Novel Reinforcement Learning based Power Control and Subchannel Selection Mechanism for Grant-Free NOMA URLLC-Enabled Systems

Duc-Dung Tran, Vu Nguyen Ha, and Symeon Chatzinotas
Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, Luxembourg
Emails: {duc.tran, vu-nguyen.ha, Symeon.Chatzinotas}@uni.lu

***Abstract*—Reducing waiting time due to scheduling process and exploiting multi-access transmission, grant-free non-orthogonal multiple access (GF-NOMA) has been considered as a promising access technology for URLLC-enabled 5G system with strict requirements on reliability and latency. However, GF-NOMA-based systems can suffer from severe interference caused by the grant-free (GF) access manner which may degrade the system performance and violate the URLLC-related requirements. To overcome this issue, the paper proposes a novel reinforcement-learning (RL)-based random access (RA) protocol based on which each device can learn from the previous decision and its corresponding performance to select the best subchannels and transmit power level for data transmission to avoid strong cross-interference. The learning-based framework is developed to maximize the system access efficiency which is defined as the ratio between the number of successful transmissions and the number of subchannels. Simulation results show that our proposed framework can improve the system access efficiency significantly in overloaded scenarios.**

***Index Terms*—Grand-free NOMA, Q-Learning, URLLC.**

I. Introduction

Providing ultra-reliable and low-latency communications (URLLC) to special services with stringent reliability and latency requirements and supporting massive access over a limited radio spectrum are two important use-cases of the fifth generation (5G) and beyond wireless networks (5GBNs) [1–6]. Recently, GF-NOMA has been demonstrated as a promising solution for these such use-cases [7]. With GF-NOMA strategy, each device can access any subchannel (SC) quickly without waiting for receiving the admission granted by the base station. Furthermore, the NOMA transmission can be exploited when there are more than one device access any SC. Hence, this technology can improve the spectrum access efficiency significantly while reduce transmission latency for the system. However, opening the spectrum for free access without admission control may cause congestion problem where there are too many devices competing a limit number of SCs. To mitigate this issue, [8, 9] proposed to modify GF-NOMA scheme by dividing a cell into several fractions and using orthogonal resource allocation among different fractions. These modified framework can reduce inter-fraction collisions but the spectrum competition among devices within the same fraction still causes severe collisions. Thus, it is important to find a smart resource access solution to further reduce the collisions and improve the system throughput.

In recent years, RL algorithms, especially Q-learning (QL), has been applied to intelligently address the collisions and severe interference in massive access scenario [10–15]. Specifically, [10] proposed a collaborative distributed QL algorithm for frame-based slotted-Aloha (SA) RA scheme to find the best RB allocation strategy for devices in order to avoid the collision in GF orthogonal multiple access systems. On different approaches, [11–15] employed QL into different GF-NOMA systems to mitigate the congestion issue and interference in overloaded scenario. However, in these works, only the scheme allowing all devices compete over all SCs is investigated and no suitable solution for URLLC-enabled systems is considered.

This paper aims to employ the RL method to develop a novel power control (PC) and SC selection mechanism for GF-NOMA URLLC-enabled system to address the collision challenge while maintaining the strict URLLC-related requirements. The proposed framework can be employed at the devices to help them select the best SC and power level for transmission to improve the system RA efficiency and requirements on reliability and latency. In this work, two different GF-NOMA transmission methods, namely with and without SDC (wSDC and woSDC) are investigated. In wSDC scheme, the SCs and devices are grouped into various clusters in each of which GF-NOMA is performed separately. While the woSDC scheme allows all devices to compete for all available SCs to perform GF-NOMA. The simulation is then demonstrated to evaluate the system performance in term of AE and convergence of the proposed learning algorithm.

II. System Model

We investigate an uplink URLLC-enabled GF-NOMA system consisting of one central-cell BS and M devices allocated randomly within a circle-cell with radius of r_0 (m), as shown in Fig. 1. In this setting, the system bandwidth of W (MHz) is divided equally into K SCs (SCs) (so-called resource blocks - RBs), i.e., the bandwidth of each SC can be expressed as $W_{\mathsf{SC}} = W/K$. Furthermore, this paper focuses on saturated traffic scenarios where data packets are always available at the beginning of each slot for all devices. GF-NOMA-based transmission scheme is assumed in this system where every SC is opened for access from all devices.

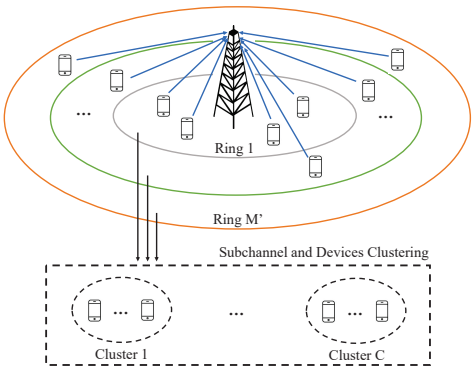

Fig. 1. Illustration of SDC in an uplink URLLC-enabled GF-NOMA system.

A. Grant-Free Access Strategy

Following the Grant-Free strategy, the SCs are opened for devices to access without scheduling processes; then, the devices can individually choose the SCs and a power level for data transmission. However, letting all device freely access all SCs in the system may result in unbalancing resource allocation where there exits one SC accessed by too many devices. In such scenario, the devices' signal may suffer the strong interference and the system's performance can be degraded. To overcome this challenge, admission-limited GF-NOMA frameworks, so-called SC and device clustering (SDC) methods, have been proposed to reduce the set of chosen-able SCs for each device and also group devices in cluster to avoid the strong cross-interference [9]. This work investigates both original and admission-limited GF-NOMA communication schemes which are named as:

- *Without SDC (woSDC)*: All K SCs are available for all devices to compete for their transmissions [10–13]. In this scheme, there is a high probability that too many devices admitted to one SC which results in performance degradation due to the heavy interference among these devices.
- *With SDC (wSDC)*: This scheme aims to reduce the interference by clustering devices and SCs for NOMA transmission [9]. Specifically, let C be the number of clusters and K' present the number of SCs in each cluster, i.e., $C = \lceil K/K' \rceil$, where $\lceil \cdot \rceil$ denotes the ceiling function. Here, there are K' SCs arranged in each of the first $C-1$ clusters while the C-th cluster consists of the remaining SCs, i.e., $K-(C-1)K'$. To access the SCs within different clusters, the devices are grouped into $M' = \lceil M/C \rceil$ rings based on their distances to the BS so that the cardinal number of each group is C or $C-1$. In particular, each of the first $M - M'(C-1)$ rings consists of C devices while the each of the remaining contains $C-1$ devices. Then, the set of devices allowed to access SCs in each cluster can be determined by randomly choosing one device from each ring.

B. NOMA Transmission Process

This section briefly overviews the NOMA transmission process over SC k. Denote $a_{m,k}$ as a binary indicator where $a_{m,k} = 1$ indicates that device m decides to access SC k and vice versa. Then, the set of devices accessing SC k can be described as $\mathcal{M}_k = \{m | a_{m,k} = 1\}$. Let M_k ($M_k < M$) be the number of devices using SC k for their data transmission. When $M_k > 1$, the power-domain NOMA transmission is employed. Let $P_{m,k}$ be transmission power due to device m over SC k. Here $P_{m,k}$ can be selected from a predetermined set $\mathcal{P} = \{P_1, P_2, ..., P_L\}$, i.e., $P_{m,k} \in \mathcal{P}$. Then, the signal received by the BS over SC k can be expressed as

$$y_k = \sum_{m \in \mathcal{M}_k} \sqrt{P_{m,k}} g_{m,k} x_{m,k} + N_0, \tag{1}$$

where $x_{m,k}$ presents the data symbol of device k, N_0 is the additive white Gaussian noise (AWGN), and $g_{m,k}$ stands for the channel coefficient of the wireless link between device k and the BS over SC k. Here, both large-scale fading and small-scale fading are considered in the channel model. In particular, the channel gain over SC k from device k to the BS is modeled as $|g_{m,k}|^2 = h_{m,k} d_m^{-\theta}$ where θ is the path loss exponent, d_m represents the distance from device m to the BS, and $h_{m,k}$ stands for the random small-scale fading factor.

Without loss of generality, one assumes $\mathcal{M}_k = \{1, \ldots, M_k\}$ where the devices are sorted in the descending order of the corresponding received power level at the BS, i.e., $P_{m,k} h_{m,k} d_m^{-\theta}$. Following NOMA principle, the devices with higher received power level are decoded earlier at the BS. In particular, the BS decodes the message of a device by considering the message of devices with lower received power level as noise [11]. It then removes this component from its observation to detect the remaining devices' messages by using successive interference cancellation (SIC) technique. Given this context, the received signal-to-interference-plus-noise ratio (SINR) of device m over SC k can be written as

$$\gamma_{m,k} = \frac{P_{m,k} h_{m,k} d_m^{-\theta}}{\underbrace{\sum_{i=1}^{m-1} \eta P_{i,k} h_{i,k} d_i^{-\theta}}_{\mathcal{I}_1} + \underbrace{\sum_{j=m+1}^{M_k} P_{j,t} h_{j,k} d_j^{-\theta}}_{\mathcal{I}_2} + \sigma^2}, \tag{2}$$

where $\mathcal{I}_1$ is the residual interference component due to SIC process, $\mathcal{I}_2$ denotes the interference caused by devices having received power level lower than device m, and η ($0 \le \eta \le 1$) represents the imperfect level of SIC process.

C. URLLC-enabled Communication

For URLLC communication, very short packets and finite blocklength (FBL) are implemented for data transmission due to the stringent requirements on latency and reliability. Consequently, Shannon's capacity formula cannot be applied for URLLC communication model since it is designed under the assumption of infinite blocklength (iFBL). According to [16], the achievable rate of device m in FBL regime for a quasi-static flat fading channel can be approximated as

$$R_{m,k} \approx W' \left[\log_2 (1 + \gamma_{m,k}) - \sqrt{\frac{V_{m,k}}{D_{max} W'}} \frac{Q^{-1}(\varepsilon_{m,k})}{\ln 2} \right], \tag{3}$$

where $V_{m,k} = (\log_2 e)^2 \left[1 - \frac{1}{(1+\gamma_{m,k})^2}\right]$ represents the channel dispersion, D_{max} denotes the maximum transmission latency, $Q^{-1}(x)$ is the inverse of the Gaussian Q-function $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$, and $\varepsilon_{m,k}$ is the decoding error probability (DEP) which can be used to evaluate the transmission reliability. Following NOMA principle, the BS needs to decode the messages of $m-1$ devices having better received power levels successfully and then removes these components from its observation before detecting the message of device m. Thus, the effective DEP of device m is expressed as

$$\varepsilon_{m,k} = 1 - \prod_{i=1}^{m-1} (1 - \varepsilon_{i,k}) + \prod_{i=1}^{m-1} (1 - \varepsilon_{i,k})\varepsilon_0, \quad (4)$$

where ε_0 is the initial DEP for NOMA process and $\varepsilon_{1,k} = \varepsilon_0$.

D. Access Efficiency Maximization

In this work context, the message of device m transmitted over SC k is successfully decoded at the BS if its achievable rate $R_{m,k}$ with URLLC requirements (i.e., D_{max} and $\varepsilon_{m,k}$) is larger than or equal to a predetermined rate threshold R_{th}, i.e., $R_{m,k} \geq R_{th}$. Then, the access attempt of device m on SC k is considered as successful access if its data stream is decoded correctly. Based on that, an access efficiency factor is introduced in this work context as

$$\mathsf{AE}(\mathbf{P}, \mathbf{A}) = \sum_{k=1}^{K} \sum_{m \in \mathcal{M}_k} \mathbb{1}_{\{R_{m,k} \geq R_{th}\}} / K, \quad (5)$$

where $\mathbb{1}_{\{\mathsf{X}\}}$ presents the indicating function which equals to one if X is true and zero otherwise; $\mathbf{P}, \mathbf{A}$ stand for the matrix of power and SC selection variables, respectively.

In this paper, we focus on develop a distributed power control and SC selection framework which can be processed at devices to maximize the system access efficiency. This problem can be stated as

$$\max_{\mathbf{P},\mathbf{A}} \mathsf{AE}(\mathbf{P}, \mathbf{A}) \text{ s.t. } a_{m,k} \in \{0, 1\}, \ \forall (m, k), \quad (6a)$$

$$P_{m,k} \in \{P_1, P_2, ..., P_L\}, \ \forall (m, k), \quad (6b)$$

Solving problem (6) is very challenging due to the coupling between binary variables $\{a_{m,k}\}$'s and discreet variables $\{P_{m,k}\}$'s, and the complicated form of $R_{m,k}$ given in (3).

III. Learning-based Random Access Mechanism

Although GF access can reduce access latency and increases the number of active devices for URLLC-enabled GF-NOMA systems, its random access nature causes collisions since multiple devices can use the same SC. To mitigate this drawback, we apply an efficient QL framework to improve the network throughput by allowing multiple devices to use the same SC while guaranteeing the required latency and DEP for devices.

QL is one of the most popular RL algorithms, which can be implemented in a distributed manner. It enables an agent to interact with the environment and learn from the previous experience in the absence of a training data-set to perform a task in a sequence of time-steps $\{1, \ldots, t, \ldots, T\}$. At each time-step t, an agent can move its state from the current state $s_t \in \mathcal{S}$ to the next state $s_{t+1} \in \mathcal{S}$ by taking an action $a_t \in \mathcal{A}$, and during this transition, the agent receives a respective reward r_{t+1}. To depict the agent-environment relationship, the agent builds an action-value function, namely Q-function. After performing action a_t, the new Q-value at each state is calculated based on the following iterative procedure [17]

$$Q_{t+1}(s_t, a_t) = (1 - \alpha) Q_t(s_t, a_t) + \alpha \left[r_{t+1} + \gamma \max_{a \in \mathcal{A}} Q_t(s_{t+1}, a) \right], \quad (7)$$

where $0 \leq \alpha \leq 1$ denotes the learning rate, $0 \leq \gamma \leq 1$ is the discount factor, and r_{t+1} represents the reward function.

To apply the above-described QL algorithm into the considered URLLC-enabled GF-NOMA system, we consider that each device is a learning agent and define $Q_t(m, a_{m,t})$ as the Q-value of device m at time-step t with the action $a_{m,t}$. The action here is the transmit power and SC selection $a_{m,t} \in \mathcal{A} = \left\{1, \ldots, lk, \ldots L\hat{K}\right\}$, where $\mathcal{A}$ is the set of possible actions, L is the maximum number of available transmit power levels (TPLs), and $\hat{K}$ is the number of available SCs depending on NOMA transmission method, i.e., woSDC or wSDC. We define $\mathcal{P} = \{P_1, \ldots, P_L\}$ as the set of TPLs. After taking the action $a_{m,t}$, the new Q-value $Q_{t+1}(m, a_{m,t})$ is updated as follows:

$$Q_{t+1}(m, a_{m,t}) = (1 - \alpha) Q_t(m, a_{m,t}) + \alpha \left[r_{m,t+1} + \gamma \max_{a \in \mathcal{A}} Q_t(m, a) \right], \quad (8)$$

where the reward function $r_{m,t+1}$ is defined as

$$r_{m,t+1} = \begin{cases} 1, & \text{if the transmission is successful,} \\ p_v, & \text{otherwise,} \end{cases} \quad (9)$$

Algorithm 1: QL-based PC and SC Selection Algorithm

Data : M, K, K', $\mathcal{P}$, D_{max}, W', r_0, δ, α, γ, R_{th}, p_v maximum number of iterations for learning process I.

Result: Q-Table for M devices.

Calculate the number of clusters C, the number of SCs and the number of devices in each cluster with respect to wSDC method;
Determine the number of available SC $\hat{K}$ for device m based on the applied NOMA transmission method, i.e., woSDC or wSDC;
Initialize $L \times \hat{K}$ zero Q-table for device m, $i \leftarrow 1$;
while $i \leq I$ **do**
    Device m ($1 \leq m \leq M$) selects an action a_m, i.e., selecting a (power transmit, SC) pair for its transmission using ϵGP or BAP method;
    Take action a_m, observe reward according to (9);
    Update Q-value according to (8);
    $i \leftarrow i + 1$;
end

in which $p_v \leq 0$ is the penalty value. For action selection, we investigate the following two methods:

- ϵ-greedy policy (ϵGP): this is a widely-used method for action selection, where device m can select an action randomly with probability ϵ or an action with highest Q-value with probability $1-\epsilon$. The probability ϵ is updated after each learning step as $\epsilon = \delta\epsilon$, where δ ($0 \leq \delta \leq 1$) is the exploration decay coefficient.
- Best action policy (BAP): this policy aims to only select the best action with highest Q-value in each learning step, where if there are many actions with the same highest Q-value, device m can choose one of these actions randomly.

The proposed mechanism is summarized in Algorithm 1. Specifically, the Q-table of device m is first initialized as a $L \times \hat{K}$ array of zeros, where $\hat{K}$ depends on the NOMA transmission method, i.e., woSDC or wSDC, as depicted in Section II. Using ϵGP or BAP methods, device m can select one SC and a TPL for its transmission. It then updates the respective Q-value based on (8). After updating the Q-value, another TPL and SC selection is performed in the next transmission. This learning process ends when each device finds the best TPL and SC to communicate with the BS.

IV. Simulation Results and Discussions

The simulation results are provided in this section to evaluate our proposed algorithm's performance. For QL setting, we select $\alpha = 0.1$, $\gamma = 0.5$, $p_v = -1$, $\delta = 0.95$, and the initial probability for ϵGP method is one, i.e., $\epsilon = 1$. In addition, the simulation parameters are set as $r_0 = 100$ m, $K = 100$, the SC bandwidth $W_{\mathsf{SC}} = 180$ KHz, $K' = 20$, $\mathcal{P} = \{0.1; 0.25; 0.5\}$ W, $\theta = 3$, $\sigma^2 = -174$ dBm, $\eta = 0.05$. The rate threshold is set as 1 bps/Hz for all transmission, i.e., $R_{th} = 1$ bps/Hz while D_{max} is selected in $\{0.5; 1; 2; 2.5; 3\}$ ms, and $\varepsilon_0 \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$.

In Fig. 2, we plot the AE versus the number of devices M. We investigate seven different access approaches including the proposed QL method with wSDC and BAP (wSDC-BAP-QL), the proposed QL method with wSDC and ϵGP (wSDC-ϵGP-QL), the proposed QL method with woSDC and BAP (woSDC-BAP-QL), Slotted-Aloha (SA), woSDC-GF-NOMA-SA, wSDC-GF-NOMA-SA, and QL-based SA (QL-SA). In SA, the devices randomly select a SC, where a collision occurs if more than one device use the same SC. In woSDC-GF-NOMA-SA and wSDC-GF-NOMA-SA, GF-NOMA scheme is applied for SA method in case of considering NOMA transmission methods woSDC and wSDC, respectively. In QL-SA, the QL algorithm is utilized to reduce the collision for SA method. As can be observed, QL-based frameworks significantly increase the system AE. Furthermore, our proposed methods outperforms the QL-SA in overloaded scenario, i.e., $M > K$. Among QL-based scheme, the wSDC ones (i.e., wSDC-BAP-QL and wSDC-ϵGP-QL) return in the higher AE than the woSDCs do. It is because when GF-NOMA is performed for small clusters separately, severe interference

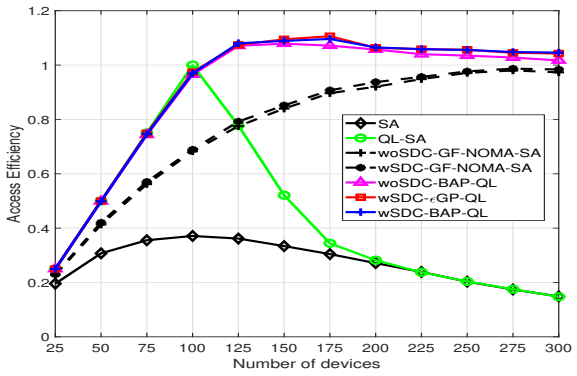

Fig. 2. AE versus number of devices with different RA methods.

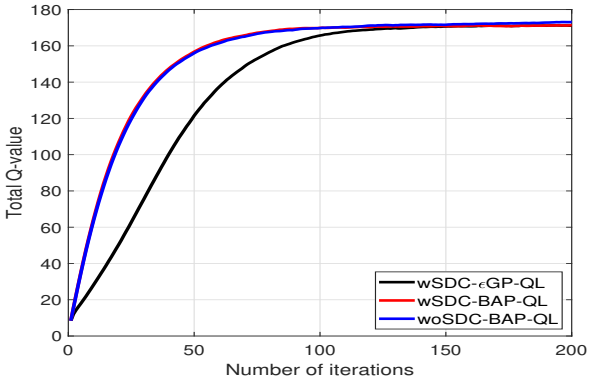

Fig. 3. Convergence of Q-learning with different RA methods.

can be avoided. Fig. 2 also indicates that wSDC-BAP-QL and wSDC-ϵGP-QL achieve the similar AE.

As mentioned earlier in Algorithm 1, the learning process of QL algorithm continues until the Q-value achieves a convergence value such that the devices find unique (transmit power, SC) pairs for their transmissions. To evaluate the convergence of the proposed QL methods, we investigate the variation of the parameter total Q-value of all devices with respect to the number of iterations in Fig. 3. We can see from this figure that although the proposed wSDC-BAP-QL and wSDC-ϵGP-QL methods bring the similar performance in terms of AE, wSDC-BAP-QL achieves the convergence faster than wSDC-ϵGP-QL. Furthermore, this figure indicates that the convergence of woSDC-BAP-QL is similar to that of wSDC-BAP-QL.

To evaluate the DEP and latency on the system performance, we analyze the change of the AE with respect to different reliability and latency thresholds in Figs. 4 and 5, respectively. Here, we use the wSDC-BAP-QL, woSDC-BAP-QL, wSDC-GF-NOMA-SA, and woSDC-GF-NOMA-SA methods for comparison. From these two figures, one can observe that the AE of the system decreases when the requirements on reliability and latency get more stringent (i.e., the reliability increases from 10^{-1} to 10^{-5} and the latency decreases from 3 ms to 0.5 ms). Furthermore, the wSDC-BAP-QL, woSDC-BAP-QL methods using QL achieve the performance better than wSDC-GF-NOMA-SA, and woSDC-GF-NOMA-SA

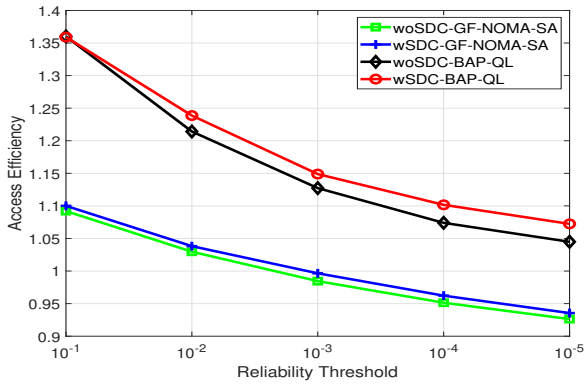

Fig. 4. AE versus reliability threshold with different RA methods.

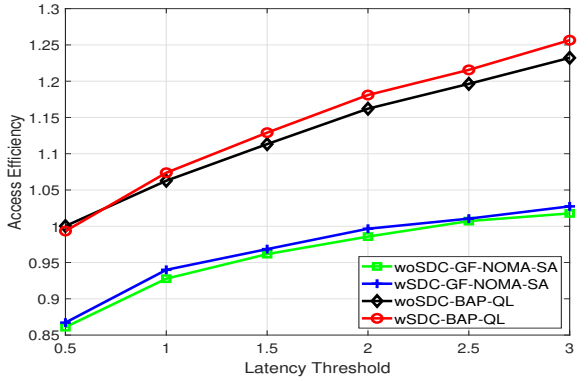

Fig. 5. AE versus latency threshold with different RA methods.

approaches without QL. In addition, these two figures also clarify the benefits of the wSDC-BAP-QL method over the woSDC-BAP-QL scheme, where wSDC-BAP-QL outperforms woSDC-BAP-QL in terms of the AE under different conditions of reliability and latency. From the above four figures, we can conclude that among the proposed QL methods, wSDC-BAP-QL brings the best performance in terms of both AE and convergence.

V. Conclusion

This paper has investigated the RA in and URLLC-enabled GF-NOMA system. To reduce severe interference and improve the AE, a SC and device clustering method has been considered for GF-NOMA transmission. Different QL methods, i.e., wSDC-BAP-QL, wSDC-ϵGP-QL, and woSDC-BAP-QL has been proposed to reduce the interference among devices and enhance the system performance in terms of AE while taking the URLLC requirements into account. Furthermore, two different action selection methods, i.e., BAP and ϵGP has been examined for learning process. Simulation results have indicated that the proposed QL methods brings better performance compared to the other approaches such as SA, woSDC-GF-NOMA-SA, wSDC-GF-NOMA-SA, and QL-SA, in overloaded scenario, i.e., the number of devices is larger than the number of available SCs. Furthermore, among the proposed QL methods, wSDC-BAP-QL achieves the best performance for both AE and convergence.

Acknowledgment

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).

References

[1] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee, and B. Shim, "Ultra-reliable and low-latency communications in 5G downlink: Physical layer aspects," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 124–130, Jun. 2018.

[2] 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.

[3] T. T. Nguyen, V. N. Ha, and L. B. Le, "Wireless scheduling for heterogeneous services with mixed numerology in 5g wireless networks," *IEEE Communications Letters*, vol. 24, no. 2, pp. 410–413, 2020.

[4] V. N. Ha, T. T. Nguyen, L. B. Le, and J.-F. Frigon, "Admission control and network slicing for multi-numerology 5g wireless networks," *IEEE Networking Letters*, vol. 2, no. 1, pp. 5–9, 2020.

[5] D.-D. Tran, S. K. Sharma, S. Chatzinotas, I. Woungang, and B. Ottersten, "Short-packet communications for MIMO NOMA systems over Nakagami-m fading: BLER and minimum blocklength analysis," *IEEE Trans. Veh. Technol.*, vol. 70, no. 4, pp. 3583–3598, 2021.

[6] Q.-V. Pham, F. Fang, V. N. Ha, M. J. Piran, M. Le, L. B. Le, W.-J. Hwang, and Z. Ding, "A survey of multi-access edge computing in 5g and beyond: Fundamentals, technology integration, and state-of-the-art," *IEEE Access*, vol. 8, pp. 116 974–117 017, 2020.

[7] A. C. Cirik, N. M. Balasubramanya, L. Lampe, G. Vos, and S. Bennett, "Toward the standardization of grant-free operation and the associated NOMA strategies in 3GPP," *IEEE Commun. Stand. Mag.*, vol. 3, no. 4, pp. 60–66, Dec. 2019.

[8] M. Shirvanimoghaddam, M. Condoluci, M. Dohler, and S. J. Johnson, "On the fundamental limits of random non-orthogonal multiple access in cellular massive IoT," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2238–2252, Jul. 2017.

[9] R. Abbas, M. Shirvanimoghaddam, Y. Li, and B. Vucetic, "A novel analytical framework for massive grant-free NOMA," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2436–2449, Nov. 2018.

[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.*, vol. 9, no. 10, pp. 1720–1724, Oct. 2020.

[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.

[13] D.-D. Tran, S. K. Sharma, and S. Chatzinotas, "BLER-based adaptive Q-learningfor efficient random access in NOMA-based mMTC networks," in *IEEE Veh. Technol. Conf. (VTC-Spring)*, Helsinki, Finland, Apr. 2021, pp. 1–5.

[14] D.-D. Tran, S. K. Sharma, S. Chatzinotas, and I. Woungang, "Q-learning-based scma for efficient random access in mmtc networks with short packets," in *IEEE Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, Helsinki, Finland, Sep. 2021, pp. 1334–1338.

[15] D. Tran, S. K. Sharma, S. Chatzinotas, and I. Woungang, "Learning-based multiplexing of grant-based and grant-free heterogeneous services with short packets," in *IEEE Global Commun. Conf. (GLOBECOM)*, Madrid, Spain, Dec. 2021, pp. 1–6.

[16] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static multiple antenna fading channels at finite blocklength," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4232–4265, Jul. 2014.

[17] 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.