Homework 4

I. Part2-Algorithm

Two types of agents are defined. Agent 1 is modelled as AgentDefined and agent 2 is modelled as AgentQ. AgentDefined has a mixed strategy of choosing clods and mcmennamins with equal probability of 0.5 each. Agent 2 uses q learning to learn a strategy. The q learning agent stores the actions taken by agent 1 in the previous games, and uses that information while making a decision.

For part 2b, agent 1's strategy is to pick clods with probability of 0.7 and mcmennamins with probability 0.3. Agent 2 uses q learning to learn an optimal policy.

II. Part 1

A. Part a

When a local reward is used. Each agent receives a reward based on the attendance on that day. The local reward is given by the following function:

$$L(z) = x_k(z)e^{x_k(z)/b} \tag{1}$$

To maximize its local reward, each agent will pick try to pick a night when attendance is less. At $x_k(z)$ = b, the value of the reward is the maximum. As we increase $x_k(z)$ beyond this value, the reward will decrease.For the system to be in nash equilibrium, it should be in a state where no agent can get a better reward if it attends the bar on a different night. The only state in which this will hold true will be when the attendance profile looks like this [7, 7, 7, 7, 7, 7]. In this state, no agent can maximize its reward by attending the bar on a different day.

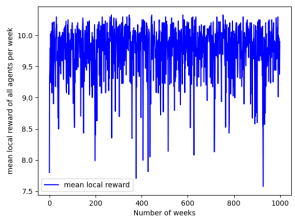

Fig. 1. Performance of local reward

B. Part b

When we use the global reward, each agent gets the global reward of the system. The individual performance of the agents don't matter, maximizing the overall global reward is desirable for all agents, as they are all working towards increasing the system's performance. The global reward will be maximum when an optimal number of agents attend on each day, that is $x_k(z)$ = b. Since b=5 gives the optimal value of reward, the agents will work towards achieving this on as many days as possible. The attendance profile for nash equilibrium will look like this- [17,5,5,5,5,5]. It will be 5 for 5 days, and the remaining agents will attend on one day. If any agent deviates from their action, the global reward they receive won't increase. Therefore, this state is a nash equilibrium.

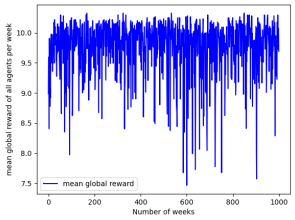

Fig. 2. Performance of global reward

C. Part c

The nash equilibrium for local and global reward are different. We can expect that the local reward will not perform as well as the global reward. In the case of local reward, the agents will end up maximizing their own reward for a particular day. Whereas, in the case of the global reward, the agents will work towards achieving optimal system reward.

D. Part d

The value of the difference reward will depend on the difference in value of the global rewards when the agent attends on that day and when it doesn't attend on that day. This reward gives each agent the incentive to improve its performance to increase the global reward. The attendance profile of [17,5,5,5,5,5] is the nash equilibrium. As each agent is rewarded when they contribute to optimize the system performance, which depends on the global reward.

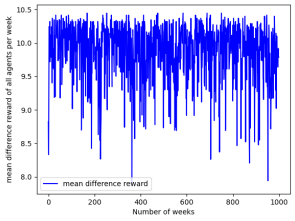

Fig. 3. Performance of difference reward

E. Part e

The nash equilibrium is the same as that of the global reward. This implies that the difference reward is expected to perform better than the local reward.

III. Part 2

A. Part a

The optimal policy agent 2 converges to is either always picking McMennamins or always picking Clods. Agent1 picks clods and mcmennamins both with probability 0.5 and agent 2 has no way of predicting or knowing what agent 1 prefers. Therefore, it either learns to always pick one. If it always picks mcmennamins and if agent1 picks that too it gets a reward of 5. Whereas, if it always picks clods and agent1 picks that too, it gets a reward of 2. So, it learns to always choose one particular bar, to increase its chances of getting a non-zero reward. Therefore, this policy maximizes agent two's payoff.

It is a Nash equilibrium when agent 1 and agent 2 both pick the same bar. When they choose differently, it is not a nash equilibrium. However, there are many instances in which it does. As seen in figures 4 and 5, a nash equilibrium is reached on most cases.

As seen in figure 4, agent 2 always picks Mcmennamins and gets a reward of 5 in most games. It gets zero in some. Figure 5 shows that agent 2 learned the policy of always picking clods and receives a reward of 2.

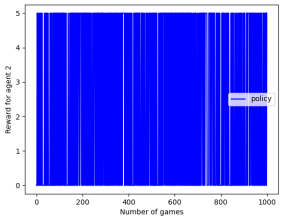

Fig. 4. Reward for agent 1 for McMennamins(always) for 1000 games

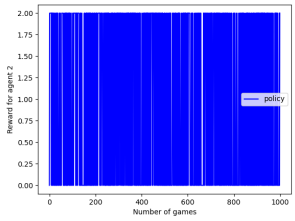

Fig. 5. Reward for agent 1 for Clods(always) for 1000 games

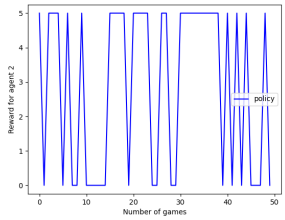

Fig. 6. Reward for agent 1 for McMennamins(always) for 50 games

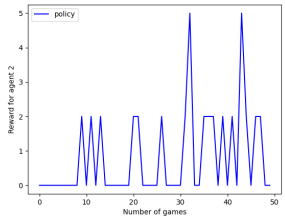

Fig. 7. Reward for agent 1 Clods(always) for 50 games

B. Part b

Let the probability of of picking mcmennmins be x for player 1 and that of picking clods be 1-x. Similarly, for player 2, probability of picking mcmennamins be y and for clods be 1-y. Expected payoff for player 1 is the sum of the payoffs of the two possible actions, multiplied with their probability of player 2 choosing those actions.

1) Player 1:

$$EP(Mc) = 5y + (1-y)0 \tag{2}$$

$$EP(C) = 0y + (1-y)2 \tag{3}$$

For player 1 to pick mcmennamins, Expected payoff of mcmennamins should be greater than that of clods.

$$\begin{aligned} 5y &> 2 - 2y \\ y &> 2/7) \end{aligned} \tag{4}$$

Similarly, for player 2 the expected payoffs can be given by-

2) Player 2:

$$EP(mc) = 2x + 0(1-x) \tag{5}$$

$$EP(C) = 0x + (1-x)5 \tag{6}$$

For player 2 to choose mcmennamins,Expected payoff of mcmennamins should be greater than that of clods.

$$\begin{aligned} 2x &> 5 - 5x \\ x &> 5/7 \end{aligned} \tag{7}$$

Therefore, for y=2/7 and x=5/7, the two players are indifferent between the two options. Hence, these probabilites are the mixed strategy nash equilibrium.

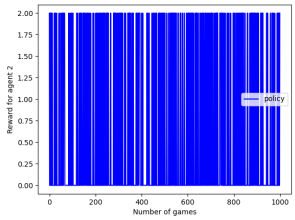

Fig. 8. Mixed Strategy reward for agent 2 for 1000 games

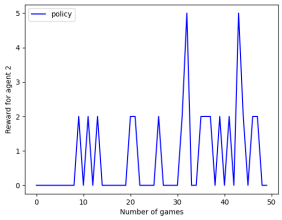

Fig. 9. Mixed Strategy reward for agent 2 for 50 games

When the mixed strategy of agent 1 is changed to 0.7 probability of choosing clods and 0.3 of choosing mcmennaminsis. The rewards of agent 2 is observed in figure 8. It shows the payoffs for agent 2 for a strategy it learns. For majority of the games, agent2 gets a reward of 2. This shows that agent2 learns the mixed strategy of picking clods with probability 0.3 and mcmennamins with probability 0.7. Therefore, in the mixed strategy equilibrium, both agents end up choosing mcmennamins. As seen in 8, agent 2 gets a reward of 2 in this equilibrium.