optimally. That is given a large number of episodes, the agent will learn the optimal policy even with a high value of epsilon. Below is a figure for different values of ϵ and $\alpha = 0.1$. We can see from the figure below that higher epsilon leads to faster convergence, however the rewards are lesser if the values of ϵ are too high. Further in case of a larger ϵ the convergence is in a larger region (as shown in figure below, the x-axis is on a logarithmic scale).



Answer 2

(a)

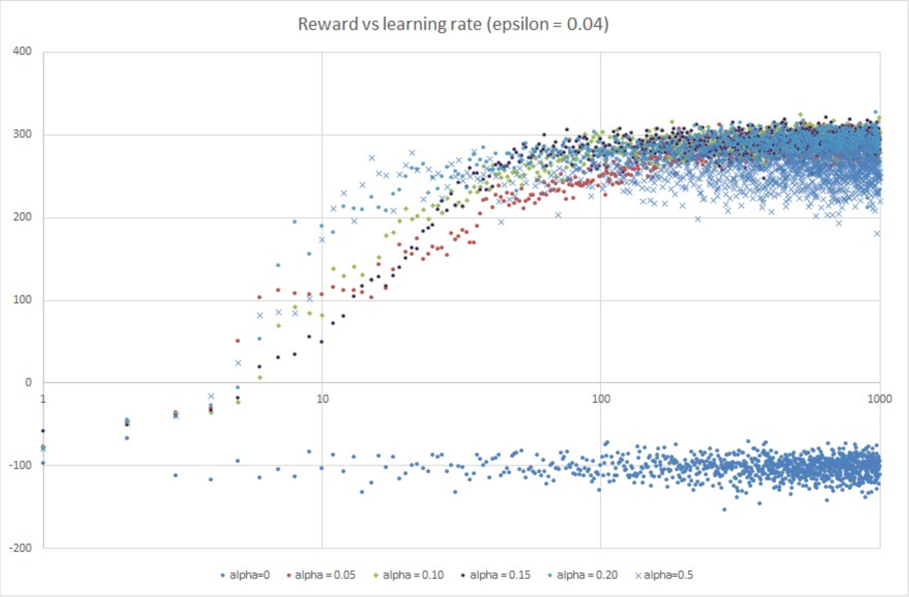
The agent performs poorly in this situation with negative rewards in majority of the episodes and there is hardly any improvement over time. This is because the agent is not aware of the position of the thief and an encounter with the thief carries a large negative penalty. While running the PolicySimulator, the agent performs poorly and a total reward of zero is observed

(b)

The performance of the agent improves considerably in this scenario. Since the agent is aware of the position of the thief, it is able to avoid any action that might lead to an encounter with it. Thus the penalty is minimized and hence the total reward in each case is higher than it was in case (b).

(c)

The optimal value, $\epsilon = 0.04$. This value ensures fast convergence in a small good region around the optimal policy. If we take values lower than this, the convergence becomes slower and if we take a higher value than this, the agent is exploring than is necessary to reach the optimal policy. For this value of ϵ the optimum value of α is 0.15. This can be seen from the graph below. Setting $\alpha = 0.15$ ensures both fast convergence around a good region. Higher values may lead to faster convergence but will result in lower rewards, whereas lower values will reduce the speed of convergence.

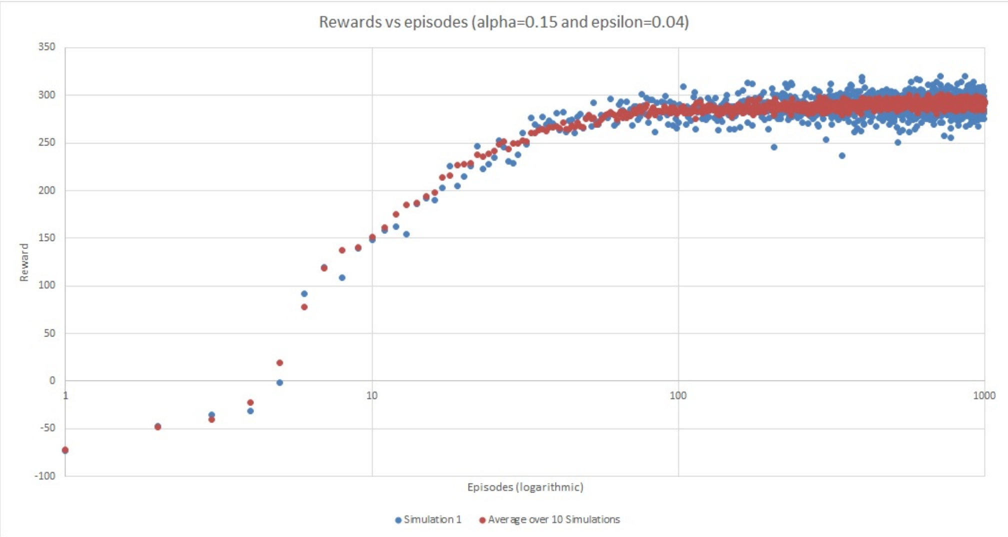


It is important to note that ideally a decaying value of both α and ϵ should be used. Since α denotes the relative confidence in new and old values. We should decrease it over time because new information is most important to us initially and gradually old information becomes important.

Similarly in case of ϵ, we want the agent to take more random actions at the beginning since the agent has not learnt anything yet. As the agent starts gaining information about the utilities of different states we would want the random actions to be minimized.

Answer 3

The plot showing the rewards for simulation 1 and average rewards over 10 simulations is shown below:



The plot shows that the convergence starts to take place at around the hundredth episode (around reward=300). The plot for the average over 10 simulations is plotted in the same graph. It is observed that now the area of convergence is much narrower. This is expected since while averaging, the deviations from the optimum policy are canceled out. If we increase the number of simulations over which we are averaging the area of convergence will become narrower.