Safe Multi-Agent Reinforcement Learning Through Neural Graph Control Barrier Functions

Extended Abstract

Ziye Deng
Southwest University
Chongqing, China
dzy2479133071@email.swu.edu.cn

Qiang Li
Southwest University
Chongqing, China
xiaokunjigiegie@email.swu.edu.cn

Mingyue Zhang
Southwest University
Chongqing, China
myzhangswu@swu.edu.cn

Wu Chen
Southwest University
Chongqing, China
chenwu@swu.edu.cn

ABSTRACT

Multi-Agent Reinforcement Learning (MARL) has shown great potential in a wide range of domains, but its trial-and-error exploration paradigm can lead to severe risks in safety-critical tasks. Existing approaches to safe MARL, including reward shaping, constrained optimization, and shielding, provide only soft guarantees and are often insufficient to ensure strict safety. In contrast, Control Barrier Functions (CBFs) from control theory offer a principled way to enforce safety by keeping system states within a safe set at all times, providing new opportunities to address safety in reinforcement learning. In this work, we propose a novel framework that integrates CBFs with MARL to guarantee safety while preserving learning performance. Comprehensive evaluations in standard benchmark environments demonstrate that our method consistently achieves zero-violation safety constraints and matches or even surpasses existing methods in terms of performance.

KEYWORDS

multi-agent reinforcement learning; control barrier functions; safe reinforcement learning; benchmark evaluation

ACM Reference Format:
Ziye Deng, Qiang Li, Mingyue Zhang, and Wu Chen. 2026. Safe Multi-Agent Reinforcement Learning Through Neural Graph Control Barrier Functions: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/KRTJ4225

1 INTRODUCTION

While MARL empowers decentralized systems, its inherent trial-and-error exploration poses severe risks [2]. Balancing safety and efficiency is critical, as failures in high-stakes domains can be catastrophic [5].

Mingyue Zhang (myzhangswu@swu.edu.cn) is the corresponding author.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/KRTJ4225

Existing methods, such as reward shaping and constrained optimization, typically function as "soft constraints." While they reduce the likelihood of unsafe behavior, they fail to provide the strict, system-wide guarantees required for safety-critical tasks. In contrast, **Control Barrier Functions (CBF)** offer a rigorous mathematical framework to enforce "zero-violation" safety by restricting system states via real-time correction [3, 4]. In this paper, we propose a novel framework integrating CBF with MARL. By employing CBF for real-time action correction, our method ensures policies strictly satisfy safety constraints without sacrificing optimality. Experiments on the Multi-Agent Particle Environment (MPE) demonstrate that our approach consistently maintains safety while matching or surpassing baselines in performance. The main contributions are: (1)**Rigorous Safety Framework:** We integrate CBFs into MARL to provide theoretical "zero-violation" safety guarantees, preventing irreversible risks during exploration; (2)**Efficient Real-Time Correction:** Our method unifies safety constraints with policy optimization through immediate action correction, preserving learning efficiency. (3) **Empirical Validation:** Systematic experiments on MPE benchmarks confirm our approach outperforms baselines in safety while maintaining competitive task performance.

2 METHOD

2.1 Problem Formulation and Architecture

We formulate the safety-critical multi-agent task as a Decentralized Partially Observable Markov Decision Process (Dec-POMDP). To balance high-level task planning with strict low-level safety guarantees, we propose a decoupled control architecture consisting of a frozen reference policy and a learnable safety correction module.

Reference Policy via MAPPO.. We first employ **Multi-Agent Proximal Policy Optimization (MAPPO)** to learn a task-oriented policy π_{ref}. Operating under the Centralized Training Decentralized Execution (CTDE) paradigm, this policy maximizes task rewards (e.g., target coverage) but is agnostic to explicit safety constraints. Once converged, π_{ref} is frozen. This ensures a stationary behavior distribution, preventing the "moving target" problem during the subsequent training of the safety module. For agent i, the nominal action is given by $u_{\text{ref},i}(t) = \pi_{\text{ref}}(o_i(t))$.

GCBF Safety Correction. We introduce a **Graph Control Barrier Function (GCBF)** module, parameterized by a Graph Neural Network (GNN), to handle variable agent counts and capture local interactions. This module predicts a safety correction $u_{\text{gcbf},i}(t)$ and explicitly learns the barrier value $h(s_i)$. The final control output is an affine combination:

$$u_i(t) = u_{\text{ref},i}(t) + \lambda\, u_{\text{gcbf},i}(t). \tag{1}$$

Here, λ scales the intervention. In safe regions, u_{gcbf} remains near zero, allowing the agent to pursue the task; near safety boundaries, it actively steers the agent back to the safe set.

2.2 Training and Optimization

The GCBF module is trained via supervised learning to distill the capabilities of a rigorous optimization-based controller into a neural network, enabling real-time execution without online solving.

QP Supervision and Sampling. We collect trajectory data using the current policy and generate "expert" safety actions u_{qp} by solving a standard CBF Quadratic Program (QP) point-wise for each state. To address the scarcity of safety-critical events, we employ a **dual-buffer strategy**, maintaining separate replay buffers for safe and unsafe transitions. Batches are sampled with a fixed ratio to ensure the model learns the safety boundary effectively.

Loss Function. The network parameters are optimized to minimize a composite loss $\mathcal{L}$ that enforces the theoretical conditions of Control Barrier Functions:

$$\mathcal{L} = \lambda_{\text{a}} L_{\text{a}} + \lambda_{\text{s}} L_{\text{s}} + \lambda_{\text{us}} L_{\text{us}} + \lambda_{\dot{h}} L_{\dot{h}}. \tag{2}$$

Action Matching (L_{a}): $\mathbb{E}[\|(u_{\text{ref}} + \lambda u_{\text{gcbf}}) - u_{qp}\|^2]$. This term forces the policy to mimic the optimal safety intervention computed by the QP solver. **State Validity ($L_{\text{s}}, L_{\text{us}}$):** These terms enforce the barrier property $h(s) > 0$ for safe states and $h(s) < 0$ for unsafe states, establishing a valid separation between safe and unsafe sets. **Forward Invariance ($L_{\dot{h}}$):** The core safety guarantee comes from enforcing the Lie derivative condition $\dot{h}(s) + \alpha h(s) \geq 0$. We approximate $\dot{h}(s)$ via discrete time-steps and penalize violations:

$$L_{\dot{h}} = \mathbb{E}\left[\max\left(0, -\left(h(s_{t+1}) - h(s_t) + \alpha h(s_t)\right)\right)\right]. \tag{3}$$

Minimizing this term ensures that if the system starts in a safe state, it remains safe indefinitely.

3 EXPERIMENTS

3.1 Setup and Baselines

We evaluate our method on a modified **MPE Simple Spread** task where N agents coordinate to cover N landmarks. We introduce strict safety constraints: (1) **Boundary:** Agents incur penalties when approaching the world edge (radius > 0.92); (2) **Collision:** A violation occurs if inter-agent distance $d_{ij} < 0.18$. Performance is measured by: **Efficiency**, i.e., mean episode return (R) and *Arrival Rate* ($R_A = C/L$), representing the fraction of covered landmarks; **Safety**, i.e., *Violation Rate* ($R_V = 1 - R_A$), representing the fraction of uncovered goals or safety breaches.

We compare against three representative algorithms using identical network backbones and budgets: (1) **MAPPO** [1]: Unconstrained PPO baseline; (2) **MACPO**: Constrained Policy Optimization using trust regions; (3) **MAPPO-Lagrangian (MAPPO-L)**: Primal-dual method with adaptive cost penalties. All constrained methods use a binary violation indicator $c_t \in \{0, 1\}$ for boundary or collision breaches.

Training spans 20M steps with episodes of length $T = 200$ (early termination upon success). We use AdamW optimizers (lr=3e−5) for both Actor and CBF networks. The backbone is a GNN (1-layer, hidden 256) with attention aggregation. Key hyperparameters include: discount $\gamma = 0$ (for CBF), soft-update $\tau = 0.5$, margin $\varepsilon = 0.02$, and loss coefficients $\lambda_{\text{unsafe/safe}} = 1.0, \lambda_{\text{action}} = 1\text{e}{-4}$.

3.2 Results and Analysis

Table 1: Comparison of Methods

Method	Return	Violation Rate	Arrival Rate
MACPO	-1019.44 ± 45.76	0.0226 ± 0.0176	0.7268 ± 0.0190
MAPPO	-900.19 ± 44.03	0.3118 ± 0.0224	0.8483 ± 0.0187
MAPPO-L	-960.30 ± 55.12	0.0841 ± 0.0173	0.8094 ± 0.0174
OURS	**-938.22 ± 44.78**	**0.0329 ± 0.0197**	**0.8293 ± 0.0179**

As shown in Table 1 , unconstrained **MAPPO** achieves the highest return and arrival rate (0.848) but suffers from severe safety failures ($R_V \approx 31\%$). Conversely, **MACPO** is the safest ($R_V \approx 2.2\%$) but overly conservative, resulting in the lowest task performance.

OURS (MAPPO+GCBF) achieves the best trade-off. It maintains a high arrival rate (0.829), closely tracking MAPPO and outperforming MAPPO-L (0.809), while drastically reducing violations to near-zero levels (3.2%), comparable to MACPO. The results demonstrate that our method reduces safety violations by $\sim$ 90% compared to MAPPO while sacrificing less than 5% in task efficiency, successfully unifying high-performance control with strict safety guarantees.

4 CONCLUSION

In this paper, we propose a safe MARL framework fusing a frozen MAPPO policy and a learnable GCBF correction module. Our decoupled architecture ensures theoretical zero-violation safety in decentralized systems without compromising near-optimal task efficiency. The lightweight, graph-based design allows the safety correction to operate via a simple forward pass, preserving real-time scalability for safety-critical domains such as robotic swarms and traffic coordination. Despite these strengths, the current framework relies on the quality of the pre-trained reference policy; a miscalibrated reference may trigger frequent safety interventions, leading to conservativeness or local deadlocks in complex topologies (e.g., narrow passages). Furthermore, the reliance on static safety hyperparameters can limit adaptability in highly dynamic scenarios.

Future work will address these limitations by exploring adaptive, state-dependent correction strengths to minimize unnecessary interventions. We also plan to investigate differentiable safety layers that incorporate constraint satisfaction directly into the policy update, as well as coordination priors to resolve multi-agent conflicts more effectively.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (No. 62402400).

REFERENCES

[1] Shangding Gu, Jakub Grudzien Kuba, Munning Wen, Ruiqing Chen, Ziyan Wang, Zheng Tian, Jun Wang, Alois Knoll, and Yaodong Yang. 2021. Multi-agent constrained policy optimisation. *arXiv preprint arXiv:2110.02793* (2021).

[2] Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, and Alois Knoll. 2024. A review of safe reinforcement learning: Methods, theories and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[3] Li Wang, Aaron D Ames, and Magnus Egerstedt. 2017. Safety barrier certificates for collisions-free multirobot systems. *IEEE Transactions on Robotics* 33, 3 (2017), 661–674.

[4] Weishu Zhan and Peter Chin. 2024. Safe multi-agent reinforcement learning for bimanual dexterous manipulation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 12420–12427.

[5] Hengjun Zhao, Quanzhong Li, Xia Zeng, and Zhiming Liu. 2022. Safe Reinforcement Learning Algorithm and Its Application in Intelligent Control for CPS. *International Journal on Food System Dynamics* 13, 4 (2022).