Sharif University of Technology
Electrical Engineering School

Advanced Neuroscience HW6

REINFORCEMENT LEARNING
LEARNING THE WATER MAZE

Armin Panjehpour
arminpp1379@gmail.com

Supervisor(s): Dr. Ghazizadeh
Sharif University, Tehran, Iran

18/05/2022

The implementation is Model based here (Actor Critic Model):

Part.1 - Path Before and After Training:

Below, you can find the map of the rat path before and after training. As you can see, with training, rat finds the target easily and reach it with low number of steps:

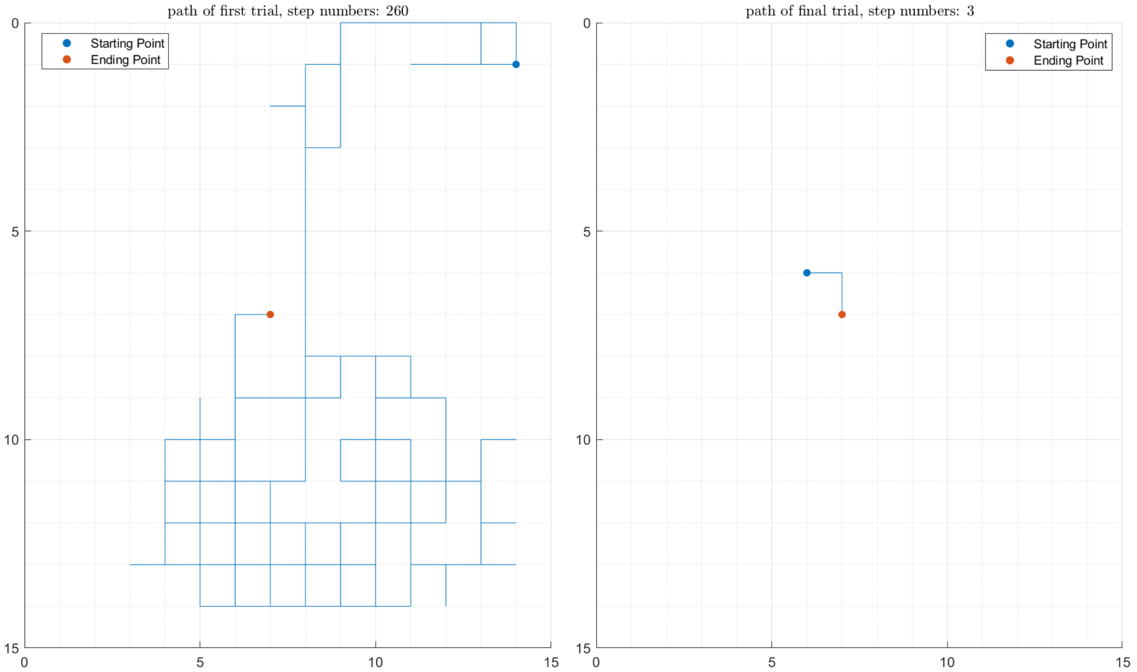

Figure 1: Before & After of Training Rat Path

To make sure this result is not due to randomization, we plot number of steps taken in each 400 trial and plot them:

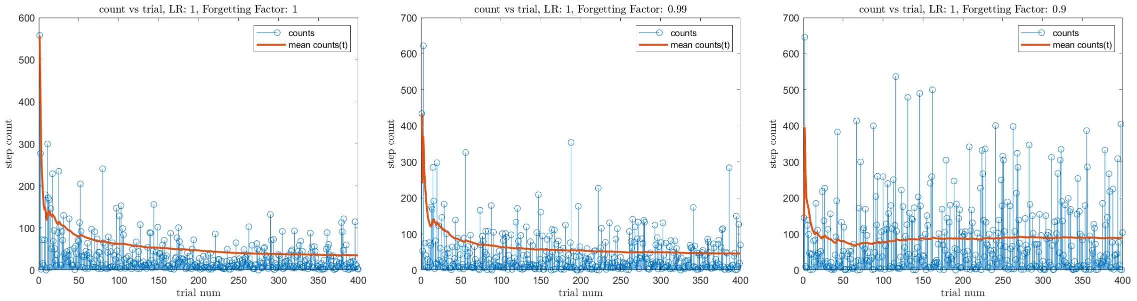

Figure 2: Steps Num. Vs Trial Num. for Three Forgetting Factors of 1, 0.99, 0.9

As you can see, as the time passes, rat reaches the target with lower number of steps. As we increase to forgetting factor from 1 to 0.99, the number of steps to reach target will increase since the rat forget about its previous findings. As you can see, the mean behavior of these plots is descending.

In order to eliminate the noisy behavior of these plots ,we could run each trial for, for example 100 iterations, and get the mean of them as the number of steps needed to reach the target for each trial but these plots are fine enough and tell the point.

Part.1.1 - Demo of Training

You can find the demo of the rat, training for two forgetting factors of 1 and 0.99 with a constant learning rate of 1 below:

Forgetting Factor = 1
Forgetting Factor = 0.99

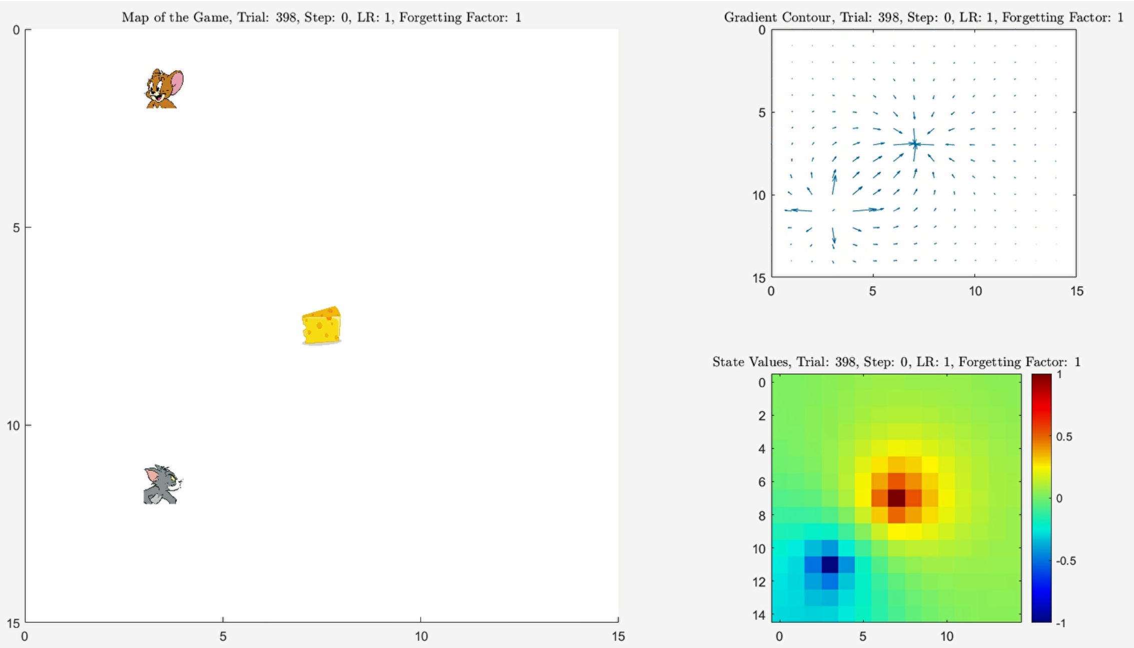

Figure 3: Snapshot of the Demo in one of the end Trials

Part.2 - Gradient & Contour Plots of Learned Values

Below you can find the plots for different learning rates, different forgetting factors, different discount rates and different number of trials.:

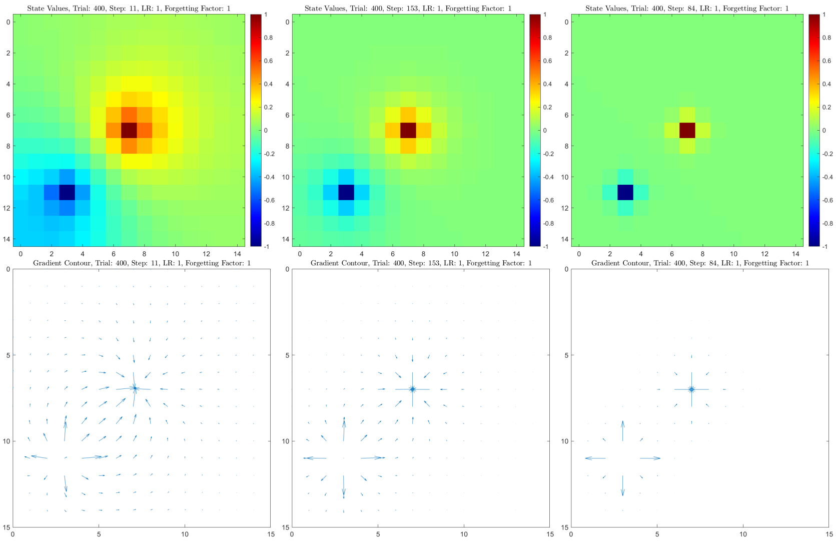

Figure 4: Learned Values Gradient & Contour Plot for Discount Rates of 1, 0.9, 0.5 for 400 Trials

As you can see in Fig.4 , by having more discount rate, the rat will learn less at the end and it will reach the target with more steps.

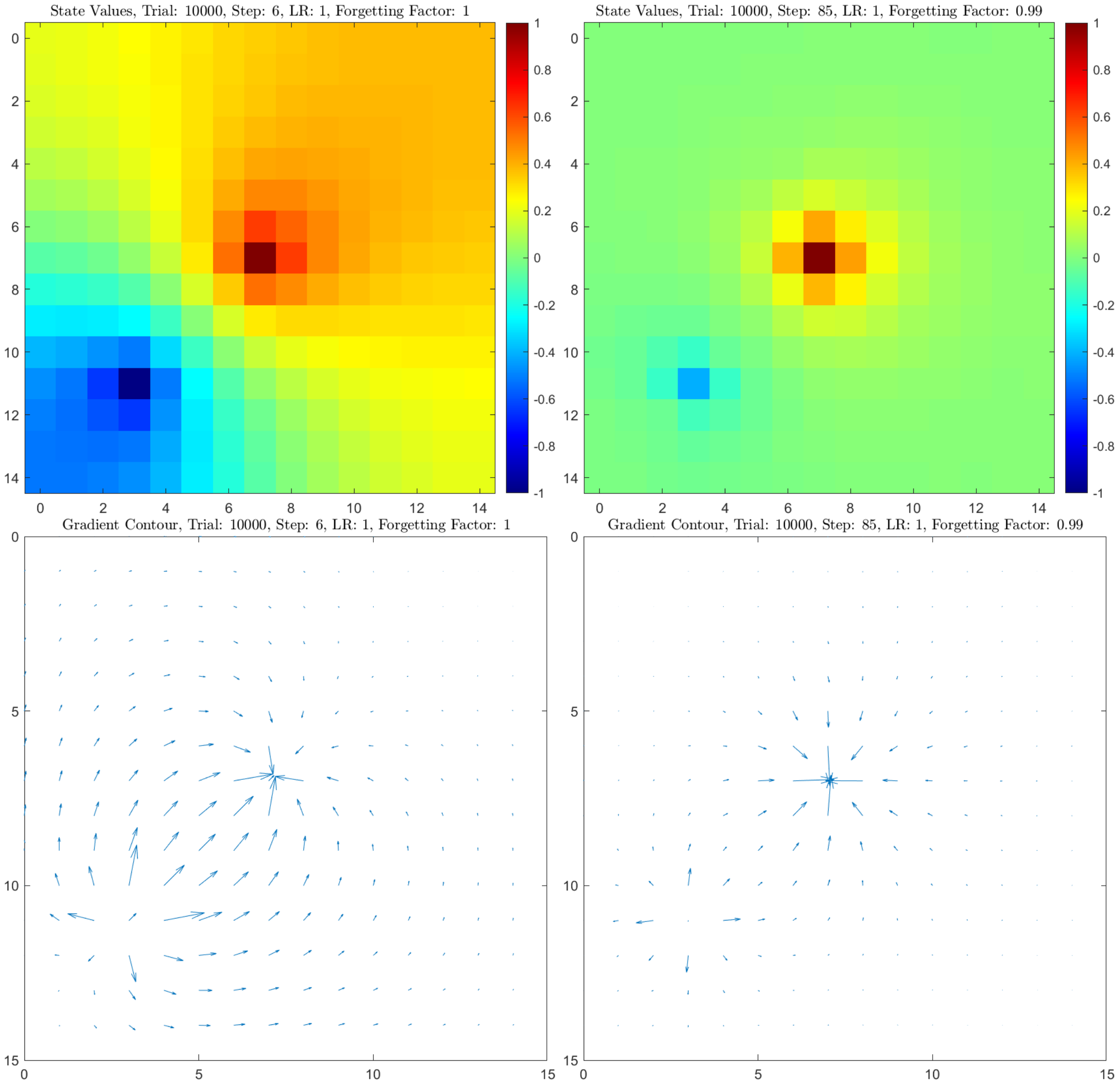

Figure 5: Learned Values Gradient & Contour Plot for Forgetting Factor of 1, 0.99 for 10000 Trials

As you can see in Fig.5 , by increasing the forgetting factor (right figure), the rat will lean slower and in a certain number of trails it will learn less than a rat without forgetting factor (=1). Another point of this plot is that the rat learn all values of map much more in 10000 trails comparing to 400 trials and it will be more sure on its decisions.

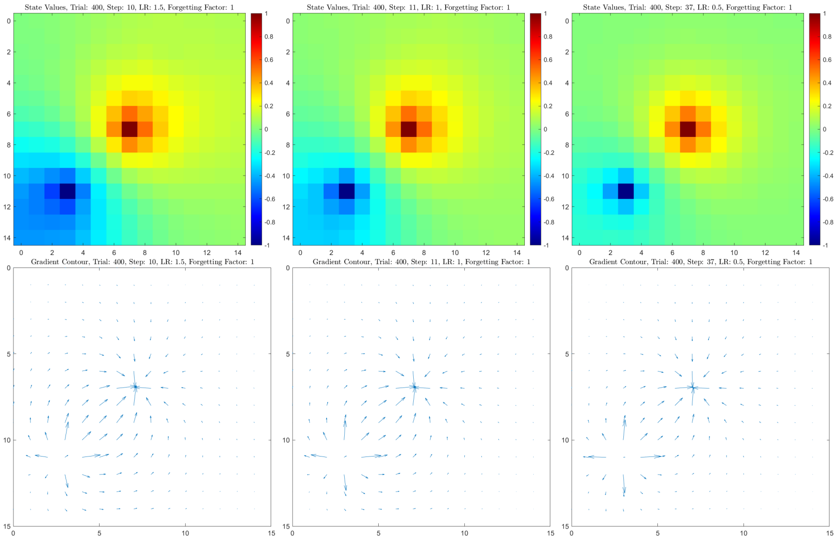

Figure 6: Learned Values Gradient & Contour Plot for Learning Rates of 1.5, 1, 0.5 for 400 Trials

As you can see in Fig.6 , the more the learning rate, the more the rat learns about the map.

Part.1.3 - Learning Rate & Discount Rate Effect on Learning Speed

Here we run the training for 400 trails each trial 100 iterations and average on all 100 iterations for each trial. Then we take the mean of the last 50 trials as a measure of convergence and plot this heat map:

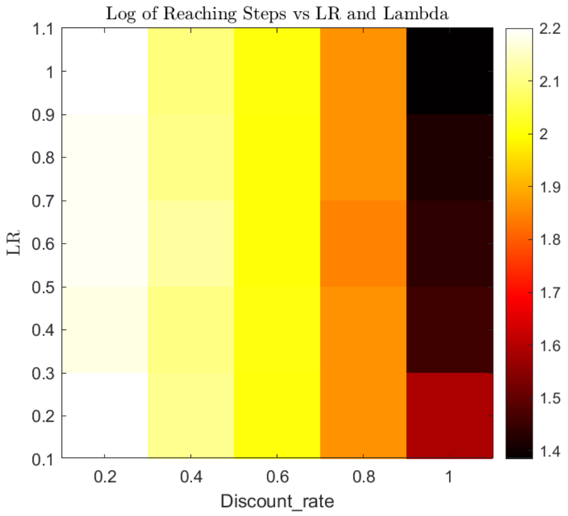

Figure 7: LR & Discount Rate Effect on Learning Speed

As you can see in high discount rates, as the learning rate decreases, the number of steps taken to reach target increases since the rat learns less. But in the lower values of discount rate, the descending pattern of learning rate is less since the hole learning becomes random but in generally, by decreasing the discount rate and learning rate, the number of steps to reach the target increases:

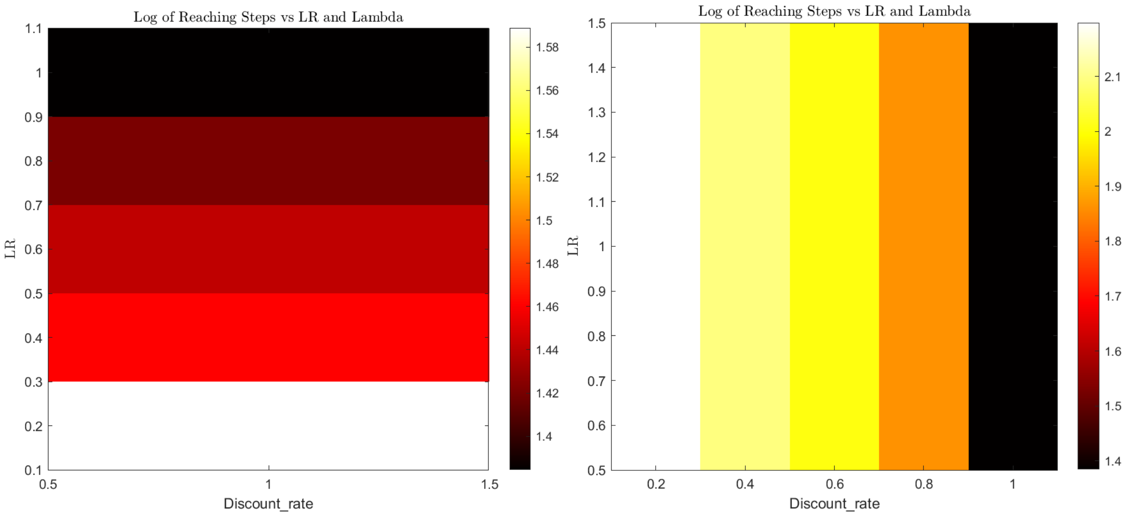

Figure 8: LR & Discount Rate Effect on Learning Speed

Part.1.4 - Two Targets with Two Different Values

Here we put one target with a value of 1 and another target with value 0.4 and here‘s the demo: Two Target

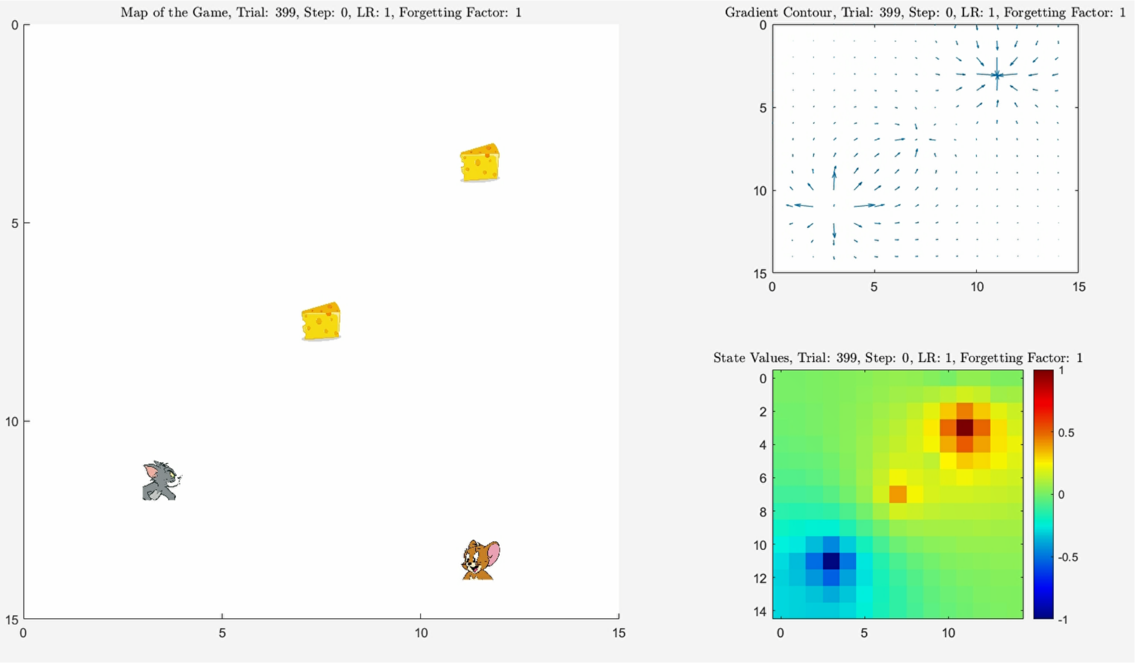

Figure 9: Snapshot of the Demo in one of the end Trials, Two Target

As we expected, the gradient vector lengths and the values around the target with the bigger value is bigger than the other target. So the rat prefer the top right target but also like the other target too.

Part.1.5 - TD Lambda

Here, we update the value of states before a state with non-zero value which we have gone to with a discount rate of 0.3. Since here we don't calculate the expected value of values, the final map is not beautiful as like in previous results but it's true. The point of this implementation is that the rat learns the maze so much faster. Find the link below and watch the demo:

TD Lambda

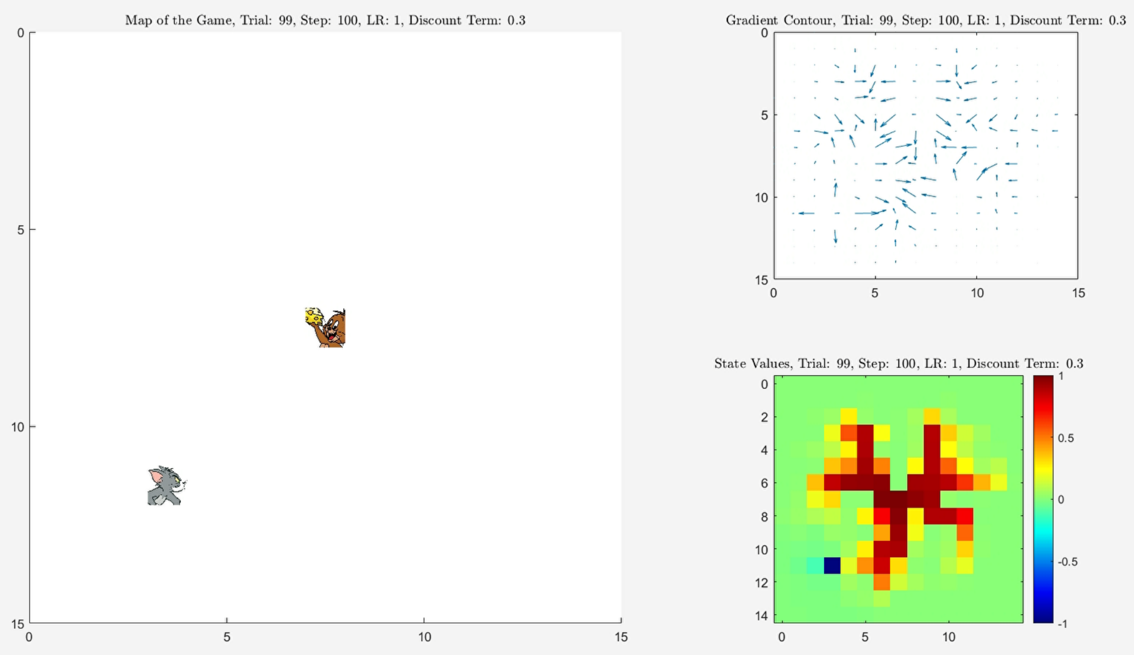

Figure 10: Snapshot of the Demo in one of the end Trials, TD Lambda, Discount Rate = 0.3

In every state with non zero value that we go, the previous moves will have an increase in their values.