ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Prof. R. Wattenhofer

Provably Robust Cooperative Multi-Agent RL

Cooperative multi-agent reinforcement learning (c-MARL) with centralized training and decentralized execution (CTDE) has achieved remarkable success in complex coordination tasks ranging from autonomous vehicle fleets to robotic swarms and distributed control systems. State-of-the-art CTDE algorithms like QMIX and MAPPO enable multiple agents to learn coordinated policies in shared environments through value decomposition—factorizing a joint Q-function into individual agent contributions. However, c-MARL systems deployed in adversarial real-world settings face a critical vulnerability: malicious agents exhibiting arbitrary behaviors can catastrophically undermine team performance. Unlike traditional machine learning settings where adversarial inputs can be filtered or detected, compromised agents in cooperative MARL directly participate in the learning process and execution, making robustness a fundamental challenge rather than a peripheral concern.

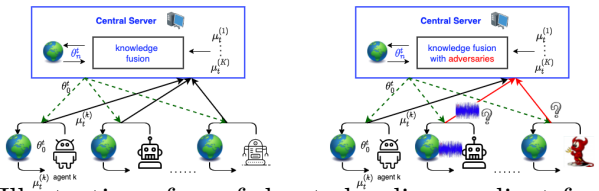

Figure 1: Illustration of our federated policy gradient framework [1]

Safety-critical applications such as autonomous driving, medical robotics, and industrial control require not only high performance but also reliability guarantees under adversarial conditions. While recent work has explored robustness in various distributed learning settings [1], cooperative MARL presents unique challenges due to the tight coupling between agents through shared environments and coordinated value functions. This project investigates fundamental questions of robustness in cooperative MARL: How can we ensure acceptable team performance when a fraction of agents are compromised? What theoretical guarantees can be provided, and how do they trade off with computational efficiency? Addressing these questions is essential for deploying multi-agent systems in real-world adversarial environments.

Interested? Please contact us for more details!

Contact

- Flint Xiaofeng Fan: xiafan@ethz.ch, ETZ G97

References

[1] Xiaofeng Fan, Yining Ma, Zhongxiang Dai, Wei Jing, Cheston Tan, and Bryan Kian Hsiang Low. Fault-tolerant federated reinforcement learning with theoretical guarantee. *Advances in neural information processing systems*, 34:1007–1021, 2021.