面向6G通信的离线多智能体强化学习：基础、应用与未来方向

ESLAM ELDEEB 和 HIRLEY ALVES

arXiv:2601.00321v1 [cs.MA] 1 Jan 2026

摘要——**包括超越5G和6G网络在内的下一代无线技术，正在为车辆编队、智慧城市和远程手术等变革性应用铺平道路。这些创新由大量相互连接的无线实体驱动，包括物联网设备、接入点、无人机和网联自动驾驶车辆；这些实体增加了网络复杂性，并对更先进的决策算法提出了更高要求。人工智能（AI）和机器学习（ML），尤其是强化学习（RL），是此类网络的关键赋能技术，可为高维度和复杂挑战提供解决方案。然而，随着网络扩展至多智能体环境，传统的在线强化学习方法面临成本、安全性和可扩展性方面的限制。离线多智能体强化学习（MARL）通过利用预先收集的数据，减少对实时交互的需求，提供了一种颇具前景的解决方案。本文介绍了一种基于保守 Q 学习（CQL）的新型离线多智能体强化学习算法，以确保训练过程安全且高效。我们进一步结合元学习来应对动态环境，并通过无线电资源管理和无人机网络中的应用场景验证该方法。我们的研究突出了离线多智能体强化学习在无线应用中的优势、局限性及未来发展方向。**

I. 引言

新一代无线技术和 6G 时代带来了车辆编队、智慧城市和远程外科手术等先进应用场景 [1]。这些应用依赖于各种无线实体的大规模部署，包括低功耗物联网设备、基站（BS）、无人机（UAV）以及网联自动驾驶车辆（CAV）。这一进展伴随着网络复杂性的增加和非线性参数的增多。因此，新一代无线技术需要高效的决策算法来取代传统方法，因为传统方法往往无法满足这些新兴应用的需求。

人工智能（AI）和机器学习（ML）有望成为下一代5G及后5G时代与6G网络的关键赋能技术。借助先进的深度神经网络和强大的函数逼近器，这些技术能够应对复杂的高维挑战。强化学习（RL）是机器学习领域强大的决策工具。它具有很强的适应性，因此非常适用于无线网络，能够在分解为离散时间步的序列环境中发挥优势 [2]。值得注意的是，

本研究由芬兰研究委员会资助的 6G Flagship 项目（资助编号 369116）提供支持。作者来自芬兰奥卢大学无线通信中心（CWC）。电子邮箱：firstname.lastname@oulu.fi。

强化学习采用无模型算法，无需依赖僵化的闭式模型，便能处理复杂的网络场景 [3]。

随着无线网络持续扩展并支持日益丰富的应用，强化学习已从单智能体系统发展到多智能体强化学习（MARL），其中多个实体在同一环境中交互 [4], [5]。例如，设想在智慧农业场景中，一群无人机为大规模部署的物联网设备提供支持，如图 1 所示。同样，智慧医院预计将引入自主机器人进行外科手术和医疗程序，从而变革医疗服务提供。在工业物联网领域，还有另一种很有前景的应用场景：通过多个接入点相连的自主机器人和传感器网络协作完成复杂任务和测量。为满足这些场景的需求，高效的无线电资源管理（RRM）至关重要。

尽管多智能体强化学习算法非常适合应对这些先进应用，但其在现实世界中的部署仍局限于研究环境。针对无线领域开发的大多数多智能体强化学习算法，包括离策略深度多智能体强化学习方法，都是依赖持续与环境交互的在线算法。然而，这种方法面临着重大挑战 [6]。在线数据收集既耗时又昂贵，在多智能体场景中尤为如此，因为所需数据量会大幅增加。此外，在线交互本身也可能存在安全隐患，因为智能体可能会采取探索性决策，进而导致严重故障或灾难性后果。

离线多智能体强化学习已成为传统在线多智能体强化学习的一种有前景的替代方案。这种方法使用静态离线数据集来优化策略，无需与环境进行实时交互。这类数据集通常通过行为策略生成。然而，离线多智能体强化学习面临的一项关键挑战是分布偏移，即数据集中的经验与算法所学习的经验之间存在不匹配。值得庆幸的是，保守 Q 学习（保守Q学习）等离线多智能体强化学习算法会在学习过程中尽量减小未见经验的影响，从而解决这一问题 [7]。在多智能体场景中，离线多智能体强化学习利用预先收集的数据，能够显著降低时间和计算资源方面的需求，因而极具价值。

本文探讨了离线多智能体强化学习在下一代无线网络中的潜力，重点关注其在无需高成本在线交互的情况下确保安全训练的能力。我们将保守Q学习扩展至多智能体设定，以应对无线环境中至关重要的分布偏移、协同和计算权衡，

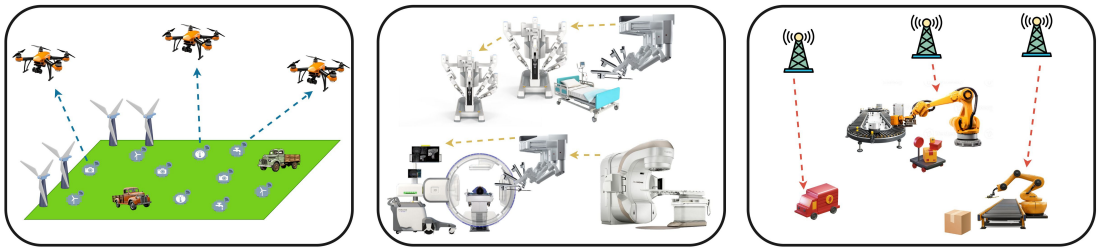

图 1. 新兴应用示意图　　面向多个智能体的 6G 网络新兴应用，包括智慧农业、智慧医院和工业物联网。

这些问题在无线环境中至关重要。此外，我们将元学习（模型无关元学习）与离线多智能体强化学习相结合，使其能够在网络目标或条件发生变化的动态场景中快速适应。我们通过无线电资源管理和无人机轨迹规划的案例研究验证了所提出的框架，证明了其实用性和有效性。本文的主要贡献如下：

- 我们详细概述无线网络中的强化学习和多智能体强化学习，重点阐述其在实际部署中的局限，尤其是在安全性、可扩展性和数据收集方面。
- 我们通过独立和集中训练、分散执行两种形式，将保守Q学习扩展到多智能体场景，以应对无线系统特有的分布偏移、协作和计算权衡问题。
- 我们将基于模型无关元学习的元学习与离线多智能体强化学习相结合，以提升在动态环境中的可扩展性和泛化能力，从而在目标变化时实现更快的适应。
- 我们通过无线领域的案例研究验证所提出的框架，研究内容涵盖无线电资源管理和无人机轨迹规划，并以开放挑战和未来研究方向作结。

II. 多智能体强化学习

A. 强化学习

在传统的强化学习框架中，问题通常被建模为马尔可夫决策过程（MDP）：智能体在环境中探索，观察环境状态，采取动作，随后环境会根据该状态下所采取的动作给予相应奖励。智能体采取恰当动作时会获得较高奖励，而选择不恰当的动作则会受到惩罚。策略是用于定义智能体在环境中的行为方式的方案，即规定在每个状态下应选择什么动作。智能体的目标是找到最优策略，使长期回报（累积奖励）最大化。状态-动作价值函数（又称 Q 函数）是一种用于估计在某个状态下采取特定动作并随后遵循该策略所获得回报的函数（查找表）。

为找到最优策略，智能体需要求解最优Q函数，以估计状态-动作对的实际回报。为此，Q学习这一无模型强化学习算法利用贝尔曼理论迭代求解最优Q函数。Q学习依靠时序差分（TD），以实际的学习率在每个时间步访问并更新Q函数。尽管它在求解各种马尔可夫决策过程方面具有很高的适用性，但随着环境维度的增加，它需要访问环境中的大多数状态-动作对，因此其性能会受到限制。

深度强化学习是一种新兴的强化学习变体，它将传统强化学习算法（如 Q学习）与深度学习（如深度神经网络）相结合，以克服大型马尔可夫决策过程的维度灾难。深度强化学习采用深度神经网络作为函数逼近器，对 Q函数进行建模和估计。深度Q网络（DQN）凭借解决围棋和 Atari 等大型博弈的高效性，彻底改变了强化学习领域。它们依赖离策略技术，将访问过的经验存入回放记忆库，并在后续时间步对其进行采样，以改进真实 Q函数的估计。DQN 已成为解决近期大多数无线通信问题的核心要素，例如无人机网络、无线电资源管理，以及云边协同物联网中的联邦强化学习方法 [8]。

B. 多智能体强化学习

随着无线通信实体的发展以及新型复杂应用的出现，无线网络迫切需要分布式多智能体解决方案。在多智能体设定中，强化学习问题演变为多智能体强化学习问题，通常被表述为多智能体马尔可夫决策过程（MA-MDP）。在多智能体马尔可夫决策过程中，多个智能体在环境中运行，获取自身观测以及整体环境状态的子集，采取动作，并接收用于评估所采取动作的奖励 [9]。

多智能体强化学习问题根据以下三个参数而有所不同：

1) 环境状态：多智能体强化学习问题根据每个智能体所看到的观测进行分类：
   - **完全可观测多智能体强化学习：** 每个智能体都能完全访问环境的整体状态，即每个智能体都会综合考虑自身观测以及其他智能体的观测。
   - **部分可观测多智能体强化学习：** 每个智能体只能观测自身的内部观测；并且

2) 目标： 多智能体强化学习问题根据环境中智能体的总体目标来定义：

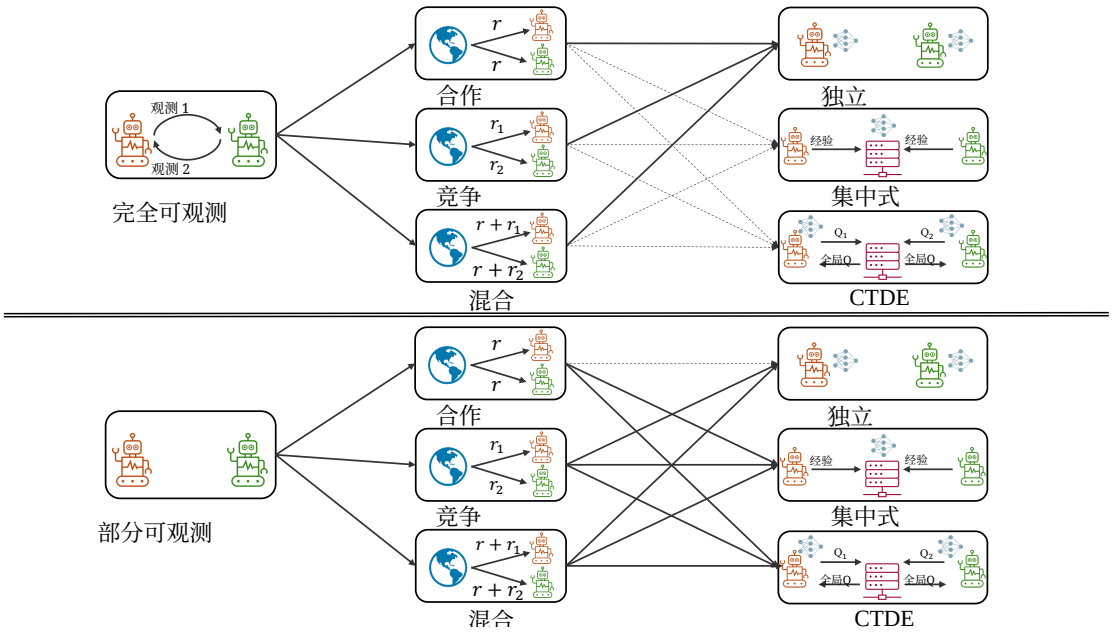

图2. 不同多智能体强化学习框架的示意图。多智能体强化学习问题因环境状态、目标和算法执行方式而异。实线表示较常采用的多智能体强化学习方案组合，虚线则表示较少采用的组合。该图还从系统层面概念性地展示了如何在无线通信和6G环境中组织多智能体强化学习训练。

- **合作多智能体强化学习：** 所有智能体为实现一个统一目标而合作。通常，所有智能体获得相同的奖励，该奖励由中央单元根据所有智能体的动作计算得出。

- **竞争性多智能体强化学习：** 智能体相互竞争，以最大化自身收益。每个智能体根据其动作获得单独的奖励。

- **混合合作与竞争多智能体强化学习：** 智能体在追求自身收益的同时，朝着共同目标进行合作。奖励由*i)*共同奖励部分和*ii)*个体收益组成；

3)算法执行：根据这一标准，多智能体强化学习问题可分为三类：

- **独立式多智能体强化学习：** 每个智能体独立优化其策略/Q函数。学习过程在每个智能体内部进行。

- **集中训练、分散执行（CTDE）多智能体强化学习：** 训练期间，每个智能体将其Q函数共享给一个中央单元，该单元利用各个Q函数估计全局Q函数。测试期间，每个智能体在内部遵循其Q函数。

- **集中式多智能体强化学习：** 所有状态和动作拼接在一起，并优化一个由各个Q函数构成的全局Q函数。学习过程在中央单元进行。

值得一提的是，集中训练、分散执行框架通常与部分可观测多智能体强化学习和合作多智能体强化学习结合使用，以克服智能体对其他智能体观测信息有限的问题。同时，独立式多智能体强化学习通常与完全可观测和合作式多智能体强化学习，或部分可观测和竞争式多智能体强化学习结合使用。图2总结了多智能体强化学习的各种形式。与单智能体 DQN 类似，多智能体 DQN（多智能体深度Q网络）中的每个 Q网络都由深度神经网络建模。除了分类之外，图2还从系统层面概念性地展示了无线网络中多智能体强化学习训练模式的组织方式，为 6G 部署场景提供了高层次指导。

因此，完全可观测多智能体强化学习比部分可观测多智能体强化学习更高效，但由于共享状态信息，需要承担更高的信令开销。同样，由于集中式情况下训练的全局Q函数，集中式多智能体强化学习比独立式多智能体强化学习更高效。然而，独立式多智能体强化学习由于信令开销较低，且学习任务分散在各智能体之间，消耗的通信和计算资源较少。集中训练、分散执行多智能体强化学习比独立式多智能体强化学习更高效，且比集中式多智能体强化学习消耗的资源更少。最后，相较于竞争型和混合合作与竞争多智能体强化学习（它们通常更常见于博弈中），合作型多智能体强化学习在无线网络中更为常见，因为无线环境中存在共同目标。

尽管针对各类多智能体强化学习变体开发高效算法取得了重大突破，但在实时无线通信应用中仍存在不足。首先，实时无线环境通常具有随机性且充满不确定性。因此，需要通过与环境进行在线交互来进行成本高昂的数据收集，而这可能不安全或不可行。其次，实时无线环境是动态的，网络配置会发生变化，例如信道特性、设备数量和网络目标。这些挑战促使人们开发能够在现实世界中安全、高效地应用的新型多智能体强化学习算法。

III. 离线多智能体强化学习

A. 离线多智能体强化学习

离线多智能体强化学习是多智能体强化学习的一个子领域，它利用离线固定数据集学习最优策略，而无需与环境进行任何在线交互。离线数据集通常使用行为策略收集，例如任何在线多智能体强化学习算法、经典算法或随机策略。使用离线数据集部署多智能体强化学习算法（如多智能体深度Q网络）通常会失败。这种失败源于数据集中收集策略与通过Q网络估计得出的学习策略之间的分布偏移。这种偏移会产生分布外（OOD）动作，即离线数据集中未得到充分覆盖的动作，从而导致乐观过高估计 [10]。在线多智能体深度Q网络中，回放缓冲区中收集到的分布外动作会通过与环境进行在线交互并准确估计这些动作的Q值来加以纠正。由于缺乏与环境的在线交互，离线多智能体深度Q网络则无法做到这一点。

离线强化学习和离线多智能体强化学习的近期进展通过明确约束动作，使其符合分布内数据，从而解决分布偏移问题。换言之，它们会惩罚分布外动作，以限制这些动作在Q函数更新中的权重。保守Q学习是一种离线强化学习/多智能体强化学习算法，它在奖励中加入正则化参数，使分布外动作产生的损失显著高于分布内动作 [11]。保守Q学习易于在深度Q网络的基础上实现，适用于任何多智能体强化学习变体。除了保守Q学习之外，行为克隆（BC）等更简单的方法会直接通过模仿数据集中的动作来学习策略，以牺牲有限的探索能力为代价换取稳定性。尽管这些方法并非本文的重点，但它们与基于保守Q学习的方法相辅相成，拓展了安全离线多智能体强化学习的工具箱。

与在线多智能体强化学习相比，离线多智能体强化学习通过避免代价高昂的实时探索，显著降低了样本复杂度，并且随着智能体数量增加，扩展能力更强。然而，离线多智能体强化学习仍依赖预先收集的数据集的多样性和质量，而在线多智能体强化学习则能借助持续的硬件反馈进行实时适应。这种权衡凸显了两种范式各自互补的优势。

B. 元学习

元学习是机器学习的一个分支，又称“学会学习”，旨在提升模型在全新、未见过的环境中的学习能力。元学习依赖于跨不同任务进行学习，这些任务被假定为来自同一任务分布，从而尽可能减少学习该分布中全新、未见过任务所需的数据和计算量。因此，相较于针对特定任务进行优化，在多个任务之间实现学习泛化，可以通过少量迭代和少量数据实现快速适应。因此，元学习可以减少所需的训练时长、计算能力和数据量。

MAML 是一种广为人知的元学习算法，它通过优化深度神经网络（例如 Q网络）的初始参数，使其能够快速适应并达到最优参数。以适应不同的任务或环境 [12]。元训练期间会抽取多个任务，并通过任务特定适应和元优化循环更新初始参数。元测试期间会抽取一个此前未见过的新任务，并在经过若干次迭代后更新已收敛的初始参数。MAML 支持少样本学习（即元测试阶段只有少量样本数据）和零样本学习（即没有可用数据）。

与单智能体情形相比，在多智能体设定中集成离线模型无关元学习会带来额外的理论和实践考量。首先，多智能体强化学习固有的非平稳性意味着每个智能体的学习都取决于其他智能体不断变化的策略，这进一步加剧了任务泛化的难度。通过将Q网络参数初始化为能够在多个相关任务上实现优化，模型无关元学习能够缓解这种非平稳性，并在网络条件变化时加速收敛到有效策略。其次，动态无线环境中的任务分布（例如信道条件或移动模式的变化）可能会显著偏移。在这种情况下，初始化仍可提供有利的偏差，有助于实现小样本适应，并增强应对分布偏移的稳健性。这凸显了模型无关元学习在动态多智能体强化学习任务中编码可迁移知识的能力。

总之，我们的框架解决了以往方法的若干局限。不同于依赖持续且往往不安全的探索的在线多智能体强化学习方法，所提出的离线多智能体强化学习方法仅依赖预先收集的数据集，从而确保训练过程更加安全。与经典的离线强化学习方法相比，我们将保守Q学习扩展到多智能体环境，从而缓解分布偏移和协作挑战。最后，通过整合元学习，我们的框架能够快速适应动态环境，而现有的多智能体强化学习或离线强化学习方法并不直接支持这一特性。具体而言，所提出的元训练过程通过在多个相关任务上进行训练，显式学习初始化参数；每个特定于任务的策略都会为优化能够在不同环境间泛化的共享参数集作出贡献。当智能体面临新的、未见过的目标或网络条件时，这种学习得到的初始化能够实现快速收敛和稳健适应。综上，这些特性使我们的方法成为适用于6G网络的一种安全、可扩展且适应性强的解决方案。

IV. 典型应用

本节基于保守Q学习框架，在三个不同的应用场景中评估所提出的离线多智能体强化学习算法：*i)* 无线电资源管理、*ii)* 无人机轨迹规划，以及*iii)*动态网络中的无人机轨迹规划。每个应用场景都代表一个部分可观测的协作式多智能体强化学习场景，其中智能体仅依靠各自的局部观测，共同实现一个目标。所提出的保守Q学习算法及所有应用场景中的Q函数均由包含两个隐藏层的神经网络实现，每个隐藏层包含256个神经元。我们将所提出的离线多智能体强化学习方法的两种变体——独立式保守Q学习（I-CQL）和集中训练、分散执行式保守Q学习（CTDE-CQL）——与基线方法进行性能比较，其中包括经典方法和

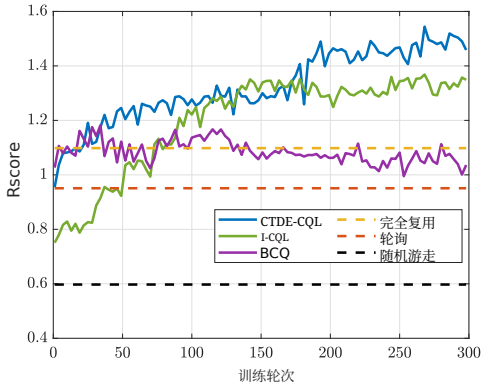

图 3. 在 100 个不同测试回合中，平均 R 分数随训练轮次数量的变化。

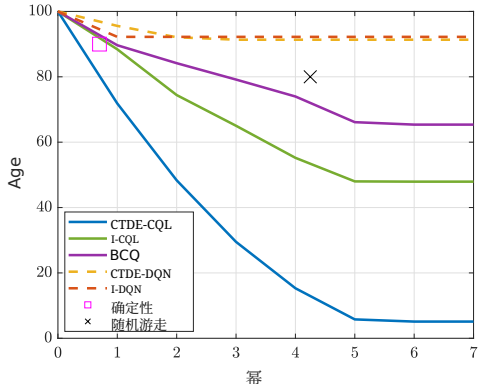

图4. 离线训练后使用不同功率因子所达到的信息年龄和传输功率。

离线多智能体强化学习方法，例如DQN和批约束Q学习（BCQ）。为提升所有实验中所收集数据集的时效性，每个回合都会随机设置设备位置、移动模式和信道条件，从而提供能够反映非平稳环境的多样化数据。

A. 案例研究 *1*：无线电资源管理

我们模拟一个 5000m × 5000m 的正方形区域，其中包含 4 个正在运行的接入点 (AP) 和 24 个用户设备 (UE)，历经 2000 个离散时间步。在所有回合中，UE / AP 的位置均随机生成，以反映实际部署中的移动性和异质性。每个 UE 以 1 m/s 的恒定速度随机移动，并在每个回合开始时分配给一个 AP。AP 根据比例公平 (PF) 因子对其关联的 UE 排序，该因子反映 UE 所体验到的服务质量和数据速率。每个 AP 观测其排名前 3 的关联 UE 的信干噪比 (SINR) 和 PF 值（局部观测），并从中选择一个 UE 提供服务（动作）。AP 协作以最大化 R-分数，该分数定义为 UE 的总和速率和第 5 百分位速率的加权和。因此，通用奖励被定义为所有 UE 的加权 PF 值和瞬时速率之和 [13]。我们将 R 分数定义为两个项的加权组合：所有用户的总和速率，以及第 5 百分位用户速率，后者反映用户之间的公平性。在评估无线电资源管理性能时，该指标兼顾了吞吐量和公平性。

我们使用在线软演员-评论家（SAC）算法获得的经验中的 20% 来生成离线数据集。选择这一比例是为了模拟现实中的离线数据收集场景，在这种场景下，大规模在线交互的成本高昂且可能存在安全风险。这一选择也表明，所提出的方法在数据质量各异且数据可用性有限的情况下仍能表现良好。我们将所提出的方法与三种知名基准方法进行比较：完全复用（full-reuse），始终选择 PF 因子最高的设备；轮询（round robin），在用户之间均匀分配资源；以及随机游走（random-walk），随机选择动作。如图 3 所示，CTDE-CQL算法优于所有基线方法，包括 I-CQL 和 BCQ。这归因于 CTDE-CQL 依赖智能体之间的 Q函数共享，从而改进了各个智能体的策略优化。相比之下，I-CQL 没有利用其他智能体的经验，导致训练性能下降。

B. 案例研究*2*：无人机轨迹规划

在此用例中，我们模拟了一个面积为1100 m × 1100 m的区域，其中随机部署了15个需要将数据传输给3架无人机的低功耗物联网设备，这些无人机以100m的固定高度飞行。在所有回合中，设备位置和无人机初始位置均随机生成，以反映实际场景中的移动性和可扩展性。主要目标是协同设计无人机轨迹，以联合最小化信息年龄（AoI）和设备的传输功率 [6]。每架无人机观测其当前位置以及每个设备的信息年龄（局部观测），并选择一个动作，该动作包括移动方向 {北、南、东、西、悬停} 以及接收数据的特定设备。奖励函数由所有设备的平均信息新鲜度和总传输功率共同构成，并通过功率因子进行缩放，以平衡信息年龄与传输功率之间的权衡。我们使用从在线DQN智能体获取的经验中的20%生成离线数据集。

本实验评估了所提出的 I-CQL 和 CTDE-CQL算法在训练后以及测试期间经过100个时间步后，在特定功率水平下实现的信息年龄。我们将它们的性能与相应的 I-DQN 和 CTDE-DQN 进行比较；后两者基于 DQN架构构建。此外，我们还与确定性模型进行比较，在该模型中，无人机沿着连接各设备的预设路径移动；并与随机游走模型进行比较，该模型随机选择动作。在图4中，坐标轴表示平均信息新鲜度与总传输功率的联合优化结果，其中奖励函数中的功率因子控制二者之间的权衡。如图4所示，确定性模型通过沿固定路径朝设备移动来节省能量，但由于时序效率低下，导致信息年龄非常高。类似地，随机游走模型任意随机选择动作，导致信息年龄和功耗都很高。基线模型 I-DQN

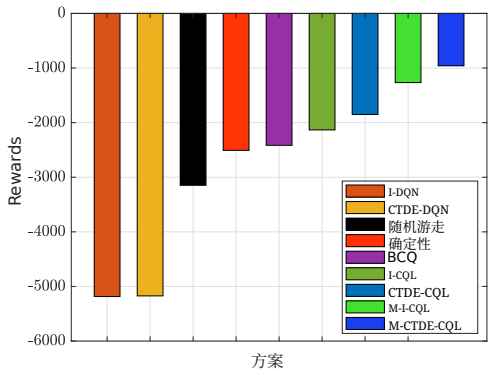

图5. 在一个采用独特功率因子的全新任务上，100个不同测试回合的收敛平均奖励。

和 CTDE-DQN 难以应对离线数据集中的分布偏移，因此其信息年龄与功率的综合表现比确定性模型和随机游走模型更差。同样，BCQ 得到的年龄-功率区域也并非最优。相比之下，所提出的 I-CQL 和 CTDE-CQL算法能够有效解决分布偏移问题，同时实现较低的信息年龄和功耗。与前一项实验一致，CTDE-CQL通过在智能体之间共享Q函数而优于 I-CQL。

C. 案例研究*3*：动态网络中的无人机轨迹规划

在本实验中，我们通过定义多个任务（环境）扩展了前述系统模型，其中每个任务都具有由奖励函数中不同功率因子决定的独特目标，从而产生不同的信息年龄–功率权衡 [14]。这构成了相互关联但并不完全相同的元训练环境，使该框架能够捕捉可迁移的结构。目标是为每个智能体确定初始Q网络权重，使其能够以尽可能少的训练迭代和较小的数据集快速适应新任务。基线模型与前一项实验中的模型相似：I-DQN、CTDE-DQN、确定性方法和随机游走。此外，我们还纳入了两种经元学习增强的模型M-I-CQL和M-CTDE-CQL，它们将模型无关元学习与保守Q学习相结合。评估指标是在使用仅由在线DQN智能体收集的5%经验构成的离线数据集训练20轮次后获得的奖励。具体实现中，每个任务都由具有不同功率因子的奖励定义，以表示信息年龄与功率之间的权衡。在元训练过程中，智能体针对采样任务调整其Q网络，并聚合更新后的参数以改进初始化。在元测试过程中，引入一个新任务，智能体从该初始化出发进行微调。这一设置表明，与随机初始化相比，模型无关元学习有助于多智能体强化学习实现更快、更稳定的适应。

如图5所示，I-DQN和CTDE-DQN因分布偏移问题而无法有效运行，而确定性模型和随机游走模型的奖励也相对较低。与前面的实验类似，BCQ取得了中等程度的次优表现。所提出的I-CQL和CTDE-CQL算法收敛表现一般，受限于训练轮次较少和数据集规模较小。这些模型随机初始化Q函数权重。相比之下，M-I-CQL和M-CTDE-CQL通过利用相似任务间的Q网络初始化（此处通过5个训练任务进行模拟），取得了高于所有其他方案的奖励。这种初始化使权重能够在几个训练步骤内优化并收敛。其中，M-CTDE-CQL表现最佳，实现了最优策略；由于智能体之间共享Q网络，其表现优于M-I-CQL。

V. 开放性挑战

尽管离线多智能体强化学习在无线领域具有广阔的应用潜力，但仍有若干挑战需要进一步探索：

- **人工智能可解释性：**随着人工智能被部署到自主无人机等实时应用中，提升人工智能可解释性至关重要。挑战在于开发透明的模型，使其能够解释强化学习智能体在动态环境中做出的决策，同时不牺牲性能。未来的工作应重点关注可解释模型和实时解释，在适应环境变化的同时保持清晰性。

- **离线数据的可扩展性：**随着智能体和网络实体数量的增加，确保收集到多样且具有代表性的离线数据仍是一项重大挑战。必须解决数据收集机制的可扩展性及其与训练过程的整合问题。此外，还应纳入对数据集规模和任务采样策略的敏感性分析，因为二者都可能在实际实现中引入偏差或影响稳定性。

- **离线-在线混合学习：**尽管完全离线的多智能体强化学习能够确保训练安全，但现实世界的 6G 环境通常需要在线微调，以适应未见过的条件。将离线预训练与轻量级在线适应相结合，对于边缘计算和无人机场景尤其重要，因为移动性、干扰和工作负载等动态因素可能快速变化。未来的研究应探索持续数据集更新和离线–在线混合学习等机制，以确保部署后的策略相关性。

- **任务分布：**在元学习中，通常假设任务是从相似的分布中采样的，但这一假设在现实应用中未必总是成立。挑战在于开发能够有效泛化的元学习模型，尤其是在动态环境中任务差异显著时。未来的研究应着重于放宽这一假设，并设计出能够适应多样且不断变化的任务分布、同时保持性能的元学习算法。

- **安全与探索：**尽管保守Q学习等离线多智能体强化学习算法通过限制未见经验的影响来解决部分安全问题，但仍需开展进一步研究，以确保这些算法在动态环境中不断演进和适应时仍然安全。

- **离线数据收集中的生成式人工智能：**生成式人工智能可以为离线强化学习创建合成数据，但确保这些数据的质量和多样性仍是一项重大挑战 [15]。未来的研究应着重改进生成模型，使其能够生成真实、多样的数据集，捕捉罕见事件并避免偏差，尤其是在复杂的多智能体环境中。

- **基础模型增强的强化学习：** 将大语言模型和多模态基础模型等预训练大型模型与强化学习相结合，可增强智能体在6G网络中解读语义信息和理解意图感知指令的能力。这种结合对于任务关键型低时延场景尤为重要，但扩展到多智能体系统仍然面临挑战。未来研究应探索通信协议，以及无需大量重新训练的高效适应方法。

VI. 结束语

本文重点介绍了离线多智能体强化学习在应对下一代无线网络（包括超越5G和6G）所带来挑战方面的变革潜力。我们借助基于保守Q学习（CQL）构建并进一步融入元学习技术的创新离线多智能体强化学习框架，提出了一种适用于动态多智能体环境、具备可扩展性和适应性的解决方案。该方法无需成本高昂且不安全的在线数据收集，同时能够在各种场景中快速优化策略。

通过全面的案例研究，包括静态和动态网络环境下的无线电资源管理与无人机轨迹规划，我们验证了所提出框架的有效性。结果表明，该框架能够有效优化复杂的高维问题，性能优于传统基准方法和现有的多智能体强化学习方法。具体而言，融入元学习展现出良好的前景，使智能体能够利用有限的训练数据和计算资源，快速适应新任务。

尽管取得了这些进展，仍有若干未解决的挑战。离线多智能体强化学习仍面临数据可扩展性、确保在多样化网络条件下实现稳健泛化，以及在安全性与探索之间取得平衡等问题。当前离线数据集的局限性和多智能体系统固有的复杂性凸显了进一步提升算法效率与性能、开发创新解决方案的必要性。此外，对任务分布假设的依赖，以及在决策过程中融入可解释性，都是未来探索的关键领域。未来的研究可以纳入更具挑战性的网络特性，例如时变的智能体间通信链路，以进一步验证所提出的框架。

总之，离线多智能体强化学习代表着将人工智能应用于无线网络的一项重大进展，为满足日益复杂且资源需求日益增长的应用场景提供了途径。通过持续应对尚未解决的挑战，离线多智能体强化学习能够在塑造无线技术的未来，以及推动智慧城市、自主系统和工业物联网等领域的创新方面发挥关键作用。

参考文献

[1] M. Z. Chowdhury, M. Shahjalal, S. Ahmed, and Y. M. Jang, “6G无线通信系统：应用、需求、技术、挑战与研究方向，” *IEEE*通信学会开放期刊，第1卷，第957–975页，2020年。[2] T. Li, K. Zhu, N. C. Luong, D. Niyato, Q. Wu, Y. Zhang, and B. Chen, “多智能体强化学习在未来互联网中的应用：全面综述，” *IEEE*通信调查与教程，第24卷，第2期，第1240–1279页，2022年。[3] P. Cheng, Y. Chen, M. Ding, Z. Chen, S. Liu, and Y.-P. P. Chen, “面向物联网网络在线资源分配的深度强化学习：技术、发展与未来挑战，” *IEEE*通信杂志，第61卷，第6期，第111–117页，2023年。[4] A. Feriani and E. Hossain, “面向人工智能赋能无线网络的单智能体与多智能体深度强化学习：教程，” *IEEE*通信调查与教程，第23卷，第2期，第1226–1252页，2021年。[5] J. Cui, Y. Liu, and A. Nallanathan, “基于多智能体强化学习的无人机网络资源分配，” *IEEE*无线通信汇刊，第19卷，第2期，第729–743页，2020年。[6] E. Eldeeb, H. Sifaou, O. Simeone, M. Shehab, and H. Alves, “保守且具风险意识的离线多智能体强化学习，” *IEEE*认知通信与网络汇刊，第1–1页，2024年。[7] S. Levine, A. Kumar, G. Tucker, and J. Fu, “离线强化学习：教程、综述与开放问题展望，” 2020年。 [ ] 2005 01643 在线]。可访问：https://arxiv.org/abs/。[8] D. Gan, X. Ge, and Q. Li, “一种基于最优传输的联邦强化学习方法，用于云边协同物联网中的资源分配，” *IEEE*物联网期刊，第11卷，第2期，第2407–2419页，2024年。[9] D. Lee, N. He, P. Kamalaruban, and V. Cevher, “强化学习优化：从单智能体到协作智能体，” *IEEE* 信号处理杂志，第 37 卷，第 3 期，第 123–135 页，2020 年。[10] W.-C. Tseng, T.-H. J. Wang, Y.-C. Lin, and P. Isola, “基于知识蒸馏的离线多智能体强化学习，”载于神经信息处理系统进展，第35卷。Curran Associates, Inc.，2022年，第226–237页。[11] A. Kumar, A. Zhou, G. Tucker, and S. Levine, “用于离线强化学习的保守 Q 学习，”载于神经信息处理系统进展，第33卷。Curran Associates, Inc.，2020年，第1179–1191页。 [在线]。可访问：https://proceedings.neurips.cc/paper files/paper/2020/file/0d2b2061826a5df3221116a5085a6052-Paper.pdf [12] S. L. Chelsea Finn, Pieter Abbeel, “用于深度网络快速适应的模型无关元学习，”第*34*届国际机器学习会议，第70卷，第1126–1135页，2017年。[13] E. Eldeeb and H. Alves, “一种用于无线电资源管理的离线多智能体强化学习框架，”2025年。 [在线]。可访问：https://arxiv.org/abs/ 2501 12991 。——，“用于及时无人机路径规划和数据收集的多智能体元离线强化学习，”2025年。 [在线]。可访问：https://arxiv.org/abs/ 2501 16098 。[14] J. Chen, B. Ganguly, Y. Xu, Y. Mei, T. Lan, and V. Aggarwal, “用于离线策略学习的深度生成模型：教程、综述与未来方向展望，”2024年。 [在线]。可访问：https://arxiv.org/abs/2402.13777

传记

ESLAM ELDEEB 于2019年获得埃及亚历山大大学电气工程理学学士学位，并于2021年获得芬兰奥卢大学理学硕士学位，随后于2022年取得博士学位。2023年期间，Eslam 曾在伦敦国王学院担任访问研究员。他目前主要从事大规模连接和超可靠低时延通信方面的研究。他的研究兴趣包括机器型通信以及面向无线通信网络的机器学习。

HIRLEY ALVES 是奥卢大学无线通信中心机器型无线通信方向的副教授，并在6G旗舰计划中负责大规模无线自动化主题。他在巴西联邦技术大学-巴拉那州立大学获得电气工程理学学士和理学硕士学位，并持有奥卢大学与UTFPR联合授予的理学博士学位。他的研究兴趣包括大规模和关键机器型通信、卫星物联网、分布式处理与学习、超越5G以及6G技术。