# 多智能体强化学习：一种理论和算法的选择性的学习方法概述

张开清\杨卓然\Tamer Basar

## 1953729 吴浩泽

## 摘要

近年来，强化学习（RL）取得了重大进展，在解决各种顺序决策方面取得了巨大成功机器学习中的问题。大多数成功的RL应用程序，例如游戏Go和Poker、机器人和自动驾驶，涉及更多不止一个单独的代理，自然属于多代理RL（MARL）领域，这是一个历史相对悠久的领域，最近由于技术进步而重新出现在单代理RL技术中。虽然在经验上取得了成功，但文献中相对缺乏泥灰岩的理论基础。在本章中，我们提供MARL的选择性概述，重点是有理论分析支持的算法。更具体地说，我们主要回顾了MARL算法的理论结果在两个有代表性的框架内，马尔可夫/随机对策和扩展形式游戏，根据其处理的任务类型，即完全合作竞争激烈，两者兼而有之。我们还介绍了这些算法的几个重要但富有挑战性的应用。与泥灰岩的现有评论正交，我们重点介绍了泥灰岩理论的几个新角度和分类法，包括学习在广泛形式的游戏中，使用网络代理分散MARL，MARL在基于策略的游戏学习方法的平均场机制，（非）收敛性，一些新的角度是从我们自己的研究努力和兴趣中推断出来的。本章的总体目标是，除了提供标记该领域的当前状态，以确定富有成效的未来研究方向泥灰岩的理论研究。我们预计这一章将继续起到刺激作用对于有兴趣研究这一激动人心而又富有挑战性的课题的研究人员。

## 1 介绍

近年来，强化学习（RL）在许多领域取得了令人瞩目的进展突出的顺序决策问题，如玩围棋，玩实时战略游戏、机器人控制、扑克牌游戏，和自动驾驶，尤其是伴随用于函数逼近的深层神经网络（DNN）的发展。有趣的是，大多数成功的应用程序都需要一个以上的代理/玩家参与哪个应系统地建模为多智能体RL（MARL）问题。明确地MARL解决了多个自治代理的顺序决策问题它们在一个共同的环境中运行，每个环境都旨在通过与环境和其他代理的交互优化自身的长期回报。除了在前面提到的热门领域中，多智能体系统的学习在其他子领域也有潜在的应用，包括网络物理系统、金融、传感器/通信网络和社会科学。在很大程度上，MARL算法可以分为三组，即完全合作、完全竞争和两者的混合，这取决于它们所处理的设置类型。特别是在协作环境中，代理协作以优化共同的长期回来而在竞争环境中，代理人的回报通常为零。混合环境包括合作代理和竞争代理，总金额返回。为不同的泥灰岩环境建模需要跨越优化理论、动态规划、博弈论和分散控制的框架，见更多详细讨论。尽管存在这些多个框架，但事实上，MARL中的挑战在不同的环境中是常见的，尤其是对于理论分析。具体来说，首先，MARL的学习目标是多维的，由于所有代理的目标不一定一致，这就带来了以下挑战处理平衡点，以及一些额外的性能标准返回优化，例如通信/协调的效率，以及对潜在敌对代理的鲁棒性。此外，由于所有代理都在根据自身利益同时改进其政策，因此每个代理所面临的环境变得不稳定。这打破或使单代理环境中大多数理论分析的基本框架无效。此外，增加的联合行动空间代理数量呈指数增长可能会导致可伸缩性问题，称为MARL的组合性质。此外，信息结构，即谁知道在泥灰岩中，更为重要的是，每个代理人都只能有限地观察其他，则可能导致局部出现次优决策规则。详细阐述潜在挑战见。事实上，一直在努力应对上述挑战。有关早期的理论和算法的全面概述，请参见泥灰岩。近年来，由于单智能体RL技术的发展，这一领域重新引起了人们的兴趣。事实上，最近出现了大量关于泥灰岩的工作，由于深度学习的发展，专注于识别新的学习标准和/或设置，或为现有设置开发新算法、运筹学和多智能体系统。然而，并非所有的努力都是在严格的理论框架下进行的基础，部分原因是对单代理深层RL理论的理解有限，部分原因是多代理设置中固有的挑战。因此必须在理论保证的情况下审查和组织MARL算法突出现有研究工作的边界，激发潜力这一主题的未来方向。在本章中，我们选择性地概述了MARL中的理论和算法，以及几个重要但极具挑战性的应用程序。更具体地说，我们关注MARL的两个代表性框架，即马尔可夫/随机博弈和广泛形式的游戏，在离散时间设置中，如在标准单代理RL中。根据上述三组，我们审查并特别关注具有收敛性和复杂性分析的MARL算法，其中大多数是最近才出现的。考虑到这一重点，我们注意到，我们的概述绝不全面。事实上，除了经典参考文献之外，还有其他一些关于泥灰岩的评论由于泥灰岩的重新出现，最近出现了这种现象。我们想要强调这些审查提供了互补的视图和分类法对我们来说：调查了专门设计用于解决对手诱导的问题的作品非平稳性是我们在§3中讨论的挑战之一；相对来说更全面，但重点是深层泥灰岩，这是一个缺乏理论的分区远的另一方面，虽然只关注协作设置中的算法但是此设置中的审查范围很广。最后，我们重点介绍了几个新的角度和分类法，这些角度和分类法在现有的泥灰岩综述中相对未充分探索，主要是由于我们自己的研究努力和兴趣。首先，我们讨论了MARL中广泛形式游戏的框架除了传统的马尔可夫博弈，甚至是简化的重复博弈；其次，我们总结了最近发展起来的一个分区的进展：带有网络代理的去中心化MARL，作为我们早期工作的外推；第三，我们将平均场制度引入泥灰岩，作为对代理人人数极其庞大的案件；第四，我们重点介绍了优化理论的一些最新进展，这些进展揭示了基于策略的MARL的方法，尤其是零和游戏。我们还回顾了部分观察环境中泥灰岩的一些文献，但没有使用深度RL作为启发解决方案。我们希望这些新角度有助于确定富有成效的未来研究方向，更重要的是，激发研究人员对泥灰岩建立严密理论基础的兴趣。路线图。本章的其余部分组织如下。在§2中，我们介绍MARL的背景：单代理RL的标准算法，以及泥灰岩。在§3中，我们总结了发展泥灰岩理论的几个挑战给单一代理对应方。一系列MARL算法，主要是理论根据担保所涉及的任务类型，在§4中对担保进行审查和组织。在§5中，我们简要介绍了算法驱动的MARL最近取得的一些成功之后是第6节中概述的结论和几个开放研究方向。

## 2背景

在本节中，我们提供了强化学习的必要背景，包括单代理和多代理设置。

#### 2.1单代理RL

对强化学习代理进行建模，通过与环境互动。环境通常被描述为一个无限水平的折扣马尔可夫决策过程（MDP），以下简称为马尔可夫决策过程决策流程2，其正式定义如下。定义2.1马尔可夫决策过程由元组（S、A、P、R、γ）定义，其中S和A分别表示状态空间和动作空间；P:S×A→ ∆（S） 表示转换任何状态的概率s∈ S到任何状态S0∈ S表示任何给定动作a∈ A.R:S×A×S→R是奖励函数，用于确定代理收到的从（s，a）过渡到s0；γ∈ [0,1）是折衷瞬时和未来奖励。作为一种标准模型，MDP已被广泛用于描述具有系统状态完全可观测性的agent决策。3每次t，代理人选择在执行操作面对系统状态st，这会导致系统崩溃过渡到st+1∼ P（·| st）在)。此外，代理还会收到即时奖励R（st在，st+1）。因此，解决MDP的目标是找到一个策略π：S→ ∆（A） ，一个映射从状态空间S到动作空间A上的分布∼ π（·| st)以及折扣累计奖励最大化因此，可以定义动作值函数（Q函数）和对于任意s，策略π下的状态值函数（V函数）∈ S和a∈ A、 这是从（s0，a0)=（s，a）和s0=s。与最优策略相对应的策略π∗ 通常被称为最优Q函数和最优状态值函数，分别地利用马尔可夫性质，可以通过动态规划/反向归纳法（如值迭代和策略迭代）获得最优策略算法，需要模型知识，即转移概率以及奖励函数的形式。另一方面，强化学习是为了这样一个最优策略，而不知道模型。RL代理从中学习策略通过与环境互动收集的经验。总的来说，RL算法可以分为两种主流类型，基于价值的方法和基于策略的方法。

##### 2.1.1基于价值的方法

设计了基于值的RL方法，以找到状态动作值的良好估计函数，即最优Q函数Qπ∗. （近似）最优策略可以然后通过Q函数估计的贪婪行为进行提取。其中一个最流行的基于值的算法是Q-learning，其中agent维护Q值函数Qˆ（s，a）的估计。从状态动作对转换时到下一个状态s0，代理收到一个付款人r，并根据以下内容更新Q函数：其中α>0是步长/学习率。在α上的某些条件下，Q-学习可以被证明几乎肯定会收敛到最优Q值函数，且状态空间和动作空间。此外，当与神经网络结合使用时近似，深度Q学习在人类水平控制应用中取得了重大的经验突破。另一种流行的基于政策价值的方法是SARSA，其收敛性是在中针对有限空间设置建立的。另一种流行的基于值的RL算法是蒙特卡罗树搜索（MCTS），通过构造搜索来估计最优值函数通过蒙特卡罗模拟生成树。明智地选择要平衡的操作的树策略探索利用用于构建和更新搜索树。最常见的树策略是应用UCB1（UCB代表置信上限）算法最初设计用于随机多臂bandit问题，用于这棵树。这就产生了流行的UCT算法。关于MCT的非渐近收敛性包括。此外，关于RL中的值函数的另一个重要任务是估计值与给定策略关联的函数（不仅仅是最优策略）。这项任务通常被称为策略评估，已经由遵循类似更新的算法来完成，命名为时间差（TD）学习。其他一些共同政策具有收敛保证的评估算法包括具有线性和非线性函数近似。有关政策评估的更详细审查，请参见。

##### 2.1.2基于策略的方法

另一种类型的RL算法直接搜索策略空间，通常是由神经网络等参数化函数逼近器估计，即逼近π（·| s）≈ πθ（·s）。因此，最直接的想法是沿着长期奖励的梯度方向更新参数，已由策略梯度（PG）方法实例化。作为这一想法的一个关键前提，PG的闭合形式如所示，其中J（θ）和Qπθ分别是政策πθ下的预期收益和Q函数，∇logπθ（a | s）是策略的得分函数，ηπθ是州占用率指标，在政策πθ下，折扣或遍历。然后，各种政策梯度方法，包括Enhanced、G（PO）MDP和actor-Critical算法，都有用不同的方法估计梯度。类似的想法也适用到连续操作设置中的确定性策略，其PG形式已派生最近由提出。除了基于梯度的策略优化方法外，其他几种策略优化方法在许多应用中都取得了最先进的性能，包括PPO，TRPO，软演员评论家。与基于值的RL方法相比，基于策略的RL方法具有更好的收敛性保证，尤其是使用神经网络进行函数逼近，它可以很容易地处理大量甚至连续的状态动作空间。此外在基于价值和策略的方法中，也存在基于线性的RL算法MDP的计划制定；参见中最近的努力。

#### 2.2多智能体 RL框架

同样，多智能体RL也解决了顺序决策问题，但不止一名特工参与其中。特别是，系统演化图1：马尔可夫决策过程的系统演化示意图，一个马尔可夫博弈和一个扩展形式博弈，对应于分别为单代理和多代理RL。具体地，在如（a）中的MDP中，代理在输出动作a之后观察状态s并从系统接收奖励r；在里面如（b）所示，所有代理都选择动作a我同时，在观察系统后状态s和接收每个个人奖励r我；在两人的广泛形式比赛中，如（c） ，代理决定选择操作a我交替地，并获得每个人的奖励r我（z） 在游戏结束时，z是终端历史。在不完美的信息案例中，玩家2不确定自己在游戏中的位置，这使得信息集为非单例。每个代理人获得的状态和报酬都受到所有代理人共同行动的影响代理人。更有趣的是，每个代理都有自己的长期奖励需要优化现在成为所有其他代理的策略的函数。这样的一般模型发现在实践中的广泛应用，请参见§5，以获取对几个突出示例的详细审查。一般来说，对于马尔可夫/随机对策和广义形式对策，存在两种看似不同但密切相关的理论框架，下面将介绍这两种理论框架。不同框架下系统的演变如图所示。

##### 2.2.1马尔可夫/随机博弈

MDP的一个直接概括是捕获多个代理的交织马尔可夫博弈（MGs），也称为随机博弈。起源于神学院文献中长期使用MGs4的框架来开发泥灰岩算法，详见§4。下面介绍形式定义。定义马尔可夫博弈由元组（N，S，{Ai）定义}我∈N，P，{R我}我∈N，γ），其中N={1，····，N}表示N>1个代理的集合，S表示所有代理观察到的状态空间，Ai表示agent i的动作空间。设A：=A1×·····×AN，则P:S×A→ ∆（S） 表示任意状态的转移概率∈ S到任何状态S0∈ S表示任何联合行动a∈ A.R我：S×A×S→ R是奖励函数，用于确定从（s，a）过渡到s的代理i0；γ∈ [0,1）是折扣系数。在时间t，每个代理i∈ N执行动作a我t型，根据系统状态st。这个然后系统转换到状态st+1，并通过R奖励每个代理i我（st在，st+1）。的目标agent i通过找到策略π来优化自身的长期回报我：S→ ∆（Ai)这样的那是我t型∼ π我（·| st)。因此，值函数V我：S→ 代理人i的R变为联合政策的作用π：S→ ∆（A） 定义为π（A | s）：=Q我∈Nπ我（a）我|s） 。特别是，对于任何联合策略π和状态s∈ S、 在哪里−i表示N中除agent i以外的所有agent的索引。因此，解决方案MG的概念与MDP的概念有所不同，因为代理不仅受其自身策略的控制，还受所有其他参与者的选择的控制游戏的一部分。最常见的解决方案概念，纳什均衡（NE）5，定义为遵循。定义2.3马尔可夫博弈的纳什均衡（N，S，{Ai}我∈N，P，{R我}我∈N，γ）是a联合保单π∗ = （π1.∗，···，πN，∗)，对于任何∈ S和i∈ N五、我πi，∗，π−我，∗（s）≥ 五、我πi，π−我，∗（s） ，对于任意π我。纳什均衡刻画了一个均衡点π∗，其中没有任何代理有任何偏离的动机。换言之，对于任何代理人，我∈ N，策略π我，∗是π的最佳响应−我，∗。作为MARL的标准学习目标，NE始终存在于有限空间中无限期折扣MGs，但通常可能不是唯一的。大多数MARL算法是为了收敛到这样一个平衡点（如果存在的话）。马尔可夫博弈的框架足够普遍，可以涵盖各种泥灰岩环境总结如下。合作设置：在完全合作的环境中，所有代理通常共享一个共同的奖励函数，i、 e.，R1=R2=···=RN=R。我们注意到该模型也被称为多智能体AI社区中的MDP（MMDP），以及控制/博弈论社区。此外，从博弈论的角度从另一个角度来看，这种合作设置也可以被视为马尔可夫势博弈的特例，势函数是公共累积函数奖励考虑到这个模型，值函数和Q函数与所有代理，从而启用单代理RL算法，例如Q-学习更新（2.1），如果所有代理作为一个决策者进行协调，则应用。全局最优解因为合作现在构成了博弈的纳什均衡。除了常见的奖励模式之外，还有一种更为通用和激增的模式对于合作，MARL考虑团队平均报酬。特别是代理人允许有不同的奖励功能，可能对每个代理都是私有的，而合作的目标是优化与平均奖励R（s、a、s0）)：=N−1.·P我∈N右我（s、a、s0)对于任何（s、a、s0)∈ S×A×S。AverageWard模型允许代理之间更多的异构性，包括上述模型作为特例。它还保护了代理之间的隐私，并促进了开发分散的MARL算法。这种异质性还需要将通信协议纳入MARL，并分析通信效率高的MARL算法。

竞争环境：

MARL中的完全竞争环境通常被建模为零和马尔可夫博弈，即P我∈N右我（s、a、s0)=0表示任何（s、a、s0)。为了便于算法分析和计算可处理性，大多数文献都集中在两个相互竞争的代理上，其中一个代理人的报酬显然是另一个代理人的损失。此外为了将应用程序引导到游戏中，零和游戏也可以作为一个模型对于鲁棒学习，由于不确定性阻碍了agent的学习过程可以被视为一个虚构的对手，在游戏中总是与代理人作对。因此，纳什均衡产生了一个稳健的策略，可以优化最坏情况下的长期回报。

混合设置：

混合设置也称为一般和游戏设置，其中没有限制是强加在目标和代理人之间的关系上的。每个代理人都是自利的，其报酬可能与其他代理人的报酬相冲突。平衡溶液概念从博弈论来看，如纳什均衡，具有最显著的影响针对此常规设置开发的算法。此外，我们还包括设置完全合作和竞争代理，例如，两个零和每个团队中都有合作代理的竞争团队，例如混合设置也是如此。

##### 2.2.2广泛形式的游戏

尽管马尔可夫博弈构成了马尔可夫的经典形式主义，但马尔可夫博弈只能处理完全观察到的情况，即代理拥有关于系统状态的完整信息st和在时间t执行的操作。尽管如此，大量MARL应用程序让只有部分可观测性的代理参与，即游戏信息不完全。Ex9马尔可夫对策对部分观测情形的张力可能是适用的，然而即使在合作环境下，也很难解决问题。相比之下，另一个名为扩展形式博弈（extensive form games）的框架可以方便地为多智能体决策建模不完美信息。此框架是植根于计算博弈论，并已证明在温和条件下允许使用多项式时间算法。下面简要介绍extensiveform游戏的框架。定义扩展形式游戏定义如下（N∪{c} ，H，Z，A，{R我}我∈N，τ，πc，S），其中N={1，…，N}表示N>1个代理的集合，c是一个称为偶然或自然的特殊代理，它有一个固定的随机策略，指定环境的随机性。此外A是代理可以执行的所有可能操作的集合，H是所有可能历史的集合，其中每个历史是从游戏开始执行的一系列操作。允许A（h）={A | ha∈ H} 表示非终结历史H之后可用的操作集。假设代理执行操作a∈ A（h）给定历史h∈ H、 这将导致一个新的历史哈∈ H在所有的历史中，Z⊆ H是终端历史的子集，表示一场比赛。为每个代理i分配一个实用程序∈ N在终端历史记录中，由函数指示R我：Z→ R、 此外，τ：H→ N∪ {c} 是指定哪个代理的标识函数在每个历史记录中执行操作。如果τ（h）=c，机会代理根据其政策πc，即a∼ πc（·| h）。此外，S是H的分区，因此对于任何S∈ S和任意h，h0∈ s、 我们有τ（h）=τ（h0)A（h）=A（h0)。换句话说，历史h和h0在对于即将采取行动的主体，即τ（h），相同的划分是无法区分的。这个S中的元素称为信息状态。直觉上，广泛形式博弈的不完全信息反映在代理无法区分同一信息集中的历史。自从我们有τ（h）=τ（h0)A（h）=A（h0)对于所有h，h0∈ s和s∈ S、 为了便于演示，在续集中∈ s、 我们让A（s）和τ（s）分别表示A（h）和τ（h）。我们还定义映射I:H→ 如果h，则让I（h）=S∈ s、 此外，我们只考虑H和A都是有限集的对策。为了简化符号，对于任意两个历史记录h、 h0级∈ H、 我们称H为H的前缀0，表示为h v h0，如果为h0可从h访问通过采取一系列行动。在这种情况下，我们称之为h0是h的后缀。此外，我们假设整个游戏具有完美的回忆功能，这意味着每个代理记住导致其当前信息状态。文献中普遍存在完全回忆的假设，这使得解决游戏的多项式时间算法的存在成为可能。更重要的是，根据著名的库恩定理[108]，在这种假设下找到纳什均衡集，将推导限制在行为均衡集就足够了映射每个信息集的策略∈ S到a（S）上的概率分布。对于任何i∈ N，让我们i={s∈ S：τ（S）=i}是agent i的信息状态集。联合策略代理的数量表示为π=（π1.，。。。，πN），其中π我：S我→ ∆（A（s））是代理人的政策i、 对于任何历史h和任何联合策略π，我们将π下h的到达概率定义为，它指定了当所有代理遵循π时创建h的概率。我们同样定义信息状态s在π下的到达概率为ηπ（s）=Ph类∈sηπ（h）。这个代理i的预期效用∈ 因此，N由P给出z∈Zηπ（Z）·R我（z） ，表示为R我（π） 为了简单起见。现在，我们准备介绍扩展形式博弈的解概念，即纳什均衡及其-近似，如下所示。定义2.5（N）表示的广义形式博弈的-纳什均衡∪{c} ，H，Z，A、 {R我}我∈N，τ，πc，S）是联合策略π∗ = （π1.∗，···，πN，∗)，因此对于任何∈ NR我（π我，∗，π−我，∗)≥ R我（π我，π−我，∗)− ，对于任何策略π我代理人i的。这里π−i表示N{i}中agent的联合策略，其中agent j采用策略πj对于所有人j∈ N \{i}。此外，如果=0，π∗构成纳什均衡。

各种设置：

广泛形式的游戏通常用于模拟非合作环境。具体来说，零和/常和效用与P我∈N右i=k，对于某些常数k对应充分竞争的环境；通用求和实用程序函数会导致混合设置。更重要的是，还可以描述不同信息结构的设置通过广泛的形式游戏。特别是，一个完美的信息游戏是这样一个游戏，其中每个信息集都是一个单体，即对于任何∈ S、 | S |=1；一个不完美的信息游戏存在s的地方∈ S、 | S |>1。换句话说，在信息不完全的情况下，用于决策的信息状态代表多个可能的历史代理无法区分它们。在各种设置中，零和不完全信息设置是主要设置将泥灰岩与广泛形式游戏联系起来的理论研究重点。它还激发了MARL算法，从而彻底改变了竞争环境扑克牌AI等应用程序。

与马尔可夫博弈的联系：

请注意，定义2.2和2.4中的两种形式是相互关联的。特别是，对于同时移动马尔可夫博弈，其他代理对动作的选择是代理未知，因此会导致不同的历史记录，这些历史记录可以聚合为这些游戏中的一个信息状态s。历史是联合行动的序列，折扣累积奖励在游戏结束时实例化效用。相反，只需设置Aj=∅ 在s状态下，对于agent j，τ（s），扩展形式对策归结为具有状态相关动作空间的马尔可夫对策。更多信息，请参见关于连接的详细讨论。备注2.6（其他泥灰岩框架）泥灰岩的其他几个理论框架存在于文献中，例如，范式和/或重复博弈，以及部分观察到的马尔可夫博弈。但是，可以查看前一个框架作为MGs的特例，具有单态；这一框架中最早的泥灰岩理论仅限于小规模问题。另一方面，后一个框架中的泥灰岩本质上具有挑战性，一般而言，导致文献中相对稀缺的理论。由于篇幅有限，我们不介绍这些任何细节的模型。我们将简要回顾其中一些模型下的MARL算法，尤其是§4中部分观察到的设置。感兴趣的读者请参阅早期回顾更多关于泥灰岩正常形态/重复游戏的讨论。

**互评人：2052225 张勤杭 互评成绩：8分**