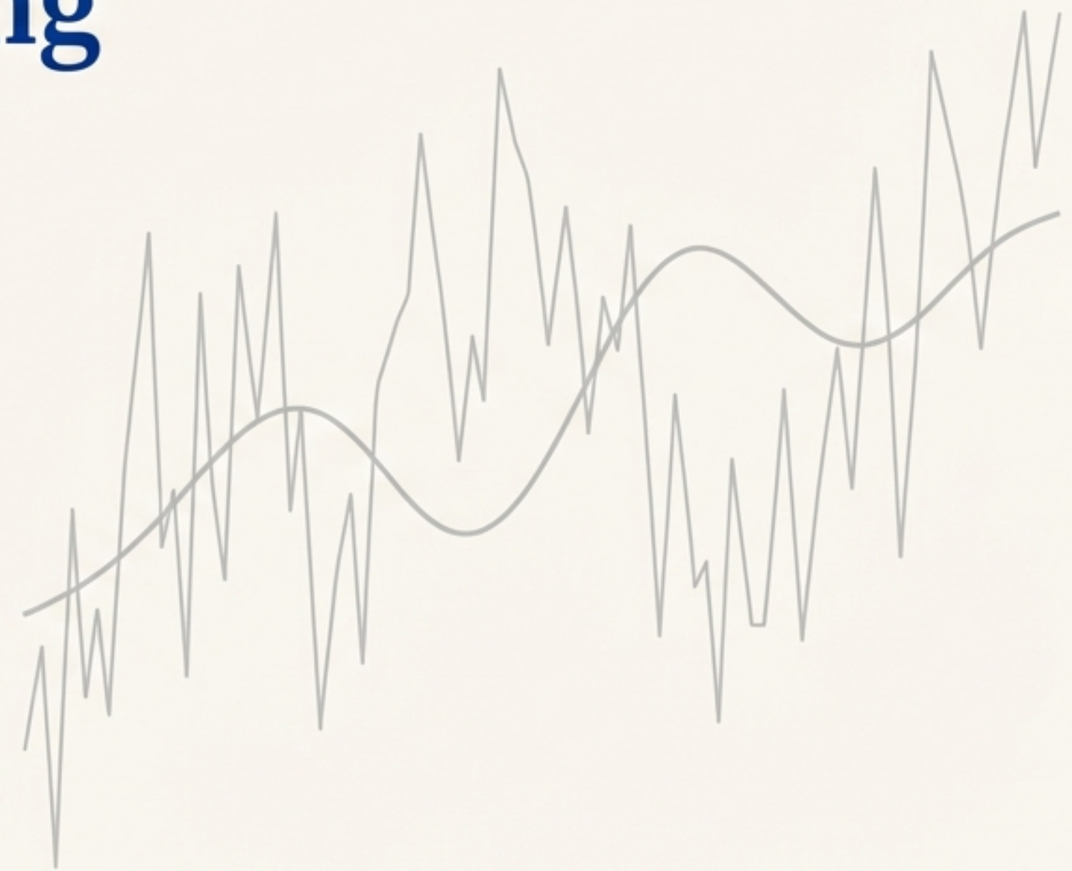
Robust Multi-Agent Reinforcement Learning via Adversarial Regularization

A Technical Overview of the ERNIE Framework

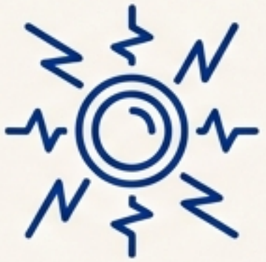


Based on the work by Alexander Bukharin, Yan Li, Yue Yu, et al.

Georgia Institute of Technology, Google, Microsoft, Ford Motor Company

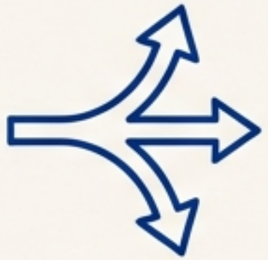
The Real-World Deployment Problem: MARL Policies are Brittle

Multi-Agent Reinforcement Learning (MARL) has achieved state-of-the-art results on complex tasks like StarCraft and traffic control. However, policies trained in fixed simulation environments often fail when deployed in the real world.



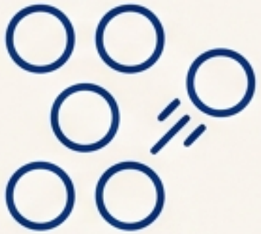
Observation Noise

Real-world sensors are imperfect. Even minor inaccuracies in state information can cause catastrophic performance degradation.



Changing Dynamics

The testing environment is rarely identical to the training environment. Slight changes in transition dynamics can lead to severe performance drops.



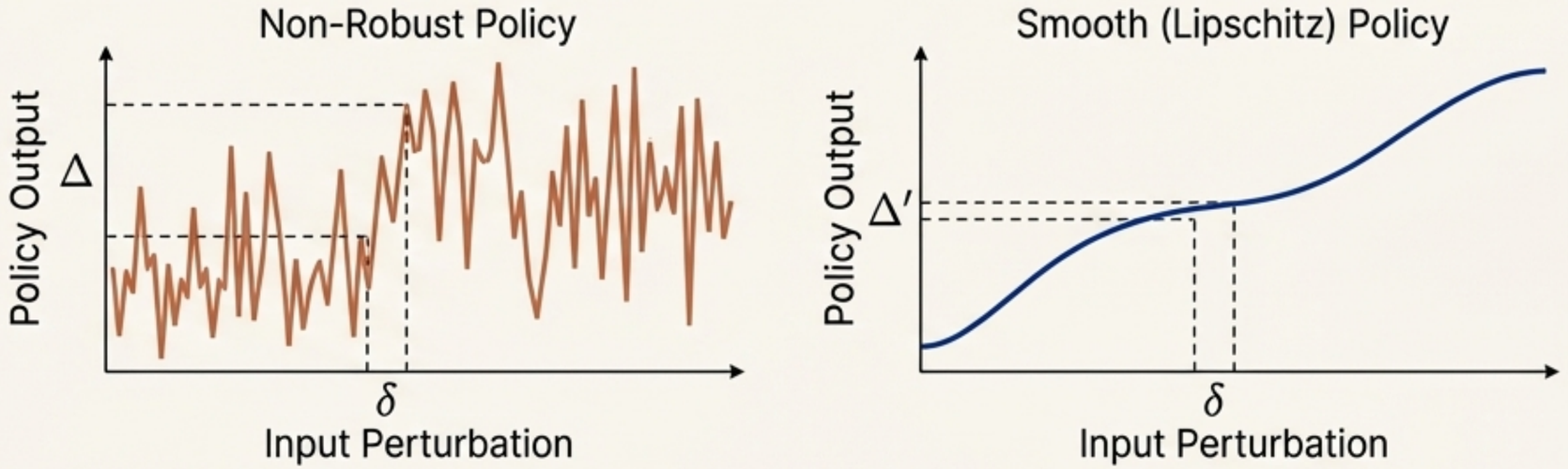
Malicious or Unpredictable Agents

A single agent acting differently than expected—maliciously or not—can destabilize the entire system through a chain reaction.

To move MARL from simulation to reality, we need a **principled approach to building robust policies.**

The Core Insight: Policy Smoothness is a Natural Precursor to Robustness

We can achieve robustness by controlling a policy's **Lipschitz constant**. A policy with a small Lipschitz constant is **"smooth"**—it does not change its output drastically in response to small changes in its input.



1. **Existence**: For any smooth environment, a smooth and close-to-optimal policy exists. We are not sacrificing performance by searching for a smooth policy. (Ref: Theorem 3.2)
2. **Guaranteed Robustness**: A policy's robustness against observation noise is **inversely** proportional to its **Lipschitz constant**. This holds *even if the environment is not smooth*. (Ref: Theorem 3.3)

Enforcing smoothness is not just an arbitrary heuristic; it is a **theoretically-grounded** strategy to **reduce the policy search space** and **directly improve robustness**.

Theoretical Foundations: Key Theorems on Smoothness and Optimality

Theorem 3.1: Smooth Environments Lead to Smooth Value Functions

If an environment's reward and transition functions are (L_r, L_P)-smooth, then the Q-function and Value function of any L_π-smooth policy are also Lipschitz continuous.

What this means: Smoothness propagates from the environment and policy to the value function itself. This suggests that in many real-world physical systems, smoothness is a natural property.

Theorem 3.2: Existence of Smooth, Near-Optimal Policies

For any smooth environment and any $\epsilon > 0$, there exists an ϵ-optimal policy π that is also $O(L_Q/\epsilon)$-smooth.

What this means: We can find a policy that is arbitrarily close to optimal while still being smooth. Searching for a smooth policy doesn't mean we have to compromise on performance.

Theorem 3.3: Smooth Policies are Provably Robust to Observation Noise

For an L_π-smooth policy, the difference in expected return between using the true state (s) and a perturbed state $(s + \delta)$ is bounded by $O(L_\pi * ||\delta||)$.

What this means: **This is the crucial link**. A smaller Lipschitz constant (L_π) directly translates to a smaller drop in performance when observations are noisy. This guarantee does not require the environment to be smooth.

The Solution: ERNIE (adv**E**rsarially **R**egularized multiage**N**t re**I**nforcement l**E**arning)

ERNIE encourages policy smoothness through **adversarial regularization**.

$$R_\pi(o_k;\theta_k) = \max_{||\delta||\leq\epsilon} D(\pi_{\theta_k}(o_k+\delta), \pi_{\theta_k}(o_k))$$

1. For each observation o_k, we find a perturbation δ.

2. This δ is **adversarially** chosen to maximize the difference D (e.g., KL divergence) between the policy's output for the original and perturbed observation.

3. By minimizing this regularizer during training, we force the policy to produce similar outputs for similar inputs, effectively minimizing its local Lipschitz constant and making it **smoother**.

Dual Benefits

Lipschitz Continuity

Directly encourages the learned policies to be smooth, improving robustness.

Rich Data Augmentation

Augmenting training data with 'worst-case' adversarial examples provides broad coverage of unseen scenarios, improving generalization.

A More Stable Algorithm: Reformulating Adversarial Training as a Stackelberg Game

The Problem with Standard Adversarial Regularization

- It is formulated as a zero-sum game between the policy (defender) and the perturbation (attacker).
- This formulation is a nonconvex-nonconcave minimax problem, which is notoriously unstable to optimize.
- In MARL, which is already prone to training instability, this effect is amplified.

The ERNIE Solution: A Leader-Follower Game

We reformulate the problem as a Stackelberg game.

Why This Works

1. **Smoother Optimization**: The leader's optimization problem becomes smoother and better-conditioned, leading to a more stable training process.
2. **Better Data Fit**: Giving the policy priority allows for a better fit to the training data compared to standard adversarial training, improving performance even in lightly perturbed environments.

Extending the Framework for Comprehensive Robustness

Robustness Against Malicious Actions (ERNIE-A)

Goal: Ensure the system is not overly dependent on the actions of any single agent.

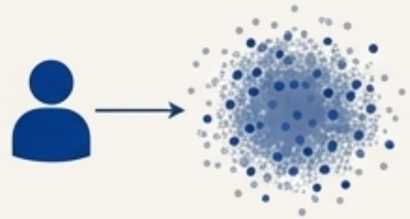
Method: Apply adversarial regularization to the joint action space.

Regularizer Intuition:

$$R_{A\omega}(s,a) = \max_{D(a,a')\leq K} ||Q(s,a;\omega) - Q(s,a';\omega)||^2$$

Explanation: We find the K agent actions that, if changed, would cause the largest change in the global Q-function. By minimizing this, we encourage the Q-function to be smooth with respect to individual agent actions, making the system robust to a few agents acting sub-optimally.

Scaling Up with Mean-Field MARL (ERNIE-MF)



Goal: Overcome the curse of many agents in large-scale systems.

Method: In mean-field MARL, agents react to an average agent's action ($\bar{a}_j$) and the distribution of states (d_s). We apply adversarial regularization directly to these mean-field terms.

Regularizer Intuition:

$$R_{QW}(s;\theta) = \max_{W(d'_s,d_s)\leq\epsilon} ||Q_\theta(s,d'_s,a) - Q_\theta(s,d_s,a)||^2$$

Explanation: We find the worst-case perturbation to the *distribution* of states (measured by Wasserstein distance) and train the Q-function to be robust against it. This provides robustness in a scalable manner.

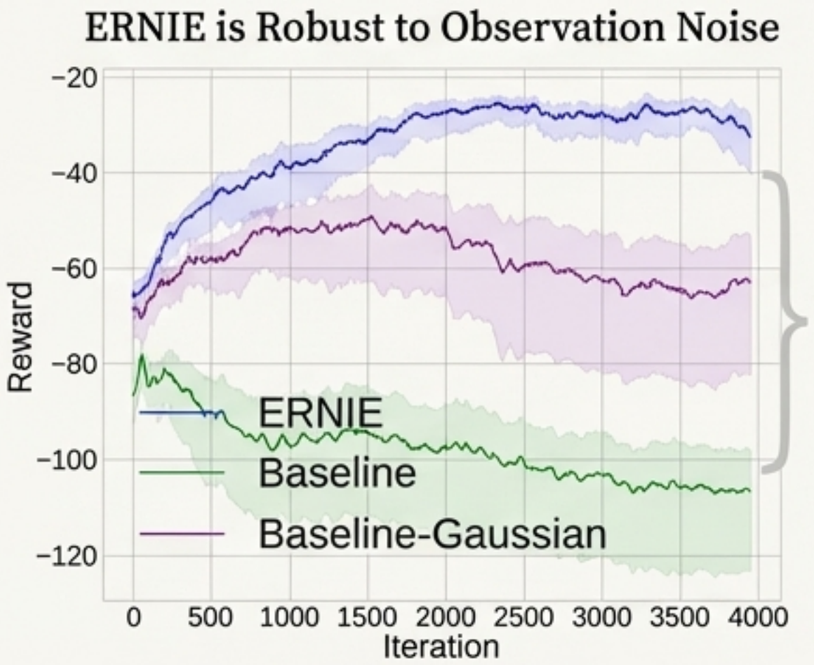
Evidence from the Field: ERNIE Excels in Dynamic Traffic Control Environments

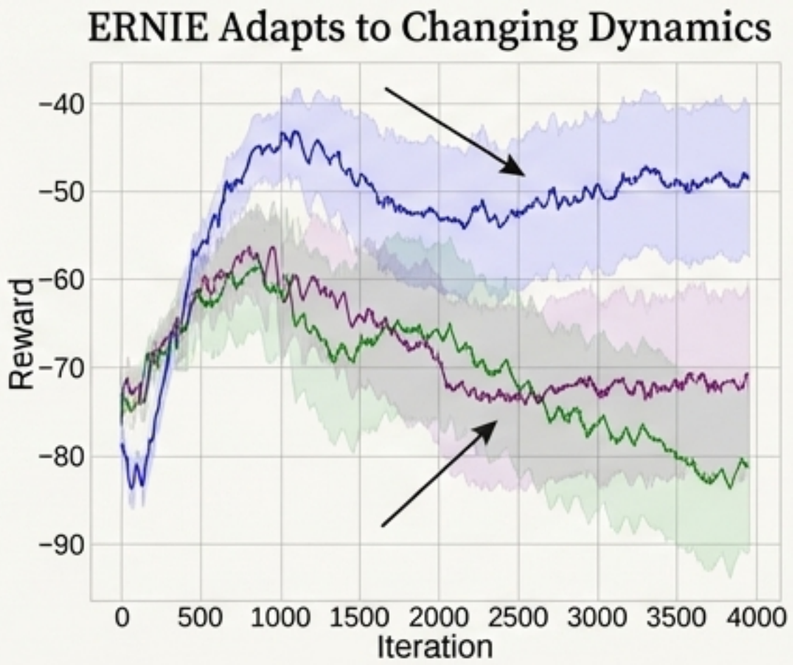
Experiment Setup

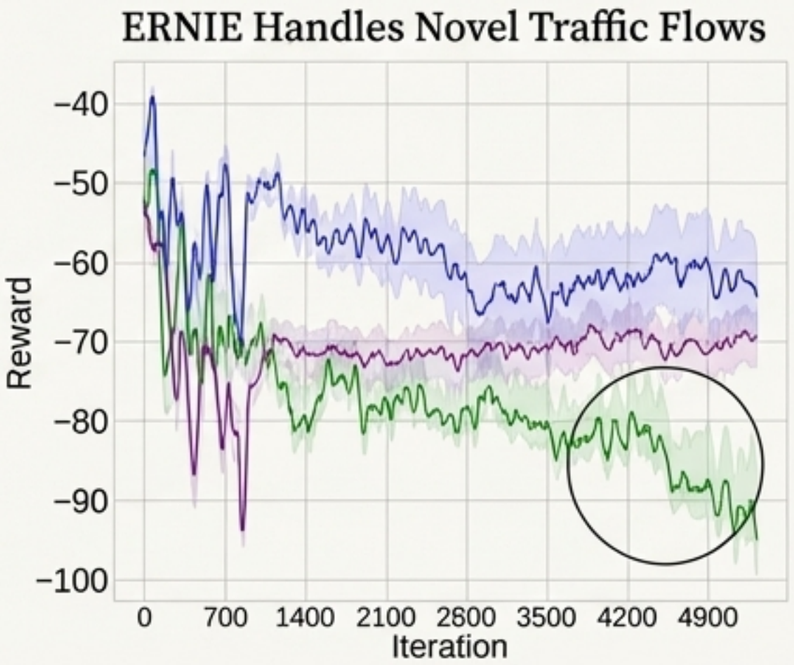
Agents are trained to control traffic lights in a 2×2 grid to minimize travel time. We then evaluate their performance when the environment changes from training to testing.

Key Findings

Standard MARL algorithms are highly sensitive to small environmental changes, while ERNIE maintains high performance.

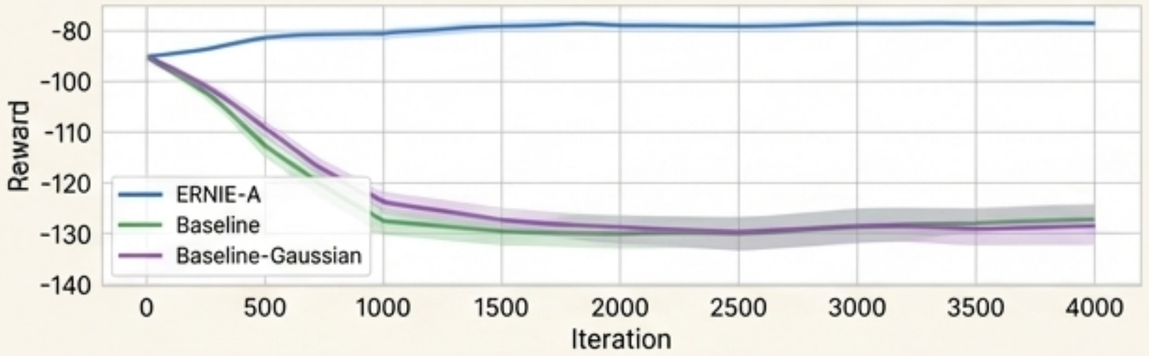






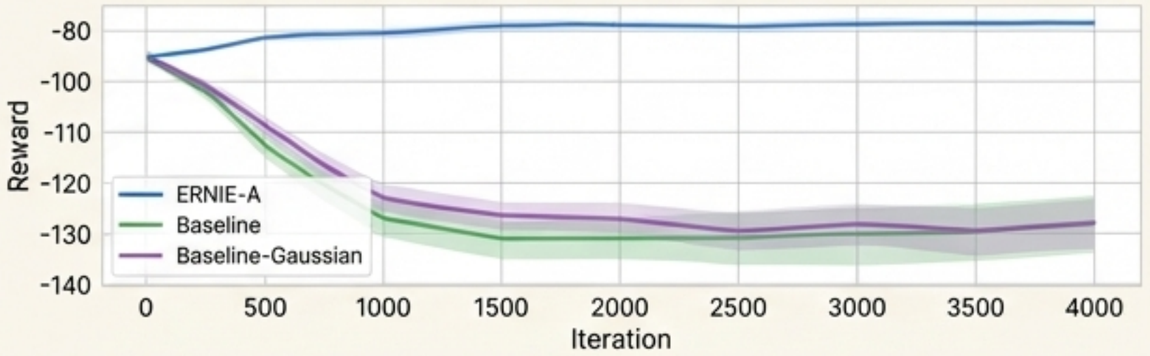
Maintaining System Integrity Against Malicious Agents and Network Changes

Robustness to Malicious Actions (3% Adversarial Actions)



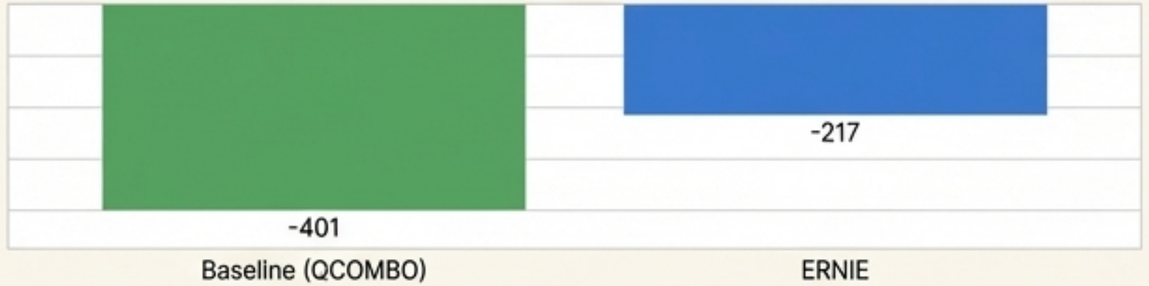
Setup: We adversarially change the action of a random agent 3% of the time.
Finding: ERNIE-A maintains high reward while baselines collapse.

Robustness to Malicious Actions (5% Adversarial Actions)



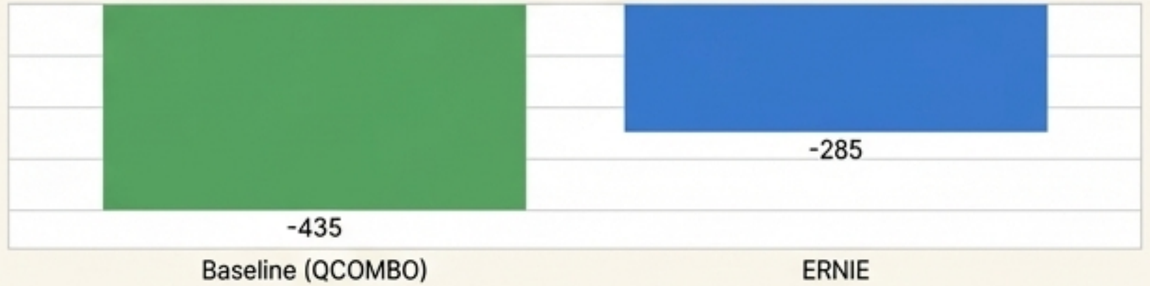
Setup: We adversarially change the action of a random agent 5% of the time.
Finding: The effect is even more pronounced with more frequent malicious actions. ERNIE-A's stability is clear.

Robustness to Larger Networks (4×4 Grid)



Setup: Policies trained on a 2x2 grid are evaluated on larger or irregular grids.
Finding: ERNIE significantiy outperforms the baseline when policies are scaled to a larger grid.

Robustness to Different Topologies (Atlanta Network)



Setup: Policies trained on a 2x2 grid are evaluated on an irregular Atlanta network.
Finding: ERNIE shows superior performance when the grid topology itself is changed.

The Stackelberg Advantage: Ablation Studies Confirm Increased Stability and Performance

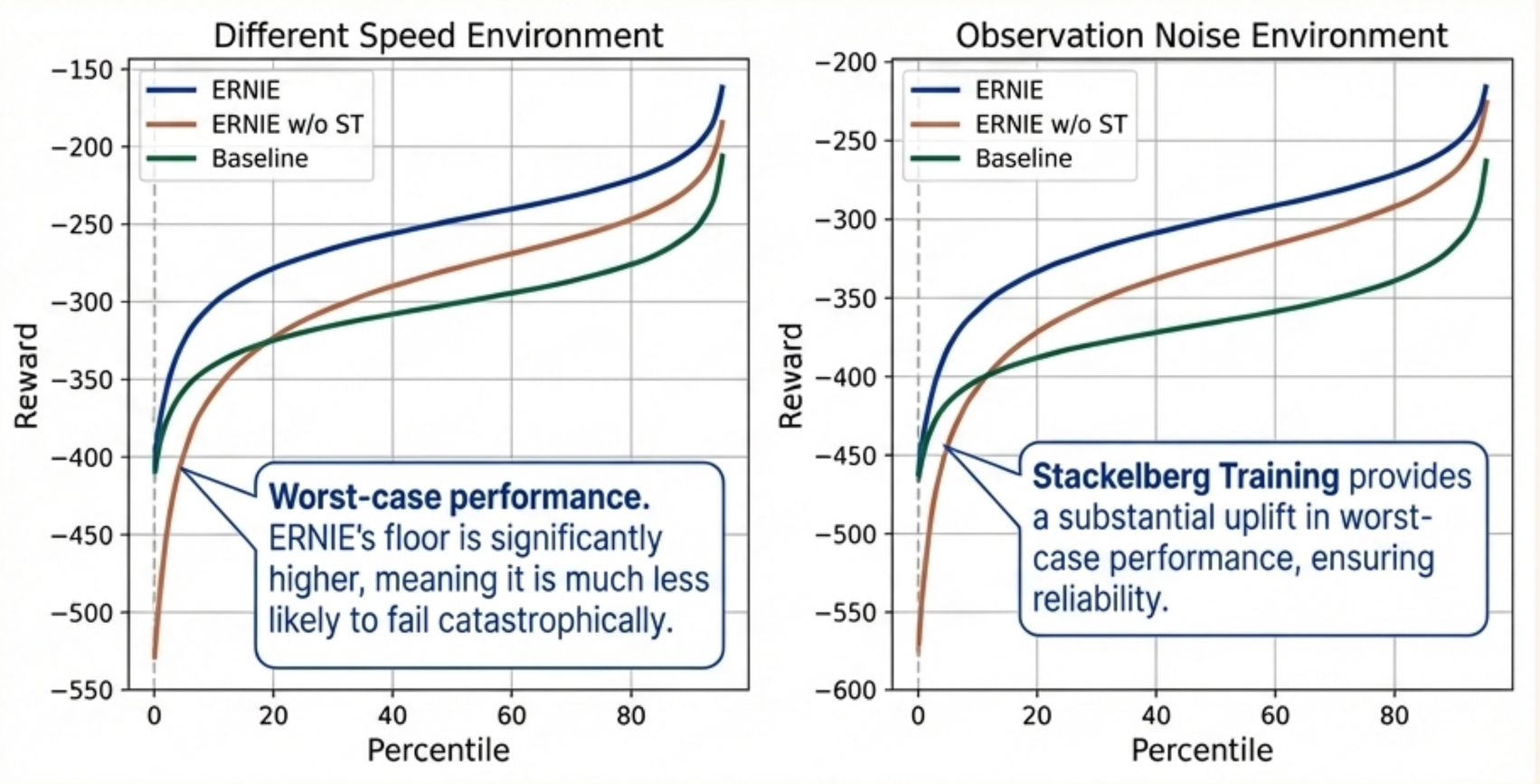
Experiment

We compare three algorithms across 10 different random initializations in perturbed environments:

1. **Baseline**: Standard MARL algorithm.
2. **ERNIE w/o ST**: ERNIE with standard adversarial regularization.
3. **ERNIE**: The full framework with Stackelberg Training (ST).

Methodology

We plot the sorted cumulative rewards (percentiles) to evaluate not just the mean performance, but the robustness to failure (i.e., the worst-case outcomes).

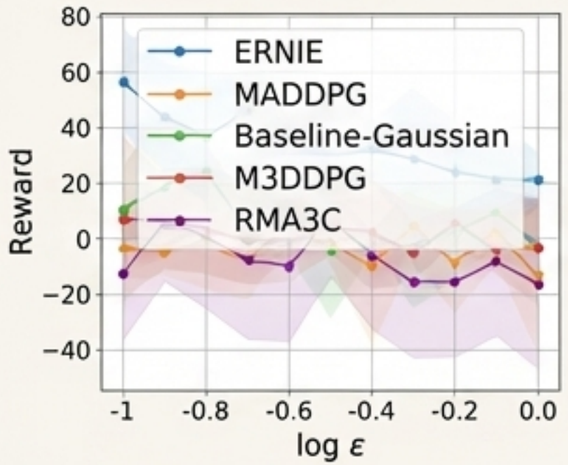


Conclusion: Stackelberg Training is not a minor tweak; it is essential for unlocking the full potential of adversarial regularization in MARL by stabilizing the training process.

Generalization Across Domains: Consistent Performance in Particle Environments

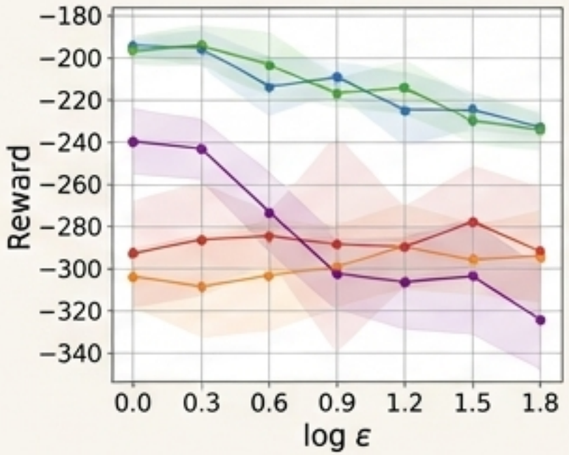
Experiment: We evaluate ERNIE (applied to MADDPG) against baselines including MADDPG, M3DDPG, and RMA3C in four cooperative particle games with varying levels of observation noise.

Covert Communication



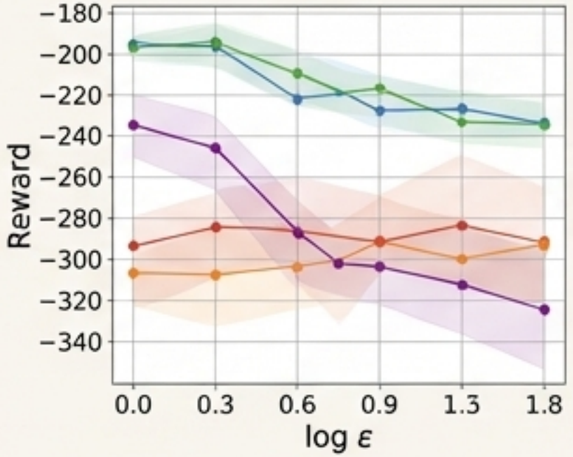
Finding: ERNIE consistently achieves the highest reward across all noise leveis, degrading gracefully.

Tag



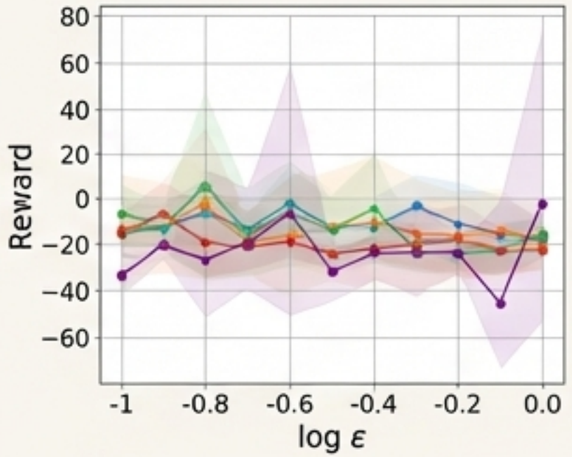
Finding: Clear separation in performance. ERNIE is substantially more robust than the baseline MADDPG and other methods.

Navigation



Finding: Clear separation in performance. ERNIE is substantially more robust than the baseline MADDPG and other other methods.

Predator Prey



Finding: ERNIE again shows superior performance and robustness to increasing observation noise.

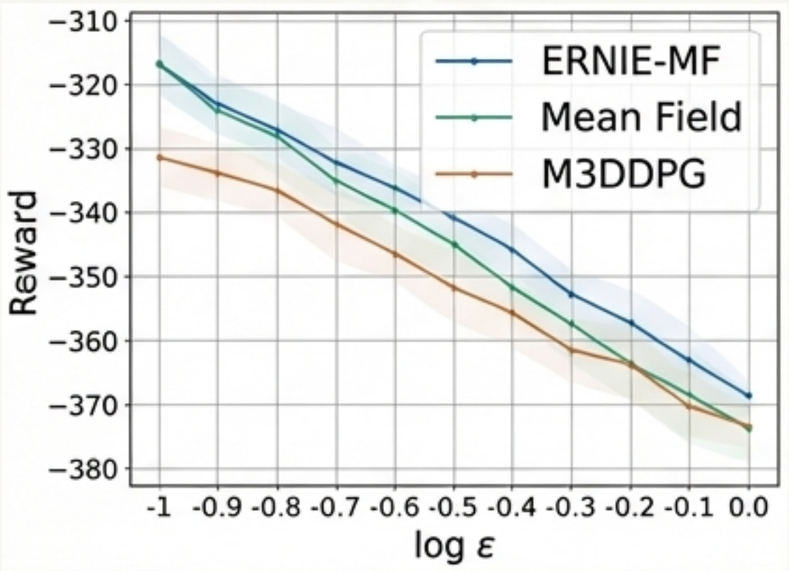
Overall Takeaway: ERNIE's performance benefit is not limited to one environment. It provides a consistent robustness advantage across a variety of multi-agent tasks.

ERNIE Scales: Robustness in the Face of Many Agents

Experiment & Key Insight

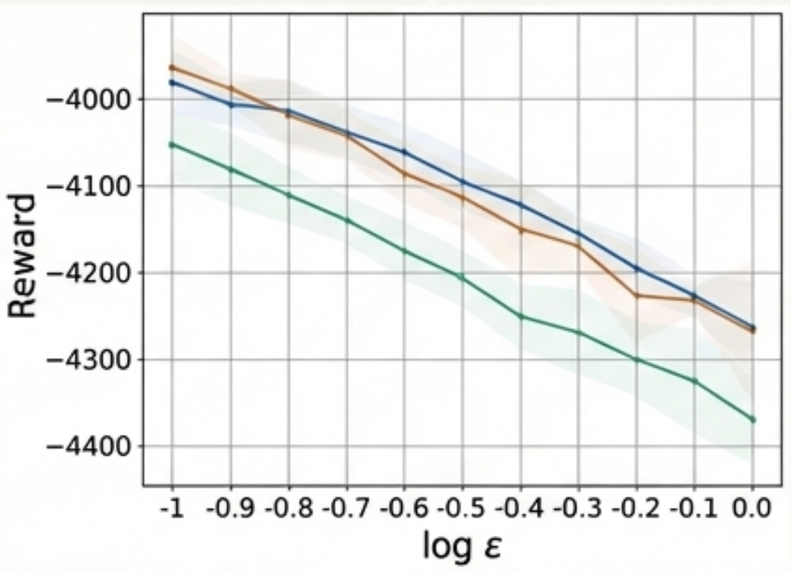
We evaluate ERNIE-MF (Mean-Field) on the cooperative navigation task with increasing numbers of agents (N=3, 6, 15) and varying levels of observation noise. The mean-field approximation avoids the curse of many agents, and ERNIE's adversarial regularization can be successfully applied to maintain robustness at scale.

N=3 Agents



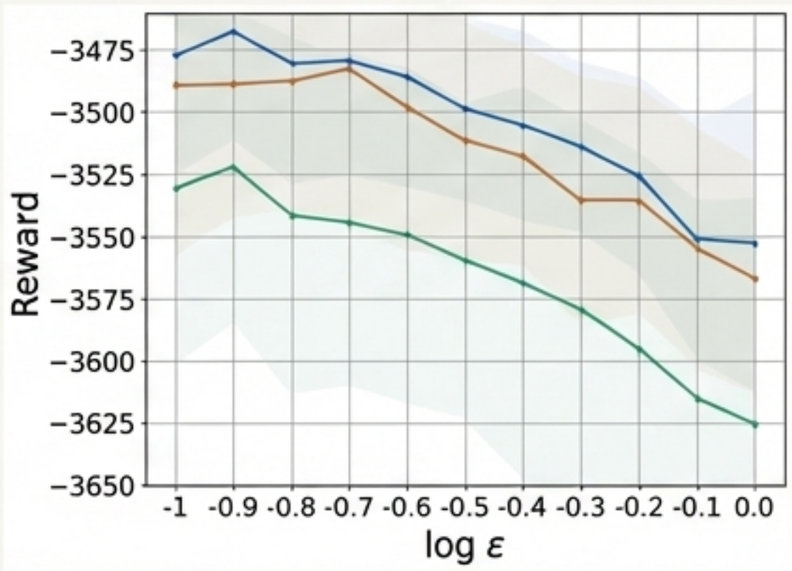
- **Finding:** ERNIE-MF shows a clear performance advantage over the Mean Field baseline and M3DDPG.

N=6 Agents



- **Finding:** The trend continues. ERNIE-MF maintains its superior robustness.

N=15 Agents



- **Finding:** Even with a large number of agents, ERNIE-MF demonstrates a higher reward and slower performance degradation as noise increases.

Conclusion: ERNIE provides a viable path for building robust policies for large-scale multi-agent systems.

A Practical Guideline for Implementation: Wider Networks Learn More Robust Policies

The Question

Our theory advocates for learning a smooth policy. Can standard neural networks effectively approximate such functions?

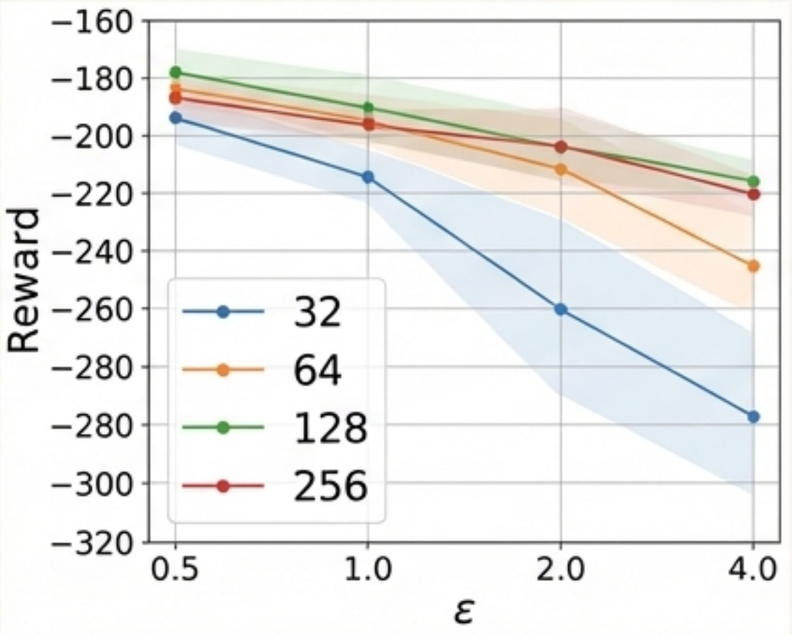
The Answer (Theorem 4.1)

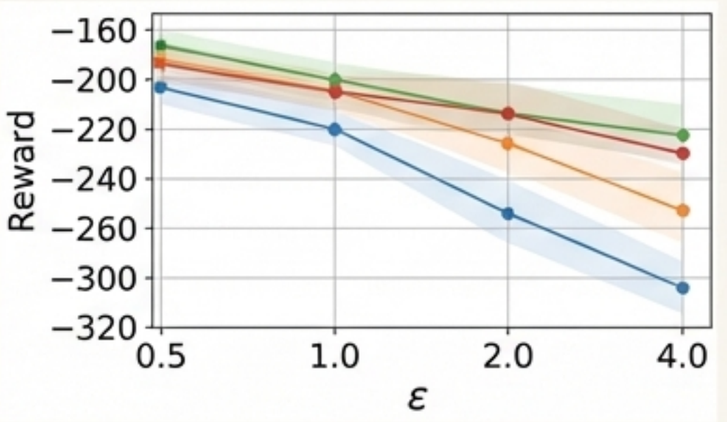
Yes. For a smooth target function, there exists a sufficiently wide ReLU network that not only approximates it well but also has a bounded Lipschitz constant.

Key takeaway:
A wider network has more capacity to represent a smooth function without being forced into a non-smooth, 'jagged' solution space.

Empirical Verification

Setup
We train ERNIE with policy networks of varying widths (32, 64, 128, 256 hidden units) and evaluate their robustness.





Analysis

As environment perturbation ε increases, the performance of narrower networks (32, 64 units) drops significantly. Wider networks (128, 256 units) are far more stable and maintain higher performance.

This empirically validates the theory: sufficient network width is crucial for learning robust policies with ERNIE.

ERNIE: A Principled Framework for Robust Multi-Agent Learning

The Narrative Recapped

- **The Problem**: Standard MARL policies are too brittle for real-world deployment.
- **The Insight**: Policy smoothness (Lipschitz continuity) is theoretically and empirically linked to robustness.
- **The Solution**: ERNIE operationalizes this insight using a stable Stackelberg formulation of adversarial regularization.
- **The Proof**: Extensive experiments demonstrate superior robustness against observation noise, dynamic environments, malicious agents, and at scale.

Summary of Contributions

1. **Advances in Theoretical Understanding**: Established the fundamental connection between Lipschitz continuity and MARL robustness.
2. **Novel Regularizers for MARL**: Developed new adversarial regularizers for observations, actions, and mean-field distributions.
3. **New Stable Algorithms**: Introduced the Stackelberg game formulation to overcome the instability of adversarial training in MARL.
4. **Comprehensive Empirical Validation**: Demonstrated state-of-the-art robustness across multiple complex environments, including traffic control and particle games.

By enforcing smoothness, ERNIE provides a practical and theoretically-grounded path towards deploying reliable and robust multi-agent systems.

Code available at: https://github.com/abukharin3/ERNIE