人机对抗中的博弈学习方法

周 雷 尹奇跃 黄凯奇

(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 近年来,人机对抗智能技术作为人工智能领域的前沿方向取得了一系列突破性的进展,如 AlphaGo 和 DeepStack 分别在围棋和二人无限注德州扑克中击败了人类专业选手.这些突破离不开博弈论和机器学习的深度结合.本文通过梳理当前人机对抗智能技术领域的重要工作,深入分析博弈论和机器学习在其中发挥的作用,总结了面向人机对抗任务的博弈学习研究框架,指出博弈论为人机对抗任务提供博弈模型和定义求解目标,机器学习帮助形成稳定高效可扩展的求解算法.具体地,本文首先介绍了人机对抗中的博弈学习方法的内涵,详细阐述了面向人机对抗任务的博弈学习研究框架,包括博弈模型构建、解概念定义、博弈解计算三个基本步骤,之后利用该框架分析了当前人机对抗智能技术领域的典型进展,最后指出了人机对抗中的博弈学习未来发展可能面临的挑战.本文梳理总结的人机对抗中的博弈学习研究框架为人机对抗智能技术领域的发展提供了方法保障和技术途径,同时也为通用人工智能的发展提供了新思路.

关键词 人工智能;人机对抗;博弈论;机器学习;博弈学习

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2022.01859

Game-Theoretic Learning in Human-Computer Gaming

ZHOU Lei YIN Qi-Yue HUANG Kai-Qi

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

Abstract Recent development in the field of human-computer gaming, one of the frontiers in artificial intelligence (AI), has witnessed a series of breakthroughs, such as AlphaGo and DeepStack beat professional human players in Go and heads-up no-limit Texas Hold'em, respectively. Such successes demonstrate synergistic interactions between game theory and machine learning. Game theory is a theoretical framework that deals with strategic interactions among multiple rational players. Combined with machine learning, it is well suited for modeling, analyzing, and solving decision-making problems in human-computer gaming tasks that often involve two or more decision-makers. Game theory based learning methods thus receive increasing attention in recent years. Besides the popular multi-agent reinforcement learning approaches, there are some other game theory based learning methods, i.e., game-theoretic learning methods, that are designed to converge to equilibria and can be dated back to the famous fictitious play proposed in 1951. In this paper, we give a selective overview of such game-theoretic learning methods in human-computer gaming. By analyzing key progresses in the field of human-computer gaming and game theory (including game-theoretic learning), we obtain a research framework for game-theoretic learning in human-computer gaming. In this framework, the role of game theory and machine learning each plays is identified: game theory provides models of strategic interactions and defines associated learning objectives (i.e., solution concepts) while machine learning helps give rise to stable,

收稿日期:2021-09-14;在线发布日期:2022-03-22.本课题得到中国科学院战略性先导科技专项(A 类)(XDA27010103)资助.**周 雷**,博士,助理研究员,主要研究方向为演化博弈论、机器学习、博弈学习.E-mail:lei.zhou@ia.ac.cn.**尹奇跃**,博士,副研究员,硕士生导师,中国计算机学会(CCF)会员,主要研究领域为机器学习、数据挖掘、博弈决策.**黄凯奇**(通信作者),博士,研究员,博士生导师,主要研究领域为计算机视觉、模式识别、认知决策.E-mail:kqhuang@nlpr.ia.ac.cn.

efficient, and scalable game solving algorithms. In detail, we first review important progresses in the field of human-computer gaming and game theory. Then, we introduce the definition of game-theoretic learning in human-computer gaming and compare it with traditional machine learning methods such as supervised learning and single-agent reinforcement learning. After that, we elaborate on its research framework. Intuitively, this research framework equivalently or approximately transforms the problem of achieving a good performance in a class of human-computer gaming tasks into the problem of solving a class of games. As we summarize, such transformation usually takes three basic steps: game model formulation, solution concept definition, and game solution computation. Employing this framework, we also analyze a recent game-theoretic learning algorithm that combines fictitious play and deep reinforcement learning called neural fictitious self-play, and also three milestones in the field of human-computer gaming, i. e., AlphaGo Zero, Libratus, and AlphaStar. At the end, we point out possible problems and challenges in the future research of game-theoretic learning in human-computer gaming, such as the definition of learning objectives in general-sum games, the interpretability of game-theoretic learning algorithms based on deep neural networks, the design of diverse environment suitable for game-theoretic learning, and the efficient solving of complex large-scale games that may exhibit non-transitive game behaviors. We believe that the research framework of game-theoretic learning in human-computer gaming offers guidance for the future development of human-computer gaming, and it also provides new perspectives on the development of artificial general intelligence.

Keywords artificial intelligence; human-computer gaming; game theory; machine learning; game-theoretic learning

1 引 言

人机对抗智能技术研究计算机博弈中机器战胜人类的方法,是当前人工智能研究领域的前沿方向,它以人机(人类与机器)和机机(机器与机器)对抗为主要形式研究不同博弈场景下,机器智能战胜人类智能的基础理论与方法技术[1]. 人机对抗智能技术通过人、机、环境之间的博弈对抗和交互学习,探索巨复杂、高动态、不确定的对抗环境下机器智能快速增长的机理和途径,以期最终达到或者超越人类智能①.

人机对抗智能技术的突破离不开机器学习的发展,机器学习主要研究如何让机器通过与数据的交互实现能力的提升[2-3]. 然而,与传统的机器学习关注单智能体(single-agent)与环境的交互不同,人机对抗智能技术研究的场景往往包含两个或两个以上智能体,也就是多智能体(multi-agent)的情形,这些智能体都拥有自己的优化目标,比如最大化自身收益. 此时,如果直接应用单智能体机器学习方法,得到的智能体(称为中心智能体)一般表现欠佳[4-5]. 这是因为传统机器学习方法假设数据的产生机制是平稳的(stationary)[6](即数据均来自于同一个分布,简称为环境的平稳性),这一假设忽略了研究场景中的其他智能体,而这些智能体也同时在进行学习,因此其行为模式会随着时间发生变化,从而破坏中心智能体所处环境的平稳性,进而导致传统机器学习方法失去理论保证[2-3]. 更为严峻的是,随着人机对抗场景中智能体数量的增加,环境非平稳的问题将会愈发凸显,多个趋利的智能体在学习的过程中相互影响的情况将不可避免.

为了处理环境非平稳的问题,有学者考虑将博弈论引入机器学习方法中[7]. 这主要是因为博弈论本身就是为了研究多个利己个体之间的策略性交互(strategic interactions)而发展的数学理论. 博弈论诞生于 1944 年 von Neumann 和 Morgenstern 合著的 Theory of Games and Economic Behavior[8]. 在完全理性的假设下,博弈论给出了一系列解概念来预测博弈最终可能的结果. 博弈论早期的大部分工作关注不同博弈场景下解概念(solution concepts)的定义、精炼(refinement)、存在性及其拥有的性质[9]. 随

① 人机对抗智能技术门户网站. 网址:http://turingai.ia.ac.cn/

着博弈论的发展，部分研究者开始研究在非完全理性的情形下，个体是否可以通过迭代学习的方式来达到这些解概念，其中著名的工作包括 Brown 提出的虚拟对局(fictitious play)[10]、Hannan 和 Blackwell 研究的无悔学习(no-regret learning，regret minimization，or Hannan consistency)[11-13]等.

近年来，得益于机器算力的提升和深度学习的兴起，人机对抗智能技术领域取得了一系列突破，如 DeepMind 团队开发的 AlphaGo[14]首次击败了人类围棋顶尖选手李世石，阿尔伯塔大学团队开发的 DeepStack[15]在二人无限注德州扑克中击败了专家级人类选手等. 在 AlphaGo 中，围棋被建模为二人零和完美信息扩展形式博弈，并利用自我对局、蒙特卡洛树搜索以及深度神经网络近似来对博弈进行求解；在 DeepStack 中，二人德州扑克被建模为二人零和非完美信息扩展形式博弈，求解方法结合了自我对局、反事实遗憾最小化算法以及深度神经网络近似. 从这些例子可以看出，人机对抗智能技术领域的突破离不开博弈论和机器学习的深度结合.

然而，虽然人机对抗智能技术领域目前取得了一系列突破，博弈论与机器学习交叉方向的研究却缺乏清晰的研究框架. 基于此，本文通过梳理人机对抗智能技术领域的重要工作，介绍了人机对抗中的博弈学习方法的内涵，总结了面向人机对抗任务的博弈学习研究框架，包括其组成要素和基本步骤，并利用该框架对人机对抗智能技术领域的典型进展进行了分析. 本文作者认为，随着人机对抗智能技术领域实验场景和测试环境逐渐接近真实场景，场景的复杂性和对抗性急剧增加，结合现代机器学习方法和博弈论的博弈学习方法将会在未来人机对抗领域的发展中发挥越来越重要的作用.

2　发展历史

自图灵测试这一人机对抗模式在 1950 年被提出[16]以来，博弈论和机器学习就在人工智能的发展中发挥着越来越重要的作用，并呈现出交叉融合的趋势. 本文梳理了人机对抗智能技术和博弈论领域开创性的工作和里程碑事件，并将其发展历史分为两条路线，一条是博弈论结合专家系统(见图 1 中绿色实线)，另一条是博弈论结合学习方法(见图 1 中橙色虚线).

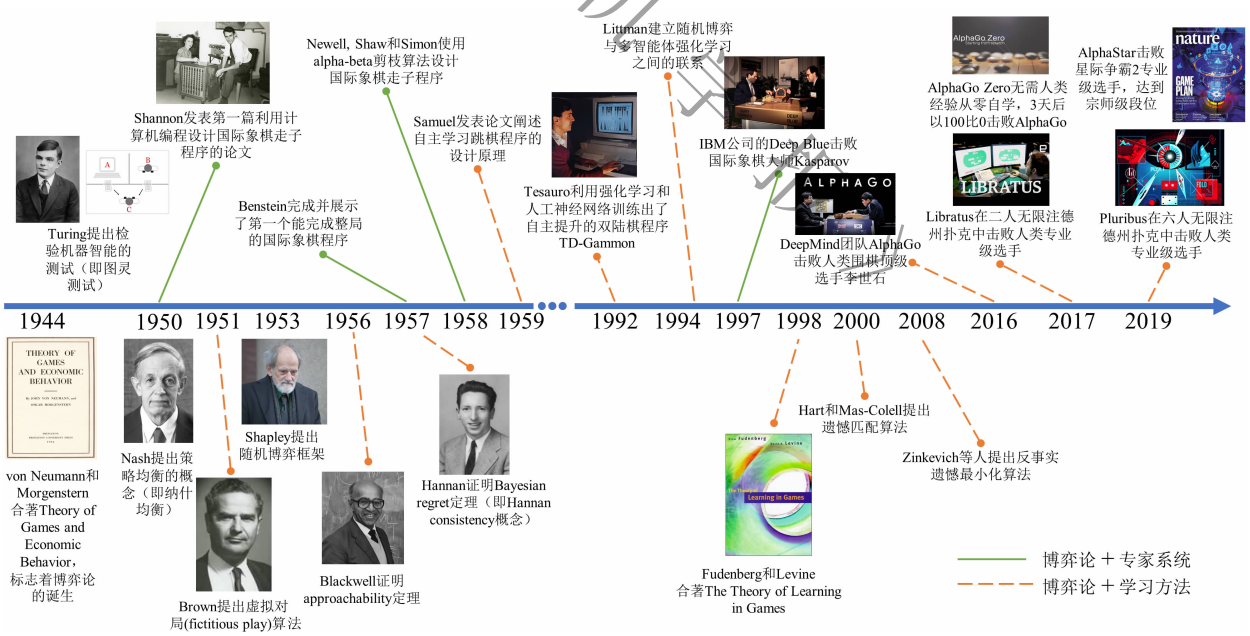

图 1　人机对抗智能技术与博弈论的发展历史

2.1　路线一：博弈论结合专家系统

在发展路线一中，为了取得较好的人机对抗表现，研究者们主要是针对基于博弈论的 min-max 树搜索算法进行优化，并结合专家经验来改进评估函数. 路线一的简要发展历程如下：

1950 年，Shannon 发表了第一篇利用编程来实现国际象棋走子程序的论文[17]，论文中 Shannon 参考 von Neumann 证明的 minimax 定理[8,18]设计了 min-max 搜索算法和局面评估函数. 对于局面评估函数的设计，参考的是如下定理：在国际象棋中，最终的结局只可能是以下三种当中的一种：(1) 不论白方如何走子，黑方有一种策略总能保证赢；(2) 不

论黑方如何走子,白方有一种策略总能保证赢;(3)黑白双方都有一种策略保证至少平局.

1956 年,Samuel 利用第一台商用计算机 IBM 701 编写了跳棋(checkers)走子程序,并在 1959 年发表论文总结了该程序的设计思想和原理[19]. 该跳棋走子程序使用了 min-max 搜索.

1957 年,Bernstein 带领的团队在 IBM 701 上完成了第一个能下完整局的国际象棋走子程序,该程序使用了 min-max 搜索,但每次最多向后搜索 4 步,每步只能考虑 7 个备选走法.

1958 年,Newell、Shaw 和 Simon 第一次在国际象棋程序中使用 alpha-beta 剪枝搜索算法[20]. Alpha-beta 剪枝算法是 min-max 搜索算法的改进,通过剪掉明显次优的子树分支,该算法极大地降低了搜索空间. 该算法最初由 McCarthy 在 1956 年提出.

此后,跳棋和国际象棋程序的优化大多围绕评估函数和搜索算法进行改进. 随着计算能力的增强,IBM 公司开发的国际象棋程序 Deep Blue 在 1997 年利用总结了大量人类经验的评估函数和强大的搜索能力击败国际象棋大师 Kasparov,一时轰动. 该事件从此成为人机对抗智能技术发展历史上的标志性事件.

2.2 路线二:博弈论结合学习方法

路线一中采用的方法很难称得上实现了机器的"学习"能力,在路线二中,研究者们试图克服机器对专家数据的过度依赖,希望能够打造自主学习的智能机器. 路线二的简要发展历程如下:

最早在人机对抗研究中引入学习的是 Samuel,他 1957 年完成的跳棋走子程序不仅使用了 min-max 搜索,同时也引入了两种"学习"机制[19]:死记硬背式学习(rote learning)和泛化式学习(learning by generalization). 前者通过存储之前下棋过程中计算得到的局面得分来减少不必要的搜索,后者则根据下棋的不同结果来更新评估函数中不同参数的系数来得到一个更好的评估函数. 此外,该论文也第一次提到了自我对局(self-play). 此后,这种通过学习来提升机器能力的思想就一直没能引起重视. 直到 1990 年前后,才陆续出现了能够学习的棋类程序. 这其中比较知名的是 1994 年 Tesauro 结合神经网络和强化学习训练出的双陆棋程序 TD-Gammon[21].

TD-Gammon 的成功引起了许多学者对学习算法的兴趣,并促成了博弈论与机器学习的初步结合,其中著名的工作是 Littman 在 1994 年正式建立了 Markov 博弈(或随机博弈)与多智能体强化学习之间的联系. 之后,Markov 博弈便作为多智能体强化学习的理论框架,启发了众多学者的研究. 同时,在该论文中 Littman 也提出了第一个多智能体强化学习算法 minimax-Q[22]. Minimax-Q 是针对二人零和博弈的学习算法,当博弈的双方都使用该算法时,最终博弈双方的策略都会收敛到二人零和博弈的最优解极大极小策略上.

值得指出的是,除了人工智能领域,博弈论领域的研究者们很早也开始了对学习方法的研究. 与人工智能领域学者的出发点不同,他们关注的是在博弈模型给定的情形下,如何设计迭代学习的规则能使个体的策略收敛到均衡. 此类方法之后被称为博弈学习(game-theoretic learning)方法. 博弈学习方法的思想最早可以追溯到 1951 年 Brown 提出的虚拟对局(fictitious play)[10],即采用迭代学习的方式来计算二人零和博弈的极大极小策略,之后著名的博弈学习方法包括无悔学习(no-regret learning)[11-13]和复制动力学(replicator dynamics)[23]. 在 1998 年,几乎与 Littman 等人同一时期,Fudenberg 和 Levine 出版了著作 The Theory of Learning in Games[24],对之前博弈学习方法的研究进行了汇总、总结和扩展. 博弈学习方法的研究为博弈论中的解概念(主要是纳什均衡)提供了非理性假设下的解释,换言之,非理性的个体在一定学习规则的指导下也能达到均衡.

此后,博弈论和机器学习领域的研究兴趣和研究内容开始交叉,逐步形成了博弈论与机器学习结合的博弈学习方法[25-30]. 相关工作包括:(1)利用强化学习方法计算博弈的解,比如 Nash-Q[31]等;(2)利用博弈论中的学习方法进行游戏 AI 的算法设计,比如针对不完美信息博弈的反事实遗憾最小化算法[28](属于无悔学习算法的一种);(3)利用机器学习加强博弈论中学习方法的可扩展性,比如虚拟自我对局(Fictitious Self-Play,FSP)[29]. 相比于传统解决单智能体与环境交互问题的机器学习方法,与博弈论结合的学习方法有两个优势:一是充分考虑了多个智能体同时最大化收益时环境的非平稳问题,学习的目标是任务的均衡解而不是让某个智能体的收益最大化;二是在满足模型的假设时,这些算法一般具有收敛的理论保证. 特别地,面向人机对抗任务,人机对抗中的博弈学习方法在此基础上添加了人机对抗任务建模,为的是能更好地利用和拓展现有的博弈学习方法来处理复杂的人机对抗任务.

近年来，随着深度学习的兴起，深度神经网络被广泛应用于人机对抗任务，形成了一系列优秀的模型和博弈学习算法[5,32-40]. 这也促进了人机对抗智能技术近期一系列的突破，包括 2016 年 AlphaGo 击败围棋 9 段选手李世石，2017 年 Libratus[30] 和 DeepStack[15] 分别在二人无限注德州扑克中击败人类专业选手以及 2019 年 AlphaStar[41] 在星际争霸 2 中击败人类顶级选手.

3　人机对抗中的博弈学习方法内涵

人机对抗中的博弈学习方法是一种面向人机对抗任务，以博弈论为理论基础、以机器学习为主要技术手段，通过智能体与环境、智能体与其他智能体的交互来获得具有良好性质（比如适应性、鲁棒性等等）博弈策略的学习方法，是实现人机对抗智能技术的核心. 具体地，人机对抗中的博弈学习方法基于博弈论建模人机对抗任务和定义学习目标，并利用机器学习方法来帮助设计高效、稳健、可扩展的学习算法以完成人机对抗任务.

为了阐述博弈学习方法与当前机器学习方法的区别与联系，本文按照系统中信息的流向以及信息产生的机制将已有的学习框架划分为一元、二元以及三元（或多元）学习. 在一元学习中，智能体从数据中获取知识，并且这个过程只涉及数据到智能体的单向信息流动，监督学习、无监督学习以及深度学习都属于一元学习（见图 2(a)）. 在二元学习中，智能体通过与环境互动得到数据，进而获取知识，与一元学习不同的是此时数据的产生不仅取决于环境也取决于智能体，即智能体决策的好坏影响它自身学习的效果，必要时智能体还需要对环境动力学进行建模，单智能体强化学习属于二元学习（见图 2(b)）. 在三元学习中，智能体通过与环境和其他智能体的交互获得数据，此时智能体学习的效果受到环境和其他智能体的共同影响，必要时智能体需要对环境动力学和其他智能体进行建模（见图 2(c)），博弈学习属于三元学习.

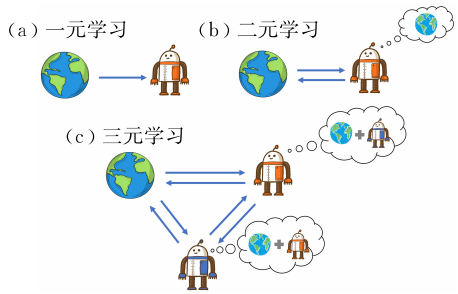

图 2　博弈学习与机器学习的区别与联系

4　人机对抗中的博弈学习研究框架

通过对博弈论和人机对抗智能技术发展历程的梳理，并结合人机对抗中的博弈学习方法的内涵，本文总结出了如图 3 所示的人机对抗中的博弈学习研究框架：人机对抗中的博弈学习研究框架以人机对抗任务为输入，首先通过博弈模型构建获得博弈模型，然后通过解概念定义得到博弈的可行解，最后通过博弈解计算输出满足需求的博弈策略组合，也就是学习任务的解. 直观来讲，人机对抗中的博弈学习研究框架将一类人机对抗任务的解决近似或等价转换为对某一类博弈问题的求解，该框架包含两个组成要素（博弈模型和博弈解）和三个基本步骤（博弈模型构建、解概念定义和博弈解计算）.

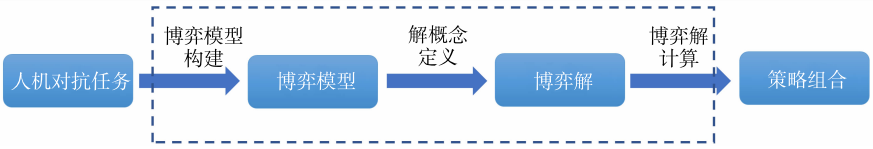

图 3　人机对抗中的博弈学习研究框架

接下来，本文就人机对抗中的博弈学习研究框架中的两个组成要素和三个基本步骤分别进行介绍.

4.1　博弈模型和博弈解

博弈模型和博弈解是人机对抗中的博弈学习研究框架中的两个重要组成要素，本文参考经典博弈论中的内容对它们进行介绍以便理解，更多的可参考文献[9,25]. 在博弈论中，经典的博弈模型有标准形式博弈（normal-form game）、扩展形式博弈（extensive-form game）、Markov 博弈（Markov game）等. 下面对它们的定义、解概念进行一一介绍.

4.1.1　标准形式博弈

标准形式博弈又称为策略形式博弈（strategic-form game），其定义如下：

定义 1.　标准形式博弈[9,25]. 标准形式博弈 G 由 $\langle N,(A_i,\mathcal{R}_i)_{i\in N}\rangle$ 表示，其中 $N=\{1,2,\cdots,n\}$ 表示

参与博弈所有个体的集合;对于任意个体 $i\in N$,A_i 是其动作集合;$A=\times_{i=1}^{n}A_i$ 表示所有个体的联合动作空间;$\mathcal{R}_i:A\to\mathbb{R}$ 表示个体 i 的收益函数,$\mathcal{R}_i(\boldsymbol{a})$ 表示在联合动作 $\boldsymbol{a}=(a_1,a_2,\cdots,a_n)$ 下,个体 i 的收益.

标准形式博弈可以看作一个 n 维矩阵,矩阵中的元素是个体在不同的联合动作下收益的向量,也就是 $(\mathcal{R}_1,\cdots,\mathcal{R}_n)$,因此标准形式博弈也称为矩阵博弈(matrix game).标准形式博弈是博弈论中最基本的博弈模型,它最直接地描述了个体的行动与收益之间的关系,并且几乎所有其他的博弈模型都可以转化为标准形式.此外,在标准形式博弈中,所有个体同时决策(或者某一个个体在做决策时不知道其他个体当前的决策结果)且只决策一次.特别地,当 $n=2$ 且对于任意的 $\boldsymbol{a}\in A$ 都有 $\mathcal{R}_1(\boldsymbol{a})+\mathcal{R}_2(\boldsymbol{a})=0$ 时,称为二人零和标准形式博弈.

定义 2. 标准形式博弈策略[9,25].给定标准形式博弈 $G=\langle N,(A_i,\mathcal{R}_i)_{i\in N}\rangle$,对于任意玩家 i,其策略 σ_i 为集合 A_i 上的概率分布 $\Delta(A_i)$;$\sigma_i(a_i)$ 表示个体 i 使用行动 $a_i\in A_i$ 的概率.

为方便说明,本文记个体 i 的策略集合为 Σ_i;所有个体策略的联合为 $\sigma=(\sigma_1,\cdots,\sigma_n)$;记在联合策略 σ 下联合行动 $\boldsymbol{a}$ 出现的概率为 $\sigma(\boldsymbol{a})=\prod_{i=1}^{N}\sigma_i(a_i)$;记除个体 i 之外,其他所有个体策略的联合为 σ_{-i},即 $\sigma_{-i}=(\sigma_1,\cdots,\sigma_{i-1},\sigma_{i+1},\cdots,\sigma_n)$.此时,标准形式博弈中的期望收益如下.

定义 3. 标准形式博弈期望收益[9,25].给定标准形式博弈 $G=\langle N,(A_i,\mathcal{R}_i)_{i\in N}\rangle$ 和联合策略 $\sigma=(\sigma_1,\cdots,\sigma_n)$,对于任意个体 i,其期望收益为

$$\overline{\mathcal{R}}_i(\sigma)=\sum_{\boldsymbol{a}\in A}\sigma(\boldsymbol{a})\mathcal{R}_i(\boldsymbol{a}) \tag{1}$$

在标准形式博弈下,经典的博弈解概念是纳什均衡[42]和近似纳什均衡(即 ϵ-纳什均衡).

定义 4. 标准形式博弈 ϵ-纳什均衡[9,25].给定标准形式博弈 $G=\langle N,(A_i,\mathcal{R}_i)_{i\in N}\rangle$ 以及 $\epsilon\geqslant 0$,博弈 G 的 ϵ-纳什均衡是一个策略组合 $\sigma^*=(\sigma_1^*,\sigma_2^*,\cdots,\sigma_n^*)$.该策略组合使得对于任意个体 $i\in N$,

$$\overline{\mathcal{R}}_i(\sigma^*)=\overline{\mathcal{R}}_i(\sigma_i,\sigma_{-i}^*)-\epsilon \tag{2}$$

对于任意的 $\sigma_i\in\Sigma_i$ 均成立.当 $\epsilon=0$ 时,近似纳什均衡就是纳什均衡.

对于标准形式博弈中纳什均衡的存在性,Nash 在 1950 年给出了如下定理:

定理 1. 标准形式博弈纳什均衡存在性[42].给定标准形式博弈 $G=\langle N,(A_i,\mathcal{R}_i)_{i\in N}\rangle$,对于任何有限标准形式博弈($N$ 以及所有 A_i 均为有限集合)都存在至少一个纳什均衡.

4.1.2 扩展形式博弈

在标准形式博弈中,所有个体只决策一次,但现实中一些博弈是需要进行多次决策之后才能知道博弈结果的,比如围棋、德州扑克等,对于这样有明显先后顺序的博弈(又称为回合制博弈)可以用扩展形式博弈来描述:

定义 5. 扩展形式博弈[9,25-27].扩展形式博弈 G 由 $\langle N,A,H,Z,(I_i,\mathcal{R}_i)_{i\in N},\chi,\rho\rangle$ 表示,其中 $N=\{1,2,\cdots,n\}$ 表示参与博弈所有个体的集合(如果博弈中包含随机事件,则添加一个额外的个体 c,此时 $N=\{1,2,\cdots,n\}\cup\{c\}$);$A$ 是博弈中玩家所有动作的集合;定义博弈历史 h 为当前已执行动作的序列 $h=(\alpha_1,\alpha_2,\cdots)$,$h_t=(h_{t-1},\alpha_{t-1})$,$h_1=\varnothing$,$\alpha_t\in A$ 表示第 t 轮个体执行的动作,$H=\{h_t\mid$博弈未在第 t 轮结束$\}$对应博弈未结束时所有可能的历史,$Z=\{h_t\mid$博弈在第 t 轮结束$\}$表示博弈结束时的所有可能历史,且 $H\cap Z=\varnothing$;$\rho:H\to N$,$\rho(h)$ 表示在博弈历史为 h 时需要做决策的个体;$\chi:H\to 2^{|A|}$,$\chi(h)$ 表示在博弈历史为 h 时个体的合法动作集合;$\mathcal{R}_i:Z\to\mathbb{R}$ 表示个体 i 在博弈结束时的收益;$I_i=\{h\in H\mid\rho(h)=i\}$ 是个体 i 所有需要做决策的节点集合,如果 $I_{i1},I_{i2},\cdots,I_{ik_i}$ 是 I_i 的一个划分,且满足对于任意的 $j\in\{1,2,\cdots,k_i\}$,对于任意的节点 $h,h'\in I_{ij}$,$\rho(h)=\rho(h')$ 并且 $\chi(h)=\chi(h')$,则称 I_{ij} 是个体 i 的一个信息集.

在以上定义中,如果所有个体的所有信息集都为单点集,则称博弈为完美信息博弈,此时每个个体在做决策时都知道当前完整的博弈历史;否则,博弈就称为不完美信息博弈.特别地,当 $n=2$ 且对于任意的博弈结束节点 $z\in Z$ 都有 $\mathcal{R}_1(z)+\mathcal{R}_2(z)=0$ 时,称为二人零和扩展形式博弈.

对于扩展形式博弈,以下定义都假设完美回忆(perfect recall),也就是到达任意一个信息集时,任何参与博弈的个体都记得它之前的所有经历(包括执行的动作和到达的信息集).完美回忆是在研究扩展形式博弈时一个较为普遍的假设,在这个假设下,根据 Kuhn 定理[43],以下策略的定义具有一般性:

定义 6. 扩展形式博弈策略[9,25-27].给定扩展形式博弈 $G=\langle N,A,H,Z,(I_i,\mathcal{R}_i)_{i\in N},\chi,\rho\rangle$,对于任意个体 i,定义其信息集的集合为 $S_i=\{I_{i1},\cdots,I_{ik_i}\}$,那么个体 i 的策略为

$$\sigma_i:S_i\to\bigcup_{j=1}^{k_i}\Delta(\chi(I_{ij})) \tag{3}$$

其中 $\sigma_i(I_{ij})$ 表示在信息集 I_{ij} 时，个体 i 使用各合法动作 $a_i\in\chi(I_{ij})$ 的概率. 特别地，当 $\pi_i:S_i\rightarrow\bigcup_{j=1}^{k_i}\chi(I_{ij})$ 时，称 π_i 为个体 i 的纯策略.

为了评估策略的表现，需要定义扩展形式博弈中不同策略的期望收益. 在定义期望收益之前，先引入如下概念：

定义 7.　到达概率[9,25-27]. 给定扩展形式博弈 $G=\langle N,A,H,Z,(I_i,\mathcal{R}_i)_{i\in N},\chi,\rho\rangle$ 和策略组合 $\sigma=(\sigma_1,\cdots,\sigma_n)$，任意节点 $h=(\alpha_1,\alpha_2,\cdots,\alpha_k)\in H\cup Z$ 在 σ 下的到达概率为

$$\beta^{\sigma}(h)=\sigma_{\rho(h_1)}(h_1,\alpha_1)\prod_{i=2}^{k}\sigma_{\rho(h')}(h',\alpha_i)\big|_{h'=(\alpha_1,\alpha_2,\cdots,\alpha_{i-1})}\tag{4}$$

其中 $\sigma_{\rho(h)}(h,a)$ 表示在节点 h，个体 $\rho(h)$ 执行动作 a 的概率；$\rho(h_1)$ 表示博弈最开始(即第 1 轮)时做决策的个体.

在以上定义的基础上，扩展形式博弈的期望收益为：

定义 8.　扩展形式博弈期望收益[25-27]. 给定扩展形式博弈 $G=\langle N,A,H,Z,(I_i,\mathcal{R}_i)_{i\in N},\chi,\rho\rangle$ 和策略组合 $\sigma=(\sigma_1,\cdots,\sigma_n)$，个体 i 在 σ 下的期望收益为

$$\overline{\mathcal{R}}_i(\sigma)=\sum_{h\in Z}\beta^{\sigma}(h)\mathcal{R}_i(h)\tag{5}$$

基于以上定义，扩展形式博弈中的近似纳什均衡定义如下：

定义 9.　扩展形式博弈 ϵ-纳什均衡[25-27]. 给定扩展形式博弈 $G=\langle N,A,H,Z,(I_i,\mathcal{R}_i)_{i\in N},\chi,\rho\rangle$ 以及 $\epsilon\geqslant 0$，博弈 G 的 ϵ-纳什均衡是一个策略组合 $\sigma^*=(\sigma_1^*,\sigma_2^*,\cdots,\sigma_n^*)$. 该策略组合使得对于任意个体 $i\in N$，

$$\overline{\mathcal{R}}_i(\sigma^*)=\overline{\mathcal{R}}_i(\sigma_i,\sigma_{-i}^*)\tag{6}$$

对于任意的 $\sigma_i\in\Sigma_i$ 均成立. 当 $\epsilon=0$ 时，近似纳什均衡就是纳什均衡.

对于扩展形式博弈中纳什均衡的存在性，有如下定理：

定理 2.　扩展形式博弈纳什均衡存在性[9,42]. 给定扩展形式博弈 $G=\langle N,A,H,Z,(I_i,\mathcal{R}_i)_{i\in N},\chi,\rho\rangle$，对于任何有限扩展形式博弈($N$ 和 A 为有限集合，任意博弈结束历史 $z\in Z$ 为有限序列)都存在至少一个纳什均衡.

4.1.3　Markov 博弈

扩展形式博弈虽然考虑了博弈状态(即博弈历史)随着决策的变化，但它假设(1)博弈状态之间的转移对每个个体来说是已知的；(2)个体的决策存在明显的先后顺序. 对于个体未知的博弈状态转移以及没有明显先后决策顺序的博弈场景，需要定义新的博弈模型. 一个较为通用的博弈模型是 Markov 博弈(或随机博弈)[22,44].

定义 10.　Markov 博弈[25-27]. Markov 博弈 G 由 $\langle N,S,(A_i,\mathcal{R}_i)_{i\in N},\mathcal{Q}\rangle$ 表示，其中 $N=\{1,2,\cdots,n\}$ 表示参与博弈所有个体的集合；S 是状态集合；对于任意个体 $i\in N$，A_i 是其动作集合；$A=\times_{i=1}^{n}A_i$ 表示所有个体的联合动作空间；$\mathcal{Q}:S\times A\times S\rightarrow[0,1]$ 是状态转移函数，$\mathcal{Q}(s,\boldsymbol{a},s')$ 表示在状态 s 下所有个体执行联合动作 $\boldsymbol{a}=(a_1,a_2,\cdots,a_n)$ 后状态转移到 s' 的概率，且满足对于任意的 $s\in S$，$\sum_{s'\in S}\mathcal{Q}(s,\boldsymbol{a},s')=1$；$\mathcal{R}_i:S\times A\rightarrow\mathbb{R}$ 表示个体 i 的收益函数，$\mathcal{R}_i(s,\boldsymbol{a})$ 表示在状态 s 下所有个体执行联合动作 $\boldsymbol{a}$ 后个体 i 的收益.

当博弈只有一个状态 $S=\{s\}$ 且所有个体只执行一次动作时，Markov 博弈就退化为了标准形式博弈. 特别地，当 $n=2$ 且对于任意的博弈状态 $s\in S$ 和任意的 $\boldsymbol{a}\in A$ 都有 $\mathcal{R}_1(s,\boldsymbol{a})+\mathcal{R}_2(s,\boldsymbol{a})=0$ 时，称为二人零和 Markov 博弈.

在 Markov 博弈中，其策略的定义如下：

定义 11.　Markov 博弈策略[25-27]. 给定 Markov 博弈 $G=\langle N,S,(A_i,\mathcal{R}_i)_{i\in N},\mathcal{Q}\rangle$，定义博弈历史为状态动作序列，即 $h_t=(s_1,\boldsymbol{a}_1,\cdots,s_{t-1},\boldsymbol{a}_{t-1},s_t)$，记 t 时刻所有可能历史的集合为 H_t(其中 $H_1=\{\varnothing\}$)；对于任意玩家 i，其策略定义为

$$\sigma_i:\bigcup_{t=1}^{T_e}H_t\rightarrow\Delta(A_i)\tag{7}$$

其中 T_e 是博弈结束的时刻，$\Delta(A_i)$ 是集合 A_i 上所有概率分布的集合.

由于标准形式博弈是 Markov 博弈的一种特例，标准形式博弈中的策略可以通过令函数(7)中的 $T_e=1$ 得到. 以上的策略是定义在任何可能的博弈历史上的，是比较一般的策略定义. 在 Markov 博弈中，还有一类相对简单的策略，这类策略只取决于当前的博弈状态，被称为平稳策略(stationary strategy).

定义 12.　Markov 博弈平稳策略[25-27]. 给定 Markov 博弈 $G=\langle N,S,(A_i,\mathcal{R}_i)_{i\in N},\mathcal{Q}\rangle$，当个体 i 的策略只取决于当前状态 s_t，即

$$\sigma_i:S\rightarrow\Delta(A_i)\tag{8}$$

时，称 σ_i 是个体 i 的平稳策略.

为了评估 Markov 博弈中不同策略的表现，可以计算策略的期望(长期)收益，其中一类常用的期望收益是累积折扣收益，也被称为值函数 $V:S\times\Sigma_1\times\cdots\times$

$\Sigma_n \to \mathbb{R}$，其中 Σ_i 表示个体 i 的平稳策略集合.

定义 13. Markov 博弈值函数[25-27]. 给定 Markov 博弈 $G=\langle N,S,(A_i,\mathcal{R}_i)_{i\in N},\mathcal{Q}\rangle$，折扣因子 $\gamma\in[0,1)$，个体 i 的值函数为

$$\mathcal{V}_i(s,\sigma_i,\sigma_{-i})=\mathbb{E}_{\boldsymbol{a}_t\sim(\sigma_i,\sigma_{-i}),s_t\sim\mathcal{Q}(s_{t-1},\boldsymbol{a}_{t-1},\cdot)}\left[\sum_{t=0}^{\infty}\gamma^t\mathcal{R}_i(s_t,\boldsymbol{a}_t)\middle|s_0=s\right] \tag{9}$$

对于平稳策略，在 Markov 博弈中可以定义比一般的纳什均衡更强的解概念，称为 Markov 完美均衡. Markov 完美均衡是纳什均衡的精炼，这里简称为 Markov 博弈纳什均衡. Markov 博弈中的近似纳什均衡有如下定义：

定义 14. Markov 博弈 ϵ-纳什均衡[25-27]. 给定 Markov 博弈 $G=\langle N,S,(A_i,\mathcal{R}_i)_{i\in N},\mathcal{Q}\rangle$、折扣因子 $\gamma\in[0,1)$ 以及 $\epsilon\geqslant 0$，博弈 G 的 ϵ-纳什均衡是一个平稳策略组合 $\sigma^*=(\sigma_1^*,\sigma_2^*,\cdots,\sigma_n^*)$，该策略组合使得对于任意个体 $i\in N$ 和状态 $s\in S$，

$$\mathcal{V}_i(s,\sigma^*)\geqslant\mathcal{V}_i(s,\sigma_i,\sigma_{-i}^*)-\epsilon \tag{10}$$

对于任意的 $\sigma_i\in\Sigma_i$ 均成立. 当 $\epsilon=0$ 时，近似纳什均衡就是纳什均衡.

对于 Markov 博弈中纳什均衡的存在性，有如下定理：

定理 3. Markov 博弈纳什均衡存在性[45]. 给定 Markov 博弈 $G=\langle N,S,(A_i,\mathcal{R}_i)_{i\in N},\mathcal{Q}\rangle$ 以及折扣因子 $\gamma\in[0,1)$，对于任何有限 Markov 博弈（N、S 以及所有 A_i 均为有限集合）都存在至少一个纳什均衡.

在以上 Markov 博弈的定义中，一个基本的假设是个体在做决策时能获知当前博弈的真实状态 s. 但在很多现实情景中，博弈的真实状态是无法获知的. 对于这样的博弈可以用部分可观测 Markov 博弈（partially observable Markov games）来进行建模：

定义 15. 部分可观测 Markov 博弈[26]. 部分可观测 Markov 博弈在 Markov 博弈 $G=\langle N,S,(A_i,\mathcal{R}_i)_{i\in N},\mathcal{Q}\rangle$ 的基础上，增加了对个体观测集以及系统观测函数的定义；对于任意个体 $i\in N$，其观测集合为 $O_i=\{o_{i1},o_{i2},\cdots,o_{ik_i}\}$，所有个体的联合观测集合为 $O=\times_{i=1}^n O_i$；系统的观测函数定义为 $\mathcal{O}$：$O\times A\times S\to[0,1]$，$\mathcal{O}(o,\boldsymbol{a},s)$ 表示在联合动作 $\boldsymbol{a}$ 且系统转移到新状态 s 时，观测到 $o\in O$ 的概率.

4.2 博弈模型构建

在人机对抗中的博弈学习研究框架中，第一步是博弈模型构建，即抽取人机对抗任务的关键特征，并将其简化和抽象，使得构建的模型对应博弈论中的某一类博弈（具体请参考 4.1 节），并保证求解该博弈等价于解决原有的任务. 在构建博弈模型时，特别需要考虑的几个任务特征有：(1) 任务中能够独立决策的个体数目；(2) 个体的行动顺序；(3) 任务的信息结构；(4) 个体之间收益的关系.

对于 4.1 节中提到的几类博弈模型，它们之间的主要区别在于(2)和(3)的定义不同. 在标准形式博弈中，所有个体只决策一次，并且在决策时不知道其他个体当前决策的结果；在 Markov 博弈中，个体会决策多次，每一次决策所有个体同时进行，当决策完成后个体获得收益，博弈状态发生转移且所有个体都能获知当前博弈的最新状态；在扩展形式博弈中，个体按照指定的先后顺序进行决策，决策时个体有可能知道也有可能不知道其他个体之前的决策信息，所有个体只有在博弈结束时才能获得收益. 三类博弈模型对应的常见人机对抗任务特征请参考表 1.

表 1 博弈模型对应的常见人机对抗任务特征

博弈模型	人机对抗任务特征			
	同时/序列决策	完美/不完美信息	对称/非对称	零和/非零和
标准形式博弈	同时决策，只进行一次	不完美信息	对于任意个体 $i,j\in N,A_i=A_j$；且对于定义在 N 上的任意置换 π，$\mathcal{R}_i(\boldsymbol{a})=\mathcal{R}_{\pi(i)}(a_{\pi^{-1}(1)},\cdots,a_{\pi^{-1}(n)})$ 时，称为对称博弈[46]；否则为非对称博弈	对于任意行动组合 $\boldsymbol{a}\in A$ 都有 $\sum_{i=1}^n\mathcal{R}_i(\boldsymbol{a})=0$ 时，为零和博弈；否则为非零和博弈
扩展形式博弈	序列决策，按顺序先后进行	当所有个体的所有信息集都为单点集时，为完美信息博弈；否则为不完美信息博弈	转换为标准形式后参考标准形式博弈对称性定义	对于任意博弈结束节点 $z\in Z$ 都有 $\sum_{i=1}^n\mathcal{R}_i(z)=0$ 时，为零和博弈；否则为非零和博弈
Markov 博弈	序列决策，每次都同时进行	不完美信息	转换为标准形式后参考标准形式博弈对称性定义	对于任意博弈状态 $s\in S$ 和任意行动组合 $\boldsymbol{a}\in A$ 都有 $\sum_{i=1}^n\mathcal{R}_i(s,\boldsymbol{a})=0$ 时，为零和博弈；否则为非零和博弈

除了特征(2)和(3)之外，特征(1)和(4)需要特别考虑的原因是它们与博弈求解的难度息息相关，其中最容易求解的是二人零和博弈(表 1 中令 $n=2$). 当博弈人数大于二或者所有个体的收益之和不始终为零(称为一般和或者非零和博弈)时，博弈的求解就可能变得相当困难. 此外，二人零和博弈在理论上也具有非常良好的性质，比如使用纳什均衡策略能保证我方在面对任意对手时收益有下界，任意一对纳什均衡策略的收益都相等，等等. 因为这些性质，二人零和博弈成为了人机对抗智能技术长期以来重点关注的博弈类型. 近期人机对抗智能技术的一系列突破，如 AlphaGo、Libratus 等也都是围绕二人零和博弈开展的. 本文会在第 5 节对这些典型应用进行详细分析.

事实上，在整个机器学习领域，对抗学习已逐渐发展为一个新的学习范式[47]，该范式将包含潜在攻击方或者本身具有对抗属性的学习任务建模为二人零和博弈，并将博弈论中的纳什均衡解作为学习目标，以此来获得鲁棒的模型，比如生成对抗网络中生成器和鉴别器之间的对抗就被 Goodfellow 等人建模为具有连续动作空间的二人零和博弈[38].

4.3　解概念定义

有了博弈模型后，人机对抗中的博弈学习研究框架的第二步就是对博弈的解概念进行定义，也就是根据希望满足的性质对博弈可能出现的结果(即策略组合的集合)给出形式化约束的过程. 博弈论中最常见的解概念性质有稳定性和安全性.

稳定性. 在这个解概念性质下，任何个体单方面地改变它自己的策略都不会增加它自己的收益. 此时，任何试图最大化自身利益的个体都没有意愿单方面地改变策略，也就体现了稳定性. 4.1 节提到的纳什均衡就是定义在稳定性上的解概念.

安全性. 在这个解概念性质下，参与博弈的个体对它的对手持最悲观的态度，即认为对手总会采取最小化我方收益的策略. 在这样的情形下，个体会评估它使用任意策略时所获得的最低收益，然后在其策略集合中选择一种能使得最低收益最大化的策略，该策略被称为极大极小策略. 采用极大极小策略的个体能保证它自身的收益不低于某一个值，也就体现了安全性. 当博弈中的所有个体都采用极大极小策略的时候，就称为博弈的极大极小解.

在二人有限零和博弈中，对于任意个体来说，极大极小策略就是最优策略，并且极大极小解就是博弈的纳什均衡，也就是说在二人零和博弈中，极大极小解既具有稳定性也具有安全性.

有较好的性质只是解概念定义的第一步，接下来需要证明解概念的存在性甚至唯一性. 一个解概念有好的性质，但只存在于一些很特殊的博弈中甚至根本不存在，那么该解概念是无法被广泛采用的；一个解概念广泛存在但不唯一，此时就涉及到博弈解选择的问题，不同的选择可能会导致非常不同的后果.

4.4　博弈解计算

定义好解概念之后，人机对抗中的博弈学习研究框架的第三步是计算满足条件的博弈解，即将定义的解概念作为计算目标，对博弈模型进行求解的过程. 博弈解的计算一般通过迭代算法实现，相关算法包括传统博弈求解方法和机器学习方法. 对于不同的博弈求解算法，一个基本的概念是计算时间复杂度，该复杂度衡量了最大可能计算时间与输入长度之间的近似函数关系. 之前的工作表明，求解二人零和博弈的纳什均衡的计算时间复杂度为 P(多项式时间复杂度)，而求解二人一般和博弈的纳什均衡的计算时间复杂度则为 PPAD[48-49]. PPAD 复杂度介于 P 和 NP(非确定性多项式时间复杂度)之间(如果 P≠NP)，属于难计算的类型. 传统的博弈求解方法包括线性规划、Lemke-Howson 算法[50]、double oracle 算法[51]、min-max 搜索算法及其改进(如 alpha-beta 剪枝、蒙特卡洛树搜索等)、虚拟对局[10]、遗憾匹配(regret matching)[52]、反事实遗憾最小化[28]等. 机器学习方法包括 minimax-Q[22]、Nash-Q[31]、虚拟自我对局[29]、PSRO[5]等.

当博弈规模较大时(比如国际象棋、围棋等拥有巨大状态空间和策略空间的博弈)，最有效的求解算法也只能试图求解近似纳什均衡(即 ϵ-纳什均衡)，即使是二人零和博弈. 求解近似纳什均衡有诸多优势：首先近似纳什均衡可能是更符合实际的解概念，也就是当个体处于均衡时，只有足够大(大于 $\epsilon>0$)的激励才能促使个体改变策略；其次，有限博弈中近似纳什均衡一定存在(基于定理 1 得到)，并且任何纳什均衡附近一定存在近似纳什均衡；最后，求解任意近似纳什均衡只需要考虑有限个策略组合，而不需要遍历整个策略空间[25,53-54].

本文现介绍几类经典的求解二人零和博弈的方法，包括蒙特卡洛树搜索、虚拟对局及其扩展、遗憾最小化算法以及 minimax-Q 算法. 其他博弈模型及对

应的机器学习类求解方法的介绍,可参考综述[26-27].

4.4.1 蒙特卡洛树搜索

对于完美信息扩展形式博弈,一个常用的求解纳什均衡(更准确地说是子博弈完美均衡)的方法是逆向回溯法(backward induction)[25].逆向回溯法从博弈的结束节点开始,不断往博弈的开始节点回溯,每一次回溯都选取使得当前决策个体在剩余子博弈中期望收益最大的行动.当整个过程结束后,就得到了博弈的一个纳什均衡.

在二人零和扩展形式博弈中,按照反向回溯法计算纳什均衡的方法被称为 min-max 搜索算法.然而,min-max 搜索算法每次都需要遍历所有节点,计算代价非常高.后续的 alpha-beta 剪枝算法虽然通过剪掉明显次优的子博弈分支节省了大量开销,但同样也只适用于小规模的博弈,对于围棋这类状态空间巨大的博弈基本束手无策.这最主要的原因是这些方法都是精确方法,基于精确的子博弈期望收益进行行动选择,求解目标对应纳什均衡.

当采用近似的方法,并将求解目标定为近似纳什均衡时,每一步需要计算的子博弈期望收益就可以通过多次蒙特卡罗仿真来估计,该估计值会随着仿真次数的增加而变得更加准确;另外,在每次选择仿真的子博弈分支时也可以根据历史信息来更有效地进行,而不是每次都遍历所有可能.在这两点的基础上,再结合一定程度的探索,就是经典的蒙特卡罗树搜索(Monte Carlo tree search)算法.在蒙特卡罗树搜索中,计算机存储的博弈节点数目随着搜索的进行逐渐增加.具体地,单次搜索包含四个过程:选择、扩展、模拟与回传.选择阶段在已存储的博弈节点中进行,选定一个节点 h 作为根节点,并从节点 h 出发,按照一定的方式选择子节点 $h_{\text{next}}=(h,\alpha)$(其中 $\alpha\in\chi(h)$),直至到达一个仍有扩展空间的节点 h'.在选择过程中需要平衡探索和利用,这其中一个代表性的方法是 UCT 算法(Upper Confidence Bounds to Trees)[55],UCT 算法将子节点的选择看作一个多臂赌博机(multi-arm bandits)问题,因此可以应用处理这类问题的经典算法 UCB1(UCB 是 Upper Confidence Bounds 的缩写).UCB1 追踪多臂赌博机每一个臂的历史平均收益并在 t 时刻选择 UCB 值最大的臂 I_t,即

$$I_t=\arg\max_{i\in\{1,\cdots,K\}}\{\overline{X}_{i,T_i(t-1)}+c_{t-1,T_i(t-1)}\} \tag{11}$$

其中 $\overline{X}_{i,T_i(t-1)}$ 为第 i 臂前 $t-1$ 步的平均收益,$T_i(t-1)$ 为前 $t-1$ 步第 i 臂被选择的总次数,$c_{t,s}=\sqrt{2\ln t/s}$ 为偏置.基于此,在蒙特卡洛仿真选择阶段,UCT 算法将上述平均收益 $\overline{X}_{i,T_i(t-1)}$ 替换为待选子节点行动价值的估计值 Q(即在当前状态下执行某一行动的期望收益),将偏置 $c_{t,s}$ 修正为 $2C_p\sqrt{\ln t/s}$(其中 C_p 是一个正常数),并将 t 和 s 分别替换为当前所处节点和待选子节点的访问次数.在扩展过程中,选择节点 h' 任意一个未扩展的子节点 h'',将其作为新节点添加到计算机的存储中.在模拟阶段,从新的节点 h'' 开始进行仿真推演,直至获得最终的输赢结果 $R_i(z)$,该步骤一般被称为 rollout.在回传阶段,基于上述输赢结果来更新根节点 h 到新添加节点 h'' 的信息,包括访问次数和 Q 值.当上述过程重复的次数足够多时,UCT 算法每次选择次优行动 α 的概率会趋于零[55-56],这也就意味着基于 UCT 的蒙特卡洛树搜索算法在理论上能逼近 min-max 搜索.

4.4.2 虚拟对局

虚拟对局[10]是博弈论中最早提出的几个博弈学习方法之一,通常用来计算几类标准形式博弈的纳什均衡,比如二人零和博弈、势博弈以及一般的(generic)$2\times m$ 矩阵博弈.在虚拟对局中,博弈重复进行,在第 t 轮博弈中,任意个体 i 根据对手的历史平均策略 $\sigma_{-i,t-1}$ 来执行最优响应,即执行 $Br_i(\sigma_{-i,t-1})=\arg\max_{a_i\in A_i}\overline{\mathcal{R}}_i(a_i,\sigma_{-i,t-1})$(如果存在多个最优响应,则随机执行其中一个).记第 t 轮博弈结束时,联合平均策略为 $\sigma_t=\times_{i=1}^n\sigma_{i,t}$,最优响应联合为 $Br(\sigma_t)=\times_{i=1}^n Br_i(\sigma_{-i,t-1})$.虚拟对局按照下式来更新平均策略:

$$\sigma_t=\frac{t-1}{t}\sigma_{t-1}+\frac{1}{t}Br(\sigma_{t-1}) \tag{12}$$

随着虚拟对局的进行,当联合平均策略收敛时,该策略就是博弈的纳什均衡.此外,当博弈是二人零和博弈、势博弈或者一般的(generic)$2\times m$ 矩阵博弈时,联合平均策略一定会收敛到纳什均衡[25].

Leslie 和 Collins 在 2006 年对以上虚拟对局进行了重要扩展,提出了一般化且弱化的虚拟对局(generalised weakened fictitious play)[57].该工作的重要扩展在于两点:一是每一轮个体不再需要精确求解最优响应,这对应了“弱化”;二是每一轮在更新平均策略时允许存在扰动且更新的步长不一定是 $1/t$,这对应了“一般化”.对于近似最优响应(即 ϵ-最优响应,$\epsilon\geq 0$),其定义为 $Br_i^{\epsilon}(\sigma_{-i})=\{\sigma_i\in\Sigma_i\mid\overline{\mathcal{R}}$

$_i(\sigma_i,\sigma_{-i})\geqslant\overline{\mathcal{R}}_i(Br_i(\sigma_{-i}),\sigma_{-i})-\epsilon\}$. 在一般化且弱化的虚拟对局中，式(12)修改为

$$\sigma_t=(1-\alpha_t)\sigma_{t-1}+\alpha_t(Br^{\epsilon_{t-1}}(\sigma_{t-1})+M_t) \quad (13)$$

其中$\lim\limits_{t\to\infty}\alpha_t=0$，$\lim\limits_{t\to\infty}\epsilon_t=0$，$\sum\limits_{t=1}^{\infty}\alpha_t=\infty$，以及对于任意的$T>0$，

$$\lim_{t\to\infty}\sup_k\left\{\left\|\sum_{m=t-1}^{k-1}\alpha_{m+1}M_{m+1}\right\|\,\middle|\,\sum_{m=t-1}^{k-1}\alpha_{m+1}\leqslant T\right\}=0 \quad (14)$$

Leslie 和 Collins 同时也证明了当博弈是二人零和博弈、势博弈或者一般的 $2\times m$ 矩阵博弈时，一般化且弱化的虚拟对局最终会使得联合平均策略收敛到纳什均衡.

在一般化且弱化的虚拟对局基础上，Heinrich 等人将其进一步拓展到扩展形式博弈，提出了 XFP 算法(full-width extensive-form fictitious play)，并证明了 XFP 与一般化且弱化的虚拟对局拥有相同的收敛性保证[29]. 在此基础上，为了更高效地处理大规模博弈场景，他们提出了虚拟自我对局算法 FSP，该算法利用强化学习来求解近似最优响应，利用监督学习来进行平均策略的更新.

4.4.3　无悔学习

无悔学习的目的是使个体学习到的策略实现遗憾最小化(regret minimization). 对于标准形式博弈 G，当个体重复博弈 T 轮时，个体 i 的平均累积遗憾值：

$$Reg_i^T=\frac{1}{T}\sum_{t=1}^{T}[\max_{\sigma_i\in\Sigma_i}\overline{\mathcal{R}}_i(\sigma_i,\sigma_{-i,t})-\overline{\mathcal{R}}_i(\sigma_t)] \quad (15)$$

遗憾最小化指的是当 $T\to\infty$ 时，上式中的 Reg_i^T 以概率 1 趋于 0. 对于平均累积遗憾值，一个重要结论是：对于二人零和博弈，当博弈进行到第 t 轮时，如果两个个体的平均累积遗憾值都小于 ϵ，那么此时博弈双方的平均策略形成 2ϵ-纳什均衡[28]. 基于此结论，在二人零和博弈中，当双方都使用无悔学习算法时，最终的平均策略以概率 1 收敛到纳什均衡.

传统的无悔学习算法面向的是标准形式博弈，可以有效求解的博弈规模较小. 对于大规模二人零和扩展形式博弈，Zinkevich 等人在 2008 年提出了反事实遗憾最小化算法(Counterfactual Regret minimization，简称 CFR)[28]，此后该算法成了为扑克类不完美信息扩展形式博弈的经典求解算法. 相对于式(15)定义的平均累积遗憾，他们定义了一种新的遗憾值，称为反事实遗憾值(counterfactual regret). 为方便说明，现定义如下记号：$\beta^{\sigma}(z|h)$ 表示给定联合策略 σ 且节点 h 已到达的情况下，博弈最终到达终止节点 z 的概率；$\beta^{\sigma_{-i}}(h)$ 表示除了个体 i 之外其他个体的策略对于到达节点 h 的概率贡献. 那么，个体 i 在其信息集 I 处的反事实遗憾值为

$$Reg_{i,\mathrm{imm}}^T(I)=\frac{1}{T}\max_{a\in\chi(I)}\sum_{t=1}^{T}\beta^{\sigma_{-i,t}}(I)(u_i(\sigma_t|_{I\to a},I)-u_i(\sigma_t,I)) \quad (16)$$

其中 $u_i(\sigma_t,I)=\sum\limits_{h\in I,z\in Z}\beta^{\sigma_t}(z|h)\mathcal{R}_i(z)$ 表示信息集 I 处个体 i 的期望收益，$u_i(\sigma_t|_{I\to a},I)$ 表示当个体 i 始终在信息集 I 处执行行动 a 时得到的期望收益. 对于非负反事实遗憾值 $Reg_{i,\mathrm{imm}}^{T,+}(I)=\max\{Reg_{i,\mathrm{imm}}^T(I),0\}$，对于任意个体 i，Zinkevich 等人证明了其平均累积遗憾值不会超过其所有信息集上的非负反事实遗憾值之和，即

$$Reg_i^T\leqslant\sum_{I\in I_i}Reg_{i,\mathrm{imm}}^{T,+}(I) \quad (17)$$

因此，最小化每一个信息集上的反事实遗憾值就同时保证了平均累积遗憾值的最小化.

类似地，定义遗憾值 $Reg_i^T(I,a)=\frac{1}{T}\sum\limits_{t=1}^{T}\beta^{\sigma_{-i,t}}(I)(u_i(\sigma_t|_{I\to a},I)-u_i(\sigma_t,I))$ 以及 $Reg_i^{T,+}(I,a)=\max\{Reg_i^T(I,a),0\}$. 反事实遗憾最小化算法按照如下遗憾匹配的方式进行策略更新：

$$\sigma_{i,T+1}=\begin{cases}\dfrac{Reg_i^{T,+}(I,a)}{\sum\limits_{a\in\chi(I)}Reg_i^{T,+}(I,a)}, & \text{当}\sum\limits_{a\in\chi(I)}Reg_i^{T,+}(I,a)>0\text{ 时}\\ \dfrac{1}{|\chi(I)|}, & \text{其他情况}\end{cases} \quad (18)$$

那么，按照式(18)更新的策略就能实现遗憾最小化.

4.4.4　Minimax-Q

根据二人零和 Markov 博弈的定义，可以得到对于任意策略组合 $\sigma=(\sigma_1,\sigma_2)$，按式(9)定义的值函数对于任意的博弈状态 $s\in S$ 都满足 $\mathcal{V}_1(s,\sigma_1,\sigma_2)=-\mathcal{V}_2(s,\sigma_1,\sigma_2)$[27]. 因此，可以定义二人零和 Markov 博弈中的最优值函数为

$$\mathcal{V}^*(s)=\max_{\sigma_1\in\Sigma_1}\min_{\sigma_2\in\Sigma_2}\mathcal{V}_1(s,\sigma_1,\sigma_2) \quad (19)$$

与经典的 Q 学习[3]类似，此时可以得到最优的行动值函数 $Q_1^*(s,a_1,a_2)=\mathcal{R}_1(s,a_1,a_2)+\gamma\sum\limits_{s'\in S}\mathcal{Q}(s,a_1,a_2,s')\mathcal{V}^*(s')$. 然后对于任意的状态 $s\in S$，就可以利用最优行动值函数就能得到个体 1 的最优策略

$$\sigma_1^*(s)=\arg\max_{\sigma_1(s,\cdot)\in\Delta(A_1)}\min_{a_2\in A_2}\sum_{a_1\in A_1}\sigma_1(s,a_1)Q_1^*(s,a_1,a_2) \quad (20)$$

同理可得个体 2 的最优策略.

基于此,Littman 在 1994 年提出了 minimax-Q 学习算法[22].与 Q 学习类似,minimax-Q 在计算时不需要知道环境动力学模型,是一种无模型方法.当博弈双方在状态 s 执行行动 (a_i,a_{-i}) 且状态转移到 s' 时,minimax-Q 按照如下方式进行 Q 值的更新:

$$Q_{i,t+1}(s,a_i,a_{-i})=Q_{i,t}(s,a_i,a_{-i})+\alpha_t(\mathcal{R}_i(s,a_i,a_{-i})+\gamma\mathcal{V}_t'(s')-Q_{i,t}(s,a_i,a_{-i})) \tag{21}$$

其中值函数

$$\mathcal{V}_t'(s)=\max_{\sigma_i(s,\cdot)\in\Delta(A_i)}\min_{a_{-i}\in A_{-i}}\sum_{a_i\in A_i}\sigma_i(s,a_i)Q_{i,t}(s,a_i,a_{-i}) \tag{22}$$

当 $t\to\infty$ 且每个状态行动组合对 (s,a_i,a_{-i}) 被访问足够多(趋于无穷)时,式(21)定义的 minimax-Q 学习算法会使每个个体 i 的 Q 函数收敛到最优行动价值函数 Q_i^*,此时就可以根据式(20)得到个体的最优策略.

4.5 案例分析

为了更清楚地阐述人机对抗中的博弈学习研究框架如何解决具体问题,这里以经典工作神经虚拟自我对局(Neural Fictitious Self-Play,简称 NFSP)[39]为例进行分析.

NFSP 是一种结合了深度学习和虚拟自我对局 FSP 的求解二人零和不完美信息扩展形式博弈的博弈学习方法,该方法被应用于有限注德州扑克(信息集数目约为 10^{14},传统博弈学习方法比如 XFP 无法处理)智能体的训练中,训练得到的智能体策略能达到当时最优算法的表现[39].具体地,NFSP 在虚拟自我对局 FSP 的基础上构建了两个深度神经网络 θ^Q 和 θ^Π,一个用来近似强化学习中的行动-价值 $Q(s,a|\theta^Q)$,另一个通过监督学习的方式来近似智能体的平均策略 $\Pi(s,a|\theta^\Pi)$;前者 $Q(s,a|\theta^Q)$ 又进一步被用于计算每一步针对对手平均策略的近似最优响应(相当于利用深度强化学习求解 MDP 中的近似最优策略).两者结合在一起就形成了每一步迭代最基本的两个运算:(1)针对对手平均策略求解近似最优响应;(2)根据该最优响应及以往的最优响应历史来更新智能体的平均策略.

根据人机对抗中的博弈学习研究框架,NFSP 针对的博弈模型是二人零和不完美信息扩展形式博弈,解概念为近似纳什均衡.按照博弈学习研究框架,如果一个人机对抗任务可以被建模为二人零和不完美信息扩展形式博弈,并且解概念定义为近似纳什均衡,那么就可以利用 NFSP 计算博弈解.基于此,目前已经有学者尝试利用 NFSP 计算即时战略游戏中的博弈策略,比如 ELF Mini-RTS[58].当然,由于无悔学习方法比如反事实遗憾最小化(CFR)算法也可以用于求解二人零和不完美信息扩展形式博弈的近似纳什均衡,根据博弈学习研究框架,以上即时战略游戏理论上也可以利用 CFR 进行求解.

本质上,人机对抗中的博弈学习研究框架为研究者提供了一个模块化的思维框架,只要博弈模型构建、解概念定义和博弈解计算这三个基本步骤能相互匹配就能形成了一种解决特定人机对抗任务建模下的博弈学习算法.

5 典型应用

上一节阐述了人机对抗中的博弈学习研究框架,本节将利用该框架对当前人机对抗智能技术领域的重要工作进行分析(如表 2 所示),这些工作基本涵盖了本文介绍的几种博弈模型,包括完美信息扩展形式博弈(围棋)、不完美信息扩展形式博弈(德州扑克)以及部分可观测 Markov 博弈(星际争霸 2).各工作的具体分析如下:

表 2 人机对抗智能典型工作的博弈学习框架

算法名称	博弈模型构建(人机对抗任务→模型)	解概念定义	博弈解计算
AlphaGo Zero	围棋→二人零和完美信息扩展形式博弈	近似纳什均衡	自我对局、蒙特卡洛树搜索、深度神经网络近似
Libratus	二人德州扑克→二人零和不完美信息扩展形式博弈	近似纳什均衡	自我对局、蒙特卡洛反事实遗憾最小化算法
AlphaStar	星际争霸 1 对 1→二人零和部分可观测 Markov 博弈	近似纳什均衡	带有优先级的虚拟自我对局、监督学习、深度神经网络近似

5.1 AlphaGo Zero

AlphaGo Zero[59]是 DeepMind 开发的围棋智能体,可以不借助于人类对抗数据从零学习达到职业人类水平.按照博弈学习框架,在该应用中,围棋被建模为二人零和完美信息扩展形式博弈,博弈解概念是近似纳什均衡,博弈解计算基于蒙特卡洛树搜索算法框架并结合了深度神经网络近似和自我对局.

围棋搜索树的平均宽度与深度分别达到 250 与 150,且盘面估值函数往往不平滑.在如此巨大的搜索树下采用传统蒙特卡洛树搜索这一近乎依赖遍历的算法,难以在计算资源与时间开销有限的条件下

获得表现优异的博弈解，比如近似纳什均衡策略. 为解决该问题，在 AlphaGo Zero 中，深度神经网络和蒙特卡洛树搜索通过相互促进、互相强化的方式来实现博弈解的计算. 一方面，深度神经网络通过自我对局的方式逐渐输出更优的走子策略和更准确的盘面胜负评估. 深度神经网络的训练采用强化学习范式，蒙特卡洛树搜索在其中扮演着策略提升与策略评估的作用：策略的评估借助于蒙特卡洛树搜索对节点行为价值 Q 的估计来进行，策略的提升基于节点行为价值通过选择潜在更优的行动来实现. 另一方面，在蒙特卡洛树搜索中（包含选择、扩展、评估与回传四个过程），深度神经网络通过输出走子（即选择博弈子节点）概率和盘面胜负的估计（即依据盘面对胜率的估计）来限制搜索的宽度和深度，以实现搜索过程中的有效剪枝. 在选择阶段，博弈子节点的选择遵循 PUCT 算法[60]，该算法能有效减少树搜索的宽度，并将计算集中于有价值的子节点上. 具体来说，PUCT 算法每一步会选择一个行为价值 Q 与 U 值之和最大的子节点，其中行为价值 Q 是多次模拟后盘面胜负值的平均，U 值可以看作一个探索项，它正比于深度神经网络输出的子节点选择概率，反比于子节点的访问次数；在扩展阶段与评估阶段，则直接通过深度神经网络依据扩展节点的盘面来预测盘面胜负，而非通过直至博弈结束的完整模拟进行盘面评估，这样做既避免了博弈树的深度扩展，也提供了更准确的盘面评估；在回传阶段，则基于传统的蒙特卡洛树搜索算法利用上述胜负预测值结合节点访问次数更新节点行为值.

5.2　Libratus

冷扑大师 Libratus[30] 为卡耐基梅隆大学开发的二人无限注德州扑克智能体，在正式比赛中战胜了顶级人类选手. 按照博弈学习框架，在此应用中，二人无限注德州扑克被建模为二人零和不完美信息扩展形式博弈，博弈解概念是近似纳什均衡，博弈解计算基于反事实遗憾最小化算法并在求解过程中通过约简问题的离线求解与更细粒度的子博弈安全实时求解实现策略的优化.

Libratus 以反事实遗憾最小化算法为基本框架求解博弈的近似纳什均衡，包括三个求解模块：离线的蓝图策略模块给出约简后的博弈中基于反事实遗憾最小化算法得到的博弈策略（用于博弈的前两个阶段 preflop 和 flop，约简过程考虑下注约简与牌面约简）；实时的子博弈安全求解模块优化蓝图策略模块在博弈的后两个阶段（turn 与 river）的行动（不进行牌面约简）；蓝图策略自主提升模块依据与人类选手的对抗情况补全约简丢失的重要决策点（博弈树的分支，依据对手实际行动构建）.

具体地，蓝图策略模块首先对完整的二人无限注德州扑克博弈进行约简，约简包括下注约简与牌面约简. 前者采用一种应用无关的参数优化算法（application-independent parameter-optimization algorithm）[61] 获得局部较优的下注集合；后者则依赖一定的领域知识，将博弈后两个阶段的牌面构型分别从 5500 万种与 240 万种约简为 250 万种与 125 万种. 约简后的博弈变得可求解（约简后博弈的决策点数目远小于约简前的决策点数目），然后采用离线计算的方式通过蒙特卡洛反事实遗憾最小化算法获得约简后博弈的近似纳什均衡策略. 实时的子博弈求解模块采用安全嵌套子博弈求解算法，用于在 turn 和 river 阶段替换粗糙的蓝图策略，其中安全性体现在子博弈求解时采用了估计最大边际子博弈求解方法（Estimated-Maxmargin subgame solving），嵌套则表示每个后续的决策点都会重复上述子博弈求解算法. 通过对非蓝图行动构建独立的解，安全嵌套子博弈求解算法得到了相比于蓝图策略更优的子博弈解. 蓝图策略自主提升模块与蓝图策略模块一样，也采用了离线计算的方式，其目的在于将观察到的未出现在蓝图策略上的分支进行补全，以得到更全面的近似纳什均衡解，从而消除比赛过程中博弈策略潜在的弱点.

5.3　AlphaStar

AlphaStar[41] 为 DeepMind 开发的星际争霸智能体，在复杂不完美信息即时战略游戏星际争霸 2 中可以达到职业人类水平. 按照博弈学习框架，在此应用中，星际争霸 1 对 1 可以看作一个二人零和部分可观测 Markov 博弈，博弈解为近似纳什均衡解，博弈解计算基于虚拟自我对局算法，并在求解过程中采用联盟训练实现对手采样和保持策略多样性，以获得鲁棒的博弈策略.

具体地，原始的虚拟自我对局算法通过对历史策略的平均计算最优响应来实现近似纳什均衡的求解，AlphaStar 在此基础上进行改进，提出了带有优先级的虚拟自我对局（Prioritized Fictitious Self-Play，PFSP），将对历史策略的平均限制在对难以获胜（$f(x)=(1-x)^p$）以及能力相近（$f(x)=x(1-x)$）的对手上，其中 x 代表智能体面对对手时的胜率. 通过对抗难以获胜的对手以及能提供课程训练的能力相近的对手，带有优先级的虚拟自我对局为智能体提供了更有价值的训练信号，避免了对抗所有历史策略的平均带来的高计算开销和弱训练效果. 此

外，AlphaStar 构建了联盟训练框架，引入智能体联盟并为联盟中不同的智能体采样针对性的对手进行训练，以保证策略的多样性，进而在一定程度上克服策略非传递性带来的训练问题. 具体来说，在智能体联盟中，所有智能体都采用深度神经网络来实现决策，智能体分为三个类别：主智能体、主利用智能体和联盟利用智能体. 主智能体基于带有优先级的虚拟自我对局采样对手进行训练，来实现策略的提升；主利用智能体仅采样当前版本主智能体以实现对主智能体弱点的挖掘；联盟利用智能体采用与主智能体相同的采样方式来从整个联盟中选择智能体进行对抗，以此来发现整个联盟的弱点. 值得一提的是，在联盟训练开始之前，智能体的初始策略是采用监督学习的方式从高水平人类对抗数据（971000 场天梯 MMR 分数大于 3500 的玩家比赛的复盘，其中还包含 16000 场 MMR 分数大于 6200 的玩家获胜的复盘）中学习到的，这极大地减少了智能体训练的代价. 其中的主要原因是：完成监督学习训练后的 AlphaStar 就已经达到了钻石段位水平，位于人类玩家排名的前 16%，这样良好的策略初始化绕过了中低水平策略中常见的循环占优情形，很大程度上避免了策略训练的循环，提升了训练效率.

6 问题与挑战

从国际象棋到围棋，从二人德州扑克到六人德州扑克，再到第一人称射击游戏和即时策略游戏，人机对抗智能技术领域的研究正逐步逼近真实环境，决策场景的复杂度和信息的不完全度也在不断上升[62]. 虽然该领域在工程实践上已经取得了较大的突破，比如六人德州扑克智能体 Pluribus[63]，但理论和原理的研究上相对滞后. 如何使理论与技术并行，这是人机对抗智能技术发展中存在的关键问题，同时也是人机对抗中的博弈学习方法面临的挑战，本文从理论和应用两方面将相关开放问题与挑战总结如下：

6.1 理论挑战

非零和博弈求解目标定义. 本文介绍的大多数博弈学习方法都只在二人零和博弈中有理论保证，这一方面是因为此时博弈的极大极小解可以作为一种最优解，另外一方面是因为在二人零和博弈中求解极大极小解属于多项式时间复杂度（P），这使得复杂博弈的求解在理论上是可行的. 然而，对于二人或多人非零和博弈，纳什均衡的求解是 PPAD 完全的，该复杂度介于多项式与非确定多项式（NP）时间复杂度之间. 如果 $P \neq NP$，复杂二人或多人一般和博弈纳什均衡的求解就是非常困难的. 更为糟糕的是，在一些非零和博弈的场景下，即使博弈参与者的策略最终能够收敛到纳什均衡，在均衡下获得的收益对这些个体来说也可能不是最优的，比如囚徒困境博弈. 此外，当博弈存在多个纳什均衡时，还存在均衡选择的问题，也就是说即使个体通过迭代学习的方式得到各自的均衡策略，这些策略的联合不一定能形成纳什均衡. 事实上，均衡选择是博弈论领域长期以来的开放问题，比如子博奕完美均衡、颤抖手均衡等解概念的提出都是为了选择更“合理”的均衡[9,25]. 此外，在本文的博弈学习研究框架中，博弈论提供博弈模型和解概念，机器学习方法帮助设计高效的求解算法. 在这个框架中，博弈解指定了学习的目标，凡是无法达到指定解概念的方法就相当于失去了理论保证. 然而，由于博弈论中最重要的解概念—纳什均衡在二人非零和及多人博弈中 PPAD 的计算复杂度，以及针对纳什均衡的一些“不可能”结论，比如 Hart 等人的一系列工作说明了没有任何一种自然动力学（自然是指个体在进行策略迭代时只知道自己的收益函数，并且只能回忆起最近有限轮博弈的行动信息，同时策略表现会随着博弈的进行逐渐提升）能在任何博弈中都使个体的最终策略收敛到纳什均衡[64]，部分学者尝试从其他角度来重新看待博弈中的学习问题，比如（1）以学习方法为中心，研究合理的学习方法最终能收敛到何处，从而定义新的解概念；（2）放弃纳什均衡这一解概念[65-67]，重新定义什么是“好“的学习方法，也就是学习方法应该具备的理想性质[68]（见表 3）. 因此，在未来博弈学习的发展中，是否应该在非零和博弈或多

表 3 学习方法理想性质

性质	说明
收敛性（convergence）	面对非平稳的学习环境时，是否能收敛到平稳策略
适应性（adaptation）	面对无法提前预知的环境变化和对手行为改变，是否具有对环境变化和对手变化的适应能力
理性（rationality）	面对任意使用平稳策略的对手，是否能收敛到最优响应
无悔性（no-regret）	面对不同的对手时，是否能实现遗憾最小化
针对性最优（targeted optimality）	给定可能的对手集合，是否能针对这个集合中的任意对手收敛到最优响应
相容性（compatibility）	面对使用相同策略的对手时，是否能获得足够高的收益

人博弈中采用纳什均衡作为求解目标，以及如何根据人机对抗中的博弈学习研究框架定义新的、合适的解概念（比如在囚徒困境博弈下，一个可能的、新的解概念可以是集体最优），这是博弈学习理论研究需要解决的重要问题.

博弈学习方法的可解释性. 博弈学习方法的可解释性贯穿整个人机对抗中的博弈学习研究框架，是人机对抗智能领域面临的重要挑战. 当前人机对抗智能领域中重要工作（比如 AlphaGo 系列、DeepStack 以及 AlphaStar）的学习方法主要结合深度学习和强化学习：深度学习通过多层神经网络中的非线性映射将原始输入逐层转换为更高、更抽象的表示，当转换足够充分时，深度学习可以表征非常复杂的函数关系[69]，这样层次化的表示使得深度学习方法善于发现高维数据中的复杂结构；强化学习通过智能体与环境的交互，在不断试错、改进的方式下，可以逐渐逼近最优策略. 近期，深度学习也逐渐被应用博弈学习算法的设计中，比如神经虚拟自我对局（NFSP）. 这很大程度上缓解了传统博弈学习方法可扩展性弱的问题（因为传统博弈学习方法主要处理的是规模较小的博弈），使得此类方法可以用于计算复杂人机对抗任务的博弈解. 这其中的关键是深度神经网络的加入实现了博弈求解时相似（但不同）状态之间的泛化，从而使得复杂博弈的求解成为可能. 因此，未来深度学习与博弈学习方法的融合发展有望成为人机对抗智能技术重点发展的方向. 虽然如此，结合深度学习的强化学习或者博弈学习方法缺点也很明显，即得到的策略往往不具有可解释性，决策过程黑箱，更为严重的是深度学习的引入可能会使得原本具有理论保证的学习方法失去理论保证，并且增加算法理论分析的困难. 因此，如何设计兼具理论保证和可解释性的博弈学习算法，是未来博弈学习理论研究需要探索的重要问题.

6.2　应用挑战

多样化博弈学习测试环境构建. 目前人机对抗智能技术领域的大多数公开测试环境都是针对机器学习领域的研究设计的，很少考虑博弈构建的问题. 即使有些测试环境可以理解为某一类博弈，但由于决策场景过于复杂，在此基础上进行博弈相关的理论分析基本不可行. 另外一方面，虽然博弈论中有成熟的博弈模型，但这些模型一般规模较小、相对简单，比如博弈论中常研究的二人二策略矩阵博弈和二人重复囚徒困境等，这些模型距离现实应用太远. 因此，为了更好地促进博弈学习理论的发展和算法的设计，需要相关领域的学者们开发合适的测试环境，在环境设计时兼顾场景的真实性以及与博弈模型的相关性，使得决策环境不至于太复杂而使得博弈模型构建和理论分析不可行，也不至于过于简单而离现实应用太远，只有这样才能逐渐填补从理论到应用的鸿沟. 目前，学术界已经有一些测试环境的设计在朝着这方面努力，比如 Deepmind 团队公开的 sequential social dilemma[70-71] 和 OpenSpiel 环境[72].

大规模复杂博弈快速求解. 对于二人零和博弈，纳什均衡解具有良好的理论性质，并且也相对容易求解（相对于非零和博弈而言）. 然而，即使是二人零和博弈，随着博弈规模（即参与者数目）和复杂度（比如单次决策可选行动数目）的增加，表征博弈的输入大小也会指数增加，这极大地增加了博弈求解的难度和计算代价. 当前，博弈学习方法结合监督学习、自我对局、联盟博弈等技术已经可以实现复杂二人零和博弈解概念的计算，但该计算往往需要消耗巨大的计算资源，比如 AlphaStar 为了实现联盟博弈的训练使用了 192 个 3 代 TPU（8 核）、12 个 TPU（128 核）以及 1800 个 CPU（28 核）进行了为期 44 天的训练，该训练过程尚不包括神经网络初始化学习时监督训练的消耗. 此外，如果一些复杂博弈没有或者仅有少量的高水平人类对抗数据，此时就需要从零开始进行自我对局训练，这将进一步增加计算代价. 主要原因是在复杂博弈中如果缺少了较好的策略初始化，那么智能体可能会因为复杂博弈中策略间广泛存在的相互克制关系而陷入策略循环，无法进一步通过自我对局实现能力提升. 因此，如何设计更好的博弈学习算法，克服自我对局早期的策略循环，实现智能体能力的稳步提升，并将其与分布式训练框架有机结合，以达到减小计算资源开销和加速智能体训练的目的，是人机对抗中的博弈学习方法应用方面需要解决的重要问题.

7　总结与展望

人机对抗智能技术是人工智能发展的前沿方向，它通过人、机、环境之间的博弈对抗和交互学习研究机器智能快速提升的基础理论与方法技术. 为了更好地促进人机对抗智能技术的发展，本文通过梳理人机对抗智能技术领域的重要工作，总结了面向人机对抗任务的博弈学习研究框架，指出了博弈论和机器学习在其中发挥的作用，阐述了人机对抗中的博弈学习方法的两个组成要素和三个基本步

骤,并利用该框架分析了领域内的重要进展.与此同时,本文就当前人机对抗中的博弈学习方法面临的理论和应用难点问题进行了介绍,包括非零和博弈求解目标定义、博弈学习方法的可解释性、多样化博弈学习测试环境构建以及大规模复杂博弈快速求解.人机对抗中的博弈学习方法是人机对抗智能技术的核心,它为人机对抗智能技术领域的发展提供了方法保障和技术途径,同时也为通用人工智能的发展提供了新思路.

参 考 文 献

[1] Huang Kai-Qi, Xing Jun-Liang, Zhang Jun-Ge, et al. Intelligent technologies of human-computer gaming. Scientia Sinica Informationis, 2020, 50(4): 540-550(in Chinese)
(黄凯奇, 兴军亮, 张俊格等. 人机对抗智能技术. 中国科学: 信息科学, 2020, 50(4): 540-550)

[2] Bishop C M. Pattern Recognition and Machine Learning. New York, USA: Springer, 2006

[3] Sutton R S, Barto A G. Reinforcement Learning: An Introduction. Cambridge, USA: MIT Press, 2018

[4] Tan M. Multi-agent reinforcement learning: Independent vs. cooperative agents//Proceedings of the International Conference on Machine Learning. Amherst, USA, 1993: 330-337

[5] Lanctot M, Zambaldi V, Gruslys A, et al. A unified game-theoretic approach to multiagent reinforcement learning//Proceedings of the International Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 4190-4203

[6] Sugiyama M, Kawanabe M. Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation. Cambridge, USA: MIT Press, 2012

[7] Hernandez-Leal P, Kaisers M, Baarslag T, et al. A survey of learning in multiagent environments: Dealing with non-stationarity. arXiv preprint arXiv:1707.09183, 2017

[8] von Neumann J, Morgenstern O. Theory of Games and Economic Behavior. Princeton, USA: Princeton University Press, 2007

[9] Maschler M, Solan E, Zamir S. Game Theory. Cambridge, UK: Cambridge University Press, 2013

[10] Brown G W. Iterative solution of games by fictitious play//Activity Analysis of Production and Allocation. New York, USA: Wiley, 1951: 374-376

[11] Blackwell D. Controlled random walks//Proceedings of the International Congress of Mathematicians. Amsterdam, Netherlands, 1954: 336-338

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

[13] Hannan J. Approximation to Bayes risk in repeated play//Contributions to the Theory of Games. Princeton, USA: Princeton University Press, 1957: 97-139

[14] 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

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

[16] Turing A M. Computing machinery and intelligence. Mind, 1950, 59(236): 433-460

[17] Shannon C E. Programming a computer for playing chess. Philosophical Magazine, 1950, 41(314): 256-275

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

[19] Samuel A L. Some studies in machine learning using the game of checkers. IBM Journal of Research and Development, 1959, 3(3): 210-229

[20] Newell A, Shaw J C, Simon H A. Chess-playing programs and the problem of complexity. IBM Journal of Research and Development, 1958, 2(4): 320-335

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

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

[23] Taylor P D, Jonker L B. Evolutionary stable strategies and game dynamics. Mathematical Biosciences, 1978, 40(1-2): 145-156

[24] Fudenberg D, Drew F, Levine D K. The Theory of Learning in Games. Cambridge, USA: MIT Press, 1998

[25] Shoham Y, Leyton-Brown K. Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations. Cambridge, UK: Cambridge University Press, 2008

[26] Yang Y, Wang J. An overview of multi-agent reinforcement learning from game theoretical perspective. arXiv preprint arXiv:2011.00583, 2020

[27] Zhang K, Yang Z, Başar T. Multi-agent reinforcement learning: A selective overview of theories and algorithms//Handbook of Reinforcement Learning and Control. Cham, Switzerland: Springer, 2021: 321-384

[28] Zinkevich M, Johanson M, Bowling M, et al. Regret minimization in games with incomplete information//Proceedings of the International Conference on Neural Information Processing Systems. Vancouver, Canada, 2008: 1729-1736

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

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

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

[32] Brown N, Bakhtin A, Lerer A, et al. Combining deep reinforcement learning and search for imperfect-information games//Proceedings of the International Conference on Neural Information Processing Systems. 2020: 8715-8725

[33] Li H, Hu K, Ge Z, et al. Double neural counterfactual regret minimization//Proceedings of the International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020

[34] Tembine H. Deep learning meets game theory: Bregman-based algorithms for interactive deep generative adversarial networks. IEEE Transactions on Cybernetics, 2020, 50(3): 1132-1145

[35] Min M, Hu R. Signatured deep fictitious play for mean field games with common noise//Proceedings of the International Conference on Machine Learning. 2021: 7736-7747

[36] Qiu S, Wei X, Ye J, et al. Provably efficient fictitious play policy optimization for zero-sum Markov games with structured transitions//Proceedings of the International Conference on Machine Learning. 2021: 8715-8725

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

[38] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets//Proceedings of the International Conference on Neural Information Processing Systems. Montreal, Canada, 2014: 2672-2680

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

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

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

[42] Nash J F. Equilibrium points in n-person games. Proceedings of the National Academy of Sciences of USA, 1950, 36(1): 48-49

[43] Kuhn H W. Extensive games and the problem of information //Contributions to the Theory of Games. Princeton, NJ: Princeton University Press, 1953: 193-216

[44] Shapley L S. Stochastic games. Proceedings of the National Academy of Sciences of USA, 1953, 39(10): 1095-1100

[45] Filar J, Vrieze K. Competitive Markov Decision Processes. New York, USA: Springer, 2012

[46] Cao Z, Yang X. Symmetric games revisited. Mathematical Social Sciences, 2018, 95: 9-18

[47] Zhou Y, Kantarcioglu M, Xi B. A survey of game theoretic approach for adversarial machine learning. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2019, 9(3): e1259

[48] Papadimitriou C H. On the complexity of the parity argument and other inefficient proofs of existence. Journal of Computer and System Sciences, 1994, 48(3): 498-532

[49] Nisan N, Roughgarden T, Tardos E, et al. Algorithmic Game Theory. Cambridge, UK: Cambridge University Press, 2007

[50] Lemke C E, Howson J J T. Equilibrium points of bimatrix games. Journal of the Society for Industrial and Applied Mathematics, 1964, 12(2): 413-423

[51] 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. Washington, USA, 2003: 536-543

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

[53] Lipton R J, Markakis E, Mehta A. Playing large games using simple strategies//Proceedings of the 4th ACM Conference on Electronic Commerce. San Diego, USA, 2003: 6-41

[54] Daskalakis C, Mehta A, Papadimitriou C. Progress in approximate Nash equilibria//Proceedings of the 8th ACM Conference on Electronic Commerce. San Diego, USA, 2007: 355-358

[55] Kocsis L, Szepesvári C. Bandit based Monte-Carlo planning//Proceedings of the European Conference on Machine Learning. Berlin, Germany, 2006: 282-293

[56] Browne C B, Powley E, Whitehouse D, et al. A survey of Monte Carlo tree search methods. IEEE Transactions on Computational Intelligence and AI in Games, 2012, 4(1): 1-43

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

[58] Kawamura K, Tsuruoka Y. Neural fictitious self-play on ELF Mini-RTS. arXiv preprint arXiv: 1902.02004, 2019

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

[60] Rosin C D. Multi-armed bandits with episode context. Annals of Mathematics and Artificial Intelligence, 2011, 61(3): 203-230

[61] Brown N, Sandholm T. Regret transfer and parameter optimization//Proceedings of the AAAI Conference on Artificial Intelligence. Québec City, Canada, 2014: 594-601

[62] Yin Qi-Yue, Zhao Mei-Jing, Ni Wan-Cheng, et al. Intelligent decision making technology and challenge of wargame. Acta Automatica Sinica, 2021, 47(1): 1-15(in Chinese)
(尹奇跃，赵美静，倪晚成等. 兵棋推演的智能决策技术与挑战. 自动化学报，2021，47(1)：1-15)

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

[64] Hart S, Mas-Colell A. Simple adaptive strategies: From regret-matching to uncoupled dynamics. Singapore: World Scientific, 2013

[65] Zinkevich M, Greenwald A, Littman M. Cyclic equilibria in Markov games//Advances in Neural Information Processing Systems. Vancouver, Canada, 2006: 1641-1648

[66] Mertikopoulos P, Zhou Z. Learning in games with continuous action sets and unknown payoff functions. Mathematical Programming, 2019, 173(1): 465-507

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

[68] Busoniu L, Babuška R, De Schutter B. Multi-agent reinforcement learning: An overview//Innovations in Multi-Agent Systems and Applications-1. Berlin, Germany: Springer, 2010: 183-221

[69] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436-444

[70] Leibo J Z, Zambaldi V, Lanctot M, et al. Multi-agent reinforcement learning in sequential social dilemmas. arXiv preprint arXiv:1702.03037, 2017

[71] Vinitsky E, Jaques N, Leibo J Z, et al. An open source implementation of sequential social dilemma games. GitHub repository: https://github.com/eugenevinitsky/sequential_social_dilemma_games, 2019

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

ZHOU Lei, Ph. D., assistant professor. His research interests include evolutionary game theory, machine learning, and game-theoretic learning.

YIN Qi-Yue, Ph. D., associate professor, M. S. supervisor. His research interests include machine learning, data mining, and decision-making in games.

HUANG Kai-Qi, Ph. D., professor, Ph. D. supervisor. His research interests include computer vision, pattern recognition, and cognitive decision-making.

Background

Human-computer gaming, which tries to uncover the underlying mechanisms for machines to surpass humans in performance, is one of the research frontiers in artificial intelligence. The research focus of human-computer gaming is usually on decision-making scenarios where multiple agents are present. To achieve superhuman performance in such multi-agent scenarios, one of the approaches is to let agents learn from experience. However, when multiple agents learn, the overall learning problem becomes difficult to handle by traditional machine learning methods. The main reason is that the premise of the theoretical guarantees in traditional machine learning methods, i.e., the stationarity of environment, is violated in multi-agent decision-making scenarios. Based on this, new machine learning methods need to be developed for multi-agent scenarios.

In past decades, various multi-agent learning methods have been proposed to deal with the above problem in multi-agent learning. In particular, researchers incorporate game theory into reinforcement learning, and develop the method of multi-agent reinforcement learning based on stochastic games. Most studies thereafter in the field of human-computer gaming focus on multi-agent reinforcement learning methods. However, to achieve superhuman performance in human-computer gaming, multi-agent reinforcement learning is just one possible approach.

In this paper, to give an overview of how game theory and machine learning contribute to multi-agent learning in human-computer gaming, we provide a research framework for game theory based machine learning methods, i.e., game-theoretic learning in human-computer gaming, which summaries possible interactions between game theory and machine learning. This research framework clarifies the roles that game theory and machine learning play in game-theoretic learning, and also provides key steps to develop game-theoretic learning algorithms in human-computer gaming. Based on this research framework, recent breakthroughs in the field of human-computer gaming are also analyzed. Last but not least, we point out open problems and possible challenges based on the research framework of game-theoretic learning to guide possible future research directions in human-computer gaming.

This work is supported by the Strategic Priority Research Program of Chinese Academy of Sciences, Grant No. XDA27010103.