# Homework 4 Question 2 writeup

Yanlin Li, Student #: 1003770305

April 5, 2021

## 2.2.1

We are exploring the Q learning algorithm that select moves for the agent. In 90% of the time, the algorithm will choose the action with the maximum Q value. At the remaining time, it will randomly choose from the four actions available (left, right, up and down). We should "break ties" when the Q values are zero for all actions by choosing uniformly from the action.
First, we will keep the original start point (6,3) and end point (1,8), run on the algorithm above. Two graphs with the first on be the number of steps required to reach the goal as a function of learning trails and the second one for policy of agents are below. (Figure 1)

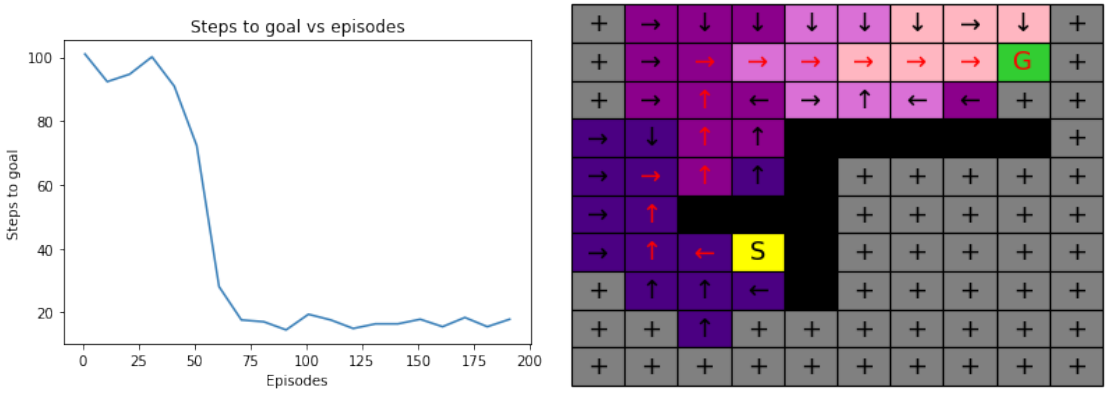


Figure 1: (a). num of steps vs. episodes, (b). policy of agents

Then we keep the start point unchanged and continue with two goal locations, (1,8) and (5,6). The reward will be 10 when we reach either of the goals. The two plots are below (Figure 2).
We can see from the plot that for the second case, there are no obvious paths around (1,8), which is a further one. Also, steps to goal continues to decrease throughout the algorithm. Thus, this algorithm is not prone to explore new paths as long as a minimum is achieved.

## 2.2.2

Now we will try different parameters for the algorithm.

### (a)

In this section, we will change the parameter $\epsilon$ for $\epsilon$-greedy algorithm. Three values will be used: $\epsilon \in \{0.1, 0.5, 0.01\}$. Here is the graph of steps to goal vs. episodes for three parameters. (Figure 3)

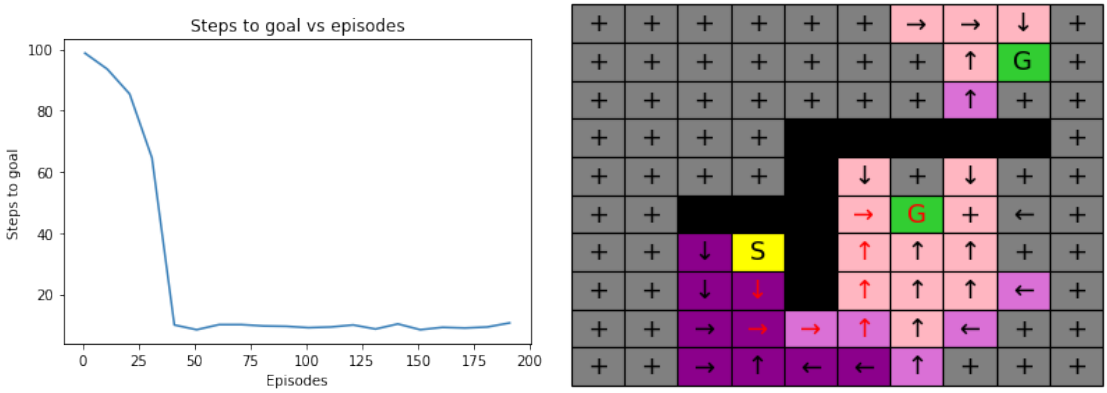


Figure 2: (a). num of steps vs. episodes, (b). policy of agents

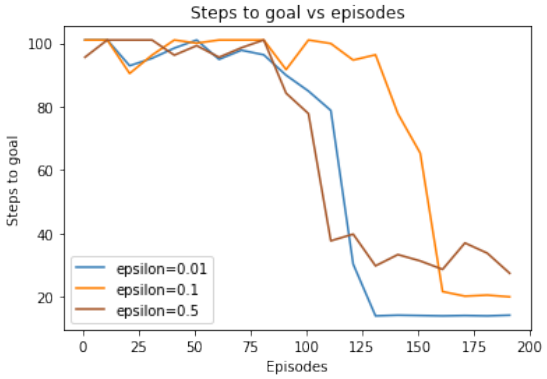


Figure 3: num of steps vs. episodes

From the graph, it is clear that when $\epsilon$ increases, the steps to goal vs. episodes line fluctuates more. It is because a larger $\epsilon$ of greedy exploration indicates higher exploration rates. The benefit from a higher exploration rate is that the algorithm can jump out of a local minimum and find more optimal paths. The cost is that it may converge slower and not necessarily stable. A lower exploration rate can let the algorithm converge stably and fast. However, it may stuck into local optimal and lose the opportunity of pursuing a better result.

## (b)

Now we will use softmax of Q-value to explore. Here we will fix three different $\beta$ values: 1, 3 and 6. To fix the initial value of $\beta$, we will keep the value of $k = 0$. Here is the graph of steps to goal vs. episodes for three parameters. (Figure 4)
Using softmax of Q-values keeps the opportunity for choosing other directions other than the maximum. A larger value of $\beta$ can cause the large ones of softmax probabilities to be larger and small ones to be smaller. So there is a large $\beta$ higher probability for the algorithm to choose the maximum, and thus decrease the ability of exploration.

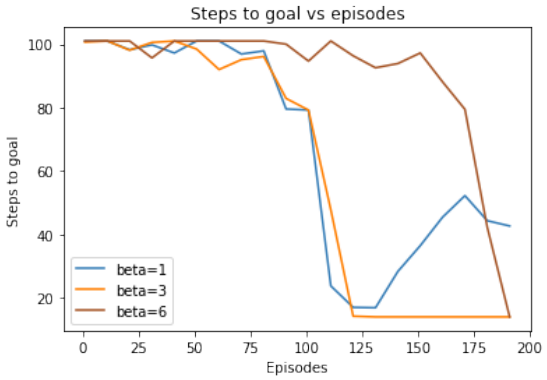


Figure 4: num of steps vs. episodes

## (c)

Now, instead of fixing the value of $\beta$, we will increase the value of $\beta$ as the number of episodes $t$ increase.

$$\beta(t) = \beta_0 e^{kt}$$

Here we will choose four different values of $k$: 0.05, 0.1, 0.25, 0.5. Also keep $\beta_0 = 1.0$. Here is the graph of steps to goal vs. episodes for four parameters. (Figure 5)

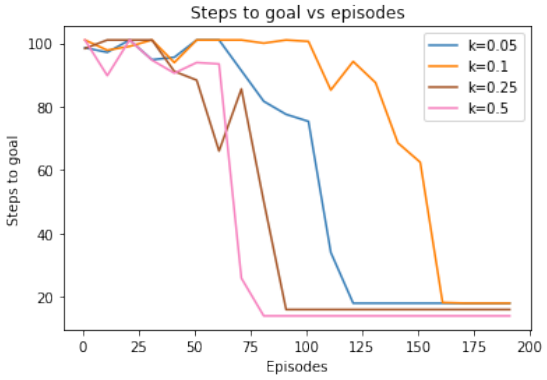


Figure 5: num of steps vs. episodes

When we employ this method, the $\beta$ value will increase as the number of episodes increase, so the certainty of the steps will increase as the algorithm proceeds. The algorithm will be decreasing likely to explore new paths. A larger value of $k$ means that the speed of this process is quicker, thus the algorithm will converge faster.

# 2.2.3

## (a)

In this part, we will make the environment stochastic, which means that the agent will move in the chosen direction only 95% of the time. Otherwise the agent will move randomly. Here is the graph of steps to goal vs. episodes (Figure 6)

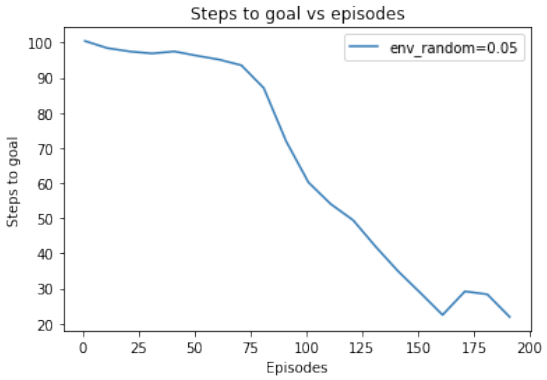


Figure 6: num of steps vs. episodes

It can be inferred from the graph that this algorithm has an ability to explore new paths, because the line fluctuates even if it has reached a minimal value.

## (b)

Now we will change $\alpha$ to be less than one ($\alpha = 0.5$). Also try four different values of probabilities that the environment performs a random action. The values are 0.05, 0.1, 0.25, 0.5. A larger value means the environment becomes more stochastic. Here is the graph of steps to goal vs. episodes for four parameters. (Figure 7)

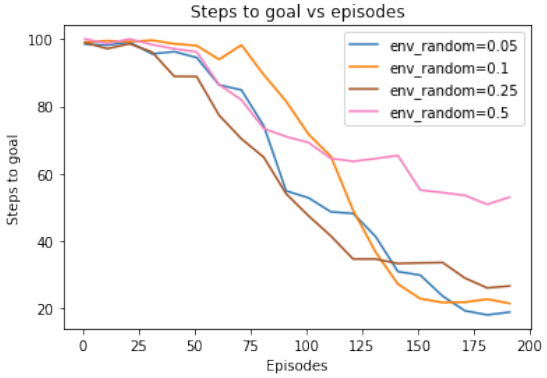


Figure 7: num of steps vs. episodes

It can be inferred from the graph that a smaller number of $p_{rand}$ (less stochastic environment) makes the algorithm to converge faster. In another word, it will take less episodes to find an optimal path.
**The codes begin on the next page.**