引用格式:陈人龙,陈嘉礼,李善琦,等. 多智能体强化学习方法综述[J]. 信息对抗技术,2024, 3(1):18-32.[CHEN Renlong, CHEN Jiali, LI Shanqi, et al. A survey of multi-agent reinforcement learning methods[J]. Information Countermeasure Technology, 2024, 3(1):18-32.(in Chinese)]

多智能体强化学习方法综述

陈人龙[1,2],陈嘉礼[1,2],李善琦[1,2],谭 营[1,2,3,4*]

(1. 北京大学机器感知与智能教育部重点实验室,北京 100871; 2. 北京大学智能学院,北京 100871;
3. 北京大学人工智能研究院,北京 100871; 4. 北京大学跨媒体通用人工智能全国重点实验室,北京 100871)

摘 要 在自动驾驶、团队配合游戏等现实场景的序列决策问题中,多智能体强化学习表现出了优秀的潜力。然而,多智能体强化学习面临着维度灾难、不稳定性、多目标性和部分可观测性等挑战。为此,概述了多智能体强化学习的概念与方法,并整理了当前研究的主要趋势和研究方向。研究趋势包括 CTDE 范式、具有循环神经单元的智能体和训练技巧。主要研究方向涵盖混合型学习方法、协同与竞争学习、通信与知识共享、适应性与鲁棒性、分层与模块化学习、基于博弈论的方法以及可解释性。未来的研究方向包括解决维度灾难问题、求解大型组合优化问题和分析多智能体强化学习算法的全局收敛性。这些研究方向将推动多智能体强化学习在实际应用中取得更大的突破。

关键词 多智能体强化学习;强化学习;多智能体系统;群体协同;维度灾难

中图分类号 TN 915 **文章编号** 2097-163X(2024)01-0018-15

文献标志码 A **DOI** 10.12399/j.issn.2097-163x.2024.01.003

A survey of multi-agent reinforcement learning methods

CHEN Renlong[1,2], CHEN Jiali[1,2], LI Shanqi[1,2], TAN Ying[1,2,3,4*]

(1. Key Laboratory of Machine Perceptron (MOE), Peking University, Beijing 100871, China;
2. School of Intelligence Science and Technology, Peking University, Beijing 100871, China;
3. Institute for Artificial Intelligence, Peking University, Beijing 100871, China;
4. National Key Laboratory of General Artificial Intelligence, Peking University, Beijing 100871, China)

Abstract In real-world scenarios such as autonomous driving and team-based cooperative games, multi-agent reinforcement learning has demonstrated significant potential in tackling sequential decision-making problems. However, it also encounters challenges including the curse of dimensionality, instability, multi-objectivity, and partial observability. This article offers an overview of the concepts and methods employed in multi-agent reinforcement learning, providing a summary of the prevailing trends and research directions in the current studies. The identified research trends comprise the CTDE paradigm, agents equipped with recurrent neural units, and various training techniques. The primary research directions encompass hybrid learning methods, cooperative and competitive learning, communication and knowledge sharing, adaptability and robustness, hierarchical and modular learning, game theoretic approaches, and interpretability. Looking ahead, future research directions entail

收稿日期:2023-02-22 **修回日期**:2023-05-04

通信作者:谭营,E-mail:ytan@pku.edu.cn

基金项目:国家重点研发计划项目(2018AAA0102301);国家自然科学基金资助项目(62250037, 62276008, 62076010)

addressing the curse of dimensionality, solving large-scale combinatorial optimization problems, and conducting analyses on the global convergence of multi-agent reinforcement learning algorithms. Pursuing these research directions will significantly contribute to further breakthroughs in the practical application of multi-agent reinforcement learning.

Keywords　multi-agent reinforcement learning; reinforcement learning; multi-agent system; swarm collaboration; curse dimensionality

0　引言

多智能体强化学习(multi-agent reinforcement learning, MARL)是近年来发展最快、最为热点的强化学习研究的分支之一。强化学习(reinforcement learning, RL)已经广泛应用于工业制造、机器人控制[1]、游戏博弈[2]等领域。在序列决策问题中，强化学习体现出了极高的有效性，特别是随着用于函数拟合的深度神经网络的发展，深度强化学习算法在棋类博弈[3]、实时战略游戏[4]、非完美信息博弈[5]和自动驾驶[6]等方面取得了极大的进步。强化学习的基本思想是通过最大化智能体(agent)从环境中获得的累计奖赏值，以学习到完成目标的最优策略。然而目前大多数在实际应用中取得优秀效果的强化学习算法通常集中在单智能体(single-agent)领域。多智能体强化学习则着重解决另一类多智能体(multi-agent)在同一环境中进行交互的任务。多智能体任务因其交互的复杂性和与现实任务贴合的紧密性，近年来受到了越来越多的关注。随着多个智能体的引入，智能体间的交互行为也产生了不同模式，这给算法设计提出了更高的要求。多智能体强化学习还面临着新的挑战，包括组合动作空间随智能体数目指数增大的维度灾难问题、智能体动作对其他智能体造成的不稳定性问题、智能体之间目标的差异性问题以及单个智能体的部分可观测性问题，等等。这些挑战也吸引着越来越多的研究者加入到对多智能体强化学习的研究中。本文对多智能体强化学习方法进行了综述，列举了主要挑战及相关研究工作，总结研究趋势并且指出了未来可能的研究方向。

1　强化学习理论

强化学习的基本交互过程如图 1 所示，即智能体与环境交互逻辑。在时刻 t，环境给出当前时刻的状态 s_t，智能体获取状态 s_t 或该状态的一个可观测分量 o_t，并根据这个输入得到当前时刻的动作 a_t，环境执行智能体给出的动作 a_t，并得到当前动作的奖励值 r_t 以及下一时刻的环境状态 s_{t+1}。因此，强化学习过程包含了一个基本的假设，即学习的目标可以被较好地解释为最大化一个特定的可累积的奖励值。

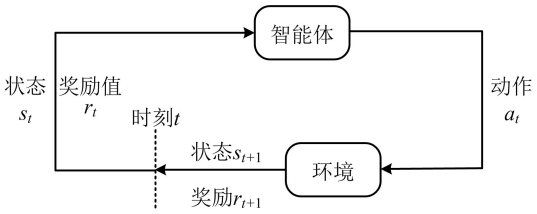

图 1　智能体与环境交互逻辑

Fig. 1　Interaction logic between angents and environment

强化学习问题可以通过一个马尔可夫决策过程(Markov decision process, MDP)[7]来建模。整个 MDP 可以被描述为一个五元组，即$\langle S, A, P, R, \gamma\rangle$。其中，$S$ 为所有环境状态的集合，$s_t \in S$ 为 t 时刻的环境状态；A 为所有可执行动作的集合，$a_t \in A$ 为 t 时刻智能体执行的动作；P 表示对所有动作产生状态转移的概率；$r \in R$ 表示环境的奖励；$\gamma \in [0,1)$ 为折扣系数，用来平衡当前和未来的奖励权重。t 时刻智能体与环境交互的操作可被归纳为：智能体接收并处理环境信息 s_t 以及 r_t，产生动作 a_t；环境接收动作 a_t，产生新状态 s_{t+1} 以及当前时刻的动作奖励。

在 MDP 中，一个状态的期望奖励(即从该状态开始直至 MDP 结束产生的累积奖励的期望)被称为该状态的价值。用函数形式进行表达，则可以记为：

$$V(s) = E\left[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots \mid s_t = s\right] \tag{1}$$

由价值函数的定义可以得到其递推形式：

$$V(s) = E\left[r_t + \gamma V(s_{t+1}) \mid s_t = s\right] \tag{2}$$

从而得到价值函数的贝尔曼方程(Bellman

equation)：

$$V(s)=r(s)+\gamma\sum_{s'\in S}p(s'\mid s)V(s') \tag{3}$$

由于动作的存在，需要额外定义一个动作价值函数(action-value function)$Q^{\pi}(s,a)$，以表征对当前状态 s 执行动作 a 得到的期望累积奖励。$Q^{\pi}(s,a)$定义如下：

$$\begin{aligned}Q^{\pi}(s_t,a_t)&=E[r_t+\lambda r_{t+1}+\lambda^2 r_{t+2}+\cdots\mid s_t,a_t]\\&=E[r_t+Q^{\pi}(s_{t+1},a_{t+1})\mid s_t,a_t]\end{aligned} \tag{4}$$

求解强化学习问题，通常有基于值函数的强化学习方法 (value-based RL)、策略梯度的强化学习方法(policy gradient RL)以及将二者结合的“演员-评论家”框架 (actor-critic structure)。

1.1 基于值函数的深度强化学习

考虑到每个状态下有多种动作可以选择，基于值函数的强化学习方法考虑在某个状态下的不同动作的价值，并根据这个价值来选择需要执行的动作，使用 $Q^{\pi}(s,a)$来表征。在基于价值的方法中，求解最优策略等价于求解最优的动作价值函数：

$$Q^{*}(s,a)=\max_{\pi}Q^{\pi}(s,a) \tag{5}$$

最优动作价值函数遵循贝尔曼最优方程(Bellman optimality equation)。最优策略可以表示为：

$$\pi^{*}=\arg\max_{a\in A}Q^{*}(s,a) \tag{6}$$

Q-Learning[8] 提出了一种更新 Q 值的方法，即：

$$\begin{aligned}Q'(s_t,a_t)\leftarrow&Q(s_t,a_t)\\&+\alpha(r_{t+1}+\lambda\max_{a}Q(s_{t+1},a)\\&-Q(s_t,a_t))\end{aligned} \tag{7}$$

然而，在很多实际任务中，状态空间的大小使得记录 Q 值的方法计算代价太大，会导致维度灾难。常用的解决维度灾难的方法为价值函数近似策略(value function approximation)，即引入一个函数 $Q(s,a)$来表示 Q 值：

$$Q(s,a)=f(s,a,\theta) \tag{8}$$

1.2 策略梯度的强化学习方法

基于值函数的 Q-Learning 方法在很多领域取得了成功的应用，但是也具有一定的局限性，主要体现在 2 个方面：1) 对连续动作的处理能力不足。由于需要遍历全部动作，得到具有最大 Q 值的动作，基于值函数的方法对处理连续动作的任务建模的难度是极大的；2) 无法解决随机策略问题，基于值函数的强化学习方法使用了确定性策略。若有些任务的最优策略是(近似)随机策略，基于值函数的方法则无法求解这类任务。

SUTTON 等[9]提出了策略梯度(policy gradient，PG)强化学习算法。与基于值函数的方法不同，策略梯度方法直接对策略进行建模和优化。在该类方法中，策略通常被建模为一个以 θ 为参数的函数 $\pi_{\theta}(a|s)$。奖励函数可以被定义为：

$$\begin{aligned}J(\theta)&=\sum_{s\in S}P^{\pi_\theta}(s)V^{\pi_\theta}(s)\\&=\sum_{s\in S}P^{\pi_\theta}(s)\sum_{a\in A}\pi_\theta(a|s)Q^{\pi_\theta}(s,a)\end{aligned} \tag{9}$$

式中，$P^{\pi_\theta}(s)$为在采用策略 $\pi_\theta(a|s)$情况下马尔可夫链的稳态分布，可以表示为：$P^{\pi_\theta}(s)=\lim_{t\to\infty}P(s_t=s\mid s_0,\pi_\theta)$。

根据强化学习的定义，需要对式(9)进行优化，然而直接计算其梯度$\nabla_\theta J(\theta)$是非常困难的。策略梯度方法证明了计算其梯度不需要对状态分布进行求导，极大简化了对式(9)求导的计算。

$$\begin{aligned}&\nabla_\theta J(\theta)\\&=\nabla_\theta\sum_{s\in S}P^{\pi_\theta}(s)\sum_{a\in A}Q^{\pi_\theta}(s,a)\pi_\theta(a|s)\\&\propto\sum_{s\in S}P^{\pi_\theta}(s)\sum_{a\in A}Q^{\pi_\theta}(s,a)\nabla_\theta\pi_\theta(a|s)\\&=\sum_{s\in S}P^{\pi_\theta}(s)\sum_{a\in A}\pi_\theta(a|s)Q^{\pi_\theta}(s,a)\frac{\nabla_\theta\pi_\theta(a|s)}{\pi_\theta(a|s)}\\&=E_{\pi_\theta}[Q^{\pi_\theta}(s,a)\nabla_\theta\lg\pi_\theta(a|s)]\end{aligned} \tag{10}$$

计算策略梯度的过程中需要用到 $Q^{\pi_\theta}(s,a)$，对 $Q^{\pi_\theta}(s,a)$的估计方式有很多种，最基本的 REINFORCE 方法采用了蒙特卡洛方法(Monte Carlo methods)进行估计。REINFORCE 方法的每次更新都使用当前策略 π_θ 与环境交互产生的采样轨迹，计算每个时刻 t 以后的折扣化奖励 $\psi_t=\sum_{t'=t}^{T}\gamma^{t'-t}r_{t'}$。其中，$T$ 为最大交互时刻。REINFORCE 算法中的策略梯度可以被表示为：

$$\nabla_\theta J(\theta)=E_{\pi_\theta}\left[\sum_{t=0}^{T}\psi_t\nabla_\theta\lg\pi_\theta(a_t|s_t)\right] \tag{11}$$

1.3 “演员-评论家”框架

上文介绍的基于值函数的方法只拟合一个动作价值函数，而策略梯度方法只学习一个策略函数。“演员-评论家”框架是一系列结合二者特点的算法的基本架构。该方法在策略梯度方法的

基础上引入值函数来帮助策略函数更好地学习。

在策略梯度方法中，策略梯度的一般形式由式(11)给出，其中，ψ_t 可以有多种表示形式：1) $\sum_{t'=0}^{T}\gamma^{t'}r_{t'}$ 为轨迹的总奖励值；2) $\sum_{t'=t}^{T}\gamma^{t'-t}r_{t'}$ 为 t 时刻之后的累积折扣奖励值；3) $\sum_{t'=t}^{T}\gamma^{t'-t}r_{t'}-b(s_t)$ 为包含基线函数(baseline function)[10]的改进形式，将 $b(s_t)$ 选取为 $V^{\pi_\theta}(s_t)$ 时，通常记为优势函数 $A^{\pi_\theta}(s_t,a_t)$；4) $r_t+\gamma V^{\pi_\theta}(s_{t+1})-V^{\pi_\theta}(s_t)$ 为时序差分残差。

使用 REINFORCE 方法中的蒙特卡洛采样得到的策略梯度估计是无偏的，但是因为采样次数的限制，通常会伴随着较大的方差。通过引入基线函数来减小方差是一个常用的改进策略。这里本文着重介绍将基线函数设置为当前状态的值函数，并引入时许差分残差来指导策略梯度学习的方法。

“演员-评论家”框架包含“演员”和“评论家”2 个部分。“演员”部分的结构和策略梯度中使用的结构一致，其参数采用策略梯度方法进行更新。“评论家”部分代表价值网络，记为 V_ω，其中 ω 为参数，用来拟合时许差分残差中的状态价值函数 $V^{\pi_\theta}(s_t)$。“评论家”网络的目标是拟合状态价值函数，由定义可知，根据时序差分方式得到的损失函数为：

$$L(\omega)=\frac{1}{2}(r+\gamma V_\omega(s_{t+1})-V_\omega(s_t))^2 \tag{12}$$

式中，将 $r+\gamma V_\omega(s_{t+1})$ 视为训练目标进行梯度截断，使用梯度下降法更新“评论家”网络的参数即可。

2　多智能体强化学习框架

与单智能体情况类似，多智能体强化学习也是在解决一个序列决策问题，但是同一时刻有不止一个智能体参与与环境的交互过程。因此，每个智能体的观测、观测的轨迹以及奖励值都会随着所有智能体的联合动作发生变化。单个智能体的长期优化目标将会对其他智能体策略的学习产生影响。由于多个智能体之间的观测范围和观测内容可能存在差异，多智能体系统的交互过程可以通过一个局部观测的马尔可夫过程(partially observable Markov decision process，POMDP)[10]来描述。

POMDP 可以被表示为一个七元组 $\langle N,S,\{A^i\},\{O^i\},P,\{R^i\},\gamma\rangle$。其中，$N=\{1,\cdots,N\}$ 表示 N 个智能体的编号，S 表示智能体无法观测到的全局状态集合，A^i 表示智能体 i 的动作集合，O^i 表示智能体 i 的观测集合，P 表示状态转移概率函数，R^i 表示智能体 i 的奖励集合，γ 表示折扣因子。在 t 时刻，智能体 i 根据自身观测 o_t^i 和自身策略 $\pi^i(a_t^i|o_t^i)$，执行动作 a_t^i，环境发生状态转移 $s_t\to s_{t+1}$ 并给智能体 i 反馈奖励 $r^i(s_t,\boldsymbol{a}_t,s_{t+1})$，其中 $\boldsymbol{a}_t=\{a_t^1,\cdots,a_t^N\}$ 表示 t 时刻所有智能体的联合动作。智能体 i 的值函数表示为：

$$V^i_{\pi^i,\pi^{-i}}(o_t^i)=E\Big[\sum_{t\geqslant 0}\gamma^t R^i(s_t,\boldsymbol{a}_t,s_{t+1})\ \Big|\ a_t^i\sim\pi^i(\cdot|o_t^i)\Big] \tag{13}$$

式中，$-i$ 表示除智能体 i 的其他智能体。由式(13)可知，单个智能体的最佳策略受到其他智能体影响，纳什均衡[11-12]常被用来解决此类问题，其定义为对于任意一个 π^i，一个纳什均衡点策略 $\pi^*=(\pi^{1,*},\cdots,\pi^{N,*})$ 满足在全部状态下对所有智能体都有 $V^i_{\pi^{i,*},\pi^{-i,*}}(o^i)\geqslant V^i_{\pi^i,\pi^{-i,*}}(o^i)$。纳什均衡点策略是满足所有智能体长期目标的最优策略，需要注意的是，纳什均衡点是不具备唯一性的，如果纳什均衡点存在的话，那么大多数多智能体强化学习算法的最终目的都是收敛到某一个纳什均衡点。多智能体系统的交互逻辑如图 2 所示。

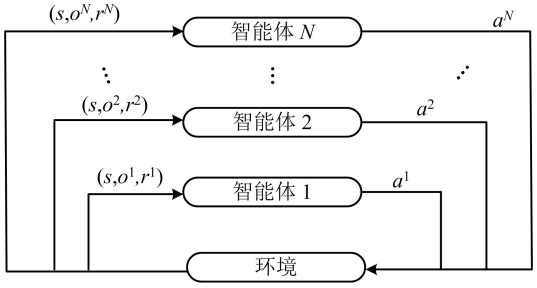

图 2　多智能体系的交互逻辑

Fig. 2　Interaction logic of multi-agent systems

此外，对多智能体场景的建模形式还包括随机博弈、局部观测随机博弈、零和局部观测随机博弈以及去中心化局部观测马尔可夫过程(Decentralized POMDP，Dec-POMDP)等，这里给出这些建模形式的简单介绍。

1) 随机博弈(stochastic game，SG)是一个多智能体的扩展 MDP 框架，用于建模多方参与的决策问题。在 SG 中，每个智能体都可以采取行动，并且环境的状态可能会受到其他智能体的影

响。SG 考虑了智能体之间的相互作用和竞争，每个智能体都追求自己的目标，并通过博弈论中的解概念来进行决策。

2）局部观测随机博弈（partially observable stochastic game，POSG）是 POMDP 和 SG 的结合，用于建模多方参与的不完全观测决策问题。在 POSG 中，每个智能体既无法直接观测到环境的状态，也无法观测其他智能体的行动和观测。POSG 考虑了智能体之间的相互作用和不完全信息，需要智能体们在不完全观测的情况下做出决策。

3）零和局部观测随机博弈（zero-sum partially observable stochastic game，Zero-Sum POSG）是一种特殊类型的 POSG，其中，智能体之间的目标是互为对立的。在 Zero-Sum POSG 中，每个智能体的奖励是互为相反数的，即一个智能体的奖励增加必然导致其他智能体奖励的减少，总奖励和为 0。这种博弈模型常见于对抗性环境中，例如棋类游戏、对策游戏和多智能体竞争环境。

4）Dec-POMDP 是一种多智能体决策问题的框架，其中，多个智能体以分布式的方式合作来解决 POMDP。每个智能体通过观测和通信来共享信息，以实现全局最优决策。Dec-POMDP 考虑了智能体之间的协作和信息共享，并通过分散的决策过程来解决整体的不完全观测问题。

通常，根据智能体之间的交互模式，多智能体强化学习可以被划分为 3 种设定模式，即合作模式、竞争模式以及混合模式。

2.1 合作模式

在完全合作模式设定中，通常所有智能体将会共享一个共同的奖励值，即 $R^1=R^2=\cdots=R^N=R$。从博弈论角度来看，这种合作模式可以被视为一种特殊的马尔可夫势博弈（Markov potential game）[13-14]，其势函数为公共的累积奖励。在这种观点中，若将所有智能体看作一个动作空间为所有智能体联合动作空间的单一智能体，则该问题将可以被视为一个单智能体强化学习问题。合作状态下的全局最优点将构成这类博弈的纳什均衡点。

此外，还有一类环境考虑了团队平均奖励[15-16]。在这类环境中，每个智能体可以有不同的奖励函数，但是整体的协作目标是将所有智能体的平均奖励最大化。这类环境直接造成了各个智能体之间的特异性，同时更符合去中心化的思想[17]，这类环境通常会鼓励智能体之间采用通信，因此基于通信的多智能体强化学习算法更青睐此类任务。

2.2 竞争模式

完全竞争模式又被称为零和马尔可夫博弈（zero-sum Markov game），即在任意时刻，所有智能体的奖励值之和为 0。为了方便理论分析，这类问题基本都聚焦于双智能体环境相互对抗[18]，其存在的意义之一是为鲁棒性学习提供理论研究的环境，可以将一方智能体视为另一方学习过程中的不确定性[19]。因此，纳什均衡点是一个优化最差情况下奖励值的鲁棒性策略。

2.3 混合模式

混合模式不再限制目标和智能体之间的关系，每个智能体都有自身的目标，它们的目标可能和其他智能体相冲突[20]。这类问题也可以由合作模式和竞争模式 2 种模式构成，例如设定 2 个在零和博弈中竞争的团队，而团队内部，则是完全合作的模式。

3 主要挑战及相关研究工作

多智能体深度强化学习近年来在许多领域中取得了较为显著的成功，但其在实践中仍然存在一系列有待解决的关键问题，主要表现在维度灾难、不稳定性、多目标性、部分可观测性 4 个方面。这些挑战制约了多智能体深度强化学习在效率、收敛性、性能等多方面的表现，因而也是未来相关研究的热点和难点。图 3 给出了本文梳理的多智能体系统的主要研究内容。

3.1 维度灾难

维度灾难[21]是一系列在分析高维数据时的反常现象。在多智能体深度强化学习中，数据的维度往往与智能体数目绑定，动作空间的大小也往往随智能体数目的增长而指数上升。因此，直接将单智能体的强化学习算法应用到多智能体场景中，可以构造出样本效率随着智能体个数的增长而指数下降的场景[22]。具体构造方法是，每个个体等概率地选取 a 和 b 2 个动作之一，当且仅当所有智能体所做的动作一致整体获得回报。可以证明，这种情况下，直接使用单智能体的策略梯度算法，得到的经验梯度和实际梯度满足：

$$P(\langle\hat{\nabla}J,\nabla J\rangle>0)\propto(0.5)^N \tag{14}$$

式中，$\hat{\nabla}J$ 代表从采样数据中求得的经验策略梯度，∇J 代表真实的策略梯度，N 代表智能体个数。从式(14)中可以看出，样本效率随智能体个数上升而指数下降。

为了实现多智能体强化学习对于智能体个数的可拓展性，往往会引入某种智能体之间的抽象结构来简化智能体之间的依赖关系。最为常见的一种抽象关系就是值函数的分解关系[23]，这一类的方法假设联合动作的值函数 $Q(s,\boldsymbol{a})$ 可以被表示成每个个体的值函数 $Q_i(o_i,a_i)$ 的函数，具体来说为：

$$Q(s,\boldsymbol{a})=f_m(Q_1(o_1,a_1),\cdots,Q_n(o_n,a_n);s) \tag{15}$$

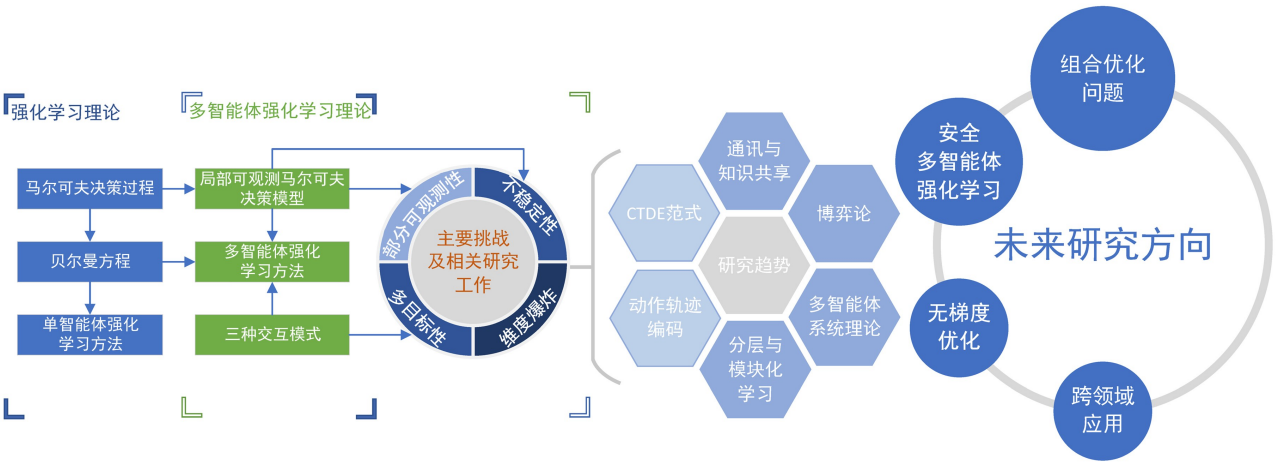

图 3　本文梳理的多智能体系的主要研究内容

Fig. 3　The main research contents of multi-agent systems discussed in this article

此类方法的关键在于假设什么样的结构将个体的值函数进行组合，即函数 f_m 的选取。VDN[24]假设联合动作值函数是个体动作值函数的和，即取 f_m 为求和函数。QMIX[25]利用个体-全局最大化(individual-global-max，IGM)假设，将联合动作值函数分解为个体值函数的单调函数形式，因此个体按照自己的值函数选择的最优动作构成的动作组合就是全局最优的动作组合，即限定 f_m 为单调函数。然而现实中，许多满足 IGM 假设的 MDP 不符合上述的分解形式，因此随后出现各种针对这一问题进行的改进方法，例如，WQMIX[26]针对 QMIX 算法可能出现的发散和低估问题提出了加权版本的 QMIX 算法；QTRAN[27]利用仿射变换得到满足 IGM 假设下真正可分解的联合动作值函数进行分解；QPLEX 则是利用对偶结构将针对值函数的 IGM 假设转化为针对优势函数的 IGM 假设，并证明了二者的等价性，从而实现对 IGM 假设的完全表达。关于值分解算法，总结见表 1 所列。

表 1　值分解算法总结

Tab. 1　Summary of value decomposition algorithms

算法名称	混合方式
VDN	$Q(s,\boldsymbol{a})=\sum_{i=1}^{n}Q_i(o_i,a_i)$
QMIX	$\forall i,\frac{\partial Q}{Q_i}\geqslant 0$
WQMIX	$\underset{w}{\Pi}Q:=\underset{q\in Q^{\mathrm{mix}}}{\operatorname{argmin}}\sum_{\boldsymbol{a}\in\boldsymbol{A}}w(s,\boldsymbol{a})(Q(s,\boldsymbol{a})-q(s,\boldsymbol{a}))^2$
QATTEN	$w_{i,h}\propto\exp(\boldsymbol{e}_i^{\mathrm{T}}\boldsymbol{W}_{k,h}^{\mathrm{T}}\boldsymbol{W}_{q,h}\boldsymbol{e}_s)$
QTRAN	$\max_{\boldsymbol{a}}Q(\boldsymbol{\tau},\boldsymbol{a})=\delta(\boldsymbol{\tau})+\sum_i Q(\tau_i,a_i)$
QPLEX	$Q_{\mathrm{tot}}(\boldsymbol{\tau},\boldsymbol{a})=\sum_{i=1}^{n}Q_i(\boldsymbol{\tau},a_i)+\sum_{i=1}^{n}(\lambda_i(\boldsymbol{\tau},\boldsymbol{a})-1)A_i(\boldsymbol{\tau},a_i)$

3.2　不稳定性

实现多智能体深度强化学习的一个最为直观的想法是，将各个体动作空间的笛卡尔积作为单智能体的动作空间，即联合动作空间，进而利用单智能体深度强化学习算法解决这一经过转化的多智能体深度强化学习问题[28]。然而正如上文所述，这类方法往往会带来维度灾难的问题，因此鲜有算法直接考虑联合动作空间当中的

动作选择。这意味着不同智能体之间的动作选择在一定程度上无法提前知晓，即多智能体深度强化学习的不稳定性问题。具体来说，单个智能体无法区分它所观测到的变化到底是来自它与环境进行的交互，还是由于其他智能体的动作选择发生了改变导致的，因此难以稳定到最优动作选择。

应对多智能体深度强化学习的不稳定性问题，通常是为单个智能体引入其他个体的动作信息，从而解耦来源于其他智能体的动作变化和环境的动态信息。MFQ 和 MFAC[29]通过引入平均场原理，使每个个体的决策不仅依赖自身动作以及自身观测，还依赖于其他个体动作的宏观影响，建模为其他个体的平均动作。其他个体可以是除该个体以外的所有个体，也可以是依照某种规则预定义的分组内的其他个体，比如距离相近的个体集合。这使得个体在决策的同时可以考虑其他个体的动作信息，进而解决多智能体深度强化学习的不稳定性问题，在大量（数百个）智能体场景下表现出色。MAVEN[30]通过采样联合隐变量 z，让每个智能体得到 z 这一共识再进行决策，因而一定程度上可以从 z 中获得其他智能体的动作选择倾向。另一方面，也有直接从方差衰减的角度出发[31]，直接对策略梯度算法的训练目标引入基线的方法，即利用方差分解公式，将集中式训练分布式执行的个体策略梯度方差进行分解，得到如下形式：

$$\begin{aligned}&\mathrm{Var}_{s_t\sim d_{\boldsymbol{\theta}}^t}\left[E_{\boldsymbol{a}_t\sim\pi_{\boldsymbol{\theta}}}\left[g_{\mathrm{C},t}^i(b)\right]\right]\\&\quad+E_{s_t\sim d_{\boldsymbol{\theta}}^t}\left[\mathrm{Var}_{\boldsymbol{a}_t^{-i}\sim\pi_{\boldsymbol{\theta}}^{-i}}\left[E_{a_t^i\sim\pi_\theta^i}\left[g_{\mathrm{C},t}^i(b)\right]\right]\right]\\&\quad+E_{\boldsymbol{a}_t^{-i}\sim\pi_{\boldsymbol{\theta}}^{-i}}\left[\mathrm{Var}_{\boldsymbol{a}_t^i\sim\pi_{\boldsymbol{\theta}}^i}\left[g_{\mathrm{C},t}^i(b)\right]\right]\end{aligned}\tag{16}$$

式中，$g_{\mathrm{C},t}^i(b)$ 代表当基线为 b 时，智能体 i 基于集中式“评论家”网络 C 得到的 t 时刻策略梯度，即 $g_{\mathrm{C},t}^i(b)=[\hat{Q}(s_t,\boldsymbol{a}_t)-b]\nabla_{\theta^i}\lg\pi_{\boldsymbol{\theta}}^i(a_t^i|s_t)$。从式(16)可以看出，第 1 项为状态引入的方差，第 2 项为其他个体的动作带来的方差，第 3 项则是由个体自身动作带来的方差。引入基线只会改变第 3 项，因此可以计算出最优基线为：

$$\begin{aligned}&b^{\mathrm{optimal}}(s,\boldsymbol{a}^{-i})\\&=\frac{E_{a^i\sim\pi_{\boldsymbol{\theta}}^i}\left[\hat{Q}(s,\boldsymbol{a}^{-i},a^i)\left\|\nabla_{\theta^i}\lg\pi_{\boldsymbol{\theta}}^i(a^i|s)\right\|^2\right]}{E_{a^i\sim\pi_{\boldsymbol{\theta}}^i}\left[\left\|\nabla_{\theta^i}\lg\pi_{\boldsymbol{\theta}}^i(a^i|s)\right\|^2\right]}\end{aligned}\tag{17}$$

与联合动作学习相反，独立学习[32]直接针对每个个体进行强化学习，将其他个体的动作影响完全视为环境动态的一部分，这也意味着这类算法往往比其他算法要承受更大的不稳定性，甚至会导致不收敛[28]。MA2QL 将同时决策的独立 Q-learning 改为顺序决策，因而后决策的智能体可以根据先决策的智能体的动作选择自己在此条件下的最优决策，进而收敛到纳什均衡点。与之类似的还有 HATRPO 和 HAPPO，分别是对 MATRPO 算法和 MAPPO 算法[33]引入顺序决策，从而增强其收敛性以及在异构智能体设定下的有效性。

3.3 多目标性

对于单智能体深度强化学习，最大化单一智能体与环境交互的回报是其唯一目标。然而在多智能体深度强化学习当中，各智能体不必要共享同一个价值函数（特别地，假如各智能体共享同一价值函数时为共同利益博弈，可以直接通过单智能体深度强化学习算法求解纳什均衡解。此时各智能体的目标为最大化自身的期望策略回报，因而是天然多目标的。针对多目标的多智能体深度强化学习算法，可以归结成理性和收敛性 2 大属性进行研究[34]，分别是从策略对其他个体动作的最优应对性，以及从策略的收敛性 2 个角度去评价以纳什均衡解作为评价标准的算法。然而在此设定下展开的多智能体学习研究的学习目标存在争议[35]，尤其是将收敛到纳什均衡作为其评价标准时，其无法保证解的最优性，也无法在存在多个均衡点时保证收敛到某个特定的均衡。针对多智能体学习任务的目标和评价，SHOHAM 等[36]提出了多智能体系统学习的 5 类目标。

1）计算。算法以计算出博弈的一种性质为目的，比如求解零和博弈的一个纳什均衡，求解一个对称博弈的纳什均衡等。这类算法不一定是效率最高的，但往往能够提供一种简单易懂且好实现的求解思路。

2）描述。算法关注建模自然个体的合作行为，比如以贝叶斯模型描述人类决策行为[37]。这类描述式的算法可以从博弈论的角度去描述自然现象，但有时现实与理论上的差别会使模型失准。

3）规范。算法主要考虑一系列的学习策略

是否相互构成均衡，比如虚拟行动模型和Q-learning模型，是否能在重复囚徒困境博弈下达成纳什均衡。这一考虑意味着其与博弈本身、博弈的时长、考虑的学习算法等都有关。

4）合作顺应。算法关注在合作场景当中，个体以何种方式调整自身策略去顺应其他个体的动作策略，以达成合作实现集体收益最大化。这种场景下往往较难从均衡的角度展开研究，个体更多地是依照既定的算法去执行而不是自由选择行为。

5）竞争。算法关注个体如何在特定的环境中以任何方式获取最大的个体回报，比如学习德州扑克[38]，这类算法并不以收敛到某个均衡为目标，而是单纯地以设计达到最大回报为唯一标准。

5类目标中，合作顺应目标与许多现实场景的实际需求相吻合，比如作战场景[39]往往以部队协同完成任务目标或歼灭敌军为合作的共同目的，又比如足球等体育竞技任务中[40]，多智能体深度强化学习也得到了广泛的研究[4,41]。为了避免维度灾难问题，将联合动作决策分解到每个智能体上，本质上形成了各个体的价值偏好。这一问题往往被称作信用分配（credit assignment）[42-43]问题。一种较为直观的信用分配方式为差分回报函数[44]，其内在逻辑为控制变量并引入微小扰动观察对输出的影响[45]。类似地，从推断角度出发的是基于反事实原理的信用分配方式，COMA[46]通过引入反事实基线计算出各智能体针对其自身动作的优势函数，从而在一定程度上消除了其他智能体对自身的动作价值误导。更多的算法是依赖神经网络直接学习最符合采样数据的信用分配结构[47]，尤其是基于集中式训练和分布式执行[48]范式设计的网络结构。QPD[49]引入神经网络研究中的梯度积分，将其应用到联合值函数网络中，提取每个个体值函数从网络结构得出的贡献值作为信用分配。此外，还有基于熵正则的一系列研究[50-51]，其主要目的是希望引入更多的随机性和更光滑的学习目标。

尽管以上将多智能体学习任务目标分为5类，但是随着研究的不断深入，不同目标间的兼容和平衡值得进一步的思考。比如在考虑竞争问题时，每个智能体会试图最大化自己的收益，因而难免存在利益冲突。其中经典的例子是囚徒困境问题，所有个体合作带来的个体收益大于存在不合作个体带来的个体收益，这一点和合作顺应问题产生了交集，且最大化个体收益在这种情况下与合作顺应不再冲突。再比如，在考虑合作顺应问题时，达成合作的方式往往又需要去考虑规范问题，即如何采取顺应策略，可以使个体策略收敛到集体利益最大的联合策略？针对此类学习目标之间的兼容和平衡问题，存在3个可能的未来方向：

1）竞争场景的计算。现有的竞争场景计算主要是通过迭代的方式求解，各智能体策略从随机策略出发，以最大化自身利益为目标进行优化。然而这种方法容易陷入局部极值，为避免陷入局部极值，一种方法是考虑维护帕累托前沿作为解集，那么算法就可以直观地理解为从点估计过渡到分布估计。然而，迭代方法对智能体数目增加难以具备拓展性，且往往效率较低。一个通过迭代方式进行竞争学习的典型示例是对抗生成网络[52]，这种生成模型可以从噪声生成图像、音频等多类数据，但训练困难是其一大问题，而这仅是2个智能体的对抗学习的一个特例。由此可见，在拓展到多智能体场景后，如何快速计算竞争个体收敛到的解是一个具有价值的开放问题。

2）描述合作顺应。描述行为的一个重要意义是在自然现象中习得有效的合作顺应策略。在准确地描述一些简单的微观行为导致的多样的宏观现象后，使实现多智能体的合作行为成为可能。这也是群体智能[53-55]研究的内容，即基于对自然现象的描述进行研究，实现多智能体的有效合作顺应，从而求解如NP、多目标搜索等复杂问题。

3）规范计算问题。现有的规范问题研究往往需要细致的分析和建模。基于对博弈问题本身以及参与博弈的智能体算法进行假设，推导出这些假设下的收敛等一系列性质。这种方式虽然可以得出相当一部分有价值的博弈策略分析结论，但难以拓展到一般的博弈策略。如何针对博弈策略和博弈问题建立合适的计算模型，将规范问题与计算问题相结合，进而得到一种可推广的规范计算方式，是未来值得研究的方向。

3.4　部分可观测性

在多智能体深度强化学习当中，部分可观测性指的是智能体具有对环境和其他智能体有限的观测信息。这在多智能体深度强化学习的实

际应用中具有重要意义，因为在实际场景中，单个智能体的能力往往会受制于其硬件条件及通讯条件。另一方面，在非合作博弈中，部分可观测性往往是一个固有条件，从而限制了个体从全局信息中直接得出最优决策。对于部分可观测性的分类可以根据局部信息的增多分为 7 个层次，从个体仅能观测到自己的奖励回报，到每个个体都了解博弈的均衡点。就学习范式而言，可以根据训练和部署时信息交换的情况分成 6 种范式。

1）共享策略的独立学习。个体共享相同的决策结构，包括策略参数、网络结构等，但个体基于其自身的独立经验进行学习，而无法访问全局信息。典型的算法包括 IQL[32]、IPPO 等。

2）独立策略的分布式学习。个体的学习完全独立，将其他个体完全视为环境动态的一部分，学习和执行期间不存在任何的直接信息交换和共享，并完全依照自身策略执行。

3）分组共享策略的独立学习。介于共享策略和独立策略的分布式学习范式，在个体之间存在以组别为单位的策略共享，也就是组内共享策略，组间独立的分布式学习范式。

4）集中式控制。执行时所有的个体都完全由唯一的中心控制器控制，这也意味着在训练阶段也是针对中心控制器的集中式训练，可以称作集中式训练集中式执行（centralized training centralized execution，CTCE）。

5）集中式训练分布式执行。这一范式一方面强调执行时个体之间的独立性，仅能通过个体的自身策略和观测而不能利用额外的信息进行辅助决策；另一方面允许甚至强调训练时的集中性，通常是通过某种中心式的聚合器将个体信息进行整合并指导训练，而个体之间的经验、观测、策略参数等在训练期间也是可以共享的。这一范式被称为集中式训练分布式执行（centralized training decentralized execution，CTDE），许多值分解算法都是基于这一范式进行设计的。

6）带通信的分布式训练。强调训练时信息交换的有限性，训练时仅允许存在通信网络连接之间的个体共享信息（且网络结构可以是时变的，一些连接可以在环境演进的过程中连接或断开）。而当实际执行时，每个个体都是独立地按照自己习得的策略进行决策，而不存在信息交换。

根据算法设计依据的范式不同，多智能体学习任务拥有的观测信息也不同，因此习得的策略理论最优表现也大相径庭。对于集中式训练集中式执行的范式而言，其拥有最多的观测信息，在理论上可以收敛到最优解。基于这一范式，原则上直接将单智能体深度强化学习应用到多智能体深度强化学习中可以达到理论最优解，但实际上由于方差过大、样本效率极低等问题难以实现。随着算力的提升和大模型的发展，集中式训练集中式执行逐渐成为近年来研究的新热点。针对这一现象，本文将集中式训练集中式执行进一步细分为 3 类，即完全集中决策、基于通信决策和基于共识决策。

1）完全集中决策。显著特点是其对联合策略进行建模，即模型会考虑建模 $\pi_\theta(a_1,\cdots,a_n|s)$。随着 Transformer 模型[56]的提出，神经网络建模复杂函数的能力大大增强。因此逐渐兴起了基于集中式训练集中式执行范式设计的多智能体深度强化学习算法。MAT[57]利用 Transformer 作为全局的信息聚合器和决策器，将多智能体在一个时间步上的行动建模为序列预测问题，以观测、已决策个体的动作、全局信息等作为输入，预测待决策个体的动作，具有较好的表现。MADT 也以类似的范式利用 Transformer 在离线强化学习场景下进行了联合式学习。

2）基于通信决策。尽管上述的几个范式均强调个体在执行时相互独立或共享策略，但是即时的信息传递也是多智能体深度强化学习考虑的一个重点，这一范式强调训练和执行时，个体之间都可以通过（动态的）网络连接来传递信息，从而个体层面透过信道获得其他个体传递的信息，使自身超越局部观测，实现更有效的决策。一系列基于通信的多智能体深度强化学习提出了基于通信的学习范式。其目的是希望在执行过程中利用智能体之间的信道，避免冲突决策动作，实现高效协同。实现基于通信的多智能体协同算法，主要需要考虑通信类型、通信策略、通信信息、信息聚合、信息利用、信道约束、沟通训练、学习方案、通信目标 9 大要素。从信息利用的角度出发，LOWE 等[58]研究了信息传递的有效性指标，指出信息的有效传递依赖于有效表出和有效接收 2 方面特点，分别代表信息的发送者需要保证其发出的信息表征了其观测或动作以及接

收者的行为需要通过某种方式受到其接收信息的影响，同时具备这 2 方面特点才能被视为有效传递信息。更多的研究关注于如何传递信息可以使得群体获得更大的收益，不同的信息传递方式体现在网络设计和训练方式上。CommNet[59]通过传递所有智能体的通信信息的平均向量，让个体的决策可以基于一个全局的平均信息进行。DIAL 和 RIAL[60]利用二值信息实现有限信息传递，从传递信息量的角度实现了更强信道约束下的通信。上述 2 种方法都是基于智能体之间的完全连接通信的。从通信对象选择而言，ATOC[61]通过使用概率门机制确定通信对象，并使用双向 LSTM 传递信息。TarMAC[62]利用注意力机制实现带权重的信息提取，从而实现指定通信对象的信息传递。

3）基于共识决策。在集中式训练集中式执行的 3 种子分类中，基于共识的决策建模最少的对间关系。从建模的复杂度上看，假设智能体的数目为 n，完全集中决策可以根据建模的复杂程度，从对所有 $O(n^2)$对对间关系建模，到对所有 $O(2^n)$组组合关系建模，实现不同复杂程度的决策逻辑。而基于通信决策，通常是基于智能体对间通信决策，根据建立的信道数目 m，可对 $O(m)$对关系进行建模。而基于共识的决策，对所有智能体的共同可见信息进行建模，因此是 $O(1)$～$O(n)$级别的建模复杂度。MAVEN[30]对所有智能体共享的策略参数隐变量 z 建模，并使智能体的动作策略依赖于隐变量 z，从而实现基于共识 z 的决策。MACKRL[63]对共同可见信息的结构进行建模，以树形结构将共识分解为智能体组之间的公共信息，相较于基于单一的隐向量建模，其假设了更丰富的共识结构，适用于更为复杂的协同场景。可以看出，相较于分布式执行的模型，集中式执行模型可以建模更为复杂的多模策略，在算力逐渐发展和场景愈发复杂的背景下，是未来值得研究的方向之一。

4　讨论

在对前文梳理的工作进行分析整理的基础上，本节总结了本文认为的当前多智能体强化学习研究的趋势特点，并指出了部分主要的多智能体强化学习的研究方向以及未来一些可能的研究方向。

4.1　研究趋势

4.1.1　CTDE 范式

该训练范式允许智能体之间在训练阶段共享信息，而在测试环境中进行完全的分布式执行。该范式既满足了最符合现实情况的分布式执行模式，又在训练阶段提供额外信息以缓解单个智能体部分可观测性带来的不稳定问题[22]和智能体之间的信用分配问题[46]。因此受到研究人员的广泛关注。

4.1.2　训练技巧

很多单智能体强化学习的研究工作指出超参数、可复现代码、随机数种子等算法外因素对算法的实验表现有不可忽视的作用[64]，这一点在多智能体强化学习领域同样适用。此外，课程学习也越来越多地被研究人员使用。课程学习可以让算法从少量智能体开始，逐步增加到更大的智能体规模；也可以从简单任务入手，逐步迁移到更困难的任务上[4]。课程学习往往会对实验表现有显著的正面作用。目前的研究工作大量使用了训练技巧，导致新提出的算法的性能提升无法被准确归因，也给其他研究人员的追踪研究带来了困难。

4.2　主要研究方向

4.2.1　混合型学习方法

为了提高多智能体系统的性能，研究者尝试将模型驱动（model-based）和模型无关（model-free）的方法相结合。混合型学习方法在学习过程中利用模型预测未来状态以指导智能体的行为，同时又可以利用模型无关方法适应不断变化的环境。这种方法可以加速学习过程，提高算法的稳定性和效率。

4.2.2　协同与竞争学习

多智能体系统中的智能体可能需要在相互协作和竞争的环境中完成任务。为了解决这种复杂场景下的学习问题，研究者提出了一些新的学习框架，例如，学习分层协同策略（learning hierarchical cooperative policies，LHCP）和基于博弈论的学习算法。这些方法有助于在协作与竞争之间找到平衡，提高多智能体系统的整体性能。

4.2.3　通信与知识共享

在多智能体系统中，有效的通信与知识共享是提高协同性能的关键。研究者正在开发新的通信协议和知识共享机制，例如，基于深度学习

的端到端通信框架(end-to-end deep learning communication frameworks)和基于图神经网络的知识共享方法。这些技术有助于实现智能体之间的信息传递和知识融合,提高多智能体系统的协同效果。

4.2.4 适应性与鲁棒性

多智能体系统面临着动态环境和不确定性因素的挑战。为了提高系统的适应性和鲁棒性,研究者正在探索新的学习算法和策略,例如,基于元学习(meta-learning)的适应性学习方法和基于安全强化学习(safe reinforcement learning)的鲁棒学习策略。适应性主要关注智能体在面临不断变化的环境和任务时,如何快速地调整自身策略以适应新的情况。鲁棒性则关注智能体在面对环境中的噪声、不确定性和异常情况时,如何保持稳定的性能。

4.2.5 分层与模块化学习

通过将复杂的任务分解为若干子任务,降低学习任务的复杂度,提高多智能体系统的可扩展性和效率。分层学习方法在 MARL 中通常采用层次化的策略结构,将任务分解为不同层次的子任务。每个层次上的智能体都有自己的策略和目标,较高层次的智能体通过协调较低层次的智能体来实现任务分配和协同。模块化学习方法在 MARL 中通常采用模块化的策略结构,将任务分解为若干相互独立的功能模块。每个模块都有自己的策略和目标,智能体通过组合不同的功能模块来实现任务的完成。

4.2.6 基于博弈论的方法

博弈论是研究多个参与者(称为玩家)之间相互作用的决策理论,尤其适用于分析智能体之间的合作与竞争关系。在多智能体环境中,非零和博弈方法可以用来研究智能体之间的协作与竞争问题。通过优化博弈均衡策略,智能体可以在合作与竞争之间达成平衡,实现整体性能的提升。重复博弈是指在一个博弈过程中,玩家多次进行相同的游戏。在多智能体环境中,智能体之间的互动往往是持续的。通过研究重复博弈,可以帮助智能体学习长期合作与竞争策略。演化博弈则关注博弈策略在长期演化过程中的变化。演化博弈可以帮助分析智能体之间策略的演化过程,为设计更高效的学习算法提供理论指导。

4.2.7 可解释性

随着智能体的学习能力不断增强,其策略和决策过程变得越来越复杂。为了确保多智能体系统的安全性、可靠性和有效性,研究者需要深入理解智能体的学习过程和决策机制。在多智能体系统中,可解释性主要关注如何设计和构建具有直观理解和分析能力的智能体。在多智能体系统中,可解释性可以帮助研究者更容易地分析和调试智能体的学习过程和策略。

4.3 未来研究方向

4.3.1 组合优化问题

需要指出的是,尽管已经有了很多成功的研究工作,多智能体强化学习领域仍旧有很多理论问题没有解决,例如,奖励不稳定性带来的泛化性问题。事实上,很多多智能体强化学习工作的成功归因于深度神经网络带来的优秀的拟合能力[22,46,65]。多智能体领域的根本问题之一是维度灾难[21,66],尽管深度神经网络能够缓解观测空间增长的问题,但随着智能体数目指数增长的联合动作空间仍是目前无法完全解决的问题。因此,如何在复杂空间中快速求解大型组合优化问题将成为多智能体强化学习中一个重要的研究问题。此外,在单智能体强化学习领域,已经有研究讨论算法的全局收敛性问题[67],而在多智能体领域,收敛性问题由于多个智能体之间存在交互变得更难以分析,目前鲜有工作研究该问题,本文认为这也将是可能的研究方向之一。

4.3.2 安全多智能体强化学习算法

另一个值得研究的方向是安全的多智能体强化学习算法。多智能体强化学习是一个贴近实际应用问题的研究领域,在实际应用中,算法的安全性也需要被纳入考量,单个智能体既需要确保整个团队取得最大收益,又希望能在训练和执行的过程中具有安全保障。这里的安全保障既包括训练目标的安全性,也应当包括数据隐私的安全性。训练目标的安全性是指即使在全合作模式下,智能体之间由于显式或隐式的任务分工不同,必定存在不完全一致的训练目标,训练目标之间不应当存在不可接受的冲突,因此需要考虑算法的抗干扰性和鲁棒性[17]。数据安全性问题在单智能体强化学习领域已经出现相关工作,如联邦学习[68]等。本文认为在多智能体领域,CTDE 范式的数据安全性研究也是可能的研究方向之一。

4.3.3 无梯度优化

多智能体强化学习和无梯度优化算法的结

合也是未来的研究方向。已有研究人员对进化算法和群体优化算法提升强化学习进行了研究，包括对奖励函数进行调整[69-70]，设置课程学习阶段等。此外，还有研究人员探索了强化学习用以提升传统基于策略的多智能体协同行为[71]，在许多稀疏奖励任务中，由专家先验知识提供的策略能够提供比纯强化学习更高的探索效率。将强化学习和人类先验知识相结合可以实现高效的探索和通过环境交互学习能力的结合。由于多智能体的群体特性，本文认为群体优化等无梯度优化方法非常适合研究智能体之间的交互行为，能够提升强化学习在多智能体领域的表现。

4.3.4　标准化测试框架

开发一系列被研究人员广泛接受的、标准化的，且能够评估多智能体算法各方面能力的测试框架（例如单智能体强化学习中的 OpenAI Gym 环境）是很有必要的，上一节中的训练技巧部分已对该问题进行了说明。

4.3.5　跨领域应用

多智能体强化学习具有广泛的跨领域应用潜力，可以将多智能体强化学习与其他领域的实际应用结合，研究任务导向的多智能体强化学习算法。这里简单列举智能交通、机器人协作、金融市场和能源管理 4 个领域。在智能交通领域，MARL 可以用于优化交通信号控制、路径规划和车辆协同等任务。多智能体系统可以对交通流量进行实时监控与分析，协调交通信号灯，以减少拥堵和提高道路利用率。此外，自动驾驶车辆可以通过多智能体学习实现协同行驶，提高行驶安全和效率。在机器人协作领域，MARL 可以用于解决多机器人协同完成任务的问题，例如仓库管理、搜救、环境监测等。通过多智能体强化学习，机器人可以学习如何有效地分配任务、协同工作，以提高任务完成速度和质量。在金融市场领域，MARL 可以应用于智能交易策略的研究。多智能体系统可以模拟市场参与者的行为，通过博弈论等方法分析市场动态，为交易策略提供有价值的信息。此外，MARL 还可以应用于风险管理、投资组合优化等金融任务。在能源管理领域，MARL 可以用于优化智能电网、微电网等能源系统的调度和运行。多智能体系统可以根据实时能源需求和供应情况，协同调度各种能源资源，以提高能源利用效率和降低运行成本。

5　结束语

多智能体强化学习在自动驾驶、团队配合游戏等领域有广阔的应用前景，但目前面临着维度灾难、不稳定性、多目标性和部分可观测性等诸多挑战。本文从多智能体强化学习的应用与理论出发，对强化学习、多智能体强化学习等方面进行概括性的介绍，针对多智能体强化学习面临的主要挑战及相关研究工作进行详细的总结与归纳，最后对多智能体强化学习未来可能的研究方向进行了展望。

参 考 文 献

[1] ZHAO W S, QUERALTA J P, WESTERLUND T. Sim-to-real transfer in deep reinforcement learning for robotics: a survey[C]//Proceedings of 2020 IEEE Symposium Series on Computational Intelligence. [S. l.]: IEEE, 2020: 737-744.

[2] SILVER D, SCHRITTWIESER J, SIMONYAN K, et al. Mastering the game of Go without human knowledge[J]. Nature, 2017, 550: 354-359.

[3] SCHRITTWIESER J, ANTONOGLOU I, HUBERT T, et al. Mastering Atari, Go, chess and shogi by planning with a learned model[J]. Nature, 2020, 588: 604-609.

[4] VINYALS O, BABUSCHKIN I, CZARNECKI W M, et al. Grandmaster level in StarCraft Ⅱ using multi-agent reinforcement learning[J]. Nature, 2019, 575: 350-354.

[5] BROWN N, SANDHOLM T. Superhuman AI for multiplayer poker[J]. Science, 2019, 365: 885-890.

[6] KIRAN B R, SOBH I, TALPAERT V, et al. Deep reinforcement learning for autonomous driving: a survey[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(6): 4909-4926.

[7] PUTERMAN M L. Markov decision processes[M]. Handbooks in Operations Research and Management Science. [S. l.]: North Holland, 1990.

[8] WATKINS C J, DAYAN P. Q-learning[J]. Machine Learning, 1992, 8(3): 279-292.

[9] SUTTON R S, MCALLESTER D, SINGH S P, et al. Policy gradient methods for reinforcement learning with function approximation[J]. Advances in Neural Information Processing Systems, 1999, 12: 1057-1063.

[10] HANSEN E A, BERNSTEIN D S, ZILBERSTEIN S. Dynamic programming for partially observable sto-

chastic games[C]//Proceedings of the 19th National Conference on Artificial Intelligence and the 16th Conference on Innovative Applications of Artificial Intelligence. [S. l. : s. n.], 2004:709-715.

[11] FILAR J, VRIEZE K. Competitive Markov decision processes[M]. New York:Springer, 2012.

[12] BASAR T, OLSDER G J. Dynamic noncooperative game theory[M]. 2nd ed. New York:Society for Industrial and Applied Mathematics, 1998.

[13] ZAZO S, MACUA S V, SÁNCHEZ-FERNÁNDEZ M, et al. Dynamic potential games with constraints: fundamentals and applications in communications[J]. IEEE Transactions on Signal Processing, 2016, 64(14):3806-3821.

[14] GONZÁLEZ-SÁNCHEZ D, HERNÁNDEZ-LERMA O. Discrete-time stochastic control and dynamic potential games: the Euler-Equation approach[M]. [S. l. : s. n.], 2013.

[15] ZHANG K Q, YANG Z R, LIU H, et al. Fully decentralized multi-agent reinforcement learning with networked agents[C]//Proceedings of the 35th International Conference on Machine Learning. [S. l. : s. n.], 2018:9340-9371.

[16] DOAN T, MAGULURI S, ROMBERG J. Finite-time analysis of distributed TD(0) with linear function approximation on multi-agent reinforcement learning [C]//Proceedings of the 36th International Conference on Machine Learning. [S. l. : s. n.], 2019:2919-2931.

[17] WAI H T, YANG Z R, WANG Z R, et al. Multi-agent reinforcement learning via double averaging primal-dual optimization[J]. Advances in Neural Information Processing Systems, 2018, 31:4649-4660.

[18] LITTMAN M L. Markov games as a framework for multi-agent reinforcement learning [M]. [S. l.]: Elsevier, 1994.

[19] JACOBSON D H. Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games[J]. IEEE Transactions on Automatic Control, 1973, 18(2):124-131.

[20] HU J L, WELLMAN M P. Nash Q-learning for general-sum stochastic games [J]. Journal of Machine Learning Research, 2003, 4:1039-1069.

[21] BELLMAN R. Dynamic programming [J]. Science, 1966, 153(3731):34-37.

[22] LOWE R, WU Y I, TAMAR A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments[J]. Advances in Neural Information Processing Systems, 2017, 30: 6382-6393.

[23] GUESTRIN C, LAGOUDAKIS M, PARR R. Coordinated reinforcement learning[C]//Proceedings of International Conference on Machine Learning. [S. l: s. n.], 2002: 227-234.

[24] SUNEHAG P, LEVER G, GRUSLYS A, et al. Value-decomposition networks for cooperative multi-agent learning based on team reward[C]//Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems. [S. l: s. n.], 2018: 2085-2087.

[25] RASHID T, SAMVELYAN M, DE WITT C S, et al. QMIX: monotonic value function factorisation for deep multi-agent reinforcement learning[C]//Proceedings of the 35th International Conference on Machine Learning. [S. l: s. n.], 2018:6846-6859.

[26] RASHID T, FARQUHAR G, PENG B, et al. Weighted QMIX: expanding monotonic value function factorisation for deep multi-agent reinforcement learning [J]. Advances in Neural Information Processing Systems, 2020, 33:10199-10210.

[27] SON K, KIM D, KANG W J, et al. QTRAN: learning to factorize with transformation for cooperative multi-agent reinforcement learning[C]//Proceedings of the 36th International Conference on Machine Learning. [S. l. : s. n.], 2019:10329-10346.

[28] CLAUS C, BOUTILIER C. The dynamics of reinforcement learning in cooperative multiagent systems [C]//Proceedings of National Conferences on Aritificial Intelligence. [S. l. : s. n.], 1999:1031-1037.

[29] YANG Y D, LUO R, LI M, et al. Mean field multi-agent reinforcement learning[C]//Proceedings of International Conference on Machine Learning. [S. l. : s. n.], 2018:5571-5580.

[30] MAHAJAN A, RASHID T, SAMVELYAN M, et al. MAVEN: multi-agent variational exploration[C]// Proceedings of the 32nd Conference on Neural Information Processing Systems. [S. l. : s. n.], 2020:7581-7592.

[31] KUBA J G, WEN M, MENG L, et al. Settling the variance of multi-agent policy gradients[J]. Advances in Neural Information Processing Systems, 2021, 34: 13458-13470.

[32] TAN M. Multi-agent reinforcement learning: independent vs. cooperative agents [C]//Proceedings of the 10th International Conference on Machine Learning. [S. l. : s. n.], 1993: 330-337.

[33] YU C, VELU A, VINITSKY E, et al. The surprising effectiveness of PPO in cooperative multi-agent games [C]//Proceedings of the 36th Conference on Neural Information Processing Systems. [S. l. : s. n.], 2022:

24611-24624.

[34] BOWLING M, VELOSO M. Rational and convergent learning in stochastic games[C]//Proceedings of the 17th International Joint Conference on Artificial Intelligence. [S. l. : s. n.], 2007:884-889.

[35] SHOHAM Y, POWERS R, GRENAGER T. Multi-agent reinforcement learning: a critical survey[R]. San Francisco: Stanford University Technical Report, 2003.

[36] SHOHAM Y, POWERS R, GRENAGER T. If multi-agent learning is the answer, what is the question? [J]. Artificial Intelligence, 2006, 171(7):365-377.

[37] KALAI E, LEHRER E. Rational learning leads to nash equilibrium[J]. Econometrica, 1993, 61:1019-1045.

[38] MORAVČÍIK M, SCHMID M, BURCH N, et al. DeepStack: expert-level artificial intelligence in heads-up no-limit poker[J]. Science, 2017, 356:508-513.

[39] SAMVELYAN M, RASHID T, DE WITT C S, et al. The StarCraft multi-agent challenge[C]//Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. [S. l. : s. n.], 2019: 2186-2188.

[40] KURACH K, RAICHUK A, STAŃCZYK P, et al. Google research football: a novel reinforcement learning environment[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S. l. : s. n.], 2020: 4501-4510.

[41] HUANG S, CHEN W, ZHANG L, et al. TiKick: toward playing multi-agent football full games from single-agent demonstrations[C]//Proceedings of the 2nd Offline Reinforcement Learning Workshop. [S. l. : s. n.], 2021.

[42] MINSKY M. Steps toward artificial intelligence[J]. Proceedings of the IRE, 1961, 49(1):8-30.

[43] CHANG Y H, HO T, KAELBLING L P. All learning is local: multi-agent learning in global reward games [C]//Proceedings of the 17th Advances in Neural Information Processing Systems. [S. l. : s. n.], 2004: 807-814.

[44] PROPER S, TUMER K. Modeling difference rewards for multiagent learning[C]//Proceedings of International Conference on Autonomous Agents and Multiagent Systems. [S. l. : s. n.], 2012:1397-1398.

[45] SUNDARARAJAN M, TALY A, YAN Q Q. Axiomatic attribution for deep networks[C]//Proceedings of the 34th International Conference on Machine Learning. [S. l. : s. n.], 2018:5109-5118.

[46] FOERSTER J N, FARQUHAR G, AFOURAS T, et al. Counterfactual multi-agent policy gradients [C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. [S. l. : s. n.], 2018:2974-2982.

[47] ZHOU M, LIU Z, SUI P, et al. Learning implicit credit assignment for cooperative multi-agent reinforcement learning[J]. Advances in Neural Information Processing Systems, 2020, 33:11853-11864.

[48] OLIEHOEK F A, SPAAN M T J, VLASSIS N. Optimal and approximate Q-value functions for decentralized POMDPs[J]. Journal of Artificial Intelligence Research, 2008, 32:289-353.

[49] YANG Y D, HAO J Y, CHEN G Y, et al. Q-value path decomposition for deep multiagent reinforcement learning[C]//Proceedings of the 37th International Conference on Machine Learning. [S. l. : s. n.], 2021: 10706-10715.

[50] IQBAL S, SHA F. Actor-attention-critic for multi-agent reinforcement learning[C]//Proceedings of the 36th International Conference on Machine Learning. [S. l. : s. n.], 2019:5261-5274.

[51] WANG J H, ZHANG Y, KIM T K, et al. Shapley Q-value: a local reward approach to solve global reward games[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence. [S. l. : s. n.], 2020: 7285-7292.

[52] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[53] KENNEDY J, EBERHART R. Particle swarm optimization[C]//Proceedings of International Conference on Neural Networks. [S. l. : s. n.], 1995: 1942-1948.

[54] DORIGO M, DI CARO G. Ant colony optimization: a new meta-heuristic[C]//Proceedings of 1999 Congress on Evolutionary Computation. [S. l.]: IEEE, 1999: 1470-1477.

[55] TAN Y, ZHU Y. Fireworks algorithm for optimization[C]//Proceedings of 2010 International Conference on Swarm Intelligence. [S. l. : s. n.], 2010: 355-364.

[56] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. [S. l. : s, n.], 2017: 6000-6010.

[57] WEN M, KUBA J G, LIN R, et al. Multi-agent reinforcement learning is a sequence modeling problem [J]. Advances in Neural Information Processing Systems, 2022, 35: 16509-16521.

[58] LOWE R, FOERSTER J, BOUREAU Y L, et al. On the pitfalls of measuring emergent communication

[C]//Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems. [S. l. :s. n.], 2019: 693-701.

[59] SUKHBAATAR S, SZLAM A, FERGUS R, et al. Learning multiagent communication with backpropagation[C]//Proceedings of the 30th Annual Conference on Neural Information Processing Systems. [S. l. :s. n.], 2016:2252-2260.

[60] FOERSTER J N, ASSAEL Y M, DE FREITAS N, et al. Learning to communicate with deep multi-agent reinforcement learning[C]//Proceedings of the 30th Annual Conference on Neural Information Processing Systems. [S. l. :s. n.], 2016:2145-2153.

[61] JIANG J, LU Z. Learning attentional communication for multi-agent cooperation[C]//Proceedings of the 32th Annual Conference on Neural Information Processing Systems. [S. l. :s. n.], 2018: 7265-7275.

[62] DAS A, GERVET T, ROMOFF J, et al. TarMAC: targeted multi-agent communication[C]//Proceedings of the 36th International Conference on Machine Learning. [S. l. :s. n.], 2019:2776-2784.

[63] DE WITT C S, FOERSTER J, FARQUHAR G, et al. Multi-agent common knowledge reinforcement learning[C]//Proceedings of Conference on Neural Information Processing Systems. [S. l. : s. n.], 2020: 9895-9907.

[64] HENDERSON P, ISLAM R, BACHMAN P, et al. Deep reinforcement learning that matters[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. [S. l. :s. n.], 2018:3207-3214.

[65] OMIDSHAFIEI S, PAZIS J, AMATO C, et al. Deep decentralized multi-task multi-agent reinforcement learning under partial observability[C]//Proceedings of the 34th International Conference on Machine Learning. [S. l. :s. n.], 2017:4108-4122.

[66] HERNANDEZ-LEAL P, KARTAL B, TAYLOR M E. A survey and critique of multiagent deep reinforcement learning[J]. Autonomous Agents and Multi-Agent Systems, 2019, 33(6): 750-797.

[67] CAI Q, YANG Z R, LEE J D, et al. Neural temporal-difference learning converges to global optima[C]//Proceedings of the 32nd Conference on Neural Information Processing Systems. [S. l. : s. n.], 2020: 11283-11294.

[68] WANG H, KAPLAN Z, NIU D, et al. Optimizing federated learning on non-IID data with reinforcement learning [C]//Proceedings of IEEE Conference on Computer Communications. [S. l.]: IEEE, 2020: 1698-1707.

[69] WANG J X, HUGHES E, FERNANDO C, et al. Evolving intrinsic motivations for altruistic behavior [C]//Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems. [S. l. :s. n.], 2019: 683-692.

[70] JADERBERG M, CZARNECKI W M, DUNNING I, et al. Human-level performance in 3D multiplayer games with population-based reinforcement learning [J]. Science, 2019, 364: 859-865.

[71] LI J, TAN Y. A two-stage imitation learning framework for the multi-target search problem in swarm robotics[J]. Neurocomputing, 2019, 334: 249-264.

作 者 简 介

陈人龙

男,1995 年生,博士研究生,研究方向为多智能体强化学习、群体机器人

E-mail:reo@pku. edu. cn

陈嘉礼

男,1998 年生,博士研究生,研究方向为多智能体强化学习、群体机器人

E-mail:chenjiali@pku. edu. cn

李善琦

男,1995 年生,硕士研究生,研究方向为多智能体强化学习、群体机器人

E-mail:lishanqi@stu. pku. edu. cn

谭 营

男,1964 年生,博士,教授,博士研究生导师,烟花算法发明人,研究方向为群体智能、群体机器人、多智能体强化学习等

E-mail:ytan@pku. edu. cn

责任编辑 董 莉