第 43 卷第 2 期 遥 测 遥 控 Vol. 43, No. 2
2022 年 3 月 Journal of Telemetry, Tracking and Command Mar. 2022

Website: ycyk.brit.com.cn Email: ycyk704@163.com

一种基于 NOMA 的 Q 学习卫星通信随机接入方法

杨伟康，许小东✉
（中国科学技术大学中科院无线光电通信重点实验室 合肥 230026）

摘要：基于非正交多址接入（NOMA）的 Q 学习（Q-Learning）随机接入方法（NORA-QL）是实现物联网中海量设备泛在接入的一项有效技术。为了解决 NORA-QL 方法仍存在的传输能效和过载容量较低的问题，提出了一种适合卫星通信网络的改进方法（I-NORA-QL）。针对传输功耗高的问题，I-NORA-QL 利用卫星广播的全局信息改进 Q 学习的学习策略，将用户发射功率用于奖励函数的构造，同时将学习速率设计为与算法迭代次数相关的衰减函数。I-NORA-QL 进一步在接入类别限制 ACB（Access Class Barring）的基础上，基于学习过程中的 Q 值特性和负载估计实现 ACB 限制因子的自适应调整以进行过载控制。仿真结果表明，提出的 I-NORA-QL 改进方法相比于现有其他方法，能够有效降低用户设备的平均功耗，且在系统过载状态下可以显著提高吞吐量。

关键词：卫星通信；随机接入；能量效率；过载控制；非正交多址接入；Q 学习
中图分类号：TN927+.2 **文献标识码：**A **文章编号：**CN11-1780(2022)02-0025-11
DOI：10.12347/j.ycyk.20210913001
引用格式：杨伟康，许小东．一种基于 NOMA 的 Q 学习卫星通信随机接入方法[J]. 遥测遥控, 2022, 43(2): 25–35.

A NOMA-based Q-learning random access scheme for satellite communications

YANG Weikang, XU Xiaodong
（CAS Key Laboratory of Wireless-Optical Communications, University of Science and Technology of China, Hefei 230026, China）

Abstract: The Non-Orthogonal Multiple Access (NOMA)-based Q-learning random access method (NORA-QL) is an effective technique to achieve ubiquitous access to a large number of devices in the Internet of Things. In order to solve the problems of low transmission energy efficiency and low overload capacity in the NORA-QL method, an improved method (I-NORA-QL) suitable for satellite communication networks is proposed. To address the problem of high transmission power consumption, I-NORA-QL improves the learning strategy of Q-learning using global information from satellite broadcasting, the transmitted power of user equipment is used in the construction of the reward function, and the learning rate is designed as a decay function related to the number of iterations of the algorithm. Furthermore, based on the Access Class Barring (ACB), I-NORA-QL realizes the adaptive adjustment of ACB barring factor based on the Q value characteristics and load estimation during the learning process to carry out overload control. Simulation results show that, compared with other existing methods, the proposed I-NORA-QL improved method can effectively reduce the average power consumption of user devices, and significantly improve the throughput under system overload state.

Key words: Satellite communications; Random access; Energy efficiency; Overload control; Non-Orthogonal Multiple Access; Q-learning
DOI：10.12347/j.ycyk.20210913001
Citation: YANG Weikang, XU Xiaodong. A NOMA-based Q-learning random access scheme for satellite communications[J]. Journal of Telemetry, Tracking and Command, 2022, 43(2): 25–35.

✉通信作者：许小东（xuxd@ustc.edu.cn）
收稿日期：2021-09-13 收修改稿日期：2021-11-01

引　言

面向机器类通信 MTC（Machine Type Communications）的卫星通信网以其全天候、广覆盖、大容量及 MTC 设备易部署等特点，受到业界广泛关注，已成为地面物联网 IOT（Internet of Things）的有益延伸和补充。随机接入方法由于灵活性高、信令开销低和易实现的特点成为 MTC 设备上行链路接入的有效解决方法。然而，海量 MTC 设备试图在短时间内以随机接入方式接入卫星网络时，不仅会造成频繁的传输冲突，导致传输时延增加，甚至网络过载，而传输失败也会严重削弱 MTC 设备的能量储备[1,2]。因此，在卫星通信网络中，研究具备有效过载控制机制的高能效随机接入方法，是改善卫星 MTC 通信网络性能的可行途径之一。

近年来，学者们致力于将强化学习 RL（Reinforcement learning）用作智能时隙选择策略，以弥补当前基于竞争的随机接入协议（如时隙 Aloha）在性能上存在的严重不足。针对蜂窝网络，文献[3]就蜂窝网络中 MTC 设备和人员如何共享随机接入信道（RACH）的问题提出了一种 Q 学习和时隙 Aloha 的组合方案。此方案通过 Q 学习智能地将时隙分配给 MTC 设备，有效避免了多用户之间的碰撞冲突。与文献[3]中的二元奖励策略不同，SHARMA[1]等人根据演进型基站 eNB（evolved nodeB）广播的 RACH 拥塞程度标识设计奖励函数，提出了一种基于协作的分布式 Q 学习方案，不仅提高了系统的吞吐量，而且明显降低了 MTC 设备的学习时间。进一步，MOON[2]等将 Q 学习和访问类别限制 ACB（Access Class Barring）相结合，通过设定一系列可能的状态和动作，将碰撞概率和平均接入延迟用于定义奖励，使得 eNB 可根据过载程度自适应调整 ACB 控制参数，进而提高系统性能。文献[4]通过空间聚类和就近接入来克服 MTC 网络拥塞问题，一方面缓解了接入拥塞程度，另一方面也降低了信号的传播损耗，节省了发送功率。此外，文献[4]还分析了学习率和奖励函数对收敛的影响，其结论对学习率和奖励函数的设置具有一定的指导作用。文献[5]针对卫星网络中多用户和多中继的卫星上行链路场景，提出了一种基于分布式 Q 学习的联合中继选择和访问控制方案。该方案中，设备的奖励由地面中继站计算和反馈而非来自卫星，在降低接入延迟上具有显著优势。

非正交多址访问 NOMA（Non-Orthogonal Multiple Access）相较于传统的正交多址接入具有更高的频谱效率，被研究用于蜂窝和卫星网络的上行链路[6-8]。文献[9]将功率域的 NOMA 技术和 Q 学习相结合，用于解决上行链路的子信道分配和功率控制问题，从而在短数据包通信中最大化 MTC 网络的能量效率。SILVA[10]等人提出了一种基于 NOMA 的 Q 学习随机接入方案（NORA-QL）用于提高系统的吞吐量而非能量效率，同时考虑了路径损耗和衰落对性能的影响。在此方案中，时隙和发射功率和被组成（时隙，功率）对（下文用（t, p）表示），设备通过不断学习，找到唯一的（t, p）对进行传输。NOMA 技术使得由于选择相同时隙而发生碰撞的多个数据包仍可能被成功解调，因此系统吞吐量相较于文献[1]有了明显提高，同时需要的来自 eNB 的反馈比特更少。与上述方案不同，文献[11]针对具有突发流量的机器通信设备的特点，采用了随机分组到达模型。为了降低具有突发流量的 MTC 设备对强化学习的影响，作者设计了基于设备活动概率的行动奖励，结合 NOMA 技术，进一步提高了设备的成功访问概率。

本文针对卫星 MTC 通信网络，在文献[10]的基础上对基于 NOMA 的 Q 学习随机接入方法（以下简称 INORA-QL）进行改进，进一步提升系统在不同负载状态下的传输能效和吞吐率。文献[10]通过引入功率域 NOMA 技术提高了系统的吞吐量，但同时也导致了更高的能量消耗，因为文献[10]中 Q 学习采用了贪婪算法，使得原本可以通过更加“智能”地学习、用低功率传输即可满足需求的设备不得不一直以更高的功率发送数据，导致能量效率低下。这对能量受限的 MTC 设备是不可接受的，具体分析将在第 2 节中详述。

为了解决能耗问题，首先依据发射功率和传输结果（成功/失败）设计奖励函数，同时根据卫星广播的全局信息改进 Q 学习的学习策略，并将学习速率设计为与迭代次数相关的函数形式，使得一帧中的每个时隙尽量只分配给单一用户，从而允许 MTC 设备尽量以低功率发送数据，实现能效优化。其次，考虑到当系统过载时，即使采用文献[10]的 NORA-QL 技术，仍会出现吞吐量迅速下降的问题。因此，本

文进一步针对过载问题设计了一种基于学习过程中的 Q 值特性和负载估计的 ACB 新机制。通过上述两项改进，本文提出的 I-NORA-QL 方法相比于文献[10]，能够在提高能量效率的同时，有效改善过载时系统的吞吐量。

1 系统模型

如图 1 所示，本文考虑采用低轨（LEO）卫星通信系统，由于其具有相对更低的传播延迟和功率消耗，更适合面向物联网的机器类通信。在卫星覆盖范围内，K 个 MTC 设备采用帧时隙 Aloha 协议尝试接入卫星网络。其中，一帧分为 N 个等长的时隙，每个时隙的长度都可满足数据包的传输要求。在一帧中，一个设备只能选择一个时隙发送数据，并且假定设备是积压的，即始终有要发送的数据包[1,10]。在每一帧的结尾，卫星发送一组反馈比特用于指示数据传输结果（成功/失败）。

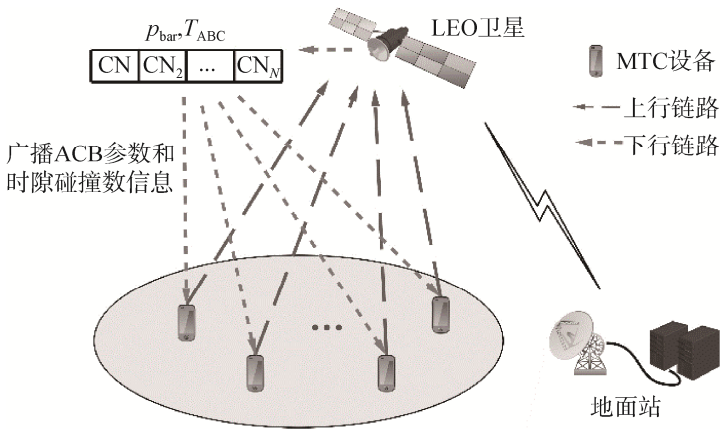


图 1　本文考虑的低轨卫星通信系统模型

Fig. 1　The model of LEO satellite communication system under consideration

此外，与现有工作相似，还假设：① 所有的 MTC 设备是全局同步的。② 卫星可对大尺度衰落和小尺度衰落的系数完美估计，因此本文不考虑衰落。③ 完美的串行干扰消除 SIC（Successive Interference Cancellation）技术，即当一个数据包被解调后，通过信号重构可完全消除它对其他数据包的解调造成的干扰。

本文方案中的随机接入的整个过程可描述为：MTC 设备通过 Q 学习算法选择一个（t, p）进行一次数据传输，地面站依据当前解调情况得到 ACB 参数和一帧中各时隙内数据包的碰撞情况，并将其通过卫星的下行链路广播给 MTC 设备，其中前者用于过载控制，后者则是结合传输结果用于设备的 Q 值更新，进而找到最优的传输策略，即最佳的（t, p）组合。

1.1　NOMA 中的 SIC 技术

在 TDMA 上行链路中，NOMA 技术允许多个设备通过使用不同的发射功率共享同一时频资源，接收机采用 SIC 技术分离不同设备，实现资源复用。具体的，接收机使用 SIC 技术先将较低功率的用户信号作为噪声，解调出信号功率较高的设备，然后从当前时隙消除较高功率信号的干扰，解调信号功率较低的设备信号。在接收端，第 n 个时隙的接收信号可表示为

$$y_n = \sum_{m=1}^{M} \sqrt{P_{m,n}} \cdot x_{m,n} + w_n \tag{1}$$

其中，M 表示一帧中选择第 n 个时隙的设备数（$1 \leqslant M \leqslant K$），$p_{m,n}$ 和 $x_{m,n}$ 分别表示第 n 个时隙中第 m 个设备的接收功率和传输信号，w_n 表示均值为 0、方差为 σ^2 的加性高斯白噪声。

本工作中，当一个接收信号的信干噪比 SINR（Signal to Interference plus Noise Ratio）大于某个阈值 γ_{th} 时，则被认为解调成功。不失一般性，假设第 n 个时隙中 M 个信号的接收功率大小顺序为 $P_{1,n} \leqslant P_{2,n} \leqslant \cdots \leqslant P_{M,n}$，则第 m 个设备的 $SINR$ 为：

$$SINR_{m,n} = \frac{P_{m,n}}{\sum_{i=1}^{m-1} P_{m,n} + \sigma^2} \tag{2}$$

其中，$\sum_{i=1}^{m-1} P_{m,n}$ 表示未解调信号的功率之和，对当前信号的解调作为干扰存在。虽然式（2）表明更多的

功率水平可有效地提高系统的吞吐量，但同时也会造成设备的发送功率指数增加，SIC 解调的解调复杂度增大[8]，因此本文只考虑了低功率 P_{L} 和高功率 P_{H} 两种功率水平，并且当 $SINR>\gamma_{\mathrm{th}}$，即可满足传输速率要求。

1.2 Q 学习

Q 学习是强化学习中的一种经典的无模型算法，具有计算复杂度较低、可在分布式场景下实现的优势[13]，适用于计算能力和能量受限的 MTC 设备中。在该算法中，每个 MTC 设备都相当于一个智能体来维护一张 Q 表，通过不断行动来获取环境的反馈，并通过 Bellman 方程来更新 Q 值。经过多次迭代，当算法收敛后，智能体就能根据 Q 值的大小找到最优的动作策略。Q 值的更新算法为：

$$Q(s,a)=Q(s,a)+\alpha[R(s,a)+\gamma\max_{a'}(s',a')-Q(s,a)] \tag{3}$$

其中，s、α 分别表示状态和动作，α 为学习速率 $(0<\alpha<1)$，R 为奖励，γ 表示折扣因子，用来衡量未来奖励对 Q 值更新的影响。本方案中 γ 被设置为 0，即不考虑未来奖励，这在考虑的场景中是合理的[1,5]。Q 值更新算法为：

$$Q(t,p)=Q(t,p)+\alpha[R(t,p)-Q(t,p)] \tag{4}$$

式中，(t,p) 表示时隙和功率组合，$Q(t,p)$ 表示任一设备 k，$\forall k\in\{1,2,\cdots,K\}$，对于 (t,p) 传输的偏好，每个设备根据前一次 Q 值和当前的奖励 R 对 Q 值进行更新。

1.3 ACB 机制

3GPP 提出采用 ACB 来控制访问尝试以减轻 LTE/LTE-A 网络的过载[12]。在该机制中，每个设备属于一个或更多的接入类别，某个设备如果属于接入允许类别，便可直接接入，否则如果属于接入受限类别，其接入将受到限制。具体的，受限设备在尝试接入时，首先产生一个介于 0 和 1 之间的随机数，如果该数不小于 eNB 广播的 ACB 限制因子，则该设备将在随后特定的限制期限内被禁止访问。

1.4 性能指标

在本文中，归一化吞吐量 T，丢包率，平均功耗和低功率数据包占比作为指标用来评估本文所提方案的性能。归一化吞吐量定义为一帧中成功解调的数据包数与时隙数 N 之比，相应的，归一化负载 G 为设备数 K 与时隙数 N 之比。丢包率定义为一帧中未成功解调的数据包与总数据包之比。平均功耗定义为 MTC 设备成功传输一个数据包的平均功耗。低功率数据包占比则定义为所有设备发送的数据包中，低功率数据包所占比例，与平均功耗共同用来评估能量效率。

2 改进方法

在这一部分，主要从能量效率和过载控制两方面改进了基于 NOMA 的 Q 学习随机接入方案[10]，降低了 MTC 设备的平均功耗，同时提高了系统过载时的吞吐量。

2.1 基于 Q 学习的随机接入

文献[1]和文献[10]中的方案作为本文工作的基准，为了便于表述，分别被命名为 Sharma-S 和 Silva-S。其中，Sharma-S 中 Q 学习算法的操作流程为：

① 在开始时，对于每个设备，Q 表被初始化为 0。

② 每次传输，设备选择最大 Q 值对应的时隙，若最大值对应的时隙有多个，则从它们中随机选择一个。

③ 设备根据传输结果和 eNB 广播的全局信息更新 Q 值。

④ 重复式（2）、式（3），直到收敛，即所有设备都找到唯一的传输时隙。

Silva-S 与 Sharma-S 不同之处主要在于式（2），即每次传输中，设备选择最大 Q 值对应的（t, p）来传输数据。此外，Silva-S 只需要依据传输结果即可进行 Q 值更新。

2.2 能效优化机制

Silva-S 中功率域 NOMA 的引入，虽然提高了系统的吞吐量，但也导致设备功耗的大幅增加，能效降低。假设某个 MTC 设备采用 Silva-S 中的方法已经学到了最佳策略（即最大 Q 值对应的（t, p））,

则其功耗增加的原因分析如下：

① 若此时该设备使用高功率，而选择的时隙中只有其一台设备，那么实际上该设备可以采用更低的发射功率，因为此时接收端的 SINR 只受白噪声的影响，但由于采用贪婪算法，因此设备将一直使用该策略，继续以高功率传输。

② 由于采用 SIC 解调技术，对于存在两个及以上数据包的时隙，不仅解调复杂度增加，更重要的是为了能够成功解调，必然有设备需选择更高的传输功率来保证其 SINR 大于某个解调阈值，因为此时功率较低的设备信号被当作噪声。而对于 Sharma-S，当算法收敛后，每个时隙只分配给一个设备，此时影响解调的只有白噪声，从而只需要更低的发射功率。

因此，如何充分利用一帧中的所有时隙，使设备更加“智能”地以低功率在唯一的时隙中传输，成为解决上述问题的关键。

为了解决上述问题，在本文的能效优化机制中，首先，时隙和功率被组成（t, p），作为 MTC 设备的动作集合。如前文所述，一帧被分为 N 个时隙，考虑 P_{L}，P_{H} 两种发射功率，因此每个 MTC 设备共有 N×2 个动作可供选择，即 Q 表的大小为 N×2，这与 Silva-S 相似。然后，重点对 Q 学习的学习策略，奖励函数 R 和学习速率 α 进行了设计和改进。

2.2.1　学习策略

在能效优化机制中，接收端的解调器首先检测一帧中每个时隙内发生碰撞的数据包数 $CN_n, n=1,2,\cdots,N$，这可以通过前导搜索器有效实现[14]，然后将此信息作为全局信息在一帧的结束时通过卫星的下行链路广播给所有设备，这与 Sharma-S 类似。此时，MTC 设备便可以了解前一次传输时所选时隙和其他时隙的数据包的碰撞情况。与 Sharma-S 区别在于，Sharma-S 利用 eNB 广播的信息作为全局成本信息来设计奖励函数，本文则利用该全局信息改进 Q 学习算法的学习策略。

学习开始后，若设备 k 选择高功率 P_{H} 在第 n 个时隙传输，如果传输成功，则：

① 当第 n 个时隙的数据包数等于 1 时，即只有设备 k 选择了该时隙，则在下一次传输时，设备 k 尝试选择低功率 P_{L} 继续在第 n 个时隙传输。

② 当第 n 个时隙的数据包数大于等于 2，即通过 SIC 技术解调成功，则在下一次传输时，设备 k 首先根据接收到的全局信息，统计空闲时隙总数 N_{idle}，然后以概率 $p(N_{\mathrm{idle}})$使用低功率 P_{L}，在 N_{idle} 个空闲时隙中随机选择一个进行传输，$p(N_{\mathrm{idle}})$依据经验可采用如下形式：

$$p(N_{\mathrm{idle}})=\max\left(1,\frac{N_{\mathrm{idle}}^2}{500}\right) \tag{5}$$

式（5）表明空闲时隙越多，设备尝试低功率传输的概率就越大，使得一帧中的每个时隙尽可能只被一个设备占用，从而允许该设备使用低功率传输也可满足传输要求。

为了平衡 Q 学习的探索（Exploration）和利用（Exploitation），定义最大探索周期 $T_{\max}^{\mathrm{explore}}$，即上述改进的学习策略在设备第一次尝试接入时启动，经过 $T_{\max}^{\mathrm{explore}}$ 学习周期后结束，从而保证算法收敛后的稳定性。需要说明的是，该方案中的 Q 学习算法在整体上采用贪婪策略，设备 k 总是选择最高 Q 值的（t, p）作为它下一次的动作，则

$$\pi_k=\arg\max_{(t,p)} Q(t,p),\forall(t,p). \tag{6}$$

其中，π_k 表示设备 k 下一次要采取的动作策略，即（t, p）的选择。

总之，本文提出的能效优化方法能充分发掘和利用卫星广播的全局信息，使 MTC 设备在这些信息的协助下，更加“智能”地学习到最优策略，在保证传输成功率的前提下，降低传输功耗。

2.2.2　奖励函数

Sharma-S 中引入时隙的拥塞水平作为奖励函数，Silva-S 中则采用一个二元的奖励，即传输成功奖励 $R=1$，传输失败奖励 $R=-1$。在本文中，发射功率被考虑用于奖励函数的设置。

$$R(k,(t,p))=\begin{cases}\log(1+\dfrac{P_{\min}}{P_t}), & \text{transmission succeeds,}\\ -\log(1+\dfrac{P_t}{P_{\max}}), & \text{transmission fails,}\\ -\log(1+\dfrac{P_{\min}}{P_{\max}}), & \text{others.}\end{cases} \tag{7}$$

其中，$P_{\max}$ 和 $P_{\min}$ 分别表示允许使用的最高和最低发射功率，P_t 表示设备选择的发射功率，本文只考虑有限多的离散功率水平。

式（7）的设计以能量效率优化为出发点，其含义为，设备选择高功率传输时，如果传输成功，奖励相对较低，这将使设备偏向选择低功率以获得更大的奖励。如果传输失败，惩罚却更高，因为此时消耗了更多的能量，同时对其他数据包的解调也造成更大的干扰。此外，正负奖励的绝对值范围都为(0,ln2)，保证所有设备具有一致的行为[4]。

2.2.3　学习速率

不同于文献[1,3,5,10,11]中将学习速率 α 取为定值，我们根据文献[4]对 α 探究结论，从经验上将其设计为关于迭代次数的指数衰减函数。

$$\alpha(t)=(r_{\text{init}}-\beta)\cdot\exp[-\frac{(t-t_0)^2}{2\sigma_r^2}]+\beta \tag{8}$$

其中，r_{init} 为初始学习速率；σ_r 为衰减因子，控制 α 的衰减速度；β 表示最小学习率[1,10]；t_0 为设备首次尝试接入的时间，t 为迭代轮次，t–t_0 表示迭代次数。

式（8）表明在学习初期，α 较大，即保留先前学习的经验较少，这有利于设备迅速将偏好转移到低发射功率上。随着迭代次数增加，α 逐渐下降到最小学习率 β，以使稳态对信道条件的微小变化（例如，偶尔的碰撞）表现出一定的鲁棒性[15]，算法 1 总结了本文提出的能效优化机制。

2.3　过载控制机制

当海量 MTC 设备同时接入卫星网络时，会造成系统过载，此时 Silva-S 的吞吐量会迅速下降，因此需要加入过载控制机制来改善其性能。由此，通过观察和研究设备在学习过程中的 Q 值特性，结合负载估计，在优化能效的基础上，又加入了一种限制因子自适应的 ACB 过载控制机制。

2.3.1　负载估计

针对随机接入中的负载估计或预测，目前已有很多工作对其进行了研究[16-18]。由于本工作主要关注 Q 学习算法的设计，因此，考虑采用通用的负载估计器[19]。假设 $\hat{G}$ 为归一化负载估计值，G 表示归一化负载实际值，则

$$\hat{G}=U\left(G-\lfloor\varepsilon G\rfloor,G+\lfloor\varepsilon G\rfloor\right) \tag{9}$$

$\varepsilon\in[0,1]$，表示负载估计误差范围，U 为离散均匀分布。本文中，为 ε 设定某个固定值来表示系统的估计误差。

2.3.2　基本原理

首先依据归一化负载 G 定义系统的负载状态，定义如下

$$LoadState=\begin{cases}\text{Low-load}, & G\leqslant G_{\text{LH}}^{\text{th}},\\ \text{High-load}, & G_{\text{LH}}^{\text{th}}<G\leqslant G_{\text{HO}}^{\text{th}},\\ \text{Overload}, & G>G_{\text{HO}}^{\text{th}}.\end{cases} \tag{10}$$

其中，$G_{\text{LH}}^{\text{th}}$、$G_{\text{LO}}^{\text{th}}$ 具体由系统参数决定。本文中它们的值分别为 1.0、2.0。$G_{\text{LH}}^{\text{th}}$ 表示设备数 K 等于时隙数 N 时的归一化吞吐量，$K/N=1.0$，$G_{\text{HO}}^{\text{th}}$ 反映系统的吞吐量 T 上限。

该机制中，当一个设备准备发送数据时首先初始化一个学习次数的阈值 $N_{\text{learn}}^{\text{th}}$。设 MTC 设备 k 经过

$N_{\text{learn}}^{\text{th}}$ 次学习后的最大 Q 值为 $Q_{\max}^{k}$，$Q_{\max}^{k}$ 在 Q 表中出现的次数为 $N_{Q\max}^{k}$。设备 k 经过 $N_{\text{learn}}^{\text{th}}$ 次学习后，如果满足 $Q_{\max}^{k}<0$ 且 $N_{Q\max}^{k}\geqslant 2$ 的条件，则该设备在随后的接入过程将受到限制，即可认为该设备被分到接入受限类，否则该设备属于接入允许类。

$Q_{\max}^{k}<0$ 且 $N_{Q\max}^{k}\geqslant 2$ 条件的含义为：设备经过 N_{learn}^{th} 次学习后依然未找到最佳的（t,p），表明此时环境可能相对“恶劣”，即系统可能处于过载状态，数据包的碰撞频繁，造成设备的学习效率很低，无法学习到最优的策略。因此，对于还未找到最优策略的设备不应总是尝试传输，这会造成更多次的传输失败，进而消耗更多的能量。

在能效优化基础上，根据负载估计和 Q 值特性，又加入一种新的 ACB 方法用于过载控制。新 ACB 方法中限制因子 p_{bar} 和设备的限制时间 T_{bar} [12]分别为

$$p_{\text{bar}}=\begin{cases}\dfrac{G_{\text{HO}}^{\text{th}}-T}{G-G_{\text{LH}}^{\text{th}}}, & 0\leqslant\dfrac{G_{\text{HO}}^{\text{th}}-T}{G-G_{\text{LH}}^{\text{th}}}<1,\\ 1, & \text{otherwise.}\end{cases}\tag{11}$$

$$T_{\text{bar}}=(0.7+0.6\times rand)\times T_{\text{ACB}}\tag{12}$$

其中，T、G 分别为归一化吞吐量和负载，T_{ACB} 为系统设定的限制时间常数，单位为帧。$rand$ 为区间[0,1]上均匀分布的随机数。由式（11）可知，当吞吐量 T 接近 $G_{\text{HO}}^{\text{th}}$ 时，限制因子 p_{bar} 接近 0，从而防止更多接入受限类的设备的接入，保证系统持续处于高吞吐量状态。T_{ACB} 和 p_{bar} 由地面站确定，并同上文提到的一帧中各时隙发生碰撞的数据包数一起由卫星广播传输给 MTC 设备。算法 2 描述了过载控制机制。

算法 1: 能效优化方法

对于设备 k，$\forall k\in\{1,2,\cdots,K\}$，outcome =1 表示前一次传输成功，$p_k^{\text{pre}}$ 和 n_k^{pre} 分别表示前一次选择的传输功率和时隙，p_k^{cut} 和 n_k^{cur} 表示当前要选择的传输功率和时隙；$N_{\text{coll}}^{\text{pre}}$ 表示前一次选择的时隙中数据包的碰撞数。

初始化 Q 表为 0，$t=t_0$；
for Every frame **do**
　if $t-t_0+1\leqslant T_{\max}^{\text{explore}}$ **then**
　　if $outcome=1$ && $p_k^{\text{pre}}=P_{\text{H}}$ **then**
　　　if $N_{\text{coll}}^{\text{pre}}=1$ **then**
　　　　$n_k^{\text{cur}}=n_k^{\text{pre}}$, $p_k^{\text{cur}}=P_{\text{L}}$;
　　　else
　　　　if $rand\leqslant p(N_{\text{idle}})$ **then**
　　　　　随机选择一个空闲时隙，并且 $p_k^{cur}=P_{\text{L}}$；
　　　　else
　　　　　根据式（6）选择(t,p);
　　　　end if
　　　end if
　　else
　　　根据式（6）选择(t,p);
　　end if
　else
　　根据式（6）选择(t,p);
　end if
　根据式（7）计算奖励；
　根据式（4）更新 Q 值；
　$t=t+1$;
end for

算法 2: 基于 Q 值特性和负载估计的 ACB 过载控制机制

对于设备 k，$\forall k\in\{1,2,\cdots,K\}$
1．初始化 $n=0$; // n 为传输次数
for Every frame **do**
　if $n\leqslant N_{\text{learn}}^{\text{th}}$ **then**
　　直接接入；
　　$n=n+1$;
　else
　　// 满足 Q 值特性条件
　　if $Q_{\max}^{k}<0$ && $N_{Q\max}^{k}\geqslant 2$
　　　if $rand\leqslant p_{\text{bar}}$ **then**
　　　　允许继续接入；
　　　　$n=n+1$;
　　　else
　　　　在随后的 T_{bar} 个帧内禁止接入；
　　　end if
　　else
　　　直接接入；
　　　$n=n+1$;
　　end if
　end if
end for

3　仿真结果

在该部分，首先验证了所提方案 I-NORA-QL 在能效优化和过载控制上的有效性，然后探究了控制 α 衰减速率的衰减因子 σ_r 对能效的影响，参数设置如表 1 所示。注意，由于发射功率 P_{L}、P_{H} 值的设置，一个时隙中最多允许两个设备同时接入，否则无法解调成功，因此吞吐量的上限 $T_{\mathrm{upper}}=2.0$，对应上文 $G_{\mathrm{HO}}^{\mathrm{th}}=2$。

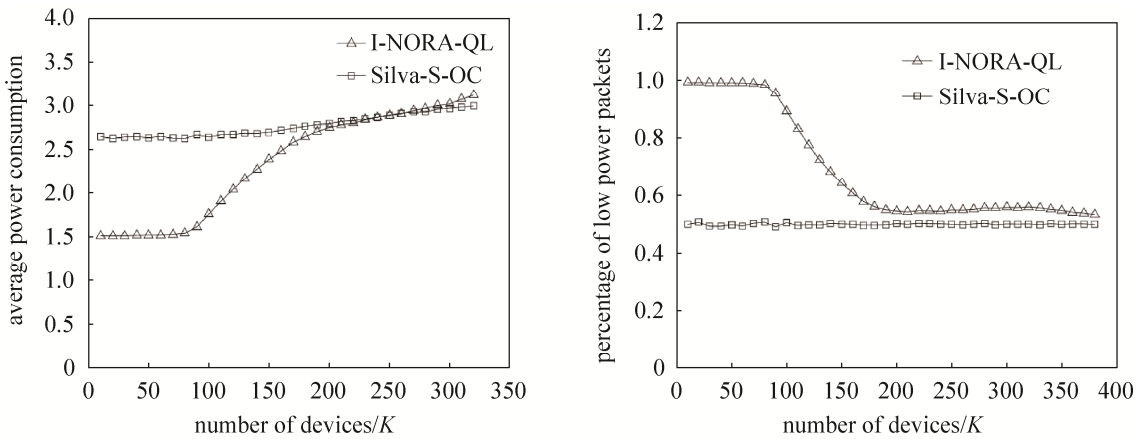
表 1　仿真参数设置

Table 1　Simulation parameter settings

参数	值
时隙数 N	100
MTC 设备数 K	10~380
发射功率 $P_{\mathrm{L}}, P_{\mathrm{H}}$	1.76 dB, 5.74 dB
噪声功率 σ^2	0 dB
SIC 解调阈值 γ_{th}	1.2 dB
最大探索周期 $T_{\max}^{\mathrm{explore}}$	10
初始学习率 r_{init}	0.5
最小学习率 β	0.1
衰减因子 σ_r	{2, 4, 6, 8, 16}
负载估计误差 ε	0.15
学习次数阈值 $N_{\mathrm{learn}}^{\mathrm{th}}$	10
限制时间 T_{ACB}	5

3.1　能效性能

为了对比的公平性，Silva-S 中也加入本文提出的过载控制机制，命名为 Silva-S-OC。每个设备成功传输 100 个数据包的平均功耗和低功率数据包占比的比较如图 2 所示。



（a）平均功耗　　（b）低功率数据包占比

（a）Average power consumption　　（b）Percentage of low power packets

图 2　平均功耗和低功率数据包占比的比较（$\sigma_r=4$，Silva-S-OC 中学习速率 $\alpha=0.1$）

Fig. 2　Comparison of average power consumption and percentage of low-power packets ($\sigma_r=4$, learning rate $\alpha=0.1$ in Silva-S-OC)

从图 2（a）中可以看出，当 $K\leqslant 80$($G\leqslant 0.8$) 时，在 I-NORA-QL 中，设备平均功耗接近设置的低功率 P_{L}=1.5（1.76 dB），对应到图 2（b），此时几乎所有设备都选择低功率传输，这是由于设备数 K 小于时隙数 N 且相差较大，使设备在探索过程中，更容易找到唯一的时隙并以低功率 P_{L} 传输数据。当 $80<K\leqslant 200$ 时，随着 K 的增加，系统的负载逐渐加大，为了保证能成功传输，设备必须以更高的概率使用高功率进行传输，使得平均功耗逐渐增加，最终与 Silva-S-OC 几乎持平。在图 2（b）中，则表现为低功率数据包的比例逐渐下降。而对于 Silva-S-OC，在非过载时（$K\leqslant 200$），无论何种负载，其平均功耗一直稳定在 2.6 左右，即高、低功率数据包几乎总是各占总数的一半，这导致平均功耗相较 I-NORA-QL 大幅度增加。当 $K>250$ 时，虽然此时 I-NORA-QL 低功率数据包占比略高于 Silva-S-OC，但平均功耗相较 Silva-S-OC 却有增加，且当 $K>300$ 时，平均功耗增加的幅度会随着设备数的增加而迅速增大，这是由于当系统过载严重时，I-NORA-QL 中设备对低功率的偏爱反而不利于整个系统的收敛，使得接入成功率下降，碰撞增加。因此，所提方案适合工作在归一化负载 $G<3.0$ 的情况，并且可在 $G<2.0$ 时较为显著地降低设备的平均功耗。

因此，I-NORA-QL 在系统非过载状态下能够降低 MTC 设备的功耗，负载较低时尤为显著，从而增加电池的使用寿命。特别地，当归一化负载 $G\leqslant 0.8$ 时，本方案等价于 Sharma-S 中设备只使用低功率 P_{L} 传输。由此，I-NORA-QL 综合了 Sharma-S 和 Silva-S 在不同负载下的优势，在提高系统吞吐量的前提下，通过改进 Q 学习算法优化 MTC 设备的能量效率，确保设备的长期工作。

3.2　衰减因子 σ_r 对能效的影响

设置 $\sigma_r=2,4,6,8,16$，得到 5 种不同衰减速率的学习速率 α，同时将固定的 $\alpha=0.1$ 作为参考，观察和分析 σ_r 与平均功耗的关系。

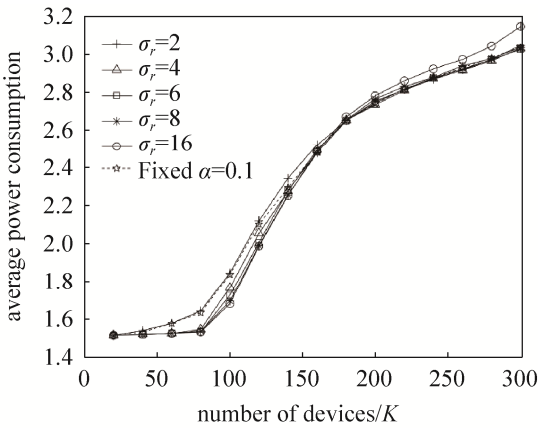


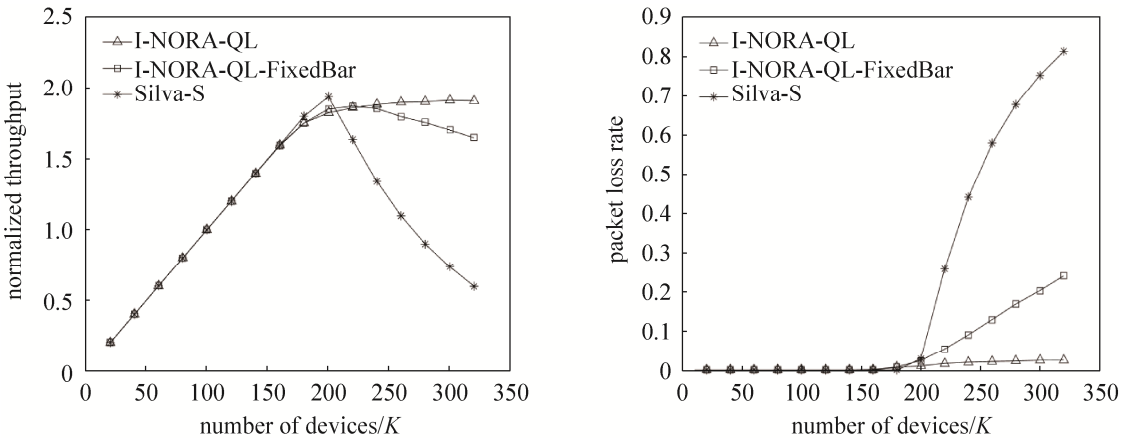
图 3　衰减因子 σ_r 对能效的影响

Fig. 3　Effect of attenuation factor σ_r on energy efficiency

如图 3 所示，首先证明了将学习速率设置为指数衰减形式在提高能效上的积极作用。当设备数 K<200 时，随着 σ_r 的增加，不同负载下设备的平均功耗相对就越低，但降低的程度在减小，这是因为 σ_r 为 $\alpha(t)$ 标准差，σ_r 越大，$\alpha(t)$ 衰减越慢，因此设备在学习阶段的初期，它的学习速率相对就更大。由式（4）可知，学习速率越大，设备保留之前学习的经验就越少，因此将更关注当前的奖励，使其更快地从当前选择的策略跳转到另一个更佳的策略，即使用低功率传输，实现能效优化。此外，由 3.1 节分析可知，随着设备数的增加，设备将以更高的概率使用高功率以保证传输成功，因此设备的平均功耗快速增加，从而也导致不同 σ_r 参数下平均功耗的差异逐渐减小。而当 $K\geqslant 200$，σ_r 取过大时（对应 $\sigma_r=16$），其平均功耗相较于其他 σ_r 值有所增加，并且增加量随着负载的增大而增大，由 2.2.3 节分析可知，一个更大的 σ_r 值将导致学习速率在更长的时间内保持较大的值，这在学习阶段的后期，会使设备对信道条件变化的鲁棒性降低，不利于系统的收敛，造成更多的碰撞和能量消耗，并且这种情况会随着负载的增大而加重。因此综合考虑，σ_r 取值区间为[6, 8]，此时可更好地平衡不同负载下设备的平均功耗。

3.3　过载控制性能

为说明 I-NORA-QL 在过载时的性能优势，我们比较了 Silva-S、I-NORA-QL-FixedBar 和 I-NORA-QL 在吞吐量和丢包率上的表现，如图 4 所示。I-NORA-QL-FixedBar 中的 ACB 限制因子设定值 p_{bar}=0.5 [20]。



（a）吞吐量比较　　（b）丢包率比较

（a）Throughput comparison　　（b）Packet loss rate comparison

图 4　吞吐量和丢包率比较（$\sigma_r=4$，Silva-S 中学习速率 $\alpha=0.1$）

Fig. 4　Comparison on throughput and packet loss rate（$\sigma_r=4$, learning rate $\alpha=0.1$ in Silva-S-OC）

当 K≤160 时，系统负载还未过高，三种方案中的所有设备几乎都能通过 Q 学习找到唯一的（t, p）对进行传输，因此，图 4（b）中的丢包率极低，接近 0。当 160<K≤200 时，系统逐渐接近过载，此时 I-NORA-QL 和 I-NORA-QL-FixedBar 吞吐量上略低于 Silva-S，这是由于这两个方案的过载控制机制被激活，使得经过 N_{learn}^{th} 学习后仍无法找到最佳策略的少量设备被禁止继续传输，而 Silva-S 中设备由于没有限制，可以进行更多次的失败和重传，直到找到最佳的学习策略，因此吞吐量有些许提高。当 K>200 时，系统过载并且逐渐加重，Silva-S 吞吐量迅速下降，丢包率迅速上升，而对于加入 ACB 过载控制的 I-NORA-QL 和 I-NORA-QL-FixedBar，经过一定的学习周期后，未找到最佳（t, p）对的设备自动地被分到接入受限类，从而减少了参与竞争的设备数，大幅降低了丢包率，同时避免了更多能量的白白消耗。随着过载程度的加重，由于 I-NORA-QL-FixedBar 中的限制因子 p_{bar} 为定值，缺少自适应能力，因此它的吞吐量逐渐低于 I-NORA-QL，而丢包率逐渐高于 I-NORA-QL。

通过以上三个方案的比较，一方面验证了将依据负载估计和 Q 值特性的 ACB 方法加入到基于 NOMA 的 Q 学习随机接入的有效性，另一方面也表明了自适应的 ACB 限制因子能够更有效地提高系统的吞吐量，降低丢包率。

为详细展示加入过载控制机制后的数据传输过程，设置 K=240，每个设备需要成功发送 100 个数据包，在一次实现过程中，传输数据的设备总数 $M_{\text{total}}^{\text{trans}}$ 和成功传输的设备数 $M_{\text{success}}^{\text{trans}}$ 的变化曲线如图 5 所示。

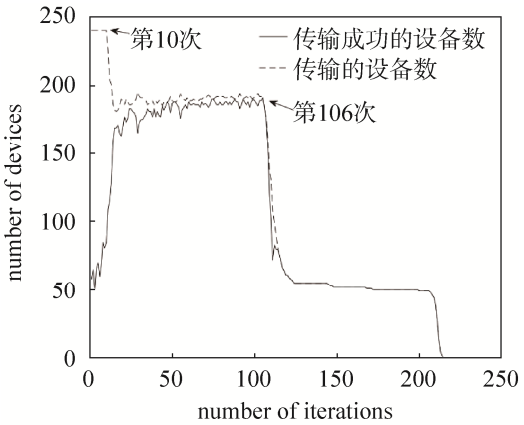


图 5　提出的过载控制方案实现过程（$\sigma_r = 4$）

Fig. 5　Implementation process of the proposed overload control scheme ($\sigma_r = 4$)

图 5 中，当 $N_{\text{learn}}^{\text{th}} > 10$，$M_{\text{total}}^{\text{trans}}$ 快速下降，$M_{\text{success}}^{\text{trans}}$ 迅速增加，原因在于部分设备被禁止继续接入，使系统负载下降，碰撞减少，进而使吞吐量迅速提高。随着迭代次数的增加，两条曲线逐渐接近，波动逐渐平缓，此时 $M_{\text{success}}^{\text{trans}} \approx M_{\text{total}}^{\text{trans}}$，表明接入成功率接近 1。第 106 次迭代后，那些一直允许接入的设备的所有数据包已经传输完毕，此时 $M_{\text{total}}^{\text{trans}}$ 迅速下降，而原先被禁止接入的设备开始重新接入，并经过短时间的学习，这些设备也都能找到最佳的（t, p），因此两条曲线快速重合，直到结束。

4　结束语

本文主要研究了 Silva-S 存在的传输能效低和系统过载时吞吐量下降的问题，并对其进行改进。功率域 NOMA 的应用虽然提高了系统的吞吐量，但也导致设备功耗的增加，降低了能量效率。为了解决这个问题，从 Q 学习本身入手，利用所需的全局信息，对 Q 学习的学习策略进行改进，同时将发射功率作为参数用于奖励函数的设置，学习速率不再固定为常量，而是被设计成随迭代轮次递减的函数形式。实验结果表明，I-NORA-QL 能够降低设备的平均功耗，在负载较低时更加显著，这对能量受限的 MTC 设备至关重要。此外，本文还探究了影响学习速率的衰减因子 σ_r 对平均功耗的影响，我们发现当 σ_r 取值在[6,8]之间时可在能效上产生更加积极的作用，并结合 Q 学习的特点分析了原因。I-NORA-QL 中的过载控制则是利用设备学习过程中的 Q 值特性和学习效率之间的关系，为设备设定学习次数的阈值，当设备学习次数超过该阈值，并且其 Q 值满足设定的条件时，该设备将被划分至接入受限类，其随后的接入过程将受 ACB 限制因子的控制，并且限制因子与系统的负载和吞吐量相关。I-NORA-QL 中的过载控制机制保证系统在过载时，依然能维持高吞吐量和低丢包率。

参考文献

[1] SHARMA S K, WANG X. Collaborative distributed Q-learning for RACH congestion minimization in cellular IoT networks[J]. IEEE Communications Letters, 2019, 23(4): 600–603.

[2] MOON J, LIM Y. Access control of MTC devices using reinforcement learning approach[C]//2017 International Conference on Information Networking(ICOIN), 2017: 641–643.

[3] BELLO L M, MITCHELL P D, GRACE D. Application of Q-learning for RACH access to support M2M traffic over a cellular network[C]//European Wireless European Wireless Conference, 2014: 1–6.

[4] HUSSAIN F, ANPALGAN A, KHWAJA A S, et al. Resource allocation and congestion control in clustered M2M communication using Q-learning[J]. European Transactions on Telecommunications, 2017, 28(4): e3039.

[5] ZHAO Bo, REN Guangliang, Dong Xiaodai, et al. Distributed Q-learning based joint relay selection and access control scheme for IoT-Oriented satellite terrestrial relay networks[J]. IEEE Communications Letters, 2021, 25(6): 1901–1905.

[6] WU Y, ZHANG N, RONG K. Non-Orthogonal random access and data transmission scheme for machine-to-machine communications in Cellular networks[J]. IEEE Access, 2020, 8: 27687–27704.

[7] ALVI S, DURRANI S, ZHOU X J. Enhancing CRDSA with transmit power diversity for machine-type communication[J]. IEEE Transactions on Vehicular Technology, 2018, 67(8): 7790–7794.

[8] CHOI J. NOMA-based random access with multichannel ALOHA[J]. IEEE Journal on Selected Areas in Communications, 2017, 35(12): 2736–2743.

[9] HAN S, XU X, LIU Z. Energy-efficient short packet communications for uplink NOMA-based massive MTC networks[J]. IEEE Transactions on Vehicular Technology , 2019, 68(12): 12066–12078.

[10] DA SILVA M, SOUZA R D, ALVES H, et al. A NOMA-based Q-learning random access method for machine type communications[J] IEEE Wireless Communications Letters, 2020, 99: 1–1.

[11] SHI Z, GAO W, LIU J, et al. Distributed Q-Learning-Assisted Grant-Free NORA for massive machine-type communications[C]// GLOBECOM 2020-2020 IEEE Global Communications Conference, 2020: 1–5.

[12] PHUYAL U, KOC A T, FONG M H, et al. Controlling access overload and signaling congestion in M2M networks[C]//2012 Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers(ASILOMAR), 2012: 591–595.

[13] JESSE C, ERIC L. Q-learning: theory and applications[J]. Annual Review of Statistics and Its Application, 2020, 7, (1): 279-301.

[14] ASINI E C, GAUDENZI R D, HERRERO O R. Contention resolution diversity slotted ALOHA(CRDSA): an enhanced random access scheme for satellite access packet networks[J]. IEEE Transactions on Wireless Communications, 2007, 6(4): 1408–1419.

[15] CHU Y, KOSUNALP S, MITCHELL P D, et al. Application of reinforcement learning to medium access control for wireless sensor networks[J]. Engineering Applications of Artificial Intelligence, 2015, 46(A): 23–32.

[16] FENG Y, REN G. LS-SVM based large capacity random access control scheme in satellite network[M]. Space Information Networks, 2018: 275–287.

[17] JIANG N, DENG Y, SIMEONE O, et al. Online supervised learning for traffic load prediction in framed-ALOHA networks[J]. IEEE Communications Letters, 2019, 23(10): 1778–1782.

[18] FEI C, JIANG B, XU K, et al. An intelligent load control-based random access scheme for space-based internet of things[J]. Sensors, 2021, 21: 1040.

[19] CHELLE H, CROSNIER M, DHAOU R, et al. Adaptive load control for IoT based on satellite communications[C]//2018 IEEE International Conference on Communications(ICC), 2018: 1–7.

[20] TELLO-OQUENDO L, LEYVA-MAYORGA I, PLA V, et al. Performance analysis and optimal access class barring parameter configuration in LTE-A networks with massive M2M traffic[J]. IEEE Transactions on Vehicular Technology, 2017, 67(4): 3505–3520.

[作者简介]

杨伟康 1998 年生，在读硕士研究生，主要研究方向为卫星通信。

许小东 1976 年生，博士，副教授，主要研究方向为无线通信与信号处理。

(本文编辑：杨秀丽)