Computer Assignment Reports Reinforcement Learning

Andreas Stasinakis & Mim Kemal Tekin

March 6, 2019

Contents

Contributors: Mim Kemal Tekin(mimte666) & Andreas Stasinakis(andst745)

In order to pass the assignment you will need to answer the following questions and upload the document to LISAM. You will also need to upload all code in .m-file format. We will correct the reports continuously so feel free to send them as soon as possible. If you meet the deadline you will have the lab part of the course reported in LADOK together with the exam. If not, you'll get the lab part reported during the re-exam period.

Question 2

Define a learning rule (equation) for the Q-function and describe how it works. (Theory, see lectures/classes)

The learning rule(equation) for the Q- function defined before is:

$$\hat{Q}(s_k, a_j) \leftarrow (1-\eta)\hat{Q}(s_k, a_j) + \eta(r + \gamma\hat{V}(s_{k+1})) = (1-\eta)\hat{Q}(s_k, a_j) + \eta(r + \gamma max_a\hat{Q}(s_{k+1}, \alpha)$$

, where $\hat{Q}(s_k, a_j)$ is the previous estimate, η is the learning rate, r is the reward from moving to state k to state $k+1$ and a_j the action we choose.

In general, the updated $\hat{Q}(s_k, a_j)$ is the sum of two proportions. The first one is the previous estimator multiplied by 1 - learning rate. The second one is again the sum of two quantities. The reward and the product of the discount factor with the estimate of future optimal value. The last summation is multiplied by the learning rate.

Question 3

3. Briefly describe your implementation, especially how you hinder the robot from exiting through the borders of a world.

In this task we implement reinforcement learng in order to train a robot to find a specific target in different worlds. The procedure is simple but powerfull. We initialize the Q function(all the formulas we need are presented in previous tasks) with zero values. We also want to put some borders to the robot. More specific, you do not want the robot to up if it is in the top of the world. In order to achieve that we put -inf all the values, given an action, we do not the robot to choose. For example if the robot is in the las left row of the world, we put the reward for going left -inf. As a result, the robot will never choose to go in this direction. One can say that instead of an initial Q function of zeros can be used random numbers. This is also correct, but in this case we have to change the reward for the terminal point equal to zero. We can skip that, starting with zeros in the initial Q function.

For each iterations(episode) we initialize the robot in a random position. The robot tries to find the goal for a given number of steps. For each step, the robot chooses a action given the Q function, its current possition and some probabilities. After going to the next state, we check if this state is valid and it is not the robot will find another valid action. Finally, we update the Q function, using the formula in previous task. The procedure stops if we find the terminal or after the total number of steps we use.

Question 4

4. Describe World 1. What is the goal of the reinforcement learning in this world? What parameters did you use to solve this world? Plot the policy and the V-function.

The first world is one static world and probably the easiest one to find the target. The goal of the the reinforcement learning in general is to find the cirle target. We should also train the robot in order to avoid the obstacle. For example when the robot start inside this purple area should leave it directly because the feedback there is negative. We can confirm that from the policy plot. We solve this world using the parameters below:

Discount factor : 0.9

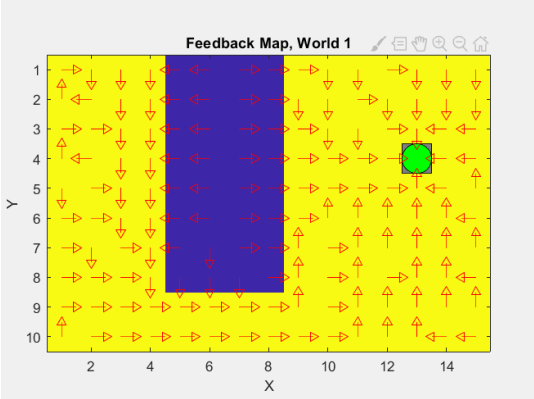
Learning rate : 0.3

Maximum steps : 300

Total episodes : 1000

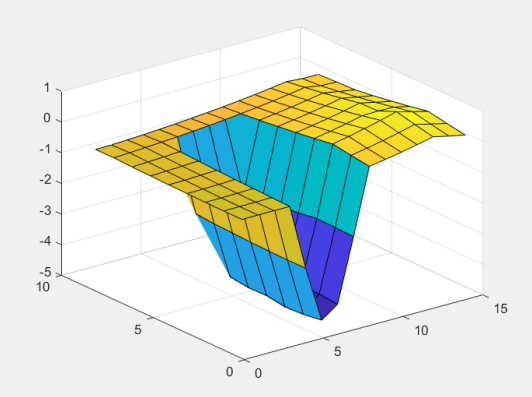
Exploration factor : 1 - number of current episode / Total episodes.

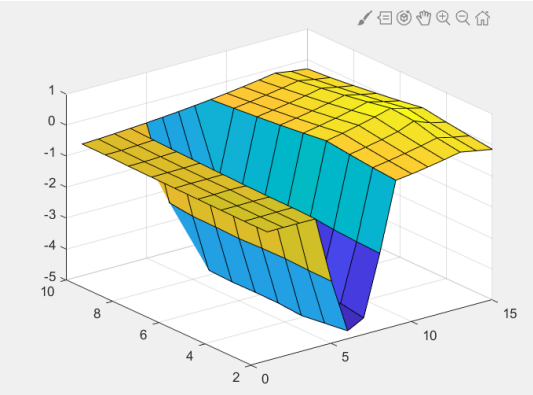
Plot for policy:



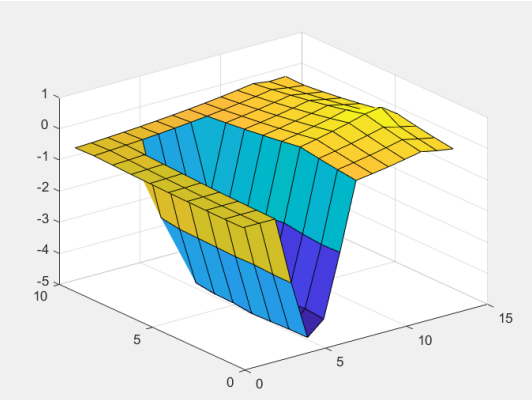
3D - Plots for Q function:

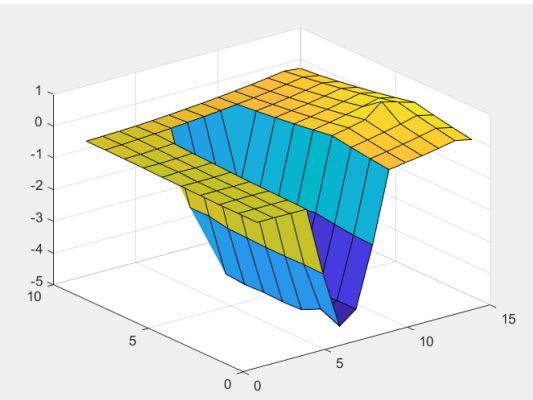
For action 1 and 2:





For action 3 and 4:





From the policy we can see that we are reaching the target for every starting point and as mentioned before when the robot starts inside the purple area, it tries to leave this area immediatly.

Question 5

Describe World 2. What is the goal of the reinforcement learning in this world? What parameters did you use to solve this world? Plot the policy and the V-function.

The second world is similar to the first one, but a randomness also occurs. That is also clear from the plots, because despite the fact that in the final policy there is no obstacle, the robot does not go directly to the target but sometimes it follows different path.

We solve this world using the parameters below:

Discount factor : 0.9

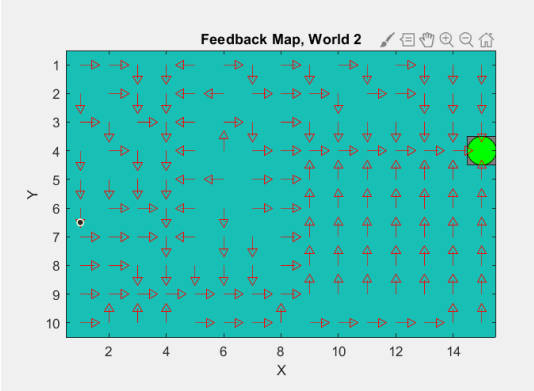
Learning rate : 0.3

Maximum steps : 500

Total episodes : 1500

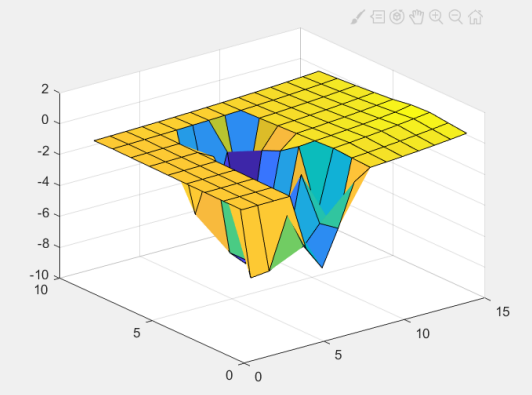
Exploration factor : 1 - number of current episode / Total episodes.

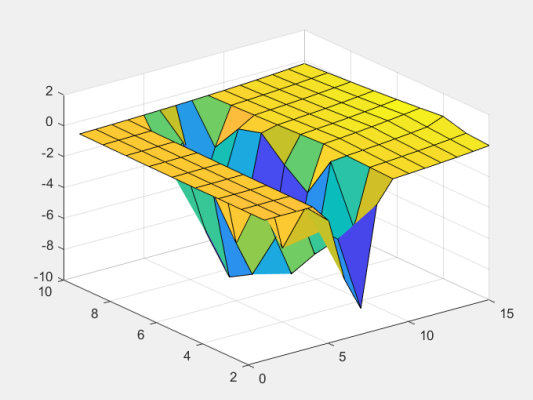
Plot for policy:



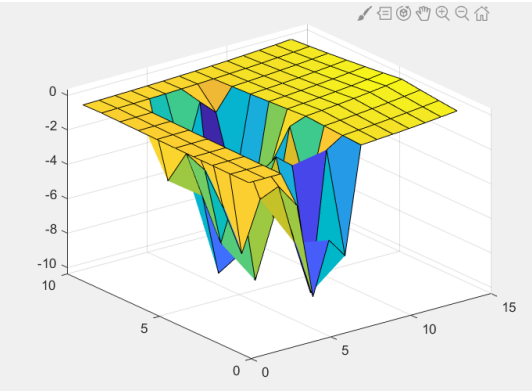
3D - Plots for Q function:

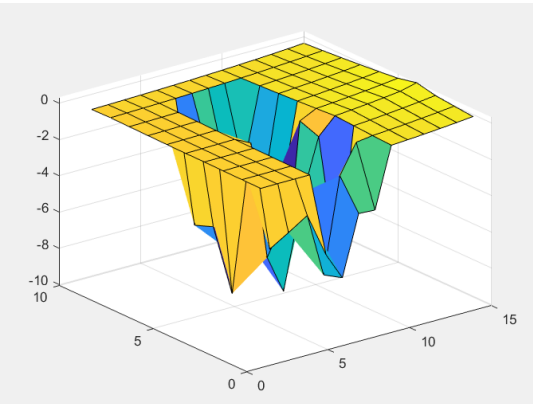
For action 1 and 2:





For action 3 and 4:





Question 6

Describe World 3. What is the goal of the reinforcement learning in this world? What parameters did you use to solve this world? Plot the policy and the V-function.

The third world is more tricky because we have two "forbitten areas" and especially bettween them, only one move can be done. Training parameters:

Discount factor : 0.9

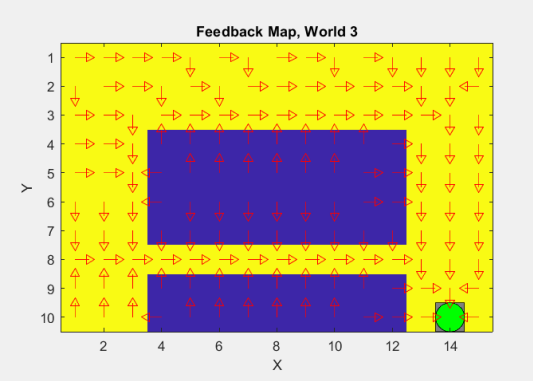
Learning rate : 0.3

Maximum steps : 500

Total episodes : 1500

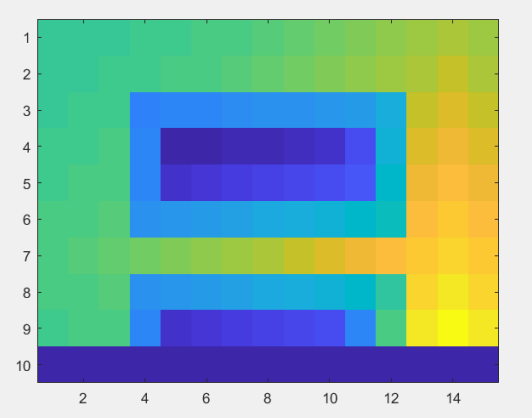
Exploration factor : 1 - number of current episode / Total episodes.

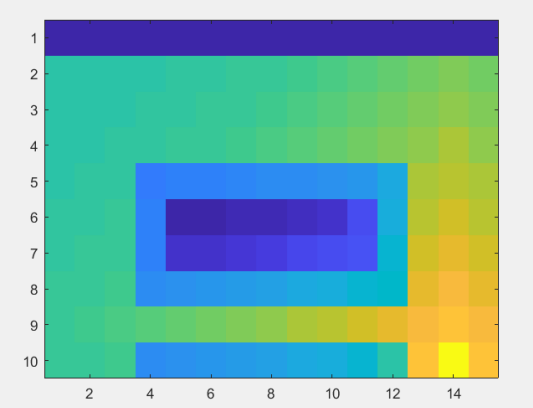
Plot for policy:



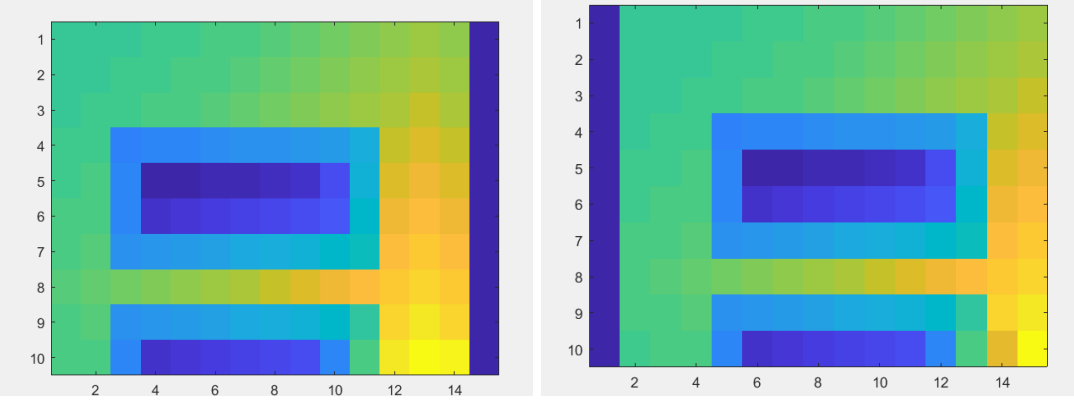
3D - Plots for Q function:

For action 1 and 2:





For action 3 and 4:



Question 7

Describe World 4. What is the goal of the reinforcement learning in this world? How is this world different from world 3, and why can this be solved using reinforcement learning? What parameters did you use to solve this world? Plot the policy and the V-function.

The forth world is the most complicated between the first 4 worlds and the one we took us time to train the robot. Again here the goal is to reach the target and do not go inside the purple areas. The difference bettween world 3 and 4 is that the goal is in the opposite side of the world and that the starting point is changing. In world number 3 starts in the same line as the shortest path, so the robot directly goes this way. In contrary, the robot in world 4 starts in other position and the path it follows is going up and after that left and down. This is happening because going up does not have any forbitten area, while going left has.

Training parameters:

Discount factor : 0.9

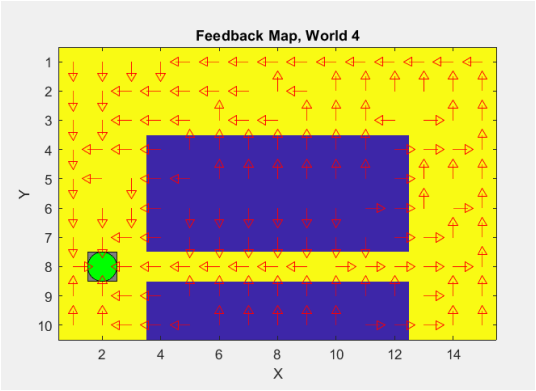
Learning rate : 0.1

Maximum steps : 500

Total episodes : 5000

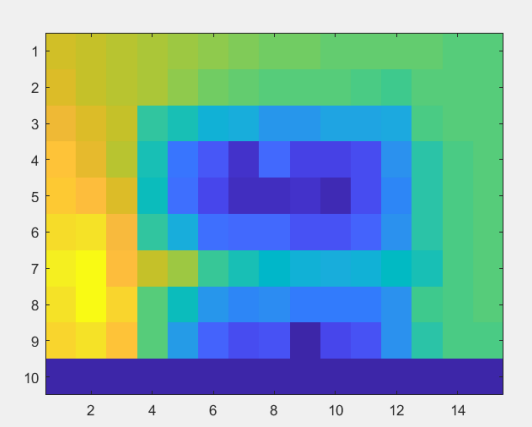
Exploration factor : 1 - number of current episode / Total episodes.

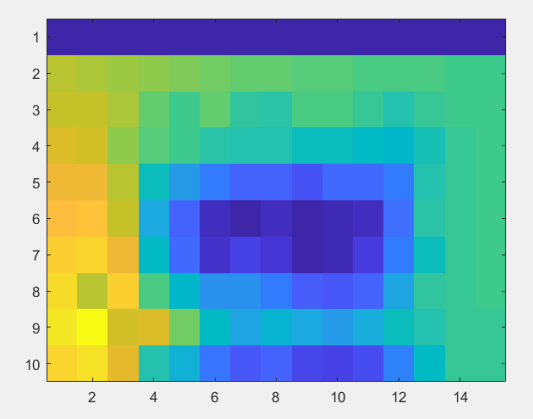
Plot for policy:



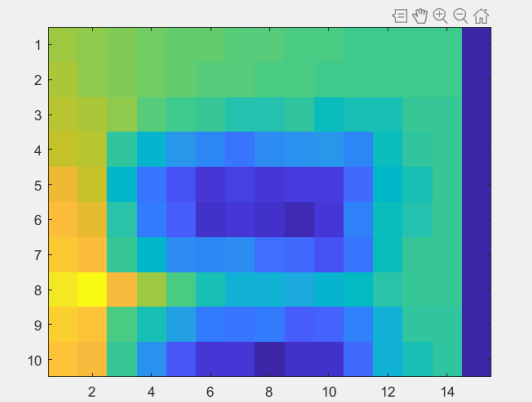
3D - Plots for Q function:

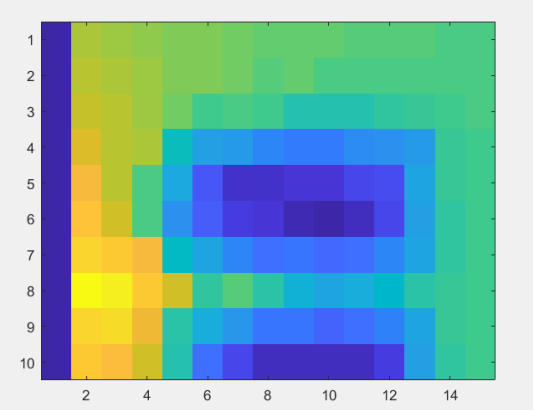
For action 1 and 2:





For action 3 and 4:





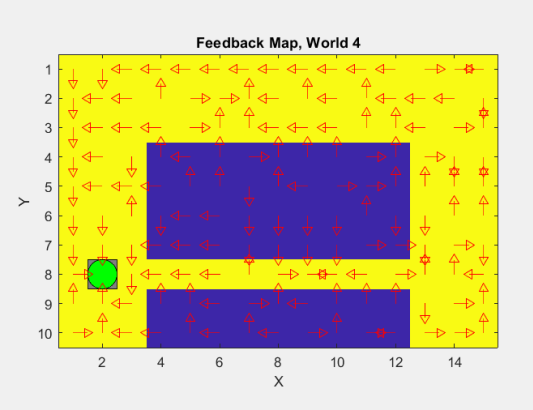
Question 8

8. Explain how the learning rate α influences the policy and V-function in each world. Use figures to make your point.

The learning parameter is really important factor for the reinforcement learning. It controls of how much of the new information will overlap the old one. More specific, a value close to 0 uses mostly the previous information, while a value close to 1 makes the model replace the oldest information with the new one.

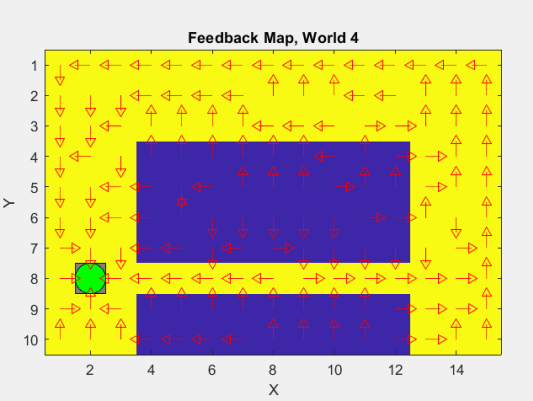
In the first 3 world the optimal learning rate is close to 0.3. The worlds are not that complicated so we can achieve our goal with this value. For the last world though we have to try many different combinations.

For $\alpha = 0.7$ and the other parameters same as before, we have the policy below:



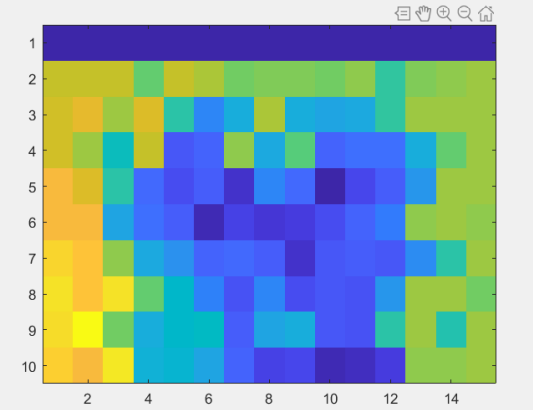
From the policy, we can observe that we have many infinite loops where the robot can not leave. This is happening because the robot does not pay attention in the previous information. Therefore, every time is taking a decision without using this information. In the end we need a lower value in order for the algorithm to converge.

For $\alpha = 0.3$:



For $\alpha = 0.3$, we have better results but stil the algorithm has not converged. That is the reason we also tried for 0.1 and we finaly keep that as the optimal learning parameter.

We also present the Q function for action 4. It seems like using a high learning parameter,in this case 0.7 again, does not gives us a clear pattern and a good optimal policy.



Question 9

9. Explain how the discount factor ?? influences the policy and V-function in each world. Use figures to make your point.

The discount factor γ determines the importance of the future rewards. More specific,

question 12

12. Can you think of any application where reinforcement learning could be of practical use? A hint is to use the Internet.

Reinforcement learning can have many applications in real world. For example, as we see during the lecture, we can use reinforcement learning in Robotics. One of the most famous achievements until now in this field is a robot which can learn policies to map raw video images to robot's actions. It is also used in popular games, such as Go. More specific, a robot trained with countless human games as a result it achieved amazing performance. One more application of reinforcement learning is the traffic light control.