Heuristics-Assisted Experience Replay Strategy for Cooperative Multi-Agent Reinforcement Learning

Extended Abstract

Yi Xie
FAET, Fudan University
Shanghai, China
yixie22@m.fudan.edu.cn

Ziqing Zhou
FAET, Fudan University
Shanghai, China
21110860021@m.fudan.edu.cn

Chun Ouyang*
FAET, Fudan University
Shanghai, China
oy_c@fudan.edu.cn

Siao Liu
FAET, Fudan University
Shanghai, China
saliu20@fudan.edu.cn

Linqiang Hu
FAET, Fudan University
Shanghai, China
18110860018@fudan.edu.cn

Zhongxue Gan*
FAET, Fudan University
Shanghai, China
ganzhongxue@fudan.edu.cn

ABSTRACT

Cooperative Multi-agent Reinforcement Learning (CMARL) has great potential for developing coordinated strategies that optimize team performance. However, common methods often fail to properly separate and utilize individual experiences due to a lack of effective team reward decomposition. The Heuristics-assisted Experience Replay Strategy (HAER) addresses this by decomposing team rewards into individual rewards and enabling efficient experience replay in MARL. By maintaining network gradient invariance, we derive a partial differential equation for the individual reward function, allowing accurate calculation of TD-errors and experience importance. The Cooperative Multi-Objective Swarm Optimization (CMOSO) algorithm is used to balance TD-errors and individual rewards for efficient learning. Extensive experiments on benchmarks demonstrate HAER's effectiveness, with up to a 17.6% performance boost in the homogeneous SMACV2 scenario and an average 8% improvement in GRF for heterogeneous agent cooperation.

KEYWORDS

Multi-Agent Reinforcement Learning, Credit Assignment

ACM Reference Format:
Yi Xie, Ziqing Zhou, Chun Ouyang*, Siao Liu, Linqiang Hu, and Zhongxue Gan*. 2025. Heuristics-Assisted Experience Replay Strategy for Cooperative Multi-Agent Reinforcement Learning: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

Multi-Agent Reinforcement Learning (MARL) enables agents to learn coordinated behaviors for common goals, with agents interacting by taking joint actions, transitioning to new global states, and receiving shared team rewards [9, 10, 16, 21]. Effective coordination is essential for maximizing cumulative team rewards [5], and exploration is a common approach in MARL [4, 8, 19]. However, relying solely on exploration can be inefficient, particularly in identifying actions leading to sparse rewards [12, 14].

Reward decomposition, which breaks down the team reward into individual contributions, is crucial for coordination [22], but sparse or uneven rewards complicate this process [7, 13]. Existing methods often rely on domain-specific rules [18], limiting generalizability across tasks [17]. Recent approaches, such as MASER [3] and DIFFER [2], offer improvements through subgoals and reward decomposition, yet they often fail to address the challenges posed by heterogeneous agents with different capabilities, which can affect reward decomposition and generalizability [1, 20].

To address the challenges in existing methods, we perform reward decomposition while preserving network gradient invariance, allowing the extraction of individual temporal difference (TD) errors and rewards. We then construct an individual experience replay buffer, denoted as $\chi_j^{\text{ind}} = \left(s_t^i, a_t^i, r_i, s_{t+1}^i, \delta_i, \sigma_i, H_i\right)$, where key factors like δ_i, σ_i, H_i, and r_i guide experience selection to optimize learning efficiency, diversity, and fairness. To balance exploration and exploitation, especially in heterogeneous and sparse reward environments, we use a Cooperative Multi-Objective Swarm Optimization (CMOSO) algorithm to optimize experience prioritization. This leads to the HAERS framework, which enhances training efficiency and agent performance in multi-agent systems.

2 PRELIMINARY

The cooperative multi-agent reinforcement learning problem is modeled as a Decentralized Partially Observable Markov Decision Process (DEC-POMDP) [6], defined by $\langle N, \boldsymbol{O}, \mathcal{A}, P, R, \gamma \rangle$, where $N = \{1, 2, \ldots, n\}$ is the set of agents, $\boldsymbol{O}$ is the joint observation space, and $\mathcal{A}$ is the joint action space. The transition function P governs state dynamics, and R is the shared team reward, with γ as the discount factor.

At each step t, each agent i observes o_t^i and selects an action a_t^i based on its policy π^i. The joint action $a_t = (a_t^1, \ldots, a_t^n)$ results in a team reward $R(o_t, a_t)$ and transitions to the next state o_{t+1} based on $P(o_{t+1}|o_t, a_t)$. The objective is to find policies $\{\pi^i\}_{i=1}^n$ that maximize the expected cumulative discounted team reward: $J(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty} \gamma^t R(o_t, a_t)\right]$.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

3 METHOD

In DEC-POMDP, individual rewards are not directly observable. To compute individual rewards while maintaining gradient invariance, we introduce the Reward Decomposition Module (RDM) within the Actor-Critic architecture. The individual reward r_i is derived from the team reward and value functions as:

$$r_i = \left(R + \gamma \tilde{V}_c - V_c\right) \frac{\partial V_c}{\partial V_i} - \gamma \tilde{V}_i + V_i,$$

where R is the immediate team reward, γ is the discount factor, V_c and $\tilde{V}_c$ are the centralized value functions, and V_i and $\tilde{V}_i$ are the local value functions for agent i. Additionally, the following metrics are computed for each agent:

$$\delta_i = r_i + \gamma \tilde{V}_i - V_i, \quad \sigma_i = \left|\tilde{V}_i - V_i\right|, \quad H_i = -\sum_{a \in \mathcal{A}} \pi_{\theta_i}(a|s_t^i) \log \pi_{\theta_i}(a|s_t^i).$$

These metrics help prioritize individual experiences, which are stored in a replay buffer:

$$\chi_j^{\text{ind}} = \left(s_t^i, a_t^i, r_i, s_{t+1}^i, \delta_i, \sigma_i, H_i\right).$$

The Experience Prioritization Module (EPM) evaluates the priority of experiences using the score:

$$e_j = w_1 |r_j| + w_2 |\delta_j| + w_3 \sigma_j + w_4 H_j + \epsilon.$$

To optimize the weighting factors w_1, w_2, w_3, w_4, the Cooperative Multi-Objective Swarm Optimization (CMOSO) algorithm is used. In this algorithm, the particles $\mathbf{x}_1$ and $\mathbf{x}_2$ represent potential solutions for the weighting factors. These particles evolve over time according to the update equations:

$$\mathbf{x}_1(t+1) = \mathbf{x}_1(t) + \eta \left[\mathbf{x}_2(t) - \mathbf{x}_1(t)\right] + \sigma(t) r_1 \Delta_2'(t),$$

$$\mathbf{x}_2(t+1) = \mathbf{x}_2(t) + \eta \left[\mathbf{x}_1(t) - \mathbf{x}_2(t)\right] + \sigma(t) r_2 \Delta_1'(t),$$

where η is the learning rate, and $\sigma(t)$ controls the exploration-exploitation balance. The values r_1 and r_2 are random variables, which introduce stochasticity to the search process, helping the algorithm explore different regions of the solution space. The terms $\Delta_1'(t)$ and $\Delta_2'(t)$ are the adjusted velocities of the particles, determining how far each particle moves in each update.

4 EMPIRICAL RESULTS

We test HAERS in both homogeneous and heterogeneous agent settings, using GRF and SMAC as benchmark environments, demonstrated in Figure1. The comparison algorithms include MAPPO [15], known for its stable policy updates and efficiency in MARL, and HAPPO [20], a trust region learning method suitable for adaptive agents in dynamic, collaborative environments. Additionally, we compare against A2PO [11], a sequential policy update method that has demonstrated strong performance across cooperative MARL benchmarks, and DIFFER [2], which focuses on reward decomposition and experience replay strategies, to evaluate their effectiveness in handling sparse rewards and uneven agent capabilities.

4.1 Hyperparameter Sensitivity Analysis

We examine the hyperparameter optimization performed by CMOSO and assess whether its convergence positively impacts the overall performance of HAERS. We visualize the training trends of

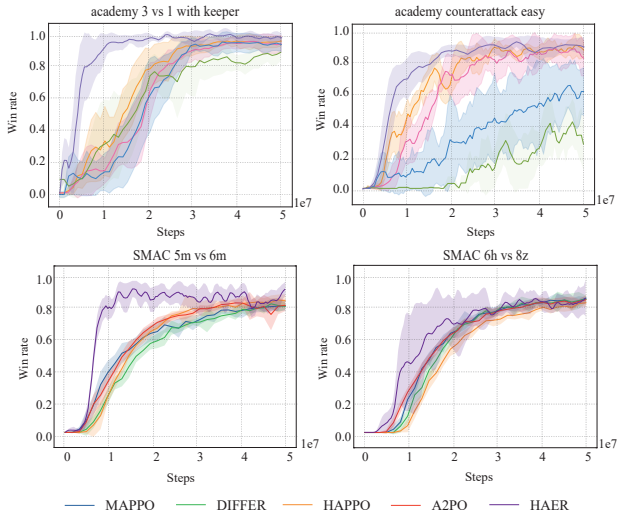

Figure 1: Performance comparison across GRF and SMAC scenarios for A2PO, MAPPO, HAPPO, DIFFER and HAERS.

w_1, w_2, w_3, w_4 over the training steps in GRF, which demonstrates that CMOSO effectively improves the performance of HAERS.

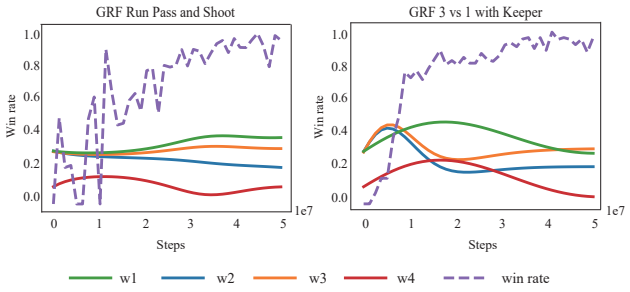

Figure 2: Hyperparameter evolution and performance in GRF: Values of w_1, w_2, w_3, w_4 across training steps.

5 CONCLUSION

This paper presents HAERS, a self-adaptive reward decomposition method for addressing sparse and uneven reward distributions in MARL. By maintaining gradient invariance, we derive a partial differential equation for the individual reward function, enabling accurate TD-error calculation and experience evaluation. Individual experience selection is framed as a multi-objective optimization problem, and a suitable algorithm balances exploration and exploitation. Experiments in benchmark environments show HAERS's effectiveness in both homogeneous and heterogeneous scenarios with sparse rewards, improving MAPPO by 12.4% in homogeneous and 10.1% in heterogeneous GRF scenarios, thus enhancing team performance through optimized reward allocation.

ACKNOWLEDGMENTS

This study was partially supported by Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0103)

REFERENCES

[1] Matteo Bettini, Ajay Shankar, and Amanda Prorok. 2023. Heterogeneous Multi-Robot Reinforcement Learning. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*. 1485–1494.

[2] Xunhan Hu, Jian Zhao, Wengang Zhou, Ruili Feng, and Houqiang Li. 2024. DIFFER: Decomposing Individual Reward for Fair Experience Replay in Multi-Agent Reinforcement Learning. *Advances in Neural Information Processing Systems* 36 (2024).

[3] Jeewon Jeon, Woojun Kim, Whiyoung Jung, and Youngchul Sung. 2022. Maser: Multi-agent reinforcement learning with subgoals generated from experience replay buffer. In *International Conference on Machine Learning*. PMLR, 10041–10052.

[4] Jiahui Li, Kun Kuang, Baoxiang Wang, Xingchen Li, Fei Wu, Jun Xiao, and Long Chen. 2024. Two heads are better than one: a simple exploration framework for efficient multi-agent reinforcement learning. *Advances in Neural Information Processing Systems* 36 (2024).

[5] Jiahui Li, Kun Kuang, Baoxiang Wang, Furui Liu, Long Chen, Changjie Fan, Fei Wu, and Jun Xiao. 2022. Deconfounded value decomposition for multi-agent reinforcement learning. In *International Conference on Machine Learning*. PMLR, 12843–12856.

[6] Michael L Littman. 1994. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*. Elsevier, 157–163.

[7] Boyin Liu, Zhiqiang Pu, Yi Pan, Jianqiang Yi, Yanyan Liang, and Du Zhang. 2023. Lazy agents: a new perspective on solving sparse reward problem in multi-agent reinforcement learning. In *International Conference on Machine Learning*. PMLR, 21937–21950.

[8] Xiangyu Liu and Ying Tan. 2022. Feudal latent space exploration for coordinated multi-agent reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems* 34, 10 (2022), 7775–7783.

[9] Boda Ning, Qing-Long Han, Zongyu Zuo, Lei Ding, Qiang Lu, and Xiaohua Ge. 2023. Fixed-Time and Prescribed-Time Consensus Control of Multiagent Systems and Its Applications: A Survey of Recent Trends and Methodologies. *IEEE Transactions on Industrial Informatics* 19, 2 (2023), 1121–1135. https://doi.org/10.1109/TII.2022.3201589

[10] Huaze Tang, Hengxi Zhang, Zhenpeng Shi, Xinlei Chen, Wenbo Ding, and Xiao-Ping Zhang. 2023. Autonomous Swarm Robot Coordination via Mean-Field Control Embedding Multi-Agent Reinforcement Learning. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 8820–8826.

[11] Xihuai Wang, Zheng Tian, Ziyu Wan, Ying Wen, Jun Wang, and Weinan Zhang. 2022. Order Matters: Agent-by-agent Policy Optimization. In *The Eleventh International Conference on Learning Representations*.

[12] Haolin Wu, Hui Li, Jianwei Zhang, Zhuang Wang, and Jianeng Zhang. 2021. Generating individual intrinsic reward for cooperative multiagent reinforcement learning. *International Journal of Advanced Robotic Systems* 18, 5 (2021), 17298814211044946.

[13] Kai Yang, Zhirui Fang, Xiu Li, and Jian Tao. 2024. CMBE: Curiosity-driven Model-Based Exploration for Multi-Agent Reinforcement Learning in Sparse Reward Settings. In *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[14] Byunghyun Yoo, Sungwon Yi, Hyunwoo Kim, Younghwan Shin, Ran Han, Seungwoo Seo, Hwa Jeon Song, Euisok Chung, and Jeongmin Yang. 2024. MuDE: Multi-agent decomposed reward-based exploration. *Neural Networks* 179 (2024), 106565.

[15] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. 2022. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems* 35 (2022), 24611–24624.

[16] Won Joon Yun, Soohyun Park, Joongheon Kim, MyungJae Shin, Soyi Jung, David A Mohaisen, and Jae-Hyun Kim. 2022. Cooperative multiagent deep reinforcement learning for reliable surveillance via autonomous multi-UAV control. *IEEE Transactions on Industrial Informatics* 18, 10 (2022), 7086–7096.

[17] Yifan Zang, Jinmin He, Kai Li, Haobo Fu, Qiang Fu, Junliang Xing, and Jian Cheng. 2024. Automatic grouping for efficient cooperative multi-agent reinforcement learning. *Advances in Neural Information Processing Systems* 36 (2024).

[18] Xianghua Zeng, Hao Peng, and Angsheng Li. 2023. Effective and stable role-based multi-agent collaboration by structural information principles. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 11772–11780.

[19] Ainur Zhaikhan and Ali H Sayed. 2024. Graph Exploration for Effective Multiagent Q-Learning. *IEEE Transactions on Neural Networks and Learning Systems* (2024).

[20] Yifan Zhong, Jakub Grudzien Kuba, Xidong Feng, Siyi Hu, Jiaming Ji, and Yaodong Yang. 2024. Heterogeneous-agent reinforcement learning. *Journal of Machine Learning Research* 25 (2024), 1–67.

[21] Ziqing Zhou, Chun Ouyang, Linqiang Hu, Yi Xie, Yuning Chen, and Zhongxue Gan. 2024. A framework for dynamical distributed flocking control in dense environments. *Expert Systems with Applications* 241 (2024), 122694.

[22] Roy Zohar, Shie Mannor, and Guy Tennenholtz. 2022. Locality matters: A scalable value decomposition approach for cooperative multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 9278–9285.