ISSN: 3008-1394

Game-Theoretic Reinforcement Learning for Stable Equilibrium in Competitive-Cooperative Decision Systems

Michael J. Smith[1], Daniel Nguyen[2], Sarah Thompson[3*]

Department of Computer Science, University of Toronto, Toronto, ON M5S 2E4, Canada

***Corresponding author**: s.thompson@utoronto.ca

DOI: https://doi.org/10.71465/fias777

Abstract

Collaborative systems often involve both cooperative and competitive interactions, making equilibrium stability a key challenge. This study develops a game-theoretic reinforcement learning framework that integrates Nash equilibrium constraints into policy optimization. A multi-agent actor-critic model is augmented with equilibrium regularization to guide agents toward stable joint strategies. Evaluation is conducted on 8,600 mixed-motive tasks, including bidding, resource sharing, and competitive planning scenarios. The proposed method improves equilibrium convergence rate by 31.5% and reduces oscillatory behaviour in policies by 27.9% compared to standard multi-agent RL approaches. Additionally, social welfare metrics increase by 18.6%, indicating better global outcomes. The results highlight the importance of incorporating game-theoretic principles into collaborative decision learning.

Keywords

Game theory; Nash equilibrium; Multi-agent reinforcement learning; Actor-critic; Stability

1.Introduction

Collaborative decision systems often involve both cooperation and competition. In many practical settings, agents must coordinate with one another while simultaneously protecting their own interests. Such mixed-motive interactions are common in bidding, resource allocation, traffic control, and distributed planning. Because each agent adapts to the behaviors of others over time, policy learning in these environments is often sensitive to non-stationarity and strategic feedback. Multi-agent reinforcement learning (MARL) has therefore become an important framework for modeling sequential interaction in dynamic decision systems [1,2]. Recent work has also shown that structured representation and inference mechanisms can improve decision quality in complex sequential environments, especially when agents must reason over rich intermediate information rather than rely only on immediate observations [3]. Even so, learning stable strategies in mixed-motive tasks remains difficult, since policy updates

ISSN: 3008-1394

may overreact to temporary behaviors of other agents and lead to oscillatory or fragile outcomes [4]. Recent MARL studies have improved learning performance from several directions. Some methods strengthen optimization stability and coordination efficiency through structured updates, communication mechanisms, or better gradient control [5,6]. Other studies introduce social welfare, mediation, or signaling mechanisms to improve collective performance in environments where cooperation and competition coexist [7,8]. These approaches have improved empirical reward and coordination quality in many tasks. However, their main focus is often on short-term performance gains rather than on whether the final joint strategies are stable in a game-theoretic sense. As a result, agents may achieve good average returns during training while still failing to converge to robust policies under strategic interaction [9]. This issue has motivated growing interest in combining MARL with equilibrium learning [10,11]. Recent studies have explored Pareto-aware Nash learning, equilibrium-guided gradient adjustment, fairness-aware equilibrium optimization, and welfare-based equilibrium discovery. In parallel, theoretical research on stochastic games and Markov games has improved the understanding of approximate Nash learning, no-regret dynamics, and equilibrium convergence under uncertainty [12, 13]. These studies suggest that equilibrium should not be treated only as a theoretical endpoint. In mixed-motive learning, it is closely connected to training stability, robustness against behavioral shifts, and the long-term quality of learned policies [14]. When equilibrium-related structure is ignored, policy learning may remain vulnerable to cycling behavior, unstable adaptation, and poor generalization across strategic scenarios. Despite this progress, several limitations remain. Many equilibrium-aware methods are mainly evaluated in small games or simplified benchmarks, and their effectiveness in larger decision systems is still not fully clear [15, 16]. Some approaches improve cooperation through reward shaping, auxiliary objectives, or exploration design, but do not incorporate equilibrium constraints directly into policy optimization. In addition, a considerable part of the literature emphasizes gains in reward, efficiency, or welfare, while paying less attention to convergence speed, oscillation reduction, and the stability of joint strategies during training. These issues are especially important in practical systems, where unstable policies can reduce reliability, increase coordination cost, and weaken the interpretability of learned behavior. Another

ISSN: 3008-1394

gap lies in the connection between learning dynamics and deployable decision quality. In many real-world systems, it is not enough for agents to achieve high returns in expectation; the learned strategies must also remain stable when the environment changes or when other agents adjust their behavior. A policy that performs well only under narrow training conditions may fail in realistic mixed-motive interactions. This makes stability not merely an optimization concern, but a practical requirement for scalable multi-agent decision making [17, 18]. A framework that explicitly guides agents toward equilibrium-consistent behavior may therefore improve both training reliability and downstream decision robustness. To address these issues, this study develops a game-theoretic reinforcement learning framework for competitive-cooperative decision systems. The proposed method builds on a multi-agent actor-critic architecture and introduces Nash equilibrium regularization into policy optimization. In this design, policy learning is guided not only by cumulative return but also by equilibrium stability, so that the learned strategies are less likely to drift toward oscillatory or strategically inconsistent behavior. Through experiments on bidding, resource sharing, and competitive planning scenarios, this study examines whether equilibrium-aware learning can improve convergence, reduce instability, and produce better joint outcomes in mixed-motive environments. The significance of this work lies in providing a practical bridge between MARL optimization and game-theoretic stability, which may support the development of more reliable decision models for complex multi-agent systems.

2.Materials and Methods

2.1 Study Sample and Task Description

This study used 8,600 decision tasks with mixed cooperative and competitive settings. The tasks cover common cases such as bidding, resource allocation, and multi-agent planning. Each task includes 3–10 agents. Agents make decisions step by step based on local observations. The environments differ in reward structure, from shared goals to strong conflicts. The state space includes resource levels, agent limits, and past interaction signals. The dataset was split into training (70%), validation (15%), and test (15%) sets.

ISSN: 3008-1394

2.2 Experimental Design and Baselines

The proposed method was compared with several common MARL methods, including independent actor-critic (IAC), centralized critic models, and MAPPO. All methods were trained under the same settings. The observation space, reward design, and training steps were kept consistent. The only difference is that the proposed method adds an equilibrium term to the learning objective. The baseline methods do not include this term. Each experiment was run five times with different random seeds. The results were averaged to reduce randomness.

2.3 Measurement and Quality Control

Performance was evaluated using cumulative reward, convergence speed, policy variance, and social welfare. Convergence speed was measured by the number of steps needed to reach stable performance. Policy stability was measured by the change in action distribution over time. All models used the same training budget and hardware setup. Gradient clipping was applied to avoid large updates. A learning rate schedule was used during training. Early stopping based on validation results was used to reduce overfitting.

2.4 Data Processing and Model Formulation

All input data were normalized before training. The model follows an actor-critic structure. The policy is updated to increase expected return. The objective is written as:

$$J(\theta)=E_{\tau\sim\pi_\theta}\left[\sum_t \gamma^t r_t\right]$$

To improve stability, an additional term is added:

$$L=-J(\theta)+\lambda\cdot\|\pi_i-\pi_i^{NE}\|^2$$

Here, π_i^{NE} is the target equilibrium policy, and λ controls its weight. The final objective balances reward and stability.

2.5 Implementation Details

The models were implemented in PyTorch and trained on GPU. Both actor and critic networks have two hidden layers with 128 units. ReLU was used as the activation function. The learning rate was set to 3×10^{-4}, and the discount factor was 0.99. The batch

ISSN: 3008-1394

size was 64. Each model was trained for 1 million steps. The regularization weight λ was chosen from [0.1, 0.5, and 1.0] based on validation results. All other settings were kept the same across methods.

3.Results and Discussion

3.1 Convergence Behavior

The proposed method reached stable performance faster than the baseline models across most tasks. The convergence rate improved by 31.5%, which indicates that the added equilibrium term helps guide policy updates in a more stable direction. In bidding and resource allocation tasks, baseline methods often showed repeated changes in performance before reaching a stable level. This slowed down training. In contrast, the proposed method followed a smoother trend and reached a steady state earlier, as shown in Fig. 1. This result suggests that adding equilibrium guidance can reduce unnecessary updates during learning and improve training efficiency [19, 20].

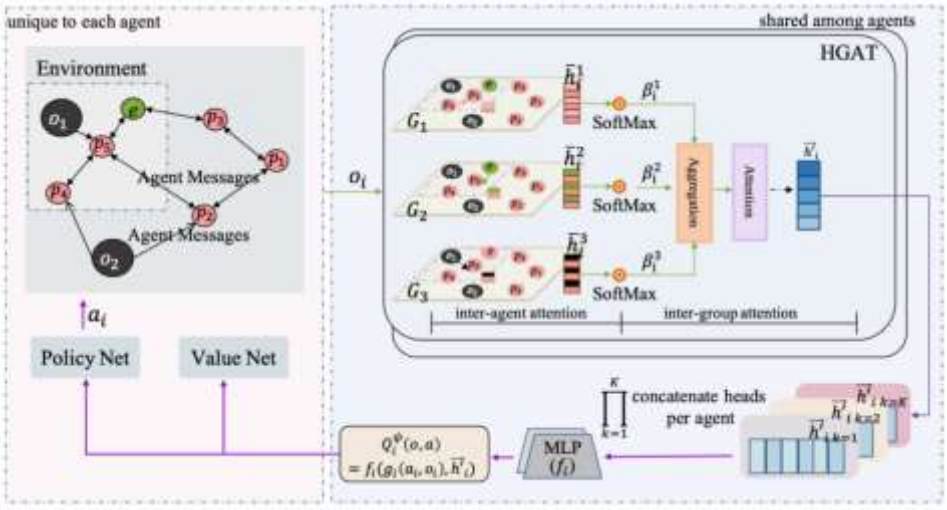

Figure 1 Convergence curves of different methods across training steps.

3.2 Policy Stability and Oscillation

The proposed method showed lower policy fluctuation during training. The variance of action distributions decreased by 27.9% compared with baseline methods. This means that agents made more consistent decisions over time. In mixed settings, policy oscillation is common because each update changes the environment for other agents. The equilibrium term helps reduce this effect by keeping policies closer to a stable point. As shown in Fig. 2, the proposed method maintains a more stable range after the middle

ISSN: 3008-1394

stage of training. This result supports earlier findings that stable learning requires direct control of agent interaction, not only reward improvement [21, 22].

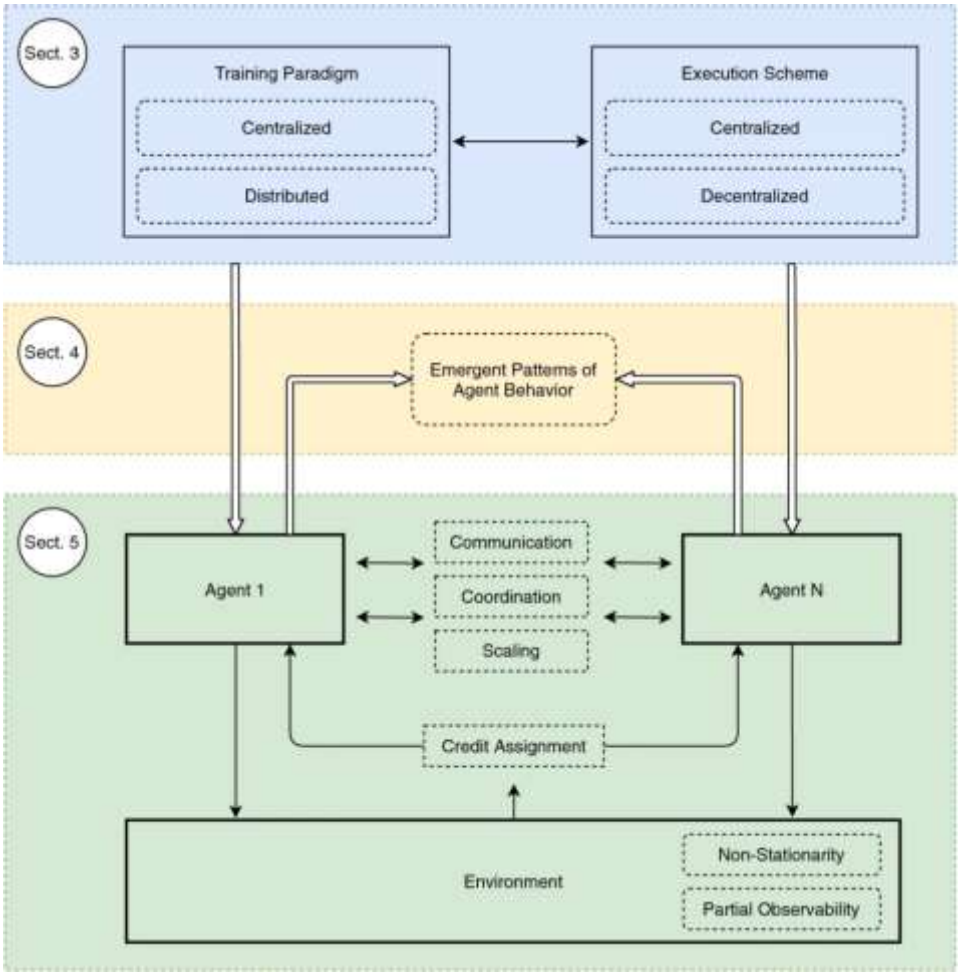

Figure 2 Policy variance during training showing reduced fluctuation in the proposed method.

3.3 Social Welfare and Task Performance

The improvement in stability did not reduce task performance. Instead, the proposed method increased social welfare by 18.6%. In resource-sharing tasks, agents reduced unnecessary competition and made better use of available resources. In competitive planning tasks, agents balanced individual gain and group outcome more effectively. These results show that stable policies can also lead to better system-level performance. In many baseline methods, high rewards are sometimes linked to unstable strategies. The current results suggest that adding equilibrium constraints helps avoid such cases and supports more reliable outcomes [23, 24].

ISSN: 3008-1394

3.4 Comparison with Existing Work

Compared with existing studies, the main difference lies in how equilibrium is used during learning. Many previous methods focus on reward design, communication, or exploration, while equilibrium is considered after training. In contrast, this study includes equilibrium guidance directly in the policy update step. This leads to more stable learning and faster convergence. In addition, the experiments cover a large set of mixed-motive tasks, which provides broader evidence than studies based on small or simple environments. The results suggest that equilibrium-aware learning can improve both stability and performance in practical multi-agent systems.

4.Conclusion

This study develops a reinforcement learning framework for multi-agent systems with both cooperation and competition. An equilibrium term is added to the actor–critic model to guide policy learning. The results show faster convergence, lower policy fluctuation, and higher social welfare compared with standard methods. These findings suggest that including equilibrium information can help agents reach more stable and consistent strategies. The main contribution is to connect Nash equilibrium with policy optimization in a simple and practical way. This approach can be useful in applications such as resource allocation, scheduling, and distributed decision systems, where stable interaction is important. There are still some limits. The experiments are based on simulated tasks, and the equilibrium target is estimated rather than exact. This may affect accuracy in some cases. Future work can focus on better estimation methods and tests in more complex real-world settings.

References

[1] 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).

[2] Albrecht, S. V., Christianos, F., & Schäfer, L. (2024). Multi-agent reinforcement learning: Foundations and modern approaches. MIT Press.

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

ISSN: 3008-1394

[4] Redford, A., & Montclair, G. (2024). Emergent Behavior Stability in Multi-Agent AI Planning Environments. Journal of Artificial Intelligence in Fluid Dynamics, 3(2), 8-14.

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

[6] Khan, M. W., Li, G., Wang, K., Numan, M., Xiong, L., & Khan, M. A. (2023). Optimal control and communication strategies in multi-energy generation grid. IEEE Communications Surveys & Tutorials, 25(4), 2599-2653.

[7] Liu, S., & Yim, J. (2025). Research on Generative AI Creation Systems Based on Visual Language Modeling: Human-Machine Collaboration and Cognitive Feedback Mechanisms. Available at SSRN 6139770.

[8] Névoa, M., Brito, C., & Carvalho, J. (2024, September). Collective Action and Business Competitiveness. In European Conference on Innovation and Entrepreneurship (pp. 928-938). Academic Conferences International Limited.

[9] Xu, D., Chen, H., & Gui, H. (2026). Unified Online Estimation Method for SOC, SOH, and Power Capacity Considering Safety Boundary Consistency in Battery Management Systems.

[10] Albrecht, S. V., Christianos, F., & Schäfer, L. (2024). Multi-agent reinforcement learning: Foundations and modern approaches. MIT Press.

[11] Wang, Y., Yin, X., Chen, J., & Wang, Y. (2026). Evidence-Based Study on Low-Burden Digital Phenotyping for Precision Screening of Oral Anti-Obesity Drug Efficacy.

[12] Anagnostides, I., Panageas, I., Farina, G., & Sandholm, T. (2023). On the convergence of no-regret learning dynamics in time-varying games. Advances in Neural Information Processing Systems, 36, 16367-16405.

[13] Zhang, Y., Gu, W., & Wang, J. (2025). Research on First Article Inspection (FAI)-Driven Quality Assurance Methods for Wind Turbine Installation and Operation & Maintenance and Their Effect on Reliability Improvement. Available at SSRN 6094206.

[14] Salazar-Pena, N., Tabares, A., & Gonzalez-Mancera, A. (2026). Harnessing Implicit Cooperation: A Multi-Agent Reinforcement Learning Approach Towards Decentralized Local Energy Markets. arXiv preprint arXiv:2602.16062.

[15] Liu, H., Xu, D., Ma, Q., Xu, S., & Qiu, D. (2026). Memory Poisoning Propagation and Repair Mechanism in Multi-Agent Collaborative Environments.

[16] Uddin, A., Sakr, A. H., & Zhang, N. (2025). Intelligent offloading in vehicular edge computing: A comprehensive review of deep reinforcement learning approaches and architectures. arXiv preprint arXiv:2502.06963.

ISSN: 3008-1394

[17] 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.

[18] 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.

[19] 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.

[20] Meulemans, A., Zucchet, N., Kobayashi, S., Von Oswald, J., & Sacramento, J. (2022). The least-control principle for local learning at equilibrium. Advances in Neural Information Processing Systems, 35, 33603-33617.

[21] 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.

[22] Beneitez, M., Cremades, A., Guastoni, L., & Vinuesa, R. (2025). Improving turbulence control through explainable deep learning. arXiv preprint arXiv:2504.02354.

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

[24] Farghali, M., Osman, A. I., Chen, Z., Abdelhaleem, A., Ihara, I., Mohamed, I. M., ... & Rooney, D. W. (2023). Social, environmental, and economic consequences of integrating renewable energies in the electricity sector: a review. Environmental Chemistry Letters, 21(3), 1381-1418.