CS747 - Assignment 4

Bhavya Bahl - 150050110

November 11, 2018

1 Task1

Initially, when the agent does not have a good estimate of the optimal policy, it takes long time to reach the goal state. This justifies the sudden increase in the number of time steps taken to reach the goal states for the beginning few episodes. As the number of episodes increase the agent learns better policy and those states which occur on the optimal path are visited more frequently and we obtain better estimates for these states. In this case, I was able to reach the minimum of 15 time steps/episode after iterating over 200 episodes. The learning rate was set to 0.5 and epsilon was decayed with the number of episodes as $\frac{0.5}{num_episode+1}$. For handling the corner cases, if the player tries to get out of the boundary, then it remains in that state and gets a reward of -1.

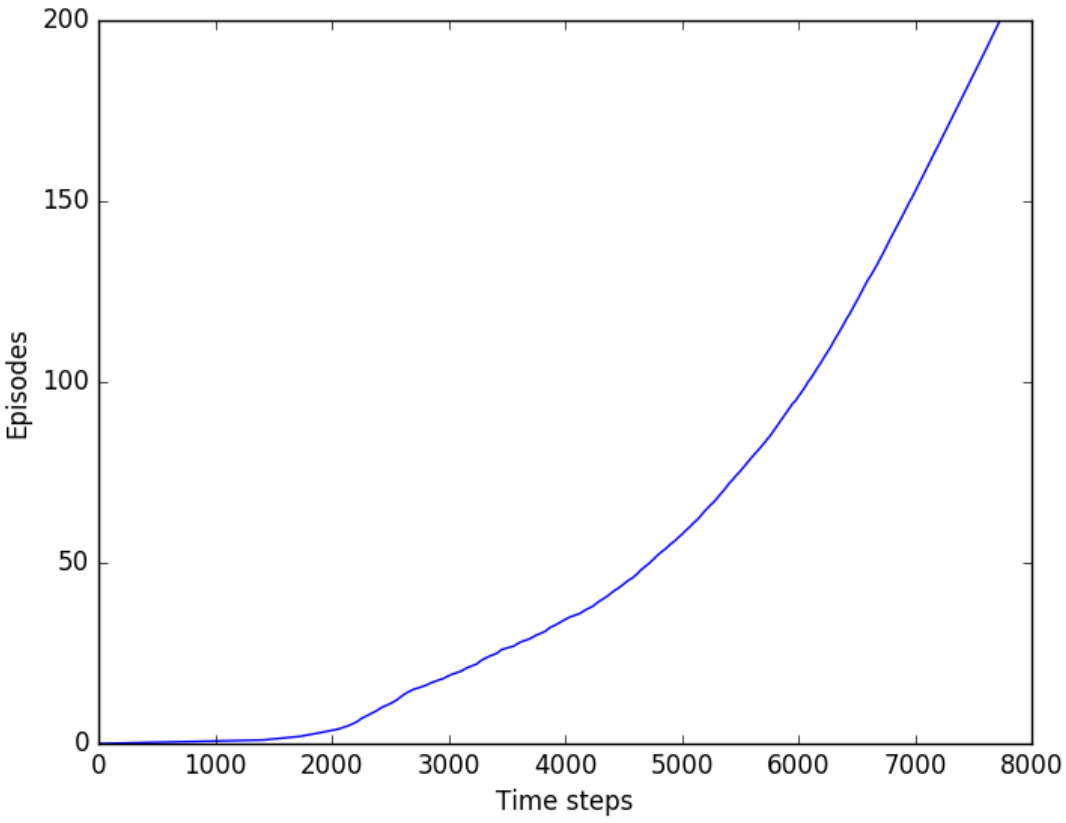

Figure 1: Task1: Episodes vs Number of time steps

2 Task2

This case is also similar to the above task. Because of more number of actions in this case, we need more data to get a better estimate of the Q values of various different actions. In this case, I was able to reach the minimum of 7 time steps/episode after iterating over 500 episodes. The same parameters for learning rate and epsilon were used as in the above case and the corner cases were also handled in the same way. It can also be seen that some of the states in the top right corner show 0 value function. This is because these states were never encountered in the trajectories and there action-value functions were never updated.

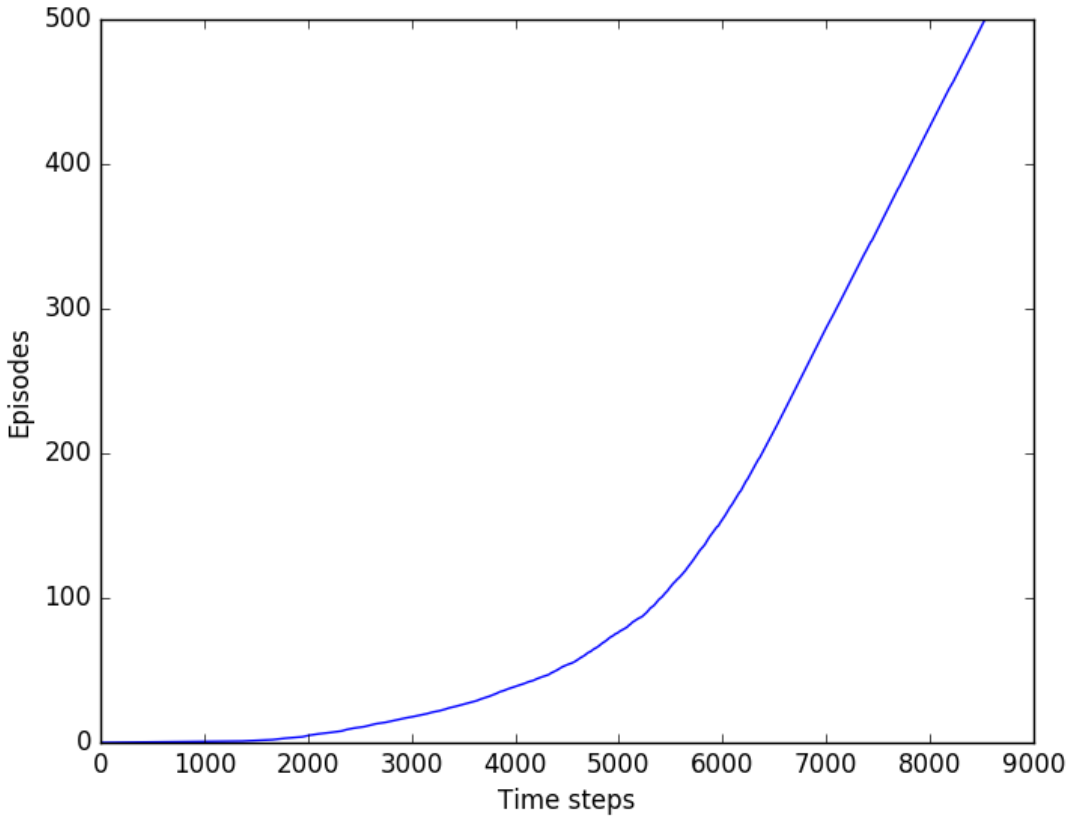

Figure 2: Task2: Episodes vs Number of time steps

3 Task3

Due to the stochasticity, this task is more exploratory in nature than the task without it. The states which were seldom observed in previous task will now be observed more frequently. Due to this fact, this task requires even more data than the previous task. Even after running the task for 10,000 episodes, the value function learned for the initial states lies in the range -12 to -18 for different random seeds. From the graph it appears that the number of time steps taken per episodes is almost a straight line due to the scale of the total number of times steps but there are minor variations as can be seen from the values. The parameters and corner cases were kept same as the previous cases.

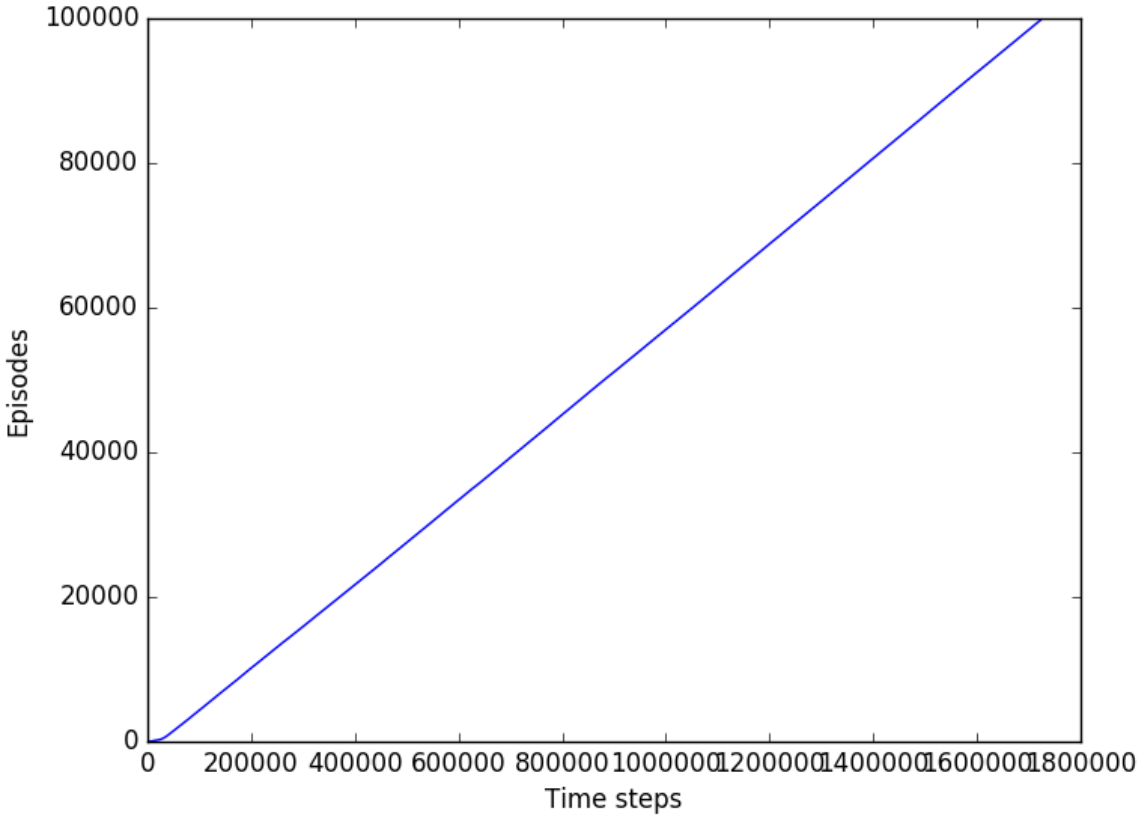

Figure 3: Task3: Episodes vs Number of time steps for 100,000 episodes