****

**本科毕业论文（设计）开题报告**

****

**题 目 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**

**指导教师 郭宏亮**

**教务处制表**

**2024　年　12 月 　5 日**

|  |  |
| --- | --- |
| **选**  **题**  **意**  **义** | **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MARL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。然而，现有的MARL方法主要基于时序差分（Temporal Difference, TD）学习的方法对智能体获得的回报的均值进行建模；对均值的建模无法描述多智能体环境中智能体获得的奖励的随机性，也就无法解决多智能体环境中随机性过大的挑战。**  **为了解决上述挑战，本项目旨在扩展TD学习，对智能体获得的回报的均值和方差同时进行建模，从而精确地描述智能体可以获得的奖励的情况；通过MARL方法将这样同时建模均值和方差的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。** |
| **国**  **内**  **外**  **研**  **究**  **现**  **状**  **概**  **述** | **由于MARL方法是在单智能体强化学习的基础上发展出来的，所以本章分为两个部分：国内外单智能体强化学习方法研究现状概述和国内外多智能体强化学习方法研究现状概述。**  1.国内外单智能体强化学习方法研究现状概述  **在单智能体强化学习中，智能体需要学习到一个策略，使得它在与环境交互的过程中可以获得尽可能大的累计奖励；在单智能体强化学习方法中，Q-learning是最流行的强化学习方法的一种[1]。Q-learning算法基于时序差分让智能体学习每个状态下每个动作可能获得的回报的均值，从而做出最优决策。通常，Q-learning方法会让智能体采用ε-贪心策略进行动作选择[2]。**  **Q-learning被Watkins博士首次提出[3]。Watkins博士[3]将智能体与环境交互并进行学习的过程建模为马尔可夫决策过程（Markov Decision Process, MDP），并将时序差分、动态规划等思想结合到一起，提出了Q-learning的概念。Q-learning通过在程序内维护一张二维的Q表格，来计算智能体在状态s下采取动作a可以获得的回报的期望；每次与环境交互时，程序都会通过状态、动作以及从环境中获得的实际奖励来更新Q表格的一部分。在Q-learning被提出之后，Watkins等人[4]证明了Q-learning算法的收敛性：假如智能体重复探索过了所有状态下所有可能的动作，那么Q-learning中的Q表格会收敛到最优的状态-动作价值函数。**  **在Q-learning提出后，许多学者针对Q-learning算法做了相关改进。Q-learning算法需要在内存中维护一张状态-动作的二维Q表格，当状态、动作数量过多或者智能体处于连续观测空间时，Q-learning算法就会产生极高的内存开销；所以Martin[5]提出了神经拟合Q值（Neural Fitted Q, NFQ）算法，用神经网络计算在状态s下执行动作a的Q值，代替从Q表格中查询在状态s下执行动作a的Q值，很大程度上节省了内存并提高了数据利用的效率。Mnih等人[6]提出了深度Q网络（Deep Q-Network, DQN），DQN可以直接将每个状态的视觉图像作为输入，并且根据状态s计算所有动作的分数，通过经验回放让智能体根据过往的经验更新网络参数；DQN中包含一个用于计算Q值的网络和一个计算目标值的网络，二者的参数不同步更新，一定程度上避免了计算的Q值过高；DQN在Atari游戏上的表现远优于当时的其它算法。**  **上述强化学习算法的目标都是找到最优的状态-动作价值函数Q，即让智能体预测在状态s下选择动作a可以获得的回报的均值。然而，在环境中存在的随机性过大的情况下，仅仅计算回报的均值是不够的。因此，Guo等人[7]提出了级联时序差分学习的方法，利用两个级联的Q表格，让智能体接连预测在状态s下选择动作a可以获得的回报的均值和方差，并推导出了Q表格更新的公式；他们的算法在实际的交通路网上高效地解决了可靠最短路径的问题。**  2. 国内外多智能体强化学习方法研究现状概述  **在MARL中，每个智能体都需要学习到自己的最优策略，使得所有智能体一起行动时多智能体系统可以获得尽可能大的累积奖励[8]。随着MARL方法广泛应用到人工智能相关领域，用于解决路径规划、博弈、任务分配等问题，越来越多的研究人员将目光聚集在了MARL方法上[9]。MARL方法可以根据智能体之间的关系大致分为三类：完全合作式、完全竞争式以及合作-竞争混合式[10]。在完全合作式MARL中，所有智能体需要相互协作以完成共同的团队任务，它们的目标是最大化团队累积奖励；在完全竞争式MARL中，每个智能体只关注自身行为，它们的目标是最大化自身累积奖励；在合作-竞争混合式MARL中，每个智能体既要考虑自身利益也要考虑团队利益，它们的优化目标是在最大化自身累积奖励的同时通过与其他智能体的合作最大化团队累计奖励。**  **对于完全合作式、完全竞争式或是合作-竞争混合式，存在一种通用的MARL方法，也就是独立Q学习（Independent Q-Learning, IQL）[11]。IQL算法中，每个智能体都会把其它智能体看作环境的一部分，也就是说，每个智能体都在解决一个单智能体强化学习任务；对于不同的任务，只需要修改每个智能体获得奖励的方式即可。然而，由于在IQL的建模中，环境中存在其它智能体，所以环境是非稳态的，算法的收敛性无法从理论上得到证明；且智能体之间无法进行任何形式的信息共享。接下来，本文将分别对完全合作式、完全竞争式以及合作-竞争混合式这三类MARL方法进行介绍。**  **在完全合作式MARL中，智能体们会因为完成团队目标而统一获得奖励，不会有个体的奖励。这种共享团队奖励的方式会带来问题，即一个智能体由于只能获得自己的局部观测，从而无法确定是因为自己的行为或者其它智能体的行为而获得奖励；除此之外，由于部分智能体可以学习到比较好的策略以完成团队任务，其它智能体就会变得“懒惰”。为了解决上述问题，Sunehag等人[12]提出了价值分解网络（Value Decomposition Networks, VDN），将团队的Q值分解为各个智能体的Q值之和，每个智能体用深度神经网络来拟合自己的Q值；这样，每个智能体就可以分配到合理的奖励值，且一定程度上避免了懒惰智能体的问题。Rashid等人在VDN的基础上提出了QMIX[13]。Rashid等人通过混合网络将团队的Q值分解为各个智能体的Q值的复杂非线性组合，并在训练中加入全局信息，使得每个智能体学习到的策略更精确；且通过保证混合网络参数的非负性，保证团队Q值与每个智能体计算的Q值的关系是单调的，从而保证集中策略和分散策略的一致性。Kyunghwan等人[14]提出了QTRAN，为价值分解提供了更多了理论证明，并且将每个智能体的价值进行进一步映射，保证每个智能体学习到的状态-价值函数可以逼近最优。**  **在完全竞争式MARL中，不同的智能体的目标通常是不同的，甚至互相之间进行博弈。Micheal[15]针对二人零和随机博弈场景提出了Minimax-Q算法，使用Q-learning中的时序差分来迭代更新minimax算法的求解公式，即让每个智能体最大化在博弈中最差情况下的期望奖励值，最终学习到纳什均衡策略。Hu[16]等人针对多智能体博弈设计了Nash Q-learning，旨在在多智能体相互竞争或合作的环境中找到纳什均衡作为各个智能体的最优策略；其中，每个智能体在给定其它智能体策略的情况下，通过求解纳什均衡，选择一个能最大化自身收益的动作，并更新状态-动作价值函数。**  **在合作-竞争混合式MARL中，智能体需要兼顾自身和团队的利益。Ryan等人[17]提出了多智能体深度确定性策略梯度（Multi-Agent Deep Deterministic Policy Gradient, MADDPG）算法；MADDPG算法是一种基于深度确定性策略梯度的扩展，用于解决多智能体环境中的协作和竞争问题。MADDPG通过独立的演员-评论家架构让每个智能体在训练中都可以了解其他智能体的策略信息，以改善学习效果和稳定性。Mnih等人[18]提出了异步优势演员-评论家（Asynchronous Advantage Actor-Critic, A3C）算法；A3C算法是一种策略梯度算法，它通过并行执行多个智能体来提高训练效率，即允许异步更新全局模型。A3C结合了价值和策略梯度方法，并使用n步回报来更新策略和价值函数，从而在复杂的围棋、星际争霸等任务中取得了显著的效果。**  **总而言之，现有的MARL方法只对智能体获得的回报的期望进行建模，从一定意义上讲，这些方法在随机性很强的多智能体环境中不会学习到鲁棒的最优策略。在单智能体强化学习方法中，有学者提出了CTD对智能体获得的回报的均值和方差同时进行建模，然而这种方法没有考虑一般强化学习中的折扣因子，只应用在了简单的交通路网环境中，并不能应用到其它领域。** |
| **主**  **要**  **研**  **究**  **内**  **容** | **本项目主要研究如何建模MARL环境中的随机性，从而使智能体学习到的策略更加鲁棒。本项目将对智能体获得的回报的均值和方差进行建模，利用回报的均值和方差的线性组合来描述MARL环境下考虑到随机性的最优策略。**  **为了同时对智能体获得的回报的均值和方差进行建模，本项目将完成一般情况下级联时序差分公式的推导，包括智能体状态-动作价值函数更新公式的推导。由于MARL环境较为复杂，通常需要神经网络对状态-动作价值函数进行拟合，所以本项目将用深度神经网络实现级联时序差分，并给出相应的公式推导。接下来，本项目将会给出一般MARL情况下基于深度神经网络的级联时序差分的公式推导，从而理论上完成对MARL情况下智能体获得的回报的均值和方差的建模。**  **在完成理论推导后，本项目将完成算法相关代码的编写及验证，并在一般的MARL环境下训练，调整算法中不同的超参数，与其它算法进行对比来测试算法的有效性；最后，将测试环境可视化，来更好地展示算法的效果。** |

|  |  |
| --- | --- |
| **拟采用**  **的研究**  **思 路**  **（方法、**  **技术路**  **线、可**  **行性论**  **证等）** | 方法 **本项目将采用一般科研工作的科研方法：**   1. **进行调研。收集并阅读相关领域的前沿文献，对文献中的方法、文献提出的方法可以解决的问题以及方法的不足进行整理，从而对相关领域有深入的了解。** 2. **确定研究目标。根据当前研究的局限性或缺陷确定研究目标。在本项目中，由于当前MARL方法没有很好的建模MARL环境下的随机性，所以本项目将对MARL环境下的随机性的建模确定为研究目标。** 3. **文献精读。在确定研究目标后，对和研究目标相关的一小部分文献更加精细的阅读，并对相关可能用到的基础知识进行学习。** 4. **确定大概的技术路线，完成可行性论证。** 5. **完成相关公式推导和理论证明，构建起一套新的方法。** 6. **编写代码实现本项目提出的方法。** 7. **对提出的方法在不同的环境中进行训练和测试，验证算法的有效性。** 8. **编写代码，对训练好的算法进行可视化展示。**  技术路线 **首先，通过知网、四川大学图书馆等方式进行文献查找，并阅读文献。在确定项目的研究目标后，采用MARL、深度神经网络等理论以及概率统计等课本知识对相关公式进行推导，并完成算法的理论证明。接下来，计划使用python完成代码的编写：采用pytorch、numpy等模块完成核心算法代码的编写，采用streamlit等模块完成可视化，采用pettingzoo、gymnasium等模块完成MARL环境的搭建；在本地或者云端服务器上搭建MARL环境，并在环境中完成算法的训练以及测试，最后将训练好的算法在web上进行可视化展示。** 3、可行性论证 **本项目具有充分的可行性，具体原因如下：**   1. **可以通过互联网和图书馆等渠道获取到所需的文献以及查询相关知识；通过相关翻译软件辅助完成文献阅读。** 2. **根据对现有文献的总结，项目的研究目标、研究方向已经确定，且具体的方法可以初步规划：推导可以预测智能体获得的回报的均值和方差的公式，并利用深度神经网络对公式进行优化，并扩展到MARL环境中。** 3. **在完成公式推导后，可以根据公式利用python完成相关编码；且现在个人计算机的计算能力可以完成算法在一些简单的MARL环境下的训练和测试；如果个人计算机的计算能力不够或计算时间过长，则考虑借助云端服务器的算力。** 4. **利用相关web展示的模块即可完成算法的可视化。** 5. **根据其它文献或者github代码，可以完成对相关文献代码的复现，即可完成算法之间的比较。**   **因此，本项目具有充分的可行性。** |
| **研**  **究**  **工**  **作**  **安**  **排**  **及**  **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**  **其中，预测智能体回报的均值和方差的级联时序差分的公式推导预计用时20天，基于深度神经网络实现级联时序差分方法的公式推导预计用时10天，将上述方法扩展到MARL环境的公式推导预计用时10天。**  **4. 2025.02.01-2025.03.20： 完成程序编写、算法训练和测试**  **其中，程序编写预计用时20天，算法训练和测试预计用时30天。**  **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** |
| **参**  **考**  **文**  **献**  **目**  **录** | 1. **Thrun, S., & Littman, M. L. (2000). Reinforcement learning: an introduction. *AI Magazine*, *21*(1), 103-103.** 2. **Jang, B., Kim, M., Harerimana, G., & Kim, J. W. (2019). Q-learning algorithms: A comprehensive classification and applications. IEEE access, 7, 133653-133667.** 3. **Watkins, C. J. C. H. (1989). Learning from delayed rewards.** 4. **Watkins, C. J., & Dayan, P. (1992). Q-learning. Machine learning, 8, 279-292.** 5. **Riedmiller, M. (2005). Neural fitted Q iteration–first experiences with a data efficient neural reinforcement learning method. In *Machine learning: ECML 2005: 16th European conference on machine learning, Porto, Portugal, October 3-7, 2005. proceedings 16* (pp. 317-328). Springer Berlin Heidelberg.** 6. **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.** 7. **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.** 8. **Oroojlooy, A., & Hajinezhad, D. (2023). A review of cooperative multi-agent deep reinforcement learning. *Applied Intelligence*, *53*(11), 13677-13722.** 9. **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.** 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. **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.** 12. **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.** 13. **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.** 14. **Son, K., Kim, D., Kang, W. J., Hostallero, D. E., & Yi, Y. (2019, May). Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *International conference on machine learning* (pp. 5887-5896). PMLR.** 15. **Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994* (pp. 157-163). Morgan Kaufmann.** 16. **Hu, J., & Wellman, M. P. (2003). Nash Q-learning for general-sum stochastic games. *Journal of machine learning research*, *4*(Nov), 1039-1069.** 17. **Lowe, R., Wu, Y. I., Tamar, A., Harb, J., Pieter Abbeel, O., & Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, *30*.** 18. **Mnih, V. (2016). Asynchronous Methods for Deep Reinforcement Learning. *arXiv preprint arXiv:1602.01783*.** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **开 题 报 告 会 议 纪 要** | | | | | |
| **时 间** | **2024年12月2日** | **地点** | **华西校区信息中心楼205** | **主持人** | **李征** |
| **参**  **会**  **教**  **师** | **姓 名** | **职务（职称）** | **姓 名** | **职务（职称）** | |
| **武岳** | **教授** |  |  | |
| **周颖杰** | **副教授** |  |  | |
| **郭宏亮** | **副教授** |  |  | |
| **会**  **议**  **记**  **录**  **摘**  **要** | **在会议中，主要确定了关于本项目的如下问题：**  **武岳老师：你认为这个项目最主要的创新点在哪里？**  **回答：考虑到MARL环境中通常会有很大的随机性，且一般的MARL方法只是针对智能体获得的回报的均值进行建模；所以我打算基于强化学习的基本知识，如贝尔曼方程、分布式部分可观测马尔可夫决策过程等，创新性地推导出针对智能体获得的回报的方差进行建模的公式。最后，将智能体获得的回报的均值和方差进行线性组合，从而在智能体的角度完成对环境中一般情况下获得的奖励（即回报的均值）和环境的随机性（即回报的方差）的建模，让智能体做出更优的决策。**  **周颖杰老师：在你的项目中具体可能要用到哪些技术路线？**  **回答：需要用到一些相关领域的基础知识和前沿技术。基础知识方面，需要用到强化学习领域的基础知识，如贝尔曼方程、分布式部分可观测马尔可夫决策过程、Q-learning等；用到概率统计、高等数学等基础数学知识；在多智能体环境下，还需要用到深度神经网络的基础知识。前沿技术方面，需要参考先进的MARL方法，如IQL、VDN等；还需要考虑单智能体强化学习方法中较为重要的文献，如TD、DQN等。**  **郭宏亮老师：为什么会想出这样的创新点？这样的创新点是否有充足的可行性？**  **回答：受CTD论文的启发，我认为对于智能体获得的回报的方差的建模可以推广到更一般的情况；然而CTD论文中只考虑了没有折扣因子和单智能体在交通路网的特殊情况，所以我计划对相关公式进行重新推导，并且与深度神经网络和MARL方法结合，完成最终的公式推导。我认为这样的方法有充足的可行性，也就是最后可以推导出在一般情况下均值和方差的状态-动作价值函数的更新公式；并且根据更新公式可以完成代码的编写；#todo环境、算力。综合以上考虑，我认为本项目的可行性很充分。**  **郭宏亮老师：你的项目最终预计将会达到什么程度？如何去评估项目中你创新算法的有效性？**  **回答：我预计最终会完成整个方法的公式推导和理论证明：包括智能体回报的方差的建模、MARL技术的应用等等；还会基于公式推导完成算法代码的编写；搭建环境，对算法进行训练和测试；最后对训练好的算法进行可视化展示。我预计将用环境中对智能体的要求以及与其它算法的对比来评估算法的有效性；比如，某个环境要求智能体以尽可能短的时间完成某任务，我就会用这个任务中完成任务的时间作为评估指标，并将算法在这个指标上与其它的先进算法进行对比，来评估本项目的算法的有效性。**  **记录人：王钰** | | | | |
| **指**  **导**  **教**  **师**  **意**  **见** | **王钰同学的研究课题具有明确的研究方向和实际应用价值。对多智能体强化学习算法的理解和提出的新算法框架显示了该同学的创新思维和对领域的深刻洞察，同时，对多智能体环境随机性的建模对该方向的研究也具有普适的意义。**  **报告中，研究背景的描述清晰，研究意义阐述充分；计划开展的研究内容具体，方法论合理，有助于系统的完成研究。预期成果的描述具体，符合科研思路，有助于指导后续的研究工作，方便后续学者的研究。**  **希望王钰同学在后续的研究中继续保持这种严谨的态度，同时保持开放的思维，以应对可能出现的新问题和挑战。**  **指导教师签名： 2024 年 12 月 6 日** | | | | |
| **备注：1.本开题报告除第3页各栏目外，其它栏目均由学生填写。2.填写各栏目时可根据内容另加附页。3.参加开题报告会议的教师不少于3人。** | | | | | |