CS747 Programming Assignment 2

190050001

October 11, 2021

Task 2

The MDP is created using one extra state which is the terminal state. Whenever the game ends in a win, loss or a draw, the next state is this terminal state with corresponding rewards of 1, 0, 0 respectively. The other states are the same as states provided in the file, and the transitions are generated by taking each possible action, seeing the output according to opponent's policy, and the next state hence reached, is formulated as a transition with probability according to the opponent's policy.

Task 3

The file task3.py was run by changing random seeds and initial players, and the policies converged in each case. We also assumed that we already have the states file for each player in the states folder. A sample output for initial player 2 and random seed 50 is given by
Iteration: 0
Iteration: 1
Iteration: 2
Player: 2, Changed states: 676
Iteration: 3
Player: 1, Changed states: 344
Player: 2, Changed states: 676
Iteration: 4
Player: 1, Changed states: 344
Player: 2, Changed states: 97
Iteration: 5
Player: 1, Changed states: 24
Player: 2, Changed states: 97
Iteration: 6
Player: 1, Changed states: 24
Player: 2, Changed states: 1
Iteration: 7
Player: 1, Changed states: 0
Player: 2, Changed states: 1
Iteration: 8
Player: 1, Changed states: 0
Player: 2, Changed states: 0
Policies Converged!
This means that the policies converged in 8 iterations. For each policy, the number of states for which the policy changed with respect to its previous policy is given by changed states, since the changed states are decreasing, we can see that the policy is indeed converging. Also, there are no changed states for first 3 iterations because there was no previous policy to compare with. This file also creates a folder by the name of policies in the same directory in which there are policy files after every iteration. It contains policy files for both the players at every iteration.
The sequence of policies generated for each player are guaranteed to converge because there exists an optimal policy for player 2 where it cannot lose. Hence, the final policy for player 2 is the optimal policy, but for player 1 there does not exist any such optimal policy, even so since the policy for player 2 is now fix the policy for player 1 also does not change.