Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Prof. R. Wattenhofer

Trust and Resilience in Distributed Reinforcement Learning

Distributed reinforcement learning (DistRL) enables agents to learn collaboratively without a central coordinator, which is crucial for settings such as multi-robot teams, sensor networks, and federated control. However, current protocols often rely on *implicit trust*—uniform mixing of neighbors' updates—which makes them fragile against unreliable or malicious participants. In distributed optimization and federated learning, robustness has been studied through Byzantine-resilient aggregation. Yet, the explicit modeling of *trust and reputation* among agents remains underexplored in DistRL.

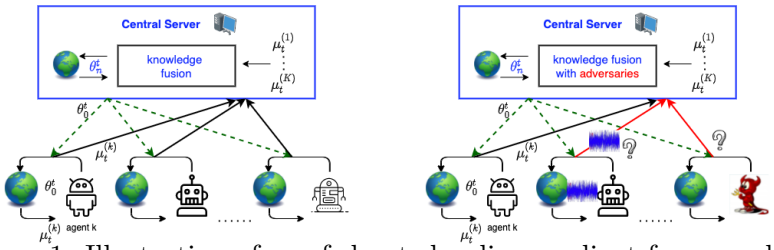

Figure 1: Illustration of our federated policy gradient framework [1]

However, in many applications, raw RL trajectories contain sensitive information (e.g., the medical records contain sensitive information about patients), and thus sharing them is prohibited. How can we collectively learn a better RL policy from distributed agents without sharing their trajectories? How do we ensure that the privacy of agents is protected? What happens if adversarial agents are present? In this project, we aim to leverage distributed computing to develop innovative techniques for distributed and decentralized reinforcement learning, and address many interesting open problems, such as privacy and robustness in RL.

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.