Deep Reinforcement Learning-Based Grant-Free NOMA Optimization for mURLLC

Yan Liu, *Member, IEEE*, Yansha Deng, *Senior Member, IEEE*, Hui Zhou, *Member, IEEE*, Maged Elkashlan, *Senior Member, IEEE*, and Arumugam Nallanathan, *Fellow, IEEE*

Abstract—Grant-free non-orthogonal multiple access (GF-NOMA) is a potential technique to support massive Ultra-Reliable and Low-Latency Communication (mURLLC) service. However, the dynamic resource configuration in GF-NOMA systems is challenging due to random traffics and collisions, that are unknown at the base station (BS). Meanwhile, joint consideration of the latency and reliability requirements makes the resource configuration of GF-NOMA for mURLLC more complex. To address this problem, we develop a novel learning framework for signature-based GF-NOMA in mURLLC service taking into account the multiple access signature collision, the UE detection, as well as the data decoding procedures for the K-repetition GF and the Proactive GF schemes. The goal of our learning framework is to maximize the long-term average number of successfully served users (UEs) under the latency constraint. We first perform a real-time repetition value configuration based on a double deep Q-Network (DDQN) and then propose a Cooperative Multi-Agent learning technique based DQN (CMA-DQN) to optimize the configuration of both the repetition values and the contention-transmission unit (CTU) numbers. Our results show the superior performance of CMA-DQN over the conventional load estimation-based uplink resource configuration approach (LE-URC) in heavy traffic and demonstrate its capability in dynamically configuring in long term for mURLLC service. In addition, with our learning optimization, the Proactive scheme always outperforms the K-repetition scheme in terms of the number of successfully served UEs, especially under the high backlog traffic scenario.

Yan Liu was with the Key Laboratory of Ministry of Education in Broadband Wireless Communication and Sensor Network Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China, and also with the School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS London, U.K. She is now with the College of Electronic and Information Engineering, Tongji University, Shanghai 201804, China (e-mail: yanliu2022@tongji.edu.cn).

Yansha Deng is with the Department of Engineering, King's College London, WC2R 2LS London, U.K. (e-mail: yansha.deng@kcl.ac.uk).

Hui Zhou is with Huawei Technology Company Ltd., Shanghai 200121, China (e-mail: zhouhui96@huawei.com).

Maged Elkashlan and Arumugam Nallanathan are with the School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS London, U.K. (e-mail: maged.elkashlan@qmul.ac.uk; a.nallanathan@qmul.ac.uk).

I. INTRODUCTION

AS A NEW and dominating service class in 6th Generation (6G) networks, massive Ultra-Reliable and Low Latency Communications (mURLLC) integrates URLLC with massive access to support massive short-packet data communications in time-sensitive wireless networks with high reliability and low access latency [2]. This requires a reliability-latency-scalability trade-off and mandates a principled and scalable framework accounting for the delay, reliability, and decision-making under uncertainty [3]. Concretely speaking, the Third Generation Partnership Project (3GPP) standard [4], [5] has defined a general URLLC requirement as: $1 - 10^{-5}$ reliability within 1ms user plane latency for 32 bytes. It is also anticipated that the device density may grow to hundred(s) of devices per cubic meter in the 6G white paper [6].

Current cellular network can hardly fulfill the joint massive connectivity, ultra-reliability, and low latency requirements in mURLLC service. To achieve low latency, *grant-free (GF) access* has been proposed [7], [8] as an alternative for traditional grant-based (GB) access due to its drawbacks in high latency and heavy signaling overhead [9]. Different from GB access, GF access allows a User Equipment (UE) to transmit its data to the Base Station (BS) in an arrive-and-go manner, without sending a scheduling request (SR) and obtaining a resource grant (RG) from the network [10]. To achieve high reliability, several GF schemes, including the *K-repetition* scheme and the *Proactive* scheme, have been proposed, where a pre-defined number (K) of consecutive replicas of the same packet are transmitted [11], [12]. To achieve massive connectivity, *non-orthogonal multiple access (NOMA)* has been proposed to synergize with GF in order to deal with the multiple access (MA) physical resource collision of contention-based[1] GF access. The signature-based GF-NOMA transmission has been proposed and discussed during 3GPP Release-14 NR Study, where NOMA signatures (e.g, codebook, pilot sequence, mapping pattern, demodulation reference signal, power, etc.) are taken as part of GF resource

---

[1]In UL GF access, the MA resources can be reserved in advance or allocated when there is a request. Preallocation of the dedicated resource is more suitable for periodic traffic with a fixed pattern, whereas contention-based GF access is more suitable for sporadic packets, as it is more efficient and flexible in terms of resource utilization [13]. Unless otherwise stated, the GF and GB access described in this work are both contention-based.

except from the traditional MA physical resource [14]. Prior to transmission, a user can randomly select one signature from a given resource pool. Then in each contention region (the basic unit of MA physical resource for GF), multiple NOMA signatures from different users will be multiplexed.

A. State-of-the-Art

In terms of analysis, a novel GF-NOMA strategy was proposed in [15], in which active devices transmitted data over a randomly selected available channel. In [16], a general GF-NOMA analytical framework was proposed to analyze the outage probability, where successive interference cancellation (SIC) are considered by treating collisions as interference. In [17], [18], [19], [20], and [21], GF-NOMA is designed empirically by directly incorporating the GF mechanism into several state-of-the-art NOMA schemes, including SCMA, MUSA, and PDMA, that are categorized according to their specially designed spreading signatures. The authors in [21] proposed a message passing algorithm to solve the problem of GF-NOMA using CS-based approaches, which improves the bit error rate (BER) performance in comparison to [22].

In terms of optimization, Machine Learning (ML), especially Reinforcement Learning (RL), are common optimization tools utilized in many works [23], [24], including Q-learning (Offline) and SARSA (Online) [25]. However, RL is inapplicable to large-scale networks as it has to explore and gain knowledge of an entire system and takes a lot of time to reach the best policy. Recently, deep learning has been introduced as a new breakthrough technique to overcome the limitations of RL, namely Deep Reinforcement Learning (DRL), including Policy Gradient (PG), Actor-Critic (AC), and Deep Q-Network (DQN) for discrete action space and the Deep Deterministic Policy Gradients (DDPG) for continuous action space [26]. Deep learning [27] and deep multi-task learning [28] werw used to solve optimization problem for GF-NOMA. However, these works assumed that each UE is pre-allocated with a unique sequence, and thus collisions are not an issue. This assumption does not hold in massive UEs settings in mURLLC, where the collision is the bottleneck of the GF-NOMA performance.

Different from these works, we aim to develop a general learning framework to optimize GF-NOMA systems for mURLLC service taking into account the MA signature collision, the UE detection as well as the data decoding procedures. Note that part of this work was presented in [1]. However, for simplicity, work [1] only took into account the K-repetition scheme without the Proactive scheme which leads to more complicated modeling, analysis, and simulations.

B. Motivations and Contributions

It is important to know that the research challenges in GF-NOMA are fundamentally different from those in GB-NOMA [29], [30]. In GB scheme, the four-step random access (RA) procedure as shown in Fig. 1 is executed by the UE to request the BS to schedule dedicated resources
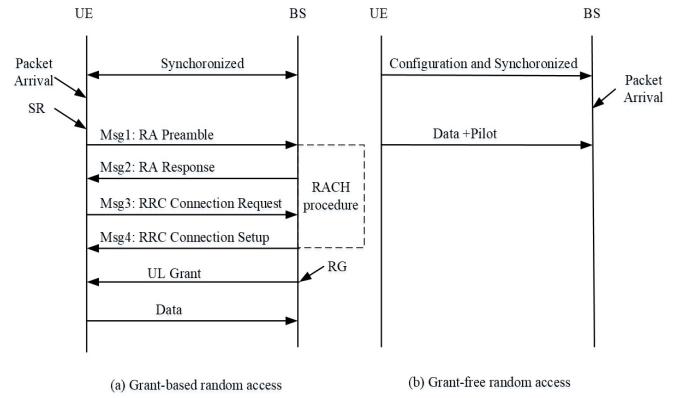

Fig. 1.   Uplink transmissions for grant-based and grant-free random access.

for data transmission, where the data transmission is more likely to be successful once the random access succeeds. While in GF scheme, the data is transmitted along with the pilot in a randomly chosen MA resource, which is unknown at the BS, and can lead to new research problems, especially for the GF-NOMA system, including: 1) the set of active users and their respective channel conditions are unknown to the BS, which prohibits the pre-configuration and the pre-assignment of resources, including pilots/preambles, power, repetition values, and etc; 2) simultaneously satisfy the reliability and latency requirements under random traffics, the optimal parameter configurations vary over different time slots, which is hard to be described by a tractable mathematical model; 3) the MA signature collision detection and the blind UE activity detection, as well as the data decoding, need to be considered, which largely impacts the resource configuration in each time slot; 4) a general optimization framework for GF-NOMA systems have never been established for various signature-based NOMA schemes.

The above mentioned challenges can hardly be solved via the traditional convex optimization method, due to the complex communication environment with the lack of tractable mathematical formulations. The complexity of the problem is compounded by the lack of prior knowledge at the BS regarding the stochastic traffic and unobservable channel statistics (i.e., random collision, and effects of physical radio including path-loss as well as fading). In the GF-NOMA system, the BS can only observe the results of both collision detection (e.g., the number of non-collision UEs and collision MA signatures) and data decoding (e.g., the number of successful decoding UEs and failure decoding UEs) in each round trip time. This historical information can be potentially used to facilitate the long-term optimization of future configurations. Even if one knew all the relevant statistics, tackling this problem in an exact manner would result in a Partially Observable Markov Decision Process (POMDP) with large state and action spaces, which is generally intractable. To deal with it, RL can be a promising tool to deal with this complex POMDP problem due to that it solely relies on the self-learning of the environment interaction, without the need to derive explicit optimization solutions based on a complex mathematical model.
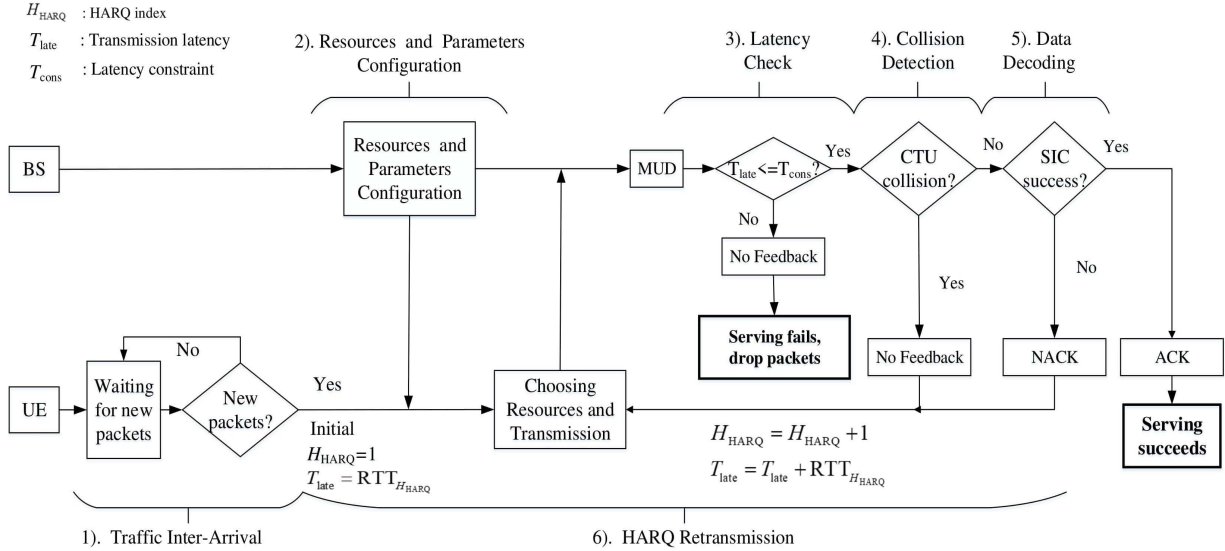
Fig. 2. Uplink GF-NOMA transmission procedure.

Our contributions can be summarized as follows:

1) We develop a general learning framework for dynamic resource configuration optimization in signature-based GF-NOMA systems for mURLLC services. In our framework, we practically simulate the random traffics, the resource selection and configuration, the transmission latency check, the collision detection, the data decoding, and the Hybrid Automatic Repeat reQuest (HARQ) retransmission procedures in this GF-NOMA system. We use this generated simulation environment to train the RL agents.

1) We first perform the repetition values dynamic optimization via developing a double Deep Q-Network (DDQN) to optimize the number of successfully served UEs under the latency constraint for the K-repetition GF scheme and the Proactive GF scheme, respectively. We then develop a Cooperative Multi-Agent learning based on DQN (CMA-DQN) to dynamically optimize both the repetition values and contention-transmission unit (CTU) numbers, which breaks down the selection in high-dimensional parameters into multiple parallel sub-tasks with a number of DQN agents cooperatively being trained to produce each parameter.

1) Through our developed learning framework, we show that the Proactive scheme outperforms the K-repetition scheme in terms of the number of successfully served UEs, especially under the high backlog traffic scenario, which are opposite to the results without optimization in our previous work [31] with only a single packet transmission. Our results also show the superior performance of CMA-DQN over the conventional load estimation-based approach (LE-URC) in heavy traffic scenarios. Our general learning framework can be extended to optimize other resource configuration problems in GF-NOMA schemes.

C. Organization

The rest of the paper is organized as follows. Section II illustrates the system model and formulates the problem. Section III illustrates the preliminary and the conventional approach. Section IV proposes Q-learning based uplink GF-NOMA resource configuration approaches. Section V elaborates the numerical results, and finally, Section VI concludes the work.

II. SYSTEM MODEL AND PROBLEM FORMULATION

We consider a single cell network consisting of a BS located at the center and a set of N UEs randomly located in an area of the plane $\mathbb{R}^2$, where the UEs are unaware of the status of each other. Once deployed, the UEs remain spatially static. The time is divided into short transmission time intervals (TTIs), and the small packets for each UE are generated according to random inter-arrival processes over the short-TTIs, which are Markovian as defined in [32] and [33] and unknown to BS. In this paper, the TTI refers to a mini-slot. The Fifth Generation (5G) New Radio (NR) introduces the concept of 'mini-slots' and supports a scalable numerology allowing the sub-carrier spacing (SCS) to be expanded up to 240 kHz. In contrast with the LTE slot consisting of 14 symbols per TTI, the number of symbols in 5G NR mini-slots ranges from 1 to 13 symbols, and the larger SCS decreases the length of each symbol further. Collectively, mini-slots and flexible numerology allow shorter slots to meet the stringent latency requirement.

A. GF-NOMA Network Model

We consider the uplink contention-based GF-NOMA over a set of preconfigured MA resources. To capture the effects of the physical radio, we consider the standard power-law path-loss model with the path-loss attenuation $r^{-\eta}$, where r is the Euclidean distance between the UE and the BS and

η is the path-loss attenuation factor. We consider a Rayleigh flat-fading environment, where the channel power gains h are exponentially distributed (i.i.d.) random variables with unit mean. We present the uplink GF-NOMA procedure in Fig. 2 following the 3GPP standard [14], [33], [34], [35], which includes 1) traffic inter-arrival, 2) resources and parameters configuration, 3) latency check; 4) collision detection, 5) data decoding, and 6) HARQ retransmissions. These six stages are explained in details in the following six subsections to illustrate the system model.

1) Traffic Inter-Arrival: We consider delay-sensitive URLLC applications in sensors-based IoT networks, which is appropriate for a scenario that a large amount of IoT devices access the network in a highly synchronized manner, e.g., triggered due to an emergency event (earthquake alarm, power outage, and fire alarms) [32]. This kind of time-varied traffic has been ignored in most works, where devices were assumed to have saturated data (i.e., the device essentially always has data to be transmitted) for simplicity. According to the 3GPP standard [33], the Beta distribution-based arrival process is recommended to model the arrival intensity during bursty traffic arrivals for a large number of UEs attempting to access the same network simultaneously during a short period of time. In this condition, the configured devices are connected and synchronized to the cell, being always ready for transmission once data arrived. More details can be found in [36]. Each device would be activated (having data arrived) at any time τ, according to a time-limited Beta probability density function as [33, Section 6.1.1]

$$p(\tau) = \frac{\tau^{\alpha-1}(T-\tau)^{\beta-1}}{T^{\alpha+\beta-1}\mathrm{Beta}(\alpha,\beta)}, \quad (1)$$

where T is the total time of the bursty traffic and Beta (α,β) is the Beta function with the constant parameters α and β [37].

Due to the nature of slotted-Aloha, a UE can only transmit at the beginning of a round trip time (RTT) as shown in Fig. 3 and Fig. 4, and the UE needs to wait for the feedback before performing retransmission, which is determined by the RTT, i.e., the time duration of the cycle from the beginning of the transmission until processing its feedback [38]. Thus, the newly activated UEs executing transmission comes from those who received a packet within the interval between the last RTT period (τ^{i-1}, τ^i). The traffic instantaneous rate in packets in a period is described by a function $p(\tau)$, so that the packets arrival rate in the ith RTT is given by

$$\mu^i = \int_{\tau_{i-1}}^{\tau_i} p(\tau)d\tau. \quad (2)$$

2) Resources and Parameters Configuration: The MA resources, repetition values, and HARQ related parameters, etc, are configured at the BS by radio resource control (RRC) signaling and L1 signaling prior to the GF access (as Type 2 GF [39]).

a) Repetition values: We consider the K-repetition GF scheme and the Proactive GF scheme as shown in Fig. 3 and Fig. 4, respectively. The repetition values for K-repetition scheme K_{Krep}^t and for Proactive scheme K_{Proa}^t are configured at the beginning of each RTT.
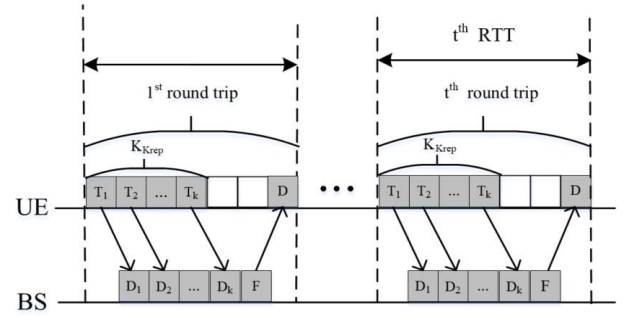

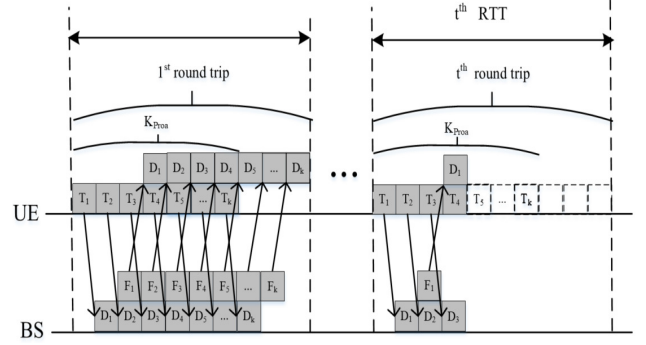Fig. 3. K-repetition GF transmission.

Fig. 4. Proactive GF transmission.

- **K-repetition scheme:** The K-repetition scheme is illustrated in Fig. 3, the UEs served by the BS are configured to autonomously transmit (T) the same packet for K_{Krep}^t repetitions in consecutive TTIs. The BS decodes (D) each repetition independently and the transmission in one RTT is successful when at least one repetition succeeds. After processing all the received K_{Krep}^T repetitions, the BS transmits the ACK/NACK feedback (F) to the UE.
- **Proactive scheme:** The Proactive scheme is illustrated in Fig. 4. Similar to the K-repetition scheme, the UEs served by the BS are configured to repeat the transmission for a maximum number of K_{Proa}^t repetitions, but can receive the feedback after each repetition. This allows the UE to terminate repetitions earlier once receiving the ACK.

Considering the small packets of mURLLC traffic, we set the packet transmission time T_{tx} as one TTI. The BS feedback time T_{fb} and the BS (UE) processing time T_{dp} and T_{up} are also assumed to be one TTI in this work same as our previous work [31]. We use TTI as a unit, but the TTI duration varies according to different subcarrier spacing. Once the repetition value is configured, the duration of one RTT is known to the UEs and the BS, which is given as

$$T_{\mathrm{RTT}}^t = K^t T_{\mathrm{tx}} + T_{\mathrm{dp}} + T_{\mathrm{fb}} + T_{\mathrm{up}} = (K^t + 3)\mathrm{TTIs}, \quad (3)$$

with $K^t = K_{\mathrm{Krep}}^t$ ($K^t = K_{\mathrm{Proa}}^t$) for the K-repetition scheme (the Proactive scheme).

b) MA resources: A *contention-transmission unit* (CTU) as shown in Fig. 5 is defined as the basic MA resource, where
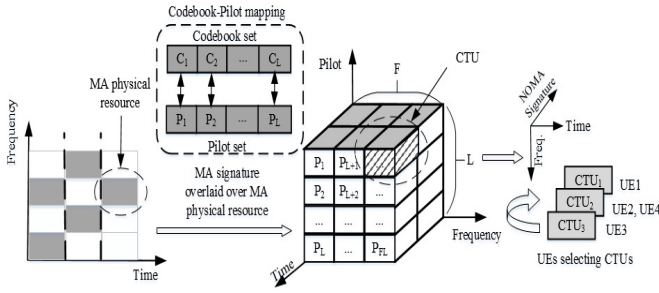
Fig. 5. GF-NOMA resource.

Collision CTU set:{1, 2, 4}
Singleton CTU set:{5, 6}
Idle CTU set:{3}

Interference UE set in RB 1 :{2, 3, 4, 7}
Interference UE set in RB 2 :{1, 5, 6, 8}
Collision UE set in RB 2 :{6, 8}
Singleton UE set in RB 2:{1, 5}

Fig. 6. Detection and Decoding in a network with L = 2 RBs, C = 6 CTUs and N = 8 UEs.

each CTU may comprise of a MA physical resource and a MA signature [40], [41]. The MA physical resources represent a set of time-frequency resource blocks (RBs). The MA signatures represent a set of pilot sequences for channel estimation and/or UE activity detection, and a set of codebooks for robust data transmission and interference whitening, etc. The receiver can estimate channels of different UEs with different pilots. But when two or more UEs transmit their data using the same MA physical resource and the same pilot sequence, the pilot collision occurs [35]. In this condition, not only the UE cannot be identified, but also the data can not be recovered or canceled since the channel condition is unknown. When two or more UEs transmit their data sharing the same MA physical resource using different pilot sequences but the same codebook, the codebook collision occurs. In this condition, the detector can still decoded these UEs' data carried over the same codebook, as long as different pilot sequences are used [18]. Although a one-to-one mapping or a many-to-one mapping between the pilot sequences and codebooks can be predefined, since it has been verified in [17] that the performance loss due to codebook collision is negligible for a real system, we focus on the pilot sequence collision and consider the one-to-one mapping as [16]. Without loss of generality, in one TTI, we consider F orthogonal RBs and each RB is overlaid with L unique codebook-pilot pairs [14]. Thus, at the beginning of each RTT, the BS configures a resource pool of $C^t = F \times L$ unique CTUs, and each UE randomly choose one CTU from the pool to transmit in this RTT.

3) Latency Check: The HARQ index H_{HARQ} is included in the pilot sequence and can be detected by the BS. At the beginning of each RTT, the HARQ index and the transmission latency T_{late} will be updated as shown in Fig. 2. For example, for the initial RTT with initial K^1, $H_{\text{HARQ}} = 1$ and $T_{\text{late}} = RTT_{H_{\text{HARQ}}=1}$, where $RTT_{H_{\text{HARQ}}}$ is calculated by using (3). After this round time trip transmission, the BS optimizes a K^2 based on the observation of the reception and configures it to the UE for the next RTT. Then, the UE updates its $H_{\text{HARQ}} = 2$ and calculated $RTT_{H_{\text{HARQ}}=2}$ by using (3) with K^2, and consequently, the transmission latency T_{late} is updated as $T_{\text{late}} = RTT_{H_{\text{HARQ}}=1} + RTT_{H_{\text{HARQ}}=2}$. When $T_{\text{late}} > T_{\text{cons}}$, the UE fails to be served and the packets will be dropped. Note that the HARQ index, as well as the transmission latency, will be updated at the beginning of each RTT instead of at the end of each RTT due to that we consider the user plane latency in this work. User plane
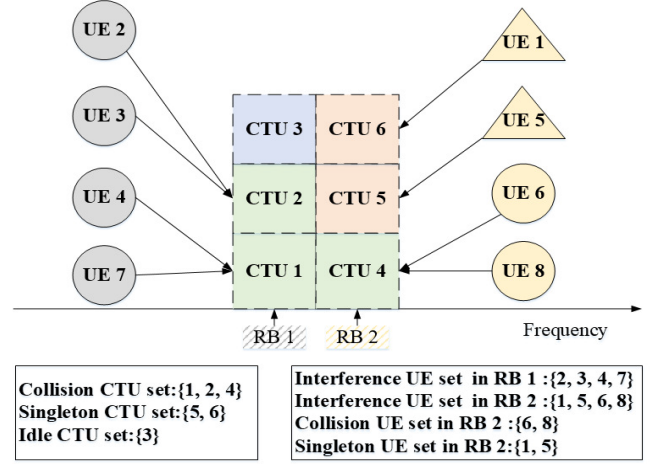
latency is defined as the one-way latency from the processing of the packet at the transmitter to the successful reception of the packet, including the transmission processing time, the transmission time, and the reception processing time. That is to say, from the UE perspective, when the UE executes this RTT, it will check transmission results after finishing the RTT. Thus, the duration of this RTT should be included when calculating the UE transmission latency.

4) Collision Detection: At each RTT, each active UE transmits its packets to the BS by randomly choosing a CTU. The BS can detect the UEs that have chosen different CTUs. However, if multiple UEs choose the same CTU, the BS cannot differentiate these UEs and therefore cannot decode the data. We categorize the CTUs into three types: an *idle* CTU is a CTU which has not been chosen by any UE; a *singleton* CTU is a CTU chosen by only one UE; and a *collision* CTU is a CTU chosen by two or more UEs [16]. One example is illustrated in Fig. 6. The UE 1 and UE 5 have chosen the unique CTU 6 and CTU 5, respectively, thus, the CTU 6 and 5 are singleton CTUs. The CTU 3 is an idle CTU. The UE 4 and UE 7 have chosen the CTU 1, the UE 2 and UE 3 have chosen the CTU 2, and the UE 6 and UE 8 have chosen the CTU 4, thus, CTU 1, 2 and 4 are collision CTUs. After collision detection at the tth RTT, the BS observes the set of singleton CTUs $\mathcal{C}_{\text{sc}}^t$, the set of idle CTUs $\mathcal{C}_{\text{ic}}^t$, and the set of collision CTUs $\mathcal{C}_{\text{cc}}^t$ as shown in orange, blue and green color, respectively, in Fig. 6.

5) Data Decoding: After detecting the UEs that have chosen the singleton CTUs (e.g., UE 1 and UE 5 shown as triangle in Fig. 6), the BS performs the successive interference cancellation (SIC) technique to decode the data of these UEs. During the decoding, the UEs that transmit in different RBs do not interfere with each other due to the orthogonality, and only UEs that transmit in the same RB cause interference, i.e., as shown in Fig. 6, the interference UE set in RB 1 is $\{2, 3, 4, 7\}$ shown in color grey and the interference UE set in RB 2 is $\{1, 5, 6, 8\}$ shown in color yellow. For SIC, the detection and decoding can be implemented through a simple algorithm. At first, the BS assumes that all the CTUs

are singleton CTUs and orders the CTUs according to their respective received powers in descending order. In the first iteration, the BS attempts to decode the strongest CTU by treating the received powers of other CTUs over the same RB as the interference. The iterative stage of SIC decoding is successful when the signal-to-interference-plus-noise ratio (SINR) in that stage is larger than the SINR threshold. If decoding is successful, the decoded signal is subtracted from the received signal.[2] If not, the BS assumes that this is a collision CTU. In the second iteration, the BS attempts to decode the second strongest CTU while regarding the previously undecoded CTU as interference. The BS can continue to follow the same steps until there are no more CTUs to decode. Thus, in the kth repetition of the tth RTT, the sth stage of SIC decoding is successful if the SINR is higher than a threshold γ_{th} [16], i.e.,

$$\text{SINR}_{f,s}^t(k)$$

$$= \frac{P h_{s,k} r_s^{-\eta}}{\sum_{m=s+1}^{N_{f,\text{sc}}^t(k)} P_m h_{m,k} r_m^{-\eta} + \sum_{n' \in \mathcal{N}_{f,\text{cc}}^t(k)} P_{n'} h_{n',k} r_{n'}^{-\eta} + \sigma^2}$$

$$\geq \gamma_{\text{th}}, \tag{4}$$

where P is the transmission power, $\mathcal{N}_{f,\text{sc}}^t$ is the set of other devices that have chosen the singleton CTUs over the fth RB, $\mathcal{N}_{f,\text{cc}}^t$ is the set of devices that have chosen the collision CTUs over the fth RB, σ^2 is the noise power.

The SIC decoding procedure for each GF scheme is described in the following.

i) K-repetition scheme: For the K-repetition scheme as shown in Fig. 3, the successful decoding event occurs at least one repetition decoding succeeds. Thus, the SIC decoding procedure follows:

1) Step 1: Start the kth repetition with the initial $k = 1$, $\mathcal{N}_{f,\text{sc}}^t$ and $\mathcal{N}_{f,\text{cc}}^t$;
2) Step 2: Decode the sth CTU with the initial $s = 1$ using (4);
3) Step 3: If the sth CTU is successfully decoded, put the decoded UE in set $\mathcal{N}_{f,\text{sd}}^t(k)$ and go to Step 4, otherwise go to Step 5;
4) Step 4: If $s \leq N_{f,\text{sc}}^t$, do $s = s + 1$, go to Step 2, otherwise go to Step 5;
5) Step 5: SIC for the kth repetition stops;
6) Step 6: If $k \leq K_{\text{Krep}}$, do $k = k + 1$, go to Step 1, otherwise go to the end.

ii) Proactive scheme: For the Proactive scheme as shown in Fig. 4, the successful decoding event occurs once the repetition decoding succeeds. The successfully decoded UEs will not transmit in the remaining repetitions in this RTT to reduce interference to other UEs. It should be noted that the ACK/NACK feedback can only be received after 3TTIs, which means the ACK feedback of the kth successful repetition can be received by the UE in the $(k+3)$th repetition and the UE stops transmission from the $(k+4)$th repetition. In addition, the

BS does not send any ACK/NACK feedback to the collision UEs. The collision USs in the kth repetition that do not receive feedback at the pre-defined timing after the UEs sent the packet (e.g., after 3TTIs) will not transmit in the remaining repetitions to reduce interference to other UEs.

1) Step 1: Initialize $k = 1$, $\mathcal{N}_{f,\text{sc}}^t$ and $\mathcal{N}_{f,\text{cc}}^t$. If $k < 4$, go to Step 3, otherwise go to Step 2;
2) Step 2: Update the $\mathcal{N}_{f,\text{sc}}^t(k) = \mathcal{N}_{f,\text{sc}}^t(k) - \mathcal{N}_{f,\text{sd}}^t(k-4)$ and $\mathcal{N}_{f,\text{cc}}^t(k) = \varnothing$;
3) Step 3: Start the kth repetition with k, $\mathcal{N}_{f,\text{sc}}^t(k)$ and $\mathcal{N}_{f,\text{cc}}^t(k)$;
4) Step 4: Decode the sth CTU with initial $s = 1$ using (4);
5) Step 5: If the sth CTU is successfully decoded, put the decoded UE in set $\mathcal{N}_{f,\text{sd}}^t(k)$ and go to Step 6, otherwise go to Step 7;
6) Step 6: If $s \leq N_{f,\text{sc}}^t$, do $s = s + 1$, go to Step 4, otherwise go to Step 7;
7) Step 7: SIC for the kth repetition stops;
8) Step 8: If $k \leq K_{\text{Pro}}$, do $k = k + 1$, go to Step 1, otherwise go to the end.

Finally, the set $\mathcal{N}_{f,\text{sd}}^t = \bigcup_{k=1}^{K_{Krep}} (\mathcal{N}_{f,\text{sd}}^t(k))$ is the successfully decoded UEs over the fth RB and $\mathcal{N}_{\text{sd}}^t = \bigcup_{f=1}^{F^t} (\mathcal{N}_{f,\text{sd}}^t)$ is the successfully decoded UEs in the tth RTT.

6) HARQ Retransmissions: We take into account the GF-NOMA HARQ retransmissions to achieve high reliability performance. However, due to the latency constraint T_{cons}, the HARQ retransmission times are limited as shown in Fig. 2. The UE determines a re-transmission or not based on the following two different scenarios.

i) when the UE receives an ACK from the BS, it means that the BS successfully detected the UE (i.e., the UE chooses the singleton CTUs) and decoded the UE's data (i.e., SIC succeeds), no further re-transmission is needed;

ii) when the UE receives a NACK from the BS, it means that the BS successfully detected the UE but failed to decode the UE's data (i.e., SIC fails). Otherwise, when the UE does not receive any feedback at the pre-defined timing after the UE sent the packet (e.g., at the end of one RTT), it means the BS failed to identify the UE, the UE determines whether to retransmit or not in the next RTT based on the transmission latency check as shown in Fig. 2.

B. Problem Formulation

Once actived in a given RTT t, a UE executes the GF-NOMA procedure, where the UE randomly chooses one of the preconfigured C^t CTUs to transmit its packets for K_{Krep}^t times or $k_{\text{Proa}}^t \leq K_{\text{Proa}}^t$ times under the K-repetition scheme and the Proactive scheme, respectively. During this RTT, the GF-NOMA fails if: *(i)* a CTU collision occurs when two or more UEs choose the same CTU (i.e., UE detection fails); or *(ii)* the SIC decoding fails (i.e., data decoding fails). Once failed, UEs decides whether to retransmit in the following RTT or not based on the transmission latency check. When $T_{\text{late}} > T_{\text{cons}}$, the UE fails to be served and its packets will be dropped. It is obvious that 1) increasing the repetition values K^t could improve the GF-NOMA success probability, but

---

[2] We assume perfect SIC with no error propagation between iterations, as in existing works like [16]. Our proposed DRL-based GF-NOMA framework can also be extended to systems with imperfect SIC cases without affecting the implement of DRL algorithms [42].

results in an increasing latency; 2) increasing CTU numbers C^t could improve the UE detection success probability, but it results in low resource utilization efficiency.

Thus, it is necessary to tackle the problem of optimizing the GF-NOMA configuration defined by parameters[3] $A^t = \{K^t, C^t\}$ for each RTT t under both the K-repetition scheme and the Proactive scheme, where K^t is the repetition value and C^t is the number of CTUs. At the beginning of each RTT t, the decision is made by the BS according to the transmission receptions $U^{t'}$ for all prior RTTs ($t' = 1, \ldots, t-1$), consisting of the following variables: the number of the collision CTUs $V_{\text{cc}}^{t'}$, the number of the idle CTUs $V_{\text{ic}}^{t'}$, the number of the singleton CTUs (non-collision UEs) $V_{\text{sc}}^{t'}$, the number of successfully served UEs (UEs that have been successfully detected and decoded) $V_{\text{sd}}^{t'}$, and the number of failure decoding UEs (UEs that have been successfully detected but not successfully decoded) $V_{\text{ud}}^{t'}$. We denote $H^t = \{O^1, O^2, \ldots, O^{t-1}\}$ with $O^{t-1} = \{U^{t-1}, A^{t-1}\}$ as the observation in each RTT t including histories of all such measurements and past actions.

At each RRT t, the BS aims at maximizing a long-term objective R_t related to the number of successfully transmitted UEs under the latency constraint $V_{\text{sd}}^{t'}$ with respect to the stochastic policy π that maps the current observation history O^t to the probabilities of selecting each possible parameters in A^t. It should be noted that though block error rate (BLER) (the ratio of the number of erroneous blocks to the total number of blocks) is a very important metric for the reliability performance of short-packet transmission [43], which is a statistical metric considered in statistical mathematical methods. While in our learning optimization framework, the RL approaches can only optimize the long-term performance metrics, such as the total number of success packets or devices, but not the statistical probability, like BLER. Actually, maximizing the number of successfully served UEs has the same performance meaning as minimizing the BLER, i.e., to increase network reliability. The optimization problem (P1) is formulated as:

$$(\text{P1}:) \quad \max_{\pi(A^t|O^t)} \quad \sum_{k=t}^{\infty} \gamma^{k-t} \mathbb{E}_\pi[V_{\text{sd}}^k] \qquad (5)$$

$$s.t. \quad T_{\text{late}} \le T_{\text{cons}}, \qquad (6)$$

where $\gamma \in [0, 1)$ is the discount factor for the performance accrued in the future RTTs, and $\gamma = 0$ means that the agent just concerns the immediate reward.

Since the dynamics of the GF-NOMA system is Markovian over the continuous RRTs, this is a Partially Observable Markov Decision Process (POMDP) problem which is generally intractable. Here, the partial observation refers to that a BS can not fully know all the information of the communication environment, including, but not limited to, the channel conditions, the UE transmission latency, the random collision process, and the traffic statistics. Approximate solutions will be discussed in Section III and VI.

---

[3]According to the UE detection and data decoding procedure described in Section II.A, for the same CTU number C^t, a large RB number F^t leads to fewer UEs in each RB, which increases the data decoding success probability. That is to say, the larger RB number, the better. Thus, we fix the RB number $F = 4$ in this work to optimize the CTU number.

III. PRELIMINARIES AND CONVENTIONAL SOLUTIONS

The optimization problem (P1) is really complicated, which cannot be easily solved via the conventional uplink resource optimization solutions, especially the dynamic optimization taking into account the latency constraint. In addition, most prior works simplified the optimization without consideration of future performance [44]. We modify the load estimation (LE) approach given in [44] via estimating based on the last number of the collision CTUs V_{cc}^{t-1} and the previous numbers of idle CTUs $V_{\text{ic}}^{t-1}, V_{\text{ic}}^{t-2}, \cdots, V_{\text{ic}}^1$. To simplify, we propose a load estimation-based uplink resource configuration (LE-URC) approach to dynamically configure the CTUs number C^t with the fixed repetition value K^t in each RTT to maximize the successfully served UEs without latency check and SIC procedure described in Section III, which is expressed as

$$(\text{P2}:) \quad \max_{\pi(C^t|O^t)} \mathbb{E}_\pi[V_{\text{sc}}^t], \qquad (7)$$

In the conventional solution, we ignore the SIC detection failure, i.e., UEs are successfully transmitted if there is no CTU collision occurs. Then the number of non-collided UEs is regarded as the upper bound of the number of successfully served UEs and is utilized as a baseline to compare with our proposed learning algorithm. Thus, V_{sc}^t is the optimization objective.

At the RTT $t-1$ we consider that $D_{\text{UE}}^{t-1} = n$ UEs randomly choose one of the available C^{t-1} CTUs with an equal probability $1/C^{t-1}$. The probability that no UE chooses a CTU c is

$$\mathbb{P}(D_c = 0|D_{\text{UE}}^{t-1} = n) = (1 - 1/C^{t-1})^n. \qquad (8)$$

The expected number of idle CTUs is given by

$$\mathbb{E}[V_{\text{ic}}^{t-1}|D_{\text{UE}}^{t-1} = n]$$
$$= \sum_{c=1}^{C^{t-1}} \mathbb{P}(D_c = 0|D_{\text{UE}}^{t-1} = n) = C^{t-1}(1 - 1/C^{t-1})^n. \qquad (9)$$

Due to that the actual number of idle CTUs V_{ic}^{t-1} can be observed at the BS, the number of active UEs in the $(t-1)$th RTT is estimated as

$$\tilde{D}_{\text{UE}}^{t-1} = f^{-1}(\mathbb{E}[V_{\text{ic}}^{t-1}|D_{\text{UE}}^{t-1} = n])$$
$$= \log_{(1-1/C^{t-1})}(V_{\text{ic}}^{t-1}/C^{t-1}). \qquad (10)$$

Next, we need to estimate the number of active UEs in the tth RTT $\tilde{D}_{\text{UE}}^t$. We use δ^t to represent the difference between the estimated numbers of UEs in the $(t-1)$th and the tth RTTs. That is $\delta^t = \tilde{D}_{\text{UE}}^t - \tilde{D}_{\text{UE}}^{t-1}$ for $t = 1, 2, \cdots$, where $\tilde{D}_{\text{UE}}^0 = 0$. According to [44], we have $\delta^t \approx \delta^{t-1}$. Therefore, the number of UEs in RTT t is estimated as

$$\tilde{D}_{\text{UE}}^t = \max\{2V_{\text{cc}}^{t-1}, \tilde{D}_{\text{UE}}^{t-1} + \delta^{t-1}\}, \qquad (11)$$

where $2V_{\text{cc}}^{t-1}$ represents that there are at least $2V_{\text{cc}}^{t-1}$ number of UEs colliding in the last RTT.

Based on the estimated number of active UEs in the tth RTT $\tilde{D}_{\text{UE}}^t$, the probability that only one UE chooses CTU c

(i.e., no collision occurs) is given by

$$\mathbb{P}(D_c = 1 | \tilde{D}_{\text{UE}}^t = n) = \binom{n}{1} 1/C^t (1 - 1/C^t)^{n-1}. \quad (12)$$

The the expected number of the successfully served UEs in the tth RTT is given as

$$V_{\text{suss}}^t(C^t) = \mathbb{E}[V_{\text{sc}}^t | \tilde{D}_{\text{UE}}^t = n]$$
$$= \sum_{c=1}^{C^t} \mathbb{P}(D_c = 1 | D_{\text{UE}}^t = n) = n(1 - 1/C^t)^{n-1}. \quad (13)$$

The maximal expected number of the successfully served UEs is obtained by choosing the number of CTUs as

$$C^{t*} = \arg\max_{C^t \in \mathcal{N}_{\text{CTU}}} V_{\text{suss}}^t(C^t). \quad (14)$$

IV. DEEP REINFORCEMENT LEARNING-BASED GF-NOMA RESOURCE CONFIGURATION

The deep reinforcement learning (DRL) is regarded as a powerful tool to address complex dynamic control problems in POMDP. In this section, we propose a Deep Q-network (DQN) based algorithm to tackle the problem (P1). The reasons in choosing DQN are that: 1) the Deep Neural Network (DNN) function approximation is able to deal with several kinds of partially observable problems [23], [45]; 2) DQN has the potential to accurately approximate the desired value function while addressing a problem with very large state spaces; 3) DQN is with high scalability, where the scale of its value function can be easily fit to a more complicated problem; 4) a variety of libraries have been established to facilitate building DNN architectures and accelerate experiments, such as Tensor-Flow, Pytorch, Theano, Keras, and etc.. To evaluate the capability of DQN in GF-NOMA, we first consider the dynamic configuration of repetition value K^t with fixed CTU numbers C^t, where the DQN agent dynamically configures the K^t at the beginning of each RTT for K-repetition and Proactive GF schemes. We then propose a cooperative multi-agent learning technique based on the DQN to optimize the configuration of both repetition value K^t and CTU numbers C^t simultaneously, which breaks down the selection in high-dimensional action space into multiple parallel sub-tasks.

A. Deep Reinforcement Learning-Based Single-Parameter Configuration

1) Reinforcement Learning Framework: To optimize the number of successfully served UEs under the latency constraint in GF-NOMA schemes, we consider a RL-agent deployed at the BS to interact with the environment in order to choose appropriate actions progressively leading to the optimization goal. We define $S \in \mathcal{S}$, $A \in \mathcal{A}$, and $R \in \mathcal{R}$ as any state, action, and reward from their corresponding sets, respectively. The RL-agent first observes the current state S^t corresponding to a set of previous observations $(O^t = \{U^{t-1}, U^{t-2}, \ldots, U^1\})$ in order to select a specific action $A^t \in \mathcal{A}(S^t)$. Here, the action A^t represents the

repetition values K^t in the tth RTT $A^t = K^t$ in this single-parameter configuration scenario and the S^t is a set of indices mapping to the current observed information $U^{t-1} = [V_{\text{cc}}^{t-1}, V_{\text{ic}}^{t-1}, V_{\text{sc}}^{t-1}, V_{\text{sd}}^{t-1}, V_{\text{ud}}^{t-1}]$. With the knowledge of the state S^T, the RL-agent chooses an action A^t from the set $\mathcal{A}$. Once an action A^t is performed, the RL-agent transits to a new observed state S^{t+1} and receives a corresponding reward R^{t+1} as the feedback from the environment, which is designed based on the new observed state S^{t+1} and guides the agent to achieve the optimization goal. As the optimization goal is to maximize the number of the successfully served UEs under the latency constraint, we define the reward R^{t+1} as

$$R^{t+1} = V_{\text{sd}}^t, \quad (15)$$

where V_{sd}^t is the observed number of successfully served UEs under the latency constraint T_{cons}.

To select an action A^t based on the current state S^t, a mapping policy $\pi(a|s)$ learned from a state-action value function $Q(s,a)$ is needed to facilitate the action selection process, which indicates probability distribution of actions with given states. Accordingly, our objective is to find an optimal value function $Q^*(s,a)$ with optimal policy $\pi^*(a|s)$. At each RTT, $Q(s,a)$ is updated based on the received reward by following

$$Q(S^t, A^t)$$
$$= Q(S^t, A^t) + \lambda[R^{t+1} + \gamma \max_{a \in \mathcal{A}} Q(S^{t+1}, a) - Q(S^t, A^t)], \quad (16)$$

where λ is a constant learning rate reflecting how fast the model adapting to the problem, $\gamma \in [0, 1)$ is the discount rate determining how current rewards affect the value function updating. After enough iterations, the BS can learn the optimal policy maximizing the long-term rewards.

2) Deep Q-Network: When the state and action spaces are large, the RL algorithm becomes expensive in terms of memory and computation complexity, which is difficult to converge to the optimal solution. To overcome this problem, DQN is proposed in [45], where the Q-learning is combined with DNN to train a sufficiently accurate state-action value function for the problems with high dimensional state space. Furthermore, the DQN algorithm utilizes the experience replay technique to enhance the convergence performance of RL. When updating the DQN algorithm, mini-batch samples are selected randomly from the experience memory as the input of the neural network, which breaks down the correlation among the training samples. In addition, through averaging the selected samples, the distribution of training samples can be smoothed, which avoids the training divergence.

In DQN algorithm, the action-state value function $Q(s.a)$ is parameterized via a function $Q(s, a, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ represents the weights matrix of a multiple layers DNN. We consider the conventional fully-connected DNN, where the neurons between two adjacent layers are fully pairwise connected. The variables in the state S^t is fed in to the DNN as the input; the Rectifier Linear Units (ReLUs) are adopted as intermediate hidden layers by utilizing the function $f(x) = \max(0, x)$;

while the output layer is consisted of linear units, which are in one-to-one correspondence with all available actions in $\mathcal{A}$.

To achieve exploitation, the forward propagation of Q-function $Q(s, a, \boldsymbol{\theta})$ is performed according to the observed state S^t. The online update of weights matrix $\boldsymbol{\theta}$ is carried out along each training episode to avoid the complexities of eligibility traces, where a double deep Q-learning (DDQN) training principle [46] is applied to reduce the overestimations of value function (i.e., sub-optimal actions obtain higher values than the optimal action). Accordingly, learning takes place over multiple training episodes, where each episode consists of several RTT periods. In each RTT, the parameter $\boldsymbol{\theta}$ of the Q-function approximator $Q(s, a, \boldsymbol{\theta})$ is updated using RMSProp optimizer [47] as

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \lambda_{\mathrm{RMS}} \nabla L^{\mathrm{DDQN}}(\boldsymbol{\theta}^t) \tag{17}$$

where $\lambda_{\mathrm{RMS}} \in (0, 1]$ is RMSProp learning rate, $\nabla L^{\mathrm{DDQN}}(\boldsymbol{\theta}^t)$ is the gradient of the loss function $L^{\mathrm{DDQN}}(\boldsymbol{\theta}^t)$ used to train the state-action value function. The gradient of the loss function is

$$\begin{aligned} &\nabla L^{\mathrm{DDQN}}(\boldsymbol{\theta}^t) \\ &= \mathrm{E}_{S^i, A^i, R^{i+1}, S^{i+1}}[(R^{i+1} + \gamma \max_{a \in \mathcal{A}} Q(S^{i+1}, a, \bar{\boldsymbol{\theta}}^t) \\ &\quad - Q(S^i, A^i, \boldsymbol{\theta}^t)) \nabla_{\boldsymbol{\theta}} Q(S^i, A^i, \boldsymbol{\theta}^t)]. \end{aligned} \tag{18}$$

We consider the application of minibatch training, instead of a single sample, to update the value function $Q(s, a, \boldsymbol{\theta})$, which improves the convergent reliability of value function $Q(s, a, \boldsymbol{\theta})$. Therefore, the expectation is taken over the minibatch, which are randomly selected from previous samples $(S_i, A_i, S_{i+1}, R_{i+1})$ for $i \in \{t - M_r, \ldots, t\}$ with M_r being the replay memory size [23]. When $t - M_r$ is negative, it represents to include samples from the previous episode. Furthermore, $\bar{\boldsymbol{\theta}}^t$ is the target Q-network in DDQN that is used to estimate the future value of the Q-function in the update rule, and $\bar{\boldsymbol{\theta}}^t$ is periodically copied from the current value $\boldsymbol{\theta}^t$ and kept unchanged for several episodes.

Through calculating the expectation of the selected previous samples in minibatch and updating the $\boldsymbol{\theta}^t$ by (17), the DQN value function $Q(s, a, \boldsymbol{\theta})$ can be obtained. The detailed DQN algorithm is presented in **Algorithm 1.**

B. Cooperative Multi-Agent Learning-Based Multi-Parameter Optimization

In practice, not only the repetition values but also the CTU numbers, influence reliability-latency performance in GF-NOMA. Fixed CTU numbers cannot adapt to the dynamic random traffic, which may violate the stringent latency requirement or lead to low resource efficiency. Thus, we study the problem (P1) of jointly optimizing the resource configuration with parameters $A^t = \{K^t, C^t\}$ to improve the network performance. The learning algorithm provided in Sec. V.A is model-free, and thus the learning structure can be extended in this multi-parameter scenario.

Due to the high capability of DQN to handle problems with massive state spaces, we consider to improve the state spaces with more observed information to support the

---

Algorithm 1 DQN-Based GF-NOMA Uplink Resource Configuration

Input: The set of repetition values in each RTT K and Operation Iteration I.

1 Algorithm hyperparameters: learning rate $\lambda_{RMS} \in (0, 1]$, discount rate $\gamma \in [0, 1)$, ϵ-greedy rate $\epsilon \in (0, 1]$, target network update frequency J;

2 Initialization of replay memory M to capacity D, the state-action value function $Q(S, A, \boldsymbol{\theta})$, the parameters of primary Q-network $\boldsymbol{\theta}$, and the target Q-network $\bar{\boldsymbol{\theta}}$;

3 **for** *Iteration ← 1 to I* **do**

4   Initialization of S^1 by executing a random action A^0 and bursty traffic arrival rate $\mu^0 = 0$;

5   **for** *t ← 1 to T* **do**

6   Update μ^0 using Eq. (2);

7   **if** $p_\epsilon < \epsilon$ **Then** select a random action A^t from $\mathcal{A}$;

8   **else** select $A^t = \arg\max_{a \in A} Q(S^t, a, \boldsymbol{\theta})$.

9   The BS broadcasts $K(A^t)$ and backlogged UEs attempt communication in the tth RTT;

10   The BS observes state S^{t+1}, and calculate the related reward R^{t+1} using Eq. (15);

11   Store transition $(S^t, A^t, R^{t+1}, S^{t+1})$ in replay memory M;

12   Sample random minibatch of transitions $(S^t, A^t, R^{t+1}, S^{t+1})$ from replay memory M;

13   Perform a gradient descent step and update parameters $\boldsymbol{\theta}$ for $Q(s, a, \boldsymbol{\theta})$ using Eq. (18);

14   Update the parameter $\bar{\boldsymbol{\theta}} = \boldsymbol{\theta}$ of the target Q-network every J steps.

15   **end**

16 **end**

---

optimization of RL-agent. Therefore, we define the current state S^t, to include information about the last M_o RTTs $(U^{t-1}, U^{t-2}, U^{t-3}, \ldots, U^{t-M_o})$, which enables the RL-agent to estimate the trend of traffic. Similar to the state spaces, the available action spaces also exponentially increases with the increment of the adjustable parameter configurations in GF-NOMA. The total number of available actions corresponds to the possible combinations of all parameter configurations.

Although the GF-NOMA configuration is managed by a central BS, breaking down the control of multiple parameters as multiple sub-tasks is sufficient to deal with the problems with unsolvable action space, which are cooperatively handled by independent Q-agents. As shown in Fig. 7, we consider multiple DQN agents that are centralized at the BS following the same structure of value function approximator as Sec. V.A. Each DQN agent controls their own action variable and receives a common reward to guarantee the objective in P1 cooperatively.

However, the common reward design also poses challenge on the evaluation of each action, because the individual effect of specific action is deeply hidden in the effects of the actions taken by all other DQN agents. For instance, a positive action taken by a agent can receive a misleading low reward due to other DQN agents' negative actions. Fortunately,
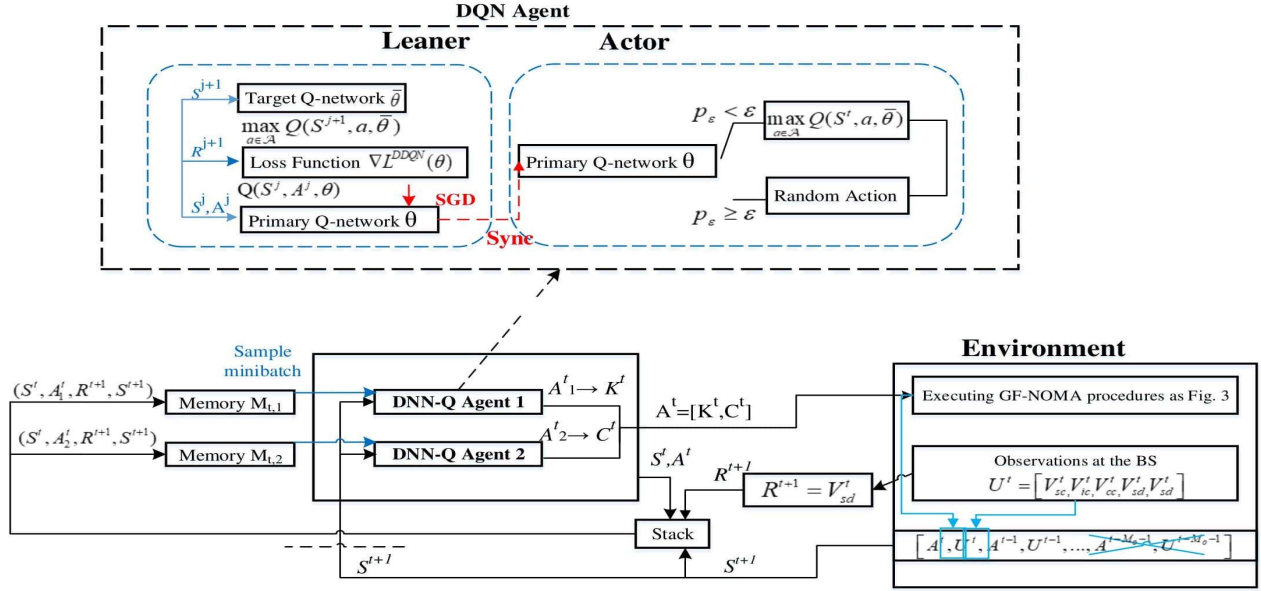
Fig. 7. The CMA-DQN agents and environment interaction in the POMDP.

in GF-NOMA scenario, all DQN agents are centralized at the BS and share full information among each other. Accordingly, we include the action selection histories of each DQN agent as part of state function, and hence, the agents are able to learn the relationship between the common reward and different combinations of actions. To do so, we define state variable S^t as

$$S^t = [A^{t-1}, U^{t-1}, A^{t-2}, U^{t-2}, \ldots, A^{t-M_o}, U^{t-M_o}], \quad (19)$$

where M_o is the number of stored observations, A^{t-1} is the set of selected action of each DQN agent in the $(t-1)$th TTI, and U^{t-1} is the set of observed transmission receptions.

In each RTT, the kth agent update the parameters $\boldsymbol{\theta}_k$ of the value function $Q(s, a_k, \boldsymbol{\theta}_k)$ using RMSProp optimizer following Eq. (17). The algorithm can be implemented following **Algorithm 1**. Different from GF-NOMA single-parameter configuration scenario in Section IV.A, it is required to initialize two primary networks $\boldsymbol{\theta}_k$, target networks $\bar{\boldsymbol{\theta}}_k$ and the replay memories M_k for each DQN agent. In step 10 of **Algorithm 1**, each agent stores their own current transactions in memory separately. In step 11 and 12 of **Algorithm 1**, the minibatch of transaction should separately be sampled from individual memory to train the corresponding DQN agent.

C. Complexity of the Proposed Algorithm

The computational complexity of the proposed algorithm consists of two aspects, i.e., the computational complexity related to the DRL model and the computational complexity related to the training process. As demonstrated in [48], the computational complexity of the DRL algorithm can be calculated as $O(\bar{m}\bar{n}\log\bar{n})$, where $\bar{m}$ is the number of layers, and $\bar{n}$ is the number of units per learning layer. In terms of the training process, the training complexity for X agents, one minibatch of I episodes with T time-steps until convergence results is of order $O(XIT)$.

TABLE I

SIMULATION PARAMETERS

Parameters	Value
Path-loss exponent η	4
Noise power σ^2	-132 dBm
Transmission power P	23 dBm
The received SINR threshold γ_{th}	-10 dB
Duration of traffic T	2 s
The set of the repetition value	$\{1, 2, 4, 6, 8\}$
The set of the CTU number	$\{12, 24, 36, 48\}$
Latency constraint	2 ms and 8 ms
Bursty traffic parameter Beta(α, β)	(2, 4)
The number of bursty UEs N	20000
Cell radius	10 km
Duration of one TTI	0.125 ms

TABLE II

LEARNING HYPERPARAMETERS

Hyperparameters	Value
Learning rate λ_{RMS}	0.0001
Minimum exploration rate ϵ	0.1
Discount rate γ	0.5
Minibatch size	32
Replay Memory	10000
Target Q-network update frequency	1000

V. SIMULATION RESULTS

In this section, we examine the effectiveness of our proposed GF-NOMA schemes with DQN algorithm via simulation and compare the results with LE-URC. We adopt the standard network parameters listed in Table I following [49], and hyperparameters for the DQN learning algorithm are listed in Table II.
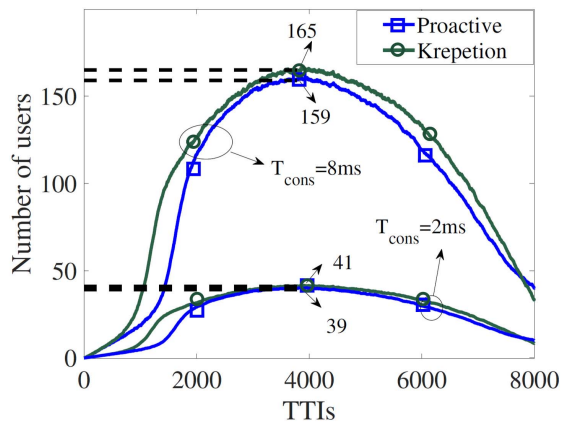
Fig. 8.   Low backlog traffic and high backlog traffic in each TTI.

All testing performance results are obtained by averaging over 1000 episodes. The BS is located at the center of a circular area with a 10 km radius, and the UEs are randomly located within the cell. Unless otherwise stated, we consider the number of bursty UEs to be $N = 20000$. The DQN is set with two hidden layers, each with 128 ReLU units. In the following, we present our simulation results of the single-repetition configuration and the multi-parameter config- uration in Section V-A and Section V-B, respectively. In the single-repetition configuration scenario, we set the number of CTU as $C = 48$. Throughout epoch, each UE has a periodical bursty traffic profile, i.e., the time limited Beta profile defined in (1) with parameters (2, 4) that has a peak around the 4000th TTI. The single-repetition configuration is optimized under the latency constraint $T_{\mathrm{cons}} = 2$ ms (low backlog traffic). The multi-parameter configuration is optimized under the latency constraint $T_{\mathrm{cons}} = 8$ ms (high backlog traffic).

Fig. 8 plots the low backlog traffic under the latency constraint $T_{\mathrm{cons}} = 2$ ms and high backlog traffic under the latency constraint $T_{\mathrm{cons}} = 8$ ms in each TTI, respectively. It should be noted that the the backlog traffic in each TTI does not only include the newly generated traffic, but also the retransmission traffic, due to the fact that the UEs are allowed to retransmit in the next RTT under the latency constraint. The results have shown that when the latency constraint increases, the backlog traffic in each TTI increase as the retransmission traffic increases. Note that the successful probability performance can be calculated by dividing the number of successful UEs by the backlog traffic UE. However, rather than the probability value, we take care of how our approaches affect the long-term metrics like the number of collision devices, non-collision devices, and decoding failure devices, and thus affect the successful UEs (or the probability value), which will be discussed in the analysis of the following results.

A. Single-Repetition Configuration for Low Backlog Traffic

Fig. 9 shows the system convergence process of the proposed DRL learning framework by plotting the aver- age received reward for the K-repetition scheme and the Proactive scheme under low backlog traffic scenario, respec-
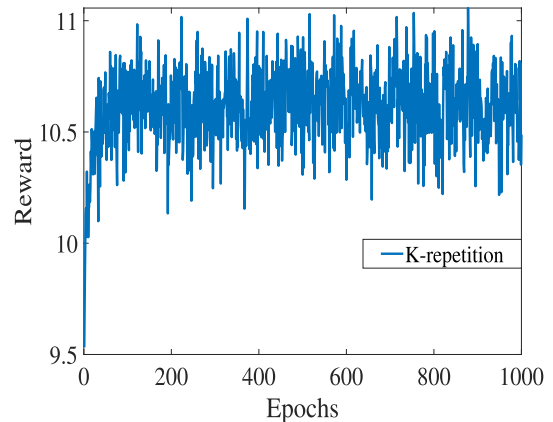
(a) K-repetition scheme
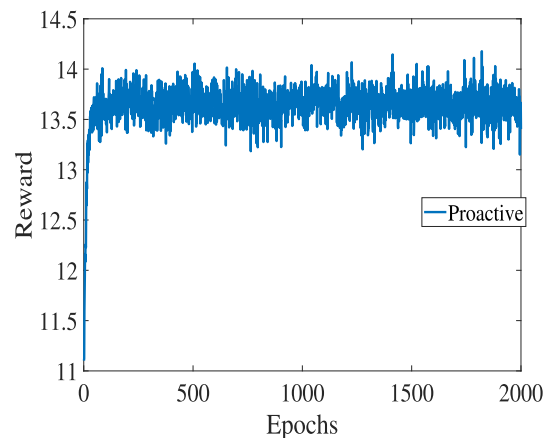
(b) Proactive scheme

Fig. 9.   Average received reward for each GF scheme.

tively. It can be intuitively seen that the proposed framework has a fast convergence speed and the episode required for system convergence is very small for both K-repetition and Proactive schemes. We can also observe that the average received reward of Proactive scheme in Fig. 9 (b) is higher than that of the K-repetition scheme in Fig. 9 (a). This is because the Proactive scheme can terminate the repeti- tion earlier and start new packet transmission with timely ACK feedback, which is able to deal with the traffic more effectively.

Fig. 10 plots the number of the successfully served UEs, the non-collision UEs, the collision UEs, and the decoding failure UEs for the K-repetition scheme and the Proactive scheme respectively. It is shown that the number of non-collision transmission UEs of both scheme is similar in the same scenario. However, the number of decoding failure UEs of the K-repetition scheme is more than that of the Proactive scheme, due to more interference caused by multiple repetitions in K-repetition scheme. That is to say, the earlier terminating of the UEs in the Proactive scheme can reduce the interference to other UEs, and thus leads to an increase in the number of successfully decoding UEs. In both Fig. 10 (a) and Fig. 10 (b), the number of collision UEs has a peak at around the 4000th TTI with the peak traffic at this time as shown in Fig. 8. In addition, the number of failure decoding UEs reaches a peak due to the peak traffic at the 4000th TTI, which leads
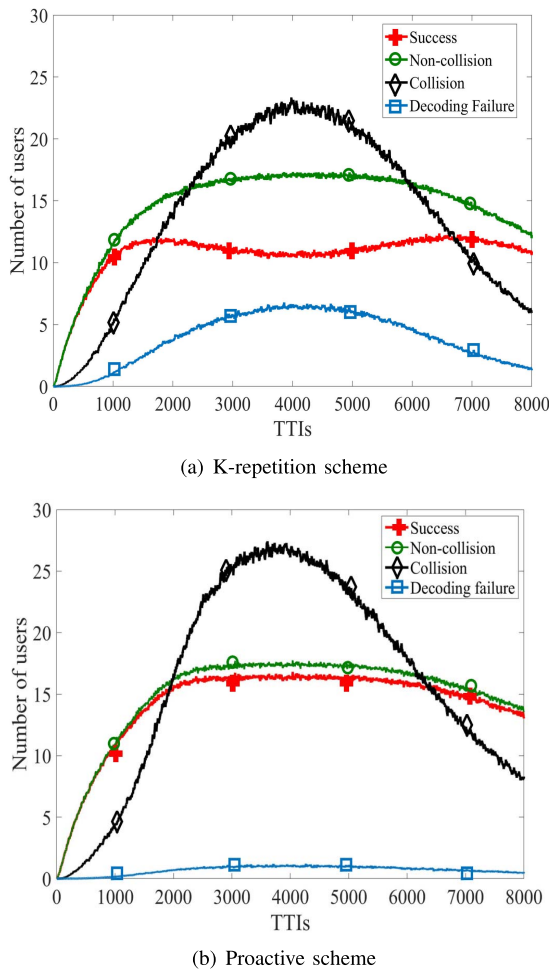
(a) K-repetition scheme

(b) Proactive scheme

Fig. 10.   The transmission results for each GF scheme.

(a) K-repetition scheme

(b) Proactive scheme

Fig. 11.   Average received reward for each GF scheme with multi-parameter configuration.

to the decrease in the number of successful UEs at that time.

B. Multi-Parameter Configuration for High Backlog Traffic

Fig. 11 shows the system convergence process of the proposed CMA-DQN by plotting the average received reward for the K-repetition scheme and the Proactive scheme with multi-parameter configuration, including the repetition values and the CTU number, under high backlog traffic scenario, respectively. It can be intuitively seen that the Proactive scheme has a little bit faster convergence speed than the K-repetition scheme. Compared to Fig. 9 under lower backlog traffic scenario, we observe that the average rewards of the two schemes decrease significantly. This is because the larger latency constraint $T_{\text{cons}} = 8$ ms leads to larger retransmission packets, thus higher backlog traffic, which results in serious traffic congestion. It is noted that the performance degradation of the K-repetition scheme is much larger than that of the Proactive scheme, and the average reward for the Proactive scheme is almost three times more than that for the K-repetition scheme, which shows the potential of the Proactive scheme in heavy traffic scenarios (mURLLC) due to timely termination.
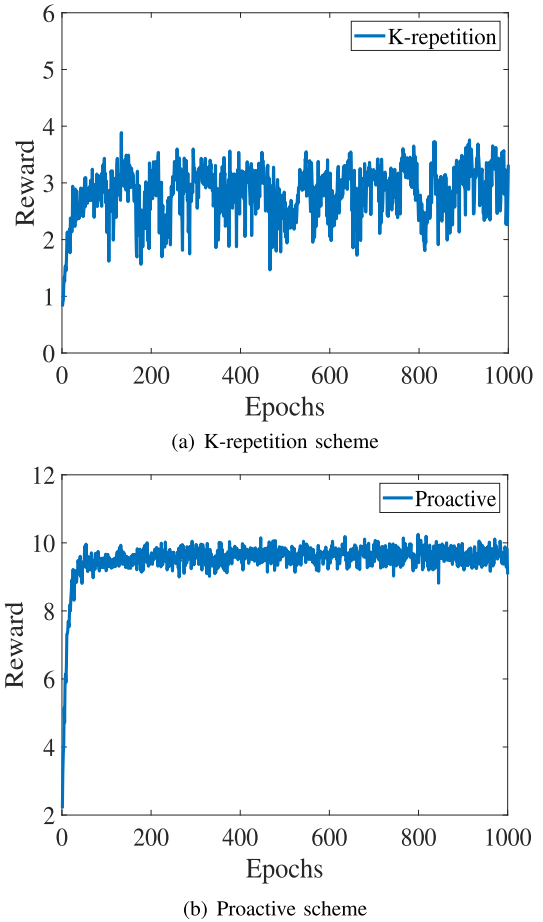
Fig. 12 plots the number of successful UEs, non-collision UEs, and decoding failure UEs for K-repetition scheme and Proactive scheme with multi-parameter configuration including the repetition values and the CTU number, respectively. Similar with Fig. 10, the number of decoding failure UEs of the K-repetition scheme is almost up to five times more than that of the Proactive scheme at the peak traffic, due to the interference caused by multiple repetitions from collision UEs. It is also noted that in both schemes, there is lower number of successful UEs in high traffic (especially in peak traffic at round 4000th TTI) and a higher number of successful UEs in low traffic (especially around 1000th and 8000th TTIs). This reveals that one design challenge for GF-NOMA transmission is to deal with the potential signature collision, which will happen in the case of random signature selection and when the number of potential users is much larger than the pool size of the NOMA signatures.

Fig. 13 plots the average number of successful UEs for the K-repetition scheme and the Proactive scheme by comparing the learning framework with fixed parameters, and with the LE-URC approach, respectively. Here, we set the fixed repetition value $K = 8$ and the CTU number $C = 48$. Our results shown that the number of successfully served UEs under the same latency constraint in our proposed learning framework is up to ten times for the K-repetition
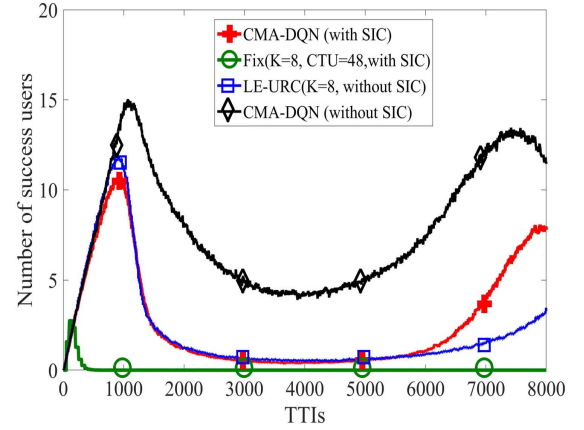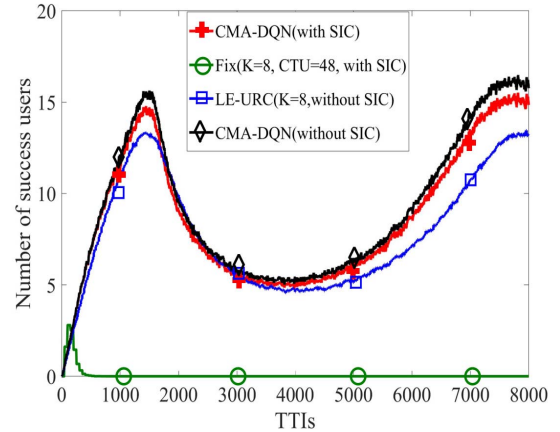
(a) K-repetition scheme
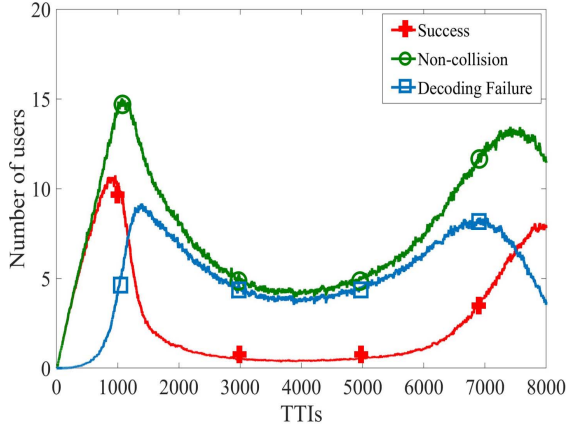
(a) K-repetition scheme
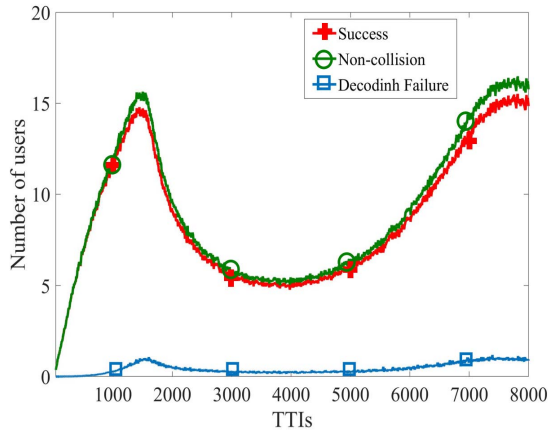
(b) Proactive scheme

Fig. 12.   The transmission results for each GF scheme.

(b) Proactive scheme

Fig. 13.   Average number of successfully served UEs for each GF scheme.

scheme, and fifteen times for the Proactive scheme, more than that with fixed repetition values and CTU numbers, respectively. In addition, since the LE-URC approach is not aware of the latency constraint and SIC procedure, the results are large at first, but still smaller than the number of non-collision UEs of CMA-DQN (without SIC). However, with increasing TTIs (above 6000), the cumulated traffic increases due to unsuccessful transmissions and retransmissions, the LE-URC method becomes worse and achieve lower number of successful UEs than that of CMA-DQN due to its ignorance in latency constraint during its optimization for one time instance. The superior performance of CMA-DQN in heavy traffic scenario also demonstrate its capability in dynamically configure lower lower repetition values and CTU numbers to alleviate the traffic congestion to obtain a long-term reward.

Fig. 14 plots the average number K of repetition values and the average number C of CTUs for each scheme that are selected by CMA-DQN and LE-URC, respectively. First, for the repetition values, it is known that increasing the repetition value increases the success probability, as it offers more opportunities to transmit. However, in overloaded traffic scenarios, the repetition also increases the collisions and
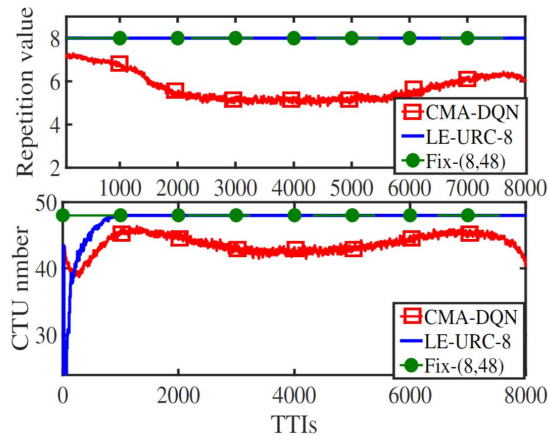
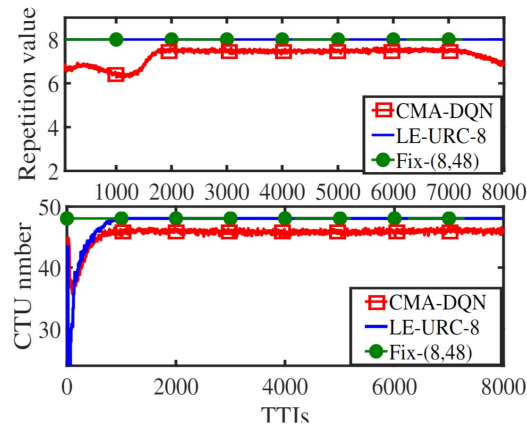wastes extra time and potential resources. We can see that in Fig. 14 (a), the repetition value of K-repetition scheme decreases first and then increases back to a higher value. This is because the agent in K-repetition scheme learns to sacrifice the current successful transmission to alleviate the traffic congestion in heavy traffic region to obtain a long-term reward, while LE-URC approach just adopts the maximum repetition value to optimize for one time instance. In Fig. 14 (b), we can see that the Proactive scheme adopts a higher and more stable repetition value due to its capability to deal with the traffic congestion. Then, for the average CTUs number, we can see that in Fig. 14 (a), the CTUs number increases during the light traffic period and decreases during the high traffic period. The chosen reason is similar to the choice of repetition value, which may be caused by the sharing of actions as observations among agents. Similarly, in Fig. 14 (b), we can see that the Proactive scheme adopts a higher and more stable average CTUs number due to its capability to deal with the traffic congestion.

The realistic network conditions can be different from the simulation environment, due to that the practical traffic and physical channel vary and can be unpredictable. This difference may lead to the inaccurate configuration that can degrade the system performance of each approach. Fortunately, the
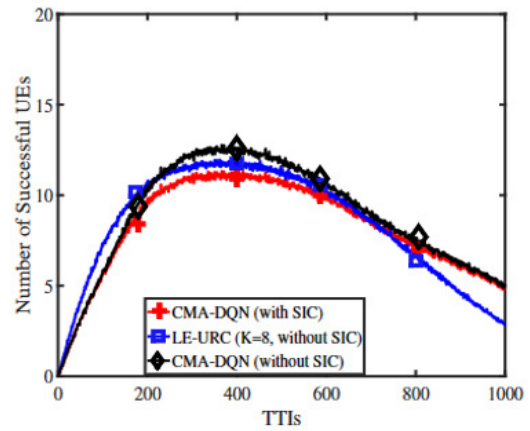
(a) K-repetition scheme

(b) Proactive scheme

Fig. 14. Actions for each GF scheme.

(a) Low Traffic

(b) High Traffic

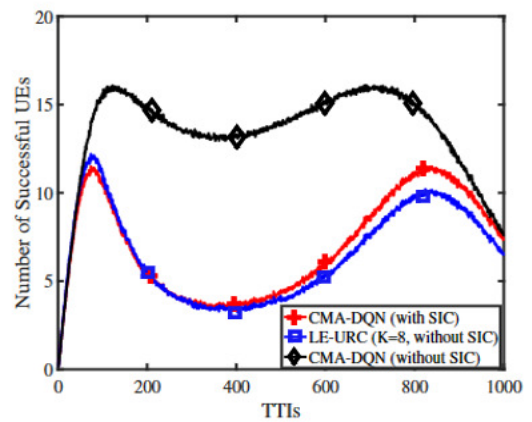Fig. 15. Average number of successfully served UEs for (a) light traffic and (b) heavy traffic.

proposed RL-based approaches can self-update after deployment according to the practical observation in GF-NOMA networks in an online manner. To model this, we use the trained CMA-DQN agents given in Fig. 14(a) (i.e., the number of bursty UEs is 20000 and the latency constraint is 8ms), and test them in a slightly modified traffic scenario as shown in Fig. 15(a) (the number of bursty UEs is 10000 for low traffic and the latency constraint is 2ms) and Fig. 15(b) (the number of bursty UEs is 30000 for high traffic and the latency constraint is 2ms). The superior performance of CMA-DQN in heavy traffic scenarios (Fig. 15(b)) demonstrates its capability to dynamically configure lower repetition values and CTU numbers to alleviate traffic congestion to obtain a long-term reward.

VI. Conclusion and Discussion

In this paper, we developed a general learning framework for dynamic resource configuration optimization in signature-based GF-NOMA systems for mURLLC service under the K-repetition GF scheme and the Proactive GF scheme. This proposed learning framework defined the observations, actions, and rewards to maximize long-term successfully served UEs under the latency constraint, which can be standardized as the collected parameters from the environment. We first performed a real-time repetition value

configuration, where a double Deep Q-Network (DDQN) was developed. We then designed a Cooperative Multi-Agent DQN (CMA-DQN) to optimize the configuration of both the repetition values and the CTU numbers for these two schemes, by dividing high-dimensional configurations into multiple parallel sub-tasks.

Our results have shown that: 1) the number of successfully served UEs under the same latency constraint in our proposed learning framework is up to ten times for the K-repetition scheme, and fifteen times for the Proactive scheme, more than that with fixed repetition values and CTU numbers; 2) with learning optimization, the Proactive scheme always outperforms the K-repetition scheme in terms of the number of successfully served UEs, especially under the high backlog traffic scenario; 3) the proposed CMA-DQN is superior to the conventional load estimation-based approach (LE-URC) that demonstrating its capability in dynamically configuring for mURLLC in heavy traffic scenarios in long term; and 4) the proposed learning framework can be extended to optimize other resource configuration problems in GF-NOMA schemes, such as retransmission times, starting offset of the grant, etc.. and for other signature-based GF-NOMA schemes with different signatures. With realistic traffic, a direct implementation of DRL may bring computational complexity and processing

delay at the BSs, so how to reduce the complexity of DRL algorithms can be considered in future work.

REFERENCES

[1] Y. Liu, Y. Deng, M. Elkashlan, and A. Nallanathan, "Cooperative deep reinforcement learning based grant-free NOMA optimization for mURLLC," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2022, pp. 1–6.

[2] X. Zhang, J. Wang, and H. V. Poor, "Statistical delay and error-rate bounded QoS provisioning for mURLLC over 6G CF M-MIMO mobile networks in the finite blocklength regime," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 652–667, Mar. 2021.

[3] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Netw.*, vol. 34, no. 3, pp. 134–142, May 2020.

[4] *Study on Scenarios and Requirements for Next Generation Access Technologies*, document TR 38.913 v16.0.0, 3GPP, Jul. 2020.

[5] *5G; Service Requirements for the 5G System*, document TS 22.261 v17.2.0, 3GPP, Mar. 2020.

[6] M. Latva-Aho and K. Leppanen, "Key drivers and research challenges for 6G ubiquitous wireless intelligence," 6G Res. Vis. 1, 6G Flagship, Univ. Oulu, Oulu, Finland, White Paper, Sep. 2019.

[7] *UL Grant-Free Transmission for URLLC*, document TSG-RAN WG1 #88, R1–1705654, 3GPP, Apr. 2017.

[8] N. Ye, H. Han, L. Zhao, and A.-H. Wang, "Uplink nonorthogonal multiple access technologies toward 5G: A survey," *Wireless Commun. Mobile Comput.*, vol. 2018, pp. 1–26, Jun. 2018.

[9] T. Taleb and A. Kunz, "Machine type communications in 3GPP networks: Potential, challenges, and solutions," *IEEE Commun. Mag.*, vol. 50, no. 3, pp. 178–184, Mar. 2012.

[10] *Considerations on Random Resource Selection*, document TSG RAN WG1 #86, R1–1608917, 3GPP, Oct. 2016.

[11] *5G; NR; Physical Layer Procedures for Data*, document TS 38.214 v15.9.0, 3GPP, Mar. 2020.

[12] *Discussion on Explicit HARQ-ACK Feedback for Configured Grant Transmission*, 3GPP document TSG RAN WG1 #96, R1-1903079, Mar. 2019.

[13] N. H. Mahmood, R. Abreu, R. Böhnke, M. Schubert, G. Berardinelli, and T. H. Jacobsen, "Uplink grant-free access solutions for URLLC services in 5G new radio," in *Proc. IEEE Int. Symp. Wireless Commun. Syst. (ISWCS)*, Aug. 2019, pp. 607–612.

[14] Y. Chen et al., "Toward the standardization of non-orthogonal multiple access for next generation wireless networks," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 19–27, Mar. 2018.

[15] 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.

[16] 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, Mar. 2019.

[17] J. Zhang et al., "PoC of SCMA-based uplink grant-free transmission in UCNC for 5G," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 6, pp. 1353–1362, Jun. 2017.

[18] K. Lai, J. Lei, Y. Deng, L. Wen, G. Chen, and W. Liu, "Analyzing uplink grant-free sparse code multiple access system in massive IoT networks," *IEEE Internet Things J.*, vol. 9, no. 7, pp. 5561–5577, Apr. 2022.

[19] X. Dai, Z. Zhang, B. Bai, S. Chen, and S. Sun, "Pattern division multiple access: A new multiple access technology for 5G," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 54–60, Apr. 2018.

[20] Z. Yuan, C. Yan, C. Yuan, and W. Li, "Blind multiple user detection for grant-free MUSA without reference signal," in *Proc. IEEE 86th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2017, pp. 1–5.

[21] C. Wei, H. Liu, Z. Zhang, J. Dang, and L. Wu, "Approximate message passing-based joint user activity and data detection for NOMA," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 640–643, Mar. 2017.

[22] B. Wang, L. Dai, T. Mir, and Z. Wang, "Joint user activity and data detection based on structured compressive sensing for NOMA," *IEEE Commun. Lett.*, vol. 20, no. 7, pp. 1473–1476, Jul. 2016.

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

[24] Y.-C. Wu, T. Q. Dinh, Y. Fu, C. Lin, and T. Q. Quek, "A hybrid dqn and optimization approach for strategy and resource allocation in MEC networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4282–4295, Jul. 2021.

[25] N. Luong et al., "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3133–3174, May 2019.

[26] N. Jiang, Y. Deng, A. Nallanathan, and J. Yuan, "A decoupled learning strategy for massive access optimization in cellular IoT networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 668–685, Mar. 2021.

[27] N. Ye, X. Li, H. Yu, A. Wang, W. Liu, and X. Hou, "Deep learning aided grant-free NOMA toward reliable low-latency access in tactile Internet of Things," *IEEE Trans. Ind. Informat.*, vol. 15, no. 5, pp. 2995–3005, Jan. 2019.

[28] N. Ye, X. Li, H. Yu, L. Zhao, W. Liu, and X. Hou, "DeepNOMA: A unified framework for NOMA using deep multi-task learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2208–2225, Jan. 2020.

[29] Y. Liu, Y. Deng, N. Jiang, M. Elkashlan, and A. Nallanathan, "Analysis of random access in NB-IoT networks with three coverage enhancement groups: A stochastic geometry approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 549–564, Jan. 2021.

[30] N. Jiang, Y. Deng, X. Kang, and A. Nallanathan, "Random access analysis for massive IoT networks under a new spatio–temporal model: A stochastic geometry approach," *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5788–5803, Nov. 2018.

[31] Y. Liu, Y. Deng, M. Elkashlan, A. Nallanathan, and G. K. Karagiannidis, "Analyzing grant-free access for URLLC service," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 741–755, Mar. 2021.

[32] *Cellular System Support for Ultra-Low Complexity and Low Throughput Internet of Things (CIoT)*, document TR 45.820 V13.1.0, Sophia Antipolis, France, 3GPP, Nov. 2015.

[33] *Study on RAN Improvements for Machine-Type Communications*, document TR 37.868 v11.0.0, 3GPP, Sep. 2011.

[34] *On MA Resource and MA Signature Configurations*, document TSG-RAN WG1 #86, R1–1609227, 3GPP, Oct. 2016.

[35] 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. Standards Mag.*, vol. 3, no. 4, pp. 60–66, Dec. 2019.

[36] J. Navarro-Ortiz, P. Romero-Diaz, S. Sendra, P. Ameigeiras, J. J. Ramos-Munoz, and J. M. Lopez-Soler, "A survey on 5G usage scenarios and traffic models," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 2, pp. 905–929, Feb. 2020.

[37] A. K. Gupta and S. Nadarajah, *Handbook of Beta Distribution and its Applications*. Boca Raton, FL, USA: CRC Press, 2004.

[38] S. Sesia, I. I. Toufik, and M. Baker, *Handbook of Beta Distribution and its Applications*. Hoboken, NJ, USA: Wiley, 2011.

[39] *Study on Physical Layer Enhancements for NR Ultra-Reliable and Low Latency Case (URLLC)*, document TR 38.824 v16.0.0, 3GPP, Mar. 2019.

[40] M. B. Shahab, R. Abbas, M. Shirvanimoghaddam, and S. J. Johnson, "Grant-free non-orthogonal multiple access for IoT: A survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1805–1838, May 2020.

[41] K. Au et al., "Uplink contention based SCMA for 5G radio access," in *IEEE Globecom Workshops (GC Wkshps)*, Dec. 2014, pp. 900–905.

[42] D.-D. Tran, V. N. Ha, and S. Chatzinotas, "Novel reinforcement learning based power control and subchannel selection mechanism for grant-free NOMA URLLC-enabled systems," in *Proc. IEEE 95th Veh. Technol. Conference: (VTC-Spring)*, Jun. 2022, pp. 1–5.

[43] Z. Shi et al., "Block error rate analysis of short-packet mobile-to-mobile communications over correlated cascaded fading channels," *IEEE Trans. Veh. Technol.*, vol. 71, no. 4, pp. 4087–4101, Apr. 2022.

[44] S. Duan, V. Shah-Mansouri, Z. Wang, and V. W. S. Wong, "D-ACB: Adaptive congestion control algorithm for bursty M2M traffic in LTE networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 9847–9861, Dec. 2016.

[45] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[46] H. V. Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Proc. 30th AAAI Conf. Artif. Intell.*, Feb. 2016, pp. 2094–2100.

[47] T. Tieleman and G. Hinton, "Lecture 6.5-RMSPROP: Divide the gradient by a running average of its recent magnitude," *COURSERA, Neural Netw. Mach. Learn.*, vol. 4, no. 2, pp. 26–31, Oct. 2012.

[48] K. He and J. Sun, "Convolutional neural networks at constrained time cost," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Oct. 2015, pp. 5353–5360.

[49] *Study on New Radio Access Technology-Physical Layer Aspects*, document TR 38.802 v14.0.0, 3GPP, Mar. 2017.

Yan Liu (Member, IEEE) received the Ph.D. degree in electrical engineering from the Queen Mary University of London, U.K., in 2021, and the Ph.D. degree in signal and information processing from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2022. From 2021 to 2022, she was a Post-Doctoral Research Associate with the Queen Mary University of London. She is currently an Assistant Professor with the College of Electronic and Information Engineering, Tongji University, Shanghai, China. Her main research interests include the field of the Internet of Things, ultra-reliability and low-latency communications, machine learning, stochastic geometry, and 5G/6G wireless networks. She received the IEEE ComSoc Student Travel Grant in IEEE International Conference on Communications (ICC) 2019 and IEEE Global Communications Conference (GLOBECOM) 2019. She has served as a Technical Program Committee Member for the IEEE Vehicular Technology Conference 2022 Spring and IEEE International Conference on Communications 2023. She is currently serving as the Guest Editor for Special Issues of *IEEE Internet of Things Magazine*.

Yansha Deng (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the Queen Mary University of London, U.K., in 2015. From 2015 to 2017, she was a Post-Doctoral Research Fellow with King's College London, U.K., where she is currently a Senior Lecturer (an Associate Professor) with the Department of Engineering. Her research interests include molecular communication and machine learning for 5G/6G wireless networks. She was a recipient of the Best Paper Awards from ICC 2016 and GLOBECOM 2017 as the first author and IEEE Communications Society Best Young Researcher Award for the Europe, Middle East, and Africa Region 2021. She also received the Exemplary Reviewers of the IEEE TRANSACTIONS ON COMMUNICATIONS in 2016 and 2017, and IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS in 2018. She has also served as a TPC Member for many IEEE conferences, such as IEEE GLOBECOM and ICC. She is currently an Associate Editor of IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE COMMUNICATIONS SURVEYS AND TUTORIALS, IEEE TRANSACTIONS ON MACHINE LEARNING IN COMMUNICATIONS AND NETWORKING, and IEEE TRANSACTIONS ON MOLECULAR, BIOLOGICAL AND MULTI-SCALE COMMUNICATIONS, a Senior Editor of the IEEE COMMUNICATION LETTERS, and the Vertical Area Editor of *IEEE Internet of Things Magazine*.

Hui Zhou (Member, IEEE) received the B.Eng. and M.A.Sc. degrees from Wuhan University, Wuhan, China, in 2016 and 2019, respectively, and the Ph.D. degree in electronic engineering from King's College London, London, U.K., in 2022. He is currently a Senior Engineer with Shanghai Huawei Technology Company Ltd. His research interests include massive MIMO, unlicensed spectrum, machine learning, and semantic communication.

Maged Elkashlan (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from The University of British Columbia in 2006. From 2007 to 2011, he was with the Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia. During this time, he held visiting faculty appointments at University of New South Wales, The University of Sydney, and University of Technology Sydney. In 2011, he joined the School of Electronic Engineering and Computer Science, Queen Mary University of London. He also holds a visiting faculty appointment at the Beijing University of Posts and Telecommunications. His research interests include the broad areas of communication theory and statistical signal processing.

He was a co-recipient of the Best Paper Awards at the IEEE International Conference on Communications (ICC) in 2016 and 2014, the International Conference on Communications and Networking in China (CHINACOM) in 2014, and the IEEE Vehicular Technology Conference (VTC-Spring) in 2013. He is an Editor of the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY and the IEEE TRANSACTIONS ON MOLECULAR, BIOLOGICAL AND MULTI-SCALE COMMUNICATIONS. He was an Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2013 to 2018 and the IEEE COMMUNICATIONS LETTERS from 2012 to 2016.

Arumugam Nallanathan (Fellow, IEEE) has been a Professor in wireless communications and the Head of the Communication Systems Research (CSR) Group, School of Electronic Engineering and Computer Science, Queen Mary University of London, since September 2017. He was with the Department of Informatics, King's College London, from December 2007 to August 2017, where he was a Professor in wireless communications, from April 2013 to August 2017, and a Visiting Professor from September 2017. He was an Assistant Professor with the Department of Electrical and Computer Engineering, National University of Singapore, from August 2000 to December 2007. His research interests include artificial intelligence for wireless systems, beyond 5G wireless networks, the Internet of Things (IoT), and molecular communications. He has published nearly 500 technical papers in scientific journals and international conferences.

He was a co-recipient of the Best Paper Awards presented at the IEEE International Conference on Communications 2016 (ICC'2016), IEEE Global Communications Conference 2017 (GLOBECOM'2017), and IEEE Vehicular Technology Conference 2018 (VTC'2018). He is also a co-recipient of IEEE Communications Society Leonard G. Abraham Prize in 2022. He is an IEEE Distinguished Lecturer. He has been selected as a Web of Science Highly Cited Researcher in 2016 and 2022 and an AI 2000 Internet of Things Most Influential Scholar in 2020. He is an Editor-at-Large for IEEE TRANSACTIONS ON COMMUNICATIONS and a Senior Editor for IEEE WIRELESS COMMUNICATIONS LETTERS. He was an Editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS (2006–2011), IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY (2006–2017), and IEEE SIGNAL PROCESSING LETTERS. He served as the Chair for the Signal Processing and Communication Electronics Technical Committee of IEEE Communications Society and the Technical Program Chair and a member of Technical Program Committees in numerous IEEE conferences. He received the IEEE Communications Society SPCE Outstanding Service Award 2012 and IEEE Communications Society RCC Outstanding Service Award 2014.