IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 24, 2022, accepted March 13, 2022, date of publication March 16, 2022, date of current version March 24, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3160156

D2D Assisted Q-Learning Random Access for NOMA-Based MTC Networks

MATHEUS V. DA SILVA[1], SAMUEL MONTEJO-SÁNCHEZ[2], (Senior Member, IEEE), RICHARD DEMO SOUZA[3], (Senior Member, IEEE), HIRLEY ALVES[1], (Member, IEEE), AND TAUFIK ABRÃO[4], (Senior Member, IEEE)

[1]Centre for Wireless Communications, University of Oulu, 90014 Oulu, Finland
[2]Programa Institucional de Fomento a la I+D+i, Universidad Tecnológica Metropolitana, Santiago 8940577, Chile
[3]Department of Electrical and Electronics Engineering, Federal University of Santa Catarina, Florianópolis 88040-900, Brazil
[4]Department of Electrical Engineering, State University of Londrina (UEL), Londrina 86057-970, Brazil

Corresponding author: Matheus V. da Silva (matheus.valentedasilva@oulu.fi)

This work was supported in part by the Brazilian National Council for Scientic and Technological Development (CNPq), Brazil; in part by the Print Coordenação de Aperfeiçoamento de Pessoal de Nível Superior-Universidade Federal de Santa Catarina (CAPES-UFSC) ‘‘Automation 4.0,’’ Brazil; in part by Rede Nacional de Ensino e Pesquisa/Ministério da Ciência, Tecnologia, Inovações e Comunicações (RNP/MCTIC), Brazil, under Grant 01245.010604/2020-14 (6G Mobile Communications Systems); in part by Fondo Nacional de Desarrollo Científico y Tecnológico (FONDECYT) Iniciación, Chile, under Grant 11200659; and in part by the Academy of Finland 6Genesis Flagship, Finland, under Grant 318927.

ABSTRACT Machine-type communications (MTC) should account for half the connections to the internet by 2030. The use case massive MTC (mMTC) allows for applications to connect a massive number of low-power and low-complexity devices, leading to challenges in resource allocation. Not only that, mMTC networks suffer under rigid random access schemes due to mMTC ultra-dense nature resulting in poor performance. In this sense, this paper proposes a Q-Learning-based random access method for massive machine-type communications, with device clustering and non-orthogonal multiple access (NOMA). The traditional NOMA implementation increases spectral efficiency, but at the same time, demands a larger Q-Table, thus slowing down convergence, which is known to be a highly detrimental effect on massive networks. We use pre-clustering through short-range device-to-device technology to mitigate this drawback, allowing devices to operate with a smaller Q-Table. Furthermore, the previous selection of partner devices allows us to implement a full-feedback-based reward mechanism so that clusters avoid time slots already successfully allocated. Additionally, to cope with the negative impact of system overload, we propose an adaptive frame size algorithm to run in the base station (BS). It allows adjusting the frame size to the network load, preventing idle slots in an underloaded scenario, and providing extra slots when the network is overloaded. The results show the great benefits in terms of throughput of the proposed method. In addition, the impact of the use of clustering and the size of the clusters, as well as the frame size adaptation, are analyzed.

INDEX TERMS mMTC, NOMA, reinforcement learning, Q-learning, 6G, random access.

I. INTRODUCTION

5G technology inherently supports critical and massive machine-type communications (MTC) [1]. The development and deployment of MTC networks has grown even more with applications such as smart cities [2] and smart industries [3].

MTC should represent half of the connections to the Internet by 2030, reaching about 14.7 billion connected devices [4]. Such a fact raises the question, how will next-generation communication systems support MTC applications?

It is undeniable that the new 6G service classes require significant physical (PHY) and medium access control (MAC) layers enhancements to ensure massive connectivity. The authors in [5] suggest the use of non-orthogonal solutions [6], channel state information (CSI) free/limited schemes [7], and coding for short packets [8] already at the PHY. Notably, they emphasize that: *(i)* the likelihood of operating with a strong line-of-sight increases with denser networks and statistical beam-forming relying on channel statistics can operate with near-optimum performance without the need for CSI acquisition [9]; and *(ii)* coding for short packets [10] becomes vital as the coding schemes for 5G (low-density-parity-check and polar codes) are not optimized for short packets.

The associate editor coordinating the review of this manuscript and approving it for publication was Jiankang Zhang.

MAC challenges include the need for modern random access (RA) schemes [11] since scheduling a vast amount of devices becomes impractical, and pure random access schemes like ALOHA have severe performance limitations [12]. Unlike pure RA schemes, intelligent RA methods can leverage the fact that collisions will happen, using them as a learning opportunity. In addition, applying successive decoding at the Base Station (BS) can resolve collisions when devices use the same slot. In this sense, non-orthogonal multiple access (NOMA) can improve resource sharing and spectral efficiency. NOMA can be considered a promising solution for massive MTC [13]. Combining NOMA and grant-free access schemes can reduce the effective device density and system overhead [14]. Furthermore, in power-domain NOMA, intelligent interference techniques can recover the transmitted signals.

However, massive MTC networks can suffer from inefficient RA schemes, as the current rigid models perform poorly in ultra-dense networks [11]. In addition, allocating transmission resources to MTC is challenging, urging intelligent schemes to learn the network characteristics. Machine learning models are often used to acquire characteristics that an explicit mathematical model can not readily obtain. For example, among the different machine learning methods, reinforcement learning is helpful in modeling various wireless communications problems [15]. Among the reinforcement learning techniques, Q-Learning stands out because of its capability of being implemented in a model-free and distributed manner [16]. A comprehensive survey in [17] discusses the issues in radio access network congestion and how machine learning techniques can improve RA in massive MTC networks, pointing out Q-Learning as a potential solution. Furthermore, Deep Q-Learning appears to improve resource allocation in wireless networks, being used with NOMA in [18] to maximize grant-free Aloha-like system throughput. Nonetheless, Deep Q-Learning can be too complex and computationally intensive for MTC devices.

A. RELATED WORK

Bello *et al.* [19] introduced a Q-Learning algorithm to conciliate RA involving human-type communication (HTC) and MTC devices in a cellular network. The MTC devices actively learn which time slots to access, avoiding collisions among MTC devices while also increasing the performance of HTC users. The reward is fed back via a single bit per time slot, which indicates the transmission's success or failure. Moreover, the back-off frame size can be dynamically adjusted according to the blocking probability experienced by the HTC users. With a focus on MTC, the authors of [20] propose a distributed Q-Learning RA algorithm using the number of collisions per time slot as a reward, where devices make independent decisions when choosing a transmitting time slot within a frame. However, the proposed approach needs substantial feedback from the BS to the devices, as sending the so-called congestion level can consume several bits per time slot, besides the unclear practical feasibility of determining the exact number of colliding devices in each time slot.

Based on [20], we propose a distributed Q-Learning algorithm exploiting power-domain NOMA in [21] for increased spectral efficiency. Each device can learn the best time slot to transmit and its power level and NOMA partners. The proposed solution achieves considerable gains in throughput and feedback complexity compared to [20], as NOMA increases the spectral efficiency. In contrast, in [21] the BS feedback contains only a single bit per time slot. However, the Q-Table size increases with the power levels, slowing convergence. Next, [22] assumes a similar setup as in [21] and considers the effect of finite blocklength and imperfect successive interference cancellation (SIC). Then, they exploit the block error rate (BLER) as a reward in order to improve the performance of the Q-Learning algorithm. However, using the BLER as a reward can make the feedback longer; while it is not clear how in practice, one could perfectly estimate the BLER of the devices in an interference scenario. In [23], the authors also introduce a Q-Learning scheduling method with SIC. However, they do so in an ad-hoc scenario.

Another work to investigate the use of power-domain NOMA and distributed Q-Learning to improve RA is [24]. The authors consider multiple power levels, multiple channels, sporadic traffic, and design a reward based on the activation probability of the devices. Even though sporadic traffic is considered after convergence, training on a saturated network is still required, and devices have to learn which are their best NOMA partners, leading to a slow convergence as in [21]. In the same line, in [25] the authors consider sporadic traffic and propose a Q-Learning RA algorithm using a reward system based on the successful transmission probability. However, [25] introduces intermittent learning, in which the devices update their Q-Tables from time to time, and non-orthogonal transmissions are supported by sparse coded multiple access (SCMA). Moreover, the authors propose an algorithm variant in which just part of the devices is involved in the learning process. For that sake, they assume that devices are grouped *a priori*, where only devices in the high activation probability group run the algorithm, reducing energy consumption and system complexity. However, this method, as [20], [22], [24], relies on the BS knowledge of how many or which devices collided when trying to access a particular resource, which can be very difficult to estimate in practice. Moreover, the reward has several bits per resource.

The work in [26] also considers a distributed Q-Learning aided RA procedure using SCMA. However, differently from [25], the devices run two separate algorithms, one for learning the time slot and another for the codebook. Moreover, the reward is different for each case, based on the congestion level [20] for the time slot and a binary variable for the codebook. Therefore, the BS feedback can be relatively large while also demanding knowledge of how many devices collided in a given resource. Finally, a collaborative Q-Learning algorithm for subcarrier assignment in wideband cognitive radio systems is introduced in [27]. The BS transmits to the secondary users, from time to time, information on the subcarriers occupied or not by other devices. Such full feedback exploitation, that considers the success or failure at a particular resource, is beneficial. It is remarkable that considering only the feedback from a single resource, devices using the methods in [20]–[22], [24]–[26] miss several learning opportunities.

B. NOVELTY AND CONTRIBUTION

In this work, we propose using Q-Learning and NOMA with clustering, alongside an adaptive frame size algorithm, to improve the throughput in massive MTC networks. Devices do not use Q-Learning to find their partners or power levels, reducing the Q-Table size, the convergence time, and the complexity. Instead, they use short-range device-to-device (D2D) communications to self-organize in clusters. Moreover, several works try to improve Q-Learning allocation methods by designing new rewards with more information, which usually leads to a large feedback size, and unrealistic or inefficient models. In this work, rather than adding information to the feedback, we better use all the information available within a simple feedback, improving collision avoidance ability and faster convergence.

Our work differs from [19], [20], [27] because, besides using Q-Learning for resource allocation purposes, we implement NOMA for improving spectral efficiency. Moreover, different from [20], [22], [24]–[26], our method requires minimal feedback from the BS, a single bit per time slot. Similar to [27] and different from [19]–[22], [24]–[26], we fully exploit the feedback sent by the BS so that devices avoid the time slots already in use. Different from [26], we implement only one learning algorithm on the device side, making use of clustering to resolve the resource sharing issue (transmit power in our case, codebook in the case of [26]). Compared to [25], the devices self-organize in small clusters, which speeds up convergence, while in [25] it is not discussed how grouping is implemented. Furthermore, unlike most related work [20]–[22], [24]–[27], we implement a frame size adaptation mechanism, which allows the method to adjust to overloaded or underloaded scenarios. Apart from the above, we consider constant learning instead of intermittent learning as in [25], as permanent learning considerably speeds up convergence. Moreover, constant learning can be used only during convergence, as in [24], if devices do not transmit periodically. Note that the learning process happens in saturated traffic, which is not typical of MTC networks. However, by reaching convergence quickly devices could then move on from a short training phase into standard operation with sporadic traffic [28]. A comparison of the proposed method with the closest literature is presented in Table 1, where the main technical scopes of our proposal are highlighted.

The contribution of this work can be summarized as follows:

- We propose a distributed NOMA Q-Learning RA method with D2D clustering, a full-feedback-based reward (fFbR) mechanism, and an adaptive frame size algorithm.
- We significantly improve the convergence speed concerning the literature by reaching the maximum throughput in a few (just over 10) iterations.
- The throughput improves, *e.g.*, 18.55% at 200 devices and 240.28% at 250 devices, compared to [21]. At the same time, we decrease the Q-Table size thrice, reducing the learning process and computational complexity accordingly.

The rest of this paper is organized as follows. Section II introduces the system model. In Section III, the NOMA power allocation is discussed, considering fixed- and dynamic-ordered SIC schemes. In Section IV, the proposed method is presented in detail. Next, numerical results are provided in Section V. Finally, the paper is concluded in Section VI. Table 2 presents a list of acronyms used in this work and the main variables are summarized in Table 3.

II. SYSTEM MODEL

Assuming a stand-alone IoT network, we consider a setup with N synchronized devices distributed uniformly around a BS in a single circular cell. Every device has L data packets ready for transmission. Medium access is based on grant free slotted Aloha, where each device can transmit in one of K time slots within a frame. All devices transmit at the same frequency and data rate and with a given transmit power, leading to one of the average received powers in $\Omega = \{\omega_0, \omega_1, \cdots, \omega_m, \cdots, \omega_{M-1}\}$, where ω_0 is a target received power that leads to a predefined operating maximum outage probability $\mathcal{O}_{\text{ref}}$ when the devices is transmitting alone (*i.e.*, without interference). Moreover, ω_m is the target received power that guarantees successful decoding of the message in the m^{th} power level, $m \in \{0, 1, \cdots, M-1\}$, given the presence of up to m interfering signals, each in one of the m smaller power levels in Ω. Note that every device has its own transmit power computed via channel inversion considering *i)* the estimation of its average path-loss using a control message broadcast by the BS between data frames assuming time division duplex reciprocity; and *ii)* the particular average power ω_i which is the intended signal to be received at the BS.

We assume the devices can find partners in its own cluster,[1] via a short-range D2D technology, as illustrated in Fig. 1. The devices within the c^{th} cluster, $c \in \{1, 2, \ldots, C\}$, share the same time slot and each device transmits at a different power as to yield one of the M possible receive powers at the BS. Therefore, we exploit D2D communication within each cluster to set up NOMA transmission of up to M devices in the same time slot. Clustering also allows evaluating slot

[1] Clustering can be implemented through short-range D2D technology, *e.g.*, Bluetooth Low Energy (BLE) allows devices to last years with a single-coin battery [29]. The discovery and connection time happens within a few milliseconds [29], rapidly forming clusters. BLE supports one-to-many devices communications [30], enabling clusters of over two devices. Besides, we expect future radios to support multiple radio access interfaces [31]. Despite the variable range of D2D technologies, here we limit the clustering range to 15 meters.

TABLE 1. Comparison of *Q*-learning schemes for adaptive RA.

Methods	NOMA/SCMA	Adaptive Frame Length	Clustering	D2D Collaboration	Exploits Full Feedback	Feedback Overhead
[19]		✓				Small
[20]						Large
[21]	✓					Small
[22]	✓					Large
[24]	✓					Large
[25]	✓		✓			Large
[26]	✓					Large
[27]					✓	Small
This work	✓	✓	✓	✓	✓	Small

A feedback is considered *small* when it uses just ACK bits and *large* when the value for each slot/resource has to be represented by several bits.

TABLE 2. List of acronyms.

Acronyms	Meaning
AWGN	Additive White Gaussian Noise
BLER	BLock Error Rate
BS	Base Station
CA	Collision Avoidance
CSI	Channel State Information
D2D	Device-to-Device
fFbR	full-Feedback-based Reward
FTX	Failed Transmissions
HTC	Human-Type Communications
IDS	IDle Slots
MAC	Medium Access Control
MDP	Markov Decision Process
MIS	Maximum number of devices In a Slot
mMTC	massive Machine-Type Communications
MTC	Machine-Type Communications
NOMA	Non-Orthogonal Multiple Access
PHY	PHYsical
RA	Random Access
RAN	Radio Access Network
SCMA	Sparse Coded Multiple Access
SIC	Successive Interference Cancellation
SINR	Signal to Interference plus Noise Ratio
SNR	Signal to Noise Ratio
SWC	Slots with Collision
TPT	Throughput

TABLE 3. List of variables.

Variable	Description
x_k	Attenuated signal
ω_m	Average receive power
Ω	Average receive power set
C	Cluster quantity
M	Cluster size
$d_{m,k}$	Device distance to the BS
$PL_{m,k}$	Device path loss
γ	Discount factor
K	Frame size
β	Future reward complexity factor
α	Learning rate
$\mathcal{O}_{\text{ref}}$	Maximum device outage probability
$\mathcal{O}_{\text{SIC}}$	Maximum sic outage probability
F	Noise figure
N_0	Noise PSD
σ^2	Noise power
X	Number of collisions
N	Number of devices
η	Path loss exponent
S	Period of frame size adaptation
$h_{m,k}$	Rayleigh fading
G_{Rx}	Receiver gain
d_0	Reference distance
R_u	Reward
y_k	Signal received at the BS
r	Spectral efficiency
G_{Tx}	Transmitter gain

allocation only for cluster heads, which is shared with the other cluster members via D2D communication, increasing the time and energy efficiency of the resource allocation process. Note that there is no inter-cluster communication. The cluster head learns solely through the feedback from the BS.

The signal received at the BS in the k^{th} time slot, coming from a single cluster of M transmitting devices, can be written as

$$y_k = \sum_{m=0}^{M-1} x_{m,k} + n_k, \tag{1}$$

with $x_{m,k}$ being the attenuated signal vector received at the BS from the m^{th} device, $m \in \{0, 1, \cdots, M-1\}$, $M \leq N$, in the k^{th} time slot, $k \in \{0, 1, \cdots, K-1\}$, subject to fading and path loss, with instantaneous received power $\omega_{m,k}|h_{m,k}|^2$, where $h_{m,k}$ is Rayleigh fading, independent and identically distributed in time and space, while $\omega_{m,k} \in \Omega$ is the average received power from the m-th device in the k^{th} time slot. Finally, n_k is the additive white Gaussian noise (AWGN), with power $\sigma^2 = FN_0B$, where N_0 is the noise power spectral density, B is the bandwidth, and F is the noise figure [32].

Moreover, path loss (PL) between the devices and the BS is determined considering a log-distance model [33],

$$\mathrm{PL}_{m,k} = \mathrm{PL}(d_0) + 10\eta \log_{10}\left(\frac{d_{m,k}}{d_0}\right) - G_{Tx} - G_{Rx}, \tag{2}$$

where $d_{m,k}$ is the distance from that device to the BS, d_0 is the reference distance, $\mathrm{PL}(d_0)$ is calculated using the Friis equation [32], η is the path loss exponent, while G_{Tx} and G_{Rx} are the transmitter and receiver antenna gains, respectively. The transmit power, $\mathrm{P}_{m,k}$, of the m_{th} device transmitting in the k_{th} time slot can then be calculated as:

$$\mathrm{P}_{m,k} = \omega_{m,k} + \mathrm{PL}_{m,k}. \tag{3}$$

In order that the message transmitted from the m^{th} device in the k^{th} slot is successfully decoded by the BS, the signal-to-interference-plus-noise ratio (SINR) at the BS, $\mathrm{SINR}_{m,k}$, has to be greater than a given threshold. In this work, we consider

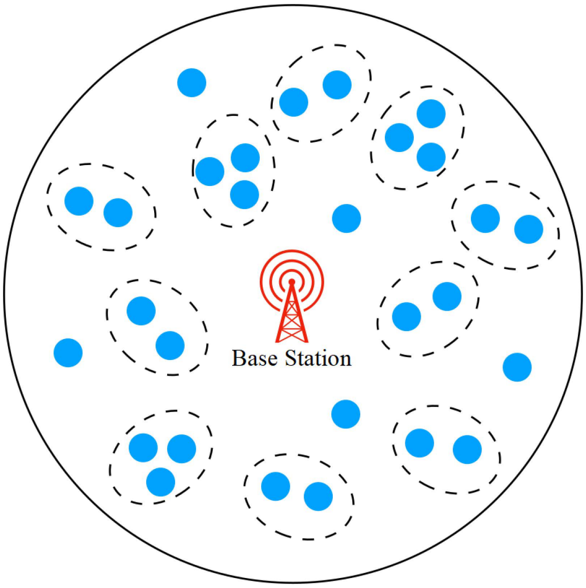

FIGURE 1. Clustered MTC network associated with a BS, N synchronous transmission devices in the uplink direction, M power levels, K time slots, and C clusters based on vicinity and low-power devices.

the Shannon capacity, so that the threshold for successful decoding is $(2^r - 1)$ where r is the spectral efficiency [33]. Consequently, the successful decoding probability from the m^{th} device in the k^{th} time slot can be expressed as follows

$$\mathcal{P}\left(\text{SINR}_{m,k} \geq 2^r - 1\right). \tag{4}$$

The BS proceeds with the successive decoding of the signals received in each time slot until they are all recovered. If that is not possible, a failure is declared. Moreover, if different clusters transmit during the same time slot, we assume that an unresolvable collision happens and a failure is also declared. Collisions, however, are not considered when calculating the information outage, as the outage assumes that the cluster is transmitting alone in a slot. This assumption allows us to design the power for differentiating devices within the same cluster, and not for surviving inter cluster collisions. Therefore, the probability that the m^{th} message is successfully decoded depends on the decoding in the presence of interferes with lower powers, but also on the previous decoding and removing of the messages with higher powers. Assuming the same target outage probability $\mathcal{O}_{\text{ref}}$ for all power levels, the information outage probability of the m^{th} message is

$$\mathcal{O}_{\text{m}} = 1 - (1 - \mathcal{O}_{\text{ref}})^{M-m}, \tag{5}$$

while the information outage probability of the last (or 0^{th}) message is the final SIC system information outage probability $\mathcal{O}_{\text{SIC}}$ considering all iterations

$$\mathcal{O}_{\text{SIC}} = \mathcal{O}_0 = 1 - (1 - \mathcal{O}_{\text{ref}})^M. \tag{6}$$

Note that the differences in average received power within the set Ω should be carefully designed to achieve the target outage probability $\mathcal{O}_{\text{ref}}$ for all SIC iterations.

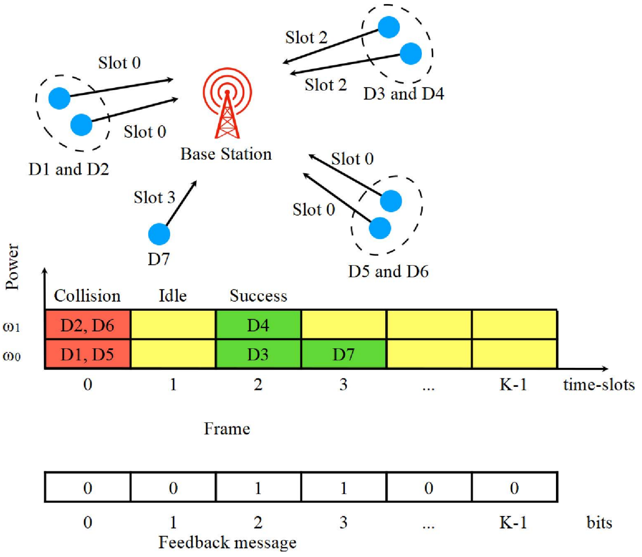

FIGURE 2. NOMA-based MTC network with random access and maximum number of devices per cluster $M = 2$.

After tentative decoding of all signals received at each time slot, the BS broadcasts a feedback message between data frames, indicating the successful decoding or not of the transmitted data. Fig. 2 shows that the feedback message is composed by K bits, each one corresponding to a time slot, so that a '1' at the k^{th} position indicates that all data packets transmitted in the k^{th} time slot were successfully decoded, and a '0' if they were not, either by message collision with different clusters, by fading, or because no device transmitted in that slot. Therefore, ACK bits indicate success or failure per time slot, not per message.

III. NOMA POWER ALLOCATION

Considering Rayleigh fading, and dropping the time-index k, the successful decoding probability of an MTC device, when it is transmitting alone (*i.e.*, free of interference), is

$$\mathcal{P}\left(\omega_0 |h_0|^2 \geq (2^r - 1)\sigma^2\right), \tag{7}$$

while, according to the Section II, this probability must be greater than or equal to $(1 - \mathcal{O}_{\text{ref}})$. Consequently, ω_0 becomes

$$\omega_0 \geq -\frac{(2^r - 1)}{\ln(1 - \mathcal{O}_{\text{ref}})}\sigma^2. \tag{8}$$

We assume that the M devices that belong to the same cluster and transmit in the same time slot are decoded in order, from highest to lowest received power, such that the strongest signal has average received power ω_{M-1}, while the weakest signal has average received power ω_0. This power allocation scheme allows us to find the NOMA partners by device proximity, *i.e.*, clustering, rather than the usual channel gain difference allocation. The choice for D2D partnering is crucial because nearby devices are able to share information and avoid the time consuming task of finding NOMA partners, greatly improving the learning process. Then, decoding all signals is possible because the BS applies SIC, reconstructing and removing the signals from the previously decoded messages. We assume that the SIC receiver can extract each signal power perfectly from the received signal when the decoding is successful. Next, we consider two decoding schemes, fixed- and dynamic-ordered SIC.

A. FIXED-ORDERED SIC

The BS determines the decoding order of this scheme by considering only the statistical CSI for each device, *i.e.*, the average received power. Then, the successful decoding probability of the m^{th} MTC device, when it is transmitting in presence of m interfering signals and given that the previous messages were successfully decoded, is[2]

$$\mathcal{P}\left(\omega_m|h_m|^2 \geq (2^r-1)\sum_{j=0}^{m-1}\omega_j|h_j|^2\right). \tag{9}$$

Therefore, considering the predefined operating maximum outage probability $\mathcal{O}_{\text{ref}}$ and following the analysis presented in [34, Eq. 32], we can establish that

$$\begin{aligned}(1-\mathcal{O}_{\text{ref}})^{-1} &= \frac{\prod_{j=0}^{m-1}\left(\frac{1}{\omega_j}+\frac{(2^r-1)}{\omega_m}\right)}{\prod_{j=0}^{m-1}\frac{1}{\omega_j}},\\ &= \prod_{j=0}^{m-1}\left(1+\frac{(2^r-1)\omega_j}{\omega_m}\right). \end{aligned}\tag{10}$$

From (10), we can estimate ω_m if the lower average received powers are known. For instance, if $M = 2$ then

$$\omega_1 \geq \frac{(2^r-1)(1-\mathcal{O}_{\text{ref}})}{\mathcal{O}_{\text{ref}}}\omega_0. \tag{11}$$

Since $\omega_m \gg \omega_{m-1}\ \forall m$, the following approximation[3] to estimate the m^{th} power level as a function of the $(m-1)^{\text{th}}$ power level is valid

$$\omega_m \gtrsim \frac{(2^r-1)(1-\mathcal{O}_{\text{ref}})}{\mathcal{O}_{\text{ref}}}\omega_{m-1}. \tag{12}$$

B. DYNAMIC-ORDERED SIC

The dynamic ordered scheme is more demanding since the BS must determine the decoding order based on the instantaneous received power of each device belonging to the cluster that operates in the current time slot. The complexity of this method increases with the number of clustered devices since it requires a more precise CSI and must resolve the decoding order in each time slot. Dynamic ordering presents a great advantage over fixed ordering for very small values of M. In contrast, such an advantage greatly diminishes for larger values, so that fixed-ordered SIC may be preferred in practice for large M. For this reason, next, we analyze in detail only the case of $M = 2$ for dynamic-ordered SIC. Note that by being able first to decode either of the two signals, *i.e.*, the one that is received at the BS with the highest power and not necessarily the one that was expected to arrive with the highest power, increases the successful decoding probability for the strongest signal, being the probability of the union of two events

$$\mathcal{P}\left(\frac{\omega_1|h_1|^2}{\omega_0|h_0|^2+\sigma^2}\geq(2^r-1)\bigcup\frac{\omega_0|h_0|^2}{\omega_1|h_1|^2+\sigma^2}\geq(2^r-1)\right). \tag{13}$$

Then, assuming an interference-limited scenario, and $r > 1$ [bps/Hz], from (11) and (13) we can obtain[4]

$$(1-\mathcal{O}_{\text{ref}}) = \frac{\omega_1/\omega_0}{\omega_1/\omega_0+(2^r-1)} + \frac{1}{1+(2^r-1)\cdot\omega_1/\omega_0}, \tag{14}$$

while after some algebraic transformations we have that

$$\omega_1 = \frac{2(2^r-1)\mathcal{O}_{\text{ref}}\omega_0}{(1-\mathcal{O}_{\text{ref}})(2^r-1)^2-\mathcal{D}(2^r-1)-(1+\mathcal{O}_{\text{ref}})}, \tag{15}$$

with

$$\mathcal{D} = \sqrt{\left((2^r-1)^2-1\right)\left((1-\mathcal{O}_{\text{ref}})^2-\frac{(1+\mathcal{O}_{\text{ref}})^2}{(2^r-1)^2}\right)}. \tag{16}$$

IV. PROPOSED METHOD

This work exploits short-range D2D communications to form device clusters together with full utilization of the BS feedback message to increase throughput and speed up convergence at the device side. Additionally, at the BS, the proposed method employs an adaptive algorithm, adjusting the frame size to the network load. The proposed method combines Q-Learning's ability to learn from the interaction with the environment to NOMA's spectral efficiency to optimize slot-allocation in a RA Aloha-like scheme. Clustering simplifies the problem of finding the optimal time slot, since up to M devices can transmit in the same time slot and the appropriate transmit power allocation is solved within each cluster. In addition, the full use of the BS feedback avoids inter-cluster collisions, speeding up convergence as clusters quickly settle down in the chosen time slots.

A. Q-LEARNING

Reinforcement learning is a family of machine learning algorithms where an agent interacts with its environment and learns from the feedback of those interactions, trying to maximize its reward [16]. Slot allocation, as with many wireless systems optimizations, can be formulated as a reinforcement learning problem [15]. Q-Learning, a well-known reinforcement learning algorithm, has been widely adopted in this context because it is model-free and can be implemented in a distributed manner [18], [20], [27]. Modeling RA in an

[2] From (8) note that in realistic scenarios ω_0 is much greater than σ^2 (*e.g.*, if $r = 2$ [bps/Hz] and $\mathcal{O}_{\text{ref}} = 10^{-2}$, then $\omega_0 = 298.5\cdot\sigma^2$), so we can neglect the contribution of σ^2 in the sum.

[3] From (11), $\omega_1 = 297\cdot\omega_0$, for $r = 2$ [bps/Hz] and $\mathcal{O}_{\text{ref}} = 10^{-2}$.

[4] For $r \leq 1$ [bps/Hz] the intersection of the events in (13) is not zero and has to be taken into account. For such low spectral efficiencies, or larger M, the adequate received power values in Ω can be obtained numerically. Moreover, for larger M the approximation in (12) works well.

MTC network as a Markov decision process (MDP) allows us to use Q-Learning. In an MDP, the agent interacts with the environment sequentially, selecting actions based on the state of the environment. The agent gets a reward based on its action and moves to the next state [16].

The Q-Learning algorithm considers the agent-environment relationship by an action-value function in the Q-table. An agent performs an action A_u, from a state S_u at each time step u, trying to maximize its reward associated with the action-value function. The Q-value update rule can be defined as [16]

$$Q(S_u, A_u) \leftarrow (1-\alpha)\, Q(S_u, A_u) + \alpha \left(R_{u+1} + \gamma \max_{a} Q(S_{u+1}, a) \right), \tag{17}$$

where $\alpha \in [0, 1]$ is the learning rate, R_{u+1} is the future reward, a is every possible action from a state, and $\gamma \in [0, 1]$ is the discount factor quantifying the importance of future rewards by multiplying the maximum Q-value available in the next time step ($\gamma = 0$ values only immediate rewards while a higher γ would aim at a better long-term reward).

We can apply the Q-Learning algorithm to our system model by considering that the agents are the cluster heads, and the environment is the network, and the state-action pair is the action of transmitting in a chosen time slot, with every cluster head having its own Q-Table. Therefore, a device has K states (*i.e,* equal to the number of time slots) with only one action for transmitting in each state, reducing the Q-Table to a $1 \times K$ vector. Hence, we can write the Q-Value for a state as $Q(k)$. The simplest way to implement the Q-Learning algorithm is to apply a greedy policy. In this way, the device always chooses the time slot with the highest Q-value. As clusters choose the best time slot for themselves, the network tends to converge, with every cluster having its own time slot. Moreover, the greedy policy also presented the best results during our simulation campaign when compared to ϵ-greedy policies. In this work, the reward value at the u^{th} time step is defined as:

$$R_u = \begin{cases} +1, & \text{successful slot} \\ -1, & \text{failed slot.} \end{cases} \tag{18}$$

B. FULL-FEEDBACK-BASED REWARD (fFbR) MECHANISM

To get the most out of the information available in the feedback message broadcast by the BS, each cluster head notified with failed transmission applies a negative reward not only to its own slot but also to every other slot that had a successful transmission, avoiding colliding with those that have already found a valid transmission slot. Hence, leading to the full exploitation of the feedback. Moreover, every cluster head notified of successful transmission will select the same time slot and refrain from updating its Q-Table, saving processing energy and simplifying the selection process. Among the works previously discussed, only in [27] the full exploitation of the feedback message was considered. However, it should be noted that this reward mechanism makes sense in the present proposal since the efficient exploitation of non-orthogonal resources by transmitting different power levels is resolved within each cluster. Otherwise, in non-clustered methods as [21], [22], this fFbR mechanism may prevent the selection of time slots with available power levels for the NOMA operation, becoming inefficient in the presence of multiple underutilized slots.

C. PROPOSED RA ALGORITHM

First, each device searches for its closest peers, *i.e.*, in clusters of up to M devices, according to the transmission range of the D2D technology being used and the device density of the cell. Although the devices may take turns in this position, we assume that the device with the largest signal-to-noise ratio (SNR) in each cluster (which is the closest to the BS) assumes the role of the cluster head. Thus, such a device is responsible for choosing the time slot, assigning power levels, and sharing this information with its partners. Note that clustering only happens at the beginning of the learning process. Then, every cluster head initializes its Q-Table following a uniformly random distribution, *i.e.,* every time slot is initially represented by a Q-Value $\in [-1, 1]$. This initialization, besides bringing an extra degree of randomness, differentiating clusters early on, can also be considered an optimistic initialization and motivates exploration [16]. The cluster heads then proceed to learn together, but in a distributed way. Each cluster head chooses the time slot with the highest Q-value and organizes itself with each device transmitting with a power that yields one of the M possible received powers at the BS.[5] Next, every device transmits its message, and the BS tries to recover them by using SIC decoding. At the end of the frame, the BS sends a feedback message with one bit per time slot, informing if the messages in that time slot were successfully decoded or not. Note that positive feedback is only given if all transmissions are successfully decoded at the determined time slot, which is made using only one bit per time slot. The cluster heads then update their Q-Tables following (17) and (18), employing the novel fFbR mechanism described above. This process repeats itself over several frames until it eventually converges.[6] The proposed RA method at the device side is summarized in Algorithm 1.

A simplified frame by frame example of the algorithm is depicted in Fig. 3, where we have three Q-Tables representing three clusters. At the first frame the Q-Values are randomly initialized. Even though each cluster has completely different

[5] The devices within a cluster can coordinate to use different powers from time to time, such that the long-term average power consumption among them becomes the same.

[6] The convergence of the Q-Learning algorithm is well known [16]. However, the convergence of multi-agent distributed Q-Learning in a competitive scenario needs further investigation. Nevertheless, [20] and [22], for example, consider convergence when the total value of Q-value stabilizes. We, on the other hand, consider convergence when there is no significant change to the throughput, as different reward systems lead to a different behavior of the total Q-value and throughput can be used as a metric across different methods.

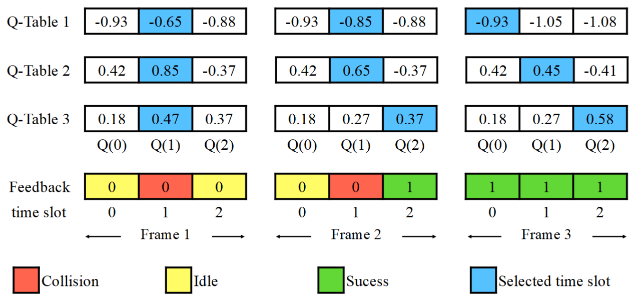

FIGURE 3. Frame by frame example with N=6, K=3 and M=2.

values, they all select time slot 1 for transmission which results in an unresolvable collision. Next, each device updates its Q-Values taking into consideration the feedback from the first frame by applying a penalty of -0.2. In the second frame, the third cluster selects time slot 2, while the first two clusters still have in time slot 1 the largest Q-Value and therefore select it for transmission. As a result, the third cluster has a successful transmission while the other two clusters collide. Finally, in the last frame of this example, clusters 1 and 2 have applied a penalty not only to the slot where they transmitted in but also to the value representing time slot 3 as a consequence of the fFbR mechanism. Thus, leading to $Q(1)$ and $Q(2)$ having Q-Values lower than -1, but keep in mind that even though the Q-Values are initialized with values between -1 and $+1$ they are not limited by these boundaries. Note that this prevents cluster 1 from selecting time slot 3 and then it selects the first time slot for transmission. Cluster 2, however, still remains at the second time slot. Thus, every cluster has successfully found its own time slot.

D. DYNAMIC FRAME SIZE ADAPTATION

Considering the fFbR mechanism and the RA algorithm above, we can conclude that the system reaches its maximum performance if and only if all clusters can be made up of M devices and the number of clusters coincides with frame size K, such that the number of devices in the system is $N = K \cdot M$. However, the numerical mismatch is not the only drawback that can prevent optimal performance, but it is also conditioned on the relative location of the nodes. The nodes distribution and the D2D communication allow all clusters to comprise M devices. Resolving such situations is beyond the scope of this paper, but two other issues can be addressed: *(i)* when the number of slots in the frame is less than the number of clusters, $K < C$, it is not possible to allocate resources to all clusters and a number of X collisions will happen; *(ii)* when $K > C$ some time slots remain idle unnecessarily, making the system temporarily inefficient. These drawbacks can be effectively solved through dynamic frame size adaptation. However, to prevent this adaptation from affecting the learning and convergence process, we propose that this adjustment be made every S frame. So, if not all clusters have found a valid time slot, then the BS detects X colliding slots and increases K by X more slots. However, suppose no collision slots are detected, and there are still unoccupied slots. In that case, $K - C$ idle slots must be removed and notified through a broadcast message of the new position of the slots that the cluster heads had previously associated. This will allow each device to send information more often, avoiding unnecessary delays. Finally, we summarize the adaptive frame size algorithm that runs only in the BS, in Algorithm 2.

Algorithm 1 NOMA-Based Distributed Q-Learning RA Method With D2D Clustering

```
Require: Devices try to find partners in the vicinity.
Require: $Q$-Table random initialized between $-1$ and 1
for Every frame do
    for Every cluster head do
        Select the time slot with the highest $Q$-value
        if More than one slot with the highest value then
            Choose randomly among them
        end if
        Transmit the chosen time slot and assigned power to its peers
    end for
    BS uses SIC to recover the transmitted messages
    BS broadcasts feedback message
    for Every cluster head do
        Update $Q$-value for using (17) and (18)
        if Transmission failed then
            for Every slot do
                if Broadcast message slot $= 1$ then
                    Update $Q$-value with (17) and $R_{u+1} = -1$
                end if
            end for
        end if
    end for
end for
```

Algorithm 2 Dynamic Frame Size Adaptation

```
for Every frame do
    if Frame mod $(S) = 0$ then
        if BS detects $X$ colliding slots then
            Add $X$ slots to the frame
        end if
        if BS does not detect colliding slots and there are free slots then
            Remove $K - C$ time slots.
            Reset learning.
        end if
    end if
end for
```

E. COMPLEXITY ANALYSIS

The complexity of reinforcement learning algorithms can be separated into three different categories. The sample complexity, the computational complexity, and the space complexity. The first two represent the number of samples and the

TABLE 4. Comparison among Sample, Computational and Space complexity [35], where $\tilde{O}(\cdot)$ represents the complexity order to attain the asymptotic convergence, and $\Theta(n)$ is the tight bound of the memory required to run the algorithm.

Complexity	Model-based	Model-free
Sample	$\tilde{O}\left(\frac{n\beta^4}{\epsilon^2}\right)$	$\tilde{O}\left(\frac{n\beta^5}{\epsilon^2}\right)$
Computational	$\tilde{O}\left(\frac{n\beta^5}{\epsilon^2}\right)$	$\tilde{O}\left(\frac{n\beta^5}{\epsilon^2}\right)$
Space	$\tilde{O}\left(\frac{n\beta^4}{\epsilon^2}\right)$	$\Theta(n)$

computational cost to reach a certain target performance (*i.e.,* achieving an ϵ-optimal action-value with high probability). The last one represents the amount of memory needed in order to run the algorithm. Note that both Algorithms 1 and 2 are, in essence, a pure model-free distributed Q-Learning with slight modifications. In this regard, the authors of [35] presented the complexities for model-free and model-based Q-Learning as shown in Table 4, where n represents the number of samples, or state-action pairs, that it takes for the algorithm to reach the ϵ-optimal solution, and β is given by $1/(1-\gamma)$. The complexities heavily rely on how many steps it takes for the algorithm to reach an optimal performance, while β evaluates how much future rewards are taken into consideration. As the importance of future rewards grow, so does the complexity, as agents need more samples to reach the optimal solution. Another important aspect to be taken into consideration is the fact that convergence for a competitive scenario in distributed Q-Learning has not been fully understood [20]. However, the fFbR mechanism allows for early convergence in just over 10 iterations greatly reducing the complexity of the algorithm. Then, devices are successfully allocated and they can keep using the learned slot. In addition, the proposed algorithm when using $\gamma = 0$ (*i.e.,* leading to $\beta = 1$ and reducing the complexity) only has a slight delay in convergence, while the method in [21] has a significant drop in performance.

F. PRACTICAL ASPECTS

We end this section with comments on some practical aspects of the techniques proposed above. Compared to the literature, the complexity added is the D2D communication needed to establish and maintain the clusters, with the cluster head sharing the chosen time slot with its partners. However, since the devices no longer have to learn their transmit power, the Q-Table is reduced to a vector of length K. At the BS, the complexity is non-negligible on SIC, but the complexity associated with the aggregated algorithm for dynamic frame size adaptation is trivial. Although it is expected that the BS would have more processing power than the devices, implementing the Q-Learning at the BS would be much more complex. The BS would be required to store and update the Q-Table for every device (or cluster), making it difficult to deploy new nodes.

The proposed method is perfectly capable of incorporating nodes not within clusters, as the learning is localized. However, a device transmitting alone ($M = 1$) will most likely use a time slot by itself. One possible way to work around this problem is by allowing devices that are not within a cluster to use the method from [21]. Thus, allowing devices that did not find a partner in the vicinity to share their slot with another device without a cluster. Another viable option can be achieved once all the clusters have their own transmission slot. In these circumstances, a well-designed protocol could allow the BS to group clusters with less than M devices into new clusters with size M.

One critical point in power domain NOMA transmission is channel estimation. In this work, we consider perfect channel knowledge at the BS, while in practice, it could be estimated through the use of orthogonal pilots sent by the devices [14]. In particular, we only need M orthogonal pilots, since that is the maximum number of superposed signals that the BS must decode per time slot. Additionally, we consider that each of the M orthogonal pilots is associated with one of the M different received power levels so that the pilot allocation is resolved at the same time that the power allocation is defined within the clusters, which is a significant practical advantage of the proposed method.

Finally, in terms of standardization, the 3rd Generation Partnership Project (3GPP) started to address D2D, or Proximity Service (ProSe), since Release 12 with relaying functionality being added in later releases. In [36] application of D2D to NB-IoT and LTE-M was further studied, however, it was not developed into a standard [37]. Nonetheless, non-3GPP radio access technologies are one of the main enablers of D2D. Besides the aforementioned BLE, the Wi-Fi Direct also allows for a direct link among devices [38].

V. RESULTS

We evaluate the performance of the proposed method by means of computer simulations, considering the system model from Section II with the parameters defined in Table 5 from typical values for IoT devices, unless stated otherwise. Dynamic SIC ordering is considered, but fixed SIC ordering with the appropriate power levels leads to the same results. The curves present the average of 30 simulation runs. The proposed method is compared to slotted Aloha and to [21]. By comparing to [21] we can, besides positioning the proposed scheme with respect to the literature,[7] incrementally investigate how each feature of the novel method (clustering and the fFbR mechanism) affects performance and the learning process. Note that in the following figures the method in [21] is called solely as [21], while [21] with clustering refers to the aforementioned method considering that devices use D2D communication for cluster formation. Finally, [21] with fFbR refers to the method in [21], but with the full-feedback reward mechanism proposed in this paper.

First, we look at the throughput, defined as the number of successful transmissions over the total number of time slots, thus measuring how well the frame is being exploited.

[7] As most of the related literature, we use the Slotted Aloha as a benchmark for the performance of the proposed method. Nonetheless, a qualitative comparison was presented in Table 1.

TABLE 5. Simulation parameters.

Parameter	Value
Bandwidth B	100 kHz
Carrier frequency	915 MHz
Cell radius	110 m
Clustering range $d_{cluster}^{max}$	15 m
Devices per cluster M	2
Discount factor γ	1
Frame size adaptation S	10
Transmitter gain G_{Tx}	2 dB
Receiver gain G_{Rx}	8 dB
Path loss exponent η	3
Noise Figure F	6 dB
Noise PSD N_0	-174 dBm/Hz
Maximum SIC Outage Probability $\mathcal{O}_{\text{SIC}}$	10^{-2}
Spectral efficiency r	2 bps/Hz
Reference distance d_0	1 m
Devices N	100-450
Messages L	100 packets
Operating Maximum Outage Probability $\mathcal{O}_{\text{ref}}$	0.005
Simulation runs	30
Number of time slots K	100
Learning rate α	0.2

Moreover, we start by considering clusters with only two devices ($M = 2$) as it is not much likely that NOMA with several layers is practical due to channel estimation and SIC imperfections, assuming all N are low-power devices. Fig. 4 shows that the proposed method can outperform every other simulated scheme, improving the throughput over [21] by 18.55% at 200 devices and by 240.28% at 250 devices, for $K = 100$ time slots. Note that the addition of clustering to [21] significantly improves throughput as devices no longer have to learn their transmission power and already have a defined partner. Another interesting behavior is that when the new fFbR mechanism is employed, the proposed method and [21] present a slow drop in throughput as the number of devices gets larger than $2 \times K$. The method in [21], on the other hand, exhibits a sharp drop after $2 \times K$. This can be attributed to the fact that the devices in [21] do not learn to avoid successful slots. Thus, in [21], when $N > M \cdot K$ devices scatter across the frame resulting in more collisions and therefore a lower throughput.

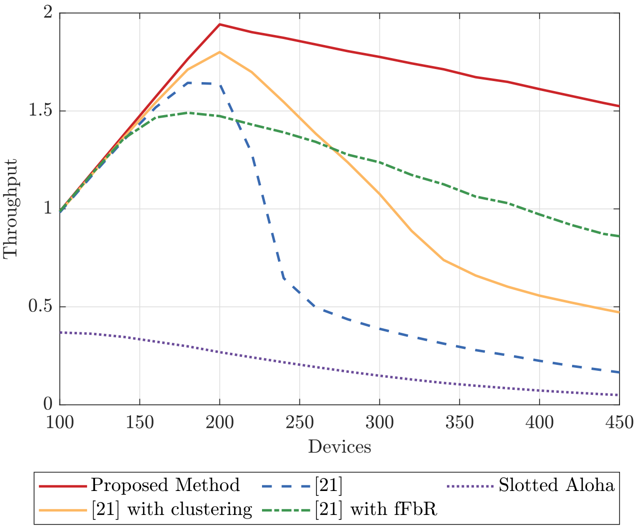

FIGURE 4. Throughput versus number of Devices and $K = 100$ time slots. Unless for Slotted Aloha, $M = 2$ devices per cluster.

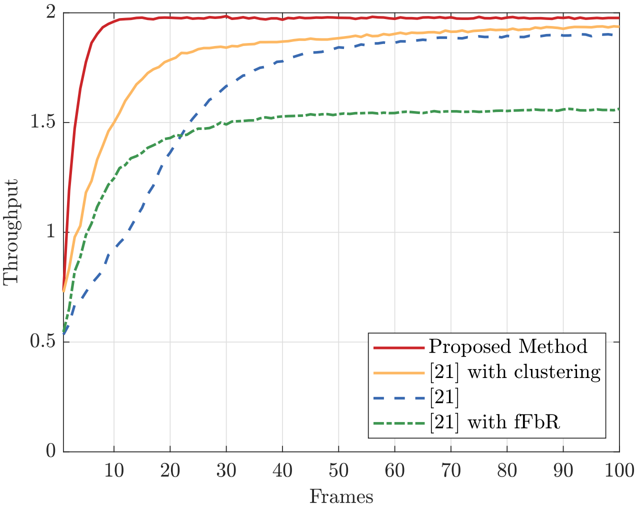

FIGURE 5. Convergence analysis for $N = 200$ devices, $K = 100$ time slots and $M = 2$ devices per cluster.

Next, we investigate the convergence in Fig. 5. We can see that the addition of clustering has a strong impact on the convergence speed. For example, [21] with clustering can reach a 1.8 throughput in about 22 frames, while it takes 44 frames for the method in [21] to cross the same threshold. Alternatively, with 10 frames, clustering enables the method in [21] to reach a 1.50 throughput while the method in [21] is still at 0.88. Another interesting takeaway from Fig. 5 is the effect of the fFbR mechanism. On one hand, the new reward scheme improves the early convergence in relation to [21]. On the other hand, it holds the maximum throughput at 1.56, with the common reward system having a better performance from 20 frames onward. This can be attributed to the fact that the new reward system penalizes the Q-values for slots that had a successful transmission, thus the devices learn to avoid slots already in use. However, the reward system cannot differentiate how many devices are successfully accessing the slot. Thus, devices end up avoiding slots that might be shared. Note that the method proposed in this work, which consists of employing both strategies (clustering and fFbR mechanism) jointly, is able to reach 1.8 throughput in just 6 frames and at 10 frames the throughput is already approximately 2, which is the ceiling for this particular network scenario. Thus, clustering and fFbR mechanisms, combined with NOMA, drastically improve the convergence speed.

To illustrate the effect of the proposed adaptive frame size algorithm. We assume $N = 300$ devices, $M = 2$ received power levels, and $K = 100$ time slots, *i.e.*, an overloaded scenario in which $\frac{N}{M \cdot K} > 1$. Moreover, $S = 10$, so that most devices have already settled in a given time slot, and, following Algorithm 2, the frame size is increased or decreased X time slots at a time. In Fig. 6, we can see that as the frame is adapted, the throughput increases getting closer to M. Note that the sharp drops in throughput, *e.g.*, around 10,

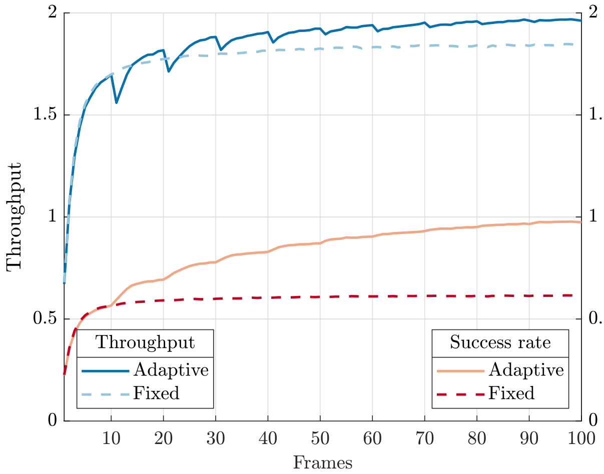

FIGURE 6. **Throughput and success rate with frame size adaptation when $N = 300$ devices, $M = 2$ devices per cluster, and $S = 10$ time slots.**

20, and 30 frames and so on, do not represent a genuine loss of messages, as the frame size increases at those points by X time slots and the amount of successful transmissions remains the same. We can better understand this by looking at the device success rate, defined as the number of successful transmissions over the total number of transmissions. Note that, while the throughput shows a little improvement as the frame size is adapted, the success rate is almost twice as high as the value it would converge to without the frame size adaptation. The success rate is a better metric for analyzing this frame adaptation, as slots can aggregate several collisions, which is not noticeable while looking only at the throughput. When the frame-size change happens, the clusters reorganize themselves within the frame, leading to a growth in the success rate. The adaptive frame size algorithm increases the throughput ceiling of the method, allowing devices to find new suitable, non-collided, slots.

Next, we analyze the average slot allocation in the 100th and last frame for the method proposed here and those used as benchmarks. We consider the average number of idle slots, the maximum number of devices in a slot, and slot with collisions. Moreover, we also include the percentage of failed transmissions and the throughput. We can see in Table 6 that the proposed method outperforms the others on every metric and, on average, does not have idle slots as it is capable of perfectly allocating every cluster to a time slot, taking full advantage of the frame size and NOMA. This allows devices to reach near full network capacity as the percentage of failed transmissions approaches the designed $\mathcal{O}_{\text{SIC}}$. It is interesting to note that the method in [21], when using the novel fFbR mechanism, is not able to discern between successful slots with one or more devices. This can be understood as we have several slots allocated to just one device, an average of 31.66 idle slots, and a few slots accumulating multiple failed transmissions for an average of 5.03 slots with 21.38 % of transmissions failed. Not only that, we can see that the new fFbR mechanism can cause slots to hoard failures in order to maintain a relatively high throughput. For example, there are over 11 devices allocated to just one time slot when the optimum allocation is $M = 2$ devices per slot, while in other time slots only one device remains operating and NOMA is not exploited. Therefore, the novel fFbR mechanism is to be used together with clustering, as proposed in this work.

TABLE 6. **Average slot allocation analysis for $N = 200$ devices and $K = 100$ time slots in the 100th frame.**

Method	IDS [a]	MIS[b]	SWC [c]	FTX % [d]	TPT[e]
[21]	2.40	3.00	2.40	4.35	1.91
[21] with fFbR	31.66	11.73	5.03	21.38	1.57
This work	0.00	2.53	0.26	1.05	1.98

[a] IDS: Idle slots, [b] MIS: Maximum number of devices in a slot, [c] SWC: Slots with collisions, [d] FTX: Failed transmissions, [e] TPT: Throughput.

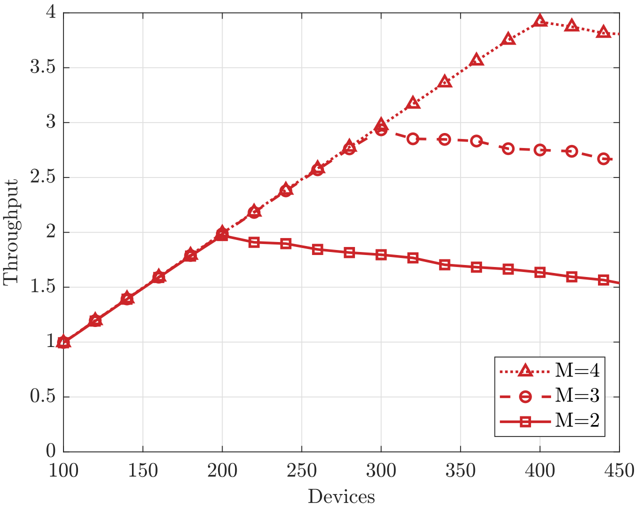

FIGURE 7. **Throughput versus number of devices, for different cluster sizes $M \in \{2, 3, 4\}$ and $K = 100$ time slots.**

Finally, we look at the throughput when more power levels are used, $M = 3$ and $M = 4$ received power levels. In Fig. 7, when we increase the number of devices per cluster, the maximum throughput also rises. The throughput reaches its peak when $N = M \cdot K$. However, note that the methods with a higher M would need higher ω's in order to become robust against fading and allow for SIC decoding. This, however, could demand prohibitively high transmit powers from the devices. Moreover, it is important to note that for larger M both channel estimation and SIC decoding become more prone to errors, even using orthogonal pilots by devices with different received power levels, and therefore in practice it is more likely that M would be small.

VI. CONCLUSION

We proposed a new Q-Learning RA method for NOMA-based MTC networks, considering: *(*i) short-range clustering, *(*ii) a full-feedback-based reward mechanism, and *(*iii) an adaptive frame structure. Clustering allows the partner selection and power allocation processes to be resolved in a distributed way and within each cluster. Thus, only the cluster heads are engaged in the distributed learning algorithm, speeding up convergence. The new reward mechanism makes the network reach its maximum performance more quickly, *e.g.*, maximum throughput for 200 devices in about 15 iterations. By fully exploiting the feedback message, devices avoid collisions with clusters that have already found their slots. Finally, the dynamic frame size adaptation algorithm allows increasing the number of slots to ensure that all clusters have their own transmission slot in overloaded situations. In contrast, the adaptation eliminates unnecessary slots in underloaded situations to favor more frequent communication.

The proposed method can be further investigated and improved by considering and analyzing different traffic models. Another possibility is to investigate the coexistence of devices with different requirements in terms of target outage probability, spectral efficiency, among other factors. Finally, this work could also be extended to address the possibility of a device requesting multiple time slots or deferring clusters to another frame as way of dealing with overload situations.

REFERENCES

[1] L. Chettri and R. Bera, "A comprehensive survey on Internet of Things (IoT) toward 5G wireless systems," *IEEE Internet Things J.*, vol. 7, no. 1, pp. 16–32, Jan. 2020.

[2] M. El-Tanab and W. Hamouda, "An overview of uplink access techniques in machine-type communications," *IEEE Netw.*, vol. 35, no. 3, pp. 246–251, May 2021.

[3] M. Vaezi, A. Azari, S. R. Khosravirad, M. Shirvanimoghaddam, M. M. Azari, D. Chasaki, and P. Popovski, "Cellular, wide-area, and non-terrestrial IoT: A survey on 5G advances and the road towards 6G," *IEEE Commun. Surveys Tuts.*, early access, Feb. 11, 2022, doi: 10.1109/COMST.2022.3151028.

[4] *Cisco Annual Internet Report 2018–2023*, CISCO, San Jose, CA, USA, Mar. 2020.

[5] N. H. Mahmood, O. López, and O. Park, *White Paper Critical Massive Machine Type Communication Towards 6G*. Oulu, Finland: Univ. Oulu, 2020. [Online]. Available: http://urn.fi/urn:isbn:9789526226781

[6] O. L. A. Lopez, H. Alves, and M. Latva-Aho, "Distributed rate control in downlink NOMA networks with reliability constraints," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5410–5423, Nov. 2019.

[7] O. L. A. López, "Massive wireless energy transfer: Enabling sustainable IoT toward 6G era," *IEEE Internet Things J.*, vol. 8, no. 11, pp. 8816–8835, Jun. 2021.

[8] J. M. D. S. Sant'Ana, A. Hoeller, R. D. Souza, S. Montejo-Sanchez, H. Alves, and M. D. Noronha-Neto, "Hybrid coded replication in Lora networks," *IEEE Trans. Ind. Informat.*, vol. 16, no. 8, pp. 5577–5585, Aug. 2020.

[9] O. L. A. Lopez, S. Montejo-Sanchez, R. D. Souza, C. B. Papadias, and H. Alves, "On CSI-free multiantenna schemes for massive RF wireless energy transfer," *IEEE Internet Things J.*, vol. 8, no. 1, pp. 278–296, Jan. 2021.

[10] S. Montejo-Sanchez, C. A. Azurdia-Meza, R. D. Souza, E. M. G. Fernandez, I. Soto, and A. Hoeller, "Coded redundant message transmission schemes for low-power wide area IoT applications," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 584–587, Apr. 2019.

[11] F. Clazzer, A. Munari, G. Liva, F. Lazaro, C. Stefanovic, and P. Popovski, "From 5G to 6G: Has the time for modern random access come?" 2019, *arXiv:1903.03063*.

[12] A. Laya, L. Alonso, and J. Alonso-Zarate, "Is the random access channel of LTE and LTE-A suitable for M2M communications? A survey of alternatives," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 4–16, 2nd Quart., 2014.

[13] Y. Ma, Z. Yuan, W. Li, and Z. Li, "Novel solutions to NOMA-based modern random access for 6G-enabled IoT," *IEEE Internet Things J.*, vol. 8, no. 20, pp. 15382–15395, Oct. 2021.

[14] Y. Yuan, S. Wang, Y. Wu, H. Vincent Poor, Z. Ding, X. You, and L. Hanzo, "NOMA for next-generation massive IoT: Performance potential and technology directions," 2021, *arXiv:2104.04911*.

[15] S. Ali, W. Saad, N. Rajatheva, and K. Chang, *6G White Paper on Machine Learning in Wireless Communication Networks*. Oulu, Finland: Univ. Oulu, 2020. [Online]. Available: http://urn.fi/urn:isbn:9789526226736

[16] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[17] S. K. Sharma and X. Wang, "Toward massive machine type communications in ultra-dense cellular IoT networks: Current issues and machine learning-assisted solutions," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 1, pp. 426–471, 1st Quart., 2020.

[18] J. Zhang, X. Tao, H. Wu, N. Zhang, and X. Zhang, "Deep reinforcement learning for throughput improvement of the uplink grant-free NOMA system," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6369–6379, Jul. 2020.

[19] L. M. Bello, P. D. Mitchell, and D. Grace, "Intelligent RACH access techniques to support M2M traffic in cellular networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8905–8918, Sep. 2018.

[20] S. K. Sharma and X. Wang, "Collaborative distributed Q-learning for RACH congestion minimization in cellular IoT networks," *IEEE Commun. Lett.*, vol. 23, no. 4, pp. 600–603, Apr. 2019.

[21] M. V. da Silva, R. D. Souza, H. Alves, and T. Abrao, "A NOMA-based Q-learning random access method for machine type communications," *IEEE Wireless Commun. Lett.*, vol. 9, no. 10, pp. 1720–1724, Oct. 2020.

[22] D.-D. Tran, S. K. Sharma, and S. Chatzinotas, "BLER-based adaptive Q-learning for efficient random access in NOMA-based mMTC networks," in *Proc. IEEE 93rd Veh. Technol. Conf.*, Apr. 2021, pp. 1–5.

[23] E. Mete and T. Girici, "Q-learning based scheduling with successive interference cancellation," *IEEE Access*, vol. 8, pp. 172034–172042, 2020.

[24] Z. Shi, W. Gao, J. Liu, N. Kato, and Y. Zhang, "Distributed Q-learning-assisted grant-free NORA for massive machine-type communications," in *Proc. Global Commun. Conf.*, 2020, pp. 1–5.

[25] J. Liu, Z. Shi, S. Zhang, and N. Kato, "Distributed Q-learning aided uplink grant-free NOMA for massive machine-type communications," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 2029–2041, Jul. 2021.

[26] J. Su, G. Ren, Q. Wang, and B. Zhao, "An SCMA-based decoupled distributed Q-Learning random access scheme for machine-type communication," *IEEE Wireless Commun. Lett.*, vol. 10, no. 8, pp. 1737–1741, Aug. 2021.

[27] Y. Zhou, F. Zhou, Y. Wu, R. Q. Hu, and Y. Wang, "Subcarrier assignment schemes based on Q-Learning in wideband cognitive radio networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 1168–1172, Jan. 2020.

[28] S. Ali, N. Rajatheva, and W. Saad, "Fast uplink grant for machine type communications: Challenges and opportunities," *IEEE Commun. Mag.*, vol. 57, no. 3, pp. 97–103, Mar. 2019.

[29] K. T. Le. (2017). *Bluetooth Low Energy and the Automotive Transformation*. [Online]. Available: https://www.ti.com/lit/wp/sway008/sway008.pdf

[30] M. Woolley. (2017). *Bluetooth Mesh Networking: Paving the Way for Smart Lighting*. [Online]. Available: https://www.bluetooth.com/bluetooth-resources/bluetooth-mesh-paving-the-way-for-smart-lighting/

[31] K. Mikhaylov, "Energy efficiency of multi-radio massive machine-type communication (MR-MMTC): Applications, challenges, and solutions," *IEEE Commun. Mag.*, vol. 57, no. 6, pp. 100–106, Jun. 2019.

[32] C. A. Balanis, *Antenna Theory: Analysis and Design*. Hoboken, NJ, USA: Wiley, 2005.

[33] A. Goldsmith, *Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[34] J. Wang, B. Xia, K. Xiao, Y. Gao, and S. Ma, "Outage performance analysis for wireless non-orthogonal multiple access systems," *IEEE Access*, vol. 6, pp. 3611–3618, 2018.

[35] M. Ghavamzadeh, H. Kappen, M. Azar, and R. Munos, "Speedy Q-learning," in *Advances in Neural Information Processing Systems*, vol. 24, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2011. [Online]. Available: https://proceedings.neurips.cc/paper/2011/file/ab1a4d0dd4d48a2ba1077c4494791306-Paper.pdf

[36] *Study on Further Enhancements to LTE Device to Device (D2D), UE to Network Relays for Internet of Things (IoT) and Wearables*, document TR 36.746 V15.1.1, 3GPP, Apr. 2018. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3110

[37] K. M. Malarski, K. D. Ballal, and S. Ruepp, "D2D-enabled failure-tolerance in cellular IoT," in *Proc. 12th Int. Conf. Netw. Future (NoF)*, Oct. 2021, pp. 1–5.

[38] K. M. Malarski, F. Moradi, K. D. Ballal, L. Dittmann, and S. Ruepp, "Internet of reliable things: Toward D2D-enabled NB-IoT," in *Proc. 5th Int. Conf. Fog Mobile Edge Comput. (FMEC)*, Apr. 2020, pp. 196–201.

MATHEUS V. DA SILVA received the B.Sc. degree in electronic engineering and the M.Sc. degree in electrical engineering from the Federal University of Santa Catarina (UFSC), Brazil, in 2019 and 2021, respectively. He is currently pursuing the Ph.D. degree with the Centre for Wireless Communications (CWC), University of Oulu, Oulu, Finland, where he is also a Researcher. His research interest includes wireless communications, with emphasis on the application of machine learning techniques to wireless networks.

SAMUEL MONTEJO-SÁNCHEZ (Senior Member, IEEE) was born in Camagüey, Cuba, in 1979. He received the B.Sc., M.Sc., and D.Sc. degrees in telecommunications from the Central University of Las Villas (UCLV), Cuba, in 2003, 2007, and 2013, respectively. From September 2003 to May 2017, he was an Associate Professor with the Department of Telecommunications, UCLV. In 2017, he held a postdoctoral position at the University of Chile. Since 2018, he has been with the Programa Institucional de Fomento a la I+D+i (PIDi) of the Universidad Tecnológica Metropolitana (UTEM), Santiago, Chile. He leads the FONDECYT Iniciación (Toward high performance wireless connectivity for IoT and beyond-5G networks) Project and is a member of the FONDECYT Regular (IoT goes to Space: Wireless Networking protocols and architectures for IoT networks served by LEO satellite constellations) and FONDEQUIP EQM180180 (Clúster Supermicro para Cómputo Científico) projects. He was a co-recipient of the 2016 Research Award from the Cuban Academy of Sciences.

RICHARD DEMO SOUZA (Senior Member, IEEE) received the D.Sc. degree in electrical engineering from the Federal University of Santa Catarina (UFSC), Brazil, in 2003. From 2004 to 2016, he was with the Federal University of Technology–Paraná (UTFPR), Brazil. Since 2017, he has been with UFSC, where he is currently a Professor. His research interests include the areas of wireless communications and signal processing. He was a co-recipient of the 2014 IEEE/IFIP Wireless Days Conference Best Paper Award and the 2016 Research Award from the Cuban Academy of Sciences and the supervisor of the awarded Best Ph.D. Thesis in Electrical Engineering in Brazil in 2014. He has served as an Editor or Associate Editor for the *Journal of Communication and Information Systems* (SBrT), the IEEE COMMUNICATIONS LETTERS, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, the IEEE TRANSACTIONS ON COMMUNICATIONS, and the IEEE INTERNET OF THINGS JOURNAL.

HIRLEY ALVES (Member, IEEE) is currently an Assistant Professor and the Head of the Machine-Type Wireless Communications Group at the 6G Flagship Centre for Wireless Communications, University of Oulu. He is actively working on massive connectivity and ultra-reliable low-latency communications for future wireless networks, 5G and 6G, full-duplex communications, and physical-layer security. He leads the URLLC activities for the 6G Flagship Program. He has received several awards and has been an organizer, chair, TPC member, and tutorial lecturer for several renowned international conferences. He was/is the General Chair of ISWCS 2019, the General Co-Chair of the 1st 6G Summit, Levi 2019, and ISWCS 2021, and the Track Chair of PIMRC 2021.

TAUFIK ABRÃO (Senior Member, IEEE) received the B.S., M.Sc., and Ph.D. degrees in electrical engineering from the Polytechnic School, University of São Paulo, São Paulo, Brazil, in 1992, 1996, and 2001, respectively. Since March 1997, he has been with the Communications Group, Department of Electrical Engineering, Londrina State University, Paraná, Brazil, where he is currently an Associate Professor of telecommunications and the Head of Telecommunication and Signal Processing Laboratory. In 2018, he was with the Connectivity Section, Aalborg University, as a Guest Researcher. In 2012, he was an Academic Visitor with the Southampton Wireless Research Group, University of Southampton, U.K. His current research interests include massive MIMO, XL-MIMO, RIS, URLLC, mMTC, optimization methods, machine learning, resource allocation, and random access protocols. He has served as an Associate Editor for the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, the IEEE SYSTEMS JOURNAL, IEEE ACCESS, IEEE COMMUNICATION SURVEYS & TUTORIALS, AEUe-Elsevier, the *IET Signal Processing*, and JCIS-SBrT, and the Executive Editor of the ETT-Wiley (2016–2021).