Homework 4: Game Theory

Josyula Gopala Krishna

I. Introduction

The El Farol bar problem is a game of constrained resources with non-cooperating agents, where the agent needs to choose a night to attend the bar, the payoff for the agent is low when the bar is overcrowded, it's optimal when the bar is uncrowded. This report considers describes the Nash Equilibrium in the El-Farol bar problem and also studies the pure and mixed strategy Nash Equilibria in a two-player game.

II. Background

A. Q-Learning

To study the nash equilirum of the El-Farol bar problem this report uses Q-learning, Q-Learning uses Bellman updates over a sequence of states, and actions to learn the game strategies. The goal of the agent is to interact with the environment and select actions in a way that maximizes the future rewards, which are discounted by a factor γ per timestep $R_t = \sum_{t'=t}^{T} \gamma^{t'-t} r_{t'}$ where T is the timestep at which game terminates.

$$Q_{i+1}(s,a) = Q_i(s,a) + \alpha \left[r + \gamma \max_{a'} Q^* \left(s', a'\right) - Q_i(s,a) \right] \tag{1}$$

In the El-Farol bar problem agent state is an estimate of the expected pay-off for choosing a night $k \in K$, and the action is choosing the night k, agent estimate for attending a night k is iteratively updated as Bellman update equation on the Q-value function as in eq.1

B. Erev-Roth Reinforcement Learning:

To learn the strategies in a two-player game this report uses the Erev-Roth Reinforcement Learning for game Erev—Roth algorithm (ERA) comes as follows. For each action $j = 1..J$ available to a learning agent over the time period t some initial level of value S_j^t that is, a propensity to select this action, is introduced. If an action k^t, selected at a time period t results in reinforcement (reward) $R^t \geq 0$, an agent's propensity to select this same action for the next time period varies by the following rule:

$$S_j^{t+1} = (1-f)S_j^t + \begin{cases} R^t(1-e), & j = k^t; \\ R^t e/(J-1), & j \neq k^t. \end{cases} \tag{2}$$

where: $f \in [0,1]$ is a recency parameter; e is an experimentation parameter (i.e. the extent to which an agent is willing to undertake trial actions in order to gain experience). In all the experiments in this report we use the following parameter values: $f = 0.9$, and $e = 0.2$ The recency parameter f determines, to what extent an agent is inclined to change his action selection as compared to a prior time period. Due to the multiplier $(1-f)$, forgetting occurs, that is, overtime the effect of past experiences tends to reduce to zero. With given propensities S_j^t an action is selected in a random manner: $k^t = j$ with a probability, calculated as follows:

$$p_j^t = S_j^t / \sum_{i=1}^{J} S_i^t. \tag{3}$$

ERA is shown to converge to Nash Equilibrium as noted in [1].

C. Nash Equilibrium

An n-player non-cooperative game is defined as a tuple $\Gamma = \left\langle N, (S_i)_{i \in N}, (u_i)_{i \in N} \right\rangle$ where $N = 1, 2, ..., n$ are the set of player, S_i is a strategy profile i.e. a set of strategies, one for each player, and u_i is the utility function, a strategy profile is a Nash equilibrium if no player can do better by unilaterally changing their strategy i.e. let S_i be the set of all possible strategies for player i, where $i = 1, \ldots, N$ Let $s^* = (s_i^*, s_{-i}^*)$ be a strategy profile, a set consisting of one strategy for each player, where s_{-i}^* denotes the $N-1$ strategies of all the players except i. Let $u_i(s_i, s_{-i}^*)$ be player i's payoff as a function of the strategies. The strategy profile s^* is a Nash equilibrium if

$$u_i(s_i^*, s_{-i}^*) \geq u_i(s_i, s_{-i}^*) \text{ for all } s_i \in S_i \tag{4}$$

A game can have more than one Nash equilibrium. Even if the equilibrium is unique, it might be weak: a player might be indifferent among several strategies given the other players' choices. It is unique and called a strict Nash equilibrium if the inequality is strict so one strategy is the unique best response.

III. Problem 1 - El-Farol Bar Problem

In this problem, we explore the Nash Equilibrium strategies of the El-Farol Bar agents, there are $n = 42$ agents, the agent has to pick at least one night $k \in [0,5]$ i.e $K = 6$, and the optimal seating capacity $b = 5$ trained using Q-Learning for various rewards using the following parameters, the training process uses ε greedy exploration for 75% of the total iterations and later uses the best strategy learned after these iterations.

$$\begin{aligned}\gamma &= 0.99\\ \alpha &= 0.99\\ \varepsilon &= 0.5\\ totalweeks(iterations) &= 1000\\ n &= no.of agents\end{aligned}$$

A. Local Reward

The Q-learning algorithm of eq. 1 is utilized with the local reward of the eq.5

$$L(z) = x_k(z) * e^{-x_k(z)/b} \tag{5}$$

After convergence, using the local rewards, the agents adopt a pure strategy, i.e. the agents choose a fixed day out of the 6 (k=6) days. The number of agents choosing a particular day is shown in Figure. 2.

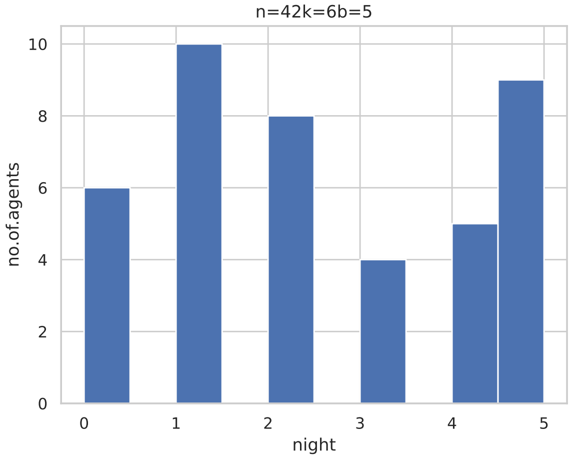

Fig. 1: No. of agents choosing a night after the q-learning converges

B. Global Reward:

Q-Learning when converged on global rewards shown in eq.6 stabilize to a Nash equilibrium where most agents choose a single night of the week while some agents have a mixed strategy this is shown in the Figure

$$G(z) = \sum_{k=1}^{K} x_k(z) * e^{-x_k/b} \tag{6}$$

Local-vs-Global Rewards and Observations: As noted the agents using local rewards converge to a pure strategy nash equilibrium while the agents using global rewards almost converge to a pure strategy i.e. although most agents choose a fixed night, they still shave a propensity to move to another night with some probability, pure strategy nash equilibria have a better-expected utility compared to mixed strategy. Therefore it is expected that using the local rewards gives a better utility maximization response compared to the global reward.

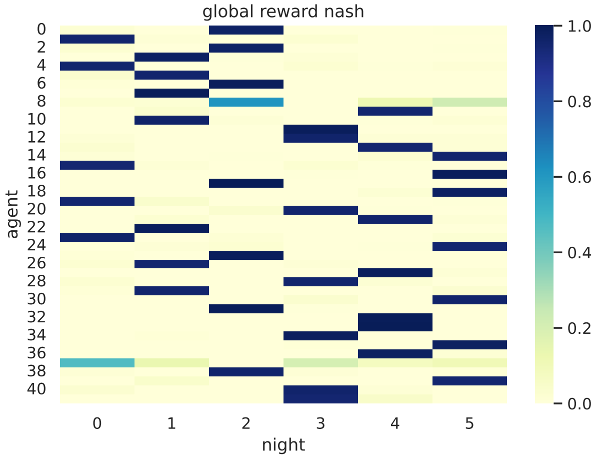

Fig. 2: Agent distribution over choice of nights using global rewards

C. Counter Factuals

In this section, we consider the counterfactual difference rewards shown in the equation.(7)

$$G(z) = x_k(z) * e^{-x_k/b} - (x_k(z) - 1) * e^{-(x_k-1)/b} \tag{7}$$

With these rewards, the agents also converge to a mixed strategy best responses. in which the agents distribute their choices over the entire week, causing congestion in the system on any given night.

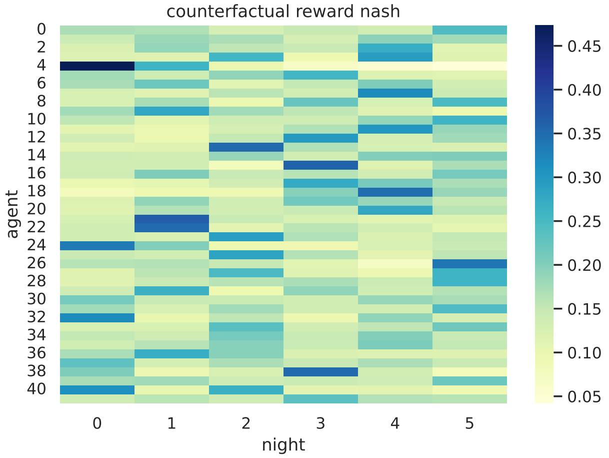

Fig. 3: Agent distribution over the choice of nights using counterfactual rewards

Local, Global and Counterfactual Rewards and Observations: From observations in the strategies developed by the agents with the three different reward types, it can be seen that the agents adopted different mixed strategy best responses to the system, since the mixed strategy response of the counterfactual rewards is having the highest variance it is expected to see that the rewards, in this case, perform the worst over a duration of 1000 weeks, however over the duration of 10000 weeks it can be seen that the counterfactual differences work better over global rewards, but not better than the local rewards, this is due to the fact that a pure strategy response always has a higher pay off than a mixed strategy.

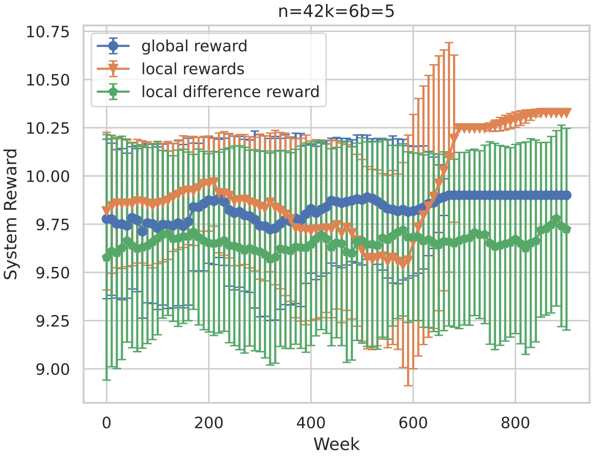

Fig. 4: (a) system performance 1000 weeks

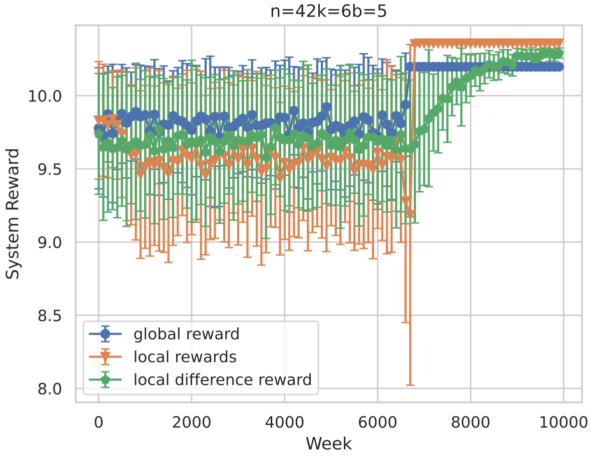

Fig. 5: (b) system performance 10000 weeks

Fig. 6: The system performance given by equation.(6) for counterfactuals is observed to better in the long run compared to global reward, however local rewards have the highest expected utility

An attendance profile showing the number of agents attending per night, using all three strategies, is presented in Figure.12, from this figure, it is clear that local rewards are converging to a pure strategy nash equilibrium where 6 or 7 agents choose to attend on any night, and they do not cross this limit, this is almost close to the best strategy of 6 agents going to the bar any night. While using global rewards, the agents achieve the optimal seating of 6 agents per night, the slightly mixed strategy makes some agents attend above the capacity of the bar, with other frequent choices being 8 and 10, using the Difference rewards with a mixed strategy distributes the agents all over the nights and this causes congestion, during most of the days, and other days it causes the bar to stay empty this mixed strategy nash equilibrium is weaker compared to the other two strategies, since the agents can choose an alternate day as likely as the current day and this causes the inconsistency in attendance and overflow above the optimum capacity, while the pure strategy Nash Equilibria of local rewards give a fixed strategy and increase the overall system utility.

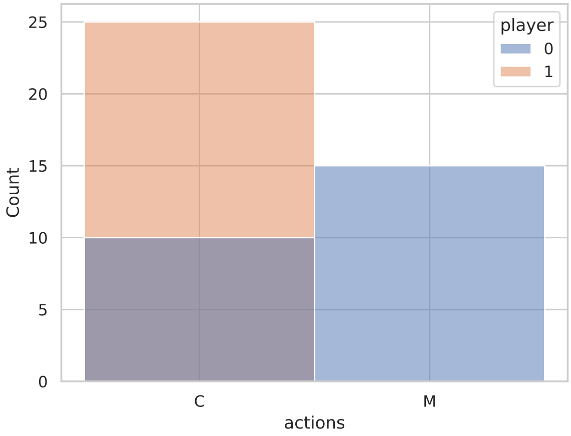

Fig. 7: (a) Histogram of actions, the figure shows that Player 2 also converges to a distribution of [1,0] over actions C and M.

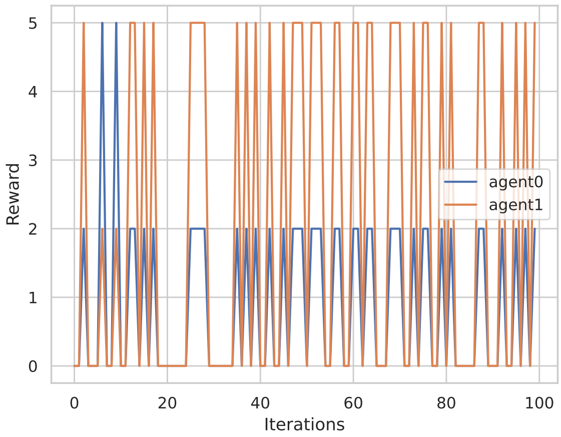

Fig. 8: (b) Reward per iteration for each player

IV. Problem 2 - Two Player Game

In this problem, we are presented with another bar problem, where the players get a reward if they go to the same bar, else they get a zero pay off. Player 1 prefers McMenamin's(M), and Player 2 prefers Clod's(C). The payoff's for player 1 are $U_1(C,C) = 2$, $U_1(M,M) = 5$, similarly for player 2 are $U_2(C,C) = 5$, $U_2(M,M) = 2$.

A. Problem 2a

We consider a case where Player 1 chooses the strategy C and M equally with a probability of 0.5 each. To maximize

Player 2's reward, Player 2 is trained using the Erev-Roth RL rules presented in Section II. B. which is a simple action value learner, using this strategy it is noted that Player 2 converges to a pure strategy of [1, 0] for P(C) and P(M), this is a result that is expected since playing strategy is the best strategy for player 2.

The actions taken by the players after convergence of the propensity updates are plotted in Figure.8

It is noted that this is a pure strategy Nash Equilibrium for Player 2.

B. Problem 2b

A mixed strategy Nash Equilibrium does exist for this game and can be computed using the Necessary and Sufficient conditions for Nash Equilibrium mentioned in [2] For mixed strategy profile $(\sigma_1^*, \ldots, \sigma_n^*)$ is a mixed strategy Nash equilibrium iff
1. $u_i\left(s_i, \sigma_{-i}^*\right)$ is the same $\forall s_i \in \delta\left(\sigma_i^*\right)$
2. $u_i\left(s_i, \sigma_{-i}^*\right) \geq u_i\left(s_i', \sigma_{-i}^*\right) \forall s_i \in \delta\left(\sigma_i^*\right) \forall s_i' \notin \delta\left(\sigma_i^*\right)$ (that is, the payoff of the player i for each pure strategy having positive probability is the same and is at least the payoff for each pure strategy having zero probability).

using these conditions gives the following system of equations in eq. 8

$$5\sigma_1^*(C) - 2\sigma_1^*(M) = 0$$
$$2\sigma_2^*(C) - 5\sigma_2^*(M) = 0$$

Solving the system of equations gives $\sigma_1^*(C) = \frac{2}{7}$ $\sigma_1^*(M) = \frac{5}{7}$ for player 1 and $\sigma_2^*(C) = \frac{5}{7}$ $\sigma_2^*(M) = \frac{2}{7}$ for player 2.

There are also pure strategy nash equilibrium for this game which is (C,C) and (M,M) for both the players.

References

[1] D. Whitehead, "The el farol bar problem revisited: Reinforcement learning in a potential game," *Edinburgh School of Economics, University of Edinburgh, ESE Discussion Papers*, 01 2008.

[2] Y. Shoham and K. Leyton-Brown, *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge, UK: Cambridge University Press, 2009.

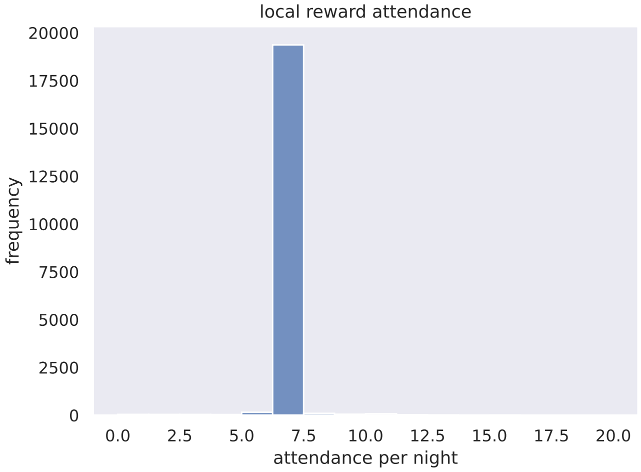

Fig. 9: a. local rewards attendance profile

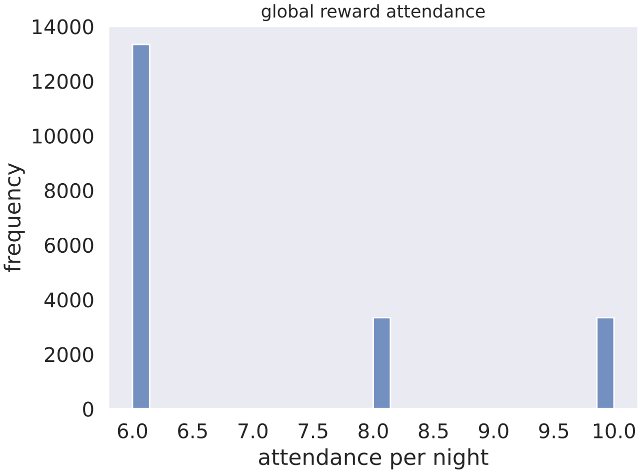

Fig. 10: b. global rewards attendance profile

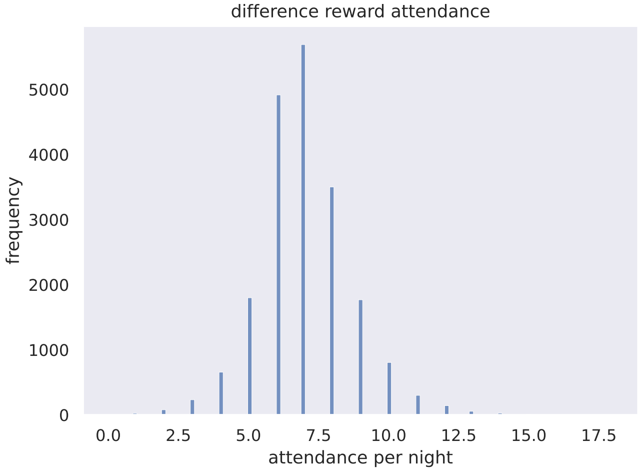

Fig. 11: c. difference reward attendance profile

Fig. 12: (a) Local rewards have a pure strategy nash equilibrium and it is observed that all days have an attendance of 7, while (b) using global reward, has most days with an attendance of 6 and others at 8 , and 10, thus reducing the system reward, (c) difference rewards have a distributed set of attendance profiles per day with a most frequent choice of 7 attendees per night, however since the agents also choose to attend en-mass to the bar, this effects the system rewards in a negative way.