开放式无线接入网中感知稀疏性的大规模智能随机接入控制：一种基于强化学习的方法

Xiao Tang、Sicong Liu（IEEE高级会员）、Xiaojiang Du（IEEE会士）和 Mohsen Guizani（IEEE会士）

arXiv:2303.02657v1 [cs.LG] 5 Mar 2023

摘要——新兴开放式无线接入网（O-RAN）中的设备大规模随机接入给接入控制与管理带来了巨大挑战。利用接入请求的突发特性，稀疏活跃用户检测（SAUD）可有效提升接入管理效率，但在海量接入请求缺乏协调的情况下，稀疏性可能会下降。为动态维持接入请求的稀疏性，本文提出一种利用接入类限制技术的强化学习（RL）辅助闭环接入控制方案，其中，RL策略通过RL智能体（即下一代节点基站（gNB））与环境之间的持续交互来确定。该方案可由O-RAN中的近实时RAN智能控制器（近实时RIC）实现，支持在不同异构垂直行业应用之间快速切换，例如海量机器类通信（mMTC）和uRLLC服务。此外，本文提出一种数据驱动的深度强化学习辅助SAUD方案，以应对状态和动作空间连续且维度极高的复杂环境，并采用经验回放缓冲区自动收集大规模数据。本文构建了一个行动者-评论家框架，将策略学习模块纳入近实时RIC。仿真结果表明，与基准方案相比，所提方案在具有海量接入请求的不同异构服务中，均能实现更高的接入效率和用户检测准确率。

索引词——开放式无线接入网、活跃用户检测、冲突缓解、压缩感知、强化学习。

一、引言

近年来，国际标准化组织和全球移动运营商开始致力于新兴开放式无线接入网（O-RAN）架构的开发与应用。O-RAN依托网络功能虚拟化和软件定义网络技术，旨在提升无线接入网的可移植性和灵活性，并打造可互操作、可编程的无线接入网产品与服务 [1]。近实时无线接入网智能控制器（near-RT RIC）是实现本地O-RAN高效网络管理的关键 [2]。

对于 O-RAN 中的异构服务，例如无线指纹识别、切换和流量引导，多种第三方控制应用（即 xAPP）运行在近实时RIC上，为不同场景提供可编程的下一代节点基站（gNB）网络管理策略以实现服务质量（QoS）优化 [3], [4]。各种应用和服务，尤其是 mMTC服务，都要求支撑高密度、大规模设备（$10^2/\text{km}^2$ 至 $10^7/\text{km}^2$）接入无线接入网，这给随机接入管理 [5], [6]带来了巨大挑战。在用户设备成功与核心网建立连接之前，gNB需要从所有可能的潜在用户中检测出活跃用户，以便向正确的用户发送响应。活跃用户检测（AUD）问题已在O-RAN工作组3中得到研究。

幸运的是，大规模用户的连接请求具有突发性和稀疏性，即在特定时隙内，需要发起网络接入请求的活跃用户仅占该网络中所有潜在用户的一小部分。这种稀疏性使一种高效技术——基于压缩感知（CS）的活跃用户检测（AUD）——得以应用，该技术也称为稀疏活跃用户检测（SAUD） [7]–[10]。然而，如果接入请求的稀疏性降低，SAUD 的性能可能会下降 [11]。例如，当接入请求数量快速增加，或网络同时支持多种异构服务时 [12],，活跃用户之间可能会发生更多冲突，从而破坏稀疏性，并对 AUD 的准确性产生不利影响。

因此，有必要设计一种有效的范式，以适当地管理和动态控制网络中用户的接入请求，从而保持接入请求的稀疏性。作为 3GPP 中的标准流量控制机制，接入类限制（ACB）技术为某些特定类别的用户规定了相应的接入优先级。当网络过载时，通过设置 ACB因子并实施基于优先级的接入请求，该技术能够有效减轻网络负担，从而在严峻条件下提升 SAUD 的性能 [13], [14]。

另一方面，随着 O-RAN 的出现，为异构垂直行业服务提供了一种全新且高效的智能接入控制方案。利用网络切片，可以将一个独立的 O-RAN 划分为多个逻辑网络，从而使近实时RIC能够根据不同的服务需求实施特定的接入控制策略。例如，车联网等 uRLLC服务对连接的可靠性更加敏感，因此更加关注 AUD 的准确性；而物联网等 mMTC服务则倾向于让更多用户接入，即使需要以牺牲可靠性为代价。网络应能够在不同的异构服务之间自适应地进行软切换，同时还要应对时变的

X. Tang和S. Liu任职于中国厦门361005厦门大学信息学院信息与通信工程系（电子邮箱：liusc@xmu.edu.cn）。X. Du任职于美国新泽西州霍博肯07030史蒂文斯理工学院电气与计算机工程系（电子邮箱：dxj@ieee.org）。Mohsen Guizani任职于阿联酋穆罕默德·本·扎耶德人工智能大学（电子邮箱：mguizani@ieee.org）。（通讯作者：Sicong Liu）

特定服务中的通信环境。

因此，为了满足多样化的需求，部署在近实时RIC上的xAPP需要自适应地调整其配置，这要求采用动态闭环决策与控制方案。机器学习（ML）和强化学习（RL）技术可以引入O-RAN [15],[16]，以便在gNB侧构建具备可编程自定义逻辑、由数据驱动的智能体，使其能够主动与时变环境交互，并针对不同业务动态配置网络接入控制策略。

为此，本文提出一种基于强化学习的框架，部署在近实时RIC上，通过自适应控制ACB因子实现动态接入管理。具体而言，本文设计了一种强化学习辅助的稀疏活跃用户检测（RL-SAUD）方案，该方案采用Q学习，通过试错获取经验并学习适当的决策策略。ACB控制过程及其与环境的闭环交互可以建模为马尔可夫决策过程（MDP），从而动态更新近实时RIC的接入控制策略。在所提出的RL-SAUD方案中，不同的ACB因子被分配给不同的基于优先级的类别，从而在允许接入概率与用户检测准确率之间实现合理权衡。

此外，为了支持对连续且高维的状态和动作空间进行有效控制，我们设计了一种采用行动者-评论家框架的深度强化学习（DRL）辅助SAUD（DRL-SAUD），以减轻状态量化和Q学习中的维数灾难所带来的影响。此外，通过预训练获取先前经验并将其存储在经验回放缓冲区中，再利用这些经验初始化深度神经网络的参数，从而加快收敛速度。随后，通过gNB智能体与网络环境之间的闭环交互，可近实时地获得接入控制的最优策略。理论分析和仿真实验均验证了所提方案的优越性。因此，本文的主要贡献总结如下。

- 构建了一种基于ACB和SAUD的大规模接入控制框架，并将其部署于O-RAN近实时RIC的xAPP中。该框架充分利用接入请求固有的稀疏性，从网络中的海量潜在用户中高效、准确地检测出活跃用户。

- 通过利用Q学习对ACB因子进行自适应闭环控制，设计了一种强化学习辅助的稀疏活跃用户检测（RL-SAUD），以实现动态接入管理。该方法能够自适应地保持接入请求的稀疏性，避免因海量接入请求或异构服务引发的严重冲突而导致稀疏性恶化，从而维持SAUD的准确性。

- 此外，本文提出一种内置行动者-评论家模块的深度强化学习辅助稀疏活跃用户检测方案（DRL-SAUD），利用先前经验加快收敛。通过训练深度神经网络实现的数据驱动范式，能够解决性能受连续状态和动作空间的量化误差及维数灾难所限的问题。

本文其余部分的组织结构如下。第二节回顾相关研究工作。第三节介绍O-RAN环境下大规模随机接入的系统模型。第四节和第五节分别详细描述所提出的、用于实现智能动态ACB因子和大规模接入控制的RL-SAUD与DRL-SAUD方案。第六节给出理论性能分析。第七节报告仿真结果，最后在第八节给出结论。

II. 相关工作

在无竞争随机接入范式中，gNB预先为潜在用户分配特定的前导序列，以加快服务恢复的平均速度[17]。有意接入网络的用户通过物理随机接入信道发送其前导码，gNB则根据已知的前导码与用户之间的分配关系检测活跃用户。该过程称为AUD，旨在检测出正确的用户，以建立随机接入连接并进行信息传输。O-RAN在现有网络接口协议的基础上不断发展，并兼容更多异构服务，因此，AUD已成为一项关键挑战。为实现可接受的AUD性能，研究人员已经探索了多种技术，例如基于接收波束成形的盲多用户检测[18] 以及小区间的协同检测 [19]。

通常，发起接入请求的活跃用户数量远少于网络中的潜在用户总数，这表明活跃流量具有稀疏性。流量的稀疏性使 AUD 问题转变为稀疏恢复问题，该问题可在压缩感知框架中表示，即 SAUD [5], [20]。因此，可以采用一些稀疏恢复方法，例如结合块稀疏贝叶斯学习的近似消息传递 [21] 以及利用多测量向量进行联合稀疏支撑集恢复 [22] ，从而降低海量接入请求情况下 AUD 的复杂度。然而，随着 O-RAN 中数据流量的激增，SAUD 的性能仍可能出现下降。与纯硬件处理相比，更好的方法是通过流量控制保留接入请求的稀疏性，以满足异构服务的服务质量需求 [23],[24]。

根据时延、连接可靠性、数据速率等方面具有不同服务质量需求的多种通信场景，用户的接入优先级可划分为多个类别。基于接入优先级的适当接入管理方案（如ACB）可以缓解流量冲突，并已广泛应用于过载网络中 [11], [13], [25], [26]。扩展随机接入方案为通过 ACB检查的设备提供接入链路，并在有额外资源可用时授予更多接入链路 [13]。此外，在资源配置信息可用时，可以推导出最优ACB因子 [25]。固定ACB控制方案无法满足

时变环境中的接入管理需求。基于优先级的ACB控制方案根据时延要求将用户划分为多个类别，并根据优先级动态调整ACB因子，以确保时延敏感服务的服务质量 [11]。自适应学习自动机方案能够充分挖掘与环境进行闭环交互的潜力，并实现对流量的有效推理和控制 [26]。

O-RAN 的出现通过开放接口释放了智能网络的可运维性。具体而言，O-RAN 中的近实时无线智能控制器使智能接入控制机制能够为多样化服务提供定制化服务。闭环控制使 xAPP 能够实时跟踪测量信息，并动态配置 gNB。在 O-RAN 中引入人工智能时，仍有一些关键问题需要解决，例如：在特定场景中应部署哪种模型、哪种数据集适合学习，以及如何为现实世界的系统设计复杂度最低的解决方案等。 [27], [28]。基于机器学习的自动化已使 O-RAN 模块能够通过黑盒模型中的内部处理与交互实现复杂控制 [29]。为了进一步使黑盒模型适用于多种场景，可以应用强化学习方法和智能体（即强化学习运维）来实现训练、测试和生命周期验证流程 [30]。强化学习运维的可行性和优越性已得到验证，并且可以轻松部署在 O-RAN 上 [31]。

通过这种方式，近实时RIC能够动态调整策略，以适应时变环境，并根据反馈信息快速、合理地做出决策。利用强化学习方法动态确定能够适应多种场景的随机接入策略，已成为一种颇具前景的解决方案 [32]–[36]。例如，文献 [32]展示了传统接入策略在大规模接入中的低效性，以及通过Q学习实现动态接入控制的可能性。文献 [33]提出了一种面向uRLLC服务的基于强化学习的方案，该方案能够在较长时间内最大限度地增加可稳定服务的节点数。文献 [34]，[35]设计了一些基于强化学习的方案，以便为各类服务自适应地分配随机接入资源。文献 [36]针对随机接入过程提出了一种支持Q学习、能够动态调整退避帧的方案，该方案可以缓解固定帧长下的接入冲突并提高网络吞吐量。

此外，在未来数据驱动的智能 O-RAN 中，需要进行精确且连续的控制，此时可以利用大规模数据集来训练增强型智能体。在这方面，可以采用广泛应用的深度神经网络和深度学习技术。经过适当构建的深度神经网络能够通过训练逼近复杂的非线性映射，并自动学习闭环控制策略中的隐式最优规则。因此，可以采用新兴且广受欢迎的深度强化学习（DRL）技术，该技术将深度神经网络与强化学习相结合，相关研究已十分广泛。例如，一种基于深度强化学习的多址接入方案以总吞吐量为目标函数，其中

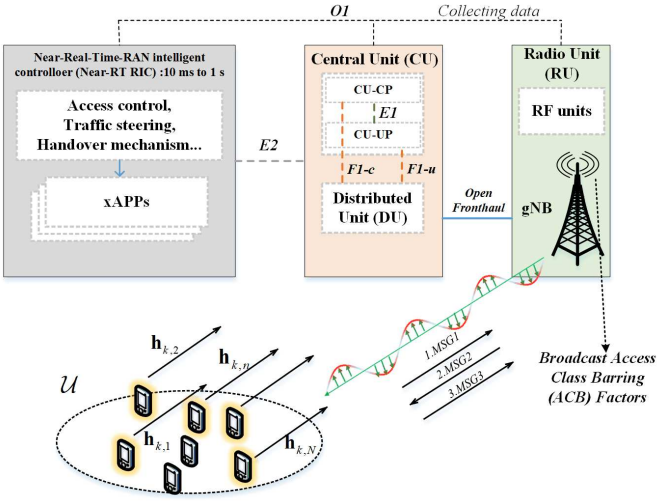


图1. O-RAN范式下的大规模随机接入模型：gNB作为智能体和物理层接入点，为大量具有不同优先级接入控制阻塞类别的用户提供服务；应具备自适应切换能力，以动态支撑异构垂直服务所需的不同服务质量和接入请求，以及时变信道环境和各种场景。

网络参数经过更新，以实现对接入时隙的连续且最优选择 [37]。基于深度 Q 网络（ DQN ）架构，一种智能节点接入控制方案能够避免频繁切换，并实现更优的控制性能 [38]。通过收集高维经验，基于 DQN 的方案能够在多种服务之间切换时自适应地确定接入竞争窗口 [39]。此外，对于异构网络中的接入控制问题，基于深度强化学习的算法能够适应不同 MAC协议和信道容量之间的不一致性 [40]。

III. 系统模型

本文所考虑的 O-RAN 范式下的大规模随机接入模型如图1所示。假设网络中共有 N 个潜在用户。令 $\mathcal{U}$ 表示所有潜在用户的集合，令 $\mathcal{U}_a$ 表示活跃用户的集合。活动指示器 α_n 用于指示第 n 个用户的活动状态，即 $\alpha_n = 1$ 表示第 n 个用户在当前时隙有接入网络的需求，否则其值为零。在与 gNB 建立连接的实际过程中，配备单天线的活跃用户会向无线电单元中配备 M 根天线的 gNB 发送一个唯一的前导码，该前导码由 gNB 预先分配给活跃用户。

在典型的无竞争随机接入链路建立过程中，活跃用户与 gNB 之间会发送三个分别称为 MSG1、MSG2 和 MSG3 的信令消息，用于传递活跃用户的唯一信息、gNB 的响应以及活跃用户的确认信息。具体而言，在时隙 t 中，活跃用户将其唯一的前导码 $\boldsymbol{\lambda}^t = [\lambda_1^t, ..., \lambda_N^t]^T$（即 MSG1）调制到$K$ 个 OFDM子载波上进行传输。

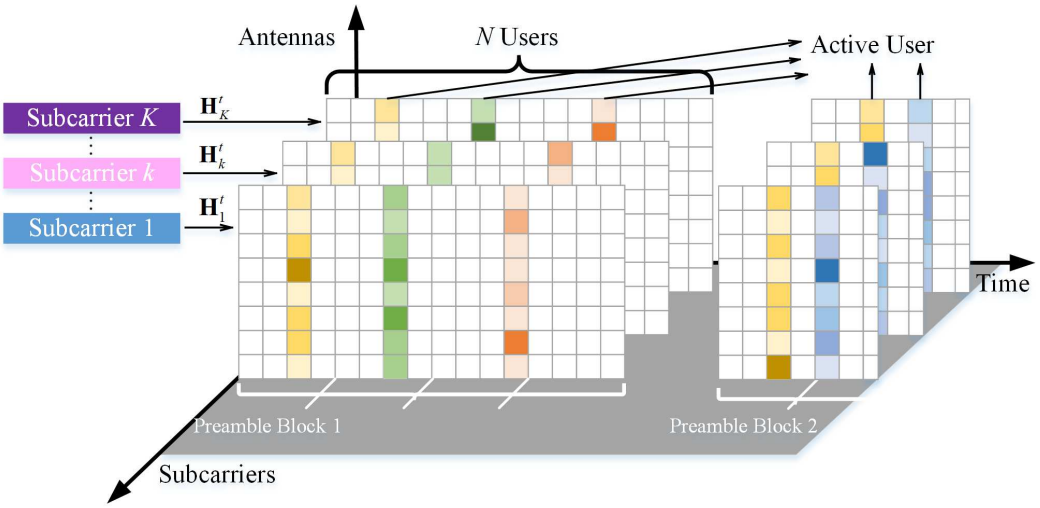


图 2. 大规模随机接入请求的可视化：具有接入请求的活跃用户发送的一系列前导码块，在时间、子载波和天线域中表示。每个前导码块的大小由网络中的潜在用户数量以及 gNB 的天线数量决定。在某个时频资源上，与网络中所有潜在用户的集合相比，具有接入请求的活跃用户子集呈现出稀疏特性。

从 N 个潜在用户到 gNB 的子载波 k 信道矩阵表示为 $\mathbf{H}_k^t = [\mathbf{h}_{k,1}^t, ..., \mathbf{h}_{k,N}^t]$，其中 $\mathbf{h}_{k,n}^t \in \mathbb{C}^M$ 表示从单天线用户 n 到 M天线 gNB 的信道响应向量。因此，gNB 在子载波 k 上接收到的测量向量 $\mathbf{y}_k^t \in \mathbb{C}^M$ 表示为

$$\mathbf{y}_k^t = \sum_{n=1}^{N} \alpha_n^t \lambda_n^t \mathbf{h}_{k,n}^t + \mathbf{z}^t = \underbrace{\mathbf{H}_k^t \mathbf{\Lambda}^t}_{\tilde{\mathbf{H}}_k^t} \boldsymbol{\alpha}^t + \mathbf{z}^t, \tag{1}$$

其中，$\mathbf{\Lambda}^t = \text{diag}\{\lambda_1^t, ..., \lambda_N^t\}$ 是一个对角矩阵，其对角元素由前导码向量 $\boldsymbol{\lambda}^t$构成，而 $\mathbf{z}^t \in \mathbb{C}^M$ 表示背景噪声。为简化符号表示，令 $\tilde{\mathbf{H}}_k^t$ 表示归一化信道矩阵 $\mathbf{H}_k^t\mathbf{\Lambda}^t$，该矩阵由原始信道矩阵 $\mathbf{H}_k^t$按前导码 $\mathbf{\Lambda}^t$归一化得到。值得注意的是，$\tilde{}$归一化信道矩阵 $\mathbf{H}_k^t$ 可视为压缩感知框架中的观测矩阵。如果式（1）中的未知活动指示向量 $\boldsymbol{\alpha}^t = [\alpha_1^t, ..., \alpha_N^t]^T$具有稀疏特性，即活跃用户数量远少于用户总数，并且观测矩阵满足受限等距性质，那么根据压缩感知理论 [5]，很有可能从测量向量 $\mathbf{y}_k^t$ 中以有界误差恢复该向量。随后，通过准确恢复活动指示向量 $\boldsymbol{\alpha}^t$ 来检测活跃用户。

如图 2 所示，具有接入请求的活跃用户在时间、子载波和天线域中发送一系列前导码块，从而实现大规模随机接入请求的机制。假设活跃用户的接入请求相对于网络中的所有潜在用户是稀疏的，则 SAUD 问题可以建模为一个稀疏恢复问题，以估计未知活动指示向量 $\boldsymbol{\alpha}^t$ 的支撑集，即非零元素的位置。为此，用于估计 $\boldsymbol{\alpha}^t$ 的稀疏恢复问题可以分解为针对每个子载波的 K 个子问题，如式（1）所示。具体而言，一个按子载波划分的活动指示向量，即 $\hat{\boldsymbol{\alpha}}_k^t = [\hat{\alpha}_{k,1}^t, ..., \hat{\alpha}_{k,N}^t]^T$，可以先通过针对每个子载波$k = 1 \cdots K$求解子问题 $\mathbf{y}_k^t = \tilde{\mathbf{H}}_k^t\boldsymbol{\alpha}_k^t + \mathbf{z}^t$ 来估计。随后，所估计的按子载波划分的活动指示向量 $\hat{\boldsymbol{\alpha}}_k^t$ 可作为投票器，对最终估计的活动指示向量 $\hat{\boldsymbol{\alpha}}^t$ 的非零位置进行投票：如果位置 n 获得的票数超过某一判决阈值 υ_{th}，则将 $\hat{\boldsymbol{\alpha}}^t$ 的第 n 个元素 $\hat{\alpha}_n^t$ 标记为 1，否则标记为 0，具体表示为

$$\hat{\alpha}_n^t = \begin{cases} 1, & \frac{1}{K}\sum_{k=1}^{K} \hat{\alpha}_{k,n}^t > \upsilon_{\text{th}}, \\ 0, & \text{otherwise}. \end{cases} \tag{2}$$

SAUD 通过式（2）检测到活跃用户后，gNB 向检测到的用户发送随机接入响应，即 MSG2。

活跃用户从发送前导码开始启动定时器窗口，以接收 gNB 的反馈。如果某个用户在该定时器窗口内成功解析出与之前 MSG1 对应的响应，则该设备会向 gNB 反馈确认字符（ACK），即 MSG3，表示连接已成功建立。否则，则视为接收失败。

IV. 强化学习辅助的稀疏活跃用户检测与 O-RAN 接入控制

SAUD的性能很大程度上取决于接入请求的稀疏性。因此，应谨慎、妥善地管理网络中潜在用户的接入流量控制。为此，本节提出一种强化学习辅助的稀疏活跃用户检测（RL-SAUD）方案，引入接入类别限制机制来协调用户的接入请求。该方案旨在合理、动态地调整ACB因子，在尽可能允许更多用户接入以满足

各种服务需求的同时，提高SAUD的检测准确率。本节将首先介绍O-RAN环境下用于流量控制的ACB机制，然后详细阐述所提出的RL-SAUD方案。

A. 用于流量控制的接入类限制（ACB）

需要注意的是，式（1）和式（2）所描述的 SAUD 问题的准确性与接入请求的稀疏度密切相关。在 O-RAN 的各种异构服务中，当并发接入请求数量极其庞大时，问题（1）中的活动指示向量可能会失去稀疏性，从而降低准确检测用户的概率，并导致接入失败。因此，gNB 需要采用合理的策略来控制接入流量，而 ACB 是实现这一目标的理想选择。

事实上，ACB 是一种流量控制机制，它根据用户的接入优先级将用户划分为多个类别。假设共有 L 个类别，类别 l 中的用户由集合 $\mathcal{U}_l$ 表示，该集合中的元素数量为 N_l。近实时 RIC 为这 L 个类别中的每个类别生成不同的 ACB 因子 $\{p_l \in [0,1]\}_{l=1}^{L}$，并通过称为 ACB检查的过程进行流量管理。具体而言，在 ACB检查过程中，某个有接入需求的活跃用户 $n \in \mathcal{U}_l$ 会在发送前导码之前随机抽取一个值 $q_n \in [0,1]$：只有当 $q_n \leq p_l$ 时，用户 n 才会发送前导码；否则，该用户将退避至预设范围内的随机抽样时间，并等待下一次接入类禁止检查过程 [41]。

在O-RAN中，各种异构服务的预训练模型存储于xAPP目录中，为智能ACB调节提供了便利。当活跃用户发起接入请求时，近实时RIC会在该目录中实例化其xAPP，并通过E2接口控制底层的集中式单元和分布式单元。最后，gNB广播ACB配置信息，以控制网络中的流量。在ACB流量控制机制的支撑下，可以合理地控制和协调同一时间段内接入网络的设备数量，这有助于保持活跃用户接入请求的稀疏性，即活动指示向量 $\boldsymbol{\alpha}^t$ 的稀疏性，如式（1）所示。

根据相关研究，SAUD的准确率随着接入请求数量的增加呈单调递减的凸曲线 [11]。在大量并发接入请求的情况下，如果不进行流量控制，即将ACB因子设为1，所有活跃用户都能通过接入类禁止检查过程并接入网络，这会使活动指示向量 $\boldsymbol{\alpha}^t$ 不再稀疏，从而导致严重冲突。在这种情况下，很难或无法根据式（1）中的接收信号准确检测并区分活跃用户。因此，应研究一种针对活跃用户的合理流量控制机制，以缓解冲突并保持接入活动的稀疏性，从而支持O-RAN中复杂时变环境下的大规模接入和多种异构服务共存。为此，我们设计了一种基于强化学习的闭环控制方案，用于调整ACB因子，并将其实时广播给网络中的所有用户，具体内容将在下一小节中详细介绍。

B. 面向大规模随机接入的强化学习辅助稀疏活跃用户检测（RL-SAUD）

首先，我们介绍智能体（即gNB）与环境（即无线接入网中的用户）之间的交互模型，该模型可视为一个马尔可夫决策过程（MDP），如图1所示。在O-RAN中，闭环控制实际上是一个MDP过程，其中xAPP在强化学习框架的支持下充当智能体，而近实时RIC则作为各种xAPP的容器。闭环接入控制过程中的交互细节如下所列。

- 近实时RIC通过开放前传接口发送接入类禁止向量 $\mathbf{p}^t$，gNB则在无线接入网中向属于不同接入类别的用户广播该向量。给定接入类禁止向量 $\mathbf{p}^t$，可以通过对所有LL类别进行汇总，近似估算允许接入的用户数量，即 $N^t_{\mathrm{p}} = \sum_{l=1} p_l^t N_l^t$。
- 通过ACB检查的活跃用户向gNB发送唯一的前导码。随后，gNB通过SAUD算法检测用户身份，并向检测到的用户发送MSG2。正确检测到的用户收到MSG2后，会向gNB返回MSG3，即确认。
- gNB通过统计反馈确认的数量，估算有效接入用户的近似数量 $N^t{}_{\mathrm{v}}$，并计算当前时隙 c^t 内的用户检测准确率 t。
- 如图1所示，近实时RIC通过E2接口提取当前环境的统计数据，更新xAPP中的流量管理策略，并执行该策略确定的动作，以控制下一时隙的ACB。

所提出的强化学习辅助 SAUD 算法总结于算法1。具体而言，为了应用 Q学习方法，类别 l 的 ACB因子被量化为 X_1 个离散等级，即 $p_l^t \in \Omega \triangleq \{i/X_1, 1 \leq i \leq X_1\}$，其中 Ω 是某一用户类别的可行动作集合，也就是 ACB因子的取值。选定的动作执行后（即广播给所有用户后），不同类别的用户会执行接入类禁止检查过程，从而生成一组活跃用户 $\mathcal{U}_a$以及相应的活动指示向量 $\boldsymbol{\alpha}^t$。然后，集合 $\mathcal{U}_a$ 中的活跃用户将与 gNB建立连接。gNB通过SAUD算法估计活动指示向量 $\hat{\boldsymbol{\alpha}}^t$，随后向检测到的活跃用户发送 MSG2。SAUD算法总结于算法2，其中 $[\tilde{\mathbf{H}}_k]^*$ 和 $[\tilde{\mathbf{H}}_k]^\dagger$ 分别表示式（1）中给出的子载波 k 的归一化信道矩阵 $\tilde{\mathbf{H}}_k$ 的共轭矩阵和摩尔-彭若斯伪逆。

在定时器窗口内成功接收 MSG2 的用户会向 gNB 返回确认，gNB 据此估计检测准确率。为便于

参考历史信息，基于强化学习的算法的状态由 ACB因子向量和前一时隙的检测准确率组成，即 $\mathbf{s}^t = [\mathbf{p}^{t-1}, c^{t-1}]$。每个状态-动作对都有一个 Q值，从而形成大小为 ${X_1}^L X_2 \times {X_1}^L$ 的 Q表，其中 X_2 是检测准确率的量化级别数 c。

智能体倾向于选择当前最优的动作，即选择使当前状态 $\mathbf{s}^t$下奖励值 $Q(\mathbf{s},\mathbf{p})$ 最大的 $\mathbf{p}^t = \mathbf{p}^*$。但这样可能会导致某些动作未被探索，从而陷入局部最优。为此，ϵ-贪心方法通过设置一个 ϵ 值，赋予采用随机策略一定的小概率，使每个可行动作都有机会被探索，其定义如下：

$$\Pr\left(\mathbf{p}^t = \mathbf{p}^*\right) = \begin{cases} 1-\epsilon, & \mathbf{p}^* = \arg\max\limits_{\mathbf{p}\in\Omega^L} Q\left(\mathbf{s}^t, \mathbf{p}\right), \\ \frac{\epsilon}{|X_1+1|}, & \text{otherwise.} \end{cases} \tag{3}$$

其中 Ω^L 为动作空间。根据式（3），选择 Q值最大的动作的概率为 $1-\epsilon$。

算法 1：强化学习辅助的流量控制稀疏活跃用户检测（RL-SAUD）

初始化：
Q表中的 Q值 $Q(\mathbf{s}^t, \mathbf{p}^t)$
初始化一个随机状态 $\mathbf{s}^0$
对于 $t = 1$、2、3.....，执行
  在状态 $\mathbf{s}^t$ 下，通过式 (3) 选择动作 $\mathbf{p}^t$
  计算 N^t_{p}
  活跃用户使用 $\mathbf{p}^t$ 执行 ACB检查，并若通过，则向 gNB 发送 MSG1
  gNB 在算法2中执行 SAUD，并反馈 MSG2
  活跃用户检测 MSG2，并将 MSG3 发送回 gNB
  gNB 统计有效接入用户的数量 N^t_{v} 通过 MSG3 中的 ACK确认
  计算检测准确率 $c^t = N^t_{\mathrm{v}}/N^t_{\mathrm{p}}$
  构建下一状态 $\mathbf{s}^{t+1} = [\mathbf{p}^t, c^t]$
  通过 (4) 获得当前系统效用 u^t
  使用 (5) 更新 Q表
结束

系统效用函数 u^t 旨在强化迭代学习过程中选择有利动作的策略，其定义如下

$$u^t = c^t \sum_{l=1}^{L} p_l^t r_l N_l^t - \rho_1 \frac{1}{L} \sum_{l=1}^{L} \left(p_l^t - \overline{p^t}\right)^2 - \rho_2 (1 - c^t), \tag{4}$$

其中，r_l 是类别 l的接入优先级分数，分数越高表示接入优先级越高。公式（4）右侧的第一项表示按接入优先级分数加权的有效接入用户数量。直观而言，如果允许更多接入优先级分数较高的用户接入，并且能够准确检测到这些用户（即有效接入），那么系统效用就应当提高。第二项和

算法2：稀疏活跃用户检测（SAUD）

输入：
信道矩阵 $\{\quad \tilde{\mathbf{H}}_k\}_{1\le k\le K}$
接收信号 $\{\mathbf{y}_k\}_{1\le k\le K}$ 和步长 s
对于 $k = 1$ 2..3..执行
  初始化：’ ’
  $F_0 = \emptyset$，$\theta = 1$，$i = 1$ 和残差 $\mathbf{r}_0 = \mathbf{y}_k$
  重复：
  $\mathcal{S}_i$ =最大化 $\left(|[\tilde{\mathbf{H}}_k]^* \mathbf{r}_{i-1}|,\ \mathrm{s} \times \theta\right)$
  $\mathcal{C}_i = F_{i-1} \cup \mathcal{S}_i$
  F =最大化 $\left(|[\tilde{\mathbf{H}}_k]^\dagger_{\mathcal{C}_i} \mathbf{y}_k|,\ \mathrm{s} \times \theta\right)$
  $\mathbf{r} = \mathbf{y}_k - [\tilde{\mathbf{H}}_k]_F [\tilde{\mathbf{H}}_k]^\dagger_F \mathbf{y}_k$
  如果满足终止条件，则
    继续
  否则，如果 $\|\mathbf{r}\|_2 > \|\mathbf{r}_{i-1}\|_2$，则
    $\theta = \theta + 1$
  else
    $F_i = F$, $\mathbf{r}_i = \mathbf{r}$，$i = i + 1$
  $\widehat{\boldsymbol{\alpha}}_k = [\tilde{\mathbf{H}}_k]^\dagger_F \mathbf{y}_k$
结束
通过 (2) 检测活跃用户 $\widehat{\boldsymbol{\alpha}}$
返回：$\widehat{\boldsymbol{\alpha}}$

第三项在公式（4）中都起到惩罚效用的作用。具体而言，公式（4）中的第二项表示 $\mathbf{p}^t$ 中各元素的方差，并由系数 ρ_1加权；该项用于惩罚忽视低优先级类别用户接入的策略。如果方差较大，则意味着某些低优先级类别的用户几乎不被允许接入网络，这并非理想的决策，尤其对于需要大量不同优先级分数用户接入的 mMTC服务而言。因此，在效用函数（4）中为系数 ρ_1 设置正值，智能体便能引导算法学习有利于 mMTC服务大规模接入控制的策略。

另一方面，对于 uRLLC服务，可靠性和稳定性至关重要。在这种情况下，(4) 中的第三项起到惩罚检测误差的作用，其权重系数为 ρ_2。因此，只需将 ρ_2 设为正值以对检测误差施加惩罚，智能体便可轻松切换到有利于在 uRLLC 服务中实现准确、可靠检测的策略。因而，适当设置两个惩罚项对应的系数 ρ_1 和 ρ_2，即可在各种异构服务的不同服务质量需求之间实现适当的权衡，并为它们之间的灵活切换提供良好支撑。

每个时隙t 结束后，都会按照贝尔曼方程更新 Q表的值，如下所示：

$$Q(\mathbf{s}^t, \mathbf{p}^t) \leftarrow (1-\varpi) Q(\mathbf{s}^t, \mathbf{p}^t) + \alpha (u^t + \beta \max_{\widehat{\mathbf{p}}\in\Omega^L} Q(\mathbf{s}^{t+1}, \widehat{\mathbf{p}})), \tag{5}$$

其中，ϖ 和 β 分别表示学习率和折扣率。Q表的更新由系统效用函数 u^t 驱动，如式（4）所示。

值得注意的是，所采用的Q学习方法

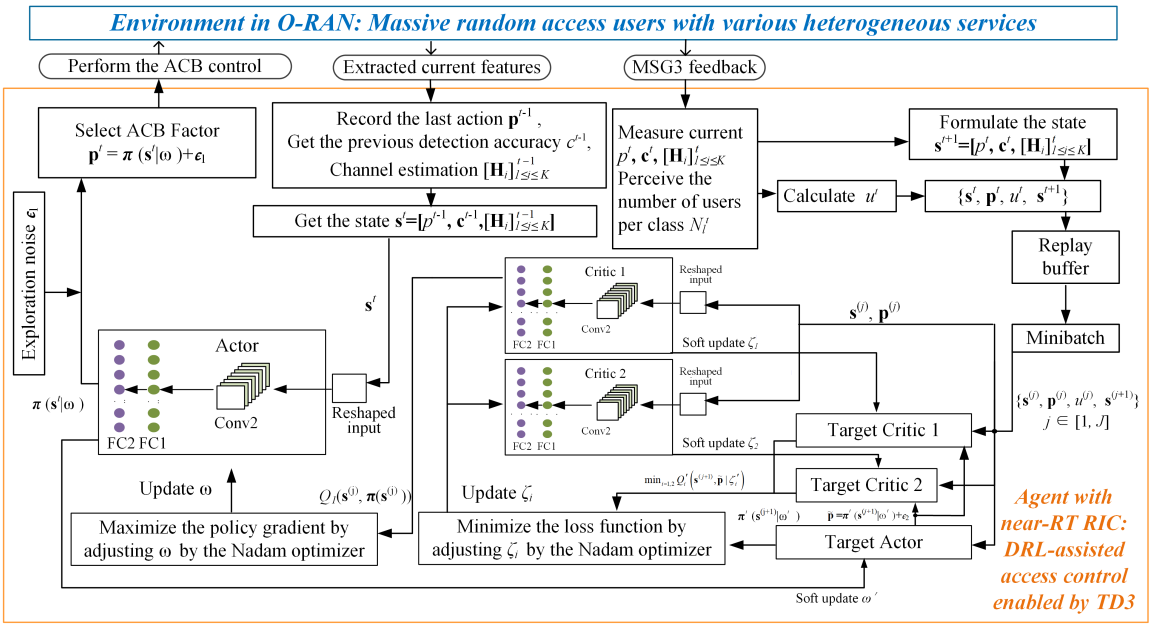


图 3. TD3支持的ACB因子动态控制架构：活跃用户通过gNB处智能体确定的ACB因子接入，这些因子由近实时RIC中的深度强化学习辅助接入控制策略提供支持。

RL-SAUD方案是一种离散动作控制方法，等价于从精确策略中进行采样。因此，应减小状态和动作的量化间隔，以传递更具体的信息，从而实现令人满意的性能 [42]。然而，这会导致Q表规模呈指数级增长，并带来过高的计算复杂度。此外，每次交互仅被学习一次，经验未能得到充分利用以支持后续学习，这限制了大数据驱动方法的潜力。因此，有必要引入深度强化学习技术，即通过训练深度神经网络实现的数据驱动范式，以解决连续状态与动作空间的离散量化和维数灾难所导致的性能限制，并充分利用以往经验，从而更快地收敛并学习最优策略。

V. 深度强化学习辅助稀疏活跃用户检测在O-RAN中的接入控制

本节介绍一种基于行动者-评论家框架的双延迟深度确定性策略梯度（TD3）算法，用于动态调整ACB因子并合理协调接入请求，从而提高有效接入用户数量和SAUD的准确性。与RL-SAUD方案相比，所提出的DRL-SAUD方案采用数据驱动方法训练深度神经网络，以解决量化损失以及高维状态和动作空间问题，并加快接入控制策略的收敛速度。事实上，DRL-SAUD方案所采用的TD3算法是一种前沿的基于DRL的算法，具有一些有利特性：i）TD3所基于的行动者-评论家框架非常有助于离线模型测试；ii）深度神经网络使状态能够参考更复杂的指标，例如信道状态信息；iii）更好地实现连续动作控制；以及iv）通过经验回放加快学习收敛。

不同于基于强化学习的Q学习算法在特定状态下选择Q表中Q值最大的动作，本文采用的基于深度强化学习的TD3方法使用行动者-评论家框架，如图 3所示。具体而言，评估网络（即评论家）处理并学习策略网络（即行动者）获得的经验，然后将Q值传递给行动者进行学习。如此一来，行动者负责动作决策，而评论家负责为动作评分。在TD3框架中，可利用行动者-评论家网络进行闭环离线训练，在将有效模型应用于真实环境之前对其进行优化。与纯在线学习方法相比，这能够显著提升模型的测试性能和用户体验，尤其是在测试阶段开始时 [43]。

采用TD3实现ACB因子动态控制的架构如图 3所示。具体而言，TD3架构共包含六个网络，包括当前的评论家1、评论家2和行动者，以及与之对应的目标网络。当前网络用于与环境实时交互，而目标网络负责为更新当前网络提供参考值。通过采用两个评论家网络，智能体可以在更新过程中选择较小的Q值，以避免高估Q值 [44]。

由 TD3 支持的 DRL-SAUD 算法在算法3中进行了详细总结。与 RL- 方案不同，状态 $\mathbf{s}^t = \left[\mathbf{p}^{t-1}, c^{t-1}, [\mathbf{H}_k]_{1\le k\le K}^{t-1}\right]$ 由此构

成：直接拼接上一时隙的实数连续值动作 p^{t-1}、检测准确率 c^{t-1} 和信道矩阵 $[\mathrm{H}_k]_{1\le k\le K}^{t-1}$，不进行量化。状态的高维特征可由行动者中的深度神经网络提取，随后通过策略 $\pi(\mathrm{s}|\omega)$确定并输出连续动作。智能体通过添加方差为 $\epsilon_1 \sim \mathcal{N}(0,\sigma)$ 的加性噪声项来实现随机探索，而非采用 ϵ-贪婪方法，因此最终执行的动作会略作调整，如 $\mathrm{p}^t = \pi(\mathrm{s}^t|\omega) + \epsilon_1$ 所示。需要注意的是，加性噪声过小会使动作不再具有探索性，而过大则会抑制对已学习策略的利用，因此可通过设置适当的加性噪声值，在探索与利用之间取得合理的权衡。

ACB检查后，活跃用户开始向 gNB 请求接入，随后 gNB 获取下一状态 s $^{t+1}$ = $\left[\ ^t\ c^t\ [\ \ k]_{1\le k\le K}^t\right]$ u^t p , ,

H 和 。包含当前状态和下一状态以及当前效用和动作的信息将被打包成一个状态转移作为经验，即 $\Im^t$ ={s^t, p^t, u^t, s^{t+1}}，并存储在经验回放缓冲区 $\mathcal{B}$ 中。经验回放缓冲区使智能体能够反复学习以往的经验，从而更快地收敛到最优模型。同时，用最新的状态转移替换过时的状态转移，以跟踪环境的变化。

当经验回放缓冲区中收集到一定数量的状态转移后，智能体会从经验回放缓冲区中随机选取 $\mathcal{J}$ 个状态转移，即 $\{\mathrm{s}^{(j)}, \mathrm{p}^{(j)}, u^{(j)}, \mathrm{s}^{(j+1)}\}, j \in [1, \mathcal{J}]$，以实时更新网络的权重。Target Actor 首先针对 $\mathrm{s}^{(j+1)}$ 计算参考动作 ~p

$$\tilde{\mathbf{p}} \leftarrow \pi'(\mathbf{s}^{(j+1)}|\omega') + \epsilon_2, \epsilon_2 \sim \mathrm{clip}(\mathcal{N}(0,\tilde{\sigma}), -g, g), \tag{6}$$

其中，策略噪声 ϵ_2 是经 $\pm g$截断的正态分布随机变量。加性策略噪声 ϵ_2可以平滑作为 Critic网络输出的 Q函数，并在动作波动时提高 Critic 提供的 Q值的可靠性。然后，Q值 $\left\{Q_i(\mathrm{s}^{(j)},\mathrm{p}^{(j)})\right\}_{i=1,2}$ 的参考值 y_r 定义为

$$y_\mathrm{r}^{(j)} \leftarrow u^{(j)} + \gamma \min_{i=1,2} Q_i{}'\left(\mathbf{s}^{(j+1)}, \tilde{\mathbf{p}}|\zeta_i{}'\right), \tag{7}$$

其中，γ 是折扣因子。Critic网络使用 Nadam优化器最小化损失函数，定义如下：

$$\zeta_i \leftarrow \arg\min_{\zeta} \frac{1}{\mathcal{J}} \sum_j \left(y_\mathrm{r}^{(j)} - Q_i(\mathbf{s}^{(j)}, \mathbf{p}^{(j)}|\zeta)\right)^2. \tag{8}$$

与基于RL的方案的更新方式不同，基于深度强化学习的方案在每个训练轮次中处理更多数据，并且新记录的经验可以在后续训练中被学习多次。

策略网络的更新速度慢于评估网络，这确保评论家在为策略更新提供评分之前，已经最小化了自身的估计误差 [43]。在本文所提出方案的设计中，行动者每更新一次，评论家就更新 d 次，并采用Nadam优化器来最大化策略梯度，如下所示

$$\omega \leftarrow \arg\max_{\omega} \frac{1}{\mathcal{J}} \nabla_{\mathbf{p}} Q_1(\mathbf{s}, \mathbf{p}|\zeta_1)|_{\mathbf{s}=\mathbf{s}^{(j)}, \mathbf{p}=\pi(\mathbf{s}^{(j)})} \nabla_\omega \pi(\mathbf{s}|\omega)|_{\mathbf{s}=\mathbf{s}^{(j)}}. \tag{9}$$

在式（9）中，$\nabla_{\mathbf{p}} Q_1$(s,p|ζ_1)表示 Q_1(s, p|ζ_1)关于 p 的梯度，而 $\nabla_\omega \pi$(s|ω)表示 π(s|ω)关于 ω的梯度。

每次更新行动者时，也会对每个目标网络执行软更新，其形式如下：

$$\zeta_i{}' = \delta\zeta_i + (1-\delta)\zeta_i{}', \omega' = \delta\omega + (1-\delta)\omega', \tag{10}$$

其中，$\delta \in (0, 1]$ 为记忆系数，可适当设置，以权衡收敛速度和准确度。

算法3：由 TD3 辅助的稀疏活跃用户检测——用于接入流量控制（DRL-SAUD）

1 初始化：
2 演员网络 π(s|ω)
3 评论家1网络 Q_1(s, p|ζ_1)，评论家2网络 Q_2(s, p|ζ_2)
4 目标网络 π'(s|ω')，$Q_1{}'$(s, p|$\zeta_1{}'$)，$Q_2{}'$(s, p|$\zeta_2{}'$)
5 重置经验回放缓冲区 $\mathcal{B}$
6 生成一个随机初始状态 s^0
$t = 1$ 的 7 , 2, 3, ……执行
8 确定 ACB 因子动作 状态 s^t 对应的状态 $\mathrm{s}^t|\omega$ 中的 $\mathrm{p}^t = \pi + \epsilon_1$
9 活跃用户发送 MSG1
10 gNB 在算法2中执行 SAUD 以检测 活跃用户 ，然后反馈 MSG2
11 检测到的活跃用户接收 MSG2 并返回 ACK
12 gNB 制定下一状态 $\mathbf{s}^{t+1} = [\mathrm{p}^t, c^t, [\mathrm{H}_k]^t{}_{1\le k\le K}]$ 并推导当前系统效用 u^t
13 将状态转移 {$\mathrm{s}^{(j)}$、$\mathrm{p}^{(j)}$、$u^{(j)}$、$\mathrm{s}^{(j+1)}$} 存储到经验回放缓冲区 缓冲区 $\mathcal{B}$
14 随机抽取 $\mathcal{J}$ 个状态转移，来自 $\mathcal{B}$
15 获取参考动作 ~p $\leftarrow \pi'(\mathrm{s}^{(j+1)}|\omega') + \epsilon_2$
16 使用式 (7) 计算参考 Q值 y_r
17 通过（8）更新 ζ_i
18 如果 (t 模 d= 0) 则
19 通过（9）更新 ω
20 软更新目标网络的权重 $\zeta_i{}'$ 和 ω' 使用（10）
21 **end**
22 $t = t + 1$
23 结束

VI. 性能分析与评估

估

本节对所提方案相关的一些重要问题进行理论性能分析与评估。首先，针对 SAUD 算法，理论分析检测准确率与用户接入请求稀疏性之间的关系。然后，对于基于强化学习的方案，推导效用函数的理论界

，并评估所提方案的计算复杂度。

A. SAUD 性能分析

通过对未知稀疏活动指示向量 [45], 进行 ℓ_1-范数最小化，可对式 (1) 中的 SAUD 问题进行凸松弛，从而得到如下凸优化问题:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}\in\mathbb{C}^N} \left\{ \frac{1}{2M} \left\| \mathbf{y} - \tilde{\mathbf{H}}\alpha \right\|_2^2 + \varepsilon_N \|\boldsymbol{\alpha}\|_1 \right\}, \tag{11}$$

在该最小化问题中，第一项表示稀疏恢复的 ℓ_2-范数误差正则化。第二项是未知稀疏活动指示向量 $\boldsymbol{\alpha}$的 ℓ_1-范数，它促使 $\boldsymbol{\alpha}$得到稀疏解。采用系数 ε_N ，以权衡稀疏恢复导致的测量误差与活跃用户的稀疏性要求，其取值为 $\varepsilon_N = \sqrt{2\varphi \log_{\eta N}^{N}}, \varphi > 2$。例如，当用户接入请求数量急剧增加时，可减小 ε_N的值，以适度放宽接入稀疏性要求。完成 SAUD 后，可从恢复得到的活动指示向量 $\hat{\boldsymbol{\alpha}}$的支撑集中获得活跃用户集合。随后，通过恢复得到的活动指示向量与真实活动指示向量之间的 ℓ_1-范数差，得到活跃用户检测准确率 c ，其表达式为:

$$c = 1 - \frac{\|\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}\|_1}{N}. \tag{12}$$

定理1. 活跃用户检测准确率 c 的下界为

$$c \geq 1 - a_1 \exp\left(-a_2 \min\left\{\|\boldsymbol{\alpha}\|_1, \log(N - \|\boldsymbol{\alpha}\|_1)\right\}\right), \tag{13}$$

若满足以下条件

$$\begin{aligned} \|\boldsymbol{\alpha}\|_{1,\max} &\geq \frac{\eta N(1-\frac{1}{\varphi})}{2\log N} \\ |\lambda_n|_{\min} &> \vartheta\varepsilon_N = \vartheta\sqrt{\frac{2\varphi\log N}{\eta N}}. \end{aligned} \tag{14}$$

证明. 详见附录A。 ■

注1. 定理1表明，活跃用户检测准确率与用户接入请求的稀疏性密切相关。(14)中的两个约束分别是对用户接入请求稀疏性的约束，以及对用户前导码$|\lambda_n|_{\min}$最小幅度的约束。$\|\boldsymbol{\alpha}\|_{1,\ \max}$ 表示网络中活动指示向量所能容忍的最大稀疏度，其满足以下约束

$$M = 2\left(\|\boldsymbol{\alpha}\|_{1,\max} + \frac{1}{\varepsilon_N^2}\right)\log\left(N - \|\boldsymbol{\alpha}\|_{1,\max}\right). \tag{15}$$

对于式（1）所示的稀疏测量模型， $\eta = M/N$ 表示压缩测量比，即测量数据大小（也就是天线数量）与稀疏向量长度（也就是潜在用户总数）之比。 ϑ 和 φ 的参数在优化过程中手动设置。

B. 基于强化学习方案的理论界与计算复杂度

为了评估所提出的基于RL和基于深度强化学习的方案的理论性能，在每个接入优先级分数均相同这一典型情况下，可以将式(4)中的系统效用推导为闭式解，具体如下述定理所示。

定理2. 若每个用户类别的接入优先级分数均相同，则最大化式(4)所示的系统效用函数u 这一问题将转化为一个具有可处理理论界的凸问题。

证明。详见附录B。 ■

备注2。当网络中用户的接入优先级没有差异，即 $L = 1$时，式（4）中的第一个惩罚项将消失，该惩罚项旨在减小不同优先级类别的ACB因子方差。在这种情况下，基于RL的方案和基于DRL的方案实际上是在动态搜索合适的接入控制策略，以逼近最优解，或逼近定理2中推导出的理论界所对应的次优解。

当不同用户类别具有不同的接入优先级分数时，网络中会引入更多不确定性。考虑到环境的时变特性，以及不同用户类别的多样化信道条件和QoS需求，无法保证原问题的最优解仍可通过闭式解求得。因此，可以利用基于RL的方案在复杂环境中于合理时间范围内搜索趋近于系统效用的次优解的能力。在基于RL的方案和基于DRL的方案的决策过程中，需要考虑两种权衡：高优先级获准接入用户数量与具有不同优先级分数的获准接入用户数量方差之间的权衡，以及获准接入用户总数与活跃用户检测准确率之间的权衡。

接下来，我们将分析所提出的基于RL的方案和基于DRL的方案的计算复杂度。根据文献 [46], 中的相关研究，当基于强化学习的算法的训练回合数为 τ ，且每个回合包含 ξ 个时隙时，其计算复杂度为

$$\mathcal{T}_1 = \mathcal{O}(\tau\xi) \tag{16}$$

若 $\tau\xi >$ poly($X_1{}^L X_2, X_1{}^L, \tau$)，其中poly($a, b, c$)是一个三次多项式，其三个根分别为 a、 b和 c。

随着状态和动作空间的量化级别提高，可行动作和状态的数量急剧增加，这使得 RL-SAUD方案在每个回合中需要更多时隙 ξ 才能收敛。作为减少量化误差所导致性能损失的副作用，动作-状态对数量的增加使得基于强化学习的算法在学习过程早期的随机探索中需要付出更多搜索和计算开销。对于 DRL-SAUD方案，深度神经网络中的浮点计算复杂度（以每秒浮点运算次数（FLOPs）衡量）是整体复杂度的主要来源。在所设计的

架构中，考虑一种网络模型，该模型包含单个具有 C_{i} 个输入通道和C_{o} 个输出通道的卷积神经网络（卷积层），以及两个全连接（FC）层。随后，通过以下定理推导所提出的 DRL-SAUD方案的计算复杂度。

定理 3. 在包含 τ 个回合、每个回合包含 ξ 个时隙的 DRL-SAUD 算法训练过程中，其计算复杂度为

$$\mathcal{T}_2 = \mathcal{O}\left(\tau\xi C_{\mathrm{i}} w_{\mathrm{i}}^2 C_{\mathrm{o}}((w_{\mathrm{i}} - v)/s + 1)^2\right) \tag{17}$$

其中， v 和 s 分别表示卷积核的大小和步幅； w_{i} 表示卷积层每个输入通道的大小。

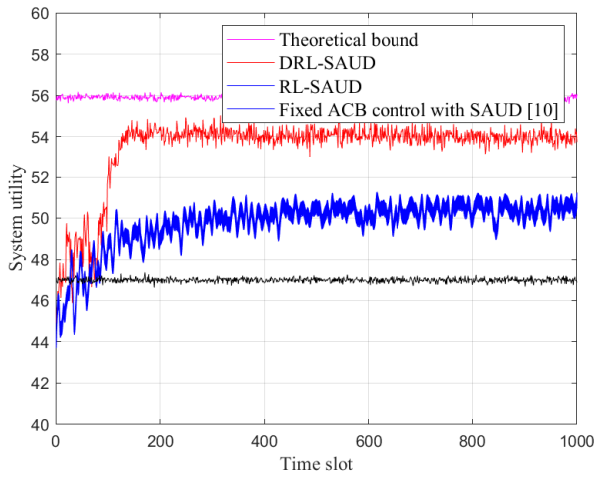
证明。详情见附录 C。 ■

注释 3. O-RAN 中开放服务的概念使制造商能够扩展其网络设计的功能，移动运营商也可以支持多种垂直行业服务共存。在实际部署多种异构服务时，神经网络的输入参数数量会显著增加。因此，仅使用简单的全连接层网络，很难分析环境和系统中的海量信息。此外，全连接层中的大量参数会降低计算速度，也更容易导致过拟合。如本定理所示，利用卷积层可以重复利用卷积核的参数，而不会带来过多的计算复杂度开销。这有助于更好地提取复杂系统的高维特征，并为智能体实现更高效的数据驱动智能方案。

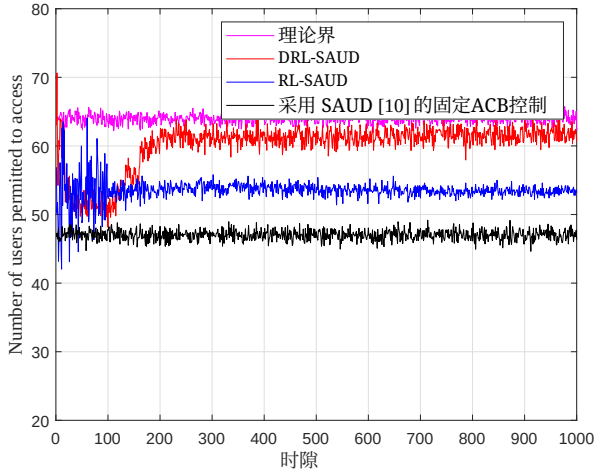
VII. 仿真结果

本节通过大量仿真评估所提出的RL-SAUD和DRL-SAUD方案的性能。为展示海量随机接入控制方案的性能，研究了用户接入数量、活跃用户检测准确率等典型指标。通过研究获准接入网络的用户数量，展示用户的接入效率和吞吐量；通过研究活跃用户检测准确率，展示接入请求的可靠性和稳定性。在不同场景以及各种异构垂直服务（如海量机器类通信和uRLLC服务）中，验证了所提方案的有效性和适应性。此外，还展示了针对具有不同接入优先级的不同用户类别，所提方案所确定的接入控制策略的变化趋势。

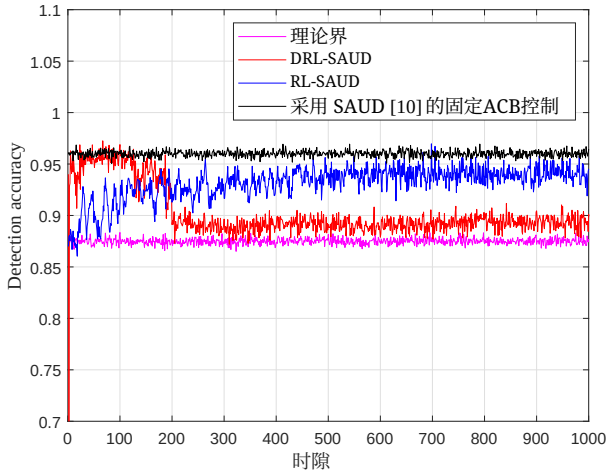
仿真配置如下：网络中的潜在用户数量设为 $N = 256$。gNB 的天线数量为$M = 128$。载波中心频率为 $2GHz$，子载波数量为 $K = 1024$。RL-SAUD方案的参数配置列于表 1。对于 DRL-SAUD方案，所用评论家的学习率



(a)

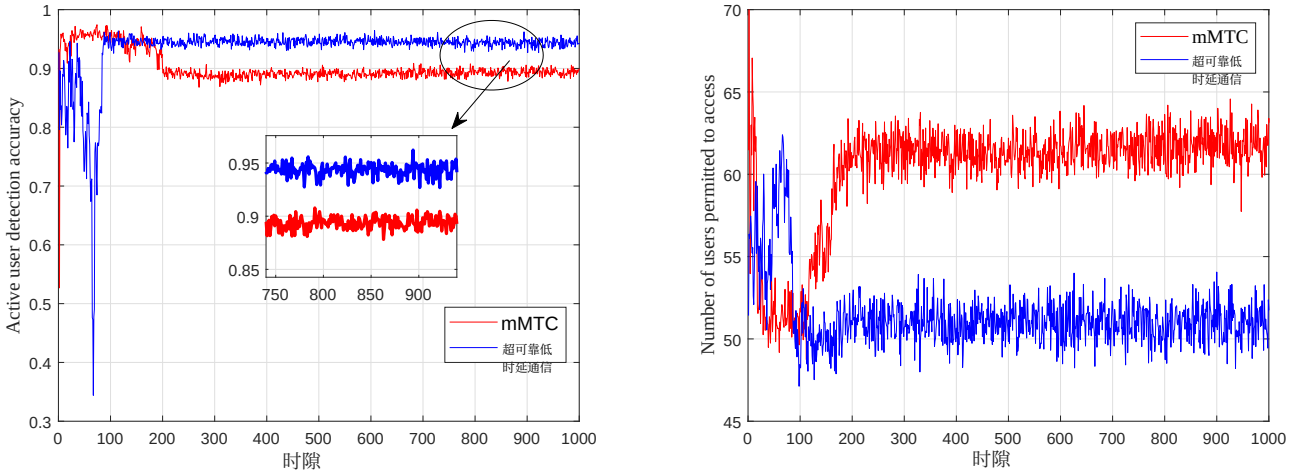


(b)



(c)

图4. 所提出的RL-SAUD和DRL-SAUD方案在以下方面的性能：(a) 系统效用；(b) 接入效率，即允许接入的用户数量；以及(c) 活跃用户检测准确率。考虑并支撑了一个由大量用户发起海量接入请求的mMTC服务。为便于比较，还展示了采用固定ACB因子与SAUD的基准方案。



(a) (b)

图5. 所提出的DRL-SAUD方案应用于uRLLC服务时与应用于mMTC服务时的性能对比：(a) 活跃用户检测准确率；(b) 允许接入网络的用户数。

在 TD3 中设为 0.00001，由于采用延迟更新，行动者的学习率设为 0.00005。行动者和评论家共用相同的网络架构，该架构由一个卷积层和两个全连接层组成。经验回放缓冲区的容量设为 1000。小批量大小为 128。为鼓励探索而加入的加性噪声方差设为 $\epsilon_1 = 0.15$。Target Actor 的策略噪声方差设为 $\epsilon_2 = 0.25$，其裁剪边界 g 设为 0.4。延迟更新时间为 $d = 4$。

图4展示了所提出的RL-SAUD和DRL-SAUD方案在系统效用、接入效率和检测准确率方面的性能，其中考虑了一个有大量用户发起海量接入请求的mMTC服务。将采用SAUD[10] 的固定ACB控制方案作为基准，并同时绘出理论界以供比较。首先，图4(a)展示了系统效用方面的性能。从图4(a)可以看出，所提出的RL-SAUD和DRL-SAUD方案显著优于采用固定ACB控制的基准方案，这验证了所提出的强化学习辅助机制在大规模接入场景中实现更优接入控制效用的有效性。结果还表明，DRL-SAUD方案的性能接近系统效用的理论界，验证了其卓越性能。

表I RL-SAUD的参数配置

参数	数值
ACB因子的量化级别数量 X_1	5
具有不同接入优先级的类别数量 L	2
检测准确率的量化级别数量 X_2	5
学习率 ϖ	0.98
折扣率 β	0.3
初始 ϵ	0.9
最终 ϵ	0.1

这也验证了在复杂环境和高维状态和动作空间下DRL辅助架构的优越性能。DRL-SAUD相较于RL-SAUD的性能增益，验证了采用结合深度神经网络的TD3架构从环境中提取更复杂信息的有效性；而RL方案的性能下降则是由量化误差造成的。

其次，图4(b)报告了接入效率的表现，该指标以获准接入网络的用户数量表示。从图4(b)可以看出，DRL-SAUD允许约62个用户接入网络，而RL-SAUD允许约53个用户接入。因此，DRL-SAUD能够支撑比RL-SAUD更多的用户，因为它克服了量化误差造成的瓶颈，并能找到更接近最优界的更优解，这有利于大量用户意图接入的mMTC服务。相比之下，采用固定ACB的基准方案仅允许48个用户接入，这表明它缺乏应对海量接入请求的灵活性。由于固定ACB因子的取值是凭经验设定的，因此在现实中复杂且时变的环境下难以找到合适的设置。

第三，图4(c)报告了活跃用户检测准确率。从图4(c)可以看出，RL-SAUD和DRL-SAUD的用户检测准确率分别约为94.93%和88.07%。尽管RL-SAUD的检测准确率相对高于DRL-SAUD，但RL-SAUD的获准接入用户数量远少于DRL-SAUD，因此其系统效用较低，如图4(a)所示。这表明，检测准确率略有下降，可以通过获准接入用户数量的大幅增加来弥补。请注意，智能体始终牢记，其目标是最大化式(4)给出的系统效用函数。因此，为了获得更高的系统效用，应寻求两者之间的最优权衡，而这一最佳权衡策略正是DRL-SAUD方案所寻求并最终收敛到的解

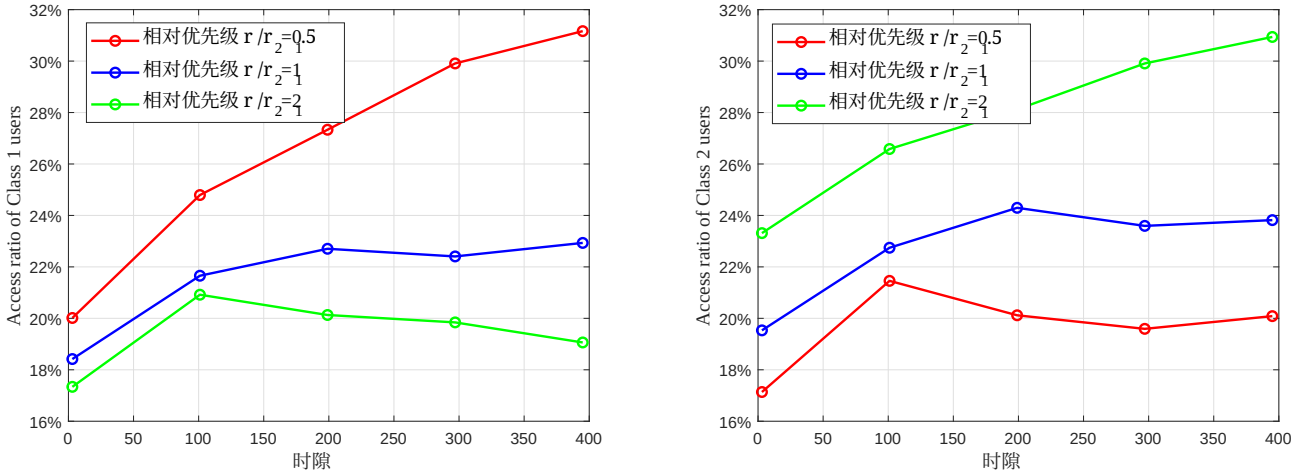


图6. 两个具有不同接入优先级分数的类别的接入率表现，即获准接入网络的用户百分比：(a) 接入优先级分数为 r_1的第1类用户的接入率；(b) 接入优先级分数为 r_2 的第2类用户的接入率。

并最终收敛至】【。

此外，从图4还可以观察到，在所提出方案的学习过程早期会进行一段动作探索。这意味着，在学习过程早期，基于强化学习的方法会尝试随机选择一些ACB因子，以高效、粗略地大范围探测环境，这有助于找到更优解，但会付出少量探索开销。

接下来，为了展示所提出方案支撑不同异构服务并在这些服务之间自适应切换的能力，我们评估了DRL-SAUD方案在超可靠低时延通信服务中的表现，并将其与海量机器类通信服务的表现进行比较，结果如图5所示。如IV-B节所述，对于超可靠低时延通信服务，通过将系数 ρ_2设为正值，激活式(4)中系统效用函数的第三项；该项对用户检测错误施加惩罚，从而促使连接具有更高的可靠性。具体而言，在图5所示的仿真中，超可靠低时延通信服务的系数设为 $\rho_2 = 100$。

图5(a)报告了DRL-SAUD方案应用于海量机器类通信服务和uRLLC服务时的活跃用户检测准确率性能。图5(a)中的结果表明，uRLLC服务的检测准确率比海量机器类通信服务高约5%。这一提升有助于智能体（即gNB）及时有效地响应在uRLLC服务中发起接入请求的用户。还可以从图5(a)中看出，在学习过程的早期阶段，uRLLC服务的检测准确率曲线出现了一个深谷，但随后迅速回升。这表明，DRL-SAUD方案首先通过初始随机探索来探测环境，然后由于检测准确率惩罚对系统效用的影响，能够迅速调整策略并收敛到优化解。

另一方面，mMTC和uRLLC服务的接入数量表现，即，允许接入的用户数量，如图5(b)所示。结果表明，所提出的方案允许约62个用户接入mMTC服务，而允许约51个用户接入uRLLC服务。这一结果意味着，所提出的方案旨在允许更多用户接入网络以使用mMTC服务，同时也决定牺牲少量接入数量，以提高uRLLC服务接入连接的可靠性和稳定性。在所提出的基于强化学习的框架中，只需将 ρ_2 设为正值，以便在式（4）的效用函数中加入关于检测误差的第二个惩罚项，智能体就能轻松地从采用有利于提高接入数量的策略的mMTC服务，切换到采用有利于精准用户检测和可靠接入连接的策略的uRLLC服务。在O-RAN范式的实际部署中，只需调整xAPP中的惩罚系数，即可实现不同异构服务之间的快速切换。

为观察所提出的DRL-SAUD方案在面对具有不同接入优先级的用户时的行为，图6报告了两个具有不同优先级的用户类别（即第1类和第2类）的接入率。某一用户类别的接入率定义为该类别中获准接入网络的用户数占该类别所有潜在用户数的百分比。令 r_2/r_1 表示所考虑的两个类别之间的相对优先级，其定义为第 2 r_2类的优先级分数与第 1 r_1类的优先级分数之比。在这种情况下，通过将相应系数设置为 $\rho_1 = 120$，系统效用（4）中的第一个惩罚项，即不同优先级类别用户之间访问数量方差所导致的惩罚项，被激活。

具体而言，图6(a)和图6(b)分别报告了第1类和第2类用户的接入率表现。为进行比较，研究了相对优先级取不同值的三种情况，即将 r_2/r_1 设为0.5、1和2。从图6(b)可以观察到，

在 $r_2/r_1 = 2$的情况下，第2类的接入率为30.9%，比 $r_2/r_1 = 1$ 和 $r_2/r_1 = 0.5$的情况分别高出9.39%和35.27%。这表明，所提出的方案已学会倾向于允许更多优先级较高的用户接入网络，以提升系统效用。对比图6(a)和图6(b)还可以发现，在 $r_2/r_1 = 1$的情况下，两类用户的接入率相近，这意味着所提出的方案已学会允许数量大致相同、接入优先级也相同的用户接入。图6中的结果验证了所提出的深度强化学习辅助方案能够适应不同优先级的用户，或具有不同服务质量需求的各种异构服务。

VIII. 结论

面对 O-RAN 场景下大规模随机接入控制的挑战，本文提出了一种基于强化学习的动态接入控制框架，可部署于 gNB 的近实时RIC中。为保持接入请求的稀疏性，从而确保超密集流量情况下 SAUD 的准确性，所提出的 RL-SAUD 方案可在闭环接入控制过程中动态调整 ACB控制策略。本文设计了一种系统效用函数，以提高获准接入网络的用户数量，并用其训练强化学习模型；此外，还采用了两个与接入率方差和检测准确率相关的惩罚项，以支持在不同异构垂直行业应用（如海量机器类通信和 uRLLC服务）之间实现适当的权衡与灵活切换。

此外，为克服基于RL的方案因离散化动作和状态而产生的量化误差，本文基于行动者-评论家架构下的 TD3 框架设计了 DRL-SAUD方案。深度神经网络能够更好地提取环境信息，并可从连续空间中选择策略和动作，从而获得更优的解决方案，使其接近最优界。利用经验回放缓冲区中的过往经验，可加快学习收敛。理论分析和仿真结果验证了所提方案在面向不同服务质量需求、不同垂直行业服务以及不同优先级用户类别的动态智能大规模接入控制中的效率、适应性和可靠性。该技术有望应用于 O-RAN范式中的 xAPP，以便为日益拥挤且日趋复杂的无线接入环境和服务提供高效有效的无线智能控制器。

附录A 定理1的证明

证明。根据文献中的相关研究 [45], ，已证明活跃用户检测准确率 c 的下界为

$$c \geq 1 - a_1 \exp\left(-a_2 \min\left\{\|\boldsymbol{\alpha}\|_1, \log(N - \|\boldsymbol{\alpha}\|_1)\right\}\right), \tag{A.1}$$

满足以下约束条件，

$$\begin{aligned} & M \geq 2\left(\|\boldsymbol{\alpha}\|_1 + \frac{1}{\varepsilon_N^2}\right)\log(N - \|\boldsymbol{\alpha}\|_1) \\ & |\lambda_n|_{\min} > \vartheta\varepsilon_N \end{aligned} \tag{A.2}$$

其中， N、 M和 $||\boldsymbol{\alpha}||_1$分别表示网络中所有潜在用户的数量、测量向量大小以及活跃用户请求的稀疏度。根据压缩感知理论 [5], ，能够被准确恢复的稀疏度应小于测量数据的数量 M，因此有$M > \|\boldsymbol{\alpha}\|_{1,\ \max}$。于是，我们可以推导出，

$$\begin{aligned} 2(||\boldsymbol{\alpha}||_1 + \frac{1}{\varepsilon_N^2})\log(N - M) &\leq 2(||\boldsymbol{\alpha}||_1 + \frac{1}{\varepsilon_N^2})\log(N - ||\boldsymbol{\alpha}||_1) \\ &\leq 2(||\boldsymbol{\alpha}||_1 + \frac{1}{\varepsilon_N^2})\log N. \end{aligned} \tag{A.3}$$

将约束条件（15）代入（A.3），可得

$$\begin{aligned} M &= 2\left(\|\boldsymbol{\alpha}\|_{1,\max} + \frac{1}{\varepsilon_N^2}\right)\log\left(N - \|\boldsymbol{\alpha}\|_{1,\max}\right) \\ &\leq 2\left(\|\boldsymbol{\alpha}\|_{1,\max} + \frac{1}{\varepsilon_N^2}\right)\log\left(N\right) \\ &\leq 2\left(\|\boldsymbol{\alpha}\|_{1,\max} + \frac{M}{2\log N}\frac{1}{\varphi}\right)\log\left(N\right), \end{aligned} \tag{A.4}$$

这等价于

$$\|\boldsymbol{\alpha}\|_{1,\max} \geq \frac{\eta N(1 - \frac{1}{\varphi})}{2\log N}. \tag{A.5}$$

因此，如果满足约束条件（14），则可由（A.1）推导出检测准确率的下界，从而完成证明。■

附录B 定理2的证明

证明。如果每个类别的接入优先级分数都相同，或者等价地，用户类别的数量仅为一个，即为$L = 1$，则网络中的潜在用户总数和ACB因子分别可表示为 $N = N_1$ 和 $p = p_1$。由于分配给不同类别的ACB因子没有差异，效用函数（4）中的第一项惩罚项消失，即

$$\begin{aligned} u &= cp_1r_1N_1 - \rho_2(1 - c) = cprN - \rho_2(1 - c) \\ &= f\left(p\right)prN - \rho_2\left(1 - f\left(p\right)\right) \\ &= f\left(p\right)\left(prN + \rho_2\right) - \rho_2, \end{aligned} \tag{B.1}$$

其中，ACB因子为 $p \in [0, 1]$，函数 $f(p)$由一条单调递减凸曲线 [11] 定义，其性质如下

$$\begin{aligned} & \frac{\partial f\left(p\right)}{\partial p} \leq 0, \frac{\partial^2 f\left(p\right)}{\partial p^2} < 0, \\ & \left.\frac{\partial f\left(p\right)}{\partial p}\right|_{p \to 0^+} \to 0^-, \\ & f\left(p\right)|_{p \to 1^-} \to 0^+. \end{aligned} \tag{B.2}$$

因此，最大化式（B.1）所给出的系统效用函数是一个凸优化问题。可以验证，

系统效用关于动作 u 的 p 一阶导数满足

$$\begin{aligned}\left.\frac{\partial u}{\partial p}\right|_{p\to 0^+} &= \frac{\partial f(p)}{\partial p}(prN+\rho_2)+rNf(p)=rNf(0^+)>0,\\ \left.\frac{\partial u}{\partial p}\right|_{p\to 1^-} &= \frac{\partial f(p)}{\partial p}(prN+\rho_2)+rNf(p)\\ &= \frac{\partial f(p)}{\partial p}(rN+\rho_2)<0.\end{aligned} \tag{B.3}$$

系统效用 u 关于动作 p 的二阶导数严格为负，如下所示:

$$\frac{\partial^2 u}{\partial p^2}=\frac{\partial^2 f(p)}{\partial p^2}(prN+\rho_2)+\frac{\partial f(p)}{\partial p}(2rN+\rho_2)<0. \tag{B.4}$$

因此，可以推导出，一阶导数 $\frac{\partial u}{\partial p}$ 在 $p\in[0,1]$ 的可行范围内有唯一的零解。因此，关于 p 的凸函数 u 在区间 $p\in[0,1]$ 内取得最大值。■

附录 C 定理 3 的证明

证明。对于图 3 所示架构中的卷积神经网络（卷积层）输入数据可重塑为尺寸为 $C_{\mathrm{i}}\times w_{\mathrm{i}}\times w_{\mathrm{i}}$ 的高维张量，其中 C_{i} 是卷积层的输入通道数。当卷积层具有 C_{o} 个输出通道，且每个通道配备一个核尺寸为 $v\times v$、步幅为 s 的卷积核时，每个卷积层消耗的浮点运算次数计算如下:

$$N_{\mathrm{FLOPs}}(\mathrm{Conv})=C_{\mathrm{i}}w_{\mathrm{i}}^2C_{\mathrm{o}}w_{\mathrm{o}}^2 \tag{C.1}$$

其中， w_{o} 是卷积层每个输出通道的尺寸。根据文献 [47]，中关于零填充模式的相关研究， w_{o} 的值与输入尺寸、卷积核尺寸和步幅有关，如下所示:

$$w_{\mathrm{o}}=\frac{w_{\mathrm{i}}-v}{s}+1 \tag{C.2}$$

其中，偏置的影响通过在（C.2）右侧加 1 来体现。由于其余两个全连接网络（FC）层的 FLOPs 远小于卷积层，因此可以忽略不计，所以基于DRL的方案每个时隙的计算复杂度为

$$N_{\mathrm{FLOPs}}(\text{time slot})=C_{\mathrm{i}}w_{\mathrm{i}}^2C_{\mathrm{o}}\left(\frac{w_{\mathrm{i}}-v}{s}+1\right)^2 \tag{C.3}$$

最后，将基于DRL的方案每个时隙的 FLOPs 数量乘以回合数和时隙数，即可推导出（17）。■

参考文献

[1] I. Afolabi、T. Taleb、K. Samdanis、A. Ksentini 和 H. Flinck，“网络切片与软件化：关于原理、赋能技术和解决方案的综述”，IEEE Communications Surveys and Tutorials，第 20 卷第 3 期，第 2429–2453 页，2018 年第三季度。

[2] S. D’Oro、M. Polese、L. Bonati、H. Cheng 和 T. Melodia，“dapps：用于实时推理与控制的分布式应用”，IEEE Communications Magazine，第 60 卷第 11 期，第 52–58 页，2022 年 11 月。[3] M. Polese、L. Bonati、S. D’Oro、S. Basagni 和 T. Melodia，“Colo-ran：在可编程实验平台上开发基于机器学习的 xAPP，以实现开放式 RAN 的闭环控制”，IEEE Transactions on Mobile Computing，第 1–14 页，2022 年 7 月。[4] A. Lacava、M. Polese、R. Sivaraj、R. Soundrarajan、B. Shanker Bhati、T. Singh、T. Zugno、F. Cuomo 和 T. Melodia，“利用开放式 RAN 架构实现 5G 网络流量引导的可编程定制智能”，第 arXiv:2209.14171 页，2022 年 9 月。[5] Z. Gao、L. Dai、S. Han、C.-L. I、Z. Wang 和 L. Hanzo，“面向下一代无线通信的压缩感知技术”，IEEE Wireless Communications，第 25 卷第 3 期，第 144–153 页，2018 年 6 月。[6] J.-C. Jiang 和 H.-M. Wang，“具有零星短分组的大规模随机接入：通过序贯消息传递联合进行活跃用户检测与信道估计”，IEEE Transactions on Wireless Communications，第 20 卷第 7 期，第 4541–4555 页，2021 年 7 月。[7] Y. Noh 和 S. Hong，“ 基于压缩感知的 MIMO 系统一比特 ADC 活跃用户检测 ”，IEEE Transactions on Vehicular Technology，第 1–6 页，2022 年 9 月。[8] J. Ahn、B. Shim 和 K. B. Lee，“面向大规模机器类通信的基于 EP 的联合活跃用户检测与信道估计”，IEEE Transactions on Communications，第 67 卷第 7 期，第 5178–5189 页，2019 年 7 月。[9] M. Ke、Z. Gao、Y. Wu、X. Gao 和 R. Schober，“基于压缩感知的自适应活跃用户检测与信道估计：大规模接入遇上大规模 MIMO”，IEEE Transactions on Signal Processing，第 68 卷，第 764–779 页，2020 年 1 月。[10] J.-P. Hong、W. Choi 和 B. D. Rao，“利用压缩感知的稀疏度控制随机多址接入”，IEEE Transactions on Wireless Communications，第 14 卷第 2 期，第 998–1010 页，2015 年 2 月。[11] Y. Sim 和 D.-H. Cho，“大规模 MTC 随机接入中基于优先级的接入类限制方案的性能分析”，IEEE Systems Journal，第 14 卷第 4 期，第 5245–5252 页，2020 年 12 月。[12] X. Du 、 D. Wu 、 W. Liu 和 Y. Fang，“ 异构移动自组织网络的多类路由与媒体接入控制”，IEEE Transactions on Vehicular Technology，第 55 卷第 1 期，第 270–277 页，2006 年 1 月。[13] F. Morvari 和 A. Ghasemi，“LTE 网络中 M2M 通信随机接入的两阶段资源分配”，IEEE Communications Letters，第 20 卷第 5 期，第 982–985 页，2016 年 5 月。[14] J. Jiao、L. Xu、S. Wu、Y. Wang、R. Lu 和 Q. Zhang，“面向大规模机器类通信的不等接入时延随机接入协议”，IEEE Transactions on Wireless Communications，第 19 卷第 9 期，第 5924–5937 页，2020 年 9 月。[15] L. Bonati、S. D’Oro、M. Polese、S. Basagni 和 T. Melodia，“面向数据驱动下一代蜂窝网络的 O-RAN 智能与学习”，IEEE Communications Magazine，第 59 卷第 10 期，第 21–27 页，2021 年 10 月。[16] S. Mollahasani、M. Erol-Kantarci 和 R. Wilson，“利用行动者-评论家学习进行 O-RAN 中的动态 CU-DU 选择与资源分配”，载于 2021 IEEE 全球通信会议（GLOBECOM），第 1–6 页，西班牙马德里，2021 年 12 月。[17] M. Centenaro、L. Vangelista、S. Saur、A. Weber 和 V. Braun，“物联网中无冲突与基于竞争的无线接入协议比较”，IEEE Transactions on Communications，第 65 卷第 9 期，第 3832–3846 页，2017 年 9 月。[18] Z. Yuan、W. Li、Y. Hu、H. Tang、J. Dai 和 Y. Ma，“基于接收波束成形的自主免授权高过载多址接入盲多用户检测”，载于 2019 IEEE 第二届 5G 世界论坛（5GWF），第 520–523 页，德国德累斯顿，2019 年 9 月。[19] Z. Chen、F. Sohrabi 和 W. Yu，“大规模随机接入的多小区稀疏活动检测：大规模 MIMO 与协作式 MIMO 对比”，IEEE Transactions on Wireless Communications，第 18 卷第 8 期，第 4060–4074 页，2019 年 8 月。[20] C. Bockelmann 、 N. Pratas 、 H. Nikopour 、 K. Au 、 T. Svensson 、 C. Stefanovic、P. Popovski 和 A. Dekorsy，“5G 中的大规模机器类通信：物理层与 MAC 层解决方案”，IEEE Communications Magazine，第 54 卷第 9 期，第 59–65 页，2016 年 9 月。[21] Y. Zhang、Q. Guo、Z. Wang、J. Xi 和 N. Wu，“基于块稀疏贝叶斯学习的免授权 NOMA 系统联合用户活动检测与信道估计”，IEEE Transactions on Vehicular Technology，第 67 卷第 10 期，第 9631–9640 页，2018 年 10 月。[22] A. Fengler、S. Haghighatshoar、P. Jung 和 G. Caire，“采用大规模 MIMO 接收机的免授权大规模随机接入”，载于 2019 年第 53 届 Asilomar 信号、系统与计算机会议，第 23–30 页，美国太平洋格罗夫

, CA, 美国 , 2019年11月。

[23] M. Bayati 和 A. Montanari，《稠密图上消息传递的动力学及其在压缩感知中的应用》，IEEE 信息论汇刊，第 57 卷，第 2 期，第 764–785 页，2011 年 2 月。[24] L. Liu 和 W. Yu，《大规模连接性与大规模 MIMO——第一部分：设备活动检测与信道估计》，IEEE 信号处理汇刊，第 66 卷，第 11 期，第 2933–2946 页，2018 年 6 月。[25] H. S. Jang、H. Jin、B. C. Jung 和 T. Q. S. Quek，《面向海量物联网的多用途接入控制：吞吐量、时延与能效》，IEEE 移动计算汇刊，第 19 卷，第 8 期，第 1984–1997 页，2020 年 8 月。[26] C. Di、B. Zhang、Q. Liang、S. Li 和 Y. Guo，《基于学习自动机的接入类限制方案：用于机器对机器通信中的大规模随机接入》，IEEE 物联网期刊，第 6 卷，第 4 期，第 6007–6017 页，2019 年 8 月。[27] H. Lee、J. Cha、D. Kwon、M. Jeong 和 I. Park，《在 O-RAN 无线智能控制器平台上托管 AI/ML 工作流》，载于《2020 IEEE GLOBECOM Workshops（GC Wkshps）》，第 1–6 页，中国台湾台北，2020 年 12 月。[28] S. D'Oro、L. Bonati、M. Polese 和 T. Melodia，《OrchestrAN：通过开放式无线接入网中的编排式智能实现网络自动化》，载于《IEEE INFOCOM 2022——IEEE 计算机通信会议》，第 270–279 页，虚拟在线会议，英国，2022 年 5 月。[29] B. Brik、K. Boutiba 和 A. Ksentini，《面向 B5G 开放式无线接入网的深度学习：演进、综述、案例研究与挑战》，IEEE 通信学会开放期刊，第 3 卷，第 228–250 页，2022 年 1 月。[30] P. Li、J. Thomas、X. Wang、A. Khalil、A. Ahmad、R. Inacio、S. Kapoor、A. Parekh、A. Doufexi、A. Shojaeifard 和 R. J. Piechocki，《RLOps：强化学习辅助开放式无线接入网的开发生命周期》，IEEE Access，第 10 卷，第 113808–113826 页，2022 年 10 月。[31] Y. Cao、S.-Y. Lien、Y.-C. Liang、K.-C. Chen 和 X. Shen，《开放式无线接入网中的用户接入控制：一种联邦深度强化学习方法》，IEEE 无线通信汇刊，第 21 卷，第 6 期，第 3721–3736 页，2022 年 6 月。[32] S. K. Sharma 和 X. Wang，《迈向超密集蜂窝物联网网络中的海量机器类通信：当前问题与机器学习辅助解决方案》，IEEE Communications Surveys and Tutorials，第 22 卷，第 1 期，第 426–471 页，2020 年第一季度。[33] N. Jiang、Y. Deng、A. Nallanathan 和 J. A. Chambers，《用于 NB-IoT 网络实时优化的强化学习》，IEEE 通信领域精选期刊，第 37 卷，第 6 期，第 1424–1440 页，2019 年 6 月。[34] D.-D. Tran、S. K. Sharma 和 S. Chatzinotas，《基于误块率的自适应 Q学习：用于基于 NOMA 的海量机器类通信网络中的高效随机接入》，载于《2021 IEEE 第 93 届车辆技术会议（VTC2021-Spring）》，第 1–5 页，虚拟在线会议，2021 年 4 月。[35] M. V. da Silva、R. D. Souza、H. Alves 和 T. Abr˜ao，《一种基于 NOMA 的 Q学习随机接入方法：用于机器类通信》，IEEE 无线通信快报，第 9 卷，第 10 期，第 1720–1724 页，2020 年 10 月。[36] L. M. Bello、P. D. Mitchell 和 D. Grace，《支持蜂窝网络中 M2M 流量的智能 RACH 接入技术》，IEEE 车辆技术汇刊，第 67 卷，第 9 期，第 8905–8918 页，2018 年 9 月。[37] Y. Yu、T. Wang 和 S. C. Liew，《面向异构无线网络的深度强化学习多址接入》，IEEE 通信领域精选期刊，第 37 卷，第 6 期，第 1277–1290 页，2019 年 6 月。[38] Y. Cao 、 S.-Y. Lien 、 Y.-C. Liang 、 K.-C. Chen 和 X. Shen，《开放式无线接入网中的用户接入控制：一种联邦深度强化学习方法》，IEEE 无线通信汇刊，第 21 卷，第 6 期，第 3721–3736 页，2022 年 6 月。[39] A. Kumar、G. Verma、C. Rao、A. Swami 和 S. Segarra，《使用深度 Q学习的自适应竞争窗口设计》，载于《ICASSP 2021——2021 IEEE 声学、语音与信号处理国际会议（ICASSP）》，第 4950–4954 页，加拿大安大略省多伦多，2021 年 6 月。[40] X. Ye、Y. Yu 和 L. Fu，《基于深度强化学习的异构网络多信道机会接入》，IEEE 无线通信汇刊，第 21 卷，第 2 期，第 794–807 页，2022 年 2 月。[41]《无线资源控制（RRC）；协议规范》，技术报告 3GPP TS 36.331 V13.0.0，2016 年 1 月。[42] B. Kiumarsi、K. G. Vamvoudakis、H. Modares 和 F. L. Lewis，《使用强化学习的最优自主控制：综述》，IEEE 神经网络与学习系统汇刊，第 29 卷，第 6 期，第 2042–2062 页，2018 年 6 月。[43] S. Fujimoto、H. van Hoof 和 D. Meger，《解决行动者-评论家方法中的函数逼近误差》，第 arXiv:1802.09477 页，2018 年 2 月。[44] X. Wang、Y. Wang、D. Shi、J. Wang 和 Z. Wang，《两阶段 WECC 复合负荷建模：一种双深度 Q学习网络方法》，IEEE智能电网汇刊，第11卷，第5期，第4331–4344页，2020年9月。[45] M. J. Wainwright，"使用受 ℓ_1 约束的二次规划（套索回归）进行高维含噪稀疏恢复的精确阈值"，IEEE信息论汇刊，第55卷，第5期，第2183–2202页，2009年4月。[46] C. Jin、Z. Allen-Zhu、S. Bubeck 和 M. I. Jordan，"Q学习是否可证明地高效？"，第 arXiv:1807.03765 页，2018年7月。[47] C. C. T. Mendes、V. Fr´emont 和 D. F. Wolf，"利用全卷积神经网络进行快速道路检测"，载于2016年IEEE国际机器人与自动化会议（ICRA），第3174–3179页，瑞典斯德哥尔摩，2016年5月。