摘要

强化学习是一种通过交互和过往经验学习的机器学习方法，已在围棋、游戏 AI 及利用人类反馈实现大模型价值对齐等领域取得突破性进展。随着强化学习研究的深入，多智能体强化学习逐渐成为研究热点，其中合作多智能体强化学习是该领域的重要方向。这一方向对于自动驾驶和智能交通等多智能体决策和控制任务有着独特的意义和价值。相较于单智能体强化学习，合作多智能体强化学习主要面对的挑战是环境不稳定性，也即多个智能体同时进行策略学习导致的相互干扰。这一挑战带来的直接影响就是智能体的策略学习过程难以收敛，策略的稳定性和性能都受到严重影响。

本文研究如何设计策略优化目标以解决环境不稳定性问题，确保策略学习的收敛性，提升智能体在多智能体合作问题中的性能。本文期望通过优化目标的设计，使得智能体的策略学习过程受到相同优化目标的引导，进而达成智能体之间的合作，确保最终策略的收敛。对合作多智能体强化学习算法的研究，根据策略执行和训练的模式，可以分成两个主要范式，其一是中心化训练去中心化执行，其二是完全去中心化。前者在训练的过程中允许中心化的信息交流，例如中心化的价值函数、通信、参数共享等，后者则在训练过程中也禁止一切智能体之间的信息交流。完全去中心化范式相较于中心化训练去中心化执行范式更加困难，但这种范式的应用场景也更加广泛。

本文针对两种范式都提出了具有收敛性保证的多智能体策略优化算法，以提升智能体在解决多智能体合作任务上的性能。在中心化训练去中心化执行的范式下，本文在策略优化目标中引入了正则项，在确保策略收敛性的同时，也对智能体的探索能力和采样效率有所提升。在完全去中心化范式下，本文从交替更新机制和代理优化目标两个角度提出了具有收敛性保证的算法，在交替更新和同步更新两种模式下都实现了策略学习的收敛和策略性能的提升。更进一步，本文还探索了完全去中心化算法在实际场景中的应用。本文的主要创新点包括以下四个方面：

首先，在中心化训练去中心化执行的场景中，本文参考单智能体强化学习中熵正则项的应用经验，在多智能体合作任务中引入了基于散度的正则项，从理论角度分析了加入这一正则项后的马尔科夫决策过程的特性，并基于这些特性提出了基于散度正则的多智能体策略优化算法，证明了该算法的单调收敛性。在中心化训练去中心化执行的范式下，既有基于策略的方法大多源于策略梯度定理，其理论依据导致其只能进行同策略更新，在多智能体环境中面临采样效率低下的挑战。即使采用异策略矫正技巧，在策略更新过程中也难以避免地引入了误差。针对这一挑战，本文从不同于策略梯度定理的角度提出了该算法的策略迭代，因而该算法能实现异策略更新。在实验中，

该算法与多种已有基线方法结合后都能提升基线方法的性能，证明了该算法的普适性，同时也验证了该算法的多种理论性质。

其次，在完全去中心化场景中，消除环境不稳定性的直观方法是采用交替更新机制，即在更新单一智能体策略时固定其他智能体策略。该机制下，智能体可在稳定环境中更新策略。基于这一机制，本文提出完全去中心化的交替策略优化算法，证明其在完全去中心化条件下具有策略收敛性，并且在实现层面只需要对价值函数进行迭代就可以完成策略提升。本文进一步分析了策略提升过程中迭代误差与迭代次数的关系，并基于这一关系给出了在交替更新机制下单一智能体实现最大幅度的策略提升所需要的迭代次数。本文也在实验中验证了完全去中心化的交替策略优化算法相较于基线方法的性能优势。

再次，虽然交替更新可确保策略收敛性，但相较于同步更新，其采样复杂度显著增加。为此，本文引入代理优化目标理论，推导出整体策略单调提升条件的新下界，并将其作为策略更新的代理优化目标。本文证明该优化目标可在完全去中心化条件下由各智能体独立优化，同时保证整体策略的单调收敛，并据此提出了完全去中心化的同步策略优化算法。实验结果也验证了这一具有收敛性保证的策略优化算法对比启发式方法的性能优势。针对代理优化目标因为估计误差而导致的平凡更新问题，本文进一步提出 f-散度策略优化通用框架，分析其整体上的理论性质，并基于该框架开发新型收敛性保证算法。

最后，在许多实际应用场景中，决策受现实条件制约必须采用去中心化方式实施。因此，构建基于实际应用的仿真环境对去中心化算法研究和相关问题的解决具有双重价值。在线广告中的自动出价是典型的完全去中心化决策场景，本文利用马尔科夫决策过程，对这一过程进行了建模，并基于自动出价过程中的合作问题，构建了一个合作多智能体仿真环境。为了提高仿真环境的真实性，本文在仿真环境的设计中应用了深度生成模型。更进一步地，本文依托该仿真环境系统评估了多种去中心化算法的性能表现，再次证明了本文提出的完全去中心化算法的有效性。

本文在中心化训练去中心化执行与完全去中心化两种场景下的多智能体合作任务中都提出了具有收敛性保证的算法，并系统评估了这些算法给智能体带来的各类提升。这为该领域的未来研究提供了有力的支持。未来的研究可以针对完全去中心化范式中策略优化仍然只能进行同策略更新等问题继续开展，得到采样效率更高的收敛性算法。

关键词：多智能体强化学习，合作，完全去中心化学习，独立策略优化

The Study of Policy Optimization Algorithms in Cooperative MARL

Kefan Su (computer application technology)
Directed by Lu Zongqing

ABSTRACT

Reinforcement learning (RL), a machine learning method that learns from interaction and experience, has achieved many breakthroughs in practical applications such as Go, game AI, and aligning large language models with human feedback. As research in reinforcement learning deepens, multi-agent reinforcement learning has gradually become focused, with cooperative multi-agent reinforcement learning being a significant setting in this field. Many decision-making and control tasks in the real world require multiple agents to collaborate, such as autonomous driving and intelligent transportation systems. Compared with single-agent RL, cooperative MARL primarily faces the challenge of environmental non-stationarity - mutual interference caused by simultaneous policy learning among multiple agents. This challenge directly results in difficulty in policy convergence, severely affect both policy stability and performance.

This thesis investigates the design of policy optimization objectives to address environmental non-stationarity, ensure policy convergence, and enhance agents' performances in cooperative MARL tasks. Through optimization objective design, this thesis aim to coordinate agents' policy learning under unified guidance, thereby achieving effective cooperation and guaranteed convergence. Current cooperative MARL algorithms can be categorized into two paradigms based on execution-training patterns: Centralized Training with Decentralized Execution (CTDE) and Decentralized Training with Decentralized Execution (DTDE) or fully decentralized Learning. CTDE permits centralized information exchange (e.g., centralized value functions, communication, parameter sharing) during training, whereas fully decentralized methods prohibit all inter-agent communication throughout both training and execution. Although more challenging, the fully decentralized paradigm can be applied in broader practical scenarios.

For both paradigms, this thesis propose multi-agent policy optimization algorithms with convergence guarantees to improve agents' performances in cooperative tasks. For the CTDE paradigm, this thesis introduce regularization terms into policy optimization objectives, en-

hancing both exploration capability and sampling efficiency while ensuring convergence. For the fully decentralized paradigm, this thesis develop convergence-guaranteed algorithms through alternating update mechanisms and surrogate objectives, achieving policy convergence and performance improvement under both alternating and synchronous update modes. Furthermore, we explore practical applications of fully decentralized algorithms. The main innovations include four aspects:

First, in CTDE scenarios, inspired by entropy regularization in single-agent RL, we propose divergence-based regularization for cooperative MARL. This thesis theoretically analyze properties of the modified Markov Decision Process (MDP) and develop corresponding policy optimization algorithms with proven monotonic convergence. Unlike existing policy gradient-based CTDE methods limited to on-policy updates (even with off-policy correction techniques), this algorithm enables off-policy updates through novel policy iteration design. Experimental results demonstrate its universal compatibility with various baseline methods and validate theoretical properties.

Second, for fully decentralized scenarios, this thesis proposes an alternating policy optimization algorithm that ensures convergence through alternating value function updates while being able to be completed in the fully decentralized condition. The relationship between iteration errors and update counts is established in this thesis, deriving optimal iteration numbers for maximal policy improvement. Experiments confirm the superiority of this alternating policy optimization algorithm over baseline methods.

Third, to address the high sample complexity of alternating updates, this thesis develops a synchronous policy optimization algorithm using surrogate objectives. By deriving novel lower bounds for monotonic policy improvement, this thesis enables decentralized independent optimization while guaranteeing convergence. Experimental results validate its advantages over heuristic methods. To address trivial updates caused by estimation errors, this thesis proposes an f-divergence policy optimization framework with theoretical guarantees. Based on this framework, a novel convergence-guaranteed algorithm with less estimation errors is proposed and verified by empirical results in experiments.

Finally, recognizing the practical need for decision-making under decentralized conditions in many applications, this thesis construct a realistic simulation environment using Markov Decision Processes (MDPs) for online advertising's auto-bidding scenario - a typical fully decentralized decision-making problem. Enhanced by deep generative models, the similarity between the simulation environment and real-world application is improved. This environment

conducts systematic evaluation of decentralized algorithms, further verifying our methods' effectiveness.

This work provides convergence-guaranteed algorithms for both CTDE and fully decentralized cooperative MARL scenarios, systematically evaluating their performance improvements. It offers substantial support for future research, particularly in addressing remaining challenges like on-policy update limitations in fully decentralized paradigms to develop higher-efficiency convergent algorithms.

KEY WORDS: Multi-Agent Reinforcement Learning, Cooperation, fully decentralized learning, independent policy optimization