SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

April 2024, Vol. 67, Iss. 4, 142102:1–142102:15
https://doi.org/10.1007/s11432-023-3853-y

Robust cooperative multi-agent reinforcement learning via multi-view message certification

Lei YUAN[1,2], Tao JIANG[1], Lihe LI[1], Feng CHEN[1],
Zongzhang ZHANG[1] & Yang YU[1,2*]

[1] *National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing* 210023*, China;*
[2] *Polixir Technologies, Nanjing* 211106*, China*

Received 10 January 2023/Revised 22 April 2023/Accepted 15 August 2023/Published online 22 March 2024

Abstract Many multi-agent scenarios require message sharing among agents to promote coordination, hastening the robustness of multi-agent communication when policies are deployed in a message perturbation environment. Major relevant studies tackle this issue under specific assumptions, like a limited number of message channels would sustain perturbations, limiting the efficiency in complex scenarios. In this paper, we take a further step in addressing this issue by learning a robust cooperative multi-agent reinforcement learning via multi-view message certification, dubbed CroMAC. Agents trained under CroMAC can obtain guaranteed lower bounds on state-action values to identify and choose the optimal action under a worst-case deviation when the received messages are perturbed. Concretely, we first model multi-agent communication as a multi-view problem, where every message stands for a view of the state. Then we extract a certificated joint message representation by a multi-view variational autoencoder (MVAE) that uses a product-of-experts inference network. For the optimization phase, we do perturbations in the latent space of the state for a certificate guarantee. Then the learned joint message representation is used to approximate the certificated state representation during training. Extensive experiments in several cooperative multi-agent benchmarks validate the effectiveness of the proposed CroMAC.

Keywords multi-agent reinforcement learning, robust communication, adversarial training, multi-view learning, message certification

1 Introduction

Many real-world problems are made up of multiple interactive agents, which could usually be modeled as a multi-agent reinforcement learning (MARL) problem [1, 2]. Further, when the agents hold a shared goal, this problem refers to cooperative MARL [3], which shows great progress in diverse domains like power management [4], multi-UAV control [5], and dynamic algorithm configuration [6]. Many methods are proposed to promote the coordination ability of MARL, including value-based methods [7–9], policy-gradient-based methods [10–12], and some variants [13–16], showing great progress in many complex and challenging benchmarks [17, 18]. Nevertheless, prior studies hugely depend on the strength of deep neural networks (DNNs), whose vulnerability might cause catastrophic results when any perturbation happens [19]. Recently this phenomenon has been tested in cooperative MARL [20], showing that a cooperative MARL system is of low robustness when encountering any perturbations (e.g., state, action, and reward).

Robustness has been widely investigated in single-agent reinforcement learning (RL) [19], and many studies have applied different techniques to various aspects to investigate it. A prior popular way is to introduce an auxiliary adversary to play against the ego-system [21–24], then model the process of policy learning as a minimax problem from the perspective of game theory [25], which may trigger performance deterioration or even unsafe behaviors when facing an unpredictable adversarial policy. Another kind of method tackles this issue by designing efficient and useful regularizers in the training process [22, 26–28], showing efficient robustness in various domains. Certificate-based methods furthermore apply

[*] Corresponding author (email: yuy@nju.edu.cn)

some techniques like vector-ϵ-ball perturbations to obtain a certificate robustness guarantee during the training and testing phases [29–31]. However, the MARL problem differs considerably from the single-agent setting, with multiple agents interacting with others [32].

For the robustness of MARL, new challenges such as scalability [33] arise as multiple agents interact with others in the training phase. For example, in the auxiliary adversary training paradigm, the action space of adversary policy may grow dramatically with respect to the number of agents in an MARL system. Studies on robust MARL should then consider both the robustness and the multi-agent specificity. Some studies design efficient mechanisms to obtain a robust policy to avoid overfitting to specific partners [34] or opponents [35], and others consider the Markov decision process (MDP) itself to get a robust policy in response to state [36], reward [37], and action [38, 39]. Nonetheless, the robustness of a communication policy is much more complex [40], as we should consider when to give what perturbations on which message channel(s) to adversarially train the communication policy. Prior studies mainly investigate the emergence of adversarial communication [41] or impose constraints like a limited number of message channels [42, 43] suffering from message perturbations. These approaches make progress somewhat, but the constraints hinder the robustness's completeness and are also far away from the real-world condition, as all the message channels could sustain perturbations [44]. Even worse, these approaches lack formal robustness guarantees or certificates between each agent's received messages and decision-making.

With this in mind, we propose to promote robustness in current multi-agent communication methods. We posit obtaining a robust communication policy where every messaging channel could suffer from perturbations at any time. Specifically, for any agent in an N-agent system, it will receive $N-1$ messages. As each message is a different view of the state, we model the message-receiving process as a multi-view (also known as "multi-modal") problem, then obtain joint message representations with robustness guarantees from each received message by a multi-view variational autoencoder (MVAE) that uses a product-of-experts inference network. For the optimization phase, we first encode the state into a latent space and do perturbations in this space to obtain a certificate relationship between the latent variable and the agents' Q-values. Then, we train the message representation by approximating the certificated latent variables, and ensure certification between each message and the agents' Q-value implicitly. As we directly impose perturbations in the latent space, the problem of specific action designing for any auxiliary adversaries can be avoided. For evaluation, we conduct extensive experiments on various cooperative multi-agent benchmarks, including hallway [45], level-based foraging [17], traffic junction [46], and two StarCraft Multi-Agent Challenge (SMAC) maps [45]. The results show that CroMAC achieves comparable or superior performance to multiple baselines. Moreover, visualization results show how CroMAC works, and more results demonstrate its high generality ability for different methods under different conditions.

2 Related work

Multi-agent communication plays a promising role in multi-agent coordination under partial observability, which considers when to communicate with whom and what contents to share [40]. The early relevant studies mainly consider designing different communication paradigms to improve communication efficiency [47, 48]. DIAL [47] is a simple communication mechanism where agents broadcast messages to all teammates, allowing the gradient to flow among agents for end-to-end training with reinforcement learning. CommNet [48] proposes an efficient centralized communication structure, where the outputs of the hidden layers from all the agents are collected and averaged to augment local observation. As the mentioned communication paradigm may cause message redundancy, some studies employ techniques such as gate mechanisms [49–51] to explicitly decide whom to communicate with, or attention mechanisms [46, 52, 53] to weigh different messages. What messages to share among agents is another crucial issue. The most naive way is only to share local observations or their embeddings [45, 47], which inevitably causes bandwidth wasting or even degrades coordination efficiency. Towards a more efficient communication protocol, some methods utilize techniques like teammate modeling to generate more succinct and efficient messages [54–56]. For the robustness of message sharing in CMARL, Blumenkamp and Prorok [41] developed a new multi-agent learning model that integrates heterogeneous, potentially self-interested policies that share a differentiable communication channel to elicit the emergence of adversarial communications. Xue et al. [43] considered multi-agent adversarial communication, learning robust communication policy when some message senders are poisoned. A recent method named AME [42] is proposed to acquire a robust communication policy when less than half of the agents in the system sustain noise and potential attackers.

Robustness in single agent reinforcement learning. Moos et al. [19] involved perturbations that occur of different aspects in single agent reinforcement learning such as state, reward, and policy. Some prior methods introduce an adversary to achieve robustness via training the ego-system and the adversary in an alternative way [21–24, 57]. RARL [21] picks out specific robot joints which the adversary acts on to find an equilibrium of the minimax objective using an alternative learning adversary. RAP [23] and GC [24] improve RARL by learning population-based augmentation to the robust RL formulation. However, while these approaches provide better robust policies, it has been shown that such approaches can negatively impact policy performance in non-adversarial scenarios. Moreover, many unsafe behaviors may be exhibited during online attacks, potentially damaging the system controlled by the learning agent if adversarial training occurs in a physical rather than a simulated environment. Other methods improve robustness by designing useful and appropriate regularizers in the loss function [26,27,58]. Zhang et al. [22] formulated the problem of decision making under adversarial attacks on state observations as SA-MDP and learned a state-adversarial policy for multiple DRL methods like DDPG and DQN. RADIAL-RL [26] trains reinforcement learning agents with improved robustness against l_p-norm bounded adversarial attacks, showing superior performance on multiple benchmarks. The mentioned approaches achieve robustness compared to adversarial training, improving the sample efficiency as they need not train an auxiliary adversary. Furthermore, these mentioned methods lack theoretical guarantee, hastening some recent certificate robustness methods [29, 30, 59, 60]. CARRL [59] develops an online certifiably robust policy that computes guaranteed lower bounds on state-action values during execution to identify and choose a robust action under a worst-case deviation in input space due to possible adversaries or noise. CROP [30] gives a solid theoretical guarantee for robust reinforcement learning and applies function smoothing techniques to train a robust policy.

Multi-view (modal) representation learning aims to learn feature representations from multi-view data using different views' information. Its main difficulty is to explicitly measure the content similarity between the heterogeneous samples. How to solve this problem roughly divides multi-view representation learning into three methods: alignment representation [61], joint representation [62], as well as shared and specific representation [63]. The key ideas of these methods are the same, which is establishing a common representation space by exploring the semantic relationship among the multi-view data. One popular and promising way is to use generative models like VAE [64], which generate this representation space in two ways: cross-view generation and joint-view generation. The former learns a conditional generative model over all views by applying techniques like conditional VAE [65]. Nevertheless, the latter learns the joint distribution of the multi-view data. For example, MVAE [66] models the joint posterior as a product-of-experts (POE), and JMVAE [67] learns a shared representation with a joint encoder. Please refer to [68, 69] for a comprehensive review. After the representation space is established by multi-view learning, some approaches use it to solve the modality missing problem [70], or obtain a compact representation from incomplete views [71]. Li et al. [72] extended the partially observable Markov decision processes (POMDPs) to support more than one observation model and proposed two solutions through observation augmentation and cross-view policy transfer in a reinforcement learning problem. DRIBO [73] leverages the sequential nature of RL to learn robust representations that encode only task-relevant information from observations based on the unsupervised multi-view setting. Kinose et al. [74] introduced a novel reinforcement learning agent for integrated recognition and control from multi-view observations. To the best of our knowledge, none of any MARL approaches use multi-view learning to train the communication policy. We take a further step in this direction to get a robust message representation.

Multi-agent robustness. Robustness also plays a promising role in MARL [20], but suffers from extra challenges that do not appear in the single-agent setting, as interactions exist among agents [75], leading to new and specific considerations such as non-stationarity [76], credit assignment [77], and scalability [33] when improving the robustness of any multi-agent system [20]. One type of relevant work aims to investigate the robustness of a learned coordination policy. Lin et al. [78] first learned an observation attacker via RL, then used it to poison one manually selected agent, showing the multi-agent system is vulnerable to observation perturbation. Guo et al. [20] recently did more comprehensive robustness testing on reward, state, and action for typical MARL methods like QMIX [8] and MAPPO [11]. As for robustness improvement in MARL, research is conducted on multiple aspects. Many prior studies focus on designing an efficient approach to learning a robust coordination policy to avoid overfitting to specific partners [34] or opponents [35]. Akin to considering the MDP in a single-agent setting (e.g., state, reward, action), R-MADDPG [37] considers the model uncertainty of an MARL system, then introduces the concept of robust Nash equilibrium. Hu et al. [38] applied a heuristic rule to investigate the robustness of MARL when some agents suffer from action mistakes, and utilized correlated equilibrium theory to learn a robust coordination policy. Robustness in multi-agent communication has also attracted some attention in recent years. Mitchell et al. [79] applied a filter based on the Gaussian process to extract valuable content from noisy messages. Tu et al. [80] studied robustness at the neural network level for secure multi-agent systems. Xue et al. [43] modeled multi-agent communication as a two-player zero-sum game and applied the policy-search response-oracle (PSRO) technique to learn a robust communication policy. The most related work to ours is ablated message ensemble (AME) [42], which assumes no more than half of the message channels in the system may be attacked, then introduces an ensemble-based defense method to achieve robustness. However, we will show that this approach performs poorly in complex scenarios, as the constraints may impede robust efficiency.

3 Problem formulation

We consider a fully cooperative MARL communication problem, which can be formally modeled as a decentralized partially observable MDP under communication (Dec-POMDP-Com) [43] and formulated as a tuple $\langle \mathcal{N}, \mathcal{S}, \mathcal{A}, P, \Omega, O, R, \gamma, \mathcal{M} \rangle$, where $\mathcal{N} = \{1, \ldots, n\}$, $\mathcal{S}$, $\mathcal{A}$, and Ω are the sets of agents, states, actions, and observations, respectively. O is the observation function, P denotes the transition function, R represents the reward function, $\gamma \in [0, 1)$ stands for the discounted factor, and $\mathcal{M}$ indicates the message set. Due to the partially observable nature of the environment, each agent $i \in \mathcal{N}$ can only obtain the local observation $o_i \in \Omega$, and hold an individual policy $\pi(a_i \mid \tau_i, m_i)$, where τ_i represents the output of a trajectory encoder (e.g., GRU [81]) which encodes $(o_i^1, a_i^1, \ldots, o_i^{t-1}, a_i^{t-1}, o_i^t)$, and $m_i \in \mathcal{M}$ is the messages received by agent i and m_{ij} represents the message transmitted from j to i. As each agent can behave as a message sender as well as a message receiver, this paper considers learning useful message representation on the receiving end, and agents only use local information (e.g., τ_i, and we use $m_{:,i}$ for generality) as message $m_{:,i}$ to share within the team. We aim to find an optimal policy under the setting where each message channel in the multi-agent system may suffer from perturbations. In line with the widely used state-adversarial MDP (SA-MDP) in single-agent RL [22, 82], we formulate this setting as a message-adversarial Dec-POMDP-Com (MA-Dec-POMDP-Com).

In an MA-Dec-POMDP-Com, we introduce a message adversary $v(m) : m \to \hat{m}$. The adversary perturbs the messages received by each agent, such that agent i takes action by $\pi(a_i|\tau_i, \hat{m}_i)$. The joint action $\boldsymbol{a} = \langle a_1, \ldots, a_n \rangle$ leads to the next state $s' \sim P(\cdot \mid s, \boldsymbol{a})$ and the global reward $R(s, \boldsymbol{a})$. The formal objective is to find a joint policy $\boldsymbol{\pi}(\boldsymbol{\tau}, \boldsymbol{a})$ to maximize the global value function $Q_{\text{tot}}^{\boldsymbol{\pi}}(\boldsymbol{\tau}, \boldsymbol{a}) = \mathbb{E}_{s,\boldsymbol{a}}[\sum_{t=0}^{\infty} \gamma^t R(s, \boldsymbol{a}) \mid s_0 = s, \boldsymbol{a_0} = \boldsymbol{a}, \boldsymbol{\pi}]$, with $\boldsymbol{\tau} = \langle \tau_1, \ldots, \tau_n \rangle$. If the adversary can perturb a message m arbitrarily without bounds, the problem becomes trivial [83]. To fit our method to the most realistic settings, we restrict the power of an adversary to a perturbation set $\mathcal{B}$, i.e., $\mathcal{B} = \{\hat{m} \mid \|m - \hat{m}\|_p \leqslant \epsilon\}$, where ϵ is the given perturbation magnitude and p determines the type of norm. Our experiments in this paper focus on $p = \infty$.

Furthermore, since $\mathcal{B}(m)$ is usually a small set nearby m, our adversary applies FGSM [84] to learn a perturbation vector Δ, and we project $m + \Delta$ to $\mathcal{B}(m)$.

4 Method

Then we obtain the bounds between the joint message representation and each message by interval bound propagation [85]. In the training phase, we first encode state $s_t \in \mathcal{S}$ into a latent variable z_{st}, then we impose perturbations in the latent space to gain a certificate guarantee between z_{st} and each state-action value $Q_i(\tau_i, z_{\text{st}} \pm \kappa\epsilon; a_i)$, where $\pm\kappa\epsilon$ represents that the variable suffers from ℓ_∞-norm perturbations within budget $\kappa\epsilon$ and κ is a constant. Finally, the joint message representation z_{msg} is optimized by approximating z_{st} via minimizing the Kullback-Leibler divergence between these two variables, endowing certification between each message and each state-action value implicitly. In the execution phase, we only use the message aggregation module and the trajectory encoder to make decisions in a decentralized way.

4.1 Multi-view multi-agent communication

We consider learning a robust communication policy in an MA-Dec-POMDP-Com. For each agent i, there are $N-1$ message channels; thus each agent receives multiple available messages about the environment. Inspired by the widely used multi-view learning [86, 87], we apply the POE [85] technique to extract joint message representations. Formally, agent i receives multiple messages m_{ij}^t from teammate $j \in \{1,\ldots,i-1,i+1,\ldots,N\}$ and let m_{ii}^t at time t denote its local history τ_i^t. We assume each message is conditioned on an unknown hidden variable z_{ij}^t; then the generation of multiple messages can be modeled as a multi-view variational autoencoder process. We then optimize the evidence lower bound (ELBO) to maximize the marginal likelihood with a message encoder $q_{\boldsymbol{\phi}_{\rm enc}}(z_{ij}^t|m_{ij}^t)$ parameterized with $\boldsymbol{\phi}_{\rm enc}$, and a message decoder $p_{\boldsymbol{\phi}_{\rm dec}}(m_{ij}^t|z_{ij}^t)$ with parameter $\boldsymbol{\phi}_{\rm dec}$:

$$\mathrm{ELBO}(m_{ij}^t) \triangleq \mathbb{E}_{q_{\boldsymbol{\phi}_{\rm enc}}(z_{ij}^t|m_{ij}^t)}\left[\log p_{\boldsymbol{\phi}_{\rm dec}}\left(m_{ij}^t \mid z_{ij}^t\right)\right] - \mathrm{KL}\left[q_{\boldsymbol{\phi}_{\rm enc}}(z_{ij}^t \mid m_{ij}^t), p(z_{ij}^t)\right], \tag{1}$$

where $\mathrm{KL}[q,p]$ is the Kullback-Leibler divergence between distributions q and p. The first term in (1) is the reconstruction likelihood, and the second term aims to guarantee that the output of the encoder is similar to the prior distribution $p(z_{ij}^t)$, and can be regarded as a regularization term. The message encoder $q_{\boldsymbol{\phi}_{\rm enc}}$ outputs parameters of an n-multivariate Gaussian distribution $\mathcal{N}(\mu_{ij}^t, \sigma_{ij}^t)$, where μ_{ij}^t and σ_{ij}^t are the mean and standard deviation of $p(z_{ij}^t)$, respectively. As all messages $\{m_{i1}^t,\ldots,m_{iN}^t\}$ are conditionally independent given the common latent variable z_i^t, we assume a generative model for all the messages in the form:

$$p(m_{i1}^t,\ldots,m_{iN}^t,z_i^t) = p(z_i^t)p(m_{i1}^t|z_i^t)p(m_{i2}^t|z_i^t)\cdots p(m_{iN}^t|z_i^t). \tag{2}$$

Then Eq. (1) can be extended as

$$\mathrm{ELBO}(m_i^t) \triangleq \mathbb{E}_{q_{\boldsymbol{\phi}_{\rm enc}}(z_i^t|m_i^t)}\left[\sum_{j=1}^N \log p_{\boldsymbol{\phi}_{\rm dec}}\left(m_{ij}^t \mid z_i^t\right)\right] - \mathrm{KL}\left[q_{\boldsymbol{\phi}_{\rm enc}}(z_i^t \mid m_i^t), p(z_i^t)\right], \tag{3}$$

where $m_i^t = \{m_{i1}^t,\ldots,m_{iN}^t\}$ is the set of messages agent i receives at time t and its local history m_{ii}^t (i.e., τ_i^t). Then, this message generation process can be treated as a multi-view representation learning problem [67]. We use the inference network $q(z_i^t \mid m_i^t)$ as a variational distribution to approximate the true posterior $p(z_i^t \mid m_i^t)$, then get the relationship among the joint- and single-view posteriors as

$$\begin{aligned}
p(z_i^t \mid m_i^t) &= \frac{p(m_i^t \mid z_i^t)p(z_i^t)}{p(m_i^t)} = \frac{p(z_i^t)}{p(m_i^t)}\prod_{j=1}^N p(m_{ij}^t \mid z_i^t) \\
&= \frac{p(z_i^t)}{p(m_i^t)}\prod_{j=1}^N \frac{p(z_i^t \mid m_{ij}^t)p(m_{ij}^t)}{p(z_i^t)} \\
&= \frac{\prod_{j=1}^N p(m_{ij}^t)}{p(m_i^t)}\frac{\prod_{j=1}^N p(z_i^t \mid m_{ij}^t)}{\prod_{j=1}^{N-1} p(z_i^t)} \\
&\propto \frac{\prod_{j=1}^N p(z_i^t \mid m_{ij}^t)}{\prod_{j=1}^{N-1} p(z_i^t)} \approx \frac{\prod_{j=1}^N [q(z_i^t \mid m_{ij}^t)p(z_i^t)]}{\prod_{j=1}^{N-1} p(z_i^t)} \\
&= p(z_i^t)\prod_{j=1}^N q(z_i^t \mid m_{ij}^t).
\end{aligned} \tag{4}$$

The last two lines in (4) hold as we use $q(z_i^t \mid m_{ij}^t)p(z_i^t)$ to approximate $p(z_i^t \mid m_{ij}^t)$ so that the inference network is composed of N neural networks $q(z_i^t \mid m_{ij}^t)$, and if each view is homogeneous, we can even replace them with only one network with shared parameters. Akin to the standard VAE [64], we apply a deep neural network (e.g., MLP) to model the message encoder $q_{\boldsymbol{\phi}_{\rm enc}}(z_i^t|m_{ij}^t)$, which outputs the parameters of the Gaussian distribution μ_{ij}, σ_{ij}^2. As for now, we can combine the multiple outputs of the message encoder in a simple analytical way: a product of Gaussian experts is itself Gaussian [88],

$$\begin{aligned}\mu_i &= \left(\sum_{j=1}^{N}\mu_{ij}T_{ij}\right)\left(\sum_{j=1}^{N}T_{ij}\right)^{-1},\\ \sigma_i^2 &= \left(\sum_{j=1}^{N}T_{ij}\right)^{-1},\end{aligned} \tag{5}$$

where μ_i and σ_i^2 are the mean and variance of the learned joint message representation's Gaussian distribution, and μ_{ij} and σ_{ij}^2 are the mean and variance of the ith agent's jth Gaussian distribution through message encoder, $T_{ij} = (\sigma_{ij}^2)^{-1}$ is the inverse of the variance. The detailed derivative process can be seen in Appendix A.

4.2 Message certificates via bound propagation

Though we have combined all the received messages into a joint message representation, the learned joint message representation still lacks a certificated guarantee with each received message under perturbation. In this part, we aim to achieve this using the interval bound propagation technique. Formally, considering agent i receives messages $m_i = \{m_{i1}, \dots, m_{iN}\}$ under perturbation of ℓ_∞-norm attack within given budget ϵ, the upper and lower bounds are $\overline{m_i} = \{\overline{m_{i1}}, \dots, \overline{m_{iN}}\} = \{m_{i1} + \boldsymbol{\epsilon}, \dots, m_{iN} + \boldsymbol{\epsilon}\}$ and $\underline{m_i} = \{\underline{m_{i1}}, \dots, \underline{m_{iN}}\} = \{m_{i1} - \boldsymbol{\epsilon}, \dots, m_{iN} - \boldsymbol{\epsilon}\}$, respectively. The averages and residuals of the upper and lower bounds can be defined as

$$\begin{aligned}\hat{\mu}_0 &= \frac{1}{2}(\overline{m_i} + \underline{m_i}) = m_i,\\ \hat{r}_0 &= \frac{1}{2}(\overline{m_i} - \underline{m_i}) = \boldsymbol{\epsilon}.\end{aligned} \tag{6}$$

Here, we use "average" instead of "mean" to distinguish it from the one in VAE, and $\hat{\mu}$ and $\hat{r}$ are used to represent averages and residuals, respectively. For simplification of notation and mathematical derivation, we assume each message encoder has only one layer fully-connected network with shared parameters, and we can use any NNs with arbitrary depths and element-wise monotonic activation functions (e.g., ReLU, Sigmoid, Tanh) by getting a reasonable bound propagation mechanism [85]. Further, we here use W_m, b_m, and W_v, b_v to represent the parameters of the two separate fully connected layers in the message encoder, which output the means and variances of the Gaussian distributions, respectively. Thus we can propagate the bounds through the one MLP layer by matrix multiplication. The averages and residuals of the bounds come to be

$$\begin{aligned}\hat{\mu}_m &= \{W_m\hat{\mu}_0(1) + b_m, \dots, W_m\hat{\mu}_0(N) + b_m\},\\ \hat{r}_m &= \{|W_m|\hat{r}_0(1), \dots, |W_m|\hat{r}_0(N)\},\\ \hat{\mu}_v &= \{W_v\hat{\mu}_0(1) + b_v, \dots, W_v\hat{\mu}_0(N) + b_v\},\\ \hat{r}_v &= \{|W_v|\hat{r}_0(1), \dots, |W_v|\hat{r}_0(N)\},\end{aligned} \tag{7}$$

where $(\hat{\mu}_m, \hat{r}_m)$ and $(\hat{\mu}_v, \hat{r}_v)$ stand for the (average, residual) pairs of the mean outputs' bounds and the variance outputs' bounds, respectively, and $|\cdot|$ is the element-wise absolute value operator. We use ReLU as the activation function which is element-wise monotonic so we can omit it as it will not affect the propagating bounds. Thus the upper and lower bounds of each single message representation (i.e., mean and variance) can be written as

$$\begin{aligned}\overline{z_m} &= \hat{\mu}_m + \hat{r}_m = \{W_m m_{i1} + b_m + |W_m|\boldsymbol{\epsilon}, \dots, W_m m_{iN} + b_m + |W_m|\boldsymbol{\epsilon}\},\\ \underline{z_m} &= \hat{\mu}_m - \hat{r}_m = \{W_m m_{i1} + b_m - |W_m|\boldsymbol{\epsilon}, \dots, W_m m_{iN} + b_m - |W_m|\boldsymbol{\epsilon}\},\\ \overline{z_v} &= \hat{\mu}_v + \hat{r}_v = \{W_v m_{i1} + b_v + |W_v|\boldsymbol{\epsilon}, \dots, W_v m_{iN} + b_v + |W_v|\boldsymbol{\epsilon}\},\\ \underline{z_v} &= \hat{\mu}_v - \hat{r}_v = \{W_v m_{i1} + b_v - |W_v|\boldsymbol{\epsilon}, \dots, W_v m_{iN} + b_v - |W_v|\boldsymbol{\epsilon}\}.\end{aligned} \tag{8}$$

Considering the relationship in (5), we then have

$$\begin{aligned} Z_M &= \left(\sum_i z_m(i) z_v(i)^{-1}\right)\left(\sum_i z_v(i)^{-1}\right)^{-1}, \\ Z_V &= \left(\sum_i z_v(i)^{-1}\right)^{-1}, \end{aligned} \tag{9}$$

where Z_M and Z_V represent the mean and variance of the joint message representation. However, we cannot get the upper and lower bounds as the POE we use is not affine neural layers. Notice that Z_V is actually the harmonic mean [89] of $\frac{z_v(i)}{N}$ (the harmonic mean of variables x_i is $H_n = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}}$) while Z_M is the weighted harmonic mean of $z_m(i)$ with weights $\frac{z_m(i)}{z_v(i)}$ (the weighted harmonic mean of x_i with weights w_i is $H_n = \frac{m_1+m_2+\cdots+m_n}{\frac{m_1}{x_1}+\frac{m_2}{x_2}+\cdots+\frac{m_n}{x_n}}$). Here, variances $z_v(i)$ are non-negative and means $z_m(i)$ can be normalized to a positive range; therefore we can scale (9) appropriately by the properties of harmonic mean to infer the upper and lower bounds. We here prove one of them for simplicity. Taking the variance term for example and simplifying W_v, b_v, Z_V, z_v as W, b, Z, z, we have

$$\begin{aligned} \overline{Z} &= \left(\sum_i \overline{z}(i)^{-1}\right)^{-1} \leqslant \frac{\mathbf{1}\cdot\max(\overline{z})}{N}, \\ \underline{Z} &= \left(\sum_i \underline{z}(i)^{-1}\right)^{-1} \geqslant \frac{\mathbf{1}\cdot\min(\underline{z})}{N}. \end{aligned} \tag{10}$$

Note that $\mathbf{1}$ is an all-1-vector with the same dimension as $\overline{Z}$. $(\leqslant, \geqslant)$ signs here act on each element of vectors, and $(\max, \min)$ operations find the maximum or minimum number of vectors of the set. We can get the upper bound of the final integration error:

$$\max(\overline{Z} - Z_{\text{true}}, Z_{\text{true}} - \underline{Z}) \leqslant \overline{Z} - \underline{Z} \leqslant \frac{\mathbf{1}\cdot(\max(\overline{z}) - \min(\underline{z}))}{N}, \tag{11}$$

where Z_{true} stands for the ground truth value. Assuming the pth element of the jth vector of $\overline{z}$ is $\max(\overline{z})$ and the qth element of the kth vector of $\underline{z}$ is $\min(\underline{z})$, we get

$$\begin{aligned} \max(\overline{z}) - \min(\underline{z}) &= (Wm_{ij} + b + |W|\boldsymbol{\epsilon})_p - (Wm_{ik} + b - |W|\boldsymbol{\epsilon})_q \\ &= W_{p,:}m_{ij} - W_{q,:}m_{ik} + b_p - b_q + (|W|_{p,:} - |W|_{q,:})\boldsymbol{\epsilon}, \end{aligned} \tag{12}$$

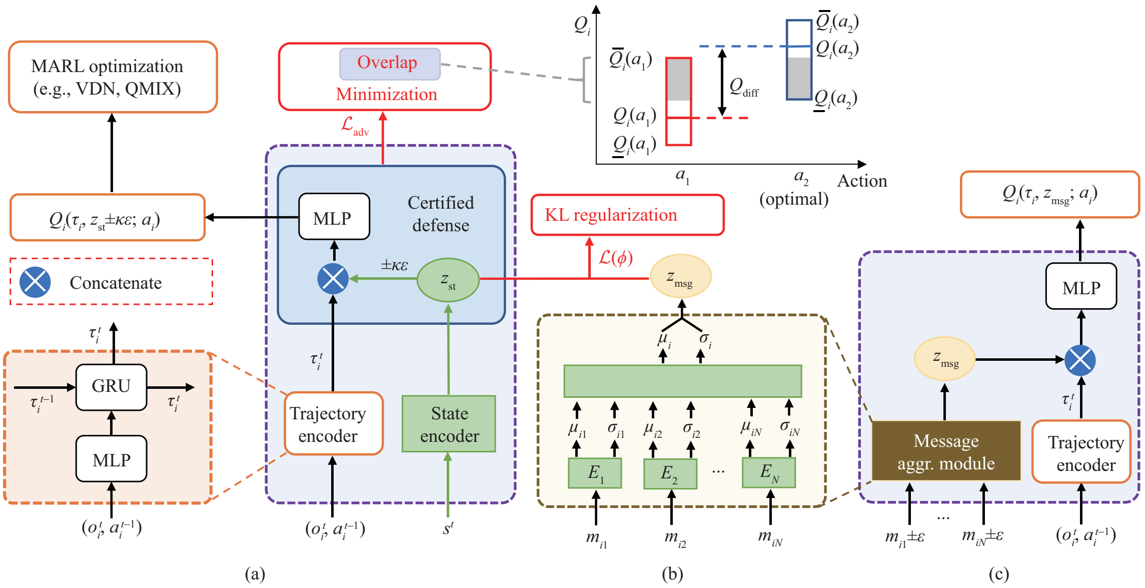
where $W_{p,:}$ means the pth row of matrix W. We can notice that the integration error can be limited to a constant $\||W|_{p,:} - |W|_{q,:}\|_1/N$ times $\boldsymbol{\epsilon}$ if W, b, m_i are bounded. Through subsequent experiments, we found that good robustness could be achieved when the integrated information is subjected to noise perturbation within the range of $\kappa\epsilon$ with only W bounded to [C_MIN,C_MAX], here C_MIN,C_MAX, and κ are hyperparameters and let C_MIN=−C_MAX.

4.3 Robustness training scheme

As we have obtained the theoretical guarantee between the received messages and the learned joint message representation, now this subsection describes how to acquire a robust communication policy. Following the popular centralized training and decentralized execution (CTDE) paradigm [90,91], during the training phase, we use a state encoder (e.g., additional VAE [64]) to encode the state s into a latent space with parameter $\boldsymbol{\psi}$:

$$\mathcal{L}(\boldsymbol{\psi}) = -\mathbb{E}_{q_{\boldsymbol{\psi}_{\text{enc}}}(\boldsymbol{z}_{\text{st}}|s)}[\log p_{\boldsymbol{\psi}_{\text{dec}}}(s \mid z_{\text{st}})] + \mathrm{KL}\left[q_{\boldsymbol{\psi}_{\text{enc}}}(\boldsymbol{z}_{\text{st}} \mid s), p(\boldsymbol{z}_{\text{st}})\right], \tag{13}$$

where the operators are similar to (1). We can then apply any robust single-agent RL algorithm to achieve robustness in the latent space of state representation. If the robustness of state representation is guaranteed, we can also ensure the robustness of joint message representation through knowledge distillation mentioned in (16). We here implement our method on RADIAL-RL [26] as it principles a



(a) (b) (c)

Figure 1 (Color online) Structure of CroMAC. (a) During the training phase, we encode the state into latent variables $z_{\rm st}$, then perturb it to gain a certificate guarantee between $z_{\rm st}$ and $Q_i(\tau_i, z_{\rm st} \pm \kappa\epsilon; a_i)$, and this process is optimized via minimizing the overlap (i.e., the gray part) weighted by the difference of Q-values. Here a_2 represents the original optimal action while a_1 is another sub-optimal action. $\overline{Q}$ and $\underline{Q}$ represent the upper bound and lower bound of Q-value under perturbation, respectively, therefore the gray overlap measures the probability of selecting action besides the original optimal action under perturbation. Since not all overlap is equally important, if two actions have similar Q-values, i.e., $Q_{\rm diff}$ is small, we can ignore the overlap as taking a different but equally good action under perturbation is acceptable. The whole process can be optimized by any value decomposition methods like QMIX [8], and the output of the message aggregation module $z_{\rm msg}$ is then used to approximate $z_{\rm st}$ by minimizing their distance (e.g., KL divergence). (b) The message aggregation module. Each message m_{ij} is encoded into a latent space via a message encoder E_j, where $j \in \{1, \ldots, i-1, i+1, \ldots, N\}$, and the parameters of E_j are regularized to obtain certificates between the joint message representation and each message. (c) After training, we use the learned message aggregation module and other shared modules like the trajectory encoder to make a decision in a decentralized way.

framework in adversarial training with strong theoretical guarantee and robustness performance. Then we can minimize the following loss to optimize each agent's individual policy:

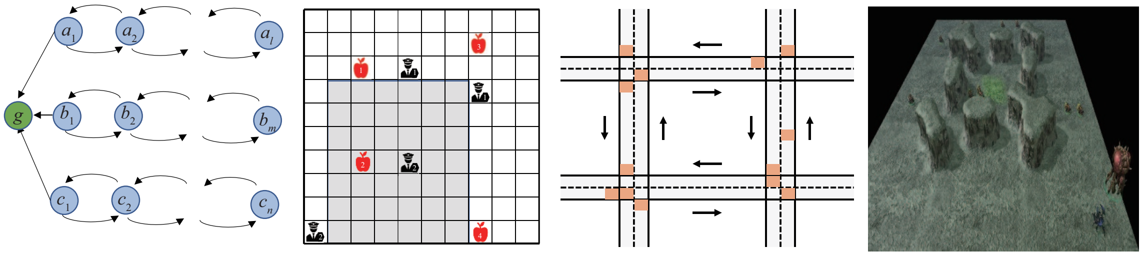
$$\mathcal{L}_{\rm adv} = \mathbb{E}_{(s,\boldsymbol{a},s',r)}\left[\sum_i \sum_y Q^i_{\rm diff}(\tau, z_{\rm st}; y) \cdot {\rm Ovl}^i(\tau, z_{\rm st}, \kappa\epsilon; y)\right], \tag{14}$$

with

$$\begin{aligned} Q^i_{\rm diff}(\tau, z_{\rm st}; y) &= \max\left(0, Q^i(\tau, z_{\rm st}; a) - Q^i(\tau, z_{\rm st}; y)\right), \\ {\rm Ovl}^i(\tau, z_{\rm st}, \kappa\epsilon; y) &= \max\left(0, \overline{Q}^i(\tau, z_{\rm st}, \kappa\epsilon; y) - \underline{Q}^i(\tau, z_{\rm st}, \kappa\epsilon; a)\right), \end{aligned} \tag{15}$$

where i is the identification of each agent, y is each action, a is the chosen action, and $\overline{Q}$ and $\underline{Q}$ can be computed by interval bound propagation under ℓ_∞-norm perturbation within budget $\kappa\epsilon$, which is readily available as there is only MLP network existing between the Q-values and $(\tau, z_{\rm st})$. Ovl represents the overlap between the bounds of two actions which can be seen in Figure 1, and $Q_{\rm diff}$ measures the relative quality between two actions as we can ignore the overlap if they are similar enough. When minimizing the weighted overlap to 0, which means even the upper bound of another action y's action-value $\overline{Q}^i(\tau, z_{\rm st}, \kappa\epsilon; y)$ is lower than the lower bound of original action a's action-value $\underline{Q}^i(\tau, z_{\rm st}, \kappa\epsilon; a)$, the agent will not change its action under perturbation, leading to a robust communication policy. We note that the model's initial training will be hindered if we add the robust loss. Therefore, it is better to start robust training after the training is stable, and we use T_r to control it. Then we optimize the joint message representation by minimizing the KL divergence between $z_{\rm st}$ and $z_{\rm msg}$ as a form of knowledge distillation, and we use only the message encoder $\boldsymbol{\phi}_{\rm enc}$ to make the inference of the joint message representation:

$$\mathcal{L}(\boldsymbol{\phi}) = {\rm KL}\left[{\rm sg}(z_{\rm st}), q_{\boldsymbol{\phi}_{\rm enc}}(z_{\rm msg} \mid m)\right], \tag{16}$$



(a) (b) (c) (d)

Figure 2 (Color online) Multiple benchmarks used in our experiments. (a) Hallway; (b) LBF; (c) traffic junction (TJ); (d) SMAC.

where $\mathrm{sg}(\cdot)$ denotes gradient stop, and z_{msg} is the joint message representations sampled from $\mathcal{N}(\mu_i, \sigma_i^2)$. Then the overall loss function is as follows:

$$\mathcal{L} = \mathcal{L}_{\text{TD}}(\boldsymbol{\theta}) + \alpha_1 \mathcal{L}(\boldsymbol{\psi}) + \alpha_2 \mathcal{L}(\boldsymbol{\phi}) + \mathbb{I}(t > T_r)\alpha_3 \mathcal{L}_{\text{adv}}, \tag{17}$$

where $\mathcal{L}_{\text{TD}}(\boldsymbol{\theta})$ is the temporal difference loss, α_1, α_2, and α_3 are adjustable hyperparameters for each loss function accordingly, and $\mathbb{I}(\cdot)$ is the indicator function. In the CTDE framework, the mixing network will be removed during the decentralized execution phase. To prevent the lazy-agent problem [7] and reduce model complexity, we make the local network have the same parameters for all agents. The pseudo-code is shown in Appendix B.

5 Experimental results

In this section, we design experiments on multiple complex scenarios to evaluate the communication robustness for the following questions: (1) How is the robustness of our proposed method on multiple benchmarks compared with baselines (Subsection 5.2)? (2) What is the generalization ability of CroMAC when encountering different message perturbations (Subsection 5.3)? (3) Can CroMAC be integrated into multiple cooperative MARL methods in different communication conditions, and how does each hyperparameter influence its performance (Subsection 5.4)?

For empirical evaluation, we compare CroMAC with multiple baselines on different cooperative tasks, including hallway [45], level-based foraging (LBF) [17], traffic junction (TJ) [46], and two maps from SMAC [45]. CroMAC is implemented on QMIX if not specified based on PyMARL[1]. All results are illustrated with mean performance and standard error on 5 random seeds. Detailed network architecture and hyperparameter choices are shown in Appendix D.

5.1 Baselines and environments

We consider multiple baselines with different communication abilities, where QMIX [8] is a value-based baseline, and no message sharing among agents, showing excellent performance on diverse multi-agent benchmarks [17]. AME [42] is a recently proposed strong method for the robustness of multi-agent communication, which assumes no more than half of the agents may suffer from message perturbations, and an ensemble-based defense approach is then introduced to realize the robustness goal. Full-Comm adopts a full communication paradigm, where each agent receives messages from all teammates at each timestep without message perturbations both in the training and testing phases, which can be seen as an upper-bound performance algorithm. For the ablation studies, we consider multiple variants of CroMAC. CroMAC w/o robust and CroMAC w/o adv are two variants of our proposed CroMAC; the former does not have the proposed robust training scheme while the latter is conducted in the non-perturbed condition in both the training and testing phases. We consider multiple benchmarks, as shown in Figure 2, where hallway [45] is a cooperative environment under partial observability, with m agents randomly initialized at different positions and required to arrive at the goal g simultaneously. We consider two scenarios with different numbers of agents and lengths of the hallway. LBF [17] is another cooperative, partially observable grid world game where agents should coordinate to collect food concurrently. As the original version focuses on exploration, here we modify it by making only one agent be able to observe the map, which needs strong communication to complete this task. TJ [46] is another popular benchmark used

[1] https://github.com/oxwhirl/pymarl.

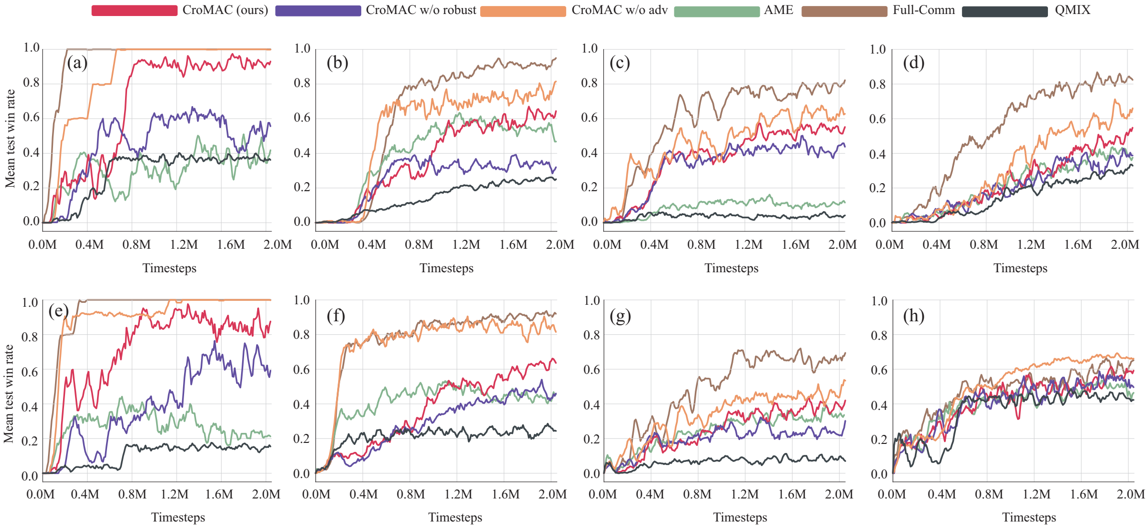


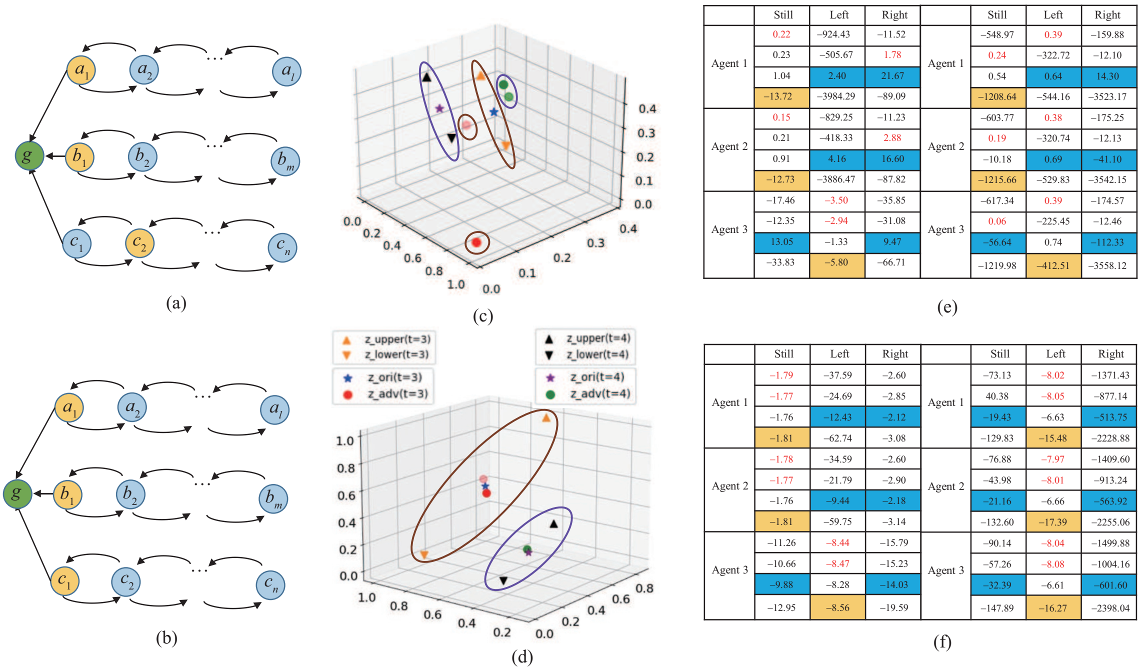
Figure 3 (Color online) Empirical results of several algorithms tested in two different perturbation conditions on benchmarks. Note that Full-Comm, CroMAC w/o adv, and QMIX are tested in perturbation-free conditions, while CroMAC, CroMAC w/o robust, and AME suffer from message perturbations when testing. (a) Hallway: 4x5x6; (b) LBF: 3p-1f; (c) TJ: slow; (d) SMAC: 1o10b_vs_1r; (e) Hallway: 3x3x4x4; (f) LBF: 4p-1f; (g) TJ: fast; (h) SMAC: 1o2r_vs_4r.

to test communication ability, where multiple cars move along two-way roads with one or more road junctions following predefined routes, and they need to drive as fast as possible while avoiding collisions. We test on the modified slow and fast scenarios where different maps have a different probability of adding new cars. Two maps named 1o_2r_vs_4r and 1o_10b_vs_1r from SMAC [45] that require efficient communication are also used to test the robustness in more complex scenarios. Details are presented in Appendix C.

5.2 Robustness comparison and analysis

We first compare CroMAC against multiple baselines to investigate the communication robustness on various benchmarks. As shown in Figure 3, QMIX achieves the most inferior performance in all environments, demonstrating that communication is needed. Full-Comm can solve all the tasks under perturbation-free conditions, showing that these tasks need communication and can be solved by a simple communication mechanism. CroMAC w/o adv, an ablation of CroMAC where testing is conducted under perturbation-free conditions, can achieve comparable coordination ability with Full-Comm, validating the specific design of our CroMAC does not cause much performance degradation for a communication goal. On the contrary, when message perturbations occur during the testing phase, it can be easily found that CroMAC w/o robust, a variant of our proposed method without an efficient robust mechanism, suffers from severe performance degradation compared with Full-Comm and CroMAC w/o adv. However, CroMAC exhibits higher robustness than others, and it surprises us that AME also suffers from severe performance degradation under perturbation, which means an unreasonable constraint for robustness training cannot be applied in complex and severe message perturbation conditions.

Furthermore, we conduct experiments on task hallway to investigate how CroMAC learns a robust communication policy. As shown in Figure 4 [92], three agents coordinate to reach the goal. When suffering from perturbations, the message representation learned by methods without a robust mechanism will go out of the upper and lower bounds, leading to an unpredictable input for the local policy. Consequently, the message perturbation influences the action selection of each agent. Take Agent 1 in Figure 4(e) as an example. It should keep still with Agent 2 at $t = 3$ to wait for Agent 3 to go left together for success. However, when suffering from perturbations, the message representation jumps out of the normal range. It unexpectedly goes right, as the according Q-value 1.78 is dominant to others (0.23 for still and -505.67 for left). On the contrary, with our robustness scheme, the message representations can be bounded in a reasonable range, leading to a robust action selection compared with the perturbation-free setting, as shown in Figure 4(f). The whole process shows our approach can obtain message certification when any perturbations happen.



(e)

	Still	Left	Right		Still	Left	Right
Agent 1	0.22	−924.43	−11.52	Agent 1	−548.97	0.39	−159.88
	0.23	−505.67	1.78		0.24	−322.72	−12.10
	1.04	2.40	21.67		0.54	0.64	14.30
	−13.72	−3984.29	−89.09		−1208.64	−544.16	−3523.17
Agent 2	0.15	−829.25	−11.23	Agent 2	−603.77	0.38	−175.25
	0.21	−418.33	2.88		0.19	−320.74	−12.13
	0.91	4.16	16.60		−10.18	0.69	−41.10
	−12.73	−3886.47	−87.82		−1215.66	−529.83	−3542.15
Agent 3	−17.46	−3.50	−35.85	Agent 3	−617.34	0.39	−174.57
	−12.35	−2.94	−31.08		0.06	−225.45	−12.46
	13.05	−1.33	9.47		−56.64	0.74	−112.33
	−33.83	−5.80	−66.71		−1219.98	−412.51	−3558.12

(f)

	Still	Left	Right		Still	Left	Right
Agent 1	−1.79	−37.59	−2.60	Agent 1	−73.13	−8.02	−1371.43
	−1.77	−24.69	−2.85		40.38	−8.05	−877.14
	−1.76	−12.43	−2.12		−19.43	−6.63	−513.75
	−1.81	−62.74	−3.08		−129.83	−15.48	−2228.88
Agent 2	−1.78	−34.59	−2.60	Agent 2	−76.88	−7.97	−1409.60
	−1.77	−21.79	−2.90		−43.98	−8.01	−913.24
	−1.76	−9.44	−2.18		−21.16	−6.66	−563.92
	−1.81	−59.75	−3.14		−132.60	−17.39	−2255.06
Agent 3	−11.26	−8.44	−15.79	Agent 3	−90.14	−8.04	−1499.88
	−10.66	−8.47	−15.23		−57.26	−8.08	−1004.16
	−9.88	−8.28	−14.03		−32.39	−6.61	−601.60
	−12.95	−8.56	−19.59		−147.89	−16.27	−2398.04

Figure 4 (Color online) Visualization results. We take $t = 3$ and 4 in hallway as shown in (a) and (b), where Agents 1 and 2 stand one step from the goal while Agent 3 needs to take two steps to reach the goal. (c) and (d) show the PCA projection [92] of the message representation $z_{\rm msg}$ for (a) and (b), with • and ⋆ represent $z_{\rm msg}$ with and without perturbations, respectively. Note that $z_{\rm msg}$ is the same for agents without perturbations, and some • are darker because multiple ones overlap together. ▲, ▼ represent the upper and lower bounds of $z_{\rm msg}$; note that ellipses of the same color represent the same time step. (e) and (f) display the Q-values (multiplied by 100 for viewing) of each agent accordingly, where the first row means the original Q-values of all actions, while the second row refers to them under perturbation with red fonts representing the selected actions in corresponding cases. The third and fourth rows show the upper and lower bounds of Q-values under ϵ-perturbation, where yellow squares are the lower bounds of Q-values over best actions while blue squares are the upper bounds of Q-values over other actions.

Table 1 Average test win rates for CroMAC and AME under different message perturbation conditions [a]. The results are averaged from 1000 test episodes among 5 random seeds, where FGSM (n) refers to methods under FGSM attack with different budgets. The details of each perturbation method are shown in Appendix D

Environment	Method	Natural	Random	PGD	FGSM (1)	FGSM (2)	FGSM	FGSM (3)	FGSM (4)
Hallway 4x5x6	CroMAC	0.93±0.06	0.91±0.11	**0.92±0.13**	**0.97±0.03**	**0.86±0.04**	**0.91±0.10**	0.60±0.31	**0.66±0.39**
	AME	0.98±0.01	0.93±0.04	0.43±0.20	0.66±0.34	0.61±0.34	0.62±0.31	0.36±0.10	0.10±0.20
	REC	**1.00±0.00**	**0.95±0.08**	0.90±0.20	0.96±0.06	0.62±0.38	0.82±0.23	**0.68±0.40**	0.41±0.43
LBF 3p-1f	CroMAC	0.71±0.05	0.72±0.03	**0.61±0.09**	**0.71±0.07**	**0.67±0.09**	**0.64±0.13**	**0.43±0.15**	**0.30±0.08**
	AME	**0.77±0.04**	0.72±0.04	0.58±0.11	0.63±0.09	0.56±0.02	0.47±0.03	0.36±0.10	0.29±0.04
TJ slow	CroMAC	**0.31±0.07**	**0.46±0.20**	**0.31±0.23**	**0.29±0.12**	**0.31±0.09**	**0.37±0.14**	**0.42±0.16**	**0.32±0.18**
	AME	0.12±0.07	0.13±0.03	0.15±0.06	0.13±0.06	0.13±0.06	0.12±0.06	0.08±0.02	0.13±0.06
SMAC 1o10b_vs_1r	CroMAC	**0.65±0.10**	**0.64±0.12**	**0.53±0.07**	**0.56±0.04**	**0.52±0.18**	**0.59±0.08**	0.41±0.14	0.34±0.03
	AME	0.38±0.12	0.52±0.24	0.51±0.17	0.44±0.13	0.45±0.20	0.38±0.02	**0.44±0.19**	**0.43±0.07**

[a] The bold data indicates that the algorithm achieves optimal performance under the given environment and message perturbation condition.

5.3 Robustness under various perturbations

As this study considers a setting where the number of attacks is fixed during the training phase, we evaluate here the generalization ability when altering the perturbation budget and encountering different perturbation methods in the testing phase. Specifically, we conduct experiments on each benchmark with the same structures and hyperparameters as Subsection 5.2 during training. As shown in Table 1, we consider eight communication situations, where "Natural" means no perturbation exists, FGSM is the training condition of the comparable approaches, and others like PGD are other conditions (details can be seen in Appendix D). We can find that AME can achieve comparable or even superiority over CroMAC in the natural setting without message perturbations and also maintain competitiveness when suffering from random perturbations, showing that AME possesses robustness for simple message per-

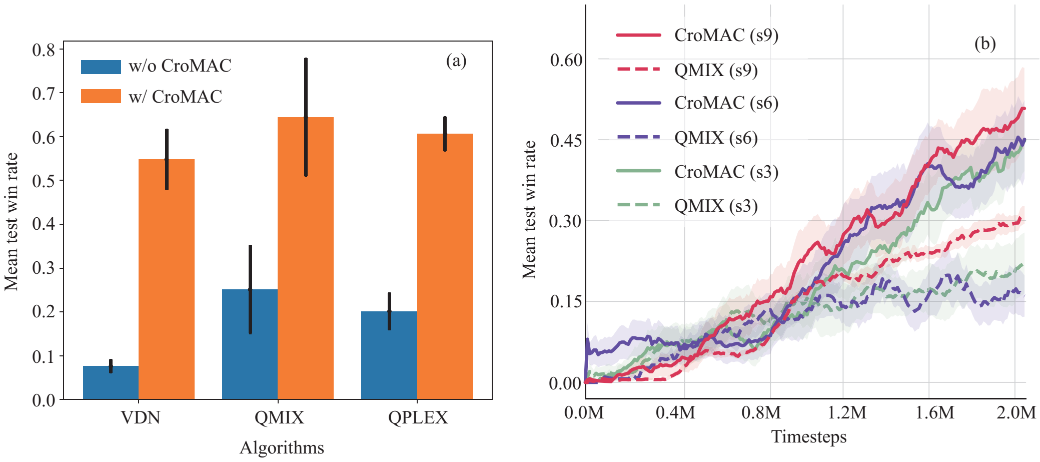


Figure 5 (Color online) Average test success rates of CroMAC implemented with different value-based MARL methods along with performance comparison with varying sights, where sn means the sight range is n and the default sight range is 9. (a) LBF: 3p-1f; (b) SMAC: 1o10b_vs_1r.

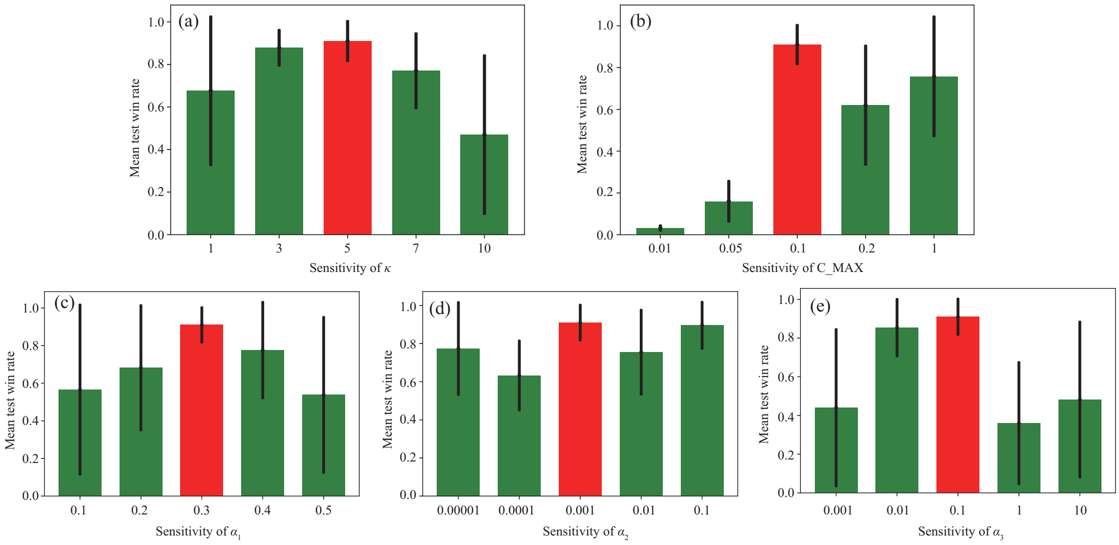


Figure 6 (Color online) Sensitivity of hyperparameters used in this paper. (a) κ; (b) C_MAX; (c) α_1; (d) α_2; (e) α_3.

turbations. However, AME sustains a drastic performance degradation when we alter the perturbation budget like FGSM (4) or other perturbation models like PGD in the hallway environment. On the other hand, our CroMAC achieves high superiority over AME in most environments under different perturbations, demonstrating its high generalization ability when encountering different perturbation budgets and perturbation methods.

5.4 Generality and parameter sensitivity

CroMAC is a robust communication training paradigm and is agnostic to specific value decomposition MARL methods and any sight conditions. Here we treat it as a plug-in module and integrate it with existing MARL value decomposition methods like VDN [7], QMIX [8], and QPLEX [9]. As shown in Figure 5(a), when integrating with CroMAC, the performance of the baselines vastly improves on scenario LBF: 3p-1f with message perturbations, indicating that the proposed training paradigm can significantly enhance robustness for various MARL methods. It is worth mentioning that QPLEX shows instability when adding robust loss and we choose the best result in the training process for comparison. Furthermore, when we alter the agents' sight range in the SMAC map 1o10b_vs_1r, the results shown in Figure 5(b) also demonstrate that CroMAC can improve the coordination robustness in different sight conditions for QMIX with communication, showing its high generality under different communication scenarios.

As CroMAC includes multiple hyperparameters, here we conduct experiments on scenario hallway: 4x5x6 to investigate how each one influences the robustness. Where κ controls the attack strength added to the latent state space. If it is too small, we cannot guarantee good robustness in the testing phase, and if it is too large, the policy may be too smooth and not optimal anymore. As shown in Figure 6(a), we can find that $\kappa = 5$ is the best choice in this scenario. Furthermore, the value range of the network weight W is used to get an approximate bound of the integration error. Figure 6(b) shows that C_MAX= 0.1 performs best. We can find the most appropriate parameters for other scenarios in the same way. More details for other scenarios are shown in Appendix D. Besides, as there are multiple hyperparameters of each loss function, we show how each adjustable hyperparameter named α_1, α_2, and α_3 influences the robustness of CroMAC. We continue to conduct experiments on the task hallway: 4x5x6 to investigate how each hyperparameter α influences the robustness. As shown in Figures 6(c)–(e), we can find that when the parameter is slightly larger or smaller, the performance may suffer corresponding degradation and the stability will also decline.

6 Conclusion and future work

Considering the great significance of robustness for real-world policy deployment and the enormous potential of MARL, this paper takes a further step towards robustness in MARL communication. We first model the multi-agent communication as a multi-view problem and apply a multi-view variational autoencoder that uses a product-of-experts inference network to obtain a joint message representation from the received messages; then a certificate guarantee between the joint message representation and each received message is obtained via interval bound propagation. For the optimization phase, we first encode the state into a latent space, and do perturbations in this space to get a certificate state representation. Then the learned joint message representation is used to approximate the certificate state representation. Extensive experimental results from multiple aspects demonstrate the efficiency of the proposed method. In terms of possible future work, as we learn the communication policy online, how we can learn a robust communication policy in offline MARL is challenging but of great value.

Acknowledgements This work was supported in part by National Key Research and Development Program of China (Grant No. 2020AAA0107200), National Natural Science Foundation of China (Grant Nos. 61921006, 61876119, 62276126), Natural Science Foundation of Jiangsu (Grant No. BK20221442), and Program B for Outstanding PhD Candidate of Nanjing University. We thank Ziqian ZHANG and Fuxiang ZHANG for their useful suggestions.

Supporting information Appendixes A–D. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Busoniu L, Babuska R, de Schutter B. A comprehensive survey of multiagent reinforcement learning. IEEE Trans Syst Man Cybern C, 2008, 38: 156–172

2 Zhang K, Yang Z, Başar T. Multi-agent reinforcement learning: a selective overview of theories and algorithms. In: Handbook of Reinforcement Learning and Control. Berlin: Springer, 2021. 321–384

3 Oroojlooy A, Hajinezhad D. A review of cooperative multi-agent deep reinforcement learning. Appl Intell, 2023, 53: 13677–13722

4 Wang J, Xu W, Gu Y, et al. Multi-agent reinforcement learning for active voltage control on power distribution networks. In: Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS), 2021. 3271–3284

5 Yun W J, Park S, Kim J, et al. Cooperative multiagent deep reinforcement learning for reliable surveillance via autonomous multi-UAV control. IEEE Trans Ind Inf, 2022, 18: 7086–7096

6 Xue K, Xu J, Yuan L, et al. Multi-agent dynamic algorithm configuration. In: Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS), 2022

7 Sunehag P, Lever G, Gruslys A, et al. Value-decomposition networks for cooperative multi-agent learning based on team reward. In: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS), 2018. 2085–2087

8 Rashid T, Samvelyan M, Schroeder C, et al. QMIX: monotonic value function factorisation for deep multi-agent reinforcement learning. In: Proceedings of the 35th International Conference on Machine Learning (ICML), 2018. 4295–4304

9 Wang J, Ren Z, Liu T, et al. QPLEX: duplex dueling multi-agent Q-learning. In: Proceedings of the International Conference on Learning Representations (ICLR), 2021

10 Lowe R, Wu Y, Tamar A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS), 2017. 6382–6393

11 Yu C, Velu A, Vinitsky E, et al. The surprising effectiveness of PPO in cooperative multi-agent games. 2021. ArXiv:2103.01955

12 Ye J, Li C, Wang J, et al. Towards global optimality in cooperative marl with sequential transformation. 2022. ArXiv:2207.11143

13 Wang Y, Han B, Wang T, et al. DOP: off-policy multi-agent decomposed policy gradients. In: Proceedings of the International Conference on Learning Representations (ICLR), 2021

14 Cao J, Yuan L, Wang J, et al. LINDA: multi-agent local information decomposition for awareness of teammates. 2021. ArXiv:2109.12508

15 Yuan L, Wang C, Wang J, et al. Multi-agent concentrative coordination with decentralized task representation. In: Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI), 2022. 599–605

16 Wen M, Kuba J G, Lin R, et al. Multi-agent reinforcement learning is a sequence modeling problem. In: Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS), 2022

17 Papoudakis G, Christianos F, Schäfer L, et al. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. In: Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS), 2021

18 Gorsane R, Mahjoub O, de Kock R, et al. Towards a standardised performance evaluation protocol for cooperative MARL. In: Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS), 2022

19 Moos J, Hansel K, Abdulsamad H, et al. Robust reinforcement learning: a review of foundations and recent advances. Mach Learn Know Extr, 2022, 4: 276–315

20 Guo J, Chen Y, Hao Y, et al. Towards comprehensive testing on the robustness of cooperative multi-agent reinforcement learning. 2022. ArXiv:2204.07932

21 Pinto L, Davidson J, Sukthankar R, et al. Robust adversarial reinforcement learning. In: Proceedings of the 34th International Conference on Machine Learning (ICML), 2017. 2817–2826

22 Zhang H, Chen H, Xiao C, et al. Robust deep reinforcement learning against adversarial perturbations on state observations. In: Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS), 2020. 21024–21037

23 Vinitsky E, Du Y, Parvate K, et al. Robust reinforcement learning using adversarial populations. 2020. ArXiv:2008.01825

24 Song Y, Schneider J. Robust reinforcement learning via genetic curriculum. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 2022. 5560–5566

25 Yu J, Gehring C, Schäfer F, et al. Robust reinforcement learning: a constrained game-theoretic approach. In: Proceedings of the 3rd Conference on Learning for Dynamics and Control (L4DC), 2021. 1242–1254

26 Oikarinen T, Zhang W, Megretski A, et al. Robust deep reinforcement learning through adversarial loss. In: Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS), 2021. 26156–26167

27 Sun Y, Zheng R, Liang Y, et al. Who is the strongest enemy? Towards optimal and efficient evasion attacks in deep RL. In: Proceedings of the 10th International Conference on Learning Representations (ICLR), 2021

28 Wu J, Vorobeychik Y. Robust deep reinforcement learning through bootstrapped opportunistic curriculum. In: Proceedings of the 39th International Conference on Machine Learning (ICML), 2022. 24177–24211

29 Everett M, Lütjens B, How J P. Certifiable robustness to adversarial state uncertainty in deep reinforcement learning. IEEE Trans Neural Netw Learn Syst, 2022, 33: 4184–4198

30 Wu F, Li L, Huang Z, et al. CROP: certifying robust policies for reinforcement learning through functional smoothing. In: Proceedings of the International Conference on Learning Representations (ICLR), 2021

31 Sun C, Kim D K, How J P. ROMAX: certifiably robust deep multiagent reinforcement learning via convex relaxation. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 2022. 5503–5510

32 Dorri A, Kanhere S S, Jurdak R. Multi-agent systems: a survey. IEEE Access, 2018, 6: 28573–28593

33 Christianos F, Papoudakis G, Rahman A, et al. Scaling multi-agent reinforcement learning with selective parameter sharing. In: Proceedings of the 38th International Conference on Machine Learning (ICML), 2021. 1989–1998

34 van der Heiden T, Salge C, Gavves E, et al. Robust multi-agent reinforcement learning with social empowerment for coordination and communication. 2020. ArXiv:2012.08255

35 Li S, Wu Y, Cui X, et al. Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2019. 4213–4220

36 Zhou Z, Liu G. RoMFAC: a robust mean-field actor-critic reinforcement learning against adversarial perturbations on states. 2022. ArXiv:2205.07229

37 Zhang K, Sun T, Tao Y, et al. Robust multi-agent reinforcement learning with model uncertainty. In: Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS), 2020. 10571–10583

38 Hu Y, Shao K, Li D, et al. Robust multi-agent reinforcement learning driven by correlated equilibrium. In: Proceedings of the International Conference on Learning Representations (ICLR), 2021

39 Hu Y, Zhang Z. Sparse adversarial attack in multi-agent reinforcement learning. 2022. ArXiv:2205.09362

40 Zhu C, Dastani M, Wang S. A survey of multi-agent reinforcement learning with communication. 2022. ArXiv:2203.08975

41 Blumenkamp J, Prorok A. The emergence of adversarial communication in multi-agent reinforcement learning. In: Proceedings of the 4th Conference on Robot Learning (CoRL), 2021. 1394–1414

42 Sun Y, Zheng R, Hassanzadeh P, et al. Certifiably robust policy learning against adversarial communication in multi-agent systems. 2022. ArXiv:2206.10158

43 Xue W, Qiu W, An B, et al. Mis-spoke or mis-lead: achieving robustness in multi-agent communicative reinforcement learning. In: Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS), 2022. 1418–1426

44 MacKay D J C. Information Theory, Inference, and Learning Algorithms. Cambridge: Cambridge University Press, 2003

45 Wang T, Wang J, Zheng C, et al. Learning nearly decomposable value functions via communication minimization. In: Proceedings of the 8th International Conference on Learning Representations (ICLR), 2020

46 Das A, Gervet T, Romoff J, et al. TarMAC: targeted multi-agent communication. In: Proceedings of the 36th International Conference on Machine Learning (ICML), 2019. 1538–1546

47 Foerster J N, Assael Y M, de Freitas N, et al. Learning to communicate with deep multi-agent reinforcement learning. In: Proceedings of the 30th International Conference on Neural Information Processing Systems (NeurIPS), 2016. 2145–2153

48 Sukhbaatar S, Szlam A, Fergus R. Learning multiagent communication with backpropagation. In: Proceedings of the 30th International Conference on Neural Information Processing Systems (NeurIPS), 2016. 2252–2260

49 Mao H, Zhang Z, Xiao Z, et al. Learning agent communication under limited bandwidth by message pruning. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2020. 5142–5149

50 Ding Z, Huang T, Lu Z. Learning individually inferred communication for multi-agent cooperation. In: Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS), 2020

51 Xue D, Yuan L, Zhang Z, et al. Efficient multi-agent communication via shapley message value. In: Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI), 2022. 578–584

52 Mao H, Zhang Z, Xiao Z, et al. Learning multi-agent communication with double attentional deep reinforcement learning. Auton Agent Multi-Ag, 2020, 34: 32

53 Wang Y, Xu J, Wang Y, et al. ToM2C: target-oriented multi-agent communication and cooperation with theory of mind. In: Proceedings of the International Conference on Learning Representations (ICLR), 2021

54 Zhang S Q, Zhang Q, Lin J. Efficient communication in multi-agent reinforcement learning via variance based control. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS), 2019. 3235–3244

55 Zhang S Q, Zhang Q, Lin J. Succinct and robust multi-agent communication with temporal message control. In: Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS), 2020. 17271–17282

56 Yuan L, Wang J, Zhang F, et al. Multi-agent incentive communication via decentralized teammate modeling. In: Proceedings of the 36th AAAI Conference on Artificial Intelligence, 2022. 9466–9474

57 Pan X, Seita D, Gao Y, et al. Risk averse robust adversarial reinforcement learning. In: Proceedings of the International Conference on Robotics and Automation (ICRA), 2019. 8522–8528

58 Liang Y, Sun Y, Zheng R, et al. Efficient adversarial training without attacking: worst-case-aware robust reinforcement learning. In: Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS), 2022

59 Qin Z, Zhang K, Chen Y, et al. Learning safe multi-agent control with decentralized neural barrier certificates. In: Proceedings of the International Conference on Learning Representations (ICLR), 2020

60 Wu F, Li L, Zhang H, et al. COPA: certifying robust policies for offline reinforcement learning against poisoning attacks. In: Proceedings of the International Conference on Learning Representations (ICLR), 2021

61 Park H, Lee S, Lee J, et al. Learning by aligning: visible-infrared person re-identification using cross-modal correspondences. In: Proceedings of the International Conference on Computer Vision (ICCV), 2021. 12046–12055

62 Chen N, Zhu J, Sun F, et al. Large-margin predictive latent subspace learning for multiview data analysis. IEEE Trans Pattern Anal Mach Intell, 2012, 34: 2365–2378

63 Xu J, Li W, Liu X, et al. Deep embedded complementary and interactive information for multi-view classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2020. 6494–6501

64 Kingma D P, Welling M. Auto-encoding variational Bayes. In: Proceedings of the International Conference on Learning Representations (ICLR), 2014

65 Sohn K, Lee H, Yan X. Learning structured output representation using deep conditional generative models. In: Proceedings of the 28th International Conference on Neural Information Processing Systems (NeurIPS), 2015. 3483–3491

66 Wu M, Goodman N. Multimodal generative models for scalable weakly-supervised learning. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS), 2018. 5580–5590

67 Suzuki M, Nakayama K, Matsuo Y. Joint multimodal learning with deep generative models. 2016. ArXiv:1611.01891

68 Yan X, Hu S, Mao Y, et al. Deep multi-view learning methods: a review. Neurocomputing, 2021, 448: 106–129

69 Bayoudh K, Knani R, Hamdaoui F, et al. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. Vis Comput, 2022, 38: 2939–2970

70 Ma M, Ren J, Zhao L, et al. SMIL: multimodal learning with severely missing modality. In: Proceedings of the 35th AAAI Conference on Artificial Intelligence, 2021. 2302–2310

71 Xu C, Tao D, Xu C. Multi-view learning with incomplete views. IEEE Trans Image Process, 2015, 24: 5812–5825

72 Li M, Wu L, Wang J, et al. Multi-view reinforcement learning. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS), 2019. 1420–1431

73 Fan J, Li W. DRIBO: robust deep reinforcement learning via multi-view information bottleneck. In: Proceedings of the 39th International Conference on Machine Learning (ICML), 2022. 6074–6102

74 Kinose A, Okada M, Okumura R, et al. Multi-view dreaming: multi-view world model with contrastive learning. 2022. ArXiv:2203.11024

75 Gronauer S, Diepold K. Multi-agent deep reinforcement learning: a survey. Artif Intell Rev, 2022, 55: 895–943

76 Papoudakis G, Christianos F, Rahman A, et al. Dealing with non-stationarity in multi-agent deep reinforcement learning. 2019. ArXiv:1906.04737

77 Wang J, Ren Z, Han B, et al. Towards understanding cooperative multi-agent Q-learning with value factorization. In: Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS), 2021. 29142–29155

78 Lin J, Dzeparoska K, Zhang S Q, et al. On the robustness of cooperative multi-agent reinforcement learning. In: Proceedings of the IEEE Security and Privacy Workshops (SPW), 2020. 62–68

79 Mitchell R, Blumenkamp J, Prorok A. Gaussian process based message filtering for robust multi-agent cooperation in the presence of adversarial communication. 2020. ArXiv:2012.00508

80 Tu J, Wang T, Wang J, et al. Adversarial attacks on multi-agent communication. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021. 7768–7777

81 Cho K, van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014. 1724–1734

82 Qiaoben Y, Ying C, Zhou X, et al. Understanding adversarial attacks on observations in deep reinforcement learning. 2021. ArXiv:2106.15860

83 Xu M, Liu Z, Huang P, et al. Trustworthy reinforcement learning against intrinsic vulnerabilities: robustness, safety, and generalizability. 2022. ArXiv:2209.08025

84 Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proceedings of the International Conference on Learning Representations (ICLR), 2015

85 Gowal S, Dvijotham K, Stanforth R, et al. On the effectiveness of interval bound propagation for training verifiably robust models. 2018. ArXiv:1810.12715

86 Li Y, Yang M, Zhang Z. A survey of multi-view representation learning. IEEE Trans Knowl Data Eng, 2018, 31: 1863–1883

87 Hwang H, Kim G H, Hong S, et al. Multi-view representation learning via total correlation objective. In: Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS), 2021. 12194–12207

88 Cao Y, Fleet D J. Generalized product of experts for automatic and principled fusion of Gaussian process predictions. 2014. ArXiv:1410.7827

89 Gautschi W. A harmonic mean inequality for the Gamma function. SIAM J Math Anal, 1974, 5: 278–281

90 Kraemer L, Banerjee B. Multi-agent reinforcement learning as a rehearsal for decentralized planning. Neurocomputing, 2016, 190: 82–94

91 Lyu X, Xiao Y, Daley B, et al. Contrasting centralized and decentralized critics in multi-agent reinforcement learning. In: Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS), 2021. 844–852

92 Tipping M E, Bishop C M. Probabilistic principal component analysis. J Royal Stat Soc B, 1999, 61: 611–622