UNIVERSITY OF ALBERTA

PMIC: Improving Multi-Agent Reinforcement Learning with Progressive Mutual Information Collaboration

Multi-Agent Reinforcement Learning

➢ Current Status:

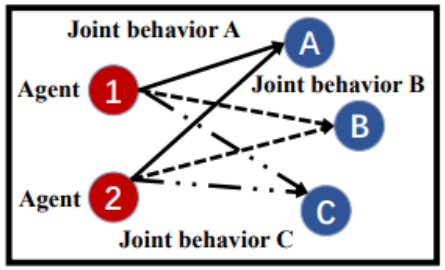
- ➢ Learning to collaborate is critical in multi-agent reinforcement learning.
- ➢ Centralized Training with Decentralized Execution (CTDE) is a mainstream frameworks (Non-stationary).
- ➢ Many CTDE-based MARL algorithms are proposed including MADDPG, MASAC, VDN and QMIX.
- ➢ Optimizing the decentralized policies of multiple agents only through reward signals is often inefficient, especially when the reward signals are stochastic or sparse. (additional machansims are often critical to facilitating effective collaboration).
- ➢ A complementary branch of works proposes to leverage the correlation or influence of agents.
- ➢ Agents with high correlation behaviors (influence) are more likely to form collaboration. Motivated by this, previous works propose to maximize the correlation of agents' behaviors to promote collaboration.

➢ Our focus:

- ➢ SIC (Chen et al., 2021) shared signals z and the joint policy (i.e., $I(z;\pi)$)
- ➢ MAVEN (Mahajan et al., 2019) shared signals z and the trajectories (i.e., $I(z;\tau)$)
- ➢ SI (Jaques et al., 2019) any two agents' action (i.e., $I(a_i;a_j|s)$)
- ➢ SI-MOA (Jaques et al., 2019) one agent' current action and the other agent' next action (i.e., $I(a_{t+1}^i;a_t^j|s_t^j)$)
- ➢ VM3-AC (Kim et al., 2020) any two agents' action (i.e., $I(a_i;a_j|s,z)$)

Can maximizing influence or correlation (MI) of agents ensure good collaboration?

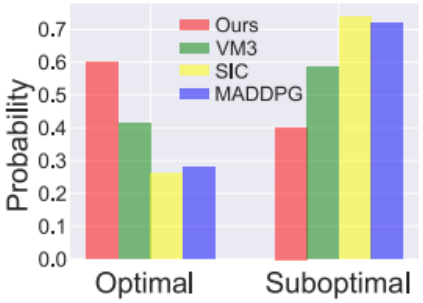
Why Can MI-based Collaboration Fail?



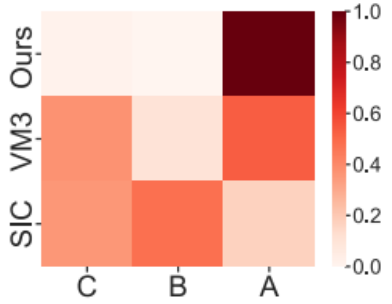
(a) A motivating example

	A	B	C	N
A	11	-30	0	-30
B	-30	7	6	-30
C	0	6	5	0
N	-30	-10	0	0

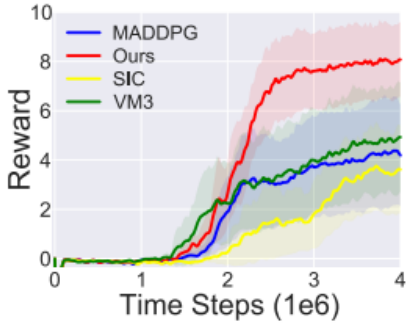
(b) Reward matrix



(c) Behavior distribution

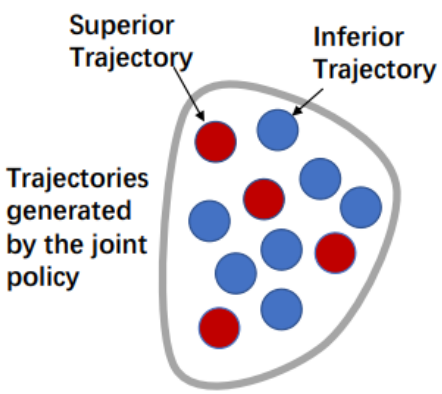


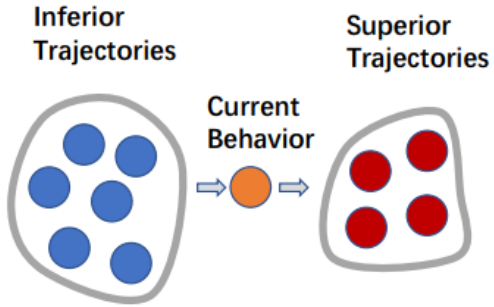
(d) Degree of correlation



(e) Learning performance

The optimal joint behavior here is to rescue target A collaboratively, while other joint behaviors lead to sub-optimal collaborations.

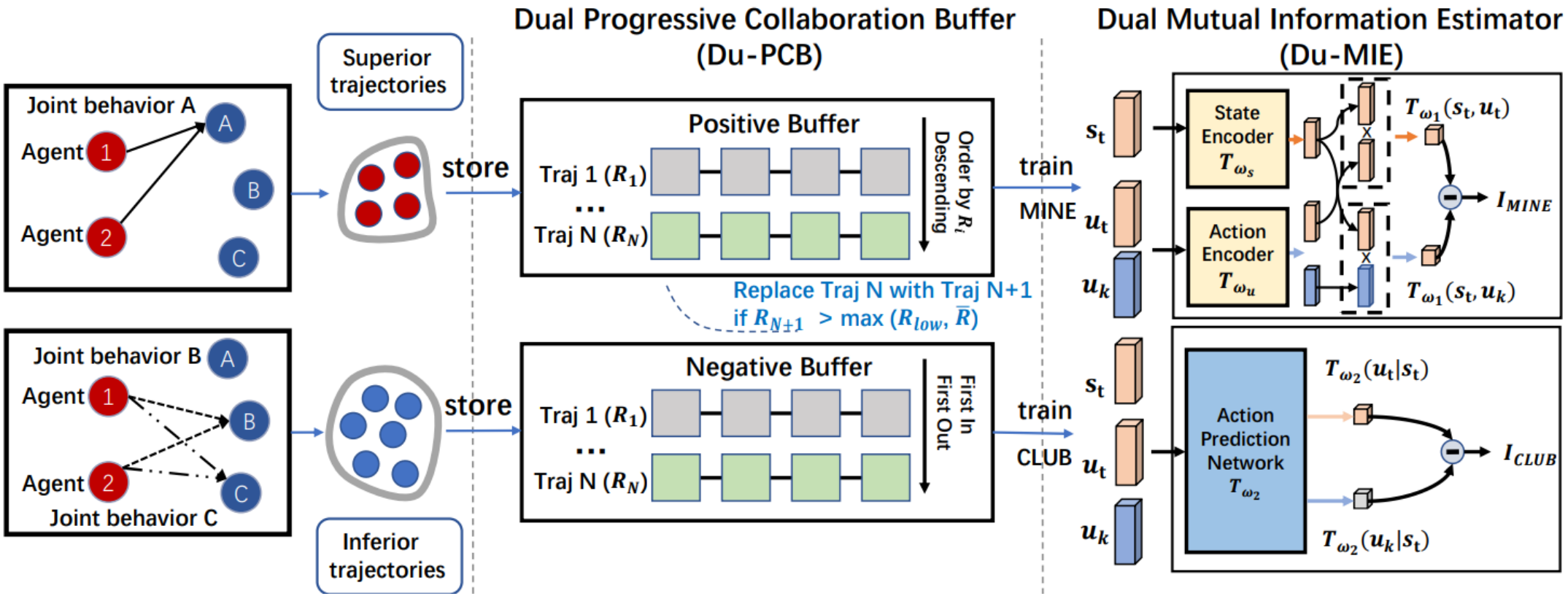




➢**Our idea:**

To achieve an ideal learning process, agents not only 1) need to enhance the correlation of their joint behaviors to form collaborations, but also 2) need to be capable of escaping from a sub-optimal collaboration to reach a better one.

Progressive Mutual Information Collaboration (PMIC)



A New Collaboration Criterion

- Previous MI forms:
  - $I(z;\tau)$
  - $I(z;\pi)$
  - $I(a_i, a_j|s,z)$, $I(a_{t+1}^i; a_t^j|s_t^j)$, $I(a_i, a_j|z)$
  - $I(s_{t+1}^j; s_t^i, a_t^i)$
- Limitations:
  - measuring correlation with the MI of any two agents' actions can be computationally infeasible with the increase of the number of agents (i.e., the scalability issue)
  - additional shared latent variables and the joint policy (or trajectories) which violates the CTDE paradigm and makes the methods fail in some real-world deployment scenarios when global communication is not available during execution
- We first propose a new criterion to measure the degree of multiagent collaboration

$$\begin{aligned} \boldsymbol{I(s;u)} &= \boldsymbol{H(u) - H(u \mid s)} \\ &= \boldsymbol{H(u) - H(u_i \mid s) - H(u_{-i} \mid u_i, s)} \end{aligned}$$

  - $\boldsymbol{H(u)}$ describes the ability to explore various behaviors of all agents (via joint actions), which could help generate diverse trajectories and avoid policy collapse when maximized
  - $\boldsymbol{-H(u_i|s)}$ measures the behavioral uncertainty of agent i, which encourages the agent to behave deterministically given global state s when minimized
  - $\boldsymbol{-H(u_{-i}|u_i, s)}$ measures the uncertainty of agent i about the actions of other agents, which implicitly characterizes the correlation between agents' behavior and will drives agents to coherent joint behaviors when minimized.

Dual Progressive Collaboration Buffer (Du-PCB)

- To achieve the main idea, we introduce Du-PCB which includes a positive and a negative buffer to dynamically keep the superior and inferior collaboration separately.
  - Positive Buffer: this buffer only keeps superior trajectories, based on the episodic return. The new trajectory return is R_k, the lowest return in the buffer is R_{low} , the current policy performance is $\bar{R}$. If $R_k > \max(R_{\text{low}}, \bar{R})$, the trajectory will be added to the buffer. When the buffer is full, DPCB overwrites the samples with the worst episodic return to ensure the collaboration patterns' quality could be monotonically increased.
  - Negative Buffer: We consider collaboration patterns which are not added into the positive buffer as inferior patterns. The negative buffer is updated according to the first-in first-out (FIFO) manner.

Dual Mutual Information Estimator (Du-MIE)

Maximize mutual information associated with superior collaboration

- To maximize the MI associated with the positive buffer, we leverage MINE to measure MI which provides a tight lower bound based on Jensen-Shannon divergence as:

$$I(s;u) \geq I_{MINE}(s;u) = \sup_{\omega_1} \underbrace{\left[\mathbb{E}_{\mathbb{P}_{su}}\left[-sp\left(-T_{\omega_1}(s,u)\right)\right] - \mathbb{E}_{\mathbb{P}_s \otimes \mathbb{P}_u}\left[sp\left(T_{\omega_1}(s,u)\right)\right]\right]}_{-\mathcal{L}(\omega_1)}$$

Minimize mutual information associated with inferior collaboration

- To ensure that agents do not fall into sub-optimal collaboration, the agents should always be as different from the suboptimal collaboration as possible:

$$I(s;u) \leq I_{CLUB}(s;u) = \underbrace{\mathbb{E}_{\mathbb{P}_{su}}\left[\log T_{\omega_2}(u \mid s)\right]}_{-\mathcal{L}(\omega_2)} - \mathbb{E}_{\mathbb{P}_s \otimes \mathbb{P}_u}\left[\log T_{\omega_2}(u \mid s)\right]$$

Integrate PMIC with MARL algorithms

- we propose a new objective function for PMIC-MARL that combines the two types of MI estimates (as additional per-step rewards) with the conventional objective:

$$J^{\mathrm{PMIC}}(\pi) = \mathbb{E}_{s,u\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t\left(r_t + r_t^{\mathrm{PMIC}}\right)\right]$$

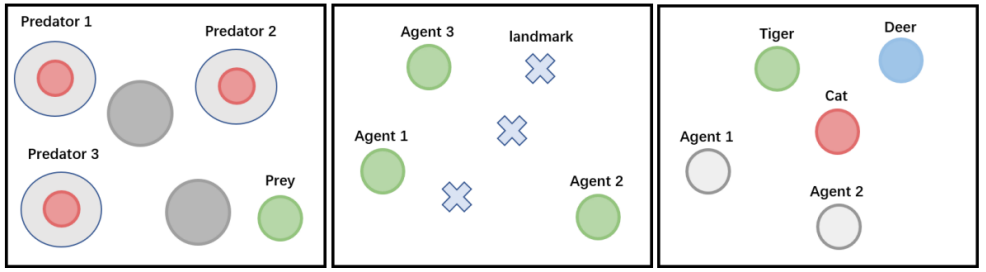
$$r_t^{\mathrm{PMIC}} = \alpha I_{\mathrm{MINE}}(s_t;u_t) - \beta I_{\mathrm{CLUB}}(s_t;u_t))$$

Algorithm 1: PMIC-MADDPG

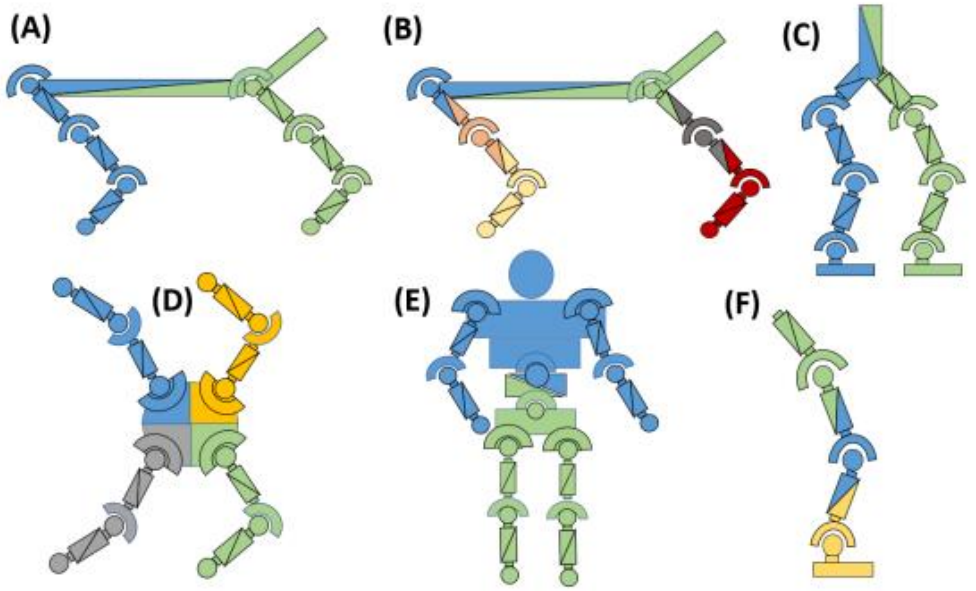
Input: the update frequency k for Du-MIE and maximum episode length T.
Initialize the critic network ϕ, n actor networks $\theta_1 ... \theta_n$ and corresponding target networks ϕ', ${\theta_1}', \ldots, {\theta_n}'$.
Initialize Du-MIE parameterized by ω_1 and ω_2.
Initialize Du-PCB and experience replay buffer $\mathcal{D}$
repeat
  for $t = 1, ..., T$ **do**
    Execute joint actions u_t via collecting $u_t^i \sim \pi_{\theta_i}(o_t^i)$.
    Receive $o_{t+1} = \{o_{t+1}^i\}_{i=1}^n$ and team reward r_t.
  Store trajectory $\nu = \{o_t, u_t, o_{t+1}, r_t\}_{t=1}^T$ to D
  if $R_\nu > \max(R_{low}, \bar{R})$ **then**
    Add ν to the positive buffer
  else
    Add ν to the negative buffer
  Update Du-MIE with Du-PCB every k steps ▷ see Eq. 4
  Update the actors and critic networks ▷ see Eq. 6
until *reaching maximum training steps*;

Experiments

- **Environments**
  - MPE (Lowe et al.,2017)
  - Multi-Agent MuJoCo (de Wittet al., 2020)
  - SMAC (Samvelyan et al., 2019)
- **Baselines**
  - MADDPG (Lowe et al.,2017)
  - VM3-AC (Kim et al., 2020)
  - MASAC (Kim et al., 2020)
  - Fac-MADDPG and COMIX (de Witt et al., 2020)
  - SIC-MADDPG (Chen et al., 2019)
  - RODE(Wang et al., 2020)

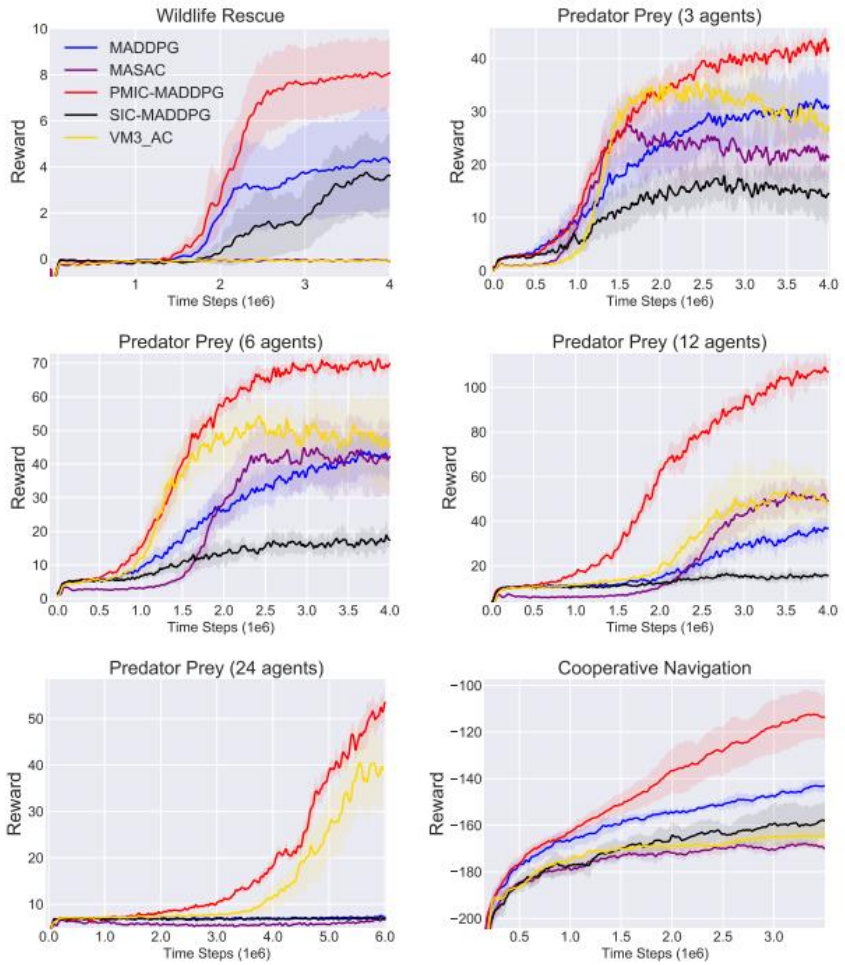


(a) Predator-Prey (b) Cooperative Navigation (c) Wildlife Rescue

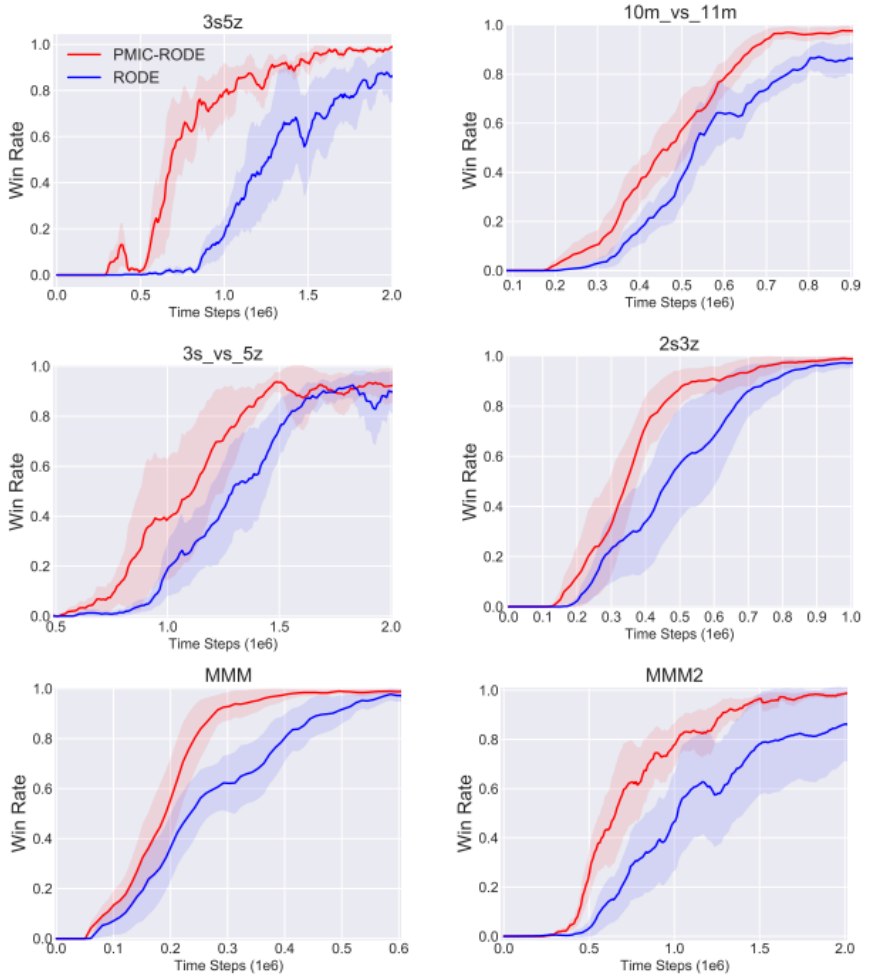


Experiments

- **Integrate PMIC with MADDPG (MPE)**

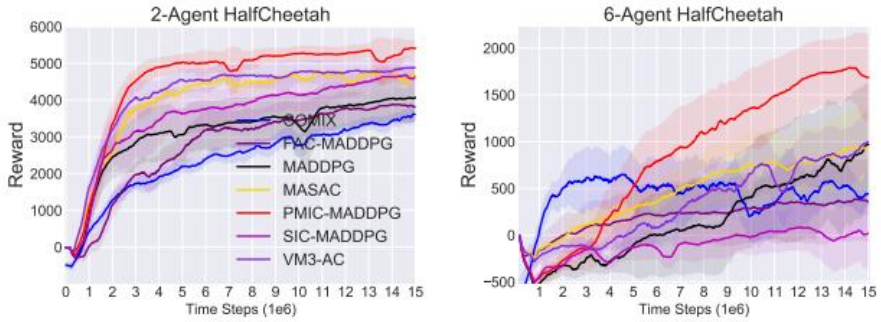


- **Integrate PMIC with RODE (SMAC)**

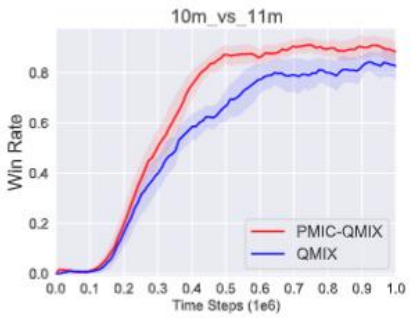


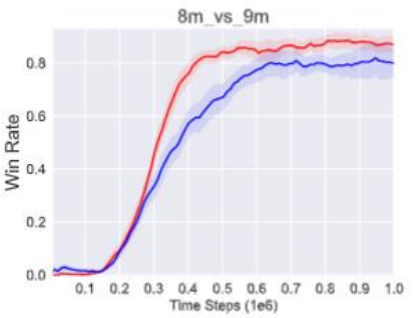
Experiments

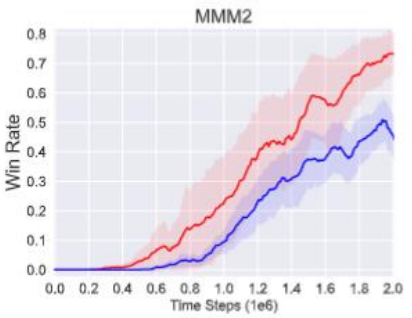
- **Integrate PMIC with MADDPG on MAMuJoCo**



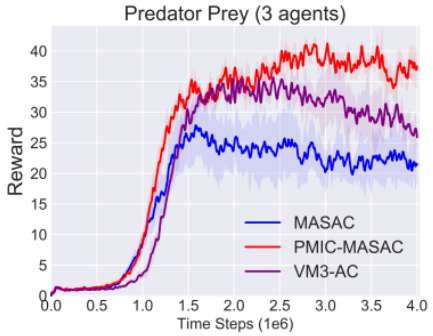
- **Integrate PMIC with QMIX on SMAC**

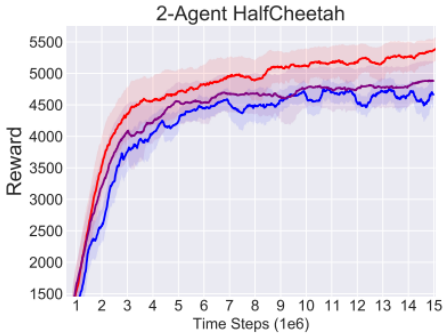


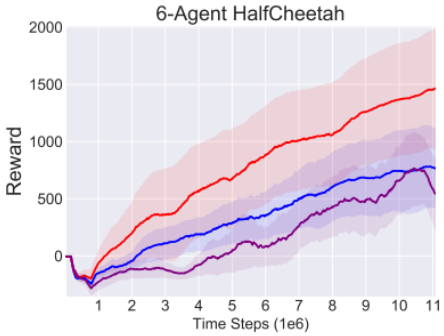




- **Integrate PMIC with MASAC on MPE and MAMuJoCo**

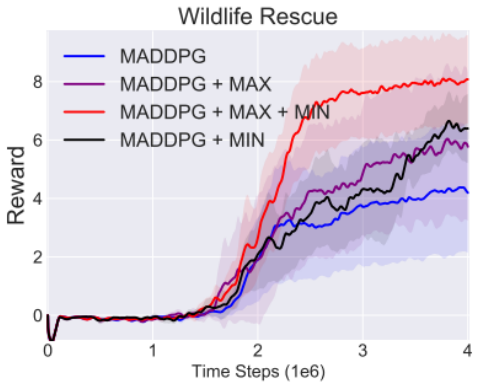


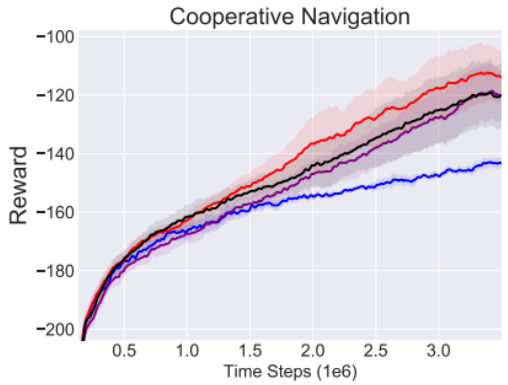




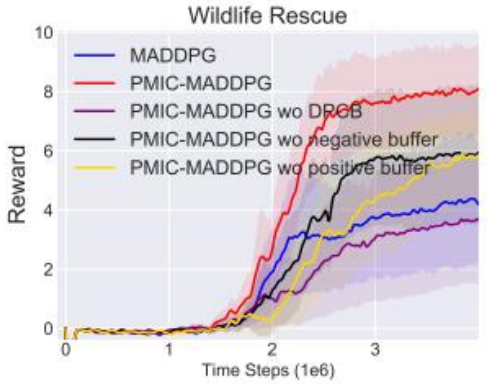
Experiments

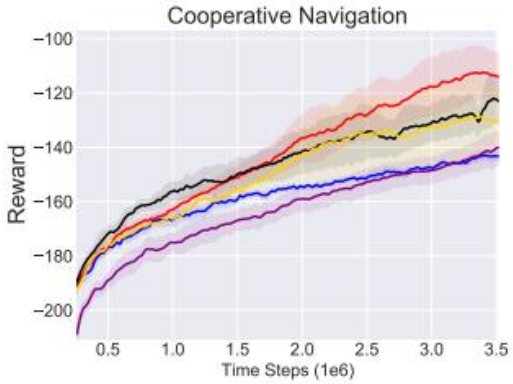
- **Ablation on MI maximization & minimization**



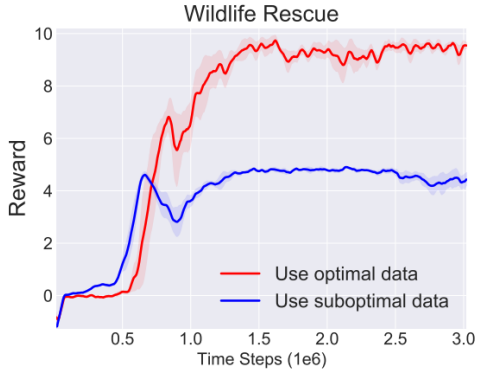


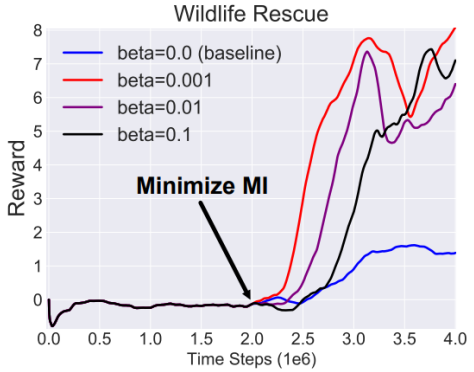
- **Ablation on Du-PCB**





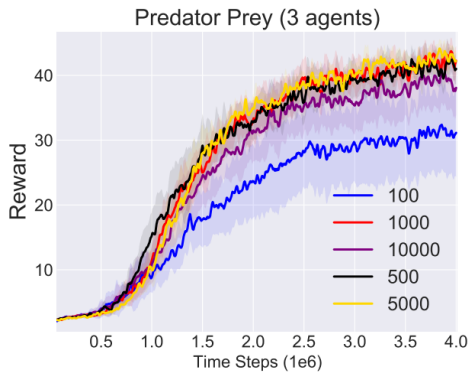
- **maximizing MI associated with optimal and suboptimal data to guide agents. Right: minimizing MI to break inferior behaviors**

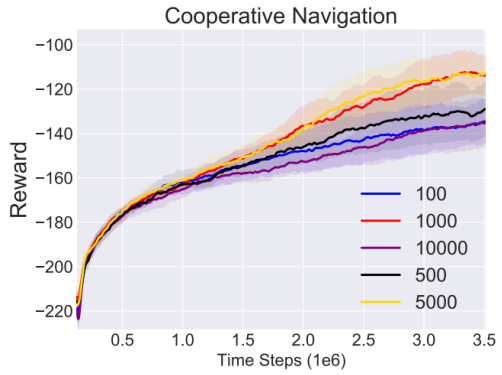




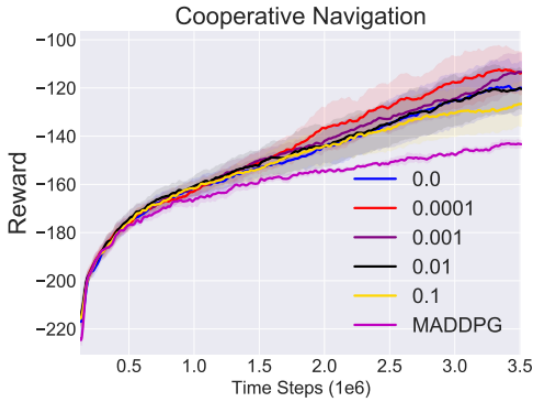
Experiments

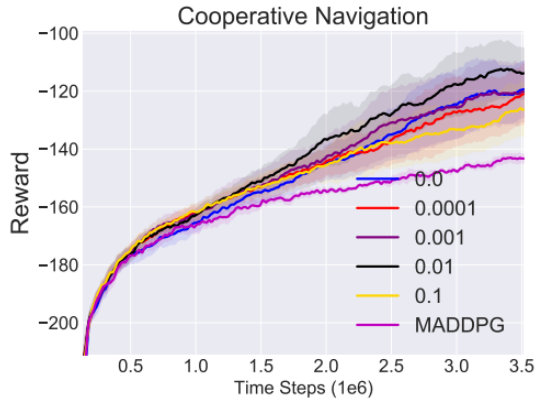
- **Influence of Du-PCB size**



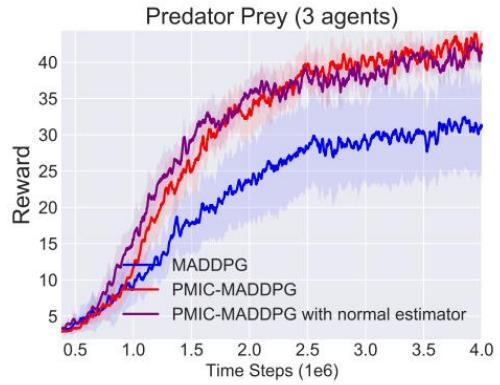


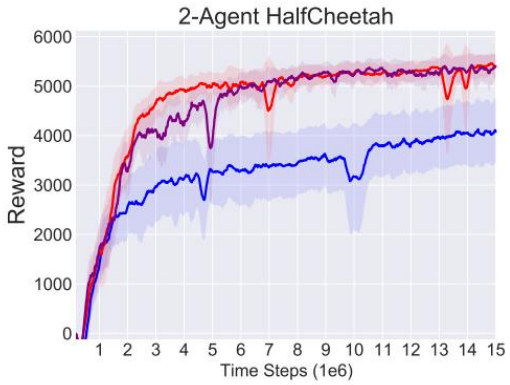
- **Ablation on α and β**



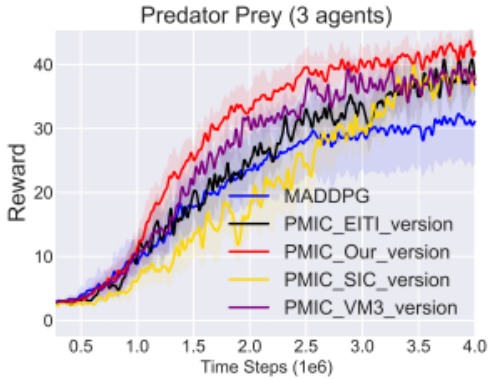


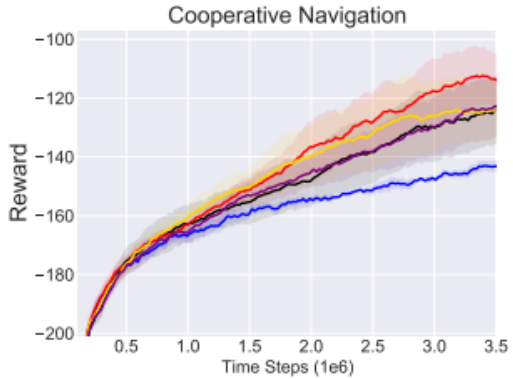
- **Comparison of PMIC with MINE and PMIC with normal estimator.**





- **Comparison of PMIC with different MI forms.**





Conclusion & Limitations & Future work

- **Conclusion：**
  - To address the potentially detrimental effects of only maximizing mutual information, we propose the PMIC framework with maximizing and minimizing MI.
  - In our experiments, we evaluate several implementations of PMIC-MARL in a wide range of cooperative environments with both continuous action space and discrete action space. The results demonstrate the effectiveness and generalization of PMIC.

- **Limitations & Future work:**
  - Lack of theory support
  - Sensitive to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ (Fixed hyperparameters)
  - More precise methods to sort data.
  - More efficient forms of mutual information to measure collaboration/influence.
  - More applications such as communication in MARL and hierarchical reinforcement learning (HRL).