情况下以有限的计算量解决协作任务，其计算复杂度与单能体 Q 学习的计算复杂度相似。然而，该算法只适用于具有非负报酬函数的确定性问题。上述算法存在一些局限，即它们都依赖于对状态的精确测量，一些还需要精确测量其他智能体的作用，并且还会受维数灾难的影响。

2.2.2 完全竞

在完全竞争的随机博弈中（对于两个智能体，即 $R_1=R$ 可以应用极小极大化原则）：在最坏情况下假设最大化一个智能体的回报，这个假设是对手将始终努力使其回报最小化。minimax-Q 算法采用极小极大原理来计算阶段游戏的策略和值，以及类似于 Q 学习的时序差分规则。下面给出智能体 1 的算法：

$$h_1(s,\cdot,\cdot)=\arg m_1(Q_t,s) \tag{19}$$

$$Q_{t+1}(s_t,a_{1,t},a_{2,t})=Q_t(s_t,a_{1,t},a_{2,t})+\alpha[r_{k+1}+\gamma m_1(Q_t,s_{t+1})-Q_t(s_t,a_{1,t},a_{2,t})] \tag{20}$$

其中，m_1 是智能体 1 的极小极大化回报

$$m_1(Q,s)=\max_{h_1(s,\cdot)}\min_{a_2}\sum_{a_1}h_1(x,a_1)Q(s,a_1,a \tag{21}$$

表 1 简单地对各种算法进行了对比和总

表 1 基于博弈的经典 MARL 算法的

Table 1 Comparison of classic MARL algorithm based on game th

算法	值函数更新	动作选择
Q-learning	$Q(s,a)\leftarrow(1-\alpha)Q(s,a)+\alpha[r+\gamma\max Q(s',a')]$	ε-贪心策略，或基于玻尔兹曼机分布的动作选择
Minimax Q	$Q(s,a,o)\leftarrow(1-\alpha)Q(s,a,o)+\alpha[r+\gamma V(s')]$ $V(s')=\max_{\pi}\min_{o}\sum aQ(s',a,o)\pi(s',a')$	基于当前学习到的策略 π 选择动作
Nash Q	$Q(s,a)\leftarrow(1-\alpha)Q(s,a)+\alpha[r_i+\gamma NashQ_i(s')]$	基于当前博弈的纳什均衡选择动作
CE-Q	$Q(s,a)\leftarrow(1-\alpha)Q(s,a)+\alpha[r_i+\gamma CEQ_i(s')]$	基于当前博弈的相关均衡选择动作

注：o 代表相互博弈的智能体的对手采取的

然而，传统的 MARL 算法多适用于小规模的问题，很有算法能应用于信息不完整或不确定的环境。提高 MARL 对实际问题的适应性是一个必不可少的研究步骤[25]。

3 MARL 的研究进

经典的 MARL 算法通常仅适用于小问题，如静态游戏小网格世界，而在现实的多智能体问题中，状态和动作空间很大甚至是连续的。很少有传统算法能够适用于不完整的、不确定的环境。可扩展性和处理不完全信息环境也是之前单智能体强化学习中未解决的问题。然而，随着深度学习的发展，单智能体的可扩展性等问题在一定程度上得到了解决。借助深度学习快速发展的红利，研究人员结合深度学习等技术，从可扩展性、智能体意图、奖励机制、环境框架等诸多方面对多智能体算法进行了改进和创新。

3.1 深度强化学习的主要进

3.1.1 深度 Q 网

深度 Q 网络（Deep Q-Network，DQN）是由 Mnih 等[26]出的。它依托强化学习中经典的 Q 学习，用一个深度网络近似价值函数为深度网络提供目标值，不断更新网络直至收敛。其中涉及 3 项关键技术：1）用经验重放技术打破了样本间的关联性，将采集到的样本先放入样本池，然后从池中随机选出一个样本用于网络训练；2）设置了目标网络来单独处理时间差分算法中的 TD 偏差，使训练的稳定性和收敛性得到极大的提高；3）利用卷积神经网络逼近行为值函数。其他学者也

其中，在状态 s 时的智能体 1 的随机策略由 $h_{1(s,\cdot)}$ 表示，点表动作参数。最优化问题可以由线性规划解决。

2.2.3 混合型任

在混合随机博弈中，智能体的奖励函数不受约束，这种式最适合自私的智能体。博弈论均衡概念在混合随机博弈中运用得最多，该类别中的大量算法仅针对静态任务。像 Q 学习这样的单智能体算法可以直接应用到混合型任务中。参数的更新需要使用所有智能体的 Q 表，因此每个智能体都要复制其他智能体的 Q 表，这要求所有的智能体使用相同的算法并且可以测量所有的动作和奖励。即使有了这些假设，当不同智能体求得的策略不唯一时，也会出现均衡选择问题。一种常用的方法是 Nash Q-learning，此外还有相关平衡 Q 学习（CE-Q）[23]或不对称 Q 学习（Asymmetric Q-learning）[24]，它们可以分别通过使用相关或 Stackelberg（前导-跟随）平衡来解决均衡问题。对于不对称 Q 学习，跟随者不需要对领导者的 Q 表进行建模，但是领导者必须知道追随者如何选择其行动。

围绕 DQN 做了许多研究和改进。Hasselt 等[27]提出了 DQN 算法，使得值估计过于乐观这一问题得到解决。Schaul 等[28]使用了经验优先回放技术，对经验的优先次序进行处理。此外，Osband 等[29]，Munos 等[30]，Francois-lavet[31]等也分别从其他角度对 DQN 提出了改进。

3.1.2 演员评论家算

演员评论家算法（Actor-Critic，AC）的架构可以追溯 30～40 年前（见图 2）。最早，Witten 于 1977 年提出了类似 AC 算法的方法；然后，Barto，Sutton 和 Anderson 等于 1983 年引入了 AC 架构。但是，AC 算法的研究难度和一些历史偶然因素使得之后学术界开始将研究重点转向基于价值的方法。之后的一段时间里，基于价值的方法和基于策略的方法都有了蓬勃的发展。AC 算法结合了两者的发展红利，在理论和实践方面再次有了长足的发展。

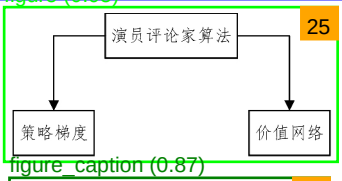

图 2 演员评论家算法的基本

Fig. 2 Basic framework of actor-criti algori

该结构包含两个网络：一个策略网络（Actor）和一个价网络（Critic）。策略网络输出动作，价值网络评判动作。策略网络通过梯度计算公式进行更新，而价值网络根据目标值进行更新。相比以值函数为中心的算法，AC 算法应用了策略