非完美信息博弈综述:对抗求解方法与对比分析

余 超[1)] 刘宗凯[1)] 胡超豪[1)] 黄凯奇[2)] 张俊格[2)]

[1)](中山大学计算机学院 广州 510006)

[2)](中国科学院自动化研究所智能系统与工程研究中心 北京 100190)

摘 要 当前,人工智能成为经济发展的新引擎,是新一轮产业变革的核心驱动力.结合人工智能与博弈论形成的新兴研究领域"博弈智能"吸引了越来越多学者的研究兴趣,并在现实生活中得到了广泛应用.作为一类典型的博弈智能,非完美信息博弈通过建模多智能体在私有信息下的博弈行为,能够刻画相较完美信息博弈更广泛的决策过程,在现实世界中具有广泛应用,例如金融贸易、商业谈判、军事对抗等.近年来,非完美信息博弈求解研究取得了突破性进展,涌现出以遗憾最小化(Regret Minimization)和最佳响应(Best Response)为核心技术的两大类离线求解方法.前者通过反省智能体过往决策以使自身策略向均衡点改进,成功解决了以德州扑克为代表的经典非完美信息博弈.后者通过特定应对方式针对对手决策以使自身策略向均衡点改进,在例如星际争霸、DOTA等大型实时战略游戏AI训练中发挥着关键作用.此外,一系列在线求解方法能够进一步实时优化离线算法求解所得的蓝图策略,使其在实时对局中得到进一步改进,成为求解非完美信息博弈的关键技术.本文将从非完美信息博弈的概念和特点切入,全面介绍这三类方法的基本原理、发展脉络和改进技巧,深入对比不同方法间的优缺点并展望未来研究方向.希望通过对非完美信息博弈求解这一研究领域的全方位细致梳理,能够进一步推动博弈智能技术向前发展,为迈向通用人工智能赋能.

关键词 非完美信息博弈;遗憾最小化;最佳响应;在线求解;强化学习

中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2024.02211

A Review of Imperfect Information Games: Adversarial Solving Methods and Comparative Analysis

YU Chao[1)] LIU Zong-Kai[1)] HU Chao-Hao[1)] HUANG Kai-Qi[2)] ZHANG Jun-Ge[2)]

[1)](*School of Computer Science And Engineering, Sun Yat-Sen University, GuangZhou* 510000)

[2)](*Center for Research Intelligent System and Engineering, Institute of Automation, Chinese Academy of Sciences, Beijing* 100190)

Abstract Artificial Intelligence (AI) has emerged as a pivotal force in the latest industrial revolution and has become a national strategic priority. The fusion of AI and game theory has given rise to "Game Intelligence" as a leading research domain. Among the diverse facets of game intelligence, Imperfect-Information Games (IIGs) stand out for their ability to simulate the strategic decision-making of multiple agents amidst private information an accurate portrayal of many real-world scenarios. Compared to perfect-information games, IIGs offer a more nuanced understanding of decision-making processes, making them applicable across various real-world domains such as financial trading, business negotiations, and military operations. Recent strides in IIG research

收稿日期:2023-08-14;在线发布日期:2024-05-23.本课题得到国家自然科学基金面上项目(No.62076259)、广东省自然科学基金(No.2023A1515012946)、中国科学院基础培育基金项目(JCPYJJ-22017)、中山大学中央高校基本科研业务费专项资金、中国科学院青年促进会项目资助.**余超**(通信作者),博士,教授,中国计算机学会(CCF)会员,主要研究领域为博弈智能、强化学习、多智能体系统.E-mail:yuchao3@mail.sysu.edu.cn.**刘宗凯**,硕士研究生,主要研究领域为博弈智能、多智能体强化学习、非完美信息博弈.**胡超豪**,硕士研究生,主要研究领域为博弈智能、多智能体强化学习、非完美信息博弈.**黄凯奇**,博士,研究员,中国计算机学会(CCF)杰出会员,主要研究领域为计算机视觉、博弈决策智能技术.**张俊格**,博士,研究员,中国计算机学会(CCF)会员,主要研究领域为博弈智能、多智能体、模式识别.

have led to the emergence of two primary streams of offline solving methods: Regret Minimization and Best Response. Regret Minimization continually refines its strategy towards equilibrium by learning from past decisions, making it particularly advantageous in scenarios with unknown or uncertain opponent strategies. On the other hand, Best Response fine-tunes its strategy towards equilibrium by devising tailored countermeasures against opponents' decisions, proving pivotal in training AI for large-scale real-time strategy games like Starcraft and DOTA. The efficacy of the Best Response approach hinges on its ability to anticipate and counteract opponents' moves. Moreover, search-based online solving methods optimize blueprint strategies in real-time, facilitating precise Nash equilibrium solutions, constituting a critical technology in IIG solving. The synergy of offline and online solving methods equips AI with the capability to navigate the intricacies of IIGs and attain optimal solutions. This survey aims to provide a comprehensive exploration of the realm of IIGs. Beginning with an elucidation of IIGs' concept and their distinguishing features, the survey offers an overview of the methods employed for their resolution. Subsequently, it delves into the fundamental principles and historical context of these methods, alongside delineating advanced techniques to enhance their efficacy. Additionally, the survey conducts an exhaustive comparison of the strengths and weaknesses of various methods, while providing insights into future research trajectories. It is our aspiration that through this comprehensive scrutiny of IIGs, this survey will drive advancements in game intelligence technology and contribute to the development of artificial intelligence.

Keywords imperfect information game; regret minimization; best response; safe search; reinforcement learning

1 引 言

近年来,随着大数据、物联网、云计算等产业的纵深推进,人工智能(Artificial Intelligence,AI)技术迅速发展并与多种应用场景深度融合.一般而言,人工智能的首要目标是设计并实现能与环境动态交互、自动进行最优决策的算法智能体.然而,现实应用中通常存在多个互联互通的智能体共同作用于同一环境,使得单一智能体处于非稳态[1](Non-stationary)的环境之中,这给智能体的决策优化带来了极大挑战.作为研究多智能体交互以及理性决策的学科,博弈论研究通过对智能体的决策目标、交互规则进行高度抽象,求解满足预定需求的均衡解(Equilibria),使得智能体能够在非平稳环境中实现动态、平衡的最优决策[2].近年来,将人工智能与博弈论结合形成"博弈智能"这一新兴研究方向,成为当下 AI 研究的前沿领域[3].

根据智能体获取信息程度的区别,博弈可分为完美信息博弈(Perfect Information Game,PIG)和非完美信息博弈(Imperfect Information Game,IIG).完美信息博弈是指智能体能够在完全知晓过往所有信息的前提下进行决策,已在棋类游戏[4-7]、Atari 游戏[8-9]等应用领域取得了巨大成功.而在例如金融经济[10](谈判、竞价、交易)、社会管理[11](网络安全、电力市场)、军事国防[12-13](资源部署、对抗推演)等更广泛的非完美信息博弈场景中,智能体需在具有隐藏信息的前提下进行决策.由于智能体无法确切还原博弈全貌,完美信息博弈求解中的反向归纳原则(Backward Induction)和以蒙特卡洛树搜索[14](Monte-Carlo Tree Search,MCTS)为代表的求解方法无法直接应用于非完美信息博弈中,导致博弈均衡求解极具挑战.近年来,非完美信息博弈研究取得了突破性进展,多项成果荣登《Nature》《Science》,成功解决了以德州扑克为代表的纸牌游戏[15-17]、多人在线战术竞技游戏[18-20]、隐藏角色扮演游戏[21]、西洋陆军棋[22]等挑战性难题,成为下一代 AI 研究的前沿领域之一[16,23].然而,近年来随着实际应用的复杂度提升,非完美信息博弈的求解研究面临着诸多挑战,其中包括以下几方面:

(1)求解的理论性.随着非完美信息博弈规模和类型拓展,博弈环境中往往涉及两个以上智能体

的合作竞争关系，例如 Diplomacy[24]、斗地主[17]、Avalon[21]等，从理论上计算这类多人一般和博弈的纳什均衡属于 PPAD-complete 问题，目前学术界普遍认为不存在高效解法[25-27]，且多人博弈中均衡解的形式仍是一个尚待研究的重大科学问题[28]. 与以上问题不同，以双人零和博弈（Two player Zero-sum Game，2p0s Game）为代表的竞争型非完美信息博弈研究具备良好的理论性能保障，已被证明存在寻找最优解（即纳什均衡）的多项式时间算法，具备可解性[29]. 此外，2p0s 博弈问题的求解为其他更为复杂的非完美信息博弈问题提供了求解思路. 例如，在多人 Diplomacy 研究中，可部署 2p0s 博弈中具有收敛性保证的算法来改进博弈策略[30]；双人德州扑克的求解算法[31]直接应用于六人德州扑克中仍具有非凡表现[32]. 本文将在第 3、4 节介绍 2p0s 博弈中两类具有理论性保证的离线求解方法.

(2)求解的泛化性. 针对不同场景的非完美信息博弈，智能体在训练环境中容易对特定组合的对手出现过拟合现象，导致其在真实评估阶段表现不佳. 因此，在训练过程中，如何挖掘其他智能体可能出现的不同行为方式以丰富自身应对策略的多样性，是非完美信息博弈应用落地面临的一大挑战. 同时，在近年的研究中，如何在各类环境中评估智能体策略的泛化性也是一个热点研究方向[33]. 本文将在 4.3 节中介绍基于种群训练的通用博弈求解框架，并探讨其对策略泛化性与多样性的影响.

(3)求解的在线适应性. 在真实对弈的过程中，智能体往往面对的是黑盒环境[34]，缺少博弈整体的先验知识，这要求算法具备一定的在线探索能力，能对博弈局势进行动态分析和重构. 同时，在实际环境中，对弈的对手往往具有非理性的特征，如何在对弈进行同时，实时地在线改变自身策略以达到期望目标，是未来非完美信息博弈扩展至更大规模场景的一大挑战. 本文将在第 5 节介绍非完美信息博弈中常用的在线求解方法.

为了厘清非完美信息博弈中各类求解算法的内核和特性，以期推动更复杂规模博弈中可靠算法的设计，本文围绕序贯双人零和博弈展开论述，分别从离线和在线角度介绍一系列求解非完美信息博弈的算法. 本文的具体内容组织如图 1 所示：第 2 节将首先梳理非完美信息博弈相关的背景知识；第 3 节将介绍基于遗憾最小化的离线求解方法，并对三类求解方法进行对比分析；第 4 节将介绍基于最佳响应的离线求解方法，在分析三类最佳响应求解方式的同时，对上述两类离线求解方法进行对比；第 5 节将介绍非完美信息博弈的在线求解方法，包括基于搜索的求解方法和对手建模求解方法，并针对在线求解与离线求解方法的特点进行对比分析；第 6 节将总结全文并展望非完美信息博弈求解研究的未来.

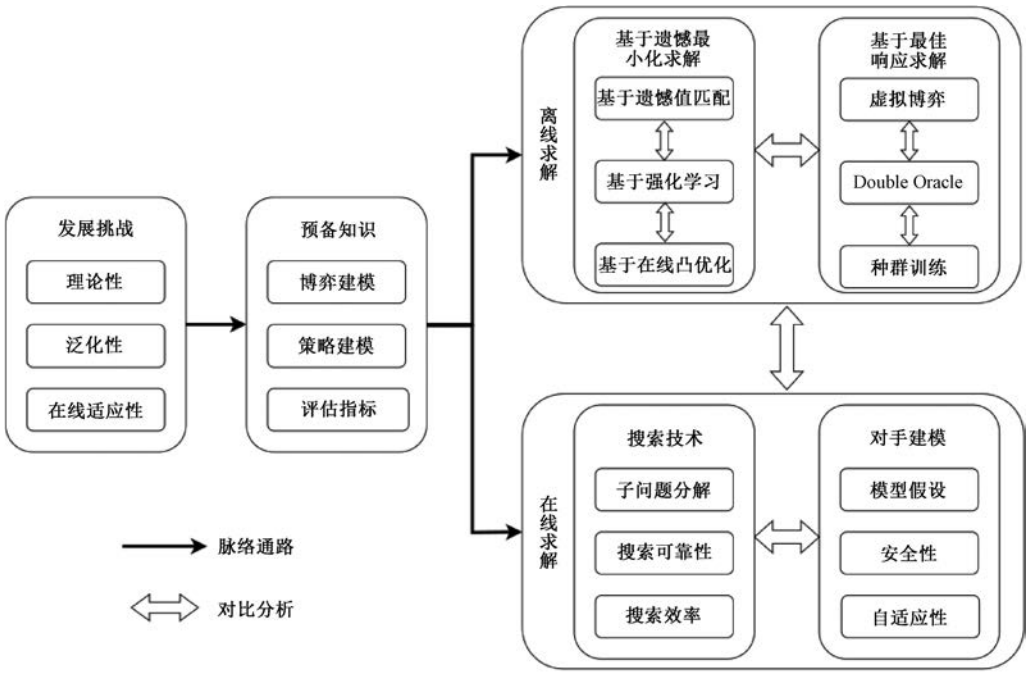

图 1　框架脉络示意图

2 背景知识

本节将梳理非完美信息博弈研究领域的背景知识和相关基础概念. 为便于读者快速查阅，表 1 中总结了文中常用的专有名词缩写.

表 1 专有名词缩写表

英文缩写	含义	英文缩写	含义
IIG (Imperfect Information Game)	非完美信息博弈	PIG (Perfect Information Game)	完美信息博弈
NFG (Norm-Form Game)	正则博弈	EFG (Extensive-Form Game)	扩展式博弈
2p0s (Two Player Zero-Sum)	双人零和	RPS (Rock-Paper-Scissors)	剪刀石头布
MCTS (Monte-Carlo Tree Search)	蒙特卡洛树搜索	FSP (Fictitious Self Play)	虚拟自博弈
NE (Nash Equilibrium)	纳什均衡	BR (Best Response)	最佳响应
CFR (Counterfactual Regret Minimization)	虚拟遗憾最小化	OCO (Online Convex Optimization)	在线凸优化

2.1 博弈建模

正则博弈(Normal-Form Game，NFG) 是一类最简单的非完美信息博弈模型. NFG 由一个玩家集合$\mathcal{N}$、动作集$\mathcal{A}$和收益矩阵$\mathcal{X}$组成. 所有玩家$\forall i\in\mathcal{N}$同时选择各自动作 $a_i\in\mathcal{A}_i$，获得矩阵$\mathcal{X}$中给定的收益. 图 2 给出了石头剪刀布 (Rock-Paper-Scissors，RPS) 正则博弈的收益矩阵.

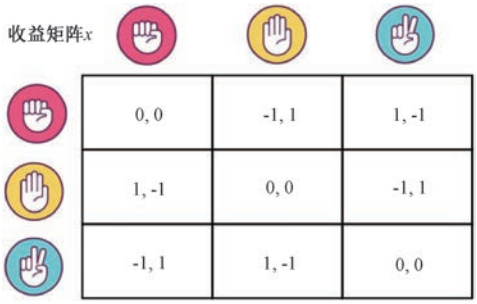

图 2 石头剪刀布博弈的正则博弈模型

尽管正则博弈要求同时选择动作，我们仍可等价地视其为玩家依次行动的交替序列决策问题. RPS 的等价序列决策模型如图 3 所示.

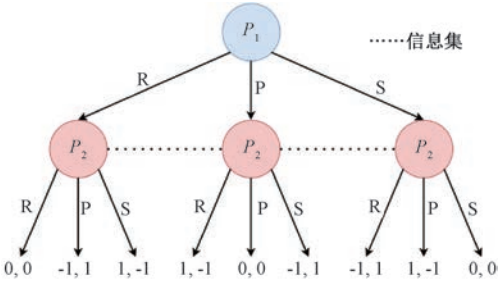

图 3 石头剪刀布博弈的扩展式博弈模型

类似图 3 的序列决策博弈被称为扩展式博弈 (Extensive-Form Game，EFG)，可由一个元组$(\mathcal{N},\mathcal{H},u,\mathcal{P},\mathcal{A},\mathcal{I})$形式化定义[2]：

(1) $\mathcal{N}$是参与博弈的玩家 (Player) 的有限集. 此外，通常还会定义一个额外的机会玩家(Chance Player) c，并将环境中能引起状态转移的随机事件视为 c 做出的动作，如发牌、掷骰子等事件，如图 4 单牌扑克博弈中圆形节点所示.

(2)$\mathcal{H}$是历史(History) h 的有限集. 如图 4 (a) 中所示，历史 h 是完整博弈树中的一个节点 (Node)，其包含当前博弈状态的公有信息和所有玩家的私有信息. 完整博弈树中的任一叶子节点也称为终止历史 z，其集合记为终止历史集 $\mathcal{Z}\subseteq\mathcal{H}$，如图 4 中三角形节点所示.

(3) 抵达终止历史 z 后，每个玩家得到收益 (Utility) $u_i:\mathcal{Z}\to\mathbb{R}$. 特别地，若 $|\mathcal{N}|=2$，且 $\forall z\in\mathcal{Z}, u_1(z)+u_2(z)=0$，则称该博弈为双人零和博弈；

(4) 所有非终止历史 $h\in\mathcal{H}\setminus\mathcal{Z}$ 都对应着某一玩家的决策点，如图 4 方形节点所示. 玩家函数 $\mathcal{P}$ 从 $\mathcal{N}\cup\{c\}$ 中为非终止历史 h 指派一个行动的玩家.

(5) 动作函数 $\mathcal{A}(h)=\{a:(h,a)\in\mathcal{H}\}$ 给出非终止历史 h 上的可选动作集合. 用 $h\cdot a$ 或者 (h,a) 表示在 h 上拼接动作 a 得到的历史，$h^1\sqsubset h^2$ 表示 h^1 是 h^2 的前缀.

(6) 在非完美信息博弈中，由于无法观测到对手私有信息，玩家 i 无法区分具有相同公共信息和玩家私有信息的历史，从而不同玩家视角下的博弈树是不同的. 将玩家 i 无法区分的历史的集合记为信息集 (Information Set) $I_i\in\mathcal{I}_i$，其中 $\mathcal{I}_i$ 是玩家 i 所有信息集的集合. 信息集 I_i 是玩家 i 视角下博弈树的一个节点，如图 4 (c)、(d) 所示. 若 $\forall I\in\mathcal{I},|I|=1$，博弈将退化为完美信息博弈.

遵循增广扩展式博弈的定义[36]，用公共信息集 $I_{pub}\in\mathcal{I}_{pub}$ 表示公共视角下博弈树的一个节点，如图 4(b) 中所示，I_{pub} 中的历史具有相同的公共信息. 定义 $I_i(I_{pub})$ 为玩家 i 与 I_{pub} 公共信息兼容的

信息集集合. 对应于图 4 中全知视角、各玩家视角和公共视角下的博弈树节点类型，图 5 展示了扩展式博弈由历史 h 、信息集 I_i 和公共信息集 I_{pub} 构成的三层信息结构及集合与元素间的包含关系.

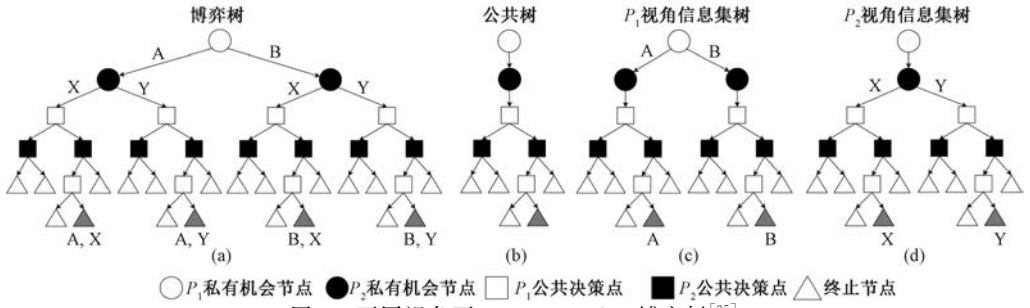

图 4　不同视角下 one-card poker 博弈树[35]

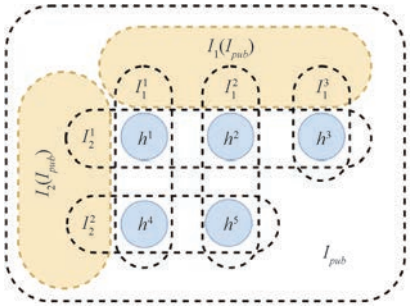

图 5　扩展式博弈信息层次结构

随着博弈论与机器学习理论的有机结合，非完美信息博弈领域借鉴机器学习中建模部分信息场景的常用模型——部分可观测马尔可夫博弈（Partically Observation Stochastic Game，POSG），在 EFG 的基础上进一步提出分解观测随机博弈（Factored-Observation Stochastic Game，FOSG）[36]，弥补 EFG 的表达缺陷. 表 2 根据博弈模型中元素的职能对 EFG、POSG 和 FOSG 三种模型进行了对比总结. 本文将主要围绕 EFG 模型下的相关研究进行论述，有关 POSG 与 FOSG 模型的相关研究进展请参考 3. 3 节中的讨论及相关文献[37-38].

表 2　EFG、POSG 和 FOSG 建模方式对比[36]

功能描述	EFG	POSG	FOSG
交互模式	交替决策	并发决策	并发决策
环境中确切的状态	历史 h	状态 s	世界状态 w
决策所依据节点	信息集 I_i	观测轨迹 τ_i	私有状态 s_i
即时反馈	—	奖励 τ	观测 $O_i(\cdot)$
优化目标	收益 u_i	累计奖励	累计收益 $R_i(h)$
环境中随机性来源	机会玩家策略 σ_c	转移概率 $T(s,a)$	转移概率 $T(w,a)$

2.2　策略建模

2.2.1　行为策略（Behaviour Strategy）

扩展式博弈中，由于无法区分信息集中的历史，玩家以信息集为单位进行决策. 行为策略 σ_i：$I \in \mathcal{I}_i \rightarrow \Delta\mathcal{A}(I)$ 确定了玩家 i 在每一信息集 I 上 $\mathcal{A}(I)$ 的概率分布. 定义行为策略集（Behaviour Strategy Profile）σ 为各玩家行为策略组成的元组 $(\sigma_1,\sigma_2,\ldots\sigma_{|\mathcal{N}|})$，其中 $\sigma_i(I,a)$ 或 $\sigma_i(h,a)$ 表示玩家 i 在信息集 I 或历史 h 采取动作 a 的概率. 作为一类特殊行为策略，纯策略（Pure Strategy）将玩家 i 的每一信息集 $I \in I_i$ 映射至单一的可选动作 $a \in \mathcal{A}(I)$. 定义纯策略空间 $\Delta_p^i \subset \sum^i$，其中 $\sum^i$ 为玩家 i 所有行为策略构成的空间. 混合策略（Mixed Strategy）Π^i 是玩家 i 纯策略空间 Δ_p^i 上的一个概率分布.

借助行为策略和纯策略的定义，可将一个扩展式博弈转换为一个等价的正则博弈[2]. 对于图 6 左上角给出的扩展式博弈，图 6（a）给出其等价的正则博弈，其中行列玩家的每一动作均为一个纯策略. 从而，在对应的正则博弈中选定某个动作相当于在原扩展式博弈中使用对应的纯策略.

2.2.2　序贯策略（Sequence-form Strategy）

行为策略具有表达简洁的优势，但也面临两方面的计算问题. 首先行为策略对于扩展式博弈中的收益通常具有非凸性，导致无法直接对整体博弈问题进行优化. 具体而言，玩家 i 在行为策略集下得到的期望收益 $u_i = \sum_{z\in\mathcal{Z}} u_i(z)P(z \mid \sigma)$，其中 $P(z \mid \sigma)$ 是在策略集 σ 下终止历史 z 的到达概率. 由于终止历史 z 上通常包含多个玩家 i 的动作，因而，策略集 σ 对于 $P(z \mid \sigma)$ 的贡献是不具备凸性的 σ 连乘形式 $\prod_{h,a\sqsubset z}\sigma(h,a)$. 其次，行为策略在信息集上的线性加和与其对应的混合策略线性加和具有不等价性.

图 6（e）给出了两个混合策略 Π_1 和 Π_2 的线性加和，与其对应的行为策略表示. 然而如图 6（b）

所示，在每一信息集上直接对行为策略进行线性加和与 6（e）中的行为策略是不等价的.

若直接将策略建模为类似 $P(z \mid \sigma)$ 中的连乘形式，可以得到一个具有凸性的策略空间[29,39]. 形式化地，定义到达概率 $\pi^{\sigma}(h)$ 为玩家根据 σ 决策时博弈到达 h 的概率：

$$\pi^{\sigma}(h)=\prod_{h'\cdot a'\sqsubset h}\sigma(h',a'). \tag{1}$$

玩家对到达概率的贡献可由决策的各项概率连乘所得：

$$\begin{aligned}\pi_i^{\sigma}(h)&=\prod_{h'\cdot a'\sqsubset h\mid P(h')=i}\sigma(h',a'),\\ \pi_{-i}^{\sigma}(h)&=\prod_{h'\cdot a'\sqsubset h\mid P(h')\neq i}\sigma(h',a').\end{aligned} \tag{2}$$

其中，$\pi_i^{\sigma}(h)$ 称为内因到达概率（Player Reach），$\pi_{-i}^{\sigma}(h)$ 为外因到达概率（External Reach）. 信息集 I 上的到达概率 $\pi_i^{\sigma}(I)$ 为信息集 I 中所有历史的到达概率之和，即满足 $\pi_i^{\sigma}(I)=\sum_{h\in I}\pi_i^{\sigma}(h)$.

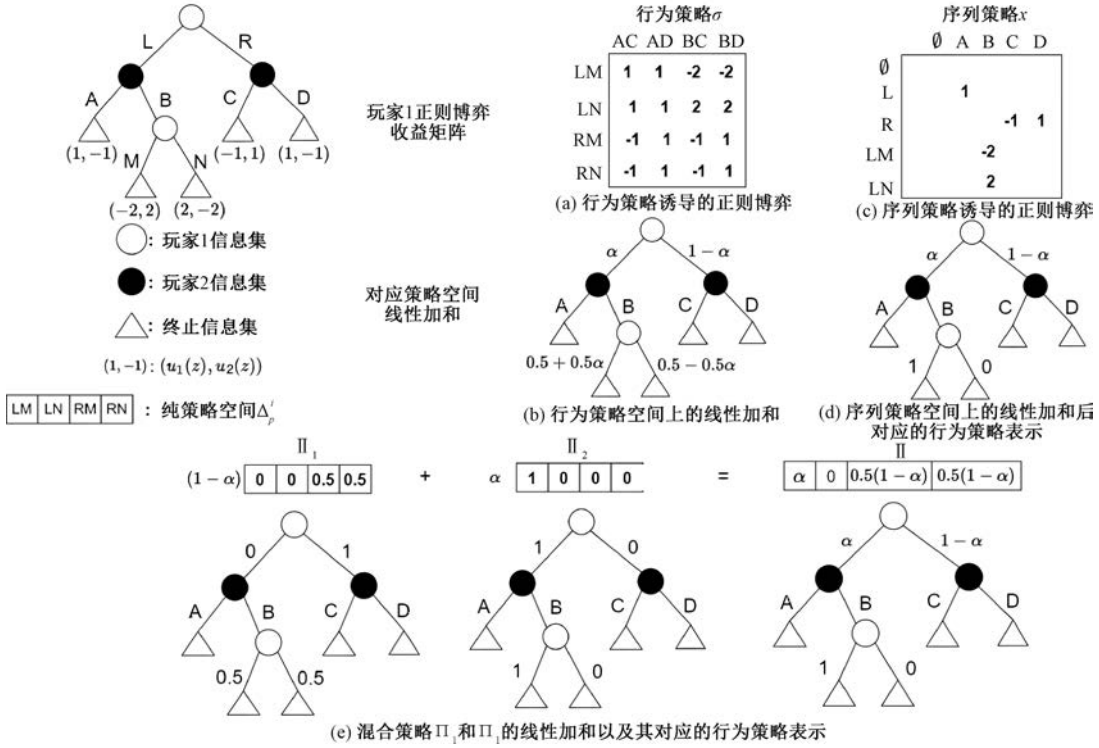

图 6 行为策略和序列策略建模对比

玩家 i 的序贯策略 $\mathcal{X}_i$ 在行为策略 σ_i 的基础上，将每一信息集 $I\in\mathcal{I}_i$ 映射至对应的内因到达概率 $\pi_i^{\sigma}(I)$，即 $\mathcal{X}_i: I\in\mathcal{I}_i\to[0,1]$. 一个良定义的序贯策略需要满足两个条件：

$$\begin{cases}①\ X_i(\varnothing)=1\\ ②\ \forall I\in\mathcal{I}_i: X_i(I)=\sum_{a\in A(I)}X_i(I\cdot a)\end{cases} \tag{3}$$

由于序贯策略是对到达概率的分配，在一些文献中也称序贯策略为到达计划[39-40]（Realization Plan）. 当且仅当两个策略能够诱导出相同的到达计划时，我们称两个策略是实现等价的[39]（Realization Equivalent）. 根据公式 1 中到达概率定义，可得行为策略 σ 与序贯策略 $\mathcal{X}$ 的关系：

$$\sigma_i(I,a)=\frac{\pi_i{}^{\sigma}(I\cdot a)}{\pi_i{}^{\sigma}(I)}=\frac{X_i(I\cdot a)}{X_i(I)} \tag{4}$$

图 6（c）给出了序贯策略诱导的正则博弈示例，其中，玩家的每一动作都是与信息集一一对应的动作序列. 在双方动作序列拼接合法的元素处，玩家获得其收益值 $u(z)$，其余非法元素均被置为 0. 相较图 6（a）图 6（c）中的正则博弈收益矩阵更为稀疏，动作空间规模也由 $|I|$ 指数相关降为线性相关. 此外，序贯策略正则博弈的收益对于动作而言是线性的，可直接利用线性规划方法进行求解[41]. 序贯策略为行为策略和混合策略建立了联系. 根据已有研究结论[40,42-43]，序贯策略的线性加和与混合策略的线性加和能够诱导出相同的行为策略，如图 6（d）所示. 因此，涉及计算行为策略在信息集上的线性加和时，可先通过计算实现等价的序贯策略进行加和，再通过公式 4 转换为行为

策略.

2.3　纳什均衡和可利用度

在非稳态的博弈环境中,算法往往以解得静态的均衡(Equilibrium)状态为目标.纳什均衡[44](Nash Equilibrium,NE)是最被广泛应用的均衡概念.在纳什均衡中,每个玩家的策略均为其余玩家策略的最佳响应(Best Response,BR),其中,最佳响应 BR^i 为玩家在其余玩家 i 策略 σ_{-i} 给定时能采取的最优策略,即 $\mathrm{BR}^i(\sigma_{-i})=argmax_{\sigma'_i}u_i(\sigma'_i,\sigma_{-i})$.可利用度(Exploitability)$e(\sigma)$ 衡量了策略与纳什均衡间的距离,常被用于策略评估:

$$e(\sigma)=\frac{1}{|\mathcal{N}|}\sum_{i\in\mathcal{N}}u_i(\mathrm{BR}(\sigma_{-i}),\sigma_{-i})-u_i(\sigma_i,\sigma_{-i}) \tag{5}$$

若某策略集 σ 的可利用度满足 $e(\sigma)\leqslant\epsilon$,则称 σ 是 ϵ 近似的纳什均衡(ϵ-Nash Equilibrium,ϵ-NE).纳什均衡在 2p0s 博弈中具有独特的地位.在一个公平的 2p0s 博弈中,无论对手采取什么策略,使用纳什均衡策略可以保证该玩家在长期期望下不输.因此,若能够设计算法计算得到纳什均衡策略,我们就可以从理论上宣称解决了该 2p0s 博弈,例如近年来被攻克的围棋[7]、双人德州扑克[31]等.

3　基于遗憾的均衡求解方法

遗憾最小化(Regret Minimization)概念起源于在线学习领域,随后逐渐发展成为求解非完美信息博弈的关键方法之一.基于事后反省(Hindsight)的逻辑,遗憾衡量了智能体当前实际采用策略与其本可以采用的策略间的差距,因而在遗憾最小化的过程,智能体的策略将得到改进.根据最小化遗憾的算法类别及领域视角,本文将基于遗憾的均衡求解方法分为基于遗憾值匹配、基于强化学习和基于在线凸优化三类.本节将首先阐明遗憾最小化与均衡求解的关联,随后依次介绍上述三类方法的发展脉络、算法优劣以及区别联系.

3.1　遗憾最小化与均衡求解

在正则重复博弈下,每一轮次 t,玩家 i 选择动作 $a\in\mathcal{A}$ 执行,并得到收益 $u^t(a)$.第 t 轮玩家的期望收益 $u^t=\sum_{a\in A}\sigma^t(a)u^t(a)$.若玩家 i 在第 t 轮结束后反省,设想在第 t 轮采取另一动作 a',则可定义未采取动作 a' 的瞬时遗憾 $r^t(a')=u^t(a')-u^t$.类似地,策略序列 $s_{a'}$ 的遗憾 $R^T(s_{a'})=\sum_{t=1}^{T}(u^t(a')-u^t)$.为便于标记,采取单一动作策略序列的遗憾可简记为 $R^T(a')$.不失一般性,定义任一策略序列 s 的遗憾 $R^T(s)=\sum_{t=1}^{T}(u^t(a_t)-u^t)$,平均遗憾 $\overline{R}^T(s)=\frac{1}{T}R^T(s)$.在某一给定的策略序列集合 S 中,定义遗憾的上界为博弈的整体遗憾 $R_i^T=\max_{s\in S}\sum_{t=1}^{T}(u^t(a)-u^t)$.图 7 以 RPS 博弈为例展示了单一动作策略序列的遗憾 $R^T(a)$ 的计算过程和结果,其中,玩家每一轮都采用某一纯策略 $\sigma^t\in\Delta_p^i=\{R,P,S\}$.

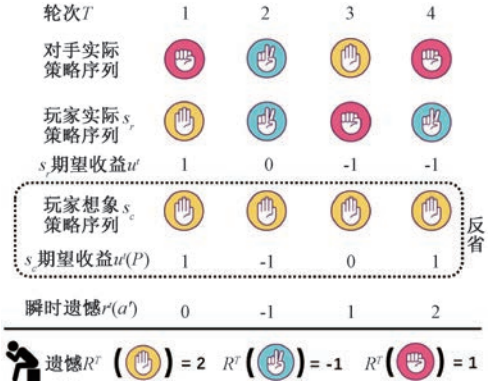

图 7　重复 RPS 博弈过程中遗憾值的分布情况

若算法对于给定的策略序列集合 S,$\forall s\in S$,$R^T(s)\in O(T)$,即平均整体遗憾随 T 的增加而趋近于某一常数,则该算法被称为是无遗憾(No-regret)的.策略集合 S 实质上是算法自省没有采取的策略集合.若 S 被给定为在每一轮都采取同一动作的策略序列集合,则该算法将保证无外部遗憾[45](External Regret),即具有 Hannan 一致性[46](Hannan Consistency).若所有玩家都采取无外部遗憾的算法,平均策略集 $\bar{\sigma}^T$ 将收敛至粗相关均衡[47-48](Coarse Correlated Equilibrium,CCE),在 2p0s 博弈中,粗相关均衡等价于纳什均衡[49].除外部遗憾外,还有其他形式的遗憾对应不同类型的均衡求解[50-52].本文关注于双人零和博弈的纳什均衡求解方法,无外部遗憾是最松弛的无遗憾条件.因此,后文将统一用遗憾指代外部遗憾.

定理 1 给出了遗憾与纳什均衡的联系.对应定理 1,一个通用的无遗憾算法更新范式如图 8 所示.

定理 1. 在双人零和博弈中,若 $\forall i\in\mathcal{P},\frac{R_i^T}{T}\leqslant\epsilon_i$,那么平均策略 $\bar{\sigma}^T$ 是 $(\epsilon_1+\epsilon_2)$-纳什均衡.

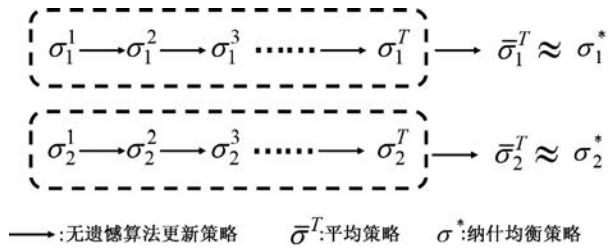

图 8 无遗憾算法的策略更新范式

3.2 基于遗憾值匹配的求解方法

遗憾值匹配算法因其易于结合采样、无参数等优点被广泛应用于非完美信息博弈的求解方法之中[53-55]，其中以虚拟遗憾最小化（Counterfactual regret minimization，CFR）为代表一系列工作，先后成功解决了双人有限注[56]、无限注[53-54]德州扑克，并在多人无限注规则中击败人类顶级玩家[32].

图 9 总结了 CFR 算法的主体发展脉络与不同变体的关键革新点，其中包括采样效率、估值方差、收敛性能与扩展规模性等方面. 本小节将遵循图 9 的结构，介绍 RM 及 CFR 算法的关键改进及分支产生的原因，更多详情请参考 CFR 算法相关综述[57].

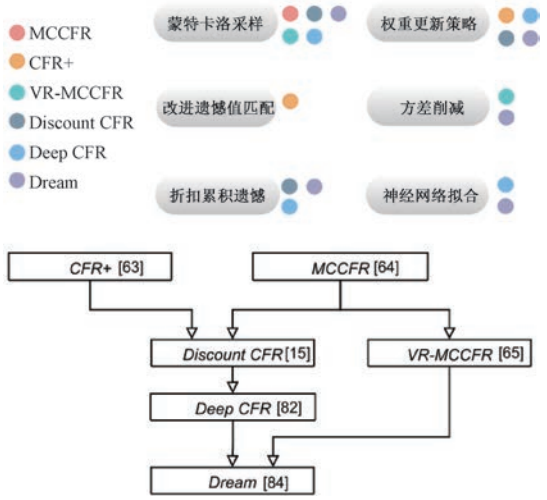

图 9 基于遗憾匹配的求解方法发展脉络

3.2.1 遗憾值匹配（Regret Matching，RM）

RM 是一种经典的在线学习算法[47]，其通过归一化所有动作的遗憾来更新策略，使玩家下轮选择动作的概率正比于当前轮次该动作的正遗憾值，如公式(6)所示：

$$\sigma^{T+1}(a)=\begin{cases}\dfrac{\lfloor R^T(a)\rfloor_+}{\sum_a \lfloor R^T(a)\rfloor_+}, & \text{if } \sum_a \lfloor R^T(a)\rfloor_+ > 0\\ \dfrac{1}{|A|}, & \text{otherwise}\end{cases} \tag{6}$$

其中 $\lfloor\cdot\rfloor_+=\max\{\cdot,0\}$. RM 保证了遗憾的上界为 $O(T^{1/2})$[48]，是一种无遗憾的算法. 图 10 展示了 RM 算法的一轮更新过程，其中玩家 1 和玩家 2 的初始策略分别为 $(\frac{1}{3},\frac{1}{3},\frac{1}{3})$ 和 $(\frac{1}{5},\frac{2}{5},\frac{2}{5})$.

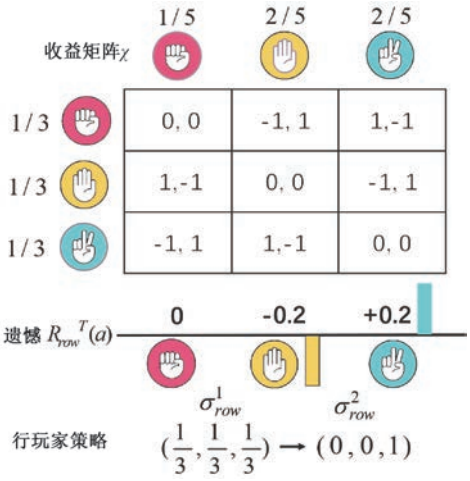

图 10 RPS 博弈中一轮 RM 算法更新流程图

RM+算法[58]在 RM 的基础上，使用中间量 $Q^T(a)=\max\{Q^{T-1}(a)+r^T(a),0\}$ 替代累积遗憾 $R^T(a)$. 从结果而言，$Q^T(a)$ 累积了最近一次遗憾值置零至当前轮次区间内的遗憾. 在实际运用中，RM+展现出远快于 RM 的收敛速度[59]，被广泛运用于后续研究中[59-60]. 近年来，一系列通过结合预测更新[61]、额外梯度更新[60]、光滑化[62]等技术的工作进一步提升了 RM 算法的收敛性能.

3.2.2 虚拟遗憾最小化（CFR）

若在 EFG 下直接应用 RM 算法，公式 6 中分母的加和要求维护所有纯策略序列的遗憾，这等同于将 EFG 转换为 NFG 再进行求解，复杂度过高. 而 CFR 算法[63]通过类似分治的思想，将最小化博弈整体遗憾的任务分解为最小化每个信息集的遗憾，使得算法复杂度降为与信息集数目线性相关.

在 EFG 中，价值被定义为 $v_i^\sigma(I)=\sum_{h\in I} v_i^\sigma(h)$，其中 $v_i^\sigma(h)=\sum_{z\in\mathcal{Z},h\sqsubset z}\pi^\sigma(z)u_i(z)$. 不同于完美信息博弈，此处历史 h 上的价值 $v^\sigma(h)$ 没有假设已经到达 h 这一前提条件，$\pi^\sigma(z)=\pi^\sigma(h)\pi^\sigma(h,z)$ 中包含 h 的到达概率 $\pi^\sigma(h)$. 这是由于一个信息集中包含多个历史，我们需要将历史的分布 $\Delta(\mathcal{H}(I))$，即 $\pi^\sigma(h)$ 考虑在内. CFR 算法进一步定义了虚拟价值（Counterfactual Value）$v_i^{\sigma,CF}(h)$：

$$v_i^{\sigma,CF}(h)=\pi_{-i}^\sigma(h)\sum_{z\in\mathcal{Z},h\sqsubset z}\pi^\sigma(h,z)u_i(z) \tag{7}$$

与价值 $v_i^\sigma(I)$ 不同，虚拟价值衡量了假定玩家 i 以概率 1 到达 h 时 h 上的价值，其本质上是 $\pi_i^\sigma(h)=1$ 条件下收益的后验分布期望. 为衡量某一动作的价值，进一步定义 $v^{\sigma,CF}(h,a)$ 表示在该历史 h 上总是选择执行动作 a 得到的虚拟价值:

$$v_i^{\sigma,CF}(h,a)=\pi_{-i}^{\sigma}(h)\sum_{z\in\mathcal{Z},h\sqsubset z}\pi^{\sigma}(h\cdot a,z)u_i(z) \tag{8}$$

由此可得信息集 I 上动作 a 的瞬时遗憾 $r_i^t(I,a)=v^{\sigma^t,CF}(I,a)-v^{\sigma^t,CF}(I)$，信息集的整体虚拟遗憾 $R_i^T(I)=\max_{a\in\mathcal{A}(I)}\lfloor R_i^T(I,a)\rfloor_+$. CFR 算法证明了最小化博弈整体遗憾的任务可分解为最小化每个信息集的遗憾，即 $R_i^T\leqslant\sum_{I\in\mathcal{I}_i}R_i^T(I)$，从而更新对象也从整体博弈的混合策略转为信息集上的行为策略，大幅降低了计算复杂度.

CFR 算法每一轮需要遍历信息集并通过 RM 算法更新行为策略. 而在计算行为策略的平均时，根据 2.2 节中实现等价的定义，我们必须以 $\pi_i^{\sigma^t}(I)$ 作为权重进行加权平均①计算 $\sigma_i^t(I)$[39]. 第 T 轮次某一信息集上的平均行为策略的计算公式为

$$\bar{\sigma}_i^T(I)=\frac{\sum_{t=1}^{T}(\pi_i^{\sigma^t}(I)\sigma_i^t(I))}{\sum_{t=1}^{T}\pi_i^{\sigma^t}(I)} \tag{9}$$

遵循图 8 中的无遗憾算法的范式，以 RPS 博弈为例，CFR 算法的流程如图 11 中所示. CFR 算法本质上是博弈树上的树遍历算法，每一轮都需要遍历所有节点并更新遗憾和策略值.

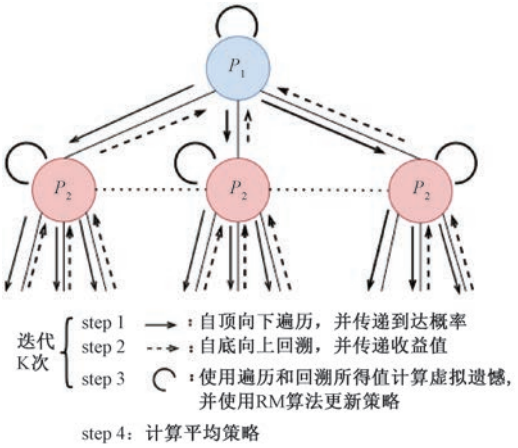

图 11　CFR 算法的更新过程的数据流

3.2.3　蒙特卡洛虚拟遗憾最小化（MCCFR）

MCCFR 将 CFR 与蒙特卡洛（Monte Carlo，MC）方法结合，每轮只访问并更新部分信息集，用估计的采样虚拟价值进行遗憾最小化[64]，解决了 CFR 算法需要遍历更新博弈树的缺陷. 在采样算法中，终止序列集合 $\mathcal{Z}$ 根据采样方案被划分成若干子集合 Q_k，定义集 $\mathcal{Q}=\{Q_1,Q_2,Q_3,\cdots,Q_n\}$. MCCFR 每一轮根据特定采样方案生成的概率分布 $\Delta(Q)$ 采样子集 Q_j，并只更新可到达 Q_j 中终止序列的信息集. 显然，若 $Q=\{\mathcal{Z}\}$，MCCFR 将退化成 CFR 算法.

不同采样方案衍生出许多 MCCFR 的变种，其中最为常用是结果采样（Outcome-Sampling，OS）和外因采样（External-Sampling，ES）. 在结果采样中，所有子集 Q 中仅包含一条终止序列 z，即 $\forall Q\in\mathcal{Q},|Q|=1,q(z)=\pi^{\xi}(z)$，其中 ξ 是一可选的采样策略，如图 12 (b) 所示. 结果采样在每一信息集上仅采样一个动作，具有向无模型（Model-Free）的算法发展的潜质. 外因采样则根据对手和机会玩家的纯策略组合来划分 $\mathcal{Q}$，同一集合 $\forall Q\in\mathcal{Q}$ 中的终止历史在对手和机会玩家上的历史上采取的动作相同 即 $\forall Q\in\mathcal{Q},\forall z_i,z_j\in Q,q(z_i)=q(z_j)=\pi_{-i}^{\sigma}(z)$. 不同采样方式可以看作不同的遍历展开方式. 结果采样在每一信息集上都不展开遍历，而外因采样会在玩家自身信息集上展开遍历，如图 12(c)所示.

尽管采样虚拟价值是虚拟价值的无偏估计，但它具有较高的估值方差，因此，方差削减技术在 MCCFR 中得到广泛使用[65]. 基于经济学中控制变量思想，VR-MCCFR 在原采样虚拟遗憾中加入基线函数 $\hat{b}_i(\sigma,I,a)$:

$$\hat{v}_i^b(\sigma,I,a)=\hat{v}_i(\sigma,I,a)-\hat{b}_i(\sigma,I,a)+b_i(\sigma,I,a) \tag{10}$$

其中 $b_i(I,a)=\mathbb{E}[\hat{b}_i(I,a)]$. 从博弈树遍历过程直观地看，结合方差削减后的采样在 $h\subseteq z,h\cdot a\subsetneq z$ 时，动作 a 的采样虚拟价值会含有基线函数部分，而不再赋 0 值，如图 12 (d)所示.

除方差削减外，还有一些 MCCFR 的重要改进. Gibson 等人指出根据玩家平均策略采样能够提升 MCCFR 在大型动作空间博弈中的收敛性能[66].

Johanson 提出向量形式 MCCFR，通过在采样时引入公共状态提升了采样效率[67]. Michael 等人将 OS-MCCFR 扩展为在线模式[68]. Jackson 和 Li 先后从不同角度提出一种介于 OS 和 ES 之间的采

① 平均本质上也是一种线性加和，公式 9 是对公式 4 的一种应用.

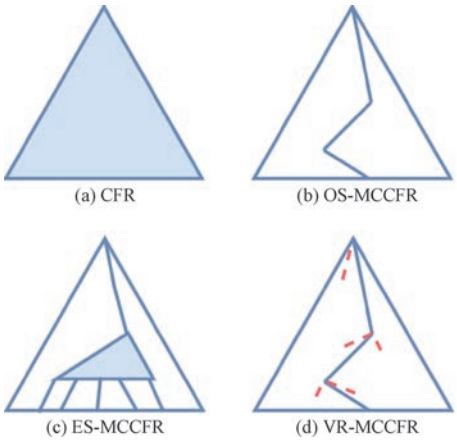

图 12 CFR 与 MCCFR 各变体间采样对比

样方式，对采样展开程度进行了合理设计[69-70].

3.2.4 虚拟遗憾最小化+(CFR+)

CFR+算法对 CFR 算法做出了三点改进. 首先，在计算平均策略时 CFR+引入轮次相关线性权重 t ：

$$\bar{\sigma}_i^T(I)=\frac{\sum_{t=1}^{T}(t\pi_i^{\sigma^t}(I)\sigma_i^t(I))}{\sum_{t=1}^{T}t\pi_i^{\sigma^t}(I)} \tag{11}$$

其次从 CFR+开始，交替更新策略方式成为 CFR 系列算法的首选. Burch 研究了在交替更新下遗憾与可利用度之间的关系，证明了交替更新的正确性[59]. 图 13 展示了交替更新策略的一般流程. 最后，CFR+采用 RM+代替了 RM 进行策略更新. 在实际运用中，CFR+算法展现出了 $O(T^{-1})$ 的渐进性能，且最终策略取得了不亚于平均策略的收敛效果.

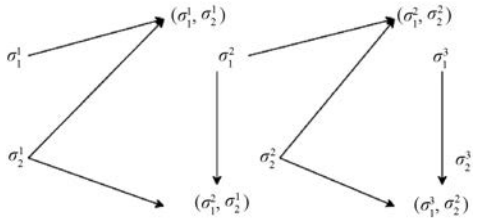

图 13 交替更新策略流程示意图[71]

DCFR 在权重设置方面继续做出了改进，在计算累积遗憾时也赋予权重[15]. 对 RM 算法轮次进行加权的做法最早在 NFG 中已有良好效果[48]. DCFR 总括性地通过三个参数 α,β,γ(DCFR$_{\alpha,\beta,\gamma}$)分别控制正累积遗憾、负累积遗憾与平均策略的加权更新. Zhang 等人提出一种动态权重设置以实现瞬时遗憾的加入能够最小化平均策略的可利用度[72]. Xu 等人通过自动机器学习（Auto ML）方式来学习 CFR 算法变种中所涉及的各类参数[73]，在算法性能提升上取得了一定成果.

3.2.5 CFR 规模扩展方法

相较线性规划等技术[41,74-75]，CFR 算法已可求解例如有限注德州扑克这类拥有 10^{13} 信息集数目的较大规模非完美信息博弈，但其仍不足以扩展至更大规模的博弈中，例如拥有 10^{161} 信息集数目的无限注德州扑克. 此外，一些博弈动作空间巨大，无法维护所有动作的遗憾值. 因此，CFR 算法需要一些技巧以扩展至更大规模的博弈. 约简和神经网络拟合是扩展算法规模的两种常规思路.

约简的方法被广泛应用于计算扑克之中[76]. 这类方法首先根据博弈规则与玩家动作集，从大型博弈抽象出策略特征相似的小型博弈进行计算，然后将小型博弈的均衡策略映射回大型博弈中. 图 14 中给出了约简方法的一般流程，其中约简与还原步骤往往需要耦合设计[77].

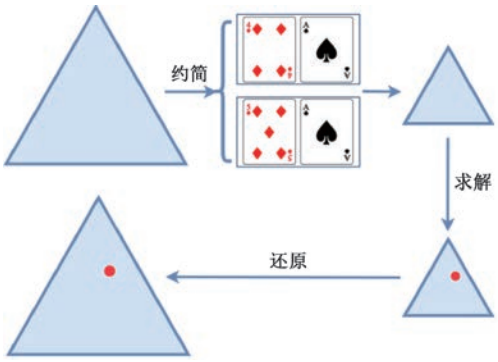

图 14 博弈约简求解流程[40]

从内容分类，抽象可以分为信息约简，动作约简和阶段约简[78-79]. 信息约简合并了对手信息相近的信息集. 在德州扑克中，对手牌力反映了这种对手信息，如图 14 中约简一步所示. 牌力的强弱一般由蒙特卡洛模拟的对阵胜率评估. 动作约简会离散化动作集或根据信息合并相近动作，例如在德州扑克中，可以将连续的下注大小离散化为分段区间. 此外，许多博弈具有明显的阶段性特征，例如扑克的下注轮次、翻牌前后等. 阶段约简划分博弈阶段分别求解，再将不同阶段策略合并. 根据约简后博弈能否保持与原博弈相同的均衡，可将约简分为无损约简和有损约简[77].

基于人工设计的约简方法往往需要特定的领域知识，具有较差的可迁移性. 随着深度学习的发展，偏向人工设计的约简方法逐渐式微，借助神经网络进行函数拟合逐渐成为扩展算法规模的首选. Keven 等人提出的 Regression CFR 是 CFR 领域最

早结合函数近似思想的算法[80-81]，其通过定义一系列信息集特征，使用回归树模型近似真实迭代产生的遗憾值. 基于这种思想，Brown 于 2018 年提出 Deep CFR[82]，通过训练两个神经网络 $V(I,a\mid\theta_i)$，$\pi(I\mid\theta_\pi)$ 分别拟合累积遗憾和平均策略，并使用缓冲池 M_V 和 M_π 来分别存储用于训练的经验. Deep 训练框架流程如图 15 所示.

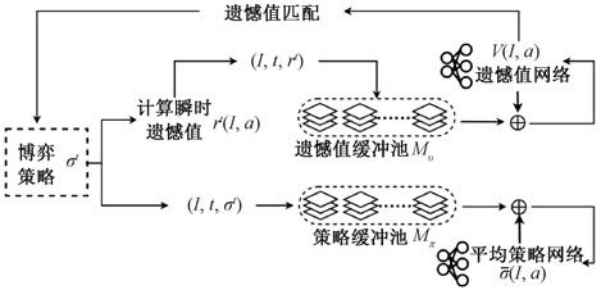

图 15　Deep CFR 算法训练框架[3]

为避免信息集访问不均匀，Deep CFR 会进行多次 ES-MCCFR 流程. 计算所得的采样虚拟遗憾 $r(I,a)$ 被存入 M_V 中用于训练累积遗憾网络. 在对手信息集上，根据 $V(I,a\mid\theta_{-i})$ 更新策略 $\sigma^t(I)$ 并进行采样，同时将经验数组 (I,t,σ^t) 存入 M_π 中用于训练平均策略网络. 在德州扑克上，Deep CFR 取得了胜过 NFSP 的效果.

同年，Li 等人提出了另一套神经网络的实现方法 Double Neural CFR[70]（DNCFR），类似地使用两个神经网络作为架构，并提出了一些针对 CFR 算法的神经网络训练技巧. Steinberger 提出 SD-CFR[83]，基于遗憾值网络直接计算平均策略，不再另设平均策略网络，降低了函数拟合误差. Steinberger 在 Deep CFR 的基础上，结合方差削减、结果采样等技巧，提出基于深度学习的无模型算法 Dream[84]，目前仍是非完美信息领域强有力的基线算法之一，其不仅在德州扑克上取得非凡表现，还在 Geister[85]、gin rummy[86] 等博弈中取得了不错的效果. McAleer 提出了 ESCHER[87]，直接训练历史价值函数对累积遗憾进行估计，并将该估计值作为基线以降低估值方差，在 Dream 的基础上进一步提升了收敛速度和稳定性. Liu 等人提出了 CFR 算法的自举更新方式[88]，使用递归值来代替累计遗憾的估计，进一步提升了神经网络的估值性能.

3.3　基于强化学习的求解方法

图 8 所展示的无遗憾算法通用流程与强化学习中的演员评论家（Actor-Critic，AC）框架十分相似. $B_{-i}(\sigma,I)$ 实质上就是贝叶斯方法中的标准化常量（normalized constant）. 他们的联系在于其都对价值做出了假设，但假设的前提条件不同. 虚拟价值 $v^{\sigma,CF}(I)$ 是在玩家 i 到达 I 的前提下的价值，对于对手 $-i$ 没有做出假设. 而在强化学习中，由于 MDP 的无后效性，状态价值是到达状态 I 条件下的价值，对所有玩家都做出了假设. 根据公式易知，强化学习中常用的优势（Advantage）函数 $A^\sigma(I,a)=q^\sigma(I,a)-v^\sigma(I)$ 与虚拟瞬时遗憾 $r(I,a)$ 同样相差了相同的系数 $\mathcal{B}_{-i}(\sigma,I)=\sum_{h\in I}\pi^\sigma_{-i}(h)$.

$$v^\sigma(I)=\frac{1}{\sum_{h\in I}\pi^\sigma_{-i}(h)}v^{\sigma,CF}(I)$$

$$q^\sigma(I,a)=\frac{1}{\sum_{h\in I}\pi^\sigma_{-i}(h)}v^{\sigma,CF}(I,a),\quad(12)$$

Srinivasan 等人进一步指出，A2C（Advantage Actor-Critic，A2C）算法等同于采用 GIGA[89] 代替 RM 作为策略更新规则的 CFR 算法. 然而 GIGA 在相同条件下并没有无遗憾的保证. 由于策略间存在相互克制关系，即博弈具有非传递性[90]（Intransitive），独立运行单智能体强化学习算法习得的策略轨迹在多智能体环境中往往会陷入循环甚至发散于纳什均衡[90-91].

基于这种联系，最早于单智能体领域涌现了一系列融入无遗憾思想的强化学习算法. Jin 等人提出，使用虚拟价值代替原先的价值在部分可观测场景中具有更好的性能[92]. Kash 等人在 Q-learning 基础上结合无遗憾思想提出了 LONR 算法，并证明了其在 MDP 中的收敛性[93]. Yasin 等人从策略迭代角度提出了 Politex 算法[94]，并给出了算法在 MDP 中的遗憾值上界. 在多智能体领域，Srinivasan 等人给出了使用 PG/AC 框架的策略更新流程（Policy Iteration，PGPI/ACPI）的遗憾值上界，并结合无遗憾算法的动态，提出了遗憾策略梯度（Regret Policy Gradient，RPG）和遗憾匹配策略梯度（Regret Matching Policy Gradient，RMPG）[95]：

$$\nabla^{\mathrm{RPG}}_\theta(I)=-\sum_a\nabla_\theta\lfloor A^{w,\theta}(I,a)\rfloor_+$$

$$\nabla^{\mathrm{RMPG}}_\theta(I)=\sum_a\nabla_\theta\pi^\theta(I,a)\lfloor A^{w,\theta}(I,a)\rfloor_+\quad(13)$$

其中由策略参数 θ 和值参数 w 估计的优势函数 $A^{w,\theta}(I,a)=q^w(I,a)-\sum_b\pi^\theta(I,b)q^w(I,b)$. Hennes 等人从演化博弈论[96]中的经典算法模仿者动态（Replicator Dynamic，RD）出发提出 NeuRD 算法[97]，分别证明了使用 Hedge[98] 作为遗憾最小

化准则的 CFR 算法和使用 Softmax 作为策略更新准则的 PG 算法与 RD 间的联系，并基于 RD 改写梯度更新式：

$$\nabla_{\theta}^{\mathrm{NeuRD}}(I)=\sum_{a}\nabla_{\theta}y^{\theta}(I,a)A^{w,\theta}(I,a) \quad (14)$$

其中 $y^{\theta}(I,a)$ 是策略网络未馈入 Softmax 层之前的 logit 值. 相较 PG 算法，NeuRD 使用 logit 值代替策略输出 $\pi^{\theta}(I,a)$，消除了梯度更新时容易受到劣势动作误导的缺陷，使 Softmax PG 在博弈中具有严格的收敛性保证[97]. Lockhart 等人提出了一种非自博弈的策略梯度算法可利用度下降（Exploitability Descent，ED）[99]，假定对手采取最佳响应策略的前提下进行策略优化，这与一类 CFR 算法的变种 CFR-BR 做法类似[100]. 虽然 ED 算法具有最终策略收敛性保证，但由于训练的非对称性以及最佳响应的粗略计算，双方玩家的策略往往不能同时收敛至纳什均衡. Morrill 等人从序贯理性（Sequential Rationality）的角度证明了上述这类基于 PGPI/ACPI 更新流程算法的无遗憾性质. 为充分发挥深度强化学习的规模可扩展性优势，Gruslys 等人提出 ARMAC 算法，通过存储历史策略镜像备份、结合离策略强化学习等技巧，有效解决无遗憾算法在涉及神经网络时的估值问题[101]. Fu 等人从相似的角度基于 AC 框架提出 ACH 算法，并在相较德州扑克具有更大信息集规模的双人麻将中取得了良好成果[102]. DeepMind 基于 NeuRD 与 FoReL 算法训练所得 DeepNash 智能体，在西洋陆军棋（Stratego）中取得与人类职业玩家匹敌的对弈能力[22].

在 EFG 模型之外，与强化学习更为接近的 POSG 模型领域之中也已涌现一系列丰富的研究工作. 在假设转移矩阵与奖励函数已知的前提下，一些早期工作如 Nash-Q[103]、Friend-or-Foe Q[104] 等研究了 POSG 中的均衡求解问题. 一系列后续工作专注于提升在这种博弈动态完全的反馈条件下的遗憾最小化求解方法的计算性能[105-107]. 但这些研究往往不能直接扩展至环境动态信息受限的场景中来. 近年来，一些工作在部分信息反馈条件下研究了算法的解耦[108]、最终策略收敛[109]、对称收敛[110]等性质，并取得了显著进展. 本文聚焦于序贯非完美信息博弈决策相关的求解方法，更多有关 POSG 领域的研究进展推荐读者阅读文献[37-38].

3.4 基于在线凸优化的求解方法

近年来，一系列通过在线凸优化（Online Convex Optimization，OCO）视角解释在线学习的文献为无遗憾的求解方法提供了新的设计框架. 在线凸优化是传统一阶方法（First-Order Methods，FOM）的在线变体，其更新式中往往带有参数并利用一阶梯度信息进行迭代更新，例如 FTRL[111] 和 OMD[112-113] 算法. 这类方法的难点之一在于如何设计距离生成函数（Distance Generate Function，DGF）以保证算法可行性. 对应于 EFG 的序贯策略表示，Hoda 等人最早提出一种特定的凸多面体类型 Treeplex，并设计了与之相关的增广 DGF 函数以满足 OCO 算法在 EFG 情景下的凸性保证[114]. Kroer[115] 和 Farina[116] 随后分别进一步研究了以熵和欧氏距离作为增广 DGF 形式的一阶算法收敛性质，并从理论上保证了优于 CFR 的收敛速度.

尽管 OCO 系列算法与无参数、离散化的 RM 算法在更新方式上看似不同，但事实上这两类算法存在密切联系. Waugh 首先提出了 RM 算法与 Dual Averaging[117] 在特定的学习率设定下是等价的[80]. Burch 对比了 RM 和 RM+算法与镜像梯度下降（Mirror Descent，MD）以及其他近端梯度下降方法[118]. Farina 于 2020 年通过布莱克威尔可达性（Blackwell Approachability），证明了 FTRL 和 OMD 可分别在一定条件下由运行 RM 与 RM+所得[61]. Liu 等人进一步在 EFG 的全局视角下证明 CFR 和 CFR+算法可被视为 FTRL 与 OMD 的一类特殊情况[119]. 基于这种联系，一系列文献从局部更新、最终收敛性、采样复杂度三个方面探索了 OCO 系列算法与 CFR（RM）算法统一通用的改进.

在局部更新方面，类似于 CFR 算法，Farina 首先提出 OCO 算法的局部更新范式，并在此基础上提出了统一保凸算子以应对任意复合凸集上的遗憾最小化流程[120]. Anagnostides 等人进一步提出了一种针对于凸包的遗憾值复合方法[121]，并能够保有 RVU 性质[122]. 后续一些工作通过遗憾值复合建立了更多 OCO 算法的局部更新收敛性保证[123-124].

在最终策略收敛性层面，一系列融入乐观（optimistic）性质的无遗憾算法近年来在各类博弈领域[125-129]取得了理论进展. 乐观信息的本质上是在一阶方法的基础上引入二阶信息对梯度进行修正，在多数博弈场景下能够引导策略向均衡点收敛[130]. 但这类方法同时也面临着高阶梯度近似困难、内循环更新增加开销等难点. 另一系列方法则

在无遗憾算法的更新式中加入额外的正则项/损失项[105,131],通过赋予优化目标额外的凸性以加快算法收敛速度甚至达到最终策略的收敛.基于这种思路,Perolat 从 FTRL 算法入手,将正则博弈中庞加莱回归的结果拓展至扩展式博弈,通过在博弈中加入策略相关的奖励使得策略具有了最终策略的收敛性[132].另一些工作将额外正则项中与模仿学习策略的约束,在大型博弈求解上取得了优秀效果[24].

在采样复杂度层面,如何在预先不知晓博弈结构只通过重复博弈进行更新的老虎机设定(bandit setting)下取得探索和利用的平衡是另一重大挑战.改进采样复杂度的工作主要可分为蒙特卡洛采样、基于模型探索以及损失函数估计三类.这三类改进的代表工作及采样复杂度汇总见表 3.

在线凸优化的视角为基于强化学习和基于遗憾值匹配的求解方法间建立了纽带.Tomar 等人从在线学习的角度入手,结合 MD 的更新方式提出了新的 PPO 算法[140].Grudzien 等人进一步提出了更为广泛的镜像学习(Mirror Learning)的框架,并指出大多数强化学习算法都可以通过这个框架推导出[141].Sokota 等人基于这种联系,提出了一种在强化学习和博弈论上通用的算法 MMD,使得不同领域间算法第一次能够在同一框架下被讨论[142].

表 3　OCO 算法的采样复杂度改进对比

方法类型	典型算法	采样复杂度
蒙特卡洛采样	Farina et al.[133]	$\tilde{\mathcal{O}}(poly(X,Y,A,B))/\varepsilon^2)$
	Bai et al.[134]	$\tilde{\mathcal{O}}((XA+YB)/\varepsilon^2)$
基于模型探索	Farina et al.[135]	—
	Zhang et al.[136]	$\tilde{\mathcal{O}}(S^2AB/\varepsilon^2)$
	Zhou et al.[137]	—
损失函数估计	Kozuno et al.[138]	$\tilde{\mathcal{O}}((X^2A+Y^2B)/\varepsilon^2)$
	Farina et al.[139]	$\tilde{\mathcal{O}}((X^4A^3+Y^4B^3)/\varepsilon^2)$

3.5　小　结

本节介绍了三类遗憾最小化的求解算法及其改进.这三类方法本质上都可被视为在线学习的不同表现形式,但在各自领域的发展过程中分化出许多算法独有的特征,表 4 对这三类方法的代表算法、特点以及优劣进行了对比总结.如何结合凸优化领域的理论保证性与强化学习领域的规模可扩展性,从而设计更为通用与高性能的智能体,是未来算法设计与研究的重要方向.

表 4　基于遗憾的均衡求解方法对比分析

算法类型	代表算法	特征概述	优势	缺陷
基于遗憾值匹配的方法	CFR[63] MCCFR[64] CFR+[56] Deep CFR[143] DREAM[84]	通过局部遗憾值分解方式,在信息集层次使用 RM 算法最小化全局遗憾	• 无参数/梯度,更新式简明 • 具有强力的收敛性质 • 易于结合采样和搜索技术 • 可与基于值强化学习关联	−适用范围小,要求博弈具有终止状态并具有完美回忆性 −难以扩展至连续动作域 −规模可扩展性较差
基于强化学习的方法	Nash-Q[103] RPG/QPG[144] NeuRD[97] ED[99] ACH[102]	在 AC 框架下,根据遗憾值和优势函数关联,运用演化博弈等知识设计强化学习更新范式	• 可对接强化学习技巧,例如通信、探索、分层等 • 利用策略梯度可应对大型决策空间 • 规模可扩展性较强	−理论保证较弱,一些结论需满足强假设 −难以实现复杂的优化更新技巧,例如二阶信息(乐观)的引入
基于在线凸优化的方法	Farina et al.[116] ei et al.[128] Liu et al.[119] MMD[142] FoReL[132]	根据在线凸优化的视角,从传统在线学习中的一阶方法入手进行策略更新和学习	• 通用求解的框架,可拓展至多数无遗憾算法 • 可利用正则项引导策略 • 通过 Treeplex 可在博弈全局进行优化	−对参数敏感,需调度超参 −受限于凸性假设,难以处理非线性问题 −从理论上仍需克服延迟奖励、噪声梯度等问题

4　基于最佳响应的均衡求解方法

基于最佳响应的方法是求解非完美信息博弈的另一大关键方法.在每次迭代中,基于最佳响应的方法通过针对对手的历史行为选择最佳响应作为行动策略以逐步达到均衡解.与基于遗憾的算法相比,基于最佳响应的均衡求解方法可被视为在每一轮都选择遗憾值最大的动作.这种理念使得该类算法无需遍历所有动作以计算遗憾值,扩展性增强.同时,由于计算最佳响应这一模块可单独对接强化学习方法,基于最佳响应的算法往往比基于遗憾值的算法更加适用于大型非完美信息博弈场景.表 5 总结了这两类算法的区别点.

本节首先从两类经典的最佳响应算法虚拟博弈和 Double Oracle 出发,分别介绍他们的原理及改

进，随后介绍通用的种群训练框架，并在此视角上对基于最佳响应的均衡求解方法进行对比分析.

表 5 基于遗憾值算法与基于最佳响应算法的比较

	基于遗憾值的算法	基于最佳响应的算法
主要算法代表	CFR 及其变种	NFSP，PSRO 及其变种
策略更新标准	所有更好的动作（better reply）	最优动作（best reply）
交叉领域	遗憾与强化学习中的优势函数（Advantage）相近；可迁移一些在线学习（online learning）中的技巧	计算最佳响应的部分可以采用强化学习求解
规模扩展难度	由于需要考虑所有动作，受制于信息集和动作规模，扩展性弱	能够借鉴强化学习扩展场景规模的经验，也可与元学习（meta-learning）结合，扩展性强
模型依赖性	往往需要知道博弈树结构，效用函数等模型信息，模型依赖性强	可利用无模型的强化学习，适用于黑箱环境，模型依赖性弱
应用场景	小型非完美信息博弈，要求动作离散，大多局限于双人零和博弈	大型非完美信息博弈，可适用于高动作规模，在双人零和博弈之外也有一些应用

4.1 虚拟博弈（Fictitious Play，FP）及其变体

作为博弈论中一类最经典的自博弈算法，虚拟博弈[145-147]（Fictitious Play，FP）在每一轮次根据过往博弈经历维护对手的经验平均策略，并针对该平均策略计算自身的最佳响应策略. 理论证明，这些最佳响应策略的平均能够在双人零和博弈、势博弈等特定博弈中收敛至均衡点[148]. 上述虚拟博弈的流程可形式化描述为 $\forall i \in \mathcal{N}$：

$$a_{i,*}^{t} \in \mathrm{BR}^{i}\left(\sigma_{-i}^{t}=\frac{1}{t}\sum_{\tau=0}^{t-1}1\{a_{-i}^{\tau}=a, a\in\mathbb{A}\}\right),$$

$$\sigma_{i}^{t+1}=\left(1-\frac{1}{t}\right)\sigma_{i}^{t}+\frac{1}{t}a_{i,*}^{t} \tag{15}$$

第一步计算针对对手过往平均策略 σ_{-i}^{t} 的最佳响应策略 BR^{i}，第二步将 BR^{i} 加入自身的过往平均策略 π_{i}^{t} 中. 与 CFR 一样，FP 的学习流程与强化学习中的 Actor-Critic 框架具有相似之处，具体而言，FP 将玩家的策略（Actors）不断向着其最新的最佳响应策略（Critics）方向改变. 在虚拟博弈中玩家只需要根据自身奖励的完整信息计算最佳响应，而无需了解对手的奖励信息.

4.1.1 广义弱虚拟博弈（Generalized Weakened Fictitious Play，GWFP）

针对 FP 在一些博弈中不收敛的问题[48,149]，已有不少针对 FP 的改进算法[150-153]. 其中广义弱虚拟博弈[154]在一系列放宽虚拟博弈收敛条件的工作[155-156]基础上，在近似计算最佳响应和更新平均策略存在扰动的情况下，证明 FP 算法依然能够收敛至均衡点，并能扩展到更多博弈场景. 记近似最佳响应为 $\mathrm{BR}_{\epsilon}(\sigma_{-i})$，其满足放宽的最优性条件 $u^{i}(\mathrm{BR}_{\epsilon}(\sigma_{-i}),\sigma_{-i}) \geqslant \sup_{\sigma\in\sum_{i}}u^{i}(\sigma,\sigma_{-i})-\epsilon$. 设轮次 t 计算策略时的扰动为 M^{t}，将公式(15)中平均策略的计算改写为：

$$\sigma_{i}^{t+1}=(1-\alpha^{t+1})\sigma_{i}^{t}+\alpha^{t+1}(\mathrm{BR}_{\epsilon}^{i}(\sigma_{-i})+M_{i}^{t+1}),\ \forall i\in N \tag{16}$$

其中 α^{t} 是一个与轮次相关的参数. 扰动 M^{t} 和近似最佳响应 $Br_{\epsilon}^{i}(\sigma_{-i})$ 的加入分别对应了 GWFP 中广义（generalized）和弱（weakened）的定义. 当 α，ϵ，M 满足一定条件时，GWFP 具有收敛性保证. 值得注意的是，FP 是 GWFP 在 $\alpha^{t}=\frac{1}{t}$，$\epsilon^{t}=0$，$M^{t}=0$ 下的一种特殊情况. GWFP 为 FP 与强化学习建立了桥梁. 首先，最佳响应 $\mathrm{BR}^{i}(\sigma^{-i})$ 的计算不再局限于精确求解，使得用强化学习来近似计算最佳响应 $\mathrm{BR}_{\epsilon}^{i}(\sigma^{-i})$ 成为可能. 其次，扰动项 M^{t} 的加入使策略探索机制成为可能，若将某种正则项作为扰动加入最佳响应的计算式中，可以得到类似强化学习中基于熵最大化的策略更新过程[157-158].

4.1.2 扩展式虚拟博弈（Extensive-Form Fictitious Play，XFP）

尽管虚拟博弈在求解正则博弈问题上取得了成功[159-161]，但在扩展式博弈，尤其是非完美信息博弈上一直缺乏有效的更新范式和较好的理论性能保证. 如 2.2 节所述，虽然可通过 Kuhn 定理[162]将扩展式博弈转变为正则博弈，但这种做法面临着策略数目指数增长的困境. Hendon 等人于 1996 年提出了 FP 在扩展式博弈下的两个变种[163]，在完美信息博弈上能够收敛至序列均衡[164]，但在非完美信息博弈上无收敛性保证. 直到 2015 年 Heinrich 等人[43]利用序贯策略实现等价的结论[29]，提出了第一个在非完美信息博弈下能够收敛的扩展式虚拟博弈（Extensive-Form Fictitious Play，XFP）算法. 形式化地，设 σ_{i}^{1} 和 σ_{i}^{2} 是玩家 i 的两个不同的行为策略集，Π_{i}^{1} 和 Π_{i}^{2} 是对应的实现等价的混合策略. 按权重组合后的行为策略 $\tilde{\sigma}(I)$ 的计算公式为

$$\tilde{\sigma}_i(I)=\sigma_i^1(I)+\frac{\alpha\pi_i^{\sigma^2}(I)}{(1-\alpha)\pi_i^{\sigma^1}(I)+\alpha\pi_i^{\sigma^2}(I)}\cdot(\sigma_i^2(I)-\sigma_i^1(I)) \tag{17}$$

根据结论，新的行为策略 $\tilde{\sigma}_i(I)$ 与混合策略 $\hat{\Pi}_i=(1-\alpha)\Pi_i^1+\alpha\Pi_i^2$ 实现等价[39]. 若将平均策略 π 和最佳响应策略 β 分别代入 σ^1 和 σ^2，便可等价实现 GWFP 更新平均策略的步骤. 而行为策略与信息集数目线性相关，大幅降低了策略数目.

4.1.3　虚拟自博弈（Fictitious Self Play，FSP）

Heinrich 于 2015 年[43]提出 FSP 算法，将从经验中学习的思想加入 XFP，充分发挥 XFP 序贯策略易于结合采样的优势，解决了 XFP 需遍历博弈树的缺陷. FSP 的算法框架设计主要分为强化学习最佳响应策略和监督学习平均策略两方面，如图 16 所示.

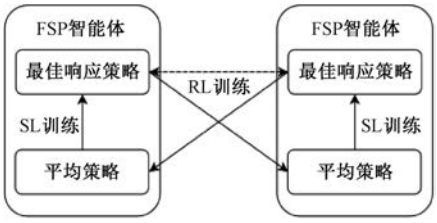

图 16　FSP 算法框架图

在最佳响应策略方面，Greenwald 等人于 2013 年指出，给定一扩展式博弈 Γ 以及行为策略集 σ，那么 $\forall i\in N$，对手策略 σ^{-i} 构成一个 MDP. $\mathcal{M}(\sigma^{-i})$[165]，其中状态为玩家 i 的信息集，奖励为博弈的回报函数，转移动态由对手策略 σ^{-i}、随机因素与博弈规则共同决定[166]. 因此，计算最佳响应策略的步骤可通过强化学习进行建模并求解. 而平均策略的学习本质上可视为以过往策略做出的行为作为标签训练一个策略模型，这与 5.2 小节对手建模的想法类似，不同之处在于此处是对于自身的平均策略进行建模. 此外，当玩家收集经验时，信息集出现于缓冲池中的概率收敛于该信息集的内因到达概率[167]，这恰好满足了在计算平均策略时需加入内因到达概率为权重的要求.

Heinrich 于 2016 进一步提出了 NFSP[168]，在 FSP 的基础上实现了以下三点改进：首先，NFSP 利用神经网络来进行 FSP 中的函数拟合. 其次，在缓冲池设计上，NFSP 采用水库采样[169]（reservoir sam-pling）代替了 FSP 中的循环队列数据结构，在定长的缓存池上能够模拟存储无限长度的策略序列. 最后，NFSP 加入了预期动态[170]（anticipatory dynamic）机制，使所有智能体能够同时进行采样和学习. 预期动态在每轮博弈开始前，以参数 η（anticipatory parameter）来分配玩家采用最佳响应策略或平均策略中的一种：

$$\sigma\leftarrow\begin{cases}\epsilon-\text{greedy}(Q), & \text{以概率 }\eta\\ \prod, & \text{以概率 }1-\eta\end{cases} \tag{18}$$

在加入 η 后，算法能够实现策略的共同演化[171]，而无需再关注策略的分配问题，解决了 FSP 在每一轮博弈中需要异步交替收集双方玩家经验的缺陷. 同时学习的特性使得 NFSP 能够节省玩家数目 $|P|$ 倍的采样开销，并能够直接运用于无法全局协调所有玩家的黑箱（black-box）环境中. 图 17 展示了 NFSP 的训练框架.

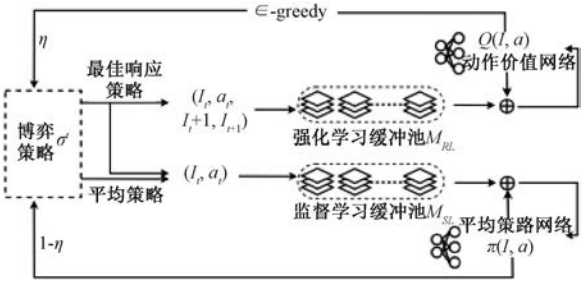

图 17　NFSP 训练框架流程图

尽管 NFSP 在众多应用场景中取得了成果[172-174]，但其仍存在尚待改进之处. 例如在最佳响应的计算过程中，NFSP 需要通过强化学习多次迭代完整博弈树，无法利用博弈树中的子结构，收敛时间较长. 在平均策略的计算过程中，NFSP 借助蓄水池采样来模拟长时段内经验回放的数据流，经验利用率低，且数据拟合的表现易出现偏差. 已有一些工作通过结合遗憾最小化[175]、蒙特卡洛采样[176]、优先经验回放[177]等方法对 NFSP 存在的问题进行改进. 未来，如何进一步提升 NFSP 的算法性能和适用领域是一个重要的研究方向.

4.2　Double Oracle 算法及其变体

Double Oracle（D. O）是一个求解零和正则博弈的迭代算法[178]. 在初始阶段，D. O 算法会为每个玩家初始化一个候选动作集 $\mathbb{A}_i^0\subset\mathbb{A}_i$. 在每一轮次中，算法针对由联合候选动作集 $\mathbb{A}^t$ 组成的子正则博弈 G^r 计算纳什均衡 σ^t. 随后，针对计算所得的纳什均衡策略 σ_{-i}^t，每个玩家选择最佳响应动作 $a_{i,*}^{t+1}\in\mathbb{A}_i$，并将其加入自己的候选动作集 $\mathbb{A}_i^{t+1}=\mathbb{A}^{t,*}{}_i\cup\{a_{i,*}^{t+1}\}$. 上述 D. O 算法流程可描述为

$$\sigma_*^t=\text{NE}(G^r),\ G^r=Generate(\mathbb{A}^t),\ a_{i,*}^{t+1}\in\text{BR}^i(\sigma_{-i,*}^t),\ \mathbb{A}_i^{t+1}=\mathbb{A}_i^t\cup\{a_{i,*}^{t+1}\},\ \forall i\in N \tag{19}$$

D. O 算法和虚拟博弈的不同之处在于，玩家的混合策略不是历史平均策略，而是由候选动作集组成的子正则博弈上的纳什均衡. 尽管在最坏的情况下，D. O 算法仍需遍历整个动作空间才能得到原始正则博弈的纳什均衡，但在实际运用中，通常只需要在小部分动作候选集上就能找到最终纳什均衡.

然而在扩展式博弈场景中，D. O 算法无法遍历随着信息集数目指数增长量级的纯策略，计算复杂度高. 为将 D. O 算法拓展至扩展式博弈场景，McAleer 等人在 2021 年提出扩展式 Double Oracle (Extensive-Form Double Oracle，XDO)算法[179]. 类似 D. O 算法，XDO 算法在每一轮次中构建一个子扩展式博弈，并通过求解该子博弈上的纳什均衡来逼近全局纳什均衡. 具体而言，XDO 算法为每个玩家额外维护一个纯策略集 Π_i^t，该集合由每个轮次中求解子博弈所得的最佳响应（纯策略）构成. 在每一轮次中，XDO 算法首先会为每个玩家创建候选动作集，其中只含有构成纯策略 $\sigma_i \in \Pi_i^t$ 的动作. 通过候选动作集在原始博弈上限制玩家的动作空间，XDO 创建一个子扩展式博弈 G^r，并计算该子博弈上的纳什均衡策略 σ_*^t. 随后，针对对手的纳什均衡策略 $\sigma_{-i,*}^t$，每个玩家计算自己的最佳响应 BR^i 并加入纯策略集 $\Pi_i^{t+1}=\Pi_i^t \cup \mathrm{BR}^i$. 其形式化流程如下：

$$\mathbb{A}_i^t(I_i)=\{a \in \mathbb{A}_i(I_i): \exists \sigma_i \in \Pi_i^t, \sigma_i(I_i,a)=1\},$$
$$\sigma_*^t=\mathrm{NE}(G^r),\ G^r=Generate(\mathbb{A}^t),$$
$$\Pi_i^{t+1}=\Pi_i^t \cup \mathrm{BR}^i(\sigma_{-i,*}^t),\ \forall i \in \{1,2\} \tag{20}$$

图 18 从左到右依次给出了 XDO 算法的三次迭代，其中实线表示候选动作集，而蓝线表示子博弈中纳什均衡策略的动作. 在第一轮中，玩家 1 和玩家 2 同时添加最佳响应到纯策略集中并构造候选动作集. 在第二轮中，针对候选动作集组成的子博弈，玩家 1 的最佳响应为 $\sigma(I_1^1)=\sigma(I_1^2)=$“右”，玩家 2 的最佳响应为 $\sigma(I_2^1)=$“右”. 在第三轮中，玩家 1 的纳什均衡策略中只有“右”单一动作，针对该子博弈上的纳什均衡，玩家 1 不改变策略，玩家 2 的最佳响应为 $\sigma(I_2^1)=$“左”. 此时玩家不再更新候选动作集，算法终止并收敛至全局纳什均衡策略. XDO 算法保证了在多次迭代后收敛至近似纳什均衡，且该迭代次数与信息集数量呈线性关系[179]，远小于 D. O 算法.

NXDO (Neural XDO)算法[179]进一步使用深度强化学习近似 XDO 算法的最佳响应计算，并使用策略作为候选动作集中的动作，称为元动作

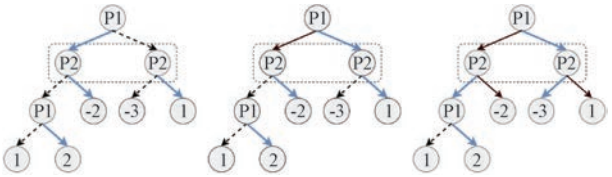

图 18 XDO 在简单游戏中的三次迭代（从左到右）[179]

(Meta-Action)：

$$\mathbb{A}_i^t(I_i)=\{1,2,\cdots,|\Pi_i^t|\},\forall I_i \in I_i,\forall i \in N \tag{21}$$

而实际使用的动作则由选择的元动作采样得到. 当构建子博弈后，NXDO 算法使用例如 NFSP[168]、Dream[84]等深度学习算法来求解该子博弈上的 ϵ — 纳什均衡 σ^t，同时使用 PPO[180]、DDQN[181]等深度强化学习算法针对 σ^t 为每个玩家计算出近似最佳响应 $\mathrm{BR}^i(\alpha_{-i}^t)$，并将其添加到当前种群中 $\Pi_i^{t+1}=\Pi_i^t \cup \mathrm{BR}^i(\alpha_{-i}^t)$. 元动作的引入很好地解决具有大型或连续动作空间的博弈问题. Online D. O (ODO)通过在 D. O 算法中融入在线学习的思想，并将 D. O 与纯策略集大小相关的最坏情况时间复杂度降低为与有效策略集相关，提升了算法的期望收敛性能[182]. 后续工作进一步结合遗憾最小化的成功实践经验，提升了 D. O 算法的采样复杂度[183].

4.3 基于博弈论原则的种群训练

近年来，基于博弈论原则的种群训练方法被应用至多个复杂大规模博弈场景，有效解决了离线训练过程中易出现的泛化性和策略多样性问题，并进一步拓展了现实应用[18,184-185]. 种群训练方法迭代地维护和训练一个不断增长的智能体种群，并利用纳什均衡等博弈论解概念或者 Elo 等评分排序系统促进种群进化以提高集体性能. Policy Space Response Oracles (PSRO)算法是种群训练方法中最为通用的求解框架. 基于元博弈(Meta-Game)的概念，PSRO 算法归纳总结了多种种群训练方法. 本节将对 PSRO 及其变体进行讨论分析.

4.3.1 元博弈

元博弈，也称经验博弈 (Empirical Game)，是经验博弈论分析的主要对象[186-187]. 在类似围棋或者星际争霸等复杂大型博弈场景中，由于底层动作空间巨大，构建所有动作间的收益矩阵是相当低效的. 与之不同，元博弈关注于玩家之间不同策略的相互影响，是原始博弈在策略层面的抽象. 元博弈使用 Π 定义一组策略（称其为种群），并将种群中

每个策略 $\pi_i \in \Pi$ 视为一个“动作”(称为元动作).针对种群中不同策略的博弈结果,可以构建收益矩阵 $\phi(\pi_i,\pi_j)=-\phi(\pi_j,\pi_i)\in[-1,1]$,其中,若 π_i 击败 π_j,则记为 $\phi(\pi_i,\pi_j)>0$;若 π_i 和 π_j 平手,则记为 $\phi(\pi_i,\pi_j)=0$;若 π_i 被 π_j 击败,则记为 $\phi(\pi_i,\pi_j)<0$. 元博弈的做法有效降低了收益矩阵的规模,且显式展现出不同策略间的强弱以及克制关系,这有利于种群训练中制定挑选对手的规则.

4.3.2　Policy Space Response Oracles (PSRO)

Policy Space Response Oracles[188] (PSRO)是一个用于求解大型双人零和博弈问题的种群训练方法,其本质是 D. O 算法在元博弈层面上的推广. PSRO 算法为每个玩家维护一个种群 Π_i^t,种群中每个智能体模型代表不同策略. PSRO 算法首先为每个玩家随机初始化一个策略 $\sigma_i^0\in\Pi_i^0$. 随后通过迭代地运用构造、求解、扩展三个阶段来训练并扩展种群. 在构造阶段,PSRO 算法根据每个玩家的种群 Π_i^t 构建一个元子博弈. 在求解阶段中,算法应用元求解器 (Meta-Solver) 在该元子博弈上求解元策略 (Meta-Strategy) α^t. 所谓元策略是指在种群中不同策略(即不同模型)上的分布,如纳什均衡、均匀分布等,元策略定义了对手采样的规则. 在扩展阶段中,算法根据对手的元策略挑选不同的策略模型作为对手,使用强化学习等方法寻找最佳响应 $\mathrm{BR}^i(\alpha_{-i}^t)$,并将该最佳响应加入种群 $\Pi_i^{t+1}=\Pi_i^t\cup\mathrm{BR}^i(\alpha_{-i}^t)$. PSRO 算法的具体流程图如图 19 所示.

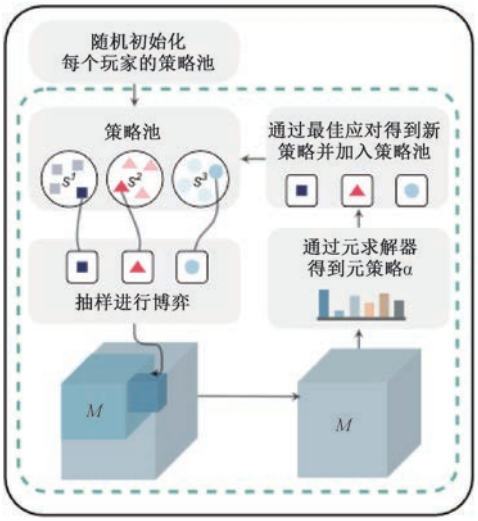

图 19　PSRO 流程图[188]

值得注意的是,当仅挑选最新加入种群的策略作为对手,即元策略为[0,…,0,1]时,PSRO 算法将退化为自博弈算法;当元策略是均匀策略时,PSRO 算法将退化为 NFSP 算法. 表 6 给出了 PSRO 类算法的汇总. PSRO 算法中元求解器和最佳响应具有多种选择. 因此,许多求解纳什均衡的算法都可被视为 PSRO 算法的特例,如自博弈[189]、广义弱虚拟博弈[154]、D. O 算法[178]等.

表 6　PSRO 类算法对比表

算法	(元) 求解器	响应函数
D. O[178]	NE	BR
GWFP[154]	UNIFORM	BR_ε
Self-play[189]	[0,…,0,1]	BR
NFSP[168]	UNIFORM	BR_ε
PSRO_N[188]	NE	BR_ε
$\mathrm{PSRO}_{\mathrm{rN}}$[190]	NE	公式 (23)
α-PSRO[191]	α-Rank	公式 (22)

4.3.3　PSRO 变体

由于 PSRO 算法的灵活性和有效性,后续研究针对不同的问题场景,进一步提出了各类变体算法,如 $\mathrm{PSRO}_{\mathrm{rN}}$[190]、α-PSRO[191]、DPP-PSRO[192]、UDM-PSRO[193]等. 由于求解多人博弈场景的纳什均衡具有极高的计算复杂度[26,194],导致 PSRO 算法仅局限于双人零和博弈. 为了解决这一问题,α-PSRO 算法使用 α-Rank[195]作为元求解器,并使用 α-Rank 解来代替纳什均衡作为解概念. α-Rank 解在多人一般和博弈场景中不仅具有多项式时间的求解复杂度,还具有唯一性,避免了均衡选择问题. 为了收敛到 α-Rank 解,α-PSRO 算法使用基于偏好的最佳响应 (Preference-Based Best Response, PBR) 作为响应函数,在多人一般和博弈场景的实验中均表现出较好的结果:

$$PBR(\alpha_{-i}^t)=\underset{\sigma_i\in\Pi_i}{\operatorname{argmax}}\mathbb{E}_{\sigma^t\sim\alpha^t}\left[1_{\phi_i(\sigma_i,\sigma_{-i}^t)>\phi_i(\sigma_i^t,\sigma_{-i}^t)}\right]. \tag{22}$$

与之类似,Marris 等人则利用相关均衡来代替纳什均衡作为解概念,提出 Joint PSRO 以解决多人一般和博弈问题[196].

在实际场景中,许多实际博弈问题都表现出强烈的非传递性[91,192-193]. 所谓非传递性是指,所有的策略 $\{\pi_i\}_{i=1}^l$ 组成一个长度为 l 的循环,使得对任意一个 $i>1$ 有 $\phi(\pi_i,\pi_{i+1})>0$,且有 $\phi(\pi_1,\pi_l)>0$. 石头剪刀布是非传递性博弈的一个经典案例. 传递性博弈则是指,若 $\phi(\pi_i,\pi_j)>0,\phi(\pi_j,\pi_k)>0$,则能推出 $\phi(\pi_i,\pi_k)>0$. 事实上,任何一个泛函式博弈 (Functional-Form Game, FFG) 都可以分解为一个传递性博弈 (Transitive Game) 和非传递性博弈[190,197] (Intransitive Game). 在这些场景

中，简单使用自博弈算法将会陷入无限的策略循环中，无法达到均衡解. 一个有效的解决思路是通过促进种群策略的多样性以帮助算法跳出当前策略循环. PSRO_{rN}、DPP-PSRO 和 UDM-PSRO 分别提出了不同的基于多样性的最佳响应函数以促进种群的多样性. PSRO_{rN} 的公式如下：

$$\text{BR}^{i}_{\text{rN}}(\alpha^{t}_{-i})=\left\{\sigma_i:\sum_{\sigma^{t}_{-i}\in\Pi^{t}_{-i}}\alpha_{-i}(\sigma^{t}_{-i})\cdot\lfloor\phi(\sigma_i,\sigma^{t}_{-i})\rfloor_{+}>0\right\} \tag{23}$$

DPP-PSRO 和 UDM-PSRO 的公式如下：

$$\text{BR}^{i}_{\text{DPP/UDM}}(\alpha^{t}_{-i})=\underset{\sigma_i}{\text{argmax}}\sum_{\sigma^{t}_{-i}\in\Pi^{t}_{-i}}\alpha_{-i}(\sigma^{t}_{-i})\cdot\phi(\sigma_i,\sigma^{t}_{-i})+\tau\cdot\text{Diversity}(\Pi^{t}_{i}\cup\{\sigma_i\}) \tag{24}$$

其中，Diversity 代指各类多样性度量指标计算公式，其衡量了一个种群中的策略多样性.

此外，PSRO 算法还有一些重要改进. 针对 PSRO 算法求解元博弈时间开销大的缺点，McAleer 等人提出了 Pipeline PSRO 算法[198]，其在 PSRO 算法的迭代流程中引入并行化思想，在大型零和博弈场景中能很快收敛到近似纳什均衡. 为了确保 PSRO 的可利用度在每个阶段能够单调下降，McAleer 等人提出适用于双人零和博弈的 anytime PSRO[199]. 为了缓解 PSRO 中的计算低效与探索低效问题，Zhou 等人提出 Efficient PSRO 算法[200]，通过引入新的元子博弈构建方法，使得 Efficient PSRO 算法不需要通过求解元子博弈就能找到最佳响应，提高计算和探索效率. Liu 提出了 NeuPL 算法[201]，使用单一条件模型表示种群策略，有效缓解了 PSRO 算法面临的存储空间开销大和基础知识学习重复的缺陷.

4.4 对比小结

本节介绍了基于最佳响应的非完美信息博弈求解算法及其改进. 从本质上看，这一类方法在设计的过程中需要回答两个问题：如何计算最佳响应以及如何利用历史的最佳响应经验以改进策略，这两个问题分别对应着表 6 的响应函数和求解器概念. 表 7 对本节所介绍的三类求解技巧进行了对比总结. 最佳响应求解方法与种群训练框架极大丰富了博弈求解的泛化性与表达性. 未来，如何融合不同博弈中训练所得的通用博弈知识，增加智能体决策的可解释性是一大研究方向.

表 7 基于最佳响应的均衡求解方法对比分析

算法类型	代表算法	特征概述	优势	缺陷
虚拟博弈及其变体	FP[145-147] GWFP[154] XFP[43] FSP[43] NFSP[168]	针对对手的经验平均策略，计算自身的最佳响应策略	·更新式简明 ·具有较强的可扩展性 ·具有较好的收敛性质	－对初值较敏感 －缺乏探索性 －易过拟合特定对战风格
Double Oracle 及其变体	D. O[178] XDO[179] ODO[179] NXDO[179]	针对对手的经验纳什均衡策略，计算自身最佳响应	·具有较快的收敛速度 ·支持高维或连续动作空间 ·具有较强的可扩展性	－局限于双人零和博弈 －无法充分利用非理性对手
Policy Space Response Oracle 及其变体	PSRO_{N}[188] α-PSRO[191] Joint PSRO[196] PSRO_{rN}[190] DPP-PSRO[192] UDM-PSRO[193]	D. O 算法在元博弈层面上的推广，并使用 DRL 计算最佳响应，是通用的种群训练	·框架灵活性、表达性强 ·具有较强的可扩展性 ·无需大量的专家先验知识 ·支持高维或连续动作空间	－最佳响应的计算复杂度高 －收敛性质缺乏理论性保证 －维护和计算种群元博弈的成本昂贵

5 非完美信息博弈在线求解方法

第 3、4 节从离线（Offline）模式的角度介绍了在非完美信息博弈中求解均衡的两类主要算法. 这些离线算法关注于在真实对弈开始前求解均衡策略，并遵循该策略进行对局. 而不同于离线方法，在线求解方法会在对弈过程进行的同时优化策略. 因此，状态只会在对局中被访问时才得以更新. 表 8 总结了在线求解方法和离线求解方法的区别.

由于约简技术、神经网络的使用，离线算法求解得到的蓝图策略（Blueprint Strategy）通常与原博弈纳什均衡仍存在一定距离. 如何在线（Online）地实时优化蓝图策略，并使其得到期望的改进，是在线求解方法的目标所在. 根据策略改进方向区分，在线求解方法可分为驱动策略进一步逼近纳什均衡的搜索（Search）方法和驱动策略竭力利用对手的对手建模（Opponent modelling）方法两类.

搜索算法往往会立足于当前状态，对潜在的未来相关状态进行推理，并借助预先推理的信息启发

表 8　在线求解方法与离线求解方法特征对比

	在线求解	离线求解
使用周期	决策/对局进行时	决策/对局开始前
模型可知性	不需要知晓博弈模型	需知晓博弈模型
求解目的	向特定方向改进策略	使策略趋近于纳什均衡策略
更新范围	可选择部分相关状态进行更新	往往更新博弈树所有状态
初始资源	可基于历史对局、对手风格等信息进行初始分析	往往从零开始 (From scratch)
关注点	关注某一玩家策略，与其他未知对手对局过程中实时改进策略	对所有玩家 (系统) 进行全局调配计算，往往采用自博弈框架

当前状态下的决策,如图 20(a)所示.搜索的实质是在当前状态上划分出一个相关状态的子问题并求解,该子问题常被称为子博弈(Subgame).子博弈的搜索过程中常可以嵌套展开新的搜索,即将子博弈视为一独立的博弈,如图 20(b)所示. 在大规模博弈中,未来状态数目巨大,子博弈搜索可在一定深度截断,转而以价值函数代替后续状态,如图 20(c) 所示. 搜索算法的研究内容主要可分为三个领域:子问题分解、搜索安全性和搜索效率.

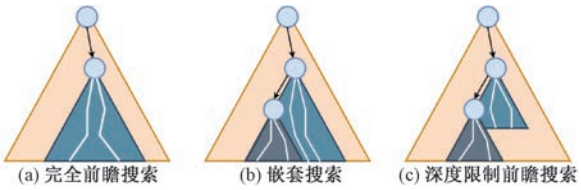

图 20　三种典型的搜索范式[71]

对手建模方法则通过构建模型来推断博弈中其他智能体的意图,预测他们未来的行动或其本身具有的特征,以便更有效地达成自身目的. 不同于搜索方法,对手建模方法往往以最大程度利用对手为目标. 一个对手建模方法的设计包括两个方面:如何建立对手模型以及如何利用模型优化策略. 本节将分别从搜索和对手建模方法出发,围绕三个领域和两个方面的设计,依次论述两类方法的发展历程,并对比分析算法优劣以及区别联系.

5.1　搜索方法

5.1.1　子博弈和价值函数

子博弈是搜索技术中子问题分解的关键概念. 子博弈被定义为在某一决策点决策时,与该决策点相关的最小的决策点集合 (闭包) 所组成的博弈,子博弈仍是一个良定义的博弈,可以独立分析. 在完美信息博弈中,状态完全可知,子博弈是一棵以当前状态为根节点的子树. 而在非完美信息博弈中,由于玩家间决策时需要循环推理对方的私有信息,不同信息集之间具有很强的相关性. 简单地说,在非完美信息博弈中,具有相同公共信息的信息集都是相关的[71]. 因此,非完美信息博弈中子博弈是一棵以当前公共状态为根节点的子树. 此外,由于非完美信息博弈中的价值计算依赖于到达概率,为便于计算,非完美信息下子博弈由以下两部分组成:

(1)以公共状态 $I_{pub} \in I_{pub}$ 为根的子树;

(2)玩家各自信息集的到达概率分布 $\Delta(I_i(I_{\text{pub}}))$.

在一些文献中[202-203],①和②的组合被称为公共信念状态 (Public Belief State, PBS). 图 21 中给出了 RPS 博弈在第二个公共状态 I^2_{pub} 上展开的子博弈.

与子博弈息息相关的是价值函数. 在 2p0s 博弈中,玩家的博弈价值 (Game Values) 被证明是唯一的[204]. 同时,由于子博弈仍然是一个良定义的博弈,2p0s 博弈的子博弈也具有唯一博弈价值. 在搜索方法中,我们可在搜索一定深度后,用子博弈的博弈价值来代替剩余的搜索过程. 在完美信息博弈中,玩家的博弈价值是一个标量. 而非完美信息博弈的公共状态中包含多个信息集. 因此,价值函数将输出一个向量. 表 9 以图 21 为例,对照给出了双方玩家在该子博弈下的信息集到达概率、价值函数以及玩家 2 在子博弈中的最佳响应策略.

5.1.2　搜索的一致性

形式化地,搜索算法 $\Omega^p: I \to \Delta(A_i(I))$ 在当前决策点 I 上将根据已知博弈序列 $p=(z_1, z_2, \ldots, z_k)$ 输出一个行为策略. 搜索的核心目的是在重复博弈过程中,仅通过改变子博弈上策略,来使蓝图策略进一步逼近纳什均衡. 因此,在重复博弈中能够保证玩家所获价值不低于采取纳什均衡所获价值的搜索算法被称为是可靠 (sound) 的[205].

根据满足可靠性的严格程度,搜索算法可分为局部一致性 (Local Consistency),全局一致性 (Global Consistency) 和强全局一致性 (Strong Global Consistency) 三种一致性等级[71,205]. 三种一致性等级的定义对比如表 10 所示. 其中,NEQ 是纳什均衡策略的集合. 局部一致性保证了在每个

表 9 RPS 博弈中玩家不同策略到达概率和价值函数对比示例

$\Delta(I_1(I^2_{pub}))$	$\Delta(I_2(I^2_{pub}))$	ν_1	ν_2	最佳响应 BR_2
(0.2,0.2,0.6)	(1)	(0,1,−1)	(0.4)	$\sigma_2=(1,0,0)$
(0.4,0.3,0.3)	(1)	(−1,0,1)	(0.1)	$\sigma_2=(0,1,0)$
$\left(\frac{1}{3}-\epsilon,\frac{1}{3}+2\epsilon,\frac{1}{3}-\epsilon\right)$	(1)	(−1,0,1)	(ϵ)	$\sigma_2=(0,0,1)$
(1/3,1/3,1/3)	(1)	$\sigma_2\cdot\begin{pmatrix}0&1&-1\\-1&0&1\\1&-1&0\end{pmatrix}$	(0)	任何行为策略 σ_2

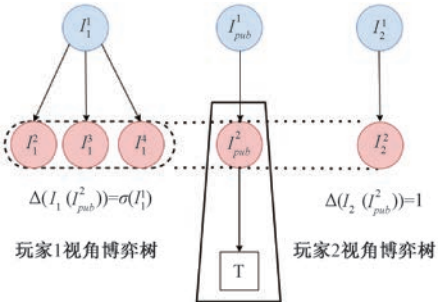

图 21 非完美信息博弈子博弈示例

决策点上，玩家都采取 NEQ 中的某一策略. 全局一致性保证了不同决策点上策略来源于 NEQ 的同一策略. 强全局一致性进一步保证了存在同一策略，搜索算法对于任意博弈序列 p 能够推荐与该策略一致的行为策略.

表 10 三层搜索一致性等级的定义

一致性等级	定义
局部一致性	$\forall p=(z_1,\cdots,z_k)\,\forall h\sqsubset z_k\,\exists\sigma\in\mathrm{NEQ}$：$\Omega^p(h)=\sigma(h)$
全局一致性	$\forall p=(z_1,\cdots,z_k)\,\exists\sigma\in\mathrm{NEQ}\,\forall h\sqsubset z_k$：$\Omega^p(h)=\sigma(h)$
强全局一致性	$\exists\sigma\in\mathrm{NEQ}\,\forall p=(z_1,\cdots,z_k)\,\forall h\sqsubset z_k$：$\Omega^p(h)=\sigma(h)$

在完美信息博弈中，状态价值等同于子博弈完美（精炼）均衡的价值. 因此，满足局部一致性并能保证子博弈完美均衡结构的搜索算法即是可靠的[206]，这也是极大极小搜索（Minimax Search）能在完美信息博弈中生效的理论基础. 然而在非完美信息博弈中，单一信息集上的独立最优决策不足以满足可靠性，子博弈上局部再求解（Resolving）所得策略与离线时全局求解所得策略并不兼容，一些局部的子博弈均衡策略在全局中反而具有很高的可利用度，这要求非完美信息博弈中一个可靠的搜索算法需满足强全局一致性条件.

5.1.3 不安全搜索

在子博弈中展开搜索时，若将到达概率归一化，可将该子博弈视为一个以机会节点为根节点的博弈，该归一化概率也被称为信念（Belief）. 为便于表述，该归一化后的子博弈被称为增广博弈（Augmented Game）. 一些工作直接在增广博弈上进行类似于极大极小算法的局部搜索以更新子博弈策略[78,207-208]. 然而这种搜索方法缺乏理论证明，在许多场景中会得到可利用度高的策略，不具有一致性. 如图 21 所示的子博弈中，即使蓝图策略已近似等于纳什均衡策略，玩家 2 在子博弈的搜索过程中也只能得出采取任何策略获得的收益都为 0 的结论，从而所有策略都具有最优性，如表 9 第 4 行所示. 但显然，并非所有策略都是原博弈的均衡策略，玩家 2 在增广博弈上没有足够信息计算均衡策略.

究其原因，在非完美信息博弈中，局部搜索虽最大化了子博弈价值，但却固定了到达子博弈的信念，这剥夺了双方玩家在子博弈之前的博弈部分改变策略的可能性. 事实上，非完美信息博弈中的局部搜索甚至不具有局部一致性，是不安全的[71].

5.1.4 安全搜索

基于不安全搜索失效的原因，设计安全搜索的一大目标是确定趋向纳什均衡策略的搜索方向. 沿该方向计算最佳策略时，能保证对手没有机会通过改变子博弈前的策略来提升其价值. 为了形式化地刻画上述理念，需要引入虚拟最佳响应（Counterfactual Best Response，CBR）和虚拟最佳响应价值[209]（Counterfactual Best response Value，CBV）的概念. CBR 在 BR 的基础上，进一步严格地规定了在玩家 i 所有信息集 I 上，采取的动作都是最优的，即 $\forall I\in I_i$：

$$v_i^{\langle CBR(\sigma_{-i}),\sigma_{-i}\rangle}(I)=\max_{a\in A(I)}v_i^{\langle CBR(\sigma_{-i}),\sigma_{-i}\rangle}(I,a)\quad(25)$$

公式 26 中的值被定义为 CBV 值，代表玩家 i 采用 $CBR(\sigma_{-i})$ 应对对手策略 σ_{-i} 得到的价值. 借助 CBR 和 CBV 的定义，尽可能减少对手玩家 $-i$

提升价值机会的目标可被转述为在蓝图策略σ的基础上，在子博弈中搜索求解策略σ'，并使得：

$$v_{-i}^{\sigma'}(I) \leqslant v_{-i}^{\langle \sigma_i, CBR(\sigma_i)\rangle}(I) = \mathrm{CBV}_{-i}^{\sigma_i}(I) \quad (26)$$

该约束最早于CFR-D算法中[210]被提出，为便于论述，我们称该约束为CFR-D约束. 通过在搜索中加入CFR-D约束，玩家i搜索所得策略σ'_i在改进方向上与原博弈纳什均衡策略是一致的[209-210]. 满足CFR-D约束的搜索算法具有强全局一致性[205]. 由于CFR-D约束针对于玩家i设立，搜索仅保留新策略σ'中玩家i的策略σ'_i[211]. 在子博弈基础部分上，安全搜索的特征需要在子博弈中加入第三部分：

(3)对手子博弈中的虚拟最佳响应价值$\mathrm{CBV}_{-i}^{\sigma_i}(I)$.

在简单博弈中，可以将公式27作为线性规划的一条约束加入求解过程. 而在扩展式博弈中，可以通过重构博弈的方式构造约束. 具体而言，重构博弈在玩家i的子博弈上增加了一个对手玩家i的虚拟决策节点，如图22所示. 在虚拟节点上对手可以选择表示进入子博弈的动作a_S，或选择动作a_T表示终止搜索进程，转而取得自己在子博弈中最佳响应玩家i的蓝图策略所得价值，即$\mathrm{CBV}_{-i}^{\sigma_i}(I)$. 直观地说，该做法通过给予对手退路，从而鞭策玩家在子博弈中精炼策略. 这种满足某种约束的重构博弈在相关文献中被称为工具博弈(Gadget Game). 工具博弈的好处在于以与原先子博弈量级相同的规模与开销实现了搜索安全性的约束.

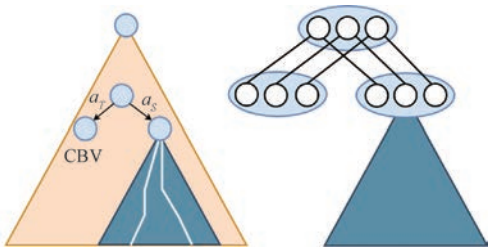

图22　CFR-D约束创建的工具博弈

另一种保证安全性的搜索技术是极大利润搜索[209](Maxmargin Search). 极大利润搜索在CBV的概念上进一步定义了玩家i在子博弈上得到改进策略σ'后获得的利润$M^{\sigma'}(I)$：

$$M^{\sigma'_i}(I) = \mathrm{CBV}_{-i}^{\sigma_i}(I) - \mathrm{CBV}_{-i}^{\sigma'_i}(I) \quad (27)$$

利润$M^{\sigma'}(I)$表达了优化策略σ'_i相较于蓝图策略σ_i精炼的程度，最大化利润等同于最小化对手应对σ'_i的CBV值$\mathrm{CBV}_{-i}^{\sigma'_i}(I)$. 满足最大利润的约束将最大程度地劝阻对手进入子博弈，因此将严格地提升子博弈上玩家的策略. 最大利润搜索也可通过创建工具博弈的方法来实现，如图23所示. 在Maxmargin约束的工具博弈上，对手奖励会被减去针对蓝图策略σ_i的CBV值. 通过赋予对手玩家在子博弈开始前选择信息集的权利，Maxmargin工具博弈保留了对手的对抗意愿，对手会倾向于选择不利于玩家的信息集，而由于零和博弈的性质，玩家自发地去利用最大利润，从而满足Maxmargin约束.

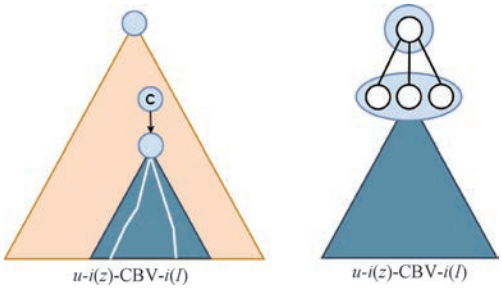

图23　Maxmargin约束创建的工具博弈

5.1.5　持续前瞻搜索和深度限制搜索

在具备一致性和安全性的基础上，完美信息博弈中一些提升搜索效率的技巧完全能够迁移至非完美信息博弈中来. 随着博弈进行，我们可在新到达的公共状态上继续展开搜索. 这种多步搜索技术被称为持续搜索[53](Continual Resolving)，在一些文献中也被称为嵌套搜索[212](Nested Search). 嵌套搜索能够重用上一公共状态所展开子博弈的中间计算结果，以便于当前子博弈的搜索计算，减少了计算开销. Sustr提出了在持续搜索的基础上结合了蒙特卡洛方法[213]，进一步提升了运算效率.

在大型博弈中的初始阶段，即使仅对子博弈展开搜索也引入了巨大开销. 因此，在达到一定深度时截断搜索过程是必要的. 深度限制搜索这种思路在完美信息博弈中已经得到了广泛应用[4-5,184]. 目前在非完美信息博弈中主要有两种实现深度限制搜索的方式. 第一种方法通过价值函数来实现. 然而在非完美信息博弈中，价值函数的向量输出特性对神经网络拟合的准确性提出了极高挑战，例如在西洋陆军棋(Stratego)等复杂博弈中，公共状态中信息集数目难以计数，无法通过神经网络进行准确估值. 另一种方法通过多值状态[211,214](Mutli-Valued States)来实现. 该方法通过为子博弈叶子节点的值函数分配多值，用于对应玩家在剩余搜索过程中可能采取的不同蓝图策略组合. 图24中给出了一套结合安全性、持续前瞻搜索和深度限制搜索三个模块的非完美信息博弈搜索范式.

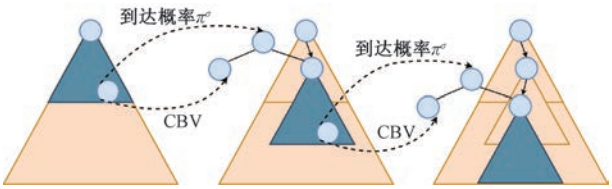

图 24 一套完整的非完美信息博弈搜索的示例[71]

5.1.6 信息集蒙特卡洛搜索（IS-MCTS）

受启发于蒙特卡洛树搜索（MCTS）算法在完美信息博弈中的成功，一些研究尝试将 MCTS 扩展至 IIG 场景中. 不同于利用约束进行推理的安全搜索，MCTS 在 IIG 中的应用采取了更为启发性的方式. 一个简单的思路是利用完美信息 MCTS，在博弈开始前通过采样真实状态进行公共状态确定化，将确定了隐藏信息的 IIG 视为若干个采样所得的 PIG 进行求解[215]. 然而，这种做法在 IIG 场景下面临三个主要挑战[216]：

(1)信息泄露（Information Leakage）；

(2)策略融合（Strategy Fusion）；

(3)非局部最优性（Non-Locality）.

信息泄露指在隐藏信息的确定化过程中，一些私有信息以采样的方式被对手知晓，这种类似“出千”的经验难以迁移至真实的在线场景中. 与采样所得 PIG 对应，IIG 信息集中的不同状态具有不同最优策略，这要求算法对不同最优策略进行融合. 非局部最优性则表现为在 IIG 中，当前状态的最优收益不能递归地由子博弈归纳得到. 上述三个挑战与 5.1.2 节中搜索一致性问题的成因相同，其根本原因在于 IIG 中不同信息集的决策相互制约影响，导致无法对其进行独立或分解分析.

Cowling 于 2012 年提出 IS-MCTS[216]，以信息集为单位聚合进行 MCTS 以保证信息的完整性，缓解了(2)带来的挑战. 然而 IS-MCTS 仍然在一些博弈中面临着(1)和(3)的挑战. 尽管如此，由于 IS-MCTS 具有实现便捷、规模可扩展、适配性高等优点，后续研究人员围绕上述三大挑战提出了一系列研究工作，并在例如斗地主[217]、桥牌[218]、电子游戏[219-220]等博弈分支因数更大的 IIG 场景中得到了广泛应用. 表 11 中总结了围绕这三个挑战由 IS-MCTS 衍生展开的一系列改进工作.

5.2 对手建模

尽管采取纳什均衡策略能够保证玩家至少获得博弈价值，但当面对非理性对手时，其策略具有模式性的瑕疵，玩家偏离均衡策略能够在对手做出错误行为时获得更多的价值. 因此，在实时对局中通过对手建模推理其他智能体意图，从而更加高效的与之互动进一步利用对手，是在线求解算法的另一类重大目标. 此外在非完美信息博弈中，对手建模有助于推断由于隐藏信息存在而更加模糊的对手意图.

表 11 IS-MCTS 的衍生改进总结

算法名	解决挑战	特征概述
PS-MCTS[221]	(1)(2)	通过后验概率更新信念以构建公共状态进行搜索
APMCTS[222]	(1)(3)	通过公共信息和值网络在搜索过程中预测对手动作
DSMCP[223]	(1)(2)	通过使用未加权的过滤信息构建公共状态，并从置信区间抽取样本进行规划
RecPIMC[224]	(2)	基于 PIMC 加入嵌套搜索
ICARUS[225]	(1)(3)	引入策略重用和信息捕获搜索技巧
MAST[226]	(1)(3)	在学习动作价值时独立于上下文以增强模拟效果
RD-ISMCTS[227]	(1)(2)	在搜索过程中根据未揭露的信息重新进行确定化
Cazenave etal[218]	(2)(3)	引入启发式 $\alpha-\mu$ 搜索条件

在如何建立对手模型的问题上，根据对于对手决策假设的严格性，文献研究可分为静态建模，递归推理和元学习三类. 静态建模假设对手策略与自身决策是无关的，这意味着对手模型的建立可以通过面向数据流的各类学习方法得以实现，例如策略构建[228]、对手风格标签[229-230]、类型推理方法[231]、行为识别[232]、对比学习[233]等. 静态建模往往作为学习算法的辅助任务同时进行. 基于循环推理的方法则假设对手也会对玩家进行一定程度的建模，从而玩家在建模对手的同时，需要将对手的建模信息考虑在内，得到更高层次的建模推理[234-235]，这与心智理论[236-237]（Theory Of Mind，TOM）和认知层次[238]（Cognitive Hierarchy）的概念是相通的. 理论上这种推理过程可以递归地持续进行，然而，推理层次的深入也意味着计算复杂度的升高，这类方法在实际使用中通常仅递归推理一层[239-240]. 基于元学习的方法假设对手在玩家学习的同时也在更新策略，从而不仅需要构建对手的当前策略模型，还需要对对手策略更新的趋势进行建模[241-242]. 然而，这类方法往往会作出对手采取与自身相同的算法进行更新的强假设. 表 12 中对这三类方法优缺点进行了对比分析. Yu 等人[243]进一步尝试放宽对于对手的假设，结合基于模型（Model-based）的强化学习将对手的相关信息融入环境建模过程之中，但其中央式的学习方法仍受限于博弈规模的扩展.

如何在大规模非完美信息博弈中高效地进行对手建模是未来可研究的方向.

表 12　三类对手建模方法的优缺点对比

建模类型	优点	缺点
静态建模	可对接面向数据流的机器学习 计算复杂度低,模型通路简单	缺乏适应性,对风格变化捕捉慢 未考虑智能体间决策的相互影响
递归推理	考虑了对手的信念认知能力 可对接心智理论和认知层次模型	高层递归推理计算复杂度高 对环境中智能体的递归推理层数需要一些假设
元学习	考虑了对手的策略学习趋势 可对接持续学习,自适应性强	元学习内外层循环开销巨大 学习趋势的预测准确度较低

在如何优化策略的问题上,安全性和自适应性是两个利用对手模型的重要性质.尽管通过对手建模能够最大化玩家对于非理性对手的收益,但偏离均衡策略意味着可利用度的上升,这同时将玩家暴露于被对手利用的风险之中.因此,一些工作着力于在利用对手的同时保障策略的安全性.Johanson 提出受限纳什响应[244],以一定概率混合对手建模策略与随机策略,避免过度偏离均衡.然而每当概率改变时,该算法都需要完全重新构建博弈策略,并不适用于在线场景.Ganzfried 在重复阶段博弈中研究了安全对手利用的相关性质[245].Moravcík[246]和 Liu[247]分别使用了类似方法在结合子博弈精炼技术的同时,实现了安全的对手利用.Slumbers 等人提出了能够最小化智能体交互时产生的潜在风险的风险厌恶均衡解形式[248].另一个研究的重点是自适应调整对手建模的能力.在样本交互受限的在线场景下,如何在对手策略风格快速改变时实现模型的调整是对手建模面临的一大挑战.Wu 通过在预训练中对阵持续生成的多样性策略,实现了在线场景下策略的快速识别和切换[249].一些类似工作通过策略蒸馏的方式,使建立的基准策略具备丰富的表达性,应对不同的对手风格[250].

尽管对手建模的研究历史最早可追溯于博弈论的早期研究[145],但随着博弈环境的复杂化,一些方法的适用性不足,对手建模仍是一个活跃的研究领域.受限于篇幅,推荐读者阅读综述[251-253]获得更详尽的信息.

5.3　对比小结

本节从在线求解方法出发,介绍 IIG 场景下的搜索技术以及对手建模方法.在线求解模块几乎被应用于所有博弈 AI 的进展中[6-8,32,53,254].尽管在线求解的研究由来已久,但非完美信息博弈中可靠、安全的在线求解范式在过去的很长一段时间内却被认为是难以实现的[68,255].直到 Deepstack[53]的提出,安全的在线求解推理技术才第一次被应用于 IIG 领域中.随后提出的 Libratus[54]、Modicum[214]、Supremus[256]等扑克 AI 都通过结合安全搜索技术,在大幅缩减了计算资源和时间的同时保证了推理的正确性,最终击败了世界顶尖的扑克玩家.通过对公共信念状态的使用,Brown 和 Schmid 先后提出了 ReBeL[202]和 GT-CFR[203]算法,建立了在 PIG 和 IIG 中通用的在线求解范式.未来,如何在非完美信息特征更加显著的博弈中避免子博弈结构复杂带来的挑战,并在将搜索思想融入模型学习的过程之中同时实现高效地在线决策和规划,是未来在线求解研究的一大方向[257].

6　总结与展望

本文首先介绍了基于遗憾和基于最佳响应求解非完美信息博弈的两大类离线求解算法的原理、发展脉络和改进技巧,并对两类算法进行了深入联系与对比.随后本文介绍了非完美信息博弈中的在线求解方法,并总结了搜索与对手建模在非完美信息博弈中的应用.结合以上所述,本节针对非完美信息博弈对抗求解框架所存在的不足,总结未来研究方向,并对发展前景作以下展望:

(1)拓展博弈问题类型.在双人零和博弈之外的博弈场景中,理论研究与实际应用间存在一定滞后性.智能体数目的增加或零和条件的放宽都会导致纳什均衡的计算复杂度过高,无法高效精确求解[25-27].事实上,即便是在三人零和博弈中,计算纳什均衡的复杂度也是 PPAD-complete 的[25-27].除此之外,多人博弈场景中的均衡选择问题[196]还会使得每个玩家难以共同达到更高的收益.种种难点导致在多人博弈场景中定义或求解最优策略仍然是一个挑战性难题.尽管如此,近年来不少工作尝试将求解双人零和博弈的算法扩展至非双人零和博弈中[191,258-260].尽管这些学习算法在非双人零和博弈中往往不具有收敛到纳什均衡的保证[49],但在例如多人扑克[32,261]、隐藏角色扮演类游戏 Avalon[21]、战略游戏 Diplomacy[24]等应用中仍然取得了

一定成果. 作为多人博弈场景下的一个特殊情况，团队对抗博弈（Team Adversary Game，TAG）问题也受到了大量关注. 团队对抗博弈建模了一组合作的智能体团队对抗一个共同对手的博弈过程，不仅在理论上取得了许多成果[262-264]，也在实际应用中也发挥了重要作用，如为多名警察抓捕罪犯提供抓捕方案[265]. 此外，利用非完美信息博弈建模智能体数目不固定的开放多智能体系统[266]（Open Multi-agent System）也是值得研究的问题. 未来，为更广泛类型博弈的求解提供理论支撑是构建可解释人工智能的关键.

(2)设计均衡解形式. 随着智能体数目增多，智能体间的关系往往同时存在竞争与合作关系，智能体需要与其他智能体合作以得到更高收益. 尽管在双人零和博弈中纳什均衡具有独特的地位，但在非双人零和博弈中，纳什均衡不一定是一个理想的均衡解选择[196]. 例如在柠檬水摊位博弈（Lemonade Stand Game)，每个玩家需同时在圆环上选定位置，其中与所有其他玩家的最小距离最大的玩家取胜. 该博弈的纳什均衡为玩家沿环均匀分布. 然而，如果玩家们独立计算纳什均衡，其策略集未必会趋于以均匀分布为特征的纳什均衡状态. 这一情况源于纳什均衡的定义隐含了智能体的策略空间不具备相关性，不仅不利于合作的发展，同时也带来了智能体间均衡选择的问题. 为应对类型更加复杂的博弈，均衡解的设计和配套的求解算法极为重要. 近年来，作为纳什均衡的泛化解形式，相关均衡[267]（Correlated Equilibrium，CE）被广泛研究[268-269]. 相较于纳什均衡，相关均衡避免了玩家间的均衡选择问题，在一般和博弈中常具有更高的社会福利[270-271]，并且被证明能在多项式时间内求解. 已有一些工作分别从最优化[268,272]、无遗憾[273-274]和最佳响应[196]的角度提出了在扩展式博弈中求解相关均衡的算法. 除相关均衡外，一些其他均衡解形式也开始受到广泛研究，例如针对于团队对抗博弈[275]的团队极大极小均衡[276-279]（Team Maxmin Equilibrium)，针对于动态马尔可夫博弈（Active Markov Game）的周期均衡[280]（Cyclic Equi libria）、动态均衡[281]（Active Equilibria）等. 如何为不同类型的博弈设计具有表达性的均衡解是迈向通用人工智能的重要一步.

(3)保证最终策略收敛性. 由于平均策略的计算需要大量过去策略的信息，同时在使用深度神经网络进行拟合时存储开销巨大，因此，最终策略的收敛对于结合深度学习和节省存储空间有重要意义[101,124]. 本文所介绍算法的收敛性保证大多针对于平均策略，然而对于最终策略而言，由于环境的非平稳性，智能体的策略往往会陷入循环解集[280]或极限环[195]，甚至陷入混沌态[282]（Chaos)，出现无法收敛的情况. 针对该问题，一系列算法最早于建模生成对抗网络的可微博弈[283]（Differentiable Games）领域中被提出并取得了良好效果，例如在梯度更新中加入二阶信息[284-285]、异步调整双方学习率[286]、惩罚策略梯度[287]等. 基于这些改进思路，近年来涌现了一系列方法来改进传统基于遗憾和基于最佳响应算法仅能保证平均策略收敛的缺陷，例如:在更新时加入某种程度上可视为近似二阶信息的“乐观”（optimistic）信息[288]；通过非自博弈的方式以达到模拟博弈双方异步的学习率[99]；基于动力系统学中的庞加莱回归理论，在原始更新式中加入额外的策略相关正则项[124]. 由于实现了最终策略的收敛，这些算法能够大幅减少计算和存储开销，其中一些算法已在大规模博弈中得到了成功应用[22]. 未来，在理论上设计更多具有最终策略收敛性的算法并加以运用，对于进一步扩展博弈规模和推动多智能体强化学习理论发展都具有重要意义.

(4)探索应用场景. 当前非完美信息博弈已在游戏应用场景中取得了突破性进展，但由于样本采样效率低、算力需求大、缺乏评估指标等问题，将现有成果应用于复杂场景仍极具挑战. 如何对游戏以外的复杂博弈场景进行合理的建模并提出高效的求解方法，成为突破下一阶段博弈智能技术发展的关键[289-291]. 当前，已有部分工作利用非完美信息博弈技术求解现实应用场景中. 例如，在无人集群协同控制问题中，通过将无人机间的协作问题建模为多人合作博弈[289]，以实现在没有中央控制器的协调下分布式地执行合作任务；在智能物联汽车优化问题中，车辆之间的交互可被建模为多人混合博弈的问题，超车、让行等决策中体现了博弈中竞争与合作的思想[290]；在大规模电力调控问题中，控制系统需要通过调整发电机与实际运行时产生的负荷扰动进行对抗而形成竞争博弈，同时在大规模分布式资源接入场景下源、荷、储多主体间需协同调度实现整个系统效能的最优化[291]. 现实世界中的应用场景往往具有隐藏信息，在非完美信息博弈的建模框架下能够被更加精确的刻画. 除此之外，博弈思想也在人工智能的各个领域发挥了关键作用，为各类学习任务提供强大的工具. 例如在生成对抗网

络[292](Generative Adversarial Networks, GAN)中,生成器和判别器之间的竞争博弈展现了博弈论的核心思想;在面临不确定性环境时,模型训练可以借鉴非完美信息博弈理论中的信息和混合策略的概念,更有效地适应复杂而动态的现实场景,从而增强模型的鲁棒性以应对对抗性攻击[293];在大语言模型的交互建模中加入例如信号博弈[294]、多智能体辩论[295]等博弈论机制,能够有效地提升大语言模型的任务表现. 未来,在如网络群体行为[11]、指挥控制系统[13]、安保布防[12]、互联网经济[296]等大规模场景下引入博弈思想并指导智能体决策,是非完美信息博弈领域研究迈向落地应用和技术验证的重要探索.

参 考 文 献

[1] Papoudakis G, Christianos F, Rahman A, et al. Dealing with nonstationarity in multi-agent deep reinforcement learning. arXiv preprint arXiv:1906.04737, 2019

[2] Osborne M J, Rubinstein A. A Course in Game Theory. Cambridge, USA: MIT press, 1994

[3] YUAN Wei-lin, LUO Jun-ren, LU Li-na, et al. Methods in adversarial intelligent game: A holistic comparative analysis from perspective of game theory and reinforcement learning. Computer Science, 2022, 49(8): 191-204. (in Chinese)
(袁唯淋, 罗俊仁, 陆丽娜, 等. 智能博弈对抗方法: 博弈论与强化学习综合视角对比分析. 计算机科学, 2022, 49(8):191-204)

[4] Anthony T, Tian Z, Barber D. Thinking fast and slow with deep learning and tree search//Proceedings of the 31st Neural Information Processing Systems(NIPS). Long Beach, USA, 2017: 5360-5370

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

[6] Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge. Nature, 2017, 550 (7676):354-359

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

[8] Campbell M, Hoane Jr A J, Hsu F h. Deep blue. Artificial Intelligence, 2002, 134(1-2):57-83

[9] Mnih V, Kavukcuoglu K, Silver D, et al. Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602, 2013

[10] Fudenberg D, Peysakhovich A. Recency, records, and recaps: Learning and nonequilibrium behavior in a simple decision problem. ACM Transactions on Economics and Computation, 2016, 4(4):23:1-23:18

[11] WANG Yuan-Zhuo, YU Jian-Ye, QIU wen, et al. Evolutionary game model and analysis methods for network group behavior. Chinese Journal of Computers, 2015, 38(2): 282-300. (in Chinese)
(王元卓, 于建业, 邱雯, 等. 网络群体行为的演化博弈模型与分析 方法. 计算机学报, 2015, 38(2): 282-300)

[12] DUAN zhe, GUO Ju-e, FENG Geng-zhong, et al. Analysis of security policy in transit hub via incomplete information games. Journal of Systems Science and Mathematical Sciences, 2019,39(10):1632-1641(in Chinese)
(段喆, 郭菊娥, 冯耕中, 等. 基于不完全信息博弈的交通枢纽安保 布防策略研究. 系统科学与数学, 2019, 39(10): 1632-1641)

[13] LI Xian-Gang, LI qiang. Technical analysis of typical intelligent game system and development prospect of intelligent command and control system. Chinese Journal of Intelligent Science and Technology, 2020, 2(1): 36-42. (in Chinese)
(李宪港, 李强. 典型智能博弈系统技术分析及指控系统智能化发 展展望. 智能科学与技术学报, 2020, 2(1): 36-42)

[14] Shoham Y, Leyton-Brown K. Multiagent systems: Algorithmic, gametheoretic, and logical foundations. Cambridge, UK: Cambridge university press, 2008

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

[16] Bard N, Foerster J N, Chandar S, et al. The Hanabi challenge: A new frontier for ai research. Artificial Intelligence, 2020, 280:103216

[17] Zha D, Xie J, Ma W, et al. Douzero: Mastering DouDizhu with self-play deep reinforcement learning//Proceedings of the 38th International Conference on Machine Learning(ICML). Virtual Event, 2021: 12333-12344

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

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

[20] Ye D, Chen G, Zhang W, et al. Towards playing full MOBA games with deep reinforcement learning//Proceedings of the 34th Neural Information Processing Systems(NIPS). Vancouver, Canada, 2020: 621-632

[21] Serrino J, Kleiman-Weiner M, Parkes D C, et al. Finding friend and foe in multi-agent games//Proceedings of the 33rd Neural Information Processing Systems(NIPS). Vancouver, Canada, 2019: 1249-1259

[22] Perolat J, De Vylder B, Hennes D, et al. Mastering the game of stratego with model-free multiagent reinforcement learning. Science, 2022, 378(6623):990-996

[23] Zhou Lei, Yin Qi-Yue, Huang Kai-Qi. Game-theoretic learn-

ing in human-computer gaming. Chinese Journal of Computers, 2022, 45(9): 1860-1876. (in Chinese)
(周雷, 尹奇跃, 黄凯奇. 人机对抗中的博弈学习方法. 计算机学报, 2022, 45(9): 1860-1876)

[24] Bakhtin A, Wu D J, Lerer A, et al. Mastering the game of no-press diplomacy via human-regularized reinforcement learning and planning//Proceedings of the 11st International Conference on Learning Representations (ICLR). Kigali, Rwanda, 2023: 1012-1020

[25] Chen X, Deng X. 3-nash is ppad-complete. Electronic Colloquium on Computational Complexity (ECCC), 2005, 134: 2-29

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

[27] Rubinstein A. Inapproximability of nash equilibrium. SIAM Journal on Computing, 2018, 47(3): 917-959

[28] Daskalakis C, Golowich N, Zhang K. The complexity of markov equilibrium in stochastic games. arXiv preprint arXiv: 2204.03991, 2022

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

[30] Bakhtin A, Wu D, Lerer A, et al. No-press diplomacy from scratch//Proceedings of the 35th Neural Information Processing Systems(NIPS). Virtual Event, 2021: 18063-18074

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

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

[33] Leibo J Z, Duéñez-Guzmán E A, Vezhnevets A, et al. Scalable evaluation of multi-agent reinforcement learning with melting pot//Meila M, Zhang T. Proceedings of the 38th International Conference on Machine Learning(ICML). Virtual Event, 2021, 139: 6187-6199

[34] Timbers F, Bard N, Lockhart E, et al. Approximate exploitability: learning a best response//Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2022: 3487-3493

[35] Johanson M, Waugh K, Bowling M H, et al. Accelerating best response calculation in large extensive games//Proceedings of the 22nd International Joint Conference on Artificial Intelligence(IJCAI). 2011: 258-265

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

[37] Zhang K, Yang Z, Basar T. Multi-agent reinforcement learning: A selective overview of theories and algorithms. Hand-Book of Reinforcement Learning and Control, 2021: 321-384

[38] Cai Y, Luo H, Wei C Y, et al. Uncoupled and convergent learning in two-player zero-sum markov games. arXiv preprint arXiv: 2303.02738, 2023

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

[40] Waugh K. Abstraction in large extensive games[Master's Thesis]. University of Alberta, Edmonton, Canada, 2009

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

[42] Kuhn H, Tucker A. Extensive Games and the Problem of Information. Princeton, USA: Princeton University Press, 1953

[43] Heinrich J, Lanctot M, Silver D. Fictitious self-play in extensive-form games//Proceedings of the 32nd International Conference on Machine Learning (ICML). Lille, France, 2015: 805-813

[44] Nash J. Non-cooperative games [Ph. D. Thesis]. Princeton University, Princeton, USA, 1951

[45] Foster D P, Vohra R V. A randomization rule for selecting forecasts. Operations Research, 1993, 41(4): 704-709

[46] Hart S, Mas-Colell A. A general class of adaptive strategies. Journal of Economic Theory, 2001, 98(1): 26-54

[47] Hart S, Mas-Colell A. A simple adaptive procedure leading to correlated equilibrium. Econometrica, 2000, 68(5): 1127-1150

[48] Cesa-Bianchi N, Lugosi G. Prediction, Learning, and Games. Cambridge, UK: Cambridge university press, 2006

[49] Cai Y, Candogan O, Daskalakis C, et al. Zero-sum polymatrix games: A generalization of minmax. Mathematics of Operations Research, 2016, 41(2): 648-655

[50] Foster D P, Vohra R V. Calibrated learning and correlated equilibrium. Games and Economic Behavior, 1997, 21 (1-2): 40

[51] Blum A, Mansour Y. From external to internal regret. Journal of Machine Learning Research, 2007, 8(6): 1307-1324

[52] Greenwald A, Li Z, Marks C. Bounds for regret-matching algorithms//Proceedings of the 9th International Symposium on Artificial Intelligence and Mathematics (ISAIM). Fort Lauderdale, USA, 2006

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

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

[55] Zhao E, Yan R, Li J, et al. Alphaholdem: High-performance artificial intelligence for heads-up no-limit poker via end-to-end reinforcement learning//Proceedings of the 36th AAAI Conference on Artificial Intelligence(AAAI). Virtual Event, 2022: 4689-4697

[56] Tammelin O, Burch N, Johanson M, et al. Solving heads-up limit texas hold'em//Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI). Buenos Aires, Argentina, 2015: 645-652

[57] Huale L, Wang X, Jia F, et al. A survey of nash equilibrium

strategy solving based on cfr. Archives of Computational Methods in Engineering, 2020, 28:2749-2760

[58] Tammelin O. Solving large imperfect information games using cfr+. arXiv preprint arXiv:1407.5042, 2014

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

[60] Farina G, Grand-Clément J, Kroer C, et al. Regret matching+:(in) stability and fast convergence in games. arXiv preprint arXiv:2305.14709, 2023

[61] Farina G, Kroer C, Sandholm T. Faster game solving via predictive blackwell approachability: Connecting regret matching and mirror descent//Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI). Virtual Event, 2021: 5363-5371

[62] Cai Y, Farina G, Grand-Clément J, et al. Last-iterate convergence properties of regret-matching algorithms in games. arXiv preprint arXiv:2311.00676, 2023

[63] Zinkevich M, Johanson M, Bowling M, et al. Regret minimization in games with incomplete information//Proceedings of the 21st Neural Information Processing Systems (NIPS). Vancouver, Canada, 2007: 1729-1736

[64] Lanctot M, Waugh K, Zinkevich M, et al. Monte carlo sampling for regret minimization in extensive games//Proceedings of the 23rd Neural Information Processing Systems (NIPS). Vancouver, Canada, 2009: 1078-1086

[65] Schmid M, Burch N, Lanctot M, et al. Variance reduction in monte carlo counterfactual regret minimization (vr-mccfr) for extensive form games using baselines//Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI). Honolulu, USA, 2019: 2157-2164

[66] Gibson R G, Burch N, Lanctot M, et al. Efficient monte carlo counterfactual regret minimization in games with many player actions. //Proceedings of the 26th Neural Information Processing Systems(NIPS). Lake Tahoe, USA, 2012: 1889-1897

[67] Johanson M, Bard N, Lanctot M, et al. Efficient nash equilibrium approximation through monte carlo counterfactual regret minimization. //Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS). Valencia, Spain, 2012: 837-846

[68] Lisý V, Lanctot M, Bowling M H. Online monte carlo counterfactual regret minimization for search in imperfect information games. //Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems(AAMAS). Istanbul, Turkey, 2015: 27-36

[69] Jackson E G. Targeted CFR//Proceedings of the AAAI Workshop on Computer Poker and Imperfect Information. San Francisco, USA, 2017

[70] Li H, Hu K, Zhang S, et al. Double neural counterfactual regret minimization//Proceedings of the 8th International Conference on Learning Representation(ICLR). Addis Ababa, Ethiopia, 2020

[71] Schmid M. Search in imperfect information games[Ph. D. Thesis]. Charles University, Prague, Czech Republic, 2021

[72] Zhang H, Lerer A, Brown N. Equilibrium finding in normal-form games via greedy regret minimization//Proceedings of 36th AAAI Conference on Artificial Intelligence (AAAI). Virtual Event, 2022: 9484-9492

[73] Xu H, Li K, Fu H, et al. Autocfr: Learning to design counterfactual regret minimization algorithms//Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI). Virtual Event, 2022: 5244-5251

[74] Koller D, Pfeffer A. Representations and solutions for game-theoretic problems. Artificial intelligence, 1997, 94(1-2): 167-215

[75] Zhang B, Sandholm T. Sparsified linear programming for zero-sum equilibrium finding//Proceedings of the 35th International Conference on Machine Learning(ICML). Virtual Event, 2020: 11256-11267

[76] Shi J, Littman M L. Abstraction methods for game theoretic poker//Proceedings of the International Conference on Computers and Games (ICCG). Hamamatsu, Japan, 2000: 333-345

[77] Gilpin A, Sandholm T. Lossless abstraction of imperfect information games. Journal of the ACM (JACM), 2007, 54(5):25-26

[78] Billings D, Burch N, Davidson A, et al. Approximating game-theoretic optimal strategies for full-scale poker//Proceedings of the 18th International Joint Conference on Artificial Intelligence(IJCAI). Acapulco, Mexico, 2003: 661-668

[79] Gilpin A. Algorithms for abstracting and solving imperfect information games[Ph. D. Thesis]. Carnegie Mellon University, Pittsburgh, USA, 2009

[80] Waugh K, Morrill D, Bagnell J A, et al. Solving games with functional regret estimation//Proceedings of the 29th AAAI Conference on Artificial Intelligence(AAAI). Austin, USA, AAAI Press, 2015: 2138-2145

[81] 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//Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems(AAMAS). Auckland, New Zealand, 2020: 339-347

[82] Brown N, Lerer A, Gross S, et al. Deep counterfactual regret minimization//Proceedings of the 36th International Conference on Machine Learning (ICML). Long Beach, USA, 2019: 793-802

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

[84] Steinberger E, Lerer A, Brown N. Dream: Deep regret minimization with advantage baselines and model-free learning. arXiv preprint arXiv:2006.10410, 2020

[85] Troillet L, Matsuzaki K. Analyzing simplified geister using

dream//Proceedings of the IEEE Conference on Games (CoG). Copenhagen, Denmark, 2021: 1-8

[86] Gallucci J, Bowser R, Kettell S, et al. Estimating card fitness for discard in gin rummy//Proceedings of the 35th AAAI Conference on Artificial Intelligence(AAAI). Virtual Event, 2021: 15503-15509

[87] McAleer S, Farina G, Lanctot M, et al. Escher: Eschewing importance sampling in games by computing a history value function to estimate regret. arXiv preprint arXiv:2206.04122, 2022

[88] Liu W, Li B, Togelius J. Model-free neural counterfactual regret minimization with bootstrap learning. IEEE Transactions on Games, 2023, 15(3): 315-325

[89] Zinkevich M. Online convex programming and generalized infinitesimal gradient ascent//Proceedings of the 20th International Conference on Machine Learning(ICML). Washington, USA, 2003: 928-936

[90] Czarnecki W M, Gidel G, Tracey B, et al. Real world games look like spinning tops//Proceedings of the 34th Neural Information Processing Systems(NIPS). Vancouver, Canada, 2020: 17443-17454

[91] Letcher A. Stability and exploitation in differentiable games [Master's Thesis]. University of Oxford, Oxford, England, 2018

[92] Jin P H, Keutzer K, Levine S. Regret minimization for partially observable deep reinforcement learning//Proceedings of the 35th International Conference on Machine Learning(ICML). Stockholm, Sweden, 2018: 2347-2356

[93] Kash I A, Sullins M, Hofmann K. Combining no-regret and qlearning//Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems(AAMAS). Auckland, New Zealand, 2020: 593-601

[94] Abbasi-Yadkori Y, Bartlett P, Bhatia K, et al. Politex: Regret bounds for policy iteration using expert prediction//Proceedings of the 36th International Conference on Machine Learning(ICML). Long Beach, USA, 2019: 3692-3702

[95] Srinivasan S, Lanctot M, Zambaldi V F, et al. Actor-critic policy optimization in partially observable multiagent environments/ /Proceedings of the 32nd Neural Information Processing Systems(NIPS). Montréal, Canada, 2018: 3426-3439

[96] Hofbauer J, Sigmund K. Evolutionary Games and Population Dynamics. Cambridge,UK: Cambridge university press, 1998

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

[98] Chaudhuri K, Freund Y, Hsu D J. A parameter-free hedging algorithm//Proceedings of the 23rd Neural Information Processing Systems(NIPS). Vancouver, Canada, 2009: 297-305

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

[100] Johanson M, Bard N, Burch N, et al. Finding optimal abstract strategies in extensive-form games//Proceedings of the 26th AAAI Conference on Artificial Intelligence(AAAI). Toronto, Canada, 2012: 1371-1379

[101] Gruslys A, Lanctot M, Munos R, et al. The advantage regret-matching actor-critic. arXiv preprint arXiv:2008.12234, 2020

[102] Fu H, Liu W, Wu S, et al. Actor-critic policy optimization in a largescale imperfect-information game//Proceedings of the 10th International Conference on Learning Representation(ICLR). Virtual Event, 2022

[103] Hu J, Wellman M P. Nash q-learning for general-sum stochastic games. J. Mach. Learn. Res., 2003, 4:1039-1069

[104] Littman M L. Friend-or-foe q-learning in general-sum games//Proceedings of the 18th International Conference on Machine Learning(ICML). Williamstown, USA, 2001: 322-328

[105] Cen S, Wei Y, Chi Y. Fast policy extragradient methods for competitive games with entropy regularization//Ranzato M, Beygelzimer A, Dauphin Y N, et al. Proceedings of the 35th Neural Information Processing Systems(NIPS). virtual event, 2021: 27952-27964

[106] Cen S, Chi Y, Du S S, et al. Faster last-iterate convergence of policy optimization in zero-sum markov games//Proceedings of The 11th International Conference on Learning Representations(ICLR). Kigali, Rwanda, 2023

[107] Yang Y, Ma C. $O(T^{-1})$ convergence of optimistic-follow-the-regularized-leader in two-player zero-sum markov games//Proceedings of The 11th International Conference on Learning Representations(ICLR). Kigali, Rwanda, 2023: 704-712

[108] Bai Y, Jin C, Yu T. Near-optimal reinforcement learning with self-play//Proceedings of the 34th Neural Information Processing Systems(NIPS). Vancouver, Canada, 2020: 2159-2170

[109] Cai Y, Oikonomou A, Zheng W. Finite-time last-iterate convergence for learning in multi-player games//Proceedings of the 36th Neural Information Processing Systems(NIPS). New Orleans, USA, 2022: 33904-33919

[110] Sayin M O, Zhang K, Leslie D S, et al. Decentralized q-learning in zero-sum markov games//Proceedings of the 35th Neural Information Processing Systems(NIPS). Virtual Event, 2021: 18320-18334

[111] Abernethy J D, Hazan E, Rakhlin A. Competing in the dark: An efficient algorithm for bandit linear optimization//Proceedings of the 21st Annual Conference on Learning Theory(COLT). Helsinki, Finland, 2008: 263-274

[112] Vovk V. Competitive on-line statistics. International Statistical Review, 2001, 69(2):213-248

[113] Beck A, Teboulle M. Mirror descent and nonlinear projected subgradient methods for convex optimization. Operations Research Letters, 2003, 31(3):167-175

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

[115] Kroer C, Waugh K, Kılınç-Karzan F, et al. Faster first-order methods for extensive-form game solving//Proceedings of the 16th ACM Conference on Economics and Computation (EC). Portland, USA, 2015: 817-834

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

[117] Nesterov Y. Primal-dual subgradient methods for convex problems. Mathematical Programming, 2009, 120(1): 221-259

[118] Burch N. Time and space: Why imperfect information games are hard[Ph. D. Thesis]. University of Alberta, Edmonton, Canada, 2018

[119] Liu W, Jiang H, Li B, et al. Equivalence analysis between counterfactual regret minimization and online mirror descent//Proceedings of the 39th International Conference on Machine Learning(ICML). Baltimore, USA, 2022: 13717-13745

[120] Farina G, Kroer C, Sandholm T. Regret circuits: Composability of regret minimizers//Proceedings of the 36th International Conference on Machine Learning(ICML). Long Beach, USA, 2019: 1863-1872

[121] Anagnostides I, Farina G, Sandholm T. Near-optimal ϕ-regret learning in extensive-form games//Proceedings of 40th International Conference on Machine Learning(ICML). Honolulu, USA, 2023: 814-839

[122] Syrgkanis V, Agarwal A, Luo H, et al. Fast convergence of regularized learning in games//Proceedings of the 29th Neural Information Processing Systems(NIPS). Montreal, Canada, 2015: 2989-2997

[123] Fiegel C, Ménard P, Kozuno T, et al. Local and adaptive mirror descents in extensive-form games. arXiv preprint arXiv:2309.00656, 2023

[124] Grand-Clément J, Kroer C. Conic blackwell algorithm: Parameterfree convex-concave saddle-point solving//Proceedings of the 34th Neural Information Processing Systems (NIPS). Virtual Event, 2021: 9587-9599

[125] Wei C, Lee C, Zhang M, et al. Linear last-iterate convergence in constrained saddle-point optimization//Proceedings of the 9th International Conference on Learning Representations(ICLR). Austria, 2021: 1212-1220

[126] Rakhlin A, Sridharan K. Optimization, learning, and games with predictable sequences//Proceedings of the 27th Neural Information Processing Systems (NIPS). Lake Tahoe, USA, 2013: 3066-3074

[127] Mokhtari A, Ozdaglar A E, Pattathil S. A unified analysis of extragradient and optimistic gradient methods for saddle point problems: Proximal point approach//Proceedings of Machine Learning Research: The 23rd International Conference on Artificial Intelligence and Statistics (AISTATS). Palermo, Italy, 2020: 1497-1507

[128] Daskalakis C, Panageas I. The limit points of (optimistic) gradient descent in min-max optimization//Proceedings of the 32nd Neural Information Processing Systems(NIPS). Montréal, Canada, 2018: 9256-9266

[129] Lee C, Kroer C, Luo H. Last-iterate convergence in extensive-form games//Proceedings of the 35th Neural Information Processing Systems (NIPS). Virtual Event, 2021: 14293-14305

[130] Jiang R, Mokhtari A. Generalized optimistic methods for convexconcave saddle point problems. arXiv preprint arXiv: 2202.09674,2022

[131] Abe K, Ariu K, Sakamoto M, et al. Last-iterate convergence with fulland noisy-information feedback in two-player zero-sum games. arXiv preprint arXiv:2208.09855, 2022

[132] Perolat J, Munos R, Lespiau J B, et al. From poincaré recurrence to convergence in imperfect information games: Finding equilibrium via regularization//Proceedings of the 38th International Conference on Machine Learning (ICML). Virtual Event, 2021: 8525-8535

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

[134] Bai Y, Jin C, Mei S, et al. Near-optimal learning of extensive-form games with imperfect information//Proceedings of the 39th International Conference on Machine Learning(ICML) Baltimore, USA, 2022: 1337-1382

[135] Farina G, Kroer C, Sandholm T. Stochastic regret minimization in extensive-form games//Proceedings of the 37th International Conference on Machine Learning(ICML). Virtual Event, 2020: 3018-3028

[136] Zhang B H, Sandholm T. Finding and certifying (near-)optimal strategies in black-box extensive-form games//Proceedings of 35th AAAI Conference on Artificial Intelligence (AAAI). Virtual Event, 2021: 5779-5788

[137] Zhou Y, Li J, Zhu J. Posterior sampling for multi-agent reinforcement learning: Solving extensive games with imperfect information//Proceedings of the 8th International Conference on Learning Representation(ICLR). Addis Ababa, Ethiopia, 2020: 6578-6586

[138] Kozuno T, Ménard P, Munos R, et al. Model-free learning for twoplayer zero-sum partially observable markov games with perfect recall. arXiv preprint arXiv:2106.06279,2021

[139] Farina G, Schmucker R, Sandholm T. Bandit linear optimization for sequential decision making and extensive-form

games//Proceedings of the 35th AAAI Conference on Artificial Intelligence(AAAI). Virtual Event, 2021: 5372-5380

[140] Tomar M, Shani L, Efroni Y, et al. Mirror descent policy optimization//Proceedings of the 10th International Conference on Learning Representation(ICLR). Virtual Event, 2022

[141] Grudzien J, De Witt C A S, Foerster J. Mirror learning: A unifying framework of policy optimisation//Proceedings of the 39th International Conference on Machine Learning(ICML). Baltimore, USA, 2022: 7825-7844

[142] Sokota S, D'Orazio R, Kolter J Z, et al. A unified approach to reinforcement learning, quantal response equilibria, and two-player zero-sum games. arXiv preprint arXiv: 2206.05825, 2022

[143] Agarap A F. Deep learning using rectified linear units (relu). arXiv preprint arXiv:1803.08375, 2018

[144] Srinivasan S, Lanctot M, Zambaldi V, et al. Actor-critic policy optimization in partially observable multiagent environments. arXiv preprint arXiv:1810.09026, 2018

[145] Brown G W. Iterative solution of games by fictitious play. Activity analysis of production and allocation, 1951, 13(1): 374-376

[146] Robinson J. An iterative method of solving a game. Annals of mathematics, 1951:296-301

[147] Berger U. Brown's original fictitious play. Journal of Economic Theory, 2007, 135(1):572-578

[148] Monderer D, Shapley L S. Fictitious play property for games with identical interests. Journal of Economic Theory, 1996, 68(1):258-265

[149] Shapley L. Some topics in two-person games. Advances in Game Theory, 1964, 52:1-29

[150] Fudenberg D, Levine D K. Consistency and cautious fictitious play. Journal of Economic Dynamics and Control, 1995, 19(5-7):1065-1089

[151] Viossat Y, Zapechelnyuk A. No-regret dynamics and fictitious play. Journal of Economic Theory, 2013, 148(2): 825-842

[152] Benaïm M, Faure M. Consistency of vanishingly smooth fictitious play. Mathematics of Operations Research, 2013, 38(3):437-450

[153] Li Z, Tewari A. Sampled fictitious play is hannan consistent. Games and Economic Behavior, 2018, 109:401-412

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

[155] Kaniovski Y M, Young H P. Learning dynamics in games with stochastic perturbations. Games and Economic Behavior, 1995, 11(2):330-363

[156] van der Genugten B. A weakened form of fictitious play in two-person zero-sum games. International Game Theory Review, 2000, 2(4): 307-328

[157] Haarnoja T, Zhou A, Abbeel P, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor//Proceedings of the 35th International Conference on Machine Learning(ICML). Stockholm, Sweden, 2018: 1856-1865

[158] Meng L, Ge Z, Tian P, et al. An efficient deep reinforcement learning algorithm for solving imperfect information extensive-form games//Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI). Washington, USA, 2023: 5823-5831

[159] Lambert Iii T J, Epelman M A, Smith R L. A fictitious play approach to large-scale optimization. Operations Research, 2005, 53(3):477-489

[160] McMahan H B, Gordon G J. A fast bundle-based anytime algorithm for poker and other convex games//Proceedings of the 10th International Conference on Artificial Intelligence and Statistics(AISTATS). San Juan, Puerto Rico, 2007: 323-330

[161] Ganzfried S, Sandholm T. Computing equilibria in multiplayer stochastic games of imperfect information//Proceedings of the 21st International Joint Conference on Artificial Intelligence(IJCAI). Pasadena, USA, 2009: 140-146

[162] Kuhn H. Extensive games and the problem of information, contributions to the theory of games ii. Annals of Mathematics Studies, 1953, 28:193-216

[163] Hendon E, Jacobsen H J, Sloth B. Fictitious play in extensive form games. Games and Economic Behavior, 1996, 15(2):177-202

[164] Kreps D M, Wilson R. Sequential equilibria. Econometrica: Journal of the Econometric Society, 1982:863-894

[165] Greenwald A, Li J, Sodomka E, et al. Solving for best responses in extensive-form games using reinforcement learning methods//Proceedings of the 1st Multidisciplinary Conference on Reinforcement Learning and Decision Making (RLDM). Princeton, USA, 2013: 116

[166] Silver D, Veness J. Monte-carlo planning in large pomdps//Proceedings of the 24th Neural Information Processing Systems(NIPS). Vancouver, Canada, 2010: 2164-2172

[167] Heinrich J. Reinforcement learning from self-play in imperfect information games [Ph. D. thesis]. University College London, London, UK, 2017

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

[169] Vitter J S. Random sampling with a reservoir. ACM Transactions on Mathematical Software (TOMS), 1985, 11(1): 37-57

[170] Shamma J S, Arslan G. Dynamic fictitious play, dynamic gradient play, and distributed convergence to nash equilibria. IEEE Transactions on Automatic Control, 2005, 50(3):312-327

[171] Weibull J W. Evolutionary game theory. Cambridge,USA: MIT press, 1997

[172] Ma X, Driggs-Campbell K, Kochenderfer M J. Improved robustness and safety for autonomous vehicle control with

adversarial reinforcement learning//The proceedings of the 29th IEEE Intelligent Vehicles Symposium (IV). Changshu, China, 2018: 1665-1671

[173] Kawamura K, Tsuruoka Y. Neural fictitious self-play on ELF mini-rts. arXiv preprint arXiv:1902.02004, 2019

[174] Xue W, Zhang Y, Li S, et al. Solving large-scale extensive-form network security games via neural fictitious self-play//Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI). Virtual Event, 2021: 3713-3720

[175] He K, Wu H, Wang Z, et al. Finding nash equilibrium for imperfect information games via fictitious play based on local regret minimization. International Journal of Intelligent Systems, 2022, 37(9):6152-6167

[176] Zhang L, Chen Y, Wang W, et al. A monte carlo neural fictitious self-play approach to approximate nash equilibrium in imperfectinformation dynamic games. Frontiers Computer Science, 2021, 15(5): 155334

[177] Li H, Qi S, Zhang J, et al. NFSP-PER: An efficient sampling nfsp-based method with prioritized experience replay//Proceedings of the 4th International Conference on Data Intelligence and Security (ICDIS). Shenzhen, China, 2022: 388-393

[178] McMahan H B, Gordon G J, Blum A. Planning in the presence of cost functions controlled by an adversary//Proceedings of the 20th International Conference on Machine Learning (ICML). Washington, USA, 2003: 536-543

[179] McAleer S, Lanier J B, Wang K A, et al. XDO: A double oracle algorithm for extensive-form games//Proceedings of the 34th Neural Information Processing Systems (NIPS). Vancouver, Canada, 2021:23128-23139

[180] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017

[181] van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double q-learning//Proceedings of the 30th AAAI Conference on Artificial Intelligence(AAAI). Phoenix, USA, 2016: 2094-2100

[182] Dinh L C, McAleer S M, Tian Z, et al. Online double oracle. Transactions on Machine Learning Research, 2022,20(1):769-790

[183] Tang X, Dinh L C, McAleer S M, et al. Regret-minimizing double oracle for extensive-form games. CoRR, 2023, abs/2304.10498

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

[185] Ye D, Liu Z, Sun M, et al. Mastering complex control in MOBA games with deep reinforcement learning//Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI). New York, USA, 2020: 6672-6679

[186] Walsh W E, Das R, Tesauro G, et al. Analyzing complex strategic interactions in multi-agent systems//AAAI-02 Workshop on GameTheoretic and Decision-Theoretic Agents. 2002: 109-118

[187] Wellman M P. Methods for empirical game-theoretic analysis//Proceedings of the 21st AAAI Conference on Artificial Intelligence(AAAI). Boston, USA, 2006: 1552-1556

[188] Lanctot M, Zambaldi V F, Gruslys A, et al. A unified game-theoretic approach to multiagent reinforcement learning//Proceedings of the 31st Neural Information Processing Systems(NIPS). Long Beach, USA, 2017: 4190-4203

[189] Fudenberg D, Drew F, Levine D K, et al. The Theory of Learning in Games. Cambridge,USA: MIT press, 1998

[190] Balduzzi D, Garnelo M, Bachrach Y, et al. Open-ended learning in symmetric zero-sum games//Proceedings of the 36th International Conference on Machine Learning (ICML). Long Beach, USA, 2019: 434-443

[191] Muller P, Omidshafiei S, Rowland M, et al. A generalized training approach for multiagent learning//Proceedings of the 8th International Conference on Learning Representation (ICLR). Addis Ababa, Ethiopia, 2020: 3759-3768

[192] Nieves N P, Yang Y, Slumbers O, et al. Modelling behavioural diversity for learning in open-ended games//Proceedings of the 38th International Conference on Machine Learning(ICML). Virtual Event, 2021: 8514-8524

[193] Liu Z, Yu C, Yang Y, et al. A unified diversity measure for multiagent reinforcement learning//Proceedings of the 36th Neural Information Processing Systems(NIPS). New Orleans, USA, 2022: 10339-10352

[194] Deng X, Li N, Mguni D, et al. On the complexity of computing markov perfect equilibrium in general-sum stochastic games. National Science Review, 2023, 10(1)

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

[196] Marris L, Muller P, Lanctot M, et al. Multi-agent training beyond zerosum with correlated equilibrium meta-solvers//Proceedings of the 38th International Conference on Machine Learning (ICML). Virtual Event, 2021: 7480-7491

[197] Balduzzi D, Tuyls K, Pérolat J, et al. Re-evaluating evaluation//Proceedings of the 32nd Neural Information Processing Systems(NIPS). Montréal, Canada, 2018: 3272-3283

[198] McAleer S, Lanier J B, Fox R, et al. Pipeline psro: A scalable approach for finding approximate nash equilibria in large games//Proceedings of the 34th Neural Information Processing Systems (NIPS). Vancouver, Canada, 2020: 20238-20248

[199] McAleer S, Wang K, Lanctot M, et al. Anytime optimal psro for twoplayer zero-sum games. arXiv preprint arXiv: 2201.07700, 2022

[200] Zhou M, Chen J, Wen Y, et al. Efficient policy space response oracles. arXiv preprint arXiv:2202.00633, 2022

[201] Liu S, Marris L, Hennes D, et al. NeuPl: Neural popula-

tion learning//The 10th International Conference on Learning Representations(ICLR). Virtual Event, 2022

[202] Brown N, Bakhtin A, Lerer A, et al. Combining deep reinforcement learning and search for imperfect-information games//Larochelle H, Ranzato M, Hadsell R, et al. Proceedings of the 34th Neural Information Processing Systems (NIPS). 2020: 17057-17069

[203] Schmid M, Moravcik M, Burch N, et al. Player of games. arXiv preprint arXiv:2112.03178,2021

[204] v Neumann J. Zur theorie der gesellschaftsspiele. Mathematische Annalen, 1928, 100(1):295-320

[205] Šustr M, Schmid M, Moravčík M, et al. Sound algorithms in imperfect information games. arXiv preprint arXiv: 2006.08740, 2020

[206] Šustr M, Schmid M, Moravčík M, et al. Sound algorithms in imperfect information games//Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems(AAMAS). Virtual Event, 2021: 1674-1676

[207] Ganzfried S, Sandholm T. Endgame solving in large imperfect information games//Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems(AAMAS). Istanbul, Turkey, 2015: 37-45

[208] Gilpin A, Sandholm T. A texas hold'em poker player based on automated abstraction and real-time equilibrium computation//Proceedings of the 5th International Conference on Autonomous Agents and Multiagent Systems (AAMAS). Hakodate, Japan, 2006: 1453-1454

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

[210] Burch N, Johanson M, Bowling M. Solving imperfect information games using decomposition//Proceedings of the 28th AAAI Conference on Artificial Intelligence(AAAI). Québec City, Canada, 2014: 602-608

[211] Brown N. Equilibrium finding for large adversarial imperfect information games[Ph. D. Thesis]. Carnegie Mellon University, Pittsburgh, USA, 2020

[212] Brown N, Sandholm T. Safe and nested subgame solving for imperfectinformation games//Proceedings of the 31st Neural Information Processing Systems (NIPS). Long Beach, USA, 2017: 689-699

[213] Sustr M, Kovarík V, Lisý V. Monte carlo continual resolving for online strategy computation in imperfect information games//Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems(AAMAS). Montreal, Canada, 2019: 224-232

[214] Brown N, Sandholm T, Amos B. Depth-limited solving for imperfectinformation games//Proceedings of the 32nd Neural Information Processing Systems(NIPS). Montréal, Canada, 2018: 7674-7685

[215] Cowling P I, Ward C D, Powley E J. Ensemble determinization in monte carlo tree search for the imperfect information card game magic: The gathering. IEEE Transactions on Computational Intelligence and AI in Games, 2012, 4(4): 241-257

[216] Cowling P I, Powley E J, Whitehouse D. Information set monte carlo tree search. IEEE Transactions on Computational Intelligence and AI in Games, 2012, 4(2):120-143

[217] Zha D, Xie J, Ma W, et al. Douzero: Mastering doudizhu with self-play deep reinforcement learning//Proceedings of the 38th International Conference on Machine Learning(ICML). Virtual Event, 2021: 12333-12344

[218] Cazenave T, Ventos V. The μ search algorithm for the game of bridge. arXiv preprint arXiv:1911.07960,2019

[219] Santos A F, Santos P A, Melo F S. Monte carlo tree search experiments in hearthstone//Proceedings of the 2017 IEEE Conference on Computational Intelligence and Games (CIG). New York, USA, 2017:272-279

[220] Jacobsen E J, Greve R, Togelius J. Monte mario: Platforming with MCTS//Proceedings of Genetic and Evolutionary Computation Conference (GECCO). Vancouver, Canada, 2014: 293-300

[221] Wang J, Zhu T, Li H, et al. Belief-state monte carlo tree search for phantom go. IEEE Transactions on Games, 2018, 10(2):139-154

[222] Zhang Y, Yan D, Shi B, et al. Combining tree search and action prediction for state-of-the-art performance in DouDiZhu//Proceedings of the 30th International Joint Conference on Artificial Intelligence(IJCAI). Montreal, Canada, 2021: 3413-3419

[223] Clark G. Deep synoptic monte-carlo planning in reconnaissance blind chess//Proceedings of the 35th Neural Information Processing Systems (NIPS). Virtual Event, 2021: 4106-4119

[224] Furtak T, Buro M. Recursive monte carlo search for imperfect information games//Proceedings IEEE Conference on Computational Inteligence in Games (CIG). Niagara Falls, Canada, 2013: 1-8

[225] Powley E J, Cowling P I, Whitehouse D. Information capture and reuse strategies in monte carlo tree search, with applications to games of hidden information. Artificial Intelligence, 2014, 217:92-116

[226] Powley E J, Whitehouse D, Cowling P I. Bandits all the way down: UCB1 as a simulation policy in monte carlo tree search//Proceedings of IEEE Conference on Computational Inteligence in Games (CIG). Niagara Falls, Canada, 2013: 1-8

[227] Goodman J. Re-determinizing information set monte carlo tree search in hanabi. arXiv preprint arXiv:1902.06075,2019

[228] Chakraborty D, Stone P. Multiagent learning in the presence of memory-bounded agents. Autonomous Agents and Multi-Agent Systems, 2014, 28(2):182-213

[229] Spronck P, den Teuling F. Player modeling in Civilization

IV//Proceedings of the 6th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment(AIIDE). Stanford, USA, 2010,6(1):180-185

[230] Synnaeve G, Bessière P. A bayesian model for opening prediction in RTS games with application to starcraft//Proceedings of IEEE Conference on Computational Intelligence and Games(CIG). Seoul, Republic of Korea, 2011: 281-288

[231] Barrett S, Stone P, Kraus S, et al. Teamwork with limited knowledge of teammates//Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI). Bellevue, USA, 2013

[232] Sukthankar G, Geib C, Bui H H, et al. Plan, Activity, and Intent Recognition: Theory and practice. Oxford, UK: Newnes, 2014

[233] Grover A, Al-Shedivat M, Gupta J K, et al. Learning policy representations in multiagent systems//Proceedings of the 35th International Conference on Machine Learning(ICML). Stockholmsmässan, Sweden, 2018: 1797-1806

[234] Dai Z, Chen Y, Low B K H, et al. R2-B2: Recursive reasoning-based bayesian optimization for no-regret learning in games//Proceedings of the 37th International Conference on Machine Learning(ICML). Virtual Event, 2020: 2291-2301

[235] Yuan L, Fu Z, Shen J, et al. Emergence of pragmatics from referential game between theory of mind agents. arXiv preprint arXiv:2001.07752,2020

[236] Jara-Ettinger J. Theory of mind as inverse reinforcement learning. Current Opinion in Behavioral Sciences, 2019, 29: 105-110

[237] Byom L J, Mutlu B. Theory of mind: Mechanisms, methods, and new directions. Frontiers in Human Neuroscience, 2013, 7:413

[238] Chong J, Ho T, Camerer C F. A generalized cognitive hierarchy model of games. Games and Economic Behavior, 2016, 99:257-274

[239] Wen Y, Yang Y, Luo R, et al. Probabilistic recursive reasoning for multi-agent reinforcement learning//Proceedings of 7th International Conference on Learning Representations (ICLR). New Orleans, USA, 2019

[240] Wen Y, Yang Y, Wang J. Modelling bounded rationality in multiagent interactions by generalized recursive reasoning//Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI). Yokohama, Japan, 2020: 414-421

[241] Foerster J N, Chen R Y, Al-Shedivat M, et al. Learning with opponentlearning awareness//Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS). Stockholm, Sweden, 2018: 122-130

[242] Kim D, Liu M, Riemer M, et al. A policy gradient algorithm for learning to learn in multiagent reinforcement learning//Proceedings of Machine Learning Research: Volume 139 Proceedings of the 38th International Conference on Machine Learning, ICML 2021. Virtual Event, 2021: 5541-5550

[243] Yu X, Jiang X Y, Zhang W, Jiang H, et al. Model-based opponent modeling//Proceedings of the 36th Neural Information Processing Systems (NIPS). New Orleans, USA, 2022: 28208-28221

[244] Johanson M, Zinkevich M, Bowling M H. Computing robust counterstrategies//Proceedings of the 21st Neural Information Processing Systems (NIPS). British Columbia, Canada, 2007: 721-728

[245] Ganzfried S, Sandholm T. Safe opponent exploitation. ACM Trans. Economics and Comput., 2015, 3(2): 8: 1-8:28

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

[247] Liu M, Wu C, Liu Q, et al. Safe opponent-exploitation subgame refinement//Proceedings of the 36th Neural Information Processing Systems (NIPS). New Orleans, USA, 2022: 27610-27622

[248] Slumbers O, Mguni D H, Blumberg S B, et al. A game-theoretic framework for managing risk in multi-agent systems//Proceedings of the International Conference on Machine Learning(ICML). Honolulu, USA, 2023,202: 32059-32087

[249] Wu Z, Li K, Xu H, et al. L2E: Learning to exploit your opponent//International Joint Conference on Neural Networks, IJCNN 2022. Padua, Italy, 2022: 1-8

[250] Rusu A A, Colmenarejo S G, Gülçehre Ç, et al. Policy distillation//Proceedings of the 4th International Conference on Learning Representations(ICLR). San Juan, Puerto Rico, 2016, 4697-4720

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

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

[253] WEI Ting Ting, YUAN Wei Lin, LUO Jun-ren. Methods in adversarial intelligent game:A holistic comparative analysis from perspective of game theory and reinforcement learning. Journal of Computer Engineering and Applications, 2022, 58: 19-29(in Chinese)
(魏婷婷, 袁唯淋, 罗俊仁. 智能博弈对抗中的对手建模方法及其应 用综述. 计算机工程与应用, 2022, 58: 19-29)

[254] Tesauro G. Td-gammon, a self-teaching backgammon program, achieves master-level play. Neural Computation, 1994, 6(2):215-219

[255] Frank I, Basin D A, Matsubara H. Finding optimal strategies for imperfect information games//Mostow J, Rich C. Proceedings of the 15th AAAI Conference on Artificial In-

telligence(AAAI). Madison, USA, 1998: 500-507

[256] Zarick R, Pellegrino B, Brown N, et al. Unlocking the potential of deep counterfactual value networks. arXiv preprint arXiv:2007.10442,2020

[257] Sokota S, Farina G, Wu D J, et al. The update equivalence framework for decision-time planning. arXiv preprint arXiv: 2304.13138, 2023

[258] Szafron D, Gibson R G, Sturtevant N R. A parameterized family of equilibrium profiles for three-player kuhn poker// Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS). Saint Paul, USA, 2013: 247-254

[259] Gibson R. Regret minimization in non-zero-sum games with applications to building champion multiplayer computer poker agents. arXiv preprint arXiv:1305.0034, 2013

[260] Berg K, Sandholm T. Exclusion method for finding nash equilibrium in multiplayer games//Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI). San Francisco, USA, 2017: 383-389

[261] Kawamura K, Mizukami N, Tsuruoka Y. Neural fictitious self-play in imperfect information games with many players//The 6th Computer Games Workshop(CGW) Held in Conjunction with the Proceedings of the 26th International Conference on Artificial Intelligence(IJCAI). Melbourne, Australia, 2017: 61-74

[262] Basilico N, Celli A, Nittis G D, et al. Team-maxmin equilibrium: Efficiency bounds and algorithms//Proceedings of the 31st AAAI Conference on Artificial Intelligence(AAAI). San Francisco, USA, 2017: 356-362

[263] Celli A, Gatti N. Computational results for extensive-form adversarial team games//Proceedings of the 32nd Neural Information Processing Systems(NIPS). Montréal, Canada, 2018: 965-972

[264] Zhang Y, An B, Subrahmanian V S. Correlation-based algorithm for team-maxmin equilibrium in multiplayer extensive-form games//Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence(IJCAI). Vienna, Austria, 2022: 606-612

[265] Basilico N, Celli A, Nittis G D, et al. Coordinating multiple defensive resources in patrolling games with alarm systems//Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems(AAMAS). São Paulo, Brazil, 2017: 678-686

[266] Rahman A, Höpner N, Christianos F, et al. Towards open ad hoc teamwork using graph-based policy learning//Proceedings of Machine Learning Research: Volume 139 Proceedings of the 38th International Conference on Machine Learning, ICML 2021. Virtual Event, 2021: 8776-8786

[267] Aumann R J. Subjectivity and correlation in randomized strategies. Journal of mathematical Economics, 1974, 1(1):67-96

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

[269] Farina G, Ling C K, Fang F, et al. Correlation in extensive-form games: Saddle-point formulation and benchmarks//Proceedings of the 33rd Neural Information Processing Systems (NIPS). Vancouver, Canada, 2019: 9229-9239

[270] Koutsoupias E, Papadimitriou C H. Worst-case equilibria. Computer Science Review, 2009, 3(2):65-69

[271] Roughgarden T, Tardos É. How bad is selfish routing? Journal of the ACM (JACM), 2002, 49(2):236-259

[272] Farina G, Bianchi T, Sandholm T. Coarse correlation in extensive form games//Proceedings of the 34th AAAI Conference on Artificial Intelligence(AAAI). New York, USA, 2020: 1934-1941

[273] Celli A. Coordination and correlation in multi-player sequential games. Polytechnic University of Milan,Milan,Italy, 2020

[274] Morrill D, D'Orazio R, Sarfati R, et al. Hindsight and sequential rationality of correlated play//Proceedings of the 35th AAAI Conference on Artificial Intelligence(AAAI). Virtual Event, 2021: 5584-5594

[275] von Stengel B, Koller D. Team-maxmin equilibria. Games and Economic Behavior, 1997, 21(1-2):309-321

[276] Basilico N, Celli A, Nittis G D, et al. Team-maxmin equilibrium: Efficiency bounds and algorithms//Proceedings of the 31st AAAI Conference on Artificial Intelligence(AAAI). San Francisco, USA, 2017: 356-362

[277] Zhang Y, An B. Converging to team-maxmin equilibria in zero-sum multiplayer games//Proceedings of the 37th International Conference on Machine Learning(ICML). Virtual Event, 2020: 11033-11043

[278] Zhang Y,An B. Computing team-maxmin equilibria in zero-sum multiplayer extensive-form games//Proceedings of the AAAI Conference on Artificial Intelligence,2020, 34(2), 2318-2325

[279] Farina G, Kroer C, Sandholm T. Better regularization for sequential decision spaces: Fast convergence rates for nash, correlated, and team equilibria//Proceedings of the 22nd Conference on Economics and Computation(EC). Budapest, Hungary, 2021: 432

[280] Zinkevich M, Greenwald A, Littman M L. Cyclic equilibria in markov games//Proceedings of the 19th Neural Information Processing Systems (NIPS). Vancouver, Canada, 2005: 1641-1648

[281] Kim D K, Riemer M, Liu M, et al. Game-theoretical perspectives on active equilibria: A preferred solution concept over nash equilibria. arXiv preprint arXiv:2210.16175, 2022

[282] Sato Y, Akiyama E, Farmer J D. Chaos in learning a simple two-person game. Proceedings of the National Academy of Sciences, 2002, 99(7): 4748-4751

[283] Balduzzi D, Racaniere S, Martens J, et al. The mechanics of n-player differentiable games//Proceedings of the 35th

International Conference on Machine Learning (ICML). Stockholm, Sweden, 2018: 363-372

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

[285] Letcher A, Foerster J N, Balduzzi D, et al. Stable opponent shaping in differentiable games//Proceedings of the 7th International Conference on Learning Representation(ICLR). New Orleans, USA, 2019

[286] Heusel M, Ramsauer H, Unterthiner T, et al. GANs trained by a two time-scale update rule converge to a local nash equilibrium//Proceedings of the 31st Neural Information Processing Systems(NIPS). Long Beach, USA, 2017: 6626-6637

[287] Gidel G, Hemmat R A, Pezeshki M, et al. Negative momentum for improved game dynamics//Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics(AISTATS). Naha, Japan, 2019: 1802-1811

[288] Farina G, Kroer C, Sandholm T. Optimistic regret minimization for extensive-form games via dilated distance-generating functions//Proceedings of the 33rd Neural Information Processing Systems (NIPS). Vancouver, Canada, 2019: 5222-5232

[289] Pham H X, La H M, Feil-Seifer D, et al. Cooperative and distributed reinforcement learning of drones for field coverage. arXiv preprint arXiv:1803.07250, 2018

[290] Kiran B R, Sobh I, Talpaert V, et al. Deep reinforcement learning for autonomous driving: A survey. arXiv preprint arXiv:2002.00444, 2020

[291] Abapour S, Nazari-Heris M, Mohammadi-Ivatloo B, et al. Game theory approaches for the solution of power system problems: A comprehensive review. Archives of Computational Methods in Engineering, 2020, 27(1):81-103

[292] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks//Proceedings of the 27th Neural Information Processing Systems (NIPS). Montreal, Canada, 2014: 2672-2680

[293] Bai T, Luo J, Zhao J, et al. Recent advances in adversarial training for adversarial robustness//Proceedings of the 24th International Joint Conference on Artificial Intelligence(IJCAI). Virtual Event, 2021: 4312-4321

[294] Jacob A P, Shen Y, Farina G, et al. The consensus game: Language model generation via equilibrium search. arXiv preprint arXiv:2310.09139, 2023

[295] Du Y, Li S, Torralba A, et al. Improving factuality and reasoning in language models through multiagent debate. arXiv preprint arXiv:2305.14325, 2023

[296] CHEN Jin-Chuan, XIA Hua-Hui, WANG Pu-Wei, et al. Detecting smart contact loopholes based on nash equilibrium. Chinese Journal of Computers, 2021, 44(1): 147-161. (in Chinese)
(陈晋川, 夏华辉, 王璞巍, 等. 基于纳什均衡的智能合约缺陷检测. 计算机学报, 2021, 44(1): 147-161)

YU Chao, Ph. D., professor. His research interests include game intelligence, reinforcement learning, and multiagent systems.

LIU Zong-Kai, M. S. candidate. His research interests include game intelligence, multi-agent reinforcement lear-ning, and imperfect information game.

HU Chao-Hao, M. S. candidate. His research interests include game intelligence, multi-agent reinforcement learning, and imperfect information game.

HUANG Kai-Qi, Ph. D., professor. His research interests include computer vision, game intelligence.

Zhang Jun-Ge, Ph. D., professor. His research interests include game intelligence, multiagent systems, and pattern recognition.

Background

In recent years, the rapid advancement in industries such as big data, the Internet of Things, and cloud computing has led to the rapid development of Artificial Intelligence (AI) technology, which has been deeply integrated into a wide range of application scenarios. The primary goal of AI is to design and implement intelligent algorithms that can interact dynamically with the environment and make optimal decisions automatically. However, in real-world applications, there are often multiple interconnected intelligent agents that interact with the same environment, which makes the environment non-stationary and thus poses a significant challenge for the decision making of each single intelligent agent. Game theory, as a discipline that studies the interactions and rational decision-making of multiple intelligent agents, is able to abstract the decision-making objectives and interaction rules of intelligent entities to a high degree. In recent years, the combination of Artificial Intelligence and game theory has formed a new research direction called "Game Intelligence" and has become a cutting-edge field in the AI research.

As a typical example of Game Intelligence, Imperfect Information Game (IIG) models the game behavior of multi-

ple intelligent agents under private information, and is able to capture a broader range of decision-making processes than Perfect Information Game (PIG). It has a wide range of applications in the real world. However, due to the fact that intelligent agents cannot accurately perceive the full picture of the game, finding equilibria in IIG is extremely challenging. In recent years, there have been significant breakthroughs in the study of IIG, with many achievements being published in prestigious journals such as Nature and Science. These successes have enabled the solution of challenging problems such as card games represented by Texas Hold'em, multiplayer online tactical games, hidden role-playing games, and stratego, making it one of the leading research fields for the next generation of AI.

The goal of this survey is to not only deepen our understanding of the field of Game Intelligence but also drive the advancement of this technology and contribute to the development of general artificial intelligence. Through conducting an in-depth comparison of the strengths and weaknesses of various methods, the survey provides an overview of the current state of research in this field, highlighting areas where further study is needed. It is hoped to offer insights and suggestions for future research that could pave the way for more advanced and sophisticated artificial intelligence technologies.

We gratefully acknowledge support from the National Natural Science Foundation of China (No. 62076259), the Fundamental and Applicational Research Funds of Guangdong Province (No. 2023A1515012946), the Basic Cultivation Fund Project of Chinese Academy of Sciences (JCPYJJ-22017), the Chinese Academy of Sciences Youth Promotion Association Project, and the Fundamental Research Funds for the Central Universities—Sun Yat-sen University.