

多智能体强化学习控制与决策研究综述

罗彪¹ 胡天萌¹ 周育豪¹ 黄廷文¹ 阳春华¹ 桂卫华¹

摘要 强化学习作为一类重要的人工智能方法, 广泛应用于解决复杂的控制和决策问题, 其在众多领域的应用已展示出巨大潜力. 近年来, 强化学习从单智能体决策逐渐扩展到多智能体协作与博弈, 形成多智能体强化学习这一研究热点. 多智能体系统由多个具有自主感知和决策能力的实体组成, 有望解决传统单智能体方法难以应对的大规模复杂问题. 多智能体强化学习不仅需要考虑到环境的动态性, 还需要应对其他智能体策略的不确定性, 从而增加学习和决策过程的复杂度. 为此, 梳理多智能体强化学习在控制与决策领域的研究, 分析其面临的主要问题与挑战, 从控制理论与自主决策两个层次综述现有的研究成果与进展, 并对未来的研究方向进行展望. 通过分析, 期望为未来多智能体强化学习的研究提供有价值的参考和启示.

关键词 强化学习, 多智能体系统, 序列决策, 协同控制, 博弈论

引用格式 罗彪, 胡天萌, 周育豪, 黄廷文, 阳春华, 桂卫华. 多智能体强化学习控制与决策研究综述. 自动化学报, 2025, 51(3): 510-539

DOI 10.16383/j.aas.c240392 **CSTR** 32138.14.j.aas.c240392

Survey on Multi-agent Reinforcement Learning for Control and Decision-making

LUO Biao¹ HU Tian-Meng¹ ZHOU Yu-Hao¹ HUANG Ting-Wen¹ YANG Chun-Hua¹ GUI Wei-Hua¹

Abstract Reinforcement learning, as an important method in artificial intelligence, has been widely used to solve complex control and decision-making problems and has shown great potential in various fields. Recently, multi-agent reinforcement learning has evolved from single-agent decision-making to multi-agent cooperation and competition, forming a research hotspot in the field. Multi-agent systems consist of multiple entities with autonomous perception and decision-making capabilities, which hold promise for solving large-scale complex problems that traditional single-agent methods struggle to address. Multi-agent reinforcement learning must not only account for the dynamic nature of the environment, but also deal with uncertainties in the strategies of other agents, adding complexity to the learning and decision-making processes. This paper reviews the research on multi-agent reinforcement learning in the fields of control and decision-making, analyzes the major problems and challenges, summarizes existing results and advances from the perspectives of control theory and decision-making, and looks forward to future research directions. Through this survey, the paper aims to provide valuable references and insights for future research in multi-agent reinforcement learning.

Key words Reinforcement learning, multi-agent system, sequential decision-making, collaborative control, game theory

Citation Luo Biao, Hu Tian-Meng, Zhou Yu-Hao, Huang Ting-Wen, Yang Chun-Hua, Gui Wei-Hua. Survey on multi-agent reinforcement learning for control and decision-making. *Acta Automatica Sinica*, 2025, 51(3): 510-539

强化学习是一类重要的控制与决策方法, 其基本思路是智能体通过与环境交互来学习如何在给定情境下做出最佳决策. 强化学习基于试错学习的概念, 智能体通过执行动作并观察结果来优化其行为, 目标是最大化长期奖励^[1]. 近年来, 强化学习解决复

杂决策问题的能力不断突破, 已经在机器人、自动驾驶、推荐系统^[2]、博弈等领域取得显著成就. 例如, DeepMind 发布的人工智能围棋程序 AlphaGo^[3] 和 AlphaGo Zero^[4], 将强化学习、监督学习和蒙特卡洛树搜索结合起来, 成功学会了超越人类围棋冠军的对弈策略, 并发现了一些新的围棋策略和走法; 在机器人控制领域, 利用分层强化学习方法训练的机器狗守门员 Mini Cheetah^[5], 能够对快速飞行的足球进行精准拦截; 在博弈领域, DeepMind 开发的 DeepNash 智能体^[6] 能够从零开始学习西洋陆军棋 Stratego 并成功达到人类专家水平, 这是一个行动和结果之间没有明显联系的不完全信息博弈问题;

收稿日期 2024-06-26 录用日期 2024-09-03

Manuscript received June 26, 2024; accepted September 3, 2024

国家自然科学基金 (62373375, U2341216) 资助

Supported by National Natural Science Foundation of China (62373375, U2341216)

本文责任编辑 杨涛

Recommended by Associate Editor YANG Tao

1. 中南大学自动化学院 长沙 410083

1. School of Automation, Central South University, Changsha 410083

在数学领域, DeepMind 推出 AlphaTensor 智能体^[7], 专注于寻找矩阵乘法的更高效算法, 该方法在多种不同大小的矩阵乘法任务中成功超越了现有的最佳算法。

在交通调度、自动驾驶、智能电网、工业控制等现实场景中, 常常存在多个实体同时参与决策过程, 这类由多个具有一定自主感知和决策能力的实体构成的系统称为多智能体系统. 通过独立个体之间的通信、协作或竞争, 多智能体系统能够实现远超单个体的复杂行为, 有望解决单智能体方法难以处理的大规模复杂问题. 为此, 多智能体强化学习扩展了传统的强化学习范式, 允许多个智能体共同在环境中学习和决策, 期望通过个体局部交互涌现出复杂的全局行为或智能. 在多智能体强化学习中, 智能体根据对环境的观测采取行动, 并根据其行动的结果获得奖励或惩罚. 然而, 与单智能体强化学习不同, 多智能体系统中每个智能体的学习和决策过程不仅受到环境的影响, 还受到其他智能体行为的影响. 因此, 在多智能体环境中, 任何一个智能体的最优策略可能都依赖于其他智能体的策略. 这种相互依赖性增加了学习过程的复杂性, 因为智能体必须考虑到其他智能体的潜在行动和策略变化.

近年来, 多智能体强化学习快速发展, 相关研究可分为两条密切相关的脉络. 一部分研究从控制理论和优化方法的角度, 面向多智能体协同控制问题, 设计有效的多智能体强化学习控制算法, 在多种控制目标下实现系统稳定与性能优化; 另一部分研究从人工智能和机器学习的角度, 面向不确定环境中的序列决策问题, 研究多智能体强化学习决策方法, 在未知环境中学习高效的多智能体协调合作或对抗博弈策略. 这两类方法从不同的角度研究多智能体控制与决策问题, 有其各自的优势与特点. 本文对多智能体强化学习控制与决策领域的问题、挑战与方法进行分析与探讨, 以期为后续相关研究提供有价值的参考. 首先, 在多智能体强化学习控制方面, 从多智能体博弈和多智能体协同两个角度, 介绍强化学习算法在各类博弈与协同控制问题中的应用; 继而, 在多智能体强化学习决策方面, 从问题的视角出发, 针对性地分析多智能体强化学习所面临的难题与挑战, 系统性地综述相应解决方案和研究进展; 最后, 介绍多智能体强化学习控制与决策方法在几类现实领域的应用研究, 并对未来可能的探索方向进行总结与展望.

1 多智能体强化学习控制

强化学习控制作为一种新兴的控制方法, 在智能系统和控制理论领域引起了广泛关注. 多智能体

强化学习控制将强化学习技术与多智能体系统的控制问题相结合, 以解决多智能体系统中的竞争与协同问题. 一方面, 多智能体博弈理论被引入强化学习框架中, 用于建模智能体之间的竞争、合作和博弈关系. 在这种情况下, 智能体需要通过学习和优化策略来应对对手的行为, 从而最大化自身收益. 另一方面, 多智能体协同控制成为研究的重点, 通过强化学习技术实现智能体之间的协同合作, 共同完成任务或达成共同目标. 在这个过程中, 智能体需要学习如何有效地分配资源、协调行动和互相协助, 以达到整体系统性能的最优化. 综合而言, 多智能体强化学习控制既涉及智能体之间的竞争与合作关系的建模和优化, 又探索了智能体如何在协同作战中实现更高效的决策和行动. 这一领域的研究不仅推动了智能系统的发展, 也为解决实际应用中的复杂多智能体系统问题提供了新的思路和方法. 本节将从多智能体博弈和多智能体协同两个方面介绍强化学习控制的相关工作, 本节框架结构关系如图 1 所示.

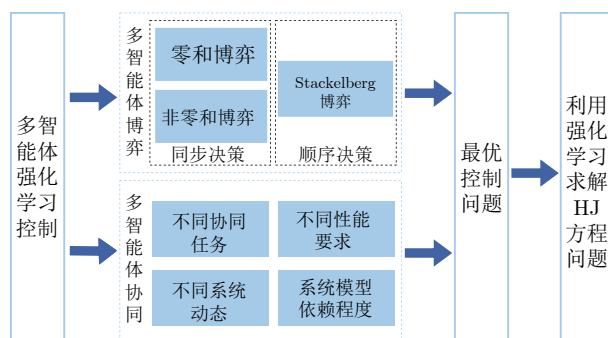


图 1 多智能体强化学习控制

Fig. 1 Multi-agent reinforcement learning control

无论是多智能体博弈还是多智能体协同, 都需要依赖智能体间的数据传输. 多智能体图论是研究多个智能体之间相互作用和通信的一种重要方法. 在此框架下, 智能体 i 表示为图的节点, 节点间的相互作用则通过图的边来描述. 这种方法在许多领域都有广泛的应用, 包括分布式控制、协作机器人、通信网络、社交网络等. 以单领航者协同一致性为例, 本文中有向图由 $\mathcal{G} = (\mathcal{V}, \Lambda, \mathcal{A})$ 表示, $\mathcal{V} = \{p_1, \dots, p_N\}$ 表示节点集合; $\Lambda(p_i, p_j) \in \mathcal{V} \times \mathcal{V}$ 表示边集; $\mathcal{A} = [a_{ij}]_{N \times N}$ 表示系统间的邻接矩阵. 当节点 p_i 能接收到节点 p_j 的信息时, 令邻接矩阵元素 $a_{ij} > 0$, 那么智能体 j 就称为智能体 i 的邻居智能体, 而 $j \in P_i = \{p_j | (p_j, p_i) \in \Lambda\}$ 则称为智能体 i 的邻居集合; 否则, $a_{ij} = 0$. 定义跟随矩阵为 $\mathcal{A}_0 = \text{diag}\{a_{10}, \dots, a_{N0}\}$, 其中若智能体 i 可以接收领航者的信息, 则 $a_{i0} = 1$; 否则 $a_{i0} = 0$. 定义度矩阵为 $\mathcal{D} = \text{diag}\{d_1, \dots,$

$d_N\}$, 其中 $d_i = \sum_{j \in P_i} a_{ij}$, 进一步将拉普拉斯矩阵定义为 $\mathcal{L} = \mathcal{D} - \mathcal{A}$. 对于无领航者协同一致性, 多智能体子系统间则不存在跟随矩阵. 此外, 多领航者协同一致性相关图论知识可见文献 [8], 在此不详细赘述.

1.1 多智能体博弈

博弈理论在多智能体系统中的应用十分广泛. 按照智能体间的相互作用关系, 可分为零和博弈与非零和博弈. 在零和博弈中, 各参与方的利益完全对立, 一方的收益等于另一方的损失, 在如此竞争对抗的环境下, 智能体需要通过精心设计的策略来最大化自身利益, 同时限制对手的收益; 而在非零和博弈中, 参与方的利益不一定完全对立, 因此需要更多地考虑合作与竞争的平衡. 此外, 与零和、非零和博弈不同的是, Stackelberg 博弈作为一种序贯博弈模型, 领导者在做出决策后会影响到跟随者的决策, 这种模型常应用于多智能体系统中的领导跟随一致性问题, 其中领导者的行动对跟随者具有示范作用, 而跟随者则会根据领导者的决策做出相应的反应. 综合利用这些博弈理论, 可以帮助多智能体系统实现资源分配的优化、协作策略以及竞争策略的制定, 从而提高系统的效率和控制性能.

1.1.1 零和博弈

基于强化学习控制的多智能体零和博弈模型是控制理论中的一个重要研究领域, 其涉及多个智能体之间如何通过博弈策略来最大化自身利益. 在这一模型中, 每个智能体都追求最大化自身的收益, 但智能体之间的决策相互影响, 形成一个总收益为零的局面, 即一方的收益等于另一方的损失. 通过强化学习控制方法, 智能体可以根据环境的反馈信息和对手的行为来动态调整自身的策略, 以适应不断变化的博弈局面. 这种学习和适应能力使得智能体能够更好地理解对手的行为模式, 从而制定出更加智能化的决策策略, 提高自身在多智能体零和博弈中的竞争优势.

特别地, 在强化学习控制研究中, 两人零和博弈与鲁棒 H_∞ 控制问题联系密切^[9-16]. H_∞ 控制问题针对系统存在外部干扰情形, 通常可将控制器视为一个最小化某一特定性能指标的博弈者, 而将未知的外部干扰或内部的不确定性看作一个最大化性能指标的另一博弈者. 此外, 根据参与博弈的玩家数量, 零和博弈可分为两人零和博弈和多人零和博弈.

两人零和博弈捕捉了两个玩家的行为, 两个玩家之间存在直接的竞争关系, 即一方的收益增加意味着另一方的收益减少. 典型的例子包括常见的博弈游戏, 如囚徒困境、石头-剪刀-布等. 对于连续时间线性系统, 双人零和博弈的求解依赖于广义博弈

代数黎卡提方程的求解. 例如, 在文献 [13] 中, 线性动态可由下列微分方程描述:

$$\begin{cases} \dot{x} = Ax + Bu + Dd \\ y = Cx \end{cases} \quad (1)$$

其中, x 为系统状态; u 与 d 代表两个控制输入 (也可看作两个玩家); y 为输出; A, B, C 和 D 为常数矩阵. 定义系统性能指标为

$$\mathcal{J}(x(0), u, d) = \int_0^\infty r(x, u, d) d\tau \quad (2)$$

其中, $r(x, u, d) = Q(x) + u^T R u - \gamma^2 \|d\|^2$, $Q(x) \geq 0$, $x(0)$ 表示系统初始时刻的状态值, R 为对称的正定矩阵, 即 $R = R^T \geq 0$, $\gamma \geq \gamma^* \geq 0$, γ^* 代表使系统稳定的 γ 的最小值.

根据上述性能指标 (2), 系统 (1) 的最优控制问题等价于最小化最优值函数 V^* , 即

$$V^*(x(0)) = \min_u \max_d \mathcal{J}(x(0), u, d)$$

并且针对最小化控制策略 u 与最大化控制策略 d , 使得性能指标函数 \mathcal{J} 满足纳什均衡, 即

$$\mathcal{J}^*(u^*, d) \leq \mathcal{J}^*(u^*, d^*) \leq \mathcal{J}^*(u, d^*)$$

式中, u^* 和 d^* 代表最优控制输入.

通过引入线性最优控制理论, 最优反馈控制输入可以表示为

$$u^*(x) = -\eta^{-1} B^T S^* x = -K^* x$$

$$d^*(x) = \gamma^{-1} D^T S^* x = L^* x$$

其中, $\eta > 0$, $K^* = \eta^{-1} B^T S^*$, $L^* = \gamma^{-1} D^T S^*$, S^* 为下列黎卡提方程的解

$$\begin{aligned} A^T S^* + S^* A + Q - \eta^{-1} S^* B B^T S^* + \\ \gamma^{-1} S^* D D^T S^* = 0 \end{aligned} \quad (3)$$

对于非线性系统, 黎卡提方程将变成哈密顿-雅可比-埃萨克斯 (Hamilton-Jacobi-Isaacs, HJI) 方程^[11, 14-15]. 例如, 对于一类连续时间仿射非线性系统, Vamvoudakis 等^[15] 提出一种在线自适应动态规划法, 求解已知连续动态非线性系统的双人零和博弈问题. 其系统动态描述为如下形式:

$$\dot{x} = f(x) + g(x)u(x) + k(x)d(x) \quad (4)$$

其中, x 为状态, $f(x)$, $g(x)$, $k(x)$ 为非线性系统动态, $u(x)$ 与 $d(x)$ 表示两种控制输入或其中一种控制输入为外部扰动. 为实现上述系统的最优控制, 作者采用如下性能指标:

$$\mathcal{J}(x(0), u, d) = \int_0^\infty r(x, u, d) dt \quad (5)$$

因此, 值函数可以定义为如下形式:

$$V(x(t), u, d) = \int_t^\infty (Q(x) + u^T R u - \gamma^2 \|d\|^2) d\tau$$

根据所定义的性能指标 (5), 实现系统 (4) 的最优控制等价于最小化最优值函数 V^* , 即

$$V^*(x(0)) = \min_u \max_d \mathcal{J}(x(0), u, d) =$$

$$\min_u \max_d \int_0^\infty (Q(x) + u^T R u - \gamma^2 \|d\|^2) d\tau$$

并且针对输入控制对 (u, d) , 使得性能指标函数 \mathcal{J} 满足纳什均衡.

通过运用最优性原理, 可得最优控制输入 $u^*(x)$ 与 $d^*(x)$

$$u^*(x) = -\frac{1}{2} R^{-1} g^T(x) \nabla V^* \quad (6)$$

$$d^*(x) = \frac{1}{2\gamma^2} k^T(x) \nabla V^* \quad (7)$$

其中, $\nabla V^* = \frac{\partial V^*}{\partial x}$, $V^*(x)$ 为如下 HJI 方程^[17] 的解:

$$\begin{aligned} H(x, \nabla V^{*\top}(x), u^*, d^*) &= Q(x) + \nabla V^{*\top}(x) f(x) - \\ &\frac{1}{4} \nabla V^{*\top}(x) g(x) R^{-1} g^T(x) \nabla V^*(x) + \\ &\frac{1}{4\gamma^2} \nabla V^{*\top}(x) k(x) k^T(x) \nabla V^*(x) = 0 \end{aligned} \quad (8)$$

可见, 上述 HJI 方程为非线性多维偏微分方程, 利用常规的方法求解非常困难. 于是, 基于强化学习的方法应运而生. 下面介绍两类典型求解 HJI 方程的算法: 策略迭代^[18] 和值迭代^[10].

在策略迭代算法中, 采用如下策略评价机制计算值函数 $V^{(i)}(x)$:

$$\begin{aligned} r(x, u^{(i)}, d^{(i)}) + [\nabla V^{(i)}(x)]^T [f(x) + \\ g(x)u^{(i)} + k(x)d^{(i)}] = 0 \end{aligned} \quad (9)$$

然后, 基于 $V^{(i)}(x)$, 使用如下同步方法进行策略改进:

$$u^{(i+1)}(x) = -\frac{1}{2} R^{-1} g^T(x) \nabla V^{(i)}(x) \quad (10)$$

$$d^{(i+1)}(x) = \frac{1}{2\gamma^2} k^T(x) \nabla V^{(i)}(x) \quad (11)$$

理论上, 上述策略迭代算法 (9) ~ (11) 等价于牛顿迭代法^[18], 进而利用 Kantorovitch 定理证明收敛性. 在算法中, 将非线性 HJI 方程转化为一系列线性方程 (9) 进行求解, 求解方程变得更加简便可行, 但在执行策略评价时依赖于系统动态模型 $f(x)$, $g(x)$, $k(x)$, 为了减小对系统模型的依赖, 借助 Vrabie 等^[10] 提出的积分强化学习机制, Wu 等^[9] 提出采用如下策略评价方法:

$$\begin{aligned} V^{(i)}(x(t)) &= \int_t^{t+T} r(x, u^{(i)}, d^{(i)}) d\tau + \\ &V^{(i)}(x(t+T)) \end{aligned} \quad (12)$$

由式 (12) 可知, 策略评价不需要系统动态模型, 而是通过采集数据 $x(t)$ 和 $x(t+T)$ 隐式地包含了系统动态演化过程. 基于式 (12), 进一步给出了脱策 (Off-policy) 学习方法^[11], 可以使用任意行为控制策略产生数据用于策略评价, 从而有效克服不充分激励问题, 提高数据利用率. 在值迭代算法^[10] 中, 采用如下方式进行值函数更新:

$$\begin{aligned} V^{(i+1)}(x(t)) &= \int_t^{t+T} r(x, u^{(i)}, d^{(i)}) d\tau + \\ &V^{(i)}(x(t+T)) \end{aligned} \quad (13)$$

值迭代算法的优点在于对初始控制策略要求低, 容易实现, 不足在于收敛速度较策略迭代慢.

此外, 还有一种基于 Actor-Critic 结构^[15] 的在线强化学习方法可求得上述 HJI 方程的解. 其核心思想是利用神经网络近似值函数

$$V^*(x) = W^{*\top} \varphi(x) + \epsilon(x)$$

其中, W^* 为神经网络的理想权值, $\varphi(x)$ 表示激活函数, $\epsilon(x)$ 为近似误差. 然而, 神经网络的理想权值难以直接获取, 往往采用其估计权值 \hat{W}_c 进行估计, 即

$$\hat{V}(x) = \hat{W}_c^\top \varphi(x) \quad (14)$$

从而最优控制输入 (6) 和 (7) 可改写为

$$\hat{u}(x) = -\frac{1}{2} R^{-1} g^T(x) \nabla \varphi \hat{W}_c \quad (15)$$

$$\hat{d}(x) = \frac{1}{2\gamma^2} k^T(x) \nabla \varphi \hat{W}_c \quad (16)$$

其中, $\nabla \varphi = \frac{\partial \varphi(x)}{\partial x}$. 由此, $\hat{V}(x)$ 可认为是 Critic 网络, $\hat{u}(x)$ 和 $\hat{d}(x)$ 则为 Actor 网络. 在式 (14) ~ (16) 中, Critic 网络和 Actor 网络共享神经网络权值 \hat{W}_c , 这意味着只需要更新 \hat{W}_c , Critic 网络和 Actor 网络将会同步进行更新, 但根据式 (15) 和式 (16) 可知, 这种机制需要部分系统动态模型信息, 即: $g(x)$ 和 $k(x)$. 当然, 在 $g(x)$ 和 $k(x)$ 未知的情况下, 可以增加两个新的 Actor 网络独立近似 $u^*(x)$ 和 $d^*(x)$, 这样就可以得到完全无模型的强化学习控制方法^[12]. 对于多人零和博弈^[19-22], 意味着有三个或更多的参与者参与博弈, 每个参与者的利益与损失之和仍然为零, 并且每个参与者的行为将影响其他参与者的收益情况. 对于多智能体系统的零和博弈问题而言, 文献 [23] 中研究了受外部扰动的单领航者多智能体线性系统的分布式最优跟踪控制问题, 利用微分博弈的

概念将分布式控制问题转化为多人零和微分图形博弈问题. 具体考虑由下列微分方程描述的多智能体系统:

$$\begin{cases} \dot{x}_i = f_i(x_i) + g_i(x_i)u_i(x_i) + k_i(x_i)d_i(x_i) \\ \dot{x}_0 = f_0(x_0) \end{cases} \quad (17)$$

其中, x_i 表示第 i 个跟随者的系统状态, x_0 为领航者的系统状态, $f_i(x_i)$, $g_i(x_i)$, $k_i(x_i)$ 代表第 i 个非线性系统动态, u_i 和 d_i 分别为控制和外部扰动输入, x_0 表示领航者的系统状态. 其领航-跟随一致性误差可以表示为

$$e_i = \sum_{j \in P_i} a_{ij}(x_i - x_j) + a_{i0}(x_i - x_0) \quad (18)$$

其中, x_j 代表第 j 个跟随者智能体的状态.

为了实现对系统 (17) 的最优跟踪鲁棒控制, 针对智能体 i 设计如下局部性能指标函数:

$$\begin{aligned} \mathcal{J}_i(e_i(0), u_i, u_{-i}, d_i, d_{-i}) = & \\ & \frac{1}{2} \int_0^\infty (e_i^\top Q_{ii} e_i + u_i^\top R_{ii} u_i + \sum_{j \in P_i} u_j^\top R_{ij} u_j - \\ & \gamma^2 d_i^\top T_{ii} d_i - \gamma^2 \sum_{j \in P_i} d_j^\top T_{ij} d_j) dt \end{aligned} \quad (19)$$

其中, u_{-i} 为智能体 i 所有的邻居控制器集合, d_{-i} 为智能体 i 所有的邻居扰动输入集合, $Q_{ii} > 0$, $R_{ii} > 0$, $R_{ij} \geq 0$, $T_{ii} > 0$, $T_{ij} \geq 0$. 进一步, 将值函数定义为

$$\begin{aligned} V_i(e_i(0), u_i, u_{-i}, d_i, d_{-i}) = & \\ & \frac{1}{2} \int_t^\infty (e_i^\top Q_{ii} e_i + u_i^\top R_{ii} u_i + \sum_{j \in P_i} u_j^\top R_{ij} u_j - \\ & \gamma^2 d_i^\top T_{ii} d_i - \gamma^2 \sum_{j \in P_i} d_j^\top T_{ij} d_j) dt \end{aligned} \quad (20)$$

考虑系统 (17), 设计合适的鲁棒控制策略, 使得系统实现下列控制目标:

1) 当 $d_i(t) = 0$ 时, 通过设计 u_i , 使得对于任意的智能体 i , 有 $\|x_i(t) - x_0(t)\| \rightarrow 0$;

2) 当 $d_i(t) \neq 0$, 满足下列有界 L_2 增益同步条件:

$$\begin{aligned} \int_0^T \|z_i(t)\|^2 dt = \int_0^T (e_i^\top Q_{ii} e_i + u_i^\top R_{ii} u_i + \\ \sum_{j \in P_i} u_j^\top R_{ij} u_j) dt \leq \\ \gamma^2 \int_0^T (d_i^\top T_{ii} d_i + \sum_{j \in P_i} d_j^\top T_{ij} d_j) dt + \beta(e_i(0)) \end{aligned}$$

其中, $z_i(t) = [e_i \ u_i \ u_{-i}]$, 有界函数 β 满足 $\beta(0) = 0$.

进一步, 上述多智能体系统的最优鲁棒跟踪控制问题可转化为如下多人零和微分博弈问题, 即

$$V_i(e_i(0)) = \min_{u_i} \max_{d_i} \mathcal{J}_i(e_i(0), u_i, u_{-i}, d_i, d_{-i})$$

基于值函数 (20), 借助 Leibniz 公式^[23] 和最优性原理可得出最优控制输入 $u^*(x)$ 以及最优扰动输入 $d^*(x)$. 进而, 多人零和微分博弈问题将转化为求解一组耦合 HJI 方程, 具体的求解过程可见文献 [20].

Jiao 等^[23] 提出一种在线的策略迭代方法求解由多智能体 H_∞ 控制问题转换的耦合 HJI 方程, 实现多智能体系统的领航-跟随一致性. Chen 等^[24] 将输出反馈同步问题通过鲁棒输出调节和强化学习进行建模, 通过零和博弈描述智能体之间的相互作用, 提出一种输出反馈策略学习算法, 利用输入-输出系统数据来实现异构多智能体系统的分布式鲁棒最优同步. 而对于受外部干扰的离散时间多智能体系统, Zhang 等^[25] 考虑了自身和局部邻居扰动信息的影响, 将优化问题转化为具有控制策略和干扰策略的零和博弈问题, 并且提出一种基于策略梯度的数据驱动迭代算法用以求解 HJI 方程. An 等^[26] 针对一类离散分数阶多智能体系统, 分别提出针对状态反馈和输出反馈的零和博弈 Q 学习算法. Ma 等^[27] 对存在执行器故障的二阶多智能体展开研究, 基于零和微分博弈方法提出一种容错控制策略, 从而保障系统的稳定性以及性能的最优性. 考虑到多智能体系统遭受服务器拒绝攻击问题, Wu 等^[28] 基于多人混合零和博弈策略, 构造一种基于神经网络的强化学习方案, 得到了多人混合零和博弈方案的纳什均衡解, 提出了具有记忆事件触发机制的协同自适应控制方法.

1.1.2 非零和博弈

在非零和博弈中, 多智能体之间的利益并不完全对立, 而是存在一定程度的合作. 基于强化学习控制的多智能体非零和博弈模型关注于如何通过协作来实现共赢, 最大化各个智能体的收益. 在这种模型中, 智能体之间可能会形成稳定的合作关系, 共同制定出有效的策略以应对环境的变化. 通过强化学习控制方法, 智能体可以学习并优化合作策略, 以最大化整体系统的收益.

在具有多个控制输入的连续时间非线性系统的最优控制问题中, 可以将系统控制器之间的相互作用看作是一个非零和博弈^[29-31], 其中每个控制器都追求自身的最大化利益或自身的最小化代价, 这意味着系统的总收益不一定为零. 以下列非线性仿射系统为例, 探究如何利用强化学习的方法求解具有多输入系统的最优控制问题:

$$\dot{x} = f(x) + \sum_{i=1}^N g_i(x)u_i(x) \quad (21)$$

其中, x 表示系统状态, u_i 表示第 i 个控制输入, $f(x)$ 和 $g_i(x)$ 为非线性函数, N 代表控制输入的个数.

为了实现系统 (21) 的最优控制问题, 针对第 i 个控制输入, 设计以下的性能指标函数

$$\mathcal{J}_i(x_0, u_1, \dots, u_N) = \int_0^\infty (x^T Q_i x + \sum_{i=1}^N u_i^T R_{ii} u_i) d\tau$$

此外, 将值函数 $V_i(x(t))$ 设计为

$$V_i(x) = \int_t^\infty (x^T Q x + \sum_{i=1}^N u_i^T R_{ii} u_i) d\tau \quad (22)$$

其中, $R_{ii} > 0$.

考虑系统 (21), 通过为 N 个控制器设计最优控制策略, 使得

$$V_i(x) = V_i^*(x) =$$

$$\min_{u_1, \dots, u_N} \left\{ \int_t^\infty (x^T Q_i x + \sum_{i=1}^N u_i^T R_{ii} u_i) d\tau \right\}$$

其中, $V_i^*(x)$ 为最优值函数. 同时使性能指标函数 \mathcal{J}_i 达到纳什均衡, 即

$$\mathcal{J}_i(u_1^*, \dots, u_i^*, \dots, u_N^*) \leq \mathcal{J}_i(u_1^*, \dots, u_i, \dots, u_N^*) \quad (23)$$

其中, $(u_1^*, \dots, u_i^*, \dots, u_N^*)$ 为最优控制输入.

基于值函数 (22), 通过借助最优性原理, 可得最优控制输入的表达式为

$$u_i^* = -\frac{1}{2} R_{ii}^{-1} g_i^T(x) \nabla V_i^*(x)$$

其中, $\nabla V_i^* = \frac{\partial V_i^*}{\partial x}$, $V_i^*(x)$ 为如下 HJI 方程的解:

$$\begin{aligned} \mathcal{H}_i(x, u_1^*, \dots, u_N^*, \nabla V_i^*) = \\ \nabla V_j^{*T} \left(f(x) - \frac{1}{2} \sum_{j=1}^N g_j(x) R_{jj}^{-1} g_j^T(x) \nabla V_j^{*T} \right) + \\ \frac{1}{4} \sum_{j=1}^N \nabla V_j^{*T} g_j(x) R_{jj}^{-T} R_{jj} R_{jj}^{-1} g_j^T(x) \nabla V_j + \\ x^T Q_i x = 0 \end{aligned} \quad (24)$$

于是, 带有 N 个输入的非线性最优控制问题由非零和博弈问题的求解转化为求解上述耦合的 HJI 方程, 具体的求解过程可以参考文献 [32-35].

然而, 在多智能体系统中, 往往将每个智能体当作玩家, 通过针对每个智能体设计最优控制策略来求解其非零和博弈问题. 以领航-跟随多智能体一致性为例, 考虑具有以下系统动态的多智能体系统:

$$\begin{cases} \dot{x}_i = f_i(x_i) + g_i(x_i) u_i(x_i) \\ \dot{x}_0 = f_0(x_0) \end{cases} \quad (25)$$

其中, 各变量定义同系统 (17), 误差定义同式 (18).

为了实现系统 (25) 最优一致性控制, 针对智能体 i 设计如下的性能指标函数:

$$\mathcal{J}_i(e_i(0), u_i, u_{-i}) = \frac{1}{2} \int_0^\infty (e_i^T Q_{ii} e_i + u_i^T R_{ii} u_i + \sum_{j \in P_i} u_j^T R_{ij} u_j) d\tau \quad (26)$$

进一步, 将其值函数定义为

$$V_i(e_i(0), u_i, u_{-i}) = \frac{1}{2} \int_t^\infty (e_i^T Q_{ii} e_i + u_i^T R_{ii} u_i + \sum_{j \in P_i} u_j^T R_{ij} u_j) d\tau \quad (27)$$

由上可知, 多智能体的零和博弈与非零和博弈不同之处在于, 非零和博弈只需要考虑智能体之间的合作关系, 无需与扰动输入互相竞争.

考虑系统 (25), 设计合适的控制策略, 使系统实现对于任意的智能体 i , 有 $\|x_i(t) - x_0(t)\| \rightarrow 0$, 同时使系统性能指标最小化.

进一步, 上述多智能体系统的最优一致性控制问题可转化为如下多人非零和微分博弈问题, 即

$$V_i(e_i(0)) = \min_{u_i} \mathcal{J}_i(e_i(0), u_i, u_{-i})$$

基于值函数 (27), 借助 Leibniz 公式和最优性原理可得出最优控制输入 $u^*(x)$. 进而, 多人零和微分博弈问题可以转化为求解一组耦合 HJI 方程.

然而不同于连续系统, 对于离散形式下多智能体的最优一致性控制^[36], 将其性能指标定义为

$$\mathcal{J}_i(e_i(k), u_i(k), u_{-i}(k)) = \sum_{l=0}^{\infty} U_i(e_i(l), u_i(l), u_{-i}(l))$$

其中, $U_i(e_i(l), u_i(l), u_{-i}(l)) = e_i^T Q_i e_i(l) + u_i^T(l) \times R_{ii} u_i(l) + \sum_{j \in P_i} u_i^T(l) R_{ij} u_j(l)$ 表示成本函数. 这代表着从当前时刻到无穷时刻区间, 每个时刻下成本函数的累积. 固定智能体 i 与其邻居的控制策略对 (u_i, u_{-i}) , 定义智能体 i 的值函数为

$$V_i(e_i(k)) = \sum_{l=k}^{\infty} U_i(e_i(l), u_i(e_i(l)), u_{-i}(e_{-i}(l)))$$

进一步, 上述值函数可以等价写成离散形式的贝尔曼方程

$$\begin{aligned} V_i(e_i(k)) = U_i(e_i(k), u_i(e_i(k)), \\ u_{-i}(e_{-i}(k))) + V_i(e_i(k+1)) \end{aligned}$$

其中, $e_{-i}(k) = \{e_j(k) : j \in P_i\}$ 表示智能体 i 邻居的

跟踪误差. 根据上述形式, 可得智能体 i 的哈密顿方程为

$$\mathcal{H}_i(e_i(k), u_i(e_i(k)), V_i) = V_i(e_i(k+1)) - V_i(e_i(k)) + U_i(e_i(k), u_i(e_i(k)), u_{-i}(e_{-i}(k)))$$

根据贝尔曼最优性原理, 智能体 i 的最优值函数满足下列耦合的哈密顿-雅可比-贝尔曼 (Hamilton-Jacobi-Bellman, HJB) 方程:

$$V_i^*(e_i(k)) = \min_{u_i(k)} [U_i(e_i(k), u_i(e_i(k)), u_{-i}^*(e_{-i}(k))) + V_i^*(e_i(k+1))] \quad (28)$$

于是, 多智能体的最优分布式一致性问题就由多人非零和博弈问题转化为求解上述 HJB 方程, 具体的解决方法可参考文献 [36-37].

此外, Vamvoudakis 等^[38] 对于具有连续动态的多智能体非零和图博弈问题, 提出仅依赖于局部信息的在线策略迭代方法, 并证明了策略迭代方法的收敛性. 对于多智能体非零和微分博弈的分类同步问题, Yang 等^[39] 同样将其转化为求解 HJB 方程问题, 提出仅使用系统数据的策略迭代算法, 从而保证了性能指标的局部最优性. Odekunle 等^[40] 针对多人非零和博弈的输出调节问题, 利用强化学习算法近似了调节方程以及耦合代数黎卡提方程的解. Wang 等^[41] 将具有外部干扰的多智能体分布式一致性控制问题建模为多重博弈问题, 其将分布式一致性问题描述为多个单智能体间的非零和博弈, 其中每个智能体被视为一个玩家, 专注于优化自身的局部性能指标函数使得整个系统实现纳什均衡; 另外, 对于每个智能体自身的鲁棒控制问题可视为控制器和外部扰动的双人零和博弈. 对于部分未知网络化系统的非零和博弈问题, Su 等^[42] 设计了一种基于模糊逻辑模型的辨识器, 重构了未知的系统动态. 此外, 将一种单 Critic 网络设计用于近似值函数和控制策略, 在求解出近似的纳什均衡解的同时, 减少了系统的计算负担.

1.1.3 Stackelberg 博弈

上述零和博弈与非零和博弈均可归类为纳什博弈类型, 主要解决同步决策问题, 即所有玩家在博弈游戏中同时做出决策. 而对于顺序决策问题, 则可建模为 Stackelberg-纳什博弈, 其在智能电网^[43]、能源感知分配^[44]、电车充电^[45] 等领域均有广泛应用. 在 Stackelberg 博弈中, 通常包括两类主要参与者, 先做出决策的领航者和根据领航者决策做出最优响应的跟随者. 多智能体 Stackelberg 博弈模型是多智能体系统中的重要研究领域, 旨在探究如何通过领航者的指导来影响跟随者的决策, 从而最大化整

体系统的收益. 通过强化学习控制方法, 领航者可以学习并优化其决策策略, 以达到最大化整体系统收益的目的. 与此同时, 跟随者也需要根据领航者的指示调整自身行为, 以实现整体系统的协调与一致. 基于强化学习控制的多智能体 Stackelberg 博弈模型旨在研究领航者和跟随者之间的合作与竞争关系, 并通过博弈策略实现整体系统性能的最优化.

首先, 本节针对一类多输入非线性系统的 Stackelberg 博弈模型^[46] 展开讨论, 其具有下列的系统动态:

$$\dot{x} = f(x) + g_0(x)u_0(t) + \sum_{i=1}^N g_i(x)u_i(t) \quad (29)$$

其中, x 为系统动态, $u_0 \in \mathcal{U}_0$ 表示领航者控制输入, $u_i \in \mathcal{U}_i$ 表示跟随者控制输入, $i \in \mathcal{F} = \{1, \dots, N\}$ 为跟随者集合, $f(x)$ 代表非线性函数, $g_0(x)$ 和 $g_i(x)$ 为控制输入矩阵.

为了实现系统的最优控制, 针对每个控制输入设计其性能指标为

$$\mathcal{J}_j(x(0), u_0, u) = \int_0^\infty r_j(x, u_0, u) d\tau$$

其中, $j \in \{0, 1, \dots, N\}$, $u = u(u_0) = \{u_i(u_0) | i \in \mathcal{F}\}$, $u_i(u_0)$ 表示基于领航者控制策略 u_0 的第 i 个跟随者做出的控制策略, $r_0(x, u_0, u) = \|x\|_{Q_0}^2 + \|u_0 + \sum_{i=1}^N C_i u_i\|_{R_0}^2$, $r_i(x, u_0, u) = \|x\|_{Q_i}^2 + \|u_i + D_i u_i\|_{R_i}^2$, C_i 表示跟随者 i 与领航者的耦合系数, D_i 表示领航者与跟随者 i 的耦合系数, $Q_0 = Q_0^T > 0$, $Q_i = Q_i^T > 0$, $R_0 = R_0^T > 0$, $R_i = R_i^T > 0$. 此外, 将每个控制输入的值函数定义为

$$V_j(x) = V_j(x, u_0, u) = \int_t^\infty r_j(x, u_0, u) d\tau \quad (30)$$

考虑系统 (29), 在保证系统稳定的情况下, 设计每个参与者的控制策略使得性能指标最小化, 且使得性能指标函数 \mathcal{J}_j 实现 Stackelberg 纳什均衡, 即如果存在一个从领导者策略空间到跟随者最优响应策略空间的映射 $\mathcal{P}_i: \mathcal{U}_0 \rightarrow \mathcal{U}_i$, $i \in \mathcal{F}$, 且其他所有跟随者选择了控制策略 $\mathcal{P}_{-i}(u_0) = \{\mathcal{P}_\iota | \iota \in \mathcal{F}, \iota \neq i\}$, 使得对于任意固定的 $u_0 \in \mathcal{U}_0$, 则下列关系成立:

$$\mathcal{J}_i(x(0), u_0, \mathcal{P}_i(u_0), \mathcal{P}_{-i}(u_0)) \leq \mathcal{J}_i(x(0), u_0, u_i(u_0), \mathcal{P}_{-i}(u_0))$$

并且如果存在策略 $\bar{u}_0 \in \mathcal{U}_0$ 满足

$$\mathcal{J}_0(x_0, \bar{u}_0, \mathcal{P}_i(\bar{u}_0), \mathcal{P}_{-i}(\bar{u}_0)) \leq \mathcal{J}_0(x_0, u_0, \mathcal{P}_i(u_0), \mathcal{P}_{-i}(u_0))$$

那么, 策略集 $\{\bar{u}_0, \bar{u}_1, \dots, \bar{u}_N\} \in \mathcal{U}_0 \times \mathcal{U}_1 \times \dots \times \mathcal{U}_N$ 即为 Stackelberg 均衡策略集, 其中, $\bar{u}_i = \mathcal{P}_i(\bar{u}_0)$.

通过使用值函数 (30) 对状态 x 求偏导, 可得系统哈密顿方程

$$\mathcal{H}_j(x, \nabla V_j(x), u_0, u) = r_j(x, u_0, u) + \nabla V_j^T(x) \left(f(x) + \sum_{\kappa=0}^N g_\kappa(x) u_\kappa \right) \quad (31)$$

其中, $\nabla V_j = \frac{\partial V_j}{\partial x}$. 根据最优值函数以及一阶最优性条件, 可得领航者与跟随者的最优控制输入

$$u_0^* = \frac{1}{2} M^{-1} \left(\sum_{i=1}^N g_i(x) D_i - g_0(x) \right)^T \nabla V_0^*(x) + \frac{1}{2} P^{-1} \sum_{i=1}^N C_i R_i^{-1} g_i^T(x) \nabla V_i^*(x) \quad (32)$$

$$u_i^*(u_0^*) = -D_i u_0^* - \frac{1}{2} R_i^{-1} g_i^T(x) \nabla V_i^*(x) \quad (33)$$

其中, $P = I_m - \sum_{i=1}^N C_i D_i$, $M = P^T R_0 P$. 将式 (32) 和式 (33) 代入式 (31), 可得下列 HJ 方程:

$$\mathcal{H}_j(x, \nabla V_j^*(x), u_0^*, u^*) = r_j(x, u_0^*, u^*) + \nabla V_j^{*T}(x) \left(f(x) + \sum_{\kappa=0}^N g_\kappa(x) u_\kappa^* \right) = 0 \quad (34)$$

因此, 关于具有多输入非线性系统的 Stackelberg 博弈问题就转化为解上述耦合的哈密顿-雅可比 (Hamilton-Jacobi, HJ) 方程问题, 详细的解决方法可参考文献 [46].

考虑具有有向通信拓扑图结构的无领航-跟随多智能体系统^[47], 其所有智能体的系统动态可由下列线性微分方程表示:

$$\dot{x}_i = A x_i + B_i u_i \quad (35)$$

其中, A 与 B_i 表示常数矩阵, x_i 表示智能体 i 的系统状态, u_i 表示智能体 i 的系统输入. 假设智能体 1 为主要智能体, 其余 $N-1$ 个为次要智能体. 主要智能体优先做出决策占主导地位; 次要智能体之后再做出决策. 此外, 值得注意的是, 主要智能体需要具备预测次要智能体可能响应的能力, 而只有一部分次要智能体能够观察到主要智能体的策略.

定义对于智能体 i 的无领航-跟随一致性误差及其一阶时间导数为

$$e_i = \sum_{j \in P_i} a_{ij} (x_i - x_j) \quad (36)$$

$$\dot{e}_i = A e_i + d_i B_i u_i - \sum_{j \in P_i} a_{ij} B_j u_j \quad (37)$$

接下来, 根据智能体在决策过程中的地位针对性地设计主要智能体性能指标为

$$J_1(e_1(0), u_1, u_{-1}) = \int_0^\infty r_1(e_1, u_1, u_{-1}) d\tau$$

$r_1(e_1, u_1, u_{-1}) = \|e_1\|_{Q_1}^2 + \|u_1 + \sum_{\kappa=2}^N b_\kappa u_\kappa\|_{R_1}^2$. 因为次要智能体能够观察到主要智能体的行为, 所以将次要智能体性能指标设计为

$$\mathcal{J}_i(e_i, u_i, u_{-i}) = \int_0^\infty r_i(e_i, u_i, u_{-i}) d\tau$$

其中, $i \in \mathcal{Y} = \{2, \dots, N\}$, $r_i(e_i, u_i, u_{-i}) = \|e_i\|_{Q_i}^2 + \|u_i + \beta_i u_{-i}\|_{R_i}^2$, $\beta_i \geq 0$, 如果第 i 个次要智能体能够观察到主要智能体的策略, $\beta_i > 0$; 否则 $\beta_i = 0$. 这意味着当主要智能体 1 为智能体 i 的邻居智能体, 即 $1 \in P_i$, 相应地, 主要智能体与跟随智能体对应的值函数为

$$V_1(x) = V_1(x, u_1, u_{-1}) = \int_t^\infty r_1(x, u_1, u_{-1}) d\tau \quad (38)$$

$$V_i(x) = V_i(x, u_i, u_{-i}) = \int_t^\infty r_i(x, u_i, u_{-i}) d\tau \quad (39)$$

考虑系统 (35), 通过设计合适的控制策略 $\{u_1^*, u_2^*, \dots, u_N^*\}$ 使得所有智能体的状态达到一致, 即 $\lim_{t \rightarrow \infty} \|x_i(t) - x_j(t)\| = 0$, 同时使性能指标达到 Stackelberg 纳什均衡.

基于式 (38) 和式 (39), 可得各智能体的哈密顿函数为

$$\mathcal{H}_i(e_i, \nabla V_i^T, u_i, u_{-i}) = r_i(e_i, u_i, u_{-i}) + \nabla V_i^T (A e_i + d_i B_i u_i - \sum_{j \in P_i} a_{ij} B_j u_j) \quad (40)$$

其中, $\nabla V_i = \frac{\partial V_i}{\partial e_i}$. 类似于式 (32) 和式 (33), 可得主要、次要智能体的最优策略为

$$u_1^* = Z_1 \sum_{j=2}^N \alpha_j d_j R_j^{-1} B_j^T \nabla V_j^* - Z_2 R_1^{-1} B_1^T \nabla V_1^* \quad (41)$$

$$u_i^*(u_1^*) = -\beta_i u_1^* - \frac{1}{2} d_i R_i^{-1} B_i^T \nabla V_i^*(x) \quad (42)$$

其中, $Z_1 = \frac{1}{2(1 - \sum_{j=2}^N \alpha_j \beta_j)}$, $Z_2 = \frac{d_1 + \sum_{j \in P_1} e_{1j} \beta_j}{2(1 - \sum_{j=2}^N \alpha_j \beta_j)^2}$, 通过选择合适的 α_j , 可以确保 $1 - \sum_{j=2}^N \alpha_j \beta_j \neq 0$. 将式 (41) 和式 (42) 代入式 (40), 可得下列方程:

$$\mathcal{H}_i(e_i, \nabla V_i^{*T}, u_i^*, u_{-i}^*) = r_i(e_i, u_i^*, u_{-i}^*) + \nabla V_i^{*T} \left(A e_i + d_i B_i u_i^* - \sum_{j \in P_i} a_{ij} B_j u_j^* \right) = 0 \quad (43)$$

于是, 顺序决策下的多智能体系统分层同步问题就

转换为求解上述方程. Lin 等^[46]针对一类多人 Stackelberg 博弈问题提出一种基于事件触发的鲁棒自适应动态规划算法, 利用在线策略迭代方法得到耦合 HJ 方程的解. 此外, Li 等^[47]提出一种基于积分强化学习的两层值迭代算法, 在不依赖于系统动态的情况下解决线性多智能体系统的最优同步问题. 通过为系统状态观测器构建辅助输入, Yan 等^[48]将原系统与观测器输入的先后作用关系建模为两人 Stackelberg 博弈, 利用积分强化学习与自适应评判学习算法达到 Stackelberg 纳什均衡的同时, 智能体间的一致性误差也实现了渐进收敛.

1.2 多智能体协同

不同于多智能体博弈, 多智能体协同控制则不需要考虑智能体间的竞争行为, 其通过多个智能体互相协调与合作, 共同完成某个复杂任务或目标. 强化学习可以为智能体提供自主学习和自适应调节的能力, 使得协同控制策略能够在动态和复杂的环境中不断优化. 通过将强化学习与协同控制相结合, 多智能体系统可以在不确定和不断变化的环境中实现高效合作. 强化学习算法使得智能体能够在不断试错的过程中学习最佳的协同策略, 并通过相互作用实现系统的全局优化. 这一融合特点显著扩展了多智能体协同控制的应用领域, 如集群机器人编队、智能交通管理和分布式传感器网络等, 使得这些领域中的智能体能够通过不断学习和调整, 更好地完成任务并提升整体系统的性能.

多智能体强化学习协同控制的研究面临着多样化的挑战和应用需求. 本节立足于不同协同任务、不同系统动态、不同系统性能要求和系统模型依赖程度等四个因素, 对多智能体强化学习控制的研究成果进行归纳, 其整体框架如图 2 所示.



图 2 多智能体强化学习协同控制

Fig. 2 Multi-agent reinforcement learning cooperative control

1.2.1 不同协同任务

典型的协同控制任务包括领航-跟随一致性控制、编队控制、包含控制等, 其在实际系统中有着广泛的应用. 例如, 在地球观测或天文探测任务中, 多

个卫星需要保持相同的姿态, 以确保观测设备能够同步对准目标. 而在集群机器人编队中, 多个机器人需要协同合作完成复杂的任务, 如搜索救援或环境监测, 这要求智能体之间能够相互协调, 保持编队形态以应对不同场景. 因此, 针对不同协同任务的特点和需求, 多智能体强化学习控制需要设计相应的算法和策略, 以实现智能体之间的有效协同.

多智能体系统中的领航-跟随一致性控制是一种常见的协同控制策略, 用于指导多个智能体以一致的方式行动. 在这种控制策略下, 领航者通常负责制定整体行动方向和目标, 跟随者则调整自身行为以与领航者保持一致, 以实现整个系统的协调运行. Li 等^[49]针对一类状态不可观测的多智能体系统提出一种基于强化学习方法的分布式输出反馈最优控制策略, 仅利用各智能体的输出顺利地实现了领航-跟随一致性. 而针对系统存在输入受限与外部干扰的情况, Zhang 等^[50]提出一种基于强化学习的在线策略迭代方法, 利用 Actor-Critic 网络得出分布式最优一致性控制问题的近似解. 对于领航跟随二部一致性问题, Li 等^[51]通过一种新的坐标转换方法将其化为传统的多智能体一致性跟踪问题, 提出基于数据的强化学习算法获得了领航-跟随双边一致性问题最优控制策略. 当系统面临间歇性的状态受限问题时, Luo 等^[52]提出一种新颖的切换函数以及改进型的坐标转换方式, 将受限的状态转换为非受限的表达形式, 在反步控制的框架下得到了近似的最优控制器.

多智能体系统的编队控制是指在多个智能体之间实现一定形态排列或规定动作模式的控制策略. 这种控制策略通常用于集群机器人、自动驾驶车辆、飞行器编队等多智能体系统中, 使智能体之间保持一定的空间关系和运动规律, 以实现协同任务的完成和系统性能的提升. Yu 等^[53]针对一类分布式高阶多智能体系统的时变编队控制问题提出一种自适应最优控制策略, 借助自适应动态规划的方法获得 HJB 方程的近似解, 其中神经网络用于近似值函数. 对于一类时变编队的低阶多智能体系统, 在文献 [54] 中, Lan 等设计了一种基于神经网络状态观测器以及简易强化学习算法的自适应最优控制策略. 对于多智能体编队控制中的通信保留与碰撞避免问题, Wang 等^[55]基于图论知识, 设计了一种虚拟分布式最优控制器, 其中 Actor-Critic 网络应用于强化学习算法的在线实现.

多智能体系统的包含控制是指通过调节系统中各个智能体的行为, 使得整个系统中的一部分或全部智能体都受到一定范围内的“包含”, 从而达到一

致性或稳定性的目标. 这种控制策略的主要目的是在多智能体系统中实现一定的集中性控制, 通过调节系统中智能体之间的相互作用, 达到约束或限制智能体运动的目的. Cheng 等^[56]针对具有多个活跃领导者以及执行器饱和的四旋翼小队, 设计了一种基于历史数据的离线强化学习算法的自适应最优控制策略, 顺利地实现了包含控制任务. 对于一类线性多智能体系统的最优鲁棒包含控制问题, Zuo 等^[57]设计了一种基于模型的离线策略迭代算法. 此外, 由于只有部分智能体才能获取领航者的状态信息, 导致难以针对其余未能直接与领航者存在通讯连接的跟随者设计分布式最优控制器. Wang 等^[58]为每个跟随者设计了一个无模型的分布式自适应观测器, 以取代最优控制器中的领航者状态, 降低多智能体系统分布式控制设计的复杂度.

1.2.2 不同系统动态

在多智能体系统的控制中, 系统动态特性的多样性带来诸多挑战. 根据不同系统的动态特性, 可以从连续时间、离散时间、线性系统和非线性系统的角度进行描述和优化. 多智能体强化学习控制需要具备强适应性和高稳定性, 以应对这些系统的快速变化和不确定性. 例如, 连续时间系统的动态特性通常表现为平滑和连续的变化. 在无人机编队中, 风速和气流等外部环境因素的变化频繁, 要求智能体能够实时调整飞行姿态以保持编队的稳定性. 与此不同, 离散时间系统则在特定时间间隔内更新状态. 例如在智能电网中, 电力负载的变化较为缓慢, 但仍需智能体在每个时间步内灵活调整, 以适应电网运行状态的变化.

对于一类具有输入饱和的线性多智能体系统, Qin 等^[59]提出一种基于数据的离线策略强化学习算法用于学习最优控制策略, 其损失函数中的控制输入项为非二次型形式. 对于使用离散时间刻画的线性多智能体系统, Mu 等^[60]提出一种基于 Q 学习的强化学习算法, 利用系统产生的数据实现一致性控制. 面对更为复杂的非线性系统动态模型, Bai 等^[61]提出一种分布式多梯度递归强化学习策略, 实现了多智能体的一致性控制目标, 其中多梯度递归方法用于神经网络学习率的整定. 当非线性多智能体间各子系统存在动态异构的情况时, Sun 等^[62]建立了一种带有折扣因子的分布式性能指标, 将多智能体的协同控制问题转化为利用强化学习方法求解 HJB 方程的问题. 另外, 在外部环境发生变化时, 原有的通信拓扑结构可能不再适用, 需要进行切换以适应新的环境条件^[63-66]. Qin 等^[67]考虑了线性异构多智能体系统在固定和动态通信拓扑结构下的输出包含

控制问题, 提出一种在线的异策略强化学习算法用以求解带折扣因子的代数黎卡提方程.

1.2.3 不同性能要求

不同系统对性能要求的差异将会影响多智能体强化学习控制的设计. 在一些安全关键型系统中, 系统性能需要具备高度鲁棒性和安全性. 因为随着工业系统变得愈加复杂化和大规模化, 执行器故障的发生在所难免^[68-71]. 然而, 一旦故障产生, 系统性能必然会受到一定程度的下降, 甚至完全失效. 因此, Zhang 等^[72]针对一类遭遇执行器故障的非线性多智能体系统, 提出基于自适应动态规划的最优容错一致性控制策略. 通过建立一个局部的故障观测器, 每个智能体的潜在故障可以得到有效估计. 对于更严重的情况——系统执行器遭受外部攻击, Xu 等^[73]提出一种基于数据的协同学习算法, 构建弹性预测器为遭受攻击的跟随者提供领导者的状态估计. 通过分别使用在线和离线的方法求解代数黎卡提方程, 得出基于强化学习的最优弹性动态输出反馈控制策略.

此外, 在一些对控制精度要求较高的系统, 如自动驾驶汽车、高速列车控制系统等, 其收敛时间和收敛精度需要受到严格的限制. 这些系统需要在极短的时间内做出准确的响应, 以确保精确性和效率性. 通过将有限时间和固定时间稳定性理论与强化学习控制相结合, Zhang 等^[74]和 Wang 等^[75]分别对多智能体最优一致性控制问题展开研究, 得到系统收敛时间的上界. 值得注意的是, 与文献 [74] 相比较, 文献 [75] 得出的系统收敛时间上界不受系统初始状态的影响. 针对一类非线性多智能体系统的编队控制问题, Zhang 等^[76]基于模糊强化学习方法提出一种预设时间的自适应最优控制策略, 其独特之处在于设计的性能指标函数同时包含预设时间编队误差变量与控制输入损失. 在事先设定的时间常数内, 系统实现了最优控制性能以及编队误差收敛到可确定的范围内.

对于一些高度集成的系统, 由于带宽与计算能力有限, 无法承受复杂外部环境下庞大的信息交互. 在这些系统中, 信息传输和处理的速度受到限制, 因此需要设计轻量级的控制算法, 并在资源有限的情况下实现系统的高效运行. 为了减少系统的通信负担同时平衡系统的控制消耗, 许多研究人员将研究兴趣投向于基于事件触发机制设计强化学习控制策略^[77-81]. Peng 等^[77]针对一类离散线性多智能体系统的领航跟随一致性控制问题, 提出基于事件触发的强化学习最优控制策略. 其中, 控制器与 Actor-Critic 网络的估计权重仅在触发时刻更新, 有效地

降低了信息的传递次数,缓解了系统的通信负担。

1.2.4 系统模型依赖程度

对系统模型依赖程度的差异性直接决定了控制策略的复杂性与实现路径,是影响多智能体强化学习控制效果的重要因素。具体而言,在某些情况下,系统模型可能是已知且准确的,这样可以直接利用模型进行强化学习算法的训练和优化。然而,在另一些情况下,系统模型可能难以获取或不完全可靠,这时需要采用模型无关的强化学习方法,例如基于策略的方法或模型无关的价值函数方法,来实现智能体之间的协同控制。因此,针对不同的系统模型依赖程度,多智能体强化学习控制需要灵活选择合适的方法和策略,以实现系统的协同控制目标。对于模型完全已知的多智能体系统最优包含控制问题, Xiao 等^[82]提出一种基于模型的在线强化学习算法,采用一种单 Critic 网络方法获得耦合 HJ 方程的解。而对于系统动态部分未知的多智能体图形博弈问题, Xiong 等^[83]为了最小化性能指标提出一种基于模型的策略迭代算法与基于数据的异策略积分强化学习算法。针对具有完全未知动态多智能体系统的最优协同问题, Zhang 等^[84]在集中式训练分散式执行的框架下提出一种异策略的强化学习控制算法用于近似 HJ 方程的解。在训练过程中,所有来自智能体的信息投入于集中化的 Critic 网络, Actor 网络则采用一种参数共享机制。在执行过程中,基于每个智能体的观察,其控制动作由训练好的执行网络给出。针对多智能体的最优同步问题, Li 等^[85]仅使用可测量的系统数据提出一种异策略的强化学习算法。这种完全不依赖于系统模型的方法,借助积分强化学习得出了异策略的贝尔曼方程。针对每个智能体,应用一种行为策略去收集数据,不断地去学习离线贝尔曼方程的解。Wang 等^[86]设计了一种基于神经网络的无模型异策略强化学习方法,解决一类非线性连续时间多智能体系统的全局协同一致性控制问题。通过使用行为策略产生的数据集与规范化的梯度下降法,得出 Actor 和 Critic 网络权重的自适应率。为了实现一类未知多智能体系统的无穷时域最优一致性, Ming 等^[87]设计了一种在线自适应强化学习方法。通过使用 Identifier-Actor-Critic 三种不同的神经网络并行学习去近似 HJB 方程的解,其中新提出的 Identifier 网络用于识别未知的系统模型。

2 多智能体强化学习决策

本节聚焦多智能体强化学习决策,这类方法使用马尔科夫决策过程来建模不确定环境中的序列决

策问题,通过在环境中进行交互式学习以优化策略的长期回报^[88]。近年来,多智能体深度强化学习充分利用各类深度神经网络的强大拟合与泛化能力,虽然在一定程度上牺牲了可解释性,但能够处理高维状态和动作空间,特别适合于复杂未知环境中的决策任务。本节首先介绍多智能体决策问题的描述与建模方法,并引入多智能体强化学习的三类主要训练架构。随后,从环境复杂性挑战、计算复杂性挑战、信用分配挑战、对抗博弈挑战四个方面,分别阐述多智能体强化学习所面临的问题和现有的解决方法。本节内容安排如图 3 所示。

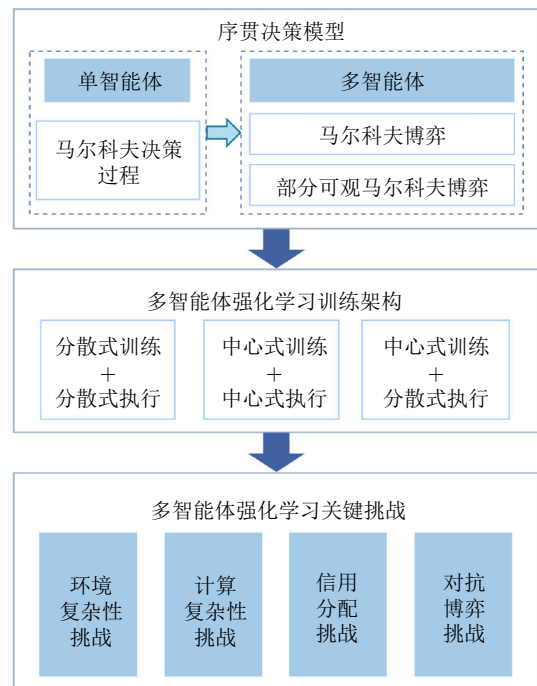


图 3 多智能体强化学习决策

Fig. 3 Multi-agent reinforcement learning decision-making

2.1 决策问题描述与建模

第 2.1 节将分别介绍三类决策模型: 马尔科夫决策过程、马尔科夫博弈、部分可观马尔科夫博弈,为后续的多智能体强化学习决策方法奠定基础。马尔科夫决策过程是建模单智能体序列决策问题的基本框架,通过状态、动作和奖励的概念来形式化决策过程; 马尔科夫博弈模型是马尔科夫决策过程在多智能体系统中的扩展,智能体分别拥有各自的状态空间、动作空间和奖励函数,环境状态变化受到多智能体联合策略影响; 部分可观马尔科夫博弈模型是马尔科夫博弈的进一步扩展,智能体仅能获得对环境的不完全观测,进一步增加了决策过程的复杂性。

2.1.1 马尔科夫决策过程

马尔科夫决策过程 (Markov decision process, MDP)^[89-90] 建立了不确定性环境下的序列决策模型, 表示为五元组: $\langle S, A, R, P, \gamma \rangle$. 其中, S 表示状态空间, 状态是对环境的描述; A 表示动作空间, 包含智能体可以采取的所有可能动作; 状态转移概率 $P(s' | s, a)$ 表示智能体在状态 s 下采取动作 a 后, 环境转移到新状态 s' 的概率; 奖励函数 $R(s, a, s')$ 定义了当环境从状态 s 通过动作 a 转移到状态 s' 时, 智能体所获得的即时奖励; $\gamma \in [0, 1]$ 表示折扣因子, 用于权衡即时奖励和未来奖励的相对重要性. 策略 $\pi(a | s)$ 定义了在某状态 s 采取某动作 a 的概率. 决策者的目标是找到最大化未来累积奖励的最优策略 π^* .

在 MDP 中, 使用价值函数和动作价值函数来评估在给定策略下处于某状态或采取某行动的期望回报. 回报 G_t 是从时间步 t 开始的累积折扣奖励:

$$G_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots = \sum_{k=0}^{\infty} \gamma^k r_{t+k} \quad (44)$$

状态价值函数 $V^\pi(s)$ 表示从状态 s 开始, 遵循策略 π 所能获得的期望折扣回报:

$$V^\pi(s) = \mathbb{E}_\pi[G_t | S_t = s] = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k} | S_t = s \right] \quad (45)$$

状态-动作价值函数 $Q^\pi(s, a)$ 表示在状态 s 下采取动作 a 后, 遵循策略 π 行动所能获得的期望折扣回报:

$$Q^\pi(s, a) = \mathbb{E}_\pi[G_t | S_t = s, A_t = a] = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k} | S_t = s, A_t = a \right] \quad (46)$$

在 MDP 的基础上, 强化学习方法大致可分为两类: 基于值函数的强化学习方法^[91-92] 致力于实现对价值函数的准确估计, 并根据最优价值函数选取动作; 基于策略的强化学习方法^[93-95] 则直接学习一个优化价值函数的参数化策略, 由策略函数输出动作或动作概率.

2.1.2 马尔科夫博弈

如果环境中存在多个决策者 (智能体), 标准马尔科夫决策过程可扩展为马尔科夫博弈模型 (Markov games, MG)^[96], 表示为一个六元组 $\langle S, \mathcal{A}, P, \{R^i\}, \mathcal{N}, \gamma \rangle$. 其中, $\mathcal{A} = A^1 \times A^2 \times \dots \times A^N$ 表示所有智能体的联合动作空间; 状态转移概率 $P(s' | s, a^1, \dots, a^N)$ 描述在状态 s 下所有智能

体选择动作 a^1, \dots, a^N 时, 环境转移到新状态 s' 的概率; \mathcal{N} 表示智能体集合; 每个智能体 i 都有各自的奖励函数 $R^i(s, a^1, \dots, a^N, s')$, 定义了当环境从状态 s 通过所有智能体的联合动作 a^1, \dots, a^N 转移到状态 s' 时, 智能体 i 所获得的即时奖励. 智能体 i 的目标是优化其策略 π^i , 使其未来累积奖励最大化. 此时, 智能体获得的奖励不仅取决于环境和其自身策略, 还取决于其他智能体的策略. 智能体 i 的价值函数表示为

$$V^{\pi^i, \pi^{-i}}(s) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k}^i | S_t = s \right] \quad (47)$$

其中, π 表示联合策略, π^{-i} 表示除智能体 i 以外其他智能体的联合策略.

在求解马尔科夫博弈问题时, 通常希望寻找纳什均衡点, 即没有任何一个智能体能够通过单方面改变自己的策略来增加期望回报. 如果对于所有的智能体 i 和所有可能的策略 π^i , 有

$$V^{\pi^{i*}, \pi^{-i*}}(s) \geq V^{\pi^i, \pi^{-i*}}(s), \quad \forall s \in S \quad (48)$$

那么联合策略 $\pi^* = (\pi^{1*}, \pi^{2*}, \dots, \pi^{N*})$ 是一个纳什均衡策略. 这表明在任何状态 s 下, 如果其他智能体的策略 π^{-i*} 不变, 智能体 i 不可能通过偏离策略 π^{i*} 来获得更高的期望回报. 在纳什均衡点, 每个智能体的策略都是对其他智能体策略的最佳响应.

2.1.3 部分可观马尔科夫博弈

部分可观马尔科夫博弈 (Partially observable Markov games, POMG)^[97] 考虑了在许多现实场景中智能体不能完全观察到整个环境状态的情况, 是对马尔科夫博弈的进一步扩展. 在这种情况下, 每个智能体只能接收到一个与环境状态有关的局部观测, 这增加了决策过程中的不确定性, 使智能体必须基于有限的信息进行决策. 部分可观马尔科夫博弈可以描述为一个元组 $\langle S, \mathcal{A}, P, \{R^i\}, \mathcal{O}, \{Z^i\}, \mathcal{N}, \gamma \rangle$, 其中, $S, \mathcal{A}, P, \{R^i\}, \mathcal{N}, \gamma$ 的定义与马尔科夫博弈相同; \mathcal{O} 表示观测空间, 是环境状态的一个子集, 对于完全可观环境有 $\mathcal{O} \equiv S$; $\{Z^i\}$ 表示观测函数, $Z^i(o^i | s, a^i)$ 定义了在执行动作 a^i 并达到状态 s 后智能体 i 接收到观测 o^i 的概率. 部分可观马尔科夫博弈的一般架构如图 4 所示.

在 POMG 中, 由于状态信息的不完全性, 智能体通常需要维护一个关于当前环境状态的信念, 称为信念状态^[98]. 信念状态 $b(s)$ 是一个概率分布, 用于描述智能体处于各个状态的概率. 根据智能体的观测历史和动作历史序列和初始信念状态, 后续任

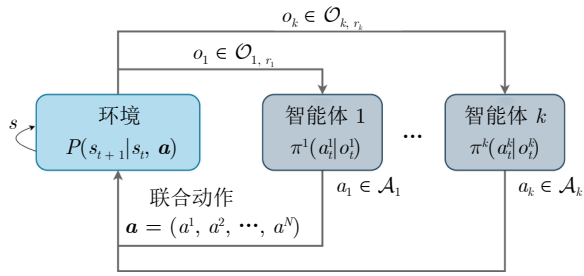


图 4 部分可观马尔科夫博弈示意图

Fig. 4 Diagram of the partially observable Markov games

意时刻的信念状态可以根据贝叶斯规则计算. 基于信念状态, 可以将部分可观 MDP 转化为信念 MDP, 使用信念状态 b 替代环境状态 s . 设 B^i 表示智能体 i 的信念空间, 则策略 π^i 依赖于信念状态, $\pi^i: B^i \rightarrow A^i$.

2.2 多智能体强化学习的一般架构

第 2.2 节分别介绍多智能体强化学习的三类训练架构: 分散式架构、中心式架构、中心式训练-分散式执行架构.

2.2.1 分散式架构

分散式架构, 即分散式训练-分散式执行 (Distributed training with decentralized execution, DTDE), 系统中的每个智能体独立地进行学习和决策. 在训练阶段, 每个智能体只能访问自己的局部观测, 并基于这些观测信息和奖励自行更新其策略或价值函数, 不与其他智能体共享参数和策略. 在分散式架构下, 策略的状态空间及动作空间维度不会随智能体数量增加而增长, 标准深度强化学习算法训练可以直接套用. 此外, 这类方法不需要全局信息, 对计算及通信资源没有特殊要求, 在灵活性和通用性方面具有优势. 例如, Tampuu 等^[99] 基于分散式架构对单智能体 DQN (Deep Q-network) 算法^[100] 进行扩展, 在一些视频游戏环境中研究了不同奖励设置下多智能体的合作与竞争行为; Chen 等^[101] 在一个无通信的多机器人场景中使用单智能体算法分散式地学习避障策略, 相比规划方法实现性能与效率的改进.

DTDE 架构的主要挑战是环境非平稳性. 如果每个智能体独立地学习, 对于系统中任一智能体而言, 其他智能体也成为环境的一部分, 状态转移概率不再仅仅取决于环境本身, 也取决于其他智能体的策略^[102]. 如果智能体的策略随时间不断变化, 将会导致环境的非平稳性, 这限制了单智能体方法在复杂多智能体问题中的直接应用. 此外, 对于 DQN、

DDPG^[103] 等依赖经验回放的强化学习方法, 这种非平稳性会导致智能体在过去采集的经验样本很快过时, 并对学习产生负面影响^[104].

2.2.2 中心式架构

中心式架构, 即中心式训练-中心式执行 (Centralized training with centralized execution, CTCE). 基于 CTCE 的方法学习一个完全中心式的策略, 将多智能体的联合状态空间映射至联合动作空间. 这相当于将多个智能体看作一个“超级智能体”, 使多智能体问题转化为单智能体问题, 不再存在环境非平稳性问题, 可以直接应用单智能体深度强化学习方法求解. 中心式的训练基于全局状态信息, 智能体的观测和经验完全共享, 有助于学习协作策略.

中心式架构所面临的主要挑战是“维度灾难”, 随着智能体数量的增加, 多智能体联合状态空间与联合动作空间都呈指数级增长. 对于高维动作空间问题, Gupta 等^[104] 采用一种动作空间分解方法, 对于由 n 个智能体组成的系统, 可以将中心式策略分解为 n 个子策略, 每个子策略都将联合状态空间映射到一个智能体的局部动作空间. 对于局部动作空间维度为 $|A|$ 的同构多智能体系统, 联合动作空间维度可从 $|A|^n$ 减少到 $n|A|$. 即使如此, 联合状态空间的维度灾难仍然不可忽视. 此外, CTCE 方法假设全局状态随时可以获取, 在训练和执行过程中依赖全局状态信息进行学习与决策. 然而, 在大多数现实场景中该假设很难满足, 在智能体之间进行不受限制的即时通信是不可能的, 这制约了 CTCE 架构的实际应用.

2.2.3 中心式训练-分散式执行架构

中心式训练-分散式执行架构 (Centralized training with decentralized execution, CTDE)^[105] 继承并改进了上述两类方法, 在策略学习过程和实际应用中使用了不同的范式. 在训练阶段, 智能体进行数据交换和访问全局信息相对容易, 因此采取中心式方案, 允许智能体访问包括其他智能体状态和动作在内的全局信息, 基于全局信息对各个智能体的局部策略进行评价和改进. 在实际部署阶段, 智能体之间的通信资源可能十分有限, 难以获取全局状态, 因此智能体仅依赖自身的局部观测信息, 通过各自的策略函数独立地选择动作. CTDE 架构结合了分散式和中心式架构的优点. 一方面, 在训练阶段, 全局信息的引入在一定程度上缓解了环境非平稳性问题, 有助于促进智能体之间的交互和协作; 另一方面, 相比 CTCE 架构, 基于局部观测的分散式策略能够缓解维度灾难的影响, 增强系统的可扩展

展性, 且放宽了对全局信息和通信能力的要求. 在 CTDE 架构下, 有两种主要的实现思路, 分别是基于 Actor-Critic^[106-107] 的方法, 以及基于值函数分解^[108] 的方法.

第 1 类方法将 CTDE 思想与强化学习中的 Actor-Critic 架构相结合. 在标准的 Actor-Critic 架构中, 使用 Critic 网络学习价值函数, 对当前策略进行评估; 使用 Actor 网络学习策略函数, 并根据 Critic 网络的估计调整策略参数. 基于 CTDE 思想, 该架构可以自然地扩展到多智能体系统中, 即中心式 Critic-分散式 Actor 架构. 中心式 Critic 网络的输入包含全局信息, 分散式 Actor 网络的输入仅有局部观测. 在学习阶段, 使用全局信息训练中心式 Critic 网络, 通过 Critic 网络的输出改进各智能体的 Actor 网络; 在执行阶段, 智能体仅依赖自己的策略独立作出决策. Lowe 等^[109] 提出的多智能体深度确定性策略梯度 (Multi-agent deep deterministic policy gradient, MADDPG) 算法是这类方法的一个典型. 在 MADDPG 中, 每个智能体 i 使用自己的策略 π^i 最大化其未来累积折扣奖励, 即

$$J(\pi^i) = E_{s, \mathbf{a} \sim \pi} [R^i(s, \mathbf{a})] \quad (49)$$

其中, $\mathbf{a} = (a^1, \dots, a^N)$ 是所有智能体的联合动作, $R^i(s, \mathbf{a})$ 是智能体 i 在状态 s 和联合动作 \mathbf{a} 下获得的即时奖励. 智能体 i 的目标是最大化 $J(\pi^i)$. 在中心式训练阶段, 采用策略梯度方法更新智能体的 Actor 网络, 即

$$\nabla_{\theta^i} J(\pi^i) = E_{s, \mathbf{a} \sim \pi} \left[\nabla_{\theta^i} \log \pi^i(a^i | o^i) Q^{\pi^i}(s, \mathbf{a}) \right] \quad (50)$$

其中, $Q^{\pi^i}(s, \mathbf{a})$ 表示中心式 Critic 网络, 用于评估 Actor 网络输出决策的质量. Critic 网络的输入包含全局信息, 其更新目标计算如下:

$$y = r^i + \gamma Q^{\pi^i}(s', \mathbf{a}')|_{a^{j'} = \pi^j(o^j)} \quad (51)$$

其中, s' 是下一个状态, $a^{j'} = \pi^j(o^j)$ 是智能体 j 在观察 o^j 下根据其策略 π^j 选择的动作, Critic 网络的损失函数为

$$L(\phi^i) = E_{s, \mathbf{a}, r, s'} \left[(Q^{\phi^i}(s, \mathbf{a}) - y)^2 \right] \quad (52)$$

Critic 网络仅在训练阶段发挥作用. 在分散式执行阶段, 每个智能体 i 根据各自的 Actor 网络 $\pi^i(o^i)$ 的输出进行决策, 其输入仅包含自身局部观测, 无需全局信息.

第 2 类 CTDE 方法采用价值函数分解, 通过将多智能体联合价值函数分解为各个智能体的局部价值函数来简化多智能体学习过程. Sunehag 等^[108] 面

向多智能体合作问题提出值分解网络 (Value decomposition network, VDN), 假设联合奖励是各智能体奖励之和, 从而将联合动作价值函数表示为局部动作价值函数之和. 考虑两个智能体, 若联合奖励函数 $r(s, \mathbf{a})$ 可以加性分解为 $r_1(o^1, a^1) + r_2(o^2, a^2)$, 那么动作价值函数可以分解为

$$Q^{\pi}(s, \mathbf{a}) = Q^1(s, \mathbf{a}) + Q^2(s, \mathbf{a}) \quad (53)$$

在联合价值函数的训练过程中, VDN 还引入了通信机制以共享局部观测信息. 在执行阶段, 由于局部价值函数的输入仅包含局部观测, 因此可以在无通信的情况下进行分布式决策. 在这种分解机制之下, 神经网络能够自动学习不同智能体的价值函数, 更准确地估计智能体的实际贡献, 并提升联合策略的性能.

目前, CTDE 架构已经成为多智能体强化学习的主流训练架构, 但 CTDE 本身并不能彻底解决多智能体学习过程中存在的诸多挑战. 此外, CTDE 架构对于训练过程中的信息共享和数据交换要求较高, 限制了其应用范畴. 在后续内容中, 本文将从几个不同方面对多智能体强化学习所面临的挑战进行分析, 并讨论相应的解决方案.

2.3 环境复杂性挑战

在多智能体系统中, 环境复杂性主要源于多个智能体的互动、环境的动态变化以及任务的多样性. 这些因素共同作用, 使得环境呈现出高度非平稳性和不确定性. 环境复杂性不仅影响算法的设计和实现, 而且也是决定多智能体系统性能的关键因素. 第 2.3 节将从环境的非平稳性、局部可观测性以及角色与任务的多样性这三个角度, 详细探讨这些挑战如何影响多智能体强化学习, 并介绍相应的解决方案.

2.3.1 环境非平稳性

在单智能体强化学习问题中, 环境的状态转移概率仅与当前状态和智能体选择的动作有关. 对智能体而言, 其外部环境模型是稳定的, 只要确定状态和动作, 就能确定状态转移概率. 虽然环境可能包含随机性, 但学习问题是平稳的. 然而, 对于多智能体系统, 环境的状态转移概率与所有智能体的策略相关, 而这些智能体在学习过程中不断更新自己的策略. 从单个智能体的角度看, 其他智能体也构成环境的一部分, 状态转移概率受到联合策略的影响而变化, 环境呈现出非平稳性, 即

$$P(s' | s, a, \pi_m) \neq P(s' | s, a, \pi_n) \quad (54)$$

其中, π_m, π_n 表示不同的多智能体联合策略.

环境非平稳性使强化学习算法面临“移动目标”问题^[110]. 马尔科夫决策过程中, 通常使用状态值函数 V 或状态-动作值函数 Q 作为决策或策略评估的依据, 而深度强化学习方法使用深度神经网络逼近价值函数. 由式 (45) 和式 (46) 可知, 价值函数表示策略的未来期望回报, 而期望回报与环境状态转移函数 P 有关. 如果 P 发生变化, 真实价值函数也将发生变化, 使学习算法在尝试逼近价值函数时的目标不断变化, 难以收敛. 此外, 环境非平稳性使策略梯度方法面临高方差问题: 智能体的奖励函数实际上是 $R(s, a^1, \dots, a^N)$, 奖励取决于所有智能体的动作, 因此仅以智能体自身动作为条件的奖励 $R(s, a^i)$ 具有很高的随机性, 并进一步导致较高的梯度方差. Lowe 等^[109] 证明: 在特定的多智能体设置中, 算法在正确的梯度方向上执行策略改进的概率随智能体数量的增加呈指数级下降.

目前, 中心式训练方法广泛应用于处理环境非平稳性. 以 MADDPG、MAAC^[111]、MAPPO^[112] 为代表的一系列方法通过中心式训练的 Critic 网络估计价值函数, 以应对非平稳性问题. 在训练阶段, Critic 网络可以访问有关其他智能体的附加信息, 其输入包括全局状态 s 和所有智能体的动作 a^1, \dots, a^N . 如果每个智能体都充分了解其他智能体的行为, 那么对单个智能体而言, 环境是平稳的. 对任意联合策略 π , 下式均成立:

$$P(s' | s, a^1, \dots, a^N) = P(s' | s, a^1, \dots, a^N, \pi) \quad (55)$$

因此, 基于这些额外知识的 Critic 网络更有可能对当前状态和策略作出准确的评价. 在此基础上, Iqbal 等^[111] 考虑系统中不同智能体之间的相互影响应有轻重缓急之分, 在 Critic 网络的设计中引入注意力机制^[113-114], 使得每个智能体的 Critic 在评估动作的价值时, 能够自适应地调整对其他智能体状态和动作的“注意力”, 并主要关注那些对当前决策至关重要的信息.

相较完全中心式方法, CTDE 架构显著放宽了对全局信息的需求, 允许智能体在实际执行阶段完全基于局部观测进行决策. 然而, 在一些现实问题中, 要求智能体共享彼此的观测和策略可能难以实现, 必须采取分布式学习方法. Tan^[115] 将表格 Q 学习^[91] 算法应用于多智能体问题, 并对比了完全独立 Q 学习和几种基于共享信息的变体, 其结果表明独立 Q 学习是有效的, 但引入共享信息有助于提升学习效果. Sen 等^[116] 在一个双人合作任务中应用独立 Q 学习, 通过实验表明智能体可以在不共享信息且不知道对方存在的情况下学会自主协作. 尽管如

此, 完全分散式的学习很难彻底解决环境非平稳性问题, Sen 等^[116] 强调在这种情况下需要仔细微调超参数, 否则很有可能影响系统的收敛性. Matignon 等^[117] 指出这类完全分散式方法的成功可能与良好的探索机制有关. 在深度强化学习领域, 独立 Q 学习等方法面临更多挑战: 一方面, 高维的状态-动作空间严重影响探索的效率; 另一方面, 深度 Q 学习等方法使用经验回放技术^[100] 提高样本效率, 而在多智能体系统中, 由于其他智能体的策略更新影响环境模型, 经验回放缓冲区中的样本会很快过时, 并干扰学习进程. Foerster 等^[118] 认为独立 Q 学习忽略了环境中其他智能体策略随时间变化的事实, 导致 Q 函数的非平稳性, 如果考虑其他智能体的策略并用于模型输入, Q 函数就是平稳的. 这与 MADDPG 的出发点相同, 但该工作并未使用真实的策略信息, 而是在经验回放中引入一种“指纹”机制, 这足以使模型能够区分经验回放缓冲区中来自不同阶段的样本, 在保证分散式学习的前提下缓解环境非平稳性问题.

此外, MADDPG 是一种确定性策略梯度方法^[94], 在价值函数的训练过程中, 算法需要根据当前策略产生下一个状态-联合动作对 (s', \mathbf{a}') , 并根据 $Q(s', \mathbf{a}')$ 更新对 $Q(s, \mathbf{a})$ 的估计. 因此, 这类方法对于全局信息的要求较为苛刻, 不仅需要已知全局状态和联合动作, 还需要访问其他智能体的策略. 对于一般的非合作或竞争场景, 要求智能体共享彼此的观测和策略是不合理的. MADDPG 的一种变体^[109] 放宽了对其他智能体策略信息的需求, 通过监督学习方式, 根据观测得到的动作信息训练近似策略网络, 用于推断其他智能体的策略. 类似地, Raileanu 等^[119] 考虑一个存在多种任务的二人博弈场景, 每个智能体拥有各自的任务. 该工作基于“推己及人”的思想设计一种意图推断机制, 根据自身策略和对手的行为, 逆向推测对方的任务, 从而降低系统的不确定性, 并提升最终策略的性能.

2.3.2 局部可观测性

局部可观测性是引起环境复杂性的另一个重要原因. 在多数现实场景中, 系统中的每个智能体只能观察到环境的一部分信息, 由于环境变化并不取决于单一智能体的局部观测, 智能体作出合理决策的难度显著提高. 在深度强化学习领域, 处理局部可观测性的常用方式是使用完整的历史观测序列代替单一时刻的局部观测作为决策依据. 这一思路与信念 MDP^[120] 有异曲同工之处: 信念 MDP 基于历史观测-动作序列和贝叶斯规则维护一个信念状态, 作为对真实状态的估计; 在深度强化学习中, 可以

利用循环神经网络^[121-122]直接处理序列数据, 实现对历史信息的“记忆”. Hausknecht 等^[123]将长短期记忆网络^[121]引入 DQN 算法中, 在不完全观测的情况下实现更好的泛化性. 对于多智能体系统, 值分解网络^[108]引入智能体的局部观测历史 h , 使用 $Q(h, a)$ 作为 $Q(s, a)$ 的近似, 使用如下的价值函数分解方案:

$$Q^\pi(s, a) = Q^1(s, a) + Q^2(s, a) \approx Q^1(h^1, a^1) + Q^2(h^2, a^2) \quad (56)$$

这允许分散式的局部价值函数以局部观测信息作为输入, 同时利用循环神经网络“记忆”历史信息, 缓解局部可观测性带来的负面影响. 此外, 在部分可观测性等因素的影响下, 智能体在环境探索过程中收到的负面反馈可能是环境随机性或其他智能体导致的, 不一定反映当前动作的质量. 对此, 滞后 Q 学习^[124]通过使用两个不同的学习率来更新 Q 值: 当观察到的回报不小于当前的 Q 值估计时, 使用正常的学习率, 否则使用较小的学习率. 这样的设置有助于避免因环境随机性或策略探索而导致 Q 值迅速下降, 使得学习过程更加稳健. Omidshafiei 等^[125]将 DRQN 与滞后 Q 学习相结合, 实现部分可观测条件下的分散式学习与决策.

缓解局部可观测挑战的另一类方法是引入通信, 允许智能体之间交换信息, 以弥补自身观测的不足. 最直观的方法是直接共享原始观测数据^[108], 也有一些方法通过共享神经网络参数^[126]或共享隐藏层输出^[127-128]实现隐式通信. 共享高维的观测或策略数据在实际应用中十分困难, 对此, 一些工作在智能体之间引入显式通信, 并由智能体在交互过程中自主学习通信的编码与解码^[129-130]. Sheng 等^[130]和 Foerster 等^[131]提出 RIAL 和 DIAL 两种通信学习算法. RIAL 结合 DRQN 与独立 Q 学习, 智能体既需要选择一个与环境交互的动作, 也需要选择一个通信动作, 并将其传递给其他智能体. DIAL 将一个智能体的通信输出直接连接到另一个智能体的价值网络输入, 梯度信号可以在不同智能体之间流动, 使智能体能够通过梯度反向传播给予彼此有关其通信行动的反馈. Peng 等^[132]提出多智能体双向协调网络, 用于连接各独立智能体的策略和价值函数, 显著增强了不同智能体之间的交流, 使多智能体学习如何作为一个团队来协调行动. 除了在智能体之间传递局部观测信息外, 也可以传递意图, 如吴俊锋等^[133]提出一种两阶段意图共享方法, 智能体在决策前与其他智能体进行通信, 互相交流意图信息, 从而促进智能体之间的协作. 在现实情况下资源有

限, 必须提高通信效率, 减少通信开销. 对此, Mao 等^[134]通过引入一种自适应门控机制, 修剪不必要的通信以保存智能体间必要的信息流; Kim 等^[135]考虑一个通信带宽有限且智能体共享通信渠道的实际场景, 研究高效的通信调度策略. Das 等^[136]提出 TarMAC, 有选择性地向特定对象发送消息, 以提高通信的效率和效果. Niu 等^[137]基于类似思路, 引入图注意力框架学习通信策略, 以更有效地处理何时通信和通信对象, 以及如何高效处理信息的问题.

2.3.3 角色与任务多样性

多智能体环境复杂性的另一个方面体现为角色和任务的多样性. 不同的智能体可能具有不同的能力或属性, 并承担不同的角色, 每个角色又可能对应着特定的行为模式和任务目标. 一方面, 在多智能体系统中引入分工和协作的机制有望提升系统的适应性和效率; 另一方面, 角色与任务多样性增加了学习过程的复杂性. Lhaksmana 等^[138]设计了一种角色建模方法, 不同的角色对应于不同的行为模式, 并使智能体通过角色之间的切换产生不同的行为. Wang 等^[139]探讨了如何在多智能体强化学习中利用角色概念优化智能体协作, 并提出一种基于角色引导的框架, 使智能体可以根据环境和任务需求, 自主地形成和适应不同的角色. 这些角色并非预先定义, 而是随着学习过程动态演化, 确保了系统的灵活性和适应性. 进一步, Wang 等^[140]在角色学习的基础上引入任务分解机制, 通过将复杂任务分解为多个子任务, 并为每个子任务分配特定角色的智能体, 算法可以更有效地管理多智能体之间的交互和协作. Hu 等^[141]基于智能体的行为模式, 使用对比学习方法生成角色编码, 并在值分解架构的混合网络中引入角色信息, 实现更好的联合价值函数估计. 此外, 方法^[142-143]主要关注如何表达和理解环境中其他实体的相互关系, 利用这些信息提升策略的可解释性, 并促进智能体间的合作和协调.

2.4 计算复杂性挑战

相比单智能体系统, 多智能体系统的状态-动作空间维度显著增高, 这造成两方面挑战. 首先, 强化学习通过试错寻找最优策略, 高效的策略探索十分关键. 在多智能体系统中, 高维的状态-动作空间导致策略探索困难、样本效率低下, 使深度强化学习面临的策略探索问题进一步加剧, 在奖励较稀疏的场景中尤为严重. 其次, 许多先进的多智能体强化学习算法都基于中心式训练架构, 将多智能体联合动作作为价值函数估计的条件. 随着参与的智能体数量增加, 状态和行动空间的维度呈指数级增长,

这导致维度灾难问题,使得学习过程需要处理的数据量和计算量急剧上升,严重影响算法的效率和实用性.第2.4节将分别从样本效率和可扩展性两个角度,阐述缓解计算复杂性挑战的方法.

2.4.1 样本效率

强化学习的样本效率直接影响策略质量和收敛速度.在多智能体系统中,由于环境动态性和维度灾难,动作与状态之间的关系十分复杂,使充分探索环境和策略所需的样本数量大幅增加.在探索不充分的情况下,策略非常容易收敛到局部最优,甚至无法收敛.对此,一种思路是利用先验知识将问题简化或降维.Wang等^[144]引入“动作语义”概念,根据动作是否直接影响其他智能体将动作空间拆解成两部分,并将不同类型的观测信息与不同的动作相关联.通过使用特定的观测信息估计特定动作的价值,该方法减少了不相关信息带来的噪声,简化学习问题并提高样本效率.另一类思路是直接改进价值函数的学习机制,通过减少估计偏差来提高效率.例如Ackermann等^[145]借鉴单智能体领域中的“双重学习”^[146]技术,提出双重中心式Critic网络以减少价值函数的过高估计偏差;Pan等^[147]通过基于正则化的价值函数更新方法来纠正这一问题.

有效的策略探索对提高样本效率和策略性能至关重要.在多智能体系统中,简单的随机探索方法可能不足以应对复杂高维的策略空间,导致学习效率低下,容易收敛到次优策略.Liu等^[148]在策略优化目标中加入一个最大熵目标,鼓励智能体探索潜在的具有更高长期回报的策略,增加策略探索的广度和深度.Na等^[149]使用一种称为“情节记忆”的机制,通过专门记录给定状态-动作对的所能产生的最高回报,有效提高样本效率.Mahajan等^[150]引入额外的潜变量用于增强多智能体联合探索.该方法在智能体价值网络中引入一个以潜变量为条件的参数层,不同的潜变量对应不同的行为模式,使多智能体以协调的方式探索环境.Liu等^[151]提出一种协作式多智能体探索机制,在学习过程中为所有智能体引入额外的共享目标,并训练智能体通过协作实现目标,以促进对不同策略的高效探索.Chen等^[152]引入一种内在动机奖励机制,基于多智能体联合状态和策略的新颖性给予额外奖励,鼓励对未知状态和策略的高效探索.

2.4.2 可扩展性

为了应对环境非平稳性,MADDPG等方法使用中心式Critic网络,其输入依赖于所有智能体的动作,这导致输入空间维度随智能体数量呈指数增长.MAAC方法对此稍作改进:每个智能体的Critic

网络仅以局部观测作为输入,所有的Critic网络通过一个共享的多头注意力结构共享信息,这使得输入空间维度随智能体数量线性增长,在一定程度上提高了可扩展性.进一步,Hao等^[153]提出一种多智能体置换机制,利用信息的置换不变性和置换等变性简化状态空间,从而提高样本效率,增强可扩展性.

平均场方法是简化多智能体问题的另一类有效手段.当智能体的数量增加时,智能体之间的复杂交互行为可能变得难以处理.对此,Yang等^[154]利用平均场表示智能体群体或者邻近智能体对单个智能体的平均影响,从而简化智能体之间的相互作用.对于智能体 k ,原价值函数可以用平均场价值函数近似:

$$Q^k(s, a) \approx Q_{\text{MF}}^k(s, a^k, \bar{a}^k) \quad (57)$$

其中, \bar{a}^k 表示平均动作,即 $\bar{a}^k = \frac{1}{N^k} \sum_{j \in N(k)} a^j$.这种方法将复杂的多智能体相互作用问题转化为更简单的两个实体间的交互问题,单个智能体学习其最优策略是基于整体多智能体团队的动态行为;同时,整体的行为也会随着各个智能体的策略更新而不断调整优化.这能够显著降低模型复杂度,提高学习效率.然而,平均场方法基于置换不变性假设,要求多智能体必须是同构的,Ganapathi等^[155]放宽了该限制,引入多类型智能体,使不同类型的智能体具有不同的任务或目标.Mondal等^[156]同样考虑类似的多类别异构多智能体问题,进一步从理论上证明这类系统可以通过平均场方法近似,并提出相应的可收敛至近似最优策略的自然策略梯度算法.

2.5 信用分配挑战

信用分配^[157]是合作式多智能体强化学习所面临的重要挑战.在多智能体系统中,环境受到所有智能体的共同影响,因此每个智能体的动作不仅影响其自身的即时奖励,还会影响其他智能体的奖励,环境反馈的奖励是所有智能体行为的复合结果.如果多个智能体具有共同的合作目标,就难以区分每个智能体对环境奖励的具体贡献.一个具有较差策略的智能体可能会因为其他智能体的良好行为从环境中接收到奖励,并错误地认为当前策略是较好的,造成“惰性智能体”现象,阻碍了联合策略的优化.因此,如何公正地将环境奖励分配给各个参与的智能体,使得它们能够根据自己行为的实际效果进行学习和调整,成为一个关键问题.

值分解方法^[108]能够有效处理信用分配问题.值分解方法不依赖于中心式训练的Critic网络,每个智能体都有各自的局部价值函数,并以局部观测历

史序列作为输入. 虽然最终的联合价值函数仍然采取中心式训练, 但这种分解机制提供了更高的灵活性, 使价值网络能够更准确地估计智能体的实际贡献, 实现较好的信用分配. 如式 (53) 所示, 标准的值函数分解方法^[108] 直接使用固定的加和形式, 这限制了模型对复杂问题的表达能力. 针对这一问题, 可将加和形式改为函数形式, 联合状态-动作价值函数 Q^{tot} 改由下式计算:

$$Q^{\text{tot}}(s, \mathbf{a}) = f(Q^1(h^1, a^1), \dots, Q^N(h^N, a^N), s) \quad (58)$$

其中, h 表示局部观测历史, $\mathbf{a} = (a^1, \dots, a^N)$ 表示联合动作空间. 式 (58) 通过引入更复杂的结构来克服 VDN 线性加和的限制, 从而更好地处理多智能体间的复杂交互作用. Rashid 等^[158] 提出 QMIX 算法, 使用一个混合网络来表示分解关系 f . 混合网络以各个智能体局部价值网络的输出作为输入, 并输出对联合状态-动作价值的估计. 为了在实际执行阶段实现分散式决策, 由联合价值函数和局部价值函数产生的最优策略必须保持一致, 即满足局部-全局一致性条件:

$$\arg \max_{\mathbf{a}} Q^{\text{tot}}(s, \mathbf{a}) = \left(\arg \max_{a^1} Q^1(h^1, a^1), \dots, \arg \max_{a^N} Q^N(h^N, a^N) \right) \quad (59)$$

QMIX 通过单调性约束来满足上述条件, 即

$$Q^{\text{tot}}(s, \mathbf{a}) \geq Q^{\text{tot}}(s, \mathbf{a}') \\ \text{若 } Q^i(s, a^i) \geq Q^i(s, a'^i), \quad \forall i \quad (60)$$

为确保混合网络满足单调性约束, QMIX 使用超网络控制混合网络的权重和偏置, 通过绝对值激活函数保证混合网络的权重非负. 类似地, Zhou 等^[159] 使用在中心式的 Critic 网络中引入基于全局状态的超网络, 实现基于策略梯度的隐式信用分配. 在 VDN 和 QMIX 基础上产生了一系列改进方法, 以提高算法对复杂价值函数的表征能力. Rashid 等^[160] 提出加权 QMIX 方法, 在损失函数中引入权重, 从而提高对最优联合策略估计的准确性. Yang 等^[161] 利用多头注意力机制近似联合价值函数, 以明确地建模各智能体对联合策略的影响. Son 等^[162] 在中心式训练过程中直接学习真实的全局价值函数 Q^{tot} , 用全局价值函数引导另一个合成价值函数 $Q^{\text{tot}'}$ 的学习, 并使两者具有相同的最优联合动作; 同时, 通过使 $Q^{\text{tot}'}$ 满足单调性条件实现分散式决策.

上述值分解方法均属于隐式信用分配, 依赖于准确的价值函数估计. 另一类是显式信用分配方法,

在奖励分配和价值函数更新中引入额外机制, 以更明确地实现良好信用分配. Foerster 等^[163] 提出反事实多智能体策略梯度 (Counterfactual multi-agent policy gradient, COMA) 算法, 借鉴差异奖励^[164] 的思想, 引入“反事实基准”评估智能体动作的实际贡献. 其主要思想是: 在评价某个智能体的动作时, 固定其他智能体的行为, 并观察改变当前智能体动作所带来的影响. 如果当前动作比特定基线更好, 那么就认为智能体的当前动作对团队有积极贡献. 在 COMA 中, 通过中心式 Critic 网络估计价值函数, 并使用平均回报作为反事实基线, 更准确地估计智能体策略的真实性能. Wang 等^[165] 使用经典的博弈理论中的沙普利值^[166] 来计算每个智能体的边际贡献, 并据此分配奖励. 边际贡献用于描述当一个参与者加入一个已存在的团队时, 对该团队总体表现所做出的额外贡献. 该方法提出沙普利 Q 值用于分配全局奖励, 从而更准确地反映每个智能体的贡献. Li 等^[167] 将反事实基线与沙普利值相结合, 进一步提高复杂任务上的算法表现. 徐诚等^[168] 提出一种基于卡尔曼滤波的奖励估计方法, 通过多智能体团队奖励估计各个智能体的局部奖励, 以实现更好的信用分配, 促进多智能体协作. Chen 等^[169] 进一步考虑一种稀疏奖励场景, 其中智能体仅在回合结束时才能收到奖励, 并提出一种结合沙普利值的时空注意力机制, 实现更低的方差和更快的收敛速度.

2.6 对抗博弈挑战

前文所讨论的方法多数针对合作式多智能体场景, 主要关注如何促进不同个体之间的协调, 共同完成目标或击败竞争对手. 而在完全竞争式任务中, 不同智能体的目标是冲突的, 每个智能体各自优化自己的累积奖励, 对方策略会对己方策略产生显著影响. 在对抗博弈场景下, 智能体不仅需要应对由于对手策略变化带来的非平稳性, 同时要避免对特定策略的过度拟合.

解决这类问题的一类方法是基于博弈论求解纳什均衡策略. Minimax Q-learning 方法^[170] 是解决二人零和随机博弈问题的经典方法. 对于任一智能体, 该方法假设对手采取对自己最不利的策略, 并在这一前提下优化自己的策略. 由于 Minimax Q-learning 采用表格式学习, 后续研究通过函数逼近器将 Minimax Q-learning 中的 Q 表扩展为 Q 函数, 以处理具有更高维状态空间的任務, 如 Zhang 等^[171] 采用线性函数逼近器来拟合 Q 函数并推导其对应的有限采样误差界. Fan 等^[172] 则采用神经网络去逼近 Q 函数, 并对其有限采样误差界给出完整的推导.

在围棋、扑克等场景中,参与者的决策有先后之分,这类问题称为扩展式博弈.对此,Heinrich等^[173]提出虚拟自博弈方法,结合虚拟博弈^[174]思想,基于强化学习方法来求解扩展式博弈下的纳什均衡策略:在每步学习迭代中,每个智能体利用强化学习算法针对敌方的平均策略求取最优响应,并以监督学习的方式用所求最优响应来更新自身的平均策略.在两人零和博弈中,智能体的平均策略可收敛到纳什均衡.进一步,Heinrich等^[175]将神经网络融合到虚拟自博弈算法中,提出神经虚拟自博弈算法,能够在无需先验知识的条件下,近似现实世界中更复杂博弈任务的纳什均衡策略.针对神经虚拟自博弈的一系列变体算法也相继提出,如异步神经虚拟自博弈和蒙特卡洛神经虚拟自博弈^[176]等,以进一步提升算法学习纳什均衡策略的收敛速度和稳定性.

在面对部分可观测的博弈问题时,一种更具通用性的框架是策略空间响应预言机(Policy space response oracle, PSRO)^[177],其基于经验博弈论的思想,维护一个策略种群,以及与策略种群对应的元采样策略,以进行元博弈分析,并利用基于强化学习的最优响应预言机不断扩展策略种群,进而扩展元博弈的规模,再对扩展后的元博弈进行分析,重复迭代,直至不再有新策略产生.由于PSRO方法在处理开放式博弈问题时的通用性和有效性,一系列工作被提出用于从不同方面提升PSRO的性能,例如McAleer等^[178]针对大型零和非完全信息博弈,提出一种扩展性较强的方法Pipeline-PSRO;Muller等^[179]将PSRO扩展到一般和多人博弈的场景,提出基于 α -rank的PSRO方法.

以上方法基于对手策略的显式建模,在特定条件下具有较强的理论支持,但不一定能获得超越均衡解的更优策略^[180].另一类隐式对手建模方法则充分利用深度网络的表征能力,通过观察对手信息,学习对手策略,从而作出有针对性的决策.He等^[181]使用神经网络分别对环境状态和对手特征进行编码,然后使用DQN进行价值函数估计,使神经网络自动发现对手的不同策略模式.Everett等^[182]引入一种切换模型,根据不同的对手模型在相应的最优响应策略之间切换,以适应对手策略的突变.Foerster等^[183]提出一种“考虑对手学习”的学习方法(Learning with apponent-learning awareness, LOLA),通过对手建模方法估计对手策略参数,然后对其参数更新进行建模,在考虑对手策略改进的前提下优化己方策略.在一些场景中,LOLA可以通过预测对方的学习来塑造对手策略,并使己方获得更高回报.

3 多智能体控制与决策应用研究

基于多智能体强化学习的控制与决策方法在诸多现实场景中均有广泛应用.本节主要讨论多智能体强化学习在机器人集群、智能交通、无人船舶控制三个领域的应用研究,并简要列举了少数其他领域的应用实例.

3.1 机器人集群

机器人集群在工业生产、物流运输、搜索救援等许多领域都有广阔的应用前景,在复杂未知的现实世界环境中实现多机器人协作仍然面临许多挑战.许多研究以机器人为背景,有针对性地改进多智能体强化学习算法,并应用于机器人集群的控制与决策任务中.第3.1节将从多机器人导航、多机器人覆盖路径规划、多机器人任务分配三个领域,介绍相关研究进展.

1) 多机器人导航

在多机器人导航问题中,机器人从各自的起点出发,目的是安全到达指定的终点,同时最小化时间或距离成本.传统的启发式方法对传感器、在线计算能力和通信资源的要求较高,许多研究开始使用多智能体强化学习方法训练具备自主导航和避障能力的多机器人集群.例如,启发式方法在寻找有效路径时往往需要实时预测其他智能体和障碍物的行为,Chen等^[101]借助分散式多智能体深度强化学习方法,将这种在线计算卸载到离线训练过程中,提高算法的实时计算效率;Long等^[184]使用分散式的强化学习方法将原始传感器测量结果直接映射到智能体的移动速度和转向命令,在大规模多机器人场景中实现有效的避障策略;Willemsen等^[185]面向多机器人系统,引入基于学习的世界模型来提高现实世界中多智能体强化学习算法的样本效率.对于更复杂的多无人机系统,Yue等^[186]利用改进的MAAC方法^[111]使无人机集群能够学习未知环境下的协同多目标跟随;Xue等^[187]设计了一种基于MADDPG^[109]的改进算法,用于协调多架无人机在复杂、未知的三维环境中安全导航.

2) 多机器人覆盖路径规划

覆盖路径规划是一种特定类型的路径规划,目标是使机器人高效地遍历整个区域,尽可能少地重复经过同一位置,同时优化行驶的总距离或时间.覆盖路径规划在自动化清洁、农业、搜索和救援等场景中有广泛的应用.Mou等^[188]研究无人机集群三维不规则地形覆盖问题,该方法基于领导者-跟随者双层架构,并在上层应用多智能体强化学习算法.Sheng等^[130]针对多机器人目标搜索问题,提出

一种基于分解的多智能体强化学习算法, 通过确保满足局部-全局一致性条件实现分散式决策. Hou 等^[189]针对大规模搜索任务, 提出一种基于多智能体强化学习算法的分布式协同搜索方法, 可以在复杂和大规模的场景中高效运行.

3) 多机器人任务分配

多机器人任务分配的目的在于高效地为多个机器人分配一系列任务, 以便优化团队整体的执行效率、减少完成时间、提高资源利用率或其他相关性能指标. Cui 等^[190]基于多智能体强化学习开发了一种多无人机通信网络动态资源分配算法, 使每架无人机自主选择其通信用户、功率级别和信道与地面用户进行通信. Wang 等^[191]将多机器人任务分配问题建模为马尔科夫决策过程, 利用图神经网络学习调度问题的特征, 并通过强化学习和模仿学习来学习多机器人调度策略. Johnson 等^[192]考虑实际生产环境中的机器人装配单元, 提出一种多智能体强化学习方法实现装配单元中的任务调度. 进一步, Paul 等^[193]引入编码器-解码器架构, 提出一种图强化学习架构用于学习任务分配策略, 并在更大规模的问题上进行验证.

3.2 智能交通

智能交通系统是未来交通的发展方向, 目的是通过使用各种感知技术和智能算法改善交通流量、减少拥堵、提高安全性和能源效率. 由于交通系统是复杂、动态的多智能体系统, 多智能体强化学习在智能交通领域具有广阔的应用前景. 第 3.2 节将从自动驾驶决策、交通信号调度、车辆协同控制三个方面, 介绍相关研究进展.

1) 自动驾驶决策

虽然许多自动驾驶算法仅针对单一车辆, 但是交通系统是一个包含多种交通参与者的多智能体系统, 自动驾驶车辆需要与行人、人类驾驶员和其他自动车辆进行交互, 环境变化不一定满足马尔科夫性. 对此, Shalev-Shwartz 等^[194]提出了在不满足马尔科夫性的环境中进行策略更新的方法, 将基于学习的驾驶策略和基于规则的轨迹规划方法相结合, 并引入分层机制以降低策略学习的方差, 提高复杂环境中的安全性. 进一步的研究考虑了多辆自动驾驶汽车组成的车队, 例如 Yu 等^[195]考虑高速公路情况下自动驾驶车队的高层决策问题, 提出一种能够协调多辆自动驾驶车辆的多智能体强化学习方法, 利用协调图显式地建模车辆之间的依赖关系, 从而降低整个决策问题的计算复杂性. Liu 等^[196]针对车队保持问题, 提出一种基于深度 Q 网络和共识算法的分布式强化学习方法, 使所有车辆学会以特定的

队形和相同的速度前进. Liang 等^[197]针对多车合作变道问题提出一种分层强化学习方法, 并引入对手建模机制在学习过程中建模其他智能体的策略, 以缓解环境非平稳性. 将策略从模拟环境转移到现实世界是一个巨大的挑战, 对此, Candela 等^[198]使用多智能体强化学习算法训练自动驾驶策略, 并提出一种将多智能体自动驾驶策略转移到现实世界的方法, 以弥补虚拟到现实的“鸿沟”.

2) 交通信号调度

多智能体强化学习广泛应用于解决复杂交通网络中的自适应交通信号调度问题. 其中, 每个交叉路口的信号灯可以视作一个智能体, 各个智能体通过观察交通流量、等待时间等环境因素来学习如何调整交通信号灯, 从而减少车辆拥堵. Chu 等^[199]提出一种完全分散式的多智能体强化学习算法, 通过提高可观察性降低每个独立智能体的策略学习难度, 并成功应用于大规模交通模拟环境中. Jiang 等^[200]在分布式强化学习基础上引入图分解机制, 提高大规模场景下的计算效率, 节省训练时间. 进一步, Wang 等^[201]引入双 Q 学习^[146]机制减少价值估计偏差, 并使用平均场方法^[154]近似多智能体之间的交互, 进一步提升了算法性能.

3) 车辆协同控制

传统的车辆控制方法通常集中在单车辆上, 但在现实道路环境中, 车辆往往需要与其他车辆协同行驶以达到更高的效率和安全性. 在多智能体强化学习中, 每辆车辆被视为一个智能体, 其目标是通过与其他车辆和环境进行交互学习, 以实现更好的协同控制. 对于智能车的巡航控制问题, Wang 等^[202]通过单 Critic 神经网络架构和存储的经验数据搭建了强化学习控制架构, 提出一种基于动态事件触发的自适应最优控制方案. 而当无人车系统遭受到服务器拒绝攻击时, Xu 等^[203]设计了一种基于感知的死区控制策略, 以减少相邻车辆之间的通信负载, 并采用异策略积分强化学习算法, 避免了对车辆动态模型的依赖. 由于无人机和无人车各自具有独特的优势和应用场景, 通过天-地协同控制可以实现无人机和无人车之间的信息共享和协同行动, 提高整体交通系统的效率和性能^[204-207]. Zhao 等^[206]针对无人机与无人车的空-地协同编队问题构造了一个鲁棒最优编队控制器来抑制系统非线性、耦合和外部扰动. 其中, 一种基于数据驱动的强化学习算法被提出, 用以更新无人机和无人车的最优控制策略.

3.3 无人船舶

强化学习控制在无人船舶中的应用是近年来人工智能技术在海洋领域中的重要应用之一. 传统的

无人船舶控制方法通常依赖于预先设计的控制策略,而强化学习控制则通过不断与环境交互,自主学习最优策略来实现船舶的智能控制和任务执行.具体而言,强化学习可以用于无人船舶的目标跟踪和避碰控制.通过将目标跟踪和避碰任务建模为强化学习问题,无人船舶可以通过学习与目标船舶的交互,动态调整自身的航向和速度,以实现有效的目标跟踪和避碰行为.此外,强化学习还可以应用于无人船舶的控制策略优化和自适应控制.通过在线学习和迭代优化,无人船舶可以不断改进自身的控制策略,适应不同的环境和任务需求,提高控制性能和鲁棒性.

Song 等^[208]针对具有外部扰动的非线性欠驱动无人船系统提出一种强化 Q 学习最优跟踪控制策略.其中,非线性动态由 Takagi-Sugeno (T-S) 模糊模型近似,同时系统初始容许控制策略假设条件也被移除,利用 Q 学习值迭代算法求解代数黎卡提方程获得最优解的存在条件.为了实现多无人船舶系统的编队碰撞避免,Chen 等^[209]提出一种基于 Actor-Critic 学习策略的最优强化学习控制算法,并使用预设性能控制技术保障系统的暂态与稳态性能.对于存在执行器故障的多无人船系统一致性控制问题,Bai 等^[210]设计一种分布式自适应强化学习控制策略提升系统的容错能力和鲁棒性,其中 Actor 与 Critic 网络分别用于近似效用函数与未知动态.进一步,Chen 等^[211]考虑一种更为严苛的情况,即当一队具有 4 自由度的欠驱动多无人船系统同时存在执行器故障、输入饱和、输入延迟、状态受限时的最优编队控制问题.一种基于强化学习算法的有限时间滑模控制器被设计来实现领航-跟随密集编队.此外,Weng 等^[212]还考虑到海-空协同编队控制问题,针对存在异构动态情况的无人船与无人机设计一种基于事件触发的最优编队控制器,提高了各子系统间的通信效率.一种新的自组织 Actor-Critic 强化学习神经网络被用于求解 HJB 方程,其神经元个数可根据系统性能进行动态调整.

3.4 其他领域应用

除上述三方面外,多智能体强化学习在许多其他领域中具有广阔的应用前景.在游戏博弈领域,多智能体强化学习的应用十分广泛,其中一项代表性工作是 DeepMind 训练的 FTW (For the win) 智能体^[213],其利用多智能体强化学习,成功学会了多人射击游戏中的团队协作策略,达到人类玩家水平.在能源管理方面,多智能体强化学习可用于电力需求响应和能源调度等任务^[214-216],以预测和管理

电力需求,优化能源的使用效率.例如,Zhang 等^[216]针对电动汽车电站的能源采购和分配问题,使用改进的 MADDPG 算法^[109]学习能源采购策略,优化经济成本和用户满意度.在资源调度方面,一些研究利用多智能体强化学习实现资源调度和任务分配^[217-219],例如:Zhao 等^[217]针对大型 GPU 集群中分布式深度学习作业的调度问题,提出基于多智能体强化学习的调度器;邝祝芳等^[220]针对移动边缘计算中的任务卸载问题,提出基于深度强化学习的资源分配算法.此外,强化学习控制在微电网中的应用同样具有显著的优势,能够有效提升微电网的运行效率、稳定性和经济性^[221-225].其中,Adibi 等^[221]针对有损电网的二次频率同步问题提出一种基于 Actor-Critic 结构的在线强化学习控制方法,能够有效地处理电阻和电感线路及负载阻抗、参数不确定性、时变负载和干扰等一系列问题,同时解放了对系统内部动态的需求.对于直流微电网的均流和电压调节问题,Dong 等^[225]设计了一种基于数据驱动的鲁棒最优一致性控制策略,通过建立一种 Q 学习方法来获取近似的最优控制策略和成本函数.针对涡轮机的控制问题^[226-228],融合强化学习可以优化控制策略达到提高发电效率、减少机械应力和延长设备寿命等目的.例如,Xie 等^[228]研究了传感器和执行器故障下风力涡轮机的俯仰与转矩控制问题,提出一种仅在增量域内近似系统模型的强化学习被动容错控制策略,保障了系统的平稳运行.

4 总结与展望

本文分别从智能控制与自主决策的角度综述了多智能体强化学习的研究进展.在多智能体强化学习控制领域,本文首先从博弈的角度出发考虑智能体间的竞争与合作的关系,分别介绍了零和博弈、非零和博弈与 Stackelberg 博弈三种主要的博弈类型,总结了典型例子与算法;随后从多智能体协同的角度,总结了强化学习控制方法在不同系统动态与需求下所取得的研究成果.在多智能体强化学习决策领域,本文从马尔科夫决策过程和马尔科夫博弈模型出发,分别介绍了多智能体序列决策的一般建模方法;总结了多智能体深度强化学习的三类训练架构;针对环境复杂性、计算复杂性、信用分配和对抗博弈四个方面,梳理了多智能体强化学习所面临的挑战,阐述了主流解决方案和最新研究进展.然后,本文选择机器人集群、智能交通、无人船舶三个重要应用场景,分别介绍了多智能体强化学习控制与决策方法的应用研究.最后,结合对当前研究

成果的分析与思考, 本文对多智能体强化学习的未来研究方向进行了一些展望。

1) 跨越从虚拟到现实的障碍

在多智能体强化学习的应用中, 从虚拟环境到现实世界的过渡是一个重大挑战。出于成本、时间和安全性的考虑, 绝大多数深度强化学习研究都依赖虚拟环境提供学习所需的交互数据。尽管多智能体强化学习方法在虚拟环境中显示出强大的潜力, 但常规虚拟环境不可能完全还原现实, 真实环境所具有的高度复杂性、动态性和不确定性很难体现在计算机仿真中, 限制了算法在现实任务中的有效性。近年来, 一些研究借助大语言模型, 构建了包含虚拟居民和可交互社会场景的环境模拟器, 如“斯坦福小镇”^[229], 以及面向机器人设计的 GRUtopia^[230] 等。“斯坦福小镇”是一个由 25 个智能体和房屋、商店、公园等公共场所组成的虚拟环境, 其中的智能体能够产生自发的个体和社交行为, 并在小镇中生活和工作。同时, 人类用户也可以通过自然语言与这些计算机智能体进行互动。未来, 这类由大语言模型驱动的虚拟环境有望成为虚拟与现实之间的桥梁, 为多智能体强化学习研究提供更加逼近现实的训练和评估平台。

另一方面, 目前的许多研究忽略了真实环境中存在的诸多约束, 例如: 虚拟环境中可以轻易获取的环境数据在现实中很难得到、分布式设备的计算和通信资源有限、智能体之间的信息传递存在延迟和干扰等^[231]。为解决这些问题, 未来研究需要关注方法的通用性和鲁棒性, 重点考虑实际资源限制和环境复杂性, 探索轻量且高效的算法, 以适应现实世界应用的需求。

2) 高效解决多目标与多任务决策

许多现实决策问题包含相互冲突的多个目标, 策略必须在多个目标之间进行权衡, 不存在绝对的最优策略。标准单目标方法处理多目标问题时会面临如下一些问题: 奖励函数的设计高度依赖直觉, 是一个繁琐的试错过程, 奖励函数与最终结果之间的关系通常是非线性的; 在设计决策系统时, 用户对不同目标的偏好可能是未知的; 目标偏好可能会因时间、具体场景、用户需求、安全约束等因素发生变化, 有时可能需要在不同策略之间切换。在单智能体领域, 多目标强化学习方法已有一些研究, 如文献 [232–234] 等; 而多智能体和多目标强化学习这一新兴交叉领域的研究较少, 相关工作如 Hu 等^[235] 针对多智能体协同决策问题, 提出一种多目标多智能体强化学习方法。现实世界决策中的另一项挑战是多任务问题, 智能体可能会面临相关但不完全相

同的多项任务。为单个任务分别学习专门的策略十分低效, 因为智能体不仅必须为每个任务存储不同的策略, 而且在实践中智能体可能需要自行判断任务特征。多任务强化学习的目的是使智能体学习可以在各种相关任务中共享和使用的通用技能, 以提高策略的适用性和泛化性。在多智能体强化学习领域, 可能面临个体和团队两个层面上的多任务问题: 在个体层面, 不同智能体可能具有不同的任务, 彼此之间需要配合或分工; 在团队层面, 智能体团队作为一个整体, 可能需要学习各类不同技能, 以解决多种团队协作任务。

总之, 多目标和多任务问题是在现实世界应用强化学习决策方法所面临的重要挑战, 而这些挑战在多智能体系统中变得更加复杂。因此, 面向多智能体强化学习, 探索高效的多目标和多任务决策方法, 是值得进一步研究的方向。

3) 提高决策安全性和可解释性

在自动驾驶、工业自动化和医疗系统等实际场景中, 安全性至关重要。这不仅要求智能体能够学习最优策略, 同时也必须保证在学习过程中和应用策略时的安全性, 以防止可能的风险和损失^[236]。然而, 深度强化学习方法通常以最大化奖励函数为目标, 这可能使智能体忽略现实场景中的安全约束, 阻碍了其在真实世界中的应用。在多智能体系统中, 环境非平稳性和局部可观测问题增加了学习难度, 也加剧了安全挑战。单个智能体要在满足自身安全约束的前提下优化奖励函数, 同时还要考虑其他智能体的行为, 以确保联合动作满足安全约束。

因此, 研究考虑约束的、安全的多智能体强化学习方法, 有助于提高多智能体强化学习在现实应用中的可行性, 是未来一个重要的研究方向。

References

- 1 Liu Quan, Zhai Jian-Wei, Zhang Zong-Chang, Zhong Shan, Zhou Qian, Zhang Peng, et al. A survey on deep reinforcement learning. *Chinese Journal of Computers*, 2018, 41(1): 1–27 (刘全, 翟建伟, 章宗长, 钟珊, 周倩, 章鹏, 等. 深度强化学习综述. 计算机学报, 2018, 41(1): 1–27)
- 2 Zhou F, Luo B, Wu Z K, Huang T W. SMONAC: Supervised multiobjective negative actor-critic for sequential recommendation. *IEEE Transactions on Neural Networks and Learning Systems*, 2024, 35(12): 18525–18537
- 3 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
- 4 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
- 5 Huang X Y, Li Z Y, Xiang Y Z, Ni Y M, Chi Y F, Li Y H, et al. Creating a dynamic quadrupedal robotic goalkeeper with reinforcement learning. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Detroit, USA: IEEE, 2023. 2715–2722

- 6 Perolat J, de Vlader B, Hennes D, Tarassov E, Strub F, de Boer V, et al. Mastering the game of stratego with model-free multiagent reinforcement learning. *Science*, 2022, **378**(6623): 990–996
- 7 Fawzi A, Balog M, Huang A, Hubert T, Romera-Paredes B, Barekatin M, et al. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 2022, **610**(7930): 47–53
- 8 Zhou Q, Wang W, Liang H J, Basin M V, Wang B H. Observer-based event-triggered fuzzy adaptive bipartite containment control of multiagent systems with input quantization. *IEEE Transactions on Fuzzy Systems*, 2021, **29**(2): 372–384
- 9 Wu H N, Luo B. Neural network based online simultaneous policy update algorithm for solving the HJI equation in nonlinear H_∞ control. *IEEE Transactions on Neural Networks and Learning Systems*, 2012, **23**(12): 1884–1895
- 10 Vrable D, Vamvoudakis K G, Lewis F L. *Optimal Adaptive Control and Differential Games by Reinforcement Learning Principles*. London: Institution of Engineering and Technology, 2012.
- 11 Luo B, Wu H N, Huang T W. Off-policy reinforcement learning for H_∞ control design. *IEEE Transactions on Cybernetics*, 2015, **45**(1): 65–76
- 12 Luo B, Huang T W, Wu H N, Yang X. Data-driven H_∞ control for nonlinear distributed parameter systems. *IEEE Transactions on Neural Networks and Learning Systems*, 2015, **26**(11): 2949–2961
- 13 Fu Y, Fu J, Chai T Y. Robust adaptive dynamic programming of two-player zero-sum games for continuous-time linear systems. *IEEE Transactions on Neural Networks and Learning Systems*, 2015, **26**(12): 3314–3319
- 14 Liu Q Y, Wang Z D, He X, Zhou D H. Event-based H_∞ consensus control of multi-agent systems with relative output feedback: The finite-horizon case. *IEEE Transactions on Automatic Control*, 2015, **60**(9): 2553–2558
- 15 Vamvoudakis K G, Lewis F L. Online solution of nonlinear two-player zero-sum games using synchronous policy iteration. *International Journal of Robust and Nonlinear Control*, 2012, **22**(13): 1460–1483
- 16 Luo B, Yang Y, Liu D R. Policy iteration Q-learning for data-based two-player zero-sum game of linear discrete-time systems. *IEEE Transactions on Cybernetics*, 2021, **51**(7): 3630–3640
- 17 van der Schaft A J. L_2 -gain analysis of nonlinear systems and nonlinear state-feedback H_∞ control. *IEEE Transactions on Automatic Control*, 1992, **37**(6): 770–784
- 18 Luo B, Wu H N. Computationally efficient simultaneous policy update algorithm for nonlinear H_∞ state feedback control with Galerkin's method. *International Journal of Robust and Nonlinear Control*, 2013, **23**(9): 991–1012
- 19 Sun J L, Long T. Event-triggered distributed zero-sum differential game for nonlinear multi-agent systems using adaptive dynamic programming. *ISA Transactions*, 2021, **110**: 39–52
- 20 Zhou Y, Zhou J L, Wen G H, Gan M G, Yang T. Distributed minmax strategy for consensus tracking in differential graphical games: A model-free approach. *IEEE Systems, Man, and Cybernetics Magazine*, 2023, **9**(4): 53–68
- 21 Sun J L, Liu C S. Distributed zero-sum differential game for multi-agent systems in strict-feedback form with input saturation and output constraint. *Neural Networks*, 2018, **106**: 8–19
- 22 Li M H, Wang D, Qiao J F. Neural critic learning for tracking control design of constrained nonlinear multi-person zero-sum games. *Neurocomputing*, 2022, **512**: 456–465
- 23 Jiao Q, Modares H, Xu S Y, Lewis F L, Vamvoudakis K G. Multi-agent zero-sum differential graphical games for disturbance rejection in distributed control. *Automatica*, 2016, **69**: 24–34
- 24 Chen C, Lewis F L, Xie K, Lyu Y, Xie S L. Distributed output data-driven optimal robust synchronization of heterogeneous multi-agent systems. *Automatica*, 2023, **153**: Article No. 111030
- 25 Zhang H, Li Y, Wang Z P, Ding Y, Yan H C. Distributed optimal control of nonlinear system based on policy gradient with external disturbance. *IEEE Transactions on Network Science and Engineering*, 2024, **11**(1): 872–885
- 26 An C L, Su H S, Chen S M. H_∞ consensus for discrete-time fractional-order multi-agent systems with disturbance via Q-learning in zero-sum games. *IEEE Transactions on Network Science and Engineering*, 2022, **9**(4): 2803–2814
- 27 Ma Y J, Meng Q Y, Jiang B, Ren H. Fault-tolerant control for second-order nonlinear systems with actuator faults via zero-sum differential game. *Engineering Applications of Artificial Intelligence*, 2023, **123**: Article No. 106342
- 28 Wu Y, Chen M, Li H Y, Chadli M. Mixed-zero-sum-game-based memory event-triggered cooperative control of heterogeneous MASs against DoS attacks. *IEEE Transactions on Cybernetics*, 2024, **54**(10): 5733–5745
- 29 Li Meng-Hua, Wang Ding, Qiao Jun-Fei. Adaptive critic control for multi-player non-zero-sum games with asymmetric constraints. *Control Theory and Applications*, 2023, **40**(9): 1562–1568
(李梦华, 王鼎, 乔俊飞. 不对称约束多人非零和博弈的自适应评判控制. 控制理论与应用, 2023, **40**(9): 1562–1568)
- 30 Lv Yong-Feng, Tian Jian-Yan, Jian Long, Ren Xue-Mei. Approximate-dynamic-programming H_∞ controls for multi-input nonlinear system. *Control Theory and Applications*, 2021, **38**(10): 1662–1670
(吕永峰, 田建艳, 菅垄, 任雪梅. 非线性多输入系统的近似动态规划 H_∞ 控制. 控制理论与应用, 2021, **38**(10): 1662–1670)
- 31 Hong Cheng-Wen, Fu Yue. Nonlinear robust approximate optimal tracking control based on adaptive dynamic programming. *Control Theory and Applications*, 2018, **35**(9): 1285–1292
(洪成文, 富月. 基于自适应动态规划的非线性鲁棒近似最优跟踪控制. 控制理论与应用, 2018, **35**(9): 1285–1292)
- 32 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
- 33 Song R Z, Lewis F L, Wei Q L. Off-policy integral reinforcement learning method to solve nonlinear continuous-time multiplayer nonzero-sum games. *IEEE Transactions on Neural Networks and Learning Systems*, 2017, **28**(3): 704–713
- 34 Kamalapurkar R, Klotz J R, Dixon W E. Concurrent learning-based approximate feedback-Nash equilibrium solution of N-player nonzero-sum differential games. *IEEE/CAA Journal of Automatica Sinica*, 2014, **1**(3): 239–247
- 35 Zhao Q T, Sun J, Wang G, Chen J. Event-triggered ADP for nonzero-sum games of unknown nonlinear systems. *IEEE Transactions on Neural Networks and Learning Systems*, 2022, **33**(5): 1905–1913
- 36 Yang X D, Zhang H, Wang Z P. Data-based optimal consensus control for multiagent systems with policy gradient reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2022, **33**(8): 3872–3883
- 37 Abouheaf M I, Lewis F L, Vamvoudakis K G, Haesaert S, Babuska R. Multi-agent discrete-time graphical games and reinforcement learning solutions. *Automatica*, 2014, **50**(12): 3038–3053
- 38 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
- 39 Yang N, Xiao J W, Xiao L, Wang Y W. Non-zero sum differential graphical game: Cluster synchronisation for multi-agents with partially unknown dynamics. *International Journal of Control*, 2019, **92**(10): 2408–2419

- 40 Odekunle A, Gao W, Davari M, Jiang Z P. Reinforcement learning and non-zero-sum game output regulation for multi-player linear uncertain systems. *Automatica*, 2020, **112**: Article No. 108672
- 41 Wang Y, Xue H W, Wen J W, Liu J F, Luan X L. Efficient off-policy Q-learning for multi-agent systems by solving dual games. *International Journal of Robust and Nonlinear Control*, 2024, **34**(6): 4193–4212
- 42 Su H G, Zhang H G, Liang Y L, Mu Y F. Online event-triggered adaptive critic design for non-zero-sum games of partially unknown networked systems. *Neurocomputing*, 2019, **368**: 84–98
- 43 Yu M M, Hong S H. A real-time demand-response algorithm for smart grids: A Stackelberg game approach. *IEEE Transactions on Smart Grid*, 2016, **7**(2): 879–888
- 44 Yang B, Li Z Y, Chen S M, Wang T, Li K Q. Stackelberg game approach for energy-aware resource allocation in data centers. *IEEE Transactions on Parallel and Distributed Systems*, 2016, **27**(12): 3646–3658
- 45 Yoon S G, Choi Y J, Park J K, Bahk S. Stackelberg-game-based demand response for at-home electric vehicle charging. *IEEE Transactions on Vehicular Technology*, 2016, **65**(6): 4172–4184
- 46 Lin M D, Zhao B, Liu D R. Event-triggered robust adaptive dynamic programming for multiplayer Stackelberg-Nash games of uncertain nonlinear systems. *IEEE Transactions on Cybernetics*, 2024, **54**(1): 273–286
- 47 Li M, Qin J H, Ma Q C, Zheng W X, Kang Y. Hierarchical optimal synchronization for linear systems via reinforcement learning: A Stackelberg-Nash game perspective. *IEEE Transactions on Neural Networks and Learning Systems*, 2021, **32**(4): 1600–1611
- 48 Yan L, Liu J H, Lai G Y, Chen C L P, Wu Z Z, Liu Z. Adaptive optimal output-feedback consensus tracking control of nonlinear multiagent systems using two-player Stackelberg game. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2024, **54**(9): 5377–5387
- 49 Li D D, Dong J X. Output-feedback optimized consensus for directed graph multi-agent systems based on reinforcement learning and subsystem error derivatives. *Information Sciences*, 2023, **649**: Article No. 119577
- 50 Zhang D F, Yao Y, Wu Z J. Reinforcement learning based optimal synchronization control for multi-agent systems with input constraints using vanishing viscosity method. *Information Sciences*, 2023, **637**: Article No. 118949
- 51 Li Q, Xia L N, Song R Z, Liu J. Leader-follower bipartite output synchronization on signed digraphs under adversarial factors via data-based reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2020, **31**(10): 4185–4195
- 52 Luo A, Zhou Q, Ren H R, Ma H, Lu R Q. Reinforcement learning-based consensus control for MASs with intermittent constraints. *Neural Networks*, 2024, **172**: Article No. 106105
- 53 Yu J L, Dong X W, Li Q D, Lv J H, Ren Z. Adaptive practical optimal time-varying formation tracking control for disturbed high-order multi-agent systems. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2022, **69**(6): 2567–2578
- 54 Lan J, Liu Y J, Yu D X, Wen G X, Tong S C, Liu L. Time-varying optimal formation control for second-order multiagent systems based on neural network observer and reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2024, **35**(3): 3144–3155
- 55 Wang Z K, Zhang L J. Distributed optimal formation tracking control based on reinforcement learning for underactuated AUVs with asymmetric constraints. *Ocean Engineering*, 2023, **280**: Article No. 114491
- 56 Cheng M, Liu H, Gao Q, Lu J H, Xia X H. Optimal containment control of a quadrotor team with active leaders via reinforcement learning. *IEEE Transactions on Cybernetics*, 2024, **54**(8): 4502–4512
- 57 Zuo S, Song Y D, Lewis F L, Davoudi A. Optimal robust output containment of unknown heterogeneous multiagent system using off-policy reinforcement learning. *IEEE Transactions on Cybernetics*, 2018, **48**(11): 3197–3207
- 58 Wang F Y, Cao A, Yin Y H, Liu Z X. Model-free containment control of fully heterogeneous linear multiagent systems. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2024, **54**(4): 2551–2562
- 59 Qin J H, Li M, Shi Y, Ma Q C, Zheng W X. Optimal synchronization control of multiagent systems with input saturation via off-policy reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2019, **30**(1): 85–96
- 60 Mu C X, Zhao Q, Gao Z K, Sun C Y. Q-learning solution for optimal consensus control of discrete-time multiagent systems using reinforcement learning. *Journal of the Franklin Institute*, 2019, **356**(13): 6946–6967
- 61 Bai W W, Li T S, Long Y, Chen C L P. Event-triggered multi-gradient recursive reinforcement learning tracking control for multiagent systems. *IEEE Transactions on Neural Networks and Learning Systems*, 2023, **34**(1): 366–379
- 62 Sun J Y, Ming Z Y. Cooperative differential game-based distributed optimal synchronization control of heterogeneous nonlinear multiagent systems. *IEEE Transactions on Cybernetics*, 2023, **53**(12): 7933–7942
- 63 Ji L H, Wang C H, Zhang C J, Wang H W, Li H Q. Optimal consensus model-free control for multi-agent systems subject to input delays and switching topologies. *Information Sciences*, 2022, **589**: 497–515
- 64 Guang W W, Wang X, Tan L H, Sun J, Huang T W. Prescribed-time optimal consensus for switched stochastic multiagent systems: Reinforcement learning strategy. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2025, **9**(1): 75–86
- 65 Wang Z S, Liu Y Y, Zhang H G. Two-layer reinforcement learning for output consensus of multiagent systems under switching topology. *IEEE Transactions on Cybernetics*, 2024, **54**(9): 5463–5472
- 66 Liu D Y, Liu H, Lv J H, Lewis F L. Time-varying formation of heterogeneous multiagent systems via reinforcement learning subject to switching topologies. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2023, **70**(6): 2550–2560
- 67 Qin J H, Ma Q C, Yu X H, Kang Y. Output containment control for heterogeneous linear multiagent systems with fixed and switching topologies. *IEEE Transactions on Cybernetics*, 2019, **49**(12): 4117–4128
- 68 Li H Y, Wu Y, Chen M. Adaptive fault-tolerant tracking control for discrete-time multiagent systems via reinforcement learning algorithm. *IEEE Transactions on Cybernetics*, 2021, **51**(3): 1163–1174
- 69 Zhao W B, Liu H, Valavanis K P, Lewis F L. Fault-tolerant formation control for heterogeneous vehicles via reinforcement learning. *IEEE Transactions on Aerospace and Electronic Systems*, 2022, **58**(4): 2796–2806
- 70 Li T S, Bai W W, Liu Q, Long Y, Chen C L P. Distributed fault-tolerant containment control protocols for the discrete-time multiagent systems via reinforcement learning method. *IEEE Transactions on Neural Networks and Learning Systems*, 2023, **34**(8): 3979–3991
- 71 Liu D H, Mao Z H, Jiang B, Xu L. Simplified ADP-based distributed event-triggered fault-tolerant control of heterogeneous nonlinear multiagent systems with full-state constraints. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2024, **71**(8): 3820–3832
- 72 Zhang Y W, Zhao B, Liu D R, Zhang S C. Distributed fault

- tolerant consensus control of nonlinear multiagent systems via adaptive dynamic programming. *IEEE Transactions on Neural Networks and Learning Systems*, 2024, **35**(7): 9041–9053
- 73 Xu Y, Wu Z G. Data-based collaborative learning for multiagent systems under distributed denial-of-service attacks. *IEEE Transactions on Cognitive and Developmental Systems*, 2024, **16**(1): 75–85
- 74 Zhang L J, Chen Y. Distributed finite-time ADP-based optimal control for nonlinear multiagent systems. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2023, **70**(12): 4534–4538
- 75 Wang P, Yu C P, Lv M L, Cao J D. Adaptive fixed-time optimal formation control for uncertain nonlinear multiagent systems using reinforcement learning. *IEEE Transactions on Network Science and Engineering*, 2024, **11**(2): 1729–1743
- 76 Zhang Y, Chadli M, Xiang Z R. Prescribed-time formation control for a class of multiagent systems via fuzzy reinforcement learning. *IEEE Transactions on Fuzzy Systems*, 2023, **31**(12): 4195–4204
- 77 Peng Z N, Luo R, Hu J P, Shi K B, Ghosh B K. Distributed optimal tracking control of discrete-time multiagent systems via event-triggered reinforcement learning. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2022, **69**(9): 3689–3700
- 78 Xu Y, Sun J, Pan Y J, Wu Z G. Optimal tracking control of heterogeneous MASs using event-driven adaptive observer and reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2024, **35**(4): 5577–5587
- 79 Tan M J, Liu Z, Chen C L P, Zhang Y, Wu Z Z. Optimized adaptive consensus tracking control for uncertain nonlinear multiagent systems using a new event-triggered communication mechanism. *Information Sciences*, 2022, **605**: 301–316
- 80 Li H Y, Wu Y, Chen M, Lu R Q. Adaptive multigradient recursive reinforcement learning event-triggered tracking control for multiagent systems. *IEEE Transactions on Neural Networks and Learning Systems*, 2023, **34**(1): 144–156
- 81 Zhao H R, Shan J J, Peng L, Yu H N. Adaptive event-triggered bipartite formation for multiagent systems via reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2024, **35**(12): 17817–17828
- 82 Xiao W B, Zhou Q, Liu Y, Li H Y, Lu R Q. Distributed reinforcement learning containment control for multiple nonholonomic mobile robots. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2022, **69**(2): 896–907
- 83 Xiong C P, Ma Q, Guo J, Lewis F L. Data-based optimal synchronization of heterogeneous multiagent systems in graphical games via reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2024, **35**(11): 15984–15992
- 84 Zhang Q C, Zhao D B, Lewis F L. Model-free reinforcement learning for fully cooperative multi-agent graphical games. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN). Rio de Janeiro, Brazil: IEEE, 2018. 1–6
- 85 Li J N, Modares H, Chai T Y, Lewis F L, Xie L H. Off-policy reinforcement learning for synchronization in multiagent graphical games. *IEEE Transactions on Neural Networks and Learning Systems*, 2017, **28**(10): 2434–2445
- 86 Wang H, Li M. Model-free reinforcement learning for fully cooperative consensus problem of nonlinear multiagent systems. *IEEE Transactions on Neural Networks and Learning Systems*, 2022, **33**(4): 1482–1491
- 87 Ming Z Y, Zhang H G, Zhang J, Xie X P. A novel actor-critic identifier architecture for nonlinear multiagent systems with gradient descent method. *Automatica*, 2023, **155**: Article No. 111128
- 88 Liang Xing-Xing, Feng Yang-He, Ma Yang, Cheng Guang-Quan, Huang Jin-Cai, Wang Qi, et al. Deep multi-agent reinforcement learning: A survey. *Acta Automatica Sinica*, 2020, **46**(12): 2537–2557
(梁星星, 冯昞赫, 马扬, 程光权, 黄金才, 王琦, 等. 多 Agent 深度强化学习综述. 自动化学报, 2020, **46**(12): 2537–2557)
- 89 Bellman R. A Markovian decision process. *Journal of Mathematics and Mechanics*, 1957, **6**(5): 679–684
- 90 Howard R A. *Dynamic Programming and Markov Processes*. Cambridge: MIT Press, 1960.
- 91 Watkins C J C H, Dayan P. Q-learning. *Machine Learning*, 1992, **8**(3): 279–292
- 92 Rummery G A, Niranjan M. *On-Line Q-Learning Using Connectionist Systems*. Cambridge: University of Cambridge, 1994.
- 93 Sutton R S, McAllester D, Singh S, Mansour Y. Policy gradient methods for reinforcement learning with function approximation. In: Proceedings of the 13th International Conference on Neural Information Processing Systems. Denver, USA: MIT Press, 1999. 1057–1063
- 94 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: ACM, 2014. I-387–I-395
- 95 Luo B, Wu Z K, Zhou F, Wang B C. Human-in-the-loop reinforcement learning in continuous-action space. *IEEE Transactions on Neural Networks and Learning Systems*, 2024, **35**(11): 15735–15744
- 96 Shapley L S. Stochastic games. *Proceedings of the National Academy of Sciences of the United States of America*, 1953, **39**(10): 1095–1100
- 97 Hansen E A, Bernstein D S, Zilberstein S. Dynamic programming for partially observable stochastic games. In: Proceedings of the 19th AAAI Conference on Artificial Intelligence. San Jose, USA: AAAI, 2004. 709–715
- 98 Smallwood R D, Sondik E J. The optimal control of partially observable Markov processes over a finite horizon. *Operations Research*, 1973, **21**(5): 1071–1088
- 99 Tampuu A, Matiisen T, Kodelja D, Kuzovkin I, Korjus K, Aru J, et al. Multiagent cooperation and competition with deep reinforcement learning. *PLoS One*, 2017, **12**(4): Article No. e0172395
- 100 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
- 101 Chen Y F, Liu M, Everett M, How J P. Decentralized non-communicating multiagent collision avoidance with deep reinforcement learning. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Singapore: IEEE, 2017. 285–292
- 102 Sun Chang-Yin, Mu Chao-Xu. Important scientific problems of multi-agent deep reinforcement learning. *Acta Automatica Sinica*, 2020, **46**(7): 1301–1312
(孙长银, 穆朝絮. 多智能体深度强化学习的若干关键科学问题. 自动化学报, 2020, **46**(7): 1301–1312)
- 103 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, USA: ICLR, 2016.
- 104 Gupta J K, Egorov M, Kochenderfer M. Cooperative multi-agent control using deep reinforcement learning. In: Proceedings of the Conference on Autonomous Agents and Multiagent Systems. São Paulo, Brazil: Springer, 2017. 66–83
- 105 Kraemer L, Banerjee B. Multi-agent reinforcement learning as a rehearsal for decentralized planning. *Neurocomputing*, 2016, **190**: 82–94
- 106 Bhatnagar S, Sutton R S, Ghavamzadeh M, Lee M. Natural actor-critic algorithms. *Automatica*, 2009, **45**(11): 2471–2482
- 107 Degris T, White M, Sutton R S. Off-policy actor-critic. In: Pro-

- ceedings of the 29th International Conference on Machine Learning. Edinburgh, UK: ACM, 2012. 179–186
- 108 Sunehag P, Lever G, Gruslly A, Czarnecki W M, Zambaldi V, Jaderberg M, et al. Value-decomposition networks for cooperative multi-agent learning based on team reward. In: Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems. Stockholm, Sweden: ACM, 2018. 2085–2087
- 109 Lowe R, Wu Y I, Tamar A, Harb J, Pieter A, 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: 2017. 6382–6393
- 110 Gronauer S, Diepold K. Multi-agent deep reinforcement learning: A survey. *Artificial Intelligence Review*, 2022, **55**(2): 895–943
- 111 Iqbal S, Sha F. Actor-attention-critic for multi-agent reinforcement learning. In: Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR, 2019. 2961–2970
- 112 Yu C, Velu A, Vnitsky E, Gao J X, Wang Y, Bayen A, et al. The surprising effectiveness of PPO in cooperative multi-agent games. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 1787
- 113 Mnih V, Heess N, Graves A, Kavukcuoglu K. Recurrent models of visual attention. In: Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press, 2014. 2204–2212
- 114 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc., 2017. 6000–6010
- 115 Tan M. Multi-agent reinforcement learning: Independent versus cooperative agents. In: Proceedings of the 10th International Conference on Machine Learning. Amherst, USA: ACM, 1993. 330–337
- 116 Sen S, Sekaran M, Hale J. Learning to coordinate without sharing information. In: Proceedings of the 12th AAAI Conference on Artificial Intelligence. Seattle, USA: AAAI, 1994. 426–431
- 117 Matignon L, Laurent G J, Le Fort-Piat N. Independent reinforcement learners in cooperative Markov games: A survey regarding coordination problems. *The Knowledge Engineering Review*, 2012, **27**(1): 1–31
- 118 Foerster J, Nardelli N, Farquhar G, Afouras T, Torr P H S, Kohli P, et al. Stabilising experience replay for deep multi-agent reinforcement learning. In: Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: ACM, 2017. 1146–1155
- 119 Raileanu R, Denton E, Szlam A, Fergus R. Modeling others using oneself in multi-agent reinforcement learning. In: Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: PMLR, 2018. 4254–4263
- 120 Kaelbling L P, Littman M L, Cassandra A R. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 1998, **101**(1–2): 99–134
- 121 Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Computation*, 1997, **9**(8): 1735–1780
- 122 Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: ACL, 2014. 1724–1734
- 123 Hausknecht M, Stone P. Deep recurrent Q-learning for partially observable MDPs. In: Proceedings of the AAAI Fall Symposium. Arlington, USA: AAAI, 2015. 29–37
- 124 Matignon L, Laurent G J, Le Fort-Piat N. Hysteretic Q-learning: An algorithm for decentralized reinforcement learning in cooperative multi-agent teams. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. San Diego, USA: IEEE, 2007. 64–69
- 125 Omidshafiei S, Pazis J, Amato C, How J P, Vian J. Deep decentralized multi-task multi-agent reinforcement learning under partial observability. In: Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: ACM, 2017. 2681–2690
- 126 Foerster J N, Assael Y M, de Freitas N, Whiteson S. Learning to communicate to solve riddles with deep distributed recurrent Q-networks. arXiv preprint arXiv: 1602.02672, 2016.
- 127 Sukhbaatar S, Fergus R, Fergus R. Learning multiagent communication with backpropagation. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: Curran Associates Inc., 2016. 2252–2260
- 128 Singh A, Jain T, Sukhbaatar S. Individualized controlled continuous communication model for multiagent cooperative and competitive tasks. In: Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: OpenReview, 2019. 1–16
- 129 Chen J D, Lan T, Joe-Wong C. RGMComm: Return gap minimization via discrete communications in multi-agent reinforcement learning. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI, 2024. 17327–17336
- 130 Sheng J J, Wang X F, Jin B, Yan J C, Li W H, Chang T H, et al. Learning structured communication for multi-agent reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 2022, **36**(2): Article No. 50
- 131 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
- 132 Peng P, Wen Y, Yang Y D, Yuan Q, Tang Z K, Long H T, et al. Multiagent bidirectionally-coordinated nets: Emergence of human-level coordination in learning to play starcraft combat games. arXiv preprint arXiv: 1703.10069, 2017.
- 133 Wu Jun-Feng, Wang Wen, Wang Liang, Tao Xian-Ping, Hu Hao, Wu Hai-Jun. Multi-agent reinforcement learning with two step intention sharing. *Chinese Journal of Computers*, 2023, **46**(9): 1820–1837
(吴俊锋, 王文, 汪亮, 陶先平, 胡昊, 吴海军. 基于两阶段意图共享的多智能体强化学习方法. 计算机学报, 2023, **46**(9): 1820–1837)
- 134 Mao H Y, Zhang Z C, Xiao Z, Gong Z B, Ni Y. Learning agent communication under limited bandwidth by message pruning. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI, 2020. 5142–5149
- 135 Kim D, Moon S, Hostallero D, Kang W J, Lee T, Son K, et al. Learning to schedule communication in multi-agent reinforcement learning. In: Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: OpenReview, 2019. 1–11
- 136 Das A, Gervet T, Romoff J, Batra D, Parikh D, Rabbat M, et al. TarMAC: Targeted multi-agent communication. In: Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: ICML, 2019. 1538–1546
- 137 Niu Y R, Paleja R, Gombolay M. Multi-agent graph-attention communication and teaming. In: Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems. Virtual Event: ACM, 2021. 964–973
- 138 Lhaksmana K M, Murakami Y, Ishida T. Role-based modeling for designing agent behavior in self-organizing multi-agent sys-

- tems. *International Journal of Software Engineering and Knowledge Engineering*, 2018, **28**(1): 79–96
- 139 Wang T H, Dong H, Lesser V, Zhang C J. ROMA: Multi-agent reinforcement learning with emergent roles. In: Proceedings of the 37th International Conference on Machine Learning. Vienna, Australia: ACM, 2020. Article No. 916
- 140 Wang T H, Gupta T, Mahajan A, Peng B, Whiteson S, Zhang C J. RODE: Learning roles to decompose multi-agent tasks. In: Proceedings of the 9th International Conference on Learning Representations. Vienna, Australia: OpenReview, 2021. 1–24
- 141 Hu Z C, Zhang Z Z, Li H X, Chen C L, Ding H Y, Wang Z. Attention-guided contrastive role representations for multi-agent reinforcement learning. In: Proceedings of the 12th International Conference on Learning Representations. Vienna, Australia: OpenReview, 2024. 1–23
- 142 Zambaldi V, Raposo D, Santoro A, Bapst V, Li Y J, Babuschkin I, et al. Relational deep reinforcement learning. arXiv preprint arXiv: 1806.01830, 2018.
- 143 Jiang H, Liu Y T, Li S Z, Zhang J Y, Xu X H, Liu D H. Diverse effective relationship exploration for cooperative multi-agent reinforcement learning. In: Proceedings of the 31st ACM International Conference on Information and Knowledge Management. Atlanta, USA: ACM, 2022. 842–851
- 144 Wang W X, Yang T P, Liu Y, Hao J Y, Hao X T, Hu Y J. Action semantics network: Considering the effects of actions in multiagent systems. In: Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia: OpenReview, 2020. 1–18
- 145 Ackermann J, Gabler V, Osa T, Sugiyama M. Reducing overestimation bias in multi-agent domains using double centralized critics. arXiv preprint arXiv: 1910.01465, 2019.
- 146 van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double Q-learning. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix, USA: AAAI, 2016. 2094–2100
- 147 Pan L, Rashid T, Peng B, Huang L B, Whiteson S. Regularized softmax deep multi-agent Q-learning. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual Event: Curran Associates Inc., 2021. Article No. 105
- 148 Liu J R, Zhong Y F, Hu S Y, Fu H B, Fu Q, Chang X J, et al. Maximum entropy heterogeneous-agent reinforcement learning. In: Proceedings of the 12th International Conference on Learning Representations. Vienna, Australia: OpenReview, 2024. 1–12
- 149 Na H, Seo Y, Moon I C. Efficient episodic memory utilization of cooperative multi-agent reinforcement learning. In: Proceedings of the 12th International Conference on Learning Representations. Vienna, Australia: OpenReview, 2024. 1–13
- 150 Mahajan A, Rashid T, Samvelyan M, Whiteson S. MAVEN: Multi-agent variational exploration. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2019. Article No. 684
- 151 Liu I J, Jain U, Yeh R A, Schwing A G. Cooperative exploration for multi-agent deep reinforcement learning. In: Proceedings of the 38th International Conference on Machine Learning. Virtual Event: ICML, 2021. 6826–6836
- 152 Chen Z H, Luo B, Hu T M, Xu X D. LJIR: Learning joint-action intrinsic reward in cooperative multi-agent reinforcement learning. *Neural Networks*, 2023, **167**: 450–459
- 153 Hao J Y, Hao X T, Mao H Y, Wang W X, Yang Y D, Li D, et al. Boosting multiagent reinforcement learning via permutation invariant and permutation equivariant networks. In: Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: OpenReview, 2023. 1–12
- 154 Yang Y D, Luo R, Li M N, Zhou M, Zhang W N, Wang J. Mean field multi-agent reinforcement learning. In: Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: ICML, 2018. 5567–5576
- 155 Subramanian S G, Poupart P, Taylor M E, Hegde N. Multi type mean field reinforcement learning. In: Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems. Auckland, New Zealand: ACM, 2020. 411–419
- 156 Mondal W U, Agarwal M, Aggarwal V, Ukkusuri S V. On the approximation of cooperative heterogeneous multi-agent reinforcement learning (MARL) using mean field control (MFC). *The Journal of Machine Learning Research*, 2022, **23**(1): Article No. 129
- 157 Chang Y H, Ho T, Kaelbling L P. All learning is local: Multi-agent learning in global reward games. In: Proceedings of the 17th International Conference on Neural Information Processing Systems. Whistler, Canada: MIT Press, 2003. 807–814
- 158 Rashid T, Samvelyan M, de Witt C S, Farquhar G, Foerster J N, Whiteson S. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In: Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: ICML, 2018. 4292–4301
- 159 Zhou M, Liu Z Y, Sui P W, Li Y X, Chung Y Y. Learning implicit credit assignment for cooperative multi-agent reinforcement learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. Article No. 994
- 160 Rashid T, Farquhar G, Peng B, Whiteson S. Weighted QMIX: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. Article No. 855
- 161 Yang Y D, Hao J Y, Liao B, Shao K, Chen G Y, Liu W L, et al. Qatten: A general framework for cooperative multiagent reinforcement learning. arXiv preprint arXiv: 2002.03939, 2020.
- 162 Son K, Kim D, Kang W J, Hostallero D, Yi Y. QTRAN: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In: Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: ICML, 2019. 5887–5896
- 163 Foerster J, Farquhar G, Afouras T, Nardelli N, Whiteson S. Counterfactual multi-agent policy gradients. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI, 2018. 2974–2982
- 164 Wolpert D H, Tumer K. Optimal payoff functions for members of collectives. *Advances in Complex Systems*, 2001, **4**(2–3): 265–279
- 165 Wang J H, Zhang Y, Kim T K, Gu Y J. Shapley Q-value: A local reward approach to solve global reward games. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI, 2020. 7285–7292
- 166 Shapley L. A value for N-person games. *Contributions to the Theory of Games II*. Princeton: Princeton University Press, 1953. 307–317
- 167 Li J H, Kuang K, Wang B X, Liu F R, Chen L, Wu F, et al. Shapley counterfactual credits for multi-agent reinforcement learning. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Virtual Event: ACM, 2021. 934–942
- 168 Xu Cheng, Yin Nan, Duan Shi-Hong, He Hao, Wang Ran. Reward-filtering-based credit assignment for multi-agent deep reinforcement learning. *Chinese Journal of Computers*, 2022, **45**(11): 2306–2320
(徐诚, 殷楠, 段世红, 何昊, 王然. 基于奖励滤波信用分配的多智能体深度强化学习算法. *计算机学报*, 2022, **45**(11): 2306–2320)
- 169 Chen S R, Zhang Z W, Yang Y D, Du Y L. STAS: Spatial-temporal return decomposition for solving sparse rewards prob-

- lems in multi-agent reinforcement learning. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI, 2024. 17337–17345
- 170 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: ACM, 1994. 157–163
- 171 Zhang K Q, Yang Z R, Liu H, Zhang T, Basar T. Finite-sample analysis for decentralized batch multiagent reinforcement learning with networked agents. *IEEE Transactions on Automatic Control*, 2021, **66**(12): 5925–5940
- 172 Fan J Q, Wang Z R, Xie Y C, Yang Z R. A theoretical analysis of deep Q-learning. In: Proceedings of the 2nd Annual Conference on Learning for Dynamics and Control. Berkeley, USA: PMLR, 2020. 486–489
- 173 Heinrich J, Lanctot M, Silver D. Fictitious self-play in extensive-form games. In: Proceedings of the 32nd International Conference on Machine Learning. Lille, France: ACM, 2015. 805–813
- 174 Berger U. Brown's original fictitious play. *Journal of Economic Theory*, 2007, **135**(1): 572–578
- 175 Heinrich J, Silver D. Deep reinforcement learning from self-play in imperfect-information games. arXiv preprint arXiv: 1603.01121, 2016.
- 176 Zhang L, Chen Y X, Wang W, Han Z L, Li S J, Pan Z J, et al. A Monte Carlo neural fictitious self-play approach to approximate Nash equilibrium in imperfect-information dynamic games. *Frontiers of Computer Science*, 2021, **15**(5): Article No. 155334
- 177 Lanctot M, Zambaldi V, Gruslys A, Lazaridou A, Tuyls K, Pérolat J, et al. A unified game-theoretic approach to multiagent reinforcement learning. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc., 2017. 4193–4206
- 178 McAleer S, Lanier J, Fox R, Baldi P. Pipeline PSRO: A scalable approach for finding approximate Nash equilibria in large games. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. Article No. 1699
- 179 Muller P, Omidshafiei S, Rowland M, Tuyls K, Pérolat J, Liu S Q, et al. A generalized training approach for multiagent learning. In: Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia: ICLR, 2020.
- 180 Xu Hao-Tian, Qin Long, Zeng Jun-Jie, Hu Yue, Zhang Qi. Research progress of opponent modeling based on deep reinforcement learning. *Journal of System Simulation*, 2023, **35**(4): 671–694
(徐浩添, 秦龙, 曾俊杰, 胡越, 张琪. 基于深度强化学习的对手建模方法研究综述. *系统仿真学报*, 2023, **35**(4): 671–694)
- 181 He H, Boyd-Graber J, Kwok K, Daumé III H. Opponent modeling in deep reinforcement learning. In: Proceedings of the 33rd International Conference on Machine Learning. New York, USA: ACM, 2016. 1804–1813
- 182 Everett R, Roberts S J. Learning against non-stationary agents with opponent modelling and deep reinforcement learning. In: Proceedings of the AAAI Spring Symposium. Palo Alto, USA: AAAI, 2018. 621–626
- 183 Foerster J, Chen R Y, Al-Shedivat M, Whiteson S, Abbeel P, Mordatch I. Learning with opponent-learning awareness. In: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems. Stockholm, Sweden: ACM, 2018. 122–130
- 184 Long P X, Fan T X, Liao X Y, Liu W X, Zhang H, Pan J. Towards optimally decentralized multi-robot collision avoidance via deep reinforcement learning. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Brisbane, Australia: IEEE, 2018. 6252–6259
- 185 Willemsen D, Coppola M, de Croon G C H E. MAMBPO: Sample-efficient multi-robot reinforcement learning using learned world models. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Prague, Czech Republic: IEEE, 2021. 5635–5640
- 186 Yue L F, Lv M L, Yan M D, Zhao X R, Wu A, Li L Y, et al. Improving cooperative multi-target tracking control for UAV swarm using multi-agent reinforcement learning. In: Proceedings of the 9th International Conference on Control, Automation and Robotics (ICCAR). Beijing, China: IEEE, 2023. 179–186
- 187 Xue Y T, Chen W S. Multi-agent deep reinforcement learning for UAVs navigation in unknown complex environment. *IEEE Transactions on Intelligent Vehicles*, 2024, **9**(1): 2290–2303
- 188 Mou Z Y, Zhang Y, Gao F F, Wang H G, Zhang T, Han Z. Deep reinforcement learning based three-dimensional area coverage with UAV swarm. *IEEE Journal on Selected Areas in Communications*, 2021, **39**(10): 3160–3176
- 189 Hou Y K, Zhao J, Zhang R Q, Cheng X, Yang L Q. UAV swarm cooperative target search: A multi-agent reinforcement learning approach. *IEEE Transactions on Intelligent Vehicles*, 2024, **9**(1): 568–578
- 190 Cui J J, Liu Y W, Nallanathan A. Multi-agent reinforcement learning-based resource allocation for UAV networks. *IEEE Transactions on Wireless Communications*, 2020, **19**(2): 729–743
- 191 Wang Z Y, Gombolay M. Learning scheduling policies for multi-robot coordination with graph attention networks. *IEEE Robotics and Automation Letters*, 2020, **5**(3): 4509–4516
- 192 Johnson D, Chen G, Lu Y Q. Multi-agent reinforcement learning for real-time dynamic production scheduling in a robot assembly cell. *IEEE Robotics and Automation Letters*, 2022, **7**(3): 7684–7691
- 193 Paul S, Ghassemi P, Chowdhury S. Learning scalable policies over graphs for multi-robot task allocation using capsule attention networks. In: Proceedings of the International Conference on Robotics and Automation (ICRA). Philadelphia, USA: IEEE, 2022. 8815–8822
- 194 Shalev-Shwartz S, Shammah S, Shashua A. Safe, multi-agent, reinforcement learning for autonomous driving. arXiv preprint arXiv: 1610.03295, 2016.
- 195 Yu C, Wang X, Xu X, Zhang M J, Ge H W, Ren J K, et al. Distributed multiagent coordinated learning for autonomous driving in highways based on dynamic coordination graphs. *IEEE Transactions on Intelligent Transportation Systems*, 2020, **21**(2): 735–748
- 196 Liu B, Ding Z T, Lv C. Platoon control of connected autonomous vehicles: A distributed reinforcement learning method by consensus. *IFAC-PapersOnLine*, 2020, **53**(2): 15241–15246
- 197 Liang Z X, Cao J N, Jiang S, Saxena D, Xu H F. Hierarchical reinforcement learning with opponent modeling for distributed multi-agent cooperation. In: Proceedings of the 42nd IEEE International Conference on Distributed Computing Systems (ICDCS). Bologna, Italy: IEEE, 2022. 884–894
- 198 Candela E, Parada L, Marques L, Georgescu T A, Demiris Y, Angeloudis P. Transferring multi-agent reinforcement learning policies for autonomous driving using sim-to-real. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Kyoto, Japan: IEEE, 2022. 8814–8820
- 199 Chu T S, Wang J, Codecà L, Li Z J. Multi-agent deep reinforcement learning for large-scale traffic signal control. *IEEE Transactions on Intelligent Transportation Systems*, 2020, **21**(3): 1086–1095
- 200 Jiang S, Huang Y F, Jafari M, Jalayer M. A distributed multi-agent reinforcement learning with graph decomposition approach for large-scale adaptive traffic signal control. *IEEE Transactions on Intelligent Transportation Systems*, 2022,

- 23(9): 14689–14701
- 201 Wang X Q, Ke L J, Qiao Z M, Chai X H. Large-scale traffic signal control using a novel multiagent reinforcement learning. *IEEE Transactions on Cybernetics*, 2021, **51**(1): 174–187
- 202 Wang K, Mu C X. Learning-based control with decentralized dynamic event-triggering for vehicle systems. *IEEE Transactions on Industrial Informatics*, 2023, **19**(3): 2629–2639
- 203 Xu Y, Wu Z G, Pan Y J. Perceptual interaction-based path tracking control of autonomous vehicles under DoS attacks: A reinforcement learning approach. *IEEE Transactions on Vehicular Technology*, 2023, **72**(11): 14028–14039
- 204 Shen G Q, Lei L, Zhang X T, Li Z L, Cai S S, Zhang L J. Multi-UAV cooperative search based on reinforcement learning with a digital twin driven training framework. *IEEE Transactions on Vehicular Technology*, 2023, **72**(7): 8354–8368
- 205 Cheng M, Liu H, Wen G H, Lv J H, Lewis F L. Data-driven time-varying formation-containment control for a heterogeneous air-ground vehicle team subject to active leaders and switching topologies. *Automatica*, 2023, **153**: Article No. 111029
- 206 Zhao W B, Liu H, Wan Y, Lin Z L. Data-driven formation control for multiple heterogeneous vehicles in air-ground coordination. *IEEE Transactions on Control of Network Systems*, 2022, **9**(4): 1851–1862
- 207 Zhao J, Yang C, Wang W D, Xu B, Li Y, Yang L Q, et al. A game-learning-based smooth path planning strategy for intelligent air-ground vehicle considering mode switching. *IEEE Transactions on Transportation Electrification*, 2022, **8**(3): 3349–3366
- 208 Song W T, Tong S C. Fuzzy optimal tracking control for nonlinear underactuated unmanned surface vehicles. *Ocean Engineering*, 2023, **287**: Article No. 115700
- 209 Chen L, Dong C, He S D, Dai S L. Adaptive optimal formation control for unmanned surface vehicles with guaranteed performance using actor-critic learning architecture. *International Journal of Robust and Nonlinear Control*, 2023, **33**(8): 4504–4522
- 210 Bai W W, Zhang W J, Cao L, Liu Q. Adaptive control for multi-agent systems with actuator fault via reinforcement learning and its application on multi-unmanned surface vehicle. *Ocean Engineering*, 2023, **280**: Article No. 114545
- 211 Chen H Z, Yan H C, Wang Y Y, Xie S R, Zhang D. Reinforcement learning-based close formation control for underactuated surface vehicle with prescribed performance and time-varying state constraints. *Ocean Engineering*, 2022, **256**: Article No. 111361
- 212 Weng P J, Tian X H, Liu H T, Mai Q. Distributed edge-based event-triggered optimal formation control for air-sea heterogeneous multiagent systems. *Ocean Engineering*, 2023, **288**: Article No. 116066
- 213 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
- 214 Xu X, Jia Y W, Xu Y, Xu Z, Chai S J, Lai C S. A multi-agent reinforcement learning-based data-driven method for home energy management. *IEEE Transactions on Smart Grid*, 2020, **11**(4): 3201–3211
- 215 Ahrarinouiri M, Rastegar M, Seifi A R. Multiagent reinforcement learning for energy management in residential buildings. *IEEE Transactions on Industrial Informatics*, 2021, **17**(1): 659–666
- 216 Zhang Y, Yang Q Y, An D, Li D H, Wu Z Z. Multistep multiagent reinforcement learning for optimal energy schedule strategy of charging stations in smart grid. *IEEE Transactions on Cybernetics*, 2023, **53**(7): 4292–4305
- 217 Zhao X Y, Wu C. Large-scale machine learning cluster scheduling via multi-agent graph reinforcement learning. *IEEE Transactions on Network and Service Management*, 2022, **19**(4): 4962–4974
- 218 Yu T, Huang J, Chang Q. Optimizing task scheduling in human-robot collaboration with deep multi-agent reinforcement learning. *Journal of Manufacturing Systems*, 2021, **60**: 487–499
- 219 Jing X, Yao X F, Liu M, Zhou J J. Multi-agent reinforcement learning based on graph convolutional network for flexible job shop scheduling. *Journal of Intelligent Manufacturing*, 2024, **35**(1): 75–93
- 220 Kuang Zhu-Fang, Chen Qing-Lin, Li Lin-Feng, Deng Xiao-Heng, Chen Zhi-Gang. Multi-user edge computing task offloading scheduling and resource allocation based on deep reinforcement learning. *Chinese Journal of Computers*, 2022, **45**(4): 812–824
(庠祝芳, 陈清林, 李林峰, 邓晓衡, 陈志刚. 基于深度强化学习的多用户边缘计算任务卸载调度与资源分配算法. 计算机学报, 2022, **45**(4): 812–824)
- 221 Adibi M, van der Woude J. Secondary frequency control of microgrids: An online reinforcement learning approach. *IEEE Transactions on Automatic Control*, 2022, **67**(9): 4824–4831
- 222 Liu Y L, Qie T H, Yu Y, Wang Y X, Chau T K, Zhang X N. A novel integral reinforcement learning-based H_∞ control strategy for proton exchange membrane fuel cell in DC Microgrids. *IEEE Transactions on Smart Grid*, 2023, **14**(3): 1668–1681
- 223 Zhang H F, Yue D, Dou C X, Xie X P, Li K, Hancke G P. Resilient optimal defensive strategy of TSK fuzzy-model-based microgrids' system via a novel reinforcement learning approach. *IEEE Transactions on Neural Networks and Learning Systems*, 2023, **34**(4): 1921–1931
- 224 Duan J J, Yi Z H, Shi D, Lin C, Lu X, Wang Z W. Reinforcement-learning-based optimal control of hybrid energy storage systems in hybrid AC-DC microgrids. *IEEE Transactions on Industrial Informatics*, 2019, **15**(9): 5355–5364
- 225 Dong X, Zhang H G, Xie X P, Ming Z Y. Data-driven distributed H_∞ current sharing consensus optimal control of DC microgrids via reinforcement learning. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2024, **71**(6): 2824–2834
- 226 Fang H Y, Zhang M G, He S P, Luan X L, Liu F, Ding Z T. Solving the zero-sum control problem for tidal turbine system: An online reinforcement learning approach. *IEEE Transactions on Cybernetics*, 2023, **53**(12): 7635–7647
- 227 Dong H Y, Zhao X W. Wind-farm power tracking via preview-based robust reinforcement learning. *IEEE Transactions on Industrial Informatics*, 2022, **18**(3): 1706–1715
- 228 Xie J J, Dong H Y, Zhao X W, Lin S Y. Wind turbine fault-tolerant control via incremental model-based reinforcement learning. *IEEE Transactions on Automation Science and Engineering*, 2025, **22**: 1958–1969
- 229 Park J S, O'Brien J, Cai C J, Morris M R, Liang P, Bernstein M S. Generative agents: Interactive simulacra of human behavior. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. San Francisco, USA: ACM, 2023. Article No. 2
- 230 Wang H Q, Chen J H, Huang W S, Ben Q W, Wang T, Mi B Y, et al. GRUtopia: Dream general robots in a city at scale. arXiv preprint arXiv: 2407.10943, 2024.
- 231 Wang Han, Yu Yang, Jiang Yuan. Review of the progress of communication-based multi-agent reinforcement learning. *Science Sinica Information*, 2022, **52**(5): 742–764
(王涵, 俞扬, 姜远. 基于通信的多智能体强化学习进展综述. 中国科学: 信息科学, 2022, **52**(5): 742–764)
- 232 Hu T M, Luo B. PA2D-MORL: Pareto ascent directional decomposition based multi-objective reinforcement learning. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI, 2024. 12547–12555
- 233 Xu M, Song Y H, Wang J Y, Qiao M L, Huo L Y, Wang Z L.

- Predicting head movement in panoramic video: A deep reinforcement learning approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, **41**(11): 2693–2708
- 234 Skalse J, Hammond L, Griffin C, Abate A. Lexicographic multi-objective reinforcement learning. In: Proceedings of the 31st International Joint Conference on Artificial Intelligence. Vienna, Austria: IJCAI, 2022. 3430–3436
- 235 Hu T M, Luo B, Yang C H, Huang T W. MO-MIX: Multi-objective multi-agent cooperative decision-making with deep reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, **45**(10): 12098–12112
- 236 Wang Xue-Song, Wang Rong-Rong, Cheng Yu-Hu. Safe reinforcement learning: A survey. *Acta Automatica Sinica*, 2023, **49**(9): 1813–1835
(王雪松, 王荣荣, 程玉虎. 安全强化学习综述. 自动化学报, 2023, **49**(9): 1813–1835)



罗彪 中南大学自动化学院教授。主要研究方向为智能控制, 强化学习, 深度学习和自主决策。本文通信作者。E-mail: biao.luo@hotmail.com

(LUO Biao Professor at the School of Automation, Central South University. His research interest covers intelligent control, reinforcement learning, deep learning, and decision-making. Corresponding author of this paper.)

(LUO Biao Professor at the School of Automation, Central South University. His research interest covers intelligent control, reinforcement learning, deep learning, and decision-making. Corresponding author of this paper.)



胡天萌 中南大学自动化学院硕士研究生。主要研究方向为强化学习, 多智能体强化学习和多目标决策。E-mail: tianmeng0824@163.com

(HU Tian-Meng Master student at the School of Automation, Central South University. His research interest covers reinforcement learning, multi-agent reinforcement learning, and multi-objective decision-making.)

(HU Tian-Meng Master student at the School of Automation, Central South University. His research interest covers reinforcement learning, multi-agent reinforcement learning, and multi-objective decision-making.)



周育豪 中南大学自动化学院博士研究生。主要研究方向为多智能体系统, 强化学习控制和自适应控制。E-mail: yuhao980603@163.com

(ZHOU Yu-Hao Ph.D. candidate at the School of Automation, Central South University. His research

interest covers multi-agent systems, reinforcement learning control, and adaptive control.)



黄廷文 中南大学自动化学院教授。主要研究方向为神经网络, 混沌动力学系统, 复杂网络和智能电网。E-mail: tingwen.huang@csu.edu.cn

(HUANG Ting-Wen Professor at the School of Automation, Central South University. His research interest covers neural networks, chaotic dynamical systems, complex networks, and smart grid.)

(HUANG Ting-Wen Professor at the School of Automation, Central South University. His research interest covers neural networks, chaotic dynamical systems, complex networks, and smart grid.)



阳春华 中南大学自动化学院教授。主要研究方向为复杂工业过程建模与优化控制, 智能自动化系统与装置。E-mail: yqh@csu.edu.cn

(YANG Chun-Hua Professor at the School of Automation, Central South University. Her research interest covers modeling and optimal control of complex industrial process, intelligent automation systems and devices.)

(YANG Chun-Hua Professor at the School of Automation, Central South University. Her research interest covers modeling and optimal control of complex industrial process, intelligent automation systems and devices.)



桂卫华 中国工程院院士, 中南大学自动化学院教授。主要研究方向为复杂工业过程建模, 优化与控制应用和故障诊断与分布式鲁棒控制。E-mail: gwh@csu.edu.cn

(GUI Wei-Hua Academician of Chinese Academy of Engineering, and professor at the School of Automation, Central South University. His research interest covers complex industrial process modeling, optimization and control applications, and fault diagnosis and distributed robust control.)

(GUI Wei-Hua Academician of Chinese Academy of Engineering, and professor at the School of Automation, Central South University. His research interest covers complex industrial process modeling, optimization and control applications, and fault diagnosis and distributed robust control.)