

**本科生毕业论文（设计）**



**多智能体均值-方差线性组合最小强化学习算法预研**

|  |  |
| --- | --- |
| 学 院： | 计算机学院 |
| 专 业： | 计算机科学与技术 |
| 学生姓名： | 王钰 |
| 学 号： | 2020141420210 |
| 年 级： | 2021级 |
| 指导教师： | 郭宏亮 |

2025年3月3日

多智能体均值-方差线性组合最小强化学习算法预研

专业 计算机科学与技术

学生姓名 王钰 指导教师 郭宏亮

**[摘要]** 当下，随着人工智能的飞速发展，多智能体强化学习（Multi-Agent Reinforcement Learning, MARL）逐步成为研究者们关注的热点。现实的应用场景中通常同时存在多个决策个体，因此，MARL方法对于构建在现实场景中更可靠的智能体有着极其重要的意义。当前的MARL方法仍然面临着多智能体环境随机性过大，从而智能体的学习不够鲁棒的挑战。多智能体环境中的随机性主要来源于两个方面：第一，由于环境的随机性，智能体们在特定状态做出特定动作的情况下，从环境中获得的奖励可能是随机的；第二，由于环境中存在多个智能体，单个智能体获得的奖励因为其它智能体的存在而具有随机性。然而，现有的MARL方法主要基于时序差分学习或策略梯度学习的方法对智能体获得的回报的均值进行建模，也就无法描述多智能体环境中智能体获得的奖励的随机性，无法让智能体更鲁棒地学习到最优决策。为了解决上述挑战，本文提出一种全新的强化学习的建模方法—多智能体级联时序差分（Multi-Agent Cascaded Temporal Difference, MACTD），对智能体获得的回报的均值和方差同时进行建模，让智能体根据回报的期望和随机性的大小做出更智能的决策。实验表明，相较于其它算法，本文提出的MACTD方法可以让智能体更鲁棒地学习到最优决策。

**[关键词]** 多智能体强化学习；环境随机性；均值-方差建模；多智能体级联时序差分

**Multi-Agent Mean-Variance Linear Combination Minimization Reinforcement Learning Algorithm Pre-research**

Computer science

Student: Yu Wang Adviser: Hongliang Guo

**[Abstract]** With the rapid advancement of artificial intelligence, Multi-Agent Reinforcement Learning (MARL) has gradually become a focal point of research. In real-world application scenarios, multiple decision-making agents often coexist simultaneously. Therefore, MARL methods hold significant importance in constructing more reliable agents for practical scenarios. However, current MARL methods still face the challenge of excessive randomness in multi-agent environments, which leads to insufficient robustness in agents' learning processes. The randomness in multi-agent environments primarily stems from two aspects. First, due to the stochastic nature of the environment, the rewards agents receive from the environment for taking specific actions in specific states may be random. Second, the presence of multiple agents in the environment introduces additional randomness, as the rewards received by a single agent are influenced by the actions of other agents. Existing MARL methods mainly model the mean of the rewards obtained by agents using Temporal Difference learning or Policy Gradient learning. These approaches fail to capture the randomness of rewards in multi-agent environments, thereby preventing agents from learning optimal decisions in a more robust manner. To address the aforementioned challenges, this paper proposes a novel reinforcement learning modeling method—Multi-Agent Cascaded Temporal Difference (MACTD). This method models both the mean and variance of the rewards obtained by agents, enabling them to make smarter decisions based on the expected rewards and the magnitude of randomness. Experiments demonstrate that, compared with other algorithms, the proposed MACTD method allows agents to learn optimal decisions more robustly.

**[Key Words]** MARL, Environmental Stochasticity, Mean-Variance Modeling, MACTD

目录

[多智能体均值-方差线性组合最小强化学习算法预研 1](#_Toc191889131)

[1 一级标题1 4](#_Toc191889132)

[1.1 二级标题 4](#_Toc191889133)

[1.1.1 三级标题 4](#_Toc191889134)

[2 第二个一级标题 4](#_Toc191889135)

[3 一级标题temp 4](#_Toc191889136)

[3.1 二级标题temp 4](#_Toc191889137)

# 背景介绍

正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正文正(小四，宋体，固定行距20磅，如果因此遇到图片插入异常，请选择 段落-间距-行距 中的单倍行距即可改正。)

# 第二个一级标题

# 一级标题temp

## 二级标题temp