# 多智能体博弈、学习与控制

**Article** · February 2025

DOI: 10.16383/j.aas.c220680



**2 authors:**

Long Wang
Peking University
**824** PUBLICATIONS **28,035** CITATIONS
SEE PROFILE

Feng Huang
**11** PUBLICATIONS **62** CITATIONS
SEE PROFILE

# 多智能体博弈、学习与控制

王 龙 [1, 2] 黄 锋 [1]

**摘 要** 近年来, 人工智能 (Artificial intelligence, AI) 技术在棋牌游戏、计算机视觉、自然语言处理和蛋白质结构解析与预测等研究领域取得了众多突破性进展, 传统学科之间的固有壁垒正在被逐步打破, 多学科深度交叉融合的态势变得越发明显. 作为现代智能科学的三个重要组成部分, 博弈论、多智能体学习与控制论自诞生之初就逐渐展现出一种“你中有我, 我中有你” 的关联关系. 特别地, 近年来在 AI 技术的促进作用下, 这三者间的交叉研究成果正呈现出一种井喷式增长的态势. 为及时反映这一学术动态和趋势, 本文对这三者的异同、联系以及最新的研究进展进行了系统梳理. 首先, 介绍了作为纽带连接这三者的四种基本博弈形式, 进而论述了对应于这四种基本博弈形式的多智能体学习方法; 然后, 按照不同的专题, 梳理了这三者交叉研究的最新进展; 最后, 对这一新兴交叉研究领域进行了总结与展望.

**关键词** 博弈论, 多智能体学习, 控制论, 强化学习, 人工智能

**引用格式** 王龙, 黄锋. 多智能体博弈、学习与控制. 自动化学报, 2023, **49**(3): 1−34

**DOI** 10.16383/j.aas.c220680

## An Interdisciplinary Survey of Multi-agent Games, Learning, and Control

WANG Long[1, 2] HUANG Feng[1]

**Abstract** In recent years, along with some ground-breaking advances made by artificial intelligence (AI) in Go, chess, video games, computer vision, natural language processing, and the analysis and prediction of protein structures, the inherent barriers of traditional disciplines are gradually being broken, and a cross-discipline wave is steadily underway in academia. As three important components of modern intelligent science, game theory, multi-agent learning, and control theory have witnessed a closely interrelated relationship from the very beginning of their establishments. Especially, with the aid of the great development of AI technologies in recent years, their interactions are becoming closer and closer, and the relevant interdisciplinary research is showing a blowout growth. To reflect this trend, in this paper, we provide a comprehensive survey of the connections, distinctions, and latest interdisciplinary research progress of games, multi-agent learning, and control. Specifically, we first introduce four different types of games, which are normally used to connect these three fields. Then, corresponding to these four types of games, some multi-agent learning methods are reviewed. Subsequently, following different research topics, we survey the latest interdisciplinary research progress of games, learning, and control. Finally, we provide a summary and an outlook for this emerging interdisciplinary field.

**Key words** Game theory, multi-agent learning, control theory, reinforcement learning, artificial intelligence

**Citation** Wang Long, Huang Feng. An interdisciplinary survey of multi-agent games, learning, and control. *Acta Automatica Sinica*, 2023, **49**(3): 1−34

近年来, 随着人工智能 (Artificial intelligence, AI) 技术的飞速发展, 博弈论在社会智能[1−2]、机器智能[3]、合作智能[4]、AI 安全[5−6] 和 AI 伦理[7−8] 等新兴交叉研究领域中扮演着越来越重要的角色. 特别地, 通过结合多智能体学习与控制论等理论方法, 博弈论已成为 AI 和自动控制领域中的一个热点研究方向.

从本质上讲, 所谓博弈论就是一类研究理性智能体 (Agent)[1] 之间策略交互的数学理论与方法[9]. 它是现代数学的一个分支, 也是运筹学的一个重要组成部分. 虽然博弈论的早期思想可以追溯到公元前, 但其作为一个独立的研究领域正式诞生的标志是 von Neumann 和 Morgenstern 于 1944 年合著的《博弈论与经济行为》(*Theory of Games and Economic Behavior*) 一书[10]. 作为多智能体系统或者分布式 AI (Distributed AI) 的一个研究主

收稿日期 2022-08-29 录用日期 2022-12-01
Manuscript received August 29, 2022; accepted December 1, 2022
国家自然科学基金 (62036002) 资助
Supported by National Natural Science Foundation of China (62036002)
本文责任编委 孙健
Recommended by Associate Editor SUN Jian
1. 北京大学系统与控制研究中心 北京 100871 2. 北京大学人工智能研究院 北京 100871
1. Center for Systems and Control, Peking University, Beijing 100871 2. Institute for Artificial Intelligence, Peking University, Beijing 100871

[1] 在本文中, 智能体也称为博弈者 (Player) 或决策者 (Decision-maker).

题[11−13], 多智能体学习主要研究多个智能体交互的策略学习问题. 从其发展历程上讲, 多智能体学习与博弈论几乎具有一样长的历史. 例如, 早在 1951 年, 文献 [14] 就提出了一类称为虚拟对弈 (Fictitious play) 的学习方法用于求解博弈的 Nash 均衡问题. 尔后, 在上世纪 80 年代后期, 伴随着学术界对演化计算、社会学习、交互学习和多智能体场景下的强化学习产生的广泛兴趣, 多智能体学习的研究开始在 AI 领域中兴起[12, 15]. 特别地, 近年来借助多智能体学习在棋牌[16−18] 和视频游戏[19−21] 等特定任务上取得的突破性进展, 这一方向在 AI 领域中再次掀起了热潮, 成为 AI 研究的核心内容之一. 与博弈论的发展历程相类似, 虽然人类进行自动控制的生产实践最早可以追溯到公元前[22], 但控制论作为一个独立的研究领域正式诞生的标志是 Wiener 于 1948 年撰写的《控制论》(*Cybernetics*) 一书[23]. 通过对比讨论 “动物智能” 与 “机器智能” 中的若干重要问题, Wiener 指出“智能的首要问题是‘学习’”[24]. 综上可见, 虽然博弈论、多智能体学习与控制论分属不同的研究领域, 但究其发展根源和轨迹, 它们从来都不是相互割裂的, 而是紧密关联、相互融合的, 展现出一种 “你中有我, 我中有你” 的景象.

反映到具体的研究中, 博弈论、多智能体学习与控制论的融合通常包含在三种不同的场景设置中. 第一种设置为智能体之间完全合作. 在该设置下, 连接博弈论、多智能体学习与控制论的一种典型博弈形式是团队博弈 (Team game)[25−26], 即所有博弈者具有一个相同收益函数的博弈形式. 在博弈论中, 团队博弈最初用于研究组织问题[25]. 而在多智能体学习中, 这类博弈主要用于处理多智能体的合作序贯决策问题, 也为发展合作型多智能体强化学习算法提供模型框架[27−28]. 考虑到分布式控制与团队博弈的信息结构问题[29] 具有诸多相似之处, 在控制论中, 团队博弈主要用于分析团队决策问题[26]. 如果博弈的所有博弈者具有相同的 (全局) 信息结构, 团队博弈中的决策问题则可以转化成一个传统的集中式控制问题; 而如果每个博弈者具有不同的 (局部) 信息结构, 团队博弈的决策问题则可以转化成一个多智能体合作控制问题或者分布式控制问题[30−31]. 虽然团队博弈最早提出于上世纪 50 年代[25], 但时至今日, 它在控制领域中仍然是一个重要的研究主题[32−34]. 第二种设置为智能体之间完全竞争. 在该设置下, 连接博弈论、多智能体学习与控制论的一种典型博弈形式是零和博弈 (Zero-sum game), 即两个博弈者具有零和收益关系 (一个博弈者的收益是另一个博弈者的损失) 的博弈形式. 在博弈论中, 零和博弈主要用于研究具有完全对立目标的博弈者间的决策问题. 而在多智能体学习中, 这类博弈主要用于处理多智能体的竞争序贯决策问题, 也为发展基于极大极小原理的多智能体强化学习算法提供模型框架[35]. 相比而言, 在控制论中, 零和博弈常用于处理含不确定因素的系统控制 (鲁棒控制) 问题[36−37]. 在该类问题中, 控制器 (Controller) 通常被视为一个最大化某一特定性能指标的博弈者; 而系统的不确定性 (如干扰、噪声) 则被视为另一个博弈者, 其目标是使控制器所最大化的性能指标最小化. 第三种设置为智能体之间既不完全合作又不完全竞争, 即混合设置. 在该设置下, 多智能体系统所形成的博弈是一个一般和博弈 (General-sum game). 对于这类博弈, Nash 均衡一般是标准的解[38]. 在多智能体学习中, 这类博弈通常用于处理一般化的多智能体序贯决策问题, 也为发展混合型的多智能体强化学习算法提供模型框架[39−40]. 相比而言, 在控制论中, 这种设置下的博弈者通常被视为是控制器, 而博弈者的策略被视为是控制律 (Control law). 为了实现一个特定的任务目标, 比如收敛到一个 Nash 均衡, 博弈者的策略更新规则或学习算法通常需要进行额外设计[30, 41].

由于学科发展等种种原因, 博弈论和自动控制的研究对象曾经一度存在一些差异. 但借助 AI 和多智能体学习等技术, 它们之间的差异如今正在慢慢变小. 博弈论的研究对象一般是理性的“智能体”, 或者是具有智能的“生命体”, 比如人和动物等[9−10]; 而自动控制的研究对象一般是“机器”, 或者是无生命的“物理对象”, 比如机器人和航空航天器等[22, 42]. 然而, 近年来在 AI 技术和信息技术的推动作用下, 传统无生命的物理对象通过机器学习等方法正在逐渐被赋予如生命体一样的智能性. 与此同时, 自动控制的研究对象也在从单纯的物理系统逐步地转向机器、人与社会等更为复杂的融合交互系统[43]. 在这一全新的交互系统中, 机器不再被视为是一种无生命的物理对象, 而是作为一种智能的载体广泛地参与到人类社会的各种交互之中, 并呈现出一种人与人、人与机器、机器与机器的混合交互景象[1−4]. 然而, 这一全新的交互系统在促进人类社会发展的同时, 也将给人类带来一些新的挑战, 比如伦理问题和安全问题等. 考虑到这些新的挑战本身大部分是由多种研究对象所引发的, 比如人这类对象可能涉及到博弈论和社会学等, 机器这类对象可能涉及到控制论和机器学习等, 所以单一的学科或研究领域都会或多或少地存在着一些不足. 为此, 博弈论、多智能体学习与控制论的交叉融合有望在这方面发

挥重要作用. 一方面, 它们交叉融合的本身可以促进各单一研究领域的发展; 另一方面, 它们涵盖的广泛理论体系可以为这一全新交互系统提供恰当的分析方法和研究工具.

当前无论是在博弈论领域、AI 领域还是在自动控制领域, 博弈与多智能体学习的交叉[13, 44−50]、博弈与控制的交叉[30, 37, 51−53]、以及多智能体学习与控制的交叉[54−56] 都是前沿的热点研究方向, 并且在这些主题下的相关工作和进展正呈现出一种井喷式增长的态势[57]. 然而, 据作者所知, 目前国内外已有的综述性文献主要集中讨论了这三者中的某两个特定领域[24, 30, 44, 47−52, 54−55, 58], 还没有文献宏观地从这三者的角度对它们的联系、区别以及最新的交叉研究成果进行全面的审视与梳理. 本文的主要目的是试图填补这一空白, 核心内容主要分为 3 节: 第 1 节主要介绍并讨论作为纽带连接博弈论、多智能体学习与控制论的四种基本博弈形式, 即标准式博弈 (Normal-form game)、演化博弈 (Evolutionary game)、随机博弈 (Stochastic game) 和不完全信息博弈 (Incomplete-information game); 第 2 节主要论述对应于这四种基本博弈形式的多智能体学习方法, 即策略学习 (Strategic learning)、学习动力学 (Learning dynamics)、强化学习 (Reinforcement learning) 和鲁棒学习 (Robust learning); 第 3 节按照不同研究专题梳理并介绍当前博弈、学习与控制的几类典型交叉研究成果. 最后, 针对这一重要前沿交叉研究领域给出总结与展望.

## 1 基本博弈形式

本节主要介绍作为纽带连接博弈论、多智能体学习与控制论的四种基本博弈形式—标准式博弈、演化博弈、随机博弈和不完全信息博弈, 并讨论它们之间的联系与区别.

### 1.1 标准式博弈

标准式博弈[10], 也称为策略式博弈 (Strategic-form game)[9, 59−60], 是描述博弈的一种基本形式. 一般地, 一个标准式博弈通常由以下三个基本要素构成: 1) 博弈者集, 即博弈的所有参与者构成的集合; 2) 博弈者的可行策略集; 3) 博弈者的收益函数 (也称为效用函数) 或者成本函数.[2] 因此, 在数学上, 一个标准式博弈 $\mathcal{G}$ 通常可以表示为一个三元组

$$\mathcal{G} = \{\mathcal{N}, \{\Omega_i\}_{i\in\mathcal{N}}, \{c_i\}_{i\in\mathcal{N}}\}, \tag{1}$$

其中 $\mathcal{N}$ 表示所有博弈者的标号构成的集合, $\Omega_i$ 表示博弈者 $i \in \mathcal{N}$ 的策略集, $c_i : \boldsymbol{\Omega} \to \mathbb{R}$ 表示博弈者 $i \in \mathcal{N}$ 的收益函数, $\boldsymbol{\Omega} := \prod_{i\in\mathcal{N}} \Omega_i$ 表示所有博弈者策略组合 (Strategy profile) 的集合. 如果集合 $\mathcal{N}$ 和 $\Omega_i$, $\forall i \in \mathcal{N}$ 是有限的 (Finite), 则称标准式博弈 $\mathcal{G}$ 是有限的. 特别地, 如果一个标准式博弈只有两个博弈参与者, 即 $\mathcal{N} = \{1, 2\}$, 则称其为两人博弈 (Two-player game). 而如果该标准式博弈具有多个博弈参与者, 则称其为多人博弈 (Multi-player game). 另外, 在标准式博弈进行过程中, 博弈者的策略选择一般是“同时的” (Simultaneous-move). 但是“同时” 并不完全意味着所有博弈者的策略选择是在同一个时间点上完成的, 而更多地强调每个博弈者在策略选择过程中并不知道其他博弈者的策略.

另外, 在一些特定的情形下, 标准式博弈的一般化形式 (1) 可以进一步简化. 例如, 对于一个两人博弈 $\mathcal{G}_m$, 其通常可以用如下的一组收益矩阵来刻画

$$\mathcal{G}_m = (A, B),$$

其中 $A = (\mathfrak{a}_{kl}) \in \mathbb{R}^{p\times q}$ 和 $B = (\mathfrak{b}_{kl}) \in \mathbb{R}^{p\times q}$ 分别表示博弈者 1 和博弈者 2 的收益矩阵, $p$ 和 $q$ 分别表示博弈者 1 和博弈者 2 的可行策略的数目. 对于收益矩阵对 $(A, B)$, 它的一种更为形象的展示方式是将其写成如下的一张收益表

| | $\pi_2^1$ | $\cdots$ | $\pi_2^q$ |
|---|---|---|---|
| $\pi_1^1$ | $(\mathfrak{a}_{11}, \mathfrak{b}_{11})$ | $\cdots$ | $(\mathfrak{a}_{1q}, \mathfrak{b}_{1q})$ |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $\pi_1^p$ | $(\mathfrak{a}_{p1}, \mathfrak{b}_{p1})$ | $\cdots$ | $(\mathfrak{a}_{pq}, \mathfrak{b}_{pq})$ |

其中 $\pi_1^k$ 和 $\pi_2^l$ 分别表示博弈者 1 (行博弈者) 的第 $k$ ($k = 1, 2, \cdots, p$) 个策略和博弈者 2 (列博弈者) 的第 $l$ ($l = 1, 2, \cdots, q$) 个策略, $\mathfrak{a}_{kl} = c_1(\pi_1^k, \pi_2^l)$ 和 $\mathfrak{b}_{kl} = c_2(\pi_1^k, \pi_2^l)$ 分别表示博弈者 1 和博弈者 2 在策略组合 $(\pi_1^k, \pi_2^l)$ 下的收益值. 特别地, 如果博弈 $\mathcal{G}_m$ 中的两个博弈者的角色是等同的 (行和列的角色可交换), 即 $p = q$ 且 $A = B^T$, 则称 $\mathcal{G}_m$ 是对称的 (Symmetric), 此时的 $\mathcal{G}_m$ 也称为矩阵博弈 (Matrix game); 否则, 则称博弈 $\mathcal{G}_m$ 是非对称的 (Asymmetric), 此时的 $\mathcal{G}_m$ 也称为双矩阵博弈 (Bi-matrix game). 由该定义可看出, 当 $\mathcal{G}_m$ 是对称的, 它可简单地用收益矩阵 $A$ 表示. 根据 $A$ 中元素大小的不同, 博弈 $\mathcal{G}_m$ 又可进一步分为多种类型. 例如, 表 1 以收益值 $\mathfrak{a}_{kl}$ ($k, l \in \{1, 2\}$) 的大小关系为依据列举了几类典型的两人两策略对称博弈.

对于一个标准式博弈 $\mathcal{G}$, 如果该博弈只进行一次, 则称其为单次博弈 (One-shot game); 否则, 如果该博弈重复地进行多次, 则称其为重复博弈 (Re-

[2] 在理性人的假设下, 博弈者最大化收益函数等价于最小化成本函数.

表 1 几类典型的两人两策略对称博弈

Table 1 Some representative examples of two-player two-strategy symmetric games

| 博弈类型 | 收益值大小关系 | 博弈类型 | 收益值大小关系 |
|---|---|---|---|
| 囚徒困境 (Prisoner′s dilemma) | $\begin{cases}\mathfrak{a}_{21}>\mathfrak{a}_{11}>\mathfrak{a}_{22}>\mathfrak{a}_{12}\\ 2\mathfrak{a}_{11}>\mathfrak{a}_{12}+\mathfrak{a}_{21}\end{cases}$ | 和谐博弈 (Harmony game) | $\mathfrak{a}_{11}>\mathfrak{a}_{21},\mathfrak{a}_{12}>\mathfrak{a}_{22}$ |
| 雪堆博弈 (Snowdrift game) | $\mathfrak{a}_{21}>\mathfrak{a}_{11},\ \mathfrak{a}_{12}>\mathfrak{a}_{22}$ | 猎鹿博弈 (Stag-hunt game) | $\mathfrak{a}_{11}>\mathfrak{a}_{21},\ \mathfrak{a}_{22}>\mathfrak{a}_{12}$ |

peated game). 另外, 根据博弈进行的总次数 $T_{tot}$, 重复博弈又可分为有限次数 (Finite horizon) 重复博弈 ( $T_{tot}<\infty$ ) 和无限次数 (Infinite horizon) 重复博弈 ( $T_{tot}=\infty$ ). 除非特别说明, 本文讨论的博弈都为无限次数的重复有限博弈 (博弈者数和策略数均为有限).

一般地, 在一个标准式博弈中, 所有博弈者通常均被设定为完全理性的 (Perfectly rational), 并且该设定是所有博弈者的共同知识 (Common knowledge). 换句话讲, 在标准式博弈中, 每个博弈者的目标都是追求自身利益最大化或者成本最小化, 并且所有博弈者都共同地知道这一点. 由于在该设定下, 博弈者之间的利益关系是非合作的, 所以该类博弈也称为非合作博弈 (Non-cooperative game)[38, 61]. 对于非合作博弈, 它的一个标准的解概念是所谓的 Nash 均衡[38, 61].

**定义 1.** 对于有限的 $n$ 人标准式博弈 $\mathcal{G}=\{\mathcal{N},\{\Omega_i\}_{i\in\mathcal{N}},\{c_i\}_{i\in\mathcal{N}}\}$, $\mathcal{N}=\{1,2,\cdots,n\}$, 如果对任意的博弈者 $i\in\mathcal{N}$, 其策略 $\pi_i^*$ 是对所有其他博弈者 $-i:=(1,\cdots,i-1,i+1,\cdots,n)$ 的策略组合 $\pi_{-i}^*:=(\pi_1^*,\cdots,\pi_{i-1}^*,\pi_{i+1}^*,\cdots,\pi_n^*)\in\boldsymbol{\Omega}_{-i}:=\prod_{j\neq i}\Omega_j$ 的一个最佳响应 (Best-response), 即

$$\pi_i^*\in\arg\max_{\pi_i\in\Omega_i}c_i(\pi_i,\pi_{-i}^*),$$

或者对任意的博弈者 $i\in\mathcal{N}$ 有

$$c_i(\pi_i^*,\pi_{-i}^*)\geq c_i(\pi_i,\pi_{-i}^*),\ \forall\pi_i\in\Omega_i,$$

则称策略组合 $(\pi_1^*,\pi_2^*,\cdots,\pi_n^*)\in\boldsymbol{\Omega}$ 是 $\mathcal{G}$ 的一个 Nash 均衡.

由上述 Nash 均衡的定义可以看出, 当所有博弈者的策略组合处在一个 Nash 均衡状态时, 任何一个博弈者在给定其他博弈者策略组合的情况下, 都无法通过单方面地改变自身的策略来提高自身的收益. 因此, Nash 均衡刻画了博弈者策略的一种 "稳定" 状态. 这点也是 Nash 均衡能够成为非合作博弈解概念的一个重要原因. 另外, 对于任意一个有限的标准式博弈, 它至少存在一个 Nash 均衡[38, 61]. 这点保证了 Nash 均衡作为非合作博弈解概念在理论上是可行的.

尽管 Nash 均衡作为非合作博弈的解概念在博弈论的发展历程中具有举足轻重的作用, 但它并不是十全十美的.[3] 首先, Nash 均衡解不仅要求博弈的参与者是完全理性的, 还要求博弈本身是完全信息的 (Complete-information), 即博弈模型中的诸如策略集和收益函数等参数信息是所有博弈者的共同知识. 对于大量现实博弈场景, 这样的要求显然是苛刻的. 一个典型的例子就是人类参与的各种博弈. 一方面, 人在获取、存储、使用和回忆信息的过程中无法保证是完全准确无误的; 另一方面, 人的表达能力是有限的, 其通过文字、数字、图表或语音等形式传递消息的过程中, 无论多么努力, 最终的效果都无法保证是完美无缺的[65]. 因此, 人类作为博弈者通常无法满足完全理性这一要求. 另外, 人类作为博弈者所参与的各种博弈通常也并不是完全信息的, 因为在人类社会中, 任何一个个体都无法完全知道其他个体的所有信息. 此外, Nash 均衡作为博弈的解本质上是一个 "静态的" 概念. 它只描述了在"稳定的" 策略组合状态下, 当其他博弈者不改变其自身策略时, 目标博弈者如何进行最佳的策略响应. 因此, 如果博弈本身或者博弈者的决策会随时间发生动态变化, 标准的 Nash 均衡则需进一步精炼才能对这种动态博弈的解进行刻画.

通过放宽上述 Nash 均衡的几点要求, 标准式博弈之后相继发展出了多种其他形式. 下面将逐一介绍其中的三种代表性形式, 即演化博弈、随机博弈和不完全信息博弈.

### 1.2 演化博弈

演化博弈是博弈论与动态演化过程相结合的一种动态博弈形式, 其起源于演化生物学家 Maynard Smith 和 Price 关于动物竞争行为的研究[66−67]. 特别地, 相比于标准式博弈, 演化博弈一般具有以下几方面的显著特征.

首先, 演化博弈在经典博弈论的基础上引入了生物学中的 "群体思维" (Population thinking)[68], 即演化博弈考虑的对象通常是一个博弈者群体, 而不仅仅是特定的几个博弈者. 因此, 演化博弈也称为群体博弈 (Population game)[69−70]. 其次, 演化博

[3] 期刊 *Artificial Intelligence* 在 2007 年组织的一个关于博弈论与多智能体学习的专刊中, 专门对 Nash 均衡在现实应用中的局限性进行了讨论[44, 62]. 另外, 文献 [63−64] 从实证等角度对 Nash 均衡进行了重新审视并提出了一些可能的替代概念.

弈中的博弈者通常被设定是有限理性的 (Bounded rationality)[71]. 所谓有限理性是指博弈者在进行决策时, 它追求的目标是一个满意的解, 而非必须是一个最优的解. 再次, 演化博弈引入了一个称为演化稳定策略 (Evolutionarily stable strategy, ESS) 的新概念[66], 从而弱化了 Nash 均衡作为博弈解的作用. 所谓 ESS 是指, 如果群体中的所有博弈者采取的策略为一个 ESS, 那么它可以抵御群体中充分小比例突变策略的入侵. 特别地, 因为每一个 ESS 都被证明是一个 Nash 均衡, 而该命题的反命题却不一定成立[72−74], 所以 ESS 是一个比 Nash 均衡更严格的概念. 最后, 同时也是最为显著的一点, 不同于传统的非合作博弈常立足于研究理性博弈者策略均衡的形成, 演化博弈更强调从系统论的角度出发, 着重于研究微观个体水平上的策略调整过程在宏观水平上所呈现出的群体动力学[69−70, 73]. 在演化博弈中, 驱动博弈者策略演化的 "动力" 是达尔文的 "优胜劣汰" 与 "适者生存" 的自然选择原理, 即具有高适应度 (Fitness) 的策略可通过个体水平上的诸如模仿、学习、复制、遗传或者传染等方式在群体中传播, 而具有低适应度的策略则倾向于在自然选择的作用下在群体中消失. 因为这些微观个体水平上的行为方式刻画了博弈者如何从一个策略更新为另一个策略, 所以在宏观群体水平上, 它们带来的一个直接结果是博弈者策略频率 (即选择每个策略的博弈者数占博弈者总数的比例) 的变化. 当这些策略频率在群体水平上随时间演化时, 在系统层面上, 它们将呈现出各种长期的动力学行为, 比如稳定和不稳定的平衡点、极限环、异宿环和混沌等[69−70, 75].

在演化博弈中, 因为个体水平上的策略更新方式通常具有多种不同的形式, 所以在群体水平上, 它们所呈现的演化动力学也将表现出多种不同的形式. 一般地, 如果按群体结构 (即博弈者群体的空间交互结构) 进行划分, 它们通常可分为无限混合均匀群体 (Infinite well-mixed population, 即博弈者总数无穷大且任意两个博弈者可以随机交互的群体)、有限混合均匀群体 (Finite well-mixed population, 即博弈者总数有限且任意两个博弈者可以随机交互的群体) 和结构群体 (Structured population, 即博弈者总数有限且每个博弈者只能与其局部邻居进行交互的群体)[74]. 然而, 如果按系统动力学进行划分, 它们一般可分为确定性演化博弈动力学和随机性演化博弈动力学. 下面将对这两类演化博弈动力学分别进行介绍.

#### 1.2.1 确定性演化博弈动力学

无限混合均匀的群体是确定性演化博弈动力学通常采用的群体结构[75]. 根据博弈者微观策略更新规则的不同, 现有的确定性演化博弈动力学主要有复制动力学 (Replicator dynamics)[76−77]、Logit 动力学 (Logit dynamics)[73]、BNN 动力学 (Brown-von Neumann-Nash dynamics)[78] 和 Smith 动力学 (Smith dynamics)[79] 等. 一般地, 这几个动力学系统可统一地用如下的一个常微分方程来描述[70]

$$\dot{\chi}_k = \sum_{l\in\tilde{\Omega}} \chi_l \eta_{lk} - \chi_k \sum_{l\in\tilde{\Omega}} \eta_{kl},\ \forall k \in \tilde{\Omega}, \tag{2}$$

其中 $\tilde{\Omega}$ 表示群体中所有博弈者的 (离散且有限的) 策略集, $\chi_k$ 表示策略 $k \in \tilde{\Omega}$ 在群体中的频率 (即策略为 $k$ 的博弈者数占博弈者总数的比例), $\eta_{kl}$ 表示任意一个博弈者从策略 $k \in \tilde{\Omega}$ 切换为策略 $l \in \tilde{\Omega}$ 的期望概率. 本质上讲, 式 (2) 是一个博弈者群体策略演化的平均场 (Mean-field) 动力学, 它刻画了一个由 $\eta_{kl}$ 描述的微观策略更新过程在宏观群体水平上所呈现的策略演化动态. 式 (2) 等号右侧第 1 项表征了所有从其他策略切换为策略 $k$ 的博弈者在群体中的期望频率, 而第 2 项表征了所有从策略 $k$ 切换为其他策略的博弈者在群体中的期望频率. 因此, 前者减去后者刻画了群体中策略为 $k$ 的博弈者频率在单位时间内的期望变化量.

特别地, 当 $\eta_{kl}$ 描述的策略更新规则采用不同的具体形式时, 式 (2) 将会随之发生相应的改变. 例如, 对应于复制动力学, $\eta_{kl}$ 描述的微观策略更新规则主要有 "成对比例模仿" (Pairwise proportional imitation)[80−81] 和 "不满意驱动模仿" (Imitation driven by dissatisfaction)[70] 两种形式. 具体地, 如果博弈者采用前者作为策略更新规则, 那么策略为 $k$ 的博弈者将以概率 $\eta_{kl} = \chi_l[h_l - h_k]_+$ 去模仿策略为 $l$ 的博弈者的策略, 其中 $h_k$ 和 $h_l$ 分别表示策略为 $k$ 和 $l$ 的博弈者的期望收益, $[\cdot]_+$ 表示斜坡函数算子 (当中括号里的变量大于零时, 经该算子得到的值为其本身, 否则得到的值为零). 相应地, 如果博弈者采用后者作为策略更新规则, 那么策略为 $k$ 的博弈者将以概率 $\eta_{kl} = \chi_l[K - h_k]$ 去模仿策略为 $l$ 的博弈者的策略, 其中 $K$ 是一个常数, 表示所有博弈者期望得到的收益. 将上述两个关于 $\eta_{kl}$ 的表达式分别代入到方程 (2) 中, 于是有复制动力学方程为

$$\dot{\chi}_k = \chi_k \left( h_k - \sum_{l\in\tilde{\Omega}} \chi_l h_l \right),\ \forall k \in \tilde{\Omega}.$$

上述复制动力学是演化博弈中使用最为广泛、研究最为透彻、角色最为重要的一种确定性动力学. 它最初由生物数学家 Taylor 和 Jonker 于 1978 年为了研究 ESS 与动态系统的渐近稳定性之间的关

系而引入的[76]. 后来, 由于人们发现复制动力学与群体生物学中的其他动力学之间存在着紧密的联系, 比如通过一个非线性变换, 复制动力学方程可等价为生态学中的 Lotka-Volterra 方程[82], 所以它在群体遗传学中的诸如性别比例演化等问题的研究中得到了广泛的应用[83]. 除此之外, 人们还发现复制动力学方程的平衡点与它对应的博弈的 Nash 均衡和 ESS 之间也存在着密切的联系[73], 比如复制动力学方程对应的博弈的每个 Nash 均衡都是它的一个平衡点; 复制动力学方程的每个 Lyapunov 稳定平衡点都是其对应的博弈的一个 Nash 均衡; 博弈的每个 ESS 都是其对应的复制动力学方程的一个渐近稳定平衡点等. 正是由于这些等价关系的存在, 复制动力学方程也被视为是一种研究博弈论的动力学方法[70, 73]. 另外, 在无穷小策略更新步长的条件下, 复制动力学方程还与强化学习具有紧密的联系. 例如, Cross 学习[84] 和 Q 学习[85−86] 的算法迭代式在无穷小策略更新步长条件下将依概率收敛为一个复制动力学方程[87−89]. 特别地, 由于近年来强化学习在视频游戏[19−21, 90−91] 和棋牌[16−18, 92−93] 等具体博弈场景中取得了许多突破性进展, 所以目前如何将演化博弈中的复制动力学方程与机器学习中的强化学习方法进行融合已成为 AI 领域中的一个研究热点[94−98].

除了上述提到的与其他研究领域建立联系之外, 近年来复制动力学本身也相继发展出了一些新形式, 并在相关研究中发挥了重要作用. 例如, 为了研究合作策略演化问题, 文献 [99−100] 基于 “自愿参与” 机制提出了一个 “可选公共品博弈” (Optional public goods game) 模型并分析了该博弈对应的复制动力学方程的性质, 而文献 [101−104] 提出了几类多人阈值博弈模型并分析了它们对应的复制动力学方程的性质. 文献 [105−108] 则提出并梳理了几类激励机制并给出了它们对应的复制动力学方程的稳定性分析. 另外, 为了将经典的复制动力学方程推广到网络结构群体中和具有随机扰动的环境中, 文献 [109] 和文献 [110−111] 分别提出了一类图 (或网络) 上的复制动力学和一类随机复制动力学. 基于类似的想法, 文献 [112−113] 则将经典的复制动力学方程从离散策略空间拓展到了连续策略空间. 除此之外, 近年来复制动力学方程还分别通过与深度学习[114]、分布式优化和控制[115]、观点动力学[116−117]、以及“环境反馈”[118] 相结合相应发展出了神经复制动力学[119]、分布式复制动力学[120−121]、基于评估网络 (Appraisal network) 的复制动力学[122] 和“博弈–环境反馈”复制动力学[123−126] 等多种新颖形式.

尽管确定性演化博弈动力学在数学上具有诸多良好性质并且能为深入理解博弈者策略的演化动态提供有效的分析工具, 但该类方法通常只能处理不受随机因素影响的混合均匀群体的策略演化动态. 然而, 在现实博弈场景中, 博弈者总数通常是有限的, 并且博弈者的策略更新过程常会受到诸如噪声等随机因素的影响. 因此, 为了研究这类具有随机扰动的有限博弈者群体的策略演化动态, 学者们提出了一类称为“随机性演化博弈动力学”[127] 的方法.

### 1.2.2 随机性演化博弈动力学

不同于确定性演化博弈动力学, 博弈者数有限的群体是随机性演化博弈动力学通常采用的群体结构[127]. 由于这一原因, 群体中博弈者的丰富度 (Abundance) 通常只能采用整数或者分数来描述. 另外, 由于在有限群体中, 博弈者的策略更新过程常会受到随机因素的影响, 所以博弈者的策略演化动态通常无法使用连续时间常微分方程来刻画, 而需借助随机过程这一数学工具. 除了这些差异之外, 随机性演化博弈动力学与确定性演化博弈动力学也存在着一些相似之处, 比如微观策略更新规则仍是影响宏观策略演化动态的一个关键因素. 从来源上讲, 目前主流的微观策略更新规则主要可分为两类: 一类来源于演化生物学, 另一类来源于社会学. 下面分别介绍这两类策略更新规则中的几种典型形式.

Moran 过程[128] 是群体遗传学中的一个经典模型, 也是随机性演化博弈动力学早期引入的一种策略更新规则[129−130]. 该过程主要借鉴了生物系统中的生死繁衍过程, 描述了一个博弈者如何从一个策略更新为另一个策略. 具体地, 在 Moran 过程中的每个时间步, 首先群体中的一个个体将以正比于其适应度的概率被随机地选定去产生一个后代. 随后, 在无变异存在的情况下, 该后代将完美地继承其父代的策略, 并随机地取代群体中的一个其他任意个体. 因为在这个过程中, 一个新个体产生的同时, 一个原有个体将被取代, 所以群体中的个体总数将始终保持不变. 特别地, 在上述策略更新过程中, 为了量化个体的适应度, 现有的文献通常将其定义为一个关于博弈者期望收益 $h$ 的函数. 例如, 目前普遍采用的两类函数形式分别为 $1-\beta+\beta h$[129−130] 和 $\exp(\beta h)$[131−132], 其中 $\beta$ 表示选择强度参数. 对于前者 $\beta\in[0,1]$, 而对于后者 $\beta\in[0,+\infty)$. 当 $\beta=0$ 时, 其代表中性选择 (Neutral selection). 在该情况下, 因为每个博弈者都有一个相同的背景适应度 (Background fitness) 或基准适应度 (Baseline fitness), 所以基于 Moran 过程每个个体被选定的概率均是相等的. 类似地, 当 $\beta\to 0$ 时, 其代表弱选择 (Weak

selection)[133]. 在该情况下, 由于个体的期望收益值只轻微地影响其适应度, 所以基于 Moran 过程每个个体被选定的概率将接近中性选择的情况. 对于 $1-\beta+\beta h$, 当 $\beta\to 1$ 时, 或者对于 $\exp(\beta h)$, 当 $\beta\to+\infty$ 时, 其代表强选择 (Strong selection). 在该情况下, 因为个体的期望收益值完全决定了其适应度, 所以基于 Moran 过程, 期望收益值越大的个体将总具有更大的机会产生后代, 并同时具有更大的机会传播其策略. 除了上述两个常用的适应度函数之外, 文献 [134] 也研究了诸如多项式函数和分段函数等其他更一般化的函数形式. 另外, 通过引入变异率参数, 文献 [135−137] 将上述标准的 Moran 过程拓展到了博弈者的策略更新过程存在变异的场景中.

不同于源自群体遗传学中的 Moran 过程, 对比较过程 (Pairwise comparison process)[109, 127, 138] 是一种基于社会学中的文化演化而提出的策略更新规则. 该规则主要刻画了"成功" 个体的策略是如何通过模仿被其他个体所复制的过程. 具体地, 对于一个有限混合均匀的博弈者群体, 基于对比较过程, 首先群体中的任意两个个体将被随机地选定, 其中一个个体作为中心博弈者 (Focal player), 而另一个个体作为被模仿者 (Role model). 随后, 中心博弈者将以一个正比于这两个个体收益之差的概率 $p_r$ 学习被模仿者个体的策略. 对于概率 $p_r$ 的函数形式, 目前一种常用的选择是 Fermi 函数, 即 $p_r=1/\{1+\exp[-\beta(h_E-h_I)]\}$, 其中 $h_I$ 和 $h_E$ 分别表示中心博弈者和被模仿者的期望收益. 除了 Fermi 函数之外, 文献 [139−140] 和文献 [134] 也分别研究了 $p_r$ 为线性函数的情形和一般化抽象函数的情形. 另外, 为了研究 Moran 过程中的适应度函数和对比较过程中的策略选择概率函数对策略演化结果的影响, 文献 [141] 提出了一个一般化形式的适应度函数和一个一般化形式的策略选择概率函数, 并给出了策略演化达到平稳状态时一个策略占优于另一个策略的充分必要条件. 受该文的启发, 文献 [142] 则进一步将上述一般化的函数思想推广到了网络化混合决策博弈场景中.

上述介绍的 Moran 过程和对比较过程都隐含地采用了有限混合均匀的群体结构假设. 然而, 通过适当的调整, 这两个策略更新规则其实也可以运用于具有网络空间结构的博弈者群体, 即所谓的复杂网络上的演化博弈[143−144]. 对应于 Moran 过程, 其在复杂网络上的两个典型版本分别是 "生–死" 过程 (Birth-death process) 和 "死–生" 过程 (Death-birth process)[109, 145−146]. 具体地, 在 "生–死" 过程中, 首先群体中的一个个体将以正比于其适应度的概率被选定去产生一个后代; 随后, 该后代将随机地取代网络中其父代邻居中的一个个体. 而在"死–生" 过程中, 首先群体中的一个个体将被随机地选定去死亡; 随后, 该死亡个体的某一个邻居将以正比于其适应度的概率被选定去将其后代产生在该死亡个体的位置上. 虽然这两个策略更新规则表面上看起来只是顺序上的差异, 但在宏观动力学行为上, 它们却会产生出截然不同的结果[146−147]. 对应于对比较过程, 其在复杂网络上的版本只需在个体选择阶段进行一些微小调整. 具体地, 在对比较过程的个体选择阶段, 首先随机地选定群体中的一个个体作为中心博弈者, 随后再随机地选定该中心博弈者的一个邻居作为被模仿者. 然后, 中心博弈者将以一个正比于这两个个体收益之差的概率 $p_r$ 去学习被模仿者个体的策略[109, 146]. 基于上述复杂网络上的演化博弈这一研究框架, 文献 [148−150] 对博弈者多样性的交互行为进行了一系列研究, 并提出了行为多样性、边多样性和博弈转移等概念. 而文献 [151−152] 基于现实的博弈者交互网络具有时序性这一特点提出了一类时序网络化系统, 并对其中的控制问题和演化博弈动力学问题进行了深入研究. 另外, 基于复杂网络上的演化博弈这一研究框架, 文献 [153]、文献 [154] 和文献 [155] 分别研究了策略的共演化 (Co-evolution) 特性、策略的自适应协调特性和博弈动力学的空间周期特性.

除了上述介绍的几种典型策略更新规则之外, 相关研究也考虑了一些其他类型的策略更新规则, 比如 "以牙还牙" (Tit-for-tat)[156]、"赢留输变" (Win-stay lose-shift)[157]、模仿 (Imitation)[81, 109, 146]、摄动最佳响应 (Perturbed best response)[158]、Wright-Fisher 过程[159−160] 和最近广受关注的基于期望的 (Aspiration-based) 策略更新规则[161−165] 等.

### 1.3 随机博弈

与演化博弈相类似, 随机博弈也是动态博弈的一种形式, 它最初是由 Shapley 于 1953 年提出并命名的[166]. 在 AI 和自动控制领域中, 随机博弈通常也称为 Markov 博弈[35]. 然而, 有别于演化博弈的一点是, 随机博弈引入了一个称为 (博弈) "状态"的全新概念.[4] 或粗略地讲, 随机博弈在每个阶段都有一个状态变量用于刻画当前所有博弈者所面临的博弈场景.

一般地, 一个随机博弈 $\mathcal{G}_s$ 通常可以表示为一个六元组

$$\mathcal{G}_s=\{\mathcal{N},\mathcal{S},\{\mathcal{A}_i\}_{i\in\mathcal{N}},\{r_i\}_{i\in\mathcal{N}},P,\Gamma\},$$

[4] 在相关文献中, "状态"也称为"系统状态"[49] 或"环境状态"[47] 等.

其中 $\mathcal{N}$ 表示所有博弈者的标号构成的集合, $\mathcal{S}$ 表示所有博弈者所处 "环境" 或 "系统" 的状态集, $\mathcal{A}_i$ 表示博弈者 $i \in \mathcal{N}$ 的行动集, $r_i : \mathcal{S} \times \boldsymbol{\mathcal{A}} \times \mathcal{S} \to \mathbb{R}$ 表示博弈者 $i \in \mathcal{N}$ 的收益函数或回报函数 (Reward function), $P : \mathcal{S} \times \boldsymbol{\mathcal{A}} \to \Delta(\mathcal{S})$ 表示状态转移概率分布函数, $\Gamma$ 表示博弈进行的时间集,[5] $\boldsymbol{\mathcal{A}} := \prod_{i\in\mathcal{N}} \mathcal{A}_i$, $\Delta(\mathcal{S})$ 表示集合 $\mathcal{S}$ 上的所有概率分布构成的集合. 在随机博弈进行的每个阶段, 系统将处于状态集 $\mathcal{S}$ 中的某一个状态 $s \in \mathcal{S}$. 随后, 依赖系统当前的状态 $s \in \mathcal{S}$, 每个博弈者 $i \in \mathcal{N}$ 将根据它的策略 (Policy) $\pi_i : \mathcal{S} \to \mathcal{A}_i$ 从其有效的行动空间 $\mathcal{A}_i$ 中选择一个行动 $\pi_i(s) = a_i \in \mathcal{A}_i$.[6] 当所有博弈者执行完它们的行动之后, 博弈将会产生两方面的结果. 一方面, 取决于系统当前的状态 $s \in \mathcal{S}$ 以及所有博弈者在当前阶段采取的行动 $\boldsymbol{a} \in \boldsymbol{\mathcal{A}}$, 系统将以概率 $P(s'|s, \boldsymbol{a})$ 从当前的状态 $s \in \mathcal{S}$ 转移到下一阶段的另一个状态 $s' \in \mathcal{S}$. 另一方面, 作为所有博弈者行动和系统状态转移的结果, 每个博弈者 $i \in \mathcal{N}$ 将从当前阶段的博弈中获得一个即时的收益值或回报值 $r_i(s, \boldsymbol{a}, s')$.

特别地, 当状态集 $\mathcal{S}$ 为一个单点集时, 随机博弈 $\mathcal{G}_s$ 将退化为一个重复的标准式博弈. 因此, 从这点上讲, 随机博弈将 von Neumann 提出的标准式博弈拓展到了动态多阶段和动态多场景中[167]. 另一方面, 当博弈者集合 $\mathcal{N}$ 只含一个博弈者时, 随机博弈 $\mathcal{G}_s$ 将退化为一个标准的 Markov 决策过程 (Markov decision process)[168−169]. 因此, 从这点上讲, 随机博弈又将 Markov 决策过程从单智能体系统拓展到了多智能体系统 (见图 1). 正是由于这一原因, 在 AI 领域中, 随机博弈也被视为是多智能体强化学习的一般性模型框架[35].

对于一个 $n$ 人随机博弈 $\mathcal{G}_s$, 此时 $\mathcal{N} = \{1, 2, \cdots, n\}$, 当分别给定所有博弈者的策略组合 $\boldsymbol{\pi} := (\pi_1, \pi_2, \cdots, \pi_n)$ 、所有博弈者的收益函数组合 $\boldsymbol{r} := (r_1, r_2, \cdots, r_n)$ 和状态转移概率分布函数 $P$ 时, 如果随机博弈 $\mathcal{G}_s$ 在离散时间集 $\Gamma$ 上进行无穷轮, 那么它在时间尺度上将会诱导出一个时间序列 $\{(s_t, \boldsymbol{a}_t, \boldsymbol{r}_t)\}_{t\in\Gamma}$, 其中 $s_t \in \mathcal{S}$ 表示系统在时刻 $t$ 的状态, $\boldsymbol{a}_t := (a_{1,t}, a_{2,t}, \cdots, a_{n,t})$ 表示所有博弈者在时刻 $t$ 的行动组合, 这里 $a_{i,t}$ 表示博弈者 $i \in \mathcal{N}$ 在时刻 $t$ 的行动, $\boldsymbol{r}_t := (r_{1,t}, r_{2,t}, \cdots, r_{n,t})$ 表示所有博弈者在时刻 $t$ 的收益值组合, $r_{i,t} = r_i(s_t, \boldsymbol{a}_t, s_{t+1})$ 表示博弈者 $i \in \mathcal{N}$ 在时刻 $t$ 的收益值. 特别地, 如果在整个博弈过程中每个博弈者都是理性的, 那么对于任意的博弈者 $i \in \mathcal{N}$, 它的目标将是找到一个策略 $\pi_i \in \Omega_i$ 使得其长期的累积期望收益 $V^i_{(\pi_i,\pi_{-i})}(s; r_i, P)$, $\forall s \in \mathcal{S}$ 最大, 这里 $\Omega_i$ 表示博弈者 $i$ 的有效策略集, $\pi_{-i} \in \boldsymbol{\Omega}_{-i} := \prod_{j\neq i} \Omega_i$ 表示博弈者 $i$ 之外的所有博弈者的策略组合. 一般地, 为了保证 $V^i_{(\pi_i,\pi_{-i})}(s; r_i, P)$ 是一个有界值函数, 其计算方式常采用如下的折扣形式[7]

$$V^i_{(\pi_i,\pi_{-i})}(s; r_i, P) = \mathbf{E}_{\boldsymbol{a}_t\sim(\pi_i,\pi_{-i})(s_t), s_{t+1}\sim P(\cdot|s_t,\boldsymbol{a}_t)} \left\{ \sum_{t=0}^{\infty} \gamma^t r_i(s_t, \boldsymbol{a}_t, s_{t+1}) \mid s_0 = s \right\}, \tag{3}$$

其中 $\mathbf{E}_{\boldsymbol{a}_t\sim(\pi_i,\pi_{-i})(s_t), s_{t+1}\sim P(\cdot|s_t,\boldsymbol{a}_t)}\{\cdot\}$ 表示对 $\boldsymbol{\pi} = (\pi_i, \pi_{-i})$ 和 $P$ 诱导出的随机过程 $\{(\boldsymbol{a}_t, s_t)\}_{t\geq 0}$ 求数学期望, $\gamma \in [0, 1)$ 是一个折扣因子.[8] 与标准 Nash 均衡的解概念相类似, 对于随机博弈 $\mathcal{G}_s$, 它的标准解是一个称为 Markov 完美均衡 (Markov perfect equilibrium) 的概念[170−171].

**定义 2.** 对于有限的 $n$ 人随机博弈 $\mathcal{G}_s = \{\mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, \{r_i\}_{i\in\mathcal{N}}, P, \Gamma\}$, $\mathcal{N} = \{1, 2, \cdots, n\}$, 如果对 $\forall i \in \mathcal{N}$ 和 $\forall s \in \mathcal{S}$ 有

$$V^i_{(\pi_i^*,\pi^*_{-i})}(s; r_i, P) \geq V^i_{(\pi_i,\pi^*_{-i})}(s; r_i, P),\ \forall \pi_i \in \Omega_i,$$

或者 $\pi_i^* \in \arg\max_{\pi_i\in\Omega_i} V^i_{(\pi_i,\pi^*_{-i})}(s; r_i, P)$, 则称策略组合 $(\pi_1^*, \pi_2^*, \cdots, \pi_n^*) \in \boldsymbol{\Omega} := \prod_{i\in\mathcal{N}} \Omega_i$ 是 $\mathcal{G}_s$ 的一个 Markov 完美均衡.

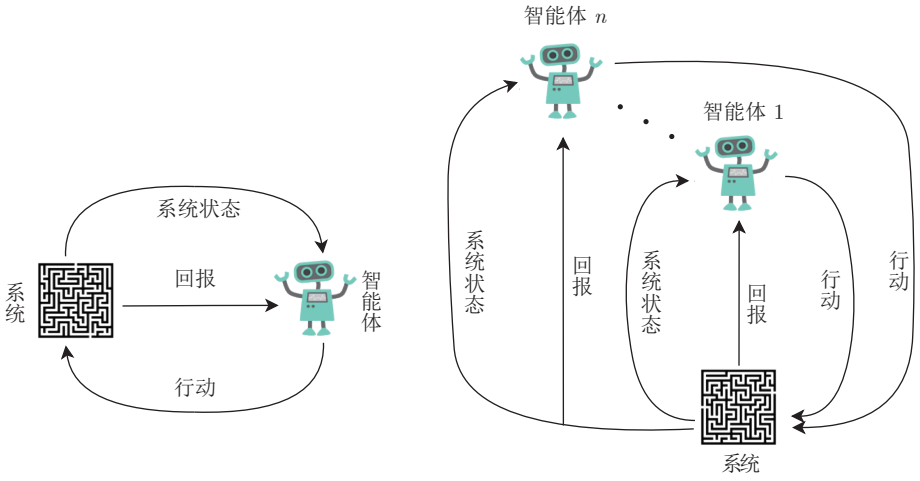
一般地, 对于一个有限的 $n$ 人随机博弈, 它的 Markov 完美均衡通常总是存在的[170]. 但是, 由于其计算复杂性问题的存在[172], 求解一个 Markov 完美均衡通常十分具有挑战性. 为此, 相关研究经常会考虑随机博弈的一些特定情形. 根据博弈者收益函数或回报函数的不同, 它们主要可分为 3 类[47, 49−50]: 1) 完全合作, 在该设置下所有博弈者具有相等的收益函数, 即 $r_i = r_j$, $\forall i, j \in \mathcal{N}$, 此时的随机博弈也称为多智能体 Markov 决策过程[27]、团队 Markov 博弈[28] 或团队随机博弈; 2) 完全竞争, 在该设置下随机博弈只有两个博弈参与者, 并且它们的收益函数满足零和关系, 即 $r_1 = -r_2$, 此时的随机博弈也称为零和随机博弈; 3) 混合型, 在该设置下所有博弈者的收益函数不受其他约束条件影响, 即此时的随机

[5] 除非特别说明, 本文只讨论 $\Gamma$ 为离散无穷的情况, 即随机博弈在离散时间集上进行无穷轮.

[6] 随机博弈中所说的 "策略" (Policy) 与演化博弈中所说的 "策略" (Strategy) 不是完全等同的概念. 在随机博弈中, 博弈者的策略是指行动集 (或纯策略集) 上的一个概率分布 (也称为混合策略); 而在演化博弈中, 个体的策略通常隐含地假定是一个确定性的纯策略. 因此, 演化博弈中所说的策略实质上是等同于随机博弈中的某一个具体的行动 (或纯策略). 另外, 为了叙述方便, 本文只讨论确定性的策略; 而对于随机性的策略 $\pi_i(\cdot|s) : \mathcal{S} \to \Delta(\mathcal{A}_i)$, $\forall i \in \mathcal{N}$, 本文中的相关结果可类似地得到.

[7] 文献 [169] 也对其他形式的计算方式进行了分析和讨论.

[8] $\gamma$ 的另外一种解释是: 在一轮博弈结束之后, 下一轮博弈继续进行的概率.




(a) Markov 决策过程　　(b) Markov 博弈

图 1　Markov 决策过程和 Markov 博弈的示意图

Fig. 1　Illustrations of Markov decision processes and Markov games

博弈为一般和随机博弈[39].

### 1.4　不完全信息博弈

上述介绍的几类博弈形式都隐含地采用了完全信息的假设, 即博弈模型的参数信息, 比如博弈者的行动集和收益函数等, 是所有博弈者的共同先验知识. 然而, 在大量现实博弈场景中, 完全信息的假设通常是一种过于理想化的要求. 通过放宽该假设, Harsanyi 提出了一类称为不完全信息博弈的模型[173]. 它为研究博弈者在具有部分模型信息条件下的决策问题提供了一套完善的数学方法与工具.

第一个成熟的不完全信息博弈模型是由 Harsanyi 于 1967 年提出并命名的 Bayesian 博弈[173]. 一般地, 一个 $n$ 人 Bayesian 博弈 $\mathcal{G}_b$ 通常可以表示为一个五元组[174]

$$\mathcal{G}_b = \{\mathcal{N}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, \{\Upsilon_i\}_{i\in\mathcal{N}}, \{\rho_i\}_{i\in\mathcal{N}}, \{g_i\}_{i\in\mathcal{N}}\},$$

其中 $\mathcal{N} = \{1, 2, \cdots, n\}$ 表示所有博弈者的标号构成的集合, $\mathcal{A}_i$ 表示博弈者 $i \in \mathcal{N}$ 的行动集, $\Upsilon_i$ 表示博弈者 $i \in \mathcal{N}$ 的类型集, $\rho_i : \Upsilon_i \to \Delta(\boldsymbol{\Upsilon}_{-i})$ 是一个条件概率分布函数, 表示博弈者 $i \in \mathcal{N}$ 在知道自己类型的情况下对其他博弈者类型的推断 (Belief), $\boldsymbol{\Upsilon}_{-i} := \prod_{j\neq i} \Upsilon_j$, $g_i : \boldsymbol{\mathcal{A}} \times \Upsilon_i \to \mathbb{R}$ 表示博弈者 $i \in \mathcal{N}$ 的收益函数或效用函数, $\boldsymbol{\mathcal{A}} := \prod_{i\in\mathcal{N}} \mathcal{A}_i$. 具体地, 在 Bayesian 博弈进行过程中, 首先一个称为“自然” (Nature) 的虚拟个体将赋予所有博弈者一个类型组合 $\boldsymbol{\iota} := (\iota_1, \iota_2, \cdots, \iota_n)$, 其中 $\iota_i \in \Upsilon_i$, $i = 1, 2, \cdots, n$ 表示博弈者 $i$ 的类型. 随后, 对于每个博弈者 $i \in \mathcal{N}$, “自然” 只告知属于其自身的类型 $\iota_i$, 而对所有其他博弈者 $-i$ 的类型进行保密. 然后, 每个博弈者 $i \in \mathcal{N}$ 将各自从其行动空间 $\mathcal{A}_i$ 中选择一个行动 $a_i \in \mathcal{A}_i$. 最后, 作为所有博弈者行动的结果, 每个博弈者 $i \in \mathcal{N}$ 将获得一个收益值 $g_i(a_1, a_2, \cdots, a_n; \iota_i)$.

特别地, 在上述 Bayesian 博弈模型中, 依赖于类型 $\iota_i \in \Upsilon_i$ 的收益函数 $g_i$ 是最为关键的一个要素, 因为它是 Bayesian 博弈模型中表示不完全信息的一个核心设置. 具体来讲, 因为每个类型 $\iota_i \in \Upsilon_i$ 都对应着博弈者 $i$ 的一个收益函数 $g_i$, 所以当博弈者 $i$ 不知道其自身的类型和其他博弈者的类型时, 它就无法明确知道其自身的收益函数和其他博弈者的收益函数. 换句话说, 当博弈者 $i$ 只知道其自身的类型时, 这等同于它只知道其自身的收益函数. 类似地, 当博弈者 $i$ 对其他博弈者的类型不确定时, 这等同于它对其他博弈者的收益函数不确定. 为了从数学上刻画这种不确定性, Bayesian 博弈假定所有博弈者的类型组合 $\boldsymbol{\iota}$ 的联合概率分布 $\rho(\boldsymbol{\iota})$ 是所有博弈者的共同知识. 基于该假定, 当“自然” 告知博弈者 $i$ 其类型为 $\iota_i$ 时, 博弈者 $i$ 于是可根据 Bayesian 法则对所有其他博弈者的类型 $\boldsymbol{\iota}_{-i} = (\iota_1, \cdots, \iota_{i-1}, \iota_{i+1}, \cdots, \iota_n)$ 进行推断, 即

$$\rho_i(\boldsymbol{\iota}_{-i} \mid \iota_i) = \frac{\rho(\boldsymbol{\iota}_{-i}, \iota_i)}{\rho(\iota_i)} = \frac{\rho(\boldsymbol{\iota}_{-i}, \iota_i)}{\sum_{\boldsymbol{\iota}'_{-i}\in\boldsymbol{\Upsilon}_{-i}} \rho(\boldsymbol{\iota}'_{-i}, \iota_i)}, \quad (4)$$

这里 $\rho_i(\boldsymbol{\iota}_{-i} \mid \iota_i)$ 表示博弈者 $i$ 在知道自己的类型为 $\iota_i$ 的情况下, 推测所有其他博弈者的类型为 $\boldsymbol{\iota}_{-i}$ 的概率.

因为在 Bayesian 博弈中, 每个博弈者的收益函数都是类型依赖的, 所以当博弈者进行决策时, 它的行动选择也将是类型依赖的. 换句话讲, 在 Bayesian 博弈中, 每个博弈者的策略都是一个关于其自身类型的函数. 基于这一考虑, Harsanyi 通过借鉴 Nash 均衡的类似定义方法, 提出了一个称为 Bayesian

均衡的博弈解概念, 并证明了该均衡解在有限 Bayesian 博弈中的存在性[173].

由上述 Bayesian 博弈类型推断的设定可以看出, 式 (4) 成立的一个关键点在于, 联合概率分布 $\rho(\boldsymbol{\iota})$ 是所有博弈者的共同知识. 然而, 由于在大量现实博弈场景中, 共同知识的假设通常是过于理想化的, 所以该假设的合理性经常受到质疑[175−176]. 为了放宽该假设, 学者们于是提出了一些解决方案. 例如, Mertens 和 Zamir 提出了一个称为"普遍类型空间" (Universal type space) 的替代概念[177]; Holmström 和 Myerson 提出了一类称为 "无分布均衡解" (Distribution-free equilibrium solution) 的方法[178]. 相较之下, Aghassi 和 Bertsimas 通过引入鲁棒优化[179] 方法, 提出了一类称为"鲁棒博弈" (Robust game) 的不完全信息博弈模型[180]. 在该模型中, 每个博弈者并不能明确地知道其收益函数, 而只能感知到其收益函数所在的一个不确定集. 借助从该不确定集得到的收益信息, 博弈者将基于鲁棒优化的方法进行行动决策, 以便获得一个"满意的"收益. 其中, "满意的" 收益是指, 每个博弈者的目标是寻找一个最优策略使得其收益值在最坏的情况 (Worst-case) 下能最大. 具体地, 对于一个由 $n$ 个博弈者参与的鲁棒博弈, 如果用 $c_i(\pi_i, \pi_{-i}; u_\alpha)$ 表示任意的博弈者 $i\,(i = 1, 2, \cdots, n)$ 在策略组合 $(\pi_i, \pi_{-i})$ 和不确定参数 $u_\alpha \in \mathcal{U}_\alpha$ 下的收益函数, 那么基于鲁棒优化方法, 任意博弈者 $i$ 的目标将是寻找在不确定参数 $u_\alpha$ 最坏的情况下, 对其他博弈者策略 $\pi_{-i} \in \boldsymbol{\Omega}_{-i}$ 的一个最佳响应 $\hat{\pi}_i$, 即

$$\hat{\pi}_i \in \arg\max_{\pi_i \in \Omega_i} \left[ \inf_{u_\alpha \in \mathcal{U}_\alpha} c_i(\pi_i, \pi_{-i}; u_\alpha) \right],$$

其中不确定参数 $u_\alpha$ 属于不确定集 $\mathcal{U}_\alpha$, 刻画了收益函数 $c_i$ 的不确定性. 由于鲁棒博弈的不确定性是借助收益函数的不确定集刻画的, 所以它完美地规避了 Bayesian 博弈关于共同知识的假定. 另外, 为了分析鲁棒博弈的性质, Aghassi 和 Bertsimas 借鉴 Nash 均衡的类似定义方法, 提出了一个鲁棒优化意义下的均衡解概念, 即所谓的"鲁棒优化均衡" (Robust-optimization equilibrium), 并证明了该均衡解在有限鲁棒博弈中的存在性[180].

然而, 在上述鲁棒博弈模型中, Aghassi 和 Bertsimas 只考虑了同时行动的 (Simultaneous-move)、有限的 (Finite)、一轮 (One-shot) 博弈场景. 之后, 为了将鲁棒博弈的思想推广到随机博弈中, Kardeş 等提出了一类称为"鲁棒随机博弈" (Robust stochastic game) 的模型[181]. 具体地, 对于一个随机博弈 $\mathcal{G}_s = \{\mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, \{r_i\}_{i\in\mathcal{N}}, P, \Gamma\}$, 如果 $\mathcal{G}_s$ 中的 $r_i,\ \forall i \in \mathcal{N}$ 和 $P$ 不是明确给定的, 而是分别属于不确定集 $\mathcal{R}_i$ 和 $\mathcal{P}$, 则称 $\mathcal{G}_s$ 为一个鲁棒随机博弈. 与 Markov 完美均衡的概念相类似, 鲁棒随机博弈在鲁棒优化意义下的均衡解是一个称为"鲁棒 Markov 完美均衡" (Robust Markov perfect equilibrium) 的概念[181].

**定义 3.** 对于有限的 $n$ 人鲁棒随机博弈 $\mathcal{G}_{rs} = \{\mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, \{r_i\}_{i\in\mathcal{N}}, P, \Gamma\}$, $\mathcal{N} = \{1, 2, \cdots, n\}$, $r_i \in \mathcal{R}_i$, $P \in \mathcal{P}$, 如果对 $\forall i \in \mathcal{N}$ 和 $\forall s \in \mathcal{S}$ 有

$$\Psi^i_{(\pi_i^*, \pi^*_{-i})}(s) \geq \Psi^i_{(\pi_i, \pi^*_{-i})}(s),\ \forall \pi_i \in \Omega_i,$$

或者 $\pi_i^* \in \arg\max_{\pi_i \in \Omega_i} \Psi^i_{(\pi_i, \pi^*_{-i})}(s)$, 则称策略组合 $(\pi_1^*, \pi_2^*, \cdots, \pi_n^*) \in \boldsymbol{\Omega} := \prod_{i\in\mathcal{N}} \Omega_i$ 是 $\mathcal{G}_{rs}$ 的一个鲁棒 Markov 完美均衡, 其中 $\Psi^i_{(\pi_i, \pi_{-i})}(s) := \inf_{r_i \in \mathcal{R}_i, P\in\mathcal{P}} V^i_{(\pi_i, \pi_{-i})}(s; r_i, P)$.

特别地, Kardeş 等发现, 当不确定集 $\mathcal{R}_i$ 和 $\mathcal{P}$ 是紧集时, 任意一个有限鲁棒随机博弈都至少存在一个鲁棒 Markov 完美均衡[181].

### 1.5 小结

作为第 1 节内容的一个小结, 图 2 从博弈者数、策略数、系统状态数和不完全信息的程度这四个维度展示了这节介绍的四类基本博弈形式的一个统一框架 (该框架引自文献 [182], 其中方框图颜色的深浅表示不完全信息程度的大小). 标准式博弈是一类传统的博弈形式, 也是现代博弈论发展的基石. 在该类博弈形式中, 一种典型的模型范式是所谓的两人两策略博弈, 比如囚徒困境博弈、雪堆博弈和猎鹿博弈等. 然而, 大量现实社会、经济和工程系统中的博弈通常都是由多个博弈者或者一个博弈者群体参与的, 比如著名的 "公共地悲剧" (The tragedy of the commons) 问题[183]. 因此, 为了研究这类具有多个博弈者参与的博弈问题, 博弈的模型范式在博弈者数这个维度上就需从两人形式转变为多人形式, 比如从囚徒困境博弈转变为公共品博弈. 进一步, 如果博弈者具有多个策略, 那么在策略这个维度上, 博弈的模型范式就需从两策略转变为多策略. 然而, 由于标准式博弈本质上是一种一轮博弈, 或者说是一种 "静态的" 博弈 (即没有考虑时间这个维度), 所以该类博弈形式通常无法对动态决策问题进行刻画. 因此, 如果在标准式博弈的基础上引入时间这个维度, 那么它将相应地转变为动态博弈. 进一步, 如果在动态博弈的基础上考虑群体和演化的作用, 那么在博弈形式上, 它就变成了演化博弈. 但是, 从本质上讲, 这类博弈仍然没有脱离标准式博弈的基本框架, 即一个由博弈者集、策略集和收益函数构成的三要素框架. 随机博弈通过在标准式博弈的基

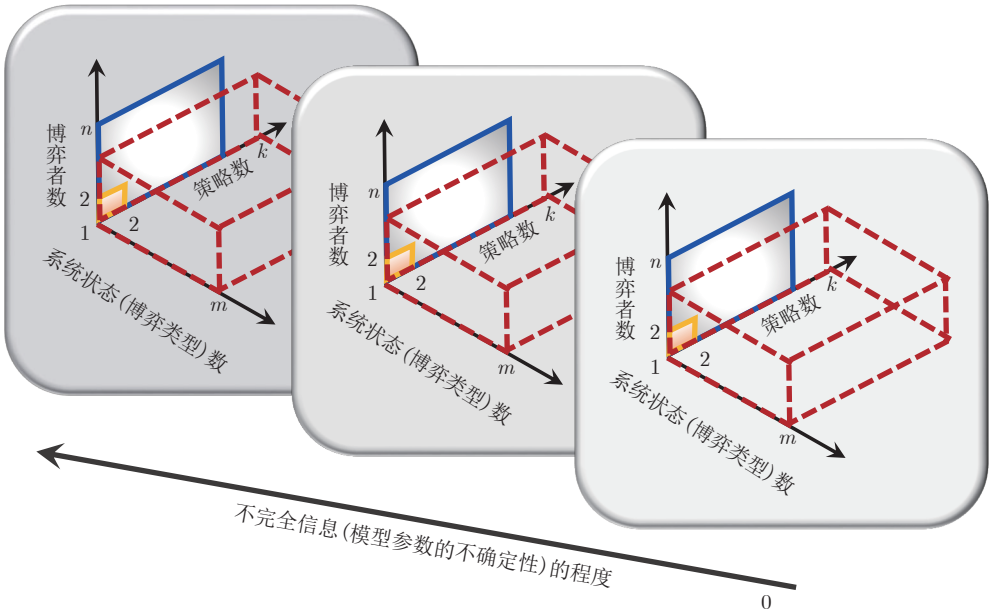



图 2　基本博弈形式的一个统一理论框架图

Fig. 2　Illustrations of a unified theoretical framework of the fundamental games

础上引入 (博弈) “状态” 这一新维度, 一方面将重复的标准式博弈拓展到了动态多类型博弈 (或状态依赖的博弈) 场景中, 另一方面又将单智能体 Markov 决策过程拓展到了多智能体系统中. 因此, 它极大地拓宽了博弈论在多智能体场景中的适用范围. 但不足的一点是, 它仍然保留了标准式博弈关于完全信息的这一假定. 在实际的博弈中, 含有不确定因素和不完全信息的场景通常是一种更为常见的情况. 由此, 通过在标准的随机博弈的基础上引入不完全信息这个维度, 鲁棒随机博弈将完全信息博弈的模型范式拓展到了不完全信息的场景中. 综上可见, 标准式博弈是博弈论的基石, 演化博弈、随机博弈和不完全信息博弈是这基石之上分别引入群体性思维与演化、博弈状态和不确定性这三个新维度之后的发展形式.

## 2　多智能体学习方法

对应于第 1 节介绍的标准式博弈、演化博弈、随机博弈和不完全信息博弈, 本节将主要论述这四类博弈形式中的多智能体学习方法, 即策略学习、学习动力学、强化学习和鲁棒学习.

### 2.1　策略学习

策略学习是一种学习对手 (Opponent) 策略的方法, 它在博弈学习理论 (The theory of learning in games) 中是一种广泛使用的经典方法[45−46].[9] 在该类方法中, 一个典型的例子是所谓的 “虚拟对弈” 过程[14, 184]. 该过程假定在一个重复的标准式博弈中, 每个博弈者所面临的对手都是平稳的 (Stationary), 即对手的策略不随时间变化. 另外, 为了描述每个博弈者对其对手策略的学习, 该过程假定每个博弈者是通过记录其对手过去所采用行动的频率 (即对手的实证策略) 来作为其对手策略的推断. 具体地, 对于一个两人重复标准式博弈 $\mathcal{G}=\{\mathcal{N},\{\mathcal{A}_i\}_{i\in\mathcal{N}},\{c_i\}_{i\in\mathcal{N}}\}$, 这里 $\mathcal{N}=\{1,2\}$, $c_i:\mathcal{A}_1\times\mathcal{A}_2\to\mathbb{R}$ 为博弈者 $i\in\mathcal{N}$ 的收益函数, 基于虚拟对弈过程, 任意的博弈者 $i$ 在第 $t$ $(t\geq 1)$ 轮博弈中对其对手 $-i$ 的策略推断为

$$\tilde{\pi}_{-i}^{t}(a)=\frac{1}{t}\sum_{\ell=0}^{t-1}\mathbb{1}\{a_{-i}^{\ell}=a\},\ a\in\mathcal{A}_{-i},$$

其中 $\tilde{\pi}_{-i}^{t}(a)$ 表示博弈者 $i$ 推测其对手 $-i$ 在第 $t$ 轮博弈中将采取行动 $a\in\mathcal{A}_{-i}$ 的概率, $\mathbb{1}\{a_{-i}^{\ell}=a\}$ 表示指标函数 (Indicator function). 如果对手 $-i$ 在第 $\ell$ $(\ell\geq 0)$ 轮博弈中采取的行动 $a_{-i}^{\ell}$ 为 $a\in\mathcal{A}_{-i}$, 那么 $\mathbb{1}\{a_{-i}^{\ell}=a\}=1$; 否则, $\mathbb{1}\{a_{-i}^{\ell}=a\}=0$. 基于该对手策略的推断 $\tilde{\pi}_{-i}^{t}(a)$, 博弈者 $i$ 在第 $t$ 轮博弈中采取的策略 $\pi_i^t$ 将为 $\tilde{\pi}_{-i}^{t}(a)$ 的一个最佳响应, 即

$$\pi_i^t\in\arg\max_{\pi_i\in\Omega_i}\mathbf{E}_{a_i\sim\pi_i,a_{-i}\sim\tilde{\pi}_{-i}^{t}}\{c_i(a_i,a_{-i})\}.$$

对于虚拟对弈学习过程, 一个重要的理论问题是, 博弈者的实证策略 $\tilde{\pi}_i^t$, $\forall i\in\mathcal{N}$ (即博弈者 $i$ 的策略推断) 是否会收敛? 如果收敛, 它是否会收敛到

[9] “策略学习”在博弈学习理论中具有十分丰富的内涵[45−46], 本文只关注其学习对手策略的这层主要含义.

一个 Nash 均衡? 这一问题也是博弈学习理论中的一个核心问题[45−46]. 对于零和博弈, Robinson 发现这个问题的答案是肯定的, 即基于虚拟对弈学习过程, 所有博弈者的实证策略将收敛到零和博弈的一个 Nash 均衡点[184]. 虽然这一肯定的答案之后也推广到了其他的博弈场景中, 比如两人两策略非零和博弈, 但在更多情况下, 虚拟对弈通常并不能保证收敛[45, 185]. 导致这一结果的一个重要原因是, 虚拟对弈过程要求对手的策略是平稳的. 然而, 在多智能体的学习环境中, 一方面, 一个博弈者在学习的同时另一个博弈者也在学习, 所以每个博弈者面临的 "外部学习环境" 都在不断变化; 另一方面, 由于对手的策略总在不断地调整, 所以每个博弈者总是需要去跟踪学习一个不断变化的"目标". 因而, 这就破坏了虚拟对弈所需的平稳性条件. 该问题也是多智能体学习中普遍存在的非平稳性 (Non-stationarity) 问题[186]. 按照复杂性程度的不同, 文献 [186] 对现有处理该类问题的方法进行了总结和归纳, 并把它们划分成了 5 类: 1) 忽视 (Ignore), 即忽视对手策略的非平稳性从而认为对手策略是平稳的; 2) 遗忘 (Forget), 即只依赖最近观察到的信息更新对手策略的推断; 3) 响应目标对手 (Respond to target opponent), 即对明确的目标对手采取最佳的策略响应; 4) 学习对手模型 (Learn opponent model), 即学习对手策略的模型并依赖该模型推导出自身的策略响应; 5) 心智理论 (The theory of mind), 即推测对手的策略并假定对手也在推测自己, 从而形成一个不断递归的推测过程. 另外, 针对学习对手模型的这类方法, 文献 [187] 将其称为对手模拟 (Opponent modelling) 并对该类方法进行了全面的总结和归纳, 而文献 [188] 从博弈者所需信息的角度给出了该类方法的一个分类框架.

由上述关于虚拟对弈的介绍可以看出, 标准的虚拟对弈过程主要适用于重复的标准式博弈. 之后, 为了将该博弈学习方法拓展到其他博弈类型中, 虚拟对弈过程也相继地发展出了一些其他的变体形式, 比如随机虚拟对弈 (Stochastic fictitious play)[45]、联合策略虚拟对弈 (Joint strategy fictitious play)[189]、分布式虚拟对弈 (Distributed fictitious play)[190−191]、采样虚拟对弈 (Sample fictitious play)[192]、神经虚拟自对弈 (Neural fictitious self-play)[193]、基于零和随机博弈的虚拟对弈[194] 和基于平均场博弈的虚拟对弈[195] 等.

## 2.2 学习动力学

学习动力学是博弈学习与系统动力学相结合的一种方法, 也是从动态系统的角度去理解博弈学习过程的一种方式. 目前, 该类方法中的大部分工作都是基于演化博弈动力学与多智能体强化学习的交叉来实现的[95]. 强化学习是一种利用经验不断试错的学习方式, 它被认为是动物学习和生物智能的基石之一[196−197]. 而多智能体强化学习是该学习方式在多智能体场景 (比如多人博弈) 下的拓展形式. 该学习方法更为细致的介绍将在第 2.3 小节中呈现, 本小节主要关注其与演化博弈动力学的结合.

Cross 学习[84] 是针对重复的标准式博弈 (也称为无状态 (Stateless) 的博弈) 提出的一种强化学习方法, 也是强化学习与演化博弈动力学产生关联的一类早期方法[87]. 具体地, 考虑一个重复的标准式博弈 $\mathcal{G}=\{\mathcal{N},\mathcal{A},\{c_i\}_{i\in\mathcal{N}}\}$, 这里假定所有博弈者共享同一个行动空间 $\mathcal{A}$ 并且 $c_i:\mathcal{A}^{|\mathcal{N}|}\to\mathbb{R}$. 基于 Cross 学习, 每个博弈者 $i\in\mathcal{N}$ 的策略更新方式为

$$\begin{cases}\pi_i^{t+1}(a)=r_i^t(a)+(1-r_i^t(a))\pi_i^t(a), & a=a_i^t\\ \pi_i^{t+1}(a)=(1-r_i^t(a_i^t))\pi_i^t(a), & \text{否则}\end{cases}\tag{5}$$

其中 $\pi_i^t(a)$ 表示博弈者 $i\in\mathcal{N}$ 在第 $t$ 轮博弈中采取行动 $a\in\mathcal{A}$ 的概率, $r_i^t(a)\in(0,1)$ 是一个随机变量, 表示博弈者 $i\in\mathcal{N}$ 在第 $t$ 轮博弈中采取行动 $a\in\mathcal{A}$ 所获得的收益, $a_i^t$ 表示博弈者 $i\in\mathcal{N}$ 在第 $t$ 轮博弈中采取的行动. 令 $\Delta\pi_i^t(a):=\pi_i^{t+1}(a)-\pi_i^t(a)$. 于是, 借助式 (5), $\Delta\pi_i^t(a)$ 的数学期望可计算为

$$\begin{aligned}\mathbf{E}\left\{\Delta\pi_i^t(a)\right\}&=\pi_i^t(a)[\mathbf{E}\{r_i^t(a)\}-\mathbf{E}\{r_i^t(a)\}\pi_i^t(a)]+\\&\sum_{a'\in\mathcal{A}\setminus\{a\}}\pi_i^t(a')[-\mathbf{E}\{r_i^t(a')\}\pi_i^t(a)]=\\&\pi_i^t(a)\left[\mathbf{E}\{r_i^t(a)\}-\sum_{a'\in\mathcal{A}}\pi_i^t(a')\mathbf{E}\{r_i^t(a')\}\right].\end{aligned}\tag{6}$$

如果假定相继两轮博弈之间的时间间隔为 $\delta\in(0,1]$, 那么第 $t$ $(t\geq 0)$ 轮博弈发生的时刻 $\tau$ 为 $\tau=t\delta$. 当 $\delta\to 0$ 时 (即在一个无穷小策略更新步长条件下), Börgers 和 Sarin 发现式 (6) 将依概率收敛为如下一个确定性复制动力学方程[87]

$$\dot{\pi}_i(a)=\pi_i(a)\left[\mathbf{E}\{r_i(a)\}-\sum_{a'\in\mathcal{A}}\pi_i(a')\mathbf{E}\{r_i(a')\}\right],$$

这里 $\pi_i(a)$ 和 $r_i(a)$ 分别表示博弈者 $i\in\mathcal{N}$ 选择行动 $a\in\mathcal{A}$ 的概率和采取行动 $a\in\mathcal{A}$ 后获得的收益, 并且它们都是关于连续时间变量的函数.

受上述研究思路的启发, 多智能体强化学习与复制动力学的交叉研究随后分别在统计物理和 AI 这两个领域中开始了独立发展, 其中文献 [88] 和文献 [89] 是它们各自的早期工作. 虽然这两篇文献的作者分别隶属于不同的研究领域 (前者为统计物理

或复杂系统, 而后者为 AI ), 但他们都基于重复的标准式博弈对强化学习中的 Q 学习的演化动力学进行了研究, 并独立地发现了 Q 学习在无穷小更新步长条件下将收敛为一个复制动力学方程的事实. 基于这两项工作, 学习动力学这一交叉研究主题后续分别沿着不同的风格在统计物理和 AI 这两个领域中继续向前发展. 具体地, 在统计物理领域中, 后续的相关工作主要专注于研究复杂博弈场景中的长期动力学行为, 比如混沌[198−200]、分叉[201] 和周期振荡[202] 等. 而在 AI 领域中, 后续的相关工作更多地关注于多智能体强化学习算法的动力学解释和新学习算法的开发, 比如频率调整的 Q 学习 (Frequency-adjusted Q-learning)[203]、加权策略学习者 (Weighted policy learner) 算法[204] 和遗憾最小化 (Regret minimization) 算法[205] 等.

目前, 虽然学习动力学的相关研究已取得了十分丰硕的成果, 并且极大地推动了多智能体强化学习与演化博弈动力学的交叉发展, 但它们也存在着一些不足之处. 例如, 目前该领域中的大部分工作所采用的博弈形式都是重复的标准式博弈 (即无状态的博弈), 并且这些工作的研究范式仍然沿袭着 Q 学习与确定性复制动力学方程的交叉. 然而, 由上一节关于演化博弈和随机博弈的介绍可以看出: 一方面, 随机性演化博弈动力学比确定性演化博弈动力学具有更宽的适用范围; 另一方面, 随机博弈比重复的标准式博弈更具一般性. 因此, 自然地产生了这样一个问题: 多智能体强化学习是否可以在随机博弈的框架下与随机性演化博弈动力学进行融合?特别地, 如果该问题可以放在具有网络空间结构的博弈者群体中进行考虑, 那么它将变得更加一般化.

为了对该问题进行研究, 文献 [206] 提出了一个网络上的多人随机博弈模型. 在该模型中, 每个博弈者只有两个可供选择的行动 C 和 D, 即所有博弈者共享同一个行动集 $\mathcal{A}=\{\mathrm{C},\mathrm{D}\}$. 基于博弈者群体的空间交互结构, 每个博弈者只能与其直接邻居进行交互并形成一个 $d$ 人随机博弈. 当系统处于状态 $s$ 时, 依赖当前所有博弈者的行动选择, 表 2 给出了任意博弈者在当前轮博弈中获得的收益. 在该表中, $\mathfrak{a}_\jmath(s)$ 和 $\mathfrak{b}_\jmath(s)$, $\jmath=0,1,\cdots,d-1$ 分别表示一个行动为 C 和 D 的博弈者, 当其 $d-1$ 个共同博弈者中有 $\jmath$ 个个体选择行动 C 且系统状态为 $s\in\mathcal{S}$ 时, 其获得的收益. 在该博弈进行过程中, 博弈者的行动决策被假定是异步的, 即在每个时间步, 群体中只有一个随机选定的博弈者更新其行动. 当该选定的博弈者进行行动更新时, 它的策略函数被设定为如下的一个 Soft-max 形式

$$\pi(s,\jmath,a;\theta,\beta)=\frac{\exp(\beta\theta^T\phi_{s,\jmath,a})}{\sum_{a'\in\mathcal{A}}\exp(\beta\theta^T\phi_{s,\jmath,a'})},\qquad(7)$$

其中 $\pi(s,\jmath,a;\theta,\beta)$ 表示当系统状态为 $s$ 且 $d-1$ 个共同博弈者中有 $\jmath$ 个个体选择行动 C 时, 行动更新的博弈者选择行动 $a\in\mathcal{A}$ 的概率, $\beta\in[0,+\infty)$ 表示选择强度参数或适应率参数[198], $\theta\in\mathbb{R}^L$ 是一个 $L$ 维的列向量, 表示策略学习参数, $\phi_{s,\jmath,a}\in\mathbb{R}^L$ 是一个 $L$ 维的特征向量, 刻画了当系统状态为 $s$ 且 $d-1$ 个共同博弈者中有 $\jmath$ 个个体选择行动 C 时, 行动更新的博弈者采取行动 $a\in\mathcal{A}$ 的特征. 在博弈进行过程中, 式 (7) 中的学习参数 $\theta$ 将根据每轮博弈的收益结果以 Actor-critic 强化学习[207−208] 的方式进行实时更新. 特别地, 文献 [206] 通过分析博弈者群体决策的长期演化结果发现, 在弱选择条件下, 群体中行动为 C 的博弈者比例大于行动为 D 的博弈者比例当且仅当存在一组依赖状态 $s\in\mathcal{S}$ 的矩阵 $\Phi_s\in\mathbb{R}^{L\times d|\mathcal{S}|}$ 和 $\Lambda_s\in\mathbb{R}^{L\times d|\mathcal{S}|}$ 以及一个常系数 $e$ 使得

$$\sum_{s\in\mathcal{S}}\mu_\pi(s)\theta^{*T}\Phi_s\boldsymbol{Y}+\sum_{s\in\mathcal{S}}\mu_\pi(s)\theta^{*T}\Lambda_s\boldsymbol{Z}+e>0,$$

其中 $\boldsymbol{Y}:=[\vec{\mathfrak{a}}(s^1),\vec{\mathfrak{a}}(s^2),\cdots,\vec{\mathfrak{a}}(s^{|\mathcal{S}|})]^T$ 和 $\boldsymbol{Z}:=[\vec{\mathfrak{b}}(s^1),\vec{\mathfrak{b}}(s^2),\cdots,\vec{\mathfrak{b}}(s^{|\mathcal{S}|})]^T$, $|\mathcal{S}|$ 表示集合 $\mathcal{S}$ 的势, $\vec{\mathfrak{a}}(s^\imath):=[\mathfrak{a}_0(s^\imath),\mathfrak{a}_1(s^\imath),\cdots,\mathfrak{a}_{d-1}(s^\imath)]$, $\vec{\mathfrak{b}}(s^\imath):=[\mathfrak{b}_{d-1}(s^\imath),\mathfrak{b}_{d-2}(s^\imath),\cdots,\mathfrak{b}_0(s^\imath)]$, $\forall s^\imath\in\mathcal{S}$, $\imath=1,2,\cdots,|\mathcal{S}|$, $\mu_\pi(\cdot)$ 表示系统状态在策略 $\pi$ 下的平稳分布, $\theta^*$ 是学习参数 $\theta$ 在博弈进行无穷轮后的极限值 (Actor-critic 强化学习算法的收敛性保证了 $\theta$ 极限值的存在). 此外, 通过比较有无学习机制作用下的数值仿真结果, 该文还发现, 学习机制的存在能够有效提升博弈者在动态社会困境博弈环境中的适应性.

## 2.3　强化学习

强化学习是人和动物普遍使用的一种学习方式, 它在不同的学科范畴内 (比如认知神经科学和计算机科学) 具有不同的解释和定义[197]. 但如果仅

表 2　$d$ 人随机博弈的收益表

Table 2　The payoff table of $d$-player stochastic games

| $d-1$ 个共同博弈者中行动为 C 的博弈者个数 | $d-1$ | $\cdots$ | $\jmath$ | $\cdots$ | 0 |
|---|---|---|---|---|---|
| 行动为 C 的博弈者的收益 | $\mathfrak{a}_{d-1}(s)$ | $\cdots$ | $\mathfrak{a}_\jmath(s)$ | $\cdots$ | $\mathfrak{a}_0(s)$ |
| 行动为 D 的博弈者的收益 | $\mathfrak{b}_{d-1}(s)$ | $\cdots$ | $\mathfrak{b}_\jmath(s)$ | $\cdots$ | $\mathfrak{b}_0(s)$ |

从计算的角度来讲, 一般认为强化学习的数学基础是 Markov 决策过程. 正如图 1 所示, 标准的 Markov 决策过程是随机博弈在单智能体场景下的退化情形. 它主要由两部分构成, 一部分称为智能体, 另一部分称为 "系统" (或"环境"). 在每一离散决策时刻 $t \geq 0$, 系统将处于状态集 $\mathcal{S}$ 中的某一个状态 $s_t \in \mathcal{S}$. 随后, 智能体将根据它的策略 $\pi : \mathcal{S} \to \mathcal{A}$ 从其行动空间 $\mathcal{A}$ 中选择一个行动 $\pi(s_t) = a_t \in \mathcal{A}$ 并执行. 作为该行动的一个直接结果, 系统将以概率 $P(s_{t+1} \mid s_t, a_t)$ 从当前状态 $s_t$ 转移到下一时刻的状态 $s_{t+1}$. 同时, 作为上述行动和系统状态转移的共同结果, 智能体将从当前的决策中获得一个即时的收益值或回报值 $r(s_t, a_t, s_{t+1})$, 其中 $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ 是该智能体的收益函数或回报函数. 上述决策过程随着时间不断地重复进行, 从而构成了一个有限时间或者无限时间的 Markov 决策过程, 这里的"Markov"是指系统状态的转移具有 Markov 属性[169].[10]

在上述序贯决策过程中, 智能体的目标是找到一个策略 $\pi : \mathcal{S} \to \mathcal{A}$ 使得其长期的累积期望收益 $V_\pi(s)$, $\forall s \in \mathcal{S}$ 最大化. 与式 (3) 相类似, 为了保证 $V_\pi(s)$ 是一个有界值函数, 其中一种常用计算方式为如下的折扣形式

$$\begin{aligned} V_\pi(s) = \mathbf{E}_{a_t \sim \pi(s_t), s_{t+1} \sim P(\cdot|s_t, a_t)} & \left\{ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}) \mid s_0 = s \right\} = \\ & \sum_{s' \in \mathcal{S}} P(s'|s, \pi(s)) \left[ r(s, \pi(s), s') + \right. \\ & \left. \gamma V_\pi(s') \right], \quad \forall s \in \mathcal{S}. \end{aligned}$$

因为 $V_\pi(s)$ 是一个关于初始状态 $s \in \mathcal{S}$ 的函数, 所以在强化学习的文献中, 它也称为状态值函数. 为了找到值函数 $V_\pi(s)$ 的最优策略 $\pi^*$, 即 $\pi^* \in \arg\max_{\pi \in \Omega} V_\pi(s)$, $\forall s \in \mathcal{S}$ (其中 $\Omega$ 是 $\pi$ 的可行集), 传统上一种常用的方法是动态规划方法[210]. 以该类方法中的值迭代 (Value iteration)[196−197, 211] 为例, 它的一个基本思路是: 首先利用一个值函数迭代过程

$$\begin{aligned} V(s) \leftarrow \max_{a \in \mathcal{A}} & \sum_{s' \in \mathcal{S}} P(s'|s, a) \left[ r(s, a, s') + \right. \\ & \left. \gamma V(s') \right], \ \forall s \in \mathcal{S}, \end{aligned} \tag{8}$$

来获得最优策略 $\pi^*$ 对应的值函数 $V^*(s) := V_{\pi^*}(s) = \max_{\pi \in \Omega} V_\pi(s)$, $\forall s \in \mathcal{S}$; 然后, 借助得到的 $V^*(s)$ 计算一个 Q 值函数 $Q^*(s, a) := Q_{\pi^*}(s, a)$, 这里 $Q_{\pi^*}(s, a)$ 是策略 $\pi^*$ 对应的 Q 值函数, 其定义为

$$\begin{aligned} Q_{\pi^*}(s, a) := \mathbf{E}_{s' \sim P(\cdot|s,a)} & \left\{ r(s, a, s') + \right. \\ & \left. \gamma V_{\pi^*}(s') \mid s_0 = s, a_0 = a \right\} = \\ & \sum_{s' \in \mathcal{S}} P(s'|s, a) \left[ r(s, a, s') + \right. \\ & \left. \gamma V^*(s') \right], \quad \forall s \in \mathcal{S}, \forall a \in \mathcal{A}. \end{aligned}$$

因为 $V^*(s)$ 满足 Bellman 最优性方程[210]

$$\begin{aligned} V^*(s) = \max_{a \in \mathcal{A}} & \sum_{s' \in \mathcal{S}} P(s'|s, a) \left[ r(s, a, s') + \right. \\ & \left. \gamma V^*(s') \right], \quad \forall s \in \mathcal{S}, \end{aligned}$$

所以 $V^*(s) = \max_{a \in \mathcal{A}} Q^*(s, a)$. 因此, 最优策略 $\pi^*$ 可选取为 $\pi^*(s) \in \arg\max_{a \in \mathcal{A}} Q^*(s, a)$, $\forall s \in \mathcal{S}$.

### 2.3.1 单智能体强化学习

虽然动态规划方法能有效地找到 Markov 决策过程的最优策略, 但由式 (8) 可以发现, 该类方法需要明确的模型信息, 比如回报函数 $r(\cdot)$ 和状态转移概率分布函数 $P(\cdot)$. 然而, 在大量工程应用, 比如自动驾驶、无人机编队和多机器人协作中, 明确的模型信息通常是一种理想化的条件. 为了弥补这一不足, 学者们于是提出了一类无模型的 (Model-free) 强化学习方法[197, 211−212], 即智能体利用实时与外部环境交互所获得的激励信号来学习一个最优策略. 基于算法迭代对象的不同, 目前主流的强化学习方法主要可分为两类[49−50]: 基于值函数的 (Value-based) 方法和基于策略的 (Policy-based) 方法.[11]

迭代一个 Q 值表 $Q(s, a)$, $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$ 使得其最终可以精确地或近似地收敛到 $Q^*(s, a)$ 是基于值函数方法的一个基本思想. 在该类方法中, 一个著名的算法是所谓的 Q 学习[85−86], 其迭代表达式为

$$\underbrace{Q(s, a)}_{\text{新的 Q 值表}} \leftarrow \underbrace{Q(s, a)}_{\text{旧的 Q 值表}} + \alpha \overbrace{\underbrace{\left[ r(s, a, s') + \gamma \max_{a' \in \mathcal{A}} Q(s', a') \right.}_{\text{时序差分目标}} - \left. \underbrace{Q(s, a)}_{\text{旧的 Q 值表}} \right]}^{\text{时序差分误差}},$$

其中 $\alpha > 0$ 为学习率参数, $r(s, a, s')$ 表示智能体采取行动 $a$ 并且系统状态从 $s$ 转移到 $s'$ 后获得的收益值 (它是系统 (或环境) 反馈给智能体的激励信号).[12] 然而, 由于随着状态空间和行动空间的增加, $Q$ 值表所需的存储空间和计算资源会呈指数增长,

[10] 从控制论的角度上讲, Markov 决策过程中的智能体一般被视为是控制器, 而"系统" (或"环境") 则被视为是受控对象. 因此, Markov 决策过程也称为"受控 Markov 过程"[209].

[11] 文献 [197] 将目前的强化学习方法划分为表格解方法 (Tabular solution method) 和近似解方法 (Approximate solution method).

[12] 在强化学习算法中, 函数 $r(\cdot)$ 的具体形式并不需要明确的已知, 它的某一个取值 $r(s, a, s')$ 是系统 (或环境) 反馈给智能体的, 并且智能体并不知道 $r(\cdot)$ 的具体形式. 为了行文简洁, 本文并没有在符号形式上突出这一点.

所以该类方法通常无法处理大规模 (Large-scale) 决策问题. 为了弥补这一不足, 目前一种可行的方法是在基于值函数的方法中引入神经网络作为函数近似 (即深度强化学习)[90, 213] 或者使用 Monte-Carlo 树搜索的方法[214−215]. 除此之外, 另一种可行的方法是基于策略的方法.

该方法的一个基本思想是, 利用参数 $\theta$ 将策略函数 $\pi$ 参数化为 $\pi_\theta$ 以便在整个策略空间中搜索最优的策略. 基于该思想, 一个基本的算法是梯度上升算法, 即沿着策略 $\pi_\theta$ 对应的长期累积期望回报 $J(\theta) := \mathbf{E}_{a_t\sim\pi_\theta(s_t),s_{t+1}\sim P(\cdot|s_t,a_t)}\{\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t,s_{t+1})\}$ 的梯度方向来搜索最优的策略

$$\theta \leftarrow \theta + \alpha\nabla_\theta J(\theta). \tag{9}$$

特别地, 在适当的条件下, 式 (9) 中的梯度 $\nabla_\theta J(\theta)$ 可由 Sutton 等推导的策略梯度定理 (Policy gradient theorem)[207−208] 给出, 即

$$\nabla_\theta J(\theta) = \mathbf{E}_{s\sim\mu_{\pi_\theta}(\cdot),a\sim\pi_\theta(s)}\left\{\nabla_\theta \log\pi_\theta(s)\cdot Q_{\pi_\theta}(s,a)\right\}, \tag{10}$$

其中 $\mu_{\pi_\theta}(\cdot)$ 是系统状态在策略 $\pi_\theta$ 下的平稳分布, $Q_{\pi_\theta}(s,a)$ 是策略 $\pi_\theta$ 对应的 Q 值函数. 由于式 (10) 中的 $Q_{\pi_\theta}(s,a)$ 是未知的, 所以在该算法的具体实施过程中, $Q_{\pi_\theta}(s,a)$ 的值通常需要借助一些近似方法来估计. 例如, 在著名的 Actor-critic 算法[207−208] 中, $Q_{\pi_\theta}(s,a)$ 的值在参数 $\theta$ 更新过程中是通过一个选定参数化形式的 Q 值函数 $Q_\omega(s,a)$ 来估计的, 其中参数 $\omega$ 的更新过程是借助时序差分学习 (Temporal-difference learning) 来实现的, 即

$$\omega \leftarrow \omega + \alpha_\omega[r(s,a,s') + \gamma Q_\omega(s',a') - Q(s,a)]\nabla_\omega Q_\omega(s,a),$$

这里 $\alpha_\omega > 0$ 表示参数 $\omega$ 的学习率. 另外, 通过改进上述基于策略梯度定理的梯度上升算法, 最近的相关文献也相继提出了一些其他形式的基于策略的学习算法, 比如确定性策略梯度算法 DPG[216]、深度确定性策略梯度算法 DDPG[217]、信赖域策略优化算法 TRPO[218]、近端策略优化算法 PPO[219]、异步优势 Actor-critic 算法 A3C[220] 和软 Actor-critic 算法 SAC[221] 等.

### 2.3.2 多智能体强化学习

标准的 Markov 决策过程刻画的只是单个智能体面临的序贯决策问题. 如果实际的决策问题是由多个智能体构成, 那么这一模型框架的适用范围将受到极大限制. 有鉴于此, 目前数学上为了描述多智能体序贯决策问题, 一种普遍使用的模型框架是 Markov 决策过程在多智能体场景下的拓展形式, 即随机博弈 (也称为 Markov 博弈). 相应地, 为了寻找随机博弈这类多智能体场景下的最优决策, 上述两类单智能体强化学习算法也可拓展到多智能体情形中.

对应于基于值函数的方法, 如果随机博弈中的博弈者是完全合作的, 即此时的博弈是一个团队随机博弈, 那么单智能体的 Q 学习运用到多智能体场景中只需将单个智能体的行动修改为所有智能体的行动组合[222−223], 即

$$Q(s,\boldsymbol{a}) \leftarrow Q(s,\boldsymbol{a}) + \alpha[r(s,\boldsymbol{a},s') + \gamma\max_{\boldsymbol{a}'\in\boldsymbol{\mathcal{A}}} Q(s',\boldsymbol{a}') - Q(s,\boldsymbol{a})],$$

其中 $\boldsymbol{a}$ 和 $\boldsymbol{\mathcal{A}}$ 分别表示所有智能体的行动组合和行动组合空间. 相较而言, 如果随机博弈中的博弈者是完全竞争的, 即此时的博弈是一个两人零和随机博弈, Minimax-Q 学习是单智能体的 Q 学习运用到完全竞争的博弈环境中的一种拓展形式[35], 即对于博弈者 1, 其 Q 值表的迭代表达式为

$$Q(s,(a_1,a_2)) \leftarrow Q(s,(a_1,a_2)) + \alpha[r(s,(a_1,a_2),s') + \gamma\max_{a_1'\in\mathcal{A}_1}\min_{a_2'\in\mathcal{A}_2} Q(s',(a_1',a_2')) - Q(s,(a_1,a_2))],$$

其中 $a_i$ 和 $\mathcal{A}_i$ 分别表示博弈者 $i$ $(i=1,2)$ 的行动和行动空间. 除此之外, 如果随机博弈中的博弈者既不是完全合作的, 又不是完全竞争的, 即此时的随机博弈为一般和随机博弈, 那么单智能体的 Q 学习运用到多智能体场景中就需要借助非合作博弈论的结果. 例如, 在 Nash-Q 算法[39] 中, 每个博弈者 $i$ 将维持如下一个 Q 值表的迭代

$$Q^i(s,\boldsymbol{a}) \leftarrow Q^i(s,\boldsymbol{a}) + \alpha[r(s,\boldsymbol{a},s') + \gamma\cdot\mathbf{eval}_i\{Q_\bullet(s',\cdot)\} - Q^i(s,\boldsymbol{a})],$$

其中 $Q^i(s,\boldsymbol{a})$ 表示博弈者 $i$ 在状态 $s$ 和所有博弈者行动组合 $\boldsymbol{a}$ 下的 Q 值函数, $\mathbf{eval}_i\{Q_\bullet(s',\cdot)\} := V^i_{\mathbf{NE}\{Q_\bullet(s',\cdot)\}}(s')$, 这里 $Q_\bullet(s',\cdot)$ 表示系统状态为 $s'$ 时的 Q 值表, $\mathbf{NE}\{Q_\bullet(s',\cdot)\}$ 表示通过 $Q_\bullet(s',\cdot)$ 计算得到的 Nash 均衡策略, $V^i_{\mathbf{NE}\{Q_\bullet(s',\cdot)\}}(s')$ 表示当所有博弈者的策略组合为 Nash 均衡策略 $\mathbf{NE}\{Q_\bullet(s',\cdot)\}$ 时, 博弈者 $i$ 从初始状态 $s'$ 出发所能获得的长期累积期望收益. 值得注意的是, 在适当的条件下, 上述三类基于值函数的多智能体强化学习算法在理论上均能保证收敛[35, 39, 222−223].

对应于基于策略的方法, 上述单智能体的梯度上升算法拓展到多智能体场景中的一个基本思路是: 首先将每个博弈者 $i$ 的策略 $\pi_i$ 借助参数 $\theta_i$ 参数化为 $\pi_{i,\theta_i}$, 然后沿着每个博弈者 $i$ 的长期累积期望回报 $J^i(\boldsymbol{\theta}) := \mathbf{E}_{\boldsymbol{a}_t\sim\boldsymbol{\pi_\theta}(s_t),s_{t+1}\sim P(\cdot|s_t,\boldsymbol{a}_t)}\{\sum_{t=0}^{\infty}\gamma^t r_i(s_t,$

$\boldsymbol{a}_t, s_{t+1})\}$ 的梯度方向更新参数 $\theta_i$, 即

$$\theta_i \leftarrow \theta_i + \alpha_i \nabla_{\theta_i} J^i(\boldsymbol{\theta}), \tag{11}$$

其中 $\alpha_i$ 表示对应于博弈者 $i$ 的学习率参数, $\boldsymbol{\theta}$ 表示所有 $\theta_i$ 构成的参数组合, $\boldsymbol{\pi_\theta}$ 表示所有博弈者的策略 $\pi_{i,\theta_i}$ 构成的策略组合. 特别地, 在适当的条件下, 式 (11) 中 $\nabla_{\theta_i} J^i(\boldsymbol{\theta})$ 可由策略梯度定理在多智能体场景下的推广形式给出[40], 即

$$\nabla_{\theta_i} J^i(\boldsymbol{\theta}) = \mathbf{E}_{s\sim\mu_{\boldsymbol{\pi_\theta}}(\cdot),\boldsymbol{a}\sim\boldsymbol{\pi_\theta}(s)} \left\{\nabla_{\theta_i} \log \pi_{i,\theta_i}(s)\cdot Q^i_{\boldsymbol{\pi_\theta}}(s,\boldsymbol{a})\right\},$$

其中 $\mu_{\boldsymbol{\pi_\theta}}(\cdot)$ 表示系统状态在策略组合 $\boldsymbol{\pi_\theta}$ 下的平稳分布, $Q^i_{\boldsymbol{\pi_\theta}}(s,\boldsymbol{a})$ 表示博弈者 $i$ 在策略组合 $\boldsymbol{\pi_\theta}$ 下的 Q 值函数.

虽然多智能体强化学习相比于单智能体强化学习具有一些显著优势, 比如具有更宽的适用范围、更高的鲁棒性和容错性, 并且可以使用通信进行经验分享等[47, 224], 但它也面临着比单智能体强化学习更多更难的挑战, 比如维数灾难 (The curse of dimensionality) 和可扩展性 (Scalability) 问题、非平稳性 (Non-stationarity) 问题、部分可观性 (Partial observability) 和不完全信息 (Incomplete information) 问题、不确定性 (Uncertainty) 和安全性 (Safety and security) 问题、协调 (Coordination) 和异质性 (Heterogeneity) 问题、可解释性 (Explainability) 问题, 以及探索与利用之间的权衡性 (Exploration-exploitation trade-off) 问题等 (这些问题的最新相关研究进展可参见文献 [47, 49−50, 186, 225−232]).

### 2.4 鲁棒学习

标准的单智能体和多智能体强化学习虽然能有效地处理 Markov 决策过程和随机博弈描述的序贯决策问题, 但它们并没有考虑诸如参数摄动、建模误差、外界干扰和结构变化等带来的系统不确定性问题. 为了进一步处理这类含不确定性的序贯决策问题, 目前一种有效的方法是在标准的强化学习算法基础上发展鲁棒方法, 即所谓的鲁棒强化学习[233].[13]

与标准的强化学习方法相类似, 鲁棒强化学习也是一类多领域交叉的研究方法, 它与鲁棒优化、鲁棒控制和博弈论都有着紧密的联系[233−234]. 具体地, 从数学上讲, 它的模型框架是一类称为鲁棒 Markov 决策过程 (也称为不确定 Markov 决策过程) 的模型[235−236]. 所谓鲁棒 Markov 决策过程是指一类含不确定参数的 Markov 决策过程. 例如, 在该类模型中, 一种典型的设置是假定标准的 Markov 决策过程的状态转移概率分布函数 $P$ 不是给定的, 而是属于一个不确定集 $\mathcal{P}$. 在该设置下, 为了获得鲁棒 Markov 决策过程的一个最优策略, 一种常用的方法是鲁棒优化方法[179] 或者鲁棒控制方法[237], 即在不确定参数最坏的情况下求解一个最优策略. 在相关文献中, 该策略也称为鲁棒最优策略. 具体来讲, 对于一个具有不确定状态转移概率分布函数 $P\in\mathcal{P}$ 的 Markov 决策过程, $\pi^*_r$ 是其一个鲁棒最优策略当且仅当

$$\pi^*_r(s) \in \arg\max_{\pi\in\Omega}\left\{\min_{P\in\mathcal{P}} \mathbf{E}_{a_t\sim\pi(s_t),s_{t+1}\sim P(\cdot|s_t,a_t)} \left\{\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t,s_{t+1}) \mid s_0=s\right\}\right\},\ \forall s\in\mathcal{S}. \tag{12}$$

传统上, 为了找到该鲁棒最优策略, 一种常用的方法是鲁棒动态规划方法[238−239], 比如鲁棒值迭代 (Robust value iteration, RVI) 算法和鲁棒策略迭代 (Robust policy iteration, RPI) 算法[238−239]. 具体地, 对于 RVI 算法, 它的值函数迭代表达式为

$$V(s) \leftarrow \max_{a\in\mathcal{A}}\left\{\min_{P\in\mathcal{P}}\sum_{s'\in\mathcal{S}} P(s'|s,a)\left[r(s,a,s')+\gamma V(s')\right]\right\},\ \forall s\in\mathcal{S}. \tag{13}$$

而对于 RPI 算法, 它的策略迭代过程主要包含两步: 第 1 步借助 Bellman 方程

$$V(s) = \min_{P\in\mathcal{P}}\left\{\sum_{s'\in\mathcal{S}} P(s'|s,\pi(s))\left[r(s,\pi(s),s')+\gamma V(s')\right]\right\},\ \forall s\in\mathcal{S},$$

求解一个给定策略 $\pi$ 对应的值函数 $V(s)$, $\forall s\in\mathcal{S}$ (该步骤的求解可通过标准的动态规划方法来实现); 第 2 步利用得到的值函数 $V(s)$ 获得一个提升的策略 $\pi'$, 即

$$\pi'(s) \in \arg\max_{a\in\mathcal{A}}\left\{\min_{P\in\mathcal{P}}\sum_{s'\in\mathcal{S}} P(s'|s,a)\left[r(s,a,s')+\gamma V(s')\right]\right\},\quad \forall s\in\mathcal{S}, \tag{14}$$

然后, 令 $\pi=\pi'$ 并回到第 1 步; 最后, 当 $\pi$ 收敛到鲁棒最优策略时, 迭代停止. 特别地, 上述 RPI 算法的思想也是著名的生成对抗网络 GAN[240] 的基本思想. 虽然当不确定集 $\mathcal{P}$ 满足“垂直性” (Rectangularity property) 时 (即对不同的状态—行动对 $(s,a)$, $P(\cdot\mid s,a)$ 所在的不确定集是相互独立的并且与历史访问的状态和行动无关), RVI 算法和 RPI 算法都

[13] 本文这部分主要专注介绍鲁棒强化学习, 未涉及鲁棒学习中的诸如鲁棒深度学习等其他内容.

将收敛到一个鲁棒最优策略[238−239], 但它们也存在着一些不足之处. 例如, RVI 算法通常收敛较慢, 而 RPI 算法由于需要求解 Bellman 方程, 计算量通常较大. 为了进一步优化这两个算法, 文献 [241] 提出了鲁棒修正策略迭代 (Robust modified policy iteration, RMPI) 算法, 进而给出了这两个算法的一个统一形式.

由式 (13) 和式 (14) 可以发现, 鲁棒动态规划方法本质上是一种极大极小化的方法, 或者说它是一种在最坏情况下求最优解的方法. 因为这类方法只考虑在不确定参数最坏的情况下求解一个最优性能指标, 所以它是一种比较保守的 (Conservative) 方法. 为了弥补这一不足, 后续的相关研究在这方面进行了一些改进. 从方法论上讲, 它们主要可分为 3 类. 第 1 类为折衷 (Trade-off) 的方法[242−244]. 该类方法的基本思想是, 首先将不确定参数名义上的 (Nominal) 性能指标和最坏情况下的性能指标进行组合, 然后将组合后的结果作为最终的目标性能指标. 第 2 类方法主要以放宽垂直性假设为目标. 例如, 通过放宽垂直性假设, 文献 [245] 和文献 [246] 分别提出了一类 $k$ 垂直性不确定集 ($k$-rectangular uncertainty set) 方法和一类因子矩阵不确定集 (Factor matrix uncertainty set) 方法. 第 3 类为分布鲁棒优化 (Distributionally robust optimization) 方法[247−248]. 该类方法的一个基本思想是, 在不确定参数 $P$ 的概率分布 $D_p$ 属于不确定集 $\mathcal{U}_d$ 的情况下, 将鲁棒最优策略 $\pi_r^*$ 所需满足的条件 (12) 修改为

$$\begin{aligned}\pi_r^*(s) \in \arg\max_{\pi\in\Omega} \min_{D_p\in\mathcal{U}_d} \mathbf{E}_{P\sim D_p}\\ \left\{\mathbf{E}_{a_t\sim\pi(s_t),s_{t+1}\sim P(\cdot|s_t,a_t)}\left\{\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t,s_{t+1})\right.\right.\\ \left.\left.\mid s_0=s\right\}\right\},\ \forall s\in\mathcal{S}.\end{aligned} \tag{15}$$

由式 (15) 可以看出, 分布鲁棒优化方法将最坏情况下的性能指标修改成了风险最小化 (Risk minimization) 或风险厌恶 (Risk averse) 的性能指标.

然而, 与标准的动态规划方法相类似, 一方面, 当 Markov 决策过程的状态空间和行动空间非常大时, 鲁棒动态规划方法同样会面临维数灾难问题; 另一方面, 鲁棒动态规划方法也需要准确的模型信息. 为了弥补这两方面的不足, 目前一种有效的方法是基于鲁棒动态规划方法发展无模型的强化学习方法. 例如, 通过引入线性函数近似, 文献 [249] 和文献 [250] 分别提出了一类鲁棒近似动态规划算法和一类鲁棒近似修正策略迭代学习算法; 而文献 [251] 提出了一类鲁棒最小二乘策略评估学习算法和一类鲁棒最小二乘策略迭代学习算法. 除了线性函数近似方法之外, 文献 [252−254] 和文献 [255] 通过借助深度神经网络这一非线性函数近似方法, 分别提出了一类鲁棒对抗强化学习算法和一类鲁棒最大后验策略优化算法. 另外, 为了求解分布鲁棒优化意义下的鲁棒最优策略, 文献 [256] 和文献 [257] 分别提出了一类分布鲁棒策略迭代学习算法和一类分布鲁棒离线强化学习算法.

上述介绍的几类鲁棒强化学习算法虽然能较好地弥补鲁棒动态规划方法关于模型信息要求的不足, 但它们大部分考虑的问题仍是单智能体的序贯决策问题. 与标准的 Markov 决策过程在多智能体场景下的随机博弈形式相类似, 鲁棒随机博弈是鲁棒 Markov 决策过程在多智能体场景下的拓展形式. 为了研究这类多智能体场景下的不确定序贯决策问题, 当前一些相关工作也提出了一些鲁棒多智能体强化学习算法. 例如, 为了找到鲁棒随机博弈的鲁棒 Markov 完美均衡, 文献 [258] 分别提出了一个鲁棒多智能体 Q 学习算法和一个鲁棒多智能体 Actor-critic 学习算法. 而文献 [259] 考虑了一类鲁棒团队随机博弈 $\mathcal{G}_{rts}=\{\mathcal{N},\mathcal{S},\{\mathcal{A}_i\}_{i\in\mathcal{N}},\bar{r},\mathcal{P},\Gamma\}$, $\mathcal{N}:=\{1,2,\cdots,n\}$. 相较于标准的随机博弈 $\mathcal{G}_s=\{\mathcal{N},\mathcal{S},\{\mathcal{A}_i\}_{i\in\mathcal{N}},\{r_i\}_{i\in\mathcal{N}},P,\Gamma\}$, 该文提出的鲁棒团队随机博弈 $\mathcal{G}_{rts}$ 具有两方面的显著差异: 一方面, 博弈中的状态转移概率分布函数 $P$ 不是给定的, 而是属于不确定集 $\mathcal{P}$, 即 $P\in\mathcal{P}$; 另一方面, 对于任意的博弈者 $i\in\mathcal{N}$, 它的决策目标是在 $P\in\mathcal{P}$ 最坏的情况下, 最大化整个团队的平均收益 $\bar{r}=\sum_{i=1}^n r_i/n$ 的长期累积期望值 $\overline{V}_{(\pi_i,\pi_{-i})}(s;P)$, 而不是最大化其自身的收益 $r_i$ 的长期累积期望值 $V^i_{(\pi_i,\pi_{-i})}(s;r_i,P)$. 类似于 $V^i_{(\pi_i,\pi_{-i})}(s;r_i,P)$ 的计算表达式 (3), $\overline{V}_{(\pi_i,\pi_{-i})}(s;P)$ 的计算表达式为

$$\begin{aligned}\overline{V}_{(\pi_i,\pi_{-i})}(s;P)=\mathbf{E}_{\boldsymbol{a}_t\sim(\pi_i,\pi_{-i})(s_t),s_{t+1}\sim P(\cdot|s_t,\boldsymbol{a}_t)}\\ \left\{\sum_{t=0}^{\infty}\gamma^t\bar{r}(s_t,\boldsymbol{a}_t,s_{t+1})\mid s_0=s\right\}.\end{aligned}$$

为了求解鲁棒团队随机博弈 $\mathcal{G}_{rst}$ 在鲁棒优化意义下的最优策略 (即在不确定参数 $P$ 最坏情况下的最优团队策略), 该文定义了一个称为 “鲁棒团队最优策略” (Robust team-optimal policy) 的博弈解概念, 即策略组合 $\boldsymbol{\pi}^*:=(\pi_1^*,\pi_2^*,\cdots,\pi_n^*)$ 是 $\mathcal{G}_{rst}$ 的一个鲁棒团队最优策略当且仅当

$$\boldsymbol{\pi}^*\in\arg\max_{\boldsymbol{\pi}}\left\{\min_{P\in\mathcal{P}}\overline{V}_{\boldsymbol{\pi}}(s;P)\right\},\ \forall s\in\mathcal{S},$$

这里 $\boldsymbol{\pi}:=(\pi_1,\pi_2,\cdots,\pi_n)$. 随后, 为了找到该鲁棒团

队最优策略, 该文基于“集中式学习分布式实施” (Centralized learning with decentralized execution) 的算法框架提出了一个鲁棒团队策略迭代学习算法, 并在适当的条件下证明了该算法的收敛性. 特别地, 相比于 RVI、RPI 和 RMPI 算法, 该文提出的算法不仅能够利用近似计算来缓解大规模决策问题中存在的维数灾难问题, 而且还具有更快的收敛速率.

### 2.5 小结

对应于上节介绍的标准式博弈、演化博弈、随机博弈和不完全信息博弈, 本节主要梳理并介绍了这 4 类基本博弈形式下的多智能体学习方法, 即策略学习、学习动力学、强化学习和鲁棒学习. 对应于标准式博弈, 本节主要介绍了策略学习这类多智能体学习方法, 并重点对虚拟对弈学习过程进行了论述. 虚拟对弈是通过记录对手博弈者过去所采用行动的频率来推断其策略的一种学习方法. 因此, 本质上讲, 它是一类数据驱动的博弈学习方法. 然而, 由于多智能体的学习环境通常会引发非平稳性问题, 所以本节还简要地回顾了一些更为复杂的策略学习方法, 比如忽视、遗忘、响应目标对手和心智理论等. 对应于演化博弈, 本节主要讨论了学习动力学这类交叉研究方法, 并重点介绍了 Cross 学习和 Q 学习在无穷小策略更新步长条件下与复制动力学方程的等价性. 此外, 该部分还简要比较了统计物理和 AI 这两个领域在该主题上的不同研究风格, 并介绍了一个多智能体强化学习与随机性演化博弈动力学相结合的工作. 对应于随机博弈, 本节主要讨论了单智能体强化学习与多智能体强化学习之间的联系与异同. 考虑到 Markov 决策过程是随机博弈在单智能体情形下的退化形式, 这部分首先介绍了几类单智能体强化学习方法; 然后, 基于随机博弈的模型框架, 集中对这几类单智能体强化学习算法在多智能体场景下的拓展形式进行了回顾. 最后, 对于不完全信息博弈, 本节主要介绍了鲁棒 Markov 决策过程和鲁棒随机博弈之间的关系, 以及求解它们的鲁棒动态规划、鲁棒强化学习和鲁棒多智能体强化学习方法.

## 3 博弈、学习与控制的交叉研究

由于目前博弈、学习与控制这一交叉研究领域仍处于发展初期, 不同主题之间的研究工作相对独立, 所以本节将主要按照不同专题的方式梳理并介绍其中的几类典型研究成果, 即基于最优控制的学习与博弈、博弈控制系统、基于矩阵半张量积的博弈控制论、分布式 Nash 均衡搜索和基于零行列式策略的收益控制.

### 3.1 基于最优控制的学习与博弈

最优控制[237, 260–261] 是一类在特定的约束条件下, 寻找控制输入使得动态系统的性能指标达到最优的数学方法. 传统上, 该类方法的两个基本数学原理分别是 Pontriagin 的极大值 (或者极小值) 原理和 Bellman 的动态规划理论[262], 其中前者为最优控制问题的最优解提供了必要条件, 而后者通过求解 Hamilton-Jacobi-Bellman (HJB) 方程为最优控制问题的最优解提供了充分条件. 然而, 由于这些传统的方法一般都需要系统的完全信息, 所以它们通常无法处理系统动力学存在不确定性的情况. 为了弥补这一不足, 目前一种有效的方法是在最优控制问题的求解过程中引入学习的技术, 比如连续时间场景下的强化学习算法[54–55, 263] 和自适应动态规划方法[264–265] 等. 该研究思路也是经典的 “学习控制系统” (Learning control system) 的基本思路[266].

具体地, 考虑一个连续时间动态系统[14]

$$\dot{x}(t) = f(x(t), u(t)), \tag{16}$$

其中 $x \in \mathcal{D}_x \subset \mathbb{R}^\kappa$ 表示系统的状态, $u \in \mathcal{D}_u \subset \mathbb{R}^\lambda$ 表示系统的控制输入或控制策略, $f : \mathcal{D}_x \times \mathcal{D}_u \to \mathbb{R}^\kappa$ 为一向量值函数. 基于动力学约束 (16), 最优控制求解的问题是, 寻找一个控制输入 $u(t) = \mu(x(t))$ 使得系统从初始状态 $x(t_0)$ 出发最小化整个时间区间 $[t_0, \infty)$ 内的累积成本

$$V_\mu(x(t_0)) = \int_{t_0}^{\infty} \exp\left(-\frac{t-t_0}{\tilde{\gamma}}\right) \psi(x(t), u(t))\mathrm{d}t,$$

其中 $V_\mu(x(t_0))$ 表示在控制策略 $u(t) = \mu(x(t))$ 下从初始状态 $x(t_0)$ 出发的值函数; $\tilde{\gamma}$ 为一正常数, 表示未来成本的折现率; 对应于系统状态 $x(t)$ 和控制输入 $u(t)$, $\psi(x(t), u(t))$ 表示系统在连续时间 $t$ 的即时成本. 对应于初始状态 $x(t_0)$, 记上述最优控制问题的最优控制输入为 $u^*(t) = \mu^*(x(t))$. 于是, 最优值函数 $V^*(x(t_0))$ 可写为

$$\begin{aligned} V^*(x(t_0)) = & \\ & \min_{u(t)} \left\{ \int_{t_0}^{\infty} \exp\left(-\frac{t-t_0}{\tilde{\gamma}}\right) \psi(x(t), u(t))\mathrm{d}t \right\} = \\ & \int_{t_0}^{\infty} \exp\left(-\frac{t-t_0}{\tilde{\gamma}}\right) \psi(x(t), u^*(t))\mathrm{d}t. \end{aligned} \tag{17}$$

将式 (17) 中的积分项按时间区间 $[t_0, t_0 + \Delta t]$ 和

[14] 除非特别说明, 本节中的符号 $t$ 专指连续时间, 而不再表示博弈进行的轮次.

$[t_0+\Delta t,\infty]$ 划分为两部分之和, 随后利用最优性原理并令 $\Delta t\to 0$ 可得 HJB 方程为

$$\frac{1}{\tilde{\gamma}}V^*(x(t))=\min_{u(t)}\left\{\psi(x(t),u(t))+\frac{\partial V^*(x(t))}{\partial x(t)}f(x(t),u(t))\right\}. \tag{18}$$

因为最优控制 $u^*(t)=\mu^*(x(t))$ 是值函数 $V^*$ 对应的控制输入, 所以如果一旦知道 $V^*$, 那么由式 (18) HJB 方程可得

$$u^*(t)=\mu^*(x(t))=\arg\min_{u(t)}\left\{\psi(x(t),u(t))+\frac{\partial V^*(x(t))}{\partial x(t)}f(x(t),u(t))\right\}.$$

然而, 由式 (18) 可以看出, 获取值函数 $V^*$ 需要求解一个偏微分方程, 这在数学上通常是困难的. 为了解决这一难题, 目前一种有效的做法是利用强化学习算法来获得 HJB 方程的一个近似解. 例如, 对于上述连续时间最优控制问题, 为了找到最优控制输入 $u^*(t)$, 文献 [263] 提出了一个连续时间 Actor-critic 强化学习算法和一个基于值函数的梯度学习算法. 另外, 为了验证这两个算法的有效性, 该文还分别在几个连续时间控制任务上进行了数值仿真并获得了良好的效果. 特别地, 当上述最优控制问题中的 $\tilde{\gamma}$、$\psi(x(t),u(t))$ 和 $f(x(t),u(t))$ 分别取为 $\tilde{\gamma}=\infty$、$\psi(x(t),u(t))=Q(x(t))+u^T(t)Ru(t)$ 和 $f(x(t),u(t))=\zeta(x(t))+\varphi(x(t))u(t)$ 时, 文献 [267] 研究了该类系统的最优调节问题 (Optimal regulation problem), 即是否可以找到一个最优控制输入使得值函数最小的同时系统的状态可收敛为 $\mathbf{0}$, 其中 $\mathbf{0}$ 表示适当维数的全 0 向量, $Q(x(t))$ 和 $R$ 分别为非负实数和正定矩阵 ($Q(x(t))$ 只当 $x(t)=0$ 时才为 0), $\zeta(x(t))\in\mathbb{R}^{\kappa}$ 和 $\varphi(x(t))\in\mathbb{R}^{\kappa\times\lambda}$ 分别表示系统的漂移动态 (Drift dynamics) 和输入动态 (Input dynamics). 为了求解这一问题, 该文基于 Actor-critic 算法架构并借助神经网络作为值函数近似, 提出了一个同策略 (On-policy) 积分强化学习算法, 并且在适当的条件下证明了该算法的收敛性. 然而, 由于同策略的算法在训练时需要大量的数据样本, 所以在实际应用中, 该类算法的数据利用率通常较低. 为了弥补这一不足, 文献 [268] 在文献 [267] 的基础上引入了经验回放 (Experience replay) 技术, 提出了一个带经验回放的同策略积分强化学习算法. 而文献 [269] 通过引入异策略 (Off-policy) 方法, 提出了一个异策略积分强化学习算法.

上述连续时间最优控制问题处理的只是单个智能体面临的控制决策问题. 当系统的智能体数由单个变为多个时, 该单智能体最优控制问题可转变为一个微分博弈 (Differential game)[270] 问题. 一般地, 一个 $n$ 人微分博弈通常可描述为如下的一个多智能体最优控制问题[271−272]

$$\begin{aligned}&\min_{u_i(t)}\ \boldsymbol{J}_i=K_i(x_v(t_f),t_f)+\\&\int_0^{t_f}L_i(x_v(t),u_1(t),\cdots,u_n(t),t)\mathrm{d}t,\\&i=1,2,\cdots,n,\\&\text{s.t. }\dot{x}_v(t)=f(x_v(t),u_1(t),\cdots,u_n(t),t),\\&x_v(0)=x_v^0,\end{aligned} \tag{19}$$

其中 $\boldsymbol{J}_i$ 表示博弈者 $i$ $(i=1,2,\cdots,n)$ 的长期代价函数, $u_i(t)\in\mathbb{R}^{\lambda_i}$ 表示博弈者 $i$ 在连续时间 $t$ 的控制输入, $x_v(t):=(x_1(t),x_2(t),\cdots,x_n(t))$, $x_i(t)\in\mathbb{R}^{\kappa_i}$ 表示博弈者 $i$ 在连续时间 $t$ 的状态, $K_i(\cdot)$ 和 $L_i(\cdot)$ 分别表示博弈者 $i$ 在终止时刻 $t_f$ 的代价函数 (Terminal cost function) 和整个时间区间 $[0,t_f]$ 的运行代价函数 (Running cost function), 并且它们都是关于各自自变量的连续函数.

特别地, 考虑上述微分博弈的一个二次型情形, 即每个博弈者 $i$ 的长期代价函数和系统动力学分别为

$$\boldsymbol{J}_i=\int_0^{\infty}Q_i(x_v(t))+\sum_{j=1}^{n}u_j^T(t)R_{ij}u_j(t)\mathrm{d}t,\quad i=1,2,\cdots,n,$$

和

$$\dot{x}_v(t)=\boldsymbol{\zeta}(x_v(t))+\sum_{j=1}^{n}\boldsymbol{\varphi}_j(x_v(t))u_j(t),$$

其中 $\boldsymbol{\zeta}(\cdot)\in\mathbb{R}^{\sum_{i=1}^n\kappa_i}$, $\boldsymbol{\varphi}_j(\cdot)\in\mathbb{R}^{\sum_{i=1}^n\kappa_i\times\lambda_j}$, $\boldsymbol{\zeta}(\mathbf{0})=\mathbf{0}$, $\boldsymbol{\zeta}(x_v(t))+\sum_{j=1}^n\boldsymbol{\varphi}_j(x_v(t))u_j(t)$ 是局部 Lipschitz 的, 对任意的 $i=1,2,\cdots,n$, $Q_i(\cdot)$ 和 $R_{ii}$ 分别是非负实数和正定矩阵, 对于 $i\neq j$, $R_{ij}$ 是非负定矩阵. 对于该二次型微分博弈, 首先定义博弈者 $i$ 在控制组合 $(u_1,u_2,\cdots,u_n)$ 下对应于系统状态 $x_v(t)$ 的值函数 $V^i(x_v(t))$ 为

$$V^i_{(u_1,\cdots,u_n)}(x_v(t))=\int_t^{\infty}Q_i(x_v(\epsilon))+\sum_{j=1}^{n}u_j^T(\epsilon)R_{ij}u_j(\epsilon)\mathrm{d}\epsilon,\quad i=1,2,\cdots,n, \tag{20}$$

其中 $V^i_{(u_1,\cdots,u_n)}(\mathbf{0})=0$, $i=1,2,\cdots,n$. 随后在式 (20) 等号两端分别关于 $t$ 求微分, 整理之后可得

$$Q_i(x_v(t))+\sum_{j=1}^{n}u_j^T(t)R_{ij}u_j(t)+$$
$$\frac{\partial V_{(u_1,\cdots,u_n)}^{i^T}(x_v(t))}{\partial x_v(t)}[\boldsymbol{\zeta}(x_v(t))+$$
$$\sum_{j=1}^{n}\boldsymbol{\varphi}_j(x_v(t))u_j(t)]=0.$$

将上式等号左端的项定义为 Hamiltonian 函数

$$H_i\left(x_v,u_1,\cdots,u_n,\frac{\partial V_{(u_1,\cdots,u_n)}^{i^T}(x_v(t))}{\partial x_v(t)}\right)=$$
$$Q_i(x_v(t))+\sum_{j=1}^{n}u_j^T(t)R_{ij}u_j(t)+$$
$$\frac{\partial V_{(u_1,\cdots,u_n)}^{i^T}(x_v(t))}{\partial x_v(t)}[\boldsymbol{\zeta}(x_v(t))+$$
$$\sum_{j=1}^{n}\boldsymbol{\varphi}_j(x_v(t))u_j(t)],$$

于是, 利用 Pontriagin 的极小值原理, 博弈者 $i$ 的最优控制输入 $u_i^*(t)$ 可写为

$$u_i^*(t)=$$
$$\arg\min_{u_i(t)}H_i\left(x_v,u_1,\cdots,u_n,\frac{\partial V_{(u_1,\cdots,u_n)}^{i^T}(x_v(t))}{\partial x_v(t)}\right)=$$
$$-\frac{1}{2}R_{ii}^{-1}\boldsymbol{\varphi}_i^T(x_v(t))\frac{\partial V_{(u_1,\cdots,u_n)}^{i}(x_v(t))}{\partial x_v(t)},$$
$$i=1,2,\cdots,n.$$

最后, 将每个博弈者 $i$ 的最优控制输入 $u_i^*(t)$ 代入到上述 Hamiltonian 函数中, 有耦合的 Hamilton-Jacobi (HJ) 方程为

$$Q_i(x_v(t))+\frac{1}{4}\sum_{j=1}^{n}\frac{\partial V_{(u_1,\cdots,u_n)}^{j^T}(x_v(t))}{\partial x_v(t)}\times$$
$$\boldsymbol{\varphi}_j(x_v(t))R_{jj}^{-1}R_{ij}R_{jj}^{-1}\boldsymbol{\varphi}_j^T(x_v(t))\times$$
$$\frac{\partial V_{(u_1,\cdots,u_n)}^{j}(x_v(t))}{\partial x_v(t)}+\frac{\partial V_{(u_1,\cdots,u_n)}^{i^T}(x_v(t))}{\partial x_v(t)}\left[\boldsymbol{\zeta}(x_v(t))-\right.$$
$$\left.\frac{1}{2}\sum_{j=1}^{n}\boldsymbol{\varphi}_j(x_v(t))R_{jj}^{-1}\boldsymbol{\varphi}_j^T(x_v(t))\frac{\partial V_{(u_1,\cdots,u_n)}^{j}(x_v(t))}{\partial x_v(t)}\right]=0.$$

类似地, 考虑到该耦合的 HJ 方程为一个偏微分方程, 文献 [273] 基于 Actor-critic 算法框架提出了一个策略迭代强化学习算法用于求解其近似解, 并证明了该算法将收敛到上述二次型微分博弈的一个 Nash 均衡.

虽然上述标准的微分博弈模型将单智能体最优控制问题推广到了多智能体系统中, 但它并没有考虑博弈者之间的空间交互关系. 为了弥补这一不足, 文献 [274−275] 进一步将基于微分博弈的强化学习框架扩展到了图博弈 (或网络博弈) 的研究中. 另外, 考虑到在自然界和人类社会中, 智能体之间的地位和分工并不是等同的, 而是具有层级结构的, 文献 [276] 研究了一类单领航者多跟随者的非线性系统 Stackelberg 博弈, 并基于值函数迭代提出了一个用于求解该博弈均衡策略的两层积分强化学习算法.

### 3.2 博弈控制系统

博弈控制系统 (Game-based control system, GBCS)[277−280] 是由两类不同的智能体组成的一个两层控制系统, 其中上层是由单个智能体构成的宏观调控者 (Regulator), 下层是由 $n$ 个博弈者形成的一个非合作微分博弈. 在进行决策时, 上层宏观调控者首先进行决策, 随后下层的 $n$ 个博弈者依从于上层宏观调控者的决策作出最大化自身利益的决策. 换句话讲, 上层宏观调控者并不实际地参与下层博弈者形成的博弈, 而是对下层的博弈进行调控以期达到系统能控 (Controllability) 和可镇定 (Stabilizability) 等整体控制目标. 由于每个下层博弈者在决策时都是追求自身利益最大化, 所以 Nash 均衡通常是下层微分博弈的一个结果.

具体地, 在该博弈控制系统中, 所有智能体状态的动力学方程为

$$\begin{cases}\dot{x}(t)=f(X(t),u(t),U(t),t),\\ \dot{x}_i(t)=f_i(X(t),u(t),U(t),t),\\ x(0)=x_0,\ x_i(0)=x_{i,0},\\ i=1,2,\cdots,n,\ t\in[0,t_f],\end{cases}\tag{21}$$

其中 $X(t):=(x(t),x_1(t),\cdots,x_n(t))$, $U(t):=(u_1(t),\cdots,u_n(t))$, $x(t)\in\mathbb{R}^{\kappa}$ 和 $x_i(t)\in\mathbb{R}^{\kappa_i}$ 分别表示上层宏观调控者和下层博弈者 $i$ 在连续时间 $t$ 的状态, $u(t)\in\mathbb{R}^{\lambda}$ 和 $u_i(t)\in\mathbb{R}^{\lambda_i}$ 分别表示上层宏观调控者和下层博弈者 $i$ 在连续时间 $t$ 的控制输入或控制策略. 每个下层博弈者 $i$ 在连续时间区间 $[0,t_f]$ 内的代价函数 $\boldsymbol{J}_i$ 为

$$\boldsymbol{J}_i=K_i(x_v(t_f),t_f)+\int_0^{t_f}L_i(X(t),U(t),t)\mathrm{d}t,$$
$$i=1,2,\cdots,n,$$

其中 $x_v(\cdot)$、$K_i(\cdot)$ 和 $L_i(\cdot)$ 的定义与式 (19) 中的定义相同.

虽然上述博弈控制系统与标准的 Stackelberg 博弈[281] 在含义上具有一些相似之处, 但由于

博弈控制系统中的上层宏观调控者是一个控制者, 而不是如同标准的 Stackelberg 博弈那样是一个博弈参与者, 所以本质上它们属于两类完全不同的系统[24, 278]. 特别地, 上述博弈控制系统更多地表现为一个控制系统而不仅仅是一个博弈系统. 因此, 对于这类系统, 一个基本的问题是系统能控性问题, 即宏观调控者是否可以通过控制输入 $u(t)$ 使得状态 $x(t)$ 能从任意初始状态 $x(0)=x_0\in\mathbb{R}^\kappa$ 被驱动到任意目标终止状态 $x(t_f)=x_{t_f}\in\mathbb{R}^\kappa$. 为了研究这一问题, 考虑上述博弈控制系统的一个一般线性二次型情形

$$\begin{cases}\dot{x}(t)=A(t)x(t)+B(t)u(t)+\\ \qquad\sum\limits_{i=1}^{n}A_i(t)x_i(t)+\sum\limits_{i=1}^{n}D_i(t)u_i(t),\\ \dot{x}_i(t)=E_i(t)x(t)+B_i(t)u(t)+\\ \qquad\sum\limits_{j=1}^{n}F_{ij}(t)x_i(t)+\sum\limits_{j=1}^{n}B_{ij}(t)u_j(t),\\ x(0)=x_0,\ x_i(0)=x_{i,0},\\ \quad i=1,2,\cdots,n,\ t\in[0,t_f],\end{cases}$$

其中每个下层博弈者 $i$ 在连续时间区间 $[0,t_f]$ 内的代价函数 $\boldsymbol{J}_i$ 为

$$\begin{aligned}\boldsymbol{J}_i=&\frac{1}{2}[x_v(t_f)]^TQ_{i,t_f}x_v(t_f)+\\ &\frac{1}{2}\int_0^{t_f}[X^T(t)Q_i(t)X(t)+u_i^T(t)R_i(t)u_i(t)]\mathrm{d}t,\\ &i=1,2,\cdots,n,\end{aligned}$$

对任意的 $t\in[0,t_f]$, $R_i(t)>0$ 、$Q_i(t)$ 和 $Q_{i,t_f}$ 是对称的, 并且系统中的所有参数矩阵都是关于时间变量 $t$ 的分段光滑的函数. 在适当的条件下, 文献 [278] 首次给出了该一般线性二次型博弈控制系统能控性的充分必要的代数条件.

由式 (21) 可以看出, 上述博弈控制系统是一类确定性的系统. 为了进一步对随机场景下的博弈控制系统进行研究, 文献 [279] 相应地提出了一类随机性博弈控制系统. 在该类系统中, 系统状态的动力学方程和每个下层博弈者 $i$ 在连续时间区间 $[0,t_f]$ 内的代价函数 $\boldsymbol{J}_i$ 分别为

$$\begin{cases}\mathrm{d}x(t)=f(X(t),u(t),t)\mathrm{d}t+\\ \qquad\sigma(X(t),u(t),t)\mathrm{d}w(t),\\ \mathrm{d}x_i(t)=f_i(x(t),x_i(t),u(t),u_i(t),t)\mathrm{d}t+\\ \qquad\sigma_i(x(t),x_i(t),u(t),u_i(t),t)\mathrm{d}w(t),\\ x(0)=x_0,\quad x_i(0)=x_{i,0},\\ i=1,2,\cdots,n,\quad t\in[0,t_f],\end{cases}$$

和

$$\boldsymbol{J}_i=\mathbf{E}\Bigg\{K_i(x_v(t_f),t_f)+\int_0^{t_f}L_i(x(t),u(t),U(t))\mathrm{d}t\Bigg\},\ i=1,2,\cdots,n,$$

其中 $w(t)$, $t\geq0$ 是一个标准的 $d$ 维 Brownian 运动. 为了研究该随机性博弈控制系统的能控性问题, 文献 [279] 分别提出了一个精确 V-能控性 (Exact V-controllability) 和一个完全能控性 (Total controllability) 的概念. 特别地, 对于一类一般线性时变博弈控制系统, 该文给出了此类系统精确 V-能控性的充分必要的代数条件, 而对于一类一般线性时不变博弈控制系统, 该文给出了此类系统完全能控性的充分必要的代数条件.

除了能控性之外, 系统是否可镇定在控制理论中也是一个十分重要的基本问题. 对于一类线性二次型微分博弈控制系统, 文献 [280] 首次研究了此类系统的镇定性问题, 即上层宏观调控者是否可以通过调节下层博弈者形成的 Nash 均衡来实现系统的镇定. 在适当的条件下, 该文首次给出了此类系统可镇定的充分必要的代数条件. 此外, 在系统参数不确定的情况下, 文献 [282−284] 通过结合微分博弈与随机自适应控制理论, 提出了一类随机自适应微分博弈系统, 并在适当的条件下, 给出了该类系统渐近地达到一个 Nash 均衡的稳定性判据.

### 3.3 基于矩阵半张量积的博弈控制论

矩阵半张量积 (Semi-tensor product of matrices)[285] 是对传统矩阵乘法的一般化推广. 它是由中国学者程代展先生首次提出并发展的一套数学方法. 一般地, 对于两个矩阵 $M_1\in\mathcal{M}_{p_1\times q_1}$ 和 $M_2\in\mathcal{M}_{p_2\times q_2}$, 它们的半张量积 $M_1\ltimes M_2$ 定义为如下的一个矩阵乘积形式

$$M_1\ltimes M_2:=(M_1\otimes I_{\hbar/q_1})(M_2\otimes I_{\hbar/p_2}),\qquad(22)$$

其中 $\mathcal{M}_{p_1\times q_1}$ 表示所有 $p_1\times q_1$ 矩阵的集合, $\ltimes$ 表示矩阵的半张量积符号, $\otimes$ 表示矩阵的 Kronecker 积符号, $\hbar=\mathrm{lcm}(q_1,p_2)$ 为 $q_1$ 和 $p_2$ 的最小公倍数, $I_k$, $k\in\mathbb{N}$ 表示 $k\times k$ 单位矩阵. 由定义式 (22) 可以看出, 当 $q_1=p_2$ 时, $M_1$ 与 $M_2$ 的半张量积将退化为通常的矩阵乘积. 特别地, 因为矩阵半张量积在数学上具有一些优良性质, 比如分配律、结合律和伪交换律等[285], 所以该方法目前在逻辑控制系统 (Logical control system)[286]、有限的标准式博弈[287−288] 和博弈控制论 (Game-theoretic control)[58, 289] 等研究领域中得到了广泛应用.

对于一个 $n$ 人有限的标准式博弈, 运用矩阵半张量积方法进行研究的一个基本思路是, 首先将每个博弈者 $i$ 的策略集 $\Omega_i$ 表示成一个向量形式的集合, 然后利用矩阵半张量积方法将博弈者策略的演化动态方程转化成一个逻辑控制系统. 具体来讲, 对于博弈者 $i$ 的策略 $\pi_i \in \Omega_i,\ i = 1, 2, \cdots, n$, 首先令 $\boldsymbol{x}_i$ 为 $\pi_i$ 的列向量表示, 这里假定策略集 $\Omega_i$ 的势为 $|\Omega_i| = k_i$. 然后, 对于 $k \times k$ 单位矩阵 $I_k$, 定义 $\Delta_k := \mathrm{Col}(I_k)$ 为 $I_k$ 的所有列构成的集合. 于是, 对应于博弈者 $i$ 策略的向量表示 $\boldsymbol{x}_i$, 策略集 $\Omega_i$ 的向量表示可写为 $\Delta_{k_i}$, 即 $\boldsymbol{x}_i \in \Delta_{k_i}$. 另外, 借助上述定义的符号, 博弈者 $i$ 的收益函数的向量形式可写为 $\boldsymbol{c}_i = \boldsymbol{v}_i^c \boldsymbol{x}$, $i = 1, 2, \cdots, n$, 其中行向量 $\boldsymbol{v}_i^c$ 称为 $\boldsymbol{c}_i$ 的结构向量, $\boldsymbol{x} := \ltimes_{i=1}^n \boldsymbol{x}_i$ 为所有博弈者策略组合的向量表示. 特别地, 如果博弈者参与的博弈是一个重复博弈并且博弈者策略的演化过程是 Markov 型的, 那么每个博弈者策略的演化动态方程可一般化地写为

$$\begin{aligned} &\boldsymbol{x}_i(t+1) = F_i(\boldsymbol{x}_1(t), \boldsymbol{x}_2(t), \cdots, \boldsymbol{x}_n(t)),\\ &\quad i = 1, 2, \cdots, n. \end{aligned} \tag{23}$$

因为对任意的 $i \in \{1, 2, \cdots, n\}$, $F_i(\cdot)$ 是一个有限集到有限集的映射, 所以存在一个对应于映射组合 $(F_1, F_2, \cdots, F_n)$ 的唯一矩阵 $M$ 使得式 (23) 可进一步改写为如下的一个逻辑系统[288]

$$\boldsymbol{x}(t+1) = M\boldsymbol{x}(t).$$

上述过程基本上展示了一个有限的标准式博弈是如何通过矩阵半张量积方法转化为一个逻辑系统的. 它也是矩阵半张量积方法应用于有限的标准式博弈的一个核心思想.

除了有限的标准式博弈之外, 势博弈 (Potential game)[287, 290] 是矩阵半张量积方法应用于博弈控制与优化问题的另一典型形式. 对于一个有限的标准式博弈 $\mathcal{G} = \{\mathcal{N}, \{\Omega_i\}_{i\in\mathcal{N}}, \{c_i\}_{i\in\mathcal{N}}\}$, $\mathcal{N} = \{1, 2, \cdots, n\}$, 如果存在一个函数 $W : \boldsymbol{\Omega} \to \mathbb{R}$ 使得对任意的 $i \in \mathcal{N}$,

$$\begin{aligned} W(\pi_i, \pi_{-i}) - W(\hat{\pi}_i, \pi_{-i}) = c_i(\pi_i, \pi_{-i}) -\\ c_i(\hat{\pi}_i, \pi_{-i}),\ \forall \pi_i, \hat{\pi}_i \in \Omega_i, \forall \pi_{-i} \in \Omega_{-i}, \end{aligned}$$

则称 $\mathcal{G}$ 为一个势博弈, $W$ 为该势博弈的势函数. 矩阵半张量积方法应用于势博弈的一个通常处理的问题是最大化一个总体性能指标 $\boldsymbol{J}(\pi_1, \pi_2, \cdots, \pi_n)$. 该问题的基本求解框架一般可分为 2 步: 第 1 步对每个博弈者的收益函数进行设计, 使得整个博弈系统成为以 $\boldsymbol{J}(\pi_1, \pi_2, \cdots, \pi_n)$ 为势函数的势博弈; 第 2 步为每个博弈者设计控制策略或学习算法, 使得当每个博弈者 $i$ 最大化其自身的收益函数 $c_i$ 时, 整体系统可收敛到 $\boldsymbol{J}(\pi_1, \pi_2, \cdots, \pi_n)$ 的最大值点 (也是势博弈的均衡点)[288]. 目前, 该研究框架已在多智能体系统的博弈控制问题[291] 和状态依赖的博弈策略学习问题[292] 中得到了成功应用.

当前, 利用矩阵半张量积方法处理博弈控制与优化问题已经取得了十分丰硕的成果, 一个更为全面的总结可参见文献 [288].

### 3.4 分布式 Nash 均衡搜索

分布式 Nash 均衡搜索是一类利用分布式控制与优化[115, 293–294] 的技术搜索博弈 Nash 均衡解的方法. 由于该类方法在移动传感器网络、智能电网以及无人机编队中具有广泛的应用前景, 所以吸引了大量控制领域学者的关注.

一般地, 由定义 1 可知, 求解一个 $n$ 人标准式博弈 $\mathcal{G} = \{\mathcal{N}, \{\Omega_i\}_{i\in\mathcal{N}}, \{c_i\}_{i\in\mathcal{N}}\}$ 的 Nash 均衡点等价于寻找一个策略组合 $(\pi_i^*, \pi_{-i}^*)$ 使得对任意的 $i \in \mathcal{N}$,

$$c_i(\pi_i^*, \pi_{-i}^*) \geq c_i(\pi_i, \pi_{-i}^*),\ \forall \pi_i \in \Omega_i.$$

特别地, 如果每个博弈者 $i \in \mathcal{N}$ 的策略集 $\Omega_i$ 依赖其对手的策略 $\pi_{-i}$, 则称上述标准的 Nash 均衡问题为广义 Nash 均衡问题 (Generalized Nash equilibrium problem)[295]. 另外, 为了保证上述 Nash 均衡问题解的存在性, 文献中经常使用的一个假设是[37, 296]: 对任意的 $i \in \mathcal{N}$, $\Omega_i$ 是一个非空凸紧集, 并且对任意给定的 $\pi_{-i} \in \Omega_{-i}$, $c_i(\pi_i, \pi_{-i})$ 是一个关于 $\pi_i \in \Omega_i$ 的连续可微的凸函数. 在该假设下, 求解博弈 $\mathcal{G}$ 的一个 Nash 均衡解可等价为求解如下的一个变分不等式 (Variational inequality) 问题[296–297]

$$(\boldsymbol{\pi}_v - \boldsymbol{\pi}_v^*)^T \boldsymbol{F}(\boldsymbol{\pi}_v^*) \geq 0,\ \forall \boldsymbol{\pi}_v \in \boldsymbol{\Omega} := \prod_{i\in\mathcal{N}} \Omega_i, \tag{24}$$

其中

$$\begin{cases} \boldsymbol{\pi}_v := [\pi_1, \pi_2, \cdots, \pi_n]^T,\\ \boldsymbol{\pi}_v^* := [\pi_1^*, \pi_2^*, \cdots, \pi_n^*]^T,\\ \boldsymbol{F}(\cdot) := [\nabla_{\pi_1} c_1(\cdot), \nabla_{\pi_2} c_2(\cdot), \cdots, \nabla_{\pi_n} c_n(\cdot)]^T. \end{cases}$$

因为博弈 $\mathcal{G}$ 中的每个博弈者 $i$ 的收益函数 $c_i(\cdot)$ 不一定是相同的, 所以 $\boldsymbol{F}(\cdot)$ 并不是传统意义上的梯度. 因此, 为了和通常的梯度相区别, 一般称 $\boldsymbol{F} : \boldsymbol{\Omega} \to \mathbb{R}^n$ 为伪梯度映射 (Pseudo-gradient map). 另外, 由文献 [297] 中的命题 1.5.8 可知, $\boldsymbol{\pi}_v^*$ 是变分不等式 (24) 的解当且仅当 $\boldsymbol{\pi}_v^*$ 是如下方程的不动点

$$\boldsymbol{\pi}_v^* = \mathrm{P}_{\boldsymbol{\Omega}}(\boldsymbol{\pi}_v^* - \boldsymbol{F}(\boldsymbol{\pi}_v^*)), \tag{25}$$

其中 $\mathrm{P}_{\boldsymbol{\Omega}}(\boldsymbol{\pi}_v)$ 表示 $\boldsymbol{\pi}_v$ 在集合 $\boldsymbol{\Omega}$ 上的 Euclidean 投影. 利用该等价性结论, 求解博弈 $\mathcal{G}$ 的 Nash 均衡则可转变成求解式 (25) 的不动点. 传统上, 为了实现该不动点的求解, 一种常用的方法是投影梯度

(Projected gradient-play) 算法. 具体地, 对应于离散时间 $k$ ($k \in \mathbb{N}$) 和连续时间 $t$ ($0 \leq t \in \mathbb{R}$), 该算法的离散时间版本和连续时间版本分别为

$$\pi_i(k+1) = \mathrm{P}_{\Omega_i}(\pi_i(k)) - \alpha_{i,k}\nabla_{\pi_i} c_i(\pi_i(k), \pi_{-i}(k)), \tag{26}$$

和

$$\dot{\pi}_i = \Xi_{\Omega_i}(\pi_i, -\nabla_{\pi_i} c_i(\pi_i, \pi_{-i})), \tag{27}$$

其中 $\pi_i(k)$ 表示博弈者 $i$ 在迭代时刻 $k$ 的策略, $\alpha_{i,k}$ 表示博弈者 $i$ 在迭代时刻 $k$ 的学习率, $\dot{\pi}_i := \mathrm{d}\pi_i(t)/\mathrm{d}t, \Xi_{\Omega_i}(\pi_i, \nu) := \lim_{\varepsilon \to 0}[\mathrm{P}_{\Omega_i}(\pi_i + \varepsilon\nu) - \pi_i]/\varepsilon$. 特别地, 如果伪梯度映射 $\boldsymbol{F}$ 是 Lipschitz 连续的且严格单调的 (或强单调的), 上述两个投影梯度算法在理论上均可收敛到博弈 $\mathcal{G}$ 的一个 Nash 均衡点[297−298].

由投影梯度算法 (26) 和 (27) 可以看出, 每个博弈者在更新自己策略的过程中不仅需要自身的收益信息而且还需要所有其他博弈者的策略信息. 这一要求在实际系统中 (尤其在网络化系统中) 通常是较难满足的. 为了弥补这一不足, 分布式 Nash 均衡搜索方法通常假定每个博弈者并不能获得所有其他博弈者的信息, 而只能通过一个通信网络来获得其直接邻居的信息. 基于这一假定, 文献 [299] 利用异步 Gossip 方法提出了一个离散时间分布式投影梯度算法, 而文献 [300] 利用伪梯度映射的增量无源性 (Passivity) 提出了一个连续时间分布式投影梯度算法. 在适当的条件下, 这两个分布式算法在理论上均能几乎必然地收敛到一个 Nash 均衡点[299−300]. 另外, 当博弈者的收益函数满足二阶连续可微条件时, 文献 [301] 基于领航者—跟随者一致性方法提出了一类连续时间分布式算法, 并利用 Lyapunov 稳定性理论证明了该算法的收敛性. 通过考虑每个博弈者的收益函数满足一个聚合映射 (Aggregation map) 并且所有博弈者的可行策略集满足一个线性耦合约束条件, 文献 [302] 基于投影动力学和非光滑跟踪动力学, 提出了一个连续时间分布式算法用于求解广义 Nash 均衡搜索问题. 受上述工作的启发, 通过考虑博弈者的可行策略集同时满足私有约束条件和耦合约束条件, 文献 [303] 提出了一个连续时间在线分布式原始—对偶 Nash 均衡搜索算法, 并在适当的条件下证明了该算法的收敛性.

虽然在特定的前提条件下, 上述分布式 Nash 均衡搜索算法在理论上均能保证收敛到一个 Nash 均衡点, 但它们绝大部分只能处理固定博弈形式 (即博弈者的收益函数是时不变的) 的 Nash 均衡搜索问题. 为了弥补这一不足, 文献 [304] 进一步对具有时变收益函数的广义 Nash 均衡搜索问题进行了研究, 并率先提出了一类在线分布式原始—对偶策略学习算法. 在通信网络满足连通性假设条件下, 该文从理论上严格证明了该算法将收敛到一个广义 Nash 均衡.

### 3.5 基于零行列式策略的收益控制

粗略地讲, 上述基于最优控制和微分博弈的博弈控制方法大部分都是通过设计或寻找使系统满足某一特定性能指标的控制输入, 来实现对博弈系统动力学的控制 (即控制博弈系统动力学的演化行为). 相比之下, 基于零行列式 (Zero determinant, ZD) 策略[305] 的博弈控制方法是一类起源于重复博弈研究, 能够实现单边控制对手期望收益的数学方法. 一般地, 依据对重复博弈的传统认识, 每个博弈者的期望收益通常会随着其对手策略的改变而改变. 然而, ZD 策略的发现几乎从根本上改变了这一传统认识[306]: 在一个重复博弈中, 如果中心博弈者的策略为一个 ZD 策略, 那么不管其对手如何进行策略选择, 中心博弈者总是能够迫使其对手的期望收益达到一个给定的值或者满足一个线性等式关系 (即控制对手的期望收益).

具体地, 考虑一个两人两策略对称重复博弈 $\mathcal{G}_{2\times 2} = \{\mathcal{N}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, \{c_i\}_{i\in\mathcal{N}}\}$, 其中 $\mathcal{N} = \{1, 2\}$ 为所有博弈者的标号构成的集合, $\mathcal{A}_i = \{\mathrm{C}, \mathrm{D}\}$, $\forall i \in \mathcal{N}$ 为博弈者 $i$ 的行动集, $c_i : \mathcal{A}_1 \times \mathcal{A}_2 \to \mathbb{R}$ 为博弈者 $i \in \mathcal{N}$ 的收益函数. 对于收益函数 $c_i$, $\forall i \in \mathcal{N}$, 其在不同行动组合下的取值所构成的收益值表为

| | C | D |
|---|---|---|
| C | $(\mathfrak{a}, \mathfrak{a})$ | $(\mathfrak{b}, \mathfrak{c})$ |
| D | $(\mathfrak{c}, \mathfrak{b})$ | $(\mathfrak{d}, \mathfrak{d})$ |

这里“行博弈者” 表示博弈者 1, “列博弈者” 表示博弈者 2. 在该博弈中, 假定所有博弈者的策略都为一步记忆策略, 即每个博弈者在当前轮博弈中的行动选择只依赖前一轮博弈中所有博弈者的行动组合. 基于该假设, 博弈者 1 和博弈者 2 的策略于是可分别写为向量 $\vec{\pi}_1 = [p_{\mathrm{CC}}, p_{\mathrm{CD}}, p_{\mathrm{DC}}, p_{\mathrm{DD}}]^T$ 和 $\vec{\pi}_2 = [q_{\mathrm{CC}}, q_{\mathrm{CD}}, q_{\mathrm{DC}}, q_{\mathrm{DD}}]^T$, 其中 $p_{a_x a_y} \in [0, 1]$ 和 $q_{a_x a_y} \in [0, 1]$ 分别表示博弈者 1 和博弈者 2 在前一轮博弈中的行动组合为 $(a_x, a_y)$, $a_x, a_y \in \mathcal{A}$ 情况下, 于当前轮博弈中选择行动 C 的概率. 在重复博弈 $\mathcal{G}_{2\times 2}$ 进行过程中, 如果两个博弈者采用的策略分别为一步记忆策略 $\vec{\pi}_1$ 和 $\vec{\pi}_2$, 那么在时间尺度上, 它们将诱导出一个定义在状态空间 $\{\mathrm{CC}, \mathrm{CD}, \mathrm{DC}, \mathrm{DD}\}$ 上的 Markov 链. 特别地, 记该 Markov 链的转移概率矩阵为 $\boldsymbol{P} = (\rho_{a_z a_w | a_x a_y}) \in [0, 1]^{4\times 4}$, 其中 $\rho_{a_z a_w | a_x a_y}(a_x, a_y, a_z, a_w \in \mathcal{A})$ 表示该 Markov 链从状态 $a_x a_y$ 转移到状态 $a_z a_w$ 的概率.

假定该 Markov 链是遍历的并且令 $\boldsymbol{\mu}$ 为其状态的平稳分布 (也是 Markov 链处在每个状态的平均时间的分布) 构成的列向量, 即 $\boldsymbol{\mu}^T\boldsymbol{P}=\boldsymbol{\mu}^T$.[15] 因为 $\boldsymbol{P}$ 是一个行随机矩阵, 所以元素全为 1 的列向量 $\boldsymbol{y}=\mathbf{1}$ 是方程 $(\boldsymbol{P}-I)\boldsymbol{y}=\mathbf{0}$ 的一个解, 即 $\boldsymbol{P}':=\boldsymbol{P}-I$ 是奇异的. 因此, $\boldsymbol{P}'$ 的行列式为 0. 另外, 由 Cramer 法则可知, $\mathrm{adj}(\boldsymbol{P}')\boldsymbol{P}'=\det(\boldsymbol{P}')I=\mathbf{0}_m$, 其中 $\mathrm{adj}(\boldsymbol{P}')$ 和 $\det(\boldsymbol{P}')$ 分别表示矩阵 $\boldsymbol{P}'$ 的伴随矩阵和行列式, $\mathbf{0}_m$ 表示元素全为 0 的矩阵. 由上述两个等式 $\boldsymbol{\mu}^T(\boldsymbol{P}-I)=\mathbf{0}$ 和 $\mathrm{adj}(\boldsymbol{P}')\boldsymbol{P}'=\mathbf{0}_m$ 可知, $\mathrm{adj}(\boldsymbol{P}')$ 的每一行均正比于 $\boldsymbol{\mu}^T$. 特别地, 对于 $\mathrm{adj}(\boldsymbol{P}')$ 的第 4 行, 由伴随矩阵的定义可知 $\boldsymbol{\mu}^T\boldsymbol{w}\equiv D(\vec{\pi}_1,\vec{\pi}_2,\boldsymbol{w})$, 其中 $\boldsymbol{w}=[w_1,\ w_2,w_3,w_4]^T\in\mathbb{R}^4$ 是任意的四维列向量, $D(\vec{\pi}_1,\vec{\pi}_2,\ \boldsymbol{w})$ 是把 $\boldsymbol{P}'$ 的第 1 列分别加到第 2、第 3 列以及把 $\boldsymbol{P}'$ 的第 4 列替换为 $\boldsymbol{w}$ 之后得到的矩阵的行列式, 即

$$\boldsymbol{\mu}^T\boldsymbol{w}\equiv D(\vec{\pi}_1,\vec{\pi}_2,\boldsymbol{w}):=$$
$$\det\left(\begin{bmatrix}-1+p_{\mathrm{CC}}q_{\mathrm{CC}} & -1+p_{\mathrm{CC}} & -1+q_{\mathrm{CC}} & w_1\\ p_{\mathrm{CD}}q_{\mathrm{DC}} & -1+p_{\mathrm{CD}} & q_{\mathrm{DC}} & w_2\\ p_{\mathrm{DC}}q_{\mathrm{CD}} & p_{\mathrm{DC}} & -1+q_{\mathrm{CD}} & w_3\\ p_{\mathrm{DD}}q_{\mathrm{DD}} & p_{\mathrm{DD}} & q_{\mathrm{DD}} & w_4\end{bmatrix}\right). \tag{28}$$

以博弈者 1 和博弈者 2 的所有可能的收益值作为元素分别构造向量 $\boldsymbol{e}_1=[\mathfrak{a},\mathfrak{b},\mathfrak{c},\mathfrak{d}]$ 和 $\boldsymbol{e}_2=[\mathfrak{a},\mathfrak{c},\mathfrak{b},\mathfrak{d}]$. 于是, 这两个博弈者在每轮博弈中所能获得的期望收益可分别计算为

$$\bar{c}_1=\frac{\boldsymbol{\mu}^T\boldsymbol{e}_1}{\boldsymbol{\mu}^T\mathbf{1}}=\frac{D(\vec{\pi}_1,\vec{\pi}_2,\boldsymbol{e}_1)}{D(\vec{\pi}_1,\vec{\pi}_2,\mathbf{1})}, \tag{29}$$

和

$$\bar{c}_2=\frac{\boldsymbol{\mu}^T\boldsymbol{e}_2}{\boldsymbol{\mu}^T\mathbf{1}}=\frac{D(\vec{\pi}_1,\vec{\pi}_2,\boldsymbol{e}_2)}{D(\vec{\pi}_1,\vec{\pi}_2,\mathbf{1})}. \tag{30}$$

因为式 (29) 和式 (30) 的计算均具有线性性, 所以对任意一组常系数 $\xi_1\in\mathbb{R}$、$\xi_2\in\mathbb{R}$ 和 $\xi_3\in\mathbb{R}$ 有

$$\xi_1\bar{c}_1+\xi_2\bar{c}_2+\xi_3=\frac{D(\vec{\pi}_1,\vec{\pi}_2,\xi_1\boldsymbol{e}_1+\xi_2\boldsymbol{e}_2+\xi_3\mathbf{1})}{D(\vec{\pi}_1,\vec{\pi}_2,\mathbf{1})}. \tag{31}$$

特别地, 如果博弈者 1 的策略向量 $\vec{\pi}_1$ 取为 $\vec{\pi}_1=\vec{\pi}^0+\xi_1\boldsymbol{e}_1+\xi_2\boldsymbol{e}_2+\xi_3\mathbf{1}$ 或者博弈者 2 的策略向量 $\vec{\pi}_2$ 取为 $\vec{\pi}_2=\vec{\pi}^0+\xi_1\boldsymbol{e}_1+\xi_2\boldsymbol{e}_2+\xi_3\mathbf{1}$, 其中 $\vec{\pi}^0:=[1,1,0,0]^T$, $\xi_1$、$\xi_2$ 和 $\xi_3$ 的选取需使得 $\vec{\pi}_1$ 或者 $\vec{\pi}_2$ 的每个分量属于 $[0,1]$, 那么由式 (28) 和式 (31) 可知

$$\xi_1\bar{c}_1+\xi_2\bar{c}_2+\xi_3=0. \tag{32}$$

考虑到上述线性等式 (32) 的成立是由于策略 $\vec{\pi}_1$ 或者 $\vec{\pi}_2$ 的选取使得式 (28) 中的行列式等于 0, 文献 [305] 由此把满足这一特性的策略称为零行列式策略. 由式 (32) 可以看出, 在重复博弈 $\mathcal{G}_{2\times2}$ 中, 任意的一个博弈者都可以通过设定自己的策略来迫使其对手的期望收益满足一个等式约束关系. 例如, 对于博弈者 1, 如果其选择的策略为 $\vec{\pi}_1=\vec{\pi}^0+\xi_2\boldsymbol{c}_2+\xi_3\mathbf{1}$ (即设置 $\xi_1=0$), 那么无论博弈者 2 如何进行策略选择, 它的期望收益将始终保持为 $\bar{c}_2=-\xi_3/\xi_2$. 特别地, 借助 $\vec{\pi}_1=\vec{\pi}^0+\xi_2\boldsymbol{e}_2+\xi_3\mathbf{1}$ 并利用 $p_{\mathrm{CC}}$ 和 $p_{\mathrm{DD}}$ 分别表示 $\xi_2$ 和 $\xi_3$, $\bar{c}_2$ 可进一步写为

$$\bar{c}_2=\frac{(1-p_{\mathrm{CC}})\mathfrak{d}+p_{\mathrm{DD}}\mathfrak{a}}{(1-p_{\mathrm{CC}})+p_{\mathrm{DD}}}.$$

除了单边控制对手的期望收益之外, 通过设定中心博弈者 (比如博弈者 1) 所采用的 ZD 策略的形式, ZD 策略还可以分别实现敲诈行为 (即中心博弈者获得一个不低于对手的期望收益)、慷慨行为 (即中心博弈者获得一个不高于对手的期望收益) 和公平行为 (即中心博弈者和对手获得一个同等的期望收益) 等[305, 307].

鉴于 ZD 策略对重复博弈研究的重要贡献, 后续关于 ZD 策略的讨论在演化博弈领域中掀起了一场研究热潮, 其中推动这一研究热潮的一个关键结果是所谓的 Akin 引理[308]. 该引理是式 (28) 的一般化形式, 也是判定一个策略是否为 ZD 策略的必要条件. 一般地, 对于一个两人两策略对称重复博弈 $\mathcal{G}_{2\times2}$, Akin 引理可具体地表述为: 如果博弈 $\mathcal{G}_{2\times2}$ 的所有博弈者的策略都为一步记忆策略并且它的所有可能的行动组合 $\{\mathrm{CC},\mathrm{CD},\mathrm{DC},\mathrm{DD}\}$ 的平稳分布为 $\boldsymbol{\mu}$, 那么对任意的博弈者 $i\in\mathcal{N}:=\{1,2\}$, 它的策略向量 $\vec{\pi}_i$ 将满足

$$\boldsymbol{\mu}^T(\vec{\pi}_i-\vec{\pi}^0)=0.$$

借助 Akin 引理的各种拓展形式, 后续的相关研究分别将 ZD 策略推广到了两人两策略折扣重复博弈[309]、两人连续行动空间折扣重复博弈[310]、多人两策略无折扣[311−312] 和有折扣重复博弈[313]、多人多策略时变折扣重复博弈[314] 等更一般化的博弈形式中. 除了这些拓展性工作之外, 一些相关的研究也讨论了 ZD 策略本身的一些属性, 比如文献 [315] 分析了 ZD 策略的鲁棒性 (即采取 ZD 策略的博弈者面对策略变化的对手是否仍然可以保持期望收益最大化), 文献 [307, 316−318] 研究了 ZD 策略的演化稳定性 (即 ZD 策略在自然选择的作用下是否能抵御其他策略的入侵), 而文献 [319] 则考虑了博弈者的错误认知对 ZD 策略形式的影响.

### 3.6 小结

按照不同专题的方式, 本节主要梳理并介绍了

[15] 文献 [305] 也证明了非遍历情况下该平稳分布的存在性.

博弈、学习与控制这一交叉研究主题的几类典型成果, 即基于最优控制的学习与博弈、博弈控制系统、基于矩阵半张量积的博弈控制论、分布式 Nash 均衡搜索和基于零行列式策略的收益控制. 在第 1 部分基于最优控制的学习与博弈中, 首先介绍了一类单智能体连续时间最优控制问题, 并回顾了求解该类问题的强化学习算法; 接着, 当单智能体最优控制问题拓展到多智能体场景中时, 进一步对微分博弈以及求解该类博弈的多智能体强化学习算法进行了介绍. 博弈控制系统是一类试图将现代控制理论与微分博弈相结合的系统. 在这部分内容中, 主要对一类确定性博弈控制系统和一类随机性博弈控制系统的能控性问题进行了介绍; 另外, 为了反映该类系统的最新研究进展, 还对一类线性二次型微分博弈控制系统的镇定性问题和一类随机自适应微分博弈系统的稳定性问题进行了简要的论述. 接着, 在基于矩阵半张量积的博弈控制论这部分内容中, 主要讨论了如何运用矩阵半张量积方法研究有限的标准式博弈和势博弈中的控制与优化问题. 然后, 针对一类 Nash 均衡求解问题, 本节主要在变分不等式方法的基础上回顾了几类基于分布式控制与优化的搜索算法. 最后, 基于对零行列式策略的讨论, 本节主要梳理并介绍了一类能单边控制对手期望收益的方法.

## 4　总结与展望

作为现代数学的一个重要分支和运筹学的一个重要组成部分, 博弈论近年来在学术界受到越来越多的关注. 特别地, 随着多智能体系统和 AI 等研究领域的兴起, 博弈论、多智能体学习与控制论的交叉融合目前已发展成为一个前沿热点研究方向. 为了及时反映这一学术动态和趋势, 本文从连接这三者的四类基本博弈形式出发, 系统地讨论了它们之间的联系与区别, 论述了它们对应的各类多智能体学习方法, 还回顾了当前几个博弈、学习与控制的交叉研究专题. 作为全文内容的一个总结, 图 3 左边子图描述了多智能体博弈、学习与控制在宏观层面上的相互联系, 右边子图展示了本文介绍的各类博弈形式、多智能体学习方法和控制论方法之间的内在关联关系 (图中的数字表示本文的章节编号).

目前, 虽然多智能体博弈、学习与控制这一新兴交叉研究领域已经取得了一些初步成果, 但总体而言, 它仍然处于发展初期, 还有巨大的发展空间. 从大的方向上看, 以下几方面未来还有待于进一步探索.

1) 无模型的 (Model-free) 博弈论或数据驱动的 (Data-driven) 博弈论. 在博弈论的传统研究中, 大部分工作都是基于特定的博弈模型或者发展新的博弈模型来开展的. 换句话讲, 在这些研究中, 博弈模型是事先设定的并且 (或部分) 为博弈者所已知的. 因此, 从方法论上讲, 它们属于基于模型的 (Model-based) 博弈论, 即“白盒” (White-box) 的方法. 然而, 在大量与博弈论相关的实际应用, 比如视频游戏、多机器人系统、自动驾驶和无人机集群中, 获取准确的模型信息通常是比较困难的. 例如, 为了实现某些实际应用中的特定任务目标, 博弈者的收益函数通常可能需要进行额外设计[41, 231, 320]. 针对这些无模型的博弈应用问题, 如果传统基于模型的博弈论方法继续被使用去求解诸如 Nash 均衡等博弈问题, 那么它将自然会面临一些理论困境. 为此, 这就十分有必要去发展一类无模型的或数据驱动的博弈论方法, 即利用博弈实时产生的数据或者博弈交互后的离线数据来实现对一个博弈问题的求解 (也就是“黑盒” (Black-box) 或者“灰盒” (Gray-box) 的方法). 目前, 虽然这方面已有一些相关工作, 比如著名的虚拟对弈方法[14, 184] 和多智能体强化学习方法, 但总体而言, 这一领域仍处于发展初期, 还有许

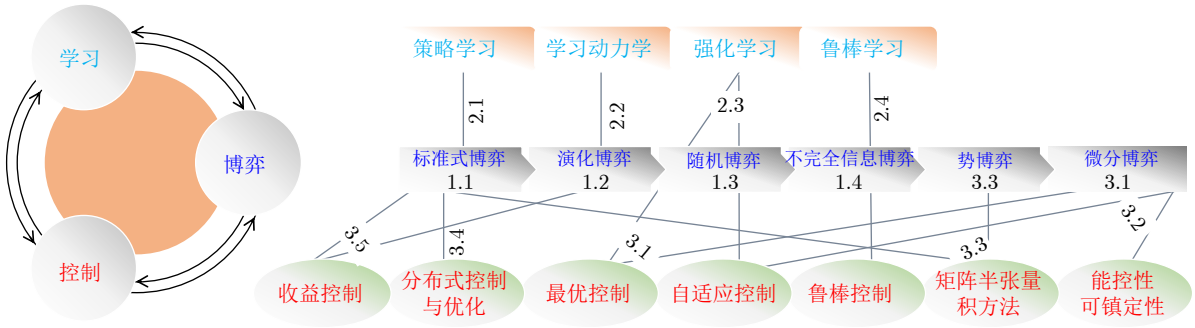



图 3　本文介绍的各类博弈形式、学习方法和控制论方法之间的内在关联关系图

Fig. 3　Illustrations of the intrinsic relationship between the games, multi-agent learning methods, and control methods presented in this paper

多不完善之处.

2) 博弈论、多智能体学习与控制论的进一步深度融合. 目前, 在自动控制领域中, 博弈、学习与控制的大部分交叉研究工作都是基于最优控制和微分博弈来开展的. 除此之外, 博弈论和控制论本质上还有许多相通之处. 例如, 如果从博弈论的角度来看控制论, 控制系统中的控制器一般可视为博弈中的决策者, 它可以改变受控系统的演化轨迹以期达到某种理想效果; 反过来, 如果从控制论的角度来看博弈论, 博弈中的博弈者一般可视为控制系统中的控制器, 它的决策目标或偏好 (Preference) 通常可认为是一种控制目标. 因此, 从控制论的角度上讲, 博弈论也被视为是一类研究交互控制器之间的合作与竞争的数学理论[30]. 另外, 作为现代控制论的一个重要组成部分, 鲁棒控制[237] 与鲁棒优化[179]、鲁棒博弈[180] 以及鲁棒学习[233] 在基本思想上均有诸多相似之处. 虽然受安全 AI 和鲁棒 AI[5−6, 321] 这一研究主题的驱动, 当前鲁棒的机器学习技术在 AI 领域中已成为一个热点研究方向, 但鲁棒控制、鲁棒优化、鲁棒博弈和鲁棒学习间的交叉研究在自动控制领域中目前仍比较稀少.

3) 建立并发展包含其他博弈形式的学习与控制的研究框架. 本文讨论的博弈形式主要涉及标准式博弈、演化博弈、随机博弈、不完全信息博弈、势博弈和微分博弈. 除了这些基本博弈形式之外, 现代博弈论中还有一些其他新颖形式, 比如扩展式博弈 (Extensive-form game)[18, 322−323]、信号博弈 (Signaling game)[324]、量子博弈 (Quantum game)[325−327] 和平均场博弈 (Mean-field game)[328−329] 等. 因此, 如何建立并发展一个包含其他博弈形式的学习与控制的研究框架将是一个十分值得深入探究的课题.

4) 博弈论、动力系统与深度学习的结合. 演化博弈动力学是博弈论与动力系统的一种结合形式. 考虑到当前在机器学习领域中, 动力系统与深度学习的结合是一个广受关注的研究课题[330−332]. 因此, 如果演化博弈动力学可以进一步与深度学习相结合, 那么这在理论上将是十分有意义的. 尽管当前这方面已有一些探索性工作, 比如神经复制动力学[119], 但总体而言, 这一研究方向目前才刚刚兴起并且在理论上还有诸多不完善之处.

5) 博弈、学习与控制的交叉研究在一些新兴领域中的应用. 博弈论自诞生以来, 已在经济学、社会学、心理学、生物学、物理学、认知科学和计算机科学等研究领域中得到了大量卓有成效的应用. 近年来, 伴随着 AI 技术的快速发展, 人们也见证着一些新兴交叉研究领域的兴起, 比如社会智能[1−2]、机器智能[3]、合作智能[4]、AI 安全[5−6] 和 AI 伦理[7−8] 等. 考虑到这些研究主题本身具有广泛的学科交叉性. 因此, 如何将博弈、学习与控制的交叉研究成果应用于这些新兴领域是未来一个十分值得探究的课题.

## References


1 McDonald K R, Pearson J M. Cognitive bots and algorithmic humans: Toward a shared understanding of social intelligence. *Current Opinion in Behavioral Sciences*, 2019, **29**: 55−62

2 Silver D, Singh S, Precup D, Sutton R S. Reward is enough. *Artificial Intelligence*, 2021, **299**: 103535

3 Rahwan I, Cebrian M, Obradovich N, Bongard J, Bonnefon J F, Breazeal C, et al. Machine behaviour. *Nature*, 2019, **568**(7753): 477−486

4 Dafoe A, Bachrach Y, Hadfield G, Horvitz E, Larson K, Graepel T. Cooperative AI: Machines must learn to find common ground. *Nature*, 2021, **593**(7857): 33−36

5 Russell S, Dewey D, Tegmark M. Research priorities for robust and beneficial artificial intelligence. *AI Magazine*, 2015, **36**(4): 105−114

6 Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D. Concrete problems in AI safety. arXiv: 1606.06565, 2016.

7 Bonnefon J F, Shariff A, Rahwan I. The social dilemma of autonomous vehicles. *Science*, 2016, **352**(6293): 1573−1576

8 Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 2019, **1**(9): 389−399

9 Myerson R B. *Game Theory: Analysis of Conflict*. Cambridge, USA: Harvard University Press, 1997.

10 von Neumann J, Morgenstern O. *Theory of Games and Economic Behavior*. Princeton, USA: Princeton University Press, 1944.

11 Wooldridge M, Jennings N R. Intelligent agents: Theory and practice. *The Knowledge Engineering Review*, 1995, **10**(2): 115−152

12 Sen S. Multiagent systems: Milestones and new horizons. *Trends in Cognitive Sciences*, 1997, **1**(9): 334−340

13 Weiss G. *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence*. Cambridge, USA: MIT Press, 1999.

14 Brown G W. Iterative solution of games by fictitious play. In: *T.C. Koopmans, editor, Activity Analysis of Production and Allocation*. New York: Wiley, 1951. 374–376

15 Tuyls K, Weiss G. Multiagent learning: Basics, challenges, and prospects. *AI Magazine*, 2012, **33**(3): 41−52

16 Moravčík M, Schmid M, Burch N, Lisý V, Morrill D, Bard N, et al. DeepStack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 2017, **356**(6337): 508−513

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

18 Brown N, Sandholm T. Superhuman AI for multiplayer poker. *Science*, 2019, **365**(6456): 885−890

19 Vinyals O, Babuschkin I, Czarnecki W M, Mathieu M, Dudzik A, Chung J, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 2019, **575**(7782): 350−354

20 Jaderberg M, Czarnecki W M, Dunning I, Marris L, Lever G, Castañeda A G, et al. Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science*, 2019, **364**(6443): 859−865

21 Wurman P R, Barrett S, Kawamoto Kenta, MacGlashan J, Subramanian K, Walsh T J, et al. Outracing champion Gran

Turismo drivers with deep reinforcement learning. *Nature*, 2022, **602**(7896): 223−228

22 Bennett S. A brief history of automatic control. *IEEE Control Systems Magazine*, 1996, **16**(3): 17−25

23 Wiener N. *Cybernetics or Control and Communication in the Animal and the Machine*. Cambridge, USA: MIT Press, 1948.

24 Guo Lei. Estimation, control, and games of dynamical systems with uncertainty. *SCIENCE CHINA: Information Science*, 2020, **50**(9): 1327−1344
(郭雷. 不确定性动态系统的估计、控制与博弈. 中国科学: 信息科学, 2020, **50**(9): 1327−1344)

25 Marschak J. Elements for a theory of teams. *Management Science*, 1955, **1**(2): 127−137

26 Ho Y C. Team decision theory and information structures in optimal control problems—Part I. *IEEE Transactions on Automatic Control*, 1972, **17**(1): 15−22

27 Boutilier C. Planning, learning and coordination in multiagent decision processes. In: Proceedings of the 6th Conference on Theoretical Aspects of Rationality and Knowledge. De Zeeuwse Stromen, The Netherlands: Morgan Kaufmann Publishers Inc, 1996. 195−210

28 Wang X F, Sandholm T. Reinforcement learning to play an optimal Nash equilibrium in team Markov games. In: Proceedings of the 15th International Conference on Neural Information Processing Systems. Vancouver, Canada: MIT Press, 2002. 1603−1610

29 Witsenhausen H S. A counterexample in stochastic optimum control. *SIAM Journal on Control*, 1968, **6**(1): 131−147

30 Marden J R, Shamma J S. Game theory and control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2018, **1**: 105−134

31 Lewis F L, Zhang H W, Hengster-Movric K, Das A. *Cooperative Control of Multi-Agent Systems: Optimal and Adaptive Design Approaches*. New York, USA: Springer, 2013.

32 Bauso D, Pesenti R. Team theory and person-by-person optimization with binary decisions. *SIAM Journal on Control and Optimization*, 2012, **50**(5): 3011−3028

33 Nayyar A, Teneketzis D. Common knowledge and sequential team problems. *IEEE Transactions on Automatic Control*, 2019, **64**(12): 5108−5115

34 Yongacoglu B, Arslan G, Yüksel S. Decentralized learning for optimality in stochastic dynamic teams and games with local control and global state information. *IEEE Transactions on Automatic Control*, 2022, **67**(10): 5230−5245

35 Littman M L. Markov games as a framework for multi-agent reinforcement learning. In: Proceedings of the 11th International Conference on Machine Learning. New Brunswick, USA: Morgan Kaufmann Publishers Inc, 1994. 157−163

36 Başar T, Bernhard P. *$H_\infty$ Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach* (Second edition). New York: Springer, 2008.

37 Başar T, Olsder G J. *Dynamic Noncooperative Game Theory* (Second edition). Philadelphia: SIAM, 1998.

38 Nash J. Equilibrium points in *n*-person games. *Proceedings of the National Academy of Sciences of the United States of America*, 1950, **36**(1): 48−49

39 Hu J L, Wellman M P. Nash Q-learning for general-sum stochastic games. *Journal of Machine Learning Research*, 2003, **4**: 1039−1069

40 Lowe R, Wu Y, Tamar A, Harb J, Abbeel P, Mordatch I. Multi-agent actor-critic for mixed cooperative-competitive environments. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc, 2017. 6382−6393

41 Li N, Marden J R. Designing games for distributed optimization. *IEEE Journal of Selected Topics in Signal Processing*, 2013, **7**(2): 230−242

42 Tsien Hsue-Shen. Engineering Cybernetics. New York: McGraw Hill Book, 1954
(钱学森 [著], 戴汝为 [译]. 工程控制论. 北京: 科学出版社, 1958.)

43 Lamnabhi-Lagarrigue F, Annaswamy A, Engell S, Isaksson A, Khargonekar P, Murray R M, et al. Systems & control for the future of humanity, research agenda: Current and future roles, impact and grand challenges. *Annual Reviews in Control*, 2017, **43**: 1−64

44 Shoham Y, Powers R, Grenager T. If multi-agent learning is the answer, what is the question? *Artificial Intelligence*, 2007, **171**(7): 365−377

45 Fudenberg D, Levine D K. *The Theory of Learning in Games*. Cambridge, USA: MIT Press, 1998.

46 Young H P. *Strategic Learning and Its Limits*. Oxford: Oxford University Press, 2004.

47 Busoniu L, Babuska R, De Schutter B. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 2008, **38**(2): 156−172

48 Nowé A, Vrancx P, De Hauwere Y M. *Game theory and multi-agent reinforcement learning. In: Marco Wiering and Martijn van Otterlo, editors, Reinforcement Learning: State-of-the-Art*, Berlin: Heidelberg, 2012. 441–470

49 Zhang K Q, Yang Z R, Başar T. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *In Kyriakos G. Vamvoudakis, Yan Wan, Frank L. Lewis, and Derya Cansever, editors, Handbook of reinforcement learning and control*, Cham, Switzerland: Springer, 2021. 321–384

50 Yang Y D, Wang J. An overview of multi-agent reinforcement learning from game theoretical perspective. arXiv: 2011.00583, 2021.

51 Ocampo-Martinez C, Quijano N. Game-theoretical methods in control of engineering systems: An introduction to the special issue. *IEEE Control Systems Magazine*, 2017, **37**(1): 30−32

52 Riehl J, Ramazi P, Cao M. A survey on the analysis and control of evolutionary matrix games. *Annual Reviews in Control*, 2018, **45**: 87−106

53 Zhang J F. Preface to special topic on games in control systems. *National Science Review*, 2020, **7**(7): 1115−1115

54 Kiumarsi B, Vamvoudakis K G, Modares H, Lewis F L. Optimal and autonomous control using reinforcement learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2018, **29**(6): 2042−2062

55 Buşoniu L, de Bruin T, Tolić D, Kober J, Palunko I. Reinforcement learning for control: Performance, stability, and deep approximators. *Annual Reviews in Control*, 2018, **46**: 8−28

56 Annaswamy A, Morari M, Pappas G J, Tomlin C, Vidal R, Zellinger M. Special issue on learning and control. *IEEE Transactions on Automatic Control*, 2022.

57 Giordano G. Tamer Başar [people in control]. *IEEE Control Systems Magazine*, 2021, **41**(6): 28−33

58 Cheng Dai-Zhan, Fu Shi-Hua. A survey on game theoretical control. *Control Theory & Applications*, 2018, **35**(5): 588−592
(程代展, 付世华. 博弈控制论简述. 控制理论与应用, 2018, **35**(5): 588−592)

59 Fudenberg D, Tirole J. *Game Theory*. Cambridge, USA: MIT Press, 1991.

60 Osborne M J, Rubinstein A. *A Course in Game Theory*. Cambridge, USA: MIT Press, 1994.

61 Nash J. Non-cooperative games. *Annals of Mathematics*, 1951, **54**(2): 286−295

62 Fudenberg D, Levine D K. An economist′s perspective on multi-agent learning. *Artificial Intelligence*, 2007, **171**(7): 378−381

63 Camerer C F. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton, USA: Princeton University Press, 2003.

64 Rubinstein A. *Modeling Bounded Rationality*. Cambridge, USA: MIT Press, 1998.

65 Williamson O E. Transaction-cost economics: The governance of contractual relations. *The Journal of Law and Economics*, 1979, **22**(2): 233−261

66 Maynard Smith J, Price G R. The logic of animal conflict. *Nature*, 1973, **246**(5427): 15−18

67 Maynard Smith J. *Evolution and the Theory of Games*. Cambridge, UK: Cambridge University Press, 1982.

68 Mayr E. *Populations, Species, and Evolution: An Abridgment of Animal Species and Evolution*. Cambridge, USA: Harvard University Press, 1970.

69 Hofbauer J, Sigmund K. *Evolutionary Games and Population Dynamics*. Cambridge, USA: Cambridge University Press, 1998.

70 Sandholm W H. *Population Games and Evolutionary Dynamics*. Cambridge, USA: MIT Press, 2010.

71 Simon H A. A behavioral model of rational choice. *The Quarterly Journal of Economics*, 1955, **69**(1): 99−118

72 Szabó G, Fáth G. Evolutionary games on graphs. *Physics Reports*, 2007, **446**(4-6): 97−216

73 Weibull J W. *Evolutionary Game Theory*. Cambridge, USA: MIT Press, 1995.

74 Nowak M A. *Evolutionary Dynamics: Exploring the Equations of Life*. Cambridge, USA: Harvard University Press, 2006.

75 Hofbauer J, Sigmund K. Evolutionary game dynamics. *Bulletin of the American Mathematical Society*, 2003, **40**(4): 479−519

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

77 Schuster P, Sigmund K. Replicator dynamics. *Journal of Theoretical Biology*, 1983, **100**(3): 533−538

78 Brown G W, von Neumann J. Solutions of games by differential equations. In: *H.W. Kuhn and A.W. Tucker, editors, Contributions to the Theory of Games (AM-24)*, Volume I. Princeton: Princeton University Press, 1950. 73–79.

79 Smith M J. The stability of a dynamic model of traffic assignment—An application of a method of Lyapunov. *Transportation Science*, 1984, **18**(3): 245−252

80 Helbing D. A mathematical model for behavioral changes by pair interactions. In: *Günter Haag, Ulrich Mueller, and Klaus G. Troitzsch, editors, Economic Evolution and Demographic Change: Formal Models in Social Sciences*. Berlin, Heidelberg: Springer, 1992. 330−348

81 Schlag K H. Why imitate, and if so, how?: A boundedly rational approach to multi-armed bandits. *Journal of Economic Theory*, 1998, **78**(1): 130−156

82 Hofbauer J. On the occurrence of limit cycles in the Volterra-Lotka equation. *Nonlinear Analysis: Theory, Methods & Applications*, 1981, **5**(9): 1003−1007

83 Dugatkin L A, Reeve H K. *Game Theory and Animal Behavior*. New York: Oxford University Press, 2000.

84 Cross J G. A stochastic learning model of economic behavior. *The Quarterly Journal of Economics*, 1973, **87**(2): 239−266

85 Watkins C J C H. Learning from Delayed Rewards [Ph. D. dissertation], King′s College, UK, 1989

86 Watkins C J C H, Dayan P. *Q*-learning. *Machine Learning*, 1992, **8**(3): 279−292

87 Börgers T, Sarin R. Learning through reinforcement and replicator dynamics. *Journal of Economic Theory*, 1997, **77**(1): 1−14

88 Sato Y, Crutchfield J P. Coupled replicator equations for the dynamics of learning in multiagent systems. *Physical Review E*, 2003, **67**(1): 015206

89 Tuyls K, Verbeeck K, Lenaerts T. A selection-mutation model for Q-learning in multi-agent systems. In: Proceedings of the 2nd International Joint Conference on Autonomous Agents and MultiAgent Systems. Melbourne, Australia: ACM, 2003. 693−700

90 Mnih V, Kavukcuoglu K, Silver D, Rusu A A, Veness J, Bellemare M G, et al. Human-level control through deep reinforcement learning. *Nature*, 2015, **518**(7540): 529−533

91 Schrittwieser J, Antonoglou I, Hubert T, Simonyan K, Sifre L, Schmitt S, et al. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, 2020, **588**(7839): 604−609

92 Silver D, Huang A, Maddison C J, Guez A, Sifre L, Van Den driessche G, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 2016, **529**(7587): 484−489

93 Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, et al. Mastering the game of Go without human knowledge. *Nature*, 2017, **550**(7676): 354−359

94 Tuyls K, Parsons S. What evolutionary game theory tells us about multiagent learning. *Artificial Intelligence*, 2007, **171**(7): 406−416

95 Bloembergen D, Tuyls K, Hennes D, Kaisers M. Evolutionary dynamics of multi-agent learning: A survey. *Journal of Artificial Intelligence Research*, 2015, **53**: 659−697

96 Mertikopoulos P, Sandholm W H. Learning in games via reinforcement and regularization. *Mathematics of Operations Research*, 2016, **41**(4): 1297−1324

97 Omidshafiei S, Tuyls K, Czarnecki W M, Santos F C, Rowland M, Connor J, et al. Navigating the landscape of multiplayer games. *Nature Communications*, 2020, **11**(1): 5603

98 Leonardos S, Piliouras G. Exploration-exploitation in multi-agent learning: Catastrophe theory meets game theory. *Artificial Intelligence*, 2022, **304**: 103653

99 Hauert C, De Monte S, Hofbauer J, Sigmund K. Replicator dynamics for optional public good games. *Journal of Theoretical Biology*, 2002, **218**(2): 187−194

100 Hauert C, De Monte S, Hofbauer J, Sigmund K. Volunteering as red queen mechanism for cooperation in public goods games. *Science*, 2002, **296**(5570): 1129−1132

101 Pacheco J M, Santos F C, Souza M O, Skyrms B. Evolutionary dynamics of collective action in *N*-person stag hunt dilemmas. *Proceedings of the Royal Society B: Biological Sciences*, 2009, **276**(1655): 315−321

102 Wang J, Fu F, Wu T, Wang L. Emergence of social cooperation in threshold public goods games with collective risk. *Physical Review E*, 2009, **80**(1): 016101

103 Souza M O, Pacheco J M, Santos F C. Evolution of cooperation under *N*-person snowdrift games. *Journal of Theoretical Biology*, 2009, **260**(4): 581−588

104 Wang J, Fu F, Wang L. Effects of heterogeneous wealth distribution on public cooperation with collective risk. *Physical Review E*, 2010, **82**(1): 016102

105 Wang Long, Cong Rui, Li Kun. Feedback mechanism in cooperation evolving. *SCIENCE CHINA: Information Science*, 2014, **44**(12): 1495−1514
(王龙, 丛睿, 李昆. 合作演化中的反馈机制. 中国科学: 信息科学, 2014, **44**(12): 1495−1514)

106 Chen X J, Sasaki T, Brännström Å, Dieckmann U. First carrot, then stick: How the adaptive hybridization of incentives promotes cooperation. *Journal of the Royal Society Interface*, 2015, **12**(102): 20140935

107 Huang F, Chen X J, Wang L. Conditional punishment is a double-edged sword in promoting cooperation. *Scientific Reports*, 2018, **8**(1): 528

108 Huang F, Chen X J, Wang L. Evolution of cooperation in a hierarchical society with corruption control. *Journal of Theoretical Biology*, 2018, **449**: 60−72

109 Ohtsuki H, Nowak M A. The replicator equation on graphs. *Journal of Theoretical Biology*, 2006, **243**(1): 86−97

110 Foster D, Young P. Stochastic evolutionary game dynamics. *Theoretical Population Biology*, 1990, **38**(2): 219−232

111 Imhof L A. The long-run behavior of the stochastic replicator dynamics. *The Annals of Applied Probability*, 2005, **15**(1B): 1019−1045

112 Cressman R. Stability of the replicator equation with continuous strategy space. *Mathematical Social Sciences*, 2005, **50**(2): 127−147

113 Galstyan A. Continuous strategy replicator dynamics for multi-agent Q-learning. *Autonomous Agents and Multi-agent Systems*, 2013, **26**(1): 37−53

114 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*, 2015, **521**(7553): 436−444

115 Nedić A, Liu J. Distributed optimization for control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2018, **1**: 77−103

116 Wang Long, Tian Ye, Du Jin-Ming. Opinion dynamics in social networks. *SCIENCE CHINA: Information Science*, 2018, **48**(1): 3−23
(王龙, 田野, 杜金铭. 社会网络上的观念动力学. 中国科学: 信息科学, 2018, **48**(1): 3−23)

117 Wu B, Du J M, Wang L. Bridging the gap between opinion dynamics and evolutionary game theory: Some equivalence results. In: Proceedings of the 39th Chinese Control Conference. Shenyang, China: IEEE, 2020. 6707−6714

118 Levins R. *Evolution in Changing Environments: Some Theoretical Explorations*. Princeton, USA: Princeton University Press, 1968.

119 Hennes D, Morrill D, Omidshafiei S, Munos R, Perolat J, Lanctot M, et al. Neural replicator dynamics: Multiagent learning via hedging policy gradients. In: Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems. Auckland, New Zealand: International Foundation for Autonomous Agents and Multiagent Systems, 2020. 492−501

120 Pantoja A, Quijano N, Passino K M. Dispatch of distributed generators using a local replicator equation. In: Proceedings of the 50th IEEE Conference on Decision and Control and European Control Conference. Orlando, USA: IEEE, 2011. 7494−7499

121 Barreiro-Gomez J, Obando G, Quijano N. Distributed population dynamics: Optimization and control applications. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2017, **47**(2): 304−314

122 Mei W J, Friedkin N E, Lewis K, Bullo F. Dynamic models of appraisal networks explaining collective learning. *IEEE Transactions on Automatic Control*, 2018, **63**(9): 2898−2912

123 Weitz J S, Eksin C, Paarporn K, Brown S P, Ratcliff W C. An oscillating tragedy of the commons in replicator dynamics with game-environment feedback. *Proceedings of the National Academy of Sciences of the United States of America*, 2016, **113**(47): E7518−E7525

124 Chen X J, Szolnoki A. Punishment and inspection for governing the commons in a feedback-evolving game. *PLoS Computational Biology*, 2018, **14**(7): e1006347

125 Wang X, Zheng Z M, Fu F. Steering eco-evolutionary game dynamics with manifold control. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2020, **476**(2233): 20190643

126 Tilman A R, Plotkin J B, Akçay E. Evolutionary games with environmental feedbacks. *Nature Communications*, 2020, **11**(1): 915

127 Traulsen A, Hauert C. Stochastic evolutionary game dynamics. In: *Heinz Georg Schuster, editor, Reviews of Nonlinear Dynamics and Complexity*. Weinheim: Wiley-VCH, 2009. 25−61

128 Moran P A P. *The Statistical Processes of Evolutionary Theory*. Oxford: Clarendon Press, 1962.

129 Nowak M A, Sasaki A, Taylor C, Fudenberg D. Emergence of cooperation and evolutionary stability in finite populations. *Nature*, 2004, **428**(6983): 646−650

130 Taylor C, Fudenberg D, Sasaki A, Nowak M A. Evolutionary game dynamics in finite populations. *Bulletin of Mathematical Biology*, 2004, **66**(6): 1621−1644

131 Traulsen A, Shoresh N, Nowak M A. Analytical results for individual and group selection of any intensity. *Bulletin of Mathematical Biology*, 2008, **70**(5): 1410−1424

132 Claussen J C, Traulsen A. Cyclic dominance and biodiversity in well-mixed populations. *Physical Review Letters*, 2008, **100**(5): 058104

133 Wild G, Traulsen A. The different limits of weak selection and the evolutionary dynamics of finite populations. *Journal of Theoretical Biology*, 2007, **247**(2): 382−390

134 Wu B, Altrock P M, Wang L, Traulsen A. Universality of weak selection. *Physical Review E*, 2010, **82**(4): 046106

135 Fudenberg D, Nowak M A, Taylor C, Imhof L A. Evolutionary game dynamics in finite populations with strong selection and weak mutation. *Theoretical Population Biology*, 2006, **70**(3): 352−363

136 Antal T, Nowak M A, Traulsen A. Strategy abundance in 2×2 games for arbitrary mutation rates. *Journal of Theoretical Biology*, 2009, **257**(2): 340−344

137 Wu B, Traulsen A, Gokhale C S. Dynamic properties of evolutionary multi-player games in finite populations. *Games*, 2013, **4**(2): 182−199

138 Traulsen A, Pacheco J M, Nowak M A. Pairwise comparison and selection temperature in evolutionary game dynamics. *Journal of Theoretical Biology*, 2007, **246**(3): 522−529

139 Traulsen A, Claussen J C, Hauert C. Coevolutionary dynamics: From finite to infinite populations. *Physical Review Letters*, 2005, **95**(23): 238701

140 Traulsen A, Claussen J C, Hauert C. Coevolutionary dynamics in large, but finite populations. *Physical Review E*, 2006, **74**(1): 011901

141 Huang F, Chen X J, Wang L. Role of the effective payoff function in evolutionary game dynamics. *EPL (Europhysics Letters)*, 2018, **124**(4): 40002

142 Huang F, Chen X J, Wang L. Evolutionary dynamics of networked multi-person games: Mixing opponent-aware and opponent-independent strategy decisions. *New Journal of Physics*, 2019, **21**(6): 063013

143 Wang Long, Fu Feng, Chen Xiao-Jie, Wang Jing, Li Zhuo-Zheng, Xie Guang-Ming, et al. Evolutionary games on complex networks. *CAAI Transactions on Intelligent Systems*, 2007, **2**(2): 1−10
(王龙, 伏锋, 陈小杰, 王靖, 李卓政, 谢广明, 等. 复杂网络上的演化博弈. 智能系统学报, 2007, **2**(2): 1−10)

144 Wang Long, Wu Bin, Du Jin-Ming, Wei Yu-Ting, Zhou Da. Spreading dynamics on complex dynamical networks. *SCIENCE CHINA: Information Science*, 2020, **50**(11): 1714−1731
(王龙, 武斌, 杜金铭, 魏钰婷, 周达. 复杂动态网络上的传播行为分析. 中国科学: 信息科学, 2020, **50**(11): 1714−1731)

145 Lieberman E, Hauert C, Nowak M A. Evolutionary dynamics on graphs. *Nature*, 2005, **433**(7023): 312−316

146 Ohtsuki H, Hauert C, Lieberman E, Nowak M A. A simple rule for the evolution of cooperation on graphs and social networks. *Nature*, 2006, **441**(7092): 502−505

147 Taylor P D, Day T, Wild G. Evolution of cooperation in a fi-

nite homogeneous graph. *Nature*, 2007, **447**(7143): 469−472

148 Su Q, Li A M, Wang L. Evolutionary dynamics under interactive diversity. *New Journal of Physics*, 2017, **19**(10): 103023

149 Su Q, Zhou L, Wang L. Evolutionary multiplayer games on graphs with edge diversity. *PLoS Computational Biology*, 2019, **15**(4): e1006947

150 Su Q, McAvoy A, Wang L, Nowak M A. Evolutionary dynamics with game transitions. *Proceedings of the National Academy of Sciences of the United States of America*, 2019, **116**(51): 25398−25404

151 Li A, Cornelius S P, Liu Y Y, Wang L, Barabási A L. The fundamental advantages of temporal networks. *Science*, 2017, **358**(6366): 1042−1046

152 Li A M, Zhou L, Su Q, Cornelius S P, Liu Y Y, Wang L, Levin S A. Evolution of cooperation on temporal networks. *Nature Communications*, 2020, **11**(1): 2259

153 Wang Long, Wu Te, Zhang Yan-Ling. Feedback mechanism in coevolutionary games. *Control Theory & Applications*, 2014, **31**(7): 823−836
(王龙, 吴特, 张艳玲. 共演化博弈中的反馈机制. 控制理论与应用, 2014, **31**(7): 823−836)

154 Wang Long, Du Jin-Ming. Evolutionary game theoretic approach to coordinated control of multi-agent systems. *Journal of Systems Science and Mathematical Sciences*, 2016, **36**(3): 302−318
(王龙, 杜金铭. 多智能体协调控制的演化博弈方法. 系统科学与数学, 2016, **36**(3): 302−318)

155 Wu T, Fu F, Wang L. Evolutionary games and spatial periodicity. arXiv: 2209.08267, 2022.

156 Axelrod R, Hamilton W D. The evolution of cooperation. *Science*, 1981, **211**(4489): 1390−1396

157 Nowak M, Sigmund K. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature*, 1993, **364**(6432): 56−58

158 Blume L E. The statistical mechanics of strategic interaction. *Games and Economic Behavior*, 1993, **5**(3): 387−424

159 Lessard S. Long-term stability from fixation probabilities in finite populations: New perspectives for ESS theory. *Theoretical Population Biology*, 2005, **68**(1): 19−27

160 Imhof L A, Nowak M A. Evolutionary game dynamics in a Wright-Fisher process. *Journal of Mathematical Biology*, 2006, **52**(5): 667−681

161 Du J M, Wu B, Altrock P M, Wang L. Aspiration dynamics of multi-player games in finite populations. *Journal of the Royal Society Interface*, 2014, **11**(94): 20140077

162 Du J M, Wu B, Wang L. Aspiration dynamics in structured population acts as if in a well-mixed one. *Scientific Reports*, 2015, **5**: 8014

163 Wu B, Zhou L. Individualised aspiration dynamics: Calculation by proofs. *PLoS Computational Biology*, 2018, **14**(9): e1006035

164 Zhou L, Wu B, Vasconcelos V V, Wang L. Simple property of heterogeneous aspiration dynamics: Beyond weak selection. *Physical Review E*, 2018, **98**(6): 062124

165 Zhou L, Wu B, Du J M, Wang L. Aspiration dynamics generate robust predictions in heterogeneous populations. *Nature Communications*, 2021, **12**(1): 3250

166 Shapley L S. Stochastic games. *Proceedings of the National Academy of Sciences of the United States of America*, 1953, **39**(10): 1095−1100

167 Solan E, Vieille N. Stochastic games. *Proceedings of the National Academy of Sciences of the United States of America*, 2015, **112**(45): 13743−13746

168 Bellman R. A Markovian decision process. *Journal of Mathematics and Mechanics*, 1957, **6**(5): 679−684

169 Puterman M L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York: John Wiley & Sons, 2014.

170 Fink A M. Equilibrium in a stochastic n-person game. *Journal of Science of the Hiroshima University, Series A-I Mathematics*, 1964, **28**(1): 89−93

171 Maskin E, Tirole J. Markov perfect equilibrium: I. Observable actions. *Journal of Economic Theory*, 2001, **100**(2): 191−219

172 Daskalakis C, Goldberg P W, Papadimitriou C H. The complexity of computing a Nash equilibrium. *SIAM Journal on Computing*, 2009, **39**(1): 195−259

173 Harsanyi J C. Games with incomplete information played by "Bayesian" players, I-III. *Management Science*, 1967, **14**(3, 5, 7): 159−182, 320−334, 486−502

174 Gibbons R. *A Primer in Game Theory*. Hoboken: Prentice Hall, 1992.

175 Dekel E, Fudenberg D. Rational behavior with payoff uncertainty. *Journal of Economic Theory*, 1990, **52**(2): 243−267

176 Morris S. The common prior assumption in economic theory. *Economics & Philosophy*, 1995, **11**(2): 227−253

177 Mertens J F, Zamir S. Formulation of Bayesian analysis for games with incomplete information. *International Journal of Game Theory*, 1985, **14**(1): 1−29

178 Holmström B, Myerson R B. Efficient and durable decision rules with incomplete information. *Econometrica*, 1983, **51**(6): 1799−1819

179 Ben-Tal A, El Ghaoui L, Nemirovski A. *Robust Optimization*. Princeton, USA: Princeton University Press, 2009.

180 Aghassi M, Bertsimas D. Robust game theory. *Mathematical Programming*, 2006, **107**(1): 231−273

181 Kardeş E, Ordóñez F, Hall R W. Discounted robust stochastic games and an application to queueing control. *Operations Research*, 2011, **59**(2): 365−382

182 Huang Feng. System dynamics and learning theory in games [Ph. D. dissertation], Peking University, China, 2022
(黄锋. 博弈系统动力学与学习理论研究 [博士学位论文], 北京大学, 中国, 2022)

183 Hardin G. The tragedy of the commons. *Science*, 1968, **162**(3859): 1243−1248

184 Robinson J. An iterative method of solving a game. *Annals of Mathematics*, 1951, **54**(2): 296−301

185 Berger U. Brown′s original fictitious play. *Journal of Economic Theory*, 2007, **135**(1): 572−578

186 Hernandez-Leal P, Kaisers M, Baarslag T, de Cote E M. A survey of learning in multiagent environments: Dealing with non-stationarity. arXiv: 1707.09183, 2017.

187 Albrecht S V, Stone P. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, 2018, **258**: 66−95

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

189 Marden J R, Arslan G, Shamma J S. Joint strategy fictitious play with inertia for potential games. *IEEE Transactions on Automatic Control*, 2009, **54**(2): 208−220

190 Swenson B, Kar S, Xavier J. Empirical centroid fictitious play: An approach for distributed learning in multi-agent games. *IEEE Transactions on Signal Processing*, 2015, **63**(15): 3888−3901

191 Eksin C, Ribeiro A. Distributed fictitious play for multiagent systems in uncertain environments. *IEEE Transactions on Automatic Control*, 2018, **63**(4): 1177−1184

192 Swenson B, Kar S, Xavier J. Single sample fictitious play. *IEEE Transactions on Automatic Control*, 2017, **62**(11):

6026−6031

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

194 Sayin M O, Parise F, Ozdaglar A. Fictitious play in zero-sum stochastic games. *SIAM Journal on Control and Optimization*, 2022, **60**(4): 2095−2114

195 Perrin S, Perolat J, Laurière M, Geist M, Elie R, Pietquin O. Fictitious play for mean field games: Continuous time analysis and applications. In: Proceedings of the 34th Conference on Neural Information Processing Systems. Vancouver, Canada, 2020. 13199−13213

196 Thorndike E L. *Animal Intelligence: Experimental Studies*. New York, USA: Macmillan, 1911.

197 Sutton R S, Barto A G. *Reinforcement Learning: An Introduction*. Cambridge, USA: MIT Press, 2018.

198 Sato Y, Akiyama E, Crutchfield J P. Stability and diversity in collective adaptation. *Physica D: Nonlinear Phenomena*, 2005, **210**(1-2): 21−57

199 Galla T, Farmer J D. Complex dynamics in learning complicated games. *Proceedings of the National Academy of Sciences of the United States of America*, 2013, **110**(4): 1232−1236

200 Barfuss W, Donges J F, Kurths J. Deterministic limit of temporal difference reinforcement learning for stochastic games. *Physical Review E*, 2019, **99**(4): 043305

201 Kianercy A, Galstyan A. Dynamics of Boltzmann $Q$ learning in two-player two-action games. *Physical Review E*, 2012, **85**(4): 041145

202 Galla T. Intrinsic noise in game dynamical learning. *Physical Review Letters*, 2009, **103**(19): 198702

203 Kaisers M, Tuyls K. Frequency adjusted multi-agent Q-learning. In: Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems, Toronto, Canada, 2010. 309−316

204 Abdallah S, Lesser V. A multiagent reinforcement learning algorithm with non-linear dynamics. *Journal of Artificial Intelligence Research*, 2008, **33**: 521−549

205 Klos T, van Ahee G J, Tuyls K. Evolutionary dynamics of regret minimization. In: Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Barcelona, Spain: Springer, 2010. 82−96

206 Huang F, Cao M, Wang L. Learning enables adaptation in cooperation for multi-player stochastic games. *Journal of the Royal Society Interface*, 2020, **17**(172): 20200639

207 Sutton R S, McAllester D, Singh S, Mansour Y. Policy gradient methods for reinforcement learning with function approximation. In: Proceedings of the 12th International Conference on Neural Information Processing Systems. Denver, USA: MIT Press, 1999. 1057−1063

208 Konda V R, Tsitsiklis J N. Actor-critic algorithms. In: Proceedings of the 12th International Conference on Neural Information Processing Systems. Denver, USA: MIT Press, 1999. 1008−1014

209 Arapostathis A, Borkar V S, Fernández-Gaucherand E, Ghosh M K, Marcus S I. Discrete-time controlled Markov processes with average cost criterion: A survey. *SIAM Journal on Control and Optimization*, 1993, **31**(2): 282−344

210 Bellman R E. *Dynamic Programming*. Princeton, USA: Princeton University Press, 1957.

211 Bertsekas D P, Tsitsiklis J N. *Neuro-Dynamic Programming*. Belmont, USA: Athena Scientific, 1996.

212 Kaelbling L P, Littman M L, Moore A W. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 1996, **4**: 237−285

213 Hessel M, Modayil J, van Hasselt H, Schaul T, Ostrovski G, Dabney W, et al. Rainbow: Combining improvements in deep reinforcement learning. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence and 30th Innovative Applications of Artificial Intelligence Conference and 8th AAAI Symposium on Educational Advances in Artificial Intelligence. New Orleans, USA: AAAI Press, 2018. 3215−3222

214 Kocsis L, Szepesvári C. Bandit based Monte-Carlo planning. In: Proceedings of the 17th European Conference on Machine Learning. Berlin, Germany: Springer, 2006. 282−293

215 Coulom R. Efficient selectivity and backup operators in Monte-Carlo tree search. In: Proceedings of the 5th International Conference on Computers and Games. Turin, Italy: Springer, 2006. 72−83

216 Silver D, Lever G, Heess N, Degris T, Wierstra D, Riedmiller M. Deterministic policy gradient algorithms. In: Proceedings of the 31st International Conference on Machine Learning. Beijing, China: JMLR.org, 2014. 387−395

217 Lillicrap T P, Hunt J J, Pritzel A, Heess N, Erez T, Tassa Y, et al. Continuous control with deep reinforcement learning. In: Proceedings of the 4th International Conference on Learning Representations. San Juan, Puerto Rico, 2016. 1−10

218 Schulman J, Levine S, Moritz P, Jordan M, Abbeel P. Trust region policy optimization. In: Proceedings of the 32nd International Conference on Machine Learning. Lille, France: JMLR.org, 2015. 1889−1897

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

220 Mnih V, Badia A P, Mirza M, Graves A, Harley T, Lillicrap T P, et al. Asynchronous methods for deep reinforcement learning. In: Proceedings of the 33rd International Conference on Machine Learning. New York, USA: JMLR.org, 2016. 1928−1937

221 Haarnoja T, Zhou A, Abbeel P, Levine S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: PMLR, 2018. 1861−1870

222 Claus C, Boutilier C. The dynamics of reinforcement learning in cooperative multiagent systems. In: Proceedings of the 15th National Conference on Artificial Intelligence and 10th Innovative Applications of Artificial Intelligence. Madison, USA: AAAI Press, 1998. 746−752

223 Littman M L. Value-function reinforcement learning in Markov games. *Cognitive Systems Research*, 2001, **2**(1): 55−66

224 Foerster J N, Assael Y M, de Freitas N, Whiteson S. Learning to communicate with deep multi-agent reinforcement learning. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: Curran Associates Inc, 2016. 2145−2153

225 García J, Fernández F. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 2015, **16**(1): 1437−1480

226 Hernandez-Leal P, Kartal B, Taylor M E. A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 2019, **33**(6): 750−797

227 Da Silva F L, Costa A H R. A survey on transfer learning for multiagent reinforcement learning systems. *Journal of Artificial Intelligence Research*, 2019, **64**: 645−703

228 Oroojlooy A, Hajinezhad D. A review of cooperative multi-agent deep reinforcement learning. *Applied Intelligence*, 2022.

229 Nguyen T T, Nguyen N D, Nahavandi S. Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications. *IEEE Transactions on Cybernetics*, 2020, **50**(9): 3826−3839

230 Heuillet A, Couthouis F, Díaz-Rodríguez N. Explainability in deep reinforcement learning. *Knowledge-Based Systems*, 2021, **214**: 106685

231 Dulac-Arnold G, Levine N, Mankowitz D J, Li J, Paduraru C, Gowal S, et al. Challenges of real-world reinforcement learning: Definitions, benchmarks and analysis. *Machine Learning*, 2021, **110**(9): 2419−2468

232 Parker-Holder J, Rajan R, Song X Y, Biedenkapp A, Miao Y J, Eimer T, et al. Automated reinforcement learning (AutoRL): A survey and open problems. *Journal of Artificial Intelligence Research*, 2022, **74**: 517−568

233 Morimoto J, Doya K. Robust reinforcement learning. *Neural Computation*, 2005, **17**(2): 335−359

234 Moos J, Hansel K, Abdulsamad H, Stark S, Clever D, Peters J. Robust reinforcement learning: A review of foundations and recent advances. *Machine Learning and Knowledge Extraction*, 2022, **4**(1): 276−315

235 Satia J K, Lave R E Jr. Markovian decision processes with uncertain transition probabilities. *Operations Research*, 1973, **21**(3): 728−740

236 White III C C, Eldeib H K. Markov decision processes with imprecise transition probabilities. *Operations Research*, 1994, **42**(4): 739−749

237 Zhou K, Doyle J C, Glover K. *Robust and Optimal Control*. Upper Saddle River, USA: Prentice Hall, 1996.

238 Nilim A, El Ghaoui L. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 2005, **53**(5): 780−798

239 Iyengar G N. Robust dynamic programming. *Mathematics of Operations Research*, 2005, **30**(2): 257−280

240 Goodfellow I J, Pouget-Abadie J, Mirza M, Xu M, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press, 2014. 2672−2680

241 Kaufman D L, Schaefer A J. Robust modified policy iteration. *INFORMS Journal on Computing*, 2013, **25**(3): 396−410

242 Xu H, Mannor S. The robustness-performance tradeoff in Markov decision processes. In: Proceedings of the 19th International Conference on Neural Information Processing Systems. Vancouver, Canada: MIT Press, 2006. 1537−1544

243 Delage E, Mannor S. Percentile optimization for Markov decision processes with parameter uncertainty. *Operations Research*, 2010, **58**(1): 203−213

244 Turchetta M, Krause A, Trimpe S. Robust model-free reinforcement learning with multi-objective Bayesian optimization. In: Proceedings of IEEE International Conference on Robotics and Automation (ICRA). Paris, France: IEEE, 2020. 10702−10708

245 Mannor S, Mebel O, Xu H. Robust MDPs with k-rectangular uncertainty. *Mathematics of Operations Research*, 2016, **41**(4): 1484−1509

246 Goyal V, Grand-Clément J. Robust Markov decision processes: Beyond rectangularity. *Mathematics of Operations Research*, 2022.

247 Gilboa I, Schmeidler D. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 1989, **18**(2): 141−153

248 Xu H, Mannor S. Distributionally robust Markov decision processes. *Mathematics of Operations Research*, 2012, **37**(2): 288−300

249 Tamar A, Mannor S, Xu H. Scaling up robust MDPs using function approximation. In: Proceedings of the 31st International Conference on Machine Learning. Beijing, China: JMLR.org, 2014. 181−189

250 Scherrer B, Ghavamzadeh M, Gabillon V, Lesner B, Geist M. Approximate modified policy iteration and its application to the game of Tetris. *The Journal of Machine Learning Research*, 2015, **16**(1): 1629−1676

251 Badrinath K P, Kalathil D. Robust reinforcement learning using least squares policy iteration with provable performance guarantees. In: Proceedings of the 38th International Conference on Machine Learning. Virtual: PMLR, 2021. 511−520

252 Pinto L, Davidson J, Sukthankar R, Gupta A. Robust adversarial reinforcement learning. In: Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: JMLR.org, 2017. 2817−2826

253 Phan T, Belzner L, Gabor T, Sedlmeier A, Ritz F, Linnhoff-Popien C. Resilient multi-agent reinforcement learning with adversarial value decomposition. In: Proceedings of the 35th AAAI Conference on Artificial Intelligence. Virtual: AAAI Press, 2021. 11308−11316

254 Everett M, Lütjens B, How J P. Certifiable robustness to adversarial state uncertainty in deep reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2022, **33**(9): 4184−4198

255 Mankowitz D J, Levine N, Jeong R, Abdolmaleki A, Springenberg J T, Shi Y Y, et al. Robust reinforcement learning for continuous control with model misspecification. In: Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia: OpenReview.net, 2020. 1−11

256 Si N, Zhang F, Zhou Z Y, Blanchet J. Distributionally robust policy evaluation and learning in offline contextual bandits. In: Proceedings of the 37th International Conference on Machine Learning. Vienna, Austria: JMLR.org, 2020. 8884−8894

257 Zhou Z Q, Bai Q X, Zhou Z Y, Qiu L H, Blanchet J H, Glynn P W. Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In: Proceedings of the 24th International Conference on Artificial Intelligence and Statistics. San Diego, USA: PMLR, 2021. 3331−3339

258 Zhang K Q, Sun T, Tao Y Z, Genc S, Mallya S, Başar T. Robust multi-agent reinforcement learning with model uncertainty. In: Proceedings of the 34th Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc, 2020. 10571–10583

259 Huang F, Cao M, Wang L. Robust optimal policies for team Markov games. arXiv: 2105.07405, 2021.

260 Lewis F L, Vrabie D L, Syrmos V L. *Optimal Control*. Hoboken, USA: John Wiley & Sons, 2012.

261 Bertsekas D P. *Reinforcement Learning and Optimal Control*. Belmont, USA: Athena Scientific, 2019.

262 Bryson A E. Optimal control—1950 to 1985. *IEEE Control Systems Magazine*, 1996, **16**(3): 26−33

263 Doya K. Reinforcement learning in continuous time and space. *Neural Computation*, 2000, **12**(1): 219−245

264 Liu D R, Wei Q L, Wang D, Yang X, Li H L. *Adaptive Dynamic Programming with Applications in Optimal Control*. Cham, Switzerland: Springer, 2017.

265 Jiang Z P, Jiang Y. Robust adaptive dynamic programming for linear and nonlinear systems: An overview. *European Journal of Control*, 2013, **19**(5): 417−425

266 Fu K S. Learning control systems—review and outlook. *IEEE Transactions on Automatic Control*, 1970, **15**(2): 210−221

267 Vrabie D, Lewis F. Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems. *Neural Networks*, 2009, **22**(3): 237−246

268 Modares H, Lewis F L, Naghibi-Sistani M B. Integral reinforcement learning and experience replay for adaptive optimal control of partially-unknown constrained-input continuous-time systems. *Automatica*, 2014, **50**(1): 193−202

269 Jiang Y, Jiang Z P. Robust adaptive dynamic programming and feedback stabilization of nonlinear systems. *IEEE Transactions on Neural Networks and Learning Systems*, 2014, **25**(5): 882−893

270 Isaacs R. *Differential Games: A Mathematical Theory with Ap-*

*plications to Warfare and Pursuit, Control and Optimization*. Mineola, USA: Dover Publications, 1999.

271 Ho Y C, Bryson A, Baron S. Differential games and optimal pursuit-evasion strategies. *IEEE Transactions on Automatic Control*, 1965, **10**(4): 385−389

272 Starr A W, Ho Y C. Nonzero-sum differential games. *Journal of Optimization Theory and Applications*, 1969, **3**(3): 184−206

273 Vamvoudakis K G, Lewis F L. Multi-player non-zero-sum games: Online adaptive learning solution of coupled Hamilton-Jacobi equations. *Automatica*, 2011, **47**(8): 1556−1569

274 Vamvoudakis K G, Lewis F L, Hudas G R. Multi-agent differential graphical games: Online adaptive learning solution for synchronization with optimality. *Automatica*, 2012, **48**(8): 1598−1611

275 Kamalapurkar R, Klotz J R, Walters P, Dixon W E. Model-based reinforcement learning in differential graphical games. *IEEE Transactions on Control of Network Systems*, 2018, **5**(1): 423−433

276 Li M, Qin J H, Freris N M, Ho D W C. Multiplayer Stackelberg-Nash game for nonlinear system via value iteration-based integral reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2022, **33**(4): 1429−1440

277 Guo Lei. Some thoughts on the development of control theory. *Journal of Systems Science and Mathematical Sciences*, 2011, **31**(9): 1014−1018
(郭雷. 关于控制理论发展的某些思考. 系统科学与数学, 2011, **31**(9): 1014−1018)

278 Zhang R R, Guo L. Controllability of Nash equilibrium in game-based control systems. *IEEE Transactions on Automatic Control*, 2019, **64**(10): 4180−4187

279 Zhang R R, Guo L. Controllability of stochastic game-based control systems. *SIAM Journal on Control and Optimization*, 2019, **57**(6): 3799−3826

280 Zhang R R, Guo L. Stabilizability of game-based control systems. *SIAM Journal on Control and Optimization*, 2021, **59**(5): 3999−4023

281 Simaan M, Cruz J B. On the Stackelberg strategy in nonzero-sum games. *Journal of Optimization Theory and Applications*, 1973, **11**(5): 533−555

282 Li Y, Guo L. Towards a theory of stochastic adaptive differential games. In: Proceedings of the 50th IEEE Conference on Decision and Control and European Control Conference. Orlando, USA: IEEE, 2011. 5041−5046

283 Yuan Shuo, Guo Lei. Stochastic adaptive dynamical games. *Scientia Sinica Mathematica*, 2016, **46**(10): 1367−1382
(袁硕, 郭雷. 随机自适应动态博弈. 中国科学: 数学, 2016, **46**(10): 1367−1382)

284 Liu N, Guo L. Stochastic adaptive linear quadratic differential games. arXiv: 2204.08869, 2022.

285 Cheng D Z, Qi H S, Zhao Y. *An Introduction to Semi-Tensor Product of Matrices and Its Applications*. Singapore: World Scientific, 2012.

286 Cheng D Z, Qi H S, Li Z Q. *Analysis and Control of Boolean Networks: A Semi-Tensor Product Approach*. London: Springer, 2011.

287 Cheng D Z. On finite potential games. *Automatica*, 2014, **50**(7): 1793−1801

288 Cheng D Z, Wu Y H, Zhao G D, Fu S H. A comprehensive survey on STP approach to finite games. *Journal of Systems Science and Complexity*, 2021, **34**(5): 1666−1680

289 Cheng D Z, He F H, Qi H S, Xu T T. Modeling, analysis and control of networked evolutionary games. *IEEE Transactions on Automatic Control*, 2015, **60**(9): 2402−2415

290 Monderer D, Shapley L S. Potential games. *Games and Economic Behavior*, 1996, **14**(1): 124−143

291 Cheng D Z, Liu T. From Boolean game to potential game. *Automatica*, 2018, **96**: 51−60

292 Li C X, Xing Y, He F H, Cheng D Z. A strategic learning algorithm for state-based games. *Automatica*, 2020, **113**: 108615

293 Yang T, Yi X L, Wu J F, Yuan Y, Wu D, Meng Z Y, et al. A survey of distributed optimization. *Annual Reviews in Control*, 2019, **47**: 278−305

294 Wang Long, Lu Kai-Hong, Guan Yong-Qiang. Distributed optimization via multi-agent systems. *Control Theory & Applications*, 2019, **36**(11): 1820−1833
(王龙, 卢开红, 关永强. 分布式优化的多智能体方法. 控制理论与应用, 2019, **36**(11): 1820−1833)

295 Facchinei F, Kanzow C. Generalized Nash equilibrium problems. *Annals of Operations Research*, 2010, **175**(1): 177−211

296 Scutari G, Palomar D P, Facchinei F, Pang J S. Convex optimization, game theory, and variational inequality theory. *IEEE Signal Processing Magazine*, 2010, **27**(3): 35−49

297 Facchinei F, Pang J S. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. New York, USA: Springer, 2003.

298 Nagurney A, Zhang D. *Projected Dynamical Systems and Variational Inequalities with Applications*. New York, USA: Springer, 1996.

299 Salehisadaghiani F, Pavel L. Distributed Nash equilibrium seeking: A gossip-based algorithm. *Automatica*, 2016, **72**: 209−216

300 Gadjov D, Pavel L. A passivity-based approach to Nash equilibrium seeking over networks. *IEEE Transactions on Automatic Control*, 2019, **64**(3): 1077−1092

301 Ye M J, Hu G Q. Distributed Nash equilibrium seeking by a consensus based approach. *IEEE Transactions on Automatic Control*, 2017, **62**(9): 4811−4818

302 Liang S, Yi P, Hong Y G. Distributed Nash equilibrium seeking for aggregative games with coupled constraints. *Automatica*, 2017, **85**: 179−185

303 Lu K H, Jing G S, Wang L. Distributed algorithms for searching generalized Nash equilibrium of noncooperative games. *IEEE Transactions on Cybernetics*, 2019, **49**(6): 2362−2371

304 Lu K H, Li G Q, Wang L. Online distributed algorithms for seeking generalized Nash equilibria in dynamic environments. *IEEE Transactions on Automatic Control*, 2021, **66**(5): 2289−2296

305 Press W H, Dyson F J. Iterated prisoner's dilemma contains strategies that dominate any evolutionary opponent. *Proceedings of the National Academy of Sciences of the United States of America*, 2012, **109**(26): 10409−10413

306 Stewart A J, Plotkin J B. Extortion and cooperation in the prisoner's dilemma. *Proceedings of the National Academy of Sciences of the United States of America*, 2012, **109**(26): 10134−10135

307 Stewart A J, Plotkin J B. From extortion to generosity, evolution in the iterated prisoner′s dilemma. *Proceedings of the National Academy of Sciences of the United States of America*, 2013, **110**(38): 15348−15353

308 Akin E. The iterated prisoner′s dilemma: Good strategies and their dynamics. In: *Assani Idris, editor, Ergodic Theory: Advances in Dynamical Systems*. Berlin: De Gruyter, 2016. 77−107

309 Hilbe C, Traulsen A, Sigmund K. Partners or rivals? Strategies for the iterated prisoner's dilemma. *Games and Economic Behavior*, 2015, **92**: 41−52

310 McAvoy A, Hauert C. Autocratic strategies for iterated games with arbitrary action spaces. *Proceedings of the National Academy of Sciences of the United States of America*, 2016, **113**(13): 3573−3578

311 Hilbe C, Wu B, Traulsen A, Nowak M A. Cooperation and control in multiplayer social dilemmas. *Proceedings of the National Academy of Sciences of the United States of America*, 2014, **111**(46): 16425−16430

312 Pan L M, Hao D, Rong Z H, Zhou T. Zero-determinant strategies in iterated public goods game. *Scientific Reports*, 2015, **5**: 13096

313 Govaert A, Cao M. Zero-determinant strategies in repeated multiplayer social dilemmas with discounted payoffs. *IEEE Transactions on Automatic Control*, 2021, **66**(10): 4575−4588

314 Tan R F, Su Q, Wu B, Wang L. Payoff control in repeated games. In: Proceedings of the 33rd Chinese Control and Decision Conference. Kunming, China: Editorial Department of Control and Decision, 2021. 997−1005

315 Chen J, Zinger A. The robustness of zero-determinant strategies in iterated prisoner′s dilemma games. *Journal of Theoretical Biology*, 2014, **357**: 46−54

316 Adami C, Hintze A. Evolutionary instability of zero-determinant strategies demonstrates that winning is not everything. *Nature Communications*, 2013, **4**(1): 2193

317 Hilbe C, Nowak M A, Sigmund K. Evolution of extortion in iterated prisoner′s dilemma games. *Proceedings of the National Academy of Sciences of the United States of America*, 2013, **110**(17): 6913−6918

318 Chen F, Wu T, Wang L. Evolutionary dynamics of zero-determinant strategies in repeated multiplayer games. *Journal of Theoretical Biology*, 2022, **549**: 111209

319 Cheng Z Y, Chen G P, Hong Y G. Misperception influence on zero-determinant strategies in iterated prisoner′s dilemma. *Scientific Reports*, 2022, **12**(1): 5174

320 Barto A G. Reinforcement learning: Connections, surprises, and challenge. *AI Magazine*, 2019, **40**(1): 3−15

321 Huang X W, Kroening D, Ruan W J, Sharp J, Sun Y C, Thamo E, et al. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review*, 2020, **37**: 100270

322 Kuhn H W. Extensive games and the problem of information. *In: Kuhn H W and Tucker A W, editors, Contributions to the Theory of Games (AM-28), Volume II*. Princeton, USA: Princeton University Press, 1953. 193−216

323 Bai Y, Jin C, Mei S, Yu T C. Near-optimal learning of extensive-form games with imperfect information. In: Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR, 2022. 1337−1382

324 Spence M. Job market signaling. *The Quarterly Journal of Economics*, 1973, **87**(3): 355−374

325 Meyer D A. Quantum strategies. *Physical Review Letters*, 1999, **82**(5): 1052−1055

326 Eisert J, Wilkens M, Lewenstein M. Quantum games and quantum strategies. *Physical Review Letters*, 1999, **83**(15): 3077−3080

327 Khan F S, Solmeyer N, Balu R, Humble T S. Quantum games: A review of the history, current state, and interpretation. *Quantum Information Processing*, 2018, **17**(11): 309

328 Huang M Y, Malhamé R P, Caines P E. Large population stochastic dynamic games: Closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle. *Communications in Information and Systems*, 2006, **6**(3): 221−252

329 Lasry J M, Lions P L. Mean field games. *Japanese Journal of Mathematics*, 2007, **2**(1): 229−260

330 Weinan E. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 2017, **5**(1): 1−11

331 Chen R T Q, Rubanova Y, Bettencourt J, Duvenaud D. Neural ordinary differential equations. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada: Curran Associates Inc, 2018. 6572−6583

332 Kidger P. On Neural Differential Equations [Ph. D. dissertation], University of Oxford, UK, 2022

**王 龙** 北京大学教授. 1992 年获得北京大学博士学位. 主要研究方向为人工智能, 博弈控制理论, 演化动力学. 本文通信作者.
E-mail: longwang@pku.edu.cn
(**WANG Long** Professor at Peking University. He received his Ph.D. degree from Peking University in 1992. His research interest covers artificial intelligence, game and control theory, and evolutionary dynamics. Corresponding author of this paper.)

**黄 锋** 北京大学博士研究生. 2016 年获得电子科技大学学士学位. 主要研究方向为博弈论、多智能体学习与控制论间的交叉.
E-mail: fenghuang@pku.edu.cn
(**HUANG Feng** Ph.D. candidate at Peking University. He received his bachelor degree from University of Electronic Science and Technology of China in 2016. His research interest covers game theory, multi-agent learning, and control theory.)