This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LWC.2020.3002691, IEEE Wireless Communications Letters

A NOMA-based Q-Learning Random Access Method for Machine Type Communications

Matheus Valente da Silva, Richard Demo Souza, *Senior Member, IEEE,*
Hirley Alves, *Member, IEEE* and Taufik Abrão, *Senior Member, IEEE,*

***Abstract*—Machine Type Communications (MTC) is a main use case of 5G and beyond wireless networks. Moreover, due to the ultra-dense nature of massive MTC networks, Random Access (RA) optimization is very challenging. A promising solution is to use machine learning methods, such as reinforcement learning, to efficiently accommodate the MTC devices in RA slots. In this sense, we propose a distributed method based on Non-Orthogonal Multiple Access (NOMA) and Q-Learning to dynamically allocate RA slots to MTC devices. Numerical results show that the proposed method can significantly improve the network throughput when compared to recent work.**

***Index Terms*—Internet-of-Things, MTC, NOMA, Q-Learning**

I. Introduction

The deployment of 5G and beyond mobile networks, including the Internet of Things (IoT), is driving the development of advanced Machine-Type-Communications (MTC) networks [1]–[3]. These networks should be able to support new applications with a massive number of devices, such as those in smart cities, Industry 4.0, etc. According to Cisco, the number of MTC devices can be as large as 3.9 billions by 2022 [4]. Many challenges arise with massive MTC (mMTC) networks, such as meeting diverse performance requirements and congestion in Radio Access Network (RAN). Moreover, mMTC networks suffer from inefficient Random Access (RA) procedures and resource allocation. Current RA protocols perform poorly in ultra-dense networks [5], leading to the need of efficient RA schemes able to handle massive requests.

In terms of standardization, 3GPP is evolving 5G to improve mobility, while aggregating physical downlink control channel (PDCCH) enhancements, and addressing new MTC use cases. Enhancements in 3GPP Rel-16 and Rel-17 [6], [7] include: **a**) 2-step Random Access Channel (RACH) to reduce latency and signaling overhead; **b**) reliability improvements; **c**) power saving techniques; **d**) enhanced support for new use cases, including industrial IoT. However, there is still plenty of room for improvement both in terms of performance and efficiency.

The authors in [8] present a comprehensive survey of the issues related with RAN congestion while introducing machine learning algorithms to improve RA for mMTC networks. Among other machine learning techniques, the reinforcement learning method known as Q-Learning stands out due to its capability of being implemented in a model-free and distributed manner [9]. The authors of [10] propose a Q-Learning based method to address the RAN congestion using the number of collisions per slot as a reward.

However, it needs substantial feedback from the base station (BS), besides the complexity of determining the number of colliding devices. Another example is [11], which attempts to conciliate the traffic load of Human-Type-Communication (HTC) with MTC devices in the RA of a cellular network, making MTC devices learn which slot to access, reducing collisions and improving throughput.

The work in [12] utilizes Q-Learning at the BS to better adapt the barring factor in an Access Class Barring (ACB) scheme. Although this method reduces the load in the network, it is a reactive solution, while the current ever increasing number of devices calls for a proactive solution. Another work that uses adaptive ACB is [13]. They propose two algorithms to minimize delay or maximize throughput. However, one needs to know the number of devices in the network and the other requires the BS to know how many devices collided in a past slot. In [14], Q-Learning is used to select the best available BS in a Long-Term Evolution (LTE) network, using throughput and delay both as QoS measurement and as the reward for the MTC devices. Whilst the MTC devices do select the best BS in this scheme, thus efficiently organizing RA within a cell, it does not deal with the growth in density of mMTC networks and therefore overload is still a problem. In [15], Non-Orthogonal Multiple Access (NOMA) [16] and Q-Learning are utilized in order to maximize energy efficiency in short packet communications. The method in [15] makes use of Q-Learning for pairing devices in sub channels. In order to maximize energy efficiency, they propose a power allocation scheme, finding the optimal transmit power for each device.

We propose the use of Q-Learning and NOMA, alongside with a power control scheme, to improve the throughput in mMTC networks. Our work differs from [8], [10], [11] because, besides using Q-Learning for slot allocation, we implement NOMA and consider the effect of path loss and fading. Unlike [10], our method requires minimal feedback from the BS, a single bit per time slot, instead of the number of contending devices per time slot [10]. Moreover, even though [12], [14] use Q-Learning in a MTC network, neither use it to improve slot allocation. The first adapts a class barring factor while the second selects the best BS for connection.

M. V. Silva and R. D. Souza are with the Department of Electrical and Electronics Engineering of the Federal University of Santa Catarina, Brazil. mvalente.silva@gmail.com, richard.demo@ufsc.br

H. Alves is with the Centre for Wireless Communications of the University of Oulu, Finland. hirley.alves@oulu.fi

T. Abrão is with the Department of Electrical Engineering, University of Londrina, Brazil, taufik@uel.br

This work has been supported in Brazil by CNPq, project PrInt CAPES-UFSC "Automation 4.0"; in Finland by Academy of Finland (Aka) 6Genesis Flagship (Gr. 318927), EE-IoT (Gr. 319008), and FIREMAN (Gr. 326301).

In [13] machine learning techniques are not used. Finally, compared to [15], we are looking to improve throughput rather than energy efficiency, although our power control scheme prevents the excessive use of power.

The main contributions of our work are: i) evaluation of the beneficial impact in RA of combining Q-Learning with NOMA; ii) a RA scheme which improves the network throughput with limited transmit power and complexity.

II. System Model

Assuming a single communication system in which all devices run the same application, we consider N synchronized devices distributed in a circular cell around a Base Station (BS). All devices transmit at the same power P_t, frequency f_c, and rate, each having L data packets ready for transmission. Medium access is based on grant free Slotted Aloha (SA), where each device transmits in one of K time-slots within a frame. There is no restriction on the quantity of devices per time-slot. After each frame the BS sends a group feedback using one bit per time-slot, informing if the transmissions were successful or not. This control message is also used to synchronize the devices. As usual, we assume that the BS acquires channel state information by means of pilots within a header contained in each transmission from the devices. Moreover, assuming a quasi-static scenario, the devices estimate the statistics of their channels using the common control message and apply channel inversion to reach a reference average power at the BS that assures a given outage probability.

The message from the m-th device, $m \in \{1, 2, \cdots, M\}$, $M \leq N$, transmitting in the k-th time slot, $k \in \{1, 2, \cdots, K\}$, is considered to be successfully decoded if the Signal to Interference plus Noise Ratio (SINR) at the BS is larger than the threshold from Shannon's capacity, so that [17]

$$\mathrm{SINR}_{m,k} \geq 2^r - 1 \tag{1}$$

where $\mathrm{SINR}_{m,k}$ is the SINR for the m-th device transmitting in the k-th time slot, and r is the spectral efficiency in bits/s/Hz. As we consider a quasi-static non-line-of-sight scenario, the asymptotic outage probability is a meaningful performance metric even in the finite blocklength regime [18].

A. NOMA

We consider the use of NOMA in the uplink, with SIC at the BS to decode colliding packets [16]. The signal received by the BS in the k-th time-slot, in the t-th frame, is

$$y_k(t) = \sum_{m=1}^{M} x_{m,k}(t) + n_k(t) \tag{2}$$

where $x_{m,k}(t)$ is the attenuated signal received at the BS from the m-th device in time-slot k, with instantaneous power $P_{m,k}$, while $n_k(t)$ is the additive white Gaussian noise.

The BS then performs SIC on the overall received signal y_k, starting from the strongest to the weakest user. Without loss of generality we assume that the users are ordered in decreasing received power from $m = 1$ to $m = M$. Then, the SINR for the m-th device after SIC becomes

$$\mathrm{SINR}_{m,k} = \frac{P_{m,k}}{\sum_{j=m+1}^{M} P_{j,k} + \overline{P}_n}, \tag{3}$$

where $P_{m,k} = h_{m,k}^2 \overline{P}_{m,k}$ is the instantaneous received power from the m-th device in the k-th time slot, $h_{m,k}$ is Rayleigh fading, which is independent and identically distributed in time and space, while $\overline{P}_{m,k}$ is the average received power, which is modelled by considering the log-distance path loss model [17],

$$\overline{P}_{m,k} = \overline{P}_{m,k}(d_0) - 10\eta \log_{10}\left(\frac{d_{m,k}}{d_0}\right), \quad [dB] \tag{4}$$

where $d_{m,k}$ is the distance from that device to the BS, d_0 is the reference distance, $\overline{P}_{m,k}(d_0)$ is calculated using the Friis equation, while η is the path loss exponent. Finally, $\overline{P}_n = FN_0B$ denotes the noise power, where N_0 is the noise power spectral density, B is the bandwidth, and F is the noise figure.

It is important to mention that [10] considers a hard collision model, in which the transmissions fail if a collision happens in a given time-slot, whatever the SINR is. The hard model limits the performance since, in many cases, it is possible to decode the strongest user in a collision [19]. Moreover, the introduction of the NOMA strategy above allows us to potentially decode all the colliding users, greatly impacting the overall throughput. As mentioned previously, the devices can apply channel inversion to reach a certain average power at the BS. However, NOMA does not work well if devices yield the same power at the BS. In order to add the needed power diversity for NOMA to work properly, we let the devices deviate $\pm\Delta$ from a reference power, which in turn is calculated so that $P_{\mathrm{ref}} - \Delta$ reaches a target outage probability. Thus, devices have three options of transmit power ($P_{ref} - \Delta$, P_{ref}, $P_{\mathrm{ref}} + \Delta$). An appropriate Δ can increase NOMA efficiency as we no longer rely on the device's position to create the diversity for NOMA paring.

III. Proposed Method

This work proposes a Q-Learning based method to optimize slot allocation taking advantage of NOMA spectral efficiency, making it possible for two devices or more to transmit at the same time-slot. This method allows the MTC devices to autonomously find NOMA pairs and their dedicated time-slot while also preventing the excessive use of transmit power.

A. Q-Learning

The use of reinforcement learning has great potential in MTC networks [8], [10]–[12], [14], [15], specially the widely adopted Q-Learning algorithm, because it is model-free and can be implemented in a distributed fashion. By modeling the RA in an MTC network as a Markov Decision Process (MDP) allows us to use Q-Learning. In an MDP the agent interacts with the environment in a sequential manner, selecting actions based on the state of the environment. The agent gets a reward based on its action and moves to the next state [9].

The Q-Learning algorithm formulates this agent-environment relationship with an action-value function,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LWC.2020.3002691, IEEE Wireless Communications Letters

the Q-table. At each time step u, while in a state S_u, an agent performs an action A_u trying to maximize its action-value function. The Q-value update rule can be defined as [9]

$$Q(S_u, A_u) \leftarrow (1-\alpha)\, Q(S_u, A_u) + \alpha \left(R_{u+1} + \gamma \max_a Q(S_{u+1}, a) \right) \tag{5}$$

where $\alpha \in [0,1]$ is the learning rate, $\gamma \in [0,1]$ is the discount factor quantifying the importance of future rewards ($\gamma = 0$ values only immediate rewards while a higher γ would aim at a better long-term reward), and R is the reward.

We can apply the Q-Learning algorithm to our system model by considering that the agents are the MTC devices, the environment is the network, while the state-action pair is the combination of the transmit power and the time-slot, with every device having its own Q-Table. The simplest way to implement the Q-Learning algorithm is to apply a greedy policy, this way the device always chooses the time-slot with the highest Q-value. Moreover, the greedy police also presented the best results during our simulation campaign when compared to ϵ-greedy policies. In this work, similar to [10], the reward is defined as the following:

$$R = \begin{cases} +1, & \text{successful transmission} \\ -1, & \text{failed transmission} \end{cases} . \tag{6}$$

The work in [10] also proposes an alternative reward using a congestion level that improves the performance of the method proposed therein. However, it requires the BS to detect how many devices collided in each time slot. The method in this paper requires only an acknowledgement bit per time slot, informing the success or not of the transmissions (irrespective of their number), which is much simpler in practice.

B. Novel RA Method: Combining Q-Learning and NOMA

For each device, the Q-table for every possible (transmit power, time-slot) pair is randomly initialized following a uniform distribution between -1 and 1. This initialization adds an extra degree of randomness, improving throughput over an all 0's initialization. Then, the devices choose the (transmit power, time-slot) pair with the highest Q-value. Next, the devices transmit their messages and the BS tries to recover them making use of SIC. At the end of the frame, the BS sends a single feedback message with one bit per time-slot, informing if the messages were successfully decoded or not. With this feedback each device updates its Q-value and proceeds to the next transmission. This process continues for several iterations (or frames), until it eventually converges[1].

The proposed method is summarized in **Algorithm 1**. Note that it adds minimal complexity at the device, requiring memory for storing one Q-Table with 3×K slots and the computational resources (calculation and memory) for (5). At the BS the increased complexity with respect to the method in [10] is the SIC decoding, which is non-negligible. However, it is not unrealistic to assume that the BS has more processing power than the devices, while, it is very unlikely that many devices successfully share the same time-slot, reducing the SIC complexity. Moreover, note that the Q-Learning implementation in **Algorithm 1** is distributed, each device updates its own Q-Table, which in turn influences their choice of (power, time-slot) in the next frame and the whole environment output. Implementing a centralized Q-Learning algorithm in the BS would be much more complex, requiring the BS to be aware of every device, storing all Q-Tables and making it more difficult to deploy new nodes. Also, implementing the Q-Learning at the BS would require extensive feedback as the BS would have to inform every device of its time-slot and power. Compared to related works that use Q-Learning, our method only requires the extra storage for the three power levels.

[1]The convergence of Q-Learning is well konwn [9], [20], however the convergence of a multi-agent distributed Q-Learning needs further investigation, which is outside the scope of this work. Nevertheless, in our extensive simulation campaign the proposed method always converged.

Algorithm 1 SIC-based Distributed Q-learning RA Method

Require: Q-Table random initialized between -1 and 1
for Every frame **do**
  for Every device **do**
    Select the (power, time-slot) with the highest Q value
    if More than one slot with the highest value **then**
      Choose randomly among them
    end if
  end for
  BS uses SIC, (3), to recover the transmitted messages
  BS broadcasts feedback message
  for Every device **do**
    Update Q-value for (power, time-slot) pair using (5)
  end for
end for

IV. Results

We investigate the performance of the proposed method via computer simulations, considering the setup in Section II, with the parameters in Table I, unless stated otherwise. The curves present the average result of 30 simulation runs. The proposed method is compared to three schemes. The first two are: i) SA, and ii) SA with NOMA. In SA the devices randomly choose the time-slot within a frame, without any feedback from the BS. In SA with NOMA the BS applies SIC decoding in order to try to recover some of the colliding packets. The third method is Collaborative Q-Learning [10]. In this method Q-Learning is used to allocate devices to slots, as discussed in Section II, but without NOMA. However, a different reward is employed, returning the congestion level of each time-slot, requiring the knowledge of how many devices collided in each time-slot. Finally the proposed method is presented using two different discount factors.

First, we look at the throughput, the number of successful transmissions over the number of time-slots; a metric of how efficiently the frame is being utilized. It is important to notice that this is a worst case scenario simulation where every device is transmitting in every frame. However, the proposed method does not require a transmission every frame. The device is free to move into sleep mode whenever necessary. Fig. 1 shows that the addition of NOMA considerably improves the throughput of SA. Moreover, SA with NOMA is able

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LWC.2020.3002691, IEEE Wireless Communications Letters

Table I
SIMULATION PARAMETERS

Parameter	Value
Bandwidth B	100 kHz
Carrier frequency f_c	915 MHz
Cell radius	133 m
Path loss exponent η	3
Power Deviation Δ	7.78 dB
Noise figure F	6 dB
Noise PSD N_0	-174 dBm/Hz
Outage Probability	0.01
Transmit power P_t	10 dBm
Spectral efficiency r	2 bits/s/Hz
Reference distance d_0	1 m
Devices N	25-300
Messages L	100
Simulation Runs	30
Time-slots K	100
Learning rate α	0.1
Discount factor γ	[0, 1]

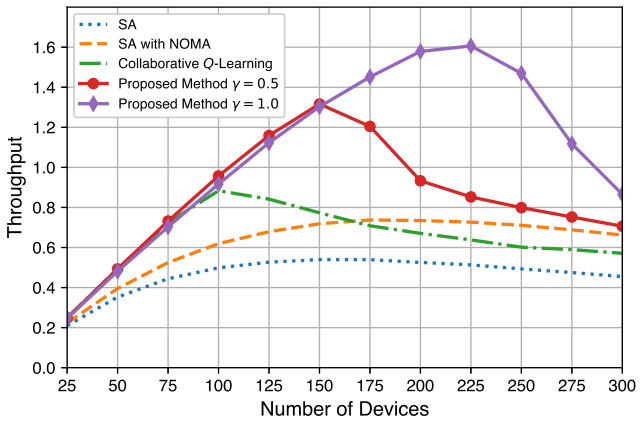


Figure 1. Throughput versus number of devices for different RA methods and the proposed scheme.

to outperform Collaborative Q-Learning from [10] when the number of devices is relatively large. However, the proposed Q-Learning method with NOMA outperforms all the other strategies, while requiring a very reduced feedback (one bit per time slot), which is much simpler in practice than the reward used in Collaborative Q-Learning [10]. Note that the peak performance occurs when $N = K$ for Collaborative Q-Learning, but with the proposed method it is obtained for $N = 2.25K$. Another interesting takeaway is that while the proposed method with $\gamma = 0.5$ performs slightly better when $N < 150$, when $\gamma = 1.0$ the performance is drastically better for $N > 150$. This can be due to the fact that γ takes future rewards into consideration. For a smaller N the devices are able to find their time-slot faster, making future rewards less important, while for a larger N the devices can take longer finding their pairs and slots making the role of γ crucial.

In order to further evaluate the performance and convergence of the proposed method, next we consider a dynamic operation setup in three stages: i) first, $N/2$ devices send $L/2$ messages, ii) then, the second half of the devices join the network and therefore in this stage N devices transmit $L/2$ messages, iii) lastly, in the third stage, as the first $N/2$ devices already transmitted their L messages, only the second half of $N/2$ devices transmit their final $L/2$ messages. Moreover, we consider two cases: $N = 130$ and $L = 200$, so that $N/2 < K$, and $N = 300$ and $L = 200$, so that $N/2 > K$. Finally, in such dynamic operation mode we can better investigate the effect of the discount factor γ in the performance of the proposed algorithm, so that we consider $\gamma \in \{0, 0.5, 1.0\}$.

As can be seen in Fig. 2-a, when $N/2 < K$, in the first stage the value of γ does not make a difference, a consequence of having more time-slots than transmitting devices. At the second stage the network is overloaded, with more devices than time-slots. In this case it is possible to see the positive effect of a larger γ, leading to faster convergence. Finally, at the third stage the network is underloaded, so that throughput decreases but again the choice of γ does not impact significantly. In Fig. 2-b $N/2 > K$, so that the network already starts with more devices than available time-slots. In this situation the advantage of a large γ is evident, converging faster and to larger values of throughput for the three stages.

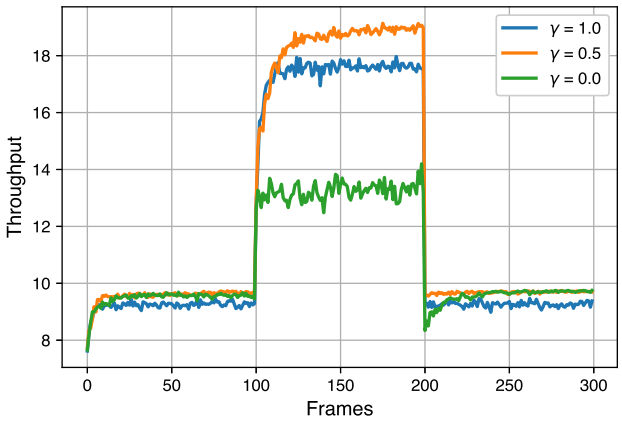
In order to better understand the behaviour of the curves in Fig. 2, recall that the discount factor γ prioritizes future rewards by softening the penalty when a collision happens, as the reward is added to the maximum Q-value weighted by γ. The curves in Fig. 2 provide a better insight on how the proposed method with $\gamma = 0.5$ is able to slightly outperform $\gamma = 1.0$ for a smaller quantity of devices. The positive effects of a smaller penalty when a collision happens can be noticed in the middle stage in Fig. 2-b, as the network suddenly becomes overloaded. A smaller γ can dismiss potential slots too quickly after collisions, making them unlikely to be utilized, resulting in a drop in the maximum system capacity and consequently in throughput, while a larger γ is able to recover the network faster and better allocate the devices resulting in a larger throughput. Moreover, we also investigated the impact of the learning rate α and found out that it is not very significant. Therefore, as $\alpha = 0.1$ has been used in [10], we then used it in all methods.

Next, we illustrate how the devices are allocated to time slots in the proposed method in Fig. 3, where the algorithm is able to allocate devices exploiting all time-slots, while also taking advantage of NOMA. Note that almost every slot is allocated to two devices providing a good distribution and there are no more than 3 devices per slot. Therefore, the algorithm was able to distribute the resources in such a way that every device has the possibility of being decoded, considering the three available power levels.

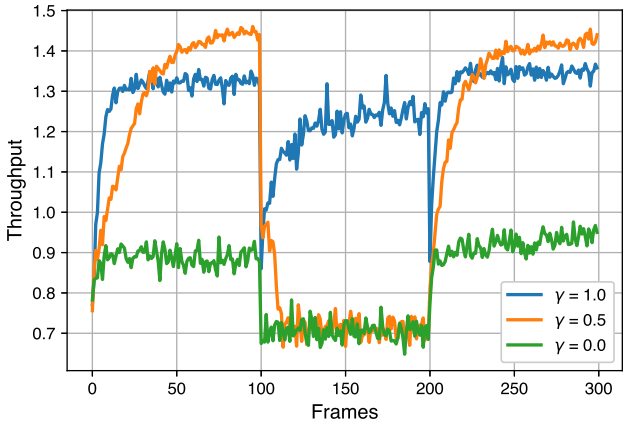
V. CONCLUSION

We introduced a novel method for RA combining the Q-Learning ability to measure uncertainties and NOMA spectral efficiency. The proposed method enables MTC devices to choose their time-slots and transmit power for improving throughput. Moreover, the method requires minimal additional complexity at the device-side, as only a simple equation has to be implemented and $3 \times K$ numerical values are stored while also preventing the device from using an unnecessary amount

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LWC.2020.3002691, IEEE Wireless Communications Letters



(a) $N/2 < K$



(b) $N/2 > K$

Figure 2. Convergence analysis as a function of the discount factor $gamma$, for a dynamic network, with a varying number of nodes.

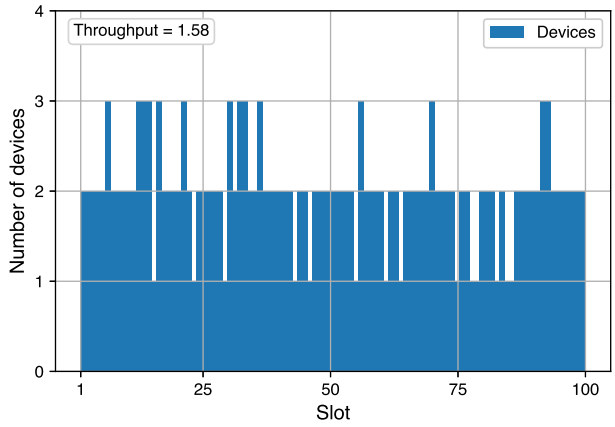


Figure 3. Allocation of devices to time slots for the proposed method.

of transmit power. From the BS it requires a very limited feedback, one bit per time slot. Simulations showed that using a larger discount factor presents the best performance when operating with a large number of devices, converging faster for a higher throughput and better handling network overload in a dynamic scenario. Furthermore, the proposed method provides significant gains in performance over other solutions.

REFERENCES

[1] H. Tullberg *et al.*, "The METIS 5G system concept: Meeting the 5G requirements," *IEEE Commun. Mag.*, vol. 54, no. 12, pp. 132–139, December 2016.

[2] B. Aazhang *et al.*, *Key drivers and research challenges for 6G ubiquitous wireless intelligence (white paper)*, 2019. [Online]. Available: http://urn.fi/urn:isbn:9789526223544

[3] Q. Bi, "Ten trends in the cellular industry and an outlook on 6G," *IEEE Communications Magazine*, vol. 57, no. 12, pp. 31–36, December 2019.

[4] CISCO, "Cisco visual networking index: Global mobile data traffic forecast update, 2017–2022," CISCO, Tech Report, Feb 2019.

[5] 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?" *arXiv*, 2019.

[6] 3GPP, "(Release 16). Technical Specification Group Services and System Aspects," 3GPP, TR 21.916 V0.4.0. Release 16, March 2020. [Online]. Available: www.3gpp.org/ftp/Specs/archive/21_series/21.916

[7] 5G Americas, "The 5G Evolution: 3GPP Releases 16-17," 3GPP, TR 1, January 2020. [Online]. Available: www.5gamericas.org/wp-content/uploads/2020/01/5G-Evolution-3GPP-R16-R17-FINAL.pdf

[8] S. K. Sharma and X. Wang, "Towards massive machine type communications in ultra-dense cellular IoT networks: Current issues and machine learning-assisted solutions," *IEEE Communications Surveys Tutorials*, pp. 1–1, 2019.

[9] R. S. Sutton and A. G. Barto, *Reinforcement learning: an introduction*. The MIT Press, 2018.

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

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

[12] Jihun Moon and Yujin Lim, "Access control of MTC devices using reinforcement learning approach," in *2017 International Conference on Information Networking (ICOIN)*, Jan 2017, pp. 641–643.

[13] L. Zhao, X. Xu, K. Zhu, S. Han, and X. Tao, "QoS-based dynamic allocation and adaptive ACB mechanism for RAN overload avoidance in MTC," in *2018 IEEE Global Communications Conference*, 2018, pp. 1–6.

[14] A. H. Mohammed, A. S. Khwaja, A. Anpalagan, and I. Woungang, "Base station selection in M2M communication using Q-learning algorithm in LTE-A networks," in *2015 IEEE 29th International Conference on Advanced Information Networking and Applications*, March 2015, pp. 17–22.

[15] S. Han, X. Xu, Z. Liu, P. Xiao, K. Moessner, X. Tao, and P. Zhang, "Energy-efficient short packet communications for uplink noma-based massive MTC networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 12, pp. 12 066–12 078, Dec 2019.

[16] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *IEEE Vehicular Technology Conference (VTC)*, 2013, pp. 1–5.

[17] A. Goldsmith, *Wireless Communications*. USA: Cambridge University Press, 2005.

[18] P. Mary, J. Gorce, A. Unsal, and H. V. Poor, "Finite blocklength information theory: What is the practical impact on wireless communications?" in *2016 IEEE Globecom Workshops (GC Wkshps)*, 2016, pp. 1–6.

[19] E. Björnson, E. de Carvalho, J. H. Sørensen, E. G. Larsson, and P. Popovski, "A random access protocol for pilot allocation in crowded massive mimo systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 4, pp. 2220–2234, April 2017.

[20] S. Kar, J. M. F. Moura, and H. V. Poor, "Distributed reinforcement learning in multi-agent networks," in *2013 5th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, Dec 2013, pp. 296–299.