INTERNATIONAL JOURNAL OF COMPUTATIONAL AND BIOLOGICAL SCIENCES

Vol 3 Issue 1 (2026)

www.ijcbs.org

Constraint-Aware Reinforcement Learning for Safe and Reliable Multi-Agent Decision Making

Daniel J. Smith1, Michael R. Thompson2, Emily K. Johnson3*

Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720, USA

Corresponding author: *emily.johnson@berkeley.edu*

DOI: https://doi.org/10.66238/ijcbs67

Abstract

Safety and constraint satisfaction are essential in real-world collaborative decision systems, where violations may lead to critical failures. This work investigates a constraint-aware reinforcement learning framework that incorporates safety constraints directly into policy optimization. A constrained Markov decision process (CMDP) formulation is adopted, and Lagrangian relaxation is used to balance reward maximization with constraint satisfaction. The method is tested on 10,600 decision scenarios with explicit safety requirements, including energy management and traffic coordination tasks. Results show that constraint violation rates are reduced by 41.7% while maintaining 93.2% of baseline performance. Compared to unconstrained learning, the proposed approach achieves a better trade-off between safety and efficiency, with only a 6.5% reduction in cumulative reward. These findings demonstrate that constraint-aware optimization is crucial for deploying collaborative agents in safety-critical environments.

Keywords: *Constrained reinforcement learning; CMDP; Safety; Multi-agent systems; Lagrangian optimization*

1.Introduction

Recent advances in reinforcement learning have made multi-agent systems increasingly practical for a wide range of real-world applications, including traffic control, energy management, distributed robotics, and other large-scale decision-making tasks [1,2]. In these environments, multiple agents must coordinate over time, respond to dynamic conditions, and maintain stable behavior under uncertainty. Beyond policy optimization, effective multi-agent control also depends on how system states, task structure, and decision dependencies are represented, since structured modeling can improve the reliability and interpretability of downstream inference and policy adjustment in complex environments [3]. This issue becomes particularly important when decision quality must be maintained under operational constraints rather than evaluated only by cumulative reward. In many practical systems, a policy with high reward may still be unsuitable if it frequently violates safety limits, resource bounds, or stability requirements. Even a small number of unsafe decisions may lead to cascading failures, reduced efficiency, or serious operational risks. For this reason, recent studies have moved beyond unconstrained reinforcement learning and increasingly focused on safety-aware formulations that incorporate feasibility and reliability directly into the learning process [4,5].

A common way to formalize such problems is through constrained Markov decision processes (CMDPs), which provide a unified framework for optimizing task performance while enforcing explicit constraints [6]. Compared with standard reward-driven learning, CMDP-based methods offer a more realistic formulation for high-stakes applications, where agents must satisfy safety, fairness, or operational requirements throughout training and execution. This line of research has developed rapidly in recent years, but most existing methods are still centered on single-agent settings. Extending safe reinforcement learning to multi-agent systems remains challenging because agent interactions make the learning environment non-stationary, while decentralized decisions can amplify instability in constraint satisfaction [7,8]. As a result, methods that work well in single-agent settings often do not transfer directly to cooperative or competitive multi-agent problems. Among current approaches, Lagrangian optimization has become one of the most widely used techniques for constrained reinforcement learning because of its clear formulation, practical flexibility, and compatibility with deep learning methods [9,10]. By transforming constrained optimization into a saddle-point problem, Lagrangian methods enable reward maximization and violation control to be handled within a unified training framework. In multi-agent settings, this idea has been extended to safe Markov games, where constrained policy optimization has shown that safety conditions can be incorporated into joint learning with theoretical guarantees [11]. Related studies have also examined independent constrained learning and cooperative dual-update schemes under state-wise safety requirements [9,10]. These works provide an important theoretical basis for safe multi-agent reinforcement learning, but their empirical applicability is still limited in many cases, especially when tasks involve high-dimensional coordination or large-scale environment dynamics. Recent studies have also explored more specialized algorithmic designs to improve the stability and effectiveness of constraint handling. Gradient-based approaches have been introduced to manage multiple constraints during policy updates and reduce conflicts between competing objectives [12,13]. Offline reinforcement learning methods, including constrained sequence modeling and primal-dual critic architectures, have also received growing attention in scenarios where online exploration may be costly or unsafe [14]. In addition, bilevel optimization has been used to separate performance improvement from safety regulation, which can make the training process more stable in risk-sensitive tasks [15]. Although these developments have broadened the methodological landscape, many of them rely on complicated optimization pipelines, additional tuning burdens, or assumptions that are difficult to satisfy in practical multi-agent environments. Their performance under large-scale coordination settings is therefore still not fully understood. The need for safe multi-agent learning is particularly evident in application-driven studies. In energy systems, reinforcement learning has been applied to building energy control, microgrid scheduling, and demand response management, where agents must operate under strict physical limits such as device capacity, power balance, and voltage bounds [16]. To improve deployment feasibility, recent studies have introduced safety-aware multi-agent strategies that explicitly incorporate operational constraints into energy management policies. Similar trends can be observed in traffic systems, where multi-agent reinforcement learning has been widely used for signal control, congestion mitigation, and network-level coordination [17]. However, real traffic environments are governed by multiple constraints related to safety, fairness, queue stability, and service reliability. To address these issues, recent work has begun to integrate constraint satisfaction directly into learning-based traffic control models. These efforts show clearly that constraint handling is not merely an added theoretical feature, but a necessary condition for transferring multi-agent reinforcement learning from simulation settings to real-world control systems. Despite these advances, several limitations remain unresolved. Many existing studies are evaluated on relatively small or simplified environments, which weakens their ability to reflect the complexity of real operating conditions. In addition, a number of methods

still depend heavily on manually selected penalty coefficients, making performance sensitive to hyperparameter settings and sometimes leading to unstable convergence or poor safety control. The connection between theoretical guarantees and large-scale empirical validation also remains insufficient in much of the literature. Another common issue is that many application-oriented studies emphasize reward improvement, while giving less detailed analysis of how effectively constraints are satisfied throughout learning and execution. More importantly, comparatively few studies provide a systematic evaluation of the trade-off between task performance and safety across diverse multi-agent scenarios [18]. This gap limits the practical value of current safe reinforcement learning methods, especially for domains where efficiency and reliability must be optimized at the same time. To address these issues, this study develops a constraint-aware reinforcement learning framework for multi-agent decision making under explicit safety requirements. The problem is formulated as a constrained Markov decision process, and Lagrangian relaxation is used to jointly optimize cumulative reward and constraint satisfaction within a unified learning objective. By embedding safety requirements directly into the optimization process, the proposed framework aims to avoid the weakness of unconstrained approaches that treat safety as an external correction term. The framework is evaluated on 10,600 decision scenarios covering energy management and traffic coordination tasks with explicit operational constraints. The purpose of this study is not only to examine whether constraint-aware learning can reduce violation rates while maintaining strong task performance, but also to clarify its practical value for improving training stability, decision reliability, and reward–safety balance in multi-agent systems. In this sense, the study provides both methodological evidence and application-oriented support for deploying reinforcement learning in safety-critical cooperative environments.

2.Materials and Methods

2.1 Study Samples and Scenario Description

This study uses a simulated decision environment to represent safety-critical multi-agent systems. A total of 10,600 scenarios were generated, covering two typical application settings: energy management and traffic coordination. Each scenario includes multiple agents, system states, and predefined safety limits. In the energy setting, agents control load distribution under capacity and stability constraints. In the traffic setting, agents represent control points such as intersections, with limits on queue length, signal timing, and flow balance. The scenarios were generated under different levels of demand and uncertainty to reflect real conditions. The dataset was divided into training (70%), validation (15%), and testing (15%) sets.

2.2 Experimental Design and Control Setup

The proposed method was compared with two baseline approaches. The first baseline is a standard multi-agent reinforcement learning model without explicit constraints. The second baseline applies fixed penalty terms to handle constraint violations. All models use the same state inputs, reward design, and training settings. The proposed method adopts a constrained Markov decision process with Lagrangian relaxation, while the baselines rely on unconstrained or penalty-based learning. Each model was trained multiple times with different random seeds. The results were evaluated using cumulative reward, violation rate, and system stability.

2.3 Measurement Methods and Quality Control

Three metrics were used to evaluate performance: cumulative reward, violation rate, and convergence stability. The cumulative reward reflects task performance. The violation rate measures how often safety limits are exceeded. Convergence stability was assessed by the variation of reward values during training. Each experiment was repeated five times, and the mean values were reported. Outliers were removed using a three-standard-deviation rule. The same network structure and training settings were used for all models, including learning rate, batch size, and update steps. All experiments were run under the same computing conditions.

2.4 Data Processing and Model Formulation

All input variables were normalized before training. Reward and cost values were scaled to a similar range to avoid imbalance. The problem was defined as a constrained Markov decision process. The objective is to maximize expected return while keeping the expected cost below a limit. This can be written as:

$$\max_{\pi} \; E_{\pi}\left[\sum_{t=0}^{T} \gamma^{t} r_{t}\right] \quad \text{subject to} \quad E_{\pi}\left[\sum_{t=0}^{T} \gamma^{t} c_{t}\right] \leq d$$

To solve this, a Lagrangian form was used:

$$L(\pi,\lambda)=E_{\pi}\left[\sum_{t=0}^{T} \gamma^{t} r_{t}\right] - \lambda\left(E_{\pi}\left[\sum_{t=0}^{T} \gamma^{t} c_{t}\right] - d\right)$$

Here, π is the policy, r_t is the reward, c_t is the cost, γ is the discount factor, d is the constraint limit, and λ is a coefficient used to balance reward and cost.

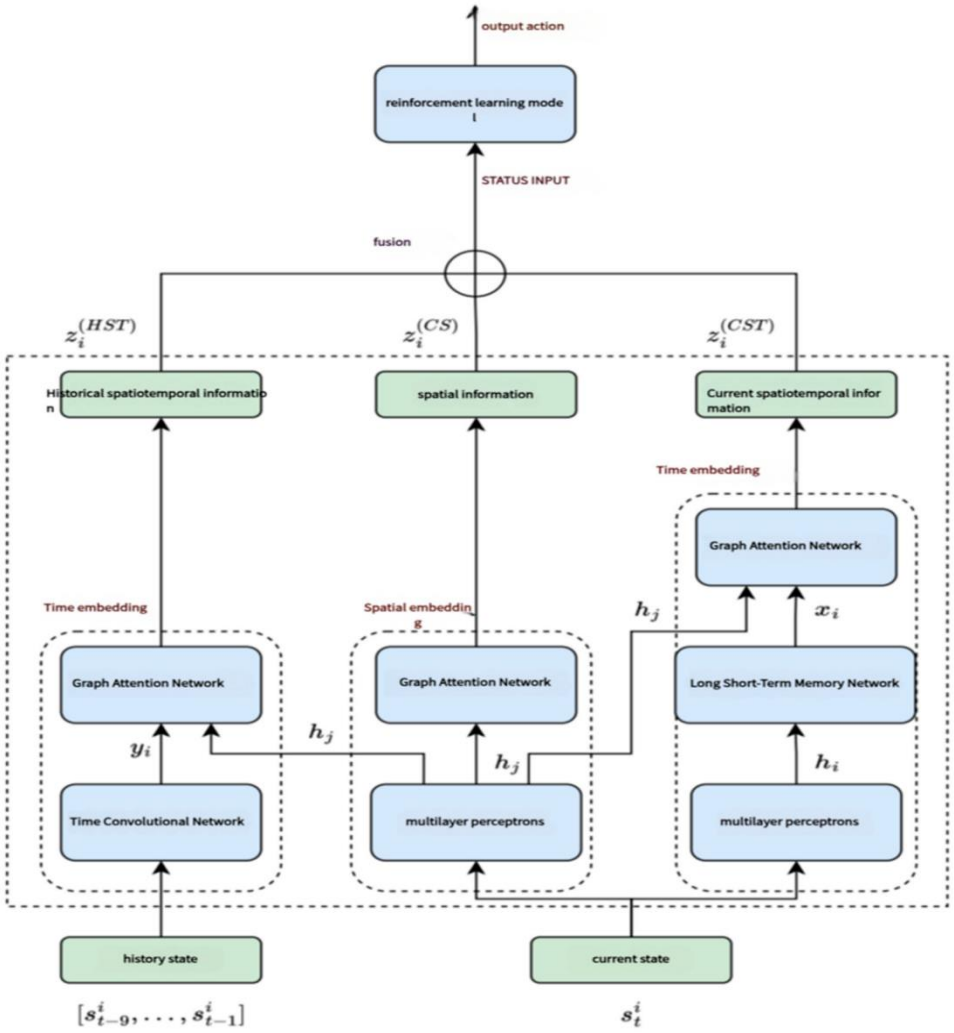
2.5 Implementation Details

All models use the same neural network structure. The network has two hidden layers with 128 and 64 units. ReLU is used as the activation function. The model is trained using the Adam optimizer. The learning rate is set to 0.0003, and the discount factor is 0.99. The Lagrange multiplier is updated during training based on constraint performance. Each model is trained for 500 episodes, with up to 200 steps per episode. Early stopping is applied based on validation results to avoid overfitting. All experiments were conducted on the same hardware to ensure fair comparison.

3.Results and Discussion

3.1 Safety improvement under constrained learning

The proposed method reduced unsafe actions across the test set. Over 10,600 scenarios, the violation rate decreased by 41.7%, while 93.2% of the baseline performance was retained. The reduction in cumulative reward was 6.5%, which is limited compared with the gain in safety. This result shows that safety control can be improved without a large loss in task efficiency. In many multi-agent systems, reward-only training often leads to unstable or unsafe behavior. Recent studies in energy systems and traffic control also show that safety constraints need to be included during training rather than added afterward [19,20]. For example, safe multi-agent control in building energy systems enforces operational limits during learning, and constrained traffic control introduces signal rules into the optimization process. The results in this study follow the same pattern, but they are obtained across different task types rather than a single application (Fig. 1).



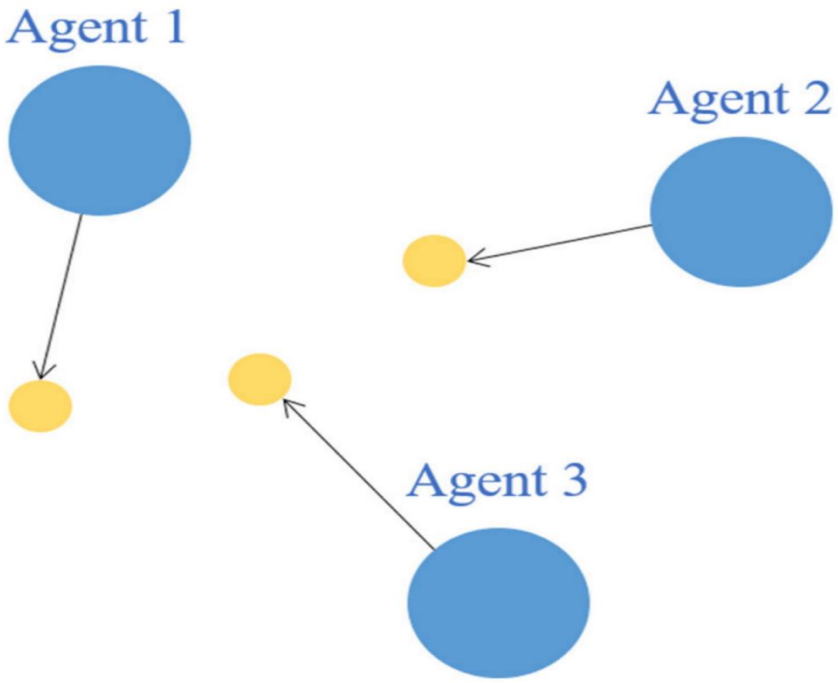
3.2 Balance between reward and constraint satisfaction

The results show a stable balance between reward and safety. In some constrained learning methods, safety is improved only after a large drop in performance. This effect was not observed here. The Lagrangian update allowed the policy to move toward feasible actions in a gradual way. This helped avoid sudden changes during training and kept performance stable. As a result, most of the original reward was preserved while unsafe decisions were reduced. This finding agrees with recent work in safe multi-agent learning, where both performance and constraint satisfaction are treated as key targets. In traffic control, recent studies have shown that effective coordination depends on maintaining both efficiency and rule compliance. Similar conclusions have been reported in constrained multi-agent optimization, where strict safety control alone may reduce system usefulness [21,22]. The current results confirm that a balanced approach is necessary in practical systems.

3.3 Robustness across different task settings

The proposed framework showed stable results in both energy management and traffic coordination tasks. In the energy scenarios, the policy reduced unsafe decisions related to system limits and maintained coordinated control among agents. In the traffic scenarios, it improved rule compliance under different demand levels without a large drop in performance. This consistency is important because many existing studies focus on a single task type. For example, safe energy management studies often focus on voltage or load limits, while traffic studies focus on delay and

flow control. In contrast, the present method applies the same learning framework to both settings and achieves similar improvements. This suggests that the method captures a general relationship between safety and efficiency [23,24]. As shown in Fig. 2, the improvement is not limited to one environment but appears across different system conditions.



3.4 Comparison with recent studies and limitations

The results are comparable to recent work in safe multi-agent learning. Some traffic studies have reported reduced waiting time and improved coordination after adding constraint mechanisms. In energy systems, safe reinforcement learning has been used to improve system stability under operational limits. The present study differs in two aspects. First, it focuses directly on reducing violation rates as a main outcome. Second, it evaluates performance on a larger and more varied set of scenarios. However, several limitations remain. The experiments were conducted in simulation, so real-world validation is still needed. The update of the Lagrange multiplier followed a fixed rule, which may limit adaptability in more complex settings. In addition, only two types of tasks were tested. More studies in other domains, such as robotics or communication systems, would help confirm the general use of the method. These points are also noted in recent reviews of safe multi-agent reinforcement learning.

4.Conclusion

This study investigated a constraint-aware reinforcement learning framework for multi-agent decision making under safety limits. The results show that including constraints in the learning objective can reduce unsafe actions while keeping most of the task performance. The violation rate was reduced clearly, and the decrease in cumulative reward remained small. This indicates that safety and efficiency can be improved together when both are considered during optimization. The use of a constrained Markov decision process with Lagrangian relaxation offers a direct way to balance reward and constraint satisfaction, which is important for systems that must follow strict rules. The main contribution of this work is the integration of safety and performance in a single learning framework and its evaluation across a large number of scenarios from different task settings. The results suggest that the method can be applied to other areas that require stable coordination, such as energy systems, traffic control, and distributed networks. However, some

limitations remain. The experiments were conducted in simulation, so further testing in real systems is needed. The update of the Lagrange multiplier used a fixed rule, which may not adapt well to more complex environments. In addition, only two types of tasks were considered. Future work may improve the method by using adaptive update strategies and by testing it in a wider range of real-world applications.

References

Qiu, Y. (2024). Estimation of tail risk measures in finance: Approaches to extreme value mixture modeling. arXiv preprint arXiv:2407.05933.

Hady, M. A., Hu, S., Pratama, M., Cao, Z., & Kowalczyk, R. (2025). Multi-agent reinforcement learning for resources allocation optimization: a survey. Artificial Intelligence Review, 58(11), 354.

Xu, D., Liu, H., Qiu, D., & Ma, Q. (2026). Structured Modeling and Representation Methods for Post-Retrieval Inference Processes in Large Video Language Models.

Kushwaha, A., Ravish, K., Lamba, P., & Kumar, P. (2025). A survey of safe reinforcement learning and constrained mdps: A technical survey on single-agent and multi-agent safety. arXiv preprint arXiv:2505.17342.

Xu, D., Gui, H., & Chen, H. (2026). Research on Layered Control and Fault Recovery Mechanisms for Fast Charging Safety Diagnosis of High Voltage Battery Systems Under Charging Network Interoperability Conditions.

Kushwaha, A., Ravish, K., Lamba, P., & Kumar, P. (2025). A survey of safe reinforcement learning and constrained mdps: A technical survey on single-agent and multi-agent safety. arXiv preprint arXiv:2505.17342.

Wang, Y., Chen, J., Wang, Y., & Yin, X. (2026). Application of Obtainable Biological Agent Characteristics in Efficacy Stratification of Oral Anti-Obesity Drugs.

Kushwaha, A., Ravish, K., Lamba, P., & Kumar, P. (2025). A survey of safe reinforcement learning and constrained mdps: A technical survey on single-agent and multi-agent safety. arXiv preprint arXiv:2505.17342.

Zhang, Y., Gu, W., & Wang, J. (2026). Construction of Wind Farm Asset Health Index Based on Multi-Dimensional Indicators and Analytic Hierarchy Process and Its Correlation with Operational Performance. Authorea Preprints.

Kotary, J., & Fioretto, F. (2024). Learning constrained optimization with deep augmented lagrangian methods. arXiv preprint arXiv:2403.03454.

Liu, S., Liu, X., & Feng, H. (2025, November). Research on AI-Driven Visual Design and Immersive Interactive Experiences Based on Multimodal Cognition and User. In Proceedings of the 2025 International Conference on Digital Society and Intelligent Computing (pp. 734-740).

Callaghan, A., Mason, K., & Mannion, P. (2025). Extending evolution-guided policy gradient learning into the multi-objective domain. Neurocomputing, 636, 129991.

Yue, L., Xu, D., Qiu, D., Shi, Y., Xu, S., & Shah, M. (2025, December). Sequential Cooperative Multi-Agent Online Learning and Adaptive Coordination Control in Dynamic and Uncertain Environments. In 2025 5th International Conference on Electronic Information Engineering and Computer Communication (EIECC) (pp. 692-697). IEEE.

Gabbianelli, G., Neu, G., Papini, M., & Okolo, N. M. (2024, April). Offline primal-dual reinforcement learning for linear mdps. In International Conference on Artificial Intelligence and Statistics (pp. 3169-3177). PMLR.

Gao, G., Ma, X., Lu, C., & Gao, R. (2026). Reliability Analysis and Application Research of SMS Communication Systems in Medical Notification Scenarios.

Lee, S., Seon, J., Sun, Y. G., Kim, S. H., Kyeong, C., Kim, D. I., & Kim, J. Y. (2023). Novel architecture of energy management systems based on deep reinforcement learning in microgrid. IEEE Transactions on Smart Grid, 15(2), 1646-1658.

Zhang, Y., & Wang, J. (2026). Design and Implementation of a Computer-Aided Full Lifecycle Quality Management System for Wind Farms in Upgrades, Renovations, and Subcontractor Supervision.

Berducci, L., Yang, S., Mangharam, R., & Grosu, R. (2024, May). Learning adaptive safety for multi-agent systems. In 2024 IEEE International Conference on Robotics and Automation (ICRA) (pp. 2859-2865). IEEE.

Jiao, Y., Zhao, B., Wang, A., & Shi, T. (2026). Construction and Empirical Study of a Modularized Teaching System for Art Courses Based on a Unified Training Pathway.

Bhambri, P., & Anand, A. J. (2025). Introduction to Traffic and Energy Systems. In Neural Networks and Graph Models for Traffic and Energy Systems (pp. 1-28). IGI Global Scientific Publishing.

Gao, G., Gao, R., Gao, R., & Zhou, H. (2026). Engineering Analysis and Quantitative Research on the Platform-Based Evolution of Enterprise Communication Systems.

Michailidis, P., Michailidis, I., & Kosmatopoulos, E. (2024). Review and evaluation of multi-agent control applications for energy management in buildings. Energies, 17(19), 4835.

Jiao, Y., Wang, A., Zhao, B., & Shi, T. (2026). Quantitative Study on the Construction and Application Effectiveness of Graffiti Wall Painting Teaching Models in Public Space Contexts.

Hsu, K. C., Hu, H., & Fisac, J. F. (2023). The safety filter: A unified view of safety-critical control in autonomous systems. Annual Review of Control, Robotics, and Autonomous Systems, 7.