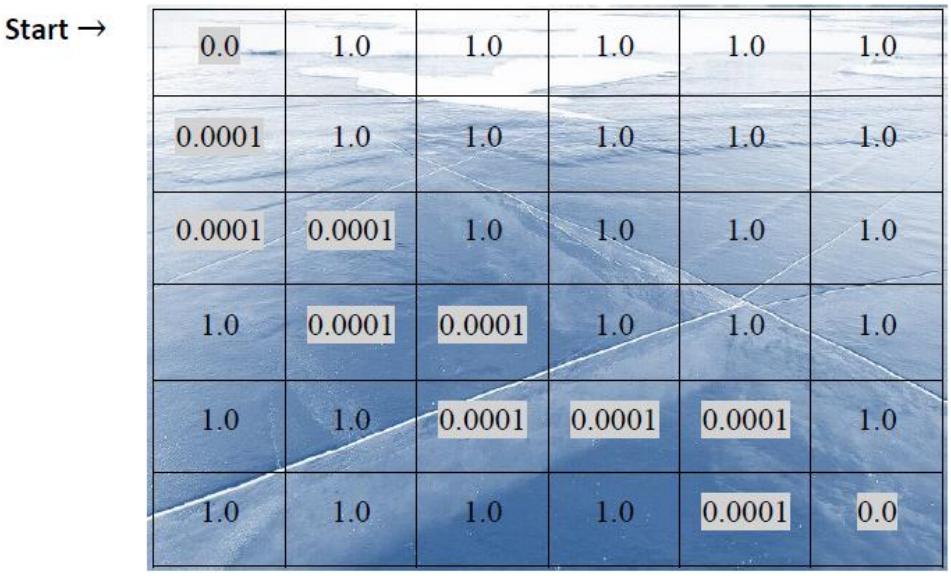
Homework 3 – RL

The purpose of this exercise is to familiarize with MDP problems, modeling and solving them. In this exercise, we implement and analyze two algorithms, value iteration and policy iteration, under different conditions.

It is in this section. Let's solve the problem on the ice lake. At first, the lake is in such a way that there is only one safe path, which is different by entering the student number for each person. That is, entering the rest of the houses is equal to breaking and disappearing. This ice lake is 6 x 6 and there are fences on all four sides of it, which makes it impossible to leave the lake. The last house is (5, 5) the target house. And you start from the point (0,0) (the property of both these houses is zero).

The initial map of the lake is placed in the attached env.py file. In the FrozenLake class in this file, complete the required functions and define any other functions it needs. You should change the model in each part, according to the requirements. The sample.py file is also attached as a code sample.



Questions

1. First, consider the environment as follows. Reward for reaching the target house 100, for each move. -1 and consider -10 points for falling into broken lake houses. Since there is very little slippage in this state, the probability of transitions is such that with a probability of 0.94, it goes to one of the other possible positions in the direction of the same action it did, and equally and smoothly. Colliding with the fence, the agent remains in place. In this part, by using the value iteration and policy iteration algorithms, you obtained the value of states, the value of state actions and the optimal policy (display the value values of states and the optimal policy on the map of the lake). Set value of the discount factor equal to 0.9

2. In this part, we are going to make changes to the map of the lake. The probability of breaking houses changes. The probability of breaking each house in the safe path is still less (0.001) and the rest of the

path has a random value between 0 and 1. On the other hand, due to slippage, it is done with a probability of 0.7 in the direction of the action and equally and to the state Random goes to one of the other possible positions. Consider the payment in this case as in the previous series. In this part, by using the value iteration and policy iteration algorithms, the value of states, the value of state actions and the optimal policy are obtained as a result (show the value values of states and the optimal policy on the map of the lake) the value of the discount factor Set equal to 0.9. Compare the obtained values and the optimal policy with the previous part and analyze and compare the convergence speed of the algorithms.

3. In this part, we want to check the effect of the profit and the limit value of theta limit in the algorithms, change the map of the environment and consider the size of the lake as 15 x 15. Apply the conditions of the second part to this new map. A) Does changing the value of theta in repeating the value of the algorithm have an effect on finding the optimal policy? What should be the value of theta to obtain an optimal policy? Explain your answer by varying θ (larger passive like 1 to near zero like 0.000001) and getting the results. b) Does the agent reach the target house with the optimal policy obtained in part a? Otherwise, change the reward value of the target house in such a way that the path starting from the first house, by adopting the optimal policy, reach the target house, analyze and justify your observations in this section.