软件学报 ISSN 1000-9825, CODEN RUXUEW
2025,36(1):107−151 [doi: 10.13328/j.cnki.jos.007212] [CSTR: 32375.14.jos.007212]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

混合博弈问题的求解与应用综述*

董绍康[1], 李 超[1], 杨 光[1], 葛振兴[1], 曹宏业[1], 陈武兵[1], 杨尚东[1,2], 陈兴国[1,2], 李文斌[1], 高 阳[1]

[1](计算机软件新技术国家重点实验室 (南京大学), 江苏 南京 210023)

[2](南京邮电大学 计算机学院、软件学院、网络空间安全学院, 江苏 南京 210023)

通信作者: 高阳, E-mail: gaoy@nju.edu.cn

摘 要: 近年来, 随着人工智能技术在序贯决策和博弈对抗等问题的应用方面取得了飞速发展, 围棋、游戏、德扑和麻将等领域取得了巨大的进步, 例如, AlphaGo、OpenAI Five、AlphaStar、DeepStack、Libratus、Pluribus 和 Suphx 等系统都在这些领域中达到或超过人类专家水平. 这些应用集中在双人、两队或者多人的零和博弈问题中, 而对于混合博弈问题的研究缺乏实质性的进展与突破. 区别于零和博弈, 混合博弈需要综合考虑个体收益、集体收益和均衡收益等诸多目标, 被广泛应用于公共资源分配、任务调度和自动驾驶等现实场景. 因此, 对于混合博弈问题的研究至关重要. 通过梳理当前混合博弈领域中的重要概念和相关工作, 深入分析国内外研究现状和未来发展方向. 具体地, 首先介绍混合博弈问题的定义与分类; 其次详细阐述博弈解概念和求解目标, 包含纳什均衡、相关均衡、帕累托最优等解概念, 最大化个体收益、最大化集体收益以及兼顾公平等求解目标; 接下来根据不同的求解目标, 分别对博弈论方法、强化学习方法以及这两种方法的结合进行详细探讨和分析; 最后介绍相关的应用场景和实验仿真环境, 并对未来研究的方向进行总结与展望.

关键词: 混合博弈; 博弈论; 强化学习

中图法分类号: TP18

中文引用格式: 董绍康, 李超, 杨光, 葛振兴, 曹宏业, 陈武兵, 杨尚东, 陈兴国, 李文斌, 高阳. 混合博弈问题的求解与应用综述. 软件学报, 2025, 36(1): 107–151. http://www.jos.org.cn/1000-9825/7212.htm

英文引用格式: Dong SK, Li C, Yang G, Ge ZX, Cao HY, Chen WB, Yang SD, Chen XG, Li WB, Gao Y. Survey on Solutions and Applications for Mixed-motive Games. Ruan Jian Xue Bao/Journal of Software, 2025, 36(1): 107–151 (in Chinese). http://www.jos.org.cn/1000-9825/7212.htm

Survey on Solutions and Applications for Mixed-motive Games

DONG Shao-Kang[1], LI Chao[1], YANG Guang[1], GE Zhen-Xing[1], CAO Hong-Ye[1], CHEN Wu-Bing[1], YANG Shang-Dong[1,2], CHEN Xing-Guo[1,2], LI Wen-Bin[1], GAO Yang[1]

[1](State Key Laboratory for Novel Software Technology (Nanjing University), Nanjing 210023, China)

[2](School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: In recent years, there has been rapid advancement in the application of artificial intelligence technology to sequential decision-making and adversarial game scenarios, resulting in significant progress in domains such as Go, games, poker, and Mahjong. Notably, systems like AlphaGo, OpenAI Five, AlphaStar, DeepStack, Libratus, Pluribus, and Suphx have achieved or surpassed human expert-level performance in these areas. While these applications primarily focus on zero-sum games involving two players, two teams, or multiple players, there has been limited substantive progress in addressing mixed-motive games. Unlike zero-sum games, mixed-motive games necessitate comprehensive consideration of individual returns, collective returns, and equilibrium. These games are extensively applied in

* 基金项目: 国家自然科学基金 (62192783, 62106100, 62206133, 62276142); 江苏省自然科学基金 (BK20221441); 江苏省产业前瞻与关键核心技术竞争项目 (BE2021028); 深圳市中央引导地方科技发展资金 (2021Szvup056); 南京大学计算机软件新技术国家重点实验室资助项目 (KFKT2022B12)
收稿时间: 2023-08-03; 修改时间: 2024-01-14; 采用时间: 2024-04-16; jos 在线出版时间: 2024-06-20
CNKI 网络首发时间: 2024-06-21

real-world applications such as public resource allocation, task scheduling, and autonomous driving, making research in this area crucial. This study offers a comprehensive overview of key concepts and relevant research in the field of mixed-motive games, providingan in-depth analysis of current trends and future directions both domestically and internationally. Specifically, this study first introduces the definition and classification of mixed-motive games. It then elaborates on game solution concepts and objectives, including Nash equilibrium, correlated equilibrium, and Pareto optimality, as well as objectives related to maximizing individual and collective gains, while considering fairness. Furthermore, the study engages in a thorough exploration and analysis of game theory methods, reinforcement learning methods, and their combination based on different solution objectives. In addition, the study discusses relevant application scenarios and experimental simulation environments before concluding with a summary and outlook on future research directions.

Key words: mixed-motive game; game theory; reinforcement learning

博弈论 (game theory) 最早在 17 世纪由数学家们构想出来解决赌博、象棋和双人纸牌游戏等零和博弈 (zero-sum game) 问题. 20 世纪 50 年代, 博弈论由 von Neumann 等人正式提出, 出版了博弈论领域的著作《博弈论与经济行为》(Theory of Games and Economic Behavior)[1]. 随着时间的推移, 博弈论逐渐扩展到非零和博弈, 即多方参与并且可以共同获益的情况. Nash 提出了纳什均衡的概念, 即在博弈中每个参与者都选择最优策略时, 没有人可以通过改变自己的策略来获得更高的收益[2]. 随后, 随着博弈论专家的深入研究产生了许多重要的概念, 如合作博弈[3]、子博弈[4]和重复博弈[5]等. 博弈论的应用领域非常广泛: 在经济学中, 博弈论被用于研究市场竞争、拍卖和定价等问题; 在政治学中, 博弈论被用于分析选举、多方合作与地缘冲突等行为; 在生物学中, 博弈论被用于研究物种进化中的合作与竞争策略. 此外, 博弈论还被应用于社会科学、计算机科学和工程学等领域. 总的来说, 博弈论的发展与应用历程丰富, 它为我们理解和解决多智能体系统 (multi-agent system, MAS) 决策问题提供了重要的工具和思维方式.

近年来, 人工智能 (artificial intelligence, AI) 和博弈论技术已经被成功应用在各种任务中, 包括游戏场景、军事作战和工业控制等. 在以上案例和应用中, 以围棋[6,7]、德扑[8–10]、麻将[11]、星际争霸[12]、Dota 2[13]、王者荣耀[14]和军事作战等为代表的双人零和或两队 (多人) 零和博弈问题, 现有方法往往通过蒙特卡洛树搜索、深度强化学习 (deep reinforcement learning, DRL) 和自博弈 (self-play) 等技术, 将人工智能和博弈论方法相结合, 最终使得智能体在以上场景中的表现能够达到或超过人类专家水平. 其次, 在网络路由[15,16]、交通指挥[17,18]、机器人控制[19], 和包括先前介绍的星际争霸多智能体挑战[20]、谷歌足球研究[21]等多智能体游戏对抗中的某一方或一队, 这类场景中往往是双人或多人的合作博弈 (cooperative game) 问题, 所有智能体会共享一个全局的奖赏函数, 需要通过合作完成一个共同目标.

而在一般的混合博弈 (mixed-motive game) 场景下, 包括公共资源分配[22–24]、任务调度[25]和自动驾驶[26,27]等应用, 不同于合作博弈问题, 此类问题中每个智能体都有独立的奖赏函数, 例如上述应用中的每个智能体分配调度的资源、行驶时间等, 因此整个系统既要考虑智能体自身的收益, 也要考虑系统的总体收益, 还要兼顾公平, 最终建立某种均衡准则和社会规范使系统维持稳定. 虽然混合博弈场景更加贴近于现实生活, 但现有的研究工作缺乏系统性的认识和实质性的突破. 例如, 对于该博弈问题解的定义与性质分析仍然沿用传统博弈的解概念, 缺乏对混合博弈特定问题的考虑; 以强化学习 (reinforcement learning, RL) 和博弈论为代表的方法对数学形式化的博弈模型和求出的收敛解没有清晰的范围和界定. 因此, 混合博弈问题亟待开展广泛研究.

目前, 已有诸多对多智能体系统、多智能体强化学习 (multi-agent reinforcement learning, MARL) 和博弈论相关的综述文章, 涉及内容也非常广泛. 首先, 关于多智能体系统的研究方向, 多智能体系统没有一个被普遍接受的定义, 在这个问题上有很多正在进行的辩论和争议[28]. 目前只能给出一些较为宽泛的定义: 多智能体系统是由多个相互作用的智能体追求某组目标或执行某组任务的系统, 这些智能体要么具有不同的信息, 要么具有不同的利益, 或者两者兼而有之[28,29]. 一些文章认为多智能体系统是分布式人工智能 (distributed artificial intelligence, DAI) 的一个子领域, 包括对智能体的架构、通信、协同、决策和学习能力的研究[30,31]. 另外, 一些与人工智能和机器学习结合的多智能体系统相关综述工作讨论了智能体同构或异构的结构, 以及智能体之间是否能够进行通信[32], 而另一个工作则提出了著名的 5 种 AI 算法议程 (agenda): 包括计算型、描述型、标准型、规定的合作型和非合作型[33]. 第二, 关于多智能体强化学习的研究方向, 相关工作对多智能体强化学习的目标和算法分类进行了

综述, 分别讨论了基于值和基于策略的方法在完全合作型、完全竞争型以及混合型任务上的应用[34], 后续相关文章又针对两种经典的扩展式博弈和随机博弈对多智能体强化学习的理论收敛性进行分析, 特别是之前工作很少涉及的扩展式博弈中的学习, 带有网络连接的独立式学习, 平均场状态下的多智能体强化学习, 基于策略的学习方法等的收敛性分析[35]. 在深度学习技术引入后, 又有许多综述文章对部分可观测环境[36]、环境非稳定性[37]、迁移学习[36,38]、强化学习可解释性[39]和强化学习探索[40]等专题进行了细致地梳理和总结. 具体来说, 部分可观测的环境表示与环境相关的完整状态信息在智能体与环境交互时是不知道的. 在这种情况下, 智能体只能观测到环境的部分信息, 并且需要在每个时间步上做出最佳决策, 这类问题通常用部分可观测的马尔可夫决策过程建模[36]. 针对环境非稳定性问题, 从复杂度不断增加的角度, 目前方法可以通过忽视、遗忘、回应对手、学习对手模型和递归推理的心智理论等方式进行解决[37]. 强化学习中关于迁移学习问题的综述, 包含了对任务差异的假设、源任务的选择、任务映射、迁移知识和允许的强化学习算法等 5 种维度的分类[38]. 强化学习可解释性的综述文章, 定义了解释的含义、讨论影响可解释性的因素、划分解释的直观性, 然后根据强化学习的特性, 将解释的内容划分为环境解释、任务解释和策略解释[39]. 最后, 为了解决强化学习中样本效率低下的问题, 强化学习探索的综述文章从单智能体和多智能体强化学习的角度对当前探索方法划分为不确定性导向探索和内在动机导向探索两个方面进行分析[40]. 近年来, 针对博弈论和多智能体强化学习技术相结合的综述文章也分析了两者的起源[41], 并且在多智能体系统中的行为涌现、智能体建模、学习合作和学习通信等方面总结目前多智能体强化学习的方法, 然后从延时反馈奖赏、自博弈和组合维度灾难等方面的挑战分析强化学习、多智能体强化学习和多智能体学习 (multi-agent learning, MAL) 之间的关系[42]. 最后一些综述文章总结了相关方法在零和博弈[43]、合作博弈[44]、混合博弈[45]、势博弈[46]、平均场博弈[46]等不同博弈类型中的应用. 相关综述文章的分类与描述如表 1 所示.

但是, 以上综述文章缺乏对混合博弈特定解概念和求解目标的拓展, 缺乏对不同求解方法的侧重和针对任务类型的描述, 缺乏对相关数学形式化博弈模型和收敛解联系与区别的讨论. 因此, 本文对混合博弈问题进行系统性梳理和分析, 以博弈基础理论为核心, 结合目前的研究现状和发展趋势, 重点介绍混合博弈问题的定义与分类, 研究从现有复杂环境到经典博弈问题的建模方法, 构建状态动作表征、奖赏函数和支付矩阵的定义方法等. 然后深入分析混合博弈问题中的解的概念与性质、总结以经典博弈论、强化学习以及两者结合的求解方法和实际应用, 最后对未来研究方向进行总结与展望.

表 1　相关综述文章的分类与描述

分类	综述文章	描述
多智能体系统	多智能体系统与分布式人工智能[28]	介绍多智能体系统和分布式人工智能的相关定义, 从基础领域到计算机科学和软件工程的相关主题
	多智能体系统的算法、博弈论和逻辑基础[29]	从优化学习算法、博弈论基础定义和概念、多智能体系统的通信、协议、联盟和机制设计等多个专题讨论
	多智能体系统简介[30,31]	从单智能体自主决策、多智能体的架构、协同、通信和协作角度对多智能体系统进行综述分析
	多智能体系统的机器学习视角[32]	讨论智能体同构和异构的问题, 以及智能体之间是否能够进行通信
	多智能体学习的问题[33]	提出多智能体学习中著名的5种AI算法议程
多智能体强化学习	多智能体强化学习综述[34]	讨论基于值和基于策略的强化学习方法在完全合作型、完全竞争型以及混合型任务上的应用
	多智能体强化学习理论和算法[35]	总结和分析扩展式博弈和随机博弈中多智能体强化学习算法的理论收敛性
	多智能体深度强化学习方法综述[36]	综述多智能体深度强化学习相关问题的研究方法, 包括环境非平稳性、部分可观测性、连续状态和动作空间、多智能体训练方案、多智能体迁移学习
	环境非稳定性[37]	从博弈论和强化学习的角度对非平稳性问题进行分析, 并将现有方法总结成5种解决方案
	迁移学习[38]	根据迁移学习的能力和目标进行分类, 并讨论迁移学习工作的未来研究方向
	强化学习可解释性[39]	定义解释的含义、讨论影响可解释性的因素、划分解释的直观性, 并从不同的解释维度进行综述分析
	强化学习探索问题[40]	从单智能体和多智能体强化学习的角度对当前探索方法划分为不确定性导向探索和内在动机导向探索的两个方面进行分析和总结

表 1 相关综述文章的分类与描述 (续)

分类	综述文章	描述
博弈论与多智能体强化学习	多智能体强化学习基础和现代方法[41]	介绍多智能体系统的概念, 多智能体强化学习的挑战和方法, 并分析博弈论和强化学习的起源
	多智能体深度强化学习的综述与批判[42]	总结多智能体强化学习的方法, 分析强化学习、多智能体强化学习和多智能体学习之间的关系
	竞争和合作的多智能体学习综述[43]	从竞争博弈和合作博弈的角度概述多智能体学习在一系列领域的研究, 包括强化学习、演化计算、博弈论、复杂系统、智能体建模和机器人
	合作型多智能体学习综述[44]	从团队学习和并发学习角度综述合作型多智能体学习方法, 探讨沟通、任务分解、可扩展性和自适应动力学方面的开放性问题
	多智能体强化学习综述[45]	在完全合作、完全竞争和混合场景中分析相关多智能体强化学习算法的代表性选择及优缺点
	多智能体强化学习的博弈论视角[46]	介绍多智能体强化学习的基本知识, 包括问题的表述、基本解决方案和存在的挑战, 并从博弈论的角度提供当前先进技术的独立评估

本文第 1 节介绍混合博弈问题的定义与分类, 并介绍相关背景知识. 第 2 节介绍从现实环境到经典博弈问题的建模方法. 第 3 节详细阐述博弈解的概念和求解目标, 包含纳什均衡、相关均衡、帕累托最优等解概念, 最大化个体收益、最大化集体收益以及兼顾公平等求解目标等. 第 4 节综述目前已有的经典博弈论方法、强化学习方法以及两者结合方法. 第 5 节介绍经典的应用场景与仿真环境. 第 6 节介绍当前混合博弈问题的挑战与发展. 最后第 7 节对全文进行总结. 全文结构如图 1 所示.

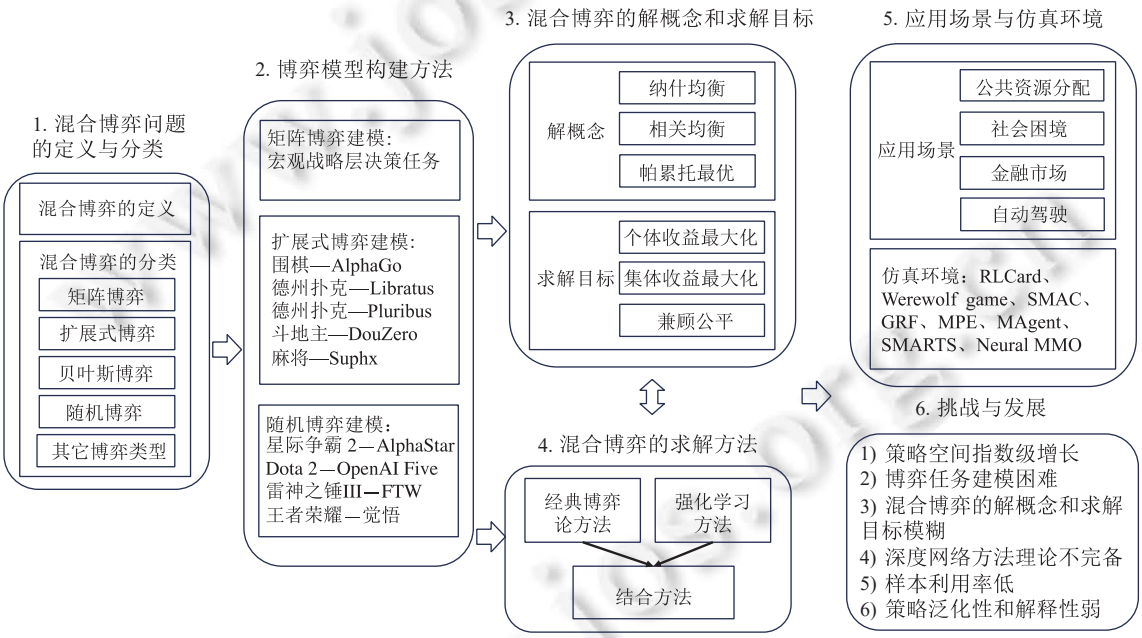


图 1 本文的整体结构图

1 混合博弈问题的定义与分类

合作博弈中所有参与者的回报或奖赏是相同的, 相互之间没有收益冲突; 在多人/多队零和博弈中, 不同人/队的参与者的收益是完全冲突和对立, 一方收益的增加必然会导致其他方收益之和的减少, 因此为纯竞争型博弈. 而混合博弈问题[47]区别于以上合作博弈和零和博弈: 第一, 参与者的收益偏好既不完全相同, 也不完全相反, 因此在这种场景中有复杂且动态变化的博弈关系, 部分参与者可能会相互竞争, 也可能会相互合作, 并且会随时间和空间

的发展而发生变化; 其次, 混合博弈中的智能体收益之和是一般和, 而不是零和或常量和, 不同博弈结果的收益总和是不同的, 一方收益的增加并不一定会导致其他方收益之和的减少, 因此这对参与者之间的隐式的合作创造了可能性; 最后, 博弈关系除了引发在参与者之间 (interpersonal) 的收益冲突, 同时也会引发个人 (intrapersonal) 心理冲突, 参与者会犹豫在下一时刻应该选择偏好合作还是竞争的策略. 下面, 我们将介绍混合博弈问题的定义.

定义 1 (混合博弈). 一个混合博弈问题可以表示为三元组 $< N, \mathcal{A} = \{A_i\}_{i\in N}, R = \{R_i\}_{i\in N} >$ [48]:

- 有限参与智能体集合 $N = \{1, \ldots, n\}$, 其中 n 为智能体的最大数量;
- 有限动作空间 $\mathcal{A} = \{A_1, \ldots, A_n\}$, 其中 A_i 为智能体 i 的动作空间;
- 奖赏函数 $R_i : \mathcal{A} \to \mathbb{R}$, 其中 $\mathcal{A}$ 是所有智能体动作的笛卡尔乘积 $\mathcal{A} = A_1 \times A_2 \times \ldots \times A_n$;
- 奖赏函数 R_i 的关系没有特殊的限制, 其和 $\sum R_i$ 为可变任意值.

奖赏函数又被称为支付函数, 下文中统一为奖赏函数. 对于零和博弈问题, 智能体奖赏函数之和 $\sum R_i = 0$; 而对于合作博弈问题, 智能体奖赏函数相等 $R_1 = R_2 = \ldots = R_n$. 根据以上对构成博弈问题的基本 3 要素空间的了解程度, 可以将博弈问题划分为完全信息博弈 (complete information game) 和非完全信息博弈 (incomplete information game)[49]. 完全信息博弈是指在博弈开始前, 所有参与智能体对博弈问题的信息结构有完全的了解且没有任何不确定性; 否则, 为非完全信息博弈. 其次, 对于博弈问题在决策时序上的不同, 可以将博弈问题分为静态博弈 (static game) 和动态博弈 (dynamic game)[49]. 静态博弈是指所有的参与智能体同时选择动作或者不知道其他智能体的具体动作, 动态博弈是指参与智能体选择动作存在先后顺序, 且先前智能体选择的动作信息能够被观测或得到. 最后, 根据博弈参与智能体是否都知道其他智能体的所有行为历史, 可以将博弈问题分为完美信息博弈 (perfect information game) 和非完美信息博弈 (imperfect information game)[49], 完美信息博弈是假设所有博弈智能体都知道其他智能体的所有行为历史, 但是不保证类似完全信息知道其他智能体的动作集合或奖赏函数等博弈结构, 而非完美信息是指不知道其他智能体的博弈具体动作. 下面我们将介绍混合博弈问题的几种经典建模类型, 具体为矩阵博弈、扩展式博弈、贝叶斯博弈、随机博弈和微分博弈、势博弈、平均场博弈、演化博弈等其他博弈类型.

1.1 矩阵博弈

矩阵博弈 (matrix game), 又称作标准式博弈 (normal-form game, NFG) 或战略式博弈 (strategic-form game, SFG), 是对多个智能体单次交互的规范性描述. 由于单次交互的特点, 矩阵博弈通常被用来描述完全信息的静态博弈问题. 首先, 我们对 2 个智能体 2 个动作的 2×2 混合博弈问题进行分析, 图 2 为建模的矩阵博弈, 其中, 智能体 1 动作为 a_1, a_2, 智能体 2 动作为 b_1, b_2.

		智能体 2	
		b_1	b_2
智能体 1	a_1	$r_1(a_1, b_1), r_2(a_1, b_1)$	$r_1(a_1, b_2), r_2(a_1, b_2)$
	a_2	$r_1(a_2, b_1), r_2(a_2, b_1)$	$r_1(a_2, b_2), r_2(a_2, b_2)$

(a) 2 智能体 2 动作的通用矩阵博弈

		智能体 2	
		b_1	b_2
智能体 1	a_1	3, 3	1, 4
	a_2	4, 1	2, 2

(b) 囚徒困境

图 2 2×2 混合博弈矩阵

图 2(a) 为通用的博弈矩阵. 即使是在用 2×2 矩阵表示的混合博弈问题中, 也有许多不同的策略类型[50], 其中包含著名的囚徒困境 (prisoner's dilemma) 博弈, 如图 2(b). 有关囚徒困境的产生原因以及纳什均衡的分析将在后续章节中具体展开.

重复矩阵博弈 (repeated matrix game) 是矩阵博弈在时序交互上的简单推广, 表示智能体会重复参与原始矩阵博弈有限次或者无限次. 在有限次重复博弈中, 参与智能体都知道博弈会持续特定且有限的回合数, 并且在进行到最大回合后博弈结束; 而在无限次重复博弈中, 参与智能体会拥有无限回合数的博弈. 重复矩阵博弈区别于传统的矩阵博弈 $< N, \mathcal{A} = \{A_i\}_{i\in N}, R = \{R_i\}_{i\in N} >$, 在每一个时刻 $t = 1, 2, \ldots, T$ 时都会进行一次博弈决策, T 是有限或者无限的. 而在每一步 t 上选择动作的时候, 智能体会根据先前的所有智能体的动作历史信息 $h^t = \{a^0, \ldots, a^{t-1}\}$ 进行策略

选择, 其中 $a^{t-1}=\{a_1^{t-1},\ldots,a_n^{t-1}\}$ 表示 n 个智能体在 $t-1$ 时刻的联合动作集合. 这种重复矩阵博弈在博弈论的应用上有诸多经典案例, 例如图 2(b) 中的囚徒困境博弈在重复有限次或无限次时就是著名的序贯社会困境 (sequential social dilemma)[51,52].

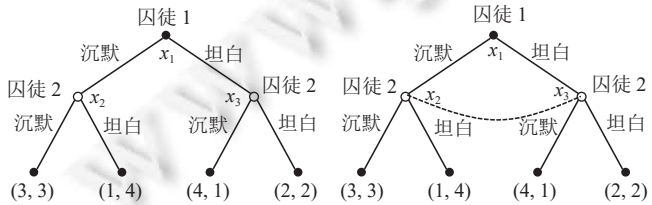
1.2 扩展式博弈

扩展式博弈 (extensive-form game, EFG)[53]区别于上述的矩阵博弈, 主要侧重博弈问题的时序分析, 考虑博弈智能体在决策过程中的先后顺序, 因此通常被用来描述完全信息的动态博弈问题. 下面介绍扩展式博弈的定义.

定义 2 (扩展式博弈 (EFG)). 一个完全信息的扩展式博弈可以表示为五元组 $\langle N,H,P,\mathcal{A},R=\{R_i\}_{i\in N}\rangle$:

- 参与智能体集合 $N=\{1,\ldots,n\}$, 其中 n 为智能体的最大数量;
- 参与智能体历史 (history) 集合 H, 一段历史的每一个组成部分都是某个智能体采取的动作, 终点历史集合由 Z 表示;
- 参与智能体的决策顺序, 参与人函数 $P(h)$ 表示在历史 $h\in H$ 后进行决策的智能体;
- 有限动作空间 $\mathcal{A}=\{A_1,\ldots,A_n\}$, 其中 A_i 为智能体 i 的动作空间;
- 奖赏函数 $R_i:Z\to\mathbb{R}$, 其中 Z 是历史集合 H 中所有的终点集合.

将图 2(b) 的囚徒困境问题进行时序化, 也就是囚徒 1 先进行决策, 囚徒 2 后进行决策, 可以将该矩阵博弈变为上述扩展式博弈的描述: (1) 参与人集合 $N=\{1,2\}$ 为囚徒 1 和囚徒 2; (2) 历史集合 $H=\{\varnothing,$(沉默), (坦白), (沉默, 沉默), (沉默, 坦白), (坦白, 沉默), (坦白, 坦白)$\}$ 等 7 段历史, 分别是从初始历史 $\varnothing$ 到 2 个囚徒都完成决策的终点历史; (3) 参与智能体的决策顺序, 根据是否知道先前决策智能体的动作可以分为完美信息和非完美信息的扩展式博弈: 完美信息表示囚徒 2 了解囚徒 1 选择的动作; 非完美信息表示囚徒 2 不了解囚徒 1 选择的动作或者 2 个囚徒是同时进行决策的; (4) 奖赏函数与矩阵博弈中的囚徒困境相同. 因此, 可以引入一种简单直观的博弈树 (game tree) 结构进行表示, 如图 3(a) 和图 3(b) 所示, 其中虚线表示非完美信息, 不了解处于哪一个决策节点.



(a) 完美信息扩展式博弈树　(b) 完美信息扩展式博弈树

		囚徒 2			
		{沉默, 沉默}	{沉默, 坦白}	{坦白, 沉默}	{坦白, 坦白}
囚徒 1	沉默	3, 3	3, 3	1, 4	1, 4
	坦白	4, 1	2, 2	4, 1	2, 2

(c) 扩展式博弈对应的博弈矩阵

图 3　扩展式博弈的博弈树结构及矩阵描述

针对完美和非完美信息, 可以引入信息集 (information set) 的概念来表示. 智能体 i 的一个信息集 I_i 表示智能体 i 决策节点的一个集合, 满足以下条件: (1) I_i 中的每个决策节点都只能是智能体 i 的决策节点; (2) 当博弈达到 I_i 中的某个决策节点时, 智能体 i 只知道应该进行决策了但是不知道具体在哪一个决策节点上. 例如在图 3(a) 中, 囚徒 2 的信息集是 $I_2\{x_2\}$ 和 $I_2\{x_3\}$, 集合中只有 1 个元素, 因此完全了解囚徒 1 的动作选择, 为完美信息博弈; 而在图 3(b) 中, 囚徒 2 的信息集是 $I_2\{x_2,x_3\}$, 集合中有 2 个元素, 因此不了解囚徒 1 的动作选择, 为非完美信息博弈. 最后根据信息集上定义决策动作, 例如囚徒 2 的动作 $\{$沉默, 坦白$\}$表示在信息集 $I_2\{x_2,x_3\}$ 上, x_2 决策节点上选择沉默, x_3 决策节点上选择坦白, 基于此可以将图 3(b) 的扩展式博弈转化为矩阵博弈的形式, 如图 3(c) 所示.

1.3 贝叶斯博弈

贝叶斯博弈 (Bayesian game)[54]是用来表示不同智能体的相关特征类型, 因为在某些场景下智能体的特征类型在博弈之前是不知道的, 所以常用来描述非完全信息博弈, 与矩阵博弈结合即可描述非完全信息静态博弈, 与扩展式博弈结合即可描述非完全信息动态博弈. 下面, 我们将介绍贝叶斯博弈的定义.

定义 3 (贝叶斯博弈). 一个贝叶斯博弈问题可以表示为四元组 $<N,\mathcal{T}=\{T_i\}_{i\in N},\mathcal{A}=\{A_i\}_{i\in N},R=\{R_i\}_{i\in N}>$, 与定义 1 不同的是多引入了类型 $\mathcal{T}$ 的概念:

● 智能体的类型 $\mathcal{T}=\{T_i\}_{i\in N}$, 其中 T_i 为智能体 i 的类型.

基于此, 概率 $p_i(t_{-i}|t_i)$ 表示智能体 i 对其他智能体 $-i$ 具体类型的推断概率, 而由于不同类型的智能体在参与博弈时会产生不同的奖赏信息, 因此称作贝叶斯博弈. 例如, 在著名的鹰鸽博弈问题中, 如果两只鸟在捕猎到食物时不知道对方鸟是鹰还是鸽的类型时, 会产生以下 4 种矩阵博弈的形式, 如图 4(a) 所示. 假设鸟 1 是鹰且双方都知道, 而鸟 2 以 p 的概率为鹰, 以 $1-p$ 的概率为鸽, 所以根据 Harsanyi 转换[54], 引入“自然”智能体可以将上述非完全信息博弈转换成完全信息但非完美信息的博弈, 其博弈树如图 4(b) 所示, 其中虚线表示非完美信息, 鸟 1 不知道“自然”智能体 0 选择的关于鸟 2 的类型, 鸟 2 不知道鸟 1 选择的是哪一个动作.

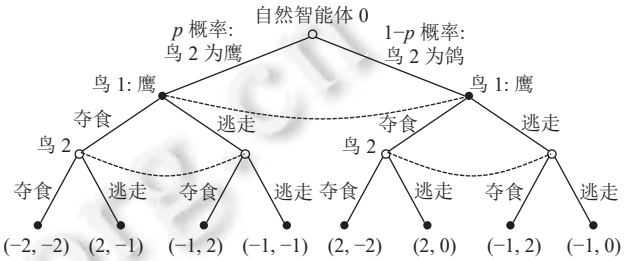
鹰-鹰	夺食	逃走
夺食	−2, −2	2, −1
逃走	−1, 2	−1, −1

鹰-鸽	夺食	逃走
夺食	2, −2	2, 0
逃走	−1, 2	−1, 0

鸽-鹰	夺食	逃走
夺食	−2, 2	2, −1
逃走	0, 2	0, −1

鸽-鸽	夺食	逃走
夺食	1, 1	2, 0
逃走	0, 2	0, 0

(a) 鹰鸽贝叶斯博弈的矩阵形式



(b) 鹰鸽贝叶斯博弈的博弈树结构

图 4 贝叶斯博弈的矩阵表示和博弈树结构

1.4 随机博弈

区别于以上所有博弈, 随机博弈 (stochastic game, SG)[55], 又被称为马尔可夫博弈 (Markov game)[56], 引入了状态 (state) 的概念. 随机博弈是定义智能体与环境的交互过程, 由一系列的阶段组成, 其中的某一个阶段称为阶段博弈 (stage game). 在一个特定的阶段或者状态下, 每一个参与智能体都会同时进行决策, 然后环境会给予奖赏反馈并转移到下一个状态, 然后重复该过程, 进行无限次或者有限次的博弈直至结束. 最后根据每个阶段的收益进行折扣结算, 每个智能体以获得最大的折扣回报. 下面, 我们将会介绍随机博弈的定义.

定义 4 (随机博弈 (SG)). 一个随机博弈可以表示为以下六元组 $<N,\mathcal{S},\mathcal{A},T,R,\gamma>$:

● 参与智能体集合 $N=\{1,\ldots,n\}$, 其中 n 为智能体的最大数量;

● 环境的状态空间 $\mathcal{S}$, 每一个阶段都会处于某个特定的状态 $s\in\mathcal{S}$;

● 智能体的动作空间 $\mathcal{A}=\{A_1,\ldots,A_n\}$, 其中 A_i 为智能体 i 的动作空间;

● 状态转移概率 $T:\mathcal{S}\times\mathcal{A}\times\mathcal{S}\to[0,1]$, 表示当前状态 $s\in\mathcal{S}$ 执行联合动作 $a\in\mathcal{A}$ 之后转移到下一个状态 $s'\in\mathcal{S}$ 的概率;

● 奖赏函数 $R_i:\mathcal{S}\times\mathcal{A}\times\mathcal{S}\to\mathbb{R}$, 其中 R_i 是智能体 i 的奖赏函数;

● 折扣因子 $\gamma\in[0,1)$, 用于计算累计回报.

因此, 每一个智能体都会学习策略 $\pi_i:\mathcal{S}\times\mathcal{A}\to[0,1]$, 使得在执行有限 T 步数或者无限步数后, 获得最大的折扣回报 $G_i=\mathbb{E}_{\pi_i}\left[\sum_{t=0}^{T}\gamma^t r_i^t\right]$ 或者 $G_i=\mathbb{E}_{\pi_i}\left[\sum_{t=0}^{\infty}\gamma^t r_i^t\right]$, 其中 r_i^t 表示智能体 i 在 t 时刻下获得的奖赏. 当然, 随机博弈和扩展式博弈有一定的相似性, 在某些条件下能够相互转换[57,58].

另外, 在随机博弈的有些情况中, 智能体并不能完全了解环境的状态信息, 只能接收到部分观测 (observation) 信息, 这些观测信息包含了有关环境的状态以及智能体动作的一些不完全信息, 我们称这种类型的博弈为部分可观测的随机博弈 (partially observable stochastic game, POSG)[59,60].

定义 5 (部分可观测的随机博弈 (POSG)). 一个部分可观测的随机博弈可以表示为以下八元组 $<N,\mathcal{S},O,\mathcal{A},T,R,Z,\gamma>$, 与定义 4 不同的是:

● 智能体的观测空间 $O=\{O_1,\ldots,O_n\}$, 其中 O_i 为智能体 i 的观测空间;

● 观测函数 $Z_i:\mathcal{A}\times\mathcal{S}\times O_i\to[0,1]$, 表示执行联合动作 $a\in\mathcal{A}$ 到状态 $s'\in\mathcal{S}$ 后智能体 i 观测到 $o_i\in O_i$ 的概率.

例如, 以自动驾驶的模拟环境 SMARTS[61]为例介绍随机博弈和部分可观测的随机博弈类型, 如图 5 所示. 首先假设智能体 i (红色驾驶车辆) 如果拥有全局视角信息, 包含路口形状、路口车辆、红绿灯信息等作为环境状态 s. 动作 a_i 为其输出的策略, 如图 5 中箭头所示的直行或者转向. 每一个参与智能体都会同时进行决策, 然后环境会给予奖赏反馈 r_i, 如果智能体顺利到达目标则奖赏为+1, 造成了车辆事故则为–1, 其他情况下则奖赏为 0, 环境随即转移到下一个状态 s', 并重复该过程. 另外, 如果每一个智能体无法获取全局视角信息, 而类似真实驾驶场景, 只能观测到自身周围区域的路况信息 o_i, 那么此时博弈类型为部分可观测的随机博弈 (https://github.com/huawei-noah/SMARTS).

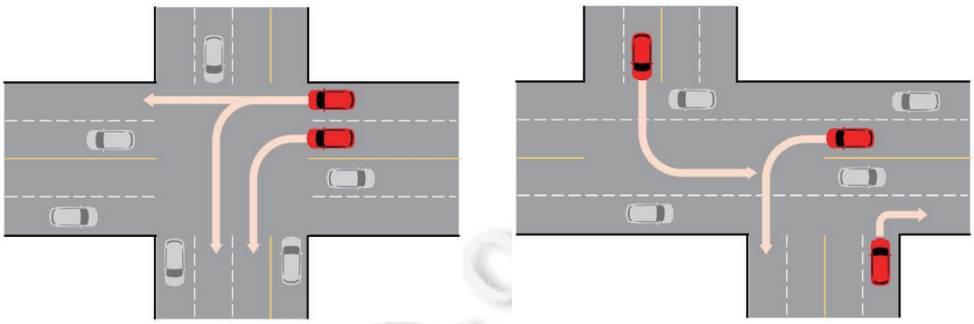


图 5 随机博弈和部分可观测的随机博弈示例: 自动驾驶

1.5 其他博弈类型

除了以上经典的博弈模型以外, 还有一些根据具体应用定义的一些博弈模型. 例如, 针对连续时间和连续动作空间的动态优化问题, 可以通过微分博弈 (differential game) 进行描述[62–64]. 其中, 参与者的决策通过微分方程刻画, 这些方程描述了参与者的长期收益函数随时间的变化. 每个参与者的目标是通过选择微分方程最优的控制变量最大化其个体收益.

在一些情况下为了能够快速求解博弈问题, 引入了势博弈 (potential game)[65]的概念, 假定存在一个全局的势函数 (potential function), 使得将所有参与博弈智能体的奖赏函数都能映射到该全局势函数中, 通过该函数就能够完全表征策略变化对每一个智能体收益的影响. 因此, 求解势函数的最大值就相当于求解出了所有智能体的均衡策略, 这大大降低了求解难度.

在智能体数量增多直到无穷大的情况下, 可以通过平均场博弈 (mean-field game)[66–68]进行表示, 其研究的是在连续时间和连续状态空间下的群体之间的相互作用和决策问题. 其中决策是通过微分方程或者动态系统的模型进行描述, 通过求解这些微分方程或者动态系统, 可以得到最优策略. 在平均场博弈中, 由于智能体的决策假设是受到群体的平均影响, 而不是直接受到其他个别智能体的影响, 从而可以很好地解决智能体数量急剧增加的问题.

另外, 博弈论与种群动态演化的过程相结合产生了一种新的动态博弈形式: 演化博弈 (evolutionary game)[69,70]. 演化博弈并不要求博弈智能体是完全理性的, 因此智能体可能对某个次优的解即可满意. 在方法上, 该博弈并不强调一个静态的均衡解, 而在于理解整个种群在时间维度上进化过程的动态均衡. 一般的演化博弈通常是基于选择 (selection) 和突变 (mutation) 进行动态演化, 选择是指利用当前回报较高的策略以吸引更多的个体加入种群, 而突变则是随机进行探索期望能够找到更好的策略. 另外一种相似的概念, 群体博弈 (population game)[71,72]则更关注群体空间维度上个体之间的相互作用, 个体都可以通过选择各自的策略来获得效用, 进而决定了整个群体在这个博弈中的效用. 群体博弈的重点是研究如何通过合作或竞争来达到群体的最优解.

以上介绍的矩阵博弈、扩展式博弈、贝叶斯博弈、随机博弈和微分博弈、势博弈、平均场博弈、演化博弈等其他博弈类型为目前大多数实际任务和相关求解方法建模的几种基础博弈类型, 其他可能存在的博弈类型不在本文综述范围内, 因此不再具体介绍和分析.

2 博弈模型构建

以上基础博弈类型是对现实场景的高度抽象, 难以直接应用于动态复杂的现实问题中. 其中, 通过工程技术建模是一种构筑数学博弈模型与现实场景的有效桥梁, 本节将重点介绍现实场景针对矩阵博弈、扩展式博弈和随机博弈等几种常见博弈模型的构建方法.

2.1 矩阵博弈建模

矩阵博弈适用于多个智能体同时做出独立决策并且互相影响的场景, 例如在市场价格竞争中[73], 每个公司可以选择高价或低价的动作, 通过现实约束可计算不同策略组合下两个公司各自的奖赏函数, 从而构建矩阵博弈; 在军事空战场景中[74], 一方智能体选择是否启动战斗机的动作, 另一方选择是否激活地对空导弹的动作, 通过数学及经验计算奖赏函数; 在节能减排与环境保护场景中[75], 双方可以选择继续排放或减少排放的动作, 通过经济发展和环境污染综合计算奖赏函数. 由此可见, 矩阵博弈的特点在于无状态转换的一次博弈, 拥有固定收益函数, 通常用于较为重要的宏观战略层决策.

2.2 扩展式博弈建模

扩展式博弈是一种用博弈树形式表示的动态博弈, 其中, 智能体的决策具有明显的先后顺序. 这类场景通常包括商业谈判、拍卖和招标等现实场景以及一些棋牌类游戏. 下面我们将具体介绍几种经典任务的扩展式博弈建模过程, 如表 2 所示.

表 2 工程化抽象的扩展式博弈模型特点

建模扩展式博弈场景	智能体数目		智能体动作类型		奖赏函数类型		
	=2	>2	连续	离散	合作	零和	混合
围棋—AlphaGo[6]	√	—	—	√	—	√	—
德州扑克—Libratus[9]	√	—	—	√	—	√	—
德州扑克—Pluribus[10]	—	√	—	√	—	√	—
斗地主—DouZero[76]	—	√	—	√	—	—	√
麻将—Suphx[11]	—	√	—	√	—	√	—

(1) 围棋—AlphaGo[6]. 围棋游戏是一种天然的扩展式博弈形式, 其难度在于庞大的动作和状态空间. AlphaGo 是于 2016 年针对双人围棋游戏开发的 AI 算法, 具体扩展式博弈建模方式如下.

- 参与智能体数目: 2 人;
- 参与智能体的决策顺序: 双方轮流;
- 动作空间: 棋盘上剩余可以落子的位置;
- 奖赏函数: 每一步奖赏为 0, 直到获胜奖赏+1, 失败奖赏–1, 为零和博弈.

(2) 德州扑克—Libratus[9]. Libratus 是于 2017 年针对无限制德州扑克游戏开发的 AI 算法, 于 2019 年扩展到 6 人版本 Pluribus[10]. 具体扩展式博弈建模方式如下.

- 参与智能体数目: 2 人 (Pluribus 存在 6 人);
- 参与智能体的决策顺序: 轮流;
- 动作空间: 跟注 (call)、加注 (bet) 和弃牌 (fold);
- 奖赏函数: 赢或输掉的金钱, 为零和博弈.

(3) 斗地主—DouZero[76]. DouZero 是于 2021 年针对 3 人斗地主游戏开发的 AI 算法. 为解决斗地主庞大的离散化决策空间, DouZero 通过网格编码方式将斗地主简化成双人零和博弈, 具体扩展式博弈建模方式如下.

- 参与智能体数目: 3 人 (1 名地主, 2 名农民);
- 参与智能体的决策顺序: 轮流;
- 动作空间: 由单牌、对子、顺子、三带一、飞机、火箭等牌型组成加上过牌共 27472 种, 可选动作遵循优

先级约束;

● 奖赏函数: 存在两种设计方式, 一种基于胜率奖赏, 即胜+1, 负–1, 另外一种基于平均分差, 为混合博弈.

(4) 麻将—Suphx[11]. Suphx 是于 2020 年针对日本麻将游戏开发的 AI 算法. 为解决多人非完美信息博弈, Suphx 通过对麻将网格编码工程化将 4 位玩家博弈问题简化成双人零和博弈 (自身和其他玩家), 然后从个体角度优化对手的最优反应, 具体扩展式博弈建模方式如下.

● 参与智能体数目: 4 人;

● 参与智能体的决策顺序: 执行动作后, 若无其他人响应则按顺序轮流;

● 动作空间: 每回合可选动作包括胡牌/吃/碰/杠;

● 奖赏函数: 直至游戏结束获得排名得分, 为零和博弈.

2.3 随机博弈建模

随机博弈是一种动态博弈, 具有多状态、时序决策的特点. 其复杂性在于综合考虑智能体间、智能体与环境的动态交互关系, 广泛存在股票投资和实时战略游戏等场景中. 下面我们将具体介绍几种经典场景的随机博弈建模过程, 如表 3 所示.

表 3 工程化抽象的随机博弈模型特点

建模随机博弈场景	智能体数目		智能体动作类型		奖赏函数类型		
	=2	>2	连续	离散	合作	零和	混合
星际争霸2—AlphaStar[12]	√	—	—	√	—	√	—
Dota 2—OpenAI Five[13]	—	√	—	√	—	—	√
雷神之锤III—FTW[77]	—	√	—	√	—	—	√
王者荣耀—绝悟[14,78]	—	√	—	√	—	—	√

(1) 星际争霸 II—AlphaStar[12]. AlphaStar 是于 2019 年针对实时战略游戏星际争霸 II 开发的 AI 算法, 玩家需要控制成百规模的单元进行对抗. 该场景的难度在于庞大的状态动作空间、以及由战争迷雾导致的不完全观测. 通过对观测、动作空间工程化处理将该问题处理成双人零和博弈问题, 具体随机博弈建模方式如下.

● 观测空间: 包含 512 维的单元实体属性、128×128 维的小地图特征、玩家状态信息以及游戏状态, 其中由战争迷雾隐藏的信息将不会出现在观测空间里;

● 动作空间: 包括可执行动作的实体单元、单元执行的动作种类、该动作执行的目标、是否将该动作放入执行序列或立马执行、是否重复执行以及执行动作指令的时机;

● 奖赏函数: 失败为−1, 平局为 0, 胜利+1, 为零和博弈.

(2) Dota 2—OpenAI Five[13]. OpenAI Five 是于 2019 年针对多人在线战斗竞技类游戏 Dota 2 开发的 AI 算法. 比赛在两个队伍间进行, 每个队伍有 5 名玩家, 双方以摧毁对方基地为获胜目标. 通过对各英雄工程化编码, 将多人混合博弈简化为队内合作博弈, 队间零和博弈, 具体随机博弈建模方式如下.

● 观测空间: 包括 189 个实体单元 (英雄、小兵、野怪、建筑物和信使等) 属性、8×8 周围地形信息、10×10 小地图信息以及前期采样的动作信息, 共计约 16 000 维输入;

● 动作空间: 包括释放技能、移动等. 将连续的动作统一化处理, 由技能释放顺序 (delay)、技能选中目标 (unit selection) 和范围偏移量构成 (offset). 其他动作由人工脚本控制, 例如升级技能与选择天赋、操控信使、购买装备等;

● 奖赏函数: 奖赏由个体奖赏 (金币获得、金币消耗、血量、英雄击杀数等) 和团队奖赏 (总胜负、塔的血量、信使死亡状态等) 加权组成, 为混合博弈.

(3) 雷神之锤 III—FTW[77]. FTW (for the win) 是于 2019 年针对第一人称多人射击竞技游戏开发的 AI 算法, 在夺旗模式中, 双方玩家以获得对方更多的旗帜为获胜目标. 尽管存在不同智能体数目的非对称博弈, 通过工程化编码将该问题简化为队内合作博弈, 队间零和博弈. 具体随机博弈建模方式如下.

● 观测空间: 该观测空间为 84×84×3 视觉图像输入;

● 动作空间: 包括改变偏航、改变俯仰角、左右扫射、前后移动、标记和跳跃, 共计 540 维动作;

● 奖赏函数: 由 13 个事件活动构成一个查找表, 为混合博弈.

(4) 王者荣耀—绝悟[14,78]. 绝悟是于 2020 年针对多人在线竞技游戏王者荣耀开发的 AI 算法. 与 Dota 2 类似, 比赛在两个队伍间进行, 每个队伍有 5 名玩家, 双方以摧毁对面的基地为获胜目标. 通过工程化编码将该问题简化为队内合作博弈, 队间零和博弈, 具体随机博弈建模方式如下.

● 观测空间: 由 4 类特征构成, 包括 8 559 维单元信息 (英雄、小兵等)、68 维游戏状态信息、560 维对手信息以及 6×17×17 空间信息;

● 动作空间: 由 3 类信息组成, 包括动作种类 (技能、回程等)、动作执行位置 (移动位置、技能释放位置)、动作释放目标单元;

● 奖赏函数: 根据 5 类相关信息构建稀疏或稠密奖赏, 包括与运营相关 (金币、经验等)、与击杀比相关 (击杀数、死亡数、助攻数等)、与伤害相关 (生命值等)、与推塔进度相关 (击毁塔楼等)、与胜负相关, 为混合博弈.

3 混合博弈的解概念和求解目标

博弈的解表示为所有智能体的一个联合策略 $\pi=\{\pi_1,\pi_2,\ldots,\pi_n\}$, 这个联合策略会综合考虑每个智能体的预期收益以及不同智能体收益的关系. 博弈的解往往通过智能体的效用值进行评估, 这里的效用值与不同的博弈类型有关. 具体地, 在矩阵博弈和扩展式博弈中智能体 i 的效用值 $U_i(\pi)$ 往往以奖赏函数 R_i 表示, 在随机博弈中往往以折扣回报 G_i 表示. 博弈论需要论证清楚解概念 (solution concept) 以及求解目标、解的存在性和唯一性问题、解的实际意义以及通过计算或者学习的方法能否保证联合策略一定收敛到对应的解等等. 因此, 系统性地理解混合博弈的解概念和求解目标是未来人工智能算法设计及其应用的基础.

传统的合作博弈, 因为所有智能体都共享一个全局的奖赏函数, 所以最直观的解就是最大化该奖赏函数的联合策略. 在零和博弈中, 智能体的收益会此消彼长, 所以直观的解就是所有智能体最大化个体收益最终达到某种均衡. 而在混合博弈中, 智能体可能会通过局部的合作和竞争获得更高的收益, 因此在该环境下会包含诸多的解概念和求解目标, 在不同的具体场景中有不同的作用. 下面, 本文将介绍混合博弈中一些经典的解概念和求解目标.

3.1 纳什均衡

纳什均衡 (Nash equilibrium)[79]是博弈论中最常用的一种解概念, 也能解决零和博弈问题. 纳什均衡是定义在最优反应 (best response) 概念的基础上. 根据以上联合策略的定义, 我们将除了智能体 i 的其他智能体的联合策略表示为 $\pi_{-i}=(\pi_1,\ldots,\pi_{i-1},\pi_{i+1},\ldots,\pi_n)$, 智能体 i 的效用值为 $U_i(\pi)$, 那么智能体 i 对其他智能体的最优反应策略 $BR_i(\pi_{-i})$ 定义为:

$$BR_i(\pi_{-i})=\arg\max_{\pi_i}U_i(\pi_i,\pi_{-i}) \tag{1}$$

纳什均衡策略 π^* 必须满足所有智能体的策略 π_i^* 都是其他智能体策略的最优反应 $\pi_i^*=BR_i(\pi_{-i}^*)$. 下面, 我们给出纳什均衡的定义.

定义 6 (纳什均衡). 博弈的纳什均衡策略 π^* 定义在乘积概率空间上, 且对任意的智能体 i 和其他策略 π_i', 都必须满足:

$$U_i(\pi_i^*,\pi_{-i}^*)\geqslant U_i(\pi_i',\pi_{-i}^*) \tag{2}$$

由此发现, 纳什均衡的意义在于任何一个智能体都不能通过背离纳什均衡调整自己的策略以获得更高的收益, 因此纳什均衡在混合博弈系统中是一个稳定解. 对于纳什均衡的存在性, 需要先介绍一下纯策略 (pure strategy) 和混合策略 (mixed strategy) 的概念. 纯策略是指参与博弈的智能体策略是在动作空间中一个确定性的动作, 而混合策略是动作空间上的一个概率分布. 任意的混合博弈都会存在至少一个纯策略或者混合策略的纳什均衡[49]. 下面, 我们通过 4 种特殊的对称 2×2 混合博弈原型[50]进行举例分析, 如图 6 所示.

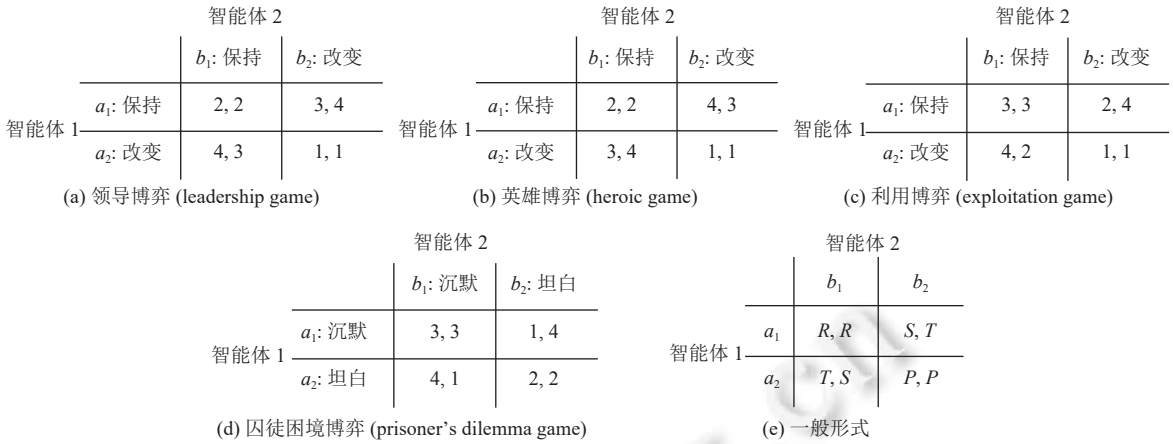


图 6　4 种特殊的对称 2×2 混合博弈原型

在图 6(e) 中可以发现, 以上 4 种类型均能表示为此一般形式, 而上述 4 种博弈中的奖赏函数 T, S, R, P 满足的条件如下所示.

- 领导博弈 (leadership game): $T > S > R > P$;
- 英雄博弈 (heroic game) : $S > T > R > P$;
- 利用博弈 (exploitation game): $T > R > S > P$;
- 囚徒困境博弈 (prisoner's dilemma game): $T > R > P > S$.

在领导博弈、英雄博弈和利用博弈中, (a_1 : 保持,b_1 : 保持) 都是自然结果, 即每一个智能体选择的极大极小值点, 但是并不是纳什均衡. 在领导博弈和英雄博弈中, 某一个智能体执行改变动作, 例如 (a_1 : 保持,b_2 : 改变) 和 (a_2 : 改变,b_1 : 保持) 都是纳什均衡, 而在领导博弈中率先执行改变动作将获得最高的收益 4, 因此称之为领导博弈; 反之, 在英雄博弈中率先执行改变动作, 对手将获得最高的收益 4, 因此称之为英雄博弈. 不难发现, 想要取得最大的收益结果需要两个智能体之间达成某种协议, 这与零和博弈中此消彼长的强对抗特点不同, 混合博弈向彼此传达意图或者协商是符合双方玩家的收益. 另外, 在利用博弈中, 虽然 (a_1 : 保持,b_2 : 改变) 和 (a_2 : 改变,b_1 : 保持) 同样都是纳什均衡, 但是选择某一个纳什均衡的时候会损害到另外一个智能体的收益. 而以上 3 种博弈中, 不存在所有智能体能够同时到达个体最大收益的联合策略, 而且对任意智能体来说都不存在占优策略 (dominant strategy). 其中占优策略的定义: 无论其他智能体采取什么策略, 占优策略都能得到比采取其他策略更好的结果. 最后, 在囚徒困境的博弈中, 可以发现 (a_2 : 坦白,b_2 : 坦白) 是纳什均衡, 但对比 (a_1 : 沉默,b_1 : 沉默) 策略所有智能体收益都是受损的, 所以换一种视角来看, 虽然纳什均衡能保证自己的策略是最优反应, 但是在全局系统上并不一定是某种"最优结果". 所以在这种情况下, 元博弈 (meta game)[80]理论提出能够在更高阶的博弈层次上解决该问题, 例如在囚徒困境中对智能体 2 进行一阶元博弈, 可以产生 4 种动作: 1) 无论智能体 1 选择什么, 智能体 2 都选择沉默; 2) 无论智能体 1 选择什么, 智能体 2 都选择坦白; 3) 和智能体 1 选择相同的动作; 4) 和智能体 1 选择相反的动作. 智能体 1 再根据智能体 2 的 4 种动作分别选择和沉默还是坦白, 又可以产生 16 种动作的二阶元博弈. 相关理论证明在 2×2 混合博弈中, 不超过二阶的元博弈可以解决这种困境问题. 而元博弈的这种概念需要了解或者推测到其他智能体的策略或者意图, 在一些重复博弈的场景中更容易刻画, 所以形成著名的以牙还牙 (tit-for-tat, TFT)[81]的策略.

最后根据第 1 节中介绍的扩展式博弈 (动态博弈)、贝叶斯博弈 (非完全信息博弈) 的例子, 分别可以采用子博弈完美纳什均衡 (subgame perfect Nash equilibrium)[82]、贝叶斯纳什均衡 (Bayesian Nash equilibrium)[83]的概念求解, 而在两者结合的非完全信息动态博弈情况下, 采用完美贝叶斯均衡 (perfect Bayesian equilibrium)[84]概念求解.

3.2　相关均衡

由上文的囚徒困境不难发现, 纳什均衡在全局系统上并不一定是某种"最优结果", 因此希望引入相关均衡

(correlated equilibrium)[85,86]的解概念最大化智能体的期望收益. 相关均衡是条件更弱和普遍的一种解概念, 而且纳什均衡是一种特殊的相关均衡. 例如在图 6(a) 的领导博弈中, 如果智能体 1 和智能体 2 都选择混合策略的纳什均衡, 不难发现有概率会产生一些非常差的 (a_2 : 改变,b_2 : 改变) 结果. 因此, 这里需要引入一种“自然信号”的概念, 例如给智能体 1 和 2 都约定好以 1/2 的概率选择 (a_1 : 保持,b_2 : 改变) 和以 1/2 的概率选择 (a_2 : 改变,b_1 : 保持), 那么每个智能体根据“自然信号”的推荐选择 1 号动作还是 2 号动作执行, 整体上的期望收益比纳什均衡大. 同样地, 在图 6(d) 的囚徒困境博弈中, 智能体 1 和 2 都约定好执行 (a_1 : 沉默,b_1 : 沉默) 动作, 整体收益高于纳什均衡 (a_2 : 坦白,b_2 : 坦白). 相比于纳什均衡选择的独立性, 相关均衡的相关性就体现在这种“自然信号”推荐选择上. 在数学上, 我们严谨地给出相关均衡定义.

定义 7 (相关均衡). 博弈的相关均衡策略是在联合动作空间 $\mathcal{A}$ 上的某种分布概率 $a \in \mathcal{D}$, 且对任意的智能体 i 和其他任意动作 a_i', 都必须满足:

$$\mathbb{E}_{a\in\mathcal{D}}[U_i(a)] \geqslant \mathbb{E}_{a\in\mathcal{D}}[U_i(a_i', a_{-i})|a_i] \tag{3}$$

由此定义可以发现, 在自然给予智能体 i 执行动作 a_i 的信号后, 其他智能体遵循该相关均衡策略的条件下, 智能体 i 选择动作 a_i 就是最优反应. 另外, 可以发现纳什均衡也是一种相关均衡, 即每个智能体都独立地进行动作选择, 这种自然信号不会提供任何其他智能体的相关信息. 更一般地, 如果去除这种提前告知动作 a_i 的选择, 智能体 i 无条件地听从自然安排, 并且也无法调整策略以获得更高的期望收益, 那么将会得到一种更加宽泛的粗糙相关均衡 (coarse correlated equilibrium)[87]概念.

定义 8 (粗糙相关均衡). 博弈的粗糙相关均衡策略是在联合动作空间 $\mathcal{A}$ 上的某种分布概率 $a \in \mathcal{D}$, 且对任意的智能体 i 和其他任意动作 a_i', 都必须满足:

$$\mathbb{E}_{a\in\mathcal{D}}[U_i(a)] \geqslant \mathbb{E}_{a\in\mathcal{D}}[U_i(a_i', a_{-i})] \tag{4}$$

从定义上看, 根据严格程度可以发现纯策略纳什均衡、混合策略纳什均衡、相关均衡、粗糙相关均衡是不断减弱的, 而且求解纳什均衡是 PPAD 完全 (complete for polynomial parity arguments on directed graphs)[88,89]的复杂度, 因此对于多人混合博弈场景求解较弱的相关均衡或者粗糙相关均衡更有实际应用价值.

3.3 帕累托最优

帕累托最优 (Pareto optimality) 最早在 19 世纪初提出, 目的是解决在经济效率与分配问题上传统纳什均衡效率低下的问题. 例如, 在图 6(e) 的囚徒困境博弈问题中, 两个智能体都选择 (a_1 : 沉默,b_1 : 沉默) 动作的收益比都选择 (a_2 : 坦白,b_2 : 坦白) 的收益更高, 但不是纳什均衡. 基于此, 帕累托最优是指无法通过改变任意动作或策略使得可以增加至少一个智能体的收益而不减少任何其他智能体的收益, 定义为如下数学形式.

定义 9 (帕累托最优). 一个策略 π 如果不被其他任何策略帕累托占优, 那么策略 π 就是帕累托最优. 其中, 帕累托占优的定义是: 如果策略 π' 被策略 π 帕累托占优, 那么必须满足以下条件:

$$\forall \text{智能体} i: U_i(\pi) \geqslant U_i(\pi') \text{ 且} \exists \text{智能体} i: U_i(\pi) > U_i(\pi') \tag{5}$$

直观地说, 如果没有其他联合策略可以提高至少一个智能体的收益, 而不降低任何其他智能体的收益, 那么联合策略就是帕累托最优. 所有帕累托最优的集合被称为帕累托前沿 (Pareto frontier). 在零和博弈中, 所有的策略都是帕累托最优, 探究帕累托最优策略是没有意义的; 而在合作博弈问题中, 帕累托最优就是全局最优的策略, 因为所有智能体的收益都是相同的. 而在混合博弈中, 帕累托最优可以用来完善均衡解的概念空间, 因为如果某个策略是纳什均衡, 但不是帕累托最优, 它仍然是存在改进空间. 但是需要注意的是, 帕累托最优和纳什均衡是完全不同的概念, 帕累托最优本身可能不是一个能被所有智能体完全接受的均衡解. 因此如何在不同的混合博弈场景中进行应用, 还需要指定具体的求解目标.

3.4 个体收益、集体收益以及兼顾公平的求解目标

以上混合博弈的解概念沿用了传统博弈的解概念, 对智能体有绝对理性的假设且主要与个体收益相关, 其中帕累托最优还会综合考虑系统的效率, 是否存在改进的空间. 但是在实际的混合博弈问题中, 智能体不一定是绝对

理性且系统中可能存在宏观调控者, 因此上述解概念不足以解决混合博弈问题. 基于此, 本文引入最大化个体收益、集体收益以及兼顾公平的 3 种求解目标, 以丰富和扩展传统博弈的解概念. 其中, 这几类求解目标与实际问题紧密结合, 并不涉及解的均衡性和存在性分析, 与整个系统和博弈结构的机制设计相关.

首先, 个体收益与利己主义 (egoism)[90–92]紧密关联, 利己主义作为一种哲学概念, 仅关注自我或自我的作用, 作为个人行动的动机和目标, 也就是说只有智能体自身决定准则的行为模式. 利己主义智能体对描述出于自身收益行事或规定其他智能体应该怎样做感兴趣. 个体收益最大化的求解目标往往在不同智能体之间是会互相影响和冲突的, 根据智能体 i 自己独立的策略 π_i, 将个体收益最大化策略定义为:

$$\pi_i = \arg\max U_i(\pi_i, \pi_{-i}), \forall \pi_{-i} \in \pi \tag{6}$$

不难发现, 如果其他智能体的策略 π_{-i} 是固定的, 那么个体收益最大化策略 π_i 就是其他智能体联合策略 π_{-i} 的最优反应, 如果所有智能体都最大化个体收益, 那么最终策略大概率会收敛到纳什均衡. 在大多数混合博弈的情况下, 过分强调个体收益的目标会造成智能体行为策略的短视, 当面对其他智能体同样利己和贪婪时将无法接受, 最后会造成整个系统不必要的损失. 例如在公共资源分配问题上, 所有智能体都共享某个草场或者河流发展畜牧业或渔业, 如果所有智能体都大肆放牧或打捞, 将会引发巨大的生态危机, 最后导致个体收益的损失反而得不偿失. 因此, 在混合博弈场景中, 仅关注个体收益的求解目标不足以解决问题.

其次, 集体收益与功利主义 (utilitarianism)[93,94]紧密相关, 功利主义作为一系列规范的哲学伦理, 主要考虑所有智能体的集体幸福或福祉. 虽然不同种类的功利主义有不同的特征, 但它们的基本思想都是在某种意义上最大化集体的效用. 因此, 与个体收益不同, 集体收益是建立在整个联合策略 π 上, 将集体收益最大化策略定义为:

$$\pi = \arg\max \sum_{i \in N} U_i(\pi) \tag{7}$$

不难发现, 集体收益最大化策略一定是满足帕累托最优, 但是反过来帕累托最优的策略不一定是集体收益最大化的策略. 因此, 在最大化集体收益的同时难免会发生损害某些特定智能体收益的情况. 在大多数混合博弈中, 过分强调集体收益的目标会造成部分智能体个体收益的损失, 最终导致这种隐式合作行为的崩溃. 仍然举例上述公共资源分配的问题, 如果一味强调草场或者河流上整体的畜牧业和渔业收益, 固定量的限制放牧和打捞的数量, 那么距离资源点近和初始拥有庞大基础产业的智能体将会在博弈中占据优势, 而那些较为弱势的智能体由于分配上的歧视将会罔顾这种系统准则, 从而导致整个系统崩溃. 因此, 在混合博弈场景中, 仅关注集体收益的求解目标也是不足以解决问题的.

最后, 为了在混合场景中保持系统的稳定, 需要引入兼顾公平的求解目标. 平等主义 (egalitarianism)[95–97]建立在社会平等的概念之上, 会考虑所有人的收益, 认为所有智能体在基本价值或社会地位上都是平等的. 因此, 绝对的公平收益是建立在联合策略 π 上的, 将公平收益最大化策略定义为:

$$\pi = \arg\max \sum_{i \in N} -\left(U_i(\pi) - \bar{U}(\pi)\right)^2 / N \tag{8}$$

其中, $\bar{U}(\pi)$ 表示所有智能体的平均收益. 由此可见, 绝对的公平收益是最小化智能体收益的方差, 但是, 严格意义上的绝对公平是没有意义的. 例如在上述的畜牧业和渔业例子中, 绝对的公平会使得智能体最终的策略趋于懒惰 (lazy), 所有的智能体都不会定期去放牧和打捞导致方差最小为 0, 这样会导致系统中资源的浪费. 所以在一些场景中, 更多的是考虑兼顾公平的目标, 例如会引入一些社会福利函数 (social welfare function)[98–101]综合考虑集体收益和平等收益, 因此有如下兼顾公平的求解目标:

$$\pi = \arg\max \prod_{i \in N} U_i(\pi) \tag{9}$$

$$\pi = \arg\max \sum_{i \in N} \omega_i U_i^{\uparrow}(\pi) \tag{10}$$

其中, $\omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$ 表示对不同智能体效用值的关注权重, 其中 $\omega_1 \geqslant \omega_2 \geqslant \ldots \geqslant \omega_N$, $\sum_{i \in N} \omega_i = 1$ 且 $U_i^{\uparrow}$ 表示对所有智能体的效用值从小到大的排序. 如果 $\omega_1 = \omega_2 = \ldots = \omega_N$, 则退化为集体收益最大化的求解目标.

4 混合博弈的求解方法

根据以上定义的解概念和求解目标, 本文将目前解决混合博弈问题的主流方法分为经典博弈论、强化学习和两者结合的 3 类方法. 首先, 为了方便对下列方法进行梳理和总结, 在方法前介绍几种基础的数学形式化模型, 其中矩阵博弈和马尔可夫决策过程 (Markov decision process, MDP) 分别是博弈论和强化学习的基础模型, 其他模型都是在其基础上演变和发展, 而定义 5 中部分可观测的随机博弈是最通用的博弈模型, 可以用来形式化建模所有的博弈论和强化学习方法. 下面, 我们将具体介绍相关方法, 并附带介绍相关形式化模型.

4.1 经典博弈论方法

经典的博弈论方法主要基于矩阵博弈和扩展式博弈模型, 本文将经典博弈论方法分成数学规划、虚拟博弈、双重预言和遗憾最小化 4 类, 每类方法的优缺点和适用范围将会在具体章节中介绍, 其中涉及与强化学习结合的方法将在第 4.3 节中介绍.

4.1.1 数学规划

数学规划 (mathematical programming) 是解决博弈问题最简单直接的方法, 在双人零和博弈问题中, 可以简化为线性规划, 但是在更一般的情况下, 往往只能通过非线性规划的方式求解. 其中, 对于定义 4 中的多人随机博弈问题 $\mathcal{G}=\langle N,\mathcal{S},\mathcal{A},T,R,\gamma\rangle$, 可以通过非线性规划的方式求解[102,103]:

$$\min_{V,\pi} f(V,\pi)=\sum_{i\in N}\sum_{s\in\mathcal{S}}\left[V_i(s)-\left(r_i(s)+\gamma T\left(s,\boldsymbol{a}\in\pi(s),s'\right)V_i(s')\right)\right] \tag{11}$$

$$\text{满足}:(1)\ \mathbb{E}_{\pi_{-i}(s)}\left[r_i(s)+\gamma\sum_{s'\in\mathcal{S}}T\left(s,\boldsymbol{a}\in\pi(s),s'\right)V_i(s')\right]\leqslant V_i(s),\quad \forall i\in N,s\in\mathcal{S} \tag{12}$$

$$(2)\ \pi_i(s,a_i)\geqslant 0,\quad \forall i\in N,s\in\mathcal{S},a_i\in A_i \tag{13}$$

$$(3)\ \sum\nolimits_{a_i\in A_i}\pi_i(s,a_i)=1,\quad \forall i\in N,s\in\mathcal{S},a_i\in A_i \tag{14}$$

其中, $V=\{V_i\}_{i\in N}$ 是所有智能体的状态值函数集合, $\pi(s)=\{\pi_i(s)\}_{i\in N}$ 是所有智能体在状态 s 上的动作分布概率, $r_i(s)$ 表示智能体 i 在状态 s 时获得的奖赏, $T(s,\boldsymbol{a}\in\pi(s),s')$ 表示环境状态转移函数. 这是一个典型的非线性规划问题, 其中优化变量是整体值函数 V 和策略 π. 目标函数旨在最小化给定策略在所有状态下的时序差分 (temporal difference) 误差, 类似于策略迭代法中的策略评估步骤. 约束 (1) 类似策略改进步骤, 约束 (2) 和 (3) 确保策略定义的合理性和完备性. 虽然, 通过数学规划的方法可以显示地计算随机博弈问题的纳什均衡, 但是在大规模智能体数量和状态动作空间的情况下很难求解, 因此下面将介绍其他基于迭代的方法进行求解.

4.1.2 虚拟博弈

虚拟博弈 (fictitious play, FP) 是博弈论中最传统的求解方法之一, 通常应用在基本的矩阵博弈问题[104]. 虚拟博弈的核心思想是通过其他智能体历史动作信息去估计其他智能体的策略分布, 并以此为基准选择最优反应. 因此, 根据公式 (1) 定义的最优反应 $BR_i(\pi_{-i})$, 智能体 i 在 $t+1$ 时刻虚拟博弈的最优策略 π^{t+1} 可以表示为如下形式:

$$\pi_i^{t+1}=\left(1-\frac{1}{t}\right)\pi_i^t+\frac{1}{t}a_i^* \tag{15}$$

$$a_i^*\in BR_i\left(\pi_{-i}^t=\frac{1}{t}\sum_{\tau=0}^{t-1}I\left\{a_{-i}^\tau=a,a\in A_{-i}\right\}\right) \tag{16}$$

其中, I 为指示函数, 如果 x 为真, 那么 $I(x)=1$, 否则 $I(x)=0$. 虚拟博弈中的初始策略通常假设为随机策略, 即 $\pi_i^0(a_i)=1/|A_i|,\forall a_i\in A_i$. 由于以上虚拟博弈中最优反应的定义是某个最优动作, 所以在混合策略纳什均衡的博弈问题中并不能保证收敛. 但是, 随着迭代次数的增加, 虚拟博弈中的时间平均或经验分布 (empirical distribution) 策略在零和博弈、势博弈和双人矩阵博弈问题中一定能够收敛到纳什均衡[105].

除此之外, 虚拟博弈在混合策略上的扩展是将上述最优反应的动作选择变成某种混合策略分布选择, 称作平滑虚拟博弈 (smooth FP) 或随机虚拟博弈 (stochastic FP)[106,107]. 因此, 智能体 i 在 $t+1$ 时刻虚拟博弈的最优反应策

略可以表示为如下形式:

$$BR_i(\pi_{-i}^t)=\frac{\exp\left(\frac{1}{\lambda}r_i(a_{i(\cdot)},\pi_{-i}^t)\right)}{\sum_{k=1}^{|A_i|}\exp\left(\frac{1}{\lambda}r_i(a_{i(k)},\pi_{-i}^t)\right)},\ \forall a_{i(k)}\in A_i \tag{17}$$

其中, 分子中的 $a_{i(\cdot)}$ 项表示最优反应策略 $BR_i(\pi_{-i}^t)$ 选择该动作的概率, λ 表示平滑参数. 相关工作证明平滑虚拟博弈在满足势博弈的条件下, 能够收敛到纳什均衡的邻域 (neighborhood of the Nash equilibrium)[108,109].

接下来, 一种广义弱化虚拟博弈 (generalized weakened FP, GWFP)[110]将传统虚拟博弈方法进行扩展, 其一是将最优反应弱化为近似最优反应, 其二是每轮更新的学习率不再是固定的 $1/t$, 而是允许存在扰动和变化. 因此, 广义弱化虚拟博弈的策略更新方式可以表示为:

$$\pi_i^{t+1}=\left(1-\alpha^{t+1}\right)\pi_i^t+\alpha^{t+1}\left(BR_i^{\epsilon_{t+1}}(\pi_{-i}^t)+M_i^{t+1}\right) \tag{18}$$

其中, $BR_i^{\epsilon_{t+1}}(\pi_{-i}^t)$ 表示 ϵ 近似纳什均衡 $BR_i^{\epsilon}(\pi_{-i})=\{\pi_i\mid r_i(\pi_i,\pi_{-i})\geqslant r_i(BR_i(\pi_{-i}),\pi_{-i})-\epsilon\}$ 且 $\lim_{t\to\infty}\epsilon^t=0$, 学习率 $\lim_{t\to\infty}\alpha^t=0$ 且 $\sum_{t=1}^{\infty}\alpha^t=\infty$ 以及对任意的 $T>0$, 满足:

$$\limsup_{t\to\infty}\left\{\left\|\sum_{\tau=t}^{k-1}\alpha^{\tau+1}M^{\tau+1}\right\|\ \text{s.t.}\ \sum_{\tau=t}^{k-1}\alpha^{\tau+1}\leqslant T\right\}=0 \tag{19}$$

由于这种广义弱化条件, 每次迭代不需要精确地求出最优反应, 因此可以与其他值迭代或策略迭代优化方法结合, 例如强化学习中的 Q 学习和行为者批评家 (actor-critic) 等方法. 同样地, 在双人矩阵博弈中, 广义弱化虚拟博弈的联合平均策略能够收敛到纳什均衡.

另外, 全宽扩展式博弈 (full-width extensive-form fictitious play, XFP)[111]算法在广义弱化虚拟博弈的基础上将 FP 在普通矩阵博弈推广到扩展式博弈问题中. 具体而言, 矩阵博弈的混合策略的加权组合可以等价实现 (realization equivalent) 行为策略: 例如, 对于 π 和 β 两个行为策略, Π 和 B 两个混合策略能够通过参数 λ_1 和参数 λ_2 等价实现 π 和 β, 其中 $\lambda_1,\lambda_2\geqslant 0$ 且 $\lambda_1+\lambda_2=1$, 那么对于任意信息集中的状态 $s\in I$, $\mu(s)$ 通过公式 (20) 可以表示混合策略 $M=\lambda_1\Pi+\lambda_2 B$ 的等价实现:

$$\mu(s)=\pi(s)+\frac{\lambda_2 x_\beta(\sigma_s)}{\lambda_1 x_\pi(\sigma_s)+\lambda_2 x_\beta(\sigma_s)}(\beta(s)-\pi(s)) \tag{20}$$

其中, σ_s 表示到达 s 的历史序列, $x_\beta(\sigma_s)$ 表示在策略 β 下实现 σ_s 序列的概率. 那么在扩展式博弈中, 智能体 i 在状态 $s\in I$ 上的策略 $\pi_i^{t+1}(s)$ 更新方式为:

$$\beta_i^{t+1}\in BR_i^{\epsilon_{t+1}}(\pi_{-i}^t) \tag{21}$$

$$\pi_i^{t+1}(s)=\pi_i^t(s)+\frac{\alpha^{t+1}x_{\beta_i^{t+1}}(\sigma_s)(\beta_i^{t+1}(s)-\pi_i^t(s))}{(1-\alpha^{t+1})x_{\pi_i^t}(\sigma_s)+\alpha^{t+1}x_{\beta_i^{t+1}}(\sigma_s)} \tag{22}$$

其中, 满足广义弱化虚拟博弈的参数条件 $M^t=0$, $\lim_{t\to\infty}\epsilon^t=0$, 学习率 $\lim_{t\to\infty}\alpha^t=0$ 且 $\sum_{t=1}^{\infty}\alpha^t=\infty$. 同样地, XFP 证明在扩展式博弈中, 平均策略能够收敛到纳什均衡. 最后, 通过引入自博弈的方法, 提出了新的虚拟自博弈 (fictitious self-play, FSP)[111]方法, 通过有限数量的样本在计算近似最优反应时使用强化学习方法, 计算平均策略更新时使用监督学习的方式, 从而解决了 XFP 在大规模博弈问题中需要对所有状态进行迭代计算的缺陷与困难.

4.1.3 双重预言

双重预言 (double oracle, DO) 最早用来解决双人零和矩阵博弈问题, 并且能够收敛到纳什均衡[112]. 具体来说, 假设智能体 1 的纯策略集合为 $\Pi=\{a^1,a^2,\ldots,a^n\}$, 智能体 2 的纯策略集合为 $C=\{c^1,c^2,\ldots,c^m\}$, 其中 n 和 m 分别表示两个智能体的纯策略数量. DO 方法通过迭代创建和求解一系列子博弈 (即具有有限纯策略集的博弈) 来近似大型零和博弈中的纳什均衡. 在任意时刻 t 上, DO 学习器都会创建和解决一个子博弈 G_t, 因为子博弈 $G_t=(\Pi_t,C_t)$ 中的纯策略数量远小于原始博弈问题, 因此能够更简单地求解. 求出该子博弈问题中的纳什均衡 (π_t^*,c_t^*) 后, 智能体 1

和 2 再根据对方的纳什均衡策略求解最优反应 (a_{t+1},c_{t+1}) 加入策略集合 Π_t,C_t, 得到 $\Pi_{t+1}=\Pi_t\cup a_{t+1},C_{t+1}=C_t\cup c_{t+1}$, 然后循环求解子博弈 $G_{t+1}=(\Pi_{t+1},C_{t+1})$, 直至 $\Pi_{t+1}=\Pi_t$ 和 $C_{t+1}=C_t$, 其中 Π_0 和 C_0 往往是随机策略. 需要注意的是, DO 方法优于数学规划方法的前提是子博弈 G_t 中的纯策略数量远小于原始博弈, 因此需要满足以下假设: 假设原始博弈 $A_{n\times m}$ 的混合纳什均衡策略为 (π^*,c^*), 那么该均衡策略的支撑 (*support*) 规模必须小于博弈的规模, 如公式 (23) 所示:

$$\max(|support(\pi^*)|,|support(c^*)|)<\min(n,m) \tag{23}$$

其中, 某混合策略 π 的支撑为 $support(\pi):=\left\{a_i\in\Pi|\pi=\sum_{i=1}^{n}\omega_i a_i,\omega_i\neq 0\right\}$, 规模 $|support(\pi)|$ 表示该策略下的纯策略组合系数不为 0 的数量. 此外, 文献 [113] 将 DO 算法推广至双人零和连续博弈问题, 证明了 DO 算法能够收敛到纳什均衡, 并且保证在有限多步中实现收敛, 在一些经典的博弈示例中的实验表现优于虚拟博弈方法. 对于双人零和博弈问题中纯策略数量过大的问题, 在线双重预言 (online double oracle, ODO)[114]方法提出将在线学习中的无遗憾 (no-regret) 分析与 DO 结合, 在每回合都用无遗憾的性质利用对手策略, 并且证明在时间 T 的维度下能够实现 $O\left(\sqrt{Tk\log(k)}\right)$ 的遗憾界, 其中 k 是有效策略 (effective strategy) 集的规模.

DO 算法虽然能在双人零和博弈中保证最终收敛到纳什均衡, 并且在一些网络安全博弈中也有广泛的应用[115,116], 但在最坏的情况下 DO 算法需要遍历整个策略空间而且无法处理复杂的扩展式博弈. 因此, 序列形式的双重预言 (sequence-form double oracle)[117]方法通过只允许智能体选定可用的动作序列来限制博弈规模, 在求解受限博弈后, 能快速找到当前解的最优反应, 从而添加新的博弈序列中. 在非完全信息的双人零和扩展式博弈中, 该方法能够保证收敛到精确的纳什均衡. 其次, 一些基于元博弈 (meta game) 概念的方法提出在策略空间上进行纳什均衡的应对和求解, 能够在扩展式博弈问题中应用, 例如策略空间的应对预言 (policy space response oracle, PSRO)[118]以及一些变体方法等. 由于该类方法引入了强化学习等技术, 我们将在第 4.3 节中具体介绍. 此外, 尽管 PSRO 类的方法能够保证收敛到近似纳什均衡, 并且可以处理连续动作的博弈问题, 但随着状态数量的增加, 它可能需要指数级的迭代次数. 因此, 提出了一种适用于双人零和博弈的扩展式双重预言 (extensive-form double oracle, XDO)[119], 该算法保证在信息状态的数量上线性收敛到近似纳什均衡. 与 PSRO 在博弈的根节点处计算混合最优反应不同, XDO 在每个信息状态下都会计算混合最优反应. 其扩展形式, 通过深度强化学习学习计算最优反应的 NXDO 版本也将在第 4.3 节中具体介绍.

4.1.4 遗憾最小化

在线凸优化 (online convex optimization, OCO)[120]中的遗憾最小化 (regret minimization) 算法是解决扩展式博弈问题的一类经典方法, 其中最具代表性的方法是在线镜像梯度 (online mirror descent, OMD)[121]和跟随正则化的领导 (follow the regularized leader, FTRL)[122,123]两类. 在双人零和扩展式博弈问题中, 假设 $\mathcal{X}$ 和 $\mathcal{Y}$ 分别是智能体 1 和 2 的策略空间, 那么最终的优化函数可以表示为求解鞍点问题 $\min\limits_{\boldsymbol{x}\in\mathcal{X}}\max\limits_{\boldsymbol{y}\in\mathcal{Y}}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{y}$, 其中 $\boldsymbol{A}$ 为智能体 1 的损失矩阵. 令 $\boldsymbol{l}=\boldsymbol{A}\boldsymbol{y}$, 那么智能体 1 的预期损失可以表示为 $\langle\boldsymbol{l},\boldsymbol{x}\rangle$, FTRL 的策略 $\boldsymbol{x}$ 的更新方式可以表示为:

$$\boldsymbol{x}^{t+1}=\arg\min_{\boldsymbol{x}\in\mathcal{X}}\left\{\boldsymbol{L}^t,\boldsymbol{x}+q^{0:t}(\boldsymbol{x})\right\} \tag{24}$$

其中, $\boldsymbol{L}^t=\sum_{k=1}^{t}\boldsymbol{l}^k$ 表示时间累计损失, $q^{0:t}(\boldsymbol{x})=\sum_{k=0}^{t}q^k(\boldsymbol{x})$ 表示正则化项, 且 $q^{0:t}(\boldsymbol{x})$ 在 $\mathcal{X}$ 空间上是可微和强凸函数. 类似地, OMD 的更新方式可以表示为:

$$\boldsymbol{x}^{t+1}=\arg\min_{\boldsymbol{x}\in\mathcal{X}}\{\boldsymbol{l}^t,\boldsymbol{x}+q^t(\boldsymbol{x})+\mathcal{B}_{q^{0:t}}(\boldsymbol{x}\|\boldsymbol{x}^t)\} \tag{25}$$

其中, $\mathcal{B}_{q^{0:t}}$ 表示布雷格曼散度 (Bregman divergence), 形式上 $\mathcal{B}_{q^{0:t}}(\boldsymbol{x}\|\boldsymbol{x}^t)=q^{0:t}(\boldsymbol{x})-q^{0:t}(\boldsymbol{x}^t)-\langle\nabla q^{0:t}(\boldsymbol{x}^t),\boldsymbol{x}-\boldsymbol{x}^t\rangle$. $q^{0:t}$ 函数也被称为距离生成函数 (distance generating function, DGF), 其中常见的一种类型称为扩大的距离生成函数 (dilated DGF)[124–126]形式为 $q^{0:t}(\boldsymbol{x})=\sum_{s\in D}\boldsymbol{x}(\sigma_s)\psi_s^t(\boldsymbol{x}(s)/\boldsymbol{x}(\sigma_s))$, 其中 $s\in D$ 表示智能体的决策节点而 σ_s 表示到达 s 节点的序列, ψ_s^t 同样是可微和强凸函数.

虽然 FTRL 和 OMD 方法有很好的理论保证, 但是在一些大规模的扩展式博弈问题中的实际表现不是很好, 而且无法应用在多人博弈场景中. 所以, 一类反事实遗憾 (counterfactual regret, CFR) 最小化方法提出在大规模的扩展式博弈问题中可以求解近似纳什均衡, 并且表现出来更快的收敛速度[127]. 首先, 定义每个智能体的反事实值

(counterfactual value), 假设智能体 i 当前的信息状态 $s \in I$, σ_s 表示当前到达 s 的序列, $T \in Z$ 表示终止状态, π 表示所有智能体的联合策略, $\mu^{\pi}(\sigma_s \to \sigma_T)$ 表示在策略 π 下从 σ_s 到达 σ_T 的概率, $r_i(T)$ 表示到达终止 T 状态的奖赏函数, 那么智能体 i 的反事实值定义为:

$$v_i(\pi, s) = \sum_{T \in Z} \mu^{\pi_{-i}}(\sigma_s)\mu^{\pi}(\sigma_s \to \sigma_T) r_i(T) \tag{26}$$

因此 CFR 方法在 t 时刻下没有选择某动作 a 的遗憾定义为:

$$regret_i^t(s,a) = v_i(\pi_t \mid s \to a, s) - v_i(\pi, s) \tag{27}$$

根据累计遗憾可以更新下个时刻的策略 π_{t+1}, 其中一类著名的方法是遗憾匹配 (regret matching, RM)[128,129], 其策略更新方式为:

$$Regret_i^T(s,a) = \sum_{t=1}^{T} regret_i^t(s,a) \tag{28}$$

$$\pi_i^{t+1}(s,a) = \begin{cases} \dfrac{Regret_i^{T,+}(s,a)}{\sum_{b \in A_i(s)} Regret_i^{T,+}(s,b)}, & \text{如果} \sum_{b \in A_i(s)} Regret_i^{T,+}(s,b) > 0 \\ \dfrac{1}{|A_i(s)|}, & \text{否则} \end{cases} \tag{29}$$

其中, + 值操作表示取大于 0 的部分: $[x]^+ = \max(0, x)$. 在遗憾匹配的基础上有一种遗憾匹配+(regret matching+, RM+)[130]方法, 有着更强的收敛性能, 其策略更新方式为:

$$Regret_i^T(s,a) = \sum_{t=1}^{T} regret_i^{t,+}(s,a) \tag{30}$$

$$\pi_i^{t+1}(s,a) = \begin{cases} \dfrac{Regret_i^{T}(s,a)}{\sum_{b \in A_i(s)} Regret_i^{T}(s,b)}, & \text{如果} \sum_{b \in A_i(s)} Regret_i^{T}(s,b) > 0 \\ \dfrac{1}{|A_i(s)|}, & \text{否则} \end{cases} \tag{31}$$

其中, 与 RM 相比, 唯一的区别是在计算每一个时刻下的遗憾时, 只记录大于 0 的部分, 因此有着更强的收敛性能. 受到 RM+算法的启发, 折扣加权 CFR (discounted CFR, DCFR)[131]方法对遗憾值和平均策略更新时的权重进行了研究, 其提出 3 个参数 α, β, γ 分别对应正遗憾值, 负遗憾值和策略平均权重. 具体来讲, 在第 t 轮更新时, 将正累积遗憾值先乘以权重 $t^{\alpha}/t^{\alpha}+1$, 负遗憾值乘以权重 $t^{\beta}/t^{\beta}+1$, 平均策略乘以权重 $(t/t+1)^{\gamma}$. 在此加权范式下, 使用 RM+的 CFR+算法 (平均策略计算时第 t 轮权重为 t) 可以看作 $\alpha=\infty$, $\beta=-\infty$, $\gamma=2$. 实验表明, 在大多数的情况下, 选取 $\alpha=\dfrac{3}{2}$, $\beta=0$, $\gamma=2$ 能取得最好的效果.

目前也有相关工作证明, 如果在 FTRL 和 OMD 在线凸优化方法中将正则化项选择为二范数, 那么策略更新的方式分别等价于 RM 和 RM+[132,133]. 除了遗憾匹配以外还有一种最小化遗憾的 Hedge[134]方法, 其中智能体 i 更新策略的方式为:

$$\pi^{t+1}(a) = \frac{\pi^t(a)\mathrm{e}^{-\eta l^t(a)}}{\sum_{a_i \in |A_i|} \pi^t(a_i)\mathrm{e}^{-\eta l^t(a_i)}}, \ \pi_0(\cdot) = \frac{1}{|A_i|} \tag{32}$$

其中, η 为温度系数, l^t 为 t 时刻的损失函数. 之后, 还有各种在 Hedge 方法下扩展的乐观 Hedge (optimistic hedge)[135–137]方法能够实现更快的收敛速度.

由于标准 CFR 方法在计算遗憾值时需要遍历整棵博弈树, 因此在大规模博弈中, 标准 CFR 方法的每轮迭代需要庞大的内存空间和计算时间, 使其难以直接应用. 例如在双人德州扑克中, 计算状态 s 的反事实值至多需要遍历 41 951 448 000 个可能的发牌情况及各动作所产生的后续状态, 这在实际使用中是难以计算的. 为了解决这一问题, CFR 方法通常与采样方法共同使用, 典型的方法如机会采样 CFR (chance-sampled CFR)[127]、蒙特卡洛 CFR

(Monte-Carlo CFR, MCCFR)[138,139]等. 这一类方法在每一轮迭代中不完全遍历整个博弈树, 并使用反事实值的无偏估计代替原本的反事实值更新累积遗憾值, 从而降低计算需求. 通过根据博弈树特点选择适当的采样方式, CFR算法可以快速有效地求解出近似纳什均衡策略. 但由于加入采样机制, MCCFR 等算法在计算遗憾值时引入了较大的方差, 影响算法实际收敛效果, 为此相关工作提出一种基于方差消减的 MCCFR (MCCFR with variance reduction, VR-MCCFR)[140]方法, 提升算法收敛速度. 同样由于 MCCFR 方法的方差特性, 结合 RM+的 MCCFR+方法在实际使用中效果表现不佳[141], 可能原因在于 RM+作为加速技巧对于方差较为敏感, 从而两者难以简单结合. 但也有相关工作表明在使用深度网络拟合遗憾值时, 引入小批次 (mini-batch) 等训练技巧, MCCFR+在一些任务上也可以取得不错的表现[142].

尽管基于采样的 MCCFR 算法有效地降低了每一轮迭代过程中的存储和计算开销, 如何在大规模博弈中采用值表的形式存储策略及遗憾值仍然是限制均衡求解的一个重要挑战. 传统的方法采用状态动作空间约简的形式[9,143], 通过合并相似的状态和动作, 降低博弈策略求解和存储空间. 但空间约简难免会导致所求策略与均衡之间存在一定差距, 且随着约简粒度的变化而变化. 为此, 结合深度网络的 CFR 方法成为近期的研究主流[8,142,144,145]. 双神经网络 CFR (double neural CFR, DNCFR)[142]与 Deep CFR[144]方法采用两个神经网络分别拟合平均遗憾值和平均策略的方式, 直接模拟 CFR 的计算流程. DeepStack[8]则仅训练一个深度反事实值网络 (deep counterfactual value network), 结合子博弈求解范式[143], 在博弈过程中实时计算实际策略. 以上基于 CFR 的方法在多人博弈场景中均能收敛到粗糙相关均衡, 并且在两人零和博弈问题中等价于纳什均衡.

以上所有经典博弈论方法的简单总结与对比如表 4 所示.

表 4 经典博弈论方法的总结与对比

学习范式	算法	博弈形式化	大于双人	深度神经网络	解概念/求解目标	博弈场景		
						合作	零和	混合
虚拟博弈	FP[104]	Matrix game	—	—	纳什均衡/个体收益	—	√	—
	Stochastic FP[106,107]	Matrix game	—	—	纳什均衡/个体收益	—	√	—
	GWFP[110]	Matrix game	—	—	纳什均衡/个体收益	—	√	—
	XFP/FSP[111]	EFG	—	—	纳什均衡/个体收益	—	√	—
双重预言	DO[112]	Matrix game	—	—	纳什均衡/个体收益	—	√	—
	ODO[114]	Matrix game	—	—	纳什均衡/个体收益	—	√	—
	Sequence-form DO[117]	EFG	—	—	纳什均衡/个体收益	—	√	—
	PSRO[118]	Matrix game	—	—	纳什均衡/个体收益	—	√	—
	XDO[119]	EFG	—	—	纳什均衡/个体收益	—	√	—
遗憾最小化	OMD[121]	EFG	—	—	纳什均衡/个体收益	—	√	—
	FTRL[122,123]	EFG	—	—	纳什均衡/个体收益	—	√	—
	CFR[127]/CFR+[130]	EFG	√	—	粗糙相关均衡/个体收益	√	√	√
	DCFR[131]	EFG	√	—	粗糙相关均衡/个体收益	√	√	√
	MCCFR[139]	EFG	√	—	粗糙相关均衡/个体收益	√	√	√
	VR-MCCFR[140]	EFG	√	—	粗糙相关均衡/个体收益	√	√	√
	Deep CFR[144]	EFG	√	√	粗糙相关均衡/个体收益	√	√	√
	DNCFR[142]	EFG	√	√	粗糙相关均衡/个体收益	√	√	√

4.2 强化学习方法

强化学习方法主要是在不断试错的过程中寻找到最优策略以最大化目标奖赏函数, 该过程通过马尔可夫决策过程 (Markov decision process, MDP) 形式化建模, 下面我们将给出 MDP 的定义.

定义 10 (马尔可夫决策过程 (MDP)). 一个马尔可夫决策过程可以表示为五元组 $\langle \mathcal{S}, \mathcal{A}, T, R, \gamma \rangle$ [146]:

- 环境的状态空间 $\mathcal{S}$, 每一个时刻都会处于某个特定的状态 $s \in \mathcal{S}$;

● 智能体的动作空间 $\mathcal{A}$;

● 状态转移概率 $T:\mathcal{S}\times\mathcal{A}\times\mathcal{S}\to[0,1]$, 表示当前状态 $s\in\mathcal{S}$ 执行动作 $a\in\mathcal{A}$ 之后转移到下一个状态 $s'\in\mathcal{S}$ 的概率;

● 奖赏函数 $R:\mathcal{S}\times\mathcal{A}\times\mathcal{S}\to\mathbb{R}$, 执行动作 $a\in\mathcal{A}$, 环境从状态 $s\in\mathcal{S}$ 转移到状态 $s'\in\mathcal{S}$ 之后, 智能体获得的奖赏 r;

● 折扣因子 $\gamma\in[0,1)$, 用于计算累计回报.

单智能体的强化学习目标就是寻找到最优的策略 $\pi:\mathcal{S}\times\mathcal{A}\to[0,1]$, 以最大化累计回报 $G=\mathbb{E}_{\pi}\left[\sum_{t=0}^{T}\gamma^t r^t\right]$. 强化学习算法可以分为基于值 (value-based) 的强化学习方法或者基于策略 (policy-based) 的强化学习方法, 并且在与深度学习技术结合后, 产生了著名的深度 Q 网络 (deep Q network, DQN)[147]和近端策略优化 (proximal policy optimization, PPO)[148]方法. 针对单智能体强化学习的 MDP 模型引入多个智能体后, 可以简单地扩展为多智能体马尔可夫决策过程 (multi-agent Markov decision process, MMDP)[149].

定义 11 (多智能体马尔可夫决策过程 (MMDP)). 一个多智能体马尔可夫决策过程可以表示为六元组 $\langle N,\mathcal{S},\mathcal{A},T,R,\gamma\rangle$, 与定义 4 随机博弈不同的是:

● 奖赏函数 $R:\mathcal{S}\times\mathcal{A}\times\mathcal{S}\to\mathbb{R}$, 其中所有智能体共享相同的奖赏函数.

由此可见, MMDP 仅是在 MDP 的基础上进行了智能体数量和动作空间的扩展, 而所有智能体会共享相同的奖赏函数, 因此只能建模合作博弈问题. 与此类似的还有去集中式马尔可夫决策过程 (decentralized Markov decision process, Dec-MDP)[150]以及去集中式部分可观测的马尔可夫决策过程 (decentralized partially observable Markov decision process, Dec-POMDP)[151], 它们主要引入的是智能体独立观测 (observation) 的概念, 具体的内容如定义 12 所示.

定义 12 (去集中式马尔可夫决策过程 (Dec-MDP) 和去集中式部分可观测的马尔可夫决策过程 (Dec-POMDP)). 一个 Dec-MDP 或 Dec-POMDP 可以表示为八元组 $\langle N,\mathcal{S},O,\mathcal{A},T,R,Z,\gamma\rangle$, 与定义 11 MMDP 不同的是:

● 智能体的观测空间 $O=\{O_1,\ldots,O_n\}$, 其中 O_i 为智能体 i 的观测空间;

● 观测函数 $Z_i:\mathcal{A}\times\mathcal{S}\times O_i\to[0,1]$, 表示执行联合动作 $a\in\mathcal{A}$ 到状态 $s'\in\mathcal{S}$ 后智能体 i 观测到 $o_i\in O_i$ 的概率.

● 如果所有智能体的观测之和能够恢复出环境的状态, 即 $P(\mathcal{S}\mid O)=1$, 那么为完全可观测, 则为 Dec-MDP; 否则是部分可观测, 为 Dec-POMDP.

不难发现, 以上定义的 MMDP、Dec-MDP 和 Dec-POMDP 都只能建模合作博弈类型, 而在 MMDP 的基础上通过加入局部网络化连接的建模方式, 可以用来建模竞争、合作和混合博弈多种类型, 具体内容如下.

定义 13 (网络化多智能体马尔可夫决策过程 (networked MMDP)). 一个网络化多智能体马尔可夫决策过程可以表示为以下七元组 $<N,\mathcal{S},\mathcal{A},T,R,\{\mathcal{G}^t\}_{t\geqslant 0},\gamma>$, 与定义 4 随机博弈不同的是:

● $\mathcal{G}^t=(N,\mathcal{E}^t)$ 表示 N 个智能体之间在 t 时刻的网络连接图, 其中 $\mathcal{E}^t$ 表示智能体的连接信息, 例如每一个顶点表示一个智能体, 如果智能体 i 和智能体 j 在 t 时刻能够交流, 那么边 $(i,j)\in\mathcal{E}^t$.

Networked MMDP 既可以用来建模合作博弈问题, 例如所有智能体都根据团队平均的奖赏 $\bar{r}(s,a,s')=\frac{1}{N}\sum_{i\in N}r_i(s,a,s')$ 最大化累计回报; 或者仅优化个人奖赏函数 $r_i(s,a,s')$ 来表征零和博弈或混合博弈问题. 而通过局部网络化, 可以引入协商的机制在混合博弈中达到某种相关均衡或者优化目标. 以上所有强化学习模型与前文介绍的博弈论模型之间的关系和对比如图 7 所示.

基于以上强化学习模型, 下面将综述目前针对混合博弈问题的一些强化学习方法, 本节将以下方法根据学习训练范式分为: 集中式学习 (centralized learning, CL)[152]方法、独立式学习 (decentralized learning/independent learning, DL/IL)[153]方法、集中式训练分布式执行 (centralized training with decentralized execution, CTDE)[150,151,154]方法和带有网络化连接的独立式学习 (decentralized learning with networked agents)[155–157]方法. 每类方法的优缺点和适用范围将会在具体章节中介绍.

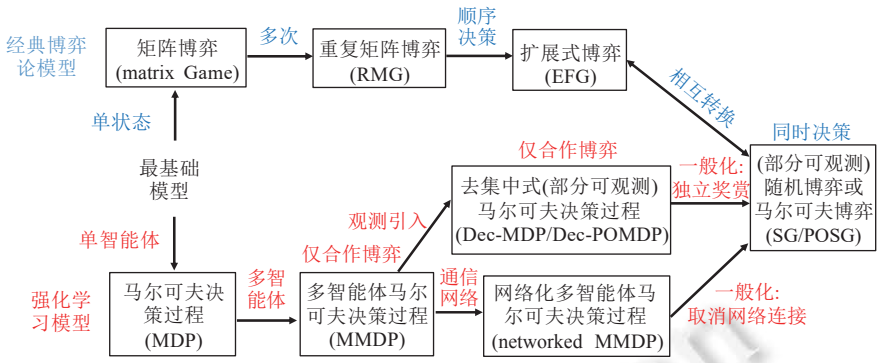


图 7 经典博弈论和强化学习形式化模型的关系与对比

4.2.1 集中式学习

集中式学习方法是将所有的智能体看作一个全局智能体, 该全局智能体在环境状态空间 $\mathcal{S}$ 和联合动作空间 $\mathcal{A}=\{A_1,\ldots,A_n\}$ 中选择联合动作 $\boldsymbol{a}\in\mathcal{A}$, 训练一个唯一的集中式策略 π, 以最大化平均奖赏或者某种均衡奖赏. 因此, 在混合博弈中, 集中式学习的方法不能保证策略最终收敛到纳什均衡, 如果使用的是平均奖赏, 那么最终的收敛解往往是帕累托最优; 如果是某种均衡或者加权形式的奖赏函数, 那么目标是最大化集体收益和兼顾公平等. 使用集中式 Q 学习方法, 其 Q 值函数更新方式如下所示:

$$Q(s,\boldsymbol{a})=Q(s,\boldsymbol{a})+\alpha\left(r+\gamma\max_{\boldsymbol{a}'\in\mathcal{A}}Q(s',\boldsymbol{a}')-Q(s,\boldsymbol{a})\right) \tag{33}$$

其中, r 为某种平均奖赏或者均衡奖赏. 因此, 最优的联合策略 $\pi(\boldsymbol{a}|s)=\text{argmax}_{\boldsymbol{a}\in\mathcal{A}}Q(s,\boldsymbol{a})$. 由此发现, 集中式学习将所有智能体视为一个全局智能体的方式, 将多智能体问题简化成了一个单智能体问题, 规避了多智能体的环境非平稳性 (non-stationarity) 和信度分配 (credit assignment) 问题. 然而, 这种方法在实践中有许多局限性, 例如随着智能体数量的增多, 联合状态动作空间会指数增长, 集中式学习的方式会随之崩溃. 目前的集中式学习方法主要有 JAL[152]、MDP-learner[158]和 MAT[159]等, 应用于合作博弈任务中, 并且最近的 MAT 工作将这种传统的合作博弈任务视为一个序贯决策的任务.

4.2.2 独立式学习

独立式学习的方法是让每一个智能体都使用单智能体的方法进行学习, 而将其他智能体视为环境的一部分. 这样, 每一个智能体 i 都通过自己局部的观测、动作和奖赏函数, 训练出一个独立的策略 π_i, 以最大化智能体的个体收益. 例如使用独立 Q 学习的方法, 每个智能体 i 的值函数 Q_i 的更新方式如下所示:

$$Q_i(o_i,a_i)=Q(o_i,a_i)+\alpha\left(r_i+\gamma\max_{a_i'\in A_i}Q_i(o_i',a_i')-Q_i(o_i,a_i)\right) \tag{34}$$

其中, r_i 为智能体 i 的独立奖赏. 因此, 每个智能体 i 的最优策略 $\pi_i(a_i|o_i)=\arg\max_{a_i\in A_i}Q(o_i,a_i)$. 此外, 如果令所有智能体的独立奖赏函数都相同 $r_1=r_2=\ldots=r_N$ 或者进行内部激励塑造 $\hat{r}_i=r_i+r_i^{\text{in}}$, 可以分别达到优化集体收益或者其他平衡收益的目标. 由此发现, 独立式学习方法虽然可以很好地解决多智能体带来的可扩展性问题, 但是由于在训练过程中受到其他智能体的策略变化而引起环境非平稳性的问题, 很难在理论上证明最终收敛解能够达到纳什均衡或相关均衡. 虽然独立式学习方法在理论上有一定的缺陷, 但是在一些场景中却能表现出较好的性能.

在第 1 类方法中, 每个智能体均使用单智能体的方法进行学习: IQL-ε [160]通过使用理想化趋近于 0 的学习率, 可以在双人双动作的零和、合作和混合博弈中取得较好的效果; 其次, 在独立深度 Q 学习 (independent DQN, IDQN)[161]方法中, 两个智能体能通过改变独立奖赏函数的方式能很好地进行乒乓球游戏的竞争或合作; 在更复杂的星际争霸多智能体挑战 (StarCraft multi-agent challenge, SMAC)[20]中, 独立近端策略优化 (independent PPO, IPPO)[162]方法也使得多智能体之间形成高效的协作策略. 这类方法可以使智能体拥有相同或者不同的奖赏函数, 因此可以在零和、合作和混合场景中均能直接应用. 除此之外, 最近的一个理想独立 Q 学习 (ideal independent Q-

learning, I2Q)[163]方法针对 MMDP 建模的完全合作型任务, 在仅使用当前状态 s 和下一个时刻状态 s' 的信息而没有全局联合动作信息的条件下使用 QSS 学习[164], 能够以完全独立式的学习方式等价于全局联合最优 Q 学习方法, 最后也通过理论和实验证明了收敛性和算法性能.

第 2 类方法, 在使用单智能体强化学习方法的同时, 引入了智能体建模的额外模块, 来改善由于其他智能体策略变化导致的环境非平稳性问题. 其中, 深度强化对手网络 (deep reinforcement opponent network, DRON)[165]通过采用两个神经网络: 一个评估自己智能体的 Q 值, 另一个学习其他智能体策略的表示. 此外, DRON 还提出了多个专家网络组合预测结果, 以获取其他智能体更加精确的策略. 深度循环策略推断 Q 网络 (deep recurrent policy inference Q-network, DRPIQN)[166]不同于 DRON 手工设计的特征表示, 直接从原始观测中学习策略特征, 并通过优化推断的对手策略和对手的真实策略之间的交叉熵损失来实现. 自我/他人建模 (self other-modeling, SOM)[167]区别于上述方法, 仅通过智能体自身的策略预测其他智能体的目标, 能够适用在各种合作、零和和混合博弈任务中. 除此之外, 一些递归推理模型在多智能体强化学习的引入可以有效考虑智能体的非完全理性, 从而进行更好的策略优化[168–170]. 其中概率递归推理 (probabilistic recursive reasoning, PR2)[168]方法将联合策略分解成自身智能体策略和其他智能体策略:

$$\pi_\theta(a_i,a_{-i}\mid s)=\pi_{\theta_i}(a_i\mid s)\rho_{\theta_{-i}}(a_{-i}\mid s,a_i) \tag{35}$$

$$\rho_{\theta_{-i}}(a_{-i}\mid s,a_i)\propto\exp(Q_i(s,a_i,a_{-i})-Q_i(s,a_i)) \tag{36}$$

类似地, 最大熵目标正则化对手模型 (regularized opponent model with maximum entropy objective, ROMMEO)[169]方法引入最大化策略熵这一目标进行优化, 将联合策略分解成以下两项:

$$\pi_\theta(a_i,a_{-i}\mid s)=\pi_{\theta_i}(a_i\mid s,a_{-i})\rho_{\theta_{-i}}(a_{-i}\mid s) \tag{37}$$

$$\rho_{\theta_{-i}}(a_{-i}\mid s,a_i)\propto\frac{\left(\sum_{\alpha_i}\exp(Q^*_{\text{soft}}(s,a_i,a_{-i})/\alpha)\right)^{\alpha}}{\exp(V^*(s))} \tag{38}$$

最后, 广义的递归推理 (generalized recursive reasoning, GR2)[170]方法将上述方法推广至 k 阶模型, 并对其他智能体进行有限理性的建模, 最终 k 阶策略可以表示为如下形式:

$$\pi_i^k\left(a_i^k\mid s\right)\propto\int_{a_{-i}^{k-1}}\left\{\pi_i^k\left(a_i^k\mid s,a_{-i}^{k-1}\right)\cdot\int_{a_i^{k-2}}\left[\rho_{-i}^{k-1}\left(a_{-i}^{k-1}\mid s,a_i^{k-2}\right)\pi_i^{k-2}\left(a_i^{k-2}\mid s\right)\right]\mathrm{d}a_i^{k-2}\right\}\mathrm{d}a_{-i}^{k-1} \tag{39}$$

其中, $\pi_i^{k-2}(a_i^{k-2}|s)$ 仍然可以按照上式继续递归推理.

第 3 类方法通过在智能体独立奖赏上进行内部激励塑造, 改变最大化个体收益的目标, 可以在混合博弈场景中实现集体收益或者亲社会合作行为的涌现. 例如, 在一些公共资源分配问题[24,171]中, 容易出现序贯社会困境的问题[51,52]: 如果每个智能体都短视地最大化个体收益, 那么最终可能并不能达到全局的“最优结果”, 例如囚徒困境. 因此一些相关工作, 开始通过多智能体强化学习的方法研究合作和竞争行为是如何随着环境资源变化而涌现的, 并揭示了现实世界中的序贯社会困境是如何影响合作[172]. 文献 [173] 指出同伴选择的方式可以促进在最大化自身收益情况下智能体之间的合作行为, 并且逐渐学习到一种报复背叛者的以牙还牙策略, 从而促进策略形成亲社会性质. 不公平厌恶 (inequity aversion, IA)[174]通过对智能体间奖赏不公平厌恶的方法设计内部激励函数 U_i, 使得智能体的奖赏既不能太低于也不能太高于其他智能体的奖赏:

$$U_i(r_1,\ldots,r_i,\ldots,r_n)=r_i-\frac{\alpha_i}{N-1}\sum_{j\neq i}\max(r_j-r_i,0)-\frac{\beta_i}{N-1}\sum_{j\neq i}\max(r_i-r_j,0) \tag{40}$$

其中, α_i 和 β_i 为影响系数. 社会影响 (social influence)[175]方法通过反事实推理来评估因果影响, 智能体 i 在 t 时刻的激励项 c_i^t 表示激励那些对其他智能体策略影响更大的动作, 并证明这种方式实际上激励动作之间具有高互信息的智能体:

$$c_i^t=\sum_{j=0,j\neq i}^{N}\left[D_{KL}\left[p\left(a_j^t\mid a_i^t,s_j^t\right)\sum_{\tilde{a}_i^t}p\left(a_j^t\mid\tilde{a}_i^t,s_j^t\right)p\left(\tilde{a}_i^t\mid s_j^t\right)\right]\right]=\sum_{j=0,j\neq i}^{N}\left[D_{KL}\left[p\left(a_j^t\mid a_i^t,s_j^t\right)\sum_{\tilde{a}_i^t}p\left(a_j^t\mid s_j^t\right)\right]\right] \tag{41}$$

社会影响方法能够在混合博弈场景中显著提高智能体的协调和沟通能力, 实现一些合作行为的涌现. 另外, 随

机不确定性社会偏好 (randomized uncertain social preferences, RUSP)[176]方法通过引入噪声随机化奖赏变换矩阵, 从而重塑单个智能体的独立奖赏, 并在实验中证明通过这种奖赏重塑而不是用单一的个人奖赏或是集体奖赏进行强化学习的训练有助于智能体形成互惠关系和小联盟. 除此之外, 学习互惠 (learning reciprocity)[177]方法将智能体分成创新者 (innovator) 和模仿者 (imitator) 两种, 创新者通过最大化个体收益进行学习, 模仿者通过模仿行为差异的内部激励和离策略修正 (off-policy correction) 机制进行学习, 最后通过互惠行为的涌现促进合作; 学习激励他人 (learning to incentivize others, LIO)[178]方法在智能体 i 看到其他智能体 j 的动作后, 赋予其额外的内部激励 $r_j^{\eta_i}(o_i^t,a_{-i}^t)$ 来影响对方策略从而促进合作行为的产生; 社会价值取向 (social value orientation, SVO)[179]方法通过用观测的奖赏倾向和目标 SVO 之间的差距作为内部激励促进合作行为的产生:

$$U_i(s,o_i,a_i)=r_i-\omega\cdot\left|\theta^{\mathrm{SVO}}-\theta\left(\vec{R}\right)\right|,\theta\left(\vec{R}\right)=\arctan\left(\frac{\bar{r}_{-i}}{r_i}\right) \tag{42}$$

其中, ω 为参数, $\bar{r}_{-i}$ 为其他智能体的平均奖赏. 综上所述, 使用内部激励塑造的方法能够在混合博弈场景中促进合作和亲社会行为的产生.

但是, 在混合博弈场景中如果过度关注集体收益而造成了个体之间收益不平等的情况, 会对新兴的集体行为产生影响, 从而降低群体成功合作的机会[180]. 因此, 第 4 类方法主要是讨论多智能体强化学习中平等主义策略的重要性和训练方式[181,182]. 首先, 严格意义上的平等主义很少被采用, 因为它会造成集体收益上巨大的损失. 嫉妒自由 (envy freeness) 证明了无嫉妒分配并不总是存在的, 因此确定嫉妒度最小的帕累托最优才是有意义的. 而在多智能体强化学习的领域中, 一种集中式学习方法 RMF (regularized maximin fairness)[183]考虑正则化的最大化公平策略, 该策略综合衡量平等主义 (最差智能体的奖赏) 和功利主义 (平均奖赏):

$$V(\pi)=\min_{i\in N}\psi(i,\pi)+\frac{\epsilon}{N}\sum\nolimits_{i\in N}\psi(i,\pi),\psi(i,\pi)=\mathbb{E}\left[\sum\nolimits_{t=0}^{\infty}\gamma^t r_i\mid\pi\right] \tag{43}$$

其中, ϵ 为权衡参数, γ 为折扣因子. 然而, 简单地最大化奖赏最低的智能体是不够的. 不公平厌恶 (inequity aversion, IA)[174]旨在获得相对公平的个体奖赏, 以促进合作和制裁背叛. 此外, 公平高效网络 (fair-efficient network, FEN)[184]提出在分层框架下学习公平策略, 并通过通信的方法与相邻智能体交换收益信息, 其中每个智能体的公平高效奖赏 $\tilde{r}_i^t$ 为:

$$\tilde{r}_i^t=\frac{\frac{\bar{u}^t}{c}}{\epsilon+\left|\frac{u_i^t}{\bar{u}^t}-1\right|},u_i^t=\frac{\sum_{j=0}^{t}r_i^t}{t},\bar{u}^t=\frac{\sum_{i=1}^{N}u_i^t}{N} \tag{44}$$

最后, 最近的一种个人导向/团队导向 (self-oriented team-oriented, SOTO)[185]方法通过构造公平的社会福利函数 U 对所有智能体的奖赏进行公平考虑 $\phi(r)=\sum_{i\in N}U(r_i)$, 然后通过独立式学习的方法求解兼顾公平的策略.

尽管独立式学习方法相对简单, 但仍然是多智能体强化学习研究的重要基准. 事实上, 它们通常可以在合作场景中产生与最先进的集中式学习或集中式训练分布式执行算法相匹配的结果, 并且可以非常便捷地应用到零和博弈和混合博弈场景[162,186].

4.2.3 集中式训练分布式执行

集中式训练分布式执行是结合了上述集中式和独立式学习方法的特点, 在集中式训练的过程中能够收集到所有智能体的观测信息和动作, 学习一个联合 Q 值函数, 从而解决多智能体场景中的环境非平稳性的问题; 其次, 在分布式执行过程中, 又能够独立输出策略, 解决了多智能体可扩展性的问题. 因此, 这类方法已成为多智能体强化学习研究的主流方向, 特别是针对合作博弈, 集中式训练分布式执行的框架更为契合.

第 1 类方法是著名的值分解 (value decomposition, VD) 方法, 例如, VDN[187], QMIX[188], W-QMIX[189], QTRAN[190]和 QPLEX[191]等, 这类方法主要通过 Dec-MDP 和 Dec-POMDP 建模, 研究如何将联合的 Q 值函数分解成每个智能体独立的 Q 值函数, 并且满足个体全局最优 (individual-global-max, IGM) 条件, 所以这类方法只能适用于所有智能体都共享一个全局奖赏函数的合作博弈问题中, 本文将不详细介绍.

第 2 类方法主要是将策略梯度的强化学习方法推广到集中式训练分布式执行的多智能体场景中, 例如反事实多智能体 (counterfactual multi-agent, COMA)[192]方法提出了一种新的多智能体行为者批评家 (Actor-Critic) 方法. 具体地, COMA 使用全局值函数与反事实基线的差异进行策略梯度更新, 其中该基线是将单个智能体的动作边缘化, 同时保持其他智能体的动作不变. 但是, 由于全局值函数是通过全局奖赏进行计算, 因此只适用于合作博弈任务. 多智能体极化策略梯度 (multi-agent polarization policy gradient, MAPPG)[193]在此基础上使用简单高效的极化函数, 实现多智能体策略梯度中的信用分配, 并且能保证策略最终收敛到帕累托最优. 其次, 多智能体深度确定性策略梯度 (multi-agent deep deterministic policy gradient, MADDPG)[194]方法和多智能体近端策略优化 (multi-agent proximal policy optimization, MAPPO)[195]方法分别将单智能体中的 DDPG[196]和 PPO[148]方法拓展成为多智能体方法, 在集中式训练过程中, 通过独立奖赏函数对每一个智能体都训练一个联合动作的值函数, 然后通过策略梯度的更新方式训练出分布式执行的策略函数. 由于每个智能体都有独立的奖赏函数, 因此可以在各种零和、合作和混合博弈场景中均能适用.

第 3 类是一些基于通信的方法, 在集中式训练的过程中允许智能体之间进行一些通信的方式获取其他智能体的相关信息, 以缓解环境非平稳性的问题, 因此目前的一些通信方法也主要是针对合作博弈任务[197−199]. 例如, 通信神经网络 (communication neural net, CommNet)[200]使用连续矢量信道接收其他智能体的信息传输汇总, 并且在每个时间步上进行多个通信周期, 由于针对合作博弈并且信息汇总通过加和平均的方式, 所以所有智能体都共享同一个策略网络, 并且在运行时允许智能体数量的动态变化, 解决了多智能体可扩展性的问题. 另外, 个体控制的连续通信模型 (individualized controlled continuous communication model, IC3Net)[201]和多智能体双向协同网络 (multiagent bidirectionally-coordinated network, BiCNet)[202]方法能够优化每个智能体的独立奖赏, 因此可以应用在混合博弈场景中, 方法细节上不同的是 IC3Net 通过带有门控机制的连续通信信道, 能很好地解决何时通信的问题. 而 BiCNet 通过双向循环神经网络在隐空间进行隐式通信, 可以有效保障智能体的私密信息.

4.2.4 带有网络化连接的独立式学习

最后在实际场景的一些博弈问题中, 智能体之间可能是非同质的, 并且有不同的行为倾向和奖赏函数, 智能体之间也希望保持自己信息的私密性, 因此集中式训练的机制不再适用. 在此基础上, 一类带有网络化连接的独立式学习方法逐渐开始在网络路由、智能交通和机器人控制等领域应用. 该类方法通过定义 13 中的 Networked MMDP 进行建模, 因为每个智能体都有独立的奖赏函数, 所以可以应用在零和博弈和混合博弈任务中. 首先, QD-learning[155]方法利用共识和创新形式的混合时间尺度随机动力学的分析技术, 使得智能体通过本地和稀疏通信网络上的相互信息交换实现合作博弈任务, 并假设每个智能体只知道其本地在线数据且智能体之间的通信网络弱连通的情况下, 证明了该方法可以收敛到最优值函数和最优固定控制策略. 其次, 带有网络连接智能体的行为者批评家 (Actor-Critic with networked agents)[157,203]方法又将局部网络连接实现近邻通信的思想扩展到行为者批评家 (Actor-Critic) 方法中, 能够在实际控制系统中处理连续状态和动作空间的问题, 并采用新提出的期望策略梯度来减小梯度估计的方差, 最后给出了线性函数逼近时算法的收敛性保证, 并通过实验仿真验证了理论结果. 不同于以上方法主要针对合作任务, Decentralized FQI[204]方法将局部网络连接与拟合 Q 迭代 (fitted Q-iteration, FQI)[205]方法结合, 能够处理多智能体场景中的非合作博弈问题, 并从理论上量化估计的动作值函数的有限样本误差, 这对于有限样本领域内的多智能体强化学习算法的严格理论具有重大意义.

以上所有强化学习方法的简单总结与对比如表 5 所示.

表 5 强化学习方法的总结与对比

学习范式	算法	博弈形式化	基于值/策略	深度神经网络	解概念/求解目标	博弈场景		
						合作	零和	混合
集中式学习	JAL[152]	Matrix game	值	—	帕累托最优/集体收益	√	—	—
	MDP-learner[158]	MMDP	值	—	帕累托最优/集体收益	√	—	—
	MAT[159]	Dec-POMDP	策略	√	帕累托最优/集体收益	√	—	—

表 5 强化学习方法的总结与对比 (续)

学习范式	算法	博弈形式化	基于值/策略	深度神经网络	解概念/求解目标	博弈场景		
						合作	零和	混合
独立式学习	IQL-ε[160]	Matrix game	值	—	个体/集体收益	√	√	√
	IDQN[161]	SG	值	√	个体/集体收益	√	√	√
	IPPO[162]	POSG/Dec-POMDP	策略	√	个体/集体收益	√	√	√
	I2Q[163]	MMDP	值	√	帕累托最优/集体收益	√	—	—
	DRON[165]	SG	值	√	个体/集体收益	√	√	√
	DPIRQN[166]	SG	值	√	个体/集体收益	√	√	√
	SOM[167]	SG	值	√	个体/集体收益	√	√	√
	PR2[168]	SG	值/策略	√	个体/集体收益	√	√	√
	ROMMEO[169]	SG	值/策略	√	个体/集体收益	√	√	√
	GR2[170]	SG	值/策略	√	个体/集体收益	√	√	√
	IA[174]	POSG	值/策略	√	兼顾公平	—	—	√
	Social influence[175]	POSG	值/策略	√	个体+集体收益	√	—	√
	RUSP[176]	POSG	策略	√	个体+集体收益	√	—	√
	Learning reciprocity[177]	POSG	策略	√	个体+集体收益	√	—	√
	LIO[178]	Matrix game/POSG	策略	√	个体+集体收益	√	—	√
	SVO[179]	POSG	策略	√	个体+集体收益	√	—	√
	FEN[184]	POSG	策略	√	兼顾公平	—	—	√
	SOTO[185]	POSG	策略	√	兼顾公平	—	—	√
集中式训练分布式执行	VD[187–191]	Dec-POMDP	值	√	集体收益	√	—	—
	COMA[192]	Dec-POMDP	策略	√	集体收益	√	—	—
	MAPPG[193]	Dec-POMDP	策略	√	帕累托最优/集体收益	√	—	—
	MADDPG[194]	POSG	策略	√	个体/集体收益	√	√	√
	MAPPO[195]	POSG	策略	√	个体/集体收益	√	√	√
	CommNet[200]	SG	策略	√	帕累托最优/集体收益	√	—	—
	IC3Net[201]	POSG	策略	√	个体/集体收益	√	√	√
	BiCNet[202]	SG	策略	√	个体/集体收益	√	√	√
带有网络化连接的独立式学习	QD-learning[155]	Networked MMDP	值	—	帕累托最优/集体收益	√	—	—
	Actor-Critic with networked agents[157,203]	Networked MMDP	策略	—	帕累托最优/集体收益	√	—	—
	Decentralized FQI[204]	Networked MMDP	值	—	个体/集体收益	√	√	√

4.3 结合方法

本节介绍博弈论与强化学习的结合方法, 主要分为基于值和基于策略的方法, 最后介绍根据元博弈、种群博弈与智能体建模的理论演变而来的其他方法. 每类方法的优缺点和适用范围将会在具体章节中介绍.

4.3.1 基于值的方法

最早的基于值的博弈论与强化学习结合方法是最小最大 Q 学习 (minimax Q-learning, minimax-Q)[206], 其主要是应用在双人的零和随机博弈问题, 其中在某个状态 s 上智能体 1 的 Q 值更新公式为:

$$Q_1(s,a_1,a_2)=(1-\alpha)Q_1(s,a_1,a_2)+\alpha(r_1+\gamma V_1(s,Q_1)) \tag{45}$$

$$V_1(s,Q_1)=\max_{\pi_1}\min_{a_2\in A_2}\sum_{a_1\in A_1}\pi_1(s,a_1)Q_1(s,a_1,a_2) \tag{46}$$

其中, V 值的计算可以通过简单的线性规划, 因此在每一次状态迭代下都要计算一次 minimax 纳什均衡, 然后进行 Q 学习的更新, 这也是最简单的将博弈论与强化学习方法的结合方式. 同样地, 在合作博弈问题中, 团队 Q 学

习 (team Q-learning, team-Q)[206]方法在计算 V 值函数是只需要最大化所有智能体的共享 Q 值函数, 然后再进行 Q 学习的更新. 而最优自适应学习 (optimal adaptive learning, OAL)[207]方法通过假设最优联合动作的奖赏为 1, 其余联合动作的奖赏为 0 的方式, 能够保证在任何合作型随机博弈问题中以 1 的概率收敛到最优纳什均衡.

随后纳什 Q 学习 (Nash Q-learning, Nash-Q)[208]推广到多人混合博弈问题, 但是其收敛到纳什均衡的条件非常苛刻, 需要在每一个状态下的阶段博弈中, 都存在全局最优点, 即所有智能体的期望收益都是最大的; 或者存在鞍点, 即当前智能体不会从偏离这个鞍点中获益, 但其他智能体可能会从偏离这个鞍点中获益. Nash-Q 学习的值更新公式为:

$$Q_i(s,a_1,a_2,\ldots,a_N)=Q_i(s,a_1,a_2,\ldots,a_N)+\alpha\left[r_i+\gamma NashQ_i(s')-Q_i(s,a_1,a_2,\ldots,a_N)\right] \tag{47}$$

其中, 也需要在每一个状态下通过二次规划的方式求解多人的纳什均衡.

或敌或友 Q 学习 (friend-or-foe Q-learning, FF-Q)[209], 在面对多人博弈问题的时候, 将所有智能体分成两队, 其中一队是敌人, 一队是队友. 因此, FF-Q 后续的更新方式类似于 minimax-Q 方法, 在所有队友或敌人的联合动作空间上进行线性规划求解 minimax 纳什均衡. 或者在多人博弈问题中, 相关 Q 学习 (correlated-Q learning, correlated-Q) 方法[210]提出通过计算相关均衡替代纳什均衡, 可以兼顾考虑个体收益、集体收益和公平收益等多个目标, 且最后能够收敛到相关均衡. 随后, 非平稳收敛策略 (non-stationary converging policies, NSCP)[211]认为不一定需要求解纳什均衡, 寻找最优反应策略是更有意义的, 因此在多人混合随机博弈问题中, 第 1 个提出针对有限类非平稳对手策略的最优学习算法. 具体地, 该算法推断出对手非平稳策略的精确模型, 并同时针对该策略创建最优反应策略.

由于基于均衡的多智能体强化学习算法均不具备较好的可扩展性, 并且利用数学规划的方式计算纳什均衡的代价高昂. 协商 Q 学习 (negotiation-based Q-learning, NegoQ)[212]方法基于不共享值函数的假设, 提出一种多步协商的过程分布式地计算纯策略纳什均衡, 均衡优超策略组 (equilibrium-dominating strategy profile, EDSP) 和弱均衡优超策略组 (non-strict EDSP) 这 3 种纯策略, 能够证明在多人混合随机博弈问题中, 学习性能和收敛速度都高于之前基于均衡的多智能体强化学习算法.

4.3.2 基于策略的方法

最早基于策略的方法主要是处理双人两动作的矩阵博弈问题, 例如在无穷小梯度上升 (infinitesimal gradient ascent, IGA)[213]方法中, 假设智能体 1 执行 a_1 动作的概率为 α, 智能体 2 执行 b_1 动作的概率为 β, 那么 α 和 β 的更新方式可以表示为:

$$\begin{cases}\alpha^{t+1}=\alpha^t+\delta_1^t\dfrac{\partial\mathbb{E}\left[r_1\mid\alpha^t,\beta^t\right]}{\partial\alpha^t}\\ \beta^{t+1}=\beta^t+\delta_2^t\dfrac{\partial\mathbb{E}\left[r_2\mid\alpha^t,\beta^t\right]}{\partial\beta^t}\end{cases} \tag{48}$$

其中, $\delta_1^t=\delta_2^t=\delta$ 表示学习率, IGA 方法证明了当 $\delta\to 0$ 时, 智能体的策略或者智能体策略的平均奖赏将会收敛到纳什均衡. 在此基础上, 赢或快速学习无穷小梯度上升 (win-or-learn-fast IGA, WoLF-IGA)[214]方法在此基础上修改了学习率的形式, 在智能体 i 赢的情况下使用较小的学习率, 而在输的情况下使用较大的学习率:

$$\delta_1^t=\begin{cases}\delta_{\min}, \text{如果} r_1(\alpha^t,\beta^t)>r_1(\alpha^{\text{Nash}},\beta^t)\\ \delta_{\max}, \text{否则}\end{cases} \tag{49}$$

$$\delta_2^t=\begin{cases}\delta_{\min}, \text{如果} r_2(\alpha^t,\beta^t)>r_2(\alpha^t,\beta^{\text{Nash}})\\ \delta_{\max}, \text{否则}\end{cases} \tag{50}$$

其中, $\alpha^{\text{Nash}}/\beta^{\text{Nash}}$ 表示在该时刻下智能体的均衡策略, 是否输赢也是和该策略进行比较. 当 $t\to\infty$, δ_1^t 和 δ_2^t 都趋近于 0 时, 该方法能够保证收敛到纳什均衡.

之后, 广义的无穷小梯度上升 (generalized IGA, GIGA)[215]方法将 IGA 方法进行推广, 能够处理双人多动作的矩阵博弈问题, 其中策略更新方式为:

$$x_i^{t+1}=\arg\min_{x\in\Pi(A_i)}x-\left(x_i^t+\eta^t r_i(a_i^t,a_{-i}^t)\right) \tag{51}$$

其中, 当 $\eta^t = 1/\sqrt{t}$ 时, GIGA 方法是满足普遍一致性 (universally consistency) 并且遗憾趋于 0. 同时, GIGA 方法可以与 WoLF 相结合成为 GIGA-WoLF[216]能够同时保证纳什均衡的策略收敛性和遗憾趋于 0 的两个指标. 而对于连续动作的矩阵博弈问题, 基于策略的方法可能会陷入局部纳什均衡[217], 甚至完全会避开全局纳什均衡[218], 因此一种基于零均值和有界方差的双重平均 (dual averaging with zero-mean and finite mean squared error)[219]方法通过引入变分稳定性 (variational stability) 的概念, 并证明了稳定均衡中局部纳什均衡具有高概率的吸引, 而全局稳定均衡具有概率为 1 的吸引, 因此平均策略肯定能够收敛到全局纳什均衡.

以上基于策略的方法都只能解决矩阵博弈问题, 而对于多人随机博弈问题, 基于策略的方法能够与策略爬山 (policy hill-climbing, PHC)[214]算法相结合进行求解. 例如赢或快速学习策略爬山 (win-or-learn-fast policy hill-climbing, WoLF-PHC)[214]和策略动态赢或快速学习 (policy dynamics WoLF, PD-WoLF)[220]方法分别使用一阶和二阶的策略信息判断输赢条件, 然后使用不同大小的学习率更新当前策略参数, 最终证明策略在自博弈的条件下能够收敛到多人随机博弈问题的纳什均衡.

4.3.3 其他演变方法

首先, 将虚拟自博弈方法与多智能体学习方法相结合, 提出了一种针对双人扩展式博弈场景的神经虚拟自博弈 (neural fictitious self-play, NFSP)[221]方法, NFSP 使用两个深度神经网络计算近似纳什均衡, 其中一个网络是通过强化学习计算近似最优反应, 另外一个使用监督学习的方式计算历史平均策略, 最后通过两种策略的凸组合来作为最终的均衡策略.

另外, 由于以上方法的原子动作的纯策略空间过大, 无法在实际问题中应用, 因此一类基于元博弈 (meta game) 概念的方法提出在策略空间上进行纳什均衡的应对和求解, 元博弈分析通常也被称为经验博弈理论分析 (empirical game-theoretic analysis, EGTA)[222]. 例如, 策略空间的应对预言 (policy space response oracle, PSRO)[118]通过扩展虚拟博弈和双重预言的概念, 引入了强化学习和种群训练 (population-based training) 的方式计算最优反应:

$$\pi_1^{t+1} = BR_1(\pi_2^t) \subseteq oracle\left(\pi_1^t, \sum_{a_2 \in A_2} \pi_2^t(a_2) \cdot \phi_{\omega_i}(a_1, a_2)\right) \tag{52}$$

其中, ϕ_{ω_i} 函数表示该种群比对手多获得的奖赏. 在此基础上, 一种基于纠正的最优反应 (rectified best response) 的 PSRO_{rN}[223]方法提出计算纳什均衡策略时, 只选择能赢得过的对手策略, 不去考虑那些赢不过的对手策略, 因此计算纠正的最优反应为:

$$\pi_1^{t+1} = BR_1(\pi_2^t) \subseteq oracle\left(\pi_1^t, \sum_{a_2 \in A_2} \pi_2^t(a_2) \cdot \phi_{\omega_i}^+(a_1, a_2)\right) \tag{53}$$

其中, $\phi_{\omega_i}^+$ 函数表示只取大于 0 的部分 $\phi_{\omega_i}^+ = \max(0, \phi_{\omega_i})$, 即不考虑那些对手较强的策略, 最终能获得更优的表现和收敛性能. 随着信息集状态数量的增加, PSRO 方法需要迭代的次数也是指数级增加, 因此神经扩展式双重预言 (neural extensive-form double oracle, NXDO)[119]没有使用 *oracle* 最佳响应, 而是使用深度强化学习算法训练的近似最优反应, 例如 PPO 或 DQN 等. 如果深度强化学习计算的近似最优反应足够接近 *oracle* 最优反应, 并且内循环求解器能找到一个足够接近的受限博弈的近似纳什均衡, 那么 NXDO 就能享有与 XDO 相同的收敛性质.

以上方法只能解决双人的矩阵博弈或者扩展式博弈, 因此 α-Rank [224]方法首先提出通过反应图 (response graph) 的方式用于解决多人的矩阵博弈问题, 其中反应图的每个节点代表一个所有智能体的联合纯策略, 然后节点之间的有向边表示, 在通过这条边改变联合策略的情况下, 至少有一个智能体能够获得更高的收益. 最后通过随机游走 (random walk) 的方式来寻找反应图中的下沉强联通分量 (sink strongly-connected components, SSCC), 并且证明该方法在多智能体矩阵博弈场景中的解是唯一并且计算复杂度是 P 完全. α-PSRO [225]方法融合了 α-Rank 和 PSRO 的优点, 通过 α-Rank 的求解器在每次博弈上求解基于偏好的最优反应 (preference-based best response, PBR), 保证在迭代结束之前能找到反应图中的 SSCC, 广泛适用于各种 Kuhn、Leduc 扑克和 MuJoCo 足球等多智能体环境.

第 2 类方法是将智能体建模应用在多智能体强化学习与博弈论中. 一种经典的方法是学习对手学习意识 (learning with opponent-learning awareness, LOLA)[226], 针对多智能体环境下由于策略学习过程导致非平稳环境的

问题, 因此需要每个智能体塑造环境中其他智能体的预期学习. 假设智能体 2 的学习过程对智能体 1 的影响如下:

$$V_1(\theta_1,\theta_2+\Delta\theta_2)\approx V_1(\theta_1,\theta_2)+(\Delta\theta_2)^T\nabla\theta_2V_1(\theta_1,\theta_2)=V_1(\theta_1,\theta_2)+\eta\cdot\nabla\theta_2V_2(\theta_1,\theta_2)\nabla\theta_2V_1(\theta_1,\theta_2) \tag{54}$$

基于此, 智能体 1 的策略更新方式为:

$$\theta_1^{t+1}=\theta_1^t+\delta\cdot\nabla\theta_1V_1(\theta_1,\theta_2)+\delta\eta\cdot(\nabla\theta_2V_1(\theta_1,\theta_2))^T\nabla\theta_1\nabla\theta_2V_2(\theta_1,\theta_2) \tag{55}$$

其中, δ 和 η 分别为智能体 1 和 2 的学习率, 最终在各种实验场景中验证了 LOLA 算法能够收敛到纳什均衡. 随后, 深度贝叶斯策略重用 (deep Bayesian policy reuse+, Deep BPR+)[227]方法研究对抗非平稳环境下智能体的有效策略检测和重用技术, 在贝叶斯策略重用 (Bayesian policy reuse+, BPR+)[228]算法的基础上, 提出了一种新的深度 BPR+算法, 该算法采用神经网络作为值函数逼近器. 同时, 通过引入了蒸馏策略网络作为 BPR+中的策略库, 实现了高效的在线策略学习和重用, 并在多个随机博弈场景中取得了更高的累计奖赏和收敛性能.

在智能体数量增加到无限多时, 普通的求解方法在面度维度灾难和策略空间指数级增长的问题下将无法应对. 最后一类基于平均场理论 (mean-field theory) 的多智能体强化学习方法, 提出将周围其他智能体的平均动作作为输入来影响自身智能体的策略学习, 与 Q 学习和 Actor-Critic 方法结合分别形成了 MFQ 和 MFAC 方法[229]. Q 值更新方式表示为:

$$Q_i^{t+1}(s,a_i,\bar{a}_i)=(1-\alpha)Q_i^t(s,a_i,\bar{a}_i)+\alpha\left[r_i+\gamma v_{i,t}^{\mathrm{MF}}(s')\right] \tag{56}$$

$$v_{i,t}^{MF}(s')=\sum_{a_i}\pi_i^t(a_i\mid s',\bar{a}_i)\cdot\mathbb{E}_{\bar{a}_i}\left[Q_i^t(s',a_i,\bar{a}_i)\right] \tag{57}$$

其中, $\bar{a}_i=\dfrac{1}{|N_i|}\sum_{k\in N_i}a_k$ 表示智能体 i 附近 $|N_i|$ 个智能体的平均动作. 最后该方法也证明在智能体数量非常大的随机博弈问题中能够收敛到纳什均衡.

以上所有结合方法的简单总结与对比如表 6 所示.

表 6 博弈论与强化学习结合方法的总结与对比

学习范式	算法	博弈形式化	大于双人	深度神经网络	解概念/求解目标	博弈场景		
						合作	零和	混合
经典基于值的算法	Minimax-Q[206]	SG	—	—	纳什均衡/个体收益	—	√	—
	Team-Q[206]	SG	√	—	帕累托最优/集体收益	√	—	—
	OAL[207]	SG	√	—	帕累托最优/集体收益	√	—	—
	Nash-Q[208]	SG	√	—	纳什均衡/个体收益	—	√	√
	FF-Q[209]	SG	√	—	纳什均衡/个体收益	√	√	√
	Correlated-Q[210]	SG	√	—	相关均衡/个体/集体/公平收益	√	√	√
	NSCP[211]	SG	√	—	纳什均衡/个体收益	√	√	√
	NegoQ[212]	SG	√	—	纯策略纳什均衡/均衡优超策略组/弱均衡优超策略组	√	√	√
经典基于策略的算法	IGA[213]	Matrix game	—	—	纳什均衡/个体收益	√	√	√
	WoLF-IGA[214]	Matrix game	—	—	纳什均衡/个体收益	√	√	√
	GIGA[215]	Matrix game	—	—	纳什均衡/个体收益	√	√	√
	GIGA-WoLF[216]	Matrix game	—	—	纳什均衡/个体收益	√	√	√
	DA[219]	Matrix game	—	—	纳什均衡/个体收益	√	√	√
	WoLF-PHC[214]	Matrix game/SG	√	—	纳什均衡/个体收益	√	√	√
	PD-WoLF[220]	Matrix game/SG	√	—	纳什均衡/个体收益	√	√	√
其他演变方法	PSRO[118]	EFG	—	√	纳什均衡/个体收益	—	√	—
	$\mathrm{PSRO}_{\mathrm{rN}}$[223]	EFG	—	√	纳什均衡/个体收益	—	√	—
	α-Rank[224]	Matrix game	√	√	纳什均衡/个体收益	√	√	√
	α-PSRO[225]	EFG	√	√	纳什均衡/个体收益	√	√	√

表 6 博弈论与强化学习结合方法的总结与对比 (续)

学习范式	算法	博弈形式化	大于双人	深度神经网络	解概念/求解目标	博弈场景		
						合作	零和	混合
其他演变方法	NXDO[119]	EFG	—	√	纳什均衡/个体收益	—	√	—
	NFSP[221]	EFG	—	√	纳什均衡/个体收益	—	√	—
	LOLA[226]	SG	√	√	纳什均衡/个体收益	√	√	√
	Deep BPR+[227]	SG	√	√	个体收益	√	√	√
	MFQ/MFAC[229]	SG	√	√	纳什均衡/个体收益	√	√	√

5 应用场景和仿真环境

混合博弈问题在现实社会中广泛存在. 作为一个复杂的社会化系统, 智能体之间的关系不仅是纯粹的零和竞争或完全合作, 更常见的是一般和的混合博弈. 因此, 在社会资源分配、金融市场以及政治协商等诸多领域, 构建高效而公平的多方协作策略具有重要意义. 该策略不仅需要确保个体相对公平的收益, 还需要确保系统的整体收益, 从而促进人类社会的和谐发展. 因此, 混合博弈不仅在学术上有重要的科研价值, 在现实中也是我们面对且亟待解决的问题. 本节将具体介绍混合博弈问题的现实应用场景以及实验仿真环境.

5.1 应用场景

混合博弈问题涉及人类政治、经济和生活等诸多方面, 下面我们将从政治上的公共资源分配、社会困境问题和规范形成, 到经济与生活上的金融市场和自动驾驶的实际问题介绍相关的应用场景.

● 公共资源分配. 人类面临资源共享和分配和确保资源的可持续利用等诸多问题. 其中, 基于混合博弈的公共池资源 (common-pool resources, CPRs)[230]分配的抽象模型表示纯自利的智能体通常无法找到社会正平衡. 例如, 每当一个智能体从这种资源中获得个体收益时, 可供其他智能体挪用的剩余资源就会略微减少, 一味获取的纯自利智能体必然会导致系统的崩溃. 特别当这种资源是一种可再生资源时, 如何权衡好资源分配和再生的问题也是人类社会中的现实问题[231]. 因此, 近年来有相关研究提出了 4 个指标: 功利指标 (utilitarian metric)、公平指标 (equality metric)、可持续指标 (sustainability metric) 以及和平指标 (peace metric) 来衡量混合博弈中的策略性能, 并在实验上从经验主义的角度出发分析得到纳什均衡在社会总和的表现上是有缺陷的, 可以通过智能体协商的方式进行调解[24]. 此外, 有研究表明种群规模的变化会对多智能体系统产生影响, 通过一个种群的平均收益来修正人口数量, 会比自博弈方法更快收敛到最优结果[232].

● 社会困境问题和社会规范的形成. 社会困境问题是当每一个智能体都不愿意奉献付出时, 整个社会就无法创新和发展. 相关工作研究社会交互下的群体智能, 发现自动生成课程可以激发群体创新, 这种创新会出现在生命体的各个层次. 群体创新的发生具有外在和内在挑战: 外在挑战是智能体策略具有遗忘性, 会抑制创新的形成; 内在挑战是社会困境的出现, 而解决社会困境的同时会产生更高阶的社会困境[233]. 总之, 仅关注人类认知能力是不够的, 还需要关心学习和进化的过程, 同时环境的多样性和群体策略多样性也是提升社会创新的关键[234].

而对于社会困境问题, 现有的解决方式是形成社会规范. 第 1 种社会规范是基于领导者和跟随者的模式: 相关研究表明, 具有较高声望的领导可能为许多情况下的合作提供基础[235]. 这种声望的传播既发生在领导者和跟随者之间, 也发生在跟随者之间. 在群体相对较小的情况下, 自然选择有利于使声望高的领导者形成更加亲社会的基因; 而在群体较大的情况下, 地位差异的有效性取决于文化传播, 而不是取决于顺从或强制规范[236]. 第 2 种是形成普遍认同的社会规范或法则[237], 例如, 形成制裁为基础的社会规范, 对违反规则的智能体进行相关的惩罚措施[238]. 同时, 一些毫无意义的规则可以对社会福利产生间接影响, 它们可以帮助智能体学习如何执行和遵守一般规范, 提高群体执行对社会福利有直接影响的规范的能力[239].

● 金融市场. 金融领域关注的是投资者的储蓄如何通过金融市场和中介机构分配给公司, 公司利用这些储蓄为其商业活动提供资金. 金融领域可以大致分为两个部分: 首先是资产定价, 这与投资者的决策有关[240]; 第 2 个是公司金融, 它与公司的决策有关[241]. 唯一可取的解决方案就是双方达成协议: 企业造成现货市场和期货市场之间

的不一致, 而金融机构需要遵守经济交易税的规范性, 无法单独进行套利, 抓住机会卖出股票以赢得可能的最大集体金额[242]. 因此金融市场中的交易也不是简单的零和博弈问题, 参与智能体也需要充分考虑到现实环境以及其他参与机构和公司的状态, 理解货币交易市场的博弈机制和规律, 才能在该混合博弈问题中获取较高利润[243,244].

● 自动驾驶. 自动驾驶汽车在道路上的行为会影响其他智能体的行为, 并受到其他智能体行为的影响, 无论是超车、协商合并还是避免事故. 这种相互依赖关系能较好地通过动态博弈描述, 将车辆的行驶规划和对其他智能体行为预测相耦合, 这一点对于自动驾驶技术的安全性和可行性具有直接影响, 因此也成为了一个开放性问题. 然而, 动态博弈对计算量的要求太高, 无法满足自动驾驶在连续状态和动作空间中的实时性决策要求[245]. 此外, 多辆车在通过一个路口时, 并非简单的零和博弈问题, 而是一个涉及先后顺序的混合博弈问题. 在自动驾驶中的司机智能体具有不同的驾驶风格和性格特征, 例如在遇到其他司机智能体变道或者转向的情况下, 自身智能体选择是抢行还是减速避让的策略, 并且司机智能体不一定是完全理性. 因此, 当前相关的研究可能需要对智能体进行有限理性的建模[246]或者递归推理智能体意图[247]等方式, 最终在确保安全的前提下, 求解纳什均衡、相关均衡、Stackelberg 均衡等概念, 或者面向合作倾向的帕累托最优[248].

5.2 仿真环境

● RLCard[249]. RLCard (https://rlcard.org/) 是一个用于多人扩展式博弈纸牌游戏中的强化学习工具包. 它支持 21 点、德扑、UNO、桥牌、斗地主和麻将等多种环境, 具有易于使用的界面和接口函数, 其中状态和动作编码简单. 不同类型的纸牌游戏在相同的程序结构下实现, 逻辑清晰, 并且提供了评估工具, 通过比赛的胜率来衡量不同算法的表现. RLCard 的目标是连接强化学习和不完全信息博弈之间的桥梁, 推动多智能体、大状态和大动作空间、稀疏奖赏领域的强化学习研究. 其中内置了 DQN、NFSP 和 CFR 等多种基本算法. 与实际纸牌游戏的差异: 基本规则与实际纸牌游戏相同, 除此之外还提供了其他高级功能, 例如可以回溯状态重新出牌, 以及缩小状态空间版本的纸牌环境, 便于算法训练和评估.

● Werewolf game. 狼人游戏 (werewolf game) 是一种典型的多人社会推断博弈 (social deduction game) 问题 (http://aiwolf.org/en/5th-international-ai-werewolf-competition). 在博弈中, 智能体必须隐藏自身信息, 这与国际象棋或围棋等完全信息博弈不同. 每个智能体通过隐藏信息且从其他智能体的对话和行为中获取秘密信息来完成自己的目标. 该博弈突出了人工智能领域尚未得到充分解决的各种问题, 例如智能体信息的不对称分布, 使用通信和交流来获得其他智能体信任, 以及使用推断来检测事实. 参与智能体将分成狼人阵营和人类阵营, 狼人阵营包括狼人和附身的角色, 人类阵营包括村民、预言家、巫师和守卫等角色, 分别具有不同的能力. 博弈将会持续几天, 白天的时候所有智能体可以投票选择一人出局, 晚上狼人阵营可以内部确认选择一人出局, 直到所有人类或狼人阵营全部出局, 此时另外一方获得胜利. 智能体在白天的交流可以通过协议通信和自然语言通信的方式进行. 与实际狼人游戏的差异: 基本规则与实际狼人游戏相同, 主要差距体现在阵营角色没有实际游戏中丰富, 智能体相互交流的自然语言通信方式的语料库单一, 并且无法体现出实际游戏场景中的人类社交和心理博弈策略.

● SMAC[20]. 星际争霸多智能体挑战 (StarCraft multi-agent challenge, SMAC) 是一种典型的多人合作博弈问题. SMAC 专注星际争霸 II 游戏的微观操作 (https://github.com/oxwhirl/smac), 其中每个本方战斗单位都由一个独立的智能体控制, 并且根据本地高维输入且部分可观测的状态采取相应动作实现智能体之间的合作行为, 而对手单位由内置算法进行控制. SMAC 提供了许多不同的场景和地图, 并且开源了一个深度多智能体强化学习框架 PyMARL, 其中包括 IQL、COMA、VDN 和 QMIX 等多种基础算法. 与实际军事对抗任务的差异: 游戏场景只简单模拟出了即时战略游戏中战斗单位对抗场景, 缺乏实际军事对抗任务中地理环境、战斗资源、作战指挥与作战通信等复杂信息的考虑, 且游戏的微操场景无法模拟出实际军事对抗任务战役级的作战目标和战术特点.

● GRF[21]. 谷歌足球研究 (Google research football, GRF) 是一种典型的多人合作博弈问题 (https://github.com/google-research/football). GRF 是一个新颖的开源强化学习环境, 提供了一个基于物理的 3D 足球模拟器, 其中智能体控制本方球员, 学习如何在本方不同智能体之间传球, 以及如何战胜内置规则策略的对手以得分. 这提供了一个具有挑战性的强化学习问题, 因为足球需要在学习不同传球、控球和射门等概念和高水平战术策略之间取得自然平衡. GRF 提供了简单易用的 API, 创建了 11 种局部对抗、过人和射门等场景, 并复现了 IMPALA、PPO 和 Ape-X

DQN 等多种基础算法. 与实际足球比赛的差异: 基本规则与实际足球比赛相同, 主要差距体现在不同球员特点不够鲜明, 球员类型和技术没有明显区分, 仿真环境中的球员动作没有实际人类球员动作丰富. 并且, 在真实足球比赛中关键的对抗阵型、对抗战术与人员替补等内容没有在仿真环境中体现.

● MPE[194]. 多智能体粒子环境 (multi-agent particle environment, MPE) 是一种典型的多人博弈问题 (https://github.com/openai/multiagent-particle-envs). 该环境由多个智能体和地标组成, 它们分布在一个空间连续、时间离散的二维世界中. 智能体观察高级特征, 比如它们的速度和相对于环境中的地标和其他智能体的位置, 同时智能体可以选择离散的动作对应于每个基本方向的移动, 或者使用连续的动作以不同的速度向任何方向移动. MPE 包括完全观测和部分可观测的竞争型、合作型和队内合作队间竞争型任务, 具体包含 6 个基础场景, 其中有纯合作的导航、通信任务, 也包含两队竞争队内合作的追捕猎物等任务, 并且提供了 DDPG、MADDPG、TRPO 和 DQN 等多种基础算法. 由于场景与 MAgent 类似, 与实际场景任务的差异见下文 MAgent 环境中介绍.

● MAgent[250]. MAgent 环境是为了建模数以万计的智能体如何进行训练评估以及群体智能的涌现, 包含了多人合作、两队竞争以及多人混合博弈场景 (https://github.com/geek-ai/MAgent). 在智能体群体之间的相互作用中, MAgent 不仅可以研究智能体最优策略的学习算法, 更重要的是可以观察和理解个体智能体的行为和社会现象, 包括沟通语言、领导力、利他主义等. MAgent 包含了追逐、收集和战争等 3 个子场景, 分别建模合作、混合和两队竞争博弈, 同时提供了参数共享的 DQN、DRQN 以及 A2C 等基础算法. 与实际场景中人类竞争合作的差异: MPE 和 MAgent 是基于虚构的仿真环境, 其设计和参数空间完全由研究者控制. 相比之下, 人类拥有更为丰富和灵活的感知和认知能力, 能够更好地适应各种情境和任务; 其次, 实际场景中存在更多的不确定性和真实感, 包括人类的情感、意图、误差因素; 最后, 人类在竞争和合作中通常需要进行复杂的社会互动, 包括协商与谈判. 这些方面在仿真环境中都没有很好的体现.

● Melting Pot[251]. 熔炉 (melting pot) 环境是集成 80 多个独特场景的多样化任务集合, 其中包含了类似 Cleanup 和 Harvest 等典型的社会困境、互利互惠、资源分配等多人混合博弈问题 (https://github.com/DeepMind/meltingpot). 智能体在所有环境中都有 6 个离散的运动动作, 或者在一些特殊场景中的攻击或者清理动作. 在所有环境中, 智能体都直接将部分可观测的图像作为状态输入. 熔炉环境不仅能够评估算法在不同智能体上取得的收益, 同时研究算法在混合博弈场景中能否涌现出群体智能, 在取得较高的社会福利条件下兼顾收益公平, 能否在零样本的新环境或者新智能体交互的情况下拥有较好的泛化性能. 同时熔炉环境提供了 A3C、V-MPO 和 OPRE 等多种基础算法. 与实际公共资源分配的差异: 熔炉环境只简单模拟了多智能体在混合博弈场景下的资源分配和再利用问题, 其设计的任务和情景非常简单, 智能体均同构没有差异, 观测空间仅为当前智能体视角下的地图观测范围, 动作空间也是预定好的简单离散动作. 相比于真实公共资源分配问题, 仿真环境缺乏对人类心理和情感的建模认知, 不同人对资源分配问题的看法和接受程度不同, 以及人类之间复杂的协商与谈判机制策略. 最后, 仿真环境也缺乏对智能体行为的道德和伦理约束.

● SMARTS[61]. 可扩展的多智能体强化学习训练学校 (scalable multi-agent RL training school, SMARTS) 模拟现实世界中自动驾驶, 是一种典型的多人混合博弈问题 (https://github.com/huawei-noah/SMARTS). SMARTS 支持可扩展的集成和分布式训练方法, 提供网络流可视化界面, 并集成了流行的 PyMARL 和 MAlib 训练框架, 支持 DQN、PPO、MAAC、MFAC、Net-Q、CommNet 和 MADDPG 等多种基础算法. 并且, SMARTS 提供了目标达到、碰撞系数、目标距离和错误行驶等诸多评价指标, 便于开发和验证强化学习和博弈论算法. 与实际自动驾驶的差异: 仿真环境中的观测空间被指定为可用传感器类型的配置子集, 包括动态对象列表、地图俯视信息、RGB 图像、自我车辆状态和道路结构等, 无法完全建模真实世界中传感器的复杂性、噪声、不确定性和动态变化; 缺乏对道路条件、天气状况、交通情况等复杂信息的环境建模; 最后, 仿真环境中车辆的动作输出是基于预定义的物理模型和算法, 无法反映人类真实的驾驶行为和不同车辆的驾驶性能.

● Neural MMO[252]. 神经大型多人在线角色扮演游戏 (neural massively multiplayer online role-playing games, Neural MMO) 旨在模拟地球世界上大量生物争夺有限资源引发的军备竞赛问题, 是一种典型的多人混合博弈问题 (https://github.com/openai/neural-mmo). 与现实世界一样, Neural MMO 环境是持续的, 并且支持大量智能体存在, 适合研究大规模的多智能体交互作用, 在实现生存的目标下群体能涌现出强大的战斗和资源导航策略. Neural

MMO 核心特征是支持大量可变数量的智能体、基于方块地形的食物和水的觅食系统、战略作战系统以及用于分析学习策略的内置可视化工具. 环境地形也是基于方块形状, 如承载食物的森林方块和草地方块. 在加入环境后, 智能体会在环境边缘的随机位置生成. 为了保持健康和血量, 智能体必须获得食物和水, 它们生命健康系数到零时会立即死亡. 在森林方块上或靠近水方块, 可以分别补充智能体的食物或水供应. 然而, 森林方块的食物供应有限, 一旦耗尽, 仅会有小概率的机会再生. 这意味着智能体必须竞争食物方块, 同时定期从无限的水方块中补充水供应. 他们可以使用 3 种攻击模式中的任何一种攻击敌方智能体, 每种攻击选项都有不同的伤害值和特点. Neural MMO 环境集成了基础的策略梯度算法, 并且提供了易于使用的 API 接口帮助实现其他强化学习及博弈论算法. 与实际生物种群生存对抗和军备竞赛的差异: 首先与上述 SMAC 仿真环境类似, 环境、状态以及动作空间的抽象化过高, 无法反映真实对抗场景的复杂性. 其次, 针对特定的生物种群生存对抗问题, 没有考虑到生物自身群体内协作、繁衍和进化的能力, 以及群体间的沟通、协商与妥协机制.

6 挑战与发展

目前混合博弈的挑战集中在问题规模、任务建模、方法求解和实际应用等从问题到应用的层层递进的维度上. 具体地, 问题规模主要是由于策略空间随着智能体数量的增加而指数级增长, 这也是多智能体博弈任务的原生挑战. 其次, 将实际问题建模成可数学形式化求解的博弈任务, 为该博弈定义合适的解概念和求解目标, 解决目前深度学习结合方法的理论完备性是将实际问题进行建模到求解的核心挑战. 最后, 实现高效的样本利用率, 提供强大的策略泛化和解释性能可以保障实际场景中的应用落地. 下面我们将具体介绍每种问题的挑战与发展.

● 策略空间指数级增长. 随着智能体数量的增加, 联合策略空间会指数级增长. 这将会导致博弈论和多智能体强化学习面临着高时间复杂度和计算成本的问题, 例如, 博弈论中求解最优反应和平均策略, 多智能体强化学习中计算联合 Q 值函数等, 因此传统的求解算法不再可行. 对于博弈论, 目前研究的发展方向主要是通过设计深度网络的计算近似最优反应和平均策略[221]、通过元博弈理论进行训练[118], 以及博弈约简和知识迁移[253]等; 而对多智能体强化学习, 则可以通过分布式训练以及分层强化学习[254]的方式降低时间复杂度和缩减策略空间等. 总的来说, 多智能体强化学习和博弈论在智能体数量增加时面临策略空间指数级增长的挑战, 研究者们正在努力寻找通过分布式训练、种群训练、近似算法和策略约简的高效求解方案, 这对于解决复杂的多智能体系统问题具有重要意义.

● 博弈任务建模困难. 多智能体博弈任务因智能体间交互关系复杂、信息部分可观测等因素通常会导致博弈模型构建困难, 真实场景与仿真模型差距较大, 策略求解复杂度高等问题. 传统机器学习领域研究者引入归纳偏置为算法模型设置额外偏好以更好挖掘底层数据的潜在规律, 辅助算法模型实现更优性能. 受此启发, 利用博弈数据归纳多智能体博弈任务知识, 包括智能体间交互知识和环境交互知识, 隐式构建博弈模型并辅助博弈策略求解有利于缓解以上问题. 针对智能体间交互知识建模, 目前有关于协同学习[255]、协同图[256]以及智能体建模[257,258]等方法; 而如何基于智能体局部观测推断环境状态信息、其他智能体观测信息以及学习环境模型仍然是一个开放性问题. 总体而言, 利用先验数据建模多智能体博弈任务, 构建真实场景与仿真模型的桥梁是求解复杂混合博弈任务的先驱.

● 混合博弈的解概念和求解目标模糊. 在多智能体混合博弈问题中, 存在着诸多的解概念和求解目标. 传统的纳什均衡和相关均衡概念, 是基于绝对理性的假设来求解最优策略. 但是, 在实际非完全信息条件下, 基于均衡的解未必是分布式决策中的“最优解”, 例如可能存在完全占优的帕累托最优策略. 其次, 在复杂的社会化系统中, 基于完全理性和均衡的假设都过于严格, 智能体个人和系统更关心的是个体收益[161]、集体收益[188]以及公平公正[174]等实际的优化目标或指标. 因此, 多智能体强化学习和博弈论在面对实际混合博弈问题时, 需要定义好建模后的博弈任务的最优解、均衡解、近似解或求解目标等基础概念, 并分析该解对应的性质, 这是未来研究混合博弈问题的基石.

● 深度网络方法理论不完备. 传统的双人零和博弈问题以及基于值表形式的强化学习方法都具有严谨的数学理论和收敛证明[206,208]. 但是, 应用到实际大规模群体连续空间决策任务中, 由于深度网络方法的引入, 导致诸多启发式和近似性的求解过程, 原始的数学理论及证明过程中的假设不再适用. 目前的博弈论和多智能体强化学习方法更关注于在实验仿真环境中的评价指标, 而忽略了研究在实际应用和新兴技术下的博弈和多智能体强化学习理论体系. 总的来说, 结合深度学习技术, 建立多智能体强化学习及博弈论方法的相关理论, 证明其最优解、均衡解、

近似解或求解目标的收敛性是未来研究方向的关键.

● 样本利用率低. 对比理论研究现状与实验仿真环境, 实际多智能体博弈任务中往往面临奖赏稀疏、观测信息不精确、环境噪声以及数据采样困难等诸多问题, 因此需要考虑如何提升样本利用效率的问题[259]. 目前, 一类方法是通过增强对环境的探索能力提升样本利用率[40], 在多智能体场景中更需要探究何时需要探索[260,261]以及智能体间怎样协同探索[262]的问题. 其次, 另一种方法可基于离线强化学习的思想通过收集到的离线轨迹数据直接进行策略的学习, 有效地提升样本利用率[263,264], 同时也能避免代价昂贵的在线交互试错. 离线强化学习也将进一步拓展多智能体在无人驾驶、工业控制以及医疗 AI 等领域的研究. 总的来说, 研究实验仿真环境与实际任务的差距, 结合深度多智能体强化学习的技术手段提升样本利用率也是未来混合博弈问题研究的重要领域.

● 策略泛化性弱. 混合博弈问题中利用博弈论和多智能体强化学习方法求解的策略面临着泛化性差的问题, 策略受限于具体的任务或场景之中, 面对新的环境无法有效地实现迁移, 这将会限制当前方法在实际混合博弈问题中的应用[265,266]. 目前已有的方法主要是从数据增广[267–269]、函数正则化[270,271]和对抗策略正则化[272,273]这 3 种方式提升强化学习算法在新环境中的泛化性能, 并且在较为成熟的单智能体强化学习领域主要考虑的是状态、奖赏、转移概率和策略这 4 个泛化问题. 当面对多智能体混合博弈问题的时候, 除上述问题外, 还应该额外考虑到新环境下智能体数量和智能体竞争合作关系的泛化问题. 总的来说, 研究强化学习训练环境和测试环境上的模型泛化能力, 结合传统机器学习领域的泛化方法并考虑多智能体系统中的特有问题, 是未来针对混合博弈问题应用落地的关键.

● 策略可解释性弱. 在混合博弈问题中, 特别是深度网络方法引入之后, 黑盒化的策略模型无法进行解释, 因此针对智能体输出的策略进行解释性验证成为一个挑战性难题. 目前单智能体强化学习的可解释研究方法主要针对人类主观因素进行考虑, 并且将现有方法划分为内在和事后可解释方式[274]. 除此之外, 一些综述文章根据强化学习本身的特征, 定义强化学习可解释的 3 个独有问题, 即环境解释、任务解释、策略解释[39]. 文献 [275] 更为细致地分析了强化学习可解释性研究中的可视化、查询解释、策略总结、人在回路以及解释验证等多种方法. 而对于更为复杂的混合博弈问题, 策略的可解释性还应该考虑到如何解释智能体间形成的隐式竞争合作关系、智能体的完全理性或有限理性以及智能体对自身收益的接受程度等. 总的来说, 未来工作应深入研究混合博弈的策略可解释性, 实现混合博弈策略溯因及人机融合策略解释, 从而提升策略的可信度以及安全性, 最终实现真实场景的应用落地.

7 总 结

混合博弈作为现实场景中最普遍的博弈类型, 已经得到了越来越广泛的关注. 本文主要针对混合博弈问题, 讨论了该问题的特点和范围, 并且综述目前的相关方法. 由于该领域存在实际问题过于复杂难以建模与量化、求解算法与技术繁多以及解概念和求解目标模糊等诸多问题, 导致当前研究缺乏统一的体系和脉络. 因此, 本文针对以上挑战进行深入的研究与总结, 具体地, 本文首先介绍混合博弈问题的定义与分类, 从现实环境到经典博弈问题的建模方法; 其次, 分析了当前混合博弈问题中的解概念和求解目标, 并基于此分析和总结目前经典博弈论、强化学习以及两者结合的方法; 最后, 讨论了混合博弈问题的具体应用场景和实验仿真环境, 分析了今后研究的挑战与发展方向. 虽然, 该领域当前的研究工作众多, 但是仍然存在博弈建模困难、解概念模糊、求解理论缺乏、方法效率低下和难以应用等诸多问题, 也是未来亟待研究和突破的重点方向.

References:

[1] von Neumann J, Morgenstern O. Theory of Games and Economic Behavior. 2nd ed., Princeton: Princeton University Press, 1947.

[2] Nash J. Non-cooperative games. Annals of Mathematics, 1951, 54(2): 286–295. [doi: 10.2307/1969529]

[3] Nash J. Two-person cooperative games. Econometrica, 1953, 21(1): 128–140. [doi: 10.2307/1906951]

[4] Bielefeld RS. Reexamination of the perfectness concept for equilibrium points in extensive games. Models of Strategic Rationality. Springer, 1988. 1–31. [doi: 10.1007/978-94-015-7774-8_1]

[5] Mertens JF. Repeated games. In: Ichiishi T, Neyman A, Tauman Y, eds. Game Theory and Applications. London: Academic Press, 1990. 77–130. [doi: 10.1016/B978-0-12-370182-4.50009-X]

[6] Silver D, Huang A, Maddison CJ, *et al*. Mastering the game of Go with deep neural networks and tree search. Nature, 2016, 529(7587):

484–489. [doi: 10.1038/nature16961]
[7] Silver D, Schrittwieser J, Simonyan K, *et al*. Mastering the game of Go without human knowledge. Nature, 2017, 550(7676): 354–359. [doi: 10.1038/nature24270]
[8] Moravčík M, Schmid M, Burch N, Lisý V, Morrill D, Bard N, Davis T, Waugh K, Johanson M, Bowling M. DeepStack: Expert-level artificial intelligence in heads-up no-limit poker. Science, 2017, 356(6337): 508–513. [doi: 10.1126/science.aam6960]
[9] Brown N, Sandholm T. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. Science, 2018, 359(6374): 418–424. [doi: 10.1126/science.aao1733]
[10] Brown N, Sandholm T. Superhuman AI for multiplayer poker. Science, 2019, 365(6456): 885–890. [doi: 10.1126/science.aay2400]
[11] Li JJ, Koyamada S, Ye QW, Liu GQ, Wang C, Yang RH, Zhao L, Qin T, Liu TY, Hon HW. Suphx: Mastering mahjong with deep reinforcement learning. arXiv:2003.13590, 2020.
[12] Vinyals O, Babuschkin I, Czarnecki WM, *et al*. Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature, 2019, 575(7782): 350–354. [doi: 10.1038/s41586-019-1724-z]
[13] Berner C, Brockman G, Chan B, *et al*. Dota 2 with large scale deep reinforcement learning. arXiv:1912.06680, 2019.
[14] Wu B. Hierarchical macro strategy model for MOBA game AI. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Hawaii: AAAI, 2019. 1206–1213. [doi: 10.1609/aaai.v33i01.33011206]
[15] Boyan JA, Littman ML. Packet routing in dynamically changing networks: A reinforcement learning approach. In: Proc. of the 6th Int'l Conf. on Neural Information Processing Systems. Denver: Morgan Kaufmann Publishers Inc., 1993. 671–678.
[16] You XY, Li XJ, Xu YD, Feng H, Zhao J, Yan HC. Toward packet routing with fully distributed multiagent deep reinforcement learning. IEEE Trans. on Systems, Man, and Cybernetics: Systems, 2022, 52(2): 855–868. [doi: 10.1109/TSMC.2020.3012832]
[17] Haydari A, Yılmaz Y. Deep reinforcement learning for intelligent transportation systems: A survey. IEEE Trans. on Intelligent Transportation Systems, 2022, 23(1): 11–32. [doi: 10.1109/TITS.2020.3008612]
[18] van der Pol E, Oliehoek FA. Coordinated deep reinforcement learners for traffic light control. In: Proc. of the 30th Conf. on Neural Information Processing Systems. Barcelona: MIT Press, 2016. 21–38.
[19] Kober J, Bagnell JA, Peters J. Reinforcement learning in robotics: A survey. The Int'l Journal of Robotics Research, 2013, 32(11): 1238–1274. [doi: 10.1177/0278364913495721]
[20] Samvelyan M, Rashid T, De Witt CS, Farquhar G, Nardelli N, Rudner TGJ, Hung CM, Torr PHS, Foerster J, Whiteson S. The StarCraft multi-agent challenge. arXiv:1902.04043, 2019.
[21] Kurach K, Raichuk A, Stańczyk P, Zajac M, Bachem O, Espeholt L, Riquelme C, Vincent D, Michalski M, Bousquet O, Gelly S. Google research football: A novel reinforcement learning environment. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI, 2020. 4501–4510. [doi: 10.1609/aaai.v34i04.5878]
[22] Gardner R, Ostrom E, Walker JM. The nature of common-pool resource problems. Rationality and Society, 1990, 2(3): 335–358. [doi: 10.1177/1043463190002003005]
[23] Ostrom E. The challenge of common-pool resources. Environment: Science and Policy for Sustainable Development, 2008, 50(4): 8–21. [doi: 10.3200/ENVT.50.4.8-21]
[24] Perolat J, Leibo JZ, Zambaldi V, Beattie C, Tuyls K, Graepel T. A multi-agent reinforcement learning model of common-pool resource appropriation. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 3646–3655.
[25] Arunarani AR, Manjula D, Sugumaran V. Task scheduling techniques in cloud computing: A literature survey. Future Generation Computer Systems, 2019, 91: 407–415. [doi: 10.1016/j.future.2018.09.014]
[26] Vinitsky E, Lichtlé N, Yang XM, Amos B, Foerster J. Nocturne: A scalable driving benchmark for bringing multi-agent learning one step closer to the real world. In: Proc. of the 36th Conf. on Neural Information Processing Systems. New Orleans: MIT Press, 2022. 3962–3974.
[27] Amini A, Wang TH, Gilitschenski I, Schwarting W, Liu ZJ, Han S, Karaman S, Rus D. Vista 2.0: An open, data-driven simulator for multimodal sensing and policy learning for autonomous vehicles. In: Proc. of the 2022 Int'l Conf. on Robotics and Automation. Philadelphia: IEEE, 2022. 2419–2426. [doi: 10.1109/ICRA46639.2022.9812276]
[28] Weiss G. Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence. Cambridge: MIT Press, 1999.
[29] Shoham Y, Leyton-Brown K. Multiagent Systems: Algorithmic, Game-theoretic, and Logical Foundations. New York: Cambridge University Press, 2008.
[30] Wooldridge M. An Introduction to Multiagent Systems. West Sussex: John Wiley & Sons, 2009.
[31] Balaji PG, Srinivasan D. An introduction to multi-agent systems. In: Srinivasan D, Jain LC, eds. Innovations in Multi-agent Systems and Applications-1. Berlin: Springer, 2010. 1–27. [doi: 10.1007/978-3-642-14435-6_1]

[32] Stone P, Veloso M. Multiagent systems: A survey from a machine learning perspective. Autonomous Robots, 2000, 8(3): 345–383. [doi: 10.1023/A:1008942012299]

[33] Shoham Y, Powers R, Grenager T. If multi-agent learning is the answer, what is the question? Artificial Intelligence, 2007, 171(7): 365–377. [doi: 10.1016/j.artint.2006.02.006]

[34] Buşoniu L, Babuška R, De Schutter B. Multi-agent reinforcement learning: An overview. In: Srinivasan D, Jain LC, eds. Innovations in Multi-agent Systems and Applications-1. Berlin: Springer, 2010. 183–221. [doi: 10.1007/978-3-642-14435-6_7]

[35] Zhang KQ, Yang ZR, Başar T. Multi-agent reinforcement learning: A selective overview of theories and algorithms. In: Vamvoudakis KG, Wan Y, Lewis FL, Cansever D, eds. Handbook of Reinforcement Learning and Control. Springer, 2021. 321–384. [doi: 10.1007/978-3-030-60990-0_12]

[36] Nguyen TT, Nguyen ND, Nahavandi S. Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications. IEEE Trans. on Cybernetics, 2020, 50(9): 3826–3839. [doi: 10.1109/TCYB.2020.2977374]

[37] Hernandez-Leal P, Kaisers M, Baarslag T, De Cote EM. A survey of learning in multiagent environments: Dealing with non-stationarity. arXiv:1707.09183, 2019.

[38] Taylor ME, Stone P. Transfer learning for reinforcement learning domains: A survey. Journal of Machine Learning Research, 2009, 10: 1633–1685.

[39] Liu X, Liu SY, Zhuang YK, Gao Y. Explainable reinforcement learning: Basic problems exploration and method survey. Ruan Jian Xue Bao/Journal of Software, 2023, 34(5): 2300–2316 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6485.htm [doi: 10.13328/j.cnki.jos.006485]

[40] Hao JY, Yang TP, Tang HY, Bai CJ, Liu JY, Meng ZP, Liu P, Wang Z. Exploration in deep reinforcement learning: From single-agent to multiagent domain. IEEE Trans. on Neural Networks and Learning Systems, 2024, 35(7): 8762–8782. [doi: 10.1109/TNNLS.2023.3236361]

[41] Stefano VA, Christianos F, Schäfer L. Multi-agent Reinforcement Learning: Foundations and Modern Approaches. London: MIT Press, 2024.

[42] Hernandez-Leal P, Kartal B, Taylor ME. A survey and critique of multiagent deep reinforcement learning. Autonomous Agents and Multi-agent Systems, 2019, 33(6): 750–797. [doi: 10.1007/s10458-019-09421-1]

[43] Hoen PJ, Tuyls K, Panait L, Luke S, La Poutré JA. An overview of cooperative and competitive multiagent learning. In: Proc. of the 1st Int'l Workshop on Learning and Adaption in Multi-agent Systems. Utrecht: Springer, 2006. 1–46. [doi: 10.1007/11691839_1]

[44] Panait L, Luke S. Cooperative multi-agent learning: The state of the art. Autonomous Agents and Multi-agent Systems, 2005, 11(3): 387–434. [doi: 10.1007/s10458-005-2631-2]

[45] Busoniu L, Babuska R, De Schutter B. A comprehensive survey of multiagent reinforcement learning. IEEE Trans. on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2008, 38(2): 156–172. [doi: 10.1109/TSMCC.2007.913919]

[46] Yang YD, Wang J. An overview of multi-agent reinforcement learning from game theoretical perspective. arXiv:2011.00583, 2021.

[47] Gallo PS Jr, McClintock CG. Cooperative and competitive behavior in mixed-motive games. Journal of Conflict Resolution, 1965, 9(1): 68–78. [doi: 10.1177/002200276500900106]

[48] Kelly A. Decision Making Using Game Theory: An Introduction for Managers. Cambridge: Cambridge University Press, 2003. 1–199. [doi: 10.1017/CBO9780511609992]

[49] Tadelis S. Game Theory: An Introduction. Princeton: Princeton University Press, 2013.

[50] Rapoport A. Exploiter, leader, hero, and martyr: The four archetypes of the 2×2 game. Behavioral Science, 1967, 12(2): 81–84. [doi: 10.1002/bs.3830120202]

[51] Kollock P. Social dilemmas: The anatomy of cooperation. Annual Review of Sociology, 1998, 24(1): 183–214. [doi: 10.1146/annurev.soc.24.1.183]

[52] Hauser OP, Hilbe C, Chatterjee K, Nowak MA. Social dilemmas among unequals. Nature, 2019, 572(7770): 524–527. [doi: 10.1038/s41586-019-1488-5]

[53] Kuhn HW. Extensive games. Proc. of the National Academy of Sciences of the United States of America, 1950, 36(10): 570–576. [doi: 10.1073/pnas.36.10.570]

[54] Harsanyi JC. Games with incomplete information played by "Bayesian" players. Management Science, 1967, 14(3): 159–183.

[55] Shapley LS. Stochastic games. Proc. of the National Academy of Sciences of the United States of America, 1953, 39(10): 1095–1100. [doi: 10.1073/pnas.39.10.1095]

[56] Littman ML. Markov games as a framework for multi-agent reinforcement learning. In: Proc. of the 11th Int'l Conf. on Machine Learning. New Brunswick: Morgan Kaufmann Publishers Inc., 1994. 157–163.

[57] Owen G. Game Theory. 4th ed., Bingley: Emerald Group Publishing, 2013.

[58] Kovařík V, Schmid M, Burch N, Bowling M, Lisý V. Rethinking formal models of partially observable multiagent decision making. Artificial Intelligence, 2022, 303: 103645. [doi: 10.1016/j.artint.2021.103645]

[59] Kaelbling LP, Littman ML, Cassandra AR. Planning and acting in partially observable stochastic domains. Artificial Intelligence, 1998, 101(1–2): 99–134. [doi: 10.1016/S0004-3702(98)00023-X]

[60] Hansen EA, Bernstein DS, Zilberstein S. Dynamic programming for partially observable stochastic games. In: Proc. of the 19th National Conf. on Artificial Intelligence. San Jose: AAAI, 2004. 709–715.

[61] Zhou M, Luo J, Villella J, Villella J, *et al*. SMARTS: Scalable multi-agent reinforcement learning training school for autonomous driving. In: Proc. of the 4th Conf. on Robot Learning. Cambridge: ML Research Press, 2020. 1–22.

[62] Pontryagin LS. On the theory of differential games. Russian Mathematical Surveys, 1966, 21(4): 193–246. [doi: 10.1070/RM1966v021n04ABEH004171]

[63] Isaacs R. Differential Games: A Mathematical Theory with Applications to Warfare and Pursuit, Control and Optimization. 2nd ed., New York: Dover Publications, 1999.

[64] Friedman A. Differential Games. 2nd ed., New York: Dover Publications, 2013.

[65] Monderer D, Shapley LS. Potential games. Games and Economic Behavior, 1996, 14(1): 124–143. [doi: 10.1006/game.1996.0044]

[66] Jovanovic B, Rosenthal RW. Anonymous sequential games. Journal of Mathematical Economics, 1988, 17(1): 77–87. [doi: 10.1016/0304-4068(88)90029-8]

[67] Lasry JM, Lions PL. Mean field games. Japanese Journal of Mathematics, 2007, 2(1): 229–260. [doi: 10.1007/s11537-007-0657-8]

[68] Caines PE, Huang MY, Malhamé RP. Large population stochastic dynamic games: Closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle. Communications in Information and Systems, 2006, 6(3): 221–252. [doi: 10.4310/CIS.2006.v6.n3.a5]

[69] Smith JM, Price GR. The logic of animal conflict. Nature, 1973, 246(5427): 15–18. [doi: 10.1038/246015a0]

[70] Smith JM. Evolution and the Theory of Games. Cambridge: Cambridge University Press, 1982. 41–45. [doi: 10.1017/CBO9780511806292]

[71] Hofbauer J, Sigmund K. Evolutionary Games and Population Dynamics. Cambridge: Cambridge University Press, 1998. [doi: 10.1017/CBO9781139173179]

[72] Sandholm WH. Population Games and Evolutionary Dynamics. Cambridge: MIT Press, 2010.

[73] Taleizadeh AA, Sadeghi R. Pricing strategies in the competitive reverse supply chains with traditional and e-channels: A game theoretic approach. Int'l Journal of Production Economics, 2019, 215: 48–60. [doi: 10.1016/j.ijpe.2018.06.011]

[74] Hamilton T, Mesic R. A Simple Game-theoretic Approach to Suppression of Enemy Defenses and other Time Critical Target Analyses. Santa Monica: Research and Development Co., 2004. 1–53.

[75] Wood PJ. Climate change and game theory. Annals of the New York Academy of Sciences, 2011, 1219(1): 153–170. [doi: 10.1111/j.1749-6632.2010.05891.x]

[76] Zha DC, Xie JR, Ma WY, Zhang S, Lian XR, Hu X, Liu J. DouZero: Mastering DouDizhu with self-play deep reinforcement learning. In: Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 12333–12344.

[77] Jaderberg M, Czarnecki WM, Dunning I, Marris L, Lever G, Castañeda AG, Beattie C, Rabinowitz NC, Morcos AS, Ruderman A, Sonnerat N, Green T, Deason L, Leibo JZ, Silver D, Hassabis D, Kavukcuoglu K, Graepel T. Human-level performance in 3D multiplayer games with population-based reinforcement learning. Science, 2019, 364(6443): 859–865. [doi: 10.1126/science.aau6249]

[78] Ye DH, Chen GB, Zhang W, Chen S, Yuan B, Liu B, Chen J, Liu Z, Qiu FH, Yu HS, Yin YT, Shi B, Wang L, Shi TF, Fu Q, Yang W, Huang LX, Liu W. Towards playing full MOBA games with deep reinforcement learning. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 621–632.

[79] Nash JF Jr. Equilibrium points in *n*-person games. Proc. of the National Academy of Sciences of the United States of America, 1950, 36(1): 48–49. [doi: 10.1073/pnas.36.1.48]

[80] Howard N. The Theory of Meta-games. Dordrecht: Springer, 1974. 167–186.

[81] Rapoport A. Prisoner's Dilemma. London: Palgrave Macmillan, 1989. 199–204.

[82] Kuhn HW, Tucker AW. Contributions to the Theory of Games. Princeton: Princeton University Press, 1953.

[83] Kajii A, Morris S. The robustness of equilibria to incomplete information. Econometrica, 1997, 65(6): 1283–1309. [doi: 10.2307/2171737]

[84] Fudenberg D, Tirole J. Perfect Bayesian equilibrium and sequential equilibrium. Journal of Economic Theory, 1991, 53(2): 236–260. [doi: 10.1016/0022-0531(91)90155-W]

[85] Aumann RJ. Subjectivity and correlation in randomized strategies. Journal of Mathematical Economics, 1974, 1(1): 67–96. [doi: 10.1016/0304-4068(74)90037-8]

[86] Aumann RJ. Correlated equilibrium as an expression of Bayesian rationality. Econometrica, 1987, 55(1): 1–18. [doi: 10.2307/1911154]

[87] Moulin H, Vial JP. Strategically zero-sum games: The class of games whose completely mixed equilibria cannot be improved upon. Int'l Journal of Game Theory, 1978, 7(3–4): 201–221. [doi: 10.1007/BF01769190]

[88] Daskalakis C, Goldberg PW, Papadimitriou CH. The complexity of computing a Nash equilibrium. Communications of the ACM, 2009, 52(2): 89–97. [doi: 10.1145/1461928.1461951]

[89] Chen X, Deng XT, Teng SH. Settling the complexity of computing two-player Nash equilibria. Journal of the ACM, 2009, 56(3): 1–57. [doi: 10.1145/1516512.1516516]

[90] Schopenhauer A. The World as Will and Idea. 5th ed., London: Kegan Paul, Trench, Trübner & Co., 1906.

[91] Weigel RH, Hessing DJ, Elffers H. Egoism: Concept, measurement and implications for deviance. Psychology, Crime and Law, 1999, 5(4): 349–378. [doi: 10.1080/10683169908401777]

[92] Becker GS. Altruism, egoism, and genetic fitness: Economics and sociobiology. Journal of Economic Literature, 1976, 14(3): 817–826.

[93] Bentham J. Of the principle of utility. 1994. https://www.blackwellpublishing.co.uk/content/BPL_Images/Content_store/Sample_chapter/9780631233510/Warnock.pdf

[94] Bentham J. Utilitarianism. London: Progressive Publishing Company, 1890. 1–29.

[95] Dutta B, Ray D. A concept of egalitarianism under participation constraints. Econometrica, 1989, 57(3): 615–635. [doi: 10.2307/1911055]

[96] Moulin H. Welfare bounds in the fair division problem. Journal of Economic Theory, 1991, 54(2): 321–337. [doi: 10.1016/0022-0531(91)90125-N]

[97] Fleurbaey M, Maniquet F. A Theory of Fairness and Social Welfare. New York: Cambridge University Press, 2011. [doi: 10.1017/CBO9780511851971]

[98] Coleman JS. The possibility of a social welfare function. The American Economic Review, 1966, 56: 1105–1122.

[99] Kaneko M, Nakamura K. The Nash social welfare function. Econometrica, 1979, 47(2): 423–435. [doi: 10.2307/1914191]

[100] Speicher T, Heidari H, Grgic-Hlaca N, Gummadi KP, Singla A, Weller A, Zafar MB. A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. In: Proc. of the 24th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. London: ACM, 2018. 2239–2248. [doi: 10.1145/3219819.3220046]

[101] Heidari H, Ferrari C, Gummadi KP, Krause A. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 1273–1283.

[102] Filar J, Vrieze K. Competitive Markov Decision Processes. 2nd ed., New York: Springer Science & Business Media, 2012.

[103] Filar JA, Schultz TA, Thuijsman F, Vrieze OJ. Nonlinear programming and stationary equilibria in stochastic games. Mathematical Programming, 1991, 50(1–3): 227–237. [doi: 10.1007/BF01594936]

[104] Berger U. Brown's original fictitious play. Journal of Economic Theory, 2007, 135(1): 572–578. [doi: 10.1016/j.jet.2005.12.010]

[105] Fudenberg D, Levine DK. The Theory of Learning in Games. Cambridge: MIT Press, 1998.

[106] Fudenberg D, Kreps DM. Learning mixed equilibria. Games and Economic Behavior, 1993, 5(3): 320–367. [doi: 10.1006/game.1993.1021]

[107] Fudenberg D, Levine DK. Consistency and cautious fictitious play. Journal of Economic Dynamics and Control, 1995, 19(5–7): 1065–1089. [doi: 10.1016/0165-1889(94)00819-4]

[108] Hofbauer J, Sandholm WH. On the global convergence of stochastic fictitious play. Econometrica, 2002, 70(6): 2265–2294. [doi: 10.1111/1468-0262.00376]

[109] Swenson B, Poor HV. Smooth fictitious play in $N\times2$ potential games. In: Proc. of the 53rd Asilomar Conf. on Signals, Systems, and Computers. Pacific Grove: IEEE, 2019. 1739–1743. [doi: 10.1109/IEEECONF44664.2019.9048995]

[110] Leslie DS, Collins EJ. Generalised weakened fictitious play. Games and Economic Behavior, 2006, 56(2): 285–298. [doi: 10.1016/j.geb.2005.08.005]

[111] Heinrich J, Lanctot M, Silver D. Fictitious self-play in extensive-form games. In: Proc. of the 32nd Int'l Conf. on Machine Learning. Lille: JMLR.org, 2015. 805–813.

[112] McMahan HB, Gordon GJ, Blum A. Planning in the presence of cost functions controlled by an adversary. In: Proc. of the 20th Int'l Conf. on Machine Learning. Washington: AAAI, 2003. 536–543.

[113] Adam L, Horčík R, Kasl T, Kroupa T. Double oracle algorithm for computing equilibria in continuous games. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. AAAI, 2021. 5070–5077. [doi: 10.1609/aaai.v35i6.16641]

[114] Dinh LC, McAleer S, Tian Z, Perez-Nieves N, Slumbers O, Mguni DH, Wang J, Bou Ammar H, Yang YD. Online double oracle. Trans. on Machine Learning Research, 2022, 10(1): 2835–8856.

[115] Jain M, Korzhyk D, Vaněk O, Conitzer V, Pěchouček M, Tambe M. A double oracle algorithm for zero-sum security games on graphs. In: Proc. of the 10th Int'l Conf. on Autonomous Agents and Multiagent Systems. Taipei: Int'l Foundation for Autonomous Agents and

Multiagent Systems, 2011. 327–334.

[116] Tsai J, Nguyen TH, Tambe M. Security games for controlling contagion. In: Proc. of the 26th AAAI Conf. on Artificial Intelligence. Toronto: AAAI, 2012. 1464–1470.

[117] Bošanský B, Kiekintveld C, Lisý V, Pěchouček M. An exact double-oracle algorithm for zero-sum extensive-form games with imperfect information. Journal of Artificial Intelligence Research, 2014, 51(1): 829–866.

[118] Lanctot M, Zambaldi V, Gruslys A, Lazaridou A, Tuyls K, Pérolat J, Silver D, Graepel T. A unified game-theoretic approach to multiagent reinforcement learning. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 4193–4206.

[119] McAleer S, Lanier J, Wang K, Baldi P, Fox R. XDO: A double oracle algorithm for extensive-form games. In: Proc. of the 34th Conf. on Neural Information Processing Systems. MIT Press, 2021. 23128–23139.

[120] Hazan E. Introduction to online convex optimization. Foundations and Trends® in Optimization, 2016, 2(3–4): 157–325. [doi: 10.1561/2400000013]

[121] Beck A, Teboulle M. Mirror descent and nonlinear projected subgradient methods for convex optimization. Operations Research Letters, 2003, 31(3): 167–175. [doi: 10.1016/S0167-6377(02)00231-6]

[122] Kalai A, Vempala S. Efficient algorithms for online decision problems. Journal of Computer and System Sciences, 2005, 71(3): 291–307. [doi: 10.1016/j.jcss.2004.10.016]

[123] Abernethy J, Hazan E, Rakhlin A. Competing in the dark: An efficient algorithm for bandit linear optimization. In: Proc. of the 21st Annual Conf. on Learning Theory. Helsinki: Springer, 2008. 263–273.

[124] Farina G, Kroer C, Sandholm T. Optimistic regret minimization for extensive-form games via dilated distance-generating functions. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 5221–5231.

[125] Hoda S, Gilpin A, Peña J, Sandholm T. Smoothing techniques for computing Nash equilibria of sequential games. Mathematics of Operations Research, 2010, 35(2): 494–512. [doi: 10.1287/moor.1100.0452]

[126] Kroer C, Waugh K, Kılınç-Karzan F, Sandholm T. Faster algorithms for extensive-form game solving via improved smoothing functions. Mathematical Programming, 2020, 179(1–2): 385–417. [doi: 10.1007/s10107-018-1336-7]

[127] Zinkevich M, Johanson M, Bowling M, Piccione C. Regret minimization in games with incomplete information. In: Proc. of the 20th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2007. 1729–1736.

[128] Hart S, Mas-Colell A. A simple adaptive procedure leading to correlated equilibrium. Econometrica, 2000, 68(5): 1127–1150. [doi: 10.1111/1468-0262.00153]

[129] Neller TW, Lanctot M. An introduction to counterfactual regret minimization. In: Proc. of the 2013 Symp. on Educational Advances in Artificial Intelligence (EAAI 2013). Washington: AAAI, 2013. 1–38.

[130] Tammelin O. Solving large imperfect information games using CFR+. arXiv:1407.5042, 2014.

[131] Brown N, Sandholm T. Solving imperfect-information games via discounted regret minimization. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Honolulu: AAAI, 2019. 1829–1836. [doi: 10.1609/aaai.v33i01.33011829]

[132] Farina G, Kroer C, Sandholm T. Faster game solving via predictive blackwell approachability: Connecting regret matching and mirror descent. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. AAAI, 2021: 5363–5371. [doi: 10.1609/aaai.v35i6.16676]

[133] Liu WM, Jiang HC, Li B, Li HQ. Equivalence analysis between counterfactual regret minimization and online mirror descent. In: Proc. of the 39th Int'l Conf. on Machine Learning. Baltimore: PMLR, 2022. 13717–13745.

[134] Cesa-Bianchi N, Lugosi G. Prediction, Learning, and Games. New York: Cambridge University Press, 2006. 1–394. [doi: 10.1017/CBO9780511546921]

[135] Rakhlin S, Sridharan K. Optimization, learning, and games with predictable sequences. In: Proc. of the 26th Int'l Conf. on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2013. 3066–3074.

[136] Syrgkanis V, Agarwal A, Luo HP, Schapire RE. Fast convergence of regularized learning in games. In: Proc. of the 28th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2015. 2989–2997.

[137] Chen X, Peng BH. Hedging in games: Faster convergence of external and swap regrets. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 18990–18999.

[138] Gibson R, Burch N, Lanctot M, Szafron D. Efficient Monte Carlo counterfactual regret minimization in games with many player actions. In: Proc. of the 25th Int'l Conf. on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2012. 1880–1888.

[139] Lanctot M, Waugh K, Zinkevich M, Bowling M. Monte Carlo sampling for regret minimization in extensive games. In: Proc. of the 22nd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2009. 1078–1086.

[140] Schmid M, Burch N, Lanctot M, Moravcik M, Kadlec R, Bowling M. Variance reduction in monte carlo counterfactual regret minimization (VR-MCCFR) for extensive form games using baselines. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence.

Honolulu: AAAI, 2019. 2157–2164. [doi: 10.1609/aaai.v33i01.33012157]

[141] Burch N. Time and space: Why imperfect information games are hard [Ph.D. Thesis]. Edmonton: University of Alberta, 2017.

[142] Li H, Hu KL, Zhang SH, Qi Y, Song L. Double neural counterfactual regret minimization. In: Proc. of the 2020 Int'l Conf. on Learning Representations. Virtual: Curran Associates Inc., 2020. 1–13.

[143] Johanson M, Burch N, Valenzano R, Bowling M. Evaluating state-space abstractions in extensive-form games. In: Proc. of the 2013 Int'l Conf. on Autonomous Agents and Multi-agent Systems. St. Paul: Int'l Foundation for Autonomous Agents and Multiagent Systems, 2013. 271–278.

[144] Brown N, Lerer A, Gross S, Sandholm T. Deep counterfactual regret minimization. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 793–802.

[145] Steinberger E. Single deep counterfactual regret minimization. arXiv:1901.07621, 2019.

[146] Bellman R. A Markovian decision process. Journal of Mathematics and Mechanics, 1957, 6(5): 679–684.

[147] Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, Petersen S, Beattie C, Sadik A, Antonoglou I, King H, Kumaran D, Wierstra D, Legg S, Hassabis D. Human-level control through deep reinforcement learning. Nature, 2015, 518(7540): 529–533. [doi: 10.1038/nature14236]

[148] Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. arXiv:1707.06347, 2017.

[149] Boutilier C. Planning, learning and coordination in multiagent decision processes. In: Proc. of the 6th Conf. on Theoretical Aspects of Rationality and Knowledge. The Netherlands: Morgan Kaufmann Publishers Inc., 1996. 195–210.

[150] Bernstein DS, Givan R, Immerman N, Zilberstein S. The complexity of decentralized control of Markov decision processes. Mathematics of Operations Research, 2002, 27(4): 819–840. [doi: 10.1287/moor.27.4.819.297]

[151] Oliehoek FA, Amato C. A Concise Introduction to Decentralized POMDPs. Cham: Springer, 2016. [doi: 10.1007/978-3-319-28929-8]

[152] Claus C, Boutilier C. The dynamics of reinforcement learning in cooperative multiagent systems. In: Proc. of the 1998 AAAI Conf. on Artificial Intelligence. Madison: AAAI, 1998. 746–752.

[153] Tan M. Multi-agent reinforcement learning: Independent versus cooperative agents. In: Proc. of the 10th Int'l Conf. on Machine Learning. Amherst: ACM, 1993. 330–337.

[154] Oliehoek FA, Spaan MTJ, Vlassis N. Optimal and approximate Q-value functions for decentralized POMDPs. Journal of Artificial Intelligence Research, 2008, 32(1): 289–353.

[155] Kar S, Moura JMF, Poor HV. QD-learning: A collaborative distributed strategy for multi-agent reinforcement learning through consensus + innovations. IEEE Trans. on Signal Processing, 2013, 61(7): 1848–1862. [doi: 10.1109/TSP.2013.2241057]

[156] Macua SV, Chen JS, Zazo S, Sayed AH. Distributed policy evaluation under multiple behavior strategies. IEEE Trans. on Automatic Control, 2015, 60(5): 1260–1274. [doi: 10.1109/TAC.2014.2368731]

[157] Zhang KQ, Yang ZR, Liu H, Zhang T, Basar T. Fully decentralized multi-agent reinforcement learning with networked agents. In: Proc. of the 35th Int'l Conf. on Machine Learning. Stockholm: ACM, 2018. 5872–5881.

[158] Kok JR, Vlassis N. Sparse cooperative Q-learning. In: Proc. of the 21st Int'l Conf. on Machine Learning. Banff: ACM, 2004. 1–8. [doi: 10.1145/1015330.1015410]

[159] Wen MN, Kuba JG, Lin RJ, Zhang WN, Wen Y, Wang J, Yang YD. Multi-agent reinforcement learning is a sequence modeling problem. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 16509–16521.

[160] Wunder M, Littman M, Babes M. Classes of multiagent Q-learning dynamics with ε-greedy exploration. In: Proc. of the 27th Int'l Conf. on Machine Learning. Haifa: Omnipress, 2010. 1167–1174.

[161] Tampuu A, Matiisen T, Kodelja D, Kuzovkin I, Korjus K, Aru J, Aru J, Vicente R. Multiagent cooperation and competition with deep reinforcement learning. PLoS ONE, 2017, 12(4): e0172395. [doi: 10.1371/journal.pone.0172395]

[162] de Witt CS, Gupta T, Makoviichuk D, Makoviychuk V, Torr PHS, Sun MF, Whiteson S. Is independent learning all you need in the StarCraft multi-agent challenge? arXiv:2011.09533, 2020.

[163] Jiang JC, Lu ZQ. I2Q: A fully decentralized Q-learning algorithm. In: Proc. of the 36th Conf. on Neural Information Processing Systems. New Orleans: MIT Press, 2022. 20469–20481.

[164] Edwards AD, Sahni H, Liu R, Hung J, Jain A, Wang R, Ecoffet A, Miconi T, Isbell C, Yosinski J. Estimating $Q(s,s')$ with deep deterministic dynamics gradients. In: Proc. of the 37th Int'l Conf. on Machine Learning. JMLR.org, 2020. 2825–2835.

[165] He H, Boyd-Graber J, Kwok K, Daumé H III. Opponent modeling in deep reinforcement learning. In: Proc. of the 33rd Int'l Conf. on Machine Learning. New York: JMLR.org, 2016. 1804–1813.

[166] Hong ZW, Su SY, Shann TY, Chang YH, Lee CY. A deep policy inference Q-network for multi-agent systems. In: Proc. of the 17th Int'l Conf. on Autonomous Agents and Multiagent Systems. Stockholm: Int'l Foundation for Autonomous Agents and Multiagent Systems,

2018. 1388–1396.

[167] Raileanu R, Denton E, Szlam A, Fergus R. Modeling others using oneself in multi-agent reinforcement learning. In: Proc. of the 35th Int'l Conf. on Machine Learning. Stockholm: ACM, 2018. 4257–4266.

[168] Wen Y, Yang YD, Luo R, Wang J, Pan W. Probabilistic recursive reasoning for multi-agent reinforcement learning. In: Proc. of the 7th Int'l Conf. on Learning Representations. New Orleans: Curran Associates Inc., 2019. 1–13.

[169] Tian Z, Wen Y, Gong ZC, Punakkath F, Zou SH, Wang J. A regularized opponent model with maximum entropy objective. In: Proc. of the 28th Int'l Joint Conf. on Artificial Intelligence. Macao: AAAI, 2019. 602–608.

[170] Wen Y, Yang YD, Wang J. Modelling bounded rationality in multi-agent interactions by generalized recursive reasoning. In: Proc. of the 29th Int'l Joint Conf. on Artificial Intelligence. Yokohama: Morgan Kaufmann, 2021. 414–421.

[171] McKee KR, Hughes E, Zhu TO, Chadwick MJ, Koster R, Castañeda AG, Beattie C, Graepel T, Botvinick M, Leibo JZ. A multi-agent reinforcement learning model of reputation and cooperation in human groups. arXiv:2103.04982, 2023.

[172] Leibo JZ, Zambaldi V, Lanctot M, Marecki J, Graepel T. Multi-agent reinforcement learning in sequential social dilemmas. In: Proc. of the 16th Int'l Conf. on Autonomous Agents and Multiagent Systems. São Paulo: Int'l Foundation for Autonomous Agents and Multiagent Systems, 2017. 464–473.

[173] Anastassacos N, Hailes S, Musolesi M. Partner selection for the emergence of cooperation in multi-agent systems using reinforcement learning. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI, 2020. 7047–7054. [doi: 10.1609/aaai.v34i05.6190]

[174] Hughes E, Leibo JZ, Phillips M, Tuyls K, Dueñez-Guzman E, Castañeda AG, Dunning I, Zhu TN, McKee K, Koster R, Roff H, Graepel T. Inequity aversion improves cooperation in intertemporal social dilemmas. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 3330–3340.

[175] Jaques N, Lazaridou A, Hughes E, Gulcehre C, Ortega PA, Strouse DJ, Leibo JZ, de Freitas N. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: ACM, 2019. 3040–3049.

[176] Baker B. Emergent reciprocity and team formation from randomized uncertain social preferences. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 15786–15799.

[177] Eccles T, Hughes E, Kramár J, Wheelwright S, Leibo JZ. Learning reciprocity in complex sequential social dilemmas. arXiv:1903.08082, 2019.

[178] Yang JC, Li A, Farajtabar M, Sunehag P, Hughes E, Zha HY. Learning to incentivize other learning agents. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 15208–15219.

[179] McKee KR, Gemp I, McWilliams B, Duéñez-Guzmán EA, Hughes E, Leibo JZ. Social diversity and social preferences in mixed-motive reinforcement learning. In: Proc. of the 19th Int'l Conf. on Autonomous Agents and Multiagent Systems. Auckland: Int'l Foundation for Autonomous Agents and Multiagent Systems, 2020. 869–877.

[180] Merhej R, Santos FP, Melo FS, Santos FC. Cooperation between independent reinforcement learners under wealth inequality and collective risks. In: Proc. of the 20th Int'l Conf. on Autonomous Agents and Multiagent Systems. Int'l Foundation for Autonomous Agents and Multiagent Systems, 2021. 898–906.

[181] Chevaleyre Y, Dunne P, Endriss U, Lang J, Lemaître M, Maudet N, Padget J, Phelps S, Rodríguez-Aguilar JA, Sousa P. Issues in multiagent resource allocation. Informatica, 2006, 30(3): 3–31.

[182] Hao JY, Leung HF. Fairness in cooperative multiagent systems. Interactions in Multiagent Systems: Fairness, Social Optimality and Individual Rationality. Berlin: Springer, 2016. 27–70. [doi: 10.1007/978-3-662-49470-7_3]

[183] Zhang CJ, Shah JA. Fairness in multi-agent sequential decision-making. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2014. 2636–2644.

[184] Jiang JC, Lu ZQ. Learning fairness in multi-agent systems. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 13854–13865.

[185] Zimmer M, Glanois C, Siddique U, Weng P. Learning fair policies in decentralized cooperative multi-agent reinforcement learning. In: Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 12967–12978.

[186] Papoudakis G, Christianos F, Schäfer L, Albrecht SV. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. In: Proc. of the 35th Conf. on Neural Information Processing Systems. MIT Press, 2021. 1–13.

[187] Sunehag P, Lever G, Gruslys A, Czarnecki WM, Zambaldi V, Jaderberg M, Lanctot M, Sonnerat N, Leibo JZ, Tuyls K, Graepel T. Value-decomposition networks for cooperative multi-agent learning based on team reward. In: Proc. of the 17th Int'l Conf. on Autonomous Agents and Multiagent Systems. Stockholm: Int'l Foundation for Autonomous Agents and Multiagent Systems, 2018. 2085–2087.

[188] Rashid T, Samvelyan M, De Witt CS, Farquhar G, Foerster J, Whiteson S. Monotonic value function factorisation for deep multi-agent reinforcement learning. The Journal of Machine Learning Research, 2020, 21(1): 7234–7284.

[189] Rashid T, Farquhar G, Peng B, Whiteson S. Weighted QMIX: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning. In: Proc. of the 34th Int'l Conf. on Neural Information Processing System. Vancouver: Curran Associates Inc., 2020. 10199–10210.

[190] Son K, Kim D, Kang WJ, Hostallero DE, Yi Y. QTRAN: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 5887–5896.

[191] Wang JH, Ren ZZ, Liu T, Yu Y, Zhang CJ. QPLEX: Duplex dueling multi-agent Q-learning. In: Proc. of the 9th Int'l Conf. on Learning Representations. Curran Associates Inc., 2021. 1–11.

[192] Foerster JN, Farquhar G, Afouras T, Nardelli N, Whiteson S. Counterfactual multi-agent policy gradients. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. New Orleans: AAAI, 2018. 2974–2982. [doi: 10.1609/aaai.v32i1.11794]

[193] Chen WB, Li WB, Liu X, Yang SD, Gao Y. Learning explicit credit assignment for cooperative multi-agent reinforcement learning via polarization policy gradient. In: Proc. of the 37th AAAI Conf. on Artificial Intelligence. Washington: AAAI, 2023. 11542–11550. [doi: 10.1609/aaai.v37i10.26364]

[194] Lowe R, Wu Y, Tamar A, Harb J, Abbeel P, Mordatch I. Multi-agent actor-critic for mixed cooperative-competitive environments. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6382–6393.

[195] Yu C, Velu A, Vinitsky E, Gao JX, Wang Y, Bayen A, Wu Y. The surprising effectiveness of PPO in cooperative multi-agent games. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 24611–24624.

[196] Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, Silver D, Wierstra D. Continuous control with deep reinforcement learning. arXiv:1509.02971, 2019.

[197] Mordatch I, Abbeel P. Emergence of grounded compositional language in multi-agent populations. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. New Orleans: AAAI, 2018. 1495–1502. [doi: 10.1609/aaai.v32i1.11492]

[198] Lazaridou A, Peysakhovich A, Baroni M. Multi-agent cooperation and the emergence of (natural) language. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: Curran Associates Inc., 2017. 1–11.

[199] Foerster JN, Assael YM, De Freitas N, Whiteson S. Learning to communicate with deep multi-agent reinforcement learning. In: Proc. of the 30th Int'l Conf. on Neural Information Processing Systems. Barcelona, 2016. 2145–2153.

[200] Sukhbaatar S, Szlam A, Fergus R. Learning multiagent communication with backpropagation. In: Proc. of the 30th Int'l Conf. on Neural Information Processing Systems. Barcelona, 2016. 2252–2260.

[201] Singh A, Jain T, Sukhbaatar S. Learning when to communicate at scale in multiagent cooperative and competitive tasks. In: Proc. of the 7th Int'l Conf. on Learning Representations. New Orleans: Curran Associates Inc., 2019. 1–16.

[202] Peng P, Wen Y, Yang YD, Yuan Q, Tang ZK, Long HT, Wang J. Multiagent bidirectionally-coordinated nets: Emergence of human-level coordination in learning to play starcraft combat games. arXiv:1703.10069, 2017.

[203] Zhang KQ, Yang ZR, Basar T. Networked multi-agent reinforcement learning in continuous spaces. In: Proc. of the 2018 IEEE Conf. on Decision and Control. Miami: IEEE, 2018. 2771–2776. [doi: 10.1109/CDC.2018.8619581]

[204] Zhang KQ, Yang ZR, Liu H, Zhang T, Başar T. Finite-sample analysis for decentralized batch multiagent reinforcement learning with networked agents. IEEE Trans. on Automatic Control, 2021, 66(12): 5925–5940. [doi: 10.1109/TAC.2021.3049345]

[205] Riedmiller M. Neural fitted Q iteration—First experiences with a data efficient neural reinforcement learning method. In: Proc. of the 16th European Conf. on Machine Learning. Porto: Springer, 2005. 317–328. [doi: 10.1007/11564096_32]

[206] Littman ML. Value-function reinforcement learning in Markov games. Cognitive Systems Research, 2001, 2(1): 55–66. [doi: 10.1016/S1389-0417(01)00015-8]

[207] Wang XF, Sandholm T. Reinforcement learning to play an optimal Nash equilibrium in team Markov games. In: Proc. of the 15th Int'l Conf. on Neural Information Processing Systems. Vancouver: MIT Press, 2002. 1603–1610.

[208] Hu JL, Wellman MP. Multiagent reinforcement learning: Theoretical framework and an algorithm. In: Proc. of the 15th Int'l Conf. on Machine Learning. Madison: ACM, 1998. 242–250.

[209] Littman ML. Friend-or-foe Q-learning in general-sum games. In: Proc. of the 18th Int'l Conf. on Machine Learning. Williamstown: ACM, 2001. 322–328.

[210] Greenwald A, Hall K, Serrano R. Correlated-Q learning. In: Proc. of the 20th Int'l Conf. on Machine Learning. Washington: ACM, 2003. 242–249.

[211] Weinberg M, Rosenschein JS. Best-response multiagent learning in non-stationary environments. In: Proc. of the 3rd Int'l Conf. on Autonomous Agents and Multiagent Systems. New York: IEEE, 2004. 506–513.

[212] Hu YJ, Gao Y, An B. Multiagent reinforcement learning with unshared value functions. IEEE Trans. on Cybernetics, 2015, 45(4):

647–662. [doi: 10.1109/TCYB.2014.2332042]

[213] Singh S, Kearns M, Mansour Y. Nash convergence of gradient dynamics in general-sum games. In: Proc. of the 16th Conf. on Uncertainty in Artificial Intelligence. Stanford: Morgan Kaufmann Publishers Inc., 2000. 541–548.

[214] Bowling M, Veloso M. Multiagent learning using a variable learning rate. Artificial Intelligence, 2002, 136(2): 215–250. [doi: 10.1016/S0004-3702(02)00121-2]

[215] Zinkevich M. Online convex programming and generalized infinitesimal gradient ascent. In: Proc. of the 20th Int'l Conf. on Machine Learning. Washington: ACM, 2003. 928–935.

[216] Bowling M. Convergence and no-regret in multiagent learning. In: Proc. of the 17th Int'l Conf. on Neural Information Processing Systems. Vancouver: MIT Press, 2004. 209–216.

[217] Ratliff LJ, Burden SA, Sastry SS. Characterization and computation of local Nash equilibria in continuous games. In: Proc. of the 51st Annual Allerton Conf. on Communication, Control, and Computing. Monticello: IEEE, 2013. 917–924. [doi: 10.1109/Allerton.2013.6736623]

[218] Mazumdar E, Ratliff LJ, Sastry SS. On gradient-based learning in continuous games. SIAM Journal on Mathematics of Data Science, 2020, 2(1): 103–131. [doi: 10.1137/18M1231298]

[219] Mertikopoulos P, Zhou ZY. Learning in games with continuous action sets and unknown payoff functions. Mathematical Programming, 2019, 173(1–2): 465–507. [doi: 10.1007/s10107-018-1254-8]

[220] Banerjee B, Peng J. Adaptive policy gradient in multiagent learning. In: Proc. of the 2nd Int'l Conf. on Autonomous Agents and Multiagent Systems. Melbourne: ACM, 2003. 686–692. [doi: 10.1145/860575.860686]

[221] Heinrich J, Silver D. Deep reinforcement learning from self-play in imperfect-information games. arXiv:1603.01121, 2016.

[222] Tuyls K, Perolat J, Lanctot M, Leibo JZ, Graepel T. A generalised method for empirical game theoretic analysis. In: Proc. of the 17th Int'l Conf. on Autonomous Agents and Multiagent Systems. Stockholm: Int'l Foundation for Autonomous Agents and Multiagent Systems, 2018. 77–85.

[223] Balduzzi D, Garnelo M, Bachrach Y, Czarnecki WM, Perolat J, Jaderberg M, Graepel T. Open-ended learning in symmetric zero-sum games. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 434–443.

[224] Omidshafiei S, Papadimitriou C, Piliouras G, Tuyls K, Rowland M, Lespiau JB, Czarnecki WM, Lanctot M, Perolat J, Munos R. α-Rank: Multi-agent evaluation by evolution. Scientific Reports, 2019, 9(1): 9937. [doi: 10.1038/s41598-019-45619-9]

[225] Muller P, Omidshafiei S, Rowland M, Tuyls K, Pérolat J, Liu SQ, Hennes D, Marris L, Lanctot M, Hughes E, Wang Z, Lever G, Heess N, Graepel T, Munos R. A generalized training approach for multiagent learning. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: Curran Associates Inc., 2020. 1–35.

[226] Foerster JN, Chen RY, Al-Shedivat M, Whiteson S, Abbeel P, Mordatch I. Learning with opponent-learning awareness. In: Proc. of the 17th Int'l Conf. on Autonomous Agents and Multiagent Systems. Stockholm: Int'l Foundation for Autonomous Agents and Multiagent Systems, 2018. 122–130.

[227] Zheng Y, Meng ZP, Hao JY, Zhang ZZ, Yang TP, Fan CJ. A deep Bayesian policy reuse approach against non-stationary agents. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 962–972.

[228] Hernandez-Leal P, Taylor ME, Rosman B, Sucar LE, de Cote EM. Identifying and tracking switching, non-stationary opponents: A Bayesian approach. In: Proc. of the 30th AAAI Conf. on Artificial Intelligence. Phoenix: AAAI, 2016. 560–566.

[229] Yang YD, Luo R, Li MN, Zhou M, Zhang WN, Wang J. Mean field multi-agent reinforcement learning. In: Proc. of the 35th Int'l Conf. on Machine Learning. Stockholm: ACM, 2018. 5571–5580.

[230] Ostrom E, Burger J, Field CB, Norgaard RB, Policansky D. Revisiting the commons: Local lessons, global challenges. Science, 1999, 284(5412): 278–282. [doi: 10.1126/science.284.5412.278]

[231] Dietz T, Ostrom E, Stern PC. The struggle to govern the commons. Science, 2003, 302(5652): 1907–1912. [doi: 10.1126/science.1091015]

[232] Leibo JZ, Perolat J, Hughes E, Wheelwright S, Marblestone AH, Duéñez-Guzmán E, Sunehag P, Dunning I, Graepel T. Malthusian reinforcement learning. In: Proc. of the 18th Int'l Conf. on Autonomous Agents and Multiagent Systems. Montreal: Int'l Foundation for Autonomous Agents and Multiagent Systems, 2019. 1099–1107.

[233] Leibo JZ, Hughes E, Lanctot M, Graepel T. Autocurricula and the emergence of innovation from social interaction: A manifesto for multi-agent intelligence research. arXiv:1903.00742, 2019.

[234] McKee KR, Leibo JZ, Beattie C, Everett R. Quantifying the effects of environment and population diversity in multi-agent reinforcement learning. Autonomous Agents and Multi-agent Systems, 2022, 36(1): 21. [doi: 10.1007/S10458-022-09548-8]

[235] Henrich J, Chudek M, Boyd R. The Big Man Mechanism: How prestige fosters cooperation and creates prosocial leaders. Philosophical Trans. of the Royal Society B: Biological Sciences, 2015, 370(1683): 20150013. [doi: 10.1098/rstb.2015.0013]

[236] Gächter S, Renner E. Leaders as role models and 'belief managers' in social dilemmas. Journal of Economic Behavior & Organization, 2018, 154: 321–334. [doi: 10.1016/j.jebo.2018.08.001]

[237] Wu J, Cao J, Wang CJ, Xie JY. Social law synthesizing method based on algorithmic mechanism design. Ruan Jian Xue Bao/Journal of Software, 2024, 35(3): 1440–1465 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6825.htm [doi: 10.13328/j.cnki.jos.006825]

[238] Vinitsky E, Köster R, Agapiou JP, Duéñez-Guzmán EA, Vezhnevets AS, Leibo JZ. A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings. Collective Intelligence, 2023, 2(2): 1–14. [doi: 10.1177/26339137231162025]

[239] Köster R, Hadfield-Menell D, Everett R, Weidinger L, Hadfield GK, Leibo JZ. Spurious normativity enhances learning of compliance and enforcement behavior in artificial agents. Proc. of the National Academy of Sciences of the United States of America, 2022, 119(3): e2106028118. [doi: 10.1073/PNAS.2106028118]

[240] Jiang D, Yuan Y, Zhang XW, Wang GR. Survey on data pricing and trading research. Ruan Jian Xue Bao/Journal of Software, 2023, 34(3): 1396–1424 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6751.htm [doi: 10.13328/j.cnki.jos.006751]

[241] Allen F, Morris S. Finance applications of game theory. Cowles Foundation Discussion Papers. 1998. https://elischolar.library.yale.edu/cowles-discussion-paper-series/1443

[242] Carfì D, Musolino F. Fair redistribution in financial markets: A game theory complete analysis. Journal of Advanced Studies in Finance, 2011, 2(2): 74–100.

[243] Bi HL, Chen YJ, Yi XJ, Wang X. Game-based user decision optimization analysis of cryptocurrency trading market. Ruan Jian Xue Bao/Journal of Software, 2023, 34(12): 5477–5500 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6798.htm [doi: 10.13328/j.cnki.jos.006798]

[244] Liang TX, Yang XP, Wang L, Han ZY. Review on financial trading system based on reinforcement learning. Ruan Jian Xue Bao/Journal of Software, 2019, 30(3): 845–864 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5689.htm [doi: 10.13328/j.cnki.jos.005689]

[245] Fisac JF, Bronstein E, Stefansson E, Sadigh D, Sastry SS, Dragan AD. Hierarchical game-theoretic planning for autonomous vehicles. In: Proc. of the 2019 Int'l Conf. on Robotics and Automation. Montreal: IEEE, 2019. 9590–9596. [doi: 10.1109/ICRA.2019.8794007]

[246] Ge ZX, Yang SD, Tian PZ, Chen ZX, Gao Y. Modeling rationality: Toward better performance against unknown agents in sequential games. IEEE Trans. on Cybernetics, 2024, 54(5): 2966–2977. [doi: 10.1109/TCYB.2022.3228812]

[247] Cui BD, Hu HY, Pineda L, Foerster JN. K-level reasoning for zero-shot coordination in Hanabi. In: Proc. of the 35th Conf. on Neural Information Processing Systems. MIT Press, 2021. 8215–8228.

[248] Schwarting W, Pierson A, Alonso-Mora J, Karaman S, Rus D. Social behavior for autonomous vehicles. Proc. of the National Academy of Sciences of the United States of America, 2019, 116(50): 24972–24978. [doi: 10.1073/pnas.1820676116]

[249] Zha DC, Lai KH, Cao YP, Huang SY, Wei RZ, Guo JY, Hu X. RLCard: A toolkit for reinforcement learning in card games. arXiv:1910.04376, 2020.

[250] Zheng LM, Yang JC, Cai H, Zhou M, Zhang WN, Wang J, Yu Y. MAgent: A many-agent reinforcement learning platform for artificial collective intelligence. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. New Orleans: AAAI, 2018. 8222–8223. [doi: 10.1609/aaai.v32i1.11371]

[251] Leibo JZ, Dueñez-Guzman EA, Vezhnevets AS, Agapiou JP, Sunehag P, Koster R, Matyas J, Beattie C, Mordatch I, Graepel T. Scalable evaluation of multi-agent reinforcement learning with melting pot. In: Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 6187–6199.

[252] Suarez J, Du YL, Isola P, Mordatch I. Neural MMO: A massively multiagent game environment for training and evaluating intelligent agents. arXiv:1903.00784, 2019.

[253] Hu YJ, Gao Y, An B. Learning in multi-agent systems with sparse interactions by knowledge transfer and game abstraction. In: Proc. of the 2015 Int'l Conf. on Autonomous Agents and Multiagent Systems. Istanbul: Int'l Foundation for Autonomous Agents and Multiagent Systems, 2015. 753–761.

[254] Huang ZG, Liu Q, Zhang LH, Cao JQ, Zhu F. Research and development on deep hierarchical reinforcement learning. Ruan Jian Xue Bao/Journal of Software, 2023, 34(2): 733–760 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6706.htm [doi: 10.13328/j.cnki.jos.006706]

[255] Guestrin C, Lagoudakis MG, Parr R. Coordinated reinforcement learning. In: Proc. of the 19th Int'l Conf. on Machine Learning. Sydney: ACM, 2002. 227–234.

[256] Böhmer W, Kurin V, Whiteson S. Deep coordination graphs. In: Proc. of the 37th Int'l Conf. on Machine Learning. JMLR.org, 2020. 980–991.

[257] Hu YJ, Gao Y, An B. Online counterfactual regret minimization in repeated imperfect information extensive games. Journal of

Computer Research and Development, 2014, 51(10): 2160–2170 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2014.20130823]

[258] Zhang MY, Jin Z, Liu K. Counterfactual regret advantage-based self-play approach for mixed cooperative-competitive multi-agent systems. Ruan Jian Xue Bao/Journal of Software, 2024, 35(2): 739–757 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6832.htm [doi: 10.13328/j.cnki.jos.006832]

[259] Zhang JW, Lü S, Zhang ZH, Yu JY, Gong XY. Survey on deep reinforcement learning methods based on sample efficiency optimization. Ruan Jian Xue Bao/Journal of Software, 2022, 33(11): 4217–4238 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6391.htm [doi: 10.13328/j.cnki.jos.006391]

[260] Pislar M, Szepesvari D, Ostrovski G, Borsa DL, Schaul T. When should agents explore? In: Proc. of the 10th Int'l Conf. on Learning Representations. Curran Associates Inc., 2022.

[261] Dong SK, Mao HY, Yang SD, Zhu SY, Li WB, Hao JY, Gao Y. WToE: Learning when to explore in multiagent reinforcement learning. IEEE Trans. on Cybernetics, 2024, 54(8): 4789–4801. [doi: 10.1109/TCYB.2023.3328732]

[262] Wang TH, Wang JH, Wu Y, Zhang CJ. Influence-based multi-agent exploration.In: Proc. of the 8th Int'l Conf. on Learning Representations. Curran Associates Inc., 2020. 1–14.

[263] Levine S, Kumar A, Tucker G, Fu J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. arXiv:2005.01643, 2020.

[264] Cao HY, Yang SD, Huo J, Chen XG, Gao Y. Enhancing OOD generalization in offline reinforcement learning with energy-based policy optimization. Frontiers in Artificial Intelligence and Applications, 2023, 372: 335–342. [doi: 10.3233/FAIA230288]

[265] Ghosh D, Rahme J, Kumar A, Zhang A, Adams RP, Levine S. Why generalization in RL is difficult: Epistemic POMDPs and implicit partial observability. In: Proc. of the 35th Conf. on Neural Information Processing Systems. MIT Press, 2021. 25502–25515.

[266] Korkmaz E. A survey analyzing generalization in deep reinforcement learning. arXiv:2401.02349, 2024.

[267] Wang KX, Kang BY, Shao J, Feng JS. Improving generalization in reinforcement learning with mixture regularization. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 7968–7978.

[268] Laskin M, Srinivas A, Abbeel P. CURL: Contrastive unsupervised representations for reinforcement learning. In: Proc. of the 37th Int'l Conf. on Machine Learning. PMLR, 2020. 5639–5650.

[269] Kostrikov I, Yarats D, Fergus R. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In: Proc. of the 9th Int'l Conf. on Learning Representations. Curran Associates Inc., 2021. 1–12.

[270] Igl M, Ciosek K, Li YZ, Tschiatschek S, Zhang C, Devlin S, Hofmann K. Generalization in reinforcement learning with selective noise injection and information bottleneck. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver, 2019. 13979–13991.

[271] Liu Z, Li XL, Kang BY, Darrell T. Regularization matters in policy optimization. In: Proc. of the 9th Int'l Conf. on Learning Representations. Curran Associates Inc., 2021. 1–12.

[272] Gleave A, Dennis M, Wild C, Kant N, Levine S, Russell S. Adversarial policies: Attacking deep reinforcement learning. In: Proc. of the 8th Int'l Conf. on LearningRepresentations. Curran Associates Inc., 2020. 1–11.

[273] Korkmaz E. Adversarial robust deep reinforcement learning requires redefining robustness. In: Proc. of the 37th AAAI Conf. on Artificial Intelligence. Washington: AAAI, 2023. 8369–8377. [doi: 10.1609/aaai.v37i7.26009]

[274] Puiutta E, Veith EMSP. Explainable reinforcement learning: A survey. In: Proc. of the 4th Int'l Cross-domain Conf. on Machine Learning and Knowledge Extraction. Dublin: Springer, 2020. 77–95. [doi: 10.1007/978-3-030-57321-8_5]

[275] Wells L, Bednarz T. Explainable AI and reinforcement learning—A systematic review of current approaches and trends. Frontiers in Artificial Intelligence, 2021, 4: 550030. [doi: 10.3389/FRAI.2021.550030]

附中文参考文献:

[39] 刘潇, 刘书洋, 庄韫恺, 高阳. 强化学习可解释性基础问题探索和方法综述. 软件学报, 2023, 34(5): 2300–2316. http://www.jos.org.cn/1000-9825/6485.htm [doi: 10.13328/j.cnki.jos.006485]

[237] 吴骏, 曹杰, 王崇骏, 谢俊元. 一种基于算法机制设计的社会法则合成方法. 软件学报, 2024, 35(3): 1440–1465. http://www.jos.org.cn/1000-9825/6825.htm [doi: 10.13328/j.cnki.jos.006825]

[240] 江东, 袁野, 张小伟, 王国仁. 数据定价与交易研究综述. 软件学报, 2023, 34(3): 1396–1424. http://www.jos.org.cn/1000-9825/6751.htm [doi: 10.13328/j.cnki.jos.006751]

[243] 毕红亮, 陈艳姣, 伊心静, 汪旭. 基于博弈的加密货币交易市场用户决策优化分析. 软件学报, 2023, 34(12): 5477–5500. http://www.jos.org.cn/1000-9825/6798.htm [doi: 10.13328/j.cnki.jos.006798]

[244] 梁天新, 杨小平, 王良, 韩镇远. 基于强化学习的金融交易系统研究与发展. 软件学报, 2019, 30(3): 845–864. http://www.jos.org.cn/1000-9825/5689.htm [doi: 10.13328/j.cnki.jos.005689]

[254] 黄志刚, 刘全, 张立华, 曹家庆, 朱斐. 深度分层强化学习研究与发展. 软件学报, 2023, 34(2): 733–760. http://www.jos.org.cn/1000-9825/6706.htm [doi: 10.13328/j.cnki.jos.006706]

[257] 胡裕靖, 高阳, 安波. 不完美信息扩展式博弈中在线虚拟遗憾最小化. 计算机研究与发展, 2014, 51(10): 2160–2170. [doi: 10.7544/issn1000-1239.2014.20130823]

[258] 张明悦, 金芝, 刘坤. 合作-竞争混合型多智能体系统的虚拟遗憾优势自博弈方法. 软件学报, 2024, 35(2): 739–757. http://www.jos.org.cn/1000-9825/6832.htm [doi: 10.13328/j.cnki.jos.006832]

[259] 张峻伟, 吕帅, 张正昊, 于佳玉, 龚晓宇. 基于样本效率优化的深度强化学习方法综述. 软件学报, 2022, 33(11): 4217–4238. http://www.jos.org.cn/1000-9825/6391.htm [doi: 10.13328/j.cnki.jos.006391]

董绍康(1996－), 男, 博士, CCF 学生会员, 主要研究领域为多智能体强化学习, 博弈论, 强化学习的探索与利用.

陈武兵(1996－), 男, 博士生, 主要研究领域为多智能体强化学习, 信度分配, 离线强化学习.

李超(1996－), 男, 博士生, CCF 学生会员, 主要研究领域为多智能体强化学习, 任务建模.

杨尚东(1990－), 男, 博士, 讲师, CCF 专业会员, 主要研究领域为强化学习, 强化学习的探索与利用.

杨光(1994－), 男, 博士生, 主要研究领域为多智能体强化学习, 强化学习值函数估计, 状态抽象.

陈兴国(1984－), 男, 博士, 讲师, CCF 专业会员, 主要研究领域为机器学习, 强化学习, 深度学习, 智能博弈.

葛振兴(1998－), 男, 博士生, CCF 学生会员, 主要研究领域为算法博弈论, 对手建模.

李文斌(1991－), 男, 博士, 副研究员, CCF 专业会员, 主要研究领域为机器学习, 元学习, 持续学习.

曹宏业(1998－), 男, 博士生, CCF 学生会员, 主要研究领域为强化学习的可解释性, 因果强化学习.

高阳(1972－), 男, 博士, 教授, CCF 杰出会员, 主要研究领域为博弈论, 多智能体强化学习, 机器学习.