TBMI26 – Computer Assignment Reports Reinforcement Learning

Deadline – March 14 2021

Shwetha Vandagadde Chandramouly
Suhani Ariga

In order to pass the assignment you will need to answer the following questions and upload the document to LISAM. Please upload the document in PDF format. **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.

1. **Define the V- and Q-function given an optimal policy. Use equations and describe what they represent. (See lectures/classes)**

   V-function:

   $$V(s_t) = \sum_{k=0}^{\infty} \gamma^k r_{t+k}$$

   Where is the optimal policy and is a discount factor. A function V(s) of the state that tells us the value of being in the state given a policy.

   Q-function:

   $$Q(s_k, a) = r(s_k, a) + \gamma V(s_{k+1})$$

   $$V(s_{k+1}) = \max_a Q(s, a)$$

   The Q-function describes the expected future reward of doing action a in state s and then following the optimal policy.

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

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

   Where is the learning rate, is the reward from the current state and is a discount factor. The updated Q function is proportional to the first term of the Q function in the current state includes the previous estimate and the second term of the Q function includes a better estimate.

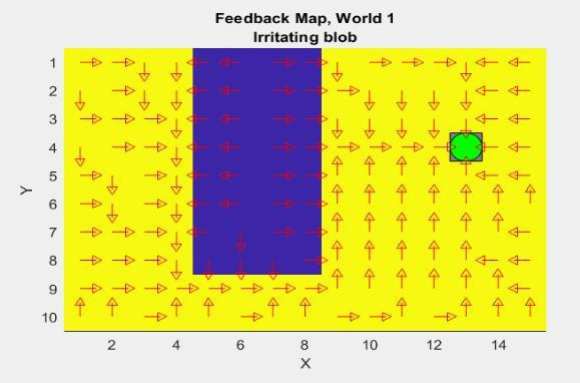
3. **Briefly describe your implementation, especially how you hinder the robot from exiting through the borders of a world.**

We initialized the world, Q-table, and hyperparameters. Q table is to store values for each action in every state and is updated using learning rule for the Q-function. Repeatedly, Q-table has been updated by choosing an action and evaluating the reward of the new state until the robot reaches its goal. The optimal policy can be found by taking the maximum Q value after training along with the 3rd dimension (action). Robot is tested by letting it to use the optimal policy.

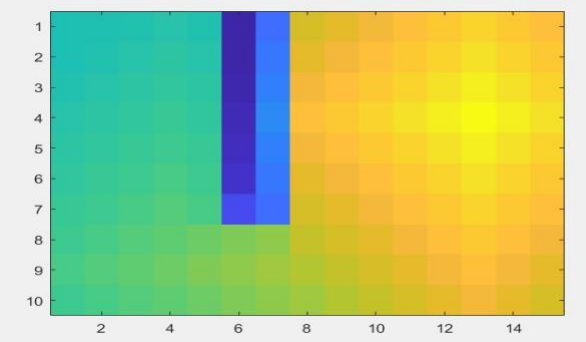
To hinder the robot from exiting through the borders of a world, we have initiated the action that would make robot to exit is set to -inf.

The aim for robot is to reach the goal (i.e green circle). The robot is given negative reward for every step in yellow/blue ground and once robot reaches the goal it stops getting negative rewards.

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.



Policy of World 1

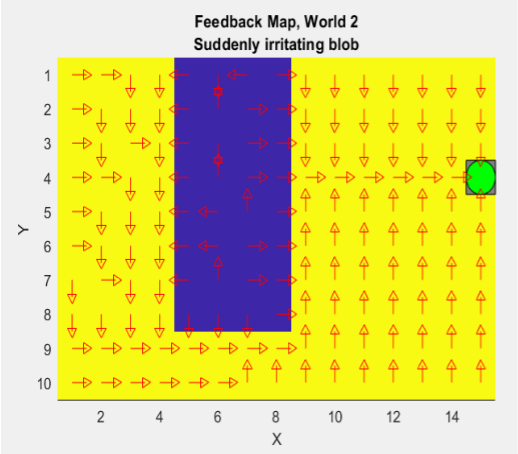


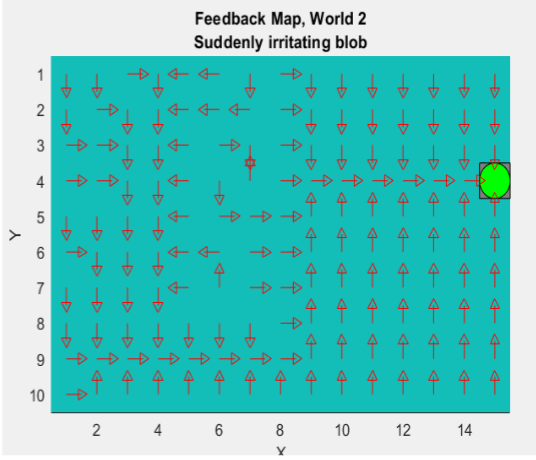
Value of World 1

World 1 is a static world and the starting position of robot is random. The goal of the reinforcement learning in this world is to take the shortest path avoiding blue ground which has negative reward and reach the goal with highest reward. The parameters used are:

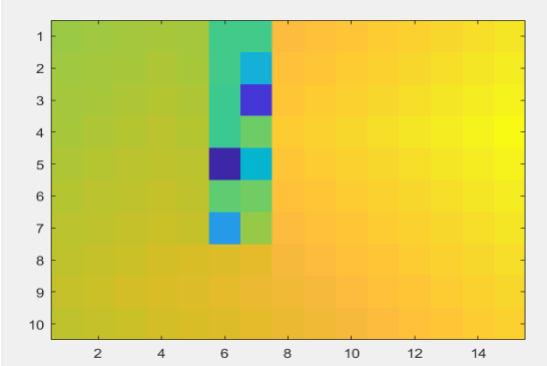
episodes = 1000;
epislon = 1;
learning_rate = 0.4;
discount = 0.95;

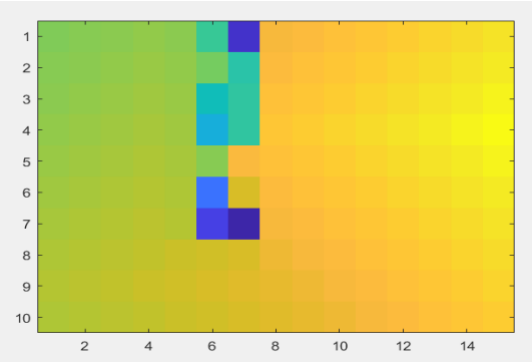
5. **Describe World 2. What is the goal of the reinforcement learning in this world? This world has a hidden trick. Describe the trick and why this can be solved with reinforcement learning. What parameters did you use to solve this world? Plot the policy and the V-function.**





Policy of World 2



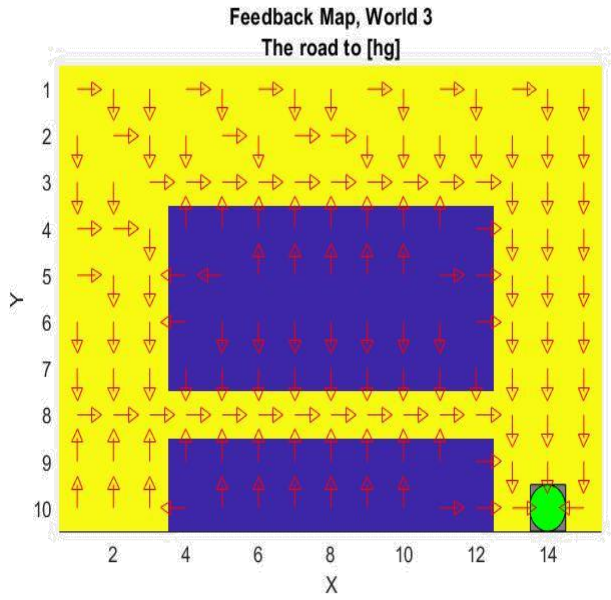


Value of World 2

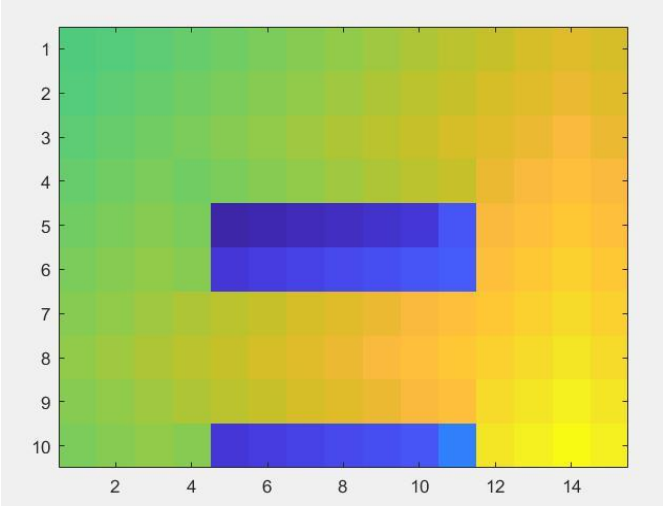
The world 2 is same as world 1, except sometimes there is a blob as shown in above plot where reward inside the blob is lower than usual. Hence, it utilises the policy which avoids passing through the blob. The starting position of the robot is random.
The goal of the reinforcement learning in this world is to reach the goal with highest reward.
The parameters used for both policies are:

episodes = 2000;
epislon = 1;
learning_rate = 0.4;
discount = 0.95;

6. **Describe World 3. What is the goal of the reinforcement learning in this world? Is it possible to get a good policy from every state in this world, and if so how? What parameters did you use to solve this world? Plot the policy and the V-function.**



Policy of World 3



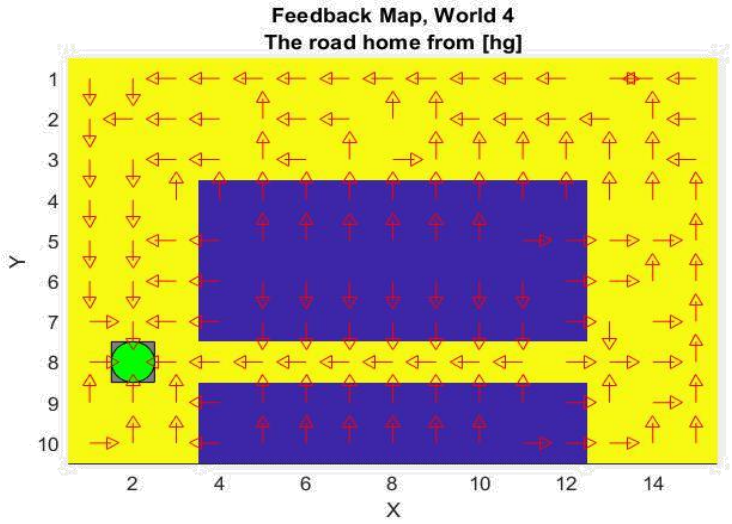
Value of World 3

The world 3 is static where both position of robot and goal is fixed.

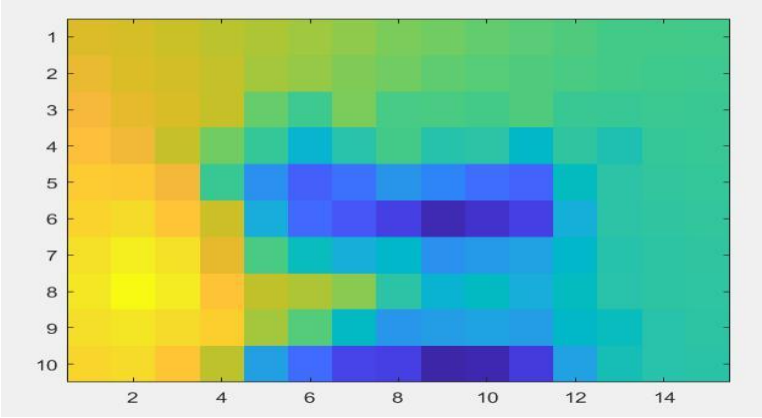
The goal of the reinforcement learning in this world is to take the shortest path avoiding blue ground which has negative reward and reach the goal with highest reward. The parameters used are:

episodes = 2000;
epislon = 1;
learning_rate = 0.4;
discount = 0.95;

7. **Describe World 4. What is the goal of the reinforcement learning in this world? This world has a hidden trick. How is it 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.**



Policy of World 4



Value of World 4

In World 4 , both position of robot and goal is fixed. As we can see in the value of world 4, it avoids the narrow path between the blobs which has low reward and also due to the risk of randomness causing robot to step to the sides. It uses longer but more secure path by following the edges of the world where random mistakes does not have big consequences. In world 4, Q learning method can find optimal policy in stochastic situations.

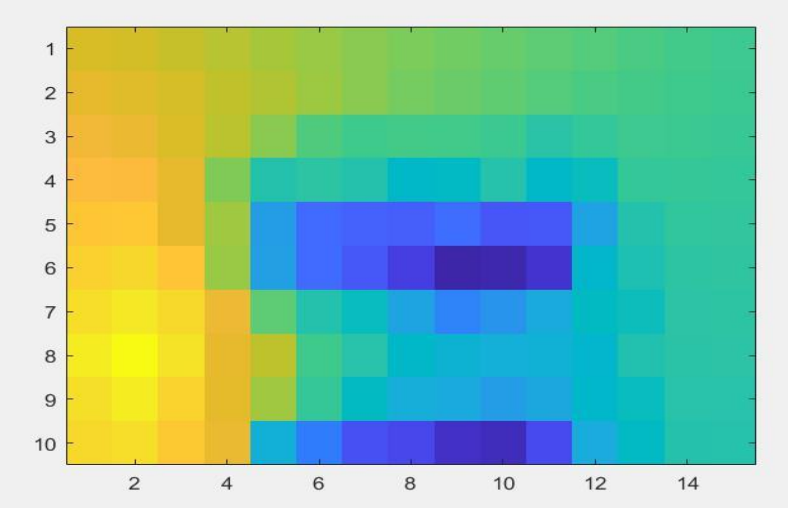
The goal of the reinforcement learning in this world is to take the shortest path avoiding blue ground which has negative reward and reach the goal with highest reward. The parameters used are:

episodes = 2000;
epislon = 1.0;
learning_rate = 0.2;
discount = 0.95;

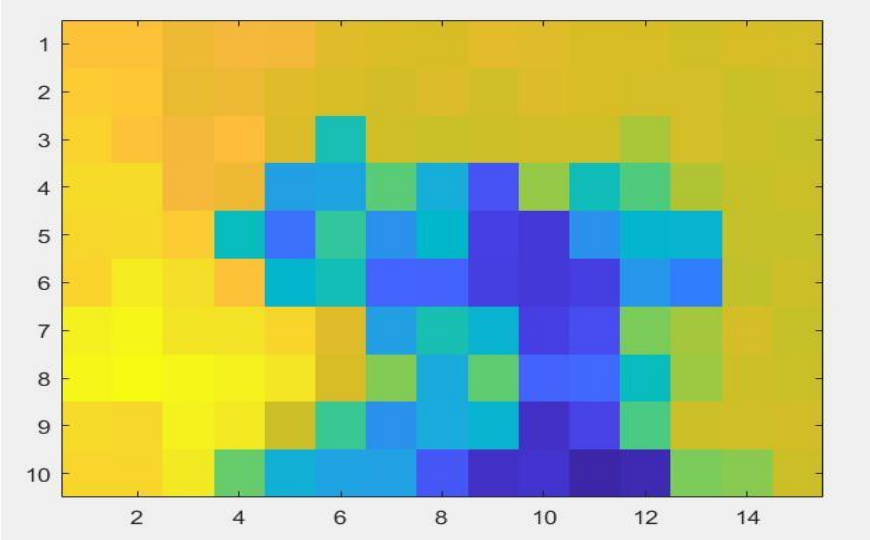
8. **Explain how the learning rate α influences the policy and V-function. Use figures to make your point.**

A learning rate close to 0 puts more emphasis on learned experience and a value close to 1 will prioritize with new information. (Borga, 2021)

World 4
Learning rate = 0.05



Learning rate = 0.9

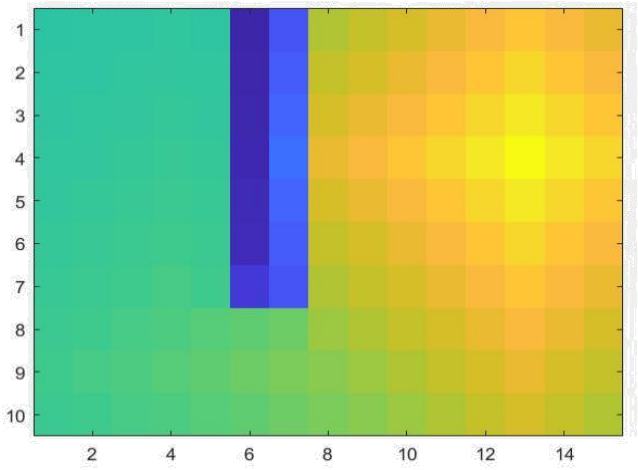


In the plots, we can see the variations of the V function with different learning rates. A low learning rate put emphasis on old Q value which takes a longer time to learn a good policy. But a very high learning rate will put emphasis on new values making Q values fluctuate. For a static world like world 1, a higher learning rate is optimal since action will always be same. But in the random world like world 4, smaller learning rate is preferred. Then the robot will consider what has worked well in the past to avoid negative rewards rather then only focusing on the quickest path.

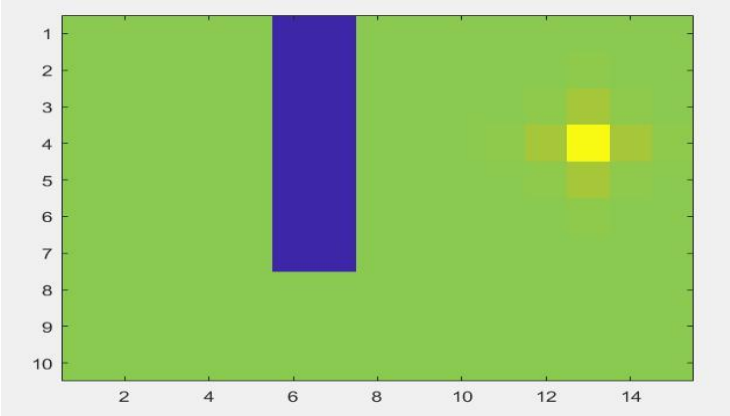
9. **Explain how the discount factor γ influences the policy and V-function. Use figures to make your point.**

γ is a discount factor that maximize the short-term rewards if it is close to 0 or maximize the long-term rewards if it is close to 1. (Borga, 2021)

World 1 Discount factor = 0.9



Discount factor = 0.1

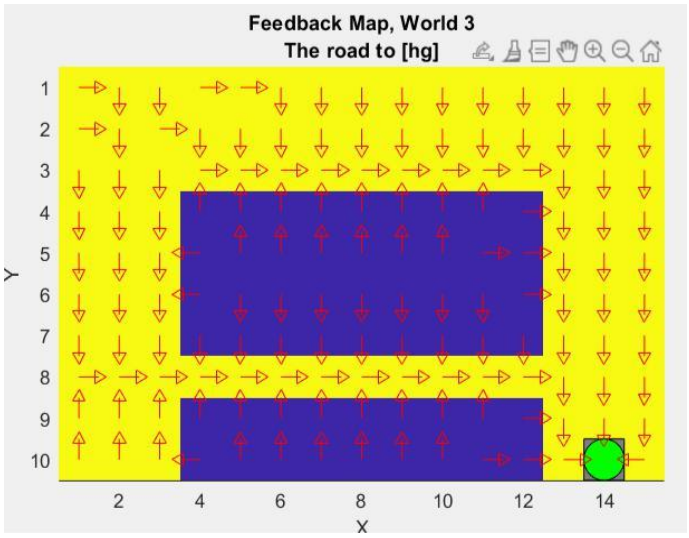


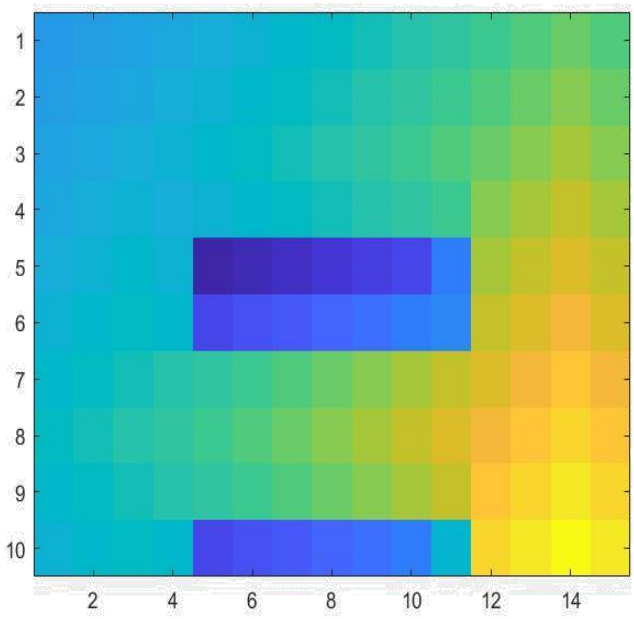
Here we can see that the 1st plot with higher discount factor will lead to values with larger magnitude having the greater impact on the value in the surrounding states and the 2nd plot with lower discount factor is darker with the large negative values in the middle area while almost every other step is valued the same, except the steps immediately next to the goal. By using the lower discount factor, the robot tends to be struck in loops since the immediate rewards are assessed profoundly. The algorithm gets more reliable and robust for long term rewards.

10. **Explain how the exploration rate ε influences the policy and V-function. Use figures to make your point. Did you use any strategy for changing ε during training?**

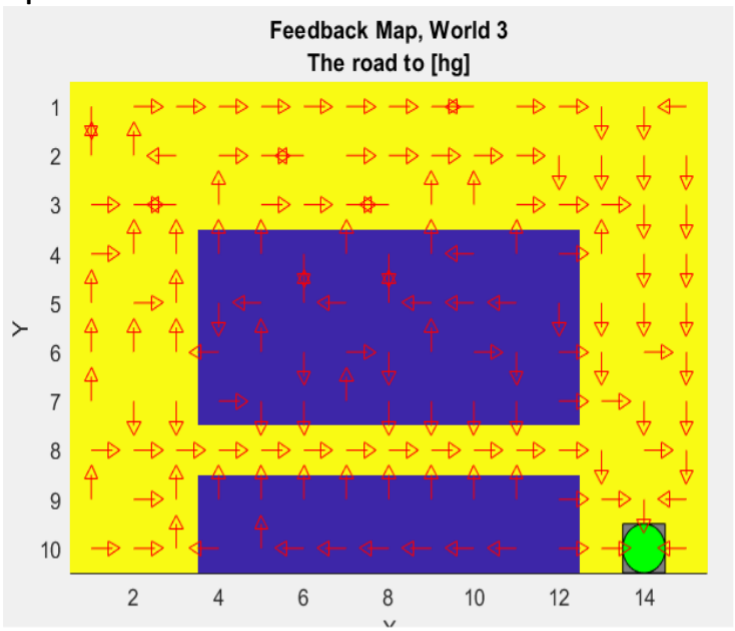
ε is the exploration factor that allows to explore new things by choosing a random action or continue using the learned experience. If ε is close to 1, it will explore and a value close to 0 will focus on already learned experience. Large ε is initialized which always explore and in each iteration its value is decreasing since it is better to use the learned experience at the end. (Borga, 2021)

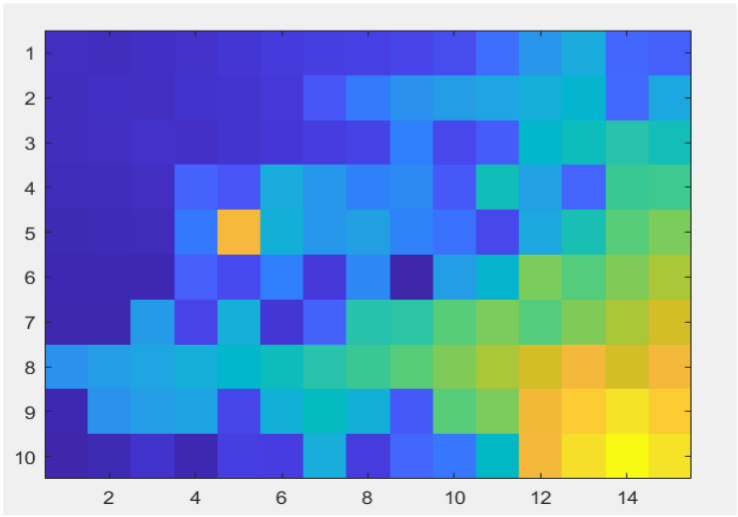
World 3 exploration rate = 0.9





Exploration rate = 0.1





When the exploration rate is low i.e 0.1 as shown in above example of world 3, many parts of the world will go unexplored, potentially missing an optimal path. When the exploration rate is high it will take more training time as it will not utilize previously learned experience. Hence, the exploration factor needs to be big in the beginning of the training so that the world is explored and the exploration rate is gradually decreasing to make robot use optimal policy and decrease the training time.

11. **What would happen if we instead of reinforcement learning were to use Dijkstra's cheapest path finding algorithm in the ''Suddenly irritating blob'' world? What about in the static ''Irritating blob'' world?**

    To create an optimal policy, if we were to run the Dijkstra's algorithm once, then we would face problems in the “Suddenly Irritating blob” world as the world is not static. But if we were to run Dijkstra's algorithm each time the world changes it would give an optimal path every time. In the static ''Irritating blob'' world if we were to run the Dijkstra's algorithm once, then it would give optimal path as the world is static.

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

    Playing computer games.
    Robotics for industrial automation.
    Machine learning and data processing.
    Traffic Light Control

References

(u.d.).
Borga, M. (2021). *Lecture 7 - Reinforcement Learning.*