# 论文（设计）选题目的、工作任务

## 选题目的

当下，随着人工智能（Artificial Intelligence, AI）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。在现实的应用场景中，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智能的智能体有着极其重要的意义。

然而，当前的多智能体强化学习方法仍然面临着多智能体环境中随机性过大的挑战。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很可能是由其它智能体的行为带来的，因此，对于单个智能体来说，获得的奖励是随机的。现有的MARL方法主要基于时序差分（Temporal Difference, TD）学习方法对智能体获得的回报的均值进行建模。然而，对均值的建模无法描述多智能体环境中的随机性，也就无法解决多智能体环境中随机性过大的挑战。

本文旨在通过一种级联时序差分（Cascaded Temporal Difference, CTD）方法对智能体获得的回报的均值和方差进行建模，从而精确地描述智能体可以获得的奖励的情况；通过价值分解网络（Value Decomposition Networks, VDNs）将CTD方法应用到多智能体环境中，从而让多智能体环境中的每个智能体可以预测将获得的回报的均值和方差，让多智能体系统拥有建模环境中随机性的能力。

## 2、工作任务

1、对CTD的公式进行推导。

2、将CTD的公式扩展到深度神经网络的情况。

3、将基于深度神经网络的CTD扩展到多智能体的情况。

4、完成整体公式推导以及代码编写。

5、构建多个多智能体强化学习环境，对提出的算法进行测试。

6、完成相关文档的编写。

# 二、目前资料收集情况（含指定参考资料）

[1]. Guo, H., Hou, X., & Peng, Q. (2021). CTD: Cascaded temporal difference learning for the mean-standard deviation shortest path problem. *IEEE Transactions on Intelligent Transportation Systems*, *23*(8), 10868-10886.

[2]. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *nature*, *518*(7540), 529-533.

[3]. Sunehag, P., Lever, G., Gruslys, A., Czarnecki, W. M., Zambaldi, V., Jaderberg, M., ... & Graepel, T. (2017). Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296*.

[4]. Rashid, T., Samvelyan, M., De Witt, C. S., Farquhar, G., Foerster, J., & Whiteson, S. (2020). Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, *21*(178), 1-51.

[5]. Tampuu, A., Matiisen, T., Kodelja, D., Kuzovkin, I., Korjus, K., Aru, J., ... & Vicente, R. (2017). Multiagent cooperation and competition with deep reinforcement learning. *PloS one*, *12*(4), e0172395.

[6]. Oroojlooy, A., & Hajinezhad, D. (2023). A review of cooperative multi-agent deep reinforcement learning. *Applied Intelligence*, *53*(11), 13677-13722.

[7]. 罗彪,胡天萌,周育豪,等.多智能体强化学习控制与决策研究综述[J/OL].自动化学报,1-30[2024-11-24].https://doi.org/10.16383/j.aas.c240392.

[8]. 李明阳,许可儿,宋志强,等.多智能体强化学习算法研究综述[J].计算机科学与探索,2024,18(08):1979-1997.

[9]. 李茹杨,彭慧民,李仁刚,等.强化学习算法与应用综述[J].计算机系统应用,2020,29(12):13-25.DOI:10.15888/j.cnki.csa.007701.

[10]. Zhang, K., Yang, Z., & Başar, T. (2021). Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, 321-384.

[11]. Tesauro, G. (1995). Temporal difference learning and TD-Gammon. *Communications of the ACM*, *38*(3), 58-68.

# 三、论文（设计）完成计划（含时间进度）

1. 2024.10.15-2024.11.25： 文献阅读、调研

2. 2024.11.25-2024.12.20： 完成开题报告

3. 2024.12.20-2025.02.01： 完成理论公式推导

4. 2025.02.01-2025.03.20： 完成程序编写、算法训练和测试

5. 2025.03.20-2025.04.15： 完成毕业论文初稿撰写

6. 2025.04.15-2025.05.01： 完成毕业论文修改

7. 2025.05.01-2025.05.06： 完成毕业论文定稿