TBMI26 – Computer Assignment Reports Reinforcement Learning

Deadline – March 15 2020

Author/-s: Vinay <vinbe289>

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 is a function of the state that tells us the value of being in the state given a policy, i.e., the expected amount of reward we get from this state by following the policy.
     $V(s_t) = \sum \gamma^k r_{t+k}$ where $0 <= \gamma <= 1$
   - Q function Q(s,a) denotes the expected future reward of doing action 'a' in state 's' and then following the optimal policy.
     $Q(s_k,a) = r(s_k,a) + \gamma V^*(s_{k+1})$

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

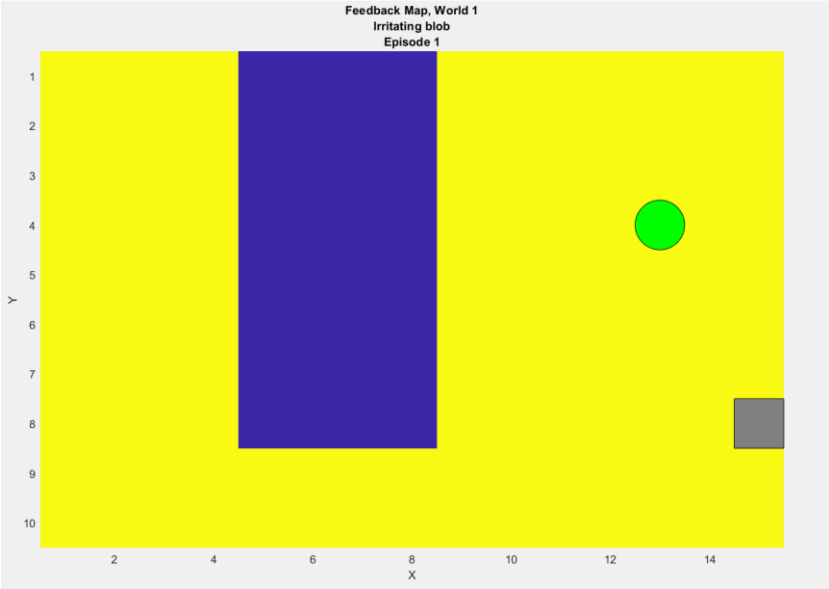
   - The Q-function is an instrument for exploration around the best policies during learning.
   - Learning the Q-function in the following way
     $Q(s_k, a_j) = (1- \eta)\, Q(s_k, a_j) + \eta\, (r + \gamma \max_a Q(s_{k+1}, a))$ i.e.
     Updated Q = Previous estimate + Better estimate
     r is the reward, η is the learning rate, γ is the weight of the reward

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

   - Here Q-function is as 3-dimensional array with the size of the world on x, y axis, and the corresponding rewards of taking a specified action in the 3rd dimension. The robot can move up, down, left and right. The world has size 10*15 which hence the Q-table is 10 * 15 * 4 multidimensional array. In the study negative feedback where used. A bad move gives high negative feedback and a good move gives low negative feedback.
   - Q table was initialized with zeros i.e. all actions are considered as equally good in the beginning. The reward of taking a step outside the world along the borders is set to negative infinity. Thus the robot never crosses the borders of the world.

- The probability parameter ε that the robot takes a random action in order to explore unknown territory, to check if this might lead to a better solution, can set by the user.
- The best action is based on V*(s) function. Here the optimal option is the action with highest reward. The probability of taking this optimal action is (1-ε) and that of random action is ε.
- We might want to explore more in the beginning of the training phase i.e. large ε and less towards the end. For each step taken by the robot until it reaches the goal the Q-function is updated and this updated Q-function is used in the next iteration.

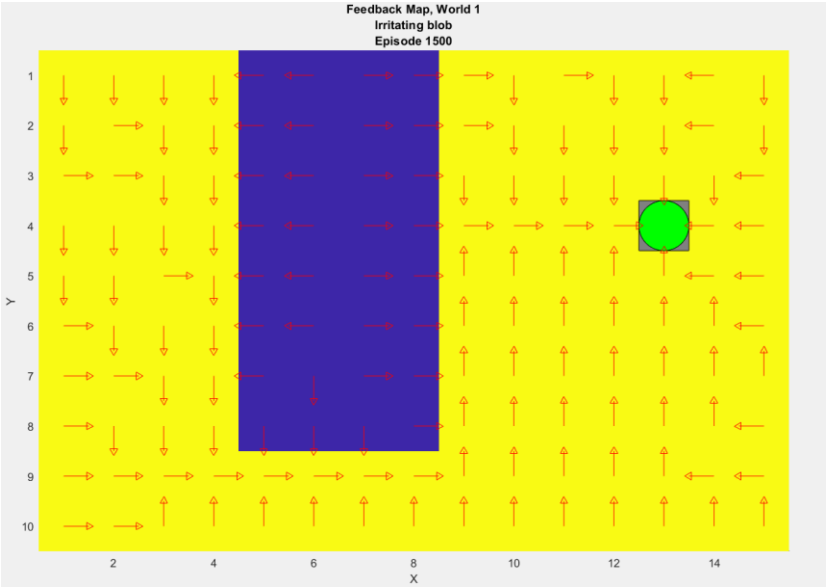
4. **Describe World 1. What is the goal of the reinforcement learning in this world? What parameters did you use to solve this world? Plot the policy and the V-function.**



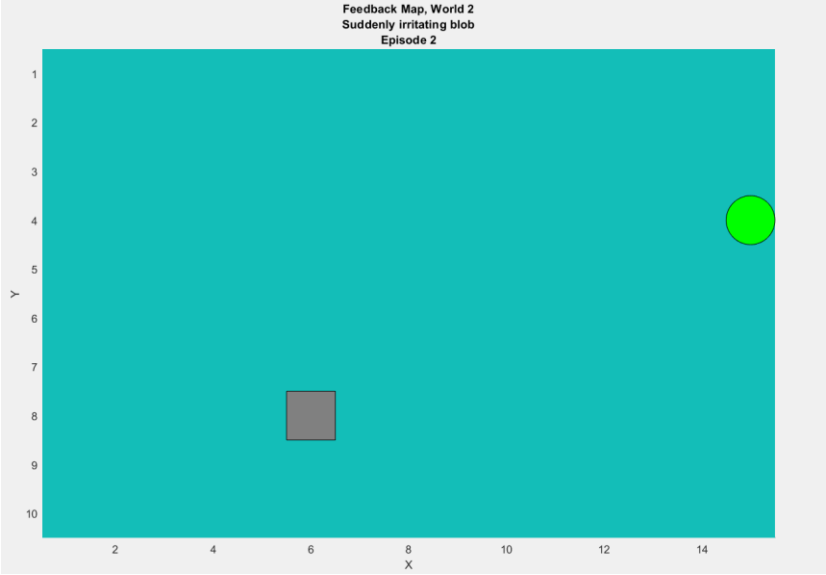
The above figure depicts world 1. The green circle is the destination and the grey square is the robot, the blue part represents a harder path to travel through and has high negative feedback. The goal of reinforcement learning is to learn an optimal policy for the robot, which is initialized at any random position in the grid, to reach the destination with highest reward.

Parameters:

- eta(learning rate): 0.2
- gamma(discount factor): 0.9
- eps(exploration parameter): 0.8
- iterations: 1500



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.

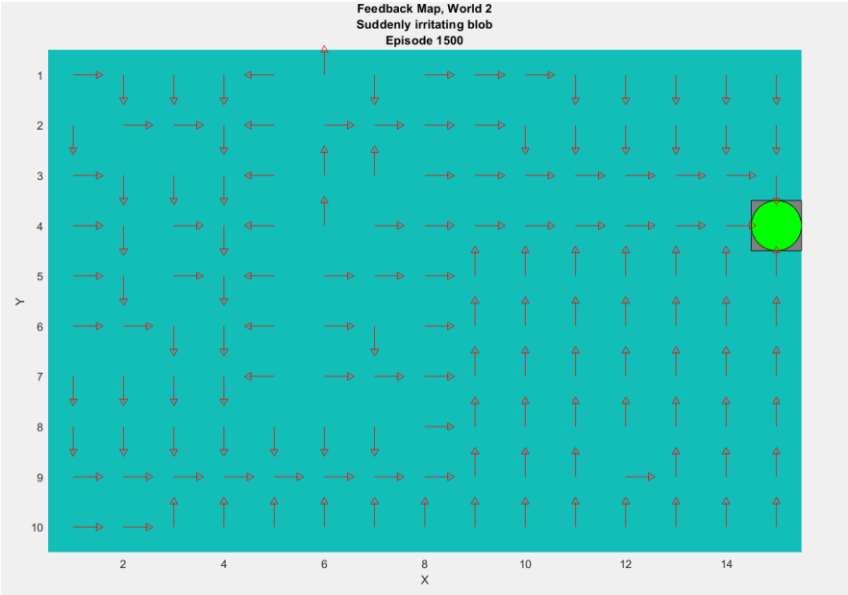


The above figure depicts world 2, the green circle is the destination and the grey square is the robot. The hidden trick here is that world 2 gets switched to world 1 and vice versa randomly. This makes learning the optimal policy harder since the best path in one world maybe the worst in the other world. The goal of reinforcement learning is to learn an optimal policy for the robot, which is initialized at any random position in the grid, to reach the destination with highest positive reward or lowest negative reward.

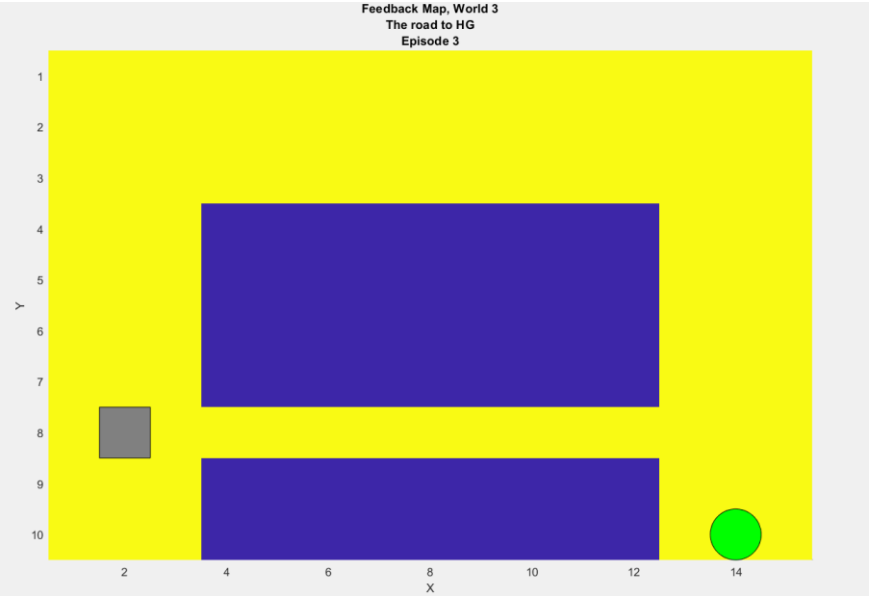
Parameters:

- eta(learning rate): 0.2
- gamma(discount factor): 0.9
- eps(exploration parameter): 0.8
- iterations: 1500

Considering the random switching between world 2 and world 1, the policies as shown in the figure below will be learnt.



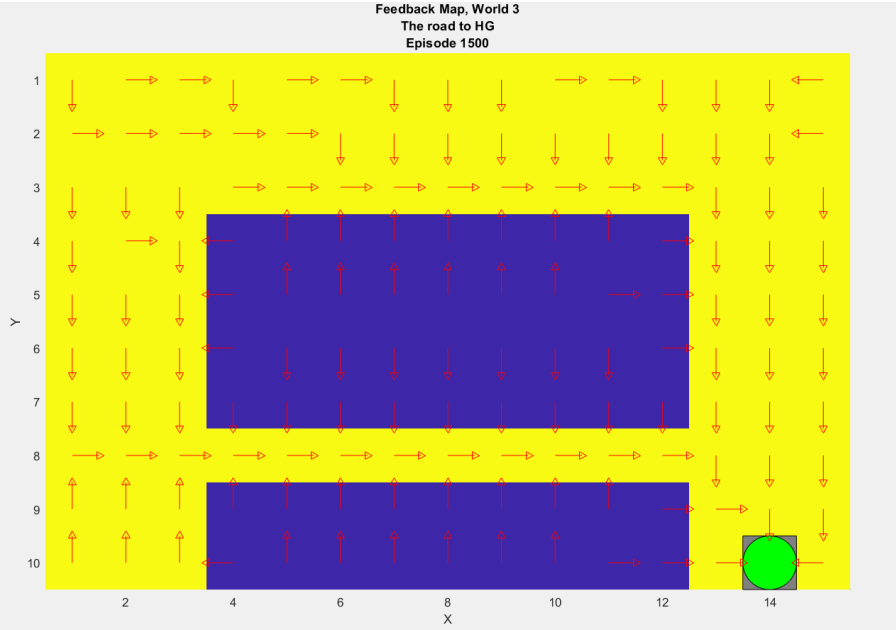
6. **Describe World 3. What is the goal of the reinforcement learning in this world? Is it possible to get a good policy from every state in this world, and if so how? What parameters did you use to solve this world? Plot the policy and the V-function.**



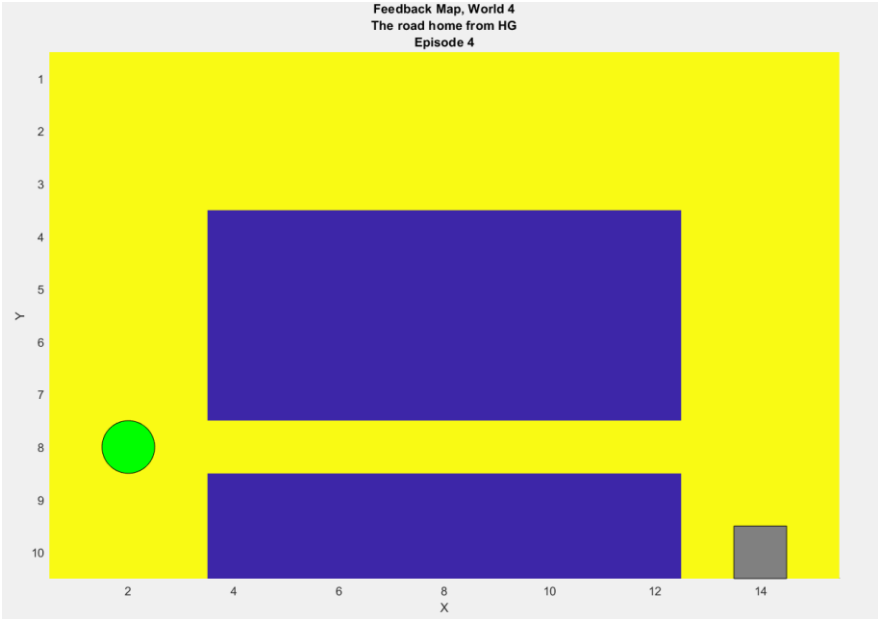
The above figure depicts world 3. The green circle is the destination and the grey square is the robot, the blue part represents a harder path to travel through and has high negative feedback. The goal of reinforcement learning is to learn an optimal policy for the robot, which is initialized at any random position in the grid, to reach the destination with highest reward. It is hard to get the best policies for the states that exist between the two blue rectangles since there are high chances of the robot slipping into the blue zone, which has high negative rewards, when the epsilon (exploration parameter) is high.

Parameters:

- eta(learning rate): 0.2
- gamma(discount factor): 0.9
- eps(exploration parameter): 0.8
- iterations: 1500



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.

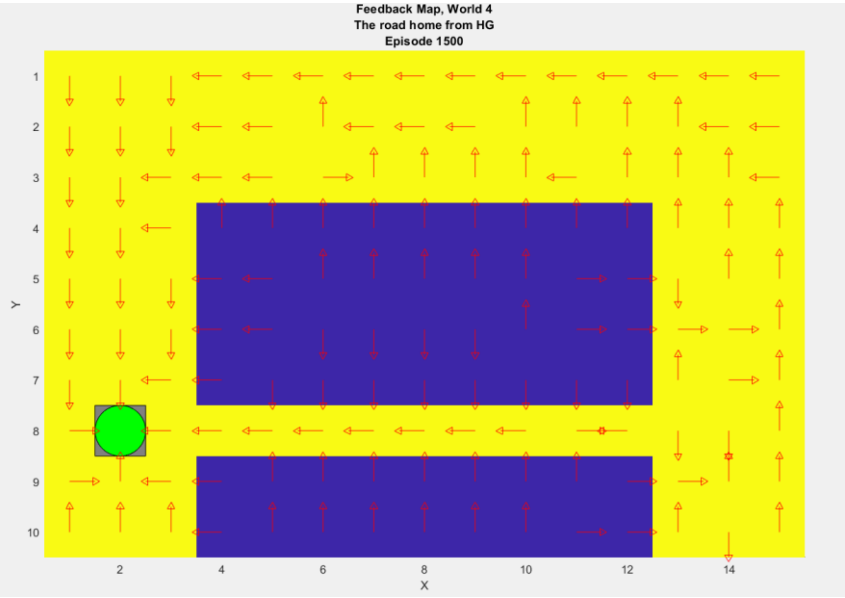


The above figure depicts world 4, the green circle is the destination and the grey square is the robot. The goal of reinforcement learning is to learn an optimal policy for the robot, which is initialized at any random position in the grid, to reach the destination with highest positive reward or lowest negative reward. This world is similar to world 3 expect

that the destination and robot position is being swapped. The hidden trick here is that the robot takes uncontrolled random steps at random intervals which simulate the alcohol intoxication after visiting HG.

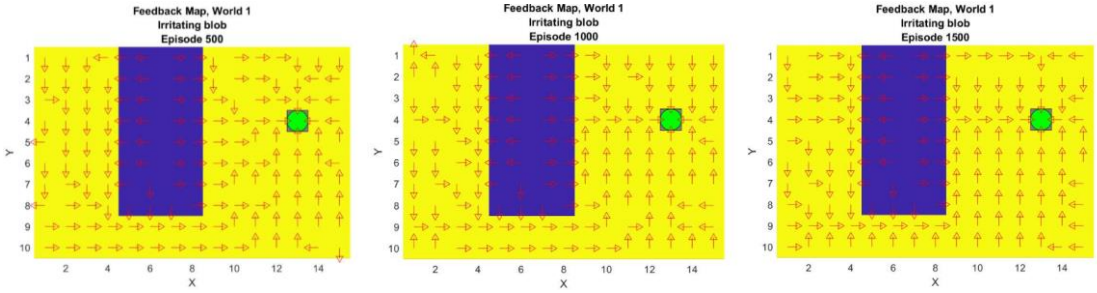
Parameters:

- eta(learning rate): 0.2
- gamma(discount factor): 0.9
- eps(exploration parameter): 0.8
- iterations: 1500

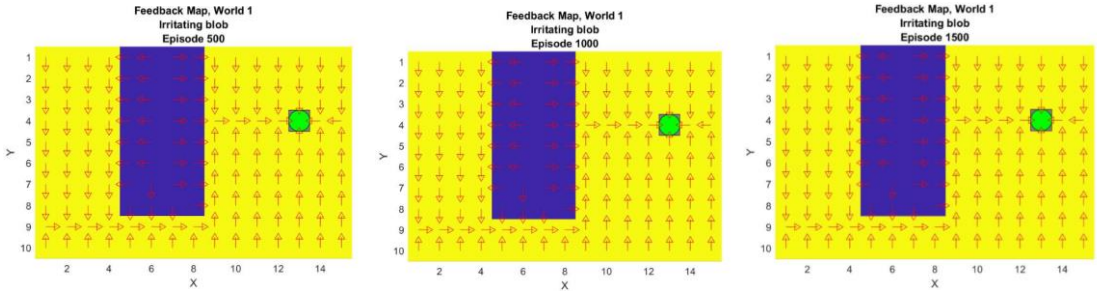


8. Explain how the learning rate α influences the policy and V-function. Use figures to make your point.

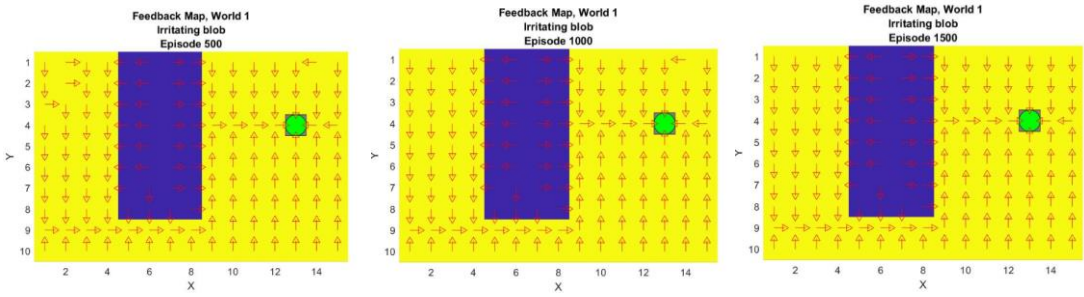
Learning rate:0.2, Discount factor:0.9, Exploration parameter:0.8



Learning rate:0.5, Discount factor:0.9, Exploration parameter:0.8



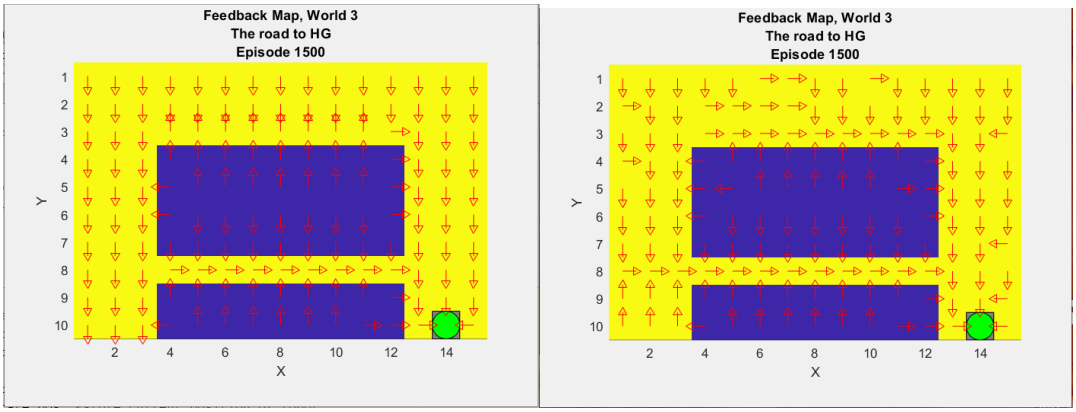
Learning rate:0.8, Discount factor:0.9, Exploration parameter:0.8



If alpha (eta in this case) learning rate is small i.e. closer to 0 preference is given to the results that is already learnt and a higher learning rate which is closer to 1 will give more preference to newly learnt results. Learning rate should be higher in a static world and lower in a dynamically changing world like world 2 so that the previously learnt policies will also be remembered.

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

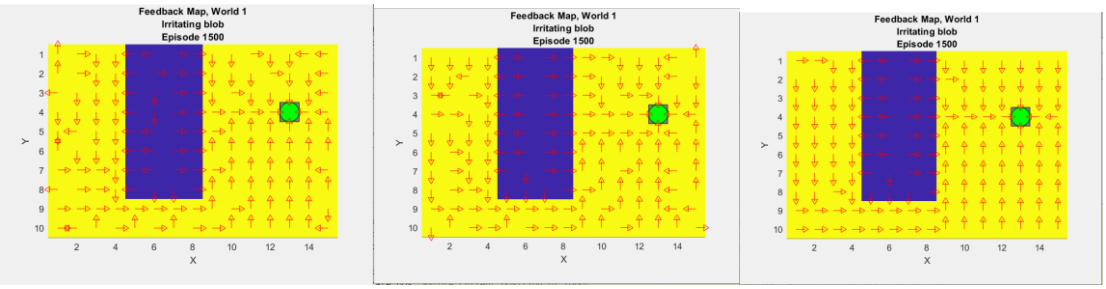
Discount factor: 0.01 Discount factor: 0.99



It is evident from the above figure that it is hard for the robot to enter the path that runs between the rectangular areas when the discount factor is low since it does not get much reward for taking that action now. Whereas when the discount factor is high the robot can enter that particular path easily and continue in the same path until it reaches the destination since it will be rewarded well later, though the immediate rewards is not much.

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?**

Epsilon:0.1 Epsilon:0.5 Epsilon:0.9



Exploration in the context of reinforcement learning refers to behavior of the agent to deviate from the previously known best path and go to a random state. If the exploration parameter is low than the robot does not explore more and sticks on to its previously known best path even though it is not the best path that exists, this is evident in the figure above. Higher epsilon means that the agent deviates from its previously best known path to explore new states and eventually ends up finding good policies.

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?**

- If we were to use Dijkstra's cheapest path finding algorithm instead of reinforcement learning, we would have to input the Q-table. The agent would not be learning the Q-table itself but just the shortest path encoded in the form of values in the Q-table.
- For Suddenly irritating blob world we can input two Q-tables i.e. one for world-1 and the other for world-2 and use the respective table for that particular current world. This would anyway not give an optimal solution to the problem.
- For irritating blob world we can input just one Q-table and learning the best path will be simple.

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

- Healthcare- Reinforcement learning can be used in medication dosing and optimization of treatment policies for those chronic, clinical trials.
- Text mining- Reinforcement learning can be used to produce highly readable summaries of long texts.
- Route optimization- In the field of Logistics optimizing the travel routes is one of the crucial tasks and reinforcement learning can be used to do it.

13. **(Optional) Try your implementation in the other available worlds 5-12. Does it work in all of them, or did you encounter any problems, and in that case how would you solve them?**