Robust Multi-Agent Coordination via Evolutionary Generation of Auxiliary Adversarial Attackers*

Lei Yuan[1,2] †, Ziqian Zhang[1] †, Ke Xue[1], Hao Yin[1], Feng Chen[1], Cong Guan[1], Lihe Li[1], Chao Qian[1], Yang Yu[1,2]

[1]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
[2]Polixir Technologies, Nanjing 210000, China
{yuanl, xuek, yinh, guanc}@lamda.nju.edu.cn, {191240076, chenf, lilh}@smail.nju.edu.cn, {qianc, yuy}@nju.edu.cn

Abstract

Cooperative multi-agent reinforcement learning (CMARL) has shown to be promising for many real-world applications. Previous works mainly focus on improving coordination ability via solving MARL-specific challenges (e.g., non-stationarity, credit assignment, scalability), but ignore the policy perturbation issue when testing in a different environment. This issue hasn't been considered in problem formulation or efficient algorithm design. To address this issue, we firstly model the problem as a limited policy adversary Dec-POMDP (LPA-Dec-POMDP), where some coordinators from a team might accidentally and unpredictably encounter a limited number of malicious action attacks, but the regular coordinators still strive for the intended goal. Then, we propose **Ro**bust **M**ulti-**A**g**e**n**t** **C**oordination via **E**volutionary Generation of Auxiliary Adversarial Attackers (ROMANCE), which enables the trained policy to encounter diversified and strong auxiliary adversarial attacks during training, thus achieving high robustness under various policy perturbations. Concretely, to avoid the ego-system overfitting to a specific attacker, we maintain a set of attackers, which is optimized to guarantee the attackers high attacking quality and behavior diversity. The goal of quality is to minimize the ego-system coordination effect, and a novel diversity regularizer based on sparse action is applied to diversify the behaviors among attackers. The ego-system is then paired with a population of attackers selected from the maintained attacker set, and alternately trained against the constantly evolving attackers. Extensive experiments on multiple scenarios from SMAC indicate our ROMANCE provides comparable or better robustness and generalization ability than other baselines.

1 Introduction

Recently, cooperative multi-agent reinforcement learning (CMARL) has attracted extensive attention (Hernandez-Leal, Kartal, and Taylor 2019; Gronauer and Diepold 2022) and shows potential in numerous domains like autonomous vehicle teams (Peng et al. 2021), multi-agent path finding (Greshler et al. 2021), multi-UAV control (Yun et al. 2022), and dynamic algorithm configuration (Xue et al. 2022b). Existing CMARL methods mainly focus on solving specific challenges such as non-stationarity (Papoudakis et al. 2019), credit assignment (Wang et al. 2021a), and scalability (Christianos et al. 2021) to improve the coordination ability in complex scenarios. Either value-based methods (Sunehag et al. 2018; Rashid et al. 2018; Wang et al. 2021b) or policy-gradient-based methods (Foerster et al. 2018; Lowe et al. 2017; Yu et al. 2022) have demonstrated remarkable coordination ability in a wide range of tasks (e.g., SMAC (Samvelyan et al. 2019) and Hanabi (Yu et al. 2022)). Despite these successes, the mainstream CMARL methods are still difficult to be applied in real world, as they mainly consider training and testing policy in a nondistinctive environment. Thus, the policy learned by those methods may suffer from a performance decrease when encountering any disagreement between training and testing (Guo et al. 2022).

Training a robust policy before deployment plays a promising role for the mentioned problem and makes excellent progress in single-agent reinforcement learning (SARL) (Moos et al. 2022; Xu et al. 2022). Previous works typically employ an adversarial training paradigm to obtain a robust policy. These methods generally model the process of policy learning as a minimax problem from the perspective of game theory (Yu et al. 2021) and optimize the policy under the worst-case situation (Pinto et al. 2017; Zhang et al. 2020a; Zhang, Wang, and Boedecker 2022). Nevertheless, the multi-agent problem is much more complex (Zhang, Yang, and Başar 2021), as multiple agents are making decisions simultaneously in the environment. Also, recent works indicate that a MARL system is usually vulnerable to any attack (Guo et al. 2022). Some MARL works study the robustness from various aspects, including the uncertainty in local observation (Lin et al. 2020), model function (Zhang et al. 2020a), and message sending (Xue et al. 2022c). The mentioned methods either focus on investigating the robustness from different aspects, or apply techniques such as heuristic rules and regularizers used in SARL to train a robust coordination policy. However, how unpredictable malicious action attacks cause policy perturbation has not been fully explored in CMARL.

In this work, we aim to develop a robust CMARL framework when malicious action attacks on some coordinators

*Corresponding author: Yang Yu.
†These authors contributed equally.

from a team exist. Concretely, we model the problem as a limited policy adversary Dec-POMDP (LPA-Dec-POMDP), where some coordinators may suffer from malicious action attacks, while the regular coordinators should still try to complete the intended goal.

Towards developing such a robust policy, we propose ROMANCE, an adversarial training paradigm based on the evolutionary generation of auxiliary attackers. Specifically, we maintain a set of attackers with high attacking quality and behavior diversity among all generated attackers to avoid the ego-system overfitting to a specific attacker, where high attack quality requires the attacker to minimize the ego-system reward, and diversity refers to generating different behaviors among attackers. A sparse action regularizer is also introduced to promote behavior diversity for different attackers. Furthermore, to prevent the attackers from being too tricky for the ego-system to complete the intended mission, we limit the total number of attacks to a fixed value. For the training of the ego-system, we pair it with a population of attackers selected from the maintained set to complete the given mission, then iteratively select and update the attacker population under the customized quality score and diversity distance. Finally, we obtain a highly robust coordination policy under different types and degrees of action perturbations.

To evaluate the proposed methods, we conduct extensive experiments on multiple maps from SMAC (Samvelyan et al. 2019) and compare ROMANCE against multiple baselines. Empirical results demonstrate that our proposed adversarial training paradigm can indeed obtain attackers with high attack ability and diverse behaviors. Also, the coordination policy trained against the population can achieve high robustness and generalization effectiveness with alternative numbers of attacks during testing. Furthermore, visualization experiments indicate how ROMANCE improves robustness under malicious action perturbations.

2 Related Work

Multi-agent reinforcement learning (MARL) has made prominent progress these years (Hernandez-Leal, Kartal, and Taylor 2019; Gronauer and Diepold 2022). Many methods have emerged as efficient ways to promote coordination among agents, and most of them can be roughly divided into policy-based and value-based methods. MADDPG (Lowe et al. 2017), COMA (Foerster et al. 2018), DOP (Wang et al. 2021c), and MAPPO (Yu et al. 2022) are typical policy gradient-based methods that explore the optimization of multi-agent policy gradient methods. MADDPG applies the CTDE (Centralized Training Decentralized Execution) paradigm to train the policies and optimizes each policy via DDPG (Lillicrap et al. 2016). COMA also applies the centralized critic to optimize the policy but employs a counterfactual model to calculate the marginal contribution of each agent in the multi-agent system. DOP takes a forward step to apply a centralized linear mixing network to decompose the global reward in a cooperative system and shows performance improvement for MADDPG and COMA significantly. Recently, MAPPO applies the widely proven learning efficiency of proximal policy optimization technique in single-agent reinforcement learning into MARL. Another category of MARL approaches, value-based methods, mainly focus on the factorization of the value function. VDN (Sunehag et al. 2018) aims to decompose the team value function into agent-wise ones by a simple additive factorization. Following the Individual-Global-Max (IGM) principle (Son et al. 2019), QMIX (Rashid et al. 2018) improves the way of value function decomposition by learning a non-linear mixing network, which approximates a monotonic function value decomposition. QPLEX (Wang et al. 2021b) takes a duplex dueling network architecture to factorize the joint value function, which achieves a full expressiveness power of IGM. Wang et al. (2021a) recently give theoretical analysis of the IGM by applying a multi-agent fitted Q-iteration algorithm. More details and advances about MARL can be seen in reviews (Zhang, Yang, and Başar 2021; Canese et al. 2021; Zhu, Dastani, and Wang 2022).

Adversarial training plays a promising role for the RL robustness (Moos et al. 2022), which involves the perturbations occurring in different cases, such as state, reward, policy, etc. These methods then train the RL policy in an adversarial way to acquire a robust policy in the worst-case situation. Robust adversarial reinforcement learning (RARL) (Pinto et al. 2017) picks out specific robot joints that the adversary acts on to find an equilibrium of the minimax objective using an alternative learning adversary. RARARL (Pan et al. 2019) takes a further step by introducing risk-averse robust adversarial reinforcement learning to train a risk-averse protagonist and a risk-seeking adversary, this approach shows substantially fewer crashes compared to agents trained without an adversary on a self-driving vehicle controller. The mentioned methods only learn a single adversary, and this approach does not consistently yield robustness to dynamics variations under standard parametrizations of the adversary. RAP (Vinitsky et al. 2020) and GC (Song and Schneider 2022) then learn population-based augmentation to the Robust RL formulation. See (Ilahi et al. 2021; Moos et al. 2022) for detailed reviews, and (Smirnova, Dohmatob, and Mary 2019; Zhang et al. 2020a,b; Oikarinen et al. 2021; Xie et al. 2022) for some recent advances.

Robust MARL has attracted widespread attention recently (Guo et al. 2022). M3DDPG (Li et al. 2019) learns a minimax extension of MADDPG (Lowe et al. 2017) and trains the MARL policy in an adversarial way, which shows potential in solving the poor local optima caused by opponents' policy altering. In order to model the uncertainty caused by the inaccurate knowledge of the model, R-MADDPG (Zhang et al. 2020c) introduces the concept of robust Nash equilibrium, and treats the uncertainty as a natural agent, demonstrating high superiority when facing reward uncertainty. For the observation perturbation of CMARL, Lin et al. (2020) learn an adversarial observation policy to attack the system, showing that the ego-system is highly vulnerable to observational perturbations. RADAR (Phan et al. 2021) learns resilient MARL policy via adversarial value decomposition. Hu and Zhang (2022) further design an action regularizer to attack the CMARL system efficiently. Xue et al. (2022c) recently consider the multi-agent adversarial communication, learning robust communication policy when some message senders are poisoned. To our knowl-

edge, no previous work has explored CMARL under LPA-Dec-POMDP, neither in problem formulation nor efficient algorithm design.

Furthermore, some other works focus on the robustness when coordinating with different teammates, referring to ad-hoc teamwork (Stone et al. 2010; Gu et al. 2022; Mirsky et al. 2022), or zero-shot coordination (ZSC) (Hu et al. 2020; Lupu et al. 2021; Xue et al. 2022a). The former methods aim at creating an autonomous agent that can efficiently and robustly collaborate with previously unknown teammates on tasks to which they are all individually capable of contributing as team members. While in the ZSC setting, a special case of ad-hoc teamwork, agents work toward a common goal and share identical rewards at each step. The introduction of adversarial attacks makes the victim an unknown teammate with regard to regular agents, while it is even more challenging because the unknown teammate might execute destructive actions. Our proposed method takes a further step toward this direction for robust CMARL.

3 Problem Formulation

This paper considers a CMARL task under the framework of Dec-POMDP (Oliehoek and Amato 2016), which is defined as a tuple $\mathcal{M} = \langle \mathcal{N}, \mathcal{S}, \mathcal{A}, P, \Omega, O, R, \gamma \rangle$. Here $\mathcal{N} = \{1, \dots, n\}$ is the set of agents, $\mathcal{S}$ is the set of global states, $\mathcal{A} = \mathcal{A}^1 \times ... \times \mathcal{A}^n$ is the set of joint actions, Ω is the set of observations, and $\gamma \in [0, 1)$ represents the discounted factor. At each time step, agent i receives the observation $o^i = O(s, i)$ and outputs the action $a^i \in \mathcal{A}^i$. The joint action $\boldsymbol{a} = (a^1, ..., a^n)$ leads to the next state $s' \sim P(\cdot|s, \boldsymbol{a})$ and a global reward $R(s, \boldsymbol{a}, s')$. To relieve the partial observability, we encode the history $(o_i^1, a_i^1, \dots, o_i^{t-1}, a_i^{t-1}, o_i^t)$ of agent i until timestep t into τ_i, then with $\boldsymbol{\tau} = \langle \tau_1, \dots, \tau_n \rangle$, the formal objective is to find a joint policy $\boldsymbol{\pi}(\boldsymbol{\tau}, \boldsymbol{a})$ which maximizes the global value function $Q_{tot}^{\boldsymbol{\pi}}(\boldsymbol{\tau}, \boldsymbol{a}) = \mathbb{E}_{s, \boldsymbol{a}} \left[\sum_{t=0}^{\infty} \gamma^t R(s, \boldsymbol{a}) \mid s_0 = s, \boldsymbol{a_0} = \boldsymbol{a}, \boldsymbol{\pi} \right]$.

We aim to optimize a policy when some coordinators from a team suffer from policy perturbation. The vulnerability of CMARL makes it difficult to tolerate an unlimited number of perturbations. To avoid the ego-system from being entirely destroyed, we assume a limited number of perturbations and formulate such setting as a LPA-Dec-POMDP:

Definition 1 (Limited Policy Adversary Dec-POMDP) *Given a Dec-POMDP $\mathcal{M} = \langle \mathcal{N}, \mathcal{S}, \mathcal{A}, P, \Omega, O, R, \gamma \rangle$, we define a limited policy adversary Dec-POMDP (LPA-Dec-POMDP) $\hat{\mathcal{M}} = \langle \mathcal{N}, \mathcal{S}, \mathcal{A}, P, K, \Omega, O, R, \gamma \rangle$ by introducing an adversarial attacker $\pi_{adv} : \mathcal{S} \times \mathcal{A} \times \mathbb{N} \to \mathcal{A}$. The attacker perturbs the ego-agents' policy by forcing the agents to execute joint action $\hat{\boldsymbol{a}} \sim \pi_{adv}(\cdot|s, \boldsymbol{a}, k)$ such that $s' \sim P(\cdot|s, \hat{\boldsymbol{a}})$, $r = R(s, \hat{\boldsymbol{a}}, s')$. Where $K \in \mathbb{N}$ is the number of attacks that meets $\sum_t \sum_{i \in \mathcal{N}} \mathbb{I}(\hat{a}_t^i \neq a_t^i) \leq K$, and $k \leq K$ indicates the current remaining attack number.*

To efficiently address the attacking problem, we introduce a class of disentangled adversarial attacker policies by decomposing a policy into two components: victim selection and policy perturbation, in Def. 2.

Definition 2 (Disentangled Adversarial Attacker Policy) *For an adversarial attacker policy π_{adv}, if there exist a victim selection function $v : \mathcal{S} \times \mathbb{N} \to \Delta(\hat{\mathcal{N}})$ and a policy perturbation function $g : \hat{\mathcal{N}} \times \mathcal{A} \times \mathbb{N} \to \Delta(\mathcal{A})$, such that the following two equations hold:*

$$\pi_{adv}(\hat{\boldsymbol{a}}|s, \boldsymbol{a}, k) = v(i|s, k) g(\hat{\boldsymbol{a}}|i, \boldsymbol{a}, k)$$
$$g(\hat{\boldsymbol{a}}|i, \boldsymbol{a}, 0) = g(\hat{\boldsymbol{a}}|null, \boldsymbol{a}, k) = \mathbb{I}(\hat{\boldsymbol{a}} = \boldsymbol{a}),$$

where $\hat{\mathcal{N}} = \mathcal{N} \cup \{null\}$, $i \sim v(\cdot|s, k)$, then we say that v and g disentangle π_{adv}.

As for the policy perturbation function, many heuristic-based methods (Pattanaik et al. 2018; Tessler, Efroni, and Mannor 2019; Sun et al. 2021) have been proposed to find adversarial perturbations for a fixed RL policy. A common attacking way is to force the victim to select the action with the minimum Q-values at some steps (Pattanaik et al. 2018). Thus, for policy perturbation function g, an effective and efficient form could be $g(\hat{\boldsymbol{a}}|i, \boldsymbol{a}, k) = g(\hat{a}^i, \boldsymbol{a}^{-i}|i, \boldsymbol{a}, k) = \mathbb{I}(\hat{a}^i = \arg\min_{a^i} Q^i(\tau^i, a^i))$, if $k > 0$ and $i \neq null$. For efficiency, we only focus on disentangled attacker policy $\pi_{adv} = v \circ g$ with a heuristic-based policy perturbation in the rest of the paper, and, without loss of generality, we suppose g performs a deterministic perturbation with $\hat{\boldsymbol{a}} = g(i, \boldsymbol{a}, k)$.

4 Method

In this section, we will explain the design of ROMANCE, a novel framework to learn a robust coordination policy under the LPA-Dec-POMDP. We first discuss the optimization object of each attacker and then show how to maintain a set of high-quality solutions with diverse behaviors by specially designed update and selection mechanisms. Finally, we propose ROMANCE, an alternating training paradigm, to improve the robustness of CMARL agents (ego-system) under policy perturbations.

4.1 Attacker Optimization Objective

In this section, we discuss how to train a population of adversarial attacker policies $P_{adv} = \{\pi_{adv}^j\}_{j=1}^{n_p} = \{v^j \circ g\}_{j=1}^{n_p}$ under a fixed joint ego-system policy $\boldsymbol{\pi}$, where n_p is the population size. We anticipate the attacker population under the goal of high quality and diversity, where the quality objective requires it to minimize the ego-system's return and diversity encourages attackers to behave differently. To achieve the mentioned goal, we first show that an individual optimal adversarial attacker could be solved under a specific MDP setting and then discuss how to solve it.

Theorem 1 *Given an LPA-Dec-POMDP $\hat{\mathcal{M}} = \langle \mathcal{N}, \mathcal{S}, \mathcal{A}, P, K, \Omega, O, R, \gamma \rangle$, a fixed joint policy $\boldsymbol{\pi}$ of the ego-system and a heuristic-based policy perturbation function g, there exists an MDP $\bar{\mathcal{M}} = (\bar{\mathcal{S}}, \bar{\mathcal{A}}, \bar{P}, \bar{R}, \gamma)$ such that the optimal adversarial attacker π_{adv}^* for $\hat{\mathcal{M}}$ is disentangled by an optimal policy v^* of $\bar{\mathcal{M}}$ and g, where $\bar{\mathcal{S}} = \mathcal{S} \times \mathbb{N}$, $\bar{s} = (s, k), \bar{s}' = (s', k'), k, k' \leq K$ indicates the remaining attack budget, $\bar{\mathcal{A}} = \mathcal{N} \cup \{null\}$, $\bar{R}(\bar{s}, \bar{a}, \bar{s}') = -R(s, \hat{\boldsymbol{a}}, s')$,*

$$\bar{P}(\bar{s}'|\bar{s}, \bar{a}) = \begin{cases} 0 & k - k' \notin \{0, 1\} \\ P(s'|s, \hat{\boldsymbol{a}}) \mathbb{I}(\hat{\boldsymbol{a}} = \boldsymbol{a}) & k - k' = 0 \\ P(s'|s, \hat{\boldsymbol{a}}) \mathbb{I}(\hat{\boldsymbol{a}} \neq \boldsymbol{a}) & k - k' = 1 \end{cases},$$

where $\bar{d}(\bar{s}_0) = d(s_0)\mathbb{I}(k_0 = K)$, d and $\bar{d}$ are distributions over initial state in $\hat{\mathcal{M}}$ and $\bar{\mathcal{M}}$, respectively, and $\boldsymbol{a} = \boldsymbol{\pi}(s)$, $\hat{\boldsymbol{a}} = g(\bar{a}, \boldsymbol{\pi}(s), k)$ are original and forced action of ego-system, respectively.

The intuition behind Thm. 1 is that the adversarial attacker is aimed at minimizing the reward earned by the ego-system. Under a fixed heuristic-based policy perturbation function g, the adversarial attacker hopes to find an optimal victim policy v to help decide which agent and when to enforce the attack. The proof can be found in Appendix. The construction of $\bar{\mathcal{M}}$ makes it possible to apply off-the-shelf DRL algorithms to solve an optimal agent v of $\bar{\mathcal{M}}$, thus deriving the optimal adversarial attacker of the LPA-Dec-POMDP.

Notice that limited numbers of attack, which could only be executed K times, making the action sparse in both agent and time dimension. Such sparse action often plays a vital role in obtaining a high reward and thus would be exploited aggressively. However, the opportunities for taking sparse action are limited. If the attacker exhausts the opportunities at the very beginning, it will lead to sub-optimality. To guide the agent to take sparse action more cautiously, Sparsity Prior Regularized Q learning (SPRQ) (Pang et al. 2021) constructs a regularized MDP by assigning a low probability to sparse action based on a reference distribution and solves the task by proposing a regularized objective:

$$\max_{v} \mathbb{E}[\sum_{t=0}^{T} \gamma^t (\bar{R}_t - \lambda D_{KL}(v(\cdot|\bar{s}_t), p_{ref}(\cdot)))], \quad (1)$$

where $\lambda \in \mathbb{R}$, p_{ref} is the reference distribution which assigns a small probability δ to sparse action "attack" and $1-\delta$ to "not attack" (i.e., $p_{ref}(\cdot) = (\frac{\delta}{|\bar{\mathcal{N}}|}, ... \frac{\delta}{|\bar{\mathcal{N}}|}, 1-\delta)$ in this case), and D_{KL} is the Kullback-Leibler (KL) divergence. As claimed in Proposition 3.1 and 3.2 of (Pang et al. 2021), the regularized optimal policy can be obtained by:

$$v(\bar{a}|\bar{s}) = \frac{p_{ref}(\bar{a}) \exp\left(\frac{\bar{Q}_v(\bar{s},\bar{a})}{\lambda}\right)}{Z(\bar{s})}, \quad (2)$$

where $\bar{Q}_v(\bar{s}, \bar{a})$ is the regularized Q function under v, and $Z(\bar{s}) = \sum_{\bar{a} \in \bar{\mathcal{A}}} p_{ref}(\bar{a}) \exp \frac{\bar{Q}_v(\bar{s},\bar{a})}{\lambda}$ is the partition function.

Following the proposed regularized Bellman optimality operator, we parameterize Q-function with ϕ and the SPRQ loss function is thus defined as follows:

$$L_{opt}(\phi) = \mathbb{E}[(\bar{Q}_\phi(\bar{s}_t, \bar{a}_t) - y)^2], \quad (3)$$

where $y = \bar{r}_t + \gamma\lambda \log\left(\mathbb{E}_{\bar{a}'_{t+1} \sim p_{ref}(\cdot)}[\exp\left(\frac{\bar{Q}_{\phi^-}(\bar{s}_{t+1}, \bar{a}'_{t+1})}{\lambda}\right)]\right)$, and ϕ^- are the parameters of the target network.

Applying the above technique can lead to an attacker with high attack efficiency, nevertheless, only one attacker can easily overfit to a specific ego-system type, still leading to poor generalization ability over unseen situation. Inspired by the wildly proved ability of Population-Based Training (PBT) (Jaderberg et al. 2019), we introduce a diversity regularization objective to ensure the effective behavior diversity of the whole population. The divergence of action distribution under same states is a reliable proxy for measuring the diversity between policies. In this way, we use Jensen-Shannon Divergence (JSD) (Fuglede and Topsøe 2004) to reflect the diversity of the population $P_{adv}(\boldsymbol{\phi}) = \{\pi_{adv}^{\phi_j}\}_{j=1}^{n_p} = \{v^{\phi_j} \circ g\}_{j=1}^{n_p}$, which can be calculated as:

$$\text{JSD}(\{v^{\phi_j}(\cdot|\bar{s})\}_{j=1}^{n_p}) = \frac{1}{n_p} \sum_{j=1}^{n_p} D_{KL}(v^{\phi_j}(\cdot|\bar{s}), \bar{v}(\cdot|\bar{s})), \quad (4)$$

where $\bar{v}(\bar{a}|\bar{s}) = \frac{1}{n_p} v^{\phi_j}(\bar{a}|\bar{s})$ is the average policy. Then, for the population, the regularization objective is defined as:

$$L_{div}(\boldsymbol{\phi}) = \mathbb{E}_{\bar{s} \sim S_a}[\text{JSD}(\{v^{\phi_j}(\cdot|\bar{s})\}_{j=1}^{n_p})], \quad (5)$$

where $S_a = \bigcup_{j=1}^{n_p} S_a^j = \bigcup_{j=1}^{n_p} \{\bar{s} | k > 0, \bar{a} \neq null, \bar{a} \sim v^{\phi_j}(\cdot|\bar{s})\}$ is the union set of states (attack points) chosen to be attacked by adversarial attackers.

Considering the mentioned sparse attack and behavior diversity, our full loss function can be derived:

$$L_{adv}(\boldsymbol{\phi}) = \frac{1}{n_p} \sum_{j=1}^{n_p} L_{opt}(\phi_j) - \alpha L_{div}(\boldsymbol{\phi}), \quad (6)$$

where L_{opt} and L_{div} are defined in Eq. (3) and Eq. (5), respectively, and α is an adjustable hyper-parameter to control the balance between quality and behavior diversity.

4.2 Evolutionary Generation of Attackers

Despite the effectiveness of PBT with the objective in Eq. (6), the ego-system may overfit to some specific types of attackers in the current population and thus forget the attacking modes occurring in the early training stage. To avoid this catastrophic result, we attempt to further improve the coverage of adversary policy space.

Among different methods, Quality-Diversity (QD) algorithms (Cully and Demiris 2017; Chatzilygeroudis et al. 2021) can obtain a set of high-quality solutions with diverse behaviors efficiently, which have recently been used to discover diverse policies (Wang, Xue, and Qian 2022), generate environments (Bhatt et al. 2022) and partners (Xue et al. 2022a) in RL. As a specific type of evolutionary algorithms (Bäck 1996), QD algorithms usually maintain an archive (i.e., a set of solutions with high-performance and diverse behaviors generated so far) and simulate the natural evolution process with iterative update and selection.

Inspired by QD algorithms, we design specialized update and selection mechanisms to generate desired auxiliary adversarial attackers. We maintain an archive Arc_{adv} with the maximum size n_a, where each individual (i.e., attacker) is assigned its quality score and behavior. Specifically, given an ego-agent joint policy $\boldsymbol{\pi}$, the quality score of adversarial attacker $\pi_{adv}^{\phi_i}$ is defined as the attacker's discounted cumulative return:

$$\text{Quality}(\pi_{adv}^{\phi_i}) = \mathbb{E}_{\bar{\boldsymbol{\tau}}}[\sum_t \gamma^t \bar{R}(\bar{s}, \bar{a}) | \boldsymbol{\pi}], \quad (7)$$

where $\bar{\boldsymbol{\tau}}$ is trajectory of attacker $\pi_{adv}^{\phi_i}$, and $\bar{R}$ is defined in Thm. 1. To describe the behavior of $\pi_{adv}^{\phi_i}$, we calculate the distance between it and another attacker $\pi_{adv}^{\phi_j}$:

$$\text{Dist}(\pi_{adv}^{\phi_i}, \pi_{adv}^{\phi_j}) = \mathbb{E}_{\bar{s} \sim S_a^{i,j}}[\text{JSD}(v^{\phi_i}(\cdot|\bar{s}), v^{\phi_j}(\cdot|\bar{s}))], \quad (8)$$

where $S_a^{i,j} = S_a^i \cup S_a^j$ is the attack points set. For simplicity, we use Quality(i) and Dist(i, j) to denote the quality score of $\pi_{adv}^{\phi_i}$ and the distance between $\pi_{adv}^{\phi_i}$ and $\pi_{adv}^{\phi_j}$, respectively, and i, j are their indexes in the archive.

In each iteration, we use fitness-based selection (Blickle and Thiele 1996) according to their quality scores to select n_p adversarial attackers from the archive and interact with the ego-system. Next, we take the optimization step described in Eq. (6) as an implicit mutation operator and derive new attackers. The archive is then updated one by one by adding the newly generated attackers. We avoid adding attackers with similar behaviors to keep the archive diverse (Cully and Mouret 2013; Wang, Xue, and Qian 2022). That is, whenever we want to add a new attacker to the archive, we first choose the most similar attacker in the archive and calculate their behavior distance. If the distance exceeds a certain threshold, the new attacker will be added. Otherwise, we keep one at random. Note that if the current archive size exceeds the capacity n_a after adding the new attacker, the oldest one will be deleted. The full procedure of the mentioned process is shown in Algo. 1 in Appendix, and we refer to such an iteration as a generation.

4.3 Robustness Training Paradigm

After obtaining a set of attackers with high quality in attacking and high diversity in behaviors, we aim to learn a robust ego-system under the existing attackers. We first investigate LPA-Dec-POMDP with a fixed adversarial attacker π_{adv} and how ego-agent joint policy could be optimized.

Theorem 2 *Given an LPA-Dec-POMDP $\hat{\mathcal{M}} = \langle \mathcal{N}, \mathcal{S}, \mathcal{A}, P, K, \Omega, O, R, \gamma \rangle$, a fixed deterministic adversarial attacker policy π_{adv}, there exists a Dec-POMDP $\tilde{\mathcal{M}} = \langle \mathcal{N}, \tilde{\mathcal{S}}, \mathcal{A}, \tilde{P}, \Omega, \tilde{O}, \tilde{R}, \gamma \rangle$, such that the optimal policy of $\tilde{\mathcal{M}}$ is the optimal policy for $\hat{\mathcal{M}}$ given π_{adv}, where $\tilde{\mathcal{S}} = \mathcal{S} \times \mathbb{N}$, $\tilde{s} = (s, k)$, $\tilde{s}' = (s', k')$, $\tilde{d}(\tilde{s}_0) = d(s_0)\mathbb{I}(k_0 = K)$, $\tilde{O}(\tilde{s}, i) = O(s, i)$, $\tilde{R}(\tilde{s}, \boldsymbol{a}, \tilde{s}') = R(s, \hat{\boldsymbol{a}}, s')$,*

$$\tilde{P}(\tilde{s}'|\tilde{s}, \boldsymbol{a}) = \begin{cases} 0 & k - k' \notin \{0, 1\} \\ P(s'|s, \hat{\boldsymbol{a}}) & otherwise \end{cases},$$

where $\tilde{d}$ and d are distributions over initial state in $\tilde{M}$ and $\hat{\mathcal{M}}$, respectively, and $\hat{\mathbf{a}} = \pi_{adv}(s, \boldsymbol{a}, k)$ indicates the executed joint action of the ego-system.

Theorem 3 *Given $\hat{\mathcal{M}} = \langle \mathcal{N}, \mathcal{S}, \mathcal{A}, P, K, \Omega, O, R, \gamma \rangle$, a stochastic adversarial attacker policy π_{adv}, there exists a Dec-POMDP $\tilde{\mathcal{M}} = \langle \mathcal{N}, \tilde{\mathcal{S}}, \mathcal{A}, \tilde{P}, \Omega, \tilde{O}, \tilde{R}, \gamma \rangle$, such that $\forall \boldsymbol{\pi}$, we have $\tilde{V}_{\boldsymbol{\pi}}(\tilde{s}) \leq \hat{V}_{\boldsymbol{\pi} \circ \pi_{adv}}(s, k)$, where $\tilde{s} = (s, k)$, $\hat{V}_{\boldsymbol{\pi} \circ \pi_{adv}}(s, k)$ denotes the state value function in the original LPA-Dec-POMDP $\hat{\mathcal{M}}$, for $\forall s \in \mathcal{S}, \forall k \in \{0, 1, ..., K\}$.*

The intuition behind Thm. 2 is that the constructed Dec-POMDP $\tilde{M}$ is functionally identical to the LPA-Dec-POMDP given the fixed π_{adv}. This theorem unveils that LPA-Dec-POMDP can be viewed as a particular version of Dec-POMDP whose policy should be robust under the "under-attack" transition and reward function. The theorem is also easy to be extended to a population version where $\pi_{adv} \sim p(P_{adv})$, where p is some distribution over the population of adversarial attackers P_{adv}. Thm. 3 illustrates that, under the circumstance where π_{adv} is a stochastic policy, the value function in the new Dec-POMDP is the lower bound of the value function of the same joint policy in the original LPA-Dec-POMDP. Related proof can be found in Appendix. The theorem reveals that by optimizing the ego-system under the constructed Dec-POMDP $\tilde{\mathcal{M}}$, we can get a robust ego-system under the Limited Adversary Dec-POMDP $\hat{\mathcal{M}}$.

Accordingly, many CMARL algorithms can be applied. Specifically, we take QMIX (Rashid et al. 2018) as the problem solver, where there exists a Q network $Q_i(\tau^i, a^i)$ for each agent i and a mixing network that takes each Q value along with the global state as input and produces the value of Q_{tot}. Under the Dec-POMDP $\tilde{\mathcal{M}}$ proposed in Thm. 3, we parameterize QMIX with θ and train it through minimizing:

$$L_{ego}(\theta) = \mathbb{E}[(Q_{tot}(\tilde{\boldsymbol{\tau}}, \boldsymbol{a}, \tilde{s}; \theta) - y_{tot})^2], \tag{9}$$

where $y_{tot} = \tilde{r} + \gamma \max_{\boldsymbol{a}'} Q_{tot}(\tilde{\boldsymbol{\tau}}', \boldsymbol{a}', \tilde{s}'; \theta^-)$, and θ^- are parameters of a periodically updated target network.

In our ROMANCE framework, we select a population of adversarial attackers from the archive, alternatively optimize the adversarial attackers or ego-system by fixing the other, and update the archive accordingly. The full algorithm of our ROMANCE can be seen in Algo. 2 in Appendix.

5 Experiments

In this section, we conduct extensive experiments to answer the following questions: 1) Can ROMANCE[1] achieve high robustness compared to other baselines in different scenarios? 2) Can ROMANCE obtain a set of attackers with high attacking quality and diversity? 3) Can ROMANCE be integrated into multiple CMARL methods, and how does each hyperparameter influence the performance of ROMANCE?

We conduct experiments on SMAC (Samvelyan et al. 2019), a widely used combat scenario of StarCraft II unit micromanagement tasks, where we train the ally units to beat enemy units controlled by the built-in AI with an unknown strategy. At each timestep, agents can move or attack any enemies and receive a global reward equal to the total damage done to enemy units. Here we consider multiple maps include maps 2s3z, 3m, 3s_vs_3z, 8m, MMM, and 1c3s5z. The detailed descriptions are presented in Appendix.

To ensure fair evaluation, we carry out all the experiments with five random seeds, and the results are presented with a 95% confidence interval. Detailed network architecture, hyperparameter setting of ROMANCE are shown in Appendix.

5.1 Competitive Results and Analysis

We implement ROMANCE based on QMIX (Rashid et al. 2018) for its widely proven coordination ability. Then, ROMANCE is compared against four baselines: the vanilla QMIX, which is obtained to complete the task without any adversarial training, RANDOM, which adds random

[1]Code is available at https://github.com/zzq-bot/ROMANCE

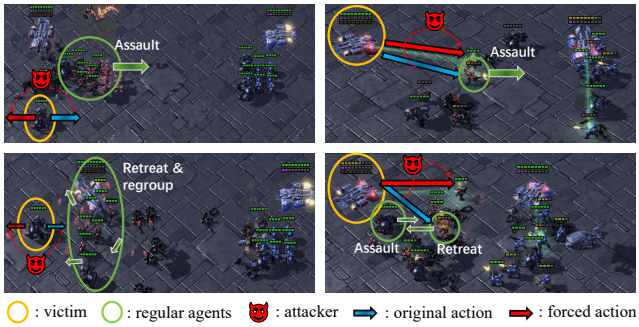


Figure 1: Visualization of action selection under unseen action attack. The first and second rows show the policy learned without and with ROMANCE, respectively.

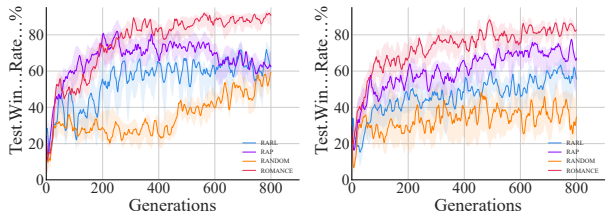
attack during training, and two strong baselines named RARL (Pinto et al. 2017) and RAP (Vinitsky et al. 2020).

RARL (Pinto et al. 2017) trains a robust agent in the presence of an adversarial attacker who applies disturbance to the system. Iteratively, the attacker learns to be an optimal destabilization policy, and then the agent learns to fulfill the original task while being robust to the adversarial attacker.

RAP (Vinitsky et al. 2020) extends RARL by introducing population based training. At each rollout, it samples an attacker uniformly from the population and trains the agent and attackers iteratively as RARL does. The introduction of population improves the generalization by forcing the agent to be robust to a wide range of attackers, thus avoiding being exploited by some specific attackers.

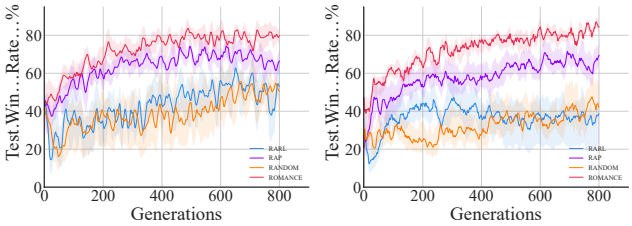
Robustness Visualization At first glance, we conduct experiments on map MMM to investigate how the training framework influences the coordination policy behavior under unpredictable attacks. As shown in Fig. 1, when one coordinator from a well-trained coordination policy suffers from an action attack, the policy without adversarial training will still take the original coordination pattern but ignore the emergent situation, resulting in a sub-optimal policy. As seen at the beginning of the battle, the survivors learned by vanilla QMIX still try to assault but ignore the attacked Marauder when it is drawn away by a malicious attack, causing severe damage to the team. Still, our ROMANCE can obtain a policy where the survivors retreat and wait for the Marauder to regroup for efficient coordination. At the middle stage of the battle, when some Marines are close to death but cannot get healed because the Medivac's action is being attacked, the survivors with full health learned by ROMANCE will charge forward to replace the dying Marine and cover him to let him get healed, while policy learned by vanilla QMIX still ramble in the map but ignore the teammates.

Training phase evaluation Fig. 2 shows the learning curves of different methods of maps 3s_vs_3z and 2s3z when facing fixed unknown attackers. ROMANCE outperforms all baselines in both maps at each generation either in terms of convergence speed or asymptotic performance. RANDOM achieves the worst performance in almost every generation, indicating that adding random policy perturbation can somehow improve the exploration ability under an environment without attack but has no effect on tasks where attackers exist. The superiority of ROMANCE over RARL and RAP demonstrates the necessity of adversarial population and diverse population training, respectively. In Fig. 3, we present the learning curves of different methods implemented on QPLEX and VDN on map 2s3z. The curves show that ROMANCE can significantly enhance the robustness of value-based MARL algorithms when they are integrated.



(a) 3s_vs_3z (b) 2s3z

Figure 2: Averaged test win rates on two maps.



(a) VDN (b) QPLEX

Figure 3: Average test win rates of VDN and QPLEX on map 2s3z during the training phase.

Robustness Comparison We here come to show whether ROMANCE can improve the coordination ability under different unpredictable attacks compared with multiple baselines. As shown in Tab. 1, we present three settings, where "Natural" means no attackers or the attack number $K = 0$ during testing, "Random Attack" indicates that every agent in the ego-system might be attacked randomly at each step, "EGA (Evolutionary Generation based Attackers)" are unseen attackers with high performance and diversity generated by our method in multiple extra runs, which could be viewed as out-of-distribution attackers for different methods. In the "Natural" setting, ROMANCE achieves comparable or better performance compared to other baselines. RARL achieves inferiority over other baselines because it aims to learn a worst-case performance, leading to a pessimistic result for the coordination ability. RAP and RANDOM show superiority over the vanilla QMIX in some maps, such as 2s3z. We believe this is because random attacks or a weak adversarial population during training can promote exploration for MARL under natural circumstances. Furthermore, when suffering from a random attack during testing (i.e., the "Random Attack" setting), vanilla

Method	Map_Name	2s3z $K=8$	3m $K=4$	3s_vs_3z $K=8$	8m $K=5$	MMM $K=8$	1c3s5z $K=6$	$+/-/\approx$
Natural	vanilla QMIX	92.8 ± 1.62	$\mathbf{97.9 \pm 1.02}$	98.3 ± 0.78	98.2 ± 0.45	95.8 ± 1.59	88.8 ± 2.13	$1/1/4$
	RARL	96.4 ± 1.19	86.0 ± 5.38	80.6 ± 27.5	95.3 ± 3.31	89.3 ± 7.01	76.9 ± 9.85	$0/4/2$
	RAP	$\mathbf{98.1 \pm 0.76}$	91.3 ± 4.93	$\mathbf{99.3 \pm 0.51}$	91.7 ± 7.96	95.3 ± 4.98	86.7 ± 10.5	$0/1/5$
	RANDOM	98.0 ± 0.60	95.3 ± 2.07	99.6 ± 0.35	$\mathbf{98.6 \pm 0.90}$	93.8 ± 7.56	93.1 ± 4.41	$1/0/5$
	ROMANCE	97.9 ± 1.34	96.0 ± 1.83	97.8 ± 1.78	94.3 ± 3.94	$\mathbf{97.1 \pm 1.49}$	$\mathbf{93.9 \pm 1.24}$	
Random Attack	vanilla QMIX	78.8 ± 1.28	$\mathbf{78.7 \pm 1.49}$	87.0 ± 0.36	66.2 ± 2.08	70.0 ± 3.97	66.6 ± 2.03	$0/5/1$
	RARL	84.3 ± 2.40	67.6 ± 5.01	70.1 ± 29.1	75.7 ± 7.00	62.2 ± 10.2	56.5 ± 10.8	$0/5/1$
	RAP	87.3 ± 1.87	73.5 ± 3.49	89.8 ± 4.81	$\mathbf{78.4 \pm 8.22}$	84.2 ± 9.05	66.8 ± 9.66	$0/1/5$
	RANDOM	83.9 ± 6.38	76.4 ± 2.27	91.9 ± 1.32	72.0 ± 3.46	72.9 ± 7.09	60.5 ± 21.3	$0/2/4$
	ROMANCE	$\mathbf{89.1 \pm 1.97}$	78.1 ± 5.13	$\mathbf{93.0 \pm 1.82}$	76.2 ± 5.36	$\mathbf{85.8 \pm 8.66}$	$\mathbf{77.9 \pm 1.96}$	
EGA	vanilla QMIX	26.7 ± 4.28	20.7 ± 2.13	30.9 ± 1.52	42.7 ± 9.79	37.9 ± 3.13	35.2 ± 8.66	$0/6/0$
	RARL	56.1 ± 11.8	86.1 ± 0.98	60.9 ± 14.2	66.3 ± 7.25	41.5 ± 11.6	35.3 ± 4.00	$0/6/0$
	RAP	64.1 ± 11.9	84.0 ± 4.27	65.1 ± 4.41	84.4 ± 8.88	74.9 ± 15.5	45.4 ± 6.83	$0/4/2$
	RANDOM	48.3 ± 17.3	66.2 ± 16.6	54.4 ± 7.83	55.6 ± 12.5	53.1 ± 6.09	43.3 ± 10.3	$0/6/0$
	ROMANCE	$\mathbf{81.6 \pm 0.84}$	$\mathbf{89.7 \pm 1.52}$	$\mathbf{90.5 \pm 1.97}$	$\mathbf{86.2 \pm 5.11}$	$\mathbf{84.0 \pm 11.5}$	$\mathbf{66.5 \pm 3.24}$	

Table 1: Average test win rates of different methods under various attack settings, where K is the number of attacks during training, "Natural" means no attack during testing, "Random Attack" indicates every agent in the ego-system may be attacked randomly, and "EGA" means our evolutionary generation based attackers. The best result of each column is highlighted in bold. The symbols '$+$', '$-$' and '$\approx$' indicate that the result is significantly superior to, inferior to, and almost equivalent to ROMANCE, respectively, according to the Wilcoxon rank-sum test (Mann and Whitney 1947) with confidence level 0.05.

Method	$K=6$	$K=7$	$K=8$	$K=9$	$K=10$	$K=11$	$K=12$	$K=14$
vanilla QMIX	59.2 ± 2.66	42.1 ± 0.81	26.7 ± 4.28	17.3 ± 0.62	12.2 ± 0.33	8.74 ± 0.14	6.42 ± 0.84	2.82 ± 0.70
RARL	72.7 ± 4.22	65.2 ± 9.11	56.1 ± 11.8	46.3 ± 12.1	38.0 ± 13.5	31.8 ± 13.1	25.9 ± 12.3	18.6 ± 10.9
RAP	81.7 ± 7.37	73.6 ± 7.46	64.1 ± 11.9	53.5 ± 11.7	42.5 ± 11.6	33.9 ± 11.4	25.8 ± 10.7	14.0 ± 7.50
RANDOM	69.3 ± 10.9	56.8 ± 12.8	48.3 ± 17.3	34.7 ± 17.3	25.5 ± 15.8	19.8 ± 14.3	14.9 ± 12.6	10.0 ± 9.69
ROMANCE	$\mathbf{89.9 \pm 1.19}$	$\mathbf{86.4 \pm 1.87}$	$\mathbf{81.6 \pm 0.84}$	$\mathbf{75.1 \pm 0.58}$	$\mathbf{66.7 \pm 1.56}$	$\mathbf{57.4 \pm 1.61}$	$\mathbf{48.6 \pm 2.60}$	$\mathbf{41.5 \pm 2.17}$

Table 2: Average test win rates of each method when the test number of attacks K changes on map 2s3z. The best result of each column is highlighted in bold, and the column for the training number (i.e., $K=8$) is highlighted as gray.

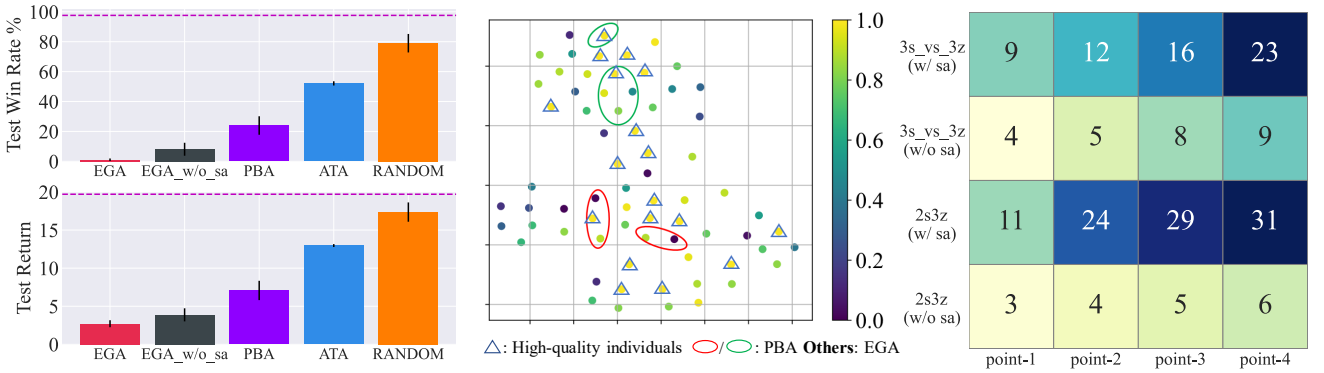
QMIX has the most remarkable performance decrease in most maps, demonstrating the multi-agent coordination policy's vulnerability without any adversarial training. Methods based on adversarial training such as RANDOM, RARL, and RAP show superiority over vanilla QMIX, indicating that adversarial training can improve the robustness of MARL policy. We further find that when encountering strong attackers (i.e., the "EGA" setting), all baselines sustain a severe performance decrease. The proposed ROMANCE achieves a high superiority over other baselines on most maps under different attack modes, indicating it can indeed learn a robust coordination policy under different policy perturbation conditions.

Beyond-limited-budget evaluation As this study considers a setting where the number of attacks is fixed during the training phase, we evaluate the generalization ability when altering the attack budget during testing. We conduct experiments on map 2s3z with the number of attacks $K=8$ during training. As shown in Tab. 2, when the budget is different from the training phase, policy learned with vanilla QMIX and RANDOM sustain a severe performance decrease even when the budget slightly changes, indicating that these two methods may overfit to the training situation and lack of generalization. RARL and RAP show superiority over vanilla QMIX and RANDOM, demonstrating that adversarial training can relieve the overfitting problem but is still inferior to ROMANCE along with the budget increasing, manifesting the high generalization ability gained by the training paradigm of ROMANCE.

5.2 Attacker Population Validation

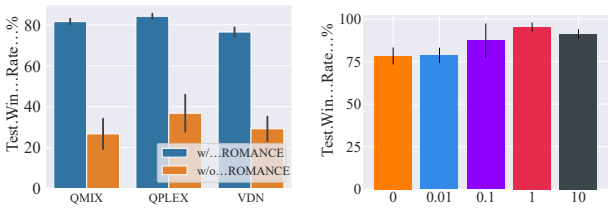
As our method needs to maintain a set of adversarial attackers, we design experiments to investigate the attackers generated by our method. As shown in Fig. 4(a), the multiple attack methods can correspondingly decrease the ego-system's performance, and our EGA (Evolutionary Generation based Attackers) show high superiority over others both in return and test win rate, demonstrating the effectiveness of the proposed training paradigm. EGA, EGA_w/o_sa, and PBA (Population-Based Attackers) outperform ATA (Alternating Training Attackers) and RANDOM, indicating that population training can indeed improve the attacking ability. Nevertheless, PBA works inefficiently, showing that only randomly initialized individuals in the population are insufficient for efficient population training. The superiority of EGA over its ablation EGA_w/o_sa demonstrates the effectiveness of sparse action regularizer.

Furthermore, we analyze the behavior representations learned by each attacking method in a two-dimensional plane using the t-SNE method (Van der Maaten and Hinton 2008). As shown in Fig. 4(b), we can discover that the traditional population-based training paradigm can only generate attackers with very limited diversity and quality, with most points gathering around in a few specific areas, while



(a) Attack quality on map 3m (b) Behavior diversity of the population (c) Sparse attack learned by ROMANCE

Figure 4: Attacker quality validation, where EGA and EGA_w/o_sa are our Evolutionary Generation of Attackers with and without sparse action regularizer, respectively; PBA and ATA refer to Population-Based Attackers and Alternate Training Attackers, respectively; RANDOM means that we select a coordinator to attack randomly; (a) The attacking quality. (b) The t-SNE projection of attackers on map 3m. (c) Attack points produced on two maps with and without sparse action regularizer.



(a) Integrative Abilities (b) Sensibility of α

Figure 5: More experiments about ROMANCE.

our EGA can find widespread points with both high quality and diversity. Fig. 4(c) shows that our sparse action regularizer can efficiently promote the sparse attacking points to disperse as much as possible within one episode, which can also implicitly promote the diversity of the attackers in a population by preventing attackers from exhausting the attack opportunities at the very beginning.

5.3 Integrative and Parameter Sensitive Studies

ROMANCE is agnostic to specific value decomposition MARL methods. We can regard it as a plug-in model and integrate it with existing MARL value decomposition methods like QPLEX (Wang et al. 2021b), QMIX (Rashid et al. 2018), and VDN (Sunehag et al. 2018). As shown in Fig. 5(a), when integrating with ROMANCE, the performance of the baselines vastly improves on map 2s3z, indicating that the proposed training paradigm can significantly enhance robustness for these value-based MARL methods.

One of the crucial elements of our framework is the hyperparameter α which controls the attack quality and diversity. We here conduct experiments to study the sensibility of α in Eq. (6) for the whole framework. As shown in Fig. 5(b), on map 3s_vs_3z, the performance is more influenced when we set $\alpha = 0$ or 0.01, which refers to optimizing the attacking quality only but almost ignoring the diversity goal, indicating the importance of behavior diversity for the population training. Nevertheless, other choices have a negligible impact on the performance. We believe this is because the customized selection and update operators of the evolution mechanism can help balance the two goals for a slightly larger α. More experimental results, such as how each parameter influences ROMANCE, are shown in Appendix.

6 Conclusion

This paper considers the robust cooperative MARL problem, where some coordinators suffer from unpredictable policy perturbation. We first formalize this problem as an LPA-Dec-POMDP, where some coordinators from a team may sustain action perturbation accidentally and unpredictably. We then propose ROMANCE, an efficient approach to learn robust multi-agent coordination via evolutionary generation of auxiliary adversarial attackers. Experimental results on robustness and generalization testing verify the effectiveness of ROMANCE, and more analysis results also confirm it from multiple aspects. As our method aims to learn a disentangled adversarial attacker policy, which demands a heuristic-based policy perturbation function, future work on more reasonable and efficient ways such as observation perturbation and automatic search for the best budget for different tasks would be of great value. Furthermore, how to design efficient and effective robust multi-agent reinforcement learning algorithms for the open-environment setting (Zhou 2022) is also valuable for the MARL community.

Acknowledgments

This work is supported by National Key Research and Development Program of China (2020AAA0107200), the National Science Foundation of China (61921006, 62022039), and the program B for Outstanding Ph.D. candidate of Nanjing University. We would like to thank Jingcheng Pang, Chengxing Jia, and the anonymous reviewers for their helpful discussions and support.

References

Bäck, T. 1996. *Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithms*. Oxford University Press.

Bhatt, V.; Tjanaka, B.; Fontaine, M. C.; and Nikolaidis, S. 2022. Deep surrogate assisted generation of environments. In *NeurIPS*.

Blickle, T.; and Thiele, L. 1996. A comparison of selection schemes used in evolutionary algorithms. *Evolutionary Computation*, 4(4): 361–394.

Canese, L.; Cardarilli, G. C.; Di Nunzio, L.; Fazzolari, R.; Giardino, D.; Re, M.; and Spanò, S. 2021. Multi-agent reinforcement learning: A review of challenges and applications. *Applied Sciences*, 11(11): 4948.

Chatzilygeroudis, K.; Cully, A.; Vassiliades, V.; and Mouret, J.-B. 2021. Quality-Diversity optimization: A novel branch of stochastic optimization. In *Black Box Optimization, Machine Learning, and No-Free Lunch Theorems*, 109–135. Springer.

Christianos, F.; Papoudakis, G.; Rahman, M. A.; and Albrecht, S. V. 2021. Scaling multi-agent reinforcement learning with selective parameter sharing. In *ICML*, 1989–1998.

Cully, A.; and Demiris, Y. 2017. Quality and diversity optimization: A unifying modular framework. *IEEE Transactions on Evolutionary Computation*, 22(2): 245–259.

Cully, A.; and Mouret, J.-B. 2013. Behavioral repertoire learning in robotics. In *GECCO*, 175–182.

Foerster, J.; Farquhar, G.; Afouras, T.; Nardelli, N.; and Whiteson, S. 2018. Counterfactual multi-agent policy gradients. In *AAAI*, 2974–2982.

Fuglede, B.; and Topsøe, F. 2004. Jensen-Shannon divergence and Hilbert space embedding. In *ISIT*.

Greshler, N.; Gordon, O.; Salzman, O.; and Shimkin, N. 2021. Cooperative multi-agent path finding: Beyond path planning and collision avoidance. In *MRS*, 20–28.

Gronauer, S.; and Diepold, K. 2022. Multi-agent deep reinforcement learning: A survey. *Artificial Intelligence Review*, 55(2): 895–943.

Gu, P.; Zhao, M.; Hao, J.; and An, B. 2022. Online ad hoc teamwork under partial observability. In *ICLR*.

Guo, J.; Chen, Y.; Hao, Y.; Yin, Z.; Yu, Y.; and Li, S. 2022. Towards comprehensive testing on the robustness of cooperative multi-agent reinforcement learning. *arXiv:2204.07932*.

Hernandez-Leal, P.; Kartal, B.; and Taylor, M. E. 2019. A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 33(6): 750–797.

Hu, H.; Lerer, A.; Peysakhovich, A.; and Foerster, J. N. 2020. "Other-Play" for zero-shot coordination. In *ICML*, 4399–4410.

Hu, Y.; and Zhang, Z. 2022. Sparse adversarial attack in multi-agent reinforcement learning. *arXiv:2205.09362*.

Ilahi, I.; Usama, M.; Qadir, J.; Janjua, M. U.; Al-Fuqaha, A.; Hoang, D. T.; and Niyato, D. 2021. Challenges and countermeasures for adversarial attacks on deep reinforcement learning. *IEEE Transactions on Artificial Intelligence*, 3(2): 90–109.

Jaderberg, M.; Czarnecki, W. M.; Dunning, I.; Marris, L.; Lever, G.; Castaneda, A. G.; Beattie, C.; Rabinowitz, N. C.; Morcos, A. S.; Ruderman, A.; et al. 2019. Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science*, 364(6443): 859–865.

Li, S.; Wu, Y.; Cui, X.; Dong, H.; Fang, F.; and Russell, S. 2019. Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient. In *AAAI*, 4213–4220.

Lillicrap, T. P.; Hunt, J. J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; and Wierstra, D. 2016. Continuous control with deep reinforcement learning. In *ICLR*.

Lin, J.; Dzeparoska, K.; Zhang, S. Q.; Leon-Garcia, A.; and Papernot, N. 2020. On the robustness of cooperative multi-agent reinforcement learning. In *SPW*, 62–68.

Lowe, R.; Wu, Y.; Tamar, A.; Abbeel, J. H. P.; and Mordatch, I. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. In *NeurIPS*, 6379–6390.

Lupu, A.; Cui, B.; Hu, H.; and Foerster, J. N. 2021. Trajectory diversity for zero-shot coordination. In *ICML*, 7204–7213.

Mann, H. B.; and Whitney, D. R. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 50–60.

Mirsky, R.; Carlucho, I.; Rahman, A.; Fosong, E.; Macke, W.; Sridharan, M.; Stone, P.; and Albrecht, S. V. 2022. A survey of ad hoc teamwork: Definitions, methods, and open problems. *arXiv:2202.10450*.

Moos, J.; Hansel, K.; Abdulsamad, H.; Stark, S.; Clever, D.; and Peters, J. 2022. Robust reinforcement learning: A review of foundations and recent advances. *Machine Learning and Knowledge Extraction*, 4(1): 276–315.

Oikarinen, T.; Zhang, W.; Megretski, A.; Daniel, L.; and Weng, T.-W. 2021. Robust deep reinforcement learning through adversarial loss. In *NeurIPS*, 26156–26167.

Oliehoek, F. A.; and Amato, C. 2016. *A Concise Introduction to Decentralized POMDPs*. Springer.

Pan, X.; Seita, D.; Gao, Y.; and Canny, J. 2019. Risk averse robust adversarial reinforcement learning. In *ICRA*, 8522–8528.

Pang, J.-C.; Xu, T.; Jiang, S.-Y.; Liu, Y.-R.; and Yu, Y. 2021. Sparsity prior regularized Q-learning for sparse action tasks. *arXiv:2105.08666*.

Papoudakis, G.; Christianos, F.; Rahman, A.; and Albrecht, S. V. 2019. Dealing with non-stationarity in multi-agent deep reinforcement learning. *arXiv:1906.04737*.

Pattanaik, A.; Tang, Z.; Liu, S.; Bommannan, G.; and Chowdhary, G. 2018. Robust deep reinforcement learning with adversarial attacks. In *AAMAS*, 2040–2042.

Peng, Z.; Li, Q.; Hui, K. M.; Liu, C.; and Zhou, B. 2021. Learning to simulate self-driven particles system with coordinated policy optimization. In *NeurIPS*, 10784–10797.

Phan, T.; Belzner, L.; Gabor, T.; Sedlmeier, A.; Ritz, F.; and Linnhoff-Popien, C. 2021. Resilient multi-agent reinforcement learning with adversarial value decomposition. In *AAAI*, 11308–11316.

Pinto, L.; Davidson, J.; Sukthankar, R.; and Gupta, A. 2017. Robust adversarial reinforcement learning. In *ICML*, 2817–2826.

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 *ICML*, 4295–4304.

Samvelyan, M.; Rashid, T.; Schroeder de Witt, C.; Farquhar, G.; Nardelli, N.; Rudner, T. G.; Hung, C.-M.; Torr, P. H.; Foerster, J.; and Whiteson, S. 2019. The StarCraft multi-agent challenge. In *AAMAS*, 2186–2188.

Smirnova, E.; Dohmatob, E.; and Mary, J. 2019. Distributionally robust reinforcement learning. *arXiv:1902.08708*.

Son, K.; Kim, D.; Kang, W. J.; Hostallero, D. E.; and Yi, Y. 2019. Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *ICML*, 5887–5896.

Song, Y.; and Schneider, J. 2022. Robust reinforcement learning via genetic curriculum. In *ICRA*, 5560–5566.

Stone, P.; Kaminka, G. A.; Kraus, S.; and Rosenschein, J. S. 2010. Ad hoc autonomous agent teams: Collaboration without pre-coordination. In *AAAI*.

Sun, Y.; Zheng, R.; Liang, Y.; and Huang, F. 2021. Who Is the strongest enemy? Towards optimal and efficient evasion attacks in deep RL. In *ICLR*.

Sunehag, P.; Lever, G.; Gruslys, A.; Czarnecki, W. M.; Zambaldi, V. F.; Jaderberg, M.; Lanctot, M.; Sonnerat, N.; Leibo, J. Z.; Tuyls, K.; and Graepel, T. 2018. Value-decomposition networks for cooperative multi-agent learning based on team reward. In *AAMAS*, 2085–2087.

Tessler, C.; Efroni, Y.; and Mannor, S. 2019. Action robust reinforcement learning and applications in continuous control. In *ICML*, 6215–6224.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11): 2579–2605.

Vinitsky, E.; Du, Y.; Parvate, K.; Jang, K.; Abbeel, P.; and Bayen, A. 2020. Robust reinforcement learning using adversarial populations. *arXiv:2008.01825*.

Wang, J.; Ren, Z.; Han, B.; Ye, J.; and Zhang, C. 2021a. Towards understanding cooperative multi-agent Q-learning with value factorization. In *NeurIPS*, 29142–29155.

Wang, J.; Ren, Z.; Liu, T.; Yu, Y.; and Zhang, C. 2021b. QPLEX: Duplex dueling multi-agent Q-learning. In *ICLR*.

Wang, Y.; Han, B.; Wang, T.; Dong, H.; and Zhang, C. 2021c. DOP: Off-policy multi-agent decomposed policy gradients. In *ICLR*.

Wang, Y.; Xue, K.; and Qian, C. 2022. Evolutionary diversity optimization with clustering-based selection for reinforcement learning. In *ICLR*.

Xie, A.; Sodhani, S.; Finn, C.; Pineau, J.; and Zhang, A. 2022. Robust policy learning over multiple uncertainty sets. *arXiv:2202.07013*.

Xu, M.; Liu, Z.; Huang, P.; Ding, W.; Cen, Z.; Li, B.; and Zhao, D. 2022. Trustworthy reinforcement learning against intrinsic vulnerabilities: Robustness, safety, and generalizability. *arXiv:2209.08025*.

Xue, K.; Wang, Y.; Yuan, L.; Guan, C.; Qian, C.; and Yu, Y. 2022a. Heterogeneous multi-agent zero-shot coordination by coevolution. *arXiv:2208.04957*.

Xue, K.; Xu, J.; Yuan, L.; Li, M.; Qian, C.; Zhang, Z.; and Yu, Y. 2022b. Multi-agent dynamic algorithm configuration. In *NeurIPS*.

Xue, W.; Qiu, W.; An, B.; Rabinovich, Z.; Obraztsova, S.; and Yeo, C. K. 2022c. Mis-spoke or mis-lead: Achieving robustness in multi-agent communicative reinforcement learning. In *AAMAS*, 1418–1426.

Yu, C.; Velu, A.; Vinitsky, E.; Gao, J.; Wang, Y.; Bayen, A.; and Wu, Y. 2022. The surprising effectiveness of PPO in cooperative multi-agent games. In *NeurIPS*.

Yu, J.; Gehring, C.; Schäfer, F.; and Anandkumar, A. 2021. Robust reinforcement learning: A constrained game-theoretic approach. In *L4DC*, 1242–1254.

Yun, W. J.; Park, S.; Kim, J.; Shin, M.; Jung, S.; Mohaisen, A.; and Kim, J.-H. 2022. Cooperative multi-agent deep reinforcement learning for reliable surveillance via autonomous multi-UAV control. *IEEE Transactions on Industrial Informatics*.

Zhang, H.; Chen, H.; Boning, D. S.; and Hsieh, C.-J. 2020a. Robust reinforcement learning on state observations with learned optimal adversary. In *ICLR*.

Zhang, H.; Chen, H.; Xiao, C.; Li, B.; Liu, M.; Boning, D. S.; and Hsieh, C. 2020b. Robust deep reinforcement learning against adversarial perturbations on state observations. In *NeurIPS*.

Zhang, K.; Sun, T.; Tao, Y.; Genc, S.; Mallya, S.; and Basar, T. 2020c. Robust multi-agent reinforcement learning with model uncertainty. In *NeurIPS*, 10571–10583.

Zhang, K.; Yang, Z.; and Başar, T. 2021. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, 321–384.

Zhang, Y.; Wang, J.; and Boedecker, J. 2022. Robust reinforcement learning in continuous control tasks with uncertainty set regularization. *arXiv:2207.02016*.

Zhou, Z.-H. 2022. Open-environment machine learning. *National Science Review*, 9(8).

Zhu, C.; Dastani, M.; and Wang, S. 2022. A survey of multi-agent reinforcement learning with communication. *arXiv:2203.08975*.