Improving Cooperative Multi-Agent Exploration via Surprise Minimization and Social Influence Maximization

Extended Abstract

Mingyang Sun
Dalian University of Technology
Dalian, China
mysun@mail.dlut.edu.cn

Yaqing Hou
Dalian University of Technology
Dalian, China
houyq@dlut.edu.cn

Jie Kang
Dalian University of Technology
Dalian, China
kangj@mail.dlut.edu.cn

Haiyin Piao
Northwestern Polytechnical University
Xi'an, China
haiyinpiao@mail.nwpu.edu.cn

Yifeng Zeng
Northumbria University
Newcastle upon Tyne, United Kingdom
yifeng.zeng@northumbria.ac.uk

Hongwei Ge
Dalian University of Technology
Dalian, China
gehw@dlut.edu.cn

Qiang Zhang
Dalian University of Technology
Dalian, China
zhangq@dlut.edu.cn

ABSTRACT

In multi-agent reinforcement learning (MARL), the uncertainty of state change and the inconsistency between agents' local observation and global information are always the main obstacles of cooperative multi-agent exploration. To address these challenges, we propose a novel MARL exploration method by combining surprise minimization and social influence maximization. Considering state entropy as a measure of surprise, surprise minimization is achieved by rewarding the individual's intrinsic motivation (or rewards) for coping with more stable and familiar situations, hence promoting the policy learning. Furthermore, we introduce mutual information between agents' actions as a regularizer to maximize the social influence via optimizing a tractable variational estimation. In this way, the agents are guided to interact positively with one another by navigating between states that favor cooperation.

KEYWORDS

Multi-Agent Reinforcement Learning, Exploration, Cooperative Multi-Agent

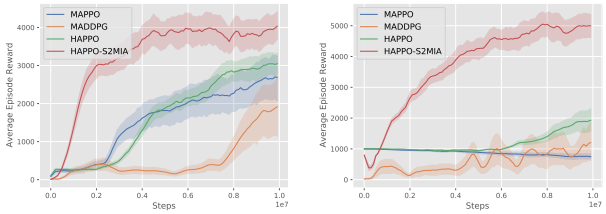
ACM Reference Format:
Mingyang Sun, Yaqing Hou, Jie Kang, Haiyin Piao, Yifeng Zeng, Hongwei Ge, and Qiang Zhang. 2023. Improving Cooperative Multi-Agent Exploration via Surprise Minimization and Social Influence Maximization: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

More recent multi-agent reinforcement learning (MARL) [2] studies have been proposed to solve the challenge of efficiently exploring unknown environments and gleaning informative experiences that could benefit the policy learning most towards optimal ones. However, while the existing exploration strategies for MARL obtain promising learning performance, they suffer from two common issues: (*a*) the partial observation and non-stationary problems [3] induce extra difficulty in the exploration measurement. (*b*) Even in coordinated exploration, the inconsistency between local and global information may exist.

This paper attempts to take a step towards solving the above issues. First, we observe that uncertainties in the environment can keep an agent in an unstable state of change, which is not conducive to exploration and learning. An example was illustrated in [1], where the environment around an agent is unstable due to weather changes, and if it builds a shelter and hides in it, it can reach a stable and predictable state in the long run. Therefore, we believe that explicitly preventing the agents from exploring states with a high degree of arbitrary uncertainty is an important prerequisite for improving the efficiency and robustness of exploration. Inspired by this, we introduce *surprise minimization* to cope with unpredictable state changes in multi-agent scenarios. However, if only surprise is minimized, it is easy for the agents to adopt negative or conservative policies, which is detrimental to their learning of cooperative behavior. Cooperative behaviors usually emerge because there is a lot of interaction between the cooperators [4]. To do this, we introduce mutual information between agents' actions as a regularizer to maximize the social influence via optimizing a tractable variational estimation. *Social influence maximization* of policies encourages agents to interact actively with each other rather than blindly performing unusual actions, thus enabling them to learn to cooperate effectively. In summary, this paper proposes a new multi-agent exploration approach by combining *surprise minimization* and *social influence maximization*, named S2MIA. A more stable state distribution reduces the range of states involved

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.



(a) Walker 3x2 (b) Ant 4x2

Figure 1: Mean performance for Multi-Agent MuJoCo tasks with 95% confidence interval shown in shaded areas.

in maximizing social influence. Conversely, social influential exploration encourages agents to navigate between the states that favor cooperation.

2 OUR METHODS

2.1 Individual Surprise Minimization

To incorporate individual surprise minimization into agents' learning, we design intrinsic rewards as an embodiment of the extra benefit an agent gets when it experiences more familiar states, based on the history of the states it experiences under the current policy. We assume an agent i learns a policy π^i_ϕ, parameterized by ϕ. The goal of surprise minimization is to minimize the entropy of its state marginal distribution under its current policy π^i_ϕ at each time step of the episode. This state entropy can be estimated by fitting an estimate of the visited state marginal $p^{\pi^i_\phi}(s^i_t)$ as each step t, given by $p_{\theta^i}(s^i_t)$, using the states during the entire episode. For a complete trajectory $\tau^i = \{s^i_1, \ldots, s^i_t \ldots, s^i_T\}$, We can get an upper bound as an estimate of the sum of the entropy of each state distribution over the whole episode:

$$\begin{aligned}\sum_{t=0}^{T} \mathcal{H}(s^i_t) &= -\sum_{t=0}^{T} \mathbb{E}_{s^i_t \sim p^{\pi^i_\phi}(s^i_t)} [\log p^{\pi^i_\phi}(s^i_t)] \\ &\leq -\sum_{t=0}^{T} \mathbb{E}_{s^i_t \sim p^{\pi^i_\phi}(s^i_t)} [\log p_{\theta^i}(s^i_t)],\end{aligned} \tag{1}$$

where the inequality becomes an equality when $p_{\theta^i}(s^i_t)$ accurately models $p^{\pi^i_\phi}(s^i_t)$. Minimizing the right-hand side of the inequality is equivalent to a new reinforcement learning maximization objective with an additional internal reward. This leads to a new reward function:

$$\tilde{r}(s^i_t) = r(s^i_t) + \alpha \log p_{\theta^i}(s^i_t), \tag{2}$$

where the coefficient α is used to control the proportion of intrinsic reward. The most basic principle is to put these two reward terms on a similar magnitude.

In principle, any appropriate model class can be selected according to the training environment to estimate the density $p_\theta(s^i)$. In our experiments, $p_\theta(s^i)$ is simply modeled as an independent Gaussian distribution for each dimension of the observation. In more complex environments, when the state of an agent or the dimensionality of observations (such as images) is large, it is advisable to employ a more sophisticated density estimator or utilize some dimensionality reduction and feature extraction techniques (such as Variational Auto-Encoders [5]).

2.2 Social Influential Exploration

In this section, we introduce social influence as a regularization term, which stimulates agents to maximize the mutual information (MI) between their actions. This regularization term $I(i, -i)$ can be added the objective function of policy optimization algorithm. By sampling sufficient joint actions, and averaging the resulting policy distribution of i in each case, we can obtain the marginal policy of i, $p(a^i_t|s^i_t) = \sum_{\boldsymbol{a}^{-i}_t} p(a^i_t|\boldsymbol{a}^{-i}_t, s^i_t) p(\boldsymbol{a}^{-i}_t|s^i_t)$, that is the decentralized policy of i without considering the actions of other agents. The discrepancy between the marginal policy of i and the conditional policy of i given $-i$'s action is a measure of the causal influence of $-i$ on i; it gives the degree to which i changes its planned action distribution because of actions of other agents. The influence regularization term to the mutual information between the actions of agents i and $-i$, which is given by

$$\begin{aligned}I(A^i; A^{-i}|s^i) &= \sum_{\mathbf{a}} p(\mathbf{a}|s^i) \log \frac{p(\mathbf{a}|s^i)}{p(a^i|s^i)p(\boldsymbol{a}^{-i}|s^i)} \\ &= \sum_{\boldsymbol{a}^{-i}} p(\boldsymbol{a}^{-i}|s^i) D_{KL}[p(a^i|\boldsymbol{a}^{-i}, s^i) \| p(a^i|s^i)].\end{aligned} \tag{3}$$

To maximize the social influence, estimating $p(a^i|\boldsymbol{a}^{-i}, s^i)$ is the main obstacle. By sampling N independent trajectories from the environment, we can perform a Monte-Carlo (MC) approximation of the MI. Unfortunately, in many multi-agent scenarios, where the problem space is often large, the amount of memory consumed by MC is often unrealistic for accurate estimation. As an alternative, for the mutual information objective, we introduce a variational posterior $q_\xi(a^i|\boldsymbol{a}^{-i}, s^i)$ via a neural network with parameters ξ to derive a tractable lower bound:

$$I(A^i; A^{-i}|s^i) \geqq \sum_{\boldsymbol{a}^{-i}} p(\boldsymbol{a}^{-i}|s^i) D_{KL}[q_\xi(a^i|\boldsymbol{a}^{-i}, s^i) \| p(a^i|s^i)]. \tag{4}$$

3 EXPERIMENTAL STUDY

To verify the adaptability of S2MIA, we apply it on Heterogeneous-Agent Proximal Policy Optimisation (HAPPO) [6]. We benchmark HAPPO-S2MIA against other existing state-of-the-art (SOTA) algorithms, which include MAPPO [9], MADDPG [7] and the original HAPPO. We consider a total of 9 tasks in 3 different scenarios, i.e., Walker, HalfCheetah and Ant, in Multi-Agent MuJoCo [8] to conduct our experiments. The partial results are shown in Fig. 1. Agents were evaluated for a total of 10 million environmental steps with the lines in the plot indicating average episode rewards and the shaded area as 95% confidence interval over 10 independent runs. The plots show that HAPPO-S2MIA performs substantially better than all rivals on every control task. A significant improvement in early learning efficiency was observed in all tasks. This suggests that under the combined effects of surprise minimization and social influence maximization, agents can discover policies that keep the task going more quickly. The final test result of HAPPO-S2MIA is better than HAPPO, which also shows that our method can explore better cooperation policies than simple entropy regularization.

REFERENCES

[1] Glen Berseth, Daniel Geng, Coline Devin, Dinesh Jayaraman, Chelsea Finn, and Sergey Levine. 2019. SMiRL: Surprise Minimizing RL in Entropic Environments. (2019).

[2] Sven Gronauer and Klaus Diepold. 2022. Multi-agent deep reinforcement learning: a survey. *Artificial Intelligence Review* (2022), 1–49.

[3] Jianye Hao, Tianpei Yang, Hongyao Tang, Chenjia Bai, Jinyi Liu, Zhaopeng Meng, Peng Liu, and Zhen Wang. 2023. Exploration in Deep Reinforcement Learning: From Single-Agent to Multiagent Domain. *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[4] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, DJ Strouse, Joel Z Leibo, and Nando De Freitas. 2019. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International Conference on Machine Learning*. PMLR, 3040–3049.

[5] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[6] Jakub Grudzien Kuba, Ruiqing Chen, Munning Wen, Ying Wen, Fanglei Sun, Jun Wang, and Yaodong Yang. 2021. Trust region policy optimisation in multi-agent reinforcement learning. *arXiv preprint arXiv:2109.11251* (2021).

[7] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. *arXiv preprint arXiv:1706.02275* (2017).

[8] Bei Peng, Tabish Rashid, Christian A Schroeder de Witt, Pierre-Alexandre Kamienny, Philip HS Torr, Wendelin Böhmer, and Shimon Whiteson. 2020. FACMAC: Factored Multi-Agent Centralised Policy Gradients. *arXiv preprint arXiv:2003.06709* (2020).

[9] C. Yu, A. Velu, E. Vinitsky, Y. Wang, and Y. Wu. 2021. The Surprising Effectiveness of MAPPO in Cooperative, Multi-Agent Games. (2021).