Agent Coordination and Reward Shaping to solve El Farol Bar problem

Hyuntaek Oh
ohhyun@oregonstate.edu
Due. Oct 29, 2024

I. Introduction

Multiagent systems (MAS) have recently become a critical field of study in artificial intelligence (AI) due to their wide range of applications including distributed control systems. Within MAS, individual agents must act both independently and cooperatively to achieve collective objects, often facing complex interactions and dependencies among each other. Agent Coordination is essential to ensure that agents can work together smoothly, maximizing the system's overall performance and avoiding conflicts that could lead to inefficiencies or even system failures.

One of the core challenges in MAS is designing a framework where agents not only achieve individual goals but also align with the group's goals. In this context, Reward Shaping has emerged as a key technique in reinforcement learning (RL) to guide agents towards desirable behaviors by strategically adjusting their reward functions. Reward shaping helps accelerate the learning process by providing intermediate rewards, enabling agents to converge on optimal policies more efficiently.

El Farol Bar Problem is a classic example in game theory and multiagent systems. It illustrates the dynamics of decision-making in situations where the benefit of an individual's choice, which is whether to go a bar or not, depends on the choices of others, a setup known as congestion game [1]. If the bar attendance is below a certain threshold, everyone who attends has a rewarding experience. On the other hand, if attendance exceeds this threshold, those who attend have a relatively less reward than others. Each agent has limited information and does not know in advance how many others will go to the bar each week. Each of the agents uses strategy based on previous attendance data to estimate whether the bar will be crowded or not.

In this problem, understanding agent coordination and reward shaping in congestion games are crucial since they uncover how individual actions influence collective outcomes in the shared environment, possibly competing for limited resources. Effective coordination helps optimize overall system efficiency and enhance each agent's experience, promoting balanced, cooperative behavior. Reward shaping can significantly impact system performance by guiding agents toward desirable behaviors more efficiently. It helps agents learn to coordinate effectively by reinforcing actions that align with system objectives, rather than just individual gains.

In this paper, we examine the influence of reward mechanisms on agent coordination to assess how different reward structures impact the agents' ability to coordinate effectively, in avoiding overcrowded nights and distributing attendance more evenly. Then, we evaluate system reward outcomes and analyze individual agent performance.

II. Method

There are N agents decide each week which of K nights to attend the bar. Each agent's goal is to choose a night that maximize their reward, considering that overcrowding reduces their reward for that night. The system reward $G(z)$ for a week depends on the attendance on each night, calculated as

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

,

where $x_k(z)$ is the number of agents attending on night k and b is a parameter representing optimal bar capacity.

The **Problem 1** in this assignment describes local reward where each agent is rewarded based on the night k they chose to attend the bar:

$$L(z) = x_k(z) e^{-x_k(z)/b}$$

.

This formula implies that an agent's reward for attending a particular night decreases as more agents choose that night, which is associated with overcrowding. Since each agent's reward depends on overall attendance for their chosen night, the local reward indirectly influences agents to avoid overcrowded nights. In terms of sensitivity, $L(z)$ is highly responsive to changes in attendance, as the reward decreases exponentially when $x_k(z)$ exceeds the optimal value determined by b. This sensitivity can incentivize agents to adjust their choices in response to past outcomes, promoting a self-regulating effect where agents gradually avoid nights with historically high attendance.

Given its structure, the local reward is likely to encourage agents to diversify their choices across different nights to avoid overcrowding. However, without an explicit mechanism for aligning individual rewards with the overall system objectives, there is a risk that agents may occasionally overcrowd certain nights. This reward structure may lead to a dynamic equilibrium where agents continually adjust their choices in response to weekly attendance fluctuations, potentially lacking stability in achieving an optimal distribution.

The **Problem 2** in the assignment is about a difference reward for each agent. The difference reward is a type of reward shaping that provides more of a balance between the degree of factoredness and learnability than do the other rewards by incorporating a counterfactual element [2]. The difference reward is designed to be aligned with the system's objectives since it encourages agents to maximize their positive impact on the system, reducing behaviors that might cause congestion. The formula of a difference reward is:

$$D_i \equiv G(z) - G(z_{-i})$$

, where D_i is a difference reward, and $G(z_{-i})$, a counterfactual term, is a system reward without the agent i's attendance. $G(z_{-i})$ is computed as if agent i had not attended the bar that week, measuring the contribution of the agent's absence. The difference reward is aligned closely with the system reward, as it encourages agents to maximize their positive impact on $G(z)$. Since $G(z_{-i})$ does not depend on agent i's states, the sensitivity of D_i to individual actions is high [2]. The difference reward provides agents with local information relative to the system, as each agent receives feedback on their contribution to $G(z)$.

In this problem, a good counterfactual c_i is the absence of the night the agent i chose. Each agent has K-dimensional action-value vector consisting of one or zero that reflects their attendance, so c_i can be a value of '-1' to represent the night agent i did not attend in the total weekly attendance.

The action-value learning process used by each agent is designed to help agents maximize their rewards over time by refining their attendance choices based on past experiences. Each agent begins with a K-dimensional action-value vector initialized to zero. This vector Q-values represents the agent's estimate of the expected reward for attending each night. Initially, without experience, all night choices are considered equal.

Each agent selects their actions based on ϵ-greedy policy. With probability ϵ, the agent explores by arbitrarily selecting a night, allowing it to discover new potential rewards. Otherwise, the agent exploits by choosing the night with the highest current Q-value, representing the night it currently expects will yield the highest reward.

Each agent updates the Q-value for the selected night using a simple value update formula:

$$Q(night) \leftarrow Q(night) + \alpha \times (reward - Q(night))$$

, where α is the learning rate, and the reward is the local reward received for attending the selected night. This update adjusts each agent's action-value estimates based on the reward experience, yielding higher rewards and reducing preference for less rewarding options.

For training, the process above repeats each week, allowing agents to refine their attendance decisions. Over time, agents learn to favor nights with higher expected rewards, maximizing overall system performance.

III. Experiments and Results

The purpose of experiments is to evaluate how different three reward mechanisms (system rewards, local rewards, and difference rewards) influence agent coordination, system performance, and individual agent rewards in the EL Farol Bar problem.

A. Experimental setup

There are two different parameter set: Case A and Case B. In the Case A, the number of agents is 25, $b = 5$, and $k = 7$. Case B has 40 agents, $b = 4$, and $k = 6$. The learning parameter α is 0.9, and the ϵ is 0.1 respectively used for the action-value learning process.

B. Results

In this section, we explore results of each case that has different parameter set.

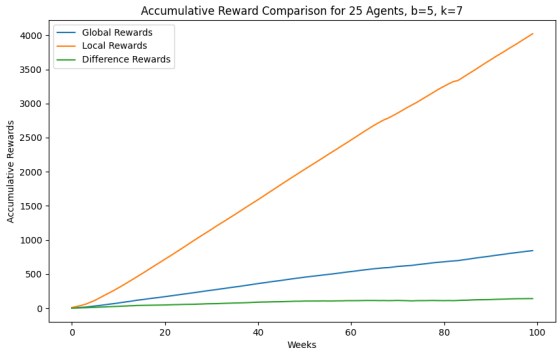

Fig. 1: Accumulative three different rewards

1) Case A: As can be seen in Fig. 1, local rewards accumulate significantly faster than the other types, reaching over 4,000 by the end of simulation. This indicates that agents focusing on maximizing their local rewards are consistently attending nights where they individually receive substantial rewards regardless of the overall system impact. The sharp increase implies that agents are acting in a way that maximizes their immediate individual rewards, leading to competition and overcrowding. The system rewards show a slower but steady increase over time, accumulating to around 800 by the end of the simulation. This trend indicates that agents make decision to provide moderate benefits to the overall system. As compared to local rewards, it suggests that individual agents may not be properly contribute to the system reward. The difference rewards show the slowest growth, remaining nearly flat with minimal accumulation over time. The lack of growth implies that agents are not receiving significant rewards based on difference rewards.

As shown in Fig. 2, the global rewards show significant fluctuation throughout the simulation, oscillating between 8 and 11. There are frequent peaks and troughs, indicating

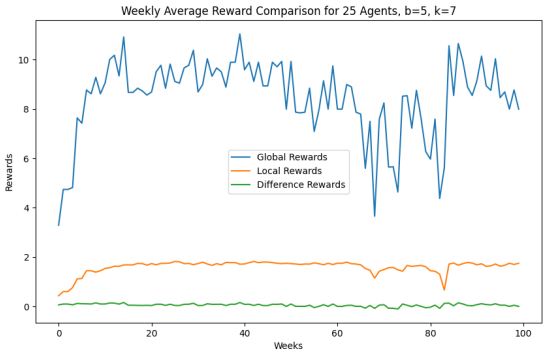

Fig. 2: Mean of three different weekly rewards

variability in system performance. The overall trend remains relatively high, and the system rewards generally effective at guiding agents towards decisions that, on average, provide decent rewards. The local rewards show a steady, stable trend around an average of 2. While stable, the lower average of local rewards, compared to global rewards, suggests that agents may not fully benefit from collective coordination. The difference rewards appear ineffective in this figure, displaying that the counterfactual approach used may not be yielding sufficient differentiation.

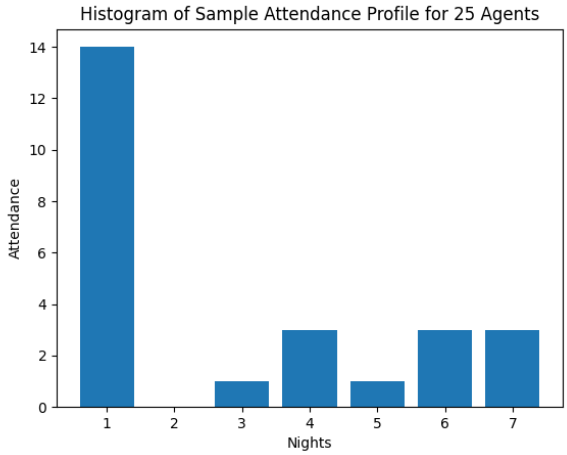

Fig. 3: a histogram of sample attendance in the second week

Fig. 3 shows a histogram of sample attendance in the second week. 14 agents (out of 25) are choosing to attend on Night 1, $k = 1$. This concentration suggests that many agents have a strong preference for this night due to not fully trained state. Attendance on other nights is much lower, with some nights receiving fewer than 2 agents. The heavy concentration on a single night and under-utilization of other nights suggest that agents are not effectively coordinating their attendance choices.

As can be seen in Fig. 4, unlike the second week, the attendance distribution across seven nights for 25 agents is more even distribution. The shift from a heavily concentrated attendance on a single night to a more distributed profile

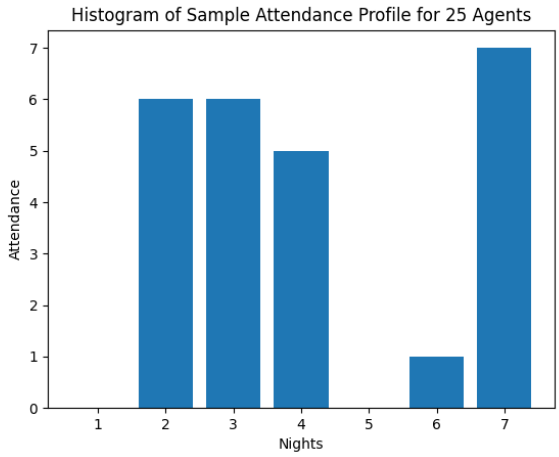

Fig. 4: a histogram of sample attendance in the last week

implies that the agents have improved in coordination over time. This may be a result of agents adapting to the rewards and learning to avoid highly popular nights. Despite the improved spread, Nights 1, 5, and 6 are still not effectively utilized.

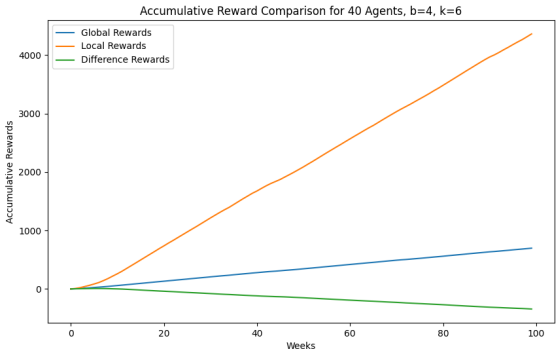

Fig. 5: Accumulative three different rewards

2) Case B: Fig. 5 displays accumulative three different rewards. Similar to Case A, the local rewards accumulate rapidly, reaching over 4,000 by the end of the simulation. This steep rise suggests that agents focusing on local rewards to achieve high individual rewards by attending the bar. The system rewards accumulate at a slow, steady rate, ending around 1,000. This means that agents are making decisions that contribute to the system but may not be fully optimized due to possible poor coordination. The difference rewards remain very low, even flat, indicating minimal impact on agent behavior.

As shown in Fig. 6, The system rewards start at a low point, increase quickly, and then stabilize around 7 to 8 weekly average reward, with occasional drops. This stability indicates that agents are generally contributing positively to the system. The local rewards are consistent, maintaining a stable average around 1.5, meaning that individual agents optimize for consistent rewards, regardless of system perfor-

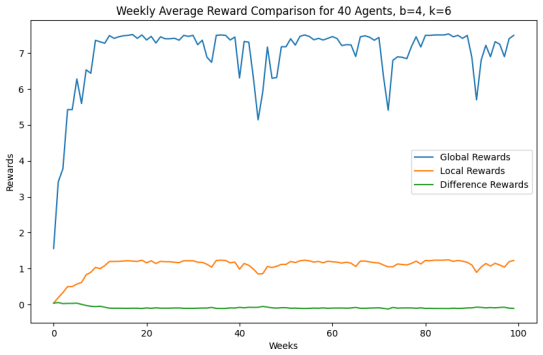

Fig. 6: Mean of three different weekly rewards

mance. The difference rewards remain close to zero, with small fluctuation.

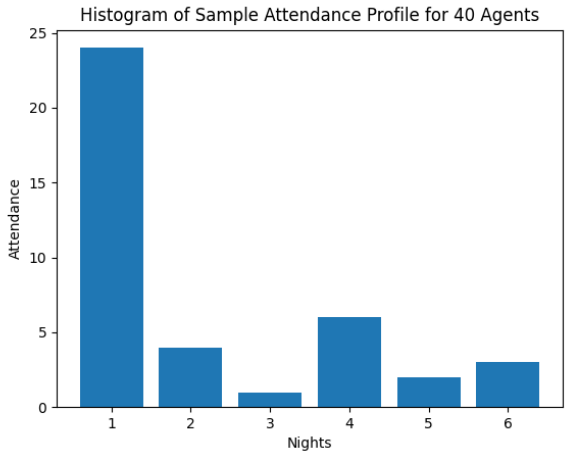

Fig. 7: a histogram of sample attendance in the second week

Similar to Case A's initial profile, in Fig. 7, many agents attend Night 1. This concentration indicates a lack of initial coordination, causing overcrowded. In addition, Nights 2 to 6 show much lower attendance, with fewer than 10 agents distributed across these nights. These observation derives from partial trained state as like Case A.

Fig. 8 shows the attendance profile in the last week. Unlike the trend in Fig. 7, the attendance is more balanced across nights, with each night receiving between 5 and 8 agents. This even spread suggests that agents have learned to distribute their attendance, adapting to avoid overcrowding and improve individual and system rewards.

IV. Conclusion

In this assignment, we investigated the effects of different reward mechanisms (system rewards, local rewards, and difference rewards) on agent coordination and system performance in the El Farol Bar problem. Our goal was to evaluate which reward structures best support coordinated behavior among agents, improve system reward outcomes, and align

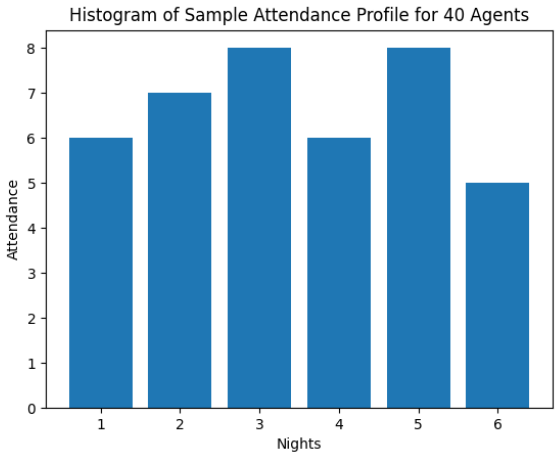

Fig. 8: a histogram of sample attendance in the last week

individual actions with the system's collective objectives. The experiments showed local rewards consistently resulted in high accumulative rewards for individual agents, but they selected the night overcrowded in the early stages of both cases. System rewards provided a moderate, stable accumulative reward. While not as high as local rewards, system rewards encouraged some degree of coordinated behavior, improving the distribution of agents across nights. Difference rewards showed minimal accumulative and average values in both cases. The results of the experiments also exhibited coordination improvement over time from the second week to the last week.

References

[1] W. B. Arthur, "Inductive reasoning and bounded rationality," vol. 84, no. 2, pp. 406–411.

[2] A. K. Agogino and K. Tumer, "Analyzing and visualizing multiagent rewards in dynamic and stochastic domains," vol. 17, no. 2, pp. 320–338.