Homework 3: El-Farol Bar Problem

Josyula Gopala Krishna

***Abstract*— This report simulates the El-Farol bar problem and learns a strategy for an agent in the system using reinforcement learning. The difference rewards are formulated as local(or nightly) and global difference rewards. Quantitative comparisons between various reward structures is provided. Results of ablation studies for q-learning parameters and their effect on system reward are documented. Factordness and sensitivity of rewards are also examined. Simulations performed suggest global counterfactual difference reward structure has a higher system performance under congestion while being robust to changes in q-learning parameters.**

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 ways of achieving an optimal collective resource allocation in such situations, where the agents do not know the strategy of other agents in the system.

II. Background

This report uses Q-Learning as a strategy to address the El-Farol Bar problem, the agent estimated pay-off per night is formulated as a q-learning update. On the assumption that each agent observes the total number of attendees on any given night, a local reward (reward per night) is designed, and the agents also maintain the most likely pay-off for attending the bar on a particular night. The bar has a maximum seating capacity of b, and attendees per night are x_k, the strategy of the agent is given by s_i, and strategies of the other agents are given by s_{-i}, K is the number of nights per week. Let $\mathcal{U}(\int_\rangle, \int_{-\rangle})$ be a local payoff for an agent, the experiments are conducted for two varieties of local payoffs.

$$U(s_i, s_{-i}) = b/x_k(z) \tag{1}$$

$$U(s_i, s_{-i}) = x_k(z) * e^{-x_k(z)/b} \tag{2}$$

and a system reward of

$$G = \sum_{k=1}^{K} x_k(z) * e^{-x_k/b} \tag{3}$$

Where $x_k(z)$ is the total number of attendees in the bar, on a night k.

In a multiagent environment, an essence for evaluating the agent's actions can be defined using Alignment(Factordness) and Sensitivity(Learnability). Alignment measures the agent behavior with respect to the system reward, it's measured as shown in eq.4

$$\mathcal{F}_{g_i} = \frac{\int_z \int_{z'} u\left[\left(g_i(z) - g_i\left(z'\right)\right)\left(G(z) - G\left(z'\right)\right)\right] dz' dz}{\int_z \int_{z'} dz' dz} \tag{4}$$

intuitively it can be seen an action taken by an agent that improves its own evaluation function also improves the global evaluation function.

Sensitivity measures agents' own response w.r.t the other agent's actions, an agent is said to be sensitive when its's fitness is not effected by the other agent's actions. It's measured as shown in eq.5

$$\lambda_{i,g_i}(z) = E_{z'}\left[\frac{\|g_i(z) - g_i\left(z - z_i + z_i'\right)\|}{\|g_i(z) - g_i\left(z' - z_i' + z_i\right)\|}\right] \tag{5}$$

In eq.4 and eq.5 g_i is the reward function z_i is the state of agent 'i' and z' is a random state vector and z_{-i} is the state of all other agents other than agent 'i' and $E_{z'}$ is the expectation over z'.

A. Method

Q-Learning is formulated for a Markov Decision Process(MDP) with a sequence of states, and actions to learn the game strategies depending on these sequences. 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.

The 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.6

$$Q_{i+1}(s,a) = Q_i(s,a) + \alpha\left[r + \gamma \max_{a'} Q^*\left(s', a'\right) - Q_i(s,a)\right] \tag{6}$$

B. Difference Rewards

A difference reward is a reward that has an implicit alignment and is defined by eq.7

$$D_i(z) = G(z) - G(z') \tag{7}$$

when $z' = z_{-i} + c_i$, it is called a counterfactual difference reward, where c_i is the counterfactual, learning with these rewards should be less noisy and insensitive to other agent's actions, difference reward can be directly used in eq.6 using $r = D_i(z)$, further variants, and results obtained for difference rewards are discussed in Section 4.

C. Training

Training is performed using Q-Learning (eq.6) and the parameters of the Q-learning are listed below

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

The state of the agent is the reward estimate for each night of the week, which is of size K and action correspond to choosing a night from any of the K nights therefore action size is K

III. Results and Analysis

A. Problem 1

In this problem, a simple local reward is used to train the agents to choose a night, in this experiment, the local rewards of eq.1 & eq.2 are used.

Using the local reward of eq.2 $g_i(z) = x_k * e^{-x_k/b}$ produces a poor performance of the system reward, the reward structure facilitates individual agent learning but does not support coordination among the agents, hence produces very noisy estimates for the agents, the variance in the rewards can be reduced using the local reward of eq.1 $g_i(z) = b/x_k$, from simulations it's noted that the performance of eq.1 is better compared to eq.2 due to the fact that eq.2 has a wider tail end, therefore, drops much slowly producing noise in the reward. This can be seen in Figure.1 both the rewards of eq.2, eq.1 are not aligned with the system reward, have poor sensitivity and alignment due to the local nature of the reward structure, this is also verified using the eq.4 and eq.5 during the simulations.

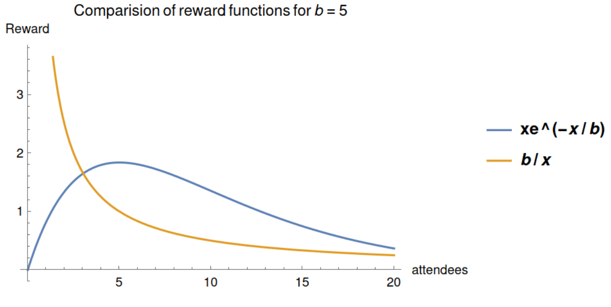

Fig. 1

In the rest of the document, experiments were performed using the exponential reward structure of eq.2 to observe the system performance.

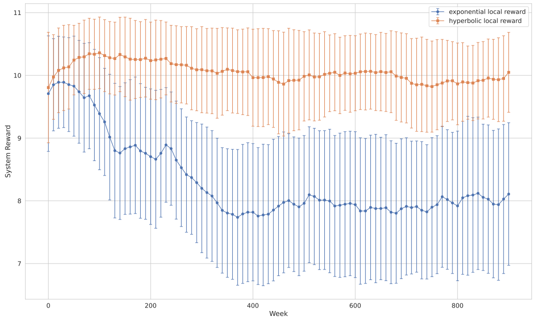

Fig. 2: System reward achieved for 1000 weeks on n=20, k=5, b=7, using the hyperbolic vs exponential reward structure shown in eq.1 and 2 respectively.

B. Problem 2

The concept of difference rewards is introduced in Section 3 eq.7, experiments are conducted on two variants of difference rewards, consider the case when $G(z) = g_i(z)$ where g_i is the local reward, then the reward in eq.9 is called the local difference reward denoted by D_l as shown in eq.9

$$D_l(z) = g_i(z) - g_i(z') \tag{8}$$

$$D_l(z) = x_k e^{-x_k/b} - z' e^{-z'/b} \tag{9}$$

a global counterfactual difference reward structure is defined on the system reward $G(z)$ is shown in eq.11

$$D_g(z) = G(z) - G(z') \tag{10}$$

$$D_g(z) = \sum_{k=0}^{K} x_k e^{-x_k/b} - \sum_{k=0}^{K} z'_k e^{-z'_k/b} \tag{11}$$

where $z' = z_{-i} + c_i$, and c_i is a counterfactual for this state, possible assignments for $c_i = x_k - 1$ and $c_i = \mathcal{U}(1, n)$ where $\mathcal{U}(1, n)$ is uniform distribution over the number of agents in the system. An advantage of both the counterfactuals is the locality of information, i.e. an agent would only need the information that is available to it at the time of computation and would not need any other information. It's observed that using the counterfactual $x_k - 1$ called "good life reward," a world without the agent, yields lesser overall performance compared to randomized counterfactual, this is since the later counterfactual accounts for more variability in the system state than the former, both counterfactuals experience weak sensitivity since the agent's over reward maximization is heavily dependent on the counterfactual state, an alternate choice for counterfactual could also be $c_k = 0$ however upon testing this counterfactual it has not demonstrated any significant performance improvements a counterfactual of $c_i = b$ has also not been of significant improvement, this can be understood as the agent perceives every other day to be occupied and hence wouldn't be able to evaluate its localized reward well. In the simulations performed it is been observed that $random > good_life > zero$ was the observed order of system performance (eq.3). By definition, counterfactual local and global difference rewards of eq.11 and eq.9 are aligned with the system but do not display high sensitivity, they are aligned since the gradient change is directly proportional to the difference $\frac{dg_i(z)}{dz} = \frac{dG(z)}{dz}$, this causes the numerator of eq.4 to be positive in all evaluations. However this may not be sensitive since $G(z)$ is not entirely dependent on the agent state but also considers other agents state, which might affect the sensitivity.

C. Problem 3

Simulation of El-Farol bar using Q-learning has been conducted for two settings n=20, b=5, k=7 and n=50, b=4, k=6, the former setting is a system that is not experiencing congestion, while the latter setting is a system of high congestion. System performance plots using randomized counterfactuals for global difference rewards in both settings are shown in Figure.3 and Figure.4. The advantage of using difference rewards is apparently observed in Figure.4 when the congestion problem has tighter constraints on b, k with a large number of agents n, global difference rewards with randomized counterfactuals produced visibly better performance than other forms of rewards.

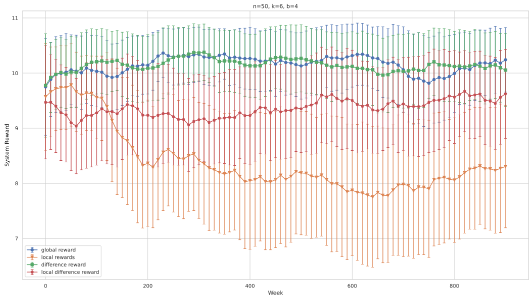

Fig. 3: plot of system rewards for n=20, k=5, b=7 difference rewards can be seen reaching global reward performance

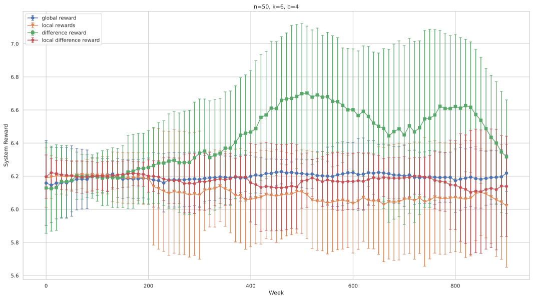

Fig. 4: plot of system rewards for n=50, k=6, b=4 global counterfactual difference rewards perform better than any other reward

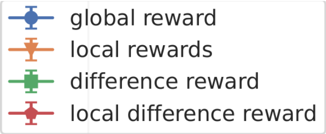

Fig. 5: Legend

The weekly histogram of attendance for all agents in the system, with best performing reward structure i.e. global counterfactual difference rewards, is plotted in Figure.6 for n=20, b=4, k=7 it is clearly observed that the agents have learned to keep the bar optimally occupied for most of the days of the week and the optimal exceeds only at the tail end of the distribution.

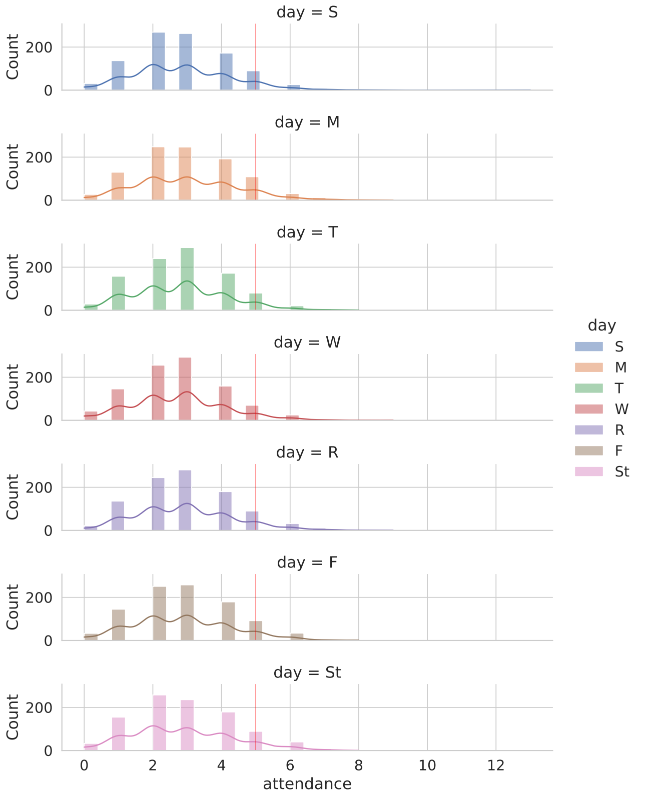

Fig. 6: plot of weekly attendance using the counterfactual global rewards for n=20, k=7, b=5, where the count of a particular number of attendees is shown on the y-axis, and the x-axis shows the number of attendees, using the counterfactual global rewards, red vertical line represents b=5

In the setting when n=50, b=5, k=6, the histogram of weekly attendance for all agents is plotted in Figure.7 in this case, it is observed that the system is congested since the agents attending the bar exceeds the optimal capacity b=5(shown as a red line) most of the time, an interesting observation is many agents choose to attend the bar on a Friday, this would improve the system reward for the agents by most frequently choosing a single night and keeping the bar less occupied for the other days.

A single agent's choice for a night to attend the bar can also be observed over 1000 weeks in Figure.8 here it is observed that the agent chooses to go all days almost equally likely due to the fact that there isn't congestion in the system and it can maximise by reward by going to the bar on all days over different weeks. However, in the case of Figure.9 it is observed that the agent is contributing to maximize

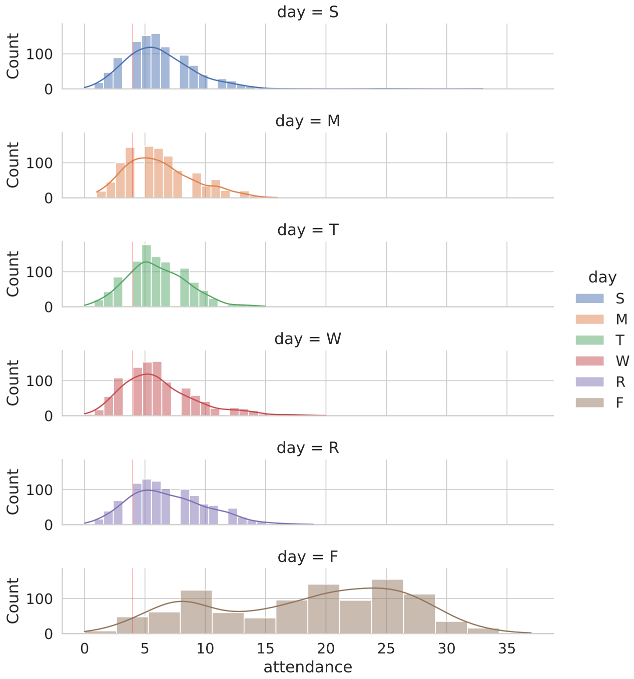

Fig. 7: plot of weekly attendance on a day over 1000 weeks for n=50, k=6, b=4, where the count of a particular number of attendees is shown on the y-axis, and the x-axis shows the number of attendees, using the counterfactual global rewards, the red vertical line represents b=4

the system reward by choosing to attend on a Friday more than any other day as observed in the collective behavior of attendees in Figure.7 which can be seen as a coordinated strategy for the agent since all the agents attending would get a lesser reward but would allow other agents to maximize their reward. It is observed that each agent takes turns attending on a Friday some weeks and increases it's reward by choosing an alternate night in the other weeks.

A randomly selected agent's choices for choosing a night after training across different reward structures are plotted in Figure.11 for n=20, b=5, k=7 which is a non-competitive case where all the reward structures, learn to equally distribute agent's choice over the week, however, from the Figure.11 it clearly observed that counterfactual global difference rewards and global rewards, achieve this better compared to other rewards formats, which see spikes and reductions, on various days. Whereas in the setting of n=50, b=4, k=6 which is seen in the Figure.10 it is observed that counterfactual global difference rewards put the agent valuation on Thursdays(R), more than any other day, this could be due to crowding on Fridays.

D. Effect of learning rate γ

Difference are observed to more robust to changes in the learning rate γ they produce almost similar system performance with a change in $\gamma = [0.3, 0.99]$, whereas the global and local rewards have shown variability to the learning it is observed that the system performance reduces

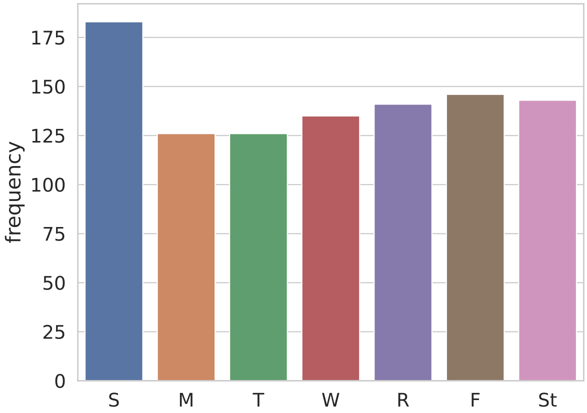

Fig. 8: bar plot of agent attendance in the week for n=20, k=7, b=5, using global counterfactual rewards for 1000 weeks or 7000 nights

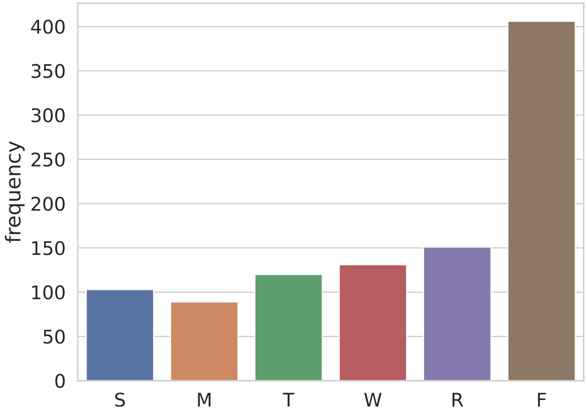

Fig. 9: bar plot of agent attendance in the week for n=50, k=6, b=4, using global counterfactual rewards for 1000 weeks or 7000 nights

as the γ is reduced, which might be accounted for the fact that difference rewards account for a possible alternate future state for maximising the expected reward. However other forms of rewards are local and subject to noise and divergence when γ is perturbed.

IV. Simulation

The code for this assignment is provided in code.zip folder, which runs the simulation, reports factordness and sensitivity of each approach along with other statistics like the system reward average, number of nights where attendance is $> b$ & $<= b$ and the average number of agents attending each night over the weeks(nash equilibrium) and also displays the graph for rewards.

```
python run_problem_3<a/b>.py
```

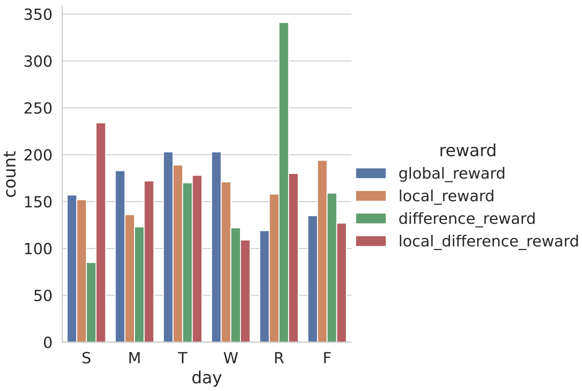

Fig. 10: Histogram of attendance choice for a particular agent in the week for n=50, k=6, b=4, over various rewards for 1000 weeks or 7000 nights

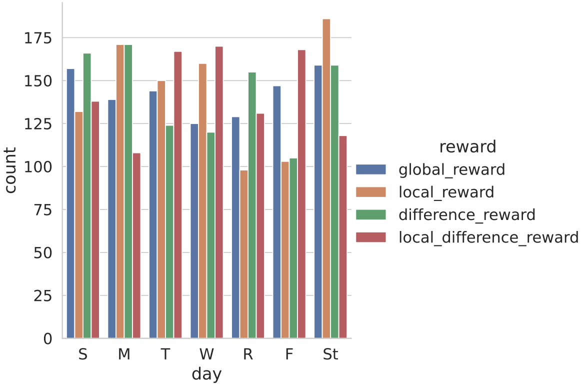

Fig. 11: Histogram of attendance choice for a particular agent in the week for n=20, b=5, k=7, over various rewards for 1000 weeks or 7000 nights

To reproduce the plots use any of the appropriate arguments for the program.

```
python generate_plots.py --histogram
/--boxplot
/--ridgeplot
/--ridgehist
```

by default demo data given in ”saved_files/n20b5k7ṗkl” is loaded and the plots are produced for problem 3a.

V. Conclusion:

Counterfactual global rewards have clearly shown an edge over other formats of rewards, however achieving the optimal system performance, with a Q-Learning-based strategy may not be effective as the number of agents increases in the system, there can be improvements in the reward formulation or learning strategies of the agent for effectively utilising difference rewards.