控制与决策
Control and Decision

计算机博弈中序贯不完美信息博弈求解研究进展

罗俊仁, 张万鹏, 苏炯铭, 魏婷婷, 陈璟

引用本文:
罗俊仁,张万鹏,苏炯铭,魏婷婷,陈. 计算机博弈中序贯不完美信息博弈求解研究进展[J]. 控制与决策, 2023, 38(10): 2721−2748.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2022.0698

多无人机协同直播场景下自适应任务卸载决策

Adaptive task offloading decision of multi−UAVs cooperation in live broadcasting scenario

控制与决策. 2021, 36(4): 974−982 https://doi.org/10.13195/j.kzyjc.2019.1104

计算机博弈中序贯不完美信息博弈求解研究进展

罗俊仁, 张万鹏, 苏炯铭, 魏婷婷, 陈 璟†

(国防科技大学 智能科学学院，长沙 410073)

摘 要: 计算机博弈是人工智能的果蝇和通用测试基准. 近年来,序贯不完美信息博弈求解一直是计算机博弈研究领域的前沿课题. 围绕计算机博弈中不完美信息博弈求解问题展开综述分析. 首先,梳理计算机博弈领域标志性突破的里程碑事件,简要介绍4类新评估基准,归纳3种研究范式,提出序贯不完美信息博弈求解研究框架;然后,着重对序贯不完美信息博弈的博弈模型和解概念进行调研,从博弈构建、子博弈和元博弈、解概念以及评估3方面进行简要介绍;接着,围绕离线策略求解,系统梳理算法博弈论、优化理论和博弈学习3大类方法,围绕在线策略求解,系统梳理对手近似式学习、对手判别式适变和对手生成式搜索3大类方法;最后,从环境、智能体(对手)和策略求解3个角度分析面临的挑战,从博弈动力学和策略空间理论、多模态对抗博弈和序贯建模、通用策略学习和离线预训练、对手建模(剥削)和反剥削、临机组队和零样本协调5方面展望未来研究前沿课题. 对于当前不完美信息博弈求解问题进行全面概述,期望能够为人工智能和博弈论领域相关研究带来启发.

关键词: 计算机博弈；不完美信息博弈；扩展式博弈；反事实后悔最小化；在线凸优化；无悔学习；对手建模

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2022.0698

引用格式: 罗俊仁, 张万鹏, 苏炯铭, 等. 计算机博弈中序贯不完美信息博弈求解研究进展 [J]. 控制与决策, 2023, 38(10): 2721-2748.

Research progress on sequential imperfect information game solving in computer games

LUO Jun-ren, ZHANG Wan-peng, SU Jiong-ming, WEI Ting-ting, CHEN Jing†

(College of Intelligence Science and Technology，National University of Defense Technology，Changsha 410073，China)

Abstract: Computer games are the drosophilae and universal benchmarks for artificial intelligence. In recent years, sequential imperfect information game solving has always been a frontier topic in the field of computer games research. Therefore, a comprehensive analysis of the imperfect information games solving problem in computer games is carried out. Firstly, it sorts out the milestones of the landmark breakthroughs in the field of computer games, briefly introduces four new evaluation benchmarks, summarizes three research paradigms, and deeply analyzes the challenges faced by games with imperfect information. Secondly, it focuses on investigating the game model and solution concept of a sequential imperfect information games, and briefly introduces it from three aspects: game formulation, sub-game and meta game, solution concept and evaluation. The offline strategy solving methods are systematically sorted from three perspectives of algorithmic game theory, optimization theory, and game theoretic learning. The online strategy solving methods are systematically sorted from three perspectives of opponent approximate learning, opponent discriminant adaptation, and opponent generative search. Finally, the challenges faced are analyzed from three perspectives: environment, agent (opponent) and strategy solving. The future research frontiers and prospects are given from five aspects: game dynamics and strategy space theory, multi-modal adversarial game and sequential modeling, general strategy learning and offline pretrain, opponent modeling (exploitation) and anti-exploitation, ad-hoc teamwork and zero-shot coordination. This paper provides a comprehensive overview of current imperfect information game solving, hoping to inspire related research in the field of aritificial intelligence and game theory.

Keywords: computer games；imperfect information game；extensive form game；counterfactual regret minimization；online convex optimization；no-regret learning；opponent modeling

收稿日期: 2022-04-26；**录用日期:** 2022-09-20.

基金项目: 国家自然科学基金项目(61806212)；湖南省研究生科研创新项目(CX20210011).

责任编委: 王光臣.

†通讯作者. E-mail: chenjing001@vip.sina.com.

*本文附带电子附录文件,可登录本刊官网该文“资源附件”区自行下载阅览.

0 引 言

计算机博弈(computer games)也称机器博弈,英语直译为计算机游戏,覆盖类型十分广泛. 传统的计算机博弈主要包括完美信息博弈(如跳棋[1]、国际象棋[2]、围棋[3]等)和不完美信息博弈(如德州扑克[4]、桥牌[5]等). 早前,研究人员将国际象棋视为“人工智能的果蝇(drosophila)”[6],将扑克作为人工智能领域的“测试平台(testbed)”[7]. 近年来,一些更加复杂的计算机博弈已然成为人工智能的新果蝇和通用测试基准,特别是2019年以来,以Alphastar(星际争霸II AI)[8]、Pluribus(德州扑克AI)[9]等为代表的人工智能程序在多类典型人机对抗比赛中战胜人类职业高手,人工智能技术在不完美信息博弈领域取得了显著突破. 当前以算法博弈论、深度强化学习和在线凸优化为基础的智能博弈对抗技术的发展,已完全能够解决围棋等两人零和完美信息博弈,而星际争霸和德州扑克等不完美信息博弈所面临的难题正牵引着人工智能领域的前沿研究.

不完美信息博弈常被用于描述多方序贯策略交互过程. 博弈对抗过程中,局中人无法获得全部或准确的信息,博弈状态和下步动作通常是不可知的,掌握的信息往往也是不对称和不完美的. 相关技术不仅可应用于量化投资[10]、拍卖机制设计[11]等社会经济生产领域,且可广泛应用于冲突分析[12]、国土安全防卫[13]和智能指控等国防军事应用领域[14]. 特别是美国国防部高级研究计划局(DARPA)先后启动了“打破游戏规则”(gamebreaker)[15]人工智能探索项目,旨在开发开放世界视频游戏的国防战争模拟游戏的人工智能程序,用于作战人员的实战训练;“面向复杂军事决策的非完美信息博弈序贯交互”(SI3-CMD)[16]项目,旨在探索自动化地利用呈指数增长的数据信息,将复杂系统的建模与推理相结合,从而辅助国防部快速认识、理解甚至预测复杂国际和军事环境中的重要事件.

近年来,面向计算机博弈的序贯不完美信息博弈求解相关研究陆续取得新突破,来自国内外不同研究组的大量前沿探索性研究从不同的视角对博弈的核心问题进行了探究. 在国内,徐心和等[17]就计算机博弈面临的各种挑战进行了分析,王亚杰等[18]重点阐述了多类博弈树搜索技术,吴军等[19]分析了3类基于随机博弈的多智能体强化学习方法,王军等[20]从博弈理论视角梳理了多智能体强化学习方法,黄凯奇等[21]提出了以博弈学习为核心的人机对抗智能理论研究框架以及关键模型,程代展等[22]利用矩阵的半张量积分析了有限博弈的向量空间分解,唐振韬等[23]围绕实时格斗游戏的智能决策方法进行了综述分析,袁唯淋等[24]梳理了以德州扑克为代表的计算机扑克智能博弈的典型算法以及关键技术. 综合而言,这些综述的内容与本文有较大不同. 本文围绕序贯不完美信息计算机博弈展开,将博弈理论基础、人工智能技术与计算机博弈紧密耦合在一起分析,力争对最新研究成果作全面归纳总结,对该领域的前沿挑战问题作探索性分析.

1 计算机博弈简介

自1943年McCulloch等[25]提出“人工神经元”网络模型,1958年Rosenblatt[26]提出“感知机”模型,1989年LeCun等[27]提出总卷积神经网络,2006年Hinton等[28]提出深度学习,2014年Goodfellow等[29]提出GAN模型,2020年OpenAI发布GPT-3预训练模

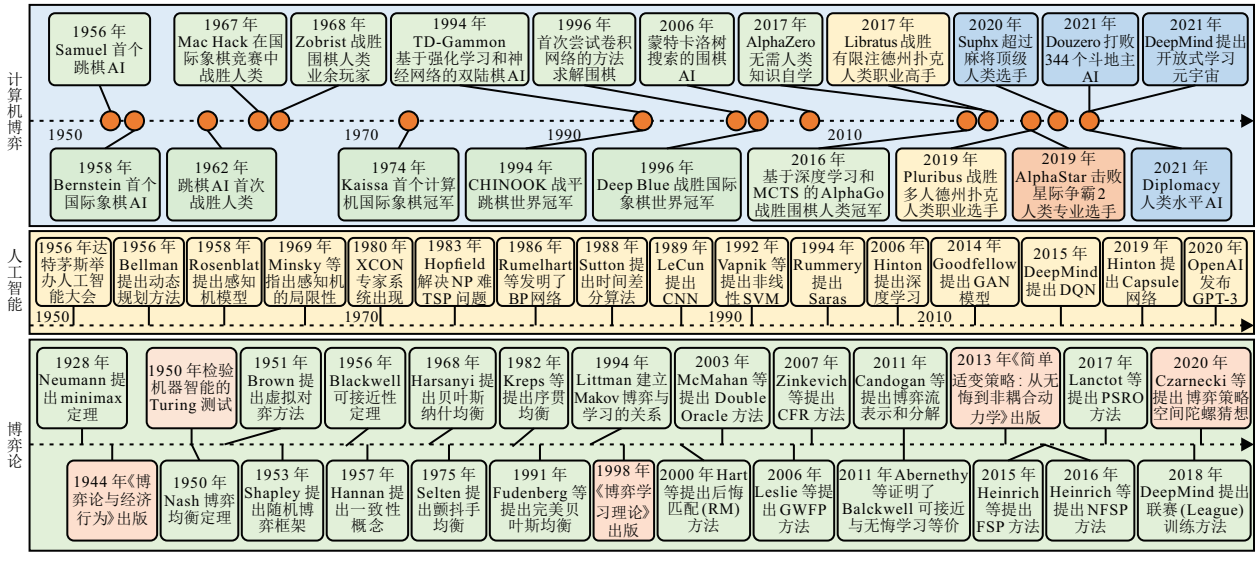

图 1 计算机博弈、人工智能和博弈论发展历史

型[30], 人工智能经历了“推理期”(定理证明、逻辑语言)、“知识期”(知识系统、专家系统兴起)和“学习期”(神经网络重新流行、统计机器学习兴起, 深度学习兴起), 期间经历 2 度寒冬 3 次复兴. 如图 1 所示, 作为博弈论研究的实验平台, 伴随人工智能近 70 年的发展, 从最初的“图灵测试”到最新的“通用人工智能”挑战, 计算机博弈的相关研究呈现出“知识与搜索”“博弈与学习”“模型与适变”3 大范式. 下面结合序贯不完美信息博弈领域的里程碑事件与博弈论领域的开创性研究, 简要介绍 3 项标志性突破工作以及 4 类新评估基准, 梳理 3 种研究范式, 提出所研究框架.

1.1　标志性突破

人工智能技术与博弈论的交叉融合发展为计算机博弈, 特别是不完美信息博弈的相关研究提供了重要基础. 从最初的 2 人对抗(如国际象棋[2]、围棋[3]、2 人德州扑克[31])到多人或团队对抗(如 6 人德州扑克[9]、麻将[32]、斗地主[33]、桥牌[5]、外交(diplomacy)[34]、花火(hanabi)[35]、刀塔 2[36]、星际争霸 II[8]、王者荣耀[37]), 不完美信息博弈人工智能 AI 的开发已然取得了比较大的突破, 几类典型计算机博弈相关特征如表 1 所示. 本文聚焦序贯不完美信息博弈求解, 不具体介绍格斗、刀塔 2、星际争霸 II 和王者荣耀等即时策略类计算机博弈.

1.1.1　国际象棋

1997 年, IBM 的 Deep Blue[2] 程序 AI 利用基于人类知识的启发式评估函数和强大的博弈树搜索方法在国际象模比赛中击败了当时的世界冠军 Kasparov, 将人工智能带回了公众视野.

表 1　典型计算机博弈特征

计算机博弈	典型 AI (算法)	策略交互	博弈类型	技术方案
国际象棋	Deep Blue[2]	序贯交替	2 人完美信息	局面评估 + 启发式搜索
围棋	AlphaGo[3]	序贯交替	2 人完美信息	策略和价值网络 + 蒙特卡洛树搜索
2 人德州扑克	DeepStack[31]	序贯交替	2 人不完美信息	反事实值估计 + 在线子博弈深度受限搜索
6 人德州扑克	Pluribus[9]	序贯交替	6 人 (1 对 5) 不完美信息	蓝图策略 + 在线子博弈深度受限搜索
麻将	Suphx[32]	序贯交替	4 人 (1 对多) 不完美信息	深度强化学习 + Oracle 引导
斗地主	DouZero[33]	序贯交替	3 人 (2 对 1) 不完美信息	深度蒙特卡洛 + 分布式强化学习
桥牌	JPS 搜索[5]	序贯交替	4 人 (2 对 2) 不完美信息	联合策略搜索 + 双明手
外交	SearchBot[34]	序贯交替	7 人 (合作-对抗) 不完美信息	监督学习 + 均衡搜索
花火	K-level 推理[35]	序贯交替	多人合作不完美信息	零样本协同 + 它对弈 (other play)
格斗	RHEAOM[38]	即时响应	2 人 (1 对 1) 不完美信息	滚动时域演化 + 对手建模
刀塔 2	OpenAI Five[36]	即时响应	2 团队 (5 对 5) 不完美信息	大规模强化学习 + 迁移
星际争霸 II	AlphaStar[8]	即时响应	2 人 (1 对 1) 不完美信息	监督学习 + 联赛训练
王者荣耀	绝悟[37]	即时响应	1 对 1 或 5 对 5 不完美信息	课程自对弈 + 学习编组

1.1.2　围　棋

2016 年, DeepMind 的 AlphaGo[3] 程序 AI 采用了 2 个单独的深度神经网络, 策略网络用于选择下一步动作, 价值网络用于预测赢率, 自对弈训练过程中采用蒙特卡洛树搜索. 在围棋比赛中先后击败韩国李世石和中国柯洁. 基于 AlphaGo 的相关研究 (不依赖人类先验知识的 AlphaZero[39]、将规划纳入学习的 MuZero[40] 等) 正式宣告完美信息博弈领域取得了重大突破. 相关研究在国内得到了多个研究机构关注, 南京大学的陈兴国等[41]、中科院自动化所唐振韬等[42]、国防大学郭圣明等[43] 均围绕围棋相关技术展开了分析.

1.1.3　德州扑克

2017 年, 阿尔伯塔大学的 Bowling 团队开发的 DeepStack[31] 程序采用深度神经网络学习反事实值和子博弈深度受限搜索, 在 2 人无限注德州扑克中首次击败人类职业选手. 2018 年, 卡内基梅隆大学的 Sandholm 团队开发的 Libratus[44] 程序采用蓝图策略求解、蓝图策略自我强化、在线子博弈安全嵌套求解方法, 在 2 人无限注德州扑克中首次击败人类顶尖职业选手. 此外, Sandholm 团队于 2019 年开发的 Pluribus[9] 程序采用蓝图策略求解和在线子博弈深度受限搜索方法, 在 6 人无限注德州扑克中首次击败人类顶尖职业选手.

1.2　新评估基准

1.2.1　斗地主、麻将和多人零和博弈

关于斗地主: Tan 等[45] 基于蒙特卡洛树搜索设计了斗地主赢率预测模型, Jiang 等[46] 利用虚拟自对弈方法构建了一个专家级斗地主 AI, Zha 等[33] 利用蒙特卡洛搜索和分布式强化学习方法构建了斗地主 AI, 李淑琴等[47] 围绕同等牌力问题展开了研究. 关于麻将: 微软亚洲研究院的 Li 等[32] 利用 Oracle 引导深度强化学习, 设计了面向日本麻将的 Suphx 程序 AI; Fu 等[48] 构建了 1 对 1 麻将的博弈模型, 提出了 ACH

策略优化方法;王亚杰等[49]结合先验知识并利用蒙特卡洛模拟设计了内陆版麻将AI. 虽然当前一些研究围绕“斗地主和麻将”等展开了探索,但是相关博弈理论研究相对滞后,如何构建可利用当前通用博弈求解技术的相应多人零和博弈模型仍然充满挑战.

1.2.2 多人德州扑克、桥牌和对抗团队博弈

传统的2人德州扑克是一类典型的不完美信息博弈,而关于多人德州扑克的相关AI实践(如Pluribus[9])虽然取得了突破,但其相关博弈理论的研究仍在探索. 其中:Pluribus首先利用德州扑克规则进行信息抽象和行动抽象,离线计算粗略的抽象博弈蓝图策略,基于蓝图策略组合,在线对抗时通过有限深度子博弈求解精细的抽象博弈策略. 关于桥牌,Tian[5]等设计了联合策略搜索方法;基于对抗团队博弈(adversarial team game, ATG),Sonzogni[50]设计了有限深度搜索方法;Derin[51]设计了策略精炼方法. 其中,对抗团队博弈可用于一类多对1博弈对抗问题的建模. 根据团队成员间3种不同的协同关系,相关博弈均衡解也有3种. 若团队成员对抗过程中可自由地私下交流,则可将团队当作单个博弈局中人,对应的均衡解为团队纳什均衡,可采用传统的CFR或线性规划方法求解;若团队成员无法交流,则团队策略需要使得团队期望收益最大化,对应的均衡解为团队极大极小均衡(team maxmin equilibrium, TME)[52];若团队成员可在博弈对抗开始前私下秘密协商战术,则对抗过程中无法显式交流,对应为带协调的团队极大极小均衡(TME with coordination device, TMECor)[53]. Zhang等[54]提出利用公共信息表示[55]构建了基于团队信念的有向图序贯决策模型. Carminati等[56]试图构建满足对抗团队博弈求解的博弈抽象、无悔学习和子博弈求解等关键技术.

1.2.3 外交和一般和博弈

外交是一类7人(合作-对抗)桌游,为典型的不完美信息博弈,在每个回合局中人可选择联合或背叛来获取更多资源. Paquette等[57]提出利用图网络和领域知识,运用监督学习、自博弈强化学习方法进行策略学习,为该领域相关研究提供了基线AI. Gray等[34]首先利用人类对抗数据训练得到蓝图策略,在线对抗时采用基于后悔最小化的一步前瞻方法进行均衡策略搜索. Anthony等[58]基于元博弈理论,采用了对抗种群的虚拟对弈策略迭代方法. 对于多人对抗博弈中允许非合作、非对抗关系存在的情形,可建模为一般和博弈(general sum game). Jacob等[59]结合正则化学习与无悔学习提出了满足输出强且类人策略的正则化搜索方法. Gemp等[60]基于偏离动机梯度和平摊估计设计了基于梯度下降的均衡近似方法. Marris等[61]设计了基于相关均衡的元博弈求解器,可用于一般和博弈的求解.

1.2.4 花火和多人合作博弈

花火(hanabi)[62]是一类需要多人(2～6人)协作的不完美信息合作游戏,局中人只能看到其他局中人手里的牌(看不到自己的牌),需要采用心智推理等方式从队友处获得信息,从而提高得分. 如何设计适当的显式或隐式交流方法、推理队友的状态是当前此类的博弈多人合作博弈(multi-player cooperative game)的重要关注点. Lerer等[63]提出利用回溯式信念更新和基于模仿学习的自举式策略搜索方法. Nekoei等[64]提出了面向连续协调问题的终生强化学习方法. Hu等[65]利用信念学习研究了零样本协调(zero shot coordination, ZSC)问题. 此外,Siu等[66]利用学习类和规划类智能体评估了当前人与AI组队的效果,指出规划类智能体表现更好.

1.3 研究范式和框架

早在1956年,Blackwell[67]证明了可接近(approachability)定理,后被广泛应用于在线凸优化领域;1957年,Hannnan[68]根据无名氏定理(folk theorem)提出了Hannan一致性问题,后被广泛应用于无悔(no regret)学习方法,研究表明,一般的无悔学习方法可收敛至相关均衡;1998年,Fundengerg等[69]出版了《博弈理论学习》,对虚拟对弈(fictitious play)、复制动态(replicator dynamics)等进行了总结;2007年,Zinkevich等[70]提出了反事实后悔最小化方法,开启了利用后悔值求解序贯不完美信息博弈的新途径;2013年,Hart等[71]出版了《简单适变策略:从无悔到非耦合动力学》,指出非耦合动力学指引下的迭代式学习可能无法收敛至纳什均衡. 根据博弈论的发展历史可梳理出计算机博弈的3种研究范式,如表2所示.

表2 计算机博弈研究范式

研究范式	假设	理论基础	解的性质	方法学
知识＋搜索	完全理性、公共知识	博弈论和最优化理论	理性、最优性	局面评估＋启发式搜索
博弈＋学习	不完全理性、无公共知识	博弈论和博弈动力学	收敛性、无悔性	博弈动力学＋迭代式学习
模型＋适变	非平稳性	博弈论和机制设计	安全性、鲁棒性	离线模型求解＋在线适变

1.3.1　知识和搜索

早期的一些研究采用了博弈论中完全理性个体假设, 主要采用专家知识辅助局面评估, 通过设计启发式引导博弈树搜索 (如 min-max 搜索、alpha-beta 剪枝) 最优解. 由于计算能力不足, 巨复杂的状态空间无法抽象、简单化, min-max 搜索需要遍历所有节点, alpha-beta 仅剪掉了部分次优分支, 该范式能够解决的问题规模十分有限, 但是这类方法仍然能够为当前一些研究工作的开展提供指引.

1.3.2　博弈和学习

博弈理论学习的研究放松了完全理性假设, 转而研究博弈动力学 (特别是耦合动力学), 个体能否通过迭代式学习方法收敛至博弈均衡解. 早在 1951 年 Brown[72] 提出了利用迭代式的虚拟对弈求解博弈. 早期的一些研究将强化学习方法与 min-max 相结合提出了 minimax-Q[73]、Nash-Q[74] 方法. 近年来, 借助深度学习的表征能力, 一些新方法相继被提出, 如神经虚拟自对弈 (neural fictitious self play, NFSP)[75]、神经复制动态[76] 等. 此外, 非耦合动力学的相关研究放松了公共知识假设, 个体仅通过己方收益和他人策略 (而非收益) 调整策略, 相关均衡或扩展式相关均衡已然成为可选解概念. 研究人员基于 “事后理性” (hindsight rationality) 分析了多种偏离类型 (deviation types)[77], 提出了多种迭代式无悔学习方法[70]. 该范式围绕博弈均衡解概念展开, 是当前博弈论与学习方法结合比较紧密的研究方向, 相关方法具有收敛性保证.

1.3.3　模型和适变

多人博弈对抗场景中, 由于环境 (对手) 的非平稳性 (non-stationary), 多均衡解面临的均衡选择问题, 使得解概念和博弈学习过程表现出 “离线耦合、在线解耦” 的状态. 根据策略求解所处的阶段, 在离线训练过程中, 可采用仿真环境模拟与对手的交互, 采样数据包含敌我双方的耦合状态, 利用大规模计算方式得到蓝图策略或利用分布式强化学习方法得到预训练模型; 在线对抗过程中, 由于仅能控制己方策略, 策略的生成处于解耦合状态, 需要采用适变的反制策略应对对手. 当前, 相关的离线模型研究有: Brown 等[44] 设计的德州扑克 AI 主要依赖强大的离线蓝图策略, 来自 DeepMind 的 Mathieu 等[78] 针对 Starcraft II 发布的离线预训练模型. 此外, 相关在线适变策略的研究包括: Ganzfried 等[79] 提出的安全策略, Johanson 等[80] 提出的鲁棒策略等.

通过回顾标志性突破和新评估基准, 基于梳理的典型研究范式, 提出序贯不完美信息博弈求解研究框架, 如图 2 所示.

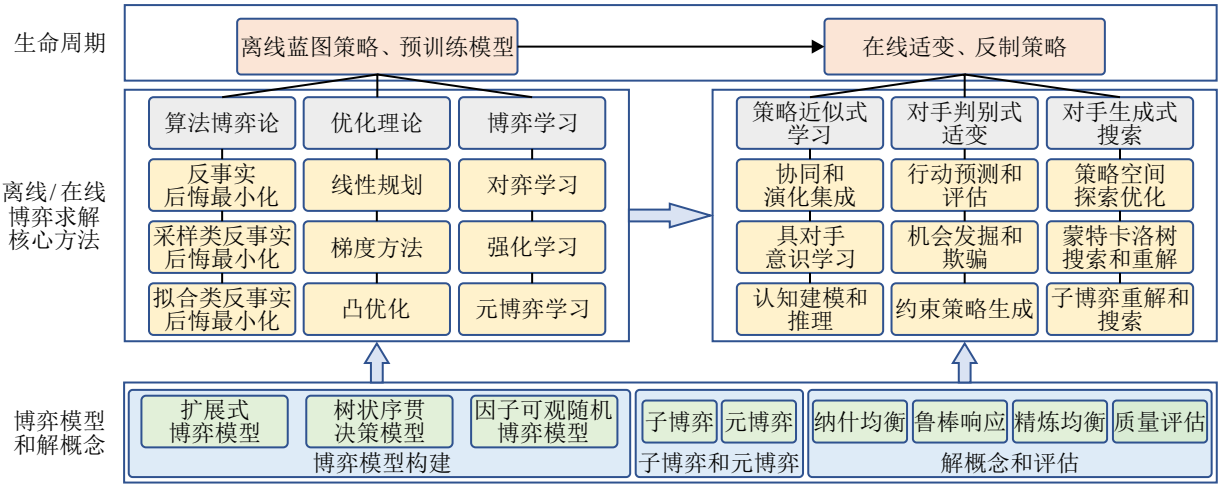

图 2　序贯不完美信息博弈求解研究框架

2　博弈模型和解概念

2.1　博弈模型构建

一个博弈模型通常包含 5 要素: 局中人 (智能体) 也称博弈方, 即博弈中独自参与决策并承担结果的决策者; 行动, 即局中人在某个决策点的决策变量; 信息, 即局中人掌握的博弈知识, 包括博弈环境、局中人的理性程度、行动和策略等; 策略, 即局中人相继决策的行动组织; 收益, 即根据特定的局中人策略组合 (strategy profile) 得出博弈结果后各局中人所获得的收益或支付. 对于常见的 2 人单次 (one shot) 博弈, 即正则式博弈或矩阵博弈, 其模型包含: 局中人集合 $N=\{1,2\}$; 局中人合法行动集 $A_1 \times A_2$; 局中人策略 $X \in \Delta(A_1), Y \in \Delta(A_2)$; 收益矩阵 $U_{|A_1|\times|A_2|}$. 本文主要研究的序贯不完美信息博弈表现出部分可观多阶段序贯交互状态, 常见的模型建模方法有 3 大类, 如表 3 所示.

表 3 序贯不完美信息博弈的3种模型比较

特点	EFGs	TFSDPs	FOSGs
任务角色	局中人	局中人, 智能体	局中人, 智能体
行动规则	策略 (strategy)	策略 (strategy)	策略 (policy, strategy)
任务时长	博弈深度	博弈深度	有限时域
环境状态	交互历史	交互历史	全局状态
客观发生	交互历史	交互历史	交互历史、轨迹
决策信息	信息集	观测信息	行动观测、交互历史
即时反馈	无	观测信息	观测信息
优化目标	收益	收益、后悔	收益、累积回报
不确定性	机会、对手	机会、对手	状态变换概率

2.1.1 扩展式博弈模型

扩展式博弈(extensive form games, EFGs)模型采用博弈树来描述序贯交互过程. 由于信息不完美, 博弈局中人无法直接观察和区分博弈的确切状态, 这些状态组成了信息集.

定义 1 扩展式博弈模型可表示为一个8元组$(H, Z, A, N, p, \pi_c, U, I)$, 其中:

1) H为交互历史集, 表示行动序列.

2) Z含结束状态的历史交互集合, 即博弈树的叶子节点, $z \in H$不是任何其他历史交互的前缀, $g \sqsubseteq h$表示g为h的前缀.

3) 对于非结束状态历史交互$h \in H \setminus Z$, $A(h) := \{a \mid ha \in H\}$为在$h$时可行行动集.

4) 局中人集合$N = \{1, N\}$, 此外, c为特殊局中人, 称为“机会”或“自然”.

5) $p : H \setminus Z \to N \bigcup \{c\}$为局中人剖分函数, 根据在$h$时由哪位局中人行动, 将非结束状态历史交互集合剖分为H_i.

6) 机会局中人的策略为行动H_c上的固定概率分布$\pi_c, \pi_c(h) \in \Delta(A(h))$.

7) 收益函数$U = (U_i)_{i \in N}$根据局中人i到达z时的收益给每个结束历史交互z赋值$U_i(z) \in \boldsymbol{R}$.

8) 不完美信息博弈G的信息集$I = (I_i)_{i \in N}$, 对于每个局中人, I_i为H_i的剖分, 若$g, h \in H_i$属于博弈$I \in I_i$, 则i无法区分它们. 对于每个$I \in I_i$, 可行行动$A(h)$对于每个$h \in I$必须一样, $A(\cdot)$可表示为$A(I) = A(h)$.

这类博弈模型通常假设局中人具有完美回忆(perfect recall), 不会遗忘历史交互信息, 任何2人带完美回忆的扩展式博弈有等价的正则式博弈形式.

2.1.2 树状序贯决策过程模型

早期的一些研究尝试利用树丛(treeplex)[81]来表示扩展式博弈. 为了便于描述树型决策空间, Farina等[82]提出利用树状序贯决策过程(tree-form sequential decision processes, TFSDPs)模型描述序贯决策问题和序贯博弈.

定义 2 树状序贯决策过程模型可表示为一个7元组$(J, A, K, S, \rho, \Sigma, p)$, 其中:

1) J为决策点集合;

2) A为合法行动的集合, A_j为在决策点$j \in J$处的合法行动;

3) K为观测点集合;

4) S为可能信号集合, S_k为在预测点$k \in K$处的可能信号;

5) ρ为变换函数, 给定$j \in J$和$a \in A_j$, $\rho(j, a)$返回决策树中下一个点$v \in J \bigcup K$(在j处选择行动a后达到的节点)或$\perp$(若决策终止), 给定$k \in K$和$s \in S_k$, $\rho(k, s)$返回决策树中下一个点$v \in J \bigcup K$(在k处预测到s后达到的节点)或$\perp$(若决策终止);

6) Σ为序列集合, 定义为$\Sigma := \{(j, a) : j \in J, a \in A_j\}$;

7) p为决策点$j \in J$的父序列, 定义为树状序贯决策过程根节点至决策节点j间的序列对(决策点-行动), 若在j前没有行动, 则$p_j = \varnothing$.

智能体的策略为合法行动集A上的分布, 其行为策略(behavioral strategies)可表示为向量$x \in \boldsymbol{R}_{\geqslant 0}^{|\Sigma|}$, 即TFSDPS的序贯策略构成一个凸多胞体(convex polytope). 一个合法的序贯策略满足$\sum\limits_{a \in A_j} x[ja] = x[p_j]$.

2.1.3 因子可观随机博弈模型

由于扩展式博弈模型中没有明确的“状态”概念, 序列式序贯决策模型采用树丛描述决策空间, 很难直接利用强化学习来学习策略. Kovařík等[83]提出了改造随机博弈模型来适用不完美信息博弈模型, 因子可观随机博弈(factored-observation stochastic games, FOSGs)模型是部分可观随机博弈的一个变体, 其中博弈包括全局状态, 转换到下一个状态的概率是所有局中人采取行动的函数, 当博弈状态发现变化后, 局中人无法直接观测到底层状态, 相反观测是可分解的, 包括私有和公共观测.

定义 3 因子可观随机博弈模型可表示为一个7元组$G = (N, W, w^o, A, T, U, O)$, 其中:

1) $N = \{1, 2\}$为局中人集合.

2) W为全局状态集合, $w^0 \in W$为指定的初始状态.

3) $A = A_1 \times A_2 \times \ldots \times A_N$为联合行动空间. 子集$A_i(w) \subset A_i$, $A(w) = A_1(w) \times A_2(w) \times \ldots \times A_N(w)$为在$w$处的合法联合行动. 对于$a \in A$, 记$a = (a_1, a_2, \ldots, a_N)$. 对于$i \in N$, $A_i(w)$或全为非空, 或全为空. 无合法行动可执行的全局状态即为结束状态.

4) 在w处执行合法联合行动a后, 状态变换函数T确定下一个$w' \sim T(w, a) \in \Delta(W)$.

5) 记$U = (U_1, U_2, \ldots, U_N)$, 其中$U_i(w, a)$为$i$在

w处执行联合行动a后获得的收益.

6) $O=(O_{\text{priv}(1)},O_{\text{priv}(2)},\dots,O_{\text{priv}(N)},O_{\text{pub}})$为观测函数, $\boldsymbol{O}=(\boldsymbol{O}_{\text{priv}(1)},\boldsymbol{O}_{\text{priv}(2)},\boldsymbol{O}_{\text{pub}})$为观测集合. 观测函数$O_{(\cdot)}:W\times A\times W\to\boldsymbol{O}_{(\cdot)}$确定$i$接受到的私有观测, 每个局中人从$w$执行$a$到达$w'$时获得的公共观测信息. 对于每个$i$, 记$O_i(w,a,w')=(O_{\text{priv}(i)}(w,a,w'),O_{\text{pub}}(w,a,w'))\in\boldsymbol{O}_i:=\boldsymbol{O}_{\text{priv}(i)}\times\boldsymbol{O}_{\text{pub}}$. 博弈开始时, 每个局中人接受到$O_i(w^0)$.

博弈从初始状态w^0开始按回合进行, 在每个回合中, 每个局中人选择一个行动$a_i\in A_i(w)$, 组成联合行动$a=(a_i)_{i\in N}$, 变换到新状态$w'\sim T(w,a)$. 由变换生成观测$O(w,a,w')$, 其中每个局中人得到$O_i(w,a,w')=(O_{\text{priv}(i)}(w,a,w'),O_{\text{pub}}(w,a,w'))$. 每个局中人被指定一个收益$U_i(w,a)$, 此过程按回合进行至结束状态.

序贯决策问题的研究主要有博弈论和强化学习两大领域, 建立不完美信息博弈的因子可观随机博弈模型有助于研究者运用强化学习领域的相关方法解决不完美信息博弈的相关问题.

当前, EFGs作为一类经典模型是其他模型的基础, TFSDPs模型可用于利用在线凸优化理论研究博弈论. FOSGs作为一种类比随机博弈的通用模型, 架构起了算法博弈理论与强化学习领域研究的桥梁. Schmid等[84]基于FOSGs结合离线自对弈学习与在线搜索方法提出了面向完美和不完美信息博弈的通用博弈求解模型.

2.2　子博弈和元博弈

2.2.1　子博弈

完美信息博弈的子博弈(subgame)是一棵子树, 对于不完美信息博弈, 子博弈的概念是建立在增强的信息集基础上的. 不完美信息的子博弈是一片“树林”, 任何局中人的增强信息集中后代关系和成员关系均是封闭的[85]. 通常将原博弈分解为主干(trunk)和子博弈. 正是由于均衡策略计算复杂, 一些研究试图用空间换时间, 先将原博弈进行分解, 分别计算子博弈的策略以及在线博弈对抗时可逆向求解原博弈策略. 这种方式推动了利用残局(endgame)信息求解原博弈均衡的相关工作的开展[86].

2.2.2　元博弈

元博弈即博弈的博弈, 是研究经验博弈理论分析(empirical game theoretic analysis, EGTA)[87]的模型基础, 可用于辅助分析博弈策略的空间形态. Balduzzi等[88]提出根据当前依靠学习得到的参数化策略, 利用泛函式博弈(functional-from game, FFG)来研究传递博弈(transitive game)和循环博弈(cyclic game).

定义4　设定W为参数化策略集合. 可用对称零和泛函式博弈$\phi:W\times W\to\boldsymbol{R}$来评估成对策略$(v,w)$, ϕ为反对称函数, $\phi(v,w)=-\phi(w,v)$, $\phi(v,w)$取值越高, 策略v越好.

假定W为有概率测度的紧集(compact set), W上的可积反对称函数构成一个向量空间, 则任何一个FFG可分解为一个传递博弈和一个循环博弈, 即FFG = Transitive Game $\oplus$ Cyclic Game. 对于传递博弈, 可借助评分函数(rating function) f描述: $\phi(v,w)=f(v)-f(w)$. 对于循环博弈, 可采用策略集积分来描述: $\int_W\phi(v,w)\cdot\mathrm{d}w=0$.

给定n个策略的种群P, 其$(n\times n)$非对称评估矩阵可表示为$A_P:=\{\phi(w_i,w_j):(w_i,w_j)\in P\times P\}=:\phi(P\otimes P)$. 从原理上而言, 需要分析策略集的泛函式博弈空间, 表示为凸包(convex hull), 有

$$G_\phi:=\mathrm{co}(\{\phi_w(\cdot):w\in W\})\subset L(W,\boldsymbol{R}),$$

其中$L(W,\boldsymbol{R})$为W上实值函数空间, 但是真实可观且可交互的为实证博弈空间, $G_P:=\mathrm{co}(\{A_P(i,\cdot)\})$[88].

一些研究认为许多泛函式博弈具有低维的隐结构, 对于给定n个策略的种群P, 其评估矩阵的秩满足$r=\mathrm{rank}(A_P)\leqslant n$, 如对于由$n$个策略构成的长循环博弈, 根据$n$的奇偶性, 其秩$\mathrm{rank}(A_P)$分别为$n-1$和$n-2$.

2.3　解概念和评估

早期解概念的相关研究主要从稳定性(是否有意愿偏离)、安全性(最低收益最大化)、存在性(是否特殊条件下存在)、唯一性(多个解可选择)等多方面属性要求出发, 假设局中人是完全理性的, 局中人i对于策略π_{-i}的最佳响应(best response, BR)可表示为$\arg\max_{\pi_i\in\Pi_i}u_i(\pi_i,\pi_{-i})$. 将局中人$i$的最佳响应值记为

$$\mathrm{BR}(\pi_i)=\min_{\pi_{-i}}u_i(\pi_i,\pi_{-i})=-\max_{\pi_{-i}}u_{-i}(\pi_i,\pi_{-i}).$$

对于局中人i, 其极大极小(maxmin)策略可表示为

$$\arg\max_{\pi_i\in\Pi_i}\min_{\pi_{-i}\in\Pi_{-i}}u_i(\pi_i,\pi_{-i})=\arg\max_{\pi_i\in\Pi_i}\mathrm{BR}(\pi_i).$$

2.3.1　纳什均衡

基于以上策略, 可定义纳什均衡策略.

定义5　不完美信息博弈的纳什均衡策略π^*是由每个局中人的最佳响应组成的策略组合(π_i,π_{-i}), 即

$$\forall i,\ u_i(\pi_i^*,\pi_{-i}^*)\geqslant\max_{\pi_i'}u_i(\pi_i',\pi_{-i}^*),$$

其中任何局中人偏离纳什均衡解均不会得到更大的利益, 反而会给对方留下可利用空间. 由于纳什均衡比较难计算, 近似纳什均衡解在实际应用中被广泛采

纳,博弈的ϵ纳什均衡满足

$$\forall i,\ u_i(\pi_i^*,\pi_{-i}^*)+\epsilon\geqslant\max_{\pi_i'}u_i(\pi_i',\pi_{-i}^*).$$

一些研究试图探索比纯策略纳什均衡(pure nash equilibrium, PNE)更一般的解概念,如混合纳什均衡(mixed nash equilibrium, MNE)、相关均衡(correlated equilibrium, CE)和粗相关均衡(coarse correlated equilibrium, CCE),4类均衡概念满足PNE $\subseteq$ MNE $\subseteq$ CE $\subseteq$ CCE. 对于不完美信息博弈,超过2个局中人[89]、非常和博弈[90]、没有完美回忆[91-92]等情形的纳什均衡(即使是近似纳什均衡)策略的求解至少是PPAD难问题[89].

对于非完美信息博弈,一些研究从非耦合动态学出发,分析了扩展式博弈的解概念,von Stengel等[93]提出了扩展式相关均衡(extensive form CE, EFCE),Celli等[94]设计了面向扩展式相关均衡的无悔学习方法,Farina等[95]提出了扩展式粗相关均衡(extensive form CCE, EFCCE),指出EFCCE为正则式粗相关均衡(normal form CCE, NFCCE)的子集,满足EFCE $\subseteq$ EFCCE $\subseteq$ NFCCE.

对于扩展式博弈,一些研究围绕Stackelberg均衡展开,Bosansky等[96]提出了使用序贯形式的混合整数规划方法求解扩展式博弈的Stackelberg均衡,Černý等[97]提出了增量策略生成方法. 此外,Cermak等[98]提出了利用相关均衡策略来求解Stackelberg扩展式相关均衡策略,Kroer等[99]提出了Stackelberg鲁棒均衡,Černý等[100]提出了考虑有限理性对手的量子(quantal) Stackelberg均衡.

2.3.2 鲁棒响应

传统的解概念均基于理性局中人假设,即使对手策略只是接近最优(如顶级专业玩家扑克比赛中),纳什均衡通常是安全的选择,因为职业玩家非常善于利用次优策略,则要想建模对手行为并予以利用十分危险. 然而,采用纳什均衡或利用对手策略的界限并不十分明确. 在真实的博弈对抗中,遇到有限理性的对手是很常见的,对手可能经常玩被支配策略[101],局中人若采用纳什均衡策略可能会失去有效利用次优策略对手获得更多奖励的机会,这会激励局中人偏离纳什均衡去利用对手的弱点. 但是,若局中人过度适应当前的对手,则由此产生的反制策略可能会使得自己被利用. 因此,追求利用对手的策略可能会导致己方策略被对手剥削,特别是遇到欺骗类诈唬[102]对手时,对手可能会在前期教会局中人适变并学习对手的策略,后期对手可能会采用针对局中人已学会策略的压制策略,这类情形常被称作“受教与被利用”[103].

早前的一些研究围绕ϵ安全策略展开,目的是确保局中人的策略在安全界限内. 随着博弈对抗持续进行,当对手犯下错误,即对手采用次优策略相当于暴露可被利用机会,这时局中人可安全地剥削对手[79]. 这类从安全角度出发的解概念主要有:约束纳什响应(restricted Nash response, RNR)[80]、数据偏置响应(data biased response, DBR)[80]、基于采样的蒙特卡洛约束纳什响应(Monte Carlo RNR, MCRNR)[104]等. 但是,每当需要重新平衡利用性(exploitation)与可利用性(exploitability)或对手切换至新的平稳策略时,需要重新求解原博弈模型.

2.3.3 精炼均衡

正如现实比赛中,选手有犯错误的可能性(类似手一颤抖至无法抓住想抓的东西),一些研究提出从纳什均衡入手,求解均衡精炼策略,试图最大限度地利用对手的错误从而利用对手. 颤抖手精炼均衡[105]是较早被提出来解决此类问题的解概念. 近年来,围绕纳什均衡的一些均衡精炼解概念逐渐引起研究人员的注意,如单侧(one-sided)量子均衡[106]、单侧准完美均衡[107]、序贯均衡[108]、扩展式完美均衡(extensive-form perfect equilibrium, EFPE)[109]、准完美均衡(quasi-perfect equilibrium, QPE)[110]等. 其中:EFPE假定局中人最大化己方收益时需要同时考虑双方可能犯错的情形,QPE假定局中人最大化己方收益时需要考虑对手在未来可能犯的错.

2.3.4 质量评估

离线策略质量的评估方法主要有可利用性[111].

定义6 不完美信息博弈的策略π对于策略π_{-i}的利用性(exploitation)描述了可赢得对当前对手的程度,可表示为$E(\pi)=\max\limits_{\pi_i\in\Pi_i}u_i(\pi_i,\pi_{-i})$.

定义7 不完美信息博弈的策略π_i的可利用性(exploitability)描述了可能输给未知对手的程度,可表示为$e(\pi_i)=u_i(\pi_i^*,\mathrm{BR}(\pi_i^*))-u_{-i}(\pi_i,\mathrm{BR}(\pi_i))$.

对于局中人角色轮转的两人零和博弈,策略组合π的可利用性表示为$e(\pi)=(u_i(\pi_i^*,\pi_{-i})+u_{-i}(\pi_i,\pi_{-i}^*))/2$,纳什均衡解的可利用性为0,$\epsilon-$NE的可利用性不超过$\epsilon$. 可利用性衡量了可被对手剥削的程度,即因未采用纳什均衡策略,对手用强有力的反制策略对其作出的惩罚. 对于多人博弈,可采用NASHCONV[112]来度量策略组合的全局可利用性$\mathrm{NASHCONV}(\pi)=\sum\limits_{i\in N}u_i((\mathrm{BR}(\pi_{-i}),\pi_{-i})-u_i(\pi_i,\pi_{-i}))$,平均可利用性为$\mathrm{NASHCONV}(\pi)/|N|$.

对于在线博弈对抗,当前一些研究主要考查了重复博弈中算法的安定性(soundness)[113]. 与离线场景

下研究策略的可利用性一致,安定性主要用于评估智能体对抗“最坏对手”时的表现,当智能体采用固定策略时,安定性与可利用性等价.

定义8　重复博弈中,对于一个(k,ϵ)-安定性在线算法,在k'次博弈对抗中,其对任意对手的平均收益至少好于采用ϵ纳什均衡策略,若算法Ω对于$\forall k\geqslant 1$是(k,ϵ)-可靠的,则称算法是ϵ-安定的,有

$$\forall k'\geqslant k\forall\Omega: E_{m\sim P^{k'}_{\Omega,\Omega}}[R(m)]\geqslant E_{m\sim P^{k'}_{\sigma,\Omega_2}}[R(m)].$$

自图灵测试被提出以来,在人机对抗中战胜人类已然成为测试人工智能AI的主要衡量准则. 人机对抗为探寻机器智能的内在生长机制和关键技术原理提供了一个极佳的实验环境和验证途径[21]. 由于各种不确定因素(模型、状态等),人类精力和水平波动,方差缩减[114]方法能够有效地减小评估过程中随机因素带来的方差. 方差分解[115]方法可有效用于度量技能(skill)、运气(chance)和非传递性(non-transitivity).

此外,当前一些评分方法也被用于人工智能AI段位的评估. 对于非传递性压制博弈,Elo[116]通过多次两两对抗的结果对所有的排名对象进行评价,TrueSkill[117]基于概率图模型对Elo评分系统从两两对抗扩展到组队对抗;对于传递性压制博弈,mElo2k[118]利用Schur和Hodge分解来构造循环压制分量矩阵的低秩近似进而扩展Elo方法,α-rank[119]通过软化(soft)策略间响应图(response graph)节点间的关系,构建Markov-Conley链,用于种群策略评估.

3　离线策略求解

当前的一些研究试图求解离线策略本质是寻找博弈的近似纳什均衡. 对于信息和动作空间比较小的不完美信息博弈一般可采用优化类方法求得精确的解析解,而对于信息和动作空间比较大的博弈,当前的主流方法是采用博弈抽象(game abtraction)方法[120],先行求解抽象后博弈的解,然后通过逆映射得到原博弈的近似均衡.

博弈抽象主要分为3大类:信息抽象、阶段抽象和动作抽象. 信息抽象[121]是不完美博弈中比较常见的一种模型简化方式,一些具有相似情境的状态信息被合并为一类,与机器学习中的聚类方法相似. 信息抽象根据关键信息的损坏程度可分为无损抽象(lossless abstraction)和有损抽象(lossy abstraction)[122]. GameShrink是一种常见的无损信息抽象方法[123],且是一个自下而上的动态算法,可合并所有有序博弈的同构节点,但是该方法存在抽象不准确,无法确定抽象类数量以及计算量大等问题. 当博弈规模如此之大以至于执行必要数量的信息抽象以使得均衡计算可行但是结果策略的性能非常差时,阶段抽象[124]可能很有用. 一个简单的阶段抽象方法将博弈分为2个半场,博弈上半场的收益是在假设博弈剩余部分一些“默认”行为的情况下计算的,博弈的下半场使用基于前半部分计算的策略更新的不完美信息的信念(根据贝叶斯规则)来求解. 但是,这种简单抽象方法不允许上半场的策略分析考虑下半场的策略情况,导致结果可能是次优策略. 动作抽象[125]是一种针对动作空间的问题简化方法,在有大型或无限动作空间的博弈中,由于动作行为的复杂性,博弈求解难度大大增加. 此时,一个好的动作抽象有助于减少搜索整个博弈空间的难度. 然而,动作抽象面临的问题是玩家会执行一些在已抽象模型外的动作. 针对这一情况,需要定义一个动作映射函数,将所有的动作映射至建模好的动作抽象上. 虽然结合抽象技术,在解决大规模不完美信息博弈问题时,利用动作或信息集层面的抽象成功地将博弈规模压缩,减小运算内存开销. 但是,抽象技术依赖于额外的领域知识,模型的通用性较差,且抽象的过程无可避免地引入误差,导致抽象诟病(pathologies)[126]. 这种误差来自抽象技术的内在缺陷——抽象的原则在于尽可能细粒度地保留博弈中“重要”的部分,合并“不重要”的部分,但是在均衡未知的前提下,无法度量和区分博弈状态是否“重要”,因此陷入了“A依赖于B,B依赖于A”的循环依赖问题. 尤其是有损抽象方法可能使得博弈树结构发生变化,抽象后的纳什均衡解不一定是原来大规模博弈问题的均衡解. 其次,手工设计的抽象粗细粒度也会对均衡的求解产生影响.

当前关于序贯非完美信息博弈离线策略求解的研究主要可分为“算法博弈论、优化理论和博弈学习”共3大类. 可将各类方法作简要区分,但是各大类间没有清晰的界限,如表4所示. 当前的一些研究对其中的等价关系进行了探索,如Blackwell可接近性和无悔学习[127]、镜像下降和后悔最小化[128]、后悔值和强化学习中优势函数[129]等.

3.1　算法博弈论

算法或计算(algorithmic or computational)博弈论是一门从博弈解概念出发,研究解的存在性、计算复杂性和机制设计的一门学科,其中后悔值和博弈动力学是算法博弈论相关研究的主要着力点. 算法博弈论类策略求解方法,本质上是无悔学习方法,关注如何设计迭代式方法驱动博弈对抗策略趋近均衡策略.

表 4 离线策略求解方法分类和示例

类别	分类	方法示例	特点
算法博弈论	反事实后悔最小化	CFR[70]、CFR+[130]	小规模博弈求解
	采样类反事实后悔最小化	MCCFR[131]、VR-MCCFR[114]	约简搜索空间
	拟合类反事实后悔最小化	f-RCFR[132]、CFVnet[31]、GT-CFR[84]	函数回归、网络拟合
优化理论	线性规划	LP[133]、Sparsified LP[134]	精确解
	梯度方法	FOMs[135]、MP[136]、MAIO[137]、OMWU[138]	梯度优化
	凸优化	EGT[139]、FTRL[128]、OMD[140]	收敛速度快
博弈学习	对弈学习	FP[72]、GWFP[141]、XFP[142]、NSFP[75]	迭代式对弈
	强化学习	DREAM[143]、Rebel[144]、ED[145]、ARMAC[146]	交互式学习
	元博弈学习	PSRO[112]、Diverse-PSRO[147]、Auto-PSRO[148]	策略评估与提升

后悔最小化(regret minimization)主要受“事后理性”驱动,其主要描述了博弈局中人通过回顾博弈对抗交互历史,变换己方策略减少后悔至无法获得更佳收益并逐步学会博弈思想. Φ-后悔(Φ-regret)[149]给出了后悔最小化方法的通用模型. 给定策略点集X和线性(仿射)变换函数集Φ(表示对原有策略的可接纳偏离),满足$\phi: X\to X$,与黑盒环境重复交互,通过2步交替操作来确保后悔值有界、收敛: 1) 下步策略: 根据后悔最小化方式输出策略$x^t\in X$. 2) 观测效用: 为后悔最小化方式提供环境反馈,通常为线性效用函数$l^t: X\to \boldsymbol{R}$,用来评估最新输出策略x^t的好坏. 其中,质量度量累计Φ-regret可定义为

$$R_{\Phi}^{\mathrm{T}}:=\max_{\hat{\phi}\in\Phi}\Big\{\sum_{t=1}^{T}(\ell^t(\hat{\phi}(\boldsymbol{x}^t))-\ell^t(\boldsymbol{x}^t))\Big\}.$$

对于变换函数的选择,可表征智能体的“理性”水平. 若Φ为点集X的全射集,则可认为是最高层级的事后理性,Φ-后悔被称为交换(swap)后悔;若Φ为点集X上的“单点偏离”($\Phi=\{\phi_{a\to b}\}_{a,b\in X}, \phi_{a\to b}: x\to b(x=a)$),则$\Phi$-后悔被称为内部(internal)后悔;若$\Phi$为点集$X$的常变换($\Phi:=\{\phi_{\hat{x}}: x\to\hat{x}\}_{\hat{x}\in X}$),则可认为局中人是极度理性的,相应$\Phi$-后悔被称为外部(external)后悔. 研究表明,外部后悔最小化方法在两人零和博弈中可确保收敛至纳什均衡,在多人一般和博弈中收敛至粗相关均衡. 此外,相似的后悔概念有与EFCE十分相关的触发(trigger)后悔[150],用于研究演化动力学的马赛克(mosaic)后悔[151]等.

后悔匹配(regret matching, RM)最初是在研究单次决策问题中提出来的. 假定表征策略分布的概率单纯形(simplex)为$\Delta^n=\{(x_1,x_2,\dots,x_n)\in\boldsymbol{R}^n_{\geqslant 0}: x_1+x_2+\dots+x_n=1\}$,则外部后悔为

$$R^{\mathrm{T}}:=\max_{\hat{\phi}\in\Delta^n}\Big\{\sum_{t=1}^{T}(\langle\ell^t,\hat{x}\rangle-\langle\ell^t,x^t\rangle)\Big\}.$$

Hart等[152]提出可利用Blackwell可接近性构造外部后悔最小化方法. 定义向量值收益函数$\boldsymbol{u}:\Delta^n\times\boldsymbol{R}^n\to\boldsymbol{R}^n, \boldsymbol{u}(\boldsymbol{x}^t,\ell^t)=\ell^t-\langle\ell^t,\boldsymbol{x}^t\rangle 1$,平均收益为$\boldsymbol{\phi}^t=\dfrac{1}{t}\displaystyle\sum_{\tau=1}^{t-1}\boldsymbol{u}(\boldsymbol{x}^\tau,\boldsymbol{y}^\tau)$,累积后悔$\boldsymbol{r}^t:=\displaystyle\sum_{\tau=1}^{t}\ell^\tau-\langle\ell^\tau,\boldsymbol{x}^\tau\rangle 1$. 根据后悔匹配方法($\boldsymbol{x}^{t+1}$与$\boldsymbol{r}^t$成比例)得到下一轮策略为

$$\boldsymbol{x}^{t+1}=\frac{[\phi^t]^+}{\|[\phi^t]^+\|_1}\in\Delta^n\Longleftrightarrow\boldsymbol{x}^{t+1}\propto[\phi^t]^+\propto[\boldsymbol{r}^t]^+.$$

为去除“差行为”对策略生成的干扰,RM+方法[153]将取负值的累积后悔视作0后悔行为. 2类方法的后悔界均为

$$R^{\mathrm{T}}\leqslant\Omega\sqrt{T},$$

其中Ω为$\|\ell^t-\langle\ell^t,\boldsymbol{x}^t\rangle 1\|_2$的最大范数.

当前,关于“下步策略”生成的方法还有乘性权重更新(multiplicative weights update, MWU)[154]、Exp4[155]、Hedge[156]等.

3.1.1 反事实后悔最小化

对于序贯不完美信息博弈,如何构造满足序贯策略空间的后悔最小化方法十分重要. Zinkevich[70]等提出了反事实后悔最小化(counterfactual regret, CFR)方法,为采用后悔最小化方法求解扩展式博弈提供了基准. Farina等[157]提出可采用凸胞上的笛卡尔积等保凸运算来设计满足序贯策略空间上的“后悔回路”(regret circuits),为处理约束类策略(对手建模、均衡精炼)求解提供了理论支撑,CFR方法是其中的一种后悔最小化方法. CFR方法以迭代方式进行,每次迭代包含如下2大步骤:

1) 下步策略: 基于后悔匹配,利用累积反事实后悔得到迭代策略σ^t,更新平均策略组合;

2) 观测效用: 基于新的平均策略组织更新反事实收益,基于反事实收益更新后悔值.

基础CFR算法在现有条件下只能解决10^{12}规模的博弈问题,其中$R^{\mathrm{T}}_{i,imm}(I)\leqslant\Delta_{u,i}\sqrt{|A_i|}/\sqrt{T}$,

从而有 $R_i^{\mathrm{T}}(I) \leqslant \Delta_{u,i}|L_i|\sqrt{|A_i|}/\sqrt{T}$, 其中 $|A_i| = \max\limits_{h:p(h)=i}|A(h)|$[131]. 为突破基础 CFR 算法的瓶颈, 将其应用于大规模不完美信息博弈问题中, CFR 的各类变体不断改进优化, 如加权的折扣 CFR (discouted CFR, DCFR)[158]、CFR+[130] 等.

3.1.2　采样类反事实后悔最小化

经典的蒙特卡洛采样方法主要包括结果 (outcome) 采样、外部 (external) 采样[131]. 2 人不完美信息博弈中, 局中人 1 的平均效用可采用 3 重线性映射 (trilinear map) 来表示, 即

$$u_1(\boldsymbol{x},\boldsymbol{y},\boldsymbol{c}) := \sum_{z\in Z} u_1(z)\cdot \boldsymbol{x}[\sigma_1(z)]\cdot \boldsymbol{y}[\sigma_2(z)]\cdot \boldsymbol{c}[\sigma_c(z)].$$

基于结果采样的蒙特卡洛反事实后悔最小化 (Monte Carlo CFR, MCCFR) 可提供效用的无偏估计. 对于 $\delta \in (0,1)$, 假定常数 M、$\tilde{M}$ 满足 $|(\ell^t)^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{x'})| \leqslant M$, $|(\tilde{\ell}^t)^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{x'})| \leqslant \tilde{M}$, 后悔值以 $1-\delta$ 的概率有界, 有

$$P\Big(R^{\mathrm{T}} \leqslant \tilde{R}^{\mathrm{T}} + (M+\tilde{M})\sqrt{2T\log\frac{1}{\delta}}\Big) \geqslant 1-\delta.$$

Farina 等[159] 从梯度估计的角度提出了利用结果采样、外部采样和平衡外部采样来构建随机后悔最小化方法. 通常这类采样类 CFR 方法面临“方差”挑战[160], 一些研究通过方差缩减 (variance reduction) 方法可有效控制对手估计的准确性[114]. Davis 等[161] 研究采用各类低方差和 0 方差基线 (baseline) 函数学习博弈策略. Schmid 等[114] 基于递归自举提供的反事实后悔无偏估计和基线提出了方差缩减 MCCFR 方法, 即 VR-MCCFR.

3.1.3　拟合类反事实后悔最小化

Greenwald 等[162] 很早提出了 (Φ, f) 后悔匹配通用框架, Φ 为变换函数, f 一般为特定连接函数 (多项式函数、指数簇函数). 基于函数拟合的方法主要有基于回归的 CFR (regression CFR, RCFR)[163]、近似后悔匹配的 f-RCFR[132] 等. 基于神经网络拟合的方法主要有面向反事实值估计的 CFVnet[31]、采用双网络的小批次 MCCFR (mini-batch MCCFR, MB-MCCFR)[164]、基于元学习的自主 CFR (autoCFR)[165]、免模型自举学习 CFR[166] 等.

由于拟合类 CFR 方法较表格式方法具有更强的表征能力, 可用于近似后悔值、累积后悔值、平均策略等. Schmid 等[84] 基于反事实后悔值和策略网络设计了满足通用不完美信息博弈求解的生长树反事实后悔最小化方法 (grow-tree CFR, GT-CFR).

3.2　优化理论

优化类方法主要是采用优化理论问题求解的思路探求博弈解的方法, 其中大多数研究与双线性鞍点问题 (bilinear saddle point problem, BSPP) 模型密切相关. 对于 2 人不完美信息博弈, 通常可直接将其转化为双线性鞍点问题这类数学模型[167] $\min\limits_{x\in X}\max\limits_{y\in Y} x^{\mathrm{T}}Uy$, 问题域由局中人的策略空间多胞体组成, 即

$$\min_{x\in X}\max_{y\in Y}\langle x, Uy\rangle = \max_{y\in Y}\min_{x\in X}\langle x, Uy\rangle.$$

其中: x 和 y 为非负策略向量, X 和 Y 为局中人的序贯策略空间的凸多胞体[133].

对于 2 人博弈, 鞍点间隙 (gap) $\epsilon(x,y) := \max\limits_{y\in Y} x^{\mathrm{T}}Uy - \min\limits_{x\in X} x^{\mathrm{T}}Uy$ 可用于度量 (x,y) 与纳什均衡间的距离, (x,y) 为纳什均衡当且仅当 $\epsilon(x,y)=0$.

Farina 等[168] 为扩展式博弈的相关均衡求解构造了双线性鞍点问题模型, Grand-Clément 等[169] 提出了免参数双线性鞍点问题求解的圆锥 Blackwell 算法.

3.2.1　线性规划

线性规划 (linear programming, LP) 是一种使用序列形式约简方法计算不完美信息博弈中的极小极大均衡的方法[133], 规划问题的规模与博弈树的大小线性相关. 设定 $X = \{\boldsymbol{x} \in \boldsymbol{R}^{|\Sigma_1|} : F_1\boldsymbol{x} = \boldsymbol{f}_1,\ \boldsymbol{x} \geqslant 0\}$, $Y = \{\boldsymbol{y} \in \boldsymbol{R}^{|\Sigma_2|} : F_2\boldsymbol{y} = \boldsymbol{f}_2,\ \boldsymbol{y} \geqslant 0\}$. 对于局中人 1 而言, 纳什均衡对应线性规划问题的解为

$$\begin{aligned} &\max\ \boldsymbol{f}_2^{\mathrm{T}}\boldsymbol{v};\\ &\text{s.t. } U^{\mathrm{T}}\boldsymbol{x} - F_2^{\mathrm{T}}\boldsymbol{v} \geqslant 0,\\ &\quad F_1\boldsymbol{x} = \boldsymbol{f}_1,\\ &\quad \boldsymbol{x} \geqslant 0. \end{aligned}$$

通常而言, 线性规划计算过程中, 矩阵的非 0 项越多, 计算时间越长. 当前, 线性规划方法无法适用于大规模的博弈问题, 特别是当博弈树的节点超过 10^8 个时, 每次迭代计算量巨大[123], 即使线性规划方法在值上实现了指数收敛至纳什均衡, 但是随着动作空间大小的增长, 这类方法仍然非常低效. Zhang 等[134] 为进一步突破线性规划求解器的限制, 提出了基于收益矩阵的稀疏化线性规划 (sparsified LP) 方法, 假定收益矩阵可表示为 $U = \hat{U} + UM^{-1}V^{\mathrm{T}}$, $\boldsymbol{w} := M^{-\mathrm{T}}U^{\mathrm{T}}\boldsymbol{x}$, 则可稀疏化, 即

$$\begin{aligned} &\max\ \boldsymbol{f}_2^{\mathrm{T}}\boldsymbol{v};\\ &\text{s.t. } \hat{U}^{\mathrm{T}}\boldsymbol{x} - F_2^{\mathrm{T}}\boldsymbol{v} + V\boldsymbol{w} \geqslant 0,\\ &\quad M^{\mathrm{T}}\boldsymbol{w} - U^{\mathrm{T}}\boldsymbol{x} = 0,\\ &\quad F_1\boldsymbol{x} = \boldsymbol{f}_1,\\ &\quad \boldsymbol{x} \geqslant 0. \end{aligned}$$

3.2.2 梯度方法

梯度方法[170]主要包括一阶方法(first order methods, FOMs)[135]、策略外梯度(policy extragradient, PE)[171]、乐观镜像下降(optimistic mirror descent, OMD)[172]、镜像近似(mirror prox, MP)[136]、随机镜像近似(stochastic mirror prox, SMP)[173]、镜像提升对抗升级对手(mirror ascent against an improved opponent, MAIO)[137]等.

FOMs可以$O(1/T)$收敛至纳什均衡,比CFR更接近纳什均衡,但是在接近实际的大规模博弈中,即使最快的FOMs收敛速度也不敌最快收敛的CFR变体,且需要精细调参,不能使用类似CFR的近似误差和采样方法. 随着信息集数量和平均信息集大小的增长,面对更大规模的问题,线性规划难以在多项式时间内求解. 在不完美信息博弈中,可通过适当的平滑实现$O(1/T)$的收敛速度和问题相关的$O(\exp(-kT))$指数收敛速度. 此外,外梯度或乐观镜像下降方法已被验证可收敛至纳什均衡,在无约束空间中可能具有指数收敛率,但是不适用于序贯博弈. 乐观乘性权重更新(optimistic MWU, OMWU)方法以$O(1/T)$的收敛速度确保了末轮迭代收敛至纳什均衡. MP方法可确保鞍点间隙满足[138]

$$\epsilon(\boldsymbol{x}^t,\boldsymbol{y}^t)\leqslant\frac{\|U\|}{2t}\sqrt{\frac{\Omega_X\Omega_Y}{\varphi_X\varphi}}.$$

3.2.3 凸优化

非平滑凸优化方法,如过间隙技术(excessive gap technique, EGT)[139]是一种加速一阶方法,其采用面向树丛的距离生成函数,将目标函数扩张为可微凸函数,沿梯度下降方向迭代产生次优解,可用于解决信息集数量多至10^8的扩展式博弈. EGT方法可确保鞍点间隙满足[174]

$$\epsilon(\boldsymbol{x}^t,\boldsymbol{y}^t)\leqslant\frac{4\|U\|}{t+1}\sqrt{\frac{\Omega_X\Omega_Y}{\varphi_X\varphi}}.$$

乐观凸优化方法,如可预测后悔最小化方法[175],其本质是一阶方法的变体,无悔学习方法的扩展,主要利用距离生成函数构造正则化项[176],与后悔最小化方法不同的是可利用损失ℓ^t的估计$m^t\in\boldsymbol{R}^n$来输出下步策略. 根据Syrgkanis等[177]提出的后悔效用变换有界(regret bounded by variation in utilities, RVU)属性,可预测后悔最小化方法的后悔界为

$$R^{\mathrm{T}}\leqslant\alpha+\beta\sum_{t=1}^{T}\|\ell^t-m^t\|_*^2-\gamma\sum_{t=1}^{T}\|x^t-x^{t-1}\|^2.$$

其中:$\alpha>0,0<\beta\leqslant\gamma,\|\cdot\|$和$\|\cdot\|_*$为一对对偶范数.

一些研究分析了正则化跟风(follow the regularized leader, FTRL)和在线镜像下降(online mirror descent, OMD)方法与传统CFR的关联关系. 其中:Waugh等[178]从一个统一的视角分析CFR与EGT方法间的关联,证明了RM和Hedge方法与对偶平均(dual averaging)等价;Farina等[128]通过可预测Blackwell可接近性理论发现了后悔匹配与镜像下降间的关联,分析了RM与FTRL、RM+以及OMD间的等价性;Liu等[140]证明了CFR-RM为未来依赖自适变FTRL的特例,CFR-RM+为未来依赖自适应OMD的特例. 虽然优化理论类方法,特别是一些梯度方法和凸优化方法取得了较好的收敛性和计算复杂性保证,但是与迭代式CFR方法相比,实用性仍显不足[153],一些研究尝试将两者结合. Farina等[175]提出的乐观后悔最小化方法收敛率为$O(T^{-1})$,Syrgkanis等[177]证明了这类正则化方法的收敛率为$O(T^{-3/4})$.

3.3 博弈学习

学习有2种解释:Fudenberg等[69]从博弈理论和经济学视角指出“计算机博弈中的学习理论主要研究由于学习、适应和/或模仿的持续非平衡过程,可能会出现何种、哪种和什么样的平衡”;Mitchell[179]从机器学习的视角指出“若计算机程序在任务T上的性能(由P衡量)随着经验E的提高而提高,则就任务T和性能度量P而言,计算机程序学习到了经验E”. 早前人工智能领域的相关研究试图利用自主学习的方法令机器掌握人类智能. 从最早的min-max搜索和自对弈(self play)方法的提出直至Littman[73]提出基于强化学习的minimax-Q学习方法,博弈论相关领域的相关研究基于虚拟对弈思路设计面向均衡收敛的迭代式学习方法,从演化博弈理论里的复制动态[69]至事后理性假设的无悔学习[156]等.

3.3.1 对弈学习

虚拟对弈方法根据与对手的博弈对抗历史数据估计对手的平均策略σ_{-i}^{t-1},计算己方的最佳响应$\mathrm{BR}(\sigma_{-i}^{t-1})=\arg\max_{a_i\in A_i}U(a_i,\sigma_{-i}^{t-1})$. 策略更新方式为

$$\sigma_i^t=\frac{t-1}{t}\sigma_i^{t-1}+\frac{1}{t}\mathrm{BR}(\sigma_i^{t-1}).$$

对于两人零和博弈,策略σ_t收敛至纳什均衡. 但是该方法每次迭代计算时均需精确计算最佳响应,策略更新步长固定. Leslie等[141]提出了通用弱化虚拟对弈(generalised weakened FP, GWFP)方法,迭代计算中采用ϵ-纳什均衡,扰动策略更新方式为

$$\sigma_i^t=(1-\alpha^t)\sigma_i^{t-1}+\alpha^t(\mathrm{BR}_i^{\epsilon^{t-1}}(\sigma_{-i}^{t-1})+M_i^t).$$

其中

$$\lim_{t\to\infty}\alpha^t=0,\ \lim_{t\to\infty}\epsilon^t=0,\ \sum_{t=1}^{\infty}\alpha^t=\infty.$$

对于扰动项满足

$$\lim_{n\to\infty}\sup_{k}\left\{\left\|\sum_{i=n}^{k-1}\alpha^{i+1}M^{i+1}\right\|:\sum_{i=n}^{k-1}\alpha_{i+1}\leqslant T\right\}=0.$$

受线性CFR[158]启发, Brown等[144]提出了虚拟线性乐观对弈(fictitious linear optimistic play, FLOP)方法, 作为一种GWFP方法, 计算最佳响应时采用一种乐观的对手平均策略估计, 即

$$\mathrm{BR}(\sigma_{-i}^{t-1})=\arg\max_{a_i\in A_i}U\Big(a_i,\frac{t-1}{t+1}\sigma_{-i}^{t-1}+\mathrm{BR}\Big(\frac{2}{t+1}\sigma_{-i}^{t-1}\Big)\Big).$$

策略更新方式为

$$\sigma_i^t=\frac{t-1}{t+1}\sigma_i^{t-1}+\frac{2}{t+1}\mathrm{BR}(\sigma_i^{t-1}).$$

在GWFP的基础上, Heinrich等[142]提出了面向扩展式博弈的扩展式虚拟对弈(extensive-form FP, XFP)和神经虚拟自对弈(neural fictitious self play, NFSP)[75]方法, 这类方法同样具有收敛性保证. Chen等[180]将NFSP与后悔最小化方法组合, 提出了ARM-NFSP. Zhang等[181]将NFSP与蒙特卡洛树搜索相结合, 提出了MC-NFSP.

3.3.2　强化学习

强化学习方法可分为基于值函数的方法、基于策略梯度的方法和基于Actor-Critic的方法. Littman[73]很早将强化学习Q值函数与minimax算法相结合设计了minimax-Q. Brown等[144]基于递归式信念学习方法ReBel, 提出了一种通用的强化学习和搜索框架. Steinberger等[143]设计了一种基于后悔值、利用优势基线(advantage baselines)函数的免模型自对弈深度强化学习方法DREAM. Marchesi等[105]给出了当前3种博弈模型中的值函数(可达最优值函数、反事实最优值函数和全局最优值函数). Schmid等[84]提出了PoG方法, 该方法使用反事实值和策略网络来学习值函数和生成策略, 采用“稳固的”自对弈方法来训练网络, 在线对抗采用基于生成树反事实后悔最小化的搜索方法. Lockhart等[145]提出了基于可利用性下降(exploitability descent, ED)的直接策略优化方法. Timbers等[182]提出了利用局部最佳响应和搜索方法可利用性近似方法. Srinivasan等[129]将后悔值与Actor-Critic框架相融合, 提出了后悔策略梯度(regret policy gradient, RPG)和后悔匹配策略梯度(regret matching policy gradient, RMPG)方法. Gruslys等[146]将优势函数与后悔匹配组合提出了基于Actor-Critic的免模型回溯式策略提升方法ARMAC.

3.3.3　元博弈学习

同样作为一类迭代式方法, McMahan等[183]提出了Double Oracle方法, 采用子博弈增量迭代方式计算均衡策略. Lanctot等[112]基于Double Oracle和强化学习方法, 提出了面向多智能体强化学习的统一博弈论方法框架, 即策略空间响应预言机(policy space response oracle, PSRO). Muller等[184]面向多人一般和博弈提出了利用α-rank(一种不需要考虑均衡选择问题的解概念)[119]与PSRO组合的通用训练方法. 通过构建纯策略组合间有向响应图来表征策略间的偏离动机, 利用一个图上的不可约噪声随机游走, 得到Markov-Conley链唯一不变平稳分布$C^{\mathrm{T}}\pi=\pi$. 有向响应图上每个节点对应一个联合纯策略组合, 为寻找图上“汇”强连通部分(sink strongly-connected components, SSCC)节点, α-rank在图上构建随机游走, 对于给定的纯策略组合$s\in S$, 假定$\sigma=(\sigma^k,s^{-k})$, $C_{s,\sigma}$为策略s到σ的变换概率, $\boldsymbol{C}_{s,s}$为策略自变换概率, 马尔可夫变换矩阵$\boldsymbol{C}$满足

$$\boldsymbol{C}_{s,\sigma}=\begin{cases}\eta\dfrac{1-\exp(-\alpha(\boldsymbol{M}^k(\sigma)-\boldsymbol{M}^k(s)))}{1-\exp(-\alpha m(\boldsymbol{M}^k(\sigma)-\boldsymbol{M}^k(s)))},\\ \quad \boldsymbol{M}^k(\sigma)\neq\boldsymbol{M}^k(s);\\ \dfrac{\eta}{m},\ \text{otherwise.}\end{cases}$$

$$\boldsymbol{C}_{s,s}=1-\sum_{\substack{k\in[K]\\ \sigma|\sigma^k\in S^k\setminus\{s^k\}}}\boldsymbol{C}_{s,\sigma}.$$

其中: $\boldsymbol{M}^k(\sigma)$为局中人k采用策略组合σ时的期望收益, $\eta=\Big(\sum_l(|S^l|-1)\Big)^{-1}$, $\alpha\geqslant 0$和$m\in\boldsymbol{N}$为3个固定参数.

近年来, 这类方法与元博弈理论紧密相关, 基于PSRO框架的多类种群学习方法已然成为当前利用学习方法求解博弈均衡解的通用方法, 如多样性PSRO(diverse-PSRO)方法[147]、自主PSRO(Auto-PSRO)方法[148]、并行PSRO(pipeline-PSRO)方法[185]、扩展式Double Oracle方法[186]、单纯形神经种群学习(simplex-neuPL)方法[187]等.

4　在线策略求解

当前关于在线策略的研究主要聚焦于重复博弈. 与离线场景不同的是, 由于在线对抗中仅能掌握己方策略, 如何制定反制对手的策略是最符合现实场景的研究点. 在线博弈对抗过程中, 可用2种方式生成己方策略: 1) 从悲观视角出发的博弈最优(game theory optimal, GTO), 即采用离线蓝图策略进行对抗; 2) 从乐观视角出发的剥削式对弈(exploitative play),

即在线发掘对手可能的弱点,最大化己方收益的方式剥削对手.

在线策略求解本质是要回答如何制定剥削对手的策略,即对手剥削(opponent exploitation)这个基本问题. 在线博弈对抗过程中,发现对手弱点并充分利用己方认知优势剥削对手. 在纳什均衡解概念中,双方均采用均衡策略,任何偏离均衡策略的一方所获得的收益将减小. 对于非完全理性的对手,其策略与纳什均衡策略间的“距离”,为其他均衡策略玩家创造了可利用性. 在求解(近似)纳什均衡解时,可利用性是均衡策略质量的衡量标准,指该策略在预期中相对于最坏情况下的对手策略所达到的少于博弈价值的量,通常也将这个差值称为一个策略的可利用程度,其衡量了对手从未采用纳什均衡策略中获利的程度,即因未采用纳什均衡策略,对手用强有力的应对策略对其作出惩罚的程度,这是对策略最坏情况下质量的度量,故其常被用来评估各种策略学习方法.

由于大规模不完美信息博弈中的均衡策略难以求解,均衡策略是一种基于对手完美理性假设的静态策略,其在面对不同对手时缺乏动态策略调整的能力,尤其是在面对次优对手时无法取得更大收益,均衡策略在多智能体环境中的有效性缺少理论保证. 对手剥削类方法则强调根据对手实际行为调整己方策略,不追求在整个博弈解空间中寻找一个最优的“不动点”,而是希望将问题分解至不同对手对抗的子博弈空间中寻找可行解,因此在对手策略不完美理性(或可利用性大于0)的假设前提下,采用对手剥削方法是更加直接有效、轻量灵活的选择.

从对手剥削的视角看,在线策略求解可划分为2个相互影响的部分: 博弈推理(game-theoretic reasoning)和反制策略(counter strategy)生成,主要表示在博弈对抗过程中,推理对手的状态和可能采取的行动,前瞻搜索己方反制策略. 类比对手建模学习方法的分类方式[188],将当前关于序贯非完美信息博弈在线策略求解的研究主要分为3大类:对手近似式学习、对手判别式适变和对手生成式搜索. 综合而言,在线策略求解方法可作简要区分,如表5所示.

表5 典型在线策略求解方法分类和示例

类别	分类	方法示例	特点
对手近似式学习	协同和演化集成	Ensemble[189]	多策略集成
	具对手意识学习	LOLA[190]	假定对手也在学习
	认知建模和推理	BPR+[191]	策略蒸馏和学习
对手判别式适变	行动预测和评估	ASHE[192]	显式建模和预测
	机会发掘和欺骗	Gift[193]	机会检测
	约束策略生成	CCFR[194]	利用约束缩减空间
对手生成式探索	策略探索和优化	IXOMD[195]	隐式策略探索
	蒙特卡洛搜索	MCCR[196]	在线采样和重求解
	子博弈重解和搜索	DLS[197]	子博弈深度有限探索

4.1 对手近似式学习

基于对手感知的对手建模方法主要通过假设对手也在学习己方模型,与对手共同学习演化是其主要思路,其中假设对手也在学习,故在对手建模过程中要考虑到这一层推理关系.

4.1.1 协同和演化集成

此类方法主要采用与对手协同交互的方式学习压制对手的策略. 如协同演化方法,利用对手模型和演化学习方法生成对抗策略池[198]. 集成学习方法采用多个已训练好的专家策略[189],对未知对手行为预测的准确率比单个分类器的结果更高,可为基于已有异构分类模型快速构造通用的对手模型提供支撑,提高对未知对手的预测性能,提高模型泛化能力.

4.1.2 具对手意识学习

相比传统对手模型学习方法,具有对手意识的学习方法通过引入一个包含对手值函数的2(高)阶项来估计对手策略的变化. Foerster等[190]提出的对手具有学习意识的LOLA学习算法,假定对手也具有学习能力,算法需要在与对手交互学习的环境下学习对手模型. Letcher等[199]将前瞻搜索与LOLA算法相结合,提出了用于微分博弈中对手建模的稳定对手塑造方法.

4.1.3 认知建模和推理

认知建模和推理方法主要是试图从认知行为学的角度,采用心智理论分析对手的行动,进而得出己方策略[200]. 基于心智理论的推理方法得到了广泛研究[201],其中递归推理[202]方法对嵌套信念进行

建模(如“我相信你相信我相信···”)并模拟对手的推理过程来预测他们的行动,递归持续推理对手的可能模型,预测行为的可能性,基于层次推理和无悔学习的方法可很好地为在线对抗学习提供支撑. 由于贝叶斯推理可为策略制定提供带信念的后验支撑,近来的一些研究主要聚焦于如何在线生成反制策略. 其中: Southey等[102]提出使用贝叶斯推理类方法可生成贝叶斯最佳响应(Bayesian best response, BBR)、最大后验响应(max a posteriori response, MAP)和汤普森响应(Thompson's response)共3种反制策略; Zheng等[203]基于贝叶斯策略重用方法(Bayesian policy reuse, BPR+)[191]提出了深度贝叶斯策略重用方法,增加的对手建模网络结合策略蒸馏方法,算法包含策略重用和新策略学习2个阶段,提高了学习新策略的效率和对对手策略判断的准确性; Yu等[204]提出了基于模型的对手建模方法将想象的对手策略与真实的对手进行相似性比较,将多种策略进行混合,以求得对手的最佳响应; Papoudakis等[205]提出了利用部分可观信息和变分自编码的局部信息智能体建模方法.

4.2 对手判别式适变

在任何复杂的多智能体系统中,智能体的最优应对策略取决于其他智能体的行为,因此,智能体的一个关键能力是了解并适应其他智能体. 博弈对抗过程中,拥有对手建模能力的智能体可通过事先预测对手可能采取的行动分析己方利弊并及时进行调整应对策略,因此在线对手建模是智能体学习博弈智能不可或缺的能力[206]. 对手建模是多智能体学习领域的一个重要研究方向[207],主要研究如何通过交互历史前瞻预测对手可能采取的序列行动. 传统的对手建模方法主要通过观察对手的行为,并构建其行为的生成模型. 这可能涉及直接估计行为的概率,或在更复杂的模型中拟合一些参数来描述智能体的行为是如何产生的. 尽管这类建模方法为其他智能体的行为提供了直接而丰富的表示,但是仍然存在重大的实际挑战. 即使可估计一个足够精确的离线显式模型,但是如何在线使用该模型调整智能体的行为自适应地应对未知对手仍然是挑战. 在2人策略博弈过程中,若想剥削对手获得更高的回报,纳什均衡不是应对次优对手(即不使用纳什均衡策略)的最佳响应策略,此时,对手建模对于在策略互动中剥削次优对手至关重要[193]. 这类方法依赖对手的显式或隐式模型,通过对手模型分析生成适应对手变化的适变反制策略.

4.2.1 行动预测和评估

行动预测和评估类方法主要是试图推理敌手可能的状态进而补全不完全可观的信息,预测对手可能采样的行动进而找到应对措施,估计对手可能的手牌牌力和赢牌潜力进而分析己方的赢率等. 这类方法包括采用聚类方法推断对手的博弈风格类型,采用频率统计类方法预测对手的动作偏好,采用神经网络类方法预测对手可采取的行动,评估对手的状态值和己方的赢率等. Li等[192]提出使用模式识别树和长短期记忆网络估计器的显式对手建模方法ASHE,其中:底层的常规模式识别树用于在线对抗中对手行为模式的识别,每轮博弈结束后保存至默认模式识别树中;上层的己方赢率估计器算和对手弃牌估计器分别用于估计己方赢率和对手弃牌率,然后决策算法给出最终反制策略.

4.2.2 机会发掘和欺骗

机会发掘和欺骗类方法主要是试图分析博弈对抗态势下对手策略是否暴露弱点、是否“有机可乘”,己方进而可极大化地剥削对手,或己方牌力差时采用“诈唬”的方法吓退强劲对手、己方牌力强时采用“慢打”的方法引诱对手下更多注等. Ganzfried[193]从安全的角度研究了重复零和博弈的对手剥削问题,给出了安全策略的完整特性,在线实时对抗过程中,通过检测对手呈现的礼物(gift),即暴露的机会,提出了一种有安全保证情况下剥削对手次优策略的方法.

4.2.3 约束策略生成

约束策略生成类方法主要是试图分析对手的可能性倾向和可能不会采取的行动,在原有的纳什均衡最优策略基础上生成限定的反制策略. 其中最佳响应剪枝方法根据判断对手的可能行动空间进行剪枝,加快算法的收敛[208]. Davis等[194]根据估计的对手私有信息和可能结局设计了策略约束,提出了基于拉格朗日松弛的约束反事实后悔最小化算法(constrained CFR, CCFR). Farina等[209]提出根据受限行动构建摄动(perturbed)扩展式博弈,用于序贯决策问题的精炼均衡求解[210]与后悔值组合回路设计[157]. Ganzfried等[211]提出的博弈最佳响应方法,通过观察对手的动作频率,并通过将近似均衡策略的信息与观察值相结合来构建对手模型,根据对手模型计算并实行最佳响应,不断进行实时更新. 如为了生成鲁棒的反制策略,可使用统计数据分析和挖掘方法分析对手的倾向性策略,基于专家知识、利用剪枝生成行动受限的纳什响应策略等. 此外, L_1[212]、L_2[211]和KL散度[213]常被用于度量对手模型. 根据每局的对抗结果(公共信息、

对手的私有信息和结果收益信息),局部最佳响应、统计数据偏差响应[80]、有限理性量子响应和安全最佳响应[79]也常用于针对对手的反制策略制定.

4.3 对手生成式搜索

这类方法不需要利用确切的对手模型,但是需要将对手行动纳入反制策略的生成过程中,分析对手可能采取的行动.

4.3.1 策略探索和优化

策略探索优化类方法主要是试图在博弈策略空间上探索剥削对手可能性的方法. Ganzfried等[214]提出基于狄利克雷(Direchlet)先验对手模型,利用贝叶斯优化模型获得对手模型的后验分布,辅助生成剥削对手的反制策略. 此外还有一些方法利用先验分布,模拟与对手的对抗交互过程并得出博弈收益结果,利用博弈策略分析方法分离出策略间的传递压制与循环压制关系,通过预先计算出多个可行对抗策略,然后利用后验采样方法或在线凸优化类方法出生反制策略. Wu等[215]提出利用元学习方法不依赖大量交互数据,而是通过交互过程少量样本学会生成对手模型. Yu等[204]提出利用基于模型的对手建模方法与贝叶斯混合的方法输出适变策略.

此外,基于无悔学习的在线凸优化方法主要试图直接利用在线交互过程中环境反馈信息(如Bandit式反馈、交互式反馈、完全信息反馈)进行策略空间探索,然后通过后验采样的方式生成在线策略. Zhou等[216]提出利用Dirichlet先验分布和贝叶斯后验采样方法. Farina等[217]提出了面临未知博弈模型的在线学习与交互式Bandit模型,将策略探索与后悔采样相结合. Kozuno等[195]提出了基于隐式探索在线镜像下降的免模型学习方法IXOMD. Bai等[218]结合Bandit反馈,提出了基于隐式探索的平衡OMD和基于无偏估计与弹性选择的平衡CFR方法. Meng等[219]提出了通用Bandit后悔最小化框架.

4.3.2 蒙特卡洛搜索和重解

在线蒙特卡洛持续重解类方法主要试图对在线对抗过程中双方状态进行分离,构造领域无关的小型博弈,进而运行蒙特卡洛采样方法重新求解在线策略. 胡裕靖等[220]提出了面向不完美信息扩展式博弈的在线CFR算法. Šustr等[196]基于在线结果采样法和信息集蒙特卡洛采样提出了蒙特卡洛持续重解(Monte Carlo continual resolving, MCCR). Ponsen等[104]提出利用蒙特卡洛树搜索和受限纳什响应生成鲁棒的反制策略. Bitan等[221]提出将人类决策预测融入基于信息集的蒙特卡洛树搜索中. Phan[222]提出了一种基于对手意识的蒙特卡洛树搜索方法,无需利用显式或先验对手模型,底层仅依赖学习到的对手意识作前向搜索,上层为基于梯度的多臂赌博机.

4.3.3 子博弈重解和搜索

子博弈精炼是在线博弈重求解的主要方法[223],传统的子博弈求解方法没有考虑生成策略的安全性,一些新的研究主要从生成策略的安全性出发,采用工具博弈(gadget)重解子博弈. 当前的一些研究表明,基于子博弈构造的工具博弈求解得到的策略可利用性有界且不会增加. 这类子博弈重解方法主要有最大边际求解[223]、安全嵌套求解[197]、子博弈精炼求解[224]、序贯工具博弈求解[225]. Brown等[197]提出在对手建模时要平衡安全与可利用性,基于安全嵌套深度有限搜索的方法可生成安全剥削对手的反制策略. 值函数有限深度搜索类方法试图利用神经网络估计最优值函数辅助在限深度搜索生成反制策略. Brown等[144]提出了基于深度强化学习的值函数方法. Milec等[225]基于值函数,利用有限深度最佳响应、限定纳什响应方法生成反制策略.

5 复杂性挑战和前沿展望

5.1 复杂性挑战

从国际象棋、围棋到德州扑克、星际争霸等,理论与技术的发展呈现交叉融合,面对“巨复杂、强对抗、高动态和多威胁”的现实应用场景、高度仿真模拟现实世界的新型计算机博弈研究环境(如博弈元宇宙)、多人博弈缺乏理论和原理支撑(如多人德州扑克)等为当前开展“环境认知”“策略求解”和“智能体(对手)建模”等相关研究提出了新挑战. 当前,智能体的能力主要依赖云原生平台、数字孪生环境和协同演化方法生成. 环境、对手(机器或人类)以及AI智能体是计算机博弈研究的主要对象,3者间博弈对抗和交互学习是机器智能协同演化生成的重要途径. 3者间共有3对关系: 智能体与环境的互生关系,即不同的

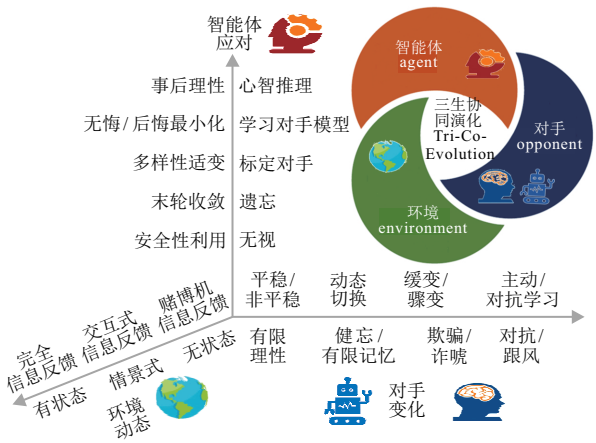

图 3 “环境-对手-智能体”的多维复杂性描述

环境需要不同的智能体策略适配; 智能体与人的共生关系, 即人机对抗研究利用智能体打败人类对手, 人机融合中研究智能体如何理解人类从而与人协作; 智能体与对手间的孪生关系, 即在博弈策略空间中, 智能体与对手策略间的压制关系是孪生共存的. “环境-对手-智能体”的多维复杂性描述如图3所示.

序贯不完美信息博弈对抗过程根据状态是否可知分为有状态(stateful)和无状态(stateless)博弈模型, 其中: 有状态模型一般会考虑博弈各方当前所处的博弈状态, 而无状态模型一般默认当局博弈中途不改变事先策略; 根据环境信息的反馈情况可分为赌博机式(bandit)反馈、交互式(interactive)反馈[217]和完全信息反馈环境模型, 其中: 完全信息反馈假定可接收到对抗过程中所有公共信息, 赌博机反馈假定仅知道每局收益信息, 交互式反馈假定需要根环境交互直至博弈结束返回收益; 根据对手的类型可分为非平稳性[226-227]或平稳性策略对手、非健忘型(non-oblivious)[228]或健忘型策略对手等, 其中非平稳性和非健忘型策略对手主要是指对手策略可变, 难以预测等. 此外, 动态对手类型包括策略切换型对手[229]、主动学习型对手[189]和对抗学习型对手[190]. 当前计算机博弈领域组织在线博弈对抗赛一般有2种形式, 即有限局重复赛(每局底注和收益独立的重复博弈)和连续锦标赛(多局底注变化且收益累计的随机博弈)[230]. 由于对抗信息不完美、状态信息空间大, 离线计算“蓝图策略”(近似纳什均衡策略)[9]时需要大量采样对手的策略空间, 博弈求解十分困难. 针对对手策略的非平稳性(风格多变、不可预测等), 智能体可采取的方法主要可分为5类: 忽略(ignore)[231]、遗忘(forget)[232]、标定(target)[233]、学习(learn)[234]和心智理论(theory of mind, ToM)[235]. 此外, 由于在线对抗过程中智能体仅能控制己方策略, “最优策略”的生成必受多个目标(理性最优、后悔有限[236]、末轮收敛[237]、安全以及剥削[79]等)牵引.

5.1.1　博弈基准和集成测试环境少

当前围绕序贯不完美信息博弈研究的基准测试主要包括RLCard[238]、Openspiel[239]等, 这些基础环境很少与现实场景问题作好映射, 很多现实场景问题无法通过简单的抽象描述为一个基准环境实例, 为了更好地引领相关领域研究应考虑博弈测试环境的设计与现实场景的相关性、相似性. 此外, 相应的集成测试环境包括PettingZoo[240]、MAVA[241], 这类环境大多依赖其他分布式强化学习框架搭建, 为众多博弈基准环境留的接口并不友好, 具体博弈问题的大规模训练仍然需要底层改造.

5.1.2　非传递性压制策略求解难

博弈策略间非传递性压制(策略$v_{t+1} \succ v_t, v_t \succ v_{t-1}$, 但是$v_{t+1} \prec v_{t-1}$), 导致一般的单种群自对弈学习方法难以迭代提升, 收敛至均衡策略. 正因为缺少数据、知识, 难以学习, 特别是自对弈对抗数据缺少多样性, 群体博弈采样效率依赖大算力, 复杂博弈对抗学习目标难确定, 导致多样性的压制策略难习得, 离线模型难以在线适配对手, 环境(对手)的非平稳导致信息不完备, 简单的策略集成存在潜在漏洞, 应对非平稳环境(对手)的在线适变策略难求解; 此外, 基于神经网络学习得到的博弈策略可解释性存疑, 很难溯因.

5.1.3　对手建模难且剥削风险大

当前大部分研究均假设对手策略是平稳的, 学习到的策略无法适应非平稳环境(对手). 根据条件反射原理, 在真实博弈对抗过程中, 人们更倾向于强化受益策略(赢了保持策略不变), 惩罚受损策略(输了改变固有策略). 如何建模非平稳对手模型、应对对手的非理性行为(欺骗[242]、诈唬[102]、合谋[243]等)均是亟待研究的问题. 此外, 当前对手剥削方法主要分为4大类: 频率统计、安全最佳响应、无悔学习和随机优化[244]. 但是这4类方法均可能面临被老练(sophsticated)对手反向剥削的陷阱, 风险比较大.

5.2　前沿展望

5.2.1　博弈动力学和策略空间理论

博弈动力学一直是研究博弈对抗过程中策略收敛至均衡解, 早前是演化博弈的主要研究内容, 近年来实证博弈理论分析方法为研究博弈动力学和策略空间形态探索提供了有效工具. Vlatakis-Gkaragkounis等[245]利用Poincaré-Bendisxon定理、Liouville定理和Poincaré并行流定理分析了隐式双线性博弈中存在循环和虚假均衡. Perolat等[246]分析了跟随正则化领导者动力学, 如何在收益中增加正则化项来确保收敛至均衡策略. 虽然Czarnecki等[247]提出了博弈策略空间的陀螺猜想, 但是并非所有策略均符合该猜想, 相关研究利用相似性会压制循环出现来解释为什么相似的竞争策略呈现出层次[248]. 策略间的传递压制与循环压制并存, 正如现实情景中如何玩好“石头、剪刀、布”一直是个挑战问题[249], 如何分析各类博弈的策略空间形态一直是一个开放式问题.

5.2.2　多模态对抗博弈和序贯建模

如何为多人德州扑克、桥牌、斗地主、麻将、连续状态空间对抗(如格斗、乒乓球)[250-251]等构建实用的

博弈模型是当前面临的现实挑战. Farina等[252]提出了0/1多面体博弈模型,尝试给出扩展式博弈与正则式博弈的统一模型. 当前围绕对抗博弈决策过程的博弈模型有多种[253],包括:零和博弈(双线性博弈、多人零和博弈、团队博弈、非完美信息博弈、线性2次微分零和博弈、非线性微分零和博弈、随机微分零和博弈)、Stackelgerg博弈(通用Stackelberg博弈、Stackelberg安全博弈、Stackelberg微分博弈)等. 围绕对抗团队博弈的相关模型以及一类更为泛化的一般和博弈模型已然成为当前研究的焦点. 此外,马尔可夫决策过程(MDP)和半马尔可夫决策过程(SMDP)常用于序贯决策建模,由于环境(对手)的非平稳性[234],导致决策过程无法满足马尔可夫性.

5.2.3 通用策略学习和离线预训练

博弈策略学习一直是学术研究的前沿课题,当前无领域知识的学习[254]、多样性学习[147]、自主课程学习[148]、元学习[165]、在线无悔学习[255]、分布式策略学习框架[256]、融合规划的学习[257]等均已成为当前的研究前沿重点. 一些研究探究了CFR与强化学习方法的结合[258]、CFR与PSRO类方法的结合[259]、自对弈与PSRO类方法的结合[260]、离线均衡求解预训练模型[261]. 此外,最新的一些研究借助注意力机制Transfomer用序列建模重构了传统的强化学习过程[262],基于序列建模的监督学习方法为离线策略大模型提供了有效支撑.

5.2.4 对手建模(剥削)和反剥削

对手建模是人工智能领域一个长期存在的研究课题. 常用的对手建模方法根据模型表示分为显式建模和隐式建模;根据所采用的学习方法分类为判别式角色或策略分类、基于目标的生成模型以及策略近似[188];根据信息结构和对抗类型分为平稳型对手、猜想型(预定义模型)对手、标定型(未知模型)对手和经验老对手[263]. 通常建模主要试图通过观察对手在不同情况下的行为来推断对手的策略,通过为对手建立一个模型来实现. 模型所需的观察次数一般取决于模型中的自由度. 在没有先验知识的情况下,一般模型的参数数量必须随着智能体策略空间的大小而增长. 对于信息集比较大的不完美信息博弈,学习对手模型所需的观察次数通常会非常大. 即使能够足够快地学习一个模型,面对模型误差,最佳响应策略是脆弱的,且在线实时计算鲁棒响应十分具有挑战性. 在线决策过程中对对手策略的有效估计十分重要. 此外,非平稳对手如何建模,剥削对手时如何确保己方策略的安全性、削减被对手设计和剥削,构建有限理性模型、防止被欺骗均十分重要.

5.2.5 临机组队和零样本协调

在知识表示方面,知识的获取通常可通过利用人类专家、先验知识和自对弈学习等方式. 常用的知识表示方法有类型法、经验回放法和任务识别法. 其中,类型法常常利用“假设类型空间”来表示队友可能的类型,假设新遇到的队友为某种类型. 早期的一些研究利用贝叶斯后验估计方法来保持一个嵌套的队友信念表示. 常用的显式类型表示主要有硬编码和决策树等方法,隐式类型表示主要有基于神经网络的学习编码方法[264]. 经验回放法通过存储交互中的转换状态,根据当前观测的转换来识别当前的队友[265]. 任务识别法将先验知识或专家经验编码包含宏观行动、可选项库[266]. 在组队协同方面,智能体需要利用同伴信息来提升己方决策. 常用的队友识别方法主要有类型推理、经验识别和任务识别等方法. 其中,类型推理方法基于交互历史信息利用贝叶斯推理出队友在“假设类型空间”上的信念. 经验识别方法与类型推理不同,主要采用相似性度量的方法. 任务识别方法将先验知识表示成任务,智能体需要推理队友当前任务. 此外,在行为选择方面,智能体需要根据组队知识和队友行为,围绕最大团队回报的目标确定己方行动. 在适应当前队友方面,智能体在交互过程中,可能接收到队友、环境、任务目标等新信息,需采用适变行动来提升协调性.

6 结 语

本文从博弈论的视角梳理了近年来计算机博弈中不完美信息博弈求解相关研究进展. 首先,结合计算机博弈、人工智能与博弈论发展历史概述了具标志性突破的里程碑事件以及新的评估基准,分析归纳了3大研究范式,提出了序贯不完美信息博弈求解研究框架,围绕序贯不完美信息博弈介绍了3种博弈模型构建方法、子博弈和元博弈以及相关解概念和评估方法;然后,主体部分对当前离线策略求解和在线策略求解各3大类9小类方法进行了全面介绍和分析;最后,分析了当前面临的3类挑战,展望了5个方面的前沿研究重点. 当前,算法博弈论、优化理论和博弈学习理论的相互融合发展,为智能博弈对抗策略求解的研究提供了土壤,博弈策略通用求解方法已然成为序贯非完美信息博弈研究的核心,可为当前开展计算机博弈智能AI设计以及通用人工智能技术提供新技术、新思路和新途径.

参考文献(References)

[1] Schaeffer J. One jump ahead: Challenging human supremacy in checkers[M]. New York: Springer Science & Business Media, 2013: 456.

[2] Newborn M. Kasparov versus Deep Blue: Computer chess comes of age[M]. Heidelberg: Springer-Verlag, 1997: 235-278.

[3] Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484-489.

[4] Bowling M, Burch N, Johanson M, et al. Heads-up limit hold'em poker is solved[J]. Science, 2015, 347(6218): 145-149.

[5] Tian Y, Gong Q, Jiang Y. Joint policy search for multi-agent collaboration with imperfect information[C]. Proceedings of the 33rd International Conference on Neural Information Processing Systems. New Orleans, 2020: 19931-19942.

[6] Ensmenger N. Is chess the drosophila of artificial intelligence? A social history of an algorithm[J]. Social Studies of Science, 2012, 42(1): 5-30.

[7] Billings D, Papp D, Schaeffer J, et al. Poker as a testbed for AI research[C]. Conference of the Canadian Society for Computational Studies of Intelligence. Heidelberg, 1998: 228-238.

[8] Vinyals O, Babuschkin I, Czarnecki W M, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning[J]. Nature, 2019, 575(7782): 350-354.

[9] Brown N, Sandholm T. Superhuman AI for multiplayer poker[J]. Science, 2019, 365(6456): 885-890.

[10] Xin B G, Ma J H, Gao Q. The complexity of an investment competition dynamical model with imperfect information in a security market[J]. Chaos, Solitons & Fractals, 2009, 42(4): 2425-2438.

[11] Bichler M, Fichtl M, Heidekrüger S, et al. Learning equilibria in symmetric auction games using artificial neural networks[J]. Nature Machine Intelligence, 2021, 3(8): 687-695.

[12] Henderson H. Cybered competition, cooperation, and conflict in a game of imperfect information[J]. The Cyber Defense Review, 2021, 6(3): 43-60.

[13] Andrew J K. Operational decision making under uncertainty: Inferential, sequential, and adversarial approaches[D]. London: Technical Report, Air Force Institute of Technology Wright-Patterson AFB OH, 2019: 173-215.

[14] Ho E, Rajagopalan A, Skvortsov A, et al. Game theory in defence applications: A review[J]. Sensors, 2022, 22(3): 1032.

[15] Moss S. DARPA gamebreaker aims to train military AI systems on open world video games[EB/OL]. (2020-06-05)[2021-12-04]. https://aibusiness.com/author.asp?section_id=789&doc_id=761294.

[16] Uppal R. DARPA's SI3-CMD to enhance military decision making through artificial intelligence and computational game theory[EB/OL]. (2020-09-27)[2021-12-04]. https://idstch.com/technology/ict/darpas-si 3-cmd-to-enhance-military-decision-making-through-artificial-intelligence-and-computational-game-theory/.

[17] 徐心和, 邓志立, 王骄, 等. 机器博弈研究面临的各种挑战[J]. 智能系统学报, 2008, 3(4): 288-293.
(Xu X H, Deng Z L, Wang J, et al. Challenging issues facing computer game research[J]. CAAI Transactions on Intelligent Systems, 2008, 3(4): 288-293.)

[18] 王亚杰, 邱虹坤, 吴燕燕, 等. 计算机博弈的研究与发展[J]. 智能系统学报, 2016, 11(6): 788-798.
(Wang Y J, Qiu H K, Wu Y Y, et al. Research and development of computer games[J]. CAAI Transactions on Intelligent Systems, 2016, 11(6): 788-798.)

[19] 吴军, 徐昕, 王健, 等. 面向多机器人系统的增强学习研究进展综述[J]. 控制与决策, 2011, 26(11): 1601-1610.
(Wu J, Xu X, Wang J, et al. Recent advances of reinforcement learning in multi-robot systems: A survey[J]. Control and Decision, 2011, 26(11): 1601-1610.)

[20] 王军, 曹雷, 陈希亮, 等. 多智能体博弈强化学习研究综述[J]. 计算机工程与应用, 2021, 57(21): 1-13.
(Wang J, Cao L, Chen X L, et al. Overview on reinforcement learning of multi-agent game[J]. Computer Engineering and Applications, 2021, 57(21): 1-13.)

[21] 黄凯奇, 兴军亮, 张俊格, 等. 人机对抗智能技术[J]. 中国科学: 信息科学, 2020, 50(4): 540-550.
(Huang K Q, Xing J L, Zhang J G, et al. Intelligent technologies of human-computer gaming[J]. Scientia Sinica: Informationis, 2020, 50(4): 540-550.)

[22] 程代展, 刘泽群. 有限博弈的矩阵半张量积方法[J]. 控制理论与应用, 2019, 36(11): 1812-1819.
(Cheng D Z, Liu Z Q. Application of semi-tensor product of matrices to finite games[J]. Control Theory & Applications, 2019, 36(11): 1812-1819.)

[23] 唐振韬, 梁荣钦, 朱圆恒, 等. 实时格斗游戏的智能决策方法[J]. 控制理论与应用, 2022, 39(6): 969-985.
(Tang Z T, Liang R Q, Zhu Y H, et al. Intelligent decision making approaches for real time fighting game[J]. Control Theory & Applications, 2022, 39(6): 969-985.)

[24] 袁唯淋, 廖志勇, 高巍, 等. 计算机扑克智能博弈研究综述[J]. 网络与信息安全学报, 2021, 7(5): 57-76.
(Yuan W L, Liao Z Y, Gao W, et al. Survey on intelligent game of computer poker[J]. Chinese Journal of Network and Information Security, 2021, 7(5): 57-76.)

[25] McCulloch W S, Pitts W. A logical calculus of the ideas immanent in nervous activity[J]. The Bulletin of Mathematical Biophysics, 1943, 5(4): 115-133.

[26] Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain[J]. Psychological Review, 1958, 65(6): 386-408.

[27] LeCun Y, Boser B, Denker J S, et al. Backpropagation

applied to handwritten zip code recognition[J]. Neural Computation, 1989, 1(4): 541-551.

[28] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7): 1527-1554.

[29] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]. Proeedings of the 27th International Conference on Neural Information Processing Systems. Montreal, 2014: 2672−2680.

[30] Floridi L, Chiriatti M. GPT-3: Its nature, scope, limits, and consequences[J]. Minds and Machines, 2020, 30(4): 681-694.

[31] Moravčík M, Schmid M, Burch N, et al. DeepStack: Expert-level artificial intelligence in heads-up no-limit poker[J]. Science, 2017, 356(6337): 508-513.

[32] Li J J, Koyamada S, Ye Q W, et al. Suphx: Mastering mahjong with deep reinforcement learning[J/OL]. 2020, arXiv: 2003.13590.

[33] Zha D, Xie J, Ma W, et al. Douzero: Mastering doudizhu with self-play deep reinforcement learning[C]. International Conference on Machine Learning. Vienna, 2021: 12333-12344.

[34] Gray J, Lerer A, Bakhtin A, et al. Human-level performance in no-press diplomacy via equilibrium search[C]. International Conference on Learning Representations. Addis Ababa, 2020: 456-459.

[35] Cui B, Hu H, Pineda L, et al. K-level reasoning for zero-shot coordination in Hanabi[C]. Proceedings of the 34th International Conference on Neural Information Processing Systems. Montreal, 2021: 8215-8228.

[36] Berner C, Brockman G, Chan B, et al. Dota 2 with large scale deep reinforcement learning[J/OL]. 2019, arXiv: 1912.06680.

[37] Ye D H, Liu Z, Sun M F, et al. Mastering complex control in MOBA games with deep reinforcement learning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(4): 6672-6679.

[38] Tang Z, Zhu Y, Zhao D, et al. Enhanced rolling horizon evolution algorithm with opponent model learning: Results for the fighting game AI competition[J]. IEEE Transactions on Games, 2020, 32(1): 56-65.

[39] Silver D, Hubert T, Schrittwieser J, et al. A general reinforcement learning algorithm that Masters chess, shogi, and Go through self-play[J]. Science, 2018, 362(6419): 1140-1144.

[40] Schrittwieser J, Antonoglou I, Hubert T, et al. Mastering Atari, Go, chess and shogi by planning with a learned model[J]. Nature, 2020, 588(7839): 604-609.

[41] 陈兴国, 俞扬. 强化学习及其在电脑围棋中的应用[J]. 自动化学报, 2016, 42(5): 685-695.
(Chen X G, Yu Y. Reinforcement learning and its application to the game of Go[J]. Acta Automatica Sinica, 2016, 42(5): 685-695.)

[42] 唐振韬, 邵坤, 赵冬斌, 等. 深度强化学习进展: 从AlphaGo到AlphaGo Zero[J]. 控制理论与应用, 2017, 34(12): 1529-1546.
(Tang Z T, Shao K, Zhao D B, et al. Recent progress of deep reinforcement learning: From AlphaGo to AlphaGo Zero[J]. Control Theory & Applications, 2017, 34(12): 1529-1546.)

[43] 郭圣明, 贺筱媛, 胡晓峰, 等. 军用信息系统智能化的挑战与趋势[J]. 控制理论与应用, 2016, 33(12): 1562-1571.
(Guo S M, He X Y, Hu X F, et al. Challenges and trends in intelligent military information system[J]. Control Theory & Applications, 2016, 33(12): 1562-1571.)

[44] Brown N, Sandholm T. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals[J]. Science, 2018, 359(6374): 418-424.

[45] Tan G Y, He Y Y, Xu H H, et al. Winning rate prediction model based on Monte Carlo tree search for computer Dou dizhu[J]. IEEE Transactions on Games, 2021, 13(2): 123-137.

[46] Jiang Q Q, Li K Z, Du B Y, et al. DeltaDou: Expert-level doudizhu AI through self-play[C]. Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, 2019: 1265-1271.

[47] 李淑琴, 陈子鹏, 郑蓝舟, 等. 竞技二打一游戏中同等牌力的研究[J]. 智能系统学报, 2021, 16(3): 466-473.
(Li S Q, Chen Z P, Zheng L Z, et al. Research on the equal card force competition system of competitive two against one game[J]. CAAI Transactions on Intelligent Systems, 2021, 16(3): 466-473.)

[48] Fu H, Liu W, Wu S, et al. Actor-critic policy optimization in a large-scale imperfect-information game[C]. International Conference on Learning Representations. Vienna, 2021: 568-570.

[49] 王亚杰, 乔继林, 梁凯, 等. 结合先验知识与蒙特卡罗模拟的麻将博弈研究[J]. 智能系统学报, 2022, 17(1): 69-78.
(Wang Y J, Qiao J L, Liang K, et al. Research on mahjong game based on prior knowledge and Monte Carlo simulation[J]. CAAI Transactions on Intelligent Systems, 2022, 17(1): 69-78.)

[50] Sonzogni S. Depth-limited approaches in adversarial team games[D]. Milan: Politecnico Di Milano, 2020: 69-78.

[51] Derin D. Strategy refinement in adversarial team games[D]. Milan: Politecnico Di Milano, 2022: 32-45.

[52] Basilico N, Celli A, de Nittis G, et al. Computing the team-maxmin equilibrium in single-team single-adversary team games[J]. Intelligenza Artificiale, 2017, 11(1): 67-79.

[53] Farina G, Celli A, Gatti N, et al. Connecting optimal ex-ante collusion in teams to extensive-form correlation: Faster algorithms and positive complexity results[C]. International Conference on Machine Learning. Vienna, 2021: 3164-3173.

[54] Zhang B H, Farina G, Sandholm T. Team belief DAG form: A concise representation for team-correlated

game-theoretic decision making[C]. ICLR Workshop on Gamification and Multiagent Solutions. Vienna, 2022: 987-995.

[55] Carminati L, Cacciamani F, Ciccone M, et al. Public information representation for adversarial team games[C]. In NeurIPS Cooperative AI Workshop. Montreal, 2021: 654-659.

[56] Carminati L, Cacciamani F, Ciccone M, et al. A marriage between adversarial team games and 2-player games: Enabling abstractions, no-regret learning, and subgame solving[C]. International Conference on Machine Learning. Baltimore, 2022: 2638-2657.

[57] Paquette P, Lu Y, Bocco S S, et al. No-press diplomacy: Modeling multi-agent gameplay[C]. Proceedings of the 32nd International Conference on Neural Information Processing Systems. Vancouver, 2019: 4474-4485.

[58] Anthony T, Eccles T, Tacchetti A, et al. Learning to play no-press diplomacy with best response policy iteration[C]. Proceedings of the 33rd International Conference on Neural Information Processing Systems. New Orleans, 2020: 17987-18003.

[59] Jacob A P, Wu D J, Farina G, et al. Modeling strong and human-like gameplay with KL-regularized search[C]. International Conference on Machine Learning. Baltimore, 2022: 9695-9728.

[60] Gemp I, Savani R, Lanctot M, et al. Sample-based approximation of Nash in large many-player games via gradient descent[C]. Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems. Auckland, 2022: 507-515.

[61] Marris L, Muller P, Lanctot M, et al. Multi-agent training beyond zero-sum with correlated equilibrium meta-solvers[C]. International Conference on Machine Learning. Vienna, 2021: 7480-7491.

[62] Hu H, Foerster J N. Simplified action decoder for deep multi-agent reinforcement learning[C]. The 8th International Conference on Learning Representations. Addis Ababa, 2020: 1-14.

[63] Lerer A, Hu H, Foerster J, et al. Improving policies via search in cooperative partially observable games[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 7187-7194.

[64] Nekoei H, Badrinaaraayanan A, Courville A, et al. Continuous coordination as a realistic scenario for lifelong learning[C]. International Conference on Machine Learning. Vienna, 2021: 8016-8024.

[65] Hu H, Lerer A, Cui B, et al. Off-belief learning[C]. International Conference on Machine Learning. Vienna, 2021: 4369-4379.

[66] Siu H C, Peña J, Chen E, et al. Evaluation of human-AI teams for learned and rule-based agents in Hanabi[C]. Proceedings of the 34th International Conference on Neural Information Processing Systems. Montreal, 2021: 16183-16195.

[67] Blackwell D. An analog of the minimax theorem for vector payoffs[J]. Pacific Journal of Mathematics, 1956, 6(1): 1-8.

[68] Hannan J. Approximation to Bayes risk in repeated play[J]. Contributions to the Theory of Games, 1957, 3(2): 97-139.

[69] Fudenberg D, Drew F, Levine D K, et al. The theory of learning in games[M]. New York: MIT Press, 1998: 56.

[70] Zinkevich M, Johanson M, Bowling M, et al. Regret minimization in games with incomplete information[C]. Proceedings of the 20th International Conference on Neural Information Processing Systems. Whistler, 2007: 1729-1736.

[71] Hart S, Mas-Colell A. Simple adaptive strategies: From regret-matching to uncoupled dynamics[M]. New York: World Scientific, 2013: 198.

[72] Brown G W. Iterative solution of games by fictitious play[J]. Activity Analysis of Production and Allocation, 1951, 13(1): 374-376.

[73] Littman M L. Markov games as a framework for multi-agent reinforcement learning[C]. Proceedings of the 7th International Conference on International Conference on Machine Learning. New Brunswick, 1994: 157-163.

[74] Hu J, Wellman M P. Nash Q-learning for general-sum stochastic games[J]. Journal of Machine Learning Research, 2003, 4(11): 1039-1069.

[75] Heinrich J. Reinforcement learning from self-play in imperfect-information games[D]. London: University College London, 2017: 64-93.

[76] Hennes D, Morrill D, Omidshafiei S, et al. Neural replicator dynamics: Multiagent learning via hedging policy gradients[C]. Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems. Auckland, 2020: 492-501.

[77] Morrill D, D'Orazio R, Lanctot M, et al. Efficient deviation types and learning for hindsight rationality in extensive-form games[C]. International Conference on Machine Learning. Vienna, 2021: 7818-7828.

[78] Mathieu M, Ozair S, Srinivasan S, et al. StarCraft II unplugged: Large scale offline reinforcement learning[C]. Deep RL Workshop NeurIPS. Montreal, 2021: 951-958.

[79] Ganzfried S, Sandholm T. Safe opponent exploitation[J]. ACM Transactions on Economics and Computation, 2015, 3(2): 1-28.

[80] Johanson M, Bowling M. Data biased robust counter strategies[C]. Artificial Intelligence and Statistics. Clearwater, 2009: 264-271.

[81] Hoda S, Gilpin A, Peña J, et al. Smoothing techniques for computing Nash equilibria of sequential games[J]. Mathematics of Operations Research, 2010, 35(2): 494-512.

[82] Farina G, Sandholm T. Model-free online learning in unknown sequential decision making problems and games[J]. Proceedings of the AAAI Conference on

Artificial Intelligence, 2021, 35(6): 5381-5390.

[83] Kovařík V, Schmid M, Burch N, et al. Rethinking formal models of partially observable multiagent decision making[J]. Artificial Intelligence, 2022, 303: 103645.

[84] Schmid M, Moravcik M, Burch N, et al. Player of games[J/OL]. 2021, arXiv: 2112.03178.

[85] Burch N, Johanson M, Bowling M. Solving imperfect information games using decomposition[C]. The 28th AAAI Conference on Artificial Intelligence. Quebec, 2014: 602-608.

[86] Zhang J J, Liu H. Building endgame data set to improve opponent modeling approach[C]. IEEE the 2nd International Conference on Data Science in Cyberspace. Shenzhen, 2017: 255-260.

[87] Wellman M P. Methods for empirical game-theoretic analysis[C]. Proceedings of the 21st National Conference on Artificial Intelligence. Boston, 2006: 1552-1555.

[88] Balduzzi D, Garnelo M, Bachrach Y, et al. Open-ended learning in symmetric zero-sum games[C]. International Conference on Machine Learning. Long Beach, 2019: 434-443.

[89] Daskalakis C, Goldberg P W, Papadimitriou C H. The complexity of computing a Nash equilibrium[J]. SIAM Journal on Computing, 2009, 39(1): 195-259.

[90] Chen X, Deng X, Teng S H. Settling the complexity of computing two-player Nash equilibria[J]. Journal of the ACM, 2009, 56(3): 1-57.

[91] Koller D, Megiddo N. The complexity of two-person zero-sum games in extensive form[J]. Games and Economic Behavior, 1992, 4(4): 528-552.

[92] Bosansky B, Cermak J, Horak K, et al. Computing maxmin strategies in extensive-form zero-sum games with imperfect recall[J]. Proceedings of the 9th International Conference on Agents and Artificial Intelligence, 2017, 1: 63-74.

[93] von Stengel B, Forges F. Extensive-form correlated equilibrium: Definition and computational complexity[J]. Mathematics of Operations Research, 2008, 33(4): 1002-1022.

[94] Celli A, Marchesi A, Farina G, et al. No-regret learning dynamics for extensive-form correlated equilibrium[C]. Proceedings of the 33rd International Conference on Neural Information Processing Systems. New Orleans, 2020: 7722-7732.

[95] Farina G, Bianchi T, Sandholm T. Coarse correlation in extensive-form games[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(2): 1934-1941.

[96] Bosansky B, Cermak J. Sequence-form algorithm for computing stackelberg equilibria in extensive-form games[C]. The 29th AAAI Conference on Artificial Intelligence. Austin, 2015: 805-811.

[97] Černý J, Boýanský B, Kiekintveld C. Incremental strategy generation for stackelberg equilibria in extensive-form games[C]. Proceedings of the ACM Conference on Economics and Computation. New York, 2018: 151-168.

[98] Cermak J, Bosansky B, Durkota K, et al. Using correlated strategies for computing stackelberg equilibria in extensive-form games[J]. Proceedisn of the AAAI Conference on Artificial Intelligence, 2016, 30(1): 439-445.

[99] Kroer C, Farina G, Sandholm T. Robust stackelberg equilibria in extensive-form games and extension to limited lookahead[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1): 1130-1137.

[100] Černý J, Lisý V, Bošanský B, et al. Computing quantal stackelberg equilibrium in extensive-form games[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(6): 5260-5268.

[101] Billings D, Burch N, Davidson A, et al. Approximating game-theoretic optimal strategies for full-scale poker[C]. Proceedings of the 18th International Joint Conference on Artificial Intelligence. San Francisco, 2003: 661-668.

[102] Southey F, Bowling M, Larson B, et al. Bayes' bluff: Opponent modelling in poker[C]. Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence. Edinburgh, 2005: 550-558.

[103] Sandholm T. Perspectives on multiagent learning[J]. Artificial Intelligence, 2007, 171(7): 382-391.

[104] Ponsen M, de Jong S, Lanctot M. Computing approximate Nash equilibria and robust best-responses using sampling[J]. Journal of Artificial Intelligence Research, 2011, 42: 575-605.

[105] Marchesi A, Gatti N. Trembling-hand perfection and correlation in sequential games[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(6): 5566-5574.

[106] Milec D, Černý J, Lisý V, et al. Complexity and algorithms for exploiting quantal opponents in large two-player games[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(6): 5575-5583.

[107] Farina G, Sandholm T. Equilibrium refinement for the age of machines: The one-sided quasi-perfect equilibrium[C]. Proceedings of the 34th International Conference on Neural Information Processing Systems. Montreal, 2021: 8845-8856.

[108] Kreps D M, Wilson R. Sequential equilibria[J]. Econometrica: Journal of the Econometric Society, 1982, 50(4): 863-894.

[109] Selten R. Reexamination of the perfectness concept for equilibrium points in extensive games[J]. International Journal of Game Theory, 1975, 4(1): 25-55.

[110] Gatti N, Gilli M, Marchesi A. A characterization of quasi-perfect equilibria[J]. Games and Economic Behavior, 2020, 122: 240-255.

[111] Johanson M, Waugh K, Bowling M, et al. Accelerating best response calculation in large extensive games[C].

The 22nd International Joint Conference on Artificial Intelligence. Barcelona, 2011: 987-993.

[112] Lanctot M, Zambaldi V, Gruslys A, et al. A unified game-theoretic approach to multiagent reinforcement learning[C]. Proceedings of the 31st International Conference on Neural Information Processing Systems. Montreal, 2017: 4193-4206.

[113] Šustr M, Schmid M, Moravčík M, et al. Sound algorithms in imperfect information games[C]. Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems. London, 2021: 1674-1676.

[114] Schmid M, Burch N, Lanctot M, et al. Variance reduction in Monte Carlo counterfactual regret minimization for extensive form games using baselines[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 2157-2164.

[115] Cloud A, Laber E B. Variance decompositions for extensive-form games[C]. IEEE Conference on Games. Copenhagen, 2021: 1-8.

[116] Pankiewicz M, Bator M. Elo rating algorithm for the purpose of measuring task difficulty in online learning environments[J]. E-Mentor, 2019, 5(82): 43-51.

[117] Herbrich R, Minka T, Graepel T. TrueSkill^{TM}: A bayesian skill rating system[C]. Proceedings of the 19th International Conference on Neural Information Processing Systems. British Columbia, 2006: 569-576.

[118] Balduzzi D, Tuyls K, Perolat J, et al. Re-evaluating evaluation[C]. Proceedings of the 32nd International Conference on Neural Information Processing Systems. Vancouver, 2018: 3272-3283.

[119] Omidshafiei S, Papadimitriou C, Piliouras G, et al. α-rank: Multi-agent evaluation by evolution[J]. Scientific Reports, 2019, 9(1): 1-29.

[120] Čermák J, Lisý V, Bošanský B. Automated construction of bounded-loss imperfect-recall abstractions in extensive-form games[J]. Artificial Intelligence, 2020, 282: 103248.

[121] Ganzfried S, Sandholm T. Potential-aware imperfect-recall abstraction with earth mover's distance in imperfect-information games[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2014, 28(1): 682-690.

[122] Sandholm T, Singh S. Lossy stochastic game abstraction with bounds[C]. Proceedings of the 13th ACM Conference on Electronic Commerce. Valencia, 2012: 880-897.

[123] Gilpin A, Sandholm T. Lossless abstraction of imperfect information games[J]. Journal of the ACM, 2007, 54(5): 25-es.

[124] Gilpin A. Algorithms for abstracting and solving imperfect information games[D]. Pittsburgh: Carnegie Mellon University, 2009: 23-106.

[125] Sandholm T. Abstraction for solving large incomplete-information games[J]. Proceedings of the 29th AAAI Conference on Artificial Intelligence, 2015, 29(1): 4127-4131.

[126] Gibson R. Regret minimization in games and the development of champion multiplayer computer poker-playing agents[D]. Edmonton: University of Alberta, 2014: 54-76.

[127] Abernethy J, Bartlett P L, Hazan E. Blackwell approachability and no-regret learning are equivalent[C]. Proceedings of the 24th Annual Conference on Learning Theory. Budapest, 2011: 27-46.

[128] Farina G, Kroer C, Sandholm T. Faster game solving via predictive Blackwell approachability: Connecting regret matching and mirror descent[J]. AAAI Workshop on Reinforcement Learning in Games, 2021, 35(6): 5363-5371.

[129] Srinivasan S, Lanctot M, Zambaldi V, et al. Actor-critic policy optimization in partially observable multiagent environments[C]. Proceedings of the 32nd International Conference on Neural Information Processing Systems. Vancouver, 2018: 3426-3439.

[130] Burch N, Moravcik M, Schmid M. Revisiting CFR+ and alternating updates[J]. Journal of Artificial Intelligence Research, 2019, 64: 429-443.

[131] Lanctot M, Waugh K, Zinkevich M, et al. Monte Carlo sampling for regret minimization in extensive games[C]. Proceedings of the 22nd International Conference on Neural Information Processing Systems. Vancouver, 2009: 1078-1086.

[132] D'Orazio R, Morrill D, Wright J R, et al. Alternative function approximation parameterizations for solving games: An analysis of f-regression counterfactual regret minimization[C]. Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems. Auckland, 2020: 339-347.

[133] von Stengel B. Efficient computation of behavior strategies[J]. Games and Economic Behavior, 1996, 14(2): 220-246.

[134] Zhang B, Sandholm T. Sparsified linear programming for zero-sum equilibrium finding[C]. International Conference on Machine Learning. Honolulu, 2020: 11256-11267.

[135] Gilpin A, Pena J, Sandholm T W. First-order algorithm with $O(\ln(1/\epsilon))$ convergence for ϵ-equilibrium in two-person zero-sum games[C]. Proceedings of the 23rd National Conference on Artificial Intelligence. Chicago, 2008: 75-82.

[136] Nemirovski A. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems[J]. SIAM Journal on Optimization, 2004, 15(1): 229-251.

[137] Munos R, Perolat J, Lespiau J B, et al. Fast computation of Nash equilibria in imperfect information games[C]. International Conference on Machine Learning. Honolulu, 2020: 7119-7129.

[138] Ben-Tal A, Nemirovski A. Lectures on modern convex optimization: Analysis, algorithms, and engineering applications[M]. New York: Society for Industrial and Applied Mathematics, 2001: 568.

[139] Yu N. Excessive gap technique in nonsmooth convex minimization[J]. SIAM Journal on Optimization, 2005, 16(1): 235-249.

[140] Liu W, Jiang H, Li B, et al. Equivalence analysis between counterfactual regret minimization and online mirror descent[C]. International Conference on Machine Learning. Baltimore, 2022: 13717-13745.

[141] Leslie D S, Collins E J. Generalised weakened fictitious play[J]. Games and Economic Behavior, 2006, 56(2): 285-298.

[142] Heinrich J, Lanctot M, Silver D. Fictitious self-play in extensive-form games[C]. International Conference on Machine Learning. Lille, 2015: 805-813.

[143] Steinberger E, Lerer A, Brown N. DREAM: Deep regret minimization with advantage baselines and model-free learning[J/OL]. 2020, arXiv: 2006.10410.

[144] Brown N, Bakhtin A, Lerer A, et al. Combining deep reinforcement learning and search for imperfect-information games[J]. Proceedings of the 33rd International Conference on Neural Information Processing Systems. New Orleans, 2020: 17057-17069.

[145] Lockhart E, Lanctot M, Pérolat J, et al. Computing approximate equilibria in sequential adversarial games by exploitability descent[C]. Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, 2019: 464-470.

[146] Gruslys A, Lanctot M, Munos R, et al. The advantage regret-matching actor-critic[J/OL]. 2020, arXiv: 2008.12234.

[147] Liu X, Jia H, Wen Y, et al. Towards unifying behavioral and response diversity for open-ended learning in zero-sum games[C]. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, 2021: 941-952.

[148] Feng X, Slumbers O, Wan Z, et al. Neural auto-curricula in two-player zero-sum games[C]. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, 2021: 3504-3517.

[149] Gordon G J, Greenwald A, Marks C. No-regret learning in convex games[C]. Proceedings of the 25th International Conference on Machine Learning. Helsinki, 2008: 360-367.

[150] Celli A, Marchesi A, Farina G, et al. No-regret learning dynamics for extensive-form correlated equilibrium[C]. Proceedings of the 34th International Conference on Neural Information Processing Systems. Montreal, 2020: 7722-7732.

[151] Piliouras G, Rowland M, Omidshafiei S, et al. Evolutionary dynamics and phi-regret minimization in games[J]. Journal of Artificial Intelligence Research, 2022, 74: 1125-1158.

[152] Hart S, Mas-Colell A. A reinforcement procedure leading to correlated equilibrium[M]. Heidelberg: Springer, 2001: 181-200.

[153] Tammelin O, Burch N, Johanson M, et al. Solving heads-up limit texas hold'em[C]. The 24th International Joint Conference on Artificial Intelligence. Buenos Aires, 2015: 658-666.

[154] Littlestone N, Warmuth M K. The weighted majority algorithm[J]. Information and Computation, 1994, 108(2): 212-261.

[155] Auer P, Cesa-Bianchi N, Freund Y, et al. Gambling in a rigged casino: The adversarial multi-armed bandit problem[C]. Proceedings of IEEE 36th Annual Foundations of Computer Science. Milwaukee, 1995: 322-331.

[156] Cesa-Bianchi N, Lugosi G. Prediction, learning, and games[M]. New York: Cambridge University Press, 2006: 658.

[157] Farina G, Kroer C, Sandholm T. Regret circuits: Composability of regret minimizers[C]. International Conference on Machine Learning. Long Beach, 2019: 1863-1872.

[158] Brown N, Sandholm T. Solving imperfect-information games via discounted regret minimization[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 1829-1836.

[159] Farina G, Kroer C, Sandholm T. Stochastic regret minimization in extensive-form games[C]. International Conference on Machine Learning. Honolulu, 2020: 3018-3028.

[160] Gibson R, Lanctot M, Burch N, et al. Generalized sampling and variance in counterfactual regret minimization[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 26(1): 1355-1361.

[161] Davis T, Schmid M, Bowling M. Low-variance and zero-variance baselines for extensive-form games[C]. International Conference on Machine Learning. Honolulu, 2020: 2392-2401.

[162] Greenwald A, Li Z, Marks C. Bounds for regret-matching algorithms[C]. ISAIM. Fort Lauderdale, 2006: 546-559.

[163] Waugh K, Morrill D, Bagnell J A, et al. Solving games with functional regret estimation[J]. Proceedings of the 29th AAAI Conference on Artificial Intelligence, 2015, 29(1): 2138-2144.

[164] Li H, Hu K, Zhang S, et al. Double neural counterfactual regret minimization[C]. International Conference on Learning Representations. Long Beach, 2019: 852-859.

[165] Xu H, Li K, Fu H, et al. AutoCFR: Learning to design counterfactual regret minimization algorithms[C]. AAAI Conference on Artificial Intelligence. Vancouver, 2022: 123-131.

[166] Liu W M, Li B, Togelius J. Model-free neural counterfactual regret minimization with bootstrap learning[J]. IEEE Transactions on Games, 2022, 48(2):

89-96.

[167] Koller D, Megiddo N, von Stengel B. Efficient computation of equilibria for extensive two-person games[J]. Games and Economic Behavior, 1996, 14(2): 247-259.

[168] Farina G, Ling C K, Fang F, et al. Correlation in extensive-form games: Saddle-point formulation and benchmarks[C]. Proceedings of the 33rd International Conference on Neural Information Processing Systems. New Orleans, 2019: 9233-9243.

[169] Grand-Clément J, Kroer C. Conic Blackwell algorithm: Parameter-free convex-concave saddle-point solving[C]. Proceedings of the 34th International Conference on Neural Information Processing Systems. Montreal, 2021: 9587-9599.

[170] Gilpin A, Hoda S, Pena J, et al. Gradient-based algorithms for finding Nash equilibria in extensive form games[C]. International Workshop on Web and Internet Economics. Heidelberg, 2007: 57-69.

[171] Azizian W, Mitliagkas I, Lacoste-Julien S, et al. A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games[C]. International Conference on Artificial Intelligence and Statistics. Palermo, 2020: 2863-2873.

[172] Rakhlin S, Sridharan K. Optimization, learning, and games with predictable sequences[C]. Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, 2013: 3066-3074.

[173] Juditsky A, Nemirovski A, Tauvel C. Solving variational inequalities with stochastic mirror-prox algorithm[J]. Stochastic Systems, 2011, 1(1): 17-58.

[174] Kroer C, Farina G, Sandholm T. Solving large sequential games with the excessive gap technique[C]. Proceedings of the 31st International Conference on Neural Information Processing Systems. Montreal, 2018: 872-882.

[175] Farina G, Kroer C, Sandholm T. Optimistic regret minimization for extensive-form games via dilated distance-generating functions[C]. Proceedings of the 32nd International Conference on Neural Information Processing Systems. Vancouver, 2019: 5221-5231.

[176] Farina G, Kroer C, Sandholm T. Better regularization for sequential decision spaces: Fast convergence rates for Nash, correlated, and team equilibria[C]. Proceedings of the 22nd ACM Conference on Economics and Computation. New York, 2021: 432.

[177] Syrgkanis V, Agarwal A, Luo H, et al. Fast convergence of regularized learning in games[C]. Proceedings of the 28th International Conference on Neural Information Processing Systems. Monteria, 2015: 2989-2997.

[178] Waugh K, Bagnell J A. A unified view of large-scale zero-sum equilibrium computation[C]. Workshops at the 29th AAAI Conference on Artificial Intelligence. Austin, 2015: 456-459.

[179] Mitchell T M, Mitchell T M. Machine learning[M]. New York: McGraw-Hill, 1997: 65.

[180] Chen Y, Zhang L, Li S, et al. Optimize neural fictitious self-play in regret minimization thinking[J/OL]. 2021, arXiv: 2104.10845.

[181] Zhang L, Chen Y X, Wang W, et al. A Monte Carlo neural fictitious self-play approach to approximate Nash equilibrium in imperfect-information dynamic games[J]. Frontiers of Computer Science, 2021, 15(5): 1-14.

[182] Timbers F, Bard N, Lockhart E, et al. Approximate exploitability: Learning a best response[C]. Proceedings of the 31st International Joint Conference on Artificial Intelligence. Vienna, 2022: 3487-3493.

[183] McMahan H B, Gordon G J, Blum A. Planning in the presence of cost functions controlled by an adversary[C]. Proceedings of the 20th International Conference on Machine Learning. Honolulu, 2003: 536-543.

[184] Muller P, Omidshafiei S, Rowland M, et al. A generalized training approach for multiagent learning[C]. The 8th International Conference on Learning Representations. Addis Ababa, 2020: 1-35.

[185] McAleer S, Lanier J B, Fox R, et al. Pipeline psro: A scalable approach for finding approximate nash equilibria in large games[C]. Proceedings of the 33rd International Conference on Neural Information Processing Systems. New Orleans, 2020: 20238-20248.

[186] McAleer S, Lanier J B, Wang K, et al. XDO: A double oracle algorithm for extensive-form games[C]. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, 2021: 23128-23139.

[187] Liu S, Lanctot M, Marris L, et al. Simplex neural population learning: Any-mixture Bayes-optimality in symmetric zero-sum games[J/OL]. 2022, arXiv: 2205.15879.

[188] Nashed S, Zilberstein S. A survey of opponent modeling in adversarial domains[J]. Journal of Artificial Intelligence Research, 2022, 73: 277-327.

[189] Ekmekci O, Sirin V. Learning strategies for opponent modeling in poker[C]. Workshops at the 27th AAAI Conference on Artificial Intelligence. Washington, 2013: 111-119.

[190] Foerster J N, Chen R Y, Al-Shedivat M, et al. Learning with opponent-learning awareness[C]. Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems. Stockholm, 2018: 122-130.

[191] Hernandez-leal P, Rosman B, Taylor M E, et al. A Bayesian approach for learning and tracking switching, non-stationary opponents[C]. Proceedings of the International Conference on Autonomous Agents & Multiagent Systems. Singapore, 2016: 1315-1316.

[192] Li X, Miikkulainen R. Opponent modeling and exploitation in poker using evolved recurrent neural networks[C]. Proceedings of the Genetic and Evolutionary Computation Conference. Jiangsu, 2018: 189-196.

[193] Ganzfried S. Computing strong game-theoretic strategies and exploiting suboptimal opponents in large games[D]. Pittsburgh: Carnegie Mellon University, 2015: 33-218.

[194] Davis T, Waugh K, Bowling M. Solving large extensive-form games with strategy constraints[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 1861-1868.

[195] Kozuno T, Ménard P, Munos R, et al. Learning in two-player zero-sum partially observable Markov games with perfect recall[C]. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, 2021: 11987-11998.

[196] Šustr M, Kovařík V, Lisý V. Monte Carlo continual resolving for online strategy computation in imperfect information games[C]. Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems. Montreal, 2019: 224-232.

[197] Brown N, Sandholm T. Safe and nested subgame solving for imperfect-information games[C]. Proceedings of the 31st International Conference on Neural Information Processing Systems. Montreal, 2017: 689-699.

[198] 张蒙, 李凯, 吴哲, 等. 一种针对德州扑克AI的对手建模与策略集成框架[J]. 自动化学报, 2022, 48(4): 1004-1017.
(Zhang M, Li K, Wu Z, et al. An opponent modeling and strategy integration framework for texas hold'em[J]. Acta Automatica Sinica, 2022, 48(4): 1004-1017.)

[199] Letcher A, Foerster J, D Balduzzi, et al. Stable opponent shaping in differentiable games[C]. The 7th International Conference on Learning Representations. New Orleans, 2019: 159-168.

[200] Rusch T, Steixner-Kumar S, Doshi P, et al. Theory of mind and decision science: Towards a typology of tasks and computational models[J]. Neuropsychologia, 2020, 146(211): 107488.

[201] Doshi P, Qu X, Goodie A S, et al. Modeling human recursive reasoning using empirically informed interactive partially observable Markov decision processes[J]. IEEE Transactions on Systems, Man, and Cybernetics — Part A: Systems and Humans, 2012, 42(6): 1529-1542.

[202] Doshi P, Gmytrasiewicz P, Durfee E. Recursively modeling other agents for decision making: A research perspective[J]. Artificial Intelligence, 2020, 279: 103202.

[203] Zheng Y, Meng Z, Hao J, et al. A deep bayesian policy reuse approach against non-stationary agents[C]. Proceedings of the 32nd International Conference on Neural Information Processing Systems. Vancouver, 2018: 962-972.

[204] Yu X, Jiang J, Jiang H, et al. Model-based opponent modeling[J/OL]. 2021, arXiv: 2108.01843.

[205] Papoudakis G, Christianos F, Albrecht S. Agent modelling under partial observability for deep reinforcement learning[C]. The 35th International Conference on Neural Information Processing Systems. Vancouver, 2021: 19210-19222.

[206] Gibson R G, Szafron D. Regret minimization in multiplayer extensive games[C]. The 22nd International Joint Conference on Artificial Intelligence. Barcelona, 2011: 159-164.

[207] Albrecht S V, Stone P. Autonomous agents modelling other agents: A comprehensive survey and open problems[J]. Artificial Intelligence, 2018, 258: 66-95.

[208] Brown N, Sandholm T. Reduced space and faster convergence in imperfect-information games via pruning[C]. International Conference on Machine Learning. Sydney, 2017: 596-604.

[209] Farina G, Kroer C, Sandholm T. Regret minimization in behaviorally-constrained zero-sum games[C]. International Conference on Machine Learning. Sydney, 2017: 1107-1116.

[210] Farina G, Kroer C, Sandholm T. Online convex optimization for sequential decision processes and extensive-form games[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 1917-1925.

[211] Ganzfried S, Sandholm T, Game theory-based opponent modeling in large imperfect-information games[C]. The 10th International Conference on Autonomous Agents and Multiagent Systems. Taipei, 2011: 533-540.

[212] Bernasconi-de-Luca M, Cacciamani F, Fioravanti S, et al. Exploiting opponents under utility constraints in sequential games[C]. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, 2021: 13177-13188.

[213] Zhang J, Wang X, Yao L, et al. Using Kullback-Leibler divergence to model opponents in poker[C]. Workshops at the 28th AAAI Conference on Artificial Intelligence. Quebec, 2014: 159-166.

[214] Ganzfried S, Sun Q. Bayesian opponent exploitation in imperfect-information games[C]. IEEE Conference on Computational Intelligence and Games. Maastricht, 2018: 1-8.

[215] Wu Z, Li K, Zhao E, et al. L2E: Learning to exploit your opponent[J/OL]. 2021, arXiv: 2102.09381.

[216] Zhou Y, Li J, Zhu J. Posterior sampling for multi-agent reinforcement learning: Solving extensive games with imperfect information[C]. International Conference on Learning Representations. New Orleans, 2019: 987-989.

[217] Farina G, Sandholm T. Model-free online learning in unknown sequential decision making problems and games[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(6): 5381-5390.

[218] Bai Y, Jin C, Mei S, et al. Near-optimal learning of extensive-form games with imperfect information[C]. ICLR Workshop on Gamification and Multiagent Solutions. Vienna, 2022: 456-459.

[219] Meng L, Gao Y. Generalized bandit regret minimizer

framework in imperfect information extensive-form game[J/OL]. 2022, arXiv: 2203.05920.

[220] 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[J]. 计算机研究与发展, 2014, 51(10): 2160-2170.
(Hu Y J, Gao Y, An B. Online counterfactual regret minimization in repeated imperfect information extensive games[J]. Journal of Computer Research and Development, 2014, 51(10): 2160-2170.)

[221] Bitan M, Kraus S. Combining prediction of human decisions with ISMCTS in imperfect information games[C]. Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems. Stockholm, 2018: 1874-1876.

[222] Phan T. Searching with opponent-awareness[J/OL]. 2021, arXiv: 2104.10508.

[223] Moravcik M, Schmid M, Ha K, et al. Refining subgames in large imperfect information games[C]. Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix, 2016: 572-578.

[224] Liu M, Wu C, Liu Q, et al. Safe opponent-exploitation subgame refinement[C]. ICLR Workshop on Gamification and Multiagent Solutions. Vienna, 2022: 123-129.

[225] Milec D, Lisý V. Continual depth-limited responses for computing counter-strategies in extensive-form games[J/OL]. 2021, arXiv: 2112.12594.

[226] Hernandez-Leal P, EMD Cote, Sucar L E. Using a priori information for fast learning against non-stationary opponents[C]. Ibero-American Conference on Artificial Intelligence. Santiago, 2014: 536-547.

[227] Everett R, Roberts S. Learning against non-stationary agents with opponent modelling and deep reinforcement learning[C]. AAAI Spring Symposium Series. Palo Alto, 2018: 456-469.

[228] Dinh L C, Mguni D H, Tran-Thanh L, et al. Online Markov decision processes with non-oblivious strategic adversary[J/OL]. 2021, arXiv: 2110.03604.

[229] Ferguson M, Devlin S, Kudenko D, et al. Player style clustering without game variables[C]. International Conference on the Foundations of Digital Games. New York, 2020: 1-4.

[230] Ganzfried S, Sandholm T. Computing equilibria by incorporating qualitative models?[C]. Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems. Toronto, 2010: 183-190.

[231] Brafman R I, Tennenholtz M. R-max-a general polynomial time algorithm for near-optimal reinforcement learning[J]. Journal of Machine Learning Research, 2002, 23(4): 213-231.

[232] Lakkaraju H, Kamar E, Caruana R, et al. Identifying unknown unknowns in the open world: Representations and policies for guided exploration[C]. The 31st AAAI Conference on Artificial Intelligence. Budapest, 2017: 2124-2132.

[233] Hernandez-Leal P, Kaisers M. Learning against sequential opponents in repeated stochastic games[C]. The 3rd Multi-Disciplinary Conference on Reinforcement Learning and Decision Making. Ann Arbor, 2017: 52-56.

[234] Hernandez-Leal P, Zhan Y S, Taylor M E, et al. Efficiently detecting switches against non-stationary opponents[J]. Autonomous Agents and Multi-Agent Systems, 2017, 31(4): 767-789.

[235] von Der O F B, Kirley M, Miller T. The minds of many: Opponent modelling in a stochastic game[C]. Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne, 2017: 3845-3851.

[236] Lanctot M, Waugh K, Zinkevich M, et al. Monte Carlo sampling for regret minimization in extensive games[C]. Proceedings of the 22nd International Conference on Neural Information Processing Systems. Vancouver, 2009: 1078-1086.

[237] Lee C W, Kroer C, Luo H. Last-iterate convergence in extensive-form games[C]. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, 2021: 14293-14305.

[238] Zha D, Lai K H, Cao Y, et al. Rlcard: A toolkit for reinforcement learning in card games[J/OL]. 2019, arXiv: 1910.04376.

[239] Lanctot M, Lockhart E, Lespiau J B, et al. OpenSpiel: A framework for reinforcement learning in games[J/OL]. 2019, arXiv: 1908.09453.

[240] Terry J K, Black B, Grammel N, et al. Pettingzoo: Gym for multi-agent reinforcement learning[C]. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, 2021: 15032-15043.

[241] Pretorius A, Tessera K, Smit A P, et al. Mava: A research framework for distributed multi-agent reinforcement learning[J/OL]. 2021, arXiv: 2107.01460.

[242] Zheng J F, Castañón D A. Dynamic network interdiction games with imperfect information and deception[C]. The 51st IEEE Conference on Decision and Control. Maui, 2012: 7758-7763.

[243] Greige L, Silva F D M, Trotter M, et al. Collusion detection in team-based multiplayer games[J/OL]. 2022, arXiv: 2203.05121.

[244] Sandholm T. Steering evolution strategically: Computational game theory and opponent exploitation for treatment planning, drug design, and synthetic biology[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2015, 29(1): 4057-4061.

[245] Vlatakis-Gkaragkounis E V, Flokas L, Piliouras G. Poincaré recurrence, cycles and spurious equilibria in gradient-descent-ascent for non-convex non-concave zero-sum games[C]. Proceedings of the 32nd International Conference on Neural Information Processing Systems. Vancouver, 2019: 10450-10461.

[246] Perolat J, Munos R, Lespiau J B, et al. From Poincaré

recurrence to convergence in imperfect information games: Finding equilibrium via regularization[C]. International Conference on Machine Learning. Vienna, 2021: 8525-8535.

[247] Czarnecki W M, Gidel G, Tracey B, et al. Real world games look like spinning tops[C]. Proceedings of the 33rd International Conference on Neural Information Processing Systems. New Orleans, 2020: 17443-17454.

[248] Cebra C, Strang A. Similarity suppresses cyclicity: Why similar competitors form hierarchies[J/OL]. 2022, arXiv: 2205.08015.

[249] Brockbank E, Vul E. Formalizing opponent modeling with the rock, paper, scissors game[J]. Games, 2021, 12(3): 70.

[250] Oh I, Rho S, Moon S, et al. Creating pro-level AI for a real-time fighting game using deep reinforcement learning[J]. IEEE Transactions on Games, 2022, 14(2): 212-220.

[251] Ding Z, Su D, Liu Q, et al. A deep reinforcement learning approach for finding non-exploitable strategies in two-player Atari games[J/OL]. 2022, arXiv: 2207.08894.

[252] Farina G, Lee C W, Luo H, et al. Kernelized multiplicative weights for 0/1-polyhedral games: Bridging the gap between learning in extensive-form and normal-form games[J/OL]. 2022, arXiv: 2202.00237.

[253] Li X, Meng M, Hong Y, et al. A survey of decision making in adversarial games[J/OL]. 2022, arXiv: 2207.07971.

[254] Zhang B, Sandholm T. Subgame solving without common knowledge[C]. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, 2021: 23993-24004.

[255] Farina G, Kroer C, Sandholm T. Online convex optimization for sequential decision processes and extensive-form games[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 1917-1925.

[256] Zhou M, Wan Z, Wang H, et al. Malib: A parallel framework for population-based multi-agent reinforcement learning[J/OL]. 2021, arXiv: 2106.07551.

[257] Boney R, Ilin A, Kannala J, et al. Learning to play imperfect-information games by imitating an oracle planner[J]. IEEE Transactions on Games, 2022, 14(2): 262-272.

[258] Li H, Wang X, Jia F, et al. RLCFR: Minimize counterfactual regret by deep reinforcement learning[J]. Expert Systems with Applications, 2022, 187: 115953.

[259] Wang X, Cerny J, Li S, et al. A unified perspective on deep equilibrium finding[J/OL]. 2022, arXiv: 2204.04930.

[260] McAleer S, Lanier J B, Wang K, et al. Self-play PSRO: Toward optimal populations in two-player zero-sum games[J/OL]. 2022, arXiv: 2207.06541.

[261] Li S, Wang X, Cerny J, et al. Offline equilibrium finding[J/OL]. 2022, arXiv: 2207.05285.

[262] Chen L, Lu K, Rajeswaran A, et al. Decision transformer: Reinforcement learning via sequence modeling[C]. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, 2021: 15084-15097.

[263] Li T, Zhao Y H, Zhu Q Y. The role of information structures in game-theoretic multi-agent learning[J]. Annual Reviews in Control, 2022, 53: 296-314.

[264] Zintgraf L, Devlin S, Ciosek K, et al. Deep interactive Bayesian reinforcement learning via meta-learning[C]. Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems. London, 2021: 1712-1714.

[265] Chen S, Andrejczuk E, Cao Z G, et al. AATEAM: Achieving the ad hoc teamwork by employing the attention mechanism[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 7095-7102.

[266] Wu S A, Wang R E, Evans J A, et al. Too many cooks: Bayesian inference for coordinating multi-agent collaboration[J]. Topics in Cognitive Science, 2021, 13(2): 414-432.

作者简介

罗俊仁(1989−), 男, 博士生, 从事不完美信息博弈、多智能体学习等研究, E-mail: luojunren17@nudt.edu.cn;

张万鹏(1981−), 男, 研究员, 博士, 从事大数据智能、智能演进等研究, E-mail: wpzhang@nudt.edu.cn;

苏炯铭(1984−), 男, 副研究员, 博士, 从事智能博弈、可解释性学习等研究, E-mail: sujiongming07@nudt.edu.cn;

魏婷婷(1997−), 女, 硕士生, 从事深度学习、对手建模等研究, E-mail: weitingting20@nudt.edu.cn;

陈璟(1972−), 男, 教授, 博士, 从事认知决策博弈、分布式智能等研究, E-mail: chenjing001@vip.sina.com.