Scalable Constrained Policy Optimization for Safe Multi-agent Reinforcement Learning

Lijun Zhang[1], Lin Li[1], Wei Wei[1]*, Huizhong Song[1], Yaodong Yang[2], Jiye Liang[1]

1. Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, School of Computer and Information Technology, Shanxi University, Taiyuan, Shanxi, China.

2. Institute for AI, Peking University, Beijing, China.

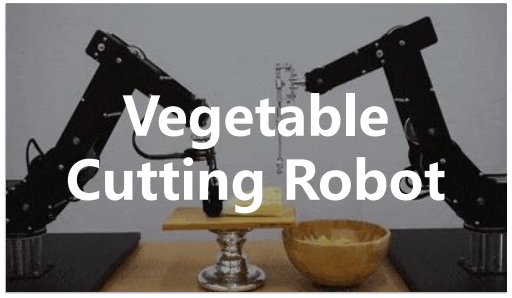
Overview

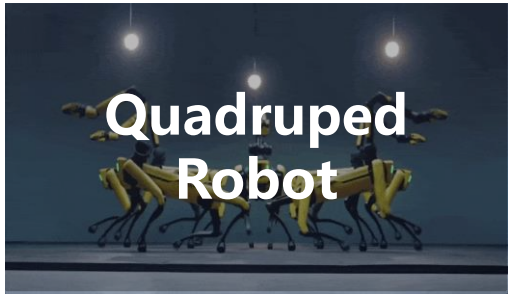
1. Challenge
2. Research Status and Innovative Method
3. Specific Implementation
4. Experiment Results

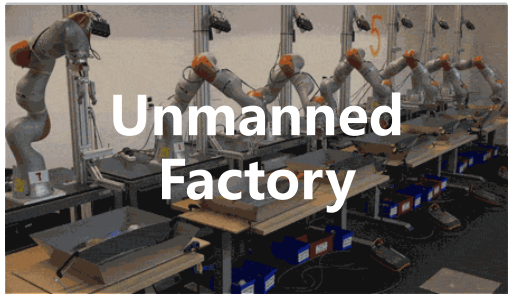
Challenge

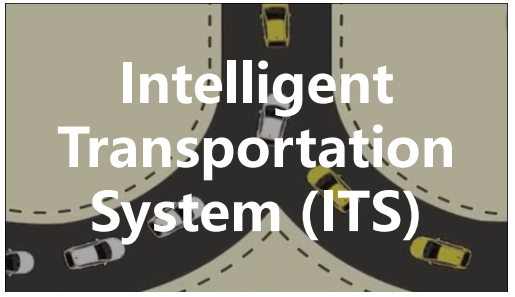
Safe Multi-agent Reinforcement Learning Problem

- The real world generally can be viewed as a multi-agent environment. A challenging problem in seeking to bring multi-agent reinforcement learning (MARL) techniques into real-world applications, such as autonomous driving and drone swarms, is how to control multiple agents safely and cooperatively to accomplish tasks.









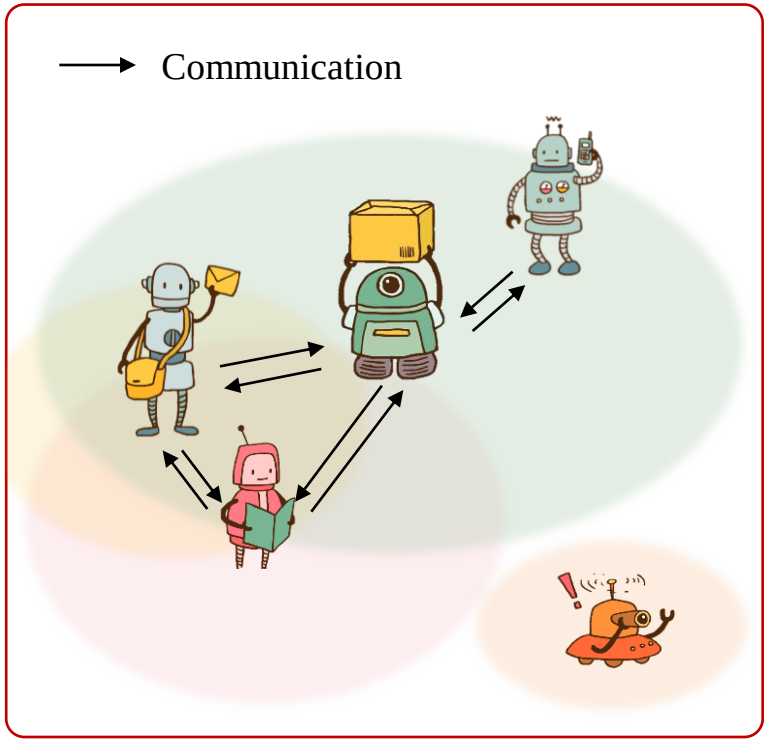
Research Status and Innovative Method

Research Status

- Most existing safe MARL methods **learn the centralized value function by introducing a global state** to guide safety cooperation. However, the global coupling arising from safety constraints and the exponential growth of the state-action space size limit their applicability in instant communication or computing resource-constrained systems and larger multi-agent systems.

Our method: Scal-MAPPO-L

- We develop **a novel scalable and theoretically-justified multi-agent constrained policy optimization method**.
- Specifically, our method has the following characteristics:
  - Policy is updated by following a sequential update scheme
  - Decentralized training with local interactions
  - The safety constraints and the joint policy improvement can be met



Specific Implementation

Problem

How to overcome the global coupling arising from safety constraints and the exponential growth of the state-action space size on the applicability of algorithms?

Solution

We introduce two assumptions about spatial correlation decay and updating the local policy based on the trust region method with a sequential update scheme.

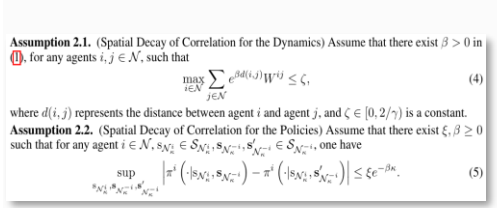
Implementation Idea

Two assumptions about spatial correlation decay are introduced.

Based on these assumptions, the maximum information loss regarding the advantage truncation is quantified.

The local policy optimization objective is provided by integrating the bounds of the trust region method and the bounds of the truncated advantage function.

A performance guarantee theorem is provided.



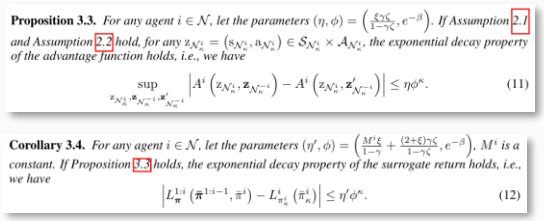
Assumption 2.1. (Spatial Decay of Correlation for the Dynamics) Assume that there exist $\beta > 0$ in (1), for any agents $i, j \in \mathcal{N}$, such that

$$\max_{i \in \mathcal{N}} \sum_{j \in \mathcal{N}} e^{\beta d(i,j)} W^{ij} \leq \zeta, \tag{4}$$

where $d(i,j)$ represents the distance between agent i and agent j, and $\zeta \in [0, 2/\gamma)$ is a constant.

Assumption 2.2. (Spatial Decay of Correlation for the Policies) Assume that there exist $\xi, \beta \geq 0$ such that for any agent $i \in \mathcal{N}$, $s_{\mathcal{N}_\kappa^i} \in \mathcal{S}_{\mathcal{N}_\kappa^i}, s_{\mathcal{N}_\kappa^{i-1}}, s'_{\mathcal{N}_\kappa^{i-1}} \in \mathcal{S}_{\mathcal{N}_\kappa^{i-1}}$, one have

$$\sup_{s_{\mathcal{N}_\kappa^i}, s_{\mathcal{N}_\kappa^{i-1}}, s'_{\mathcal{N}_\kappa^{i-1}}} \left| \pi^i\left(\cdot | s_{\mathcal{N}_\kappa^i}, s_{\mathcal{N}_\kappa^{i-1}}\right) - \pi^i\left(\cdot | s_{\mathcal{N}_\kappa^i}, s'_{\mathcal{N}_\kappa^{i-1}}\right) \right| \leq \xi e^{-\beta\kappa}. \tag{5}$$

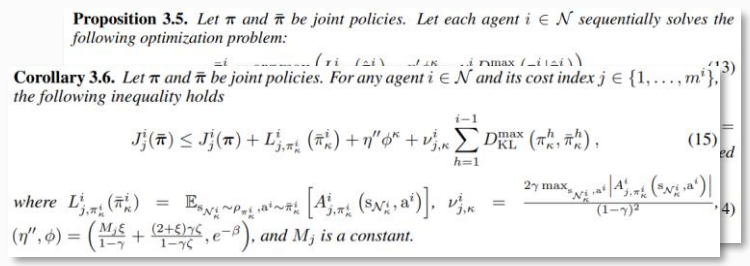


Proposition 3.3. *For any agent $i \in \mathcal{N}$, let the parameters $(\eta, \phi) = \left(\frac{\xi\gamma\zeta}{1-\gamma\zeta}, e^{-\beta}\right)$. If Assumption 2.1 and Assumption 2.2 hold, for any $z_{\mathcal{N}_\kappa^i} = \left(s_{\mathcal{N}_\kappa^i}, a_{\mathcal{N}_\kappa^i}\right) \in \mathcal{S}_{\mathcal{N}_\kappa^i} \times \mathcal{A}_{\mathcal{N}_\kappa^i}$, the exponential decay property of the advantage function holds, i.e., we have*

$$\sup_{z_{\mathcal{N}_\kappa^i}, z'_{\mathcal{N}_\kappa^{i-1}}} \left| A^i\left(z_{\mathcal{N}_\kappa^i}, z_{\mathcal{N}_\kappa^{i-1}}\right) - A^i\left(z_{\mathcal{N}_\kappa^i}, z'_{\mathcal{N}_\kappa^{i-1}}\right) \right| \leq \eta\phi^\kappa. \tag{11}$$

Corollary 3.4. *For any agent $i \in \mathcal{N}$, let the parameters $(\eta', \phi) = \left(\frac{M^i\xi}{1-\gamma} + \frac{(2+\xi)\gamma\zeta}{1-\gamma\zeta}, e^{-\beta}\right)$, M^i is a constant. If Proposition 3.3 holds, the exponential decay property of the surrogate return holds, i.e., we have*

$$\left| L^{1:i}_{\bar{\boldsymbol{\pi}}}\left(\bar{\boldsymbol{\pi}}^{1:i-1}, \bar{\pi}^i\right) - L^i_{\pi^i_\kappa}\left(\bar{\pi}^i_\kappa\right) \right| \leq \eta'\phi^\kappa. \tag{12}$$

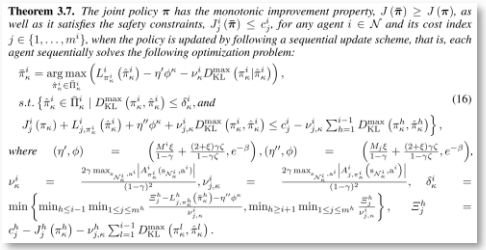


Proposition 3.5. *Let $\boldsymbol{\pi}$ and $\bar{\boldsymbol{\pi}}$ be joint policies. Let each agent $i \in \mathcal{N}$ sequentially solves the following optimization problem:*

Corollary 3.6. *Let $\boldsymbol{\pi}$ and $\bar{\boldsymbol{\pi}}$ be joint policies. For any agent $i \in \mathcal{N}$ and its cost index $j \in \{1, \ldots, m^i\}$, the following inequality holds*

$$J^i_j(\bar{\boldsymbol{\pi}}) \leq J^i_j(\boldsymbol{\pi}) + L^i_{j,\pi^i_\kappa}\left(\bar{\pi}^i_\kappa\right) + \eta''\phi^\kappa + \nu^i_{j,\kappa} \sum_{h=1}^{i-1} D^{\max}_{\mathrm{KL}}\left(\pi^h_\kappa, \bar{\pi}^h_\kappa\right), \tag{15}$$

where $L^i_{j,\pi^i_\kappa}\left(\bar{\pi}^i_\kappa\right) = \mathbb{E}_{s_{\mathcal{N}^i_\kappa} \sim \rho_{\pi^i_\kappa}, a^i \sim \bar{\pi}^i_\kappa}\left[A^i_{j,\pi^i_\kappa}\left(s_{\mathcal{N}^i_\kappa}, a^i\right)\right]$, $\nu^i_{j,\kappa} = \frac{2\gamma \max_{s_{\mathcal{N}^i_\kappa}, a^i}\left|A^i_{j,\pi^i_\kappa}\left(s_{\mathcal{N}^i_\kappa}, a^i\right)\right|}{(1-\gamma)^2}$, $(\eta'', \phi) = \left(\frac{M_j\xi}{1-\gamma} + \frac{(2+\xi)\gamma\zeta}{1-\gamma\zeta}, e^{-\beta}\right)$, *and M_j is a constant.*



Theorem 3.7. *The joint policy $\boldsymbol{\pi}$ has the monotonic improvement property, $J(\bar{\boldsymbol{\pi}}) \geq J(\boldsymbol{\pi})$, as well as it satisfies the safety constraints, $J^j_j(\bar{\boldsymbol{\pi}}) \leq c^j_j$, for any agent $i \in \mathcal{N}$ and its cost index $j \in \{1, \ldots, m^i\}$, when the policy is updated by following a sequential update scheme, that is, each agent sequentially solves the following optimization problem:*

$$\begin{aligned} &\bar{\pi}^i_\kappa = \arg\max_{\bar{\pi}^i_\kappa \in \Pi^i_\kappa} \left(L_{\pi^i_\kappa}\left(\bar{\pi}^i_\kappa\right) - \eta'\phi^\kappa - \nu^i_\kappa D^{\max}_{\mathrm{KL}}\left(\pi^i_\kappa | \bar{\pi}^i_\kappa\right)\right), \\ &s.t.\ \left\{\bar{\pi}^i_\kappa \in \Pi^i_\kappa \mid D^{\max}_{\mathrm{KL}}\left(\pi^i_\kappa, \bar{\pi}^i_\kappa\right) \leq \delta^i_\kappa, \text{and} \right. \\ &\left. J^j_j\left(\pi_\kappa\right) + L^j_{j,\pi^i_\kappa}\left(\bar{\pi}^i_\kappa\right) + \eta''\phi^\kappa + \nu^j_{j,\kappa} D^{\max}_{\mathrm{KL}}\left(\pi^i_\kappa, \bar{\pi}^i_\kappa\right) \leq c^j_j - \nu^j_{j,\kappa} \textstyle\sum_{h=1}^{i-1} D^{\max}_{\mathrm{KL}}\left(\pi^h_\kappa, \bar{\pi}^h_\kappa\right)\right\}, \end{aligned} \tag{16}$$

where $(\eta', \phi) = \left(\frac{M^i\xi}{1-\gamma} + \frac{(2+\xi)\gamma\zeta}{1-\gamma\zeta}, e^{-\beta}\right), (\eta'', \phi) = \left(\frac{M_j\xi}{1-\gamma} + \frac{(2+\xi)\gamma\zeta}{1-\gamma\zeta}, e^{-\beta}\right)$, $\nu^i_\kappa = \frac{2\gamma \max_{s_{\mathcal{N}^i_\kappa}, a^i}\left|A^i_{\pi^i_\kappa}\left(s_{\mathcal{N}^i_\kappa}, a^i\right)\right|}{(1-\gamma)^2}$, $\nu^j_{j,\kappa} = \frac{2\gamma \max_{s_{\mathcal{N}^i_\kappa}, a^i}\left|A^j_{j,\pi^i_\kappa}\left(s_{\mathcal{N}^i_\kappa}, a^i\right)\right|}{(1-\gamma)^2}$, $\delta^i_\kappa = \min\left\{\min_{h \leq i-1} \min_{1 \leq j \leq m^h} \frac{\Xi^h_j - L^h_{j,\pi^h_\kappa}\left(\bar{\pi}^h_\kappa\right) - \eta''\phi^\kappa}{\nu^h_{j,\kappa}}, \min_{h \geq i+1} \min_{1 \leq j \leq m^h} \frac{\Xi^h_j}{\nu^h_{j,\kappa}}\right\}$, $\Xi^h_j = c^h_j - J^h_j\left(\pi^h_\kappa\right) - \nu^h_{j,\kappa} \sum_{h=1}^{i-1} D^{\max}_{\mathrm{KL}}\left(\pi^h_\kappa, \bar{\pi}^h_\kappa\right)$.

Challenge 1: How to quantify the information loss regarding the advantage truncation?

Challenge 2: How to ensure the local policy updates are not overly conservative?

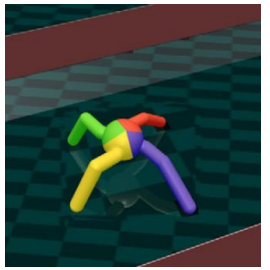
Challenge 3: How to prove that the method can improve reward performance and satisfy safe constraints?

Experiment Results

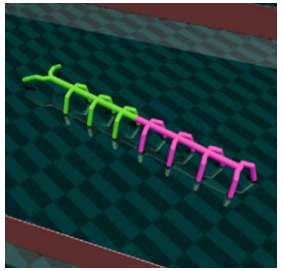
Main results (under Safe MAMuJoCo environment)

With the truncation parameter $\kappa \geq 3$, we can observe that the performance of Scal-MAPPO-L improves considerably and can approach or even outperform MAPPO-L in some environments.

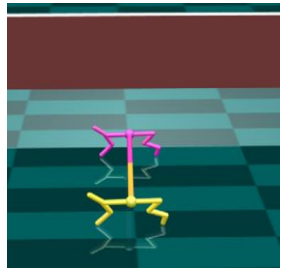
Safe 4 × 2 Ant task



Safe 2 × 4 Manyagent



Safe p1p-couple HalfCh



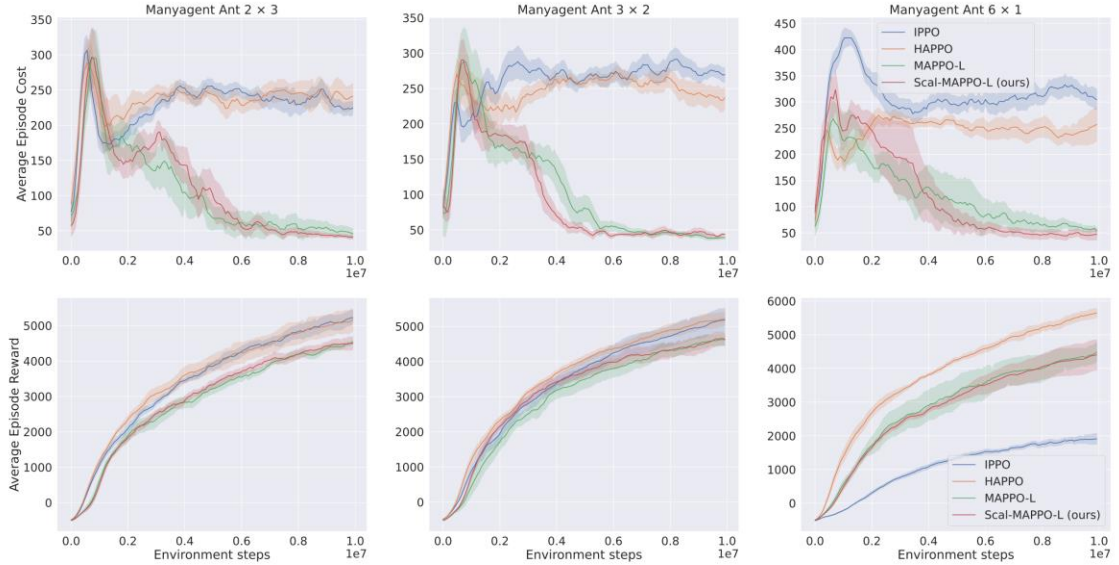


Figure 1: Performance comparisons in terms of cost and reward on three Safe ManyAgent Ant tasks. Each column subfigure represents a different task, and we plot the cost curves (the lower the better) in the upper row and the reward curves (the higher the better) in the bottom row for each task.

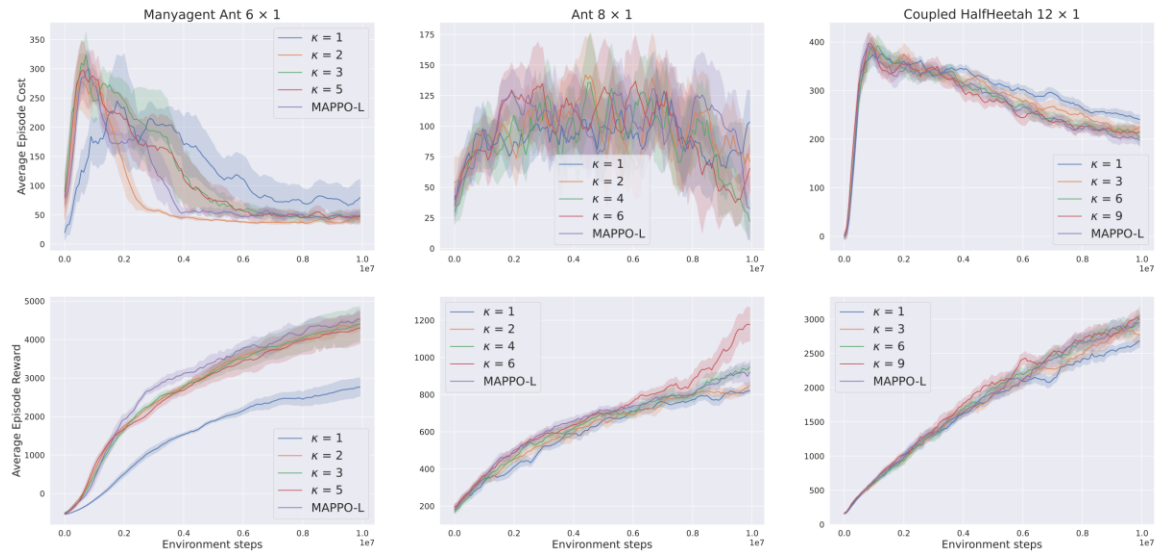


Figure 2: Performance comparisons in terms of cost and reward on Safe ManyAgent Ant task, Safe Ant task, and Safe Coupled HalfCheetah task. In each task, the performance of Scal-MAPPO-L with different κ and MAPPO-L are demonstrated.

Lijun Zhang, Lin Li, Wei Wei*, Huizhong Song, Yaodong Yang, Jiye Liang. Scalable Constrained Policy Optimization for Safe Multi-agent Reinforcement Learning. ***NeurIPS 2024*.**

Thanks for Listening

DM&KD Lab

—数据挖掘与知识发现实验室—