Certifiably Robust Multi-Agent Reinforcement Learning against Adversarial Communication

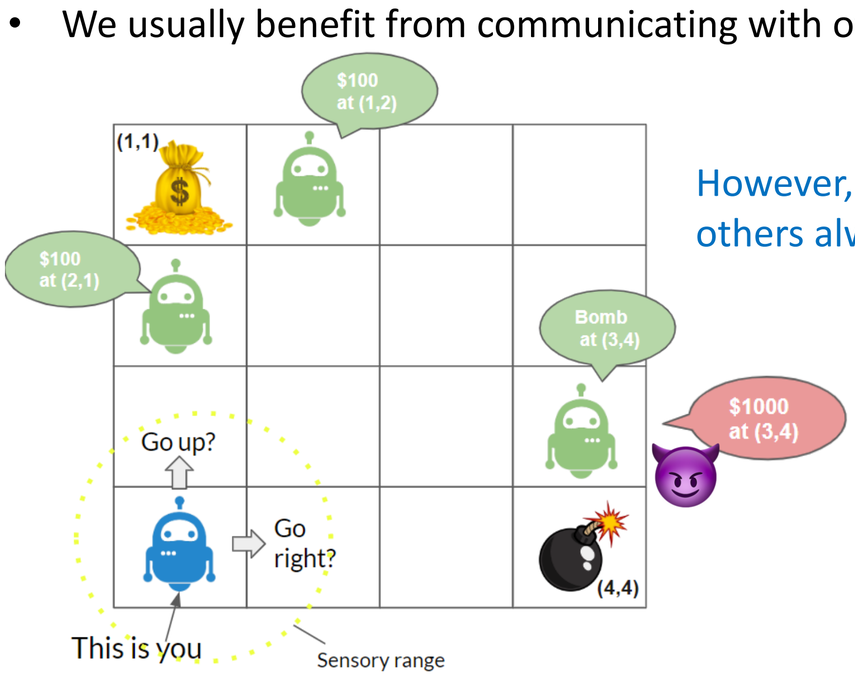
Yanchao Sun[†] (*ycs@umd.edu*), Ruijie Zheng[†], Parisa Hassanzadeh[‡], Yongyuan Liang[ɗ], Soheil Feizi[†], Sumitra Ganesh[‡], Furong Huang[†]

[†]University of Maryland, College Park [‡]JPMorgan AI Research [ɗ]Sun Yat-sen University

Introduction & Problem Setup

❑ Motivation: Adversarial Communication in MARL

- We usually benefit from communicating with other agents.



However, is trusting others always safe?

- A communication-dependent policy may be *unsafe* when communications get perturbed, deliberately or not.

➢ Communication is a double-edged sword in multi-agent systems!

❑ Setup: Communicative MARL with Test-time Attack

- A partially-observable environment where agents are trained to communicate.
- Clean and safe training-time communication.
- Probably perturbed test-time communication.

❑ Goal: Make Communication-driven Policies Robust

Challenging because …

- Communication attacks can be *stealthy*, and can take *any form* (e.g. replace a word in a sentence).
- Attackers can be *adaptive* to the defender's policy.
- There can be *multiple attackers* collaboratively perturbing communication messages.
- Trade-off between natural performance and robustness.

No trust → no communication benefit Trust all → unsafe to perturbations

➢ Our goal: let the agent benefit from benign communication while being robust under adversarial communication.

Proposed Method

❑ Algorithm: Ablated Message Ensemble (AME)

- Basic notations

  $\boldsymbol{N}$ **agents** in the system (assume symmetry for simplicity)
  At every step, each agent receives $\boldsymbol{N-1}$ **messages** from all others.

- Basic assumption

  In deployment (test time), adversaries may arbitrarily perturb up to C out of $N-1$ messages, where $C < \frac{N-1}{2}$.

  Good News: 1) up to half of messages can be adversarial.
  2) arbitrary perturbations are allowed.

- Algorithm Design

  ➢ Idea: making decisions based on the **consensus** of communication messages.

Training Time (no attack)

For each agent, train a message-ablation policy

$$\hat{\pi}: \Gamma \times M^k \to A$$

(Obs. history; k ≤ $N-1$ messages)

such that $\hat{\pi}$ obtains high reward with randomly ablated size-k message subset $m^k \sim m$.

Test Time (under attack)

Make decisions with a message-ensemble policy

$$\tilde{\pi}: \Gamma \times M^{N-1} \to A$$

where for a discrete action space

$$\tilde{\pi}(\tau, m) \coloneqq \text{argmax}_a \sum_{\forall m^k} 1\left(\hat{\pi}(\tau, m^k) = a\right)$$

and for a continuous action space

$$\tilde{\pi}(\tau, m) \coloneqq \text{Median}\left\{\hat{\pi}(\tau, m^k), \forall m^k\right\}$$

Remarks:
(1) training a message-ablation policy is computationally efficient (input space smaller than the original one).
(2) During test time, $\tilde{\pi}$ traverses through $\binom{N-1}{k}$ message subsets (m^k), which is not expensive when N is relatively small.
(3) If N is large, there is a partial-sample version of AME, in which $\tilde{\pi}$ samples $D \leq \binom{N-1}{k}$ message subsets. The guarantee still holds with a probability corresponding to D.
(4) k is a hyper-parameter that controls the trade-off between natural performance and robustness. A smaller k trades natural reward off for robustness.

❑ Theoretical Guarantees: Certifiable Robustness

Under mild conditions (for the selection of k), it is guaranteed that

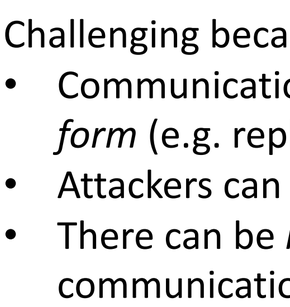
- The message-ensemble policy selects a "safe action" that is suggested by some benign messages.
- Cumulative reward $R_{attacked}(\tilde{\pi}) \approx R_{natural}(\hat{\pi})$.

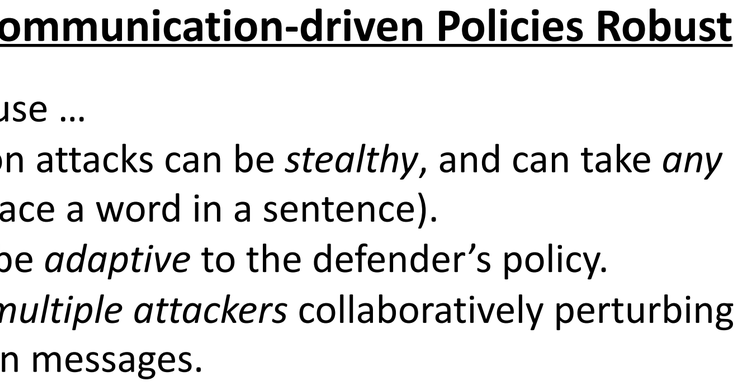
Experiment Results

➢ AME makes agents robust while still benefiting from communication.

➢ Robustness hold even when theoretical conditions are not satisfied.

○ Environment 1: FoodCollector

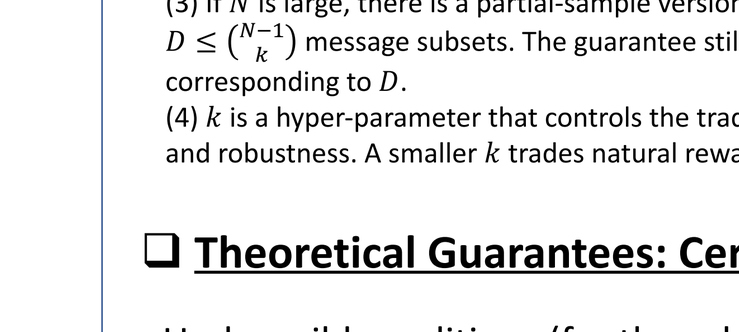




Scenarios where guarantee holds

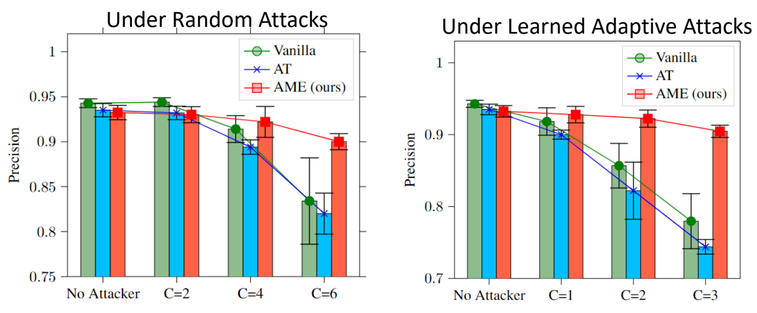
Scenarios where guarantee does not hold

○ Environment 2: InvertoryManager



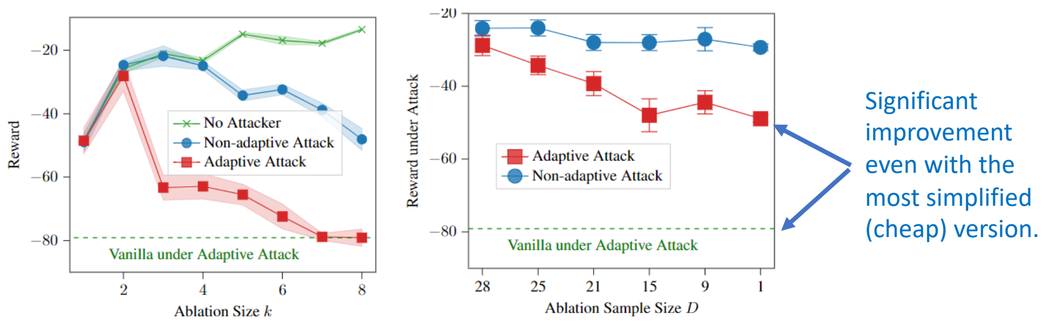
- **Vanilla**: train with no defense
- **AT** (Adversarial Training): train attacker and agent alternately
- AME (ours): ensemble-based (k=2)

○ Environment 3: MARL Classification on MNIST



➢ AME works well with a wide range of hyperparameter settings.

k controls the trade-off between performance and robustness



Significant improvement even with the most simplified (cheap) version.

ArXiv link: *https://arxiv.org/abs/2206.10158*