大规模对抗博弈理论基础与求解方法

杨耀东[1] 邓小铁[1] 程郁琨[2] 等
[1]北京大学
[2]江南大学

关键词：博弈论 强化学习 群体智能

引言

博弈，描述的是某一系统内多方参与者在复杂交互过程中形成利益矛盾，并以此为主要驱动力展开策略优化的过程。其中，对抗博弈尤指多方之间存在利益竞争关系的交互模式。对抗博弈的思想和特征根植于工业、农业、军事、国防及科学技术等多领域的发展历程中，已成为人类社会演进的核心驱动力和衡量人工智能发展的重要标准。因此，研究以对抗关系为主的博弈相关理论，并挖掘其中深刻的科学原理，将极大促进社会各行各业的发展，也将为实现通用人工智能构建宏观的组织架构提供关键的技术工具。

现阶段对抗博弈的相关研究成果已在诸多社会场景中得到应用，体现出其价值与意义。在国家层面的博弈中，各国之间的利益冲突往往需要通过合理的外交手段协调策略，使其接近均衡，从而达成各方满意的利益分配方案和博弈稳定局面。在军事战争领域，建设和升级未来信息化、智能化、海陆空一体化的国防战略体系与战术模式，尤其是赋予大规模军备设施在复杂场景中的自主博弈能力，已成为当前国防竞争力的核心体现之一；在股票交易、广告拍卖和量化风控等场景中，天然存在着多方博弈模型，多个投资者之间的利益冲突与股价波动往往存在着复杂的博弈关系。通过分析对抗博弈模型，可以更好地辅助预测市场趋势、制定有效的投资策略，从而提高企业和个体的经济效益。在市场推广与竞标投标中，善用博弈思想推动投标策略的优化，可极大提高广告商和相关平台的曝光率。在交通网络和能源规划设计中，采用博弈式方案能够在疏通车流拥堵、优化资源调度等任务中具备良好的高效性和安全性。

除了人类社会中的博弈，人类与自然界之间的博弈也广泛存在。比如，大象的生存环境往往受到人类或其他野生动物活动（例如走私动物、非法狩猎、环境退化、生存竞争等）的影响。而这些问题又涉及多个利益相关主体，如野生动物保护机构、当地政府、资源开发商等。在这种复杂的背景下，大规模对抗博弈模型能够建立各利益主体之间的博弈关系，探索物种保护的潜在矛盾与博弈空间，进而优化保护措施，提高野生动物生存环境的可持续性。正是这些实际的场景应用与深刻的科学洞察相互促进，加速了大规模对抗博弈的潜力在社会与自然双重背景下的充分释放。

对抗博弈的主要研究目标是针对具有现实意义场景下的博弈问题，从理论上刻画不同均衡解概念的性质，证明相应的计算复杂度，并在此基础上设计近似求解算法，建立完整科学的策略评估体系。

其研究难点体现在博弈场景中参与方数量庞大、合作竞争关系耦合、信息残缺、通信受限、策略空间高维及博弈结构复杂等方面（见图 1）。近年来，研究者们已尝试融合算法博弈论、控制理论、多智能体强化学习等多领域的先进成果，为大规模对抗博弈的相关研究带来了飞速的发展和新颖的研究思路。

本文将从对抗博弈的基本解概分析、均衡近似求解方法及策略量化评估三个方面出发，简要回顾大规模对抗博弈研究的主要问题及经典方法，针对目前大规模博弈中存在的挑战，着重介绍以深度强化学习为代表的人工智能技术的发展给大规模对抗博弈研究带来的新问题与新方法，并对未来研究方向进行展望。

对抗博弈基础理论

建立大规模对抗博弈的基本均衡解概念和复杂性分析是设计求解算法的基础。本节将首先围绕大规模对抗博弈的均衡理论展开描述，以团队极大极小均衡（Team Maxmin Equilibrium，TME）的解概念为基础，分别在正则式博弈和扩展式博弈中延伸出 TME 的几个重要变体均衡解概念，并进一步介绍国内外的研究现状及发展动态。

团队极大极小均衡解概念

在探究两队对抗博弈过程中，团队极大极小均衡是一类常用的均衡解概念。该概念最初由 Stengel 和 Koller 在文献 [1] 中提出，主要适用于多对一的两队对抗博弈，团队中的所有成员拥有相同的效用函数。

在正则式博弈中，TME 是指在对手采取最优对抗策略的情况下，团队采取联合策略以使团队收益最大化时达到的均衡状态。在该均衡下，团队无法通过改变策略获得更高的收益。研究 TME 的意义在于，它可以帮助团队在对抗中达到最佳平衡，使得即使对手采取最优的对抗策略，团队也不会获得

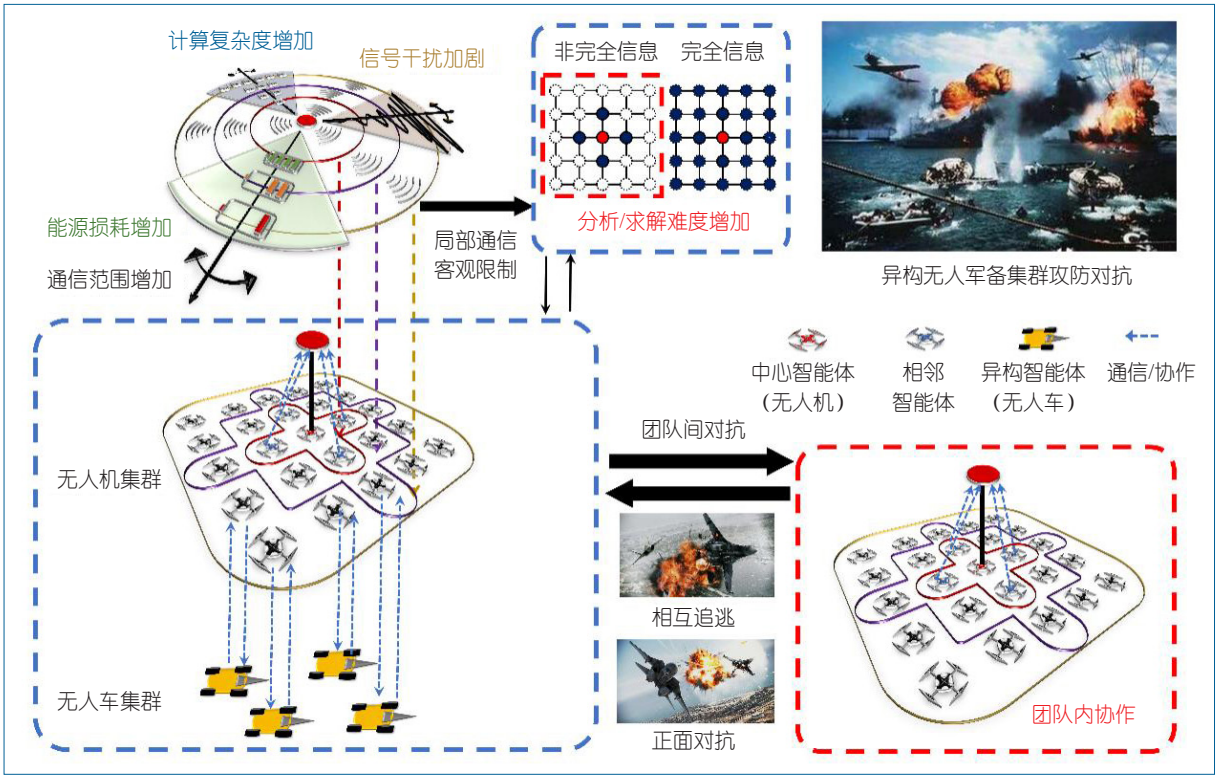


图1　大规模对抗博弈主要面向场景及困难

比 TME 更低的收益。此外，由于 TME 的团队收益通常是唯一的，避免了在多个均衡点之间进行选择的困难，使其成为研究多智能体对抗博弈的重要解概念之一。然而，计算 TME 是一项具有挑战性的任务，属于非线性规划问题。文献 [2] 首次证明了 TME 的计算复杂度为 FNP- 难（一类无法在多项式时间内验证解的函数问题，与决策类 NP- 难问题相对应，指的是函数类问题的复杂度）。通常情况下，人们利用全局优化技术 [3]，通过构建非凸规划 [1] 的方法求解 TME，但这种方法很难推广至大规模博弈。为了解决这个问题，Zhang 和 An [4] 提出了一种基于统计学习的方法，该方法可以在理论上收敛至 TME。另一方面，在对抗博弈中，TME 不满足极大极小定理 [1]，这使得在双人零和博弈纳什均衡（Nash Equilibrium，NE）计算中使用的传统技术几乎无法用于两队对抗博弈中求解 TME，也使得其不具有鞍点性质，难以进行理论分析和算法设计，因此需要引入适当的 TME 变体均衡解概念分析和求解大规模对抗博弈。

团队极大极小均衡解概念的变体

与 TME 不同，团队相关极大极小均衡（Correlated Team-Maxmin Equilibrium，CTME）[5] 满足极大极小定理，考虑了团队内参与者可以相互关联他们策略的情况，即他们可以在正则式零和对抗博弈中同步行动。当团队所有参与者被视为一个整体时，正则式零和对抗博弈可退化为双人零和博弈，CTME 也在性质上等同于纳什均衡，可使用线性规划进行计算，计算难度比 TME 低。在正则式零和对抗博弈中，设 v_{NE} 为达到纳什均衡策略时的团队收益值，v_{TME} 为达到 TME 策略时的团队收益值，v_{CTME} 为达到 CTME 策略时的团队收益值。则有：

$$v_{\mathrm{NE}} \leqslant v_{\mathrm{TME}} \leqslant v_{\mathrm{CTME}}$$

甚至在一些特殊情况下有 $\frac{v_{\mathrm{TME}}}{v_{\mathrm{NE}}} \to \infty$，$\frac{v_{\mathrm{CTME}}}{v_{\mathrm{TME}}} \to \infty$ [5]。这表示在对抗博弈中，某些纳什均衡策略可能是任意差的。

在扩展式博弈中，TME 同样是指在对手采取最优对抗策略的情况下，团队采取联合策略以使团队收益最大化时所达到的均衡状态。TME 策略可以通过非凸优化的方式计算。在扩展式博弈中，TME 同样不满足极大极小定理，不具有鞍点性质，难以进行理论分析和算法设计，因此需要引入适当的满足极大极小定理的 TME 变体均衡解概念进行分析。

因此，研究人员尝试采用事前协调和统计学习思路解决多对一对抗拓展形式博弈的 TME 问题 [6~8]。带有协同特性的团队极大极小均衡（Team-Maxmin Equilibrium with Coordination device，TMECor）[9] 也是扩展式博弈中一个重要的均衡解概念。TMECor 中允许团队内参与者在博弈开始前通过某些协作设备进行沟通，该问题在非完美信息扩展式对抗博弈下是 FNP- 难的 [6]，但是当参与者的信息不对称且有限时，文献 [10] 指出仍然存在高效算法求解。在这种情况下，团队内部所有参与者相当于一个整体，TMECor 退化为双人零和博弈中的 NE，可以通过求解线性规划计算 TMECor。

另一个重要的解概念是可通信的团队极大极小均衡（Team-Maxmin Equilibrium with Communication device，TMECom）[9]，团队内部所有参与者都可以在博弈开始前和博弈进行中通过通信设备进行实时交流，该通信设备可接收团队内参与者观察到的信息，并根据每个参与者的信息集发送与博弈策略和收益相关的建议。TMECom 也是扩展式零和对抗博弈中的纳什均衡，当团队内所有参与者被视为单个玩家时，TMECom 具有双人零和博弈中的纳什均衡属性，因此可以采用与双人零和博弈中求解纳什均衡相同的线性规划方式计算 TMECom。

在扩展式零和对抗博弈中，设 v_{TME} 为达到 TME 时团队的收益值，v_{TMECor} 为达到 TMECor 时团队的收益值，v_{TMECom} 为达到 TMECom 时团队的收益值。此时有 $v_{\mathrm{TME}} \leqslant v_{\mathrm{TMECor}} \leqslant v_{\mathrm{TMECom}}$。特别需要注意的是，在一些特殊情况下有 $\frac{v_{\mathrm{TMECom}}}{v_{\mathrm{TMECor}}} \to \infty$，$\frac{v_{\mathrm{TMECor}}}{v_{\mathrm{TME}}} \to \infty$ [9]。这表示在扩展式对抗博弈中，TMECom 是团队收益最高的纳什均衡解概念。

因此，在大规模对抗博弈中，当达到均衡时，不同均衡解的效用关系为：

$$v_{\mathrm{NE}} \leqslant v_{\mathrm{TME}} \leqslant v_{\mathrm{TMECor}} \leqslant v_{\mathrm{TMECom}}$$

表1 对抗博弈中的不同解概念性质对比

团队对抗博弈类型	通信方式	均衡解概念	凸问题	双线性鞍点问题	极大极小定理	团队收益	计算复杂度
正则式博弈	团队成员之间无通信	TME	×	×	×	低	FNP-难
	团队成员之间有通信	CTME	√	√	√	最高	NP-难
扩展式博弈	团队成员之间无通信	TME	×	×	×	低	FNP-难
	仅博弈开始前团队成员之间有通信	TMECor	√	√	√	较高	FNP-难
	博弈开始前和博弈中团队成员之间有通信	TMECom	√	√	√	最高	NP-难

为了证明 $v_{\mathrm{NE}} \leqslant v_{\mathrm{TME}}$，我们可以证明当己方采取了团队最大化最小收益策略时，达到的 NE 便是 TME。如果己方没有采取团队最大化最小收益策略，则可证明己方最坏情况下获得的收益低于采取团队最大化最小收益策略时的收益；然而此时扩展成 NE 时，敌方修改策略无法提升自己的收益，将必然采取对己方最不利的情况，使得己方收益小于 TME 的收益。而关于 TMECor 与 TME 之间的关系，不难看出，若在博弈开始前不允许队内交流，则 TMECor 与 TME 一致，因此有 $v_{\mathrm{TME}} \leqslant v_{\mathrm{TMECor}}$。类似地，如果在博弈开始后不允许队内交流，则 TMECom 等价于 TMECor，所以 $v_{\mathrm{TMECor}} \leqslant v_{\mathrm{TMECom}}$。表 1 总结了几种不同的解概念在对抗博弈中的性质，其中 TEMCor 和 TEMCom 在计算理论和博弈理论方面都具有更好的性质，满足极大极小定理，因此更适合作为分析和求解大规模对抗博弈的均衡解概念 [11]。

多对一两队对抗博弈研究进展

在现有大规模对抗博弈的研究中，尽管已经有一些针对解概念的分析及求解算法的研究，但目前关于两队对抗博弈纳什均衡的研究还未取得充分进展，已有的结果仅集中在多对一对抗博弈的情形中。文献 [12] 首次研究了多对一的两队对抗标准式博弈的纳什均衡计算问题，利用关于连续局部搜索（Continuous Local Search, CLS）– 困难复杂性的结果，证明了求解多对一两队对抗纳什均衡是 CLS- 完全的。同时，利用线性规划对偶理论，文献 [12] 也证明了团队的任意一个 ε - 近似稳定策略可以在多项式时间内扩展成一个 $O(\varepsilon)$- 近似的纳什均衡解。文献 [13] 首次讨论了多对一情况下的两队对抗马尔可夫博弈均衡问题。该文献设计了 IPGmax 算法，在每轮决策过程中，唯一的对手根据团队成员前一轮的策略组合计算最优反应，并获得其当前轮的策略；而团队成员则根据对手当前轮的最优反应，执行独立策略梯度（Independent Policy Gradient，IPG）来调整自己的策略。此外，文献 [13] 提出了一个 FPTAS 算法，用来求解多对一两队对抗马尔可夫博弈的近似纳什均衡。这些算法和方法的应用，可以帮助团队在对抗中达到最佳平衡，避免在多个均衡点之间进行选择的困难，提高整个团队的收益。

总体而言，TME 均衡虽然是最基础的均衡解概念，但其不满足极大极小定理，缺乏较好的理论分析和计算特性。因此，在实际的大规模对抗博弈中，研究者更倾向于分析和计算 TME 的相关变体，如 TMECom、TEMCor 等。

基于强化学习等方法的均衡近似求解

作为大规模对抗博弈的求解范式之一，多智能体强化学习与单智能体强化学习有诸多不同。当大量智能体在一个相同的开放式环境中共同存在时，其复杂的动态交互模式、非完全的观测机制，都在一定程度上破坏了马尔可夫决策过程的稳态性假设，使得仅以最大化累计收益为学习目标的策略求解难以实现。针对此问题，Littman 在 20 世纪 90 年代提出了以纳什均衡为学习目标的马尔可夫博弈模型 [14]，为多智能体强化学习的研究引入了新的指导框架。这一框架与近年来蓬勃发展的深度学习技术进一步融合，利用神经网络拟合不同场景下多智能体强化学习问题中的策略、价值、信息等诸多功能函数，缓解了大规模对抗博弈的分析与近似求解难

度，这一思想或将是解决高维空间复杂关系的多智能体学习问题的核心之一。

在大规模对抗博弈中，与 TME 相比，TMECom 和 TMECor 满足极大极小定理，具有更好的理论性质和计算特性，更适合作为大规模对抗博弈的均衡解概念进行求解，且具有双人零和博弈纳什均衡的特性，因此在团队之间可采用求解双人零和博弈的方法对大规模对抗博弈中 TMECom 等解概念进行近似求解，本节将介绍策略空间博弈方法及其他求解方法。

策略空间博弈方法

虽然研究者们基于马尔可夫博弈框架先后提出了诸如 MiniMax-Q 学习 [14]、Nash-Q 学习 [15] 等一系列多智能体强化学习算法，但受限于传统博弈均衡求解的计算复杂度，这些算法和模型只适用于一些简单的小规模合作和竞争任务中，难以满足真实世界中更复杂的大规模对抗博弈场景的需求。因此，自 2016 年以来，研究者们结合深度学习技术，突破了传统博弈论仅在状态 – 动作空间分析策略的瓶颈，提出了参数化策略空间的博弈求解方法，从更宏观的策略空间层面分析大规模对抗博弈的不同均衡解 [14, 16]。

本质上来看，策略空间博弈方法是一种宏观求解方法，通过在策略空间建立与原始状态 – 动作空间上等价的分析模型进行求解。此方法借助主流的自动化对手选择方法元博弈（meta-game）实现均衡分析，并可通过自适应地挑选合适的任务和对手以使智能体学习到鲁棒且多样的策略。具体来讲，参与方都会建立一个策略池供策略采样使用，在博弈的每一回合中，参与方首先对策略池进行离散采样，自动化地选择博弈对手；然后对现有对手求解最佳应对策略，扩充策略池，并在策略空间层面模拟博弈构建收益矩阵进行均衡的求解与分析。不断迭代以上过程直到策略池无可扩充的新策略，即可得到近似均衡解。这一方法可在连续高维的原始博弈的状态 – 动作空间进行高效的分析，并可扩充至大规模对抗博弈的求解中，充分降低了分析和求解的复杂度，为应对复杂场景中的大规模对抗博弈问题，构建高效的自动化对手训练框架提供了新的示范。

策略空间博弈方法具有以下几个优点：首先，策略空间博弈方法可以降低原始博弈中基于状态 – 动作空间进行分析带来的超高计算复杂度。这一优势在求解大规模对抗博弈问题时尤为突出。其次，在求解最佳响应策略的过程中，策略空间博弈方法可以灵活地引入各类策略度量作为学习目标函数之一。例如，可以引入几何度量和矩阵分析工具进一步衡量现有策略的多样性，并优化智能体对整个策略空间的探索效率。这样可以更好地处理快速多变的多类型对手策略。最后，策略空间博弈方法可充分利用策略空间博弈分析和策略池评估的结果，设计灵活的算法进行更高效的策略优化和扩展，以加速均衡的收敛。实际上，策略空间博弈方法的三个主要构成——元博弈求解、最佳响应策略求解以及策略池扩充，都可以依据不同的场景特征和问题属性进行针对性设计。这为研究者提供了灵活的可优化模块和充分的研究空间，成为针对大规模对抗博弈未来研究的课题之一。

其他求解方法

围绕策略空间博弈方法这一基础框架与范式，研究者们展开了进一步的深入探索，使用了不同的元博弈求解器 α-Rank[7] 得到均衡元策略；提出 Pipeline-PSRO[17]，通过分布式并行框架加速种群策略的训练；提出 Efficient-PSRO[18]、Anytime-PSRO[19] 自适应地构建不同的子博弈加速种群学习；借助元学习技术，进一步提出 Auto-PSRO[20]，在不依赖博弈论先验知识的前提下让神经网络学会自动地选择对手，构建子博弈并进行求解；除此之外，根据一般博弈的几何结构假说 [21]，需要策略种群具备一定的多样性才能快速跨越非传递性区域并收敛至纳什均衡，因此研究者们针对策略种群的多样性也开展了一系列的研究，如利用行列式点过程度量策略的多样性（如 DDP-PSRO[22]），将策略空间的多样性分为行为多样性和收益多样性（如 PSRO w.BD&RD[23]），并进行了策略多样性度量的统一（如 UDM-PSRO[24]）。

虽然 PSRO（Policy Space Response Oracles，策略空间预言机）框架在理论和实践中都被证明可以求解一般性的多智能博弈问题，但还是存在一些缺陷 ：随着博弈的进行，策略种群逐步扩张，新策略只能通过计算密集型的方法重复求解得到，并且大量策略并不参与最终收敛得到的混合纳什均衡，当面对大规模博弈问题时，计算复杂度和成本都巨大。因此有研究者提出 Neural PL[25, 26]，利用条件神经网络逼近策略种群，将所有策略统一为一个大策略，实现了轻量化的均衡求解以及种群策略的统一演化。其次，该框架对于非完美信息博弈仍然不能很好地进行求解，因此有研究者基于经典的非完美信息博弈求解算法框架 CFR[27] 和 Deep-CFR[28]，将 PSRO 和 CFR 的优势组件相结合提出了 CFR-PSRO[29]。进一步地，有研究者提出了一种博弈动力学修正的方法 DeepNash[30]（见图 2），其在每一个训练周期对奖励函数进行修正从而改变博弈动力学，能够快速引导策略收敛至纳什均衡。研究者成功地将其应用在大规模的复杂非完美信息博弈游戏西洋陆军棋中 [31]，该方法在降低了 PSRO 种群计算复杂度的同时为非完美信息博弈求解提供了新的思路。

总而言之，策略空间博弈方法可以将原始复杂的博弈抽象到策略空间上，进行宏观分析 ；构建自适应的对手生成和选择框架，进行灵活的策略池评估和优化，为开放式场景下大规模对抗博弈的求解提供了统一、可靠的算法方案，而在某些特殊类型的场景中，其他类型方法可以更好地进行针对性求解。

策略量化评估

从学术角度来看，对多智能体强化学习策略进行量化评估和度量，有助于研究模式的标准化，以及算法开发和设计的加速。本节将围绕这一主题，介绍针对一般博弈求解算法的策略评估标准 Elo 及针对策略空间博弈的非传递性及多样性评估标准。

阿帕德·埃洛（Arpad Elo）在 1959 年开发的 Elo 评级是最著名的通用评估标准之一，它被广泛用于足球、网球、围棋、象棋以及各类电子游戏中，可应用于本文上节提到的任何一类算法及策略的评估。Elo 评级通过为每位博弈参与者分配一个数字评级分数，反映参与者的历史博弈能力。这个评级分数可以根据历史累计胜败进行计算，并根据对手的能力进行加权组合。随着现实世界中对大规模非完全信息多智能体策略评估的需求增长，研究者们进一步基于 Elo 评级分析，设计出更具适应性的多维评价体系 [32]。除此之外，也诞生了诸多新的评价体系，如 Glicko 评级 [33]、α-Rank[34] 等。

另外，有研究者提出将真实世界中的博弈抽象为一个函数形式博弈（Functional-Form Game，FFG），并进一步分解为传递性博弈（transitive game）与非传递性博弈（non-transitive game）的组合。其中，传递性博弈可解释为 ：若有 A 战胜 B，B 战胜 C，则可得出结论 A 能战胜 C ；反之，若由“ A

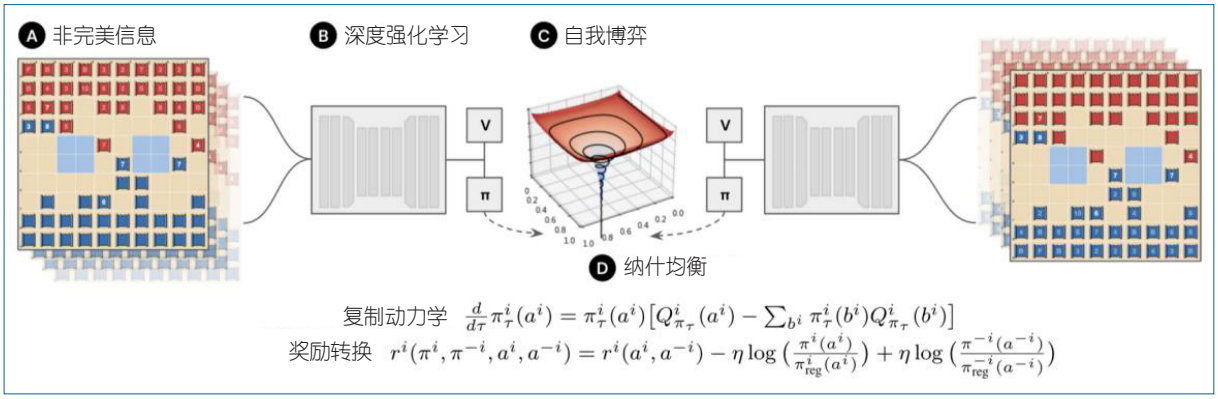


图2　DeepNash方法的整体结构[30]

战胜 B，B 战胜 C ”无法得出 A 能战胜 C，甚至出现 C 战胜 A 的情况，则称为非传递性博弈。很少有博弈是纯传递性或传非传递性的。进行策略空间的度量对理解博弈中的传递性和非传递性有重要的意义。

基于以上的博弈分解，在策略空间博弈视角下，有研究者对真实世界中博弈结构的几何进行了可视化 [21]，提出了博弈的陀螺结构假说。有研究者对国际象棋和中国象棋的博弈结构进行了度量 [35]，以此为基础分析博弈求解算法和策略空间多样性和均衡解的关系。以中国象棋游戏为例，其博弈的几何结构如图 3 左侧所示，它类似于一个旋转的陀螺结构，左上角展示了 AlphaZero 算法 [36] 在求解并探索中国象棋策略时产生的非传递性博弈实例，图中粉色区域是在博弈中非传递性最大的区域，需要较高的策略多样性才可以快速跨越从而接近纳什均衡。研究者们在 PSRO 算法的基础之上，进一步提出了 DDP-PSRO[22]，利用行列式点过程度量策略空间的多样性 ；并提出 UDM-PSRO[24] 将策略空间的多样性分为行为多样性和收益多样性，并对策略多样性的度量进行了统一。

图 3 右侧描绘了现实世界中的中国象棋游戏结构，对应于左侧的图。其中 x 轴表示纳什簇大小，反映非传递性的强度 ；y 轴表示 Elo 评级，反映传递性的强度。Elo 评级（右上角）和 RPS（Rock-Paper-Scissor，剪刀 – 石头 – 布）循环（右下角）的直方图比较了传递性程度和非传递性程度，非传递性程度可以通过 RPS 的循环数衡量。

这种分析方法可以有效地提示研究者对策略空间进行度量，并依据度量结果选择合适的算法，处理大规模对抗博弈中棘手的非传递性区域策略学习与探索。

发展趋势与展望

在多年的发展历程中，对抗博弈的研究广泛汲取了跨学科领域的研究成果，融合了理论计算机科学、控制工程、深度强化学习等多领域前沿技术，推动了一系列基础理论的进步、求解算法的发展和应用场景的扩充。特别是在大规模对抗博弈的研究中，不同子方向的研究者均借助深度神经网络强大的拟合能力取得了一定的进展。然而，该领域仍然有很多科学问题亟待解决，也是未来研究中有待重点关注的几大难题 ：

1. 如何进行大规模对抗博弈不同均衡解的性质刻画并完成计算复杂性和可学习性的分析。在大规模对抗博弈中，对建立的博弈模型进行纳什均衡性质的刻画，研究均衡是否唯一、是否多项式可计算以及可学习性如何等，是设计高效求解均衡的强化

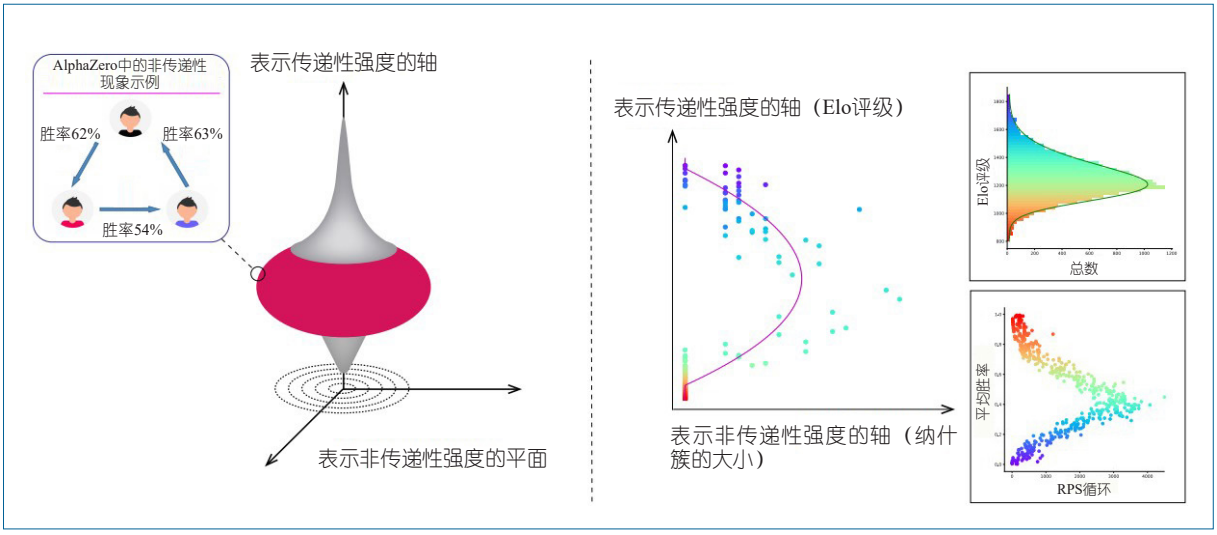
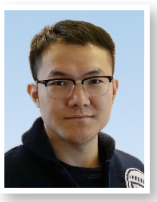


图3　中国象棋游戏中的策略空间几何结构

学习算法的关键。在刻画纳什均衡的过程中，需要解决团队内部奖励设计、策略多样性等问题，以实现既定均衡目标。

2. 非完全信息条件下大规模智能体对抗博弈的近似纳什均衡高效求解。在实现深度强化学习算法时，需要考虑如何在非完全信息博弈中构建高效的策略种群，快速跨越非传递性区域；如何在通信受限的场景中构建自适应的动态通信网络，实现有限信息价值的最大化利用；如何针对异构多智能体系统进行合理的分层信用分配，实现单调改进的策略学习模式，实现智能体间高效的协作和对抗。

3. 如何针对大规模对抗博弈研究领域构建统一的标准评价体系。针对大规模对抗博弈的研究领域缺乏标准评价体系，难以衡量所设计算法的优劣和策略空间的多样性等问题，需要构建统一的标准评价体系。该评价体系需要考虑在策略空间层面评估求解策略的多样性、鲁棒性和泛化性等性质，使该领域的研究模式标准化。同时，该评价体系需要能够针对不同博弈类型的不同结构针对性地选择求解算法，从而更好地指导算法的设计与开发。 ■

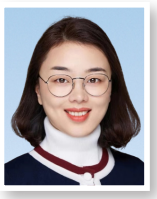


杨耀东

CCF专业会员。北京大学人工智能研究院助理教授，伦敦国王大学客座助理教授。主要研究方向为强化学习、博弈论和多智能体系统。yaodong.yang@pku.edu.cn

邓小铁

CCF专业会员。北京大学讲席教授。ACM/IEEE Fellow，欧洲科学院院士。主要研究方向为算法及博弈论、互联网经济、在线算法、并行计算。xiaotie@pku.edu.cn



程郁琨

CCF专业会员。江南大学教授。主要研究方向为区块链技术与应用、计算经济学、算法博弈论和组合优化等。ykcheng@amss.ac.cn

其他作者：马成栋

参考文献

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

[2] Borgs C, Chayes J, Immorlica N, et al. The myth of the folk theorem[J]. *Games and Economic Behavior*, 2010, 70(1):34-43.

[3] Zhang Y, An B. Converging to team-maxmin equilibria in zero-sum multiplayer games[C]// *International Conference on Machine Learning*. 2020.

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

[5] Basilico N, Celli A, De Nittis G, et al. Team-maxmin equilibrium: efficiency bounds and algorithms[C]// *Proceedings of the AAAI Conference on Artificial Intelligence*. 2017, 31(1).

[6] Chu F, Halpern J. On the np-completeness of finding an optimal strategy in games with common payoffs[J]. *Int. J. Game Theory,* 2001, 30(1):99-106.

[7] Omidshafiei S, Papadimitriou C, Piliouras G, et al. α-rank: Multi-agent evaluation by evolution[J]. *Scientific Reports*, 2019, 9(1): 9937.

[8] Jin C, Liu Q, Wang Y, et al. V-Learning A Simple, Efficient, Decentralized Algorithm for Multiagent RL[C]// ICLR2022.

[9] Celli A, Gatti N. Computational results for extensive-form adversarial team games[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32(1).

[10]Zhang Y, An B, Cerný J. Computing ex ante coordinated team-maxmin equilibria in zero-sum multiplayer extensive-form games[C]// National Conference on Artificial Intelligence. 2021.

[11]Farina G. Game-Theoretic Decision Making in Imperfect-Information Games Learning Dynamics, Equilibrium Computation, and Complexity[J].

[12]Yakov Babichenko and Aviad Rubinstein. Settling the complexity of Nash equilibrium in congestion games. STOC2021.

[13]Fivos Kalogiannis, Ioannis Anagnostides, Ioannis Panageas, et al. Efficiently Computing Nash Equilibria in Adversarial Team Markov Games. arXiv:2208.02204.

[14]Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, et al. A unified game-theoretic approach to multiagent reinforcement learning. NeurIPS2017.

[15]Junjing Hu, Michael Wellman. Nash Q-learning for

general-sum stochastic games. Journal of machine learning research, 1039-1069, 2003.
[16]Max Jaderberg, Valentin Dalibard, Simon Osindero, et al. Population Based Training of Neural Networks. arXiv:1711.09846.
[17]McAleer S, Lanier J B, Fox R, et al. Pipeline psro: A scalable approach for finding approximate nash equilibria in large games[J]. Advances in neural information processing systems, 2020, 33: 20238-20248.
[18]Zhou M, Chen J, Wen Y, et al. Efficient Policy Space Response Oracles[J]. arXiv preprint arXiv:2202.00633, 2022.
[19]McAleer S, Wang K, Lanctot M, et al. Anytime optimal psro for two-player zero-sum games[J]. arXiv preprint arXiv:2201.07700, 2022.
[20]Feng X, Slumbers O, Wan Z, et al. Neural auto-curricula in two-player zero-sum games[J]. Advances in Neural Information Processing Systems, 2021, 34: 3504-3517.
[21]Wojciech Czarnecki, Gauthier Gidel, Brendan Tracey, et al. Real world games look like spinning tops. NeurIPS2020.
[22]Perez-Nieves N, Yang Y, Slumbers O, et al. Modelling behavioural diversity for learning in open-ended games[C]//International Conference on Machine Learning. PMLR, 2021: 8514-8524.
[23]Liu X, Jia H, Wen Y, et al. Unifying behavioral and response diversity for open-ended learning in zero-sum games[J]. arXiv preprint arXiv:2106.04958, 2021.
[24]Liu Z, Yu C, Yang Y, et al. A Unified Diversity Measure for Multiagent Reinforcement Learning[C]//Advances in Neural Information Processing Systems.
[25]Liu S, Marris L, Hennes D, et al. NeuPL: Neural population learning[J]. arXiv preprint arXiv:2202.07415, 2022.
[26]Liu S, Lanctot M, Marris L, et al. Simplex Neural Population Learning: Any-Mixture Bayes-Optimality in Symmetric Zero-sum Games[C]//International Conference on Machine Learning. PMLR, 2022: 13793-13806.
[27]Zinkevich M, Johanson M, Bowling M, et al. Regret minimization in games with incomplete information[J]. Advances in neural information processing systems, 2007, 20.
[28]Brown N, Lerer A, Gross S, et al. Deep counterfactual regret minimization[C]//International conference on machine learning. PMLR, 2019: 793-802.
[29]Wang X, Cerny J, Li S, et al. A unified perspective on deep equilibrium finding[J]. arXiv preprint arXiv:2204.04930, 2022.
[30]Perolat J, Munos R, Lespiau J B, et al. From Poincaré recurrence to convergence in imperfect information games: Finding equilibrium via regularization[C]// International Conference on Machine Learning. PMLR, 2021: 8525-8535.
[31]Perolat J, De Vylder B, Hennes D, et al. Mastering the game of Stratego with model-free multiagent reinforcement learning[J]. Science, 2022, 378(6623): 990-996.
[32]David Balduzzi, Karl Tuyls, Julien Perolat, et al. Re-evaluating evaluation.NerIPS2018.
[33]Johannes Heinrich and David Silver. Deep reinforcement learning from self-play in imperfect-information games. arXiv:1603.01121, 2016.
[34]G Brown. Iterative solution of games by fictitious play. Act. Anal. Prod Allocation, 13(1): 374, 1951.
[35]Ricky Sanjaya, Jun Wang, Yaodong Yang. Measuring the non-transitivity in chess. Algorithms, 15(5): 152, 2022.
[36]Hammer P. Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm[J]. 2018.