Coordinate Agents via Policy Optimization

Xihuai Wang (王锡淮)
上海交通大学
https://xihuai18.github.io

Aug. 8 2023 at RLChina

A2PO. ICLR 2023. https://arxiv.org/abs/2302.06205

Content

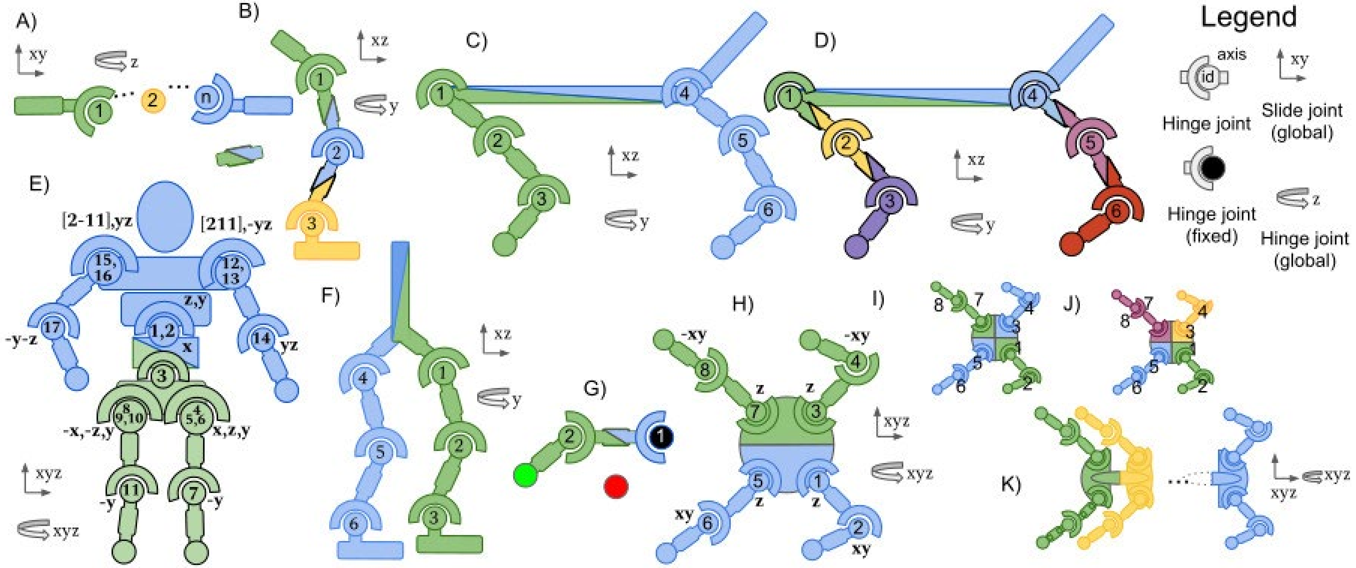
1. Brief Introduction about Cooperative MARL and Trust-region Methods Recap
2. Extensions of Trust-region Methods in Cooperative MARL
   - MAPPO - Reconstructing State Representation
   - CoPPO - Coordinating the Joint Policy
3. Sequential Policy Optimization – From Non-stationarity to Monotonic Improvement
4. Agent-by-agent Policy Optimization

Cooperative Multi-agent Scenarios

- SMAC (StarCraft Multi-Agent Challenge)

(a) 2c_vs_64zg (b) 3s_vs_5z (c) corridor

- MAMuJoCo (Multi-agent MuJoCo)



Goal: Learn **a policy for each agent** that all agents **together achieve the goal of the system**.

Decentralized Execution **Shared Reward Function**

Modelled by a Dec-MDP $(\mathcal{S}, \{\mathcal{A}^i\}_{i\in\mathcal{N}}, r, \mathcal{T}, \gamma)$

- $\mathcal{N} = \{1, \ldots, n\}$ is the set of agents;
- $\mathcal{S}$ is the state space;
- $\mathcal{A} = \mathcal{A}^1 \times \cdots \times \mathcal{A}^n$ is the joint action space, where $\mathcal{A}^i$ is the action space of agent i ;
- $r : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ is the reward function;
- $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto [0, 1]$ is the dynamics function;
- $\gamma \in [0, 1)$ is the reward discount factor.

Goal: $\max_{\boldsymbol{\pi}} \mathbb{E}_{\tau\sim(\mathcal{T},\boldsymbol{\pi})}[\sum_{t=0}^{\infty} \gamma^t r(s_t, \boldsymbol{a_t})]$

$$\boldsymbol{\pi}(\cdot|s_t) = \pi^1(\cdot|s_t) \times \ldots \times \pi^n(\cdot|s_t),\ \tau = \{(s_0, \boldsymbol{a}_0), (s_1, \boldsymbol{a}_1), \ldots\}$$

Trust-region Methods Recap

(Performance Difference Lemma) For any two policies $\pi, \bar{\pi}$, we have

$$\begin{aligned}
\mathcal{J}(\bar{\pi}) - \mathcal{J}(\pi) &= \mathcal{J}(\bar{\pi}) - \mathbb{E}_{s_0 \sim \mu}\left[V^{\pi}(s_0)\right] \\
&= \mathcal{J}(\bar{\pi}) - \mathbb{E}_{\tau \sim (\mu, \bar{\pi})}\left[V^{\pi}(s_0)\right] \\
&= \mathcal{J}(\bar{\pi}) - \mathbb{E}_{\tau \sim (\mu, \bar{\pi})}\left[\sum_{t=0}^{\infty} \gamma^t V^{\pi}(s_t) - \sum_{t=1}^{\infty} \gamma^t V^{\pi}(s_t)\right] \\
&= \mathcal{J}(\bar{\pi}) - \mathbb{E}_{\tau \sim (\mu, \bar{\pi})}\left[\sum_{t=0}^{\infty} \gamma^t V^{\pi}(s_t) - \sum_{t=0}^{\infty} \gamma^{t+1} V^{\pi}(s_{t+1})\right] \\
&= \mathbb{E}_{\tau \sim (\mu, \bar{\pi})}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right] - \mathbb{E}_{\tau \sim (\mu, \bar{\pi})}\left[\sum_{t=0}^{\infty} \gamma^t V^{\pi}(s_t) - \sum_{t=0}^{\infty} \gamma^{t+1} V^{\pi}(s_{t+1})\right] \\
&= \mathbb{E}_{\tau \sim (\mu, \bar{\pi})}\left[\gamma^t \left(r(s_t, a_t) + \gamma V^{\pi}(s_{t+1}) - V^{\pi}(s_t)\right)\right] \\
&= \mathbb{E}_{\tau \sim (\mu, \bar{\pi})}\left[\gamma^t A^{\pi}(s_t, a_t)\right] \\
&= \mathbb{E}_{(s,a) \sim (d^{\bar{\pi}}, \bar{\pi})}\left[A^{\pi}(s_t, a_t)\right]
\end{aligned}$$

The normalized state distribution $d^{\pi_\theta}(s) = \frac{1}{1-\gamma} \sum_{t=0}^{\infty} \gamma^t P^{\pi_\theta}(s_t = s)$

- Performance Difference Lemma indicates that the return of a new policy (target policy) can be represented by the old policy, with the access to the new policy's occupancy measure (impractical) and the new policy itself (practical).

$$\pi : \text{Old Policy} \quad \bar{\pi} : \text{New Policy}$$

- To approximate the new policy's occupancy measure, we need π and $\bar{\pi}$ to be similar, e.g., small $D_{TV}^{max}(\pi \| \bar{\pi}) = \max_s D_{TV}(\pi(\cdot|s) \| \bar{\pi}(\cdot|s)))$

$$\mathbb{E}_{(s,a)\sim(d^{\bar{\pi}},\bar{\pi})}\left[A^{\pi}(s_t, a_t)\right]$$

$$\mathbb{E}_{(s,a)\sim(d^{\pi},\bar{\pi})}\left[A^{\pi}(s_t, a_t)\right] = \mathbb{E}_{(s,a)\sim(d^{\pi},\pi)}\left[\frac{\bar{\pi}(a_t|s_t)}{\pi(a_t|s_t)} A^{\pi}(s_t, a_t)\right]$$

- Surrogate Objective $\mathcal{L}_{\pi}(\bar{\pi}) = \mathcal{J}(\pi) + \mathbb{E}_{(s,a)\sim(d^{\pi},\bar{\pi})}\left[A^{\pi}(s_t, a_t)\right]$

- Why is PPO/TRPO effective?

(Monotonic Improvement Bound) Given $\alpha = D_{TV}^{max}(\pi \| \bar{\pi})$, $\epsilon = \max_{s,a} |A^{\pi}(s,a)|$, and $\mathcal{L}_{\pi}(\bar{\pi}) = \mathcal{J}(\pi) + \mathbb{E}_{(s,a)\sim(d^{\pi},\bar{\pi})}\left[A^{\pi}(s_t, a_t)\right]$, we have:

$$\mathcal{J}(\bar{\pi}) \geq \mathcal{J}(\pi) - \frac{4\epsilon}{1-\gamma}\alpha$$

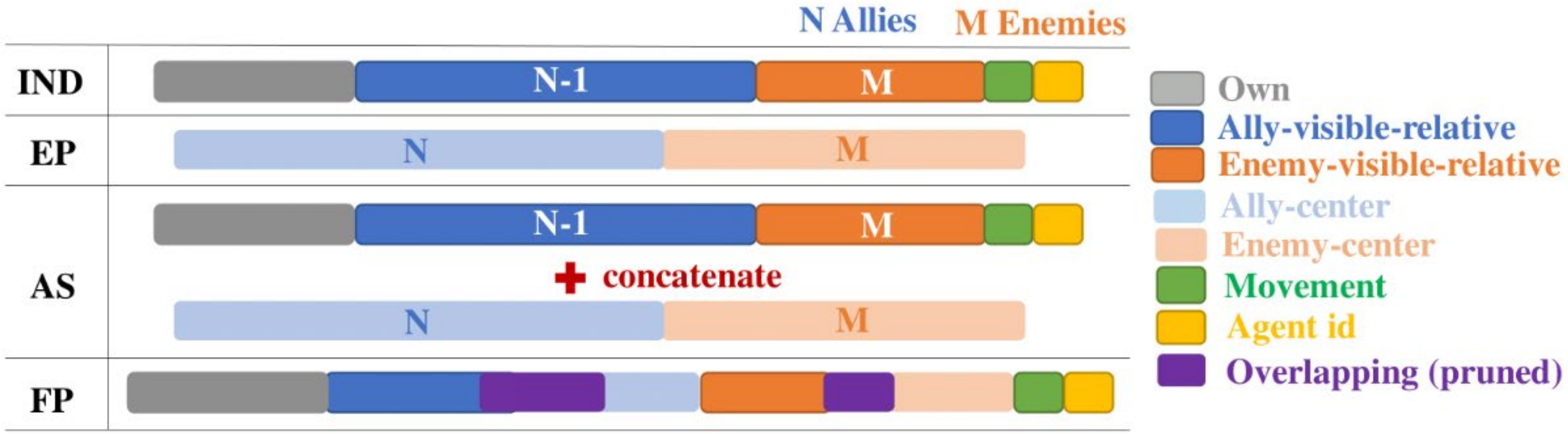
- The performance of the target policy can be monotonic improved by maximizing the righthand side, which is feasible.
- Maximization Objective of PPO

$$\mathbb{E}_{(s,a)\sim(d^{\pi},\pi)}\left[\min\left(\frac{\bar{\pi}}{\pi}A^{\pi}(s,a), \mathrm{clip}\left(\frac{\bar{\pi}}{\pi}, 1 \pm \epsilon\right)A^{\pi}(s,a)\right)\right]$$

A variant of Theorem 1 in *Schulman, John, et al. "Trust region policy optimization." International conference on machine learning. PMLR, 2015.*

Trust-region Methods in Cooperative MARL

- Multi-agent PPO (MAPPO)
  - State Construction



  - Implementation Tricks
  - Surrogate Objective

$$\mathcal{J}(\pi) + \frac{1}{n}\frac{1}{1-\gamma}\sum_{i=1}^{n}\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi}},\boldsymbol{\pi})}\left[\frac{\bar{\pi}^i}{\pi^i}A^{\boldsymbol{\pi}}(s,\boldsymbol{a})\right]$$

Yu, Chao, et al. "The surprising effectiveness of ppo in cooperative multi-agent games." *Advances in Neural Information Processing Systems* 35 (2022): 24611-24624.

- Coordinated PPO (CoPPO)
  - Surrogate Objective of MAPPO

$$\mathcal{J}(\pi) + \frac{1}{n}\frac{1}{1-\gamma}\sum_{i=1}^{n}\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi}},\boldsymbol{\pi})}\left[\frac{\bar{\pi}^i}{\pi^i}A^{\boldsymbol{\pi}}(s,\boldsymbol{a})\right]$$

  - Local constraints on individual policies $\overset{?}{\to}$ A Controllable constraint on the joint action

Corollary 2 *For all s,* $D_{TV}(\boldsymbol{\pi}(\cdot|s)\|\bar{\boldsymbol{\pi}}(\cdot|s)) \leq \sum_{i=1}^{n} D_{TV}(\pi^i(\cdot|s)\|\bar{\pi}^i(\cdot|s))$.

- Directly restrict the joint policy difference

(Multi-agent Performance Difference Lemma) Given any joint policies $\boldsymbol{\pi}$ *and* $\bar{\boldsymbol{\pi}}$*, the difference between the performance of the two joint policies can be expressed as :*

$$\mathcal{J}(\bar{\boldsymbol{\pi}}) - \mathcal{J}(\boldsymbol{\pi}) = \frac{1}{1-\gamma}\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\bar{\boldsymbol{\pi}}},\bar{\boldsymbol{\pi}})}\left[A^{\boldsymbol{\pi}}(s,\boldsymbol{a})\right]$$

Wu, Zifan, et al. "Coordinated proximal policy optimization." *Advances in Neural Information Processing Systems* 34 (2021): 26437-26448.

- Approximating $d^{\bar{\boldsymbol{\pi}}}$ by $d^{\boldsymbol{\pi}}$, similarly as in PPO, the surrogate objective of CoPPO:

$$\mathcal{J}(\boldsymbol{\pi}) + \frac{1}{1-\gamma}\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi}},\boldsymbol{\pi})}\left[\frac{\bar{\boldsymbol{\pi}}}{\boldsymbol{\pi}}A^{\boldsymbol{\pi}}(s,\boldsymbol{a})\right]$$

- Monotonic improvement of the joint policy

$$\left|\mathcal{J}(\bar{\boldsymbol{\pi}}) - \mathcal{J}(\boldsymbol{\pi}) - \frac{1}{1-\gamma}\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi}},\bar{\boldsymbol{\pi}})}[A^{\boldsymbol{\pi}}]\right|$$
$$\leq 4\epsilon\sum_{i=1}^{n}\alpha^i\left(\frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\sum_{j=1}^{n}\alpha^j)}\right) \quad < \quad 4\epsilon\sum_{i=1}^{n}\frac{\alpha^i}{1-\gamma}$$

Monotonic improvement bound of MAPPO[1]

- Clip the joint action, where the outer clip limits the influence of other agents and reduce the variance

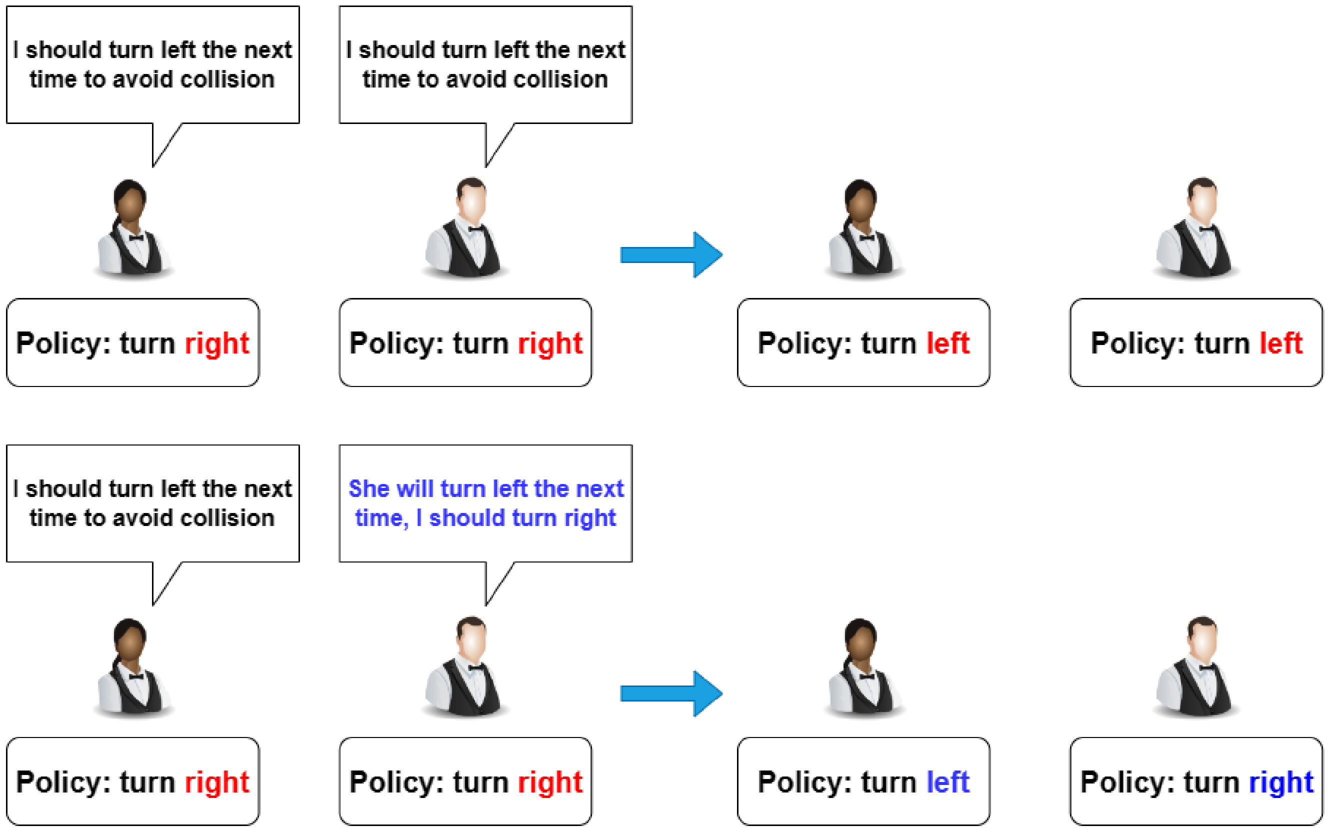
$$\max_{\pi^i}\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi}},\boldsymbol{\pi})}\left[\min\left(l(s,\boldsymbol{a})A^{\boldsymbol{\pi}}, \mathrm{clip}\left(l(s,\boldsymbol{a}), 1\pm\epsilon^{\text{inner}}\right)A^{\boldsymbol{\pi}}\right)\right]$$

$$l(s,\boldsymbol{a}) = \frac{\bar{\pi}^i(a^i|s)}{\pi^i(a^i|s)}\,\mathrm{clip}(\prod_{j\in -i}\frac{\bar{\pi}^j(a^j|s)}{\pi^j(a^j|s)}, 1\pm\epsilon^{\text{outer}})$$

1. Proof without considering the parameter sharing technique.

Sequential Policy Optimization – From Non-stationarity

- MAPPO and CoPPO update the agents simultaneously, that is, all agents perform policy improvement at the same time and **cannot observe the change of other agents.**
- The simultaneous update scheme brings about the non-stationarity problem, i.e., the environment dynamic changes from one agent's perspective as **other agents also change their policies.**



- Sequential Update scheme: Agents sequentially perform policy update in a given order, the incoming agents **are allowed to perceive changes made by preceding agents.**
- Alleviate the problems brought by simultaneous update scheme.

- We formulate the update process in sequential policy update scheme as (assume agents are updated in the order $1, 2, \ldots, n$):

$$\boldsymbol{\pi} = \hat{\boldsymbol{\pi}}^0 \xrightarrow[\text{Update } \pi^1]{\max_{\pi^1} \mathcal{L}_{\boldsymbol{\pi}}\left(\hat{\boldsymbol{\pi}}^1\right)} \hat{\boldsymbol{\pi}}^1 \rightarrow \cdots \rightarrow \hat{\boldsymbol{\pi}}^{n-1} \xrightarrow[\text{Update } \pi^n]{\max_{\pi^n} \mathcal{L}_{\hat{\boldsymbol{\pi}}^{n-1}}(\hat{\boldsymbol{\pi}}^n)} \hat{\boldsymbol{\pi}}^n = \bar{\boldsymbol{\pi}}$$

where $\hat{\boldsymbol{\pi}}^i = \bar{\pi}^1 \times \ldots \times \bar{\pi}^i \times \pi^{i+1} \times \ldots \times \pi^n$ is the joint policy while updating agent i, $\mathcal{L}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i)$ is the surrogate objective of agent i, and we denote the preceding agents of agent i as a set e^i.

- More on non-stationarity: an analysis on the state transition shift

(Non-stationarity Decomposition)
Given the state transition shift $\Delta^{\bar{\pi}^1,\ldots,\bar{\pi}^{i-1},\pi^i,\ldots,\pi^n}_{\pi^1,\ldots,\pi^n}(s'|s) = \sum_{\boldsymbol{a}}[\mathcal{T}(s'|s,\boldsymbol{a})(\hat{\boldsymbol{\pi}}^{i-1}(\boldsymbol{a}|s) - \boldsymbol{\pi}(\boldsymbol{a}|s))]$*,*
the following decomposition holds:

$$\Delta^{\bar{\pi}^1,\ldots,\bar{\pi}^{i-1},\pi^i,\ldots,\pi^n}_{\pi^1,\ldots,\pi^n} = \Delta^{\bar{\pi}^1,\pi^2,\ldots,\pi^n}_{\pi^1,\ldots,\pi^n} + \Delta^{\bar{\pi}^1,\bar{\pi}^2,\pi^3,\ldots,\pi^n}_{\bar{\pi}^1,\pi^2,\ldots,\pi^n} + \cdots + \Delta^{\bar{\pi}^1,\ldots,\bar{\pi}^{i-1},\pi^i,\ldots,\pi^n}_{\bar{\pi}^1,\ldots,\bar{\pi}^{i-2},\pi^{i-1},\ldots,\pi^n}$$

- The total state transition shift encountered by agent i can be decomposed into the sum of state transition shift caused by each agent whose policy has been updated.
- Sequential update scheme presents a new perspective of tackling the non-stationarity problem.

Sun, Mingfei, et al. "Monotonic improvement guarantees under non-stationarity for decentralized ppo." *AAMAS.* 2022

Sequential Policy Optimization – to Monotonic Improvement

- Recap the multi-agent performance difference lemma, we derive a variant for sequential update:

$$\mathcal{J}(\hat{\boldsymbol{\pi}}^{i}) - \mathcal{J}(\hat{\boldsymbol{\pi}}^{i-1}) = \frac{1}{1-\gamma}\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\hat{\boldsymbol{\pi}}^{i}},\hat{\boldsymbol{\pi}}^{i})}\left[A^{\hat{\boldsymbol{\pi}}^{i-1}}(s,\boldsymbol{a})\right]$$

- Directly, an intuitive surrogate objective is obtained by approximating $d^{\hat{\boldsymbol{\pi}}^{i}}$ using $d^{\boldsymbol{\pi}^{i}}$ and constraining the change between the joint policies:

$$\mathcal{L}^{I}_{\hat{\boldsymbol{\pi}}^{i-1}}\left(\hat{\boldsymbol{\pi}}^{i}\right) = \mathcal{J}(\boldsymbol{\pi}) + \mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi}},\boldsymbol{\pi})}\left[\frac{\hat{\boldsymbol{\pi}}^{i}}{\boldsymbol{\pi}}A^{\boldsymbol{\pi}}(s,\boldsymbol{a})\right]$$

- Can agent i achieve monotonic improvement? **No!**

$$\left|\mathcal{J}\left(\hat{\boldsymbol{\pi}}^i\right)-\mathcal{L}^I_{\hat{\boldsymbol{\pi}}^{i-1}}\left(\hat{\boldsymbol{\pi}}^i\right)\right| \leq 2\epsilon\alpha^i\left(\frac{3}{1-\gamma}-\frac{2}{1-\gamma\left(1-\sum_{j\in\left(e^i\cup\{i\}\right)}\alpha^j\right)}\right)+\frac{\overbrace{2\epsilon\sum_{j\in e^i}\alpha^j}^{\text{Uncontrllable by agent } i}}{1-\gamma}$$

  - Implies that the target policy may not get improved even if α^i is well constrained, since the uncontrollable term could be too large.

- Why?
  - Review the policy iteration in sequential update scheme and performance difference lemma:

$$\hat{\boldsymbol{\pi}}^{i-1} \xrightarrow[\text{Update } \pi^i]{\max_{\pi^i}\mathcal{L}_{\hat{\boldsymbol{\pi}}^{i-1}}\left(\hat{\boldsymbol{\pi}}^i\right)} \hat{\boldsymbol{\pi}}^i$$

$$\mathcal{J}(\hat{\boldsymbol{\pi}}^i)-\mathcal{J}(\hat{\boldsymbol{\pi}}^{i-1})=\frac{1}{1-\gamma}\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\hat{\boldsymbol{\pi}}^i},\boldsymbol{\pi})}\left[\frac{\hat{\boldsymbol{\pi}}^i}{\boldsymbol{\pi}}A^{\hat{\boldsymbol{\pi}}^{i-1}}(s,\boldsymbol{a})\right]$$

$$\mathcal{L}^I_{\hat{\boldsymbol{\pi}}^{i-1}}\left(\hat{\boldsymbol{\pi}}^i\right)=\mathcal{J}(\boldsymbol{\pi})+\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi}},\boldsymbol{\pi})}\left[\frac{\hat{\boldsymbol{\pi}}^i}{\boldsymbol{\pi}}A^{\boldsymbol{\pi}}(s,\boldsymbol{a})\right]$$

$\hat{\boldsymbol{\pi}}^i$ **should be evaluated by** $A^{\hat{\boldsymbol{\pi}}^{i-1}}$ **instead of** $A^{\boldsymbol{\pi}}$

- How about Heterogeneous-agent PPO (HAPPO) ?
  - $\mathcal{L}^{I}_{\hat{\boldsymbol{\pi}}^{i-1}}\left(\hat{\pi}^{i}\right)$ **is equivalent to the surrogate objective of HAPPO.**

Multi-agent state-action value function:

$$Q_{\boldsymbol{\pi}}^{i_{1:m}}\left(s, \boldsymbol{a}^{i_{1:m}}\right) \triangleq \mathbb{E}_{\mathbf{a}^{-i_{1:m}} \sim \boldsymbol{\pi}^{-i_{1:m}}}\left[Q_{\boldsymbol{\pi}}\left(s, \boldsymbol{a}^{i_{1:m}}, \mathbf{a}^{-i_{1:m}}\right)\right]$$

- $i_{1:m}$ denotes an ordered subset $\{i_1, \ldots, i_m\}$ of $\mathcal{N}$, and $-i_{1:m}$ refers to its complement.
- i_k refers to the k^{th} agent in the ordered subset.

Multi-agent advantage function:

$$A_{\boldsymbol{\pi}}^{i_{1:m}}\left(s, \boldsymbol{a}^{j_{1:k}}, \boldsymbol{a}^{i_{1:m}}\right) \triangleq Q_{\boldsymbol{\pi}}^{j_{1:k}, i_{1:m}}\left(s, \boldsymbol{a}^{j_{1:k}}, \boldsymbol{a}^{i_{1:m}}\right) - Q_{\boldsymbol{\pi}}^{j_{1:k}}\left(s, \boldsymbol{a}^{j_{1:k}}\right)$$

- $j_{1:k}$ and $i_{1:m}$ are disjoint sets.

1. Kuba, Jakub Grudzien, et al. "Trust region policy optimisation in multi-agent reinforcement learning." ICLR. 2022.
2. Yaodong Yang's talk. https://www.techbeat.net/talk-info?id=715.

Lemma 1 (Multi-Agent Advantage Decomposition). *In any cooperative Markov games, given a joint policy $\boldsymbol{\pi}$, for any state s, and any agent subset $i_{1:m}$, the below equations holds.*

$$A_{\boldsymbol{\pi}}^{i_{1:m}}\left(s, \boldsymbol{a}^{i_{1:m}}\right) = \sum_{j=1}^{m} A_{\boldsymbol{\pi}}^{i_j}\left(s, \boldsymbol{a}^{i_{1:j-1}}, a^{i_j}\right).$$

- We can re-derive the HAPPO surrogate objectives:

$$\frac{1}{1-\gamma}\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi}},\boldsymbol{\pi})}\left[\frac{\hat{\boldsymbol{\pi}}^{\mathcal{N}}}{\boldsymbol{\pi}^{\mathcal{N}}}A^{\boldsymbol{\pi}^{\mathcal{N}}}(s,\boldsymbol{a})\right]$$
$$=\frac{1}{1-\gamma}\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi}},\boldsymbol{\pi})}\left[\frac{\bar{\boldsymbol{\pi}}}{\boldsymbol{\pi}}A^{\boldsymbol{\pi}^{i_{1:n}}}(s,\boldsymbol{a})\right]$$
$$=\frac{1}{1-\gamma}\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi}},\boldsymbol{\pi})}\left[\frac{\bar{\boldsymbol{\pi}}}{\boldsymbol{\pi}}A^{\boldsymbol{\pi}^{i_{1:n}}}(s,\boldsymbol{a})\right]$$
$$=\frac{1}{1-\gamma}\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi}},\boldsymbol{\pi})}\left[\frac{\bar{\boldsymbol{\pi}}}{\boldsymbol{\pi}}\sum_{j=1}^{n}A^{i_j}(s,a^{i_{1:j-1}},a^{i_j})\right]$$
$$=\frac{1}{1-\gamma}\sum_{j=1}^{n}\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi}},\boldsymbol{\pi})}\left[\frac{\bar{\boldsymbol{\pi}}}{\boldsymbol{\pi}}A^{i_j}(s,a^{i_{1:j-1}},a^{i_j})\right]$$
$$\overset{(a)}{=}\frac{1}{1-\gamma}\sum_{j=1}^{n}\mathbb{E}_{(s,\boldsymbol{a}^{i_{1:j}})\sim(d^{\boldsymbol{\pi}},\boldsymbol{\pi}^{i_{1:j}})}\left[\frac{\bar{\boldsymbol{\pi}}^{i_{1:j}}}{\boldsymbol{\pi}^{i_{1:j}}}A^{i_j}(s,a^{i_{1:j-1}},a^{i_j})\right]$$
$$\overset{(b)}{=}\frac{1}{1-\gamma}\sum_{j=1}^{n}\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi}},\boldsymbol{\pi})}\left[\frac{\hat{\boldsymbol{\pi}}^{i_j}-\hat{\boldsymbol{\pi}}^{i_{j-1}}}{\boldsymbol{\pi}}A^{\boldsymbol{\pi}}(s,\boldsymbol{a})\right]$$

(a): Eliminate the non-random variables
(b): Substitute the definition of the advantage function

- $\hat{\boldsymbol{\pi}}^{i_{j-1}}$ is a constant while updating agent i_j
- The surrogate objective of agent i_j becomes $\frac{1}{1-\gamma}\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi}},\boldsymbol{\pi})}\left[\frac{\hat{\boldsymbol{\pi}}^{i_j}}{\boldsymbol{\pi}}A^{\boldsymbol{\pi}}(s,\boldsymbol{a})\right]$
- Given the order $1,\ldots,n$, we recover $\frac{1}{1-\gamma}\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi}},\boldsymbol{\pi})}\left[\frac{\hat{\boldsymbol{\pi}}^{j}}{\boldsymbol{\pi}}A^{\boldsymbol{\pi}}(s,\boldsymbol{a})\right]=\mathcal{L}^{I}_{\hat{\boldsymbol{\pi}}^{j-1}}\left(\hat{\boldsymbol{\pi}}^{j}\right)$
- **HAPPO also fails in guarantee the monotonic improvement of a single agent.**

$$\left|\mathcal{J}\left(\hat{\boldsymbol{\pi}}^i\right)-\mathcal{L}^I_{\hat{\boldsymbol{\pi}}^{i-1}}\left(\hat{\boldsymbol{\pi}}^i\right)\right| \leq 2\epsilon\alpha^i\left(\frac{3}{1-\gamma}-\frac{2}{1-\gamma\left(1-\sum_{j\in\left(e^i\cup\{i\}\right)}\alpha^j\right)}\right)+\frac{\overbrace{2\epsilon\sum_{j\in e^i}\alpha^j}^{\text{Uncontrllable by agent } i}}{1-\gamma}$$

- The uncontrollable term is caused by one **ignoring how the updating of its preceding agents' policies influences its advantage function.** We investigate reducing the uncontrollable term in policy evaluation.

- Preceding-agent Off-policy Correction (PreOPC):

$$A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}(s_t,\boldsymbol{a}_t)=\delta_t+\sum_{k\geq 1}\gamma^k\Big(\prod_{j=1}^{k}\lambda\min\Big(1.0,\frac{\hat{\boldsymbol{\pi}}^{i-1}(\boldsymbol{a}_{t+j}|s_{t+j})}{\boldsymbol{\pi}(\boldsymbol{a}_{t+j}|s_{t+j})}\Big)\Big)\delta_{t+k}$$

$$\delta_t=r(s_t,\boldsymbol{a}_t)+\gamma V(s_{t+1})-V(s_t)$$

  - We also prove that $A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}$ converges to $A^{\hat{\boldsymbol{\pi}}^{i-1}}$ with probability 1 as the agent i update its value function.

- Retain Monotonic Improvement Bound
  - With PreOPC, the surrogate objective of agent i becomes:

$$\mathcal{L}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i) = \mathcal{J}(\hat{\boldsymbol{\pi}}^{i-1}) + \frac{1}{1-\gamma}\mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi}},\hat{\boldsymbol{\pi}}^i)}[A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}(s,\boldsymbol{a})]$$

(Single Agent Monotonic Bound) For agent i , we have:

$$\left|\mathcal{J}(\hat{\boldsymbol{\pi}}^i) - \mathcal{L}_{\hat{\boldsymbol{\pi}}^{i-1}}(\hat{\boldsymbol{\pi}}^i)\right| \le 4\epsilon^i\alpha^i\Big(\frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\sum_{j\in(e^i\cup\{i\})}\alpha^j)}\Big) + \frac{\xi^i}{1-\gamma} \ ,$$

where $\xi^i = \max_{s,\boldsymbol{a}}\left|A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}(s,\boldsymbol{a}) - A^{\hat{\boldsymbol{\pi}}^{i-1}}(s,\boldsymbol{a})\right|$ converges to 0 with probability 1 as the agent i updates its value function.

- **We retain the monotonic improvement guarantee of a single agent!**

(Multi Agent Monotonic Bound) For agent $i \in \mathcal{N}$, we have:

$$|\mathcal{J}(\bar{\boldsymbol{\pi}}) - \mathcal{G}_{\boldsymbol{\pi}}(\bar{\boldsymbol{\pi}})| \le 4\epsilon \sum_{i=1}^{n} \alpha^i \left(\frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\sum_{j\in(e^i\cup\{i\})}\alpha^j)}\right) + \frac{\sum_{i=1}^{n}\xi^i}{1-\gamma}$$

Algorithm	Update	Monotonic Bound
MAPPO	Simultaneous	$4\epsilon\sum_{i=1}^{n}\frac{\alpha^i}{1-\gamma}$
CoPPO	Simultaneous	$4\epsilon\sum_{i=1}^{n}\alpha^i(\frac{1}{1-\gamma}-\frac{1}{1-\gamma(1-\sum_{j=1}^{n}\alpha^j)})$
HAPPO	Sequential	$4\epsilon\sum_{i=1}^{n}\alpha^i(\frac{1}{1-\gamma}-\frac{1}{1-\gamma(1-\sum_{j=1}^{n}\alpha^j)})$
		Single Agent: No Guarantee
A2PO (ours)	Sequential	$4\epsilon\sum_{i=1}^{n}\alpha^i(\frac{1}{1-\gamma}-\frac{1}{1-\gamma(1-\sum_{j\in(e^i\cup\{i\})}\alpha^j)})+\frac{\sum_{i=1}^{n}\xi^i}{1-\gamma}$
		Single Agent: $4\epsilon^i\alpha^i(\frac{1}{1-\gamma}-\frac{1}{1-\gamma(1-\sum_{j\in(e^i\cup\{i\})}\alpha^j)})+\frac{\xi^i}{1-\gamma}$

Algorithm	Update	Monotonic Bound
MAPPO	Simultaneous	$4\epsilon\sum_{i=1}^{n}\frac{\alpha^i}{1-\gamma}$
CoPPO	Simultaneous	$4\epsilon\sum_{i=1}^{n}\alpha^i(\frac{1}{1-\gamma}-\frac{1}{1-\gamma(1-\sum_{j=1}^{n}\alpha^j)})$
HAPPO	Sequential	$4\epsilon\sum_{i=1}^{n}\alpha^i(\frac{1}{1-\gamma}-\frac{1}{1-\gamma(1-\sum_{j=1}^{n}\alpha^j)})$ Single Agent: No Guarantee
A2PO (ours)	Sequential	$4\epsilon\sum_{i=1}^{n}\alpha^i(\frac{1}{1-\gamma}-\frac{1}{1-\gamma(1-\sum_{j\in(e^i\cup\{i\})}\alpha^j)})+\frac{\sum_{i=1}^{n}\xi^i}{1-\gamma}$ Single Agent: $4\epsilon^i\alpha^i(\frac{1}{1-\gamma}-\frac{1}{1-\gamma(1-\sum_{j\in(e^i\cup\{i\})}\alpha^j)})+\frac{\xi^i}{1-\gamma}$

- Given that $$-\frac{1}{1-\gamma(1-\sum_{j\in(e^i\cup\{i\})}\alpha^j)} < -\frac{1}{1-\gamma(1-\sum_{j=1}^{n}\alpha^j)}$$
- Considering that $\forall i\in\mathcal{N},\ \xi^i$ converges to 0, we **get tighter monotonic improvement bound** compared to previous trust region methods in multi-agent scenarios. **A tighter bound improves target expected performance by optimizing the surrogate objective more effectively.**

Agent-by-agent Policy Optimization

- The practical objective of updating agent i becomes:

$$\mathcal{L}_{\hat{\boldsymbol{\pi}}^{\hat{i}-1}}\left(\hat{\boldsymbol{\pi}}^{i}\right) = \mathbb{E}_{(s,\boldsymbol{a})\sim(d^{\boldsymbol{\pi}},\boldsymbol{\pi})}\left[\min\left(l(s,\boldsymbol{a})A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}, \mathrm{clip}\left(l(s,\boldsymbol{a}), 1\pm\epsilon^{i}\right)A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{i-1}}\right)\right]$$

$$\text{where } l(s,\boldsymbol{a}) = \frac{\bar{\pi}^{i}(a^{i}|s)}{\pi^{i}(a^{i}|s)}g(s,\boldsymbol{a}),\ \text{ and } g(s,\boldsymbol{a}) = \mathrm{clip}\left(\prod_{j\in e^{i}}\frac{\bar{\pi}^{j}(a^{j}|s)}{\pi^{j}(a^{j}|s)}, 1\pm\frac{\epsilon^{i}}{2}\right)$$

- We have obtained a surrogate objective with theoretical strengths.
- How to maximize such objective more effectively?
  1. Formulated as maximization with coordinate ascent $\rightarrow$ the agents updating order matters.
  2. Further reduce the influence of the non-stationarity problem.

- **Semi-greedy Agent Selection Rule**
  - Select the agent to update in order k by:

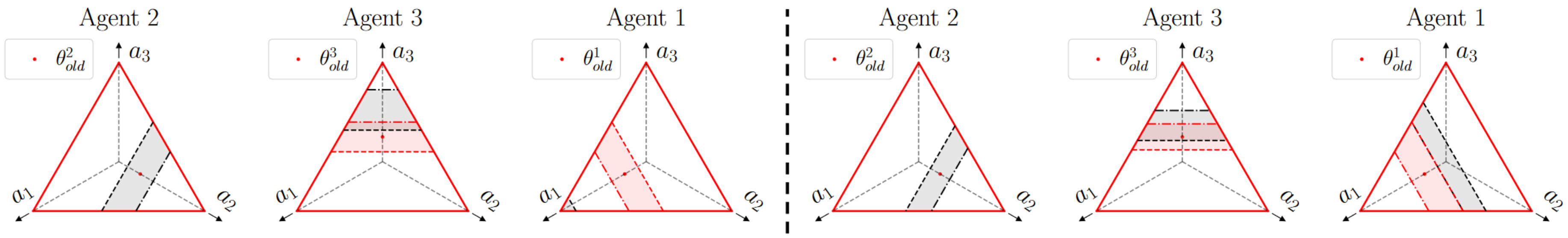
$$\begin{cases} \mathcal{R}(k) = \arg\max_{i \in (\mathcal{N}-e)} \mathbb{E}_{s,a^i}[|A^{\boldsymbol{\pi},\hat{\boldsymbol{\pi}}^{\mathcal{R}(k-1)}}|], & k\%2=0 \\ \mathcal{R}(k) \sim \mathcal{U}(\mathcal{N}-e), & k\%2=1 \end{cases}, \text{ where } e = \{\mathcal{R}(1), \ldots, \mathcal{R}(k-1)\}$$

- **Adaptive Clipping Parameter**
  - From *Non-stationarity Decomposition*, agents with higher priorities contribute more to the non-stationarity problem.

$$\Delta_{\pi^1,\ldots,\pi^n}^{\bar{\pi}^1,\ldots,\bar{\pi}^{i-1},\pi^i,\ldots,\pi^n} = \Delta_{\pi^1,\ldots,\pi^n}^{\bar{\pi}^1,\pi^2,\ldots,\pi^n} + \Delta_{\bar{\pi}^1,\pi^2,\ldots,\pi^n}^{\bar{\pi}^1,\bar{\pi}^2,\pi^3,\ldots,\pi^n} + \cdots + \Delta_{\bar{\pi}^1,\ldots,\bar{\pi}^{i-2},\pi^{i-1},\ldots,\pi^n}^{\bar{\pi}^1,\ldots,\bar{\pi}^{i-1},\pi^i,\ldots,\pi^n}$$

  - Adjust the clipping parameters according to the agent order, leading to more balanced and sufficient clipping ranges.

$$\mathcal{C}(\epsilon, k) = \epsilon \cdot c_\epsilon + \epsilon \cdot (1 - c_\epsilon) \cdot k/n$$



Experiments

- StarCraftII Multi-agent Challenge (SMAC)
- Multi-agent MuJoCo (MA-MuJoCo)
- Google Research Football Full-game Scenarios
- Training Duration

- **StarCraftII Multi-agent Challenge (SMAC)**

Table 5: Median win rates and standard deviations on SMAC tasks. 'w/ PS' means the algorithm is implemented as parameter sharing

Map	Difficulty	MAPPO w/ PS	CoPPO w/ PS	HAPPO w/ PS	A2PO w/ PS	Qmix w/ PS
MMM	Easy	96.9(0.988)	96.9(1.25)	95.3(2.48)	**100(1.07)**	95.3(2.5)
3s_vs_5z	Hard	**100(1.17)**	**100(2.08)**	**100(0.659)**	**100(0.534)**	98.4(2.4)
2c_vs_64zg	Hard	**98.4(1.74)**	96.9(0.521)	96.9(0.521)	96.9(0.659)	92.2(4.0)
3s5z	Hard	84.4(4.39)	92.2(2.35)	92.2(1.74)	**98.4(1.04)**	88.3(2.9)
5m_vs_6m	Hard	84.4(2.77)	84.4(2.12)	87.5(2.51)	**90.6(3.06)**	75.8(3.7)
8m_vs_9m	Hard	84.4(2.39)	84.4(2.04)	96.9(3.78)	**100(1.04)**	92.2(2.0)
10m_vs_11m	Hard	93.8(18.7)	96.9(2.6)	98.4(2.99)	**100(0.521)**	95.3(1.0)
6h_vs_8z	Super Hard	87.5(1.53)	**90.6(0.765)**	87.5(1.49)	**90.6(1.32)**	9.4(2.0)
3s5z_vs_3s6z	Super Hard	82.8(19.2)	84.4(2.9)	37.5(13.2)	**93.8(19.8)**	82.8(5.3)
MMM2	Super Hard	90.6(8.89)	90.6(6.93)	51.6(9.01)	**98.4(1.25)**	87.5(2.6)
27m_vs_30m	Super Hard	93.8(3.75)	93.8(2.2)	90.6(4.77)	**100(1.55)**	39.1(9.8)
corridor	Super Hard	96.9(0)	**100(0.659)**	96.9(0.96)	**100(0)**	84.4(2.5)
overall	/	91.1(5.46)	92.6(2.2)	85.9(3.68)	**97.4(2.65)**	78.4(3.6)

- **Multi-agent MuJoCo (MA-MuJoCo)**

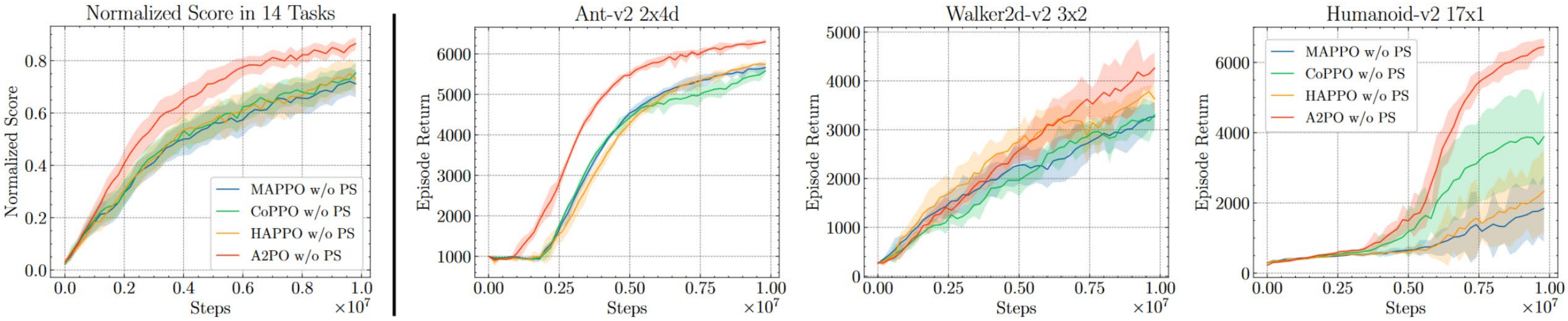


Figure 3: Experiments in MA-MuJoCo. **Left**: Normalized scores on all the 14 tasks. **Right**: Comparisons of averaged return on selected tasks. The number of robot joints increases from left to right.

• Google Research Football Full-game Scenarios

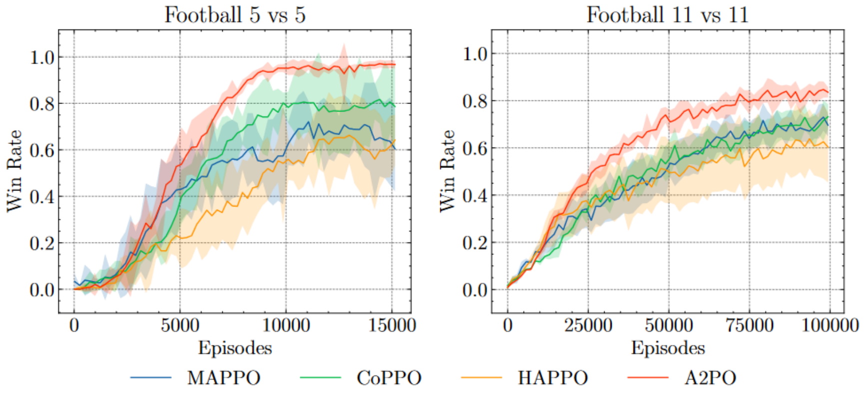
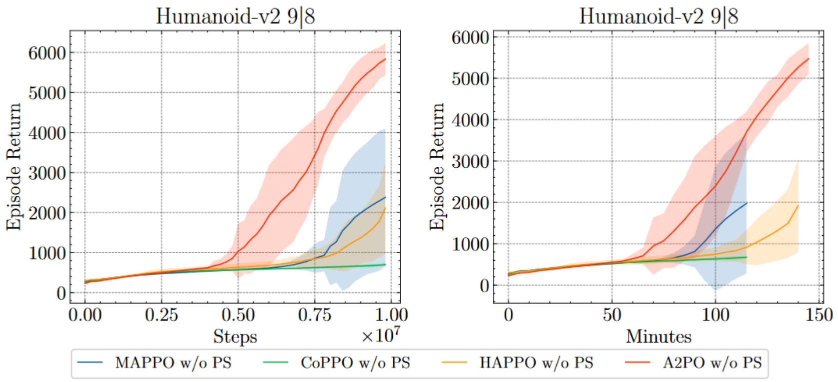


Figure 4: Averaged win rate on the Google Research Football full-game scenarios.

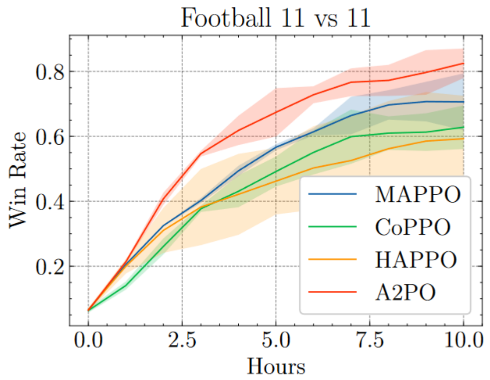
Table 3: Learned behaviors on the Google Research Football 5-vs-5 scenario. Bigger values are better except fot the 'Lost' metric.

Metric	MAPPO	CoPPO	HAPPO	A2PO
Assist	0.04(0.02)	0.19(0.08)	0.07(0.05)	**0.56(0.20)**
Goal	1.95(1.17)	4.42(2.08)	2.68(0.86)	**9.01(0.95)**
Lost	**0.49(0.11)**	0.74(0.33)	1.04(0.12)	0.78(0.15)
Pass	1.52(0.13)	3.44(1.04)	4.03(1.97)	**6.42(2.23)**
Pass Rate	19.3(10.0)	35.0(10.3)	48.9(25.7)	**67.1(11.7)**

• Training Duration



(a) Comparison on Humanoid 9|8 over both environment steps and training time.



(b) Comparison on GRF 11-vs-11 scenario.

Table 6: The comparison of training duration. The format of the first line in a cell is: Training time(Sampling time+Updating Time). The second line of a cell represents the time normalized.

Task	MAPPO	CoPPO	HAPPO	A2PO
3s5z	3h29m(3h3m+0h26m) 1.00(0.87 + 0.13)	3h33m(3h6m+0h27m) 1.02(0.89 + 0.13)	3h49m(3h7m+0h42m) 1.10(0.89 + 0.20)	4h32m(3h41m+0h51m) 1.30(1.06 + 0.25)
27m vs 30m	13h23m(8h31m + 4h52m) 1.00(0.64 + 0.36)	13h19m(8h24m + 4h55m) 1.00(0.63 + 0.37)	16h2m(8h20m + 7h42m) 1.20(0.62 + 0.58)	15h53m(8h7m + 7h46m) 1.19(0.61 + 0.58)
Humanoid 9|8	2h0m(1h45m + 0h15m) 1.00(0.87 + 0.13)	1h58m(1h43m + 0h15m) 0.99(0.86 + 0.13)	2h15m(1h45m + 0h30m) 1.12(0.87 + 0.25)	2h31m(2h0m + 0h31m) 1.26(1.00 + 0.26)
Ant 4x2	6h42m(6h16m + 0h26m) 1.00(0.93 + 0.07)	6h45m(6h19m + 0h26m) 1.01(0.94 + 0.07)	7h29m(6h5m + 1h24m) 1.12(0.91 + 0.21)	7h2m(5h34m + 1h28m) 1.05(0.83 + 0.22)
Humanoid 17x1	12h9m(10h6m + 2h3m) 1.00(0.83 + 0.17)	17h7m(15h5m + 2h2m) 1.41(1.24 + 0.17)	16h55m(11h2m + 5h53m) 1.39(0.91 + 0.48)	19h25m(11h59m + 7h26m) 1.60(0.99 + 0.61)
Football 5vs5	34h46m(32h47m + 1h59m) 1.00(0.94 + 0.06)	32h46m(30h49m + 1h57m) 0.94(0.89 + 0.06)	39h26m(31h54m + 7h32m) 1.13(0.92 + 0.22)	37h26m(30h2m + 7h24m) 1.08(0.86 + 0.21)

Summary

1. Brief introduction of Cooperative MARL
2. Serial Progress:
   - MAPPO: PPO in CTDE scheme
   - CoPPO: Coordinate the agents via the joint policy
   - HAPPO: Advantage function decomposition
3. How to Retain Monotonic Improvement Guarantee in Sequential Policy Optimization and Tighten the Monotonic Improvement Bound
4. A Practical Algorithm: Agent-by-agent Policy Optimization
5. More Efficient Surrogate Objective Maximization