This paper has been accepted for publication in 2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring), Helsinki, Finland, 25-28 April 2021.

BLER-based Adaptive Q-learning for Efficient Random Access in NOMA-based mMTC Networks

Duc-Dung Tran, Shree Krishna Sharma, and Symeon Chatzinotas
Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, Luxembourg
Email: {duc.tran, shree.sharma, Symeon.Chatzinotas}@uni.lu

***Abstract*—The ever-increasing number of machine-type communications (MTC) devices and the limited available radio resources are leading to a crucial issue of radio access network (RAN) congestion in upcoming 5G and beyond wireless networks. Thus, it is crucial to investigate novel techniques to minimize RAN congestion in massive MTC (mMTC) networks while taking the underlying short-packet communications (SPC) into account. In this paper, we propose an adaptive Q-learning (AQL) algorithm based on block error rate (BLER), an important metric in SPC, for a non-orthogonal multiple access (NOMA) based mMTC system. The proposed method aims to efficiently accommodate MTC devices to the available random access (RA) slots in order to significantly reduce the possible collisions, and subsequently to enhance the system throughput. Furthermore, in order to obtain more practical insights on the system design, the scenario of imperfect successive interference cancellation (ISIC) is considered as compared to the widely-used perfect SIC assumption. The performance of the proposed AQL method is compared with the recent Q-learning solutions in the literature in terms of system throughput over a range of parameters such as the number of devices, blocklength, and residual interference caused by ISIC, along with its convergence evaluation. Our simulation results illustrate the superiority of the proposed method over the existing techniques, in the scenarios where the number of devices is higher than the number of available RA time-slots.**

***Index Terms*—BLER, MTC, NOMA, Q-Learning, short-packet communications.**

I. Introduction

The fifth generation (5G) and beyond wireless networks (5GBNs) are expected to obtain a remarkable increase in coverage, data rates, and connection density with ultra-high reliability and ultra-low latency compared to the fourth generation (4G) wireless networks [1]. These heterogeneous and stringent requirements of 5GBNs are driving the strong development of machine-type communications (MTC) networks. The main benefit of these networks is able to support novel applications such as Internet of Things (IoT), Industry 4.0, and Tactile Internet with an extremely large number of devices (could be around 125 billion devices by 2030, according to IHS Markit forecast) [1]. However, the ever-increasing number of devices is leading to critical challenges for massive MTC (mMTC) networks related to Radio Access Network (RAN) congestion and fulfilling the diverse requirements of heterogeneous devices/services. Furthermore, to achieve the unprecedented requirements of reliability and latency, the traditional analytic methods based on Shannon theorem using long data-packets are no longer suitable for the MTC devices with short-packet communications (SPC) [2]. Therefore, 5GBNs are demanding for novel transmission methods, which can effectively support SPC [3]. In this paper, we investigate the RAN congestion issue in SPC-enabled mMTC networks, with the objective of minimizing random access channel (RACH) congestion.

In recent years, random access (RA) scheme based on machine learning (ML) has emerged as a promising solution to avoid RAN congestion in ultra-dense cellular networks including mMTC [1]. The conventional RA methods such as access class barring, slotted RA, MTC-specific backoff, separation of RA resources and paging-based RA [4]; prioritized RA, grouped-based RA, and code-expanded RA [5], are mostly reactive methods performed in a centralized manner. In contrast, the ML-based RA can bring the ability of learning the system variations for MTC devices, where some promising ML techniques such as Q-learning, can be implemented in a model-free and distributed way [6].

Taking the ML-based RA into account, some research works have been recently conducted under various scenarios [7–12]. Specifically, the authors in [7] considered a learning based RACH access scheme enabling the coexistence of human-type-communication (HTC) and MTC devices in a cellular network, in which Q-learning algorithm is utilized to intelligently assign access time-slots to the devices. The work in [8] applied Q-learning to access class barring (ACB) scheme at the base station (BS) to better adjust the value of ACB factor. In [9], Q-learning was utilized to support MTC devices in order to choose the best BS among the available BSs. Furthermore, the authors in [10] proposed a RAN congestion avoidance scheme based on collaborative distributed Q-learning to improve the performance in terms of throughput, in which the reward function is set by using the congestion level per time-slot.

To further improve the ability of RAN congestion avoidance in mMTC networks, the use of Q-learning and non-orthogonal multiple access (NOMA) was considered in [11]. Specifically, in the proposed method in [11], each MTC device selects the time-slot and transmit power for its transmission using Q-learning to improve the system throughput. However, the works [7–11] did not investigate SPC, which is considered as a potential paradigm to meet the stringent requirements of reliability and latency from 5G and beyond applications. Furthermore, the work in [11] considered perfect successive interference cancellation (PSIC), which is an ideal assumption

This work was supported by the National Research Fund (FNR), Luxembourg under the CORE project 5G-Sky (Grant C19/IS/13713801), and ERC-funded project Agnostic (Grant 742648).

for NOMA transmission. Given this context, Han et al. [12] proposed a power allocation solution to maximize energy efficiency for SPC in NOMA-based mMTC networks. Furthermore, the authors applied Q-learning to allocate devices to different subchannels such that the number of devices sharing the same subchannel does not exceed a predetermined threshold. However, the work in [12] did not investigate the resource allocation optimization problem (i.e., transmit power and subchannel/time-slot) based on Q-learning to enhance the throughput of NOMA-based mMTC networks with SPC.

In this paper, we investigate the combination of Q-learning and NOMA-based SPC for time-slot and power allocation to address RACH congestion problem in mMTC networks. Specifically, we utilize block error rate (BLER), an important performance metric to characterize the SPC-based systems, as a global cost during the learning process. Given this context, we propose an adaptive Q-learning (AQL) algorithm for a NOMA-based mMTC network, namely BLER-NOMA-AQL. The main contributions of this paper are briefly summarized as follows: i) we propose a BLER-NOMA-AQL scheme for time-slot and power allocation to address the RACH congestion problem in a NOMA-based mMTC network; ii) we study the effects of imperfect SIC (ISIC) on the proposed method in terms of the system throughput as compared to the widely-used PSIC assumption; iii) we analyze and compare the throughput performance of the proposed BLER-NOMA-AQL method with some recently proposed techniques in the literature, along with its convergence analysis.

The remainder of the paper is organized as follows. Section II depicts the system model in detail. Section III presents the Q-learning algorithm for RA, effective BLER for SPC-based NOMA transmission, and the proposed BLER-NOMA-AQL scheme. Section IV describes the numerical results. Finally, Section V concludes this paper.

II. System Model

We investigate an uplink NOMA-based mMTC network consisting of one cellular base station (BS) and M MTC devices, as depicted in Fig. 1. In this network, the devices transmit their short data packets to the BS through a frame-based slotted aloha (SA) scheme, as utilized in [10, 11]. We assume that each frame has T available time-slots, denoted by $\mathbb{T} = \{1; 2; \ldots; T\}$, and each device has L data packets ready for transmission. Following the principle of SA scheme, each device transmits a single packet to the BS by using one of T time-slots within a frame. After each frame, the BS transmits a feedback bit to the devices in order to inform their transmission outcomes (i.e., success or failure) [10, 11]. In addition, this control message can be used for synchronizing the devices [11]. Note that in SA scheme, multiple devices can select the same time-slot for their transmissions, which may result in a collision. In the conventional SA, a collision occurs when there are more than one device selecting the same time-slot. In contrast, the NOMA-based SA which is considered in this paper can serve multiple devices in one time-slot by utilizing different transmit power levels [11].

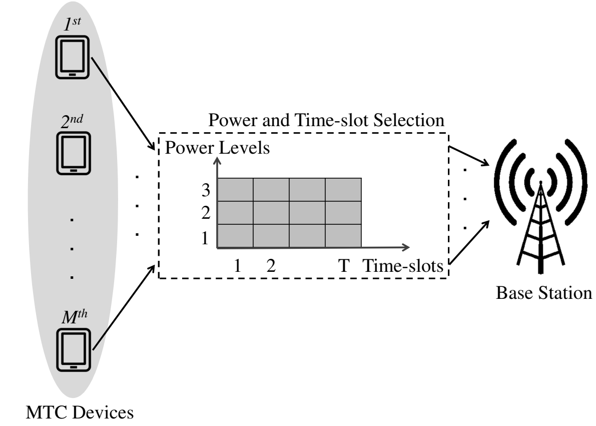

Fig. 1. Model of an uplink NOMA-based mMTC network under SPC.

Let us assume that there are $\hat{M}$ ($\hat{M} < M$) devices using the same t-th ($1 \leq t \leq T$) time-slot to transmit their messages to the BS. Given the NOMA principle, the received signal at the BS in the t-th time-slot of the i-th frame is given by [11]

$$y_t\left(i\right) = \sum_{m=1}^{\hat{M}} h_{m,t}\left(i\right) \sqrt{P_{m,t} d_m^{-\theta}} x_{m,t}\left(i\right) + n_t\left(i\right), \quad (1)$$

where $h_{m,t}(i)$ denotes the Rayleigh channel coefficient of the link from device m to the BS in time-slot t, which is assumed to be constant during frame i under a quasi-static scenario [11]; $P_{m,t}$ is the transmit power of device m in time-slot t; d_m is the distance from device m to the BS; θ denotes the path loss exponent; $x_{m,t}(i)$ denotes the message of device m in time-slot t; and $n_t\left(i\right) \sim \mathcal{CN}\left(0, \sigma^2\right)$ is the additive white Gaussian noise (AWGN).

At the BS, SIC is utilized to detect multi-user data by treating the messages of weaker devices as noise when decoding the message of a stronger device [12]. Specifically, the device set $\left\{1, \ldots, \hat{M}\right\}$ is assumed to be ordered in decreasing received power, i.e., $P_{m,t}\lambda_{m,t}$, where $\lambda_{m,t} = \left|h_{m,t}\right|^2 d_m^{-\theta}$. In this paper, we use three transmit power levels, denoted by $\mathbb{P} = \{P_t - \delta,\ P_t,\ P_t + \delta\}$, where P_t is the reference power and δ denotes the power deviation, so that NOMA can work properly, as discussed in [11]. Thus, each device will select one of these transmit power levels for its transmission. Furthermore, unlike [11] considering PSIC, this paper investigates a more practical scenario by taking ISIC into account. From (1), the instantaneous signal-to-interference-plus-noise ratio (SINR) of device m is given by

$$\gamma_{m,t} = \frac{P_{m,t}\lambda_{m,t}}{I_{m,t} + \hat{I}_{m,t} + \sigma_2}, \quad (2)$$

where $I_{m,t} = \sum\limits_{j=m+1}^{\hat{M}} P_{j,t}\lambda_{j,t}$ denotes the interference of device m, $\hat{I}_{m,t} = \sum\limits_{j=1}^{m-1} \eta P_{j,t}\lambda_{j,t}$ is the residual interference component due to ISIC [12], and $0 \leq \eta \leq 1$ represents the level of residual interference caused by ISIC.

In this paper, we consider a scenario that the message of device m is successfully decoded if its instantaneous SINR is larger than or equal to a threshold [11], γ_{th}, i.e.,

$$\gamma_{m,t} \geq \gamma_{th}, \tag{3}$$

where $\gamma_{th} = 2^{r_{th}} - 1$ and r_{th} denotes the minimum spectral efficiency threshold for successful transmission.

III. Proposed Adaptive Q-Learning Method

In this section, the proposed BLER-NOMA-AQL method is presented. This approach aims to effectively allocate the power and time-slot for MTC devices utilizing the effective BLER of each device as the global cost during the learning process. Furthermore, the analysis of effective BLER for the considered NOMA-based MTC system is provided in Sec. III-B.

A. Q-learning for Random Access

The application of reinforcement learning, especially Q-learning, to MTC networks has recently gained great attention [1]. It can be implemented in a distributed manner and supports MTC devices to learn from the previous experiment by interacting with the environment. The RA in an MTC network can be modeled as Markov Decision Process (MDP). Given MDP principle, an agent can interact with the environment to move from the current state to the next state by performing an appropriate action and receive a respective reward [6].

With Q-learning, the agent builds its own action-value function, so-called the Q-table, to depict the agent-environment relationship. The simplest way to select the action is to utilize a greedy policy. Given this policy, at time step k and state $s_k \in \mathcal{S}$, an agent selects an action $a_k \in \mathcal{A}$ with the highest Q-value. As a result, the agent moves to the next state s_{k+1} and receives a reward r_{k+1}. After performing action a_k, the new Q-value for state-action pair (s_k, a_k), i.e., $Q_{k+1}(s_k, a_k)$, is updated based on the following iterative procedure [6, 10]

$$\begin{aligned} Q_{k+1}(s_k, a_k) &= (1-\alpha_k)\, Q_k(s_k, a_k) \\ &\quad + \alpha_k \left[R_{k+1} + \gamma \max_{a} Q_k(s_{k+1}, a) \right], \end{aligned} \tag{4}$$

where $0 \leq \alpha \leq 1$ denotes the learning rate applied at the k-th time step, $0 \leq \gamma \leq 1$ is the discount factor, and R_{k+1} represents the reward function defined as follows [10]:

$$R_{k+1} = \begin{cases} 1, & \text{for successful transmission,} \\ p_f, & \text{for otherwise.} \end{cases}, \tag{5}$$

where $p_f = -1$ denotes the penalty function.

The adoption of Q-learning algorithm to our system model can be achieved by considering the devices as the agents, the investigated network as the environment, and the combination of the transmit power and time-slot as the state-action pair. Specifically, a device is in state $s(p,t) \in \mathcal{S}$ ($p \in \mathbb{P}$ and $t \in \mathbb{T}$) if it occupies a (transmit power, time-slot) pair (p,t). An action $a(s,s') \in \mathcal{A}$ is a transition from a certain state s to a target state s', where a device changes its selection from one (transmit power, time-slot) pair to another one. In this regard, each device will use the greedy policy to build its own Q-table. In particular, at the beginning, all Q-values, for every (transmit power, time-slot) pair, are initialized to zero. Then, each device selects randomly a (transmit power, time-slot) pair for its transmission. Next, each device updates Q-value of selected (transmit power, time-slot) pair based on its successful or unsuccessful transmission outcome by using (4). After the first frame, each device chooses a (transmit power, time-slot) pair with the highest Q-value for its next transmission. This learning process continues in several frames till the convergence, i.e., all devices find unique (transmit power, time-slot) pairs for their transmissions, is observed.

B. Effective BLER for SPC-based NOMA System

Considering SPC in the investigated NOMA-based MTC network, the instantaneous BLER of decoding the message of the m-th device at the BS in the t-th time-slot is given by [3]

$$\hat{\varepsilon}_{m,t} \approx Q\left(\frac{\log_2(1+\gamma_{m,t}) - b_m/B_m + \log_2 B_m/2B_m}{\sqrt{v_{m,t}/B_m}} \right), \tag{6}$$

where $Q(x) = \int\limits_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$ is the Gaussian Q-function, $v_{m,t} = (\log_2 e)^2 \left[1 - \frac{1}{(1+\gamma_{m,t})^2} \right]$ represents the channel dispersion, B_m and b_m denote the blocklength and the number of information bits to device m, respectively.

It is noted that the BS needs to decode the messages of the former $(m-1)$ stronger devices before detecting the message of the m-th device. Therefore, the effective BLER for the m-th device is given by

$$\varepsilon_{m,t} = 1 - \mathcal{E}_{j,t} + \mathcal{E}_{j,t} \hat{\varepsilon}_{m,t}, \tag{7}$$

where $\mathcal{E}_{j,t} = \prod\limits_{j=1}^{m-1} (1 - \hat{\varepsilon}_{j,t})$ and $\hat{\varepsilon}_{j,t}$ is calculated from (6).

C. Proposed BLER-NOMA-AQL Scheme

Based on (2) and (6), we observe that the SINR of the devices decreases when the number of devices increases due to the higher interference. This causes increase in the BLER, hence, leading to the reduction of reliability level in short-packet transmission. Therefore, awareness of reliability level is necessary so that the devices can select the best (transmit power, time-slot) pairs for their transmission in case the large inter-user interference is observed in chosen time-slots.

We consider the reliability level of transmission process based on the instantaneous BLER of each device calculated at the BS. Thus, for proposed AQL algorithm, we define the reward function similar to (5), where the penalty function p_f is defined as

$$p_f = -\varepsilon_{m,t}. \tag{8}$$

The devices update their Q-table by using (4) with the reward and penalty functions defined in (5) and (8), respectively. The value of $\varepsilon_{m,t}$ to be used in (8) is obtained from the effective BLER given by (7). The detail of the proposed method is depicted in Algorithm 1.

Algorithm 1: Proposed BLER-NOMA-AQL algorithm for minimizing RACH congestion in NOMA-based MTC networks.

```
Data : $M$, $T$, $P_t$, $\delta$, $b_m$, $B_m$, $\gamma_{th}$, number of
       iterations/frames for learning process $K$.
Result: Q-Table for $M$ devices.
Initialize $3 \times T$ zero Q-table for all devices, $k \leftarrow 1$;
while $k \leq K$ do
    Device $m$ ($1 \leq m \leq M$) selects an action $a_m$, i.e.,
      selecting a (power transmit, time-slot) pair for its
      transmission, with highest Q-value;
    if $size(a_m) > 1$ then
        Choose from $a_m$ an action randomly;
    end
    Take action $a_m$, observe reward according to (5)
      and (8);
    Update Q-value according to (4);
    $k \leftarrow k + 1$;
end
```

IV. Numerical Results

In this section, we provide the performance analysis in terms of throughput of the proposed AQL method over a range of system parameters via numerical results. For the proposed AQL algorithm, we set the learning rate $\alpha = 0.1$, the discount factor $\gamma = 0.5$, the reward and penalty functions are defined in (5) and (8), respectively. In addition, the predetermined simulation parameters are set as follows [11, 12]: the time-slots $T = 150$, the reference power $P_t = 10$ dBm, the power deviation $\delta = 7.78$ dB, device m ($1 \leq m \leq M$) are randomly deployed around the BS with the distance $d_m \leq 120$ m, the path loss exponent $\theta = 3$, the noise variance $\sigma^2 = -174$ dBm, the number of messages $L = 100$, the spectral efficiency $r_{th} = 2$ bits/s/Hz. All devices have the same number of information bits $b = 40$ and blocklength B.

To evaluate the system performance, we use throughput metric which is defined as the number of successful transmissions over the number of available time-slots [11]. In Fig. 2, we plot the throughput versus the number of devices (M) with different schemes. Herein, we compare our proposed BLER-NOMA-AQL method to the following three schemes: SA, SA with NOMA (NOMA-SA), and NOMA-based QL (NOMA-QL) [11]. In SA, the devices randomly select the time-slot within a frame, where a collision occurs when there are more than one device utilizing the same time-slot [10]. In NOMA-SA, NOMA is used to support the decoding of the collided messages. In NOMA-QL, both QL and NOMA are applied to enable the devices to choose the best time-slots for their transmissions using the binary reward defined in (5), where the penalty function $p_f = -1$. In contrast to NOMA-QL, our proposed method utilizes BLER as the global cost for the learning process to define the penalty function in (8).

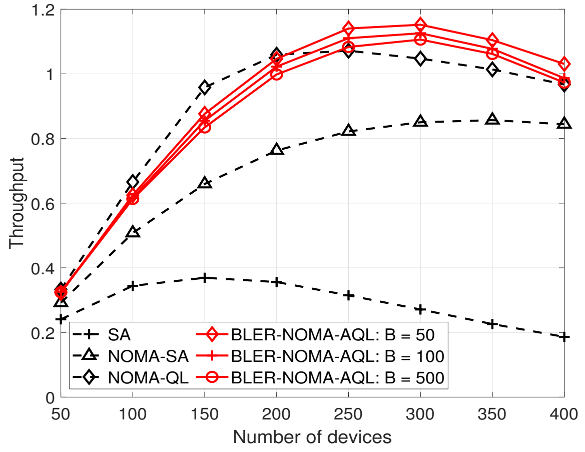

Fig. 2. Throughput versus number of devices with different blocklength and RA methods, where $\eta = 0.05$.

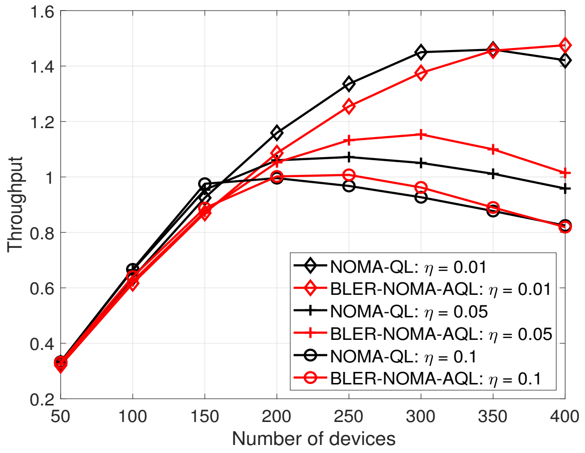

Fig. 3. Throughput versus number of devices with different values of η and RA methods, where $B = 50$.

Fig. 2 shows that the throughput of SA can be significantly improved by utilizing NOMA. This figure also indicates that using Q-learning brings better performance in terms of throughput than the case without Q-learning. Furthermore, in comparison with NOMA-QL, our proposed BLER-NOMA-AQL method outperforms in high M regime, but it obtains slightly lower performance in small M area. This can be due to the opposite relationship between SINR and BLER, as depicted in (6). For a smaller M, the SINR of the devices are higher due to the decrease in the interference. This leads to the reduction of BLER, making it less crucial. Meanwhile, for a higher M, the interference is larger and the SINR decreases, leading to the increase in BLER and its role becomes important for learning process of the proposed AQL algorithm. When considering long-packets for transmission as utilized in [11], BLER can be ignored. However, it is an important performance metric to be considered in SPC. With the growing role of SPC in new applications of 5GBNs, the proposed method could be a promising solution to mitigate RAN congestion considering the ever-increasing number of devices in mMTC networks. It is noteworthy to mention that under the given conditions, the peak throughput occurs when $M = 1.67T$ for NOMA-QL, whereas it is achieved when $M = 2T$ for our proposed method

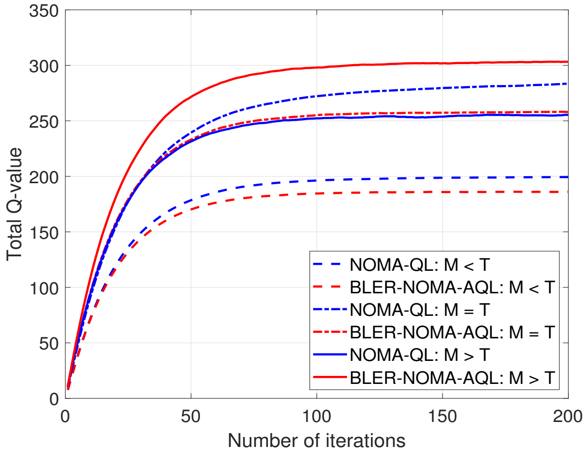

Fig. 4. Convergence of Q-learning with different NOMA-based RA methods, $\eta = 0.05$ and $B = 50$.

as reflected in Fig. 2. Furthermore, compared to NOMA-QL, our proposed scheme can improve the throughput with 6.42% when $M = 1.67T$ and 10.01% when $M = 2T$.

To evaluate the effect of blocklength (B) on the system performance, we depict the variation of throughput obtained from the proposed BLER-NOMA-AQL scheme with different values of B, as also depicted in Fig. 2. One can observe from this figure that the lower throughput can be achieved when B increases. This can be explained by the fact the increase in the value of B leads to the reduction of BLER, making its importance reduced when implementing AQL algorithm.

In practice, it is difficult to achieve perfect SIC which is used in [11]. Therefore, it is reasonable to consider ISIC when evaluating the benefits of different NOMA-based Q-learning algorithms for RAN congestion problem. Taking the effect of ISIC on throughput performance into account, Fig. 3 plots the throughput of NOMA-QL and BLER-NOMA-AQL versus M with the different values of the residual interference level obtained after ISIC η. This figure shows that the throughput values achieved in case of using NOMA-QL scheme and our proposed BLER-NOMA-AQL method decrease with the increase in η due to the resulting higher interference, leading to the reduction of SINR. However, in this case, our proposed BLER-NOMA-AQL method still achieves higher peak throughput than NOMA-QL in high M regime.

Considering the learning process of the Q-learning algorithm, the achieved Q-value will gradually converge to a certain value when the devices find unique (transmit power, time-slot) pairs for their transmissions [6, 10]. To achieve the further analysis of our proposed BLER-NOMA-AQL method, we evaluate this convergence via the parameter total Q-value of all learning MTC devices, as depicted in Fig. 4. Specifically, we consider three cases: $M < T$, $M = T$, and $M > T$ ($M \in \{100, 150, 200\}$ and $T = 150$). For $M \leq T$, the proposed method obtains slightly lower total Q-value leading to the slower convergence ability compared to NOMA-QL. Meanwhile, for $M > T$, our proposed method gradually converges to a total Q-value much higher than NOMA-QL. This confirms the results achieved from Figs. 2 and 3, where our proposed method outperforms the NOMA-QL when M is relatively higher than T.

V. Conclusion

This paper has proposed a novel AQL method to address the issue of RACH congestion in SPC-enabled NOMA-based mMTC networks. In contrast to the Q-learning schemes considered in the literature, the proposed AQL utilizes BLER as the global cost in the learning process. Numerical results have shown that the proposed AQL method provides better performance compared to the existing schemes (SA, NOMA-SA, and NOMA-QL) when M is higher than the number of available time-slots, making it important as the number of devices is ever-increasing in beyond 5G-mMTC networks. Furthermore, the impact of ISIC on throughput performance of the proposed scheme has been analyzed and it has been noted that the throughput decreases with the increase in the value of residual interference caused by the ISIC. In our future work, we plan to evaluate the performance of the proposed AQL method for other SPC scenarios such as sparse code division multiple access (SCMA)-based mMTC networks.

References

[1] 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, Firstquarter 2020.

[2] G. Durisi, T. Koch, and P. Popovski, “Toward massive, ultra-reliable, and low-latency wireless communication with short packets,” *Proc. IEEE*, vol. 104, no. 9, pp. 1711–1726, Sep. 2016.

[3] Y. Polyanskiy, H. V. Poor, and S. Verdu, “Channel coding rate in the finite blocklength regime,” *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[4] “Study on RAN improvements for machine-type communications,” 3GPP, techreport, Sep. 2011, TR 37.868.

[5] M. S. Ali, E. Hossain, and D. I. Kim, “LTE/LTE-A random access formassive machine-type communications in smart cities,” *IEEE Commun. Mag.*, vol. 55, no. 1, pp. 76–83, Jan. 2017.

[6] R. Li, Z. Zhao, X. Zhou, G. Ding, Y. Chen, Z. Wang, and H. Zhang, “Intelligent 5G: When cellular networks meet artificial intelligence,” *IEEE Wireless Commun.*, vol. 24, no. 5, pp. 175–183, Oct. 2017.

[7] L. M. Bello, P. Mitchell, and D. Grace, “Application of Q-learning for RACH access to support M2M traffic over a cellular network,” in *Eur. Wireless Conf.*, Barcelona, Spain, May 2014.

[8] J. Moon and Y. Lim, “Access control of MTC devices using reinforcement learning approach,” in *Int. Conf. Inf. Netw. (ICOIN)*, Da Nang, Vietnam, Jan. 2017.

[9] 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 *Int. Conf. Adv. Inf. Netw. Appl.*, Gwangiu, South Korea, Mar. 2015.

[10] 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.

[11] M. V. da Silva, R. D. Souza, H. Alves, and T. Abrão, “A NOMA-based Q-learning random access method for machine type communications,” *IEEE Wireless Commun. Lett.*, Jun. 2020, Early Access.

[12] 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 Trans. Veh. Technol.*, vol. 68, no. 12, pp. 12 066–12 078, Dec. 2019.