ROB 538
Multiagent Systems
Homework 3: Agent Coordination and Reward Shaping

Timothy Connor (TC) Drury
Oregon State University
Corvallis, OR 97331
druryt@oregonstate.edu

***Abstract*—This paper presents a multiagent learning framework inspired by Arthur's El Farol Bar problem, a classic congestion game, to investigate the performance of reward-based strategies in decentralized environments. This work on the Bar problem focuses on simulating the behavior of agents choosing among multiple nights with the goal of optimizing individual and system rewards while avoiding overcrowding. We compare three reward mechanisms: system-wide rewards, local rewards, and difference rewards. The simulation examines agent behavior and reward outcomes over time, incorporating exploration, exploitation, and reward shaping. Through analysis of agent coordination, we observe how the different reward structures impact overall attendance patterns, reward distribution, and agent decision-making. This assignment contributes to understanding of reward dynamics in multiagent systems and highlights the importance of reward structuring in influencing agent cooperation and system efficiency.**

***Index Terms*—Congestion Games, Decentralized Learning, Agent Coordination, Reward Shaping, Reinforcement Learning**

I. Introduction

This report addresses three key problems that examine reward-based strategies in a multiagent congestion game, inspired by Arthur's El Farol Bar problem. The focus is on evaluating how different reward mechanisms influence agent behavior and overall system performance.

The report addresses the following problems:

1) Problem 1: We consider a simple local reward where each agent is rewarded based on the night they choose to attend. We will discuss the alignment and sensitivity of this reward structure and predict how effectively it encourages agents to avoid overcrowding.

2) Problem 2: We will look at a difference reward for each agent looking at how the effect of that agent attending another night would affect the system. One counterfactual being if the agent attended the night with the max attendance and another the least attendance. We will analyze the alignment and sensitivity of the resulting difference rewards.

3) Problem 3: In this area, we perform a simulation to analyze the performance of different reward structures in the context of a multiagent system attending a bar with limited capacity. The simulation is set up under two distinct scenarios, each characterized by varying numbers of agents, capacities, and nights available for attendance. Specifically, the first scenario consists of 25 agents, a bar capacity of 5, and 7 nights available, while the second scenario features 40 agents, a capacity of 4, and 6 nights.

The primary objective is to evaluate the effects of local rewards, difference rewards, and global rewards on agent attendance patterns and system efficiency. By plotting the performance of these three reward mechanisms, we can gain insights into how each structure influences agent behavior and overall system dynamics. Additionally, we will analyze and discuss the resulting attendance profiles to understand the implications of reward dynamics on agent cooperation and the effectiveness of strategies to mitigate overcrowding.

A. Background

In multiagent systems where agents must make decisions in shared environments with limited resources, reward structures play a critical role in shaping agent behavior. This report focuses on three primary types of rewards: global, local, and difference rewards. Each reward system affects agents' decision-making processes differently, influencing the overall system dynamics and efficiency. Below, we describe each reward type and its associated equation.

1) Global Reward: The global reward structure assigns the same reward to all agents, regardless of their individual actions, based on the overall system performance. In the context of this problem, the global reward $G(z)$ for a given week z is calculated as a function of the total attendance across all nights, $x_k(z)$, where k represents the different nights available. The formula for the global reward is given as:

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

where $x_k(z)$ is the attendance on night k, and b represents the bar's capacity. This structure encourages agents to avoid overcrowding, as the reward diminishes exponentially with

higher attendance.

2) Local Reward: In contrast to the global reward, the local reward structure rewards agents individually based on their chosen night's attendance. Each agent's local reward $L_i(z)$ depends on the number of agents $x_k(z)$ attending night k and is calculated as:

$$L_i(z) = x_k(z) \cdot e^{-\frac{x_k(z)}{b}}$$

where $x_k(z)$ is the attendance on night k chosen by agent i, and b is the bar's capacity. The local reward encourages agents to select less crowded nights, as the reward decreases when attendance increases beyond the capacity threshold.

3) Difference Reward: The difference reward is designed to isolate the contribution of an individual agent to the overall system's performance. It is computed by considering a counterfactual scenario in which the agent does not participate in the system, allowing the agent to see how their action specifically impacts the global reward. The difference reward $D_i(z)$ for agent i is given by:

The difference reward $D_i(z)$ for agent i helps evaluate the agent's contribution to the system's overall performance by comparing the actual outcome to a counterfactual scenario. It is defined as:

$$D_i(z) = G(z) - G(z|x_i = 0)$$

where $G(z|x_i = 0)$ represents the global reward if agent i had not attended any night. This formulation aligns individual incentives with the overall system performance.

We explore two additional counterfactuals to assess how different attendance choices impact the system:

1. Minimum Attendance Night: The counterfactual assumes that agent i attends the night with the fewest attendees. In this case, the counterfactual is represented as $x_i = \min$, reflecting how the system would perform if agent i had chosen the least crowded night. This scenario helps agents understand how avoiding crowded nights affects the overall reward.

$$D_i(z) = G(z) - G(z|x_i = \min)$$

2. Maximum Attendance Night: Here, the counterfactual assumes that agent i attends the most crowded night, represented as $x_i = \max$. This approach assesses the system's performance if agent i had contributed to overcrowding.

$$D_i(z) = G(z) - G(z|x_i = \max)$$

By evaluating these two counterfactuals, agents can better understand how their choices between attending the least or most crowded nights impact the system. This helps analyze the alignment of individual actions with the overall global reward and the sensitivity of agent decisions to different levels of congestion.

Each of these reward structures has distinct implications for agent behavior. The global reward promotes collective coordination but can result in agents failing to recognize their individual contribution. Local rewards focus on individual performance but may lead to overcrowding on popular nights. Difference rewards aim to balance both, guiding agents to make decisions that align personal and system-wide objectives.

II. Methodology

At the core of this simulation is the `run_simulation` method. This method initiates the simulation for a predetermined number of agents, each facing the challenge of attending one of the multiple nights while maximizing their individual and collective rewards. The simulation takes in parameters such as `time_limit`, `num_nights`, `b`, and `epsilon_init`. The `b` parameter is the capacity of the bar, which influences agents' rewards. With each iteration, agents explore their options based on their Q-values, which represent the expected rewards for attending each night. The exploration-exploitation trade-off is managed through an epsilon-greedy strategy, with `epsilon_init=0.05` determining the initial likelihood of exploration versus exploitation. The formal Q-learning update equation used in this simulation is given by:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left(r + \gamma \max_{a'} Q(s',a') - Q(s,a) \right) \quad (1)$$

where $Q(s,a)$ is the Q-value for state s and action a, α is the learning rate set to 0.1, r is the reward (local, global, or difference) received after taking action a in state s which is the attendance, γ is the discount factor set to 0.9, and $\max_{a'} Q(s',a')$ represents the maximum estimated Q-value for the next state s' based on the possible rewards from attending a different night.

The `sample_action` method is employed during each simulation step, sampling their actions based on their learned Q-values. The Q-values are updated continuously, allowing agents to adapt their strategies over time in response to their rewards and the observed attendance patterns of other agents.

To evaluate the simulation outcomes, the `plot_results` and `plot_attendance` method visualizes the average performance throughout the simulation and average attendance. These include average local rewards, global rewards, and difference rewards. The attendance bargraph shows the average with stadard deviation error bars. The rewards and the attendance will provide insights into how agent interactions shape the overall dynamics of the system.

In summary, the simulation will effectively captures the interplay between individual decision-making and collective behavior within a dynamic setting. Through the iterative learning process, agents refine their strategies based on previous outcomes, highlighting the complexities inherent in the decision-making processes in multiagent systems.

III. Results

In this section, we present the results of the simulations conducted for each problem outlined in the methods. The

results include visualizations that illustrate the performance of agents based on different reward systems, their attendance patterns, and the overall dynamics of the environment.

Problem 1: Local Reward Structure

For Problem 1, we analyzed a simple local reward structure where agents are rewarded based on their chosen night to attend the bar. The performance of agents utilizing this local reward structure across 1000 weeks is illustrated in Figure 1.

Fig. 1. Performance of agents using the local reward structure.

Simulation with 25 agents, 5 capacity, 7 nights, and 1000 weeks
Initial system reward: 3.417906841411997
Max system reward: 10.857213821019087
Max system reward at iteration 42
Final system reward: 6.945832019490588

The results indicate that agents' preference for attending nights with higher attendance can lead to a decrease in overall system value. This behavior can be from agents pursuing greater individual rewards associated with busier nights. However, this more personal tendency creates a misalignment between individual incentives and the overall system objective, resulting in suboptimal outcomes. Sensitivity to initial fluctuations is evident, as agents frequently switch their choices in the early stages of the simulation, leading to a high degree of variance before gradually converging to a suboptimal solution over time. This dynamic underscores the challenge of achieving alignment between individual and collective goals, as agents continue to seek immediate rewards without fully accounting for the broader system impact.

Problem 2: Difference Reward Derivation

In Problem 2, we derived a difference reward for each agent based on their performance relative to a counterfactual scenario. Figure 2 illustrates the performance of difference rewards derived from the baseline counterfactual where the agent would not attend any night. Figure 3 shows the performance of difference reward driven agents with a counterfactual where the agent would attend the least crowded night instead. Figure 4 displays the contrast where the difference reward is based on the the agent looking at if they attended the over-crowded night. The performance of agents utilizing the difference reward structure was shown across 1000 weeks.

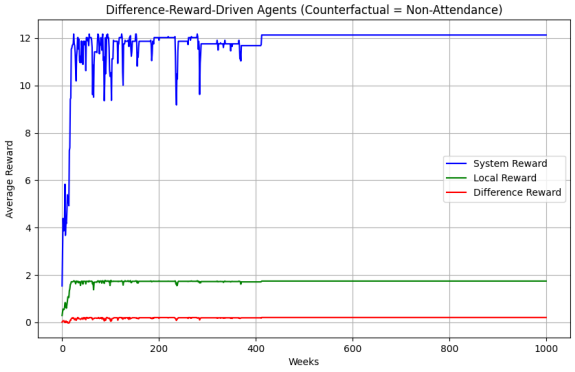

Fig. 2. Analysis of difference rewards with Non-Attendance Baseline Counterfactual.

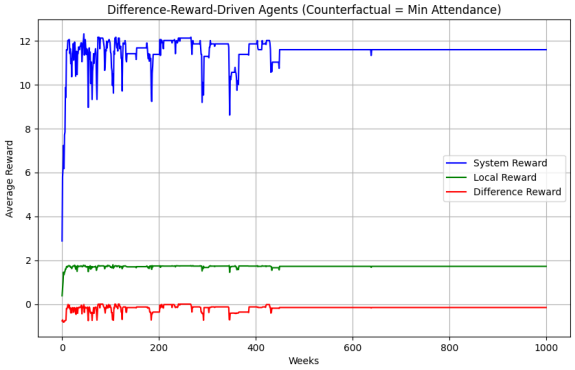

Fig. 3. Analysis of difference rewards with varying Min-Attendance Counterfactual.

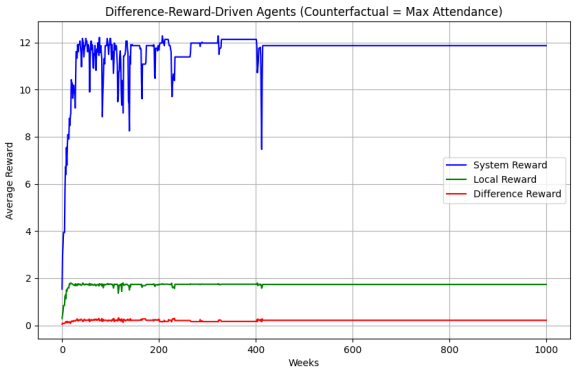

Fig. 4. Analysis of difference rewards with Max-Attendance Counterfactual.

The findings indicate that the minimum counterfactual aligns more closely with global system dynamics, promoting

rewards that benefit the overall system when agents select nights that contribute positively to attendance. This alignment not only encourages agents to consider the collective impact of their choices but also enhances sensitivity to global performance metrics, ultimately leading to decisions that are beneficial for collective outcomes.

In contrast, the maximum counterfactual, while exhibiting anti-alignment with overall system performance, aids agents in making better decisions by increasing sensitivity to the consequences of attending busier nights. This occurs because the maximum counterfactual highlights scenarios where attendance is high, potentially signaling to agents that such choices, although initially attractive, may not be optimal for their long-term rewards.

Overall, these results demonstrate that the minimum counterfactual approach provides a more coherent strategy for agents to navigate the complexities of attendance, balancing individual motivations with the collective welfare of the system. The maximum counterfactual, despite being more aligned than localized rewards, still yields better outcomes than local rewards due to the higher sensitivity than the locally driven agents.

Problem 3: Simulation Results

The simulations for Problem 3 were conducted under two specific scenarios:

- Case 1: 25 agents, $b = 5$, and $k = 7$.
- Case 2: 40 agents, $b = 4$, and $k = 6$.

Figures 5 and 7 present the performance of three agent rewards (G, difference, and local) alongside barographs of attendance profiles for each case in Figures 6 and 8.

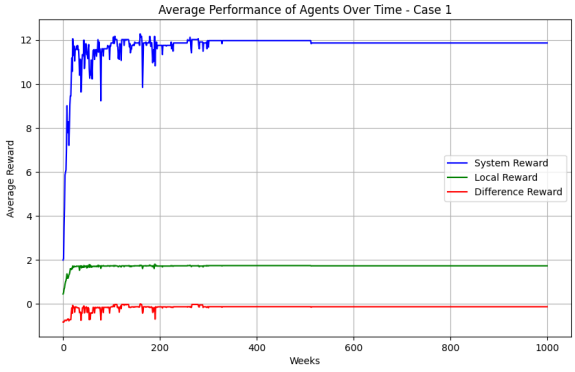

Fig. 5. Performance of agents for Case 1 (25 agents, $b = 5$, $k = 7$).

IV. Conclusion

For both sets of parameters, (25 agents with b=5 and k=7) and (40 agents with b=4 and k=6), we evaluated the performance of three reward structures: global, difference, and local rewards.

Global Rewards: In both scenarios, global rewards produced more consistent system-wide optimization. Agents under this

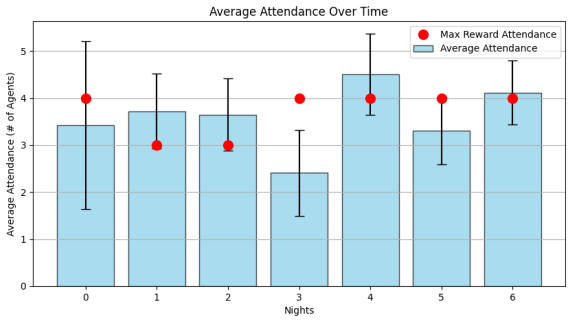

Fig. 6. Average attendance of agents for Case 1 (25 agents, $b = 5$, $k = 7$).

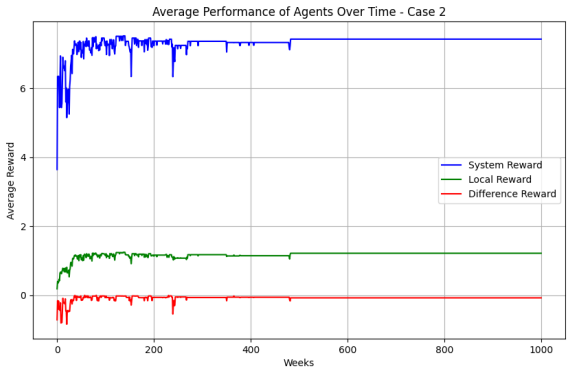

Fig. 7. Performance of agents for Case 2 (40 agents, $b = 4$, $k = 6$).

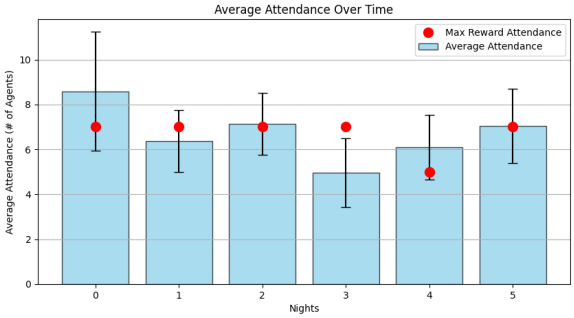

Fig. 8. Average attendance profiles for Case 2 (40 agents, $b = 4$, $k = 6$).

reward structure generally converged toward attendance profiles that maximized overall system performance. This was reflected in relatively stable outcomes, with fewer large fluctuations in individual behavior, leading to a higher system reward. However, some variance in attendance was still observed, especially at the beginning of the simulation.

Difference Rewards: Difference rewards led to improved outcomes compared to local rewards, as agents received more tailored feedback based on their impact on the system. When the counterfactual scenario (what would have happened if the agent had taken a different action) was better than the actual chosen action, agents received a negative difference

reward, which served as corrective feedback. This allowed agents to adjust and improve their future decisions. In both cases, the difference rewards promoted better alignment between individual agent actions and overall system objectives. Agents avoided overly crowded nights while still benefiting from their attendance choices, reducing system inefficiencies. The alignment was clearer in the 25-agent case, where there was more room to spread out across nights. However, as the number of agents increased (40 agents), the sensitivity of difference rewards became more pronounced, with agents adjusting their behavior more dynamically.

Local Rewards: The local rewards performed the worst in both cases. Agents driven by local rewards showed higher sensitivity to immediate attendance fluctuations, with little regard for the overall system value. This misalignment often led to agents overcrowding certain nights, especially in the scenario with 40 agents, where the smaller capacity ($b = 4$) made it harder for agents to avoid congestion. As a result, local rewards led to suboptimal outcomes for the system, with lower overall system rewards and greater variability in attendance profiles.

Overall, these findings emphasize the importance of reward structure design in multiagent systems. Global rewards offer the most effective alignment between individual actions and collective objectives, ensuring optimal system-wide outcomes. Difference rewards help mitigate some of the misalignment seen in local rewards by providing counterfactual feedback, but still fall short of the system-wide performance achieved under global rewards. Local rewards, though effective in promoting short-term individual gains, ultimately lead to suboptimal outcomes when individual incentives are misalignment with global system objectives.

V. Acknowledgments

I would like to acknowledge the assistance of ChatGPT, an AI language model developed by OpenAI, for providing support in refining grammar and enhancing sentence structure throughout this document [1].

References

[1] OpenAI, "ChatGPT: A language model for conversational AI," 2023. [Online]. Available: https://www.openai.com/chatgpt. [Accessed: -Oct-2024].

[2] Devlin, S., Yliniemi, L., Kudenko, D., & Turner, K. (2014). Potential-based difference rewards for multiagent reinforcement learning. In *13th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2014* (pp. 165–172). International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS).