软件开发环境国家重点实验室
State Key Laboratory of Software Development Environment

北京航空航天大學 BEIHANG UNIVERSITY

北京大學 PEKING UNIVERSITY

北京通用人工智能研究院
Beijing Institute for General Artificial Intelligence

BYZANTINE ROBUST COOPERATIVE MULTI-AGENT REINFORCEMENT LEARNING AS A BAYESIAN GAME

Paper ID 871

Simin Li, Jun Guo, Jingqiao Xiu, Ruixiao Xu, Xin Yu,
Jiakai Wang, Aishan Liu, Yaodong Yang, Xianglong Liu

Introduction

Cooperative Multi-Agent Reinforcement Learning Algorithms

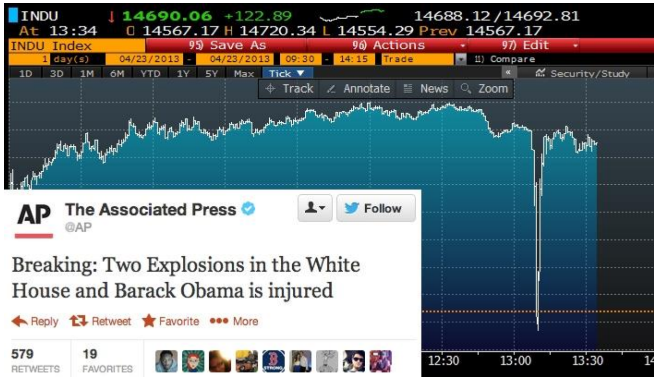
Hardware Error

Software Error

Unknown Error

Adversarial Attack

Real world agents can get out of control and not fully cooperative!



Financial loss in stock market

Life loss in autonomous driving

How to design robust c-MARL algorithms against these uncertainties?

Introduction

- How to design robust c-MARL algorithms?
  - Problem formulation
  - Optimal equilibrium concept
  - Practical algorithm

Contributions

- We theoretically formulate Byzantine adversaries in c-MARL as a BARDec-POMDP, and concurrently pursues robustness and cooperation by targeting an ex interim equilibrium.

- To achieve this equilibrium, we devise an actor-critic algorithm that ensures almost sure convergence under certain conditions.

- Our method exhibits greater resilience against a broad spectrum of adversaries on three c-MARL environments.

Methods

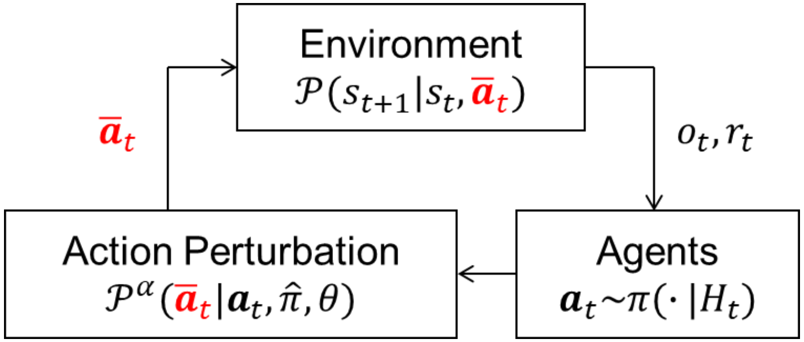
☐Problem Formulation

- How to model this problem?
- Model uncertainty of attacker as **type** in Bayesian game
- Formulation： Bayesian Adversarial Robust Dec-POMDP

$$\hat{\mathcal{G}} := \langle \mathcal{N}, \mathcal{S}, \Theta, \mathcal{O}, O, \mathcal{A}, \mathcal{P}^{\alpha}, \mathcal{P}, R, \gamma \rangle$$

- $\Theta \in \{0,1\}$： **type space**
  - $\theta^i = 0$: agent is cooperative
  - $\theta^i = 1$: agent is adversary
- $\mathcal{P}^{\alpha}$： characterize **attack process**

$$\mathcal{P}^{\alpha}(\overline{\mathbf{a}}_t|\mathbf{a}_t, \hat{\pi}, \theta) = \prod_{i \in \mathcal{N}} \hat{\pi}^i(\cdot|H_t^i, \theta) \cdot \theta^i + \delta(\overline{a}_t^i - a_t^i) \cdot (1 - \theta^i)$$



Framework for our Defense

Methods

Optional Solution: *ex interim* robust Markov perfect Bayesian equilibrium (RMPBE)

Previous Solution: *ex ante* RMPBE

***ex ante* RMPBE:** $$(\pi_*^{EA}(\cdot|H), \hat{\pi}_*^{EA}(\cdot|H,\theta)) \in \arg\max_{\pi(\cdot|H)} \mathbb{E}_{p(\theta)}\Big[\min_{\hat{\pi}(\cdot|H,\theta)} V_\theta(s)\Big],$$

with $V_\theta(s) = \sum_{\overline{\mathbf{a}}\in\mathcal{A}} \mathcal{P}^\alpha(\overline{\mathbf{a}}|\mathbf{a},\hat{\pi},\theta)\prod_{i\in\mathcal{N}}\pi^i(a^i|H^i)(R(s,\overline{\mathbf{a}}) + \gamma\sum_{s'\in\mathcal{S}}\mathcal{P}(s'|s,\overline{\mathbf{a}})V_\theta(s'))$.

- Cannot balance equilibrium between robustness and cooperation

Our Solution: *ex interim* RMPBE

- Consistency: at each timestep, each agent update belief using Bayes' rule
- Sequential Rationality: each policy maximizes value function under current belief

***ex interim* RMPBE:** $$(\pi_*^{EI}(\cdot|H,b), \hat{\pi}_*^{EI}(\cdot|H,\theta)) \in \arg\max_{\pi(\cdot|H,b)} \mathbb{E}_{p(\theta|H)}\Big[\min_{\hat{\pi}(\cdot|H,\theta)} V_\theta(s))\Big]$$

Proposition 2.2 (Existence of RMPBE). Assume a BARDec-POMDP of finite agents, finite set of state, observation and action space, agents use stationary policies, the type space Θ is a compact set, then *ex ante* and *ex interim* mixed strategy robust Markov perfect Bayesian equilibrium exists.

Proposition 2.3. Under Assumption 2.3, given finite type space and the prior of each type is not zero, as $t \to \infty$, $\pi_*^{EI}(\cdot|H_t, b_t)$ weakly dominates $\pi_*^{EA}(\cdot|H_t)$ under the worst-case adversary.

Methods

☐Algorithms

- How to achieve this ex interim RMPBE?

- **Value function update:** Robust Harsanyi-Bellman equation

$$Q_*^i(s, \overline{\mathbf{a}}, b^i) = \max_{\pi(\cdot|H,b)} \min_{\hat{\pi}(\cdot|H,\theta)} R(s, \overline{\mathbf{a}}) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s, \overline{\mathbf{a}}) \sum_{\theta \in \Theta} p(\theta|H'^i) \sum_{\overline{\mathbf{a}}' \in \mathcal{A}} \overline{\pi}(\overline{\mathbf{a}}'|H', b', \theta) Q_*^i(s', \overline{\mathbf{a}}', b'^i).$$

☐Q function updated by Robust Harsanyi-Bellman equation converge to optimal value

- **Policy Gradient Theorem:** Update robust agent and adversary

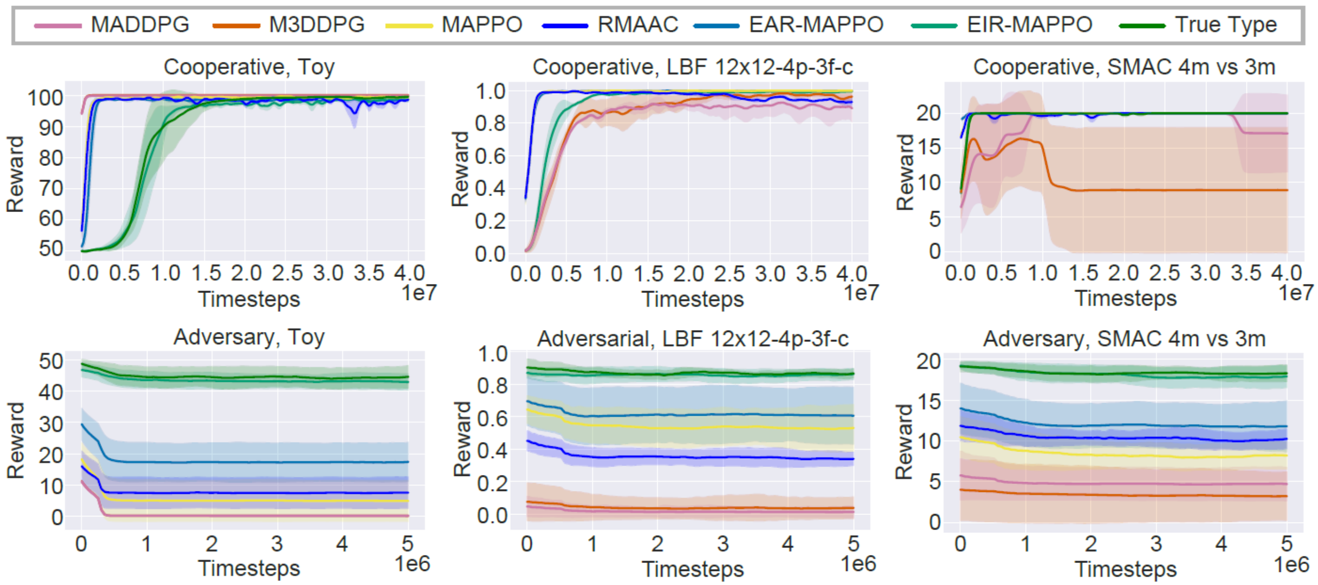
$$\nabla_{\phi^i} J^i(\phi^i) = \mathbb{E}_{s\sim\rho^{\overline{\pi}}(s), \overline{\mathbf{a}}\sim\overline{\pi}_{\phi,\hat{\phi}}(\overline{\mathbf{a}}|H,b,\theta)} \left[(1-\theta^i) \nabla \log \pi_{\phi^i}(a^i|H^i, b^i) Q^i(s, \overline{\mathbf{a}}, b^i) \right],$$

$$\nabla_{\hat{\phi}^i} J^i(\hat{\phi}^i) = \mathbb{E}_{s\sim\rho^{\overline{\pi}}(s), \overline{\mathbf{a}}\sim\overline{\pi}_{\phi,\hat{\phi}}(\overline{\mathbf{a}}|H,b,\theta)} \left[-\theta^i \nabla \log \hat{\pi}_{\hat{\phi}^i}(\hat{a}^i|H^i, \theta) Q^i(s, \overline{\mathbf{a}}, b^i) \right].$$

☐Provable convergence under two-timescale update

Experiments

■Is our EIR-MAPPO algorithm more robust?



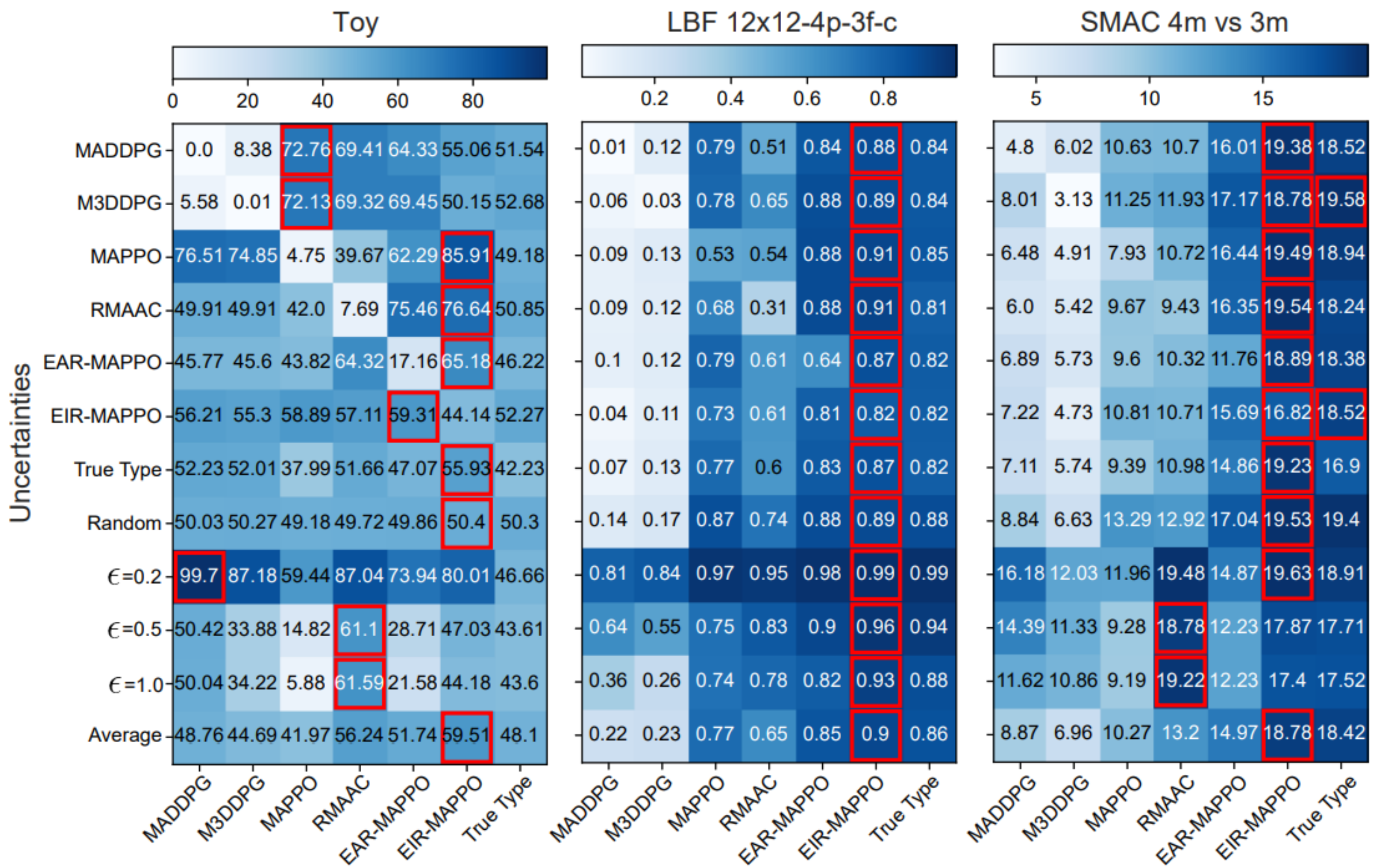
We achieve higher robustness under worst-case adversary, without harming cooperation

(a) MADDPG/M3DDPG (b) MAPPO/RMAAC (c) EAR-MAPPO (d) EIR-MAPPO

We learn intricate micromanagement skills under attack, including kiting and focus fire in SMAC

Experiments

■Is EIR-MAPPO robust under various attack scenarios?



EIR-MAPPO is robust against 11 unseen attacks, including non-oblivious adversaries, random allies, observation-based attacks, and transfer-based attacks.

Thanks For Your Interest!