**Artificial Intelligence Laboratory 4: Reinforcement Learning**

**Rasmita Samantaray, Jazib Ahmed**

The code is written for both Policy Iteration and TD-learning (Task1, Task2, Task3) along with the extra credit (Task-4) solution.

**Task 1:**

* Two more grids, ‘grid2’ and ‘grid3’ are created and the result is analyzed with using the value iteration, policy iteration, TD-learning and Q-learning.

Grid-2 (immediate rewards =0)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  | Wall |  | -1 |  |
|  | Wall | 1 | Wall | 10 |
|  |  |  |  |  |
| -10 | -10 | -10 | -10 | -10 |

Grid-3

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | 1 |
|  |  |  | -1 |
|  |  | Wall |  |

**Task-2:**

Policy Iteration and TD-Learning is implemented.

**Task-3:**

1. In Value iteration, the maximum utility value rarely changes and takes long time to converge than policy iteration. The policy iteration is optimal and can converge faster than value iteration.

For grid1 – Value Iteration converged at step 24

Policy Iteration converged at step 5

For grid2- Value Iteration converged at step 75

Policy Iteration converged at step 10

For grid3- Value Iteration converged at step 189

Policy Iteration converged at step 5

1. Depending upon the immediate reward values the agent decides to take risk or no-risks at all. Hence the optimal policies change as per the immediate reward values.

e.g If immediate reward R(s) = -2.0, The agent heads straight for the nearest exit (the result from both value and policy iteration as shown below), even if the exit=-1

[['R' 'R' 'R' 'G']

['U' 'W' 'R' 'G']

['R' 'R' 'R' 'U']]

1. TD-Learning and Q-Learning are model free learning. Value and Policy Iterations are model-based learning where we know the model (T and Rewards). TD-Learning uses a fixed policy to get the optimal utility matrix. Q-Learning gives both optimal policy and utility matrix.
2. Alpha is the learning parameter and depending on the epsilon value, the agent explores randomly or on current policy.

Alpha is different for all the 3 grid values.

1. Episode is the length of the simulation or, the number of iterations. At the end of the episode TD or, Q-learning converges optimally. TD-Learning needs more number episodes than Q-leaning to converge optimally.

The result of value iteration, policy iteration, TD learning and Q-earning is attached for grid1, grid2, grid3.



**Task-4: (Extra Credit)**

Here the diagonal moves are possible as per the below diagram.

0 (L)

1 (R)

2 (U)

3 (D)

5 (UR)

7 (DR)

6 (DL)

4 (UL)

* 8 directions: [‘L’, ‘R’, ‘U’, ‘D’, ‘UL’, ‘UR’, ‘DL’, ‘DR’]

Here the probability of going UP with Up action is (1-noise) i.e. 0.8

Probability of going UP-Left (UL) with the Up action is noise/2 i.e. 0.1

Probability of going UP-Right (UR) with the Up action is noise/2 i.e. 0.1

The result of the Value and Policy Iterations for ‘grid-1’ using diagonal movement,

