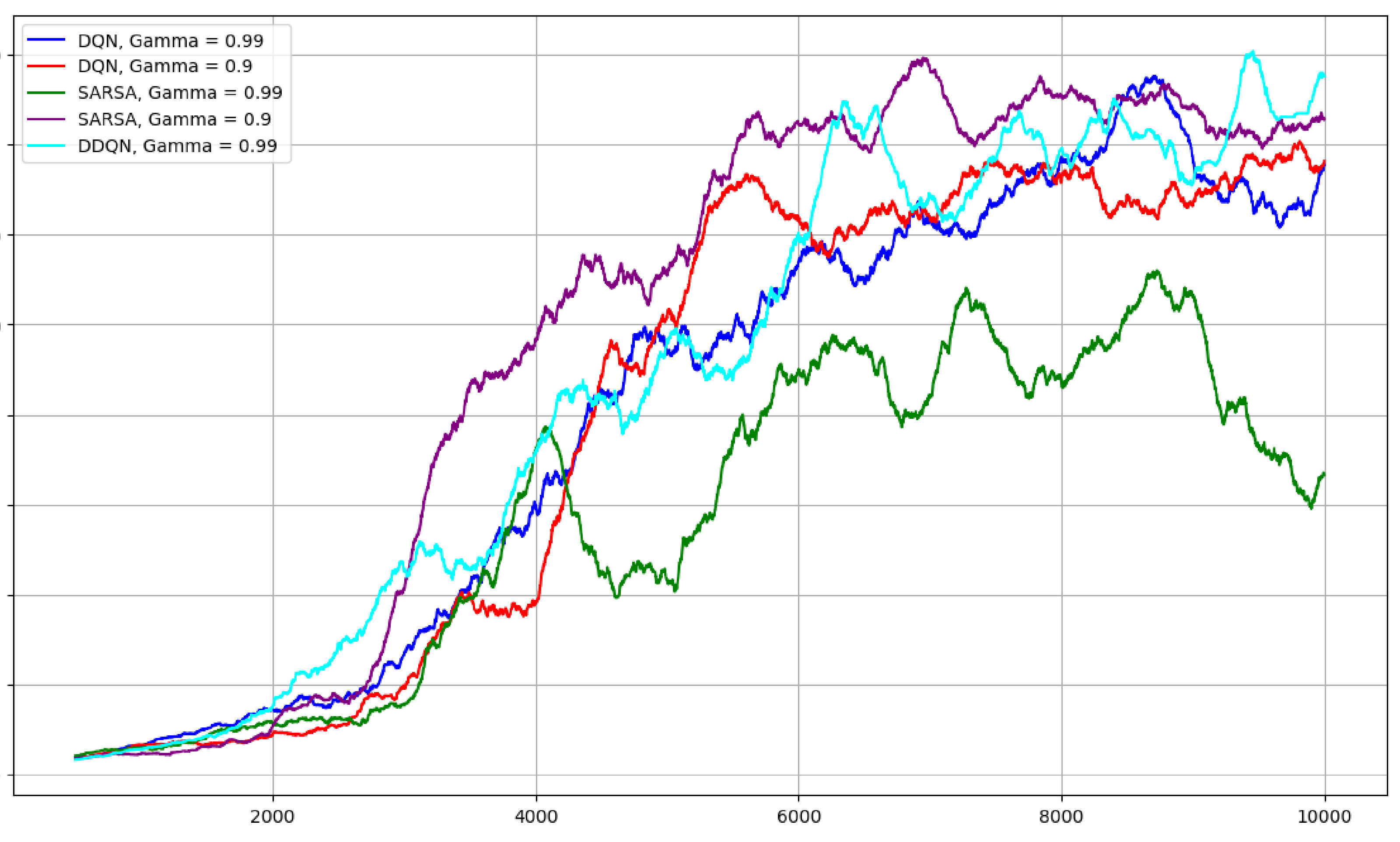
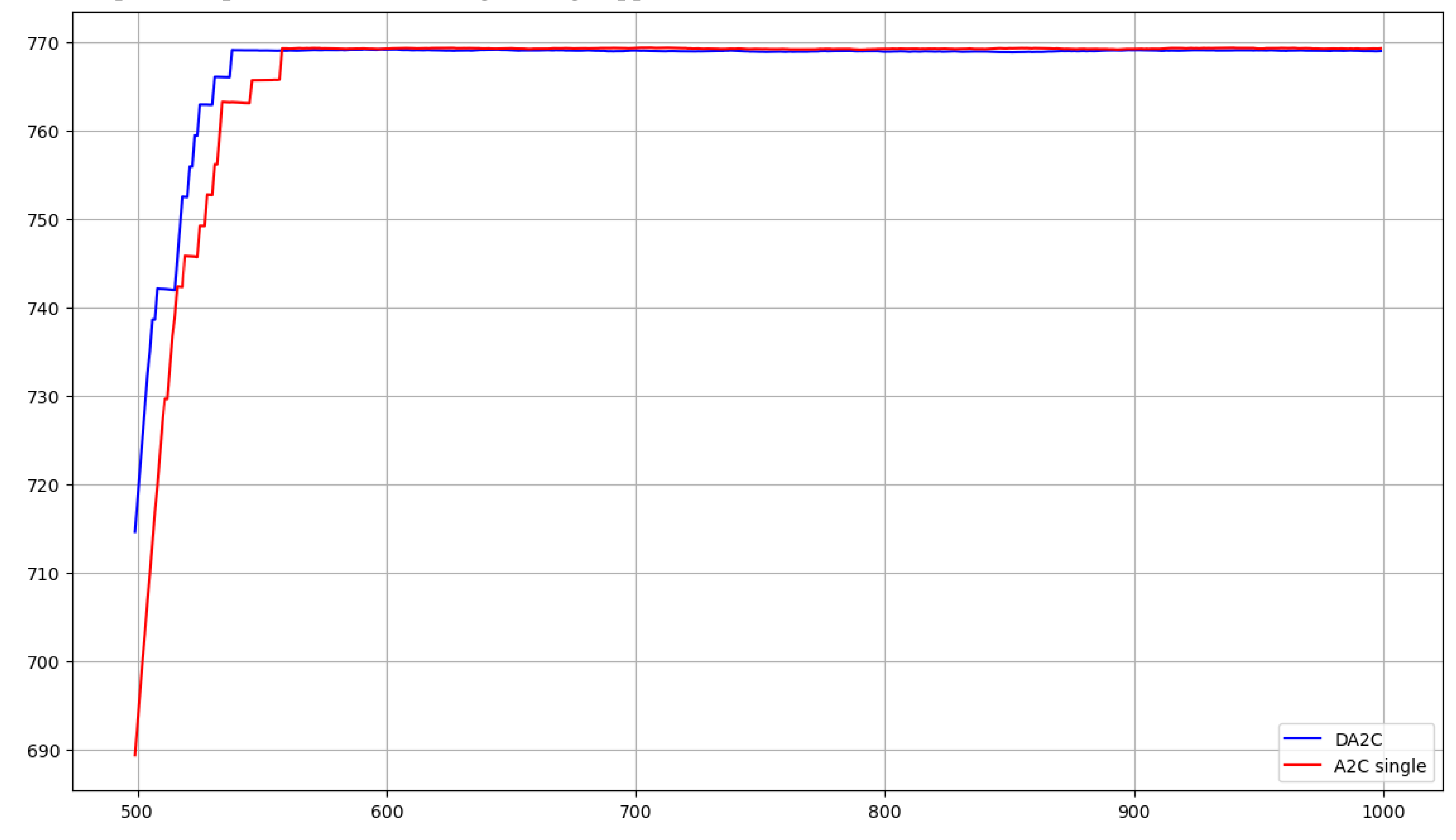
DRL-Based Movement Learning for UAV Base-Stations

Student: Mohammadsadra Rastravesh
Prof. Hamed Kebriaei
ECE Department, University of Tehran

Results



Results of Value-Based Algorithms. The figure illustrates the episode rewards versus the number of episodes, plotted with a moving average applied to smooth the reward curve.

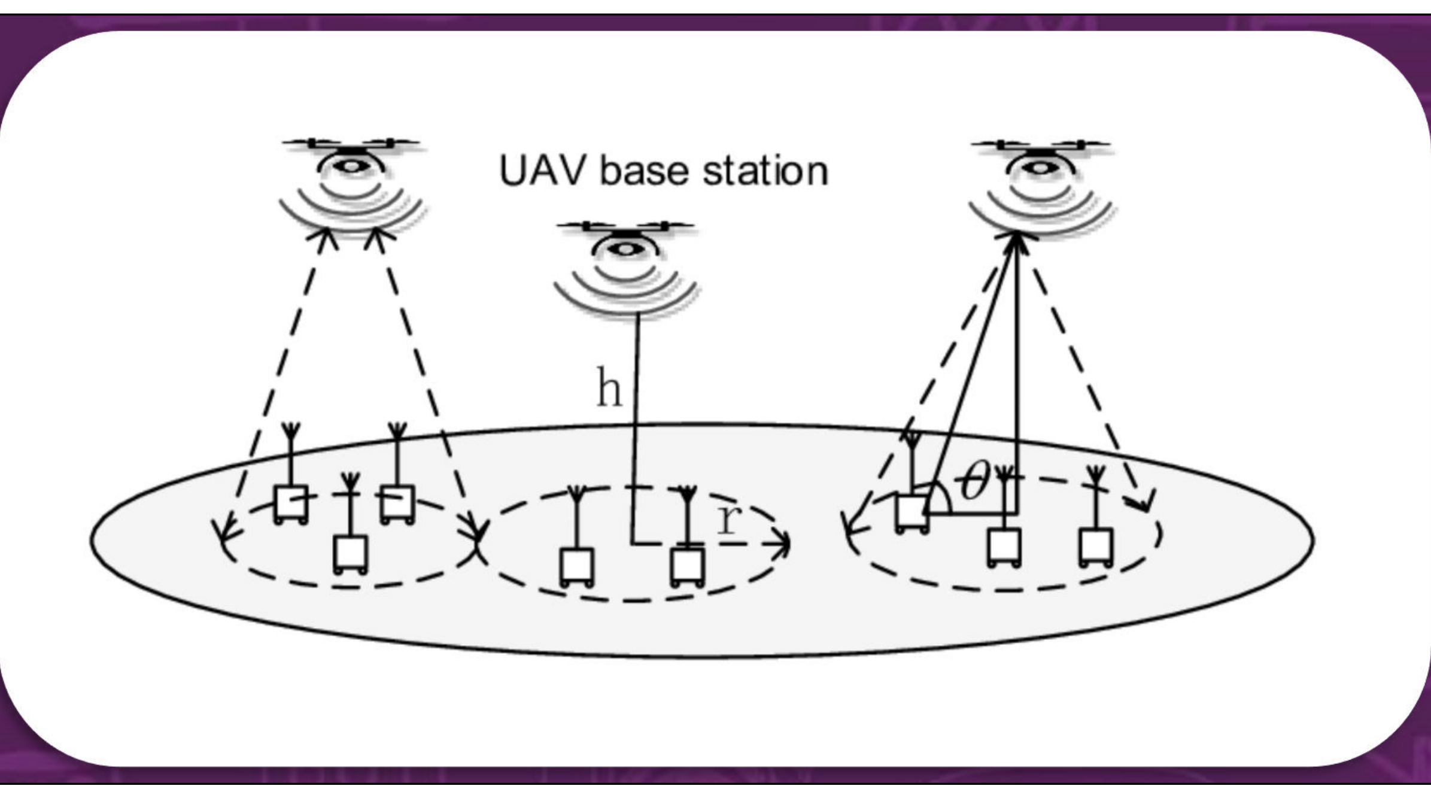


Results of Policy-Based Algorithms. The plot shows the episode rewards as a function of the number of episodes, with a moving average applied to highlight the overall learning trend.

Project Summary

With the commercialization and mass production of quadcopters, new applications are continuously emerging. One such application is employing them as aerial radio stations. In this role, quadcopters can be deployed in various scenarios, such as areas experiencing main service outages, regions with temporary high traffic, or zones requiring temporary network setup. The objective in these problems is to maximize the weighted service rate for users, where the weights represent each user's service history and are designed to ensure fairness in resource allocation. The decision variables include both the resource allocation strategy and the position of each quadcopter at specific time intervals.

In this project, we aimed to solve the optimization problem using reinforcement learning, considering both single-agent and multi-agent frameworks. The scenario includes five stationary users and two mobile base stations (quadcopters).

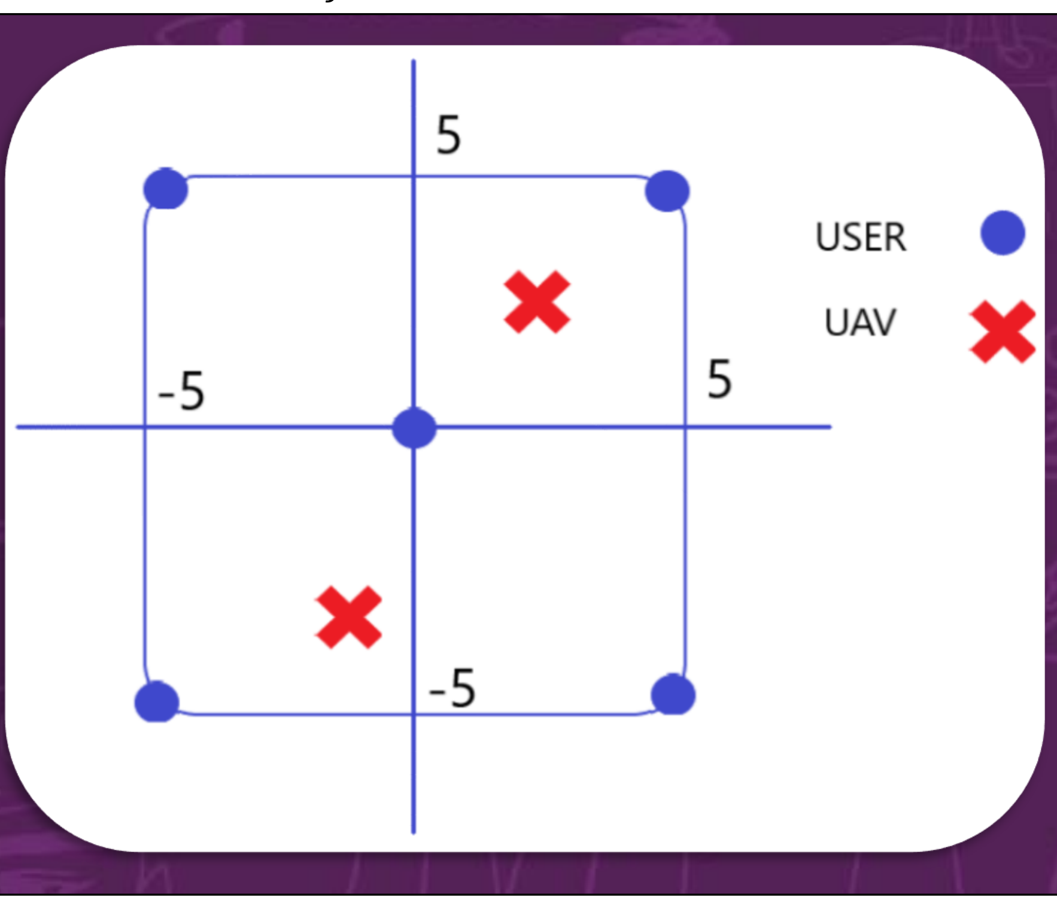


$$\max_{q} \sum_{i=1}^{M=2} \sum_{j=1}^{K=5} \alpha_{ij} K_j \log_2 \left(1 + \frac{\frac{\rho_0}{H^2 + \left\| q_i - w_j \right\|_2^2}}{\sum_{\substack{l=1 \\ l \neq i}}^{M} \frac{\rho_0}{H^2 + \left\| q_l - w_j \right\|_2^2} + \sigma^2} \right)$$

(1) $\|q_i - q_{i0}\|_2^2 \leq {S_{max}}^2$: Max displacement Constraint

(2) $\|q_i - q_i\|_2^2 \geq {d_{min}}^2$: Safety Constraint

H: *constant Height of Stations*, ρ_0: *reference power of a station, here* 1
w_j: *location of jth user*, q_i: *location of ith station*, K_j: *priority of jth user*
σ^2: *thermal noise, here* 0.01, α_{ij}: *connectivity between user j and station i*, 0 *or* 1



The workspace is a 10×10 square in which agents (base stations) fly at a fixed altitude. Each agent has five discrete actions: move 2 units left, right, up, or down, or remain stationary. If an action would cross the boundary, the agent is projected back inside. Because agent motion is bounded, Constraint 1 is non-binding. Rewards are assigned according to the optimization objective; collisions or violations of the safety buffer (set to 5 meters) incur a 1,000-unit penalty.

Single-Agent Setting. Actions for both agents are chosen jointly, yielding 25 joint actions, and a single Q-function is learned over the joint state–action space. We evaluated DQN, Deep SARSA, and A2C.

Multi-Agent Setting. Each agent learns its own function approximator and independently selects among its 5 actions. We experimented with Distributed Q-Learning and fully decentralized MARL with networked agents.

Observations

- Value-Based algorithms require significantly more training time compared to Policy-Based algorithms. Policy-Based methods such as A2C, which use a neural network to directly select actions, adapt to the environment much faster. The difference is substantial: A2C converged in about 3 minutes, whereas SARSA required nearly 9 hours to reach convergence.
- SARSA explores the environment more extensively than Q-Learning and thus converges more slowly. However, once convergence occurs, SARSA generally achieves superior performance and is better suited for non-stationary environments.
- Multi-agent learning algorithms do not necessarily yield worse results than single-agent methods; in fact, they can sometimes achieve even better performance.
- Reducing the discount factor (γ) limits the agent's long-term foresight in maximizing rewards, but it accelerates convergence.

References

1. Wu, Q., Zeng, Y., & Zhang, R. (2018). Joint trajectory and communication design for multi-UAV enabled wireless networks.
2. Qin, Z., Liu, Z., Han, G., Lin, C., Guo, L., & Xie, L. (2021). Distributed UAV-BSs trajectory optimization for user-level fair communication service with multi-agent deep reinforcement learning. IEEE Transactions on Vehicular Technology, 70(12), 12290-12301.
3. Zhang, K., Yang, Z., Liu, H., Zhang, T., & Basar, T. (2018, July). Fully decentralized multi-agent reinforcement learning with networked agents. In International conference on machine learning (pp. 5872-5881). PMLR.