

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

Pengyi Li¹ Hongyao Tang¹ Tianpei Yang^{1,2} Xiaotian Hao¹ Tong Sang¹ Yan Zheng¹ Jianye Hao¹
 Matthew E.Taylor² Wenyuan Tao¹ Zhen Wang³

Abstract

Learning to collaborate is critical in Multi-Agent Reinforcement Learning (MARL). Previous works promote collaboration by maximizing the correlation of agents' behaviors, which is typically characterized by Mutual Information (MI) in different forms. However, we reveal sub-optimal collaborative behaviors also emerge with strong correlations, and simply maximizing the MI can, surprisingly, *hinder* the learning towards better collaboration. To address this issue, we propose a novel MARL framework, called Progressive Mutual Information Collaboration (PMIC), for more effective MI-driven collaboration. PMIC uses a new collaboration criterion measured by the MI between global states and joint actions. Based on this criterion, the key idea of PMIC is maximizing the MI associated with superior collaborative behaviors and minimizing the MI associated with inferior ones. The two MI objectives play complementary roles by facilitating better collaborations while avoiding falling into sub-optimal ones. Experiments on a wide range of MARL benchmarks show the superior performance of PMIC compared with other algorithms.

1. Introduction

With the potential to solve complex real-world problems, Multi-Agent Reinforcement Learning (MARL) has attracted much attention in recent years (Lyu et al., 2022; Yang et al., 2021b; Zheng et al., 2020; Wang et al., 2020c; Hernandez-Leal et al., 2019) and has been applied to many practical domains like Game AI (Peng et al., 2017), Robotics Con-

¹College of Intelligence and Computing, Tianjin University, China ²University of Alberta, Canada ³Northwestern Polytechnical University, China. Correspondence to: Yan Zheng <yanzheng@tju.edu.cn>, Jianye Hao <jianye.hao@tju.edu.cn>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

trol (Matignon et al., 2012), Transportation (Li et al., 2019). However, efficiently achieving collaboration and learning optimal policies still remains challenging in MARL (Liu et al., 2020; Wen et al., 2019; Yang et al., 2021a; Zheng et al., 2018a;b).

Centralized Training with Decentralized Execution (CTDE) (Rashid et al., 2018; Sunehag et al., 2017) is a popular MARL paradigm, adopted to promote collaboration among agents. During centralized training, agents are granted access to other agents' information and possibly the global state, while during decentralized execution, agents make decisions independently based on their individual policies. There are many CTDE-based MARL algorithms being proposed, including MADDPG (Lowe et al., 2017), MASAC (Kim et al., 2020), VDN (Sunehag et al., 2017), and QMIX (Rashid et al., 2018). However, although global information is incorporated during centralized CTDE training, optimizing the decentralized policies of multiple agents only through reward signals is often inefficient, especially when the reward signals are stochastic or sparse — *therefore, additional mechanisms are often critical to facilitating effective collaboration*. A complementing branch of works proposes to leverage the correlation or influence of agents (Jaques et al., 2018; 2019; Xie et al., 2020; Liu et al., 2020; Merhej & Chetouani, 2021). The intuition behind these works is that if agents make decisions that account for their influence on the behaviors of other agents, the problem of non-stationarity could be mitigated, and thus agents are more likely to achieve collaboration.

Several works (Chen et al., 2021; Mahajan et al., 2019; Kim et al., 2020) proposed to maximize the correlation of agents' behaviors to promote collaboration, which commonly quantifies the correlation of agents' behaviors by the mutual information (MI). Unfortunately, these previous works overlook the fact that agents with a high degree of collaboration may not necessarily generate high rewards. In complex environments, there exist multiple types of collaborations differing in return when agents achieve them — *simply maximizing the MI of agents' behaviors cannot guarantee high-quality collaboration* because agents in sub-optimal collaborations can also have a high degree of correlation.

Furthermore, maximizing MI exacerbates the problem: an agent can easily overfit its strategy to the behaviors of other agents (Zhang et al., 2019; Lanctot et al., 2017).

To solve the problem, this paper proposes a novel framework, called **Progressive Mutual Information Collaboration** (PMIC). In PMIC, a new collaboration criterion is measured by the MI between global states and joint actions, freeing us from relying on additional global input (existing in previous works) and addressing the scalability issue. Based on the new criterion, PMIC uses two main components to promote agents' collaboration. The first component is the *Dual Progressive Collaboration Buffer* (Du-PCB), which includes a positive and a negative buffer to dynamically maintain data about superior and inferior collaborations, respectively. The second component is the *Dual Mutual Information Estimator* (Du-MIE), which employs two MI neural estimators to estimate the MI of global states and joint actions for transitions in Du-PCB. In particular, one estimator is trained on the positive buffer to provide the lower bound of MI and the other is trained on the negative buffer to provide the upper bound of MI. By maximizing the lower bound and minimizing the upper bound, the agents can progressively break the current sub-optimal collaboration and learn to achieve better ones, which thus promotes an efficient and stable learning process. Importantly, PMIC is general and can be easily combined with existing MARL algorithms. Our experiments show that PMIC significantly accelerates existing MARL algorithms, outperforming other baseline algorithms on a wide range of MARL benchmarks.

In summary, our contributions are threefold: First, we reveal that simply maximizing the correlation of agents without distinguishing what kind of behaviors are expected can hinder the learning towards better collaborations. Secondly, we propose a novel framework PMIC to solve the problem, including a new collaboration criterion and a progressive MI estimation designed by Du-MIE and Du-PCB. Last but not least, we build PMIC on many MARL algorithms such as MADDPG (Lowe et al., 2017), MASAC (Kim et al., 2020), RODE (Wang et al., 2020a) and show their superior performance by comparing them with other competitive methods on a wide range of MARL benchmarks.

2. Background

This section presents the necessary background to understand PMIC and its relationship to existing works.

2.1. Preliminaries

We consider a fully cooperative multi-agent task where a team of agents are situated in a stochastic, partially observable environment, it can be modeled as a *decentralised partially observable Markov decision process* (Dec-POMDP)

(Oliehoek & Amato, 2016), which can be defined as a tuple: $\langle \mathcal{N}, \mathcal{S}, \mathcal{U}, \mathcal{O}, \mathcal{T}, \mathcal{R}, \gamma \rangle$. Here $\mathcal{N} = \{1, \dots, N\}$ denotes the set of N agents. In Dec-POMDP, the full state of the environment $s_t \in \mathcal{S}$ cannot be observed by agents at each time step t . Each agent $i \in \mathcal{N}$ can only observe its individual observation o_t^i determined by observation function $\mathcal{O}(s_t, i)$, each agent i uses a stochastic policy π_i to choose actions $u_t^i \sim \pi_i(\cdot | o_t^i)$, yielding the joint action $u_t = \{u_t^i\}_{i=1}^N \in \mathcal{U}$. After executing u_t in state s_t , the environment transits to the next state s_{t+1} according to transition function $\mathcal{T}(s_t, u_t)$ and agents receive a common reward r_t from $\mathcal{R}(s_t, u_t)$, with a discount factor $\gamma \in [0, 1]$. We denote the joint policy as $\pi = (\pi_1, \pi_2, \dots, \pi_N) \in \Pi$, where Π is the joint policy space. In cooperative MARL, the collaborative team aims to find a joint policy to maximize the total expected discounted return, denoted by $J(\pi) = \mathbb{E}_\pi [\sum_{t=0}^{\infty} \gamma^t r_t]$.

2.2. Related Work

Centralized training & decentralized execution (CTDE) has been a major paradigm in recent efforts in MARL. For example, MADDPG (Lowe et al., 2017) uses a centralized critic to train decentralized policies. VDN (Sunehag et al., 2017), QMIX (Rashid et al., 2018), MAAC (Iqbal & Sha, 2019), COMIX, and FAC-MADDPG (de Witt et al., 2020) achieve CTDE through value function factorisation.

MI-based collaboration MARL: Many existing algorithms explicitly maximize the correlation or influence of agents to facilitate collaboration, where the correlation or influence is often quantified by the MI of the agent's behavior. For example, Signal Instructed Coordination (SIC) (Chen et al., 2021) takes a holistic view and facilitates collaboration by increasing the MI of the agent's behavior and the joint policy. Specifically, SIC extracts the information of the joint policy into the latent variables z (sampled from a predefined distribution), which are then used as the input of agents' policy networks. SIC then maximizes the MI of each agent's behaviors and the latent variables z to improve the correlation of agents. Based on the latent variables, the agents can know what kind of joint policy the whole team is executing and what kind of action should be selected. Multi-agent Variational Exploration (MAVEN) (Mahajan et al., 2019) shares a similar idea with SIC, except that MAVEN extracts the latent variables about joint policy information from the initial global state and maximizes the mutual information of future trajectories and the latent variables. However, one common drawback of both methods is the shared latent variable required during decentralized execution violates the CTDE paradigm. This makes algorithms fail in some real-world deployment scenarios where global communication is not available. EITI (Wang et al., 2020b) leverages MI to capture the influence between one agent's current actions/states and the other agents' transitions in grid environments. SI (Jaques et al., 2019) proposes a social

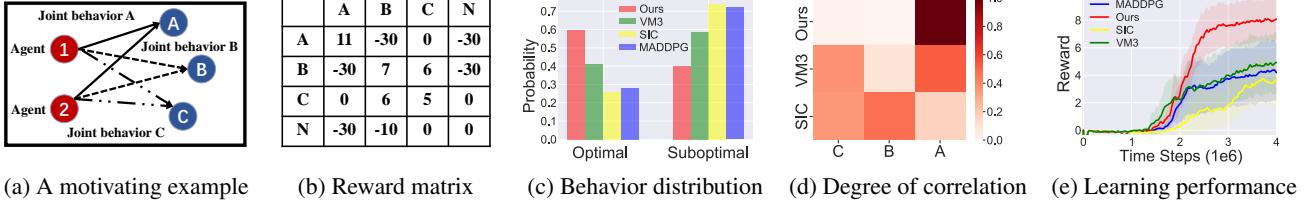


Figure 1: (a) A motivating example: *Wildlife Rescue* (Lowe et al., 2017). The optimal joint behavior here is to rescue target A collaboratively, while other joint behaviors lead to sub-optimal collaborations. (b) The reward matrix for the agents capturing different targets at the end of the game, where ‘N’ means an agent does not catch any target. (c) The probability of optimal and sub-optimal joint behaviors learned by different algorithms, averaged over 10k (1k x 10 seeds) episodes. (d) The degree of correlation is measured by different algorithms for different joint behaviors. (e) The performance of episodic rewards averaged over 10 seeds with 95% confidence regions for different algorithms.

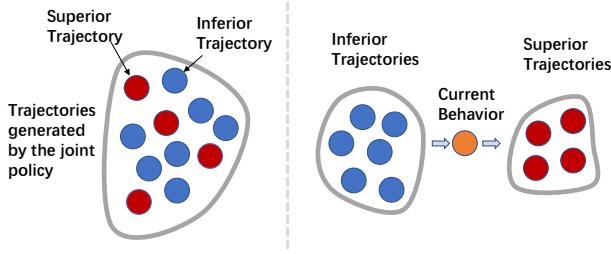


Figure 2: *Left*: Trajectories collected during the learning process contain a mix-up of different joint behaviors. *Right*: An ideal learning process that identifies the distinctions and performs progressive improvement.

influence intrinsic reward measured by the MI between any two agents’ actions to achieve coordination in sequential social dilemmas. SI-MOA (Jaques et al., 2019) extends this idea to CTDE by modeling the other agents’ actions and promotes coordination by the MI between one agent’s current action and the other agent’ next action. VM3-AC (Kim et al., 2020) also tries to extend SI to CTDE. VM3-AC modifies policy iteration based on the MI of any two agents’ current action and introduces additional input to explicitly represent the relation of agents’ policies, and VM3 achieves better performance than previous methods.

Unfortunately, these methods may fall into the trap of blindly maximizing MI, resulting in a high degree of correlation, but failing to achieve high-performing policies. The next section will explain this failure mode, while Section 4 will provide a solution to avoid the problem.

3. Why Can MI-based Collaboration Fail?

This section motivates our approach, showing why blindly enhancing the correlation of agents’ behaviors can lead to agents falling into sub-optimal collaborations. Figure 1a illustrates a motivating example where two agents need to collaborate to rescue three targets (i.e., A, B, and C), and

receive a team reward (Figure 1b). The joint behaviors of the two agents in this game are various, and Figure 1a presents three kinds of joint behaviors. Apparently, the expected joint behavior is that both agents collaborate to rescue target A, since this achieves the highest reward.

Due to the stochasticity of the environment and the learning dynamics of agents’ policies, trajectories of different behaviors are collected by the two agents’ joint policy, as shown in the left of Figure 2. These trajectories mix up many joint behaviors of a high collaboration degree, which have distinct outcomes (e.g., rescue B or C). Intuitively, to achieve an ideal learning process as depicted in the right of Figure 2, 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. The contradiction between the two objectives indicates that the correlation should be enhanced and loosened in an adaptive manner.

Existing MARL methods with MI-based collaboration only focus on the first point, maximizing MI while neglecting the second. In principle, this can be problematic because the consistent enhancement of behavioral correlation can prevent learning other joint behaviors. One natural solution is to enhance the correlation of agents in superior trajectories and reduce the correlation in inferior ones. Following this idea, in our PMIC framework, we maintain the superior and inferior trajectories separately. PMIC then maximizes the MI associated with the superior trajectories and minimizes the MI associated with the inferior trajectories. Thus, agents learn to form superior joint behavior and avoid inferior ones.

We verify our analysis by empirically evaluating the policies achieved by different methods in our motivating example. VM3 (Kim et al., 2020) and SIC (Chen et al., 2021) are representative methods that use MI maximizing approaches and MADDPG serves as a basic reference. Figure 1c shows

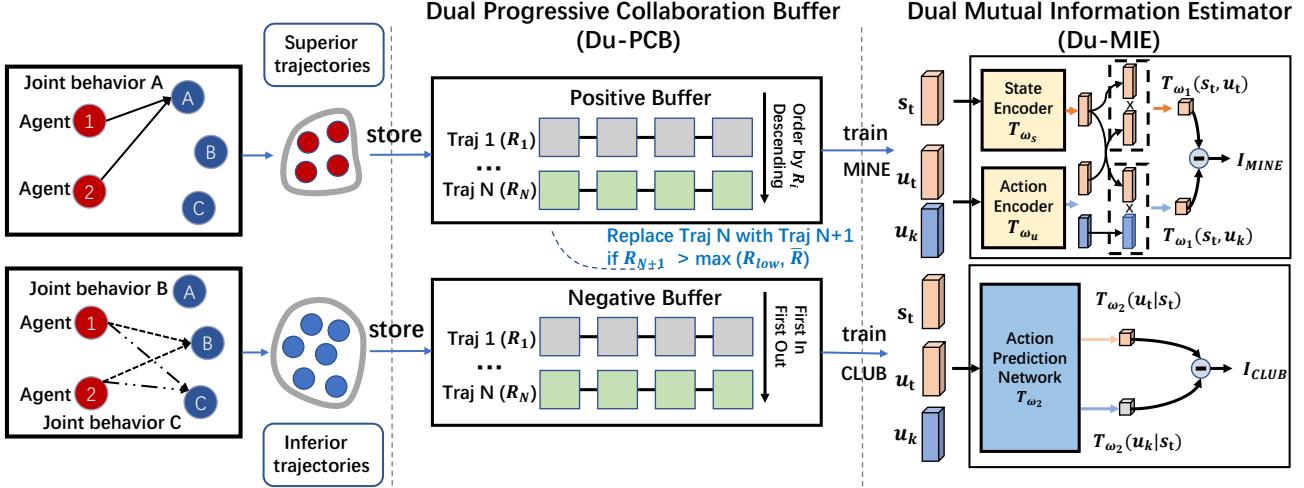


Figure 3: An overall illustration of Progressive Mutual Information Estimation, consisting of two main components: 1) Du-PCB maintains superior and inferior trajectories separately in a progressive manner and 2) Du-MIE estimates the collaboration criterion $I(s; u)$ from Equation 1 using I_{MINE} and I_{CLUB} .

the probability of converging to the optimal and sub-optimal joint behaviors. We can see that the methods maximizing the correlation of agents indiscriminately (i.e., SIC and VM3) have a greater probability of falling into the sub-optimal joint behaviors than PMIC (our algorithm). In turn, PMIC achieves higher rewards during the learning process (Figure 1e). Furthermore, Figure 1d shows the degrees of correlation over three joint behaviors measured by different methods with corresponding MI forms. The results show that sub-optimal joint behaviors (rescuing B and C) can also gain high degrees of correlation (even higher than rescuing A) in VM3 and SIC. In contrast, the degree of correlation for the optimal joint behavior in PMIC is significantly higher than in other methods. This further explains the results shown in Figures 1c and 1e. Overall, these empirical results in our motivating example demonstrate that only maximizing MI can make agents fall into sub-optimal collaboration behaviors and prevent them from learning the optimal ones, revealing the necessity of breaking the correlation over sub-optimal behaviors during the learning process. We detail our methodology and deeper experimental studies in the following sections.

4. Progressive Mutual Information Collaboration for MARL

This section introduces our framework, Progressive Mutual Information Collaboration (PMIC) to improve cooperative MARL based on our discovery in the previous section. The key idea of PMIC is to identify superior and inferior collaboration behaviors during the learning process, and encourage agents to achieve superior behaviors while avoiding sticking to inferior ones. This dual guidance of joint policy learning

is imposed in a progressive manner as the learning proceeds. We first propose a new MI-based collaboration criterion (Sec. 4.1) for measuring the collaboration degree of agents. Second, we introduce our approach to realize the dual collaboration guidance by progressive MI estimation (Sec. 4.2). Third, we show how to integrate PMIC into general MARL algorithms (Sec. 4.3).

4.1. A New Collaboration Criterion

Previous works measure the correlation of agents using MI in different forms. However, they often suffer from at least one of the following limitations. First, measuring correlation with the MI of any two agents' actions (Jaques et al., 2019; Kim et al., 2020) can be computationally infeasible with the increase in the number of agents (i.e., the scalability issue). Second, other methods (Mahajan et al., 2019; Chen et al., 2021) leverage the MI of 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.

To resolve these problems, we propose a new criterion to measure the degree of multiagent collaboration, which is defined as the mutual information between global state s and joint action u , which is formulated as follows:

$$\begin{aligned} I(s; u) &= H(u) - H(u | s) \\ &= H(u) - H(u_i | s) - H(u_{-i} | u_i, s), \end{aligned} \quad (1)$$

where $H(\cdot)$ and $H(\cdot | \cdot)$ denote the entropy and conditional entropy respectively, u_i is the action of any agent i , and u_{-i} is the joint action of all agents except i . Note that $I(s; u)$ can be decomposed into three distinct terms: (1)

$H(u)$ 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; (2) $H(u_i|s)$ measures the behavioral uncertainty of agent i , which encourages the agent to behave deterministically given global state s when minimized; (3) $H(u_{-i}|u_i, s)$ measures the uncertainty of agent i about the actions of other agents, which implicitly characterizes the correlation between agents' behavior and will drive agents to coherent joint behaviors when minimized. Overall, $I(s; u)$ can serve as a quantitative measure of collaboration, which can be optimized to incentivize agents to enhance or break different joint behaviors.

Compared with the aforementioned MI-based criteria, our collaboration criterion obeys the CTDE paradigm because it does not incorporate extra latent variables. Moreover, the MI measurement of our criterion is free of calculating the MI for all possible two agents, thus it does not suffer from the scalability issue as the number of agents increases. In the following subsection, we introduce how $I(s; u)$ is estimated and used as collaboration guidance.

4.2. Progressive Mutual Information Estimation

We now introduce dual collaboration guidance by progressive estimation of the collaboration criterion $I(s; u)$ to help agents learn to achieve better collaboration and avoid sub-optimal collaborations. This is achieved by two components, which are illustrated in Figure 3. The first component, Dual Progressive Collaboration Buffer (Du-PCB), stores superior and inferior trajectories in separate buffers, which correspond to joint behaviors to achieve and avoid, respectively. The second component, Dual Mutual Information Estimator (Du-MIE), provides MI estimates of $I(s; u)$ as quantitative signals of dual collaboration guidance. We detail the two components below.

Dual Progressive Collaboration Buffer (Du-PCB) consists of a positive buffer D^+ and a negative buffer D^- to store superior and inferior trajectories respectively. To identify the superior trajectories, we use the average return \bar{R} of the most recent M episodes as a measurement. We denote the trajectory with the lowest return in the positive buffer as R_{low} . For each episode k , we store a trajectory with return R_k if $R_k > \max(R_{\text{low}}, \bar{R})$ until D^+ is full; then Du-PCB overwrites the trajectories with return R_{low} . This ensures the quality of trajectories stored in D^+ monotonically increases during the learning process. In contrast, trajectories with returns $R_k \leq \max(R_{\text{low}}, \bar{R})$ are stored in D^- in a First-In-First-Out (FIFO) manner, since most recent inferior behaviors are more needed to be avoided. By this means, Du-PCB maintains the trajectories of both superior and inferior joint behaviors progressively according to the policy learned at present.

Dual Mutual Information Estimator (Du-MIE) will be used to estimate the collaboration criterion $I(s; u)$ described in Section 4.1, based on the trajectories stored in Du-PCB. MINE (Belghazi et al., 2018) will estimate the lower bound of $I(s; u)$ for maximization and CLUB (Cheng et al., 2020) will estimate the upper bound of $I(s; u)$ for minimization, based on the positive buffer D^+ and negative buffer D^- , respectively. To be concrete, MINE approximates the lower bound of $I(s; u)$ based on samples in D^+ as follows:

$$I(s; u) \geq I_{\text{MINE}}(s; u) = \sup_{\omega_1 \in \Omega} \underbrace{\mathbb{E}_{\mathbb{P}_{SU}} [-sp(-T_{\omega_1}(s_t, u_t))] - \mathbb{E}_{\mathbb{P}_S \otimes \mathbb{P}_U} [sp(T_{\omega_1}(s_t, u_k))]}_{-\mathcal{L}(\omega_1)}, \quad (2)$$

where \mathbb{P}_{SU} is state-action joint distribution, \mathbb{P}_S and \mathbb{P}_U are the marginals. The samples can be obtained by sampling s_t, u_t pairs jointly, u_k solely from D^+ . T_{ω_1} is a neural network with parameters $\omega_1 \in \Omega$ that outputs a scalar and soft-plus function $sp(z) = \log(1 + \exp(z))$. By contrast, CLUB approximates the upper bound of $I(s; u)$:

$$I(s; u) \leq I_{\text{CLUB}}(s; u) = \underbrace{\mathbb{E}_{\mathbb{P}_{SU}} [\log T_{\omega_2}(u_t | s_t)] - \mathbb{E}_{\mathbb{P}_S \otimes \mathbb{P}_U} [\log T_{\omega_2}(u_k | s_t)]}_{-\mathcal{L}(\omega_2)}, \quad (3)$$

where T_{ω_2} is a neural network with parameters ω_2 that approximates the conditional distribution. The joint and marginals are similarly sampled as in MINE, but are based on the negative buffer D^- .

The training losses of MINE and CLUB neural estimators are $\mathcal{L}(\omega_1)$ and $\mathcal{L}(\omega_2)$, as defined in Equations 2 and 3. Thus, the total loss of Du-MIE is:

$$\mathcal{L}_{\text{Du-MIE}}(\omega) = \mathcal{L}(\omega_1) + \mathcal{L}(\omega_2), \quad (4)$$

where $\omega = (\omega_1, \omega_2)$. The architecture of MINE and CLUB are illustrated in Figure 3. The complete process of training is detailed in Appendix B. After training, given state and joint-action samples, we can use MINE and CLUB to estimate the upper and lower bounds of MI.

Since MINE is trained following Equation (2) based on the samples from the positive buffer, only trajectories that resemble the joint behavior of superior collaboration have large MI estimates calculated by MINE; it is similar to CLUB. Therefore, given the interaction samples collected by the joint policy of current agents, MINE and CLUB can provide effective signals in guiding agents' behaviors towards or away from superior and inferior ones, progressively. In the following subsection, we show how PMIC functions with MARL for more efficient learning.

4.3. Integration of PMIC and MARL

With the MI estimations introduced in the previous subsection, now we aim to find the joint policy that maximizes the expected discounted return and follows the progressive collaboration guidance by optimizing the MI estimates from Du-MIE. In particular, 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 $J(\pi)$:

$$J^{\text{PMIC}}(\pi) = \mathbb{E}_{s, u \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t (r_t + r_t^{\text{PMIC}}) \right], \quad (5)$$

where $r_t^{\text{PMIC}} = \alpha I_{\text{MINE}}(s_t; u_t) - \beta I_{\text{CLUB}}(s_t; u_t)$, and α, β are the hyperparameters that weight the impact of MI guidance.

In principle, PMIC-MARL is a general framework that can be implemented with different MARL algorithms. We use PMIC-MADDPG for a representative demonstration, building upon MADDPG (Lowe et al., 2017). We also implement PMIC-RODE, built on RODE (Wang et al., 2020a), and study it in our experiments. The pseudo-code of PMIC-MADDPG is shown in Algorithm 1. In each episode, agents interact with the environment and store the trajectory samples into experience replay D (Lines 6-9). The trajectory is added to the positive buffer or the negative buffers in Du-PCB according to its return (Lines 10-13). Every k steps, Du-MIE is trained with the samples from Du-PCB (Line 14) to reflect current superior and inferior joint behaviors. Lastly, PMIC-MADDPG updates the centralized critic and the actors according to $J^{\text{PMIC}}(\pi)$ (Line 15), by minimizing the loss functions \mathcal{L}_Q and \mathcal{L}_π , defined below:

$$\begin{aligned} \mathcal{L}_Q(\phi) &= \mathbb{E}_{s_t, u_t, r_t, s_{t+1} \sim \mathcal{D}} \left[(\hat{y} - Q_\phi(s_t, u_t))^2 \right]; \\ \mathcal{L}_\pi(\theta) &= \mathbb{E}_{s_t \sim \mathcal{D}} [-Q_\phi(s_t, \pi_\theta(\cdot | s_t))], \end{aligned} \quad (6)$$

where $\hat{y} = r_t + r_t^{\text{PMIC}} + \gamma Q_{\phi'}(s_{t+1}, \pi_{\theta'}(s_{t+1}))$, $\theta = (\theta_1, \dots, \theta_n)$ and ϕ', θ' are the parameters of corresponding target networks.

Overall, we provide the technical details of PMIC, as well as how to combine PMIC with MARL algorithms, which we then evaluate empirically in the following section.

5. Experiments

This section empirically evaluates PMIC on multiple multiagent tasks to answer the following research questions (RQs):

RQ1 (Performance) Can PMIC effectively achieve collaboration and outperform related baselines? Is PMIC a generic framework?

Algorithm 1: PMIC-MADDPG

```

1 Input: the update frequency  $k$  for Du-MIE, maximum
   episode length  $T$ , hyperparameters  $\alpha$  and  $\beta$  to balance the
   effects of maximizing and minimizing MI.
2 Initialize the critic network  $\phi$ ,  $n$  actor networks  $\theta_1 \dots \theta_n$  and
   corresponding target networks  $\phi', \theta_1', \dots, \theta_n'$ .
3 Initialize Du-MIE parameterized by  $\omega_1$  and  $\omega_2$ .
4 Initialize Du-PCB and experience replay buffer  $\mathcal{D}$ 
5 repeat
6   for  $t = 1, \dots, T$  do
7     Execute joint actions  $u_t$  via collecting  $u_t^i \sim \pi_{\theta_i}(o_t^i)$ .
8     Receive  $o_{t+1} = \{o_{t+1}^i\}_{i=1}^n$  and team reward  $r_t$ .
9     Store trajectory  $\nu = \{o_t, u_t, o_{t+1}, r_t\}_{t=1}^T$  to  $\mathcal{D}$ 
10    if  $R_\nu > \max(R_{\text{low}}, \bar{R})$  then
11      Add  $\nu$  to the positive buffer
12    else
13      Add  $\nu$  to the negative buffer
14    Update Du-MIE with Du-PCB every  $k$  steps ▷ see Eq. 4
15    Update the actors and critic networks ▷ see Eq. 6
16 until reaching maximum training steps;

```

RQ2 (Superiority of Dual MI) Is maximizing and minimizing $I(s; u)$ necessary? Is $I(s; u)$ effective for performance improvements?

RQ3 (Necessity of Du-PCB) Does Du-PCB significantly improve performance over a normal replay buffer?

5.1. Benchmarks & Baselines

For a comprehensive comparative study, we evaluate our algorithms on both discrete and continuous action spaces. For the continuous action space, we consider the Multi-Agent Particle Environment (MPE) and the Multi-Agent MuJoCo benchmark, comparing PMIC-MADDPG with six advanced algorithms as baselines: SIC-MADDPG (Chen et al., 2021), VM3-AC (Kim et al., 2020), MASAC (Kim et al., 2020), MADDPG (Lowe et al., 2017), FacMADDPG (de Witt et al., 2020), and COMIX (de Witt et al., 2020), where FacMADDPG and COMIX are the state-of-the-art (SOTA) algorithms in Multi-Agent MuJoCo benchmark (de Witt et al., 2020). The StarCraft II micromanagement (SMAC) benchmark (Samvelyan et al., 2019) is also considered, which has high complexity of control and requires learning policies in large discrete action space, and we compare our PMIC-RODE algorithm with the current SOTA algorithm, RODE (Wang et al., 2020a). More experiments to integrate PMIC with MASAC and QMIX (Rashid et al., 2018) are provided in Appendix E. The environment description is provided in Appendix A.

For all baseline algorithms, we use the official code if available or reproduce it according to the original papers. Hyperparameters have been fine-tuned in all environments. For a fair comparison, we use the same structures to avoid the influence of different structures on the results. Further im-

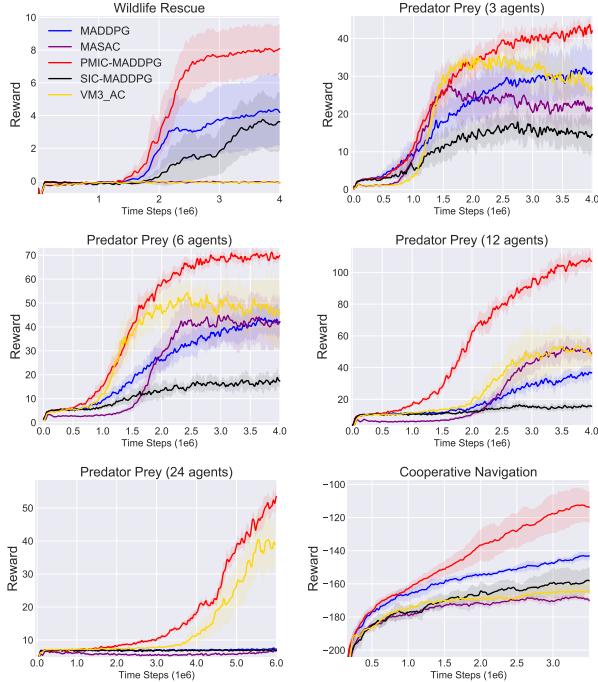


Figure 4: Comparisons of averaged return on MPE.

plementation details are in Appendix G.

5.2. Performance (RQ1)

We first evaluate the performance on 6 environments of MPE with 10 different random seeds: Wildlife Rescue, Cooperative Navigation, and Partial Observation Cooperative Predator Prey with 3, 6, 12, and 24 predators (where the agents control predators and the policy of prey is fixed). The results in Figure 4 show that PMIC-MADDPG outperforms other methods across all tasks. Both VM3-AC and SIC-MADDPG receive lower rewards than PMIC-MADDPG, indicating that simply maximizing MI does not guarantee high-performing collaboration behaviors. PMIC can also deliver significant performance gains as the number of agents increases, while some of the other algorithms cannot learn to collaborate effectively. For example, on Predator Prey with 24 agents, only PMIC-MADDPG and VM3-AC can learn collaborative behaviors (but VM3-AC requires three times more time than PMIC-MADDPG). For Wildlife Rescue, results show that the MASAC-related algorithms (e.g., MASAC and VM3-AC) can not make agents capture any wildlife, which is due to underestimation of using double Q-learning. To remove the influence of irrelevant factors, in the motivating example, we use single Q-learning instead. Details of the experiments are in Appendix H.

We further evaluate PMIC on 2 tasks of Multi-Agent MuJoCo benchmark with 10 different random seeds. In these tasks, agents need to cooperate in robot control and different

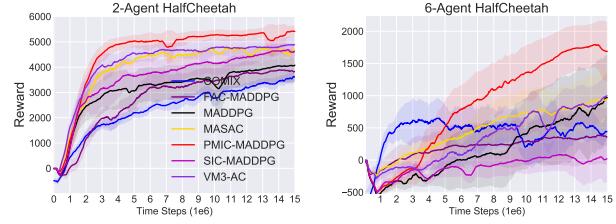


Figure 5: Comparisons of averaged return on MA-MuJoCo.

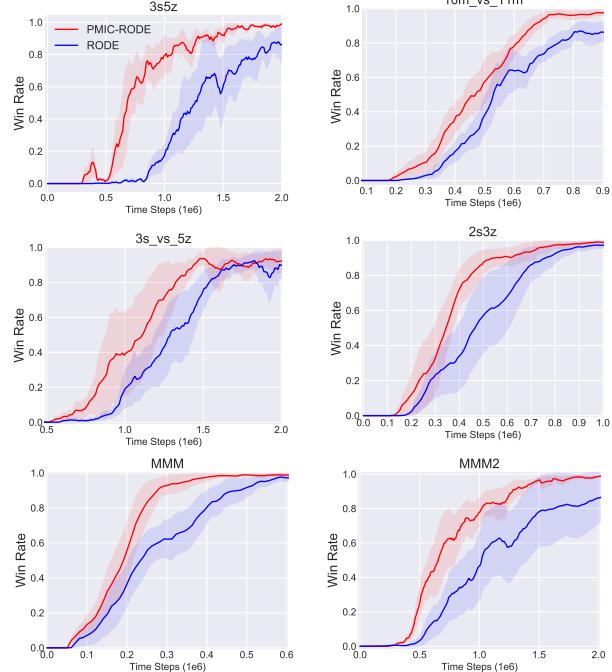


Figure 6: Comparisons of averaged test win rate on SMAC.

agents control different joints of the robot. In our experimental setting, agents do not share information, which is the most difficult setting in Multi-Agent MuJoCo. The results in Figure 5 show that PMIC-MADDPG outperforms other baselines (COMIX, Fac-MADDPG, SIC-MADDPG, VM3AC, and MASAC), which demonstrates the benefits of PMIC in challenging continuous control tasks. We perform significance analysis (t-test) and prove that PMIC-MADDPG achieves statistically significant advantages over other baseline algorithms.

For SMAC, we evaluate PMIC on 6 maps with 5 different random seeds. Note that RODE is the SOTA algorithm on SMAC. The results in Figure 6 show that PMIC can provide an improvement over RODE, reaching convergence faster and achieving higher performance.

In summary, these experiments show that PMIC is an effective framework that can be integrated with multiple algorithms and can provide significant improvements in both continuous and discrete action space tasks. More experi-

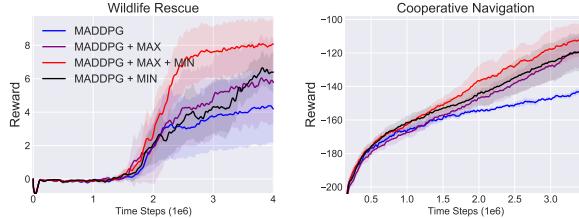


Figure 7: Ablation on MI maximization & minimization.

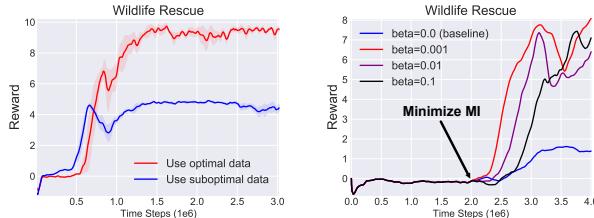


Figure 8: *Left*: maximizing $I(s; u)$ associated with optimal and suboptimal data to guide agents. *Right*: minimizing $I(s; u)$ to break inferior behaviors .

ments, such as PMIC-MASAC and PMIC-QMIX, are included in Appendix E.

5.3. Superiority of Dual MI (RQ2)

To answer RQ2, we first provide an ablation study on maximizing and minimizing $I(s; u)$ to investigate the effect of the two mechanisms. The results in Figure 7 show that maximizing and minimizing $I(s; u)$ can achieve faster convergence and higher performance, relative to only maximizing or only minimizing $I(s; u)$.

To verify whether the MINE estimator that maximizes $I(s; u)$ can guide agents correctly, we collect optimal (i.e., two agents both catch A) and sub-optimal (i.e., two agents both catch C) trajectories to train MINE respectively, then use the trained MINE to guide an initialized MADDPG from the beginning. As shown in the left of Figure 8, training MINE with optimal trajectories can quickly guide the algorithm to learn the optimal joint behavior, which indicates that maximizing $I(s; u)$ on superior trajectories can guide agents correctly and quickly. However, if MINE operates on sub-optimal trajectories, the policy converges to the sub-optima. This indicates that only maximizing MI without distinguishing the quality of trajectories can make agents fall into inferior ones. This further verifies our assumption in the motivation section and also indicates the necessity of designing Du-PCB.

Next, we examine whether the CLUB estimator that minimizes $I(s; u)$ can escape sub-optima and help agents achieve high-quality collaboration. We test on Wildlife Rescue and find a policy in a sub-optima (e.g., rescuing

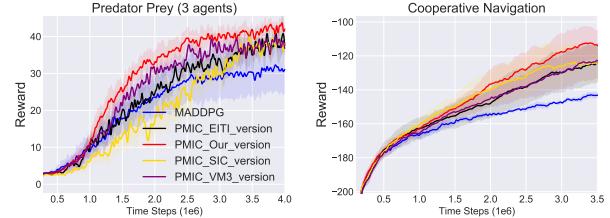


Figure 9: Comparisons of different MI under PMIC. EITI version measures the MI of agents' current actions/states and other agents' next states. SIC version measures the MI of the shared latent variables and agents' joint policy VM3 version measures the MI of any two agents' actions.

animals with lower rewards). After 2 million time steps, we begin minimizing $I(s; u)$ for the policy with different β . Figure 8 (right) shows that the training curve gradually increases and achieves higher rewards than baseline. This indicates that minimizing $I(s; u)$ can help the algorithm escape from sub-optima and discover better ones.

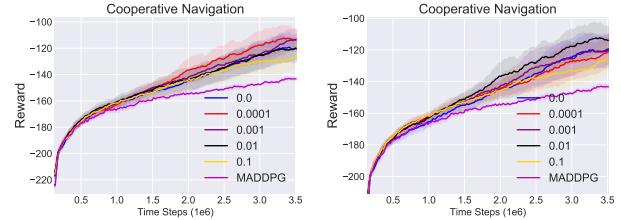


Figure 10: Parameter analysis on α (left) and β (right).

We further provide the parameter analysis on α and β . The results are shown in Figure 10. The experiment proves that appropriately sized values of α and β are critical for algorithm improvement. The performance loss occurs if β or α is too large or too small. Thus, for different environments, we need to adjust α and β to achieve the best performances. Intuitively, if the environment has multiple kinds of collaboration (e.g., Wildlife Rescue), one may scale up β to prevent falling into the inferior ones; if the environment has a smooth underlying path from inferior to superior collaboration, one may scale up α to enhance guidance and accelerate learning.

To finish answering RQ2, we compare different MI measurements under PMIC. Results in Figure 9 show that our proposed measure is more effective than other alternatives. Besides, we test the sensitivity of different parameters such as the size of Du-PCB, α , and β . Due to the space limitation, these ablation experiments are put in Appendix E.

5.4. Necessity of Du-PCB (RQ3)

To address RQ3, we consider whether the positive and negative buffers in Du-PCB can be replaced with a normal replay

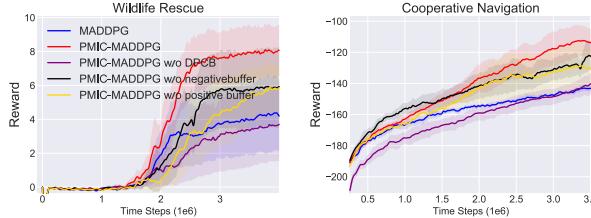


Figure 11: Ablation experiments on Du-PCB. "PMIC-MADDPG w/o negative buffer" means we use a regular buffer to train CLUB, i.e., without data filtering as introduced in Sec. 4.2

buffer. Results in Figure 11 show that Du-PCB is indeed more effective. If we put all these trajectories into one normal buffer, MINE and CLUB cannot distinguish what behaviors are more favorable, in a result MINE and CLUB cannot provide useful guidance for agents.

6. Conclusion & Future work

To address the potentially detrimental effect of only maximizing mutual information, we propose the PMIC framework, consisting of a newly proposed collaboration criterion measured by the MI between global states and joint actions; Du-PCB to progressively maintain superior and inferior trajectories; and Du-MIE to maintain two MI estimates of our new criterion from samples in Du-PCB. In addition to maximizing the expected discounted return, PMIC-MARL maximizes the MI estimates associated with superior collaboration to guide agents to facilitate better collaboration, while minimizing the MI estimates associated with inferior collaborations, effectively encouraging exploration while avoiding sub-optimal collaborations. 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.

For limitations, firstly our work is empirical proof of the effectiveness of the PMIC idea and we provide no theory on optimality, convergence, and complexity. Secondly, the choice of hyperparameters α and β has a significant effect on the performances. Thirdly, the current approach to distinguishing data is relatively simple and straightforward. In the future, we would like to improve PMIC to solve the above limitations by theoretical support, automatic hyperparameter tuning technology, and more accurate methods to distinguish data. In addition, we would like to further develop PMIC by applying the idea to other fields, such as communication in MARL (Sun et al., 2020) and hierarchical reinforcement learning (HRL) (Tang et al., 2018). For example, in HRL, the maximization-minimization of MI could help lower-level agents achieve effective execution based on a target provided by an upper-level policy. On the

other hand, extending PMIC with advanced techniques, like hybrid action space (Li et al., 2021), evolutionary RL (Shen et al., 2020), or prior human knowledge (Zhang et al., 2020), is worth further study.

Acknowledgements

The work is supported by the National Natural Science Foundation of China (Grant Nos.: 62106172, U1836214) and the new Generation of Artificial Intelligence Science and Technology Major Project of Tianjin under grant: 19ZXZNGX00010 and the Science and Technology on Information Systems Engineering Laboratory (Grant No. WDZC20205250407).

References

- Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, D. Mutual information neural estimation. In *International Conference on Machine Learning*, pp. 531–540, 2018.
- Chen, L., Guo, H., Du, Y., Fang, F., Zhang, H., Zhang, W., and Yu, Y. Signal instructed coordination in cooperative multi-agent reinforcement learning. In *Distributed Artificial Intelligence*, volume 13170, pp. 185–205, 2021.
- Cheng, P., Hao, W., Dai, S., Liu, J., Gan, Z., and Carin, L. Club: A contrastive log-ratio upper bound of mutual information. In *International Conference on Machine Learning*, pp. 1779–1788. PMLR, 2020.
- de Witt, C. S., Peng, B., Kamienny, P.-A., Torr, P., Böhmer, W., and Whiteson, S. Deep multi-agent reinforcement learning for decentralized continuous cooperative control. *arXiv preprint arXiv:2003.06709*, 2020.
- Guo, Y., Oh, J., Singh, S., and Lee, H. Generative adversarial self-imitation learning. *arXiv preprint arXiv:1812.00950*, 2018.
- Hernandez-Leal, P., Kartal, B., and Taylor, M. E. A survey and critique of multiagent deep reinforcement learning. *Journal of Autonomous Agents and Multiagent Systems*, 33:750–797, October 2019. doi: <https://doi.org/10.1007/s10458-019-09421-1>.
- Hu, J., Jiang, S., Harding, S. A., Wu, H., and Liao, S.-w. Rethinking the implementation tricks and monotonicity constraint in cooperative multi-agent reinforcement learning. *arXiv e-prints*, pp. arXiv-2102, 2021.
- Iqbal, S. and Sha, F. Actor-attention-critic for multi-agent reinforcement learning. In *International Conference on Machine Learning*, pp. 2961–2970. PMLR, 2019.

- Jaques, N., Lazaridou, A., Hughes, E., Gülc̄ehre, Ç., Ortega, P. A., Strouse, D., Leibo, J. Z., and de Freitas, N. Intrinsic social motivation via causal influence in multi-agent RL. *arXiv preprint arXiv:1810.08647*, 2018.
- Jaques, N., Lazaridou, A., Hughes, E., Gulcehre, C., Ortega, P., Strouse, D., Leibo, J. Z., and De Freitas, N. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International Conference on Machine Learning*, pp. 3040–3049. PMLR, 2019.
- Kim, W., Jung, W., Cho, M., and Sung, Y. A maximum mutual information framework for multi-agent reinforcement learning. *arXiv preprint arXiv:2006.02732*, 2020.
- Lanctot, M., Zambaldi, V., Gruslys, A., Lazaridou, A., Tuyls, K., Pérolat, J., Silver, D., and Graepel, T. A unified game-theoretic approach to multiagent reinforcement learning. *arXiv preprint arXiv:1711.00832*, 2017.
- Li, B., Tang, H., Zheng, Y., Hao, J., Li, P., Wang, Z., Meng, Z., and Wang, L. Hyar: Addressing discrete-continuous action reinforcement learning via hybrid action representation. *CoRR*, abs/2109.05490, 2021.
- Li, M., Qin, Z., Jiao, Y., Yang, Y., Wang, J., Wang, C., Wu, G., and Ye, J. Efficient ridesharing order dispatching with mean field multi-agent reinforcement learning. In *The World Wide Web Conference*, pp. 983–994, 2019.
- Liu, M., Zhou, M., Zhang, W., Zhuang, Y., Wang, J., Liu, W., and Yu, Y. Multi-agent interactions modeling with correlated policies. *arXiv preprint arXiv:2001.03415*, 2020.
- Lowe, R., Wu, Y. I., Tamar, A., Harb, J., Abbeel, O. P., and Mordatch, I. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in neural information processing systems*, pp. 6379–6390, 2017.
- Lyu, X., Baisero, A., Xiao, Y., and Amato, C. A deeper understanding of state-based critics in multi-agent reinforcement learning. *arXiv preprint arXiv:2201.01221*, 2022.
- Mahajan, A., Rashid, T., Samvelyan, M., and Whiteson, S. Maven: Multi-agent variational exploration. In *Advances in Neural Information Processing Systems*, pp. 7613–7624, 2019.
- Matignon, L., Jeanpierre, L., and Mouaddib, A. Coordinated multi-robot exploration under communication constraints using decentralized markov decision processes. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.
- Merhej, R. and Chetouani, M. Lief: Learning to influence through evaluative feedback. In *Adaptive and Learning Agents Workshop (AAMAS 2021)*, 2021.
- Oliehoek, F. A. and Amato, C. *A concise introduction to decentralized POMDPs*. Springer, 2016.
- Peng, P., Wen, Y., Yang, Y., Yuan, Q., Tang, Z., Long, H., and Wang, J. Multiagent bidirectionally-coordinated nets: Emergence of human-level coordination in learning to play starcraft combat games. *arXiv preprint arXiv:1703.10069*, 2017.
- Rashid, T., Samvelyan, M., De Witt, C. S., Farquhar, G., Foerster, J., and Whiteson, S. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. *arXiv preprint arXiv:1803.11485*, 2018.
- Samvelyan, M., Rashid, T., de Witt, C. S., Farquhar, G., Nardelli, N., Rudner, T. G. J., Hung, C., Torr, P. H. S., Foerster, J. N., and Whiteson, S. The starcraft multi-agent challenge. In *International Conference on Autonomous Agents and MultiAgent Systems*, pp. 2186–2188, 2019.
- Shen, R., Zheng, Y., Hao, J., Meng, Z., Chen, Y., Fan, C., and Liu, Y. Generating behavior-diverse game ais with evolutionary multi-objective deep reinforcement learning. In Bessiere, C. (ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pp. 3371–3377. ijcai.org, 2020.
- Sun, J., Zheng, Y., Hao, J., Meng, Z., Chen, Y., Fan, C., and Liu, Y. Continuous multiagent control using collective behavior entropy for large-scale home energy management. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 922–929. AAAI Press, 2020.
- Sunehag, P., Lever, G., Gruslys, A., Czarnecki, W. M., Zambaldi, V., Jaderberg, M., Lanctot, M., Sonnerat, N., Leibo, J. Z., Tuyls, K., et al. Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296*, 2017.
- Tang, H., Hao, J., Lv, T., Chen, Y., Zhang, Z., Jia, H., Ren, C., Zheng, Y., Meng, Z., Fan, C., et al. Hierarchical deep multiagent reinforcement learning with temporal abstraction. *arXiv preprint arXiv:1809.09332*, 2018.
- Wang, T., Gupta, T., Mahajan, A., Peng, B., Whiteson, S., and Zhang, C. Rode: Learning roles to decompose multi-agent tasks. *arXiv preprint arXiv:2010.01523*, 2020a.
- Wang, T., Wang, J., Wu, Y., and Zhang, C. Influence-based multi-agent exploration. In *International Conference on Learning Representations*, 2020b.

- Wang, W., Yang, T., Liu, Y., Hao, J., Hao, X., Hu, Y., Chen, Y., Fan, C., and Gao, Y. Action semantics network: Considering the effects of actions in multiagent systems. In *Proceedings of the 8th International Conference on Learning Representations*, 2020c.
- Wen, Y., Yang, Y., Luo, R., Wang, J., and Pan, W. Probabilistic recursive reasoning for multi-agent reinforcement learning. *arXiv preprint arXiv:1901.09207*, 2019.
- Xie, A., Losey, D. P., Tolsma, R., Finn, C., and Sadigh, D. Learning latent representations to influence multi-agent interaction. *arXiv preprint arXiv:2011.06619*, 2020.
- Yang, T., Tang, H., Bai, C., Liu, J., Hao, J., Meng, Z., and Liu, P. Exploration in deep reinforcement learning: A comprehensive survey. *CoRR*, abs/2109.06668, 2021a.
- Yang, T., Wang, W., Tang, H., Hao, J., Meng, Z., Mao, H., Li, D., Liu, W., Chen, Y., Hu, Y., Fan, C., and Zhang, C. An efficient transfer learning framework for multiagent reinforcement learning. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems 34*, pp. 17037–17048, 2021b.
- Zhang, P., Hao, J., Wang, W., Tang, H., Ma, Y., Duan, Y., and Zheng, Y. Kogun: Accelerating deep reinforcement learning via integrating human suboptimal knowledge. In Bessiere, C. (ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pp. 2291–2297, 2020.
- Zhang, S. Q., Zhang, Q., and Lin, J. Efficient communication in multi-agent reinforcement learning via variance based control. *arXiv preprint arXiv:1909.02682*, 2019.
- Zheng, Y., Meng, Z., Hao, J., and Zhang, Z. Weighted double deep multiagent reinforcement learning in stochastic cooperative environments. In Geng, X. and Kang, B. (eds.), *PRICAI 2018: Trends in Artificial Intelligence - 15th Pacific Rim International Conference on Artificial Intelligence, Nanjing, China, August 28-31, 2018, Proceedings, Part II*, volume 11013 of *Lecture Notes in Computer Science*, pp. 421–429. Springer, 2018a.
- Zheng, Y., Meng, Z., Hao, J., Zhang, Z., Yang, T., and Fan, C. A deep bayesian policy reuse approach against non-stationary agents. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 962–972, 2018b.
- Zheng, Y., Hao, J., Zhang, Z., Meng, Z., and Hao, X. Efficient multiagent policy optimization based on weighted estimators in stochastic cooperative environments. *J. Comput. Sci. Technol.*, 35(2):268–280, 2020.

A. Environment Details

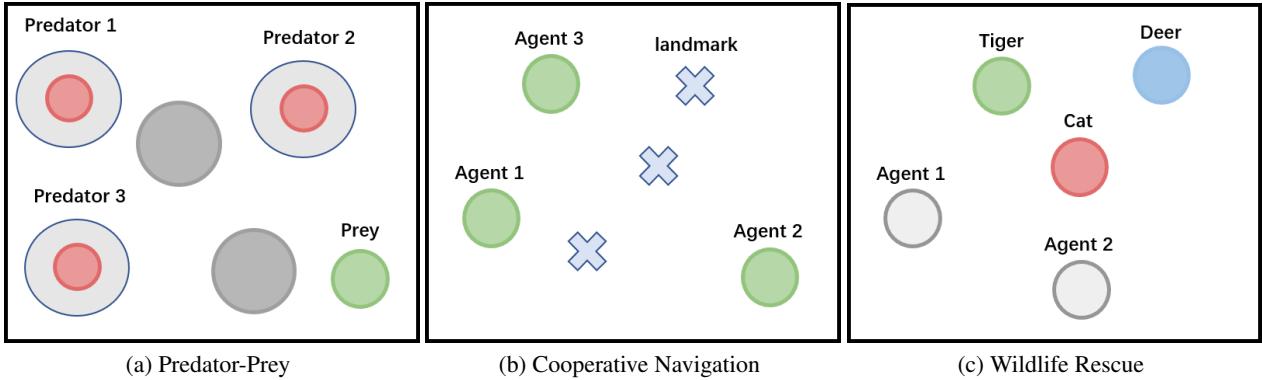


Figure 12: Multi-Agent Particle Environment.

Predator Prey: N slower cooperating agents chase the faster adversary around a randomly generated environment with L large landmarks impeding the way. In our setting, agents control the predators to chase the prey, the policy of prey is fixed. The agents are partially observable, the observation radius of each predator is 0.25. Only when the predator captures the prey, it can get the reward 10. we set N to 3, 6, 12, 24 separately. Each game has 25 steps.

Cooperative Navigation: Agents must cooperate through physical actions to reach a set of L landmarks. In our setting, agents receive a shared reward which is the sum of the minimum distance of the landmarks from any agent, and the agents who collide with each other receive negative reward -1. Besides, all agents receive 1 if all landmarks are occupied. Each game has 25 steps.

Wildlife Rescue (the motivating example): N agents must cooperate to rescue M wildlife with different risks and rewards. We provide a control time T_c . When an agent catches up with an animal, the agent can control T_c seconds to wait for other agents to arrive. Each game has T steps. In the motivating example and experiment, we leverage the same setting. We set T_c to 8, T to 60. The specific reward at the end of each episode is set in Table 1.

Table 1: The reward matrix of the rescue agents at the end of each episode. Both agents receive the same reward.

reward \ agent 2	tiger	deer	cat	on the road
agent 1				
tiger	11	-30	0	-30
deer	-30	7	6	-30
cat	0	6	5	0
on the road	-30	-10	0	0

Multi-Agent MuJoCo (de Witt et al., 2020) is a range of challenging continuous multi-agent control tasks of Multi-Agent MuJoCo benchmark suite. Multi-Agent MuJoCo is designed for decentralized cooperative continuous multi-agent robotic control. The structure is shown in Figure 13. For evaluation, we use the reward setting of the original paper, but we set k to zero for all environments. k controls how much information each agent can observe from its adjacent agents. When k is zero, it means each agent can only observe information about its own joints, which is the hardest setting for Multi-Agent MuJoCo.

For SMAC, we use the latest version 2.4.10, and default settings for all maps.

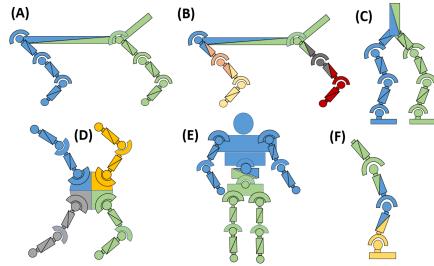


Figure 13: The structure of Multi-Agent MuJoCo.

B. Process of training MINE and CLUB

In this section, we detail the process of training MINE and CLUB.

First, we introduce the process of training MINE. We train MINE based on the positive buffer of Du-PCB. Figure 3 illustrates the detailed process of estimating the lower bound. We design the network T with parameters ω_1 as a state encoder E_{ω_s} and an action encoder E_{ω_u} . MINE takes a batch of $(s_t, u_t) \sim \mathbb{P}_{SU}$ and $(s_t, u_k) \sim \mathbb{P}_S \otimes \mathbb{P}_U$ as inputs and encodes the data as vectors $E_{\omega_s}(s_t)$, $E_{\omega_u}(u_t)$, $E_{\omega_u}(u_k)$ with the same dimension. Then we take the inner product of $E_{\omega_s}(s_t)$ and $E_{\omega_u}(u_t)$, $E_{\omega_s}(s_t)$ and $E_{\omega_u}(u_k)$ to get $T_{\omega_1}(s_t, u_t)$ and $T_{\omega_1}(s_t, u_k)$ (e.g., scores) separately. Finally we do the subtraction operation on the scores to get the estimate of $I_{\text{MINE}}(s, u)$ according to Equation (2). In summary, MINE is trained by sampling data from the positive buffer of Du-PCB to minimize $\mathcal{L}(\omega_1)$ in Equation (2).

Then we details the process of training CLUB. We train CLUB based on the negative buffer of Du-PCB. Figure 3 illustrates the detailed process of estimating the upper bound. We design the network T with parameters ω_2 as action prediction network. The train process of CLUB is different from MINE, we only need to sample data from the joint distribution \mathbb{P}_{SU} . Then we directly minimize the loss $\mathcal{L}(\omega_2)$ based on the samples to optimize CLUB.

C. Integrate PMIC with RODE

RODE is a role-based algorithm which decomposes the joint action space into a restricted role action space to reduce the primitive action-observation spaces. In RODE, each task can be decomposed into a sub-task which has a smaller action-observation space, and each sub-task is associated with a role ρ . To demonstrate the generalization and effectiveness of PMIC, we design PMIC-RODE (i.e., integrate PMIC with RODE). To integrate with PMIC, we take the joint role ρ into consideration. Thus we measure the MI between the joint actions and global states with the joint role ρ (e.g., $I(u; \rho, s)$). Since the selection of actions for the agents needs to consider the selected role and the observations. $I(u; \rho, s)$ has a similar effort with $I(u; s)$ in PMIC-MADDPG where the decisions are only based on the observations. Different from PMIC-MADDPG, we set $r_t^{\text{PMIC}} = \alpha I_{\text{MINE}}(s_t; \rho_t, u_t) - \beta I_{\text{CLUB}}(s_t; \rho_t, u_t)$. r_t^{PMIC} is used in the same way as in PMIC-MADDPG. To achieve the goal in Equation 5, we only need to modify the Q_{tot} as follows:

$$\mathcal{L} = \mathbb{E}_{s_t, u_t, r_t, s_{t+1} \sim \mathcal{D}} \left[(y_{rode} - Q_{tot}(s_t, u_t))^2 \right], \quad (7)$$

where $y_{rode} = r_t + r_t^{\text{PMIC}} + \gamma \max_{a_{t+1}} \bar{Q}_{tot}(s_{t+1}, \pi_{\theta'}(s_{t+1}))$ and Q_{tot} is a QMIX-style (Rashid et al., 2018) mixing network to estimate the total value of the state and actions and \bar{Q}_{tot} is the target network of Q_{tot} . Du-MIE is updated according to Equation 4. Since RODE is QMIX-based method, credit assignment is also applied to the MI signals. The MI signals over superior trajectories and inferior trajectories can reward and punish the agents based on their contribution, which can help agents to find the optimal joint behaviors.

PMIC can be easily combined with existing MARL algorithms. For example, to integrate PMIC with RODE, we only need to initialize the networks of RODE in line 2 of Algorithm 1 and retain other components. All other processes remain the same except changing $I(u; s)$ to $I(u; \rho, s)$. Finally, we update the value network according to Equation 7 and others are the same with RODE to replace the line 15 in Algorithm 1. The same applies in combination with other algorithms.

D. Experimental Settings

For experiments on MPE and SMAC, we use episodic rewards as the criterion for adding data to Du-PCB.

On Multi-Agent MuJoCo, we can more accurately collect superior trajectories. Since the episode reward of one trajectory is high, there is no guarantee that there are no poor sub-trajectories that affect Du-PCB. Thus we leverage a more fine-grained filtering method to filter the trajectories on Multi-Agent MuJoCo. We leverage sub-trajectories to replace the complete trajectory to accurately filter the superior and inferior trajectories. In particular, we set the length of sub-trajectories to 50 steps and use the total reward of each sub-trajectory as the criterion for adding data to Du-PCB. The update frequency of MINE and CLUB is 1 on all tasks.

The experiments on MPE are carried out on Intel(R) Xeon(R) Gold 5117 CPU @ 2.00GHz. The experiments on Multi-Agent MuJoCo are carried out on NVIDIA GTX 2080 Ti GPU with Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz. The experiments on SMAC are carried out on Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz and Intel(R) Xeon(R) CPU E5-2679 v4 @ 2.50GHz.

E. Additional Experiments

In this section, we present more experiments to help better understand our framework. Six questions are raised and more experiments about PMIC-MASAC and PMIC-QMIX are provided:

RQ1: Is Du-PCB buffer size sensitive?

RQ2: Is only maximizing $I(s; \pi(\cdot|s))$ better than other MI forms?

RQ3: What is the time consumption of PMIC-MADDPG?

RQ4: Can MINE be replaced by normal MI estimator?

RQ5: Can MINE trained with the positive buffer of Du-PCB provide a good guide?

RQ6: How to choose the k (Update frequency of MI estimators)?

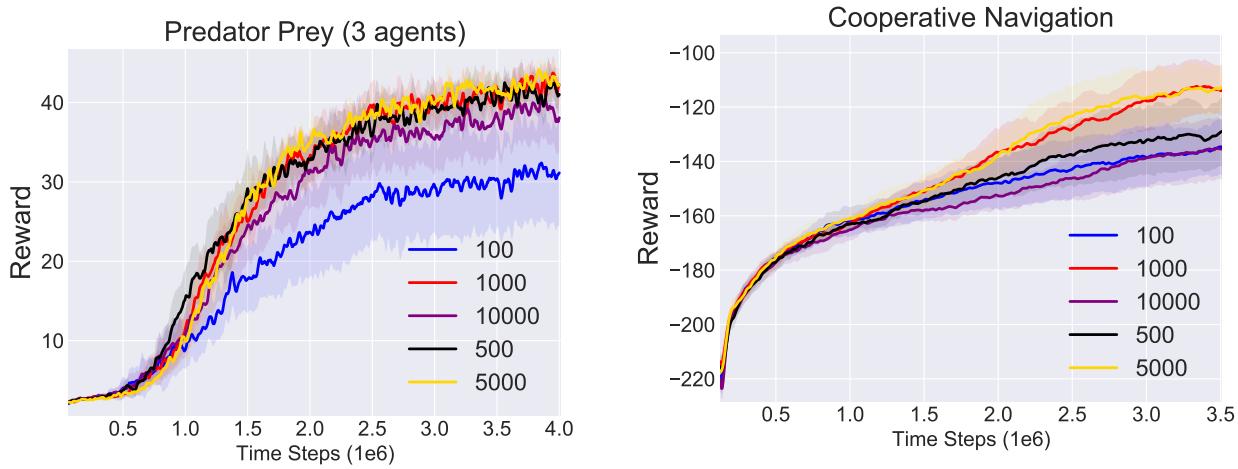


Figure 14: Influence of Du-PCB size.

To answer RQ1, we design experiments on MPE and adjust the size of Du-PCB. Specifically, we test 5 different buffer sizes (100, 500, 1000, 5000, 10000) for positive buffer and negative buffer of Du-PCB. The results shown in Figure 14 indicate that both too large size and too small size are harmful to the performance. A buffer size that is too small prevents MINE and CLUB from characterizing trajectories well, while too large size hinders the optimality of Du-PCB which makes it impossible for MINE and CLUB to adapt quickly to the new joint behavior. Therefore, an appropriate size is important.

To answer RQ2, we only maximize MI with different forms to give a comparison. To avoid differences caused by some mechanisms (e.g., double Q) or structures (e.g., MINE), we apply different forms of MI to MADDPG and leverage MINE as the MI estimator and other settings remain the same. The hyperparameters have been fine-tuned. The results are shown in Figure 15. Maximize MI of the global states and the joint actions is better than other MI forms in terms of the final performance. This demonstrates that our proposed method of maximizing the MI of global states and joint actions is more effective than other MI forms. Besides, we also find that maximizing MI in environments with sparse reward brings performance degradation to the original algorithm. The reason is: In environments with sparse reward, the MI signals play a greater role in guiding than reward signals. Beside, maximizing MI has the problem of easily making agents fall into sub-optimal collaborations, thus agents quickly fall into sub-optimal collaborations.

To answer RQ3, we evaluate the time consumption of different algorithms on 6-Agent HalfCheetah. The experiment is carried out on NVIDIA GTX 2080 Ti GPU with Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz. We evaluate the time consumption of each algorithm individually with no additional programs running on the device. Each result is the average of 10 time-consuming calculations. The time consumption includes the time consumption of the execution phase and the time consumption of the centralized training phase. The results are shown in Table 2. PMIC-MADDPG brings less time consuming than other methods.

To answer RQ4, we evaluate the impact of MINE on the performance of the algorithm. Because we use MINE for the

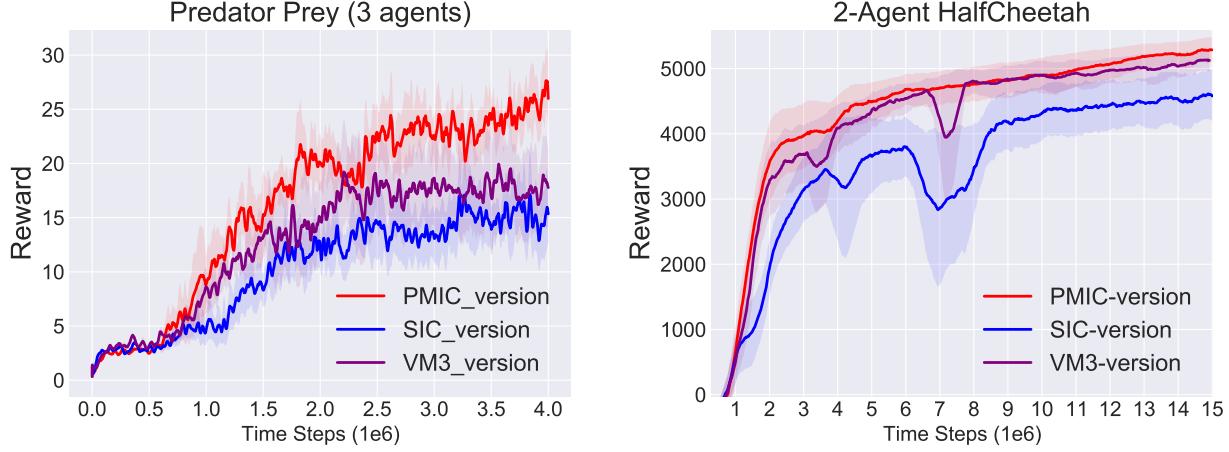


Figure 15: Comparison of only maximizing mutual information with different forms. PMIC-version leverages MI of global states and joint actions. VM3-version leverages the MI of any two agents’ actions among agents. SIC-version leverages the MI of z and the joint policy.

Table 2: Time consumption of different algorithms on 6-Agent HalfCheetah every 1000 time steps.

algorithm	PMIC-MADDPG	MADDPG	SIC-MADDPG	VM3-AC
seconds	30.96	24.43	31.01	182.58
algorithm	COMIX	Fac-MADDPG	MASAC	
seconds	34.43	87.13	41.68	

first time in the field of multi-agent reinforcement learning, thus we need to provide a comparison of MINE and the normal mutual information estimation method. The results are shown in Figure 16. We can see that the performance of both methods is similar, which demonstrates that the effectiveness of PMIC is introduced by the mechanism of the maximization-minimization MI and new collaboration criterion, rather than MINE.

To answer RQ5, we further analyze whether MINE trained with the positive buffer in Du-PCB could act as a good guide, which requires the MINE to estimate a large value over superior trajectories and a small value over inferior trajectories. We train MADDPG in 2-Agent Walker. During training, we save the positive buffer every 100000 steps, which ensures that the trajectories saved each time are better than the previous time. Then we save the MINE every 100000 steps and use it to estimate MI of trajectories saved in the positive buffer. The estimated MI of different positive buffer by different MINE are plotted on Figure 17. The larger the index, the newer the MINE and buffer. When we use Du-PCB to train MINE, the color gradually darkens from left to right in each line, indicating that MINE has a higher estimation value for the better trajectories. However, when training MINE in the same way with VM3-AC (Kim et al., 2020), MINE estimates the same value of MI for both inferior and superior trajectories. From the perspective of visualization, the MI of superior coordination estimated by MINE with Du-PCB is large so that such coordination can be more encouraged, which can give accurate guidance to superior joint behavior.

We train Du-MIE at the same frequency as the critic in all our experiments and it works well. **To answer RQ6**, we complete the hyperparameter analysis on k by changing this frequency to 0.5x and 2x of the critic (with other settings unchanged) and evaluating in *Predator Prey (3 agents)*. As shown in Fig.18, PMIC learns faster and achieves slightly better performance when increasing the update frequency of Du-MIE to 2x (red). Increasing the frequency of training Du-MIE provides more accurate estimates, but it also brings additional time consumption, which needs to be traded off.

Integrate PMIC with MASAC: we integrate PMIC with MASAC to further verify the generalisation and effectiveness of PMIC. The results are shown in Figure 19. PMIC also has significant improvement on MASAC. Beside, PMIC-MASAC is better than the other MASAC-based method, VM3-AC. The effectiveness and generalisation of PMIC in continuous action space is more convincing by the above experiments.

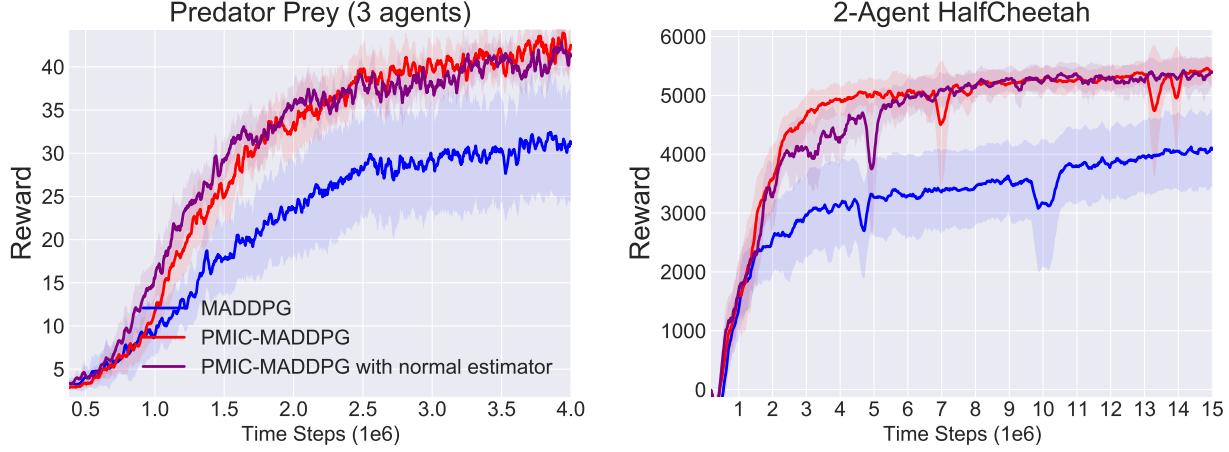


Figure 16: Comparison of PMIC with MINE and PMIC with normal estimator.

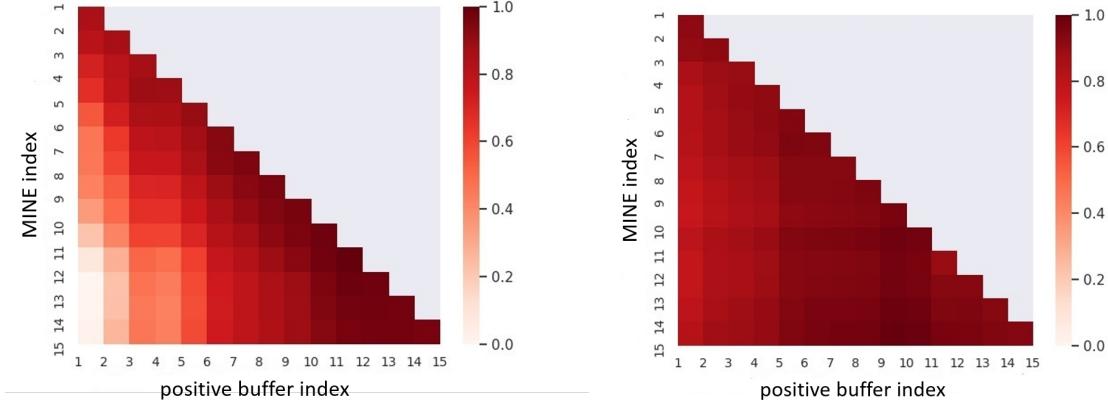


Figure 17: Visualization results of the MI estimated by different MINE for different positive buffer (The left uses our method, the right uses the regular method). X-axis denotes the index of positive buffer. Y-axis denotes the index of MINE. The value in coordinates (x, y) represents the MI estimated by the MINE with index y for the buffer with index x .

Integrate PMIC with QMIX: To better evaluate the generalization ability, we integrate PMIC with QMIX based on PyMARL2 (Hu et al., 2021) and evaluate PMIC-QMIX and QMIX on 3 maps. We run QMIX and PMIC-QMIX both in serial mode, and other settings remain the same as in the original paper. Fig. 20 shows that PMIC can also improve QMIX. This further verifies the generalization and effectiveness of PMIC on the discrete action space.

F. Comparison with Other Related Algorithms

In this section, we give a new question: Whether PMIC can be replaced by a discriminator, which can also achieve the purpose of guiding agents to better collaboration. To answer the question, more experiments are carried out. This idea of using a discriminator is similar to Generative Adversarial Self-Imitation Learning (GASIL) (Guo et al., 2018) where a discriminator is trained to discriminate between superior trajectories and inferior trajectories, while the policy learns to fool the discriminator by imitating superior trajectories. Specifically, GASIL maintains a good trajectory experience. The discriminator scores the good trajectory to 1.0 and scores the trajectory generated by current policy to 0.0. The results are shown in Figure 21 which indicates that our MI based architecture PMIC is more effective than GASIL, which may be mainly due to the fact: 1) GASIL only has a guiding role based on the discriminator. The inferior collaboration data is not

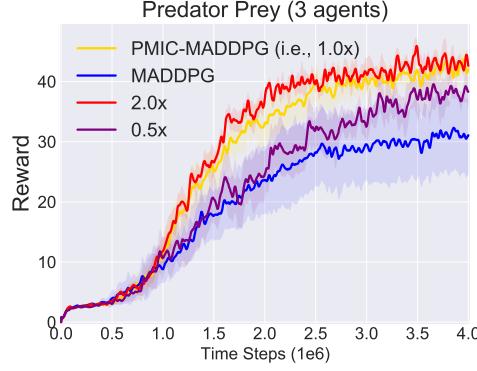


Figure 18: Hyperparameter analysis of k on *Predator Prey (3 agents)*.

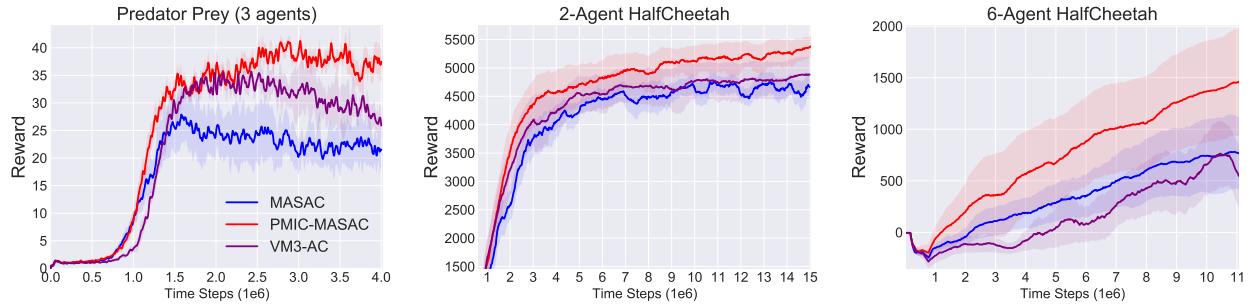


Figure 19: Comparisons of MASAC-based methods on MPE and Multi-Agent MuJoCo.

utilized to encourage agents exploration. 2) The discriminator is difficult to train and suffers from model collapse.

G. Hyperparameters

The hyperparameters for network architecture:

1: On MPE, we use two fully connected layers comprised of units 64 with ReLU nonlinearity and a final layer with tanh to output actions as the policy network for each agent, the critic network adopts the same architecture as the policy network except tanh for the final layer.

2: For Multi-Agent MuJoCo, the policy networks and critic network use the same architecture as those on MPE, but with 200 and 100 units for two fully connected layers.

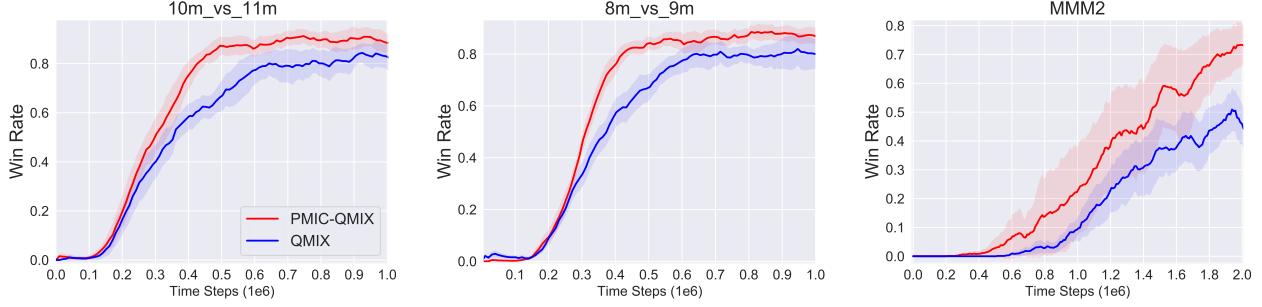


Figure 20: Comparisons of PMIC-QMIX and QMIX on SMAC

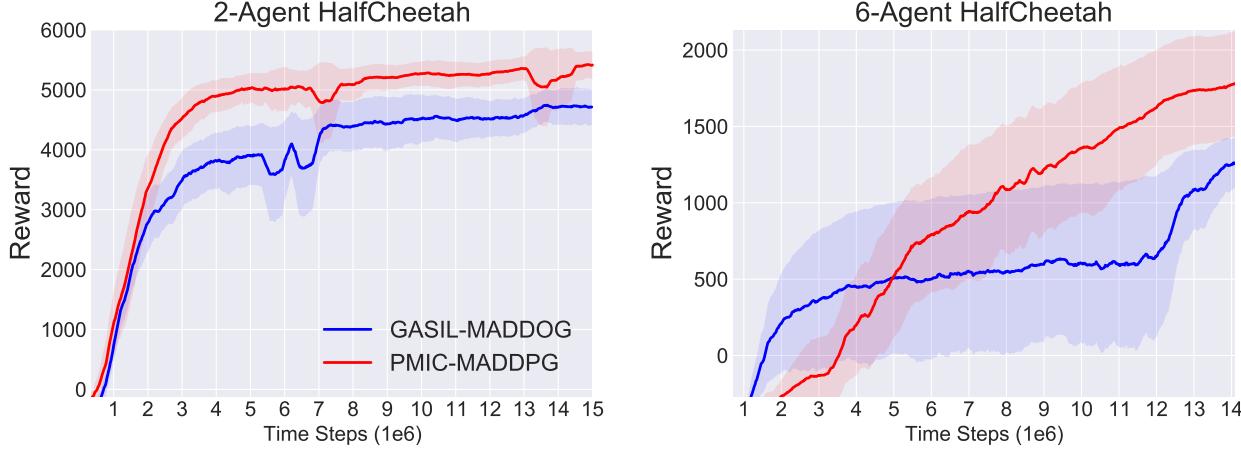


Figure 21: Comparisons between leveraging PMIC and GASIL.

3: For SMAC, we use code provided by RODE, the parameters and details are the same as the original paper.

For a fair comparison, all algorithms' network architecture remain consistent. In addition, we do not use the attention mechanism in the critic for all algorithms. For MADDPG, we use a centralized critic network and N policy networks. For MINE, we use two fully connected layers (100 units on Multi-Agent MuJoCo and 64 on MPE) with Leaky ReLU nonlinearity to encode global states and joint actions respectively, the results are obtained by a dot product of the embeddings of global states and joint actions. For CLUB, we use two fully connected layers (50 units on Multi-Agent MuJoCo and 32 on MPE) with Leaky ReLU nonlinearity to encode global states. For SMAC, we need to encode ρ , thus CLUB uses two fully connected layers (64 units) to encode ρ and the global states, then the outputs are concatenated to form a vector of 128 dimensions, and use the vector to predict the mean and variance of the global actions. MINE uses two fully connected layers (32 units) to encode ρ and the global states and uses a fully connected layer (64 units) to encode the joint action. The results are obtained by a dot product of the embeddings of the new vector (e.g., combine vectors of the global states and the joint role ρ) and joint actions. The architectures used to calculate MI in other MI-related algorithms (SIC-MADDPG and VM3-AC) use the same number of units and the activation function of MINE and the other parts are consistent with the setting in their original papers.

For SIC-MADDPG and VM3-AC, there are two parameters to adjust: the dimension of z and α . For SIC-MADDPG, we select z 's dimension from $[2, 3, 5, 8, 10, 15, 20]$ following the setting in original paper on MPE, $[2, 3, 4, 5, 8, 10, 15]$ on Multi-Agent MuJoCo and select α from $[0.1, 0.01, 0.001, 0.0001, 0.00001]$ following the setting in original paper. For VM3-AC, we select z 's dimension from $[2, 4, 8]$ following the setting in original paper and select α from $[0.1, 0.01, 0.001, 0.0001, 0.00001]$. For MASAC, we adjust β from $[1.0, 0.1, 0.01, 0.001, 0.0001]$, 0.00001 to control the entropy. For PMIC, we need to adjust α and β , we select from $[1.0, 0.1, 0.01, 0.001, 0.0001]$. The final choice of α , β and dimension of z is shown in Table 3. For FacMADDPG and COMIX, we use the official code and the parameters of the original paper.

For other hyperparameters on MPE and Multi-Agent MuJoCo, 1×10^{-3} for the critic and 1×10^{-4} for the actors on MPE except 1×10^{-2} on Wildlife Rescue and use Adam optimizer with learning rate 1×10^{-3} for the critic and 1×10^{-4} for the actors on Multi-Agent MuJoCo. For MINE and CLUB, the learning rate is 1×10^{-4} on all environments except 1×10^{-3} on Wildlife Rescue. The discounted factor γ and τ are 0.99 and 0.002 on Multi-Agent MuJoCo and 0.95 and 0.001 on MPE. Replay buffer size is 1×10^6 on MPE and Multi-Agent MuJoCo except 3×10^5 on Wildlife Rescue. The batch size is 1024 on MPE and 100 on Multi-Agent MuJoCo. The size of positive buffer and negative buffer of Du-PCB is 1000 on MPE except 6000 on Wildlife Rescue. 5000 for positive buffer and negative buffer on Multi-Agent MuJoCo. 500 for the positive buffer and 3000 for the negative buffer on SMAC.

For hyperparameters of RODE, all parameters remain the same as in the code provided in the original paper. RODE has two main adjusted hyperparameters: The number of role clusters and role interval. Number of role clusters is used to control the number of role types. The role interval decides how frequently the action spaces change and may have a critical influence

on the performance. All parameter settings are consistent with the original paper. The selection of α and β is shown in Table 4. The batch size to update MINE and CLUB is 128. We apply the maximization and minimization of MI after 100000 timesteps on SMAC.

Table 3: Selection of α , β and dimension of z for different algorithms on MPE and Multi-Agent MuJoCo.

Env name	SIC-MADDPG	PMIC-MADDPG	VM3-AC	MASAC
	αz dim	$\alpha \beta$	αz dim	β
Predator Prey (3 Agents)	0.00001 2	0.01 0.1	0.1 4	0.1
Predator Prey (6 Agents)	0.0001 8	0.01 0.1	0.01 8	0.1
Predator Prey (12 Agents)	0.0001 3	0.1 0.1	0.01 4	0.1
Predator Prey (24 Agents)	0.0001 2	0.01 0.1	0.01 4	0.1
Cooperative Navigation	0.0001 2	0.0001 0.01	0.1 4	0.01
Wildlife Rescue	0.001 3	0.001 0.1	0.001 2	0.001
2 agents HalfCheetah	0.0001 3	0.1 0.0001	0.1 8	0.01
6 agents HalfCheetah	0.0001 2	0.1 0.001	0.1 2	0.01

Table 4: Selection of α , β on SMAC.

Map name	PMIC-RODE
	$\alpha \beta$
MMM2	0.0001 0.0001
MMM	0.0001 0.001
2s3z	0.001 0.1
3s_vs_5z	0.001 0.01
3s5z	0.001 0.01
10_vs_11m	0.01 0.01

H. Experiments about MASAC-related Algorithms on Wildlife Rescue

Through experiments, we find that the MASAC-related algorithms can not learn positive rewards on Wildlife Rescue environment. Thus we experiment MASAC on Wildlife Rescue environment to find the reason.

There are two differences between MASAC and MADDPG: $\beta H(\pi)$ and double Q mechanism. Firstly, we make adjustments to β and find that no matter how we adjust β , we can not get positive rewards, even if β is set to 0.0. Secondly, we change the network architecture of MASAC by removing the double Q mechanism, then find that the algorithm can get positive rewards. Thus we hypothesize that the main reason why MASAC-related algorithms can not learn to cooperate on Wildlife Rescue is caused by the double Q mechanism.

To further verify our hypothesis, we add double Q to MADDPG and find significant performance degradation and slower convergence, which proves that the double Q mechanism is the main factor. We give some explanations, the Wildlife Rescue is characterized by high punishment for miss-coordination and low positive rewards. Double Q prevents overestimation but can lead to underestimation, thus using double Q might ignore less frequent positive rewards, leading to underestimation.

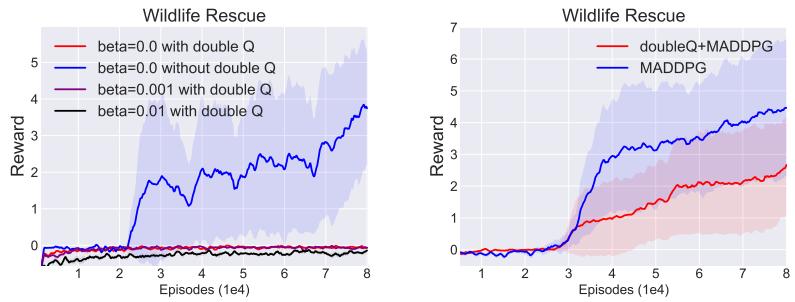


Figure 22: Ablation experiments of βH and double Q on Wildlife Rescue.