AIR: Unifying Individual and Collective Exploration in Cooperative Multi-Agent Reinforcement Learning

Guangchong Zhou[1,2], **Zeren Zhang**[1,2], **Guoliang Fan**[1]

[1]The Key Laboratory of Cognition and Decision Intelligence for Complex Systems,
Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
{zhouguangchong2021, zhangzeren2021, guoliang.fan}@ia.ac.cn

Abstract

Exploration in cooperative multi-agent reinforcement learning (MARL) remains challenging for value-based agents due to the absence of an explicit policy. Existing approaches include individual exploration based on uncertainty towards the system and collective exploration through behavioral diversity among agents. However, the introduction of additional structures often leads to reduced training efficiency and infeasible integration of these methods. In this paper, we propose Adaptive exploration via Identity Recognition (AIR), which consists of two adversarial components: a classifier that recognizes agent identities from their trajectories, and an action selector that adaptively adjusts the mode and degree of exploration. We theoretically prove that AIR can facilitate both individual and collective exploration during training, and experiments also demonstrate the efficiency and effectiveness of AIR across various tasks.

Code — https://github.com/Jugg1er/AIR

Introduction

Multi-agent reinforcement learning (MARL) has achieved outstanding results in complex cooperative tasks, such as flocking control (Xu et al. 2018; Gu et al. 2023), autonomous driving (Shamsoshoara et al. 2019; Zhang et al. 2023b), and sensor network (Zhao et al. 2023). To deal with the partial observability and the vast joint spaces of agents, most MARL methods follow the *centralized training and decentralized execution* (CTDE) paradigm. Especially, value-based approaches (Sunehag et al. 2017; Rashid et al. 2018; Yang et al. 2020; Wang et al. 2020a) enjoy high sample efficiency and great scalability, ultimately achieving superior performance in popular benchmarks. However, most value-based methods suffer from insufficient exploration due to the adoption of the vanilla ϵ-greedy exploration strategy, and how to effectively enhance the exploration of value-based agents in cooperative tasks remains a challenging problem.

In some studies, the exploration of an agent is guided by its uncertainty towards the system. Zhang et al. (2023a) define the uncertainty as the correlation between an agent's action and other agents' observations, based on which they adjust exploration rate ϵ dynamically. EITI & EDTI (Wang et al. 2019) and EMC (Zheng et al. 2021) refer to the concept of "curiosity", which is measured by the model prediction error in practice and used as an intrinsic reward to facilitate exploration. These approaches introduce perturbations into individual policies but do not explicitly consider coordination, and we call them individual exploration.

Compared to single-agent domain, multi-agent cooperative tasks typically require various skills and have vastly larger policy space, rendering comprehensive exploration infeasible for a single agent. An ideal solution is to keep the diversity of agents' behaviors, which enables individuals to acquire distinct skills, explore disparate regions of the policy space, and share their experiences by parameter sharing technique. Recently, a group of methods revealed the importance of behavioral diversity in collective exploration. As shown in Figure 1(a), all agents learn to compete for the ball and form terrible cooperation, while the globally optimal policy in Figure 1(b) induces different behaviors among agents for tacit cooperation (Li et al. 2021). Such sub-optimal situations arise exclusively in multi-agent tasks and can not be alleviated through individual exploration. MAVEN (Mahajan et al. 2019) proposes committed exploration that learns a diverse ensemble of monotonic approximations with the help of a latent space to explore. CDS (Li et al. 2021) maximizes an information-theoretical regularizer to produce diverse individualized behaviors for extensive exploration. In role-based methods (Wang et al. 2020b,c; Yang et al. 2022; Zhou et al. 2023), agents are required to deal with distinct subtasks based on their assigned roles, thereby reducing conflicts and accelerating the formation of cooperation.

Both individual skills and group collaboration are crucial in cooperative tasks. However, even without methodological conflicts, a direct integration of the above individual and collective exploration methods is infeasible, as excessive extra modules make the overall model too heavy to train. To the best of our knowledge, no research has integrated individual and collective exploration into a unified framework. To address this gap, we propose Adaptive exploration via Individual Recognition (AIR). We first learn an identity classifier that distinguishes corresponding agents based on given trajectories. Based on the classifier, we conduct theoretical analysis and design a unified exploration mechanism that can switch the exploration mode between individual and col-

(a) Sub-optimal policy.   (b) Optimal policy.
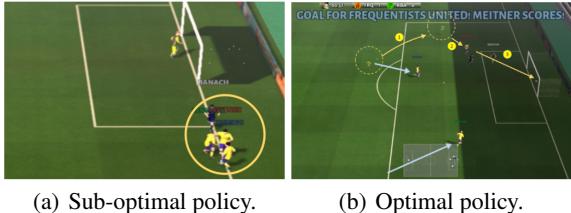
Figure 1: The importance of behavioral diversity in Google Research Football. (a) Agents all compete for the ball, exhibiting homogeneous behaviors and poor coordination. (b) Agents behave differently to achieve coordination.

lective exploration according to the training stage as well as tuning the degree of exploration. Our main contributions can be summarized as follows:

- We analyze the impact and limitations of traditional exploration strategies within multi-agent systems and reveal that multi-agent exploration accounts for both individual behaviors and the joint behaviors of agents.

- We propose a novel multi-agent exploration framework for value-based agents called AIR, which enables dynamic adjustments of exploration during training. To the best of our knowledge, we are the very first to address individual and collective exploration in a single unified framework.

- Our proposed method incorporates only the minimum additional structure and is computationally efficient. Experiments demonstrate the remarkable effectiveness of AIR across various multi-agent tasks.

Related Works

Value-based MARL Methods

Independent Q-learning (Tampuu et al. 2017) directly migrates DQN into multi-agent tasks, but it completely ignores the interactions between agents, leading to various training issues. Value decomposition factorizes the global value into individual ones and alleviates the instability during training. VDN (Sunehag et al. 2017) and QMIX (Rashid et al. 2018) aggregate individual Q-values into a global one through additive and monotonic functions respectively. Qatten (Yang et al. 2020) replaces the dense network in QMIX with the attention mechanism. To further enhance the capability of the mixing network to express richer joint action-value function classes, QTRAN (Son et al. 2019) constructs two soft regularizations to align the greedy action selections between the joint and individual value functions. Weighted-QMIX (Rashid et al. 2020) introduces a weighting mechanism into the projection of monotonic value factorization to place more importance on better joint actions. QPLEX (Wang et al. 2020a) proposes a duplex dueling network architecture to loose the monotonicity constraints in QMIX. QPro (Mei, Zhou, and Lan 2024) casts the factorization problem as regret minimization over the projection weights of different state-action values.

Exploration for Value-based Agents in MARL

Despite their success, the simple ϵ-greedy exploration strategy used in the above methods has been found ineffective in solving coordination tasks with complex state and reward transitions. Some studies adjust the intensity of individual exploration dynamically according to the metric of uncertainty towards the system. SMMAE (Zhang et al. 2023a) refers to the correlation between each agent's action and others' observations. EITI & EDTI (Wang et al. 2019) quantify and characterize the influence of one agent's behavior on others using mutual information and the difference of expected returns respectively, while EMC (Zheng et al. 2021) uses prediction errors of individual Q-values to capture the novelty of states and the influence from other agents.

Other studies have realized the importance of behavioral diversity in exploration in multi-agent tasks. MAVEN (Mahajan et al. 2019) uses a hierarchical policy to produce a shared latent variable and learns several state-action value functions for each agent to explore different trajectories. CDS (Li et al. 2021) proposes a novel information-theoretical objective to maximize the mutual information between agents' identities and trajectories to encourage diverse individualized behaviors. Role-based methods (Wang et al. 2020b,c; Yang et al. 2022; Zhou et al. 2023) decompose the complex task to subtasks, in which each agent search in a sub-region of the entire observation-action space, thereby reducing training complexity. The drawback of these methods lies in their introduction of additional modules to the policy network, which increases the complexity of the decision-making process and consequently reduces both training and execution efficiency.

Preliminaries

Dec-POMDP

A fully cooperative multi-agent system (MAS) is typically represented by a decentralized partially observable Markov decision process (Dec-POMDP) (Oliehoek and Amato 2016), which is composed of a tuple $G = \langle \mathcal{S}, \mathcal{U}, \mathcal{P}, \mathcal{Z}, r, \mathcal{O}, n, \gamma \rangle$. At each time-step, the current global state of the environment is denoted by $s \in \mathcal{S}$, while each agent $a \in \mathcal{A} := \{1, \ldots, n\}$ only receives a unique local observation $z_a \in \mathcal{Z}$ generated by the observation function $\mathcal{O}(s, a) : \mathcal{S} \times \mathcal{A} \to \mathcal{Z}$. Subsequently, every agent a selects an action $u_a \in \mathcal{U}$, and all individual actions are combined to form the joint action $\boldsymbol{u} = [u_1, \ldots, u_n] \in \mathcal{U} \equiv \mathcal{U}^n$. The interaction between the joint action $\boldsymbol{u}$ and the current state s leads to a change in the environment to state s' as dictated by the state transition function $\mathcal{P}(s'|s, \boldsymbol{u}) : \mathcal{S} \times \mathcal{U} \times \mathcal{S} \to [0, 1]$. All agents in the Dec-POMDP share the same global reward function $r(s, \boldsymbol{u}) : \mathcal{S} \times \mathcal{U} \to \mathbb{R}$, and $\gamma \in [0, 1)$ represents the discount factor.

Value Decomposition

Credit assignment is a key problem in cooperative MARL which helps each agent to discern its contribution to the collective performance. To solve this, value decomposition (VD) methods assume that each agent has a specific value function, and the integration of all individual functions

creates the joint value function. To guarantee that the optimal action for each agent aligns with the global optimal joint action, most value decomposition methods satisfy the *Individual Global Max* (IGM) assumption:

$$\arg\max_{\boldsymbol{u}} Q_{tot}(\boldsymbol{\tau}, \boldsymbol{u}) = \arg\max_{u_a} Q_a(\tau_a, u_a), \quad \forall a \in \mathcal{A}$$

where $Q_{tot} = f(Q_1, ..., Q_n)$, $Q_1, ..., Q_n$ denote the individual Q-values, and f is the mixing function. Various value decomposition methods, including VDN (Sunehag et al. 2017), QMIX (Rashid et al. 2018) and so on, arise from variations in the construction of f. The model is commonly learnt by the temporal difference (TD) error below, in which Q_{tot}^- is a periodically updated target network of Q_{tot}.

$$\mathcal{L}_{\text{TD}} = \left[r_t + \gamma \max_{\boldsymbol{u}_{t+1}} Q_{tot}^-(\boldsymbol{\tau}_{t+1}, \boldsymbol{u}_{t+1}) - Q_{tot}(\boldsymbol{\tau}_t, \boldsymbol{u}_t) \right]^2$$

Method

Compared to single-agent reinforcement learning, exploration in MARL also necessitates behavioral diversity across agents. In this section, we introduce a novel strategy called Adaptive exploration via Individual Recognition (AIR), which integrates individual and collective exploration into a unified framework. We first formulate a discrepancy metric of trajectories to measure the behavioral diversity among different agents and obtain the learning objective for collective exploration by theoretical derivation. With this objective, we further design an adversarial strategy to stimulate each agent to deviate from a fixed policy and explore novel actions. Finally, we integrate the collective and individual exploration into a unified framework.

Diverse Behaviors across Agents

We begin the derivation of AIR by defining the concept of *trajectory visit distribution*:

Definition 1 (Trajectory visit distribution). In a multi-agent system, the distribution of the trajectory experienced by the agent k is called individual trajectory visit distribution and is defined as:

$$\rho^k(\tau_T, u_T) = \rho(\tau_T, u_T | \text{agent\_id} = k)$$
$$= \rho(o_0, u_0, o_1, u_1, ..., o_T, u_T | z_k)$$
$$= \prod_{t=0}^{T} \pi^k(u_t | o_t) \sum_{s_t} P(s_t) O(o_t | s_t, k), \quad (1)$$

where $P(s_t)$ is the probability of the state being s_t at time step t, $O(o_t | s_t, k)$ is the probability of agent k receiving observation o_t under state s_t, and π^k is the policy of agent k. The trajectory visit distribution of the overall system containing n agents is called system trajectory visit distribution and is defined as:

$$\rho = \frac{1}{n} \sum_{k=1}^{n} \rho^k. \quad (2)$$

The trajectory visit distribution, which indicates the probability of encountering a specific trajectory with a given policy from initial state s_0, encapsulates the diversity of policies

naturally. Therefore, the diverse behaviors of agents can be assessed by measuring the diversity of trajectories generated by distinct individual policies. The Kullback-Leibler (KL) divergence is a commonly used metric for comparing the discrepancy between distributions, by which we define the difference between an agent's policy and other agents' policies as below:

Definition 2 (Difference between individual policies). In a multi-agent system with N agents, the difference between the policy of agent k and the policies of other agents can be defined by the discrepancy of their generated trajectories:

$$\mathcal{D}_{\text{KL}} \left[\rho^k || \rho \right] = \mathcal{D}_{\text{KL}} \left[\rho(\tau, u | z_k) || \rho(\tau, u) \right] \quad (3)$$

According to Equation 3, we improve the difference between an individual trajectory visit distribution and the system trajectory visit distribution to enforce diverse behaviors of agents. However, it is completely infeasible to estimate $\rho(\tau, u)$ directly due to an intractably huge input space. We notice that Equation 3 can be expressed by the form of entropy discrepancy during training:

$$\mathbb{E}_z \left[\mathcal{D}_{\text{KL}} \left(\rho(\tau, u | z) || \rho(\tau, u) \right) \right] = \mathcal{H}(\rho) - \mathcal{H}(\rho | z), \quad (4)$$

in which $\mathcal{H}(\rho)$ is the Shannon Entropy of trajectory visit distribution. The proof of Equation 4 can be found in Appendix. With the mutual information theory, it can be further transformed into:

$$\mathcal{H}(\rho) - \mathcal{H}(\rho | z) = \mathcal{H}(z) - \mathcal{H}(z | \rho) \quad (5)$$

This optimization target is easier to solve in practice. As each agent k has the same number of sampled trajectories in a training step, $\rho(z = k)$ follows a uniform distribution and thus $\mathcal{H}(z) = -\sum_{k=1}^{n} \frac{1}{n} \log \frac{1}{n} = \log n$ is a constant. The second term $\mathcal{H}(z | \rho)$ implies that the identity of an agent can be inferred given its trajectory and selected action. We approximate true posterior $\rho(z | \tau, u)$ with a learned discriminator $q_\zeta(z | \tau, u)$ and optimize it using the evidence lower bound (ELBO).

$$\mathcal{H}(z) - \mathcal{H}(z | \rho) = \mathbb{E}_{\tau, u, z} \left[\log \rho(z | \tau, u) \right] + \log n$$
$$= \mathbb{E}_{\tau, u} \left[\mathcal{D}_{\text{KL}} \left(\rho(z | \tau, u) || q_\zeta(z | \tau, u) \right) \right]$$
$$+ \mathbb{E}_{\tau, u, z} \left[\log q_\zeta(z | \tau, u) \right] + \log n$$
$$\geq \mathbb{E}_{\tau, u, z} \left[\log q_\zeta(z | \tau, u) \right] + \log n, \quad (6)$$

where we exploit the non-negativity of KL divergence. Therefore, with the sub-goal of promoting behavioral diversity, the policy of each agent k can be expressed as below:

$$u^* = \arg\max_{u} \left[Q^k(\tau, u) + \alpha \log q_\zeta(z_k | \tau, u) \right], \quad (7)$$

where α is a positive temperature controlling the degree of collective exploration. For more precise approximation, we further tighten the lower bound by minimizing $\mathcal{D}_{\text{KL}} \left(\rho(z | \tau, u) || q_\zeta(z | \tau, u) \right)$ and derive the gradient of q_ζ as shown in Equation 8.

$$\nabla_\zeta \mathbb{E}_{\tau, u} \left[\mathcal{D}_{\text{KL}} \left(\rho(z | \tau, u) || q_\zeta(z | \tau, u) \right) \right]$$
$$= -\mathbb{E}_{\tau, u} \left[\nabla_\zeta \log q_\zeta(z | \tau, u) \right]. \quad (8)$$

Individual Exploration of Value-based Agents

In essence, exploration at the individual level drives deviation from established policies, prompting agents to take unfamiliar and unknown actions. A common individual exploration strategy used in policy gradient methods is to add the entropy of policy π_θ as a regularization term in the optimization target $J(\theta)$. While it can adjust the degree of exploration dynamically during training, this approach does not work for VD methods because the value-based agent makes decision according to its value estimations towards the given observation and possible actions instead of an explicit policy distribution.

To facilitate exploration for VD methods, we introduce a novel adversarial mechanism somewhat similar to generative adversarial networks (GAN) (Goodfellow et al. 2014). Under the CTDE framework, we introduce two adversarial components: a centralized identity classifier trained to identify the agent responsible for generating the given trajectory-action pair (we directly reuse the discriminator q_ζ for collective exploration in practice), and a decentralized action selector that attempts to mislead the classifier. This adversarial mechanism can be interpreted by the following relationship (the proof is in Appendix):

$$p(z_k|\tau_T, u_T) = \frac{\prod\limits_{t=0}^{T} \pi^k(u_t|o_t)}{\prod\limits_{t=0}^{T} \pi^k(u_t|o_t) + \sum\limits_{\substack{i=1\\i\neq k}}^{n} \prod\limits_{t=0}^{T} \pi^i(u_t|o_t)} \quad (9)$$

According to Equation 9, the higher the probability that an action u_t in the trajectory is selected by the policy, the higher the likelihood that the classifier will correctly identify the source agent k of the trajectory, and vice versa. Thus, even without an explicit policy function, we can leverage the classifier to measure action selection probabilities and encourage the agent to explore low-probability actions, achieving a similar effect to the entropy-based approach. Without distorting the individual Q-value estimation, the action selector promotes the agent k to choose unexplored actions with low posterior probabilities $q_\zeta(z_k|\tau, u)$. We formally describe how the action selector works as:

$$u = \arg\max_u \tilde{Q}^k(\tau, u)$$
$$= \arg\max_u \left[Q^k(\tau, u) - \alpha \log q_\zeta(z_k|\tau, u) \right] \quad (10)$$

Equation 10 has a very similar form to Equation 7, and we take Equation 10 as the standard form. The value estimations of less exploited actions are raised more to increase their chances of being chosen when $\alpha > 0$, thereby enhancing the individual exploration of the agent. When $\alpha < 0$, it turns to collective exploration to induce diverse and cooperative behaviors among agents.

Adaptive Temperature

In Equation 10, the tuning of temperature value α is closely related to the magnitude of Q-value estimation, which is further determined by the magnitude of received rewards that differs not only across tasks but also over the training process due to the changing policy. Therefore, choosing the optimal temperature α^* is non-trivial. Besides, forcing the temperature to a fixed value during training is a poor solution, since the mode and degree of exploration are expected to be adjusted adaptively at different training stages. To dynamically tune the temperature during training, we formulate the learning target as a constrained optimization problem for the agent k:

$$\max_{\pi_{0:T}^k} \mathbb{E}_\rho \left[\sum_{t=0}^{T} r(\boldsymbol{\tau}_t, \boldsymbol{u}_t) \right] \equiv \max_{\pi_{0:T}^k} \mathbb{E}_{\rho^k} \left[\sum_{t=0}^{T} r(\tau_t, u_t) \right]$$
$$\text{s.t. } \mathbb{E}_{(\tau_t, u_t)\sim\rho^k} \left[-\log q_\zeta(z_k|\tau_t, u_t) \right] \geq \mathcal{H} \quad \forall t, \quad (11)$$

where $r(\boldsymbol{\tau}_t, \boldsymbol{u}_t)$ is the global reward function and $r(\tau_t, u_t)$ is the factored individual reward at time step t under IGM assumption, and $\mathcal{H}$ is a minimum expected entropy. Since the policy and trajectory at t would affect the future calculation, we rewrite the objective in a recursive form to perform dynamic programming backward:

$$\max_{\pi_0^k} \left(\mathbb{E}[r(\tau_0, u_0)] + \max_{\pi_1^k} \left(\mathbb{E}[...] + \mathbb{E}_{\pi_T^k}[r(\tau_T, u_T)] \right) \right), \quad (12)$$

subjected to the same constraint in Equation 11. To focus on the temperature, we turn to the dual problem instead. We start with the last time step T. With the constraint in Equation 11,

$$\max_{\pi_T^k} \mathbb{E}_{(\tau_T, u_T)\sim\rho^k} [r(\tau_T, u_T)] \quad (13)$$
$$= \min_{\alpha_T \geq 0} \max_{\pi_T^k} \mathbb{E}[r(\tau_T, u_T) - \alpha_T \log q_\zeta(z_k|\tau_T, u_T)] - \alpha_T \mathcal{H},$$

where α_T is the dual variable. As π_T^k is in the form of Equation 10 and only related to q_ζ when the Q-value function is fixed, we can replace π_T^k in the subscripts of Equation 13 with q_ζ. Here we employ strong duality since the objective is linear and the constraint in Equation 11 is a convex function with respect to q_ζ. With the correspondence between the optimal policy deduced by q_ζ^* and the temperature value α_T, the optimal dual variable α_T^* is solved by

$$\arg\max_{\alpha_T} \mathbb{E}_{(\tau_t, u_t)\sim q_\zeta^*} [\alpha_T \log q_\zeta(z_k|\tau_T, u_T) + \alpha_T \mathcal{H}]. \quad (14)$$

To simplify notation, we refer to the soft Bellman equation (Haarnoja et al. 2017)

$$Q^*(\tau_t, u_t) = r(\tau_t, u_t) + \mathbb{E}_{\rho^k}[Q^*(\tau_{t+1}, u_{t+1}) - \alpha \log q_\zeta(z_k|\tau_{t+1}, u_{t+1})] \quad (15)$$

with $Q_T^*(\tau_T, u_T) = \mathbb{E}[r(\tau_T, u_T)]$. We now take Equation 12 a step further using the dual problem and get:

$$\max_{\pi_{T-1}^k} \left(\mathbb{E}[r(\tau_{T-1}, u_{T-1})] + \max_{\pi_T^k} \mathbb{E}[r(\tau_T, u_T)] \right)$$
$$= \max_{\pi_{T-1}^k} (Q^*(\tau_{T-1}, u_{T-1}) - \alpha_T \mathcal{H}) \quad (16)$$
$$= \min_{\alpha_{T-1}\geq 0} \max_{T-1} \left(\mathbb{E}[Q^*(\tau_{T-1}, u_{T-1})] - \mathbb{E}[\alpha_{T-1} \log q_\zeta(z_k|\tau_{T-1}, u_{T-1})] - \alpha_{T-1}\mathcal{H} \right) + \alpha_T^* \mathcal{H}.$$

Algorithm 1: Adversarial Identity Recognition (AIR)

Parameters: Value decomposition framework θ, target network $\theta^- = \theta$, identity classifier ζ, initial temperature α_0.

1: Initialize an empty replay buffer $\mathcal{D}$.
2: **while** *training* **do**
3:    **for** $episode \leftarrow 1$ **to** M **do**
4:       Initialize $\boldsymbol{E} = \emptyset$.
5:       **for** each time step t **do**
6:          **for** each agent k **do**
7:             Select action based on $\tilde{Q}^k$ in Equation 10.
8:             Store the transition in $\boldsymbol{E}$.
9:          **end for**
10:       **end for**
11:       Add the episodic data $\boldsymbol{E}$ to $\mathcal{D}$.
12:    **end for**
13:    Sample a mini-batch $\mathcal{B}$ from $\mathcal{D}$.
14:    Follow the value decomposition method to update θ,
15:    $\zeta \leftarrow \zeta + \lambda_\zeta \hat{\nabla}_\zeta J(\zeta)$,
16:    $\alpha \leftarrow \alpha + \lambda_\alpha \hat{\nabla}_\alpha J(\alpha)$,
17:    Update target network $\theta^- \leftarrow \theta$ periodically.
18: **end while**

Thus Equation 11 can be optimized recursively from back to front. After obtaining the optimal Q-function Q^* and the corresponding optimal policy π_t^*, the optimal dual variable α_t^* can be solved by:

$$\alpha_t^* = \arg\max_{\alpha_t} \mathbb{E}_{u_t \sim \pi_t^*} \left[\alpha_t \log q_\zeta(z_k|\tau_t, u_t) + \alpha_t \bar{\mathcal{H}} \right]. \tag{17}$$

Equation 17 provides a theoretical approach for calculating the optimal temperature, but it is not feasible in practice since we can not obtain Q^* directly but iteratively update the Q-function to approximate it. Therefore, we refer to dual gradient descent (Boyd and Vandenberghe 2004), which alternates between finding the optimal values of primal variables with fixed dual variables and applying gradient descent on dual variables for one step. It is also noted that while optimizing with respect to the primal variables fully is impractical, a truncated version that performs incomplete optimization (even for a single gradient step) can be shown to converge under convexity assumptions. Although the assumptions are not strictly satisfied, Haarnoja et al. (2018) still extended the technique to neural networks and found it working in practice. Thus, we compute gradients for α according to the objective in Equation 18. $\mathcal{H}$ is a hyperparameter and is calculated as the running mean of $-\log q_\zeta(z_k|\tau_t, u_t)$ in practice. The overall algorithm is displayed in Algorithm 1.

$$J(\alpha) = \mathbb{E}_{u_t \sim \pi_t^k} \left[\alpha \log q_\zeta(z_k|\tau_t, u_t) + \alpha \bar{\mathcal{H}} \right] \tag{18}$$

Experiments

In this section, we conduct a large set of experiments to validate the efficiency and effectiveness of AIR across various scenarios. To be consistent with previous methods (Mahajan et al. 2019; Wang et al. 2020c; Yang et al. 2022), we similarly employ QMIX (Rashid et al. 2018) as the

backbone while using AIR for exploration. Several state-of-the-art MARL methods are selected as baselines, including QMIX (Rashid et al. 2018), QPLEX (Wang et al. 2020a), and recognized works about multi-agent exploration (MAVEN (Mahajan et al. 2019), RODE (Wang et al. 2020c), LDSA (Yang et al. 2022)). All the baselines are implemented with the open-source codes in their original papers. We run each algorithm on 5 different random seeds for every task to alleviate the randomness in experiments.

StarCraft II Multi-Agent Challenges

StarCraft II Multi-Agent Challenges (SMAC) (Samvelyan et al. 2019) is a popular benchmark for cooperative MARL. There are two armies of units in SMAC. Each ally unit is controlled by a decentralized agent that can only act based on its local observation and the enemy units are controlled by built-in handcrafted heuristic rules. The performance of an MARL method is assessed by its test win rates in the SMAC scenarios. In our experiments, the difficulty level of built-in AI is set to 7 (very hard), and the version of StarCraft II engine is 4.6.2 instead of the simpler 4.10.

We present experiment results on 6 *Hard* and *Super Hard* SMAC scenarios in Figure 2. The solid line represents the median win rates during training, with the 25-75% percentiles being shaded. Although the win rate in *corridor* is slightly lower than RODE, AIR reaches the best performances in 5 out of all 6 scenarios with both the highest win rates and the earliest rise-ups. It is noteworthy that different scenarios necessitate distinct algorithmic capabilities. For example, agents in *corridor* should adopt specialized roles alternately, with some drawing enemies and others focusing fire on enemies, which requires the algorithm to induce diverse agent behaviors. *3s5z_vs_3s6z* and *6h_vs_8z* are two hardest scenarios that require extensive exploration (Hu et al. 2021). While other multi-agent exploration methods (MAVEN, RODE and LDSA) only work on *corridor* and completely fail in *3s5z_vs_3s6z* and *6h_vs_8z*, we are delighted to find that our proposed method exhibits remarkable sample efficiency and effectiveness across scenarios, which strongly demonstrates the superiority of AIR's multi-agent exploration strategy.

Google Research Football

Experiments in numerous previous studies have revealed that algorithms claiming state-of-the-art performance often exhibit significant degradation when applied to different environments. To assess the generalizability of AIR to other environments, we conducted further experiments in Google Research Football (GRF) (Kurach et al. 2020). GRF is a user-friendly platform for researchers to explore and develop novel RL algorithms with customized scenarios, which necessitates the acquisition of sophisticated strategies involving teamwork, ball control, and tactical decision-making.

We choose three official scenarios from Football Academy, and the experiment results are displayed in Figure 3. *Academy_pass_and_shoot_with_keeper* is a relatively easy scenario where AIR performs closely with the best baselines. *Academy_corner* and *academy_run_pass_and_shoot_with_keeper* are two harder
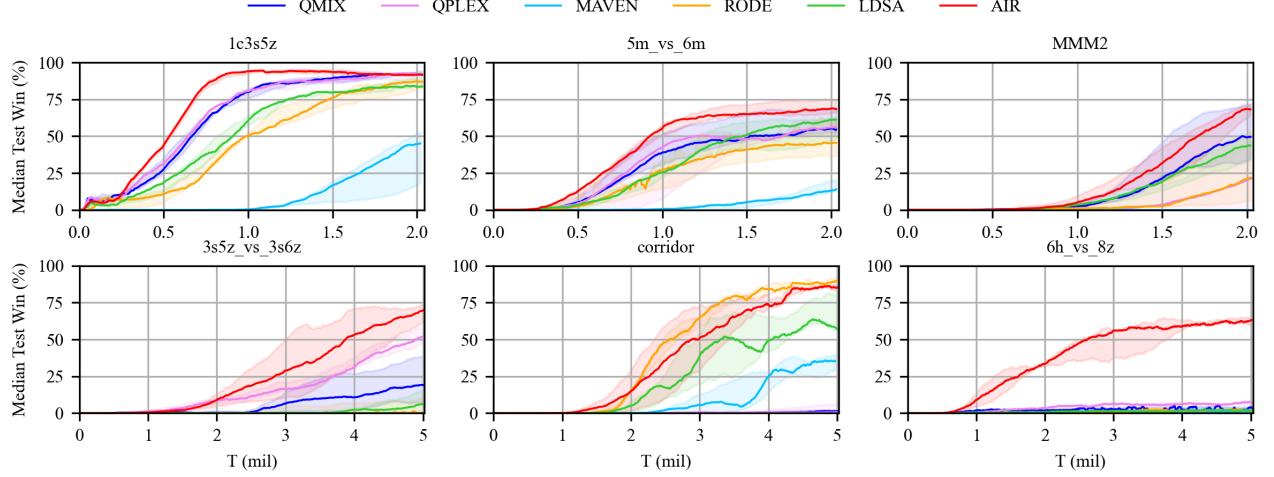
Figure 2: Experiment results of AIR and baselines on SMAC.
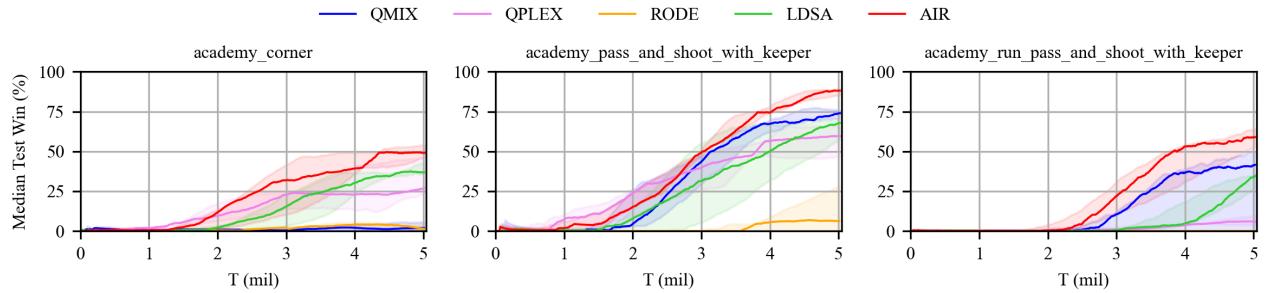
Figure 3: Experiment results of AIR and baselines on GRF.

scenarios that require tacit cooperation of agents. We can see that most of the baselines can address only one of the two scenarios, whereas AIR achieves significantly superior performance in both. Although RODE is considered a state-of-the-art method and even outperforms AIR in SMAC *corridor*, it experiences a substantial performance decline in GRF. In contrast, AIR maintains its superiority in both environments. Therefore, we can conclude that AIR not only exhibits exceptional performance across diverse scenarios within the same environment but also maintains its superiority when applied to different environments.

Case Study

In order to study how AIR adjusts the exploration during training, we dive into a training session conducted on the *corridor* scenario in SMAC. We record the win rate and temperature values during training, and we reduce the high-dimensional observation vectors to 2D space via the t-SNE (Van der Maaten and Hinton 2008) algorithm for dimensionality reduction to visualize the distributions of agents' trajectories at different training steps.

According to Figure 4, the temperature α is positive during the first half of the training process, thus the individual exploration is enhanced. At $T = 300k$, the region visited
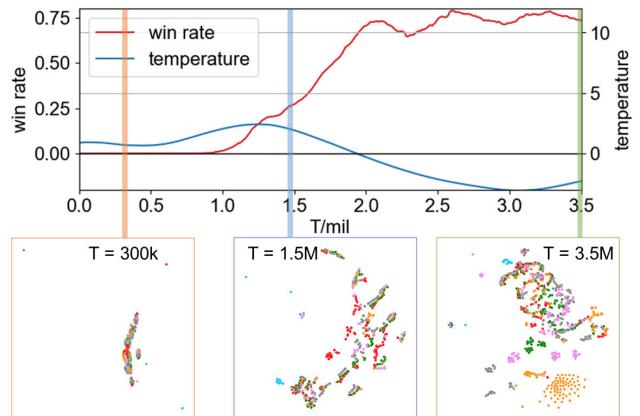

Figure 4: The training session on SMAC *corridor*. **Up**: The curves that depict the changes of win rate and temperature value during training. **Down**: The 2D t-SNE visualizations of the agents' trajectories at different training steps.
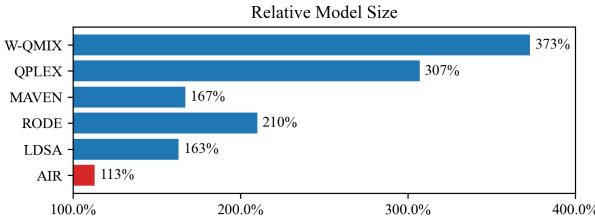
Figure 5: The relative model sizes of algorithms.

Figure 6: The experiment results of AIR and its variants with distinct exploration modes.

Figure 7: The experiment results of AIR and QMIX variants with varied degrees of individual exploration.

by the agents is very limited, with few noises distributed away from the cluster. The degree of individual exploration reaches the peak at around $T = 1.5M$, when all the agents visit a much wider region than they did at $T = 300k$ steps, which confirms an extensive individual exploration between the two timestamps. After 2 million training steps, the temperature turns negative and the exploration mode is switched to collective exploration. At $T = 3.5M$, the trajectories of different agents exhibit initial proximity at the beginning of an episode, but subsequently diverge and become readily distinguishable, while at $T = 1.5M$ the trajectories of agents are largely intertwined. This suggests an obvious increase in the diversity of agents' behaviors. Therefore, we can conclude that AIR integrates individual and collective exploration and can dynamically adjust the mode and degree of exploration during training.

Ablation Study

To ensure a fair comparison in performance, it is necessary to exclude the effect of model size. We set the model size of QMIX as the baseline (100%) and display the relative model sizes of other algorithms with respect to QMIX in Figure 5. Despite utilizing an extremely lightweight model, AIR still achieves significant results across tasks, demonstrating its great efficiency and superiority.

The primary advantage of AIR over other exploration methods for cooperative MARL lies in its integration of individual and collective exploration within a unified framework, as well as allowing for dynamic adjustment of exploration modes and intensity based on training progress. We first validate the benefits of unifying the two types of exploration and design two variants, AIR-in and AIR-co, which constrain the temperature value to be positive and negative during training, so the mode of exploration is set as **in**dividual and **co**llective respectively. We conduct experiments in SMAC *MMM2* and *6h_vs_8z* scenarios, and the results are shown in Figure 6. When AIR incorporates only a single mode of exploration (AIR-in or AIR-co), its performance is substantially inferior to the original AIR, which integrates both individual and collective exploration modes. Therefore, the integration of individual and collective exploration within a unified framework is a key factor contributing to AIR's superior performance.

Furthermore, we investigate the impact of dynamically adjusting the exploration degree (i.e., the temperature value) on the training process with a fixed exploration mode. However, selecting a proper temperature value α is non-trivial as mentioned above. Drawing from the study by Hu et al.
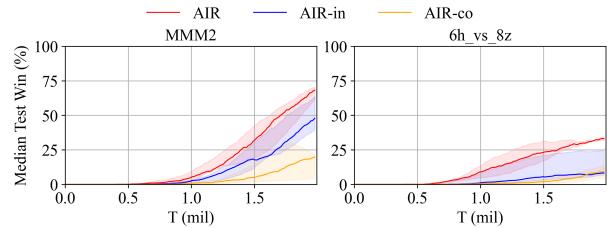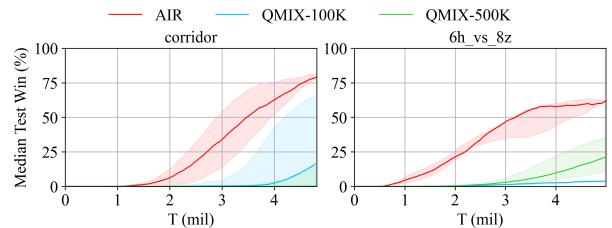
(2021), we use QMIX with ϵ-greedy exploration strategy as the control group and adjust the degree of individual exploration by switching the value of ϵ anneal period between $100K$ and $500K$. As shown in Figure 7, with more intensive exploration, QMIX-500K performs better than QMIX-100K in *6h_vs_8z*, but it would completely fail in *corridor* as over-exploration ruins the exploitation of agents. Meanwhile, the degree of exploration in AIR can be dynamically adapted to the specific task and training process, rendering the success of AIR in both scenarios. Evidently, the mechanism of AIR for dynamically adjusting exploration intensity is of great superiority.

Conclusion

Exploration for value-based agents in cooperative multi-agent tasks remains a tough problem, partly due to the absence of an explicit policy. Previous works have made some progress mainly through individual exploration based on agent's uncertainty towards the system or collective exploration focusing on agents' behavioral diversity. In this paper, we propose Adaptive exploration via Identity Recognition (AIR), which is the very first approach to integrate both individual and collective in a unified framework. The core module in AIR is an identity classifier that distinguishes corresponding agents based on given trajectories. According to our theoretical analysis, the accuracy of the classifier is highly related to the exploration of agents. Consequently, we design a unified exploration mechanism that can switch the exploration mode between individual and collective exploration according to the training stage as well as tuning the degree of exploration. We conduct extensive experiments and studies on varied scenarios, and the results strongly demonstrate the superiority of our proposed method.

Acknowledgements

References

Boyd, S.; and Vandenberghe, L. 2004. *Convex optimization.* Cambridge university press.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Gu, S.; Kuba, J. G.; Chen, Y.; Du, Y.; Yang, L.; Knoll, A.; and Yang, Y. 2023. Safe multi-agent reinforcement learning for multi-robot control. *Artificial Intelligence*, 319: 103905.

Haarnoja, T.; Tang, H.; Abbeel, P.; and Levine, S. 2017. Reinforcement learning with deep energy-based policies. In *International conference on machine learning*, 1352–1361. PMLR.

Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, 1861–1870. PMLR.

Hu, J.; Jiang, S.; Harding, S. A.; Wu, H.; and Liao, S.-w. 2021. Rethinking the implementation tricks and monotonicity constraint in cooperative multi-agent reinforcement learning. *arXiv preprint arXiv:2102.03479*.

Kurach, K.; Raichuk, A.; Stańczyk, P.; Zajac, M.; Bachem, O.; Espeholt, L.; Riquelme, C.; Vincent, D.; Michalski, M.; Bousquet, O.; and Gelly, S. 2020. Google Research Football: A Novel Reinforcement Learning Environment. arXiv:1907.11180.

Li, C.; Wang, T.; Wu, C.; Zhao, Q.; Yang, J.; and Zhang, C. 2021. Celebrating diversity in shared multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34: 3991–4002.

Mahajan, A.; Rashid, T.; Samvelyan, M.; and Whiteson, S. 2019. Maven: Multi-agent variational exploration. *Advances in Neural Information Processing Systems*, 32.

Mei, Y.; Zhou, H.; and Lan, T. 2024. Projection-Optimal Monotonic Value Function Factorization in Multi-Agent Reinforcement Learning. In *AAMAS*, 2381–2383.

Oliehoek, F. A.; and Amato, C. 2016. *A concise introduction to decentralized POMDPs.* Springer.

Rashid, T.; Farquhar, G.; Peng, B.; and Whiteson, S. 2020. Weighted qmix: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning. *Advances in neural information processing systems*, 33: 10199–10210.

Rashid, T.; Samvelyan, M.; Schroeder, C.; Farquhar, G.; Foerster, J.; and Whiteson, S. 2018. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *International conference on machine learning*, 4295–4304. PMLR.

Samvelyan, M.; Rashid, T.; De Witt, C. S.; Farquhar, G.; Nardelli, N.; Rudner, T. G.; Hung, C.-M.; Torr, P. H.; Foerster, J.; and Whiteson, S. 2019. The starcraft multi-agent challenge. *arXiv preprint arXiv:1902.04043*.

Shamsoshoara, A.; Khaledi, M.; Afghah, F.; Razi, A.; and Ashdown, J. 2019. Distributed cooperative spectrum sharing in uav networks using multi-agent reinforcement learning. In *2019 16th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, 1–6. IEEE.

Son, K.; Kim, D.; Kang, W. J.; Hostallero, D.; and Yi, Y. 2019. QTRAN: Learning to Factorize with Transformation for Cooperative Multi-Agent Reinforcement Learning. *CoRR*, abs/1905.05408.

Sunehag, P.; Lever, G.; Gruslys, A.; Czarnecki, W. M.; Zambaldi, V.; Jaderberg, M.; Lanctot, M.; Sonnerat, N.; Leibo, J. Z.; Tuyls, K.; et al. 2017. Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296*.

Tampuu, A.; Matiisen, T.; Kodelja, D.; Kuzovkin, I.; Korjus, K.; Aru, J.; Aru, J.; and Vicente, R. 2017. Multiagent cooperation and competition with deep reinforcement learning. *PloS one*, 12(4): e0172395.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, J.; Ren, Z.; Liu, T.; Yu, Y.; and Zhang, C. 2020a. Qplex: Duplex dueling multi-agent q-learning. *arXiv preprint arXiv:2008.01062*.

Wang, T.; Dong, H.; Lesser, V.; and Zhang, C. 2020b. Roma: Multi-agent reinforcement learning with emergent roles. *arXiv preprint arXiv:2003.08039*.

Wang, T.; Gupta, T.; Mahajan, A.; Peng, B.; Whiteson, S.; and Zhang, C. 2020c. Rode: Learning roles to decompose multi-agent tasks. *arXiv preprint arXiv:2010.01523*.

Wang, T.; Wang, J.; Wu, Y.; and Zhang, C. 2019. Influence-based multi-agent exploration. *arXiv preprint arXiv:1910.05512*.

Xu, Z.; Lyu, Y.; Pan, Q.; Hu, J.; Zhao, C.; and Liu, S. 2018. Multi-vehicle flocking control with deep deterministic policy gradient method. In *2018 IEEE 14th International Conference on Control and Automation (ICCA)*, 306–311. IEEE.

Yang, M.; Zhao, J.; Hu, X.; Zhou, W.; Zhu, J.; and Li, H. 2022. Ldsa: Learning dynamic subtask assignment in cooperative multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 35: 1698–1710.

Yang, Y.; Hao, J.; Liao, B.; Shao, K.; Chen, G.; Liu, W.; and Tang, H. 2020. Qatten: A general framework for cooperative multiagent reinforcement learning. *arXiv preprint arXiv:2002.03939*.

Zhang, S.; Cao, J.; Yuan, L.; Yu, Y.; and Zhan, D.-C. 2023a. Self-motivated multi-agent exploration. *arXiv preprint arXiv:2301.02083*.

Zhang, Z.; Han, S.; Wang, J.; and Miao, F. 2023b. Spatial-temporal-aware safe multi-agent reinforcement learning of connected autonomous vehicles in challenging scenarios. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 5574–5580. IEEE.

Zhao, G.; Lin, K.; Chapman, D.; Metje, N.; and Hao, T. 2023. Optimizing energy efficiency of LoRaWAN-based wireless underground sensor networks: A multi-agent reinforcement learning approach. *Internet of Things*, 22: 100776.

Zheng, L.; Chen, J.; Wang, J.; He, J.; Hu, Y.; Chen, Y.; Fan, C.; Gao, Y.; and Zhang, C. 2021. Episodic multi-agent reinforcement learning with curiosity-driven exploration. *Advances in Neural Information Processing Systems*, 34: 3757–3769.

Zhou, G.; Xu, Z.; Zhang, Z.; and Fan, G. 2023. SORA: Improving Multi-agent Cooperation with a Soft Role Assignment Mechanism. In *International Conference on Neural Information Processing*, 319–331. Springer.