Joint Intrinsic Motivation for Coordinated Exploration in Multi-Agent Deep Reinforcement Learning

Extended Abstract

Maxime Toquebiau
ECE Paris & Sorbonne Université, CNRS, ISIR
F-75005 Paris, France
maxime.toquebiau@gmail.com

Nicolas Bredeche*
Sorbonne Université, CNRS, ISIR
F-75005 Paris, France
nicolas.bredeche@sorbonne-universite.fr

Faïz Benamar
Sorbonne Université, CNRS, ISIR
F-75005 Paris, France
faiz.ben_amar@sorbonne-universite.fr

Jae-Yun Jun*
ECE Paris
Paris, France
jaeyunjk@gmail.com

ABSTRACT

Multi-agent deep reinforcement learning (MADRL) often struggles to learn strongly coordinated tasks, as performance depends not only on one agent's behavior but rather on the joint behavior of multiple agents. In this context, a group of agents can benefit from actively exploring different joint strategies to determine the most efficient one. In this paper, we propose an approach for rewarding strategies where agents collectively exhibit novel behaviors. We present JIM (Joint Intrinsic Motivation), a multi-agent intrinsic motivation method that rewards joint trajectories based on a centralized measure of novelty. We show how JIM can be used to improve state-of-the-art MADRL methods in a highly coordinated task, demonstrating the crucial role of coordinated exploration.

KEYWORDS

Multi-agent Systems; Deep Reinforcement Learning; Intrinsic Motivation

ACM Reference Format:
Maxime Toquebiau, Nicolas Bredeche, Faïz Benamar, and Jae-Yun Jun. 2024. Joint Intrinsic Motivation for Coordinated Exploration in Multi-Agent Deep Reinforcement Learning: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

While MADRL methods are able to solve increasingly more complex tasks [7, 10, 14, 16], they still struggle in setups that require a high degree of coordination between agents. Strongly coordinated tasks can be seen as multi-agent exploration problems, where the objective is to find a specific coordinated strategy through very sparse positive reward signals. Such hard exploration problems have been a challenge for classical reinforcement learning algorithms that use random exploration strategies (e.g., ϵ-greedy) [2, 11]. Moreover, the multi-agent setting makes the exploration task even harder, as some local behavior may appear trivial and worthless from an agent's point of view, while it is actually valuable from the perspective of the whole multi-agent system (MAS).

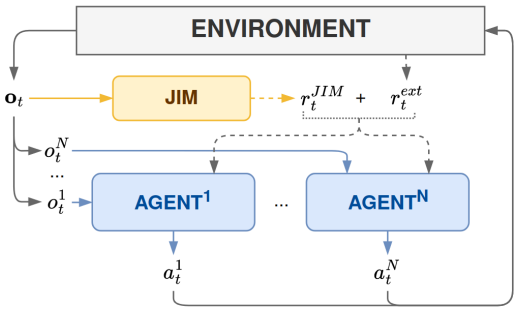

Figure 1: Architecture for Joint Intrinsic Motivation. A single module (JIM) generates the intrinsic reward for all agents, based on the joint observation.

We propose to solve this issue by inciting agents to explore new coordinated behaviors with a joint intrinsic motivation (JIM) mechanism (depicted in Figure 1). JIM takes inspiration from single-agent intrinsic motivation methods [1, 3, 4, 12, 13, 17] to define a metric for measuring novelty of joint observations. This metric is used as an intrinsic reward that motivates agents to coordinately explore their environment. Intrinsic motivation has already been used in multi-agent settings [5, 6, 8, 15], but previous methods always compute localized intrinsic rewards and rarely study exploration of the environment. With JIM, we propose the first approach that uses joint observations in an intrinsic motivation algorithm to incite coordinated exploration.

2 METHOD

Following previous works on intrinsically motivated exploration [1], we design a novelty metric that combines two exploration criteria working at different timescales. First, the *life-long exploration criterion* N_{LLEC} captures how novel is the current observation with respect to all observations since the beginning of training. This

*Authors contributed equally to the paper.

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

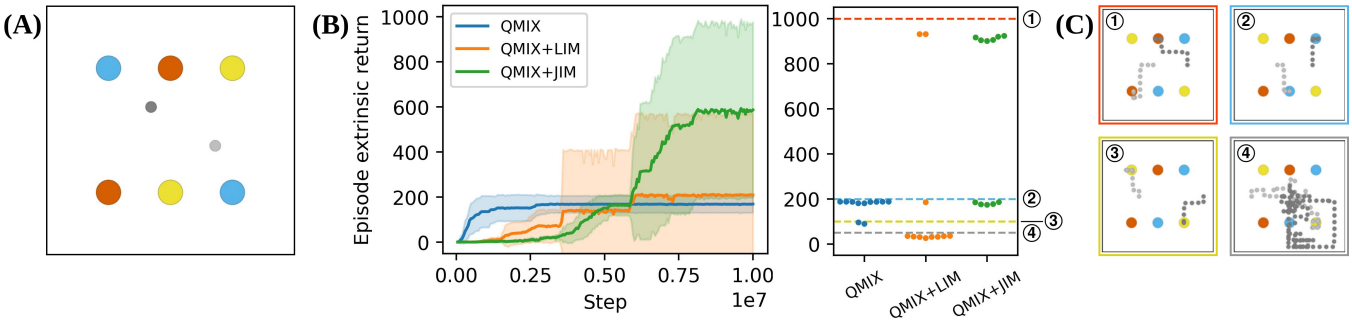

Figure 2: Experiments in the coordinated placement task. (A) Screenshot of the environment, showing the colored landmarks and the two agents. (B) Results of training the three variants of QMIX, with training curves (mean and standard deviation for 11 runs each) on the left and performance of all independent runs at the last iteration of training on the right. Dashed lines indicate different levels of strategy. Example trajectories for each of these strategies are shown in (C).

criterion, inspired by NovelD [17], motivates agents to search for never-experienced parts of the state space. Second, the *episodic exploration criterion* N_{EEC} captures the difference between the current observation and all previous observations in the current episode, following an elliptical bonus defined in [4]. This incites agents to have diverse trajectories. See the full version of this paper for more details on N_{LLEC} and N_{EEC}[1].

Our **Joint Intrinsic Motivation (JIM)** algorithm uses these two criteria to reward agents with the novelty of the consecutive joint observations $\mathbf{o}_t = \{o_t^i\}_{0 \leq i \leq N}$ and $\mathbf{o}_{t+1}$:

$$r_t^{JIM}(\mathbf{o}_t, \mathbf{o}_{t+1}) = N_{LLEC}(\mathbf{o}_t, \mathbf{o}_{t+1}) \times N_{EEC}(\mathbf{o}_{t+1}). \quad (1)$$

Therefore, at each time step, agents receive the augmented reward $r_t = r_t^{ext} + \beta r_t^{JIM}$, with the extrinsic reward from the environment r_t^{ext} and the hyper-parameter β controlling the weight of r_t^{JIM}. This intrinsic reward exploits centralized information during training to motivate coordinated exploration of the joint-observation space. Thus, it can be used to augment any MADRL algorithm that fits in the centralized training with decentralized execution (CTDE) paradigm.

3 EXPERIMENTS

To study how JIM helps multi-agent learning, we use the Multi-agent Particle Environment (MPE) [9] and design a coordination task. The coordinated placement scenario, shown in Figure 2A, has two sets of three colored landmarks. The reward given at each time step depends on the placement of the two agents on the landmarks: two agents capturing both red landmarks gives +10, both on blue gives +2, both on yellow gives +1, and only one agent on either blue or yellow gives +0.5. Blue and yellow landmarks act as deceptive positions requiring only one agent to generate a small reward. Careful exploration of the environment will allow agents to see that red landmarks are the optimal choice.

In this setup, we implement JIM with the state-of-the-art algorithm QMIX [14]. We compare this new algorithm, termed QMIX+JIM, with both the original QMIX with no intrinsic motivation and with a variant using local intrinsic motivation only (QMIX+LIM). LIM uses the same reward definition as JIM (see Eq.1), but it takes as input the local observations instead of the joint observations. Therefore, each agent has its own intrinsic reward related to its local observations.

Results are shown in Figure 2B and demonstrate the importance of coordinated exploration. QMIX alone always goes for either blue or yellow landmarks. This indicates that without actively exploring the environment, QMIX falls into the deceptive rewards trap and is unable to find the optimal strategy. QMIX+LIM seems slightly better than QMIX on average, but the individual run performance shows that LIM arguably performs worse. While two runs manage to find the optimal strategy, LIM often performs poorly with only one agent on a blue or yellow landmark. This demonstrates that exploring the space of local observations can be helpful, but can also mislead agents into focusing on local immediate rewards. JIM helps QMIX to find the optimal strategy more often, with more than half of the runs where the optimal strategy is learned. When agents do not find the optimal strategy, they stick with the best sub-optimal strategy with both agents on blue. This shows the benefits of actively looking for novel joint observations. As they contain all the information to understand the sparse reward signals, exploring the joint-observation space allows to better learn the task.

4 DISCUSSIONS

In this paper, we present an algorithm for joint intrinsic motivation (JIM) that can be used to enhance any MADRL algorithm in the CTDE paradigm. To the best of our knowledge, this is the first approach that rewards agents for exploring the joint observation space. We demonstrate that this is crucial to more reliably solve strongly coordinated tasks. In further experiments (see full paper[1]), we show that these results hold with more agents and present an ablation study that showcases the need for combining the two exploration criteria presented in Section 2. Overall, we think that these results should promote the value of using joint observation for computing intrinsic rewards in multi-agent setups.

ACKNOWLEDGMENTS

The authors appreciate ECE Paris for financing the Lambda Quad Max Deep Learning server, which was employed to obtain the results illustrated in the present work.

[1]https://arxiv.org/abs/2402.03972

REFERENCES

[1] Adrià Puigdomènech Badia, Pablo Sprechmann, Alex Vitvitskyi, Daniel Guo, Bilal Piot, Steven Kapturowski, Olivier Tieleman, Martin Arjovsky, Alexander Pritzel, Andrew Bolt, and Charles Blundell. 2020. Never Give Up: Learning Directed Exploration Strategies. In *8th International Conference on Learning Representations*.

[2] Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. 2016. Unifying Count-Based Exploration and Intrinsic Motivation. In *Advances in Neural Information Processing Systems*, Vol. 29. 1–9.

[3] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. 2019. Exploration by Random Network Distillation. In *7th International Conference on Learning Representations*.

[4] Mikael Henaff, Roberta Raileanu, Minqi Jiang, and Tim Rocktäschel. 2022. Exploration via Elliptical Episodic Bonuses. In *Advances in Neural Information Processing Systems*, Vol. 35. 37631–37646.

[5] Shariq Iqbal and Fei Sha. 2019. Coordinated Exploration via Intrinsic Rewards for Multi-Agent Reinforcement Learning. In *arXiv:1905.12127*.

[6] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro A. Ortega, DJ Strouse, Joel Z. Leibo, and Nando de Freitas. 2019. Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning. In *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97. 3040–3049.

[7] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. 2017. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. In *Advances in Neural Information Processing Systems*, Vol. 30. 1–12.

[8] Zixian Ma, Rose E Wang, Li Fei-Fei, Michael S. Bernstein, and Ranjay Krishna. 2022. ELIGN: Expectation Alignment as a Multi-Agent Intrinsic Reward. In *Advances in Neural Information Processing Systems*, Vol. 35. 8304–8317.

[9] Igor Mordatch and Pieter Abbeel. 2018. Emergence of Grounded Compositional Language in Multi-Agent Populations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32. 1495–1502.

[10] OpenAI, Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, Rafal Józefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov, Henrique P. d. O. Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski, and Susan Zhang. 2019. Dota 2 with Large Scale Deep Reinforcement Learning. In *arXiv:1912.06680*.

[11] Georg Ostrovski, Marc G. Bellemare, Aäron van den Oord, and Rémi Munos. 2017. Count-Based Exploration with Neural Density Models. In *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70. 2721–2730.

[12] Pierre-Yves Oudeyer and Frederic Kaplan. 2007. What is Intrinsic Motivation? A Typology of Computational Approaches. In *Frontiers in neurorobotics*, Vol. 1. 6.

[13] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. 2017. Curiosity-driven Exploration by Self-supervised Prediction. In *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70. 2778–2787.

[14] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. 2018. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. In *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80. 4295–4304.

[15] Tonghan Wang, Jianhao Wang, Yi Wu, and Chongjie Zhang. 2020. Influence-Based Multi-Agent Exploration. In *8th International Conference on Learning Representations*.

[16] Chao Yu, Akash Velu, Eugene Vinitsky, Yu Wang, Alexandre Bayen, and Yi Wu. 2021. The Surprising Effectiveness of PPO in Cooperative, Multi-Agent Games. In *arXiv:2103.01955*.

[17] Tianjun Zhang, Huazhe Xu, Xiaolong Wang, Yi Wu, Kurt Keutzer, Joseph E Gonzalez, and Yuandong Tian. 2021. NovelD: A Simple yet Effective Exploration Criterion. In *Advances in Neural Information Processing Systems*, Vol. 34. 25217–25230.