EE5904 NEURAL NETWORKS

Project 2: Q-Learning for World Grid Navigation

Name: Yu Shixin

Matriculation Number.: A0195017E

Date:2019/4/21

Task 1:

The goal of this task is to find the optimal way from initial state (1) to reach the goal state (100), with the given reward matrix.

Process:

I will run this algorithm 10 times, for each time, and there are 3000 trials. For each trial, we use Exploitation and Exploration two methods to decide the next step. Different action will get different Q-value. I use the rand function to make a random number, and compare it with the exploration probability ϵ_k to decide to choose which method.

Finally, in each run, l will maximize the Q-value table to find an optimal way. The optimal route shown in Fig.1, and the **total reward = 1869.7448**.

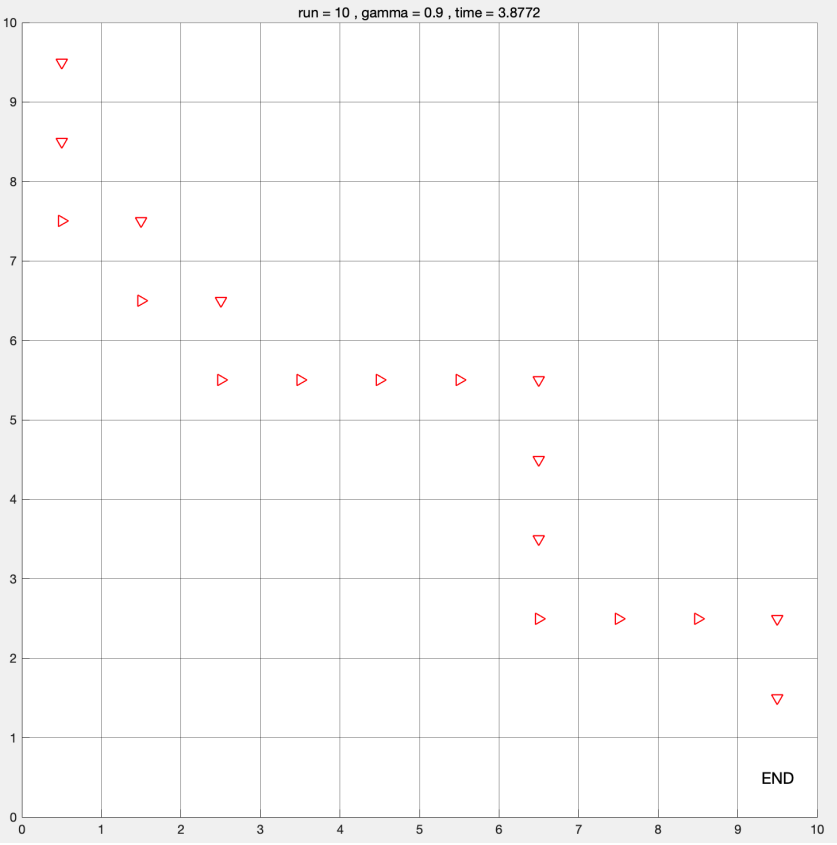

Figure.1 The optimal route

I try different discount factor γ, exploration probability ϵ_k, and learning rate α_k, and the results shown as below:

Table.1 PARAMETER VALUES AND PERFORMANCE OF Q-LEARNING

ϵ_k, α_k	**No. of goal-reached runs**		**Execution time (sec.)**	
	$\gamma = 0.5$	$\gamma = 0.9$	$\gamma = 0.5$	$\gamma = 0.9$
$\frac{1}{k}$	0	0	/	/
$\frac{100}{100+k}$	0	**10**	/	**2.7912**
$\frac{1+\log k}{k}$	0	**1**	/	**3.925**
$\frac{1+5\log k}{k}$	0	**8**	/	**6.15**

According to table.1, we can find that only 3 situations we can get an optimal solution. Among them, $\frac{1+\log k}{k}$ can not get a goal-reached route per 10 runs. Most of time, it can not find the optimal solution. Thus, I decide to discuss the impact of parameter learning rate α_k , discount factor γ and exploration probability ϵ_k.

1. For learning rate α_k. The reason why $\frac{100}{100+k}$, $\frac{1+\log k}{k}$ and $\frac{1+5\log k}{k}$ can get result is that all of them guarantee a large value more than threshold (0.005), when the step is too large. So the far step also can provide an appropriate figure to update the Q-value. For $\frac{1}{k}$, with the increase of K, their value approach 0 very fast. Thus, it provides very limited help for far step.

2. For discount factor γ. The larger the discount factor is, the more effect the next step has. It help us to find the larger reward given by next step.

3. For exploration probability ϵ_k. The larger exploration probability is, the more randomness the choice of Q-value has.

In conclusion, for learning rate, we should choose the function who can keep its value stable, or in very little decrease. For discount factor, it can be chose a larger one. As for exploration probability, it must be a proper value, not too large and not too small, and I will discuss about how to choose exploration probability more in task2.

Task 2:

In order to choose a learning rate, discount rate and exploration probability wisely, I try different value of these three parameters.

Table.2 Gamma=0.9

Learning rate	Task1.mat (sec)	Own.mat (sec)
$\frac{200}{200+k}$	2.3552(10)	1.4771(6)
$\frac{300}{300+k}$	1.4396(10)	1.3369(6)
$\frac{400}{400+k}$	1.2844(10)	1.395(6)
$\frac{500}{500+k}$	1.0703(10)	0.84855(9)
$\frac{1000}{1000+k}$	0.91131(10)	0.71513(9)
$\frac{2000}{2000+k}$	0.63943(10)	0.67877(9)
$\frac{2500}{2500+k}$	**0.62689(10)**	**0.47695(10)**
$\frac{3000}{3000+k}$	0.74668(10)	0.48328(10)

Table.3 Learning rate = $\frac{2500}{2500+k}$

Discount factor	Task1.mat (sec)	Own.mat (sec)
0.6	/	/
0.7	0.72735(10)	/
0.8	0.70327(10)	0.51761(9)
0.9	0.65568(10)	0.43416(9)
0. 95	**0.65174(10)**	**0.40181(10)**
0. 99	/	/

Based on Gamma=0.95 & Learning rate = $\frac{2500}{2500+k}$, I change the different exploration probability, which is not same with the learning rate.

Table.4 Gamma=0.95 & Learning rate = $\frac{2500}{2500+k}$

Exploration probability	Task1.mat (sec)	Own.mat (sec)
0.4	4.2334(10)	0.27053(1)
0. 5	0.43493(10)	0.22206(5)
0. 6	**0.28136(10)**	**0.24848(10)**
0. 7	0.31143(10)	0.29156(10)

According to table.2 & table.3, when Gamma=0.95 & Learning rate = $\frac{2500}{2500+k}$, we can get best performance. Then I try to change exploration probability to make it different with learning rate. According to table.4, we can get best performance, when exploration probability=0.6. Thus, I choose learning rate function ($\frac{2500}{2500+k}$), discount factor (0.95) and exploration probability=0.6. This .m files is saved in the RL_main.m.