Received 30 May 2023; accepted 23 June 2023. Date of publication 3 July 2023; date of current version 27 July 2023.

Digital Object Identifier 10.1109/OJCOMS.2023.3291689

Multi-Agent DRL Approach for Energy-Efficient Resource Allocation in URLLC-Enabled Grant-Free NOMA Systems

DUC-DUNG TRAN [1] (Member, IEEE), SHREE KRISHNA SHARMA [1] (Senior Member, IEEE), VU NGUYEN HA [1] (Member, IEEE), SYMEON CHATZINOTAS [1] (Fellow, IEEE), AND ISAAC WOUNGANG [2] (Senior Member, IEEE)

[1] Interdisciplinary Center for Security, Reliability and Trust, University of Luxembourg, 1855 Luxembourg City, Luxembourg

[2] Department of Computer Science, Toronto Metropolitan University, Toronto, ON M5B 2K3, Canada

CORRESPONDING AUTHOR: D.-D. TRAN (e-mail: duc.tran@uni.lu)

This work was supported in part by the FNR-Funded Project CORE 5G-Sky under Grant C19/IS/13713801.

ABSTRACT Grant-free non-orthogonal multiple access (GF-NOMA) has emerged as a promising access technology for the fifth generation and beyond wireless networks that enable ultra-reliable and low-latency communications (URLLC) to ensure low access latency and high connectivity density. Furthermore, designing energy-efficient (EE) resource allocation strategies is a crucial aspect of future cellular system development. Taking these goals into account, this paper proposes an EE sub-channel and power allocation strategy for URLLC-enabled GF-NOMA (URLLC-GF-NOMA) systems based on multi-agent (MA) deep reinforcement learning (MADRL). In particular, the URLLC-GF-NOMA methods using MA dueling double deep Q network (MA3DQN), MA double deep Q network (MA2DQN), and MA deep Q network (MADQN) techniques are designed to enable users to select the most appropriate sub-channel and transmission power for their communications. The aim is to build an efficient MADRL-based solution, ensuring rapid convergence with small signaling overhead, to maximize the network EE while fulfilling the URLLC requirements of all users. Simulation results show that the MADQN and MA2DQN methods, which have lower complexity than MA3DQN, are more appropriate for the URLLC-GF-NOMA systems under consideration. Moreover, our proposed methods exhibit superior convergence characteristics, a reduction in signaling overhead, and enhanced EE performance compared to other benchmark strategies.

INDEX TERMS Energy efficiency, grant-free NOMA, multi-agent deep reinforcement learning, URLLC.

I. INTRODUCTION

ULTRA-RELIABLE and low-latency communications (URLLC) is one of the most critical services of the fifth generation (5G) and beyond wireless networks [1], [2]. It is expected to support mission-critical Internet of Things (IoT) applications, such as smart city, remote surgery, intelligent transportation, and vehicle-to-everything (V2X) communications, with stringent reliability and latency requirements. Specifically, a general URLLC condition for a one-way radio is defined as 99.999% target reliability and 1 ms latency [3], [4]. Due to the unprecedented constraints of high reliability and low latency, the packet lengths of URLLC messages are generally ultra-short. Thus, the channel's blocklength is finite, requiring a thorough analysis of achievable rate and decoding error probability. These considerations can be ignored in traditional wireless communication schemes that mostly focus on the Shannon channel capacity under the assumption of infinite blocklength [3]. Therefore, URLLC-enabled systems require a new transmission method. In this regard, short-packet communications (SPC) in finite blocklength (FBL) regime could be a promising approach to meet the URLLC requirements [3], [5].

Furthermore, one of the major challenges in 5G and beyond wireless networks is supporting massive access

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

over a limited radio spectrum [6]. To resolve this challenge, non-orthogonal multiple access (NOMA) has been demonstrated as a promising solution [7]. One of the latest NOMA techniques is the grant-free (GF) NOMA (GF-NOMA), where users can communicate with the base station (BS) simultaneously and quickly on the same time-frequency resource block (RB) without the need for a demand-assigned access from the BS [8]. This access method can improve the spectrum access efficiency and reduce the transmission latency for the system. The application of NOMA to URLLC-enabled systems has also been considered in recent years [9], [10], [11] to further enhance the system performance.

GF transmission has been proposed for 5G new radio (NR) as a promising solution to reduce the latency in URLLC and massive access scenarios [7], [12]. In GF URLLC, a user can communicate with the base station in an arrive-and-go manner without the need to schedule the requests and uplink resource grants, thereby reducing the latency. However, the random nature of the GF access might lead to congestion, as multiple users could potentially access the same RB. The GF-NOMA can mitigate this issue by enabling many users to share the same RBs. However, because the GF access is random, a larger number of users can occupy one RB simultaneously, which may lead to severe interference in GF-NOMA systems and degrade the system performance. This demands an intelligent resource allocation approach for GF-NOMA networks to optimize the system performance. Machine learning (ML), which is recognized as one of the potential technologies for the next generation wireless networks [13], could be an enabling solution to address the above problem. The underlying principle of ML is to learn from the observed data or surrounding environment in order to make optimal decisions in complex, dynamic, and uncertain large-scale environments. ML techniques including supervised learning [14], unsupervised learning, and reinforcement learning (RL) [15], [16], have been recently investigated in order to address various issues in wireless communication schemes such as channel estimation and signal detection, beamforming design, resource allocation, and system security.

A. RELATED WORKS

Recently, the combination of NOMA and URLLC has been investigated in several works [9], [10], [11] to increase connectivity and guarantee the reliability and latency requirements for wireless networks. Specifically, these works considered multiple-input multiple-output (MIMO) and multiple-input single-output (MISO) schemes for URLLC-enabled systems to improve the system performance in terms of reliability and latency. The works proposed user-pairing methods based on the power-domain NOMA principle to enhance connectivity and reduce interference. However, the above works did not examine the GF access method, which can support massive access and reduce the transmission latency for wireless systems requiring high reliability and low latency.

Taking GF transmission into account, the works in [17], [18] studied GF access for OMA. In the GF-OMA scheme, users can select RBs randomly, and each RB is used strictly by a single user for successful reception. This limitation may lead to severe collisions when the number of users is much higher than the number of available RBs. To overcome this challenge, GF-NOMA has emerged as a promising technology for massive access by allowing multiple users to access the same RB based on the power-domain NOMA [7]. In particular, the users occupying the same RB are distinguished by different received power levels, and multi-user data can be decoded at the receivers by utilizing the successive interference cancellation (SIC). The traditional contention-based GF-NOMA schemes are implemented by dividing a cell area into multiple fractions and using the orthogonal resource allocation among those fractions to reduce the inter-fraction collisions [8], [19]. Nevertheless, the spectrum competition among users within the same fraction is still high, resulting in severe interference and reducing system performance. Thus, it is important to find a smart congestion control method to reduce the collisions and improve the long-term system performance.

Intelligent features are an important aspect of future cellular networks, and many current research works have applied RL-based algorithms to address the collisions and severe interference in massive access scenarios [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31]. Specifically, Sharma and Wang [20] proposed a collaborative distributed Q-learning algorithm for the frame-based slotted-Aloha (SA) random access (RA) scheme to find the best resource block allocation strategy for IoT users, in order to avoid collisions in GF-OMA-based IoT systems. The authors in [21], [22], [23], [24] investigated the application of Q-learning to different GF-NOMA scenarios with/without SPC to mitigate the congestion and interference in overloaded systems, where the number of users is larger than the number of available RBs. However, RL-based algorithms such as Q-learning are not suitable for large high-dimensional state-action spaces [13], making them inadequate for addressing the network optimization problems in complex and large-scale scenarios of future wireless networks.

To overcome the aforementioned challenges, recent studies have been applying deep RL (DRL) to address the complex resource allocation problems and optimize system performance [25], [26], [27], [28], [29], [30], [31]. In particular, the work in [25] proposed a DRL framework to find an optimal resource management strategy for GF-OMA systems and address dynamic spectrum access issues. In [26], a DRL algorithm based on generative adversarial networks was proposed to minimize power consumption while ensuring high reliability and low latency for orthogonal frequency division multiple access (OFDMA) systems. To further improve the spectral access efficiency and enhance the system performance, DRL-based GF-NOMA schemes were

investigated in [27], [28], [29], [30], [31] under different scenarios. Specifically, the work [27] investigated a pilot sequence-based GF-NOMA system and proposed a centralized training distributed execution multi-agent (MA) DRL (MADRL) solution to maximize the network throughput (number of successfully served users). Additionally, different MADRL-based dynamic resource allocation strategies for power-domain GF-NOMA systems were investigated in [28], [29] to maximize the system throughput [28] and sum rate [29]. In [30], [31], DRL-based methods were proposed for GF-NOMA systems enabling massive URLLC (mURLLC) to maximize the long-term average throughput.

B. CONTRIBUTIONS

Unlike the aforementioned works on GF-NOMA systems, this paper investigates an MADRL-based resource allocation strategy aimed at maximizing the energy efficiency (EE) while satisfying the users' requirements on reliability and latency for URLLC-enabled GF-NOMA (URLLC-GF-NOMA) systems. Given the stringent requirements of reliability and latency of URLLC users, there is a demand for an efficient and rapid communication protocol. Therefore, our focus is on constructing an effective distributed MADRL-based solution that achieves both EE and rapid convergence with minimal signaling overhead. The approach is designed to reduce the information exchange between the environment and agents, based on which the lower processing latency for URLLC users can be achieved. Indeed, we consider a GF-NOMA scenario where the users compete for the RBs, i.e., subchannels (SCs) and transmission power levels (TPLs), to communicate with the BS by randomly selecting one SC and one TPL for their transmissions. Following the NOMA principle, the users utilizing the same SC are distinguished by their received power at the BS, and their messages are decoded in an orderly manner using SIC [8]. However, with its random access nature, GF-NOMA may cause severe interference since too many users can select the same SC, leading to the system performance degradation. To overcome this drawback, we utilize DRL techniques to enable the users to find the most suitable SCs and TPLs for their transmissions, optimizing the network EE, and fulfilling the URLLC requirements of all users. Thus, the main contributions of this paper are summarized as follows:

- Given that EE is an important factor due to users' energy limitations, we investigate the problem of maximizing the long-term average EE for URLLC-GF-NOMA systems. The goal must be achieved while also ensuring the strict requirements of users in terms of reliability and latency, which necessitates a rapid and efficient transmission protocol. Building on this EE maximization problem, we further investigate the objectives of maximizing the sum rate and minimizing power consumption to clarify the benefits of the proposed problem in balancing the achievable sum rate against power consumption for energy-limited users.
- We develop three distributed MADRL-based resource allocation methods to address the considered problem: MA Dueling Double Deep Q Network (MA3DQN), MA Double Deep Q Network (MA2DQN), and MA Deep Q Network (MADQN). Within this context, the MADRL frameworks are designed to provide energy-efficient learning-based solutions which ensure rapid convergence and minimal signaling overhead, ultimately reducing the processing latency for URLLC users.
- We provide a performance comparison between the proposed mechanisms and other benchmark schemes to clarify the benefits of the former in terms of convergence property and EE performance. Additionally, we evaluate the effects of different state-action spaces, URLLC requirements, and the number of users on the achieved rewards and EE performance. The provided numerical results prove that the proposed solutions outperform other benchmark schemes, achieving higher EE, faster convergence, and reduced signaling overhead.

The remainder of the paper is organized as follows. Section II presents the system model, URLLC method, and the EE maximization problem. Section III describes the MADRL-based solution of the EE optimization problem for the considered URLLC-GF-NOMA system. Section IV provides the obtained simulation results and discussions. Finally, Section V concludes this paper. For clarity, we provide a summary of the main notations and symbols used in this paper in Table 1.

II. SYSTEM MODEL

We consider an uplink URLLC-GF-NOMA system consisting of one base station (BS) and a set of M URLLC users, denoted by $\mathcal{M}$, allocated uniformly around the BS within a circle-cell radius of r_c (m), as shown in Fig. 1. The system bandwidth is equally divided into a set of K orthogonal SCs, denoted by $\mathcal{K}$, to serve the users. Moreover, the GF-NOMA transmission strategy is utilized to improve the spectrum access efficiency and guarantee strict requirements of the URLLC users in overloaded scenarios, i.e., $M > K$. Following this transmission scheme, the users utilize the available SCs to communicate with the BS, and multiple users can share the same SC based on the power-domain NOMA principle [7].

In 5G new radio (5G-NR) networks, the SC's bandwidth is defined as 2^{ν} times of SC's bandwidth in 4G systems (i.e., 180 kHz), where $\nu \in \{0, 1, 2, 3, 4\}$ denotes the numerology index which stands for the various SC types in order to support different services [32], [33]. In particular, the SC with higher bandwidth is used for URLLC service while other services such as enhanced mobile broadband (eMBB) and massive machine type communications (mMTC) can utilize the numerology with smaller SC spacing. Given this context, this paper considers a scenario where the total bandwidth is divided into a set of SCs, i.e., $\mathcal{K}$, serving the URLLC users, and the bandwidth of SCs is defined as $W = 2^{\nu} \times 180$ (kHz).

TABLE 1. Main notations and symbols.

Notation	Description
$\mathcal{M}$	The set of users
$\mathcal{K}$	The set of SCs
L	The number of TPLs
P_{max}	The maximum transmission power of users
r_c	The cell radius
W	SC bandwidth
ν	The numerology index
$x_m^{(k)}(t)$	Binary SC allocation variable for user m over SC k in time-slot (TS) t
$\|\cdot\|$	The absolute value
$\mathcal{CN}\left(0,\sigma^2\right)$	A scalar complex Gaussian distribution with zero mean and variance σ^2
$P_m^{(k)}(t)$	The transmission power of user m over SC k in TS t
$\hat{P}_l$	The l-th TPL
$h_m^{(k)}(t)$	The channel coefficient of the link from user m to the BS over SC k in TS t
$u_m^{(k)}(t)$	The message of user m transmitted on SC k in TS t
$\mathbb{E}[\cdot]$	The expectation operator
$Q(x)$	The Gaussian Q-function
$Q^{-1}(x)$	The inverse of the Gaussian Q-function
$\gamma_m^{(k)}(t)$	The received signal-to-interference-plus-noise ratio (SINR) according to user m over SC k in TS t
$R_m^{(k)}(t)$	The achievable rate of user m over SC k in TS t
n_b	The packet size
τ	The transmission latency
ε_m	The decoding error probability of user m
$\mathcal{E}(t)$	Energy efficiency in TS t
α	The learning rate
γ	The discount factor
$s_m(t)$	The network state of agent m at TS t
$a_m(t)$	The action of agent (user) m at TS t
$r(t)$	The reward function at TS t

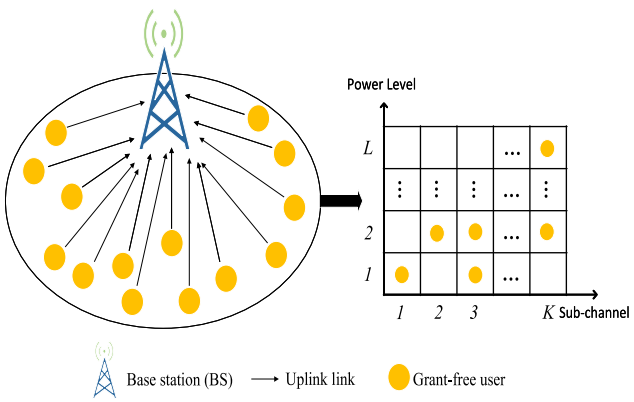


FIGURE 1. Illustration of an uplink URLLC-GF-NOMA system.

A. UPLINK GF-NOMA TRANSMISSION PROCESS

Under the GF strategy, the users are free to choose the SCs for their transmissions without any scheduling instructions from the BS. However, this can lead to severe collision issues as too many users may select the same SCs. To mitigate this drawback, the NOMA technique can be applied, where multiple users can access the same SC. Considering the NOMA transmission process over SC k ($k \in \mathcal{K}$) in time slot (TS) t, we denote $x_m^{(k)}(t)$ as a binary SC allocation variable, where $x_m^{(k)}(t) = 1$ if user m occupies SC k and $x_m^{(t)} = 0$ otherwise. The set of users occupying SC k in TS t is described as $\mathcal{M}^{(k)}(t) = \{m|x_m^{(k)}(t) = 1, m \in \mathcal{M}\}$. Let M_k be the number of users using SC k in TS t, i.e., $\sum_{k=1}^{K} M_k = M$. Then, the received signal at the BS over SC k in TS t is given by

$$y^{(k)}(t) = \sum_{m=1}^{M_k} \sqrt{P_m^{(k)}(t)} h_m^{(k)}(t) u_m^{(k)}(t) + n(t), \tag{1}$$

where $n(t) \sim \mathcal{CN}(0, \sigma^2)$ is the additive white Gaussian noise (AWGN), $P_m^{(k)}(t)$ and $u_m^{(k)}(t)$ denote the transmission power and the transmitted message of user m over SC k in TS t, respectively. Herein, the transmission power is defined as $P_m^{(k)}(t) = 0$ if $x_m^{(k)}(t) = 0$, otherwise, $P_m^{(k)}(t) \neq 0$. Besides, $h_m^{(k)}(t)$ represents the channel coefficient between user m and the BS over SC k in TS t.

We assume that the users using SC k are sorted in the descending order of the corresponding received power level at the BS, i.e., $\mathcal{P}_1^{(k)}(t) \geq \cdots \geq \mathcal{P}_{M_k}^{(k)}(t)$, where $\mathcal{P}_m^{(k)}(t) = P_m^{(k)}(t)|h_m^{(k)}(t)|^2$. Following the NOMA principle, the messages of the users with higher received power level are decoded earlier at the BS. Specifically, the BS decodes the message of a user by treating the messages of users with lower received power level as noise [11], [34]. It then reconstructs and removes this component from the received signal to decode the remaining users' messages successively by using the SIC technique. Accordingly, the received signal-to-interference-plus-noise ratio (SINR) of user m over SC k in TS t is expressed as

$$\gamma_m^{(k)}(t) = \frac{\mathcal{P}_m^{(k)}(t)}{\sum_{i=m+1}^{M_k} \mathcal{P}_i^{(k)}(t) + \sigma^2}. \tag{2}$$

B. URLLC COMMUNICATION MODEL

Due to the stringent low-latency requirement of URLLC communication, very short packets and finite blocklength (FBL) is implemented for data transmission, so-called short-packet communications (SPC). Consequently, the Shannon-related capacity formula cannot be applied to the URLLC communication model since it is designed under the assumption of the infinite block length (iFBL). According to [5], the achievable rate of user m over SC k in the FBL regime for a quasi-static flat fading channel can be approximated as

$$R_m^{(k)}(t) \approx W\left[\log_2\left(1+\gamma_m^{(k)}(t)\right) - \sqrt{\frac{v_m^{(k)}(t)}{\tau W}} Q^{-1}(\varepsilon_m)\right], \tag{3}$$

where $v_m^{(k)}(t) = 1 - \frac{1}{(1+\gamma_m^{(k)}(t))^2}$ is the channel dispersion, τ denotes the transmission latency threshold, ε_m is the decoding error probability, and $Q^{-1}(x)$ represents the inverse of the

Gaussian Q-function $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$. Based on (3), one can define an SNR threshold for user m trying to transmit one packet over one SC k in each transmission TS that satisfies the URLLC requirements (i.e., τ and ε_m) as [35]

$$\hat{\gamma}_m = 2^{\frac{n_b}{\tau W} + \frac{Q^{-1}(\varepsilon_m)}{\ln 2\sqrt{\tau W}}} - 1, \tag{4}$$

where n_b (bits) is the packet size. From (4), the target rate for the transmission of user m can be defined as

$$\hat{R}_m \approx W\left[\log_2(1+\hat{\gamma}_m) - \sqrt{\frac{\hat{v}_m}{\tau W}} Q^{-1}(\varepsilon_m)\right], \tag{5}$$

where $\hat{v}_m = 1 - \frac{1}{(1+\hat{\gamma}_m)^2}$. Similar to [28], [35], we assume that each user m can transmit its packet only once. As the interference over an SC increases, the likelihood of packet drops escalates. Specifically, a successful transmission occurs if $R_m^{(k)}(t) \geq \hat{R}_m$; otherwise, any deviation from this condition results in a failed transmission, i.e., a dropped packet.

C. ENERGY EFFICIENCY MAXIMIZATION

Energy efficiency (EE) is considered one of the major goals in 5G and beyond wireless networks [36]. Furthermore, the majority of mobile devices operate on limited battery power [36], resulting in the need to design energy-efficient communication methods. To address this concern, we first define an EE factor with the purpose of ensuring the achievable rate requirement while reducing the power consumption for the system as follows:

$$\mathcal{E}(t) = \frac{\sum_{k=1}^{K}\sum_{m=1}^{M_k} x_m^{(k)}(t) R_m^{(k)}(t)}{MP_c + \sum_{k=1}^{K}\sum_{m=1}^{M_k} P_m^{(k)}(t)}, \tag{6}$$

where P_c denotes the circuit power consumption. In what follows, the work focuses on designing an effective distributed power control and SC assignment strategy for URLLC-GF-NOMA systems to maximize the average EE while ensuring the URLLC requirements of all users. This can have a direct impact on the overall sustainability and cost-effectiveness of the considered networks. The design objective can be cast by the following problem:

$$\max_{x,P} \ \mathbb{E}_t[\mathcal{E}(t)] \tag{7a}$$

$$\text{s.t.} \ \sum_{k=1}^{K} x_m^{(k)}(t) R_m^{(k)}(t) \geq \hat{R}_m, \quad \forall m \tag{7b}$$

$$\mathcal{P}_1^{(k)}(t) \geq \mathcal{P}_2^{(k)}(t) \geq \cdots \geq \mathcal{P}_{M_k}^{(k)}(t), \quad \forall k \tag{7c}$$

$$\sum_{k=1}^{K} x_m^{(k)}(t) \leq 1, \quad \forall m \tag{7d}$$

$$\sum_{k=1}^{K} P_m^{(k)}(t) \leq P_{\max}, \quad \forall m, \tag{7e}$$

where $\mathbb{E}_t[\cdot]$ is the expectation operation over TSs, $\mathbf{x}$ and $\mathbf{P}$ denote the SC assignment and power control strategies, respectively. The constraint (7b) represents the rate condition to guarantee the users' URLLC requirements. The constraint (7c) ensures the NOMA-based multi-user decoding process. The constraint (7d) implies that each user selects at most one SC. The constraint (7e) shows the users' power budget.

Remark 1: It is noteworthy that the EE maximization problem defined in (7) can also include the objectives of maximizing the sum rate and minimizing the power consumption. These objectives can be attained by setting the denominator and numerator as 1, respectively. Thus, the considered scenario represents a general case where an efficient solution, striking the trade-off between the achievable sum rate and power consumption, can be achieved. Further evaluation on this matter is provided in Section IV.

III. MADRL-BASED ENERGY EFFICIENCY RESOURCE ALLOCATION SOLUTION FOR URLLC-GF-NOMA SYSTEMS

The problem described in (7) is challenging to solve due to its non-convex nature and NP-hard complexity. Moreover, with the GF access method, the users can select their preferred SC and transmission power independently in each TS without requiring admission approval from the BS. While this feature can reduce the access latency and increase the connectivity density, it also necessitates a decentralized optimization solution. Therefore, to effectively address the problem stated in (7), we consider an MADRL-based method, which can be implemented in a distributed manner.

A. MADRL FRAMEWORK

RL is one of the machine learning methods that enable a learning agent to achieve its specific goal with the best long-term reward by interacting with the environment in a trial-and-error manner [29]. In particular, an agent interacts with the environment by taking an action selected from its action space at the current state. It then receives a respective reward and moves to a new state. These procedures are repeated until convergence is observed, where the learning policy of the agent achieves an optimal value in terms of average reward. This learning process can be formulated as a Markov decision process (MDP) with a tuple of four elements $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P})$, defined as follows:

- $\mathcal{S}$: The set of states in the environment, where $s(t) \in \mathcal{S}$ denotes the state of an agent at TS t.
- $\mathcal{A}$: The set of actions that an agent can take, where $a(t) \in \mathcal{A}$ is the action of an agent at TS t.
- $\mathcal{R}$: The reward function, where $r(t)$ represents the immediate reward of the agent at TS t by performing action $a(t)$ in state $s(t)$.
- $\mathcal{P}$: The probability distribution function of the state transition, where $\mathcal{P}(s(t), s(t+1))$ denotes the state transition probability from state $s(t)$ to state $s(t+1)$.

In the considered URLLC-GF-NOMA system, the behavior of all users (i.e., transmission power and SC selection) can be modeled as an MA MDP (MAMDP), which is denoted by $(\{\mathcal{S}\}_{m=1}^{M}, \{\mathcal{A}\}_{m=1}^{M}, \mathcal{R}, \mathcal{P})$. Unlike a single-agent DRL related to the learning process of only one single agent, our proposed

MADRL-based model involves a set of agents $\mathcal{M}$, where all agents operate autonomously and concurrently in a sharing environment. In particular, each agent m observes its current state $s_m(t) \in \mathcal{S}_m$ from the environment and performs an action $a_m(t)$ chosen from its own action space $\mathcal{A}_m$. The joint action of all agents can be formulated as $a(t) = \{a_1(t), a_2(t), \ldots, a_M(t)\}$. The agent m then moves from the current state $s_m(t)$ to a new state $s_m(t+1)$. All agents then receive a reward of $r(t+1)$ and perform an update of their current policy according to the feedback from the environment. It is worth noting that each agent having a distinct reward may result in selfish behavior, leading to a reduction of the global network performance [37]. Therefore, we assume that all agents have a common reward to obtain the global optimum. The main elements of the proposed MADRL approach are defined as follows:

- *State:* Due to users' independence and URLLC requirements, the state of agent (user) $m \in \mathcal{M}$ is designed only based on the local information available at this agent to reduce the processing latency and the signaling overhead in information exchange between the agent and environment. Specifically, the state of agent m in TS t can be defined as the combination of SC index and transmission power value it selected in the previous TS $t-1$, which is expressed as

$$s(t) = \left\{k_m(t-1), P_m^{k_m(t-1)}(t-1)\right\}, \tag{8}$$

  where $k_m(t-1)$ and $P_m^{k_m(t-1)}(t-1)$ are the selected SC index and transmission power of agent m. Since the users' selection of SC and transmission power will impact the overall EE, it is reasonable to include this information in the defined state. From (8), the state of agent m has a cardinality of 2. It is noteworthy that the state definition in (8) differs from those in recent related works on GF-NOMA systems, which require a large signaling overhead in information exchange between the environment and the agents during the learning process [28], [29]. A performance comparison between different state definitions will be provided in Section IV.
- *Action:* At the beginning of TS t, agent m selects an SC and transmission power for its transmission. As a feasible solution, the discrete power domain has been widely used for the learning-based GF-NOMA systems in the literature [21], [27], [29]. This approach can ensure stable convergence and reduce the computational complexity of the distributed learning models conducted by the users who have limited computational resources. Given this context, we consider a discrete action space, where the power is quantized into L levels which are determined as $\hat{P}_l = lP_{max}/L$, $l \in \{1, 2, \ldots, L\}$, where $\hat{P}_l$ is the l-th TPL. Thus, the action of user m in TS t is defined as

$$a_m(t) \in \mathcal{A}_m = \{1, \ldots, kl, \ldots, KL\}, \tag{9}$$

  where $a_m(t) = kl$ indicates that agent m selects SC k and TPL l in TS t. Thus, the action space size of agent m is KL and the overall action space size of all agents is determined as $(KL)^M$.
- *Reward:* After all agents take their chosen actions, they receive an immediate reward from the environment reflecting if their transmissions are successful or not, i.e., if all constraints in the problem (7) are satisfied or not. In the MADRL frameworks, both centralized and decentralized rewards can be considered to build learning models. The centralized-reward mechanism yields a common reward to all agents, whereas in decentralized-reward schemes, each agent receives a distinct reward. However, the decentralized-reward strategy can lead to selfish behavior among agents. They may compete with others to maximize their own rewards, which potentially results in a degradation of overall system performance. To circumvent this issue, a common reward can be implemented to align the agents towards a shared global objective [37]. Since the objective is to maximize the network EE, we use the achieved EE to formulate the reward function. Furthermore, all agents receive the same reward with the aim of achieving the common objective, i.e., optimizing the network EE and guaranteeing URLLC requirements of all users. Thus, the reward function is defined as

$$r(t) = \begin{cases} \mathcal{E}(t), & \text{if all constraints in the problem (7) are satisfied,} \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

  Based on the reward function defined in (10), it becomes apparent that inappropriate user actions, such as an excessive number of users choosing the same SC, may degrade the system's EE. Consequently, the users will receive a low reward. Throughout the learning process, users explore the environment to find the best policies that will maximize their reward, ultimately leading to optimal EE performance.

The objective of RL algorithms is to find a policy π to maximize the expected reward [38]. Considering the Q-learning algorithm - a popular RL technique, the expected reward achieved by agent m after taking action a_m in state s_m following a policy π can be determined based on the action-value function (or Q-value function) as

$$Q_\pi(s_m, a_m) = \mathbb{E}_\pi\left[\hat{r}(t) | s_m(t) = s_m, a_m(t) = a_m\right], \tag{11}$$

where $\mathbb{E}[\cdot]$ denotes the expectation operator and $\hat{r}(t)$ is the long-term discounted cumulative reward which is given by

$$\hat{r}(t) = \sum_{k=0}^{\infty} \gamma^k r(t+k+1), \tag{12}$$

where γ is the discount factor that determines the weight of the future reward. Based on (11), the optimal Q-function can be calculated as

$$Q^*(s_m, a_m) = \max_\pi Q_\pi(s_m, a_m). \tag{13}$$

Through the Q-learning method, the optimal policy can be found based on the available information $(s_m(t), a_m(t), r(t), s_m(t+1))$. The update equation of the Q-value function of agent m can be expressed as [38]

$$Q(s_m(t), a_m(t)) = Q(s_m(t), a_m(t)) + \alpha\big[y_m(t) - Q(s_m(t), a_m(t))\big], \quad (14)$$

where $y_m(t) = r(t) + \gamma \max_a Q(s_m(t+1), a)$ and $\alpha \in [0, 1]$ is the learning rate.

Although the Q-learning method has been widely adopted in wireless networks for resource management purposes, it only works well under small state-action spaces, which limits its applicability. Its practicality diminishes as the problem size increases, primarily due to two key factors [29]: (i) the need for a lookup table to store Q-values for every possible state-action pair becomes unmanageable in terms of storage complexity when dealing with large-scale problems; and (ii) with a larger state space, many states are rarely visited, resulting in decreased performance. To overcome this drawback, we consider DRL techniques to efficiently solve the proposed problem in (7). In the DRL method, a deep neural network (DNN) is integrated into the framework of Q-learning to reduce the memory size and computational complexity by calibrating and training the DNN's different layers to define the best action for each state instead of using a large storage space (i.e., Q-table) to store all Q-values [39]. In this paper, we propose MADRL-based EE URLLC-GF-NOMA methods, where different DRL techniques including deep Q network (DQN), double DQN (2DQN), and dueling 2DQN (3DQN), are investigated.[1]

B. PROPOSED MADRL ALGORITHMS FOR URLLC-GF-NOMA SYSTEMS

1) MADQN-BASED APPROACH

In this section, we consider a MADQN-based URLLC-GF-NOMA approach. With this method, each agent constructs its own DQN model that consists of two different DNNs: the online and target networks, as depicted in Fig. 2. Specifically, in each TS t, agent m uses the online network for Q-function approximation $Q(s_m(t), a_m(t); \theta_m)$ to select an action $a_m(t) \in \mathcal{A}_m$ at state $s_m(t) \in \mathcal{S}_m$. Here, θ_m represents the parameters (weights) of the agent m's online network. Meanwhile, the target network is used to stabilize the learning process, and its parameters $\hat{\theta}_m$ are updated by copying the parameters θ_m of the online network after a certain number of TSs, which is also known as the parameter update frequency F.

Regarding the action selection at each state, one should consider the trade-off between exploration and exploitation during the learning process to achieve the optimal policy. Given this context, the ϵ-greedy policy can be used for action selection to obtain a balance between the exploitation of the best Q-value function and the environmental exploration [38]. In particular, the ϵ-greedy policy selects an action based on two conditions:

$$a_m(t) = \begin{cases} \text{random action}, & \text{with probability } \varepsilon \\ \arg\max_{a \in \mathcal{A}_m}\{\mathcal{Q}_m(t)\}, & \text{with probability } 1-\varepsilon \end{cases}, \quad (15)$$

where $\mathcal{Q}_m(t) = Q(s_m(t), a; \theta_m)$. Herein, the parameter ϵ determines the level of exploration, and it is usually set to decrease over time to reduce the exploration rate as the learning progresses.

During the learning process, MADQN approach uses the experience replay strategy to achieve learning stability, where the transition in the form of a tuple $(s_m(t), a_m(t), r(t), s_m(t+1))$ is stored in the experience replay memory of each agent m. At each iteration, a mini-batch of experiences is sampled uniformly to train the learning model and update the parameters of the online network θ_m with the purpose of minimizing the loss function defined as

$$L_m(\theta_m) = \big[y_m(t) - Q(s_m(t), a_m(t); \theta_m)\big]^2, \quad (16)$$

where $y_m(t)$ is the target value calculated from the target network as follows:

$$y_m(t) = r(t) + \gamma \max_{a \in \mathcal{A}_m} Q\Big(s_m(t+1), a; \hat{\theta}_m\Big). \quad (17)$$

Given the DQN model of each agent mentioned above, the proposed MADQN-based URLLC-GF-NOMA approach is summarized in Algorithm 1. In particular, in TS t, each agent m observes its current state $s_m(t) \in \mathcal{S}_m$ and takes an independently action $a_m(t) \in \mathcal{A}_m$ selected based on the ϵ-greedy policy in (15). After performing the chosen action, agent m receives a common reward $r(t)$ based on the achieved EE and moves to a new state $s_m(t+1)$. It then stores an experience tuple of $(s_m(t), a_m(t), r(t), s_m(t+1))$ into its experience replay memory, and a minibatch of experiences is sampled for training the online network. The parameters of the online network θ_m are then updated to minimize the loss function in (16) by using the stochastic gradient method, where the target value is given by (17). After a predetermined number

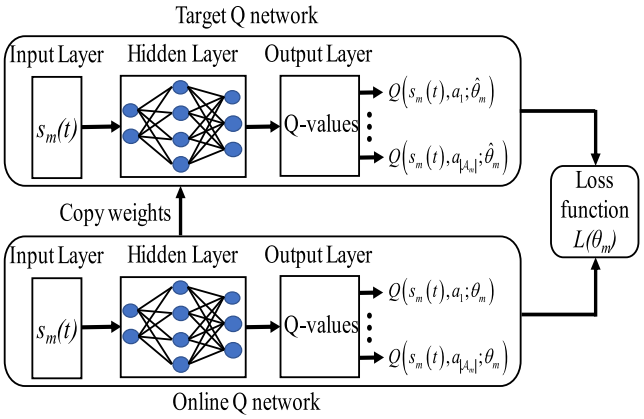


FIGURE 2. Illustration of DQN/2DQN model.

1. Besides DRL algorithms based on Q-learning and DNN, tile coding and on-policy learning could also be promising methods to achieve an effective solution and analytical convergence. This would be a noteworthy issue to investigate in future work.

Algorithm 1 MADRL-Based Energy Efficiency Optimization Algorithm for URLLC-GF-NOMA Systems

Initialize online Q network with random parameters θ_m, $\forall m \in \mathcal{M}$.
Initialize target Q network with parameters $\hat{\theta}_m = \theta_m$, $\forall m \in \mathcal{M}$.
for $e = 1, 2, \ldots, E$ **do**
  Initialize the network state $s_m(t)$, $\forall m$.
  for $t = 1, 2, \ldots, T$ **do**
    All agents select their actions $a_m(t) \in \mathcal{A}_m$, $\forall m$, based on the ϵ-greed policy in (15).
    All agents take their actions, receive a common reward $r(t)$, and move to the next state $s_m(t+1)$.
    for $m = 1, 2, \ldots, M$ **do**
      Store an experience tuple of $(s_m(t), a_m(t), r(t), s_m(t+1))$ to the replay memory of agent m.
      Randomly sample a mini-batch of experience from the replay memory for training.
      Determine the loss function $L(\theta_m)$ as follows:
        - **MADQN approach:** Using (16) and (17).
        - **MA2DQN approach:** Using (16) and (18).
        - **MA3DQN approach:** Using (16) and (18),
      where the Q-value (action-value) functions are calculated by utilizing (19).
      Update θ_m by using stochastic gradient to minimize $L(\theta_m)$.
      Update $\hat{\theta}_m$ as $\hat{\theta}_m = \theta_m$ after every F TSs.
    end for
  end for
end for

of TSs, the parameters of the target network $\hat{\theta}_m$ are updated by copying θ_m. The above training process continues until reaching a predefined number of episodes guaranteeing the algorithm's convergence.

2) MA2DQN-BASED APPROACH

From (17), one can observe that the MADQN approach based on DQN model using the same Q-value function for both tasks, i.e., action selection, $\max_{a\in\mathcal{A}_m} Q(s_m(t+1), a; \hat{\theta}_m)$, and action estimation, $Q(s_m(t+1), a; \hat{\theta}_m)$. This can lead to an unstable learning process since the Q-value function is estimated over-optimistically. To mitigate this issue, we investigate an MA2DQN-based URLLC-GF-NOMA approach, where 2DQN model is considered [40], as shown in Fig. 2. In this method, the action selection and evaluation are decoupled to avoid the overestimation issue by replacing the target value in (17) with the following one

$$y_m(t) = r(t) + \gamma Q\left(s_m(t+1), \arg\max_{a\in\mathcal{A}_m} \mathcal{Q}_m(t+1); \hat{\theta}_m\right), \quad (18)$$

where $\mathcal{Q}_m(t+1) = Q(s_m(t+1), a; \theta_m)$. As can be seen from (18) that the online network $Q(s, a; \theta_m)$ is used for the action selection, whereas the target network $Q(s, a; \hat{\theta}_m)$ is applied to estimate the action. The MA2DQN-based URLLC-GF-NOMA algorithm is also summarized in Algorithm 1 with **MA2DQN** remark in **Step 11**.

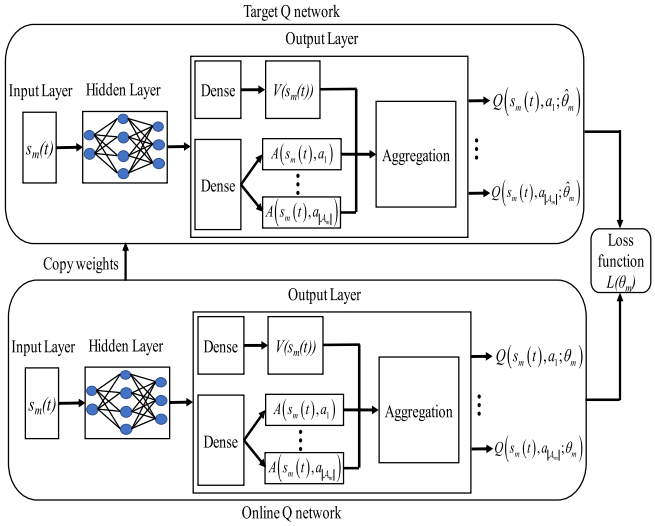


FIGURE 3. Illustration of 3DQN model.

3) MA3DQN-BASED APPROACH

An MA3DQN-based URLLC-GF-NOMA approach is studied in this section. This method uses a 3DQN model whose structure is depicted in Fig. 3, to speed up the convergence and improve the learning efficiency [41]. Following MA3DQN approach, each agent m creates its own 3DQN model based on 2DQN, where the last layer of the 2DQN model is split into two parts to evaluate the state value function (SVF) $V(s_m(t))$ and the advantage function (AF) $A(s_m(t), a_m(t))$. Herein, the SVF $V(s_m)$ is used for estimating the quality (goodness or badness) of a given state $s_m(t)$, allowing the agent to evaluate the long-term potential of being in that state. Meanwhile, the AF $A(s_m(t), a_m(t))$ captures how much better or worse a specific action is compared to other actions in state $s_m(t)$. This allows the agent to choose the best action to take in a given state. The two parts are then combined to produce the final action-value function $Q(s_m(t), a_m(t); \theta_m, \theta_m^V, \theta_m^A)$ that is used to select actions in the environment. Here, θ_m^V and θ_m^A denote the parameters according to SVF-related and AF-related parts, respectively. Given this context, the action-value function determined by agent m for a given state $s_m(t)$ and action $a_m(t)$ is calculated as follows:

$$Q\left(s_m(t), a_m(t); \theta_m, \theta_m^V, \theta_m^A\right) = V(s_m(t)) + A(s_m(t), a_m(t)) - \frac{1}{|\mathcal{A}_m|}\sum_{a\in\mathcal{A}_m} A(s_m(t), a_m(t)), \quad (19)$$

where the last term of the right-hand side of (19) is the mean of the AF over all actions. It is subtracted from the

AF $A(s_m(t), a_m(t))$ of a specific action to ensure that the AF is centered around zero, making it easier to train the network. This approach improves the convergence and stability of the network and enables the effective separation of the estimation of SVF and AF, resulting in better performance compared to DQN and 2DQN architectures. The MA3DQN-based URLLC-GF-NOMA approach is also cast by Algorithm 1 under the designation **MA3DQN** mentioned in **Step 11**.

C. ANALYSIS OF THE PROPOSED METHODS

1) COMPLEXITY ANALYSIS

Let H, N_h, and I_s be the number of training layers (input, hidden, and output layers), the number of neurons in layer h, and the size of the input layer. For each TS, the computational complexity of URLLC-GF-NOMA algorithms based on MADQN and MA2DQN can be calculated by

$$C_{\mathsf{TS}} = \mathcal{O}(X), \tag{20}$$

where $X = I_s N_1 + \sum_{h=1}^{H-1} N_h N_{h+1}$. For the training phase with M agents, E episodes, and T TSs, the computational complexities of the algorithms can be given by

$$C_{\mathsf{MADQN}} = C_{\mathsf{MA2DQN}} = MET \times C_{\mathsf{TS}} = \mathcal{O}(METX). \tag{21}$$

Taking the MA3DQN-based URLLC-GF-NOMA algorithm into account, it has higher complexity than MADQN and MA2DQN-based algorithms due to the implementation of the dueling network architecture. Specifically, its complexity can be determined as

$$C_{\mathsf{MA3DQN}} = \mathcal{O}(MET(X + N_{H-1})). \tag{22}$$

2) CONVERGENCE ANALYSIS

The convergence of a multi-agent system relies on whether the combined strategy of the agents ultimately approaches the optimal state (Nash equilibrium), ensuring the stability of the solution. In this paper, we propose URLLC-GF-NOMA methods based on MADQN, MA2DQN, and MA3DQN, which combine the conventional Q-learning and neural networks. To analyze the convergence of these methods, two key aspects need to be addressed [42]: (i) demonstrating the ability of the conventional Q-learning to converge to the optimal state, and (ii) verifying that the neural network approach effectively identifies or approximates the nonlinear Q-values generated by the general Q-learning iteration as depicted in equation (14). In particular, it has been shown in [43] that the conventional Q-learning algorithm guarantees the attainment of the optimal state when the learning rate α_t satisfies $0 \leq \alpha_t \leq 1$, $\sum_t \alpha_t = \infty$, and $\sum_t \alpha_t^2 < \infty$. Additionally, based on [44], it is established that the neural network can approximate any nonlinear continuous function when adequately sized and suitably initialized. Thus, the convergence of our proposed methods can be guaranteed. It is noteworthy that as mentioned in [45], the theoretical analysis of the neural network's size and initial conditions for ensuring its convergence before training poses challenges due to the complex quantitative relationship between the network convergence and hyperparameters. Therefore, we utilize simulations to demonstrate the convergence of our proposed methods.

3) SOLUTION ANALYSIS

To clarify the difference between the scenario considered in this paper and the ones investigated in related works on RL-based GF-NOMA [27], [28], [29], [31], this section provides a solution summary examined in these works, as shown in Table 2. As can be seen from this table, different DRL frameworks have been proposed to address the unique problems of GF-NOMA systems effectively. In delay-sensitive RL-based systems, signaling overhead is a key performance indicator. It is defined as the number of information bits needed to feed back the channel status data, SC indicators, and the transmission power of a specific user over an SC [46]. Also, the total number of users and SCs, and the exchange of states as well as rewards between the agents and environment can affect the signaling overhead. Higher signaling overhead results in larger processing latency for users.

Following [46], it is assumed that transmitting a continuous value of channel status, data rate, and reward requires 16 bits. Additionally, 1 bit is allocated for acknowledgment (ACK) feedback, 2 bits for decoding status, and 4 bits for the SC indicator, transmission power, and other relevant parameters. The work [27] produces a large signaling overhead because it depends on the decoding status of $\hat{K}$ pilot sequences, users' average throughput, and parameters (weights) of the centralized-training MADRL model transmitted from the BS to users who build local DRL models for distributed execution. These parameters depend on the number of input, hidden, and output layers (A) and the number of neurons per layer (N_a, $1 \leq a \leq A$). In addition, large signaling overhead can be observed in [28], [29] due to the inclusion of various feedback information. This includes the channel status and ACK information of each user [28], as well as users' data rate [29]. In [31], the BS decides the actions for users (the selection of repetition value and contention transmission unit (CTU)), hence, the signaling overhead depends on the feedback information from the BS to the users regarding the selected actions for the transmission of each user. Note that V_{cc}, V_{ic}, V_{sc}, V_{sd}, and V_{ud} used in Table 2 stand for the number of collision CTUs, idle CTUs, singleton CTUs, successfully served users, and failure decoding users, respectively. In our method, only the reward feedback is required to reduce the signaling overhead, but still guarantee an effective learning solution. Consequently, the signaling overhead is determined by the reward feedback.

IV. SIMULATION RESULTS

In this section, the simulation results are provided to evaluate the performance of the proposed MADRL-based resource allocation methods for the considered URLLC-GF-NOMA system. The simulations were performed on an Intel core

TABLE 2. Solution summary of related works.

References	[27]	[28]	[29]	[31]	Our paper
Optimization Problem	Throughput	Throughput	Sum rate	Throughput	Energy efficiency
Solution	Centralized-training and distributed-execution MADRL	Distributed MADRL	Distributed MADRL	Centralized MADRL	Distributed MADRL
State	Decoding states and average throughput	User's action, CSI, and ACK	Users' achievable rate	V_{cc}, V_{ic}, V_{sc}, V_{sd}, V_{ud}	User's selected SC index and TPL
Action	Pilot sequence	SC and TPL	SC and TPL	Repetition value and CTU	SC and TPL
Reward	Throughput	Throughput	Sum rate	Throughput	Energy efficiency
Signaling Overhead	$\underbrace{2\hat{K}+M}_{\text{State}}+\underbrace{\sum_{a=1}^{A}4N_a}_{\text{Parameters}}$	$\underbrace{16KM+M}_{\text{State}}+\underbrace{4}_{\text{Reward}}$	$\underbrace{16KM}_{\text{State}}+\underbrace{16}_{\text{Reward}}$	$\underbrace{8M}_{\text{Action}}$	$\underbrace{16}_{\text{Reward}}$

i7-8665U CPU with 1.9 GHz frequency, 16 GB of random access memory (RAM), and 64-bit Windows 10 operating system. The learning models were considered with three hidden layers, including 256, 128, and 64 neurons. The experimental parameters are provided in Tables 3. Besides the proposed URLLC-GF-NOMA approaches based on MADQN, MA2DQN, and MA3DQN, we further investigate the following methods for comparison purpose.

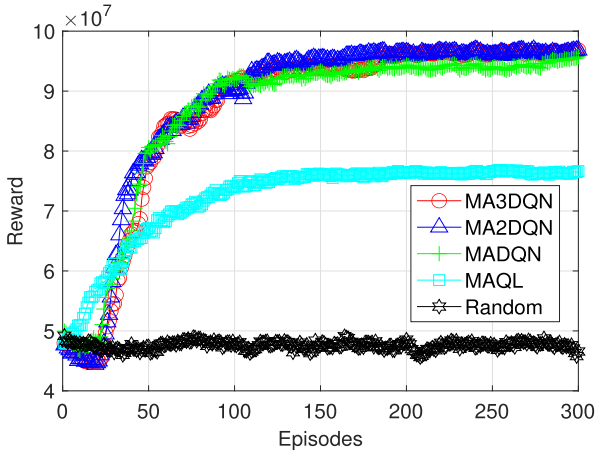
- *MA Q-learning (MAQL) [21]:* MAQL is applied for GF-NOMA systems in [21]. With this scheme, each agent builds its own Q-table to store Q-values of all possible state-action combinations during learning process.
- *Random approach:* In this scheme, users randomly select SC and TPL for their transmissions without learning.
- *Exhaustive search (ES):* This method determines the optimal solution through exploration of the entire network space in every TS.
- *GF-OMA method:* This method explores GF-OMA scheme, where the users utilize distinct frequency/time domains for their transmissions [47].
- *Different state spaces [28], [29]:* Various state spaces for MADRL-based GF-NOMA systems introduced in [28], [29] are also considered to assess the proposed methods' efficiency in terms of convergence property and signaling overhead. Specifically, the network state defined for agent *m* in [28], named State 1, consists of its action, its channel gains over all SCs, and its transmission outcome. Meanwhile, the work [29] defines agent *m*'s state, so-called State 2, as the combination of the achievable rates of all agents.

Fig. 4(a) shows the convergence behavior during the training phase of the URLLC-GF-NOMA approaches based on MA3DQN, MA2DQN, MADQN, MAQL, and Random schemes by plotting the reward achieved by all agents with respect to the various number of episodes. As can be observed from this figure, the Random method achieves the worst performance (i.e., lowest reward) as compared to other schemes. This is because the users randomly select SC and

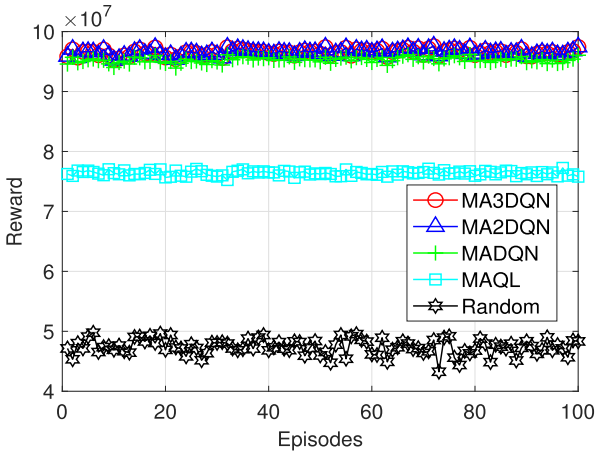
TABLE 3. Simulation parameters.

Parameters	Value
Cell radius (r_c)	500 m
Channel model	Rayleigh
Number of users (M)	$\{2; 4; 6; 8; 10\}$
Number of SCs (K)	$\{2; 3\}$
Reliability requirement (ε_m)	$\{10^{-1}, 10^{-3}, 10^{-5}, 10^{-7}\}$
Latency threshold (τ)	$\{0.5; 1; 1.5; 2\}$ ms
Numerology index (ν)	2
Number of transmit power levels (L)	$\{2; 4; 6; 8; 10\}$
Circuit power consumption (P_c)	0.05 W
Noise power (σ^2)	-174 dBm/Hz
Packet size (n_b)	256
Number of episodes (E)	500
Number of learning steps (T)	100
Number of hidden layers	3
Number of neurons per hidden layer	$\{256, 128, 64\}$
ϵ-greedy policy	$\epsilon = 1$ and $\epsilon_{min} = 0.001$
Learning rate (α)	0.001
Discount factor (γ)	0.9
Optimizer	Adam

TPL when using this method. It is, therefore, difficult for them to find the best SC and TPL for their transmissions to optimize the network performance and guarantee URLLC requirements. Among the remaining approaches, the MAQL scheme outperforms the Random method thanks to the application of the Q-learning algorithm, but still achieves worse performance than others. This highlights the constraint of the Q-learning method when applied to a dynamic environment with an extremely large state-action space. Taking our proposed URLLC-GF-NOMA methods (i.e., MA3DQN, MA2DQN, and MADQN) into account, they are superior to the MAQL and Random methods, while achieving the same learning behavior and comparable rewards in this simulation. After the training phase, the testing phase is conducted to evaluate the training results, where the users always select the best action with the highest Q-value based on their learning



(a) Training



(b) Testing

FIGURE 4. Convergence analysis with different approaches, where $M = 4$, $K = 2$, $L = 7$.

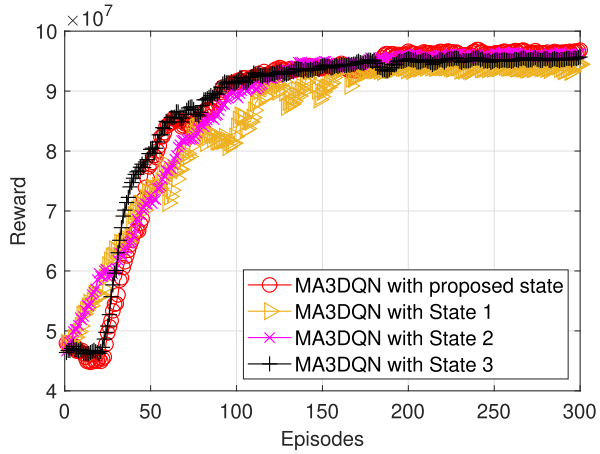
results under new network conditions (network states and channels). The simulation results for the testing phase are provided in Fig. 4(b), where the testing process is performed over 100 episodes. This figure shows that during the testing phase, the learning-based methods (MA3DQN, MA2DQN, MADQN, and MAQL) can guarantee the convergence they achieved in the training phase.

In Fig. 5, we plot the variation of the achieved reward versus the number of episodes when using the MA3DQN approach with different network state definitions. This is to evaluate the efficiency of our proposed methods in terms of convergence and signaling overhead. Specifically, we investigate two network states used for GF-NOMA systems in [28], [29], namely State 1 and State 2, as mentioned earlier. In addition, a channel-based state definition, so-called State 3, is also investigated, where only the channel state information (CSI) of each user is used to define its state. One can see from Fig. 5 that the method utilizing the proposed state in (8) attains rewards comparable to the method that uses State 2 and State 3, and larger than the method utilizing State 1. Furthermore, the proposed state demands lower signaling overhead than State 1, State 2, and State 3. In particular, the proposed state only requires the agents to know their own selected SC index and transmission power value, which are available at the agent. Thus, the environment only needs to provide feedback to the agents regarding their transmission outcomes (i.e., reward), which is used for the training process. Meanwhile, State 1 requires the agents to also have knowledge of their own channel quality and incorporate transmission results into their state information. This unnecessarily increases the input data for the agents' learning model. On the other hand, State 2 requires agents to grasp the achievable rates of all users. This necessitates significant information exchange between the environment and the agents, resulting in high signaling overhead. Moreover, State 3 demands for additional information exchange between the agents and the BS to achieve the CSI, increasing the signaling overhead but does not contribute to further improving the learning process and the system performance in our considered scenario.

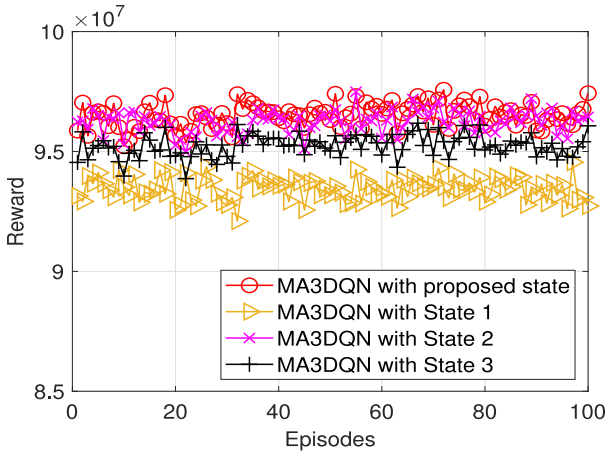
Fig. 6(a) and Fig. 6(b) illustrate the effect of small and large state-action spaces (i.e., number of users (M), SCs (K), and TPLs (L_p)) on the achieved rewards, respectively. Herein, the MA3DQN, MA2DQN, and MADQN approaches using the proposed state and State 2 are considered. As demonstrated by these figures, the methods using the proposed state and those employing State 2 have similar learning behavior and achieve comparable reward values in the small state-action space. However, in the large state-action space, the methods utilizing the proposed state outperform those using State 2. This is because by utilizing the proposed state, the state-action space of the considered methods is significantly reduced compared to that of the methods employing State 2, resulting in a faster learning process and higher achieved rewards for the methods using the proposed state.

Fig. 6(b) also illustrates that the MA3DQN method outperforms the MA2DQN and MADQN methods in the large state-action space generated by State 2. This is due to the MA3DQN approach's ability to rapidly identify optimal actions and important states, leading to better learning outcomes than the MA2DQN and MADQN techniques. The enhanced performance of MA3DQN is achieved by the separation of state and action networks at the last layer of the DNNs model used in these schemes. On the other hand, when the proposed state is employed, it results in a considerably smaller state-action space than State 2, even with an increase in M, K, and L_p, resulting in faster learning. As a result, the MA3DQN, MA2DQN, and MADQN methods employing the proposed state achieve comparable learning outcomes. Thus, the MA3DQN method is developed for problems with a larger state-action space, whereas the MA2DQN and MADQN methods, with a simpler network design, are suitable for problems with smaller state-action spaces.

To evaluate the effect of the URLLC requirements (i.e., ε_m and τ) on the system performance, we plot the variation of the achieved reward versus the number of episodes with

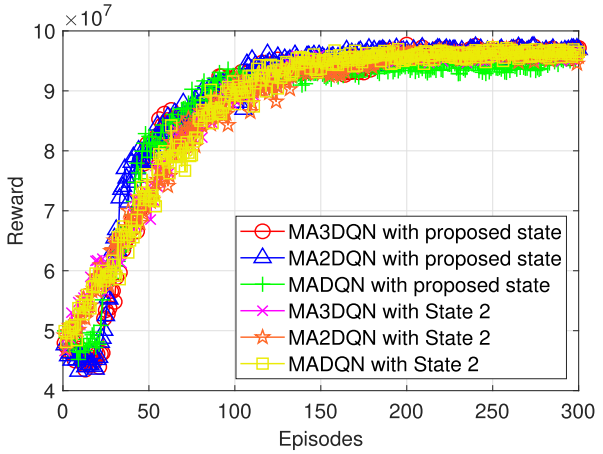


(a) Training

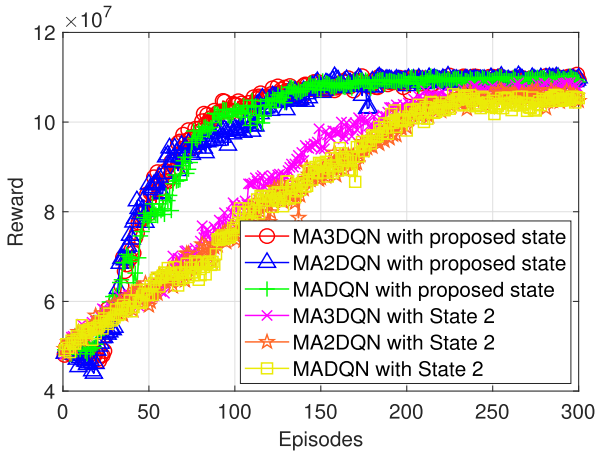


(b) Testing

FIGURE 5. Convergence analysis with different network states and MA3DQN method, where $M = 4$, $K = 2$, $L = 7$.



(a) Small state-action space: $M = 4$, $K = 2$, $L = 7$.



(b) Large state-action space: $M = 8$, $K = 3$, $L = 10$.

FIGURE 6. Effect of state-action spaces on the achieved reward with different approaches.

different value sets of (ε_m, τ), while using the MA3DQN method in Fig. 7. This figure indicates that the achieved reward can converge to a greater value when the lower URLLC requirements are set; for instance, the reliability decreases (i.e., ε_m increases from 10^{-7} to 10^{-1}), and the latency threshold is degraded (i.e., τ increases from 0.5 ms to 2 ms). This can be explained by the fact that the minimum data rate threshold based on (5) gets higher with the increase in the URLLC requirements. It is, thus, more difficult to obtain the rate constraint required to fulfill the URLLC conditions in this case, leading to an EE performance degradation.

Fig. 8 shows the performance comparison in terms of the achieved reward between the methods using GF-NOMA and GF-OMA. For the GF-OMA scheme, each user occupies a distinct resource block and the system bandwidth W is equally divided among the users [47]. Observing Fig. 8 reveals that the methods utilizing GF-NOMA obtain greater reward gains compared to those utilizing GF-OMA. This can be attributed to the performance degradation that occurs in the latter due to the splitting of bandwidth resources among users in the OMA scheme. Moreover, this figure illustrates that in both GF-NOMA and GF-OMA scenarios, the achieved rewards are comparable for the proposed

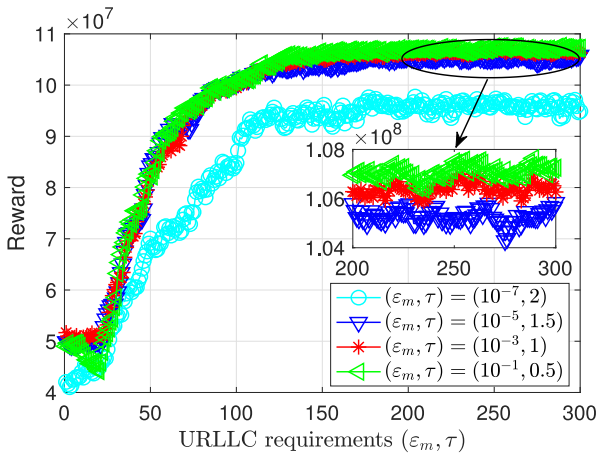


FIGURE 7. Effect of URLLC requirements (ε_m, τ) on the achieved reward, where $M = 4$, $K = 2$, and $L = 10$.

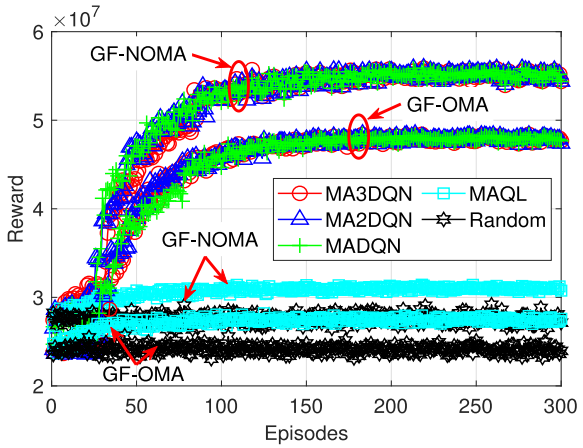


FIGURE 8. Performance comparison between the methods using GF-OMA and GF-NOMA, where $M = 4$, $L = 10$.

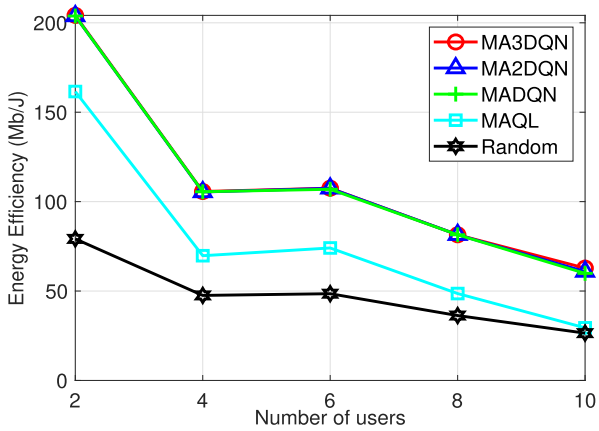


FIGURE 9. Effect of number of users on the EE performance with different approaches, where $K = 2$, $L = 10$.

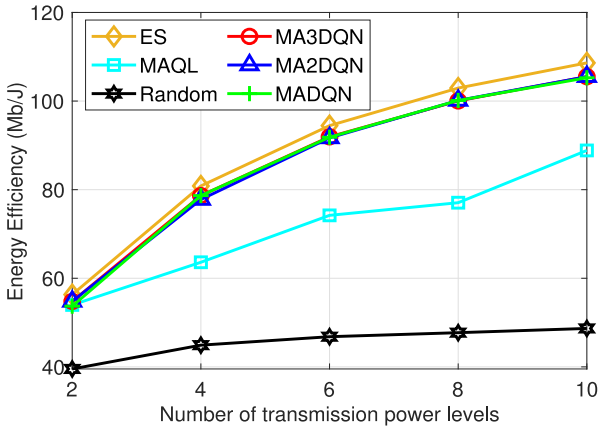


FIGURE 10. EE performance comparison between different methods, where $M = 4$, $K = 2$.

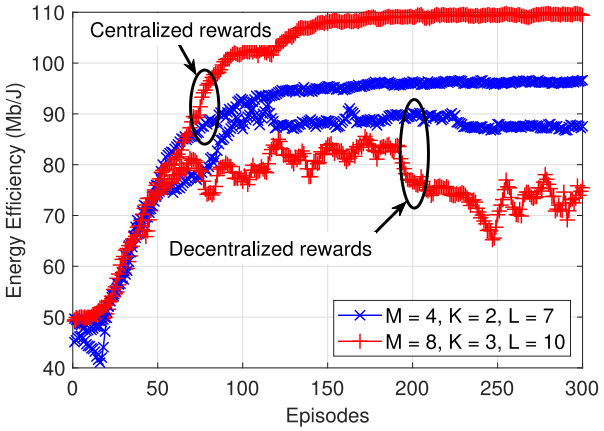


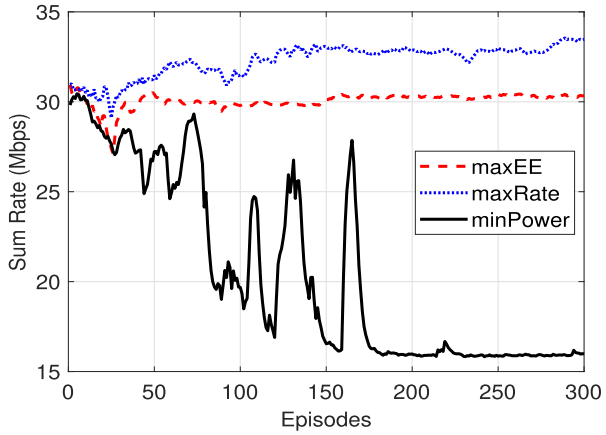
FIGURE 11. EE performance of MADQN method with centralized and decentralized rewards.

MA3DQN, MA2DQN, and MADQN methods, and these approaches outperform the MAQL and Random schemes.

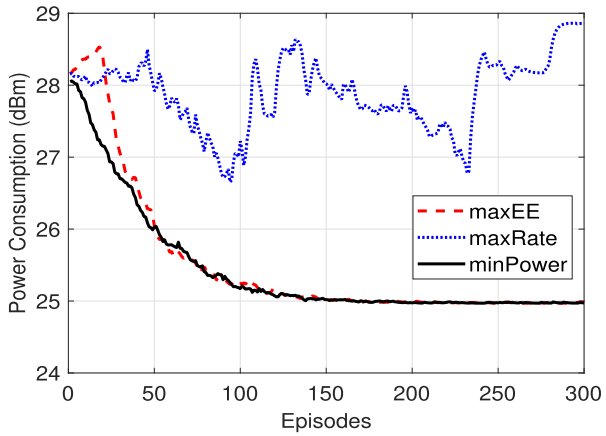
Fig. 9 depicts the variation of the average EE with respect to the number of users (M) for different methods. As observed from this figure, the EE performance decreases as the value of M gets higher since the growth of the number of users sharing the same SCs in this case leads to stronger interference. In addition, the proposed MA3DQN, MA2DQN, and MADQN methods yield better EE performance than the MAQL and Random methods when M increased. Furthermore, they achieve comparable EE gains under the different values of M. As mentioned earlier in the previous results, this is because the proposed approaches produce a small state-action space for each agent, accelerating their learning process and leading to equivalent EE performance.

Fig. 10 provides an EE performance comparison between the investigated methods (i.e., MA3DQN, MA2DQN, MADQN, MAQL, and Random) and an optimal solution obtained through the ES method by plotting the achieved EE versus the number of TPLs. The ES method finds the largest EE by traversing all possible actions in the network in every TS. As illustrated in Fig. 10, the EE values achieved by the MA3DQN, MA2DQN, and MADQN methods are close to those of the ES method and significantly exceed those of the MAQL and Random approaches. It is noteworthy that the ES method is infeasible for large network spaces since it requires exploring the entire network space, leading to high computational complexity. To address this issue, the proposed URLLC-GF-NOMA methods based on MA3DQN, MA2DQN, and MADQN enable the users to interact with the wireless environment and learn from their accumulated experiences to rapidly achieve a near-optimal solution without visiting the entire network space.

Fig. 11 provides an EE performance comparison between MADQN methods using centralized and decentralized rewards with different values of M, K, and L. Specifically, the centralized reward is defined in (10), whereas the decentralized reward implies that each agent can receive a distinct reward depending on its own transmission outcome. In particular, with the objective of maximizing EE, the decentralized reward of each agent m can be defined as $r_m(t) = R_m^{(k)}(t)/P_m^{(k)}(t)$ if its transmission is successful (i.e., $R_m^{(k)}(t) \geq \hat{R}_m$) and $r_m(t) = 0$ otherwise. Herein, $P_m^{(k)}(t)$,



(a) Achievable sum rate.



(b) Power consumption.

FIGURE 12. Achievable sum rate and power consumption of different problems, where $M = 4$, $K = 2$, and $L = 7$.

$R_m^{(k)}(t)$, and $\hat{R}_m$ are defined in (1), (3), and (5), respectively. As can be seen from Fig. 11, the EE performance achieved by using decentralized rewards is much smaller than the cases using centralized rewards. This is due to the fact that employing decentralized rewards can lead to the selfish behavior of agents, where they may compete with each other to maximize their own objective instead of the common one, i.e., maximizing the overall EE while guaranteeing the URLLC requirements of all users. Therefore, a significant global EE performance degradation can be observed as shown in Fig. 11.

As mentioned earlier in Section II-C, the problems of maximizing the achievable sum rate, named as maxRate, and minimizing the power consumption, so-called minPower, can also be investigated based on the EE maximization problem, denoted by maxEE, defined in (7). Herein, maxRate and minPower are achieved by setting the denominator and numerator of (6) as 1, respectively. Given this context, Figs. 12(a) and 12(b) depict the achievable sum rate and the power consumption versus learning episodes for different problems, including maxEE, maxRate, and minPower, respectively. These figures demonstrate that maxRate can obtain the highest sum rate but with the largest power consumption since it only focuses on maximizing the sum rate, leading to high power consumption. Meanwhile, minPower can achieve minimum power consumption but results in a poor achievable sum rate due to its power minimization objective. On the other hand, the proposed maxEE problem can achieve a high sum rate close to that obtained by maxRate while minimizing the users' power consumption. Thus, maxEE outperforms maxRate and minPower in guaranteeing the trade-off between the achievable sum rate and power consumption for energy-limited users.

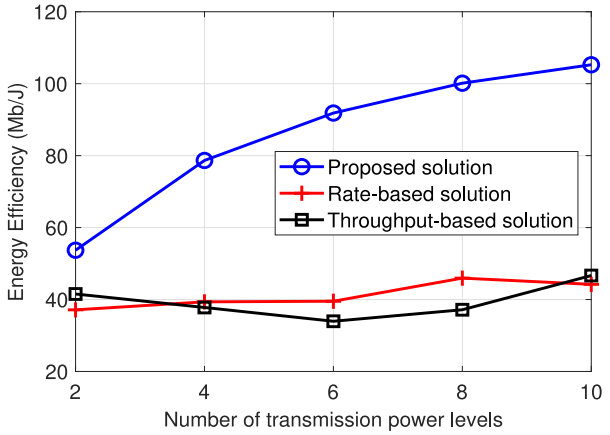


FIGURE 13. EE performance of different MADRL solutions for GF-NOMA systems, where $M = 4$ and $K = 2$.

Fig. 13 provides the EE performance of different MADRL frameworks proposed for GF-NOMA systems including our proposed solution, throughput-based solution [28], and rate-based solution [29]. As can be seen from this figure, our proposed solution achieves much better EE performance than throughput-based and rate-based solutions. This is because our proposed solution aims to maximize EE with minimum transmission power to save energy for those users with limited energy resources. In contrast, the throughput-based method tries to maximize network throughput, hence, higher transmission power than necessary can be used to ensure the successful decoding of the users' messages. Meanwhile, the rate-based solution focuses on maximizing data rate with large transmission power resulting in EE performance reduction.

To clarify the benefits of received power-based decoding order, Fig. 14 shows the EE comparison between received power-based and rate-based SIC methods during the learning process. Here, we consider that the predetermined rate demand of user m ($1 \leq m \leq M$) is set as m bps/Hz. Considering the rate-based SIC method, the message of the user with lower rate demand will be decoded earlier at the BS. This is because the user having its signal decoded earlier would suffer stronger interference and achieve a smaller data rate. As can be observed from Fig. 14, the received power-based SIC outperforms the rate-based SIC in terms of EE. The reason behind this result is that the decoding order in the received power-based SIC method is more flexible than that

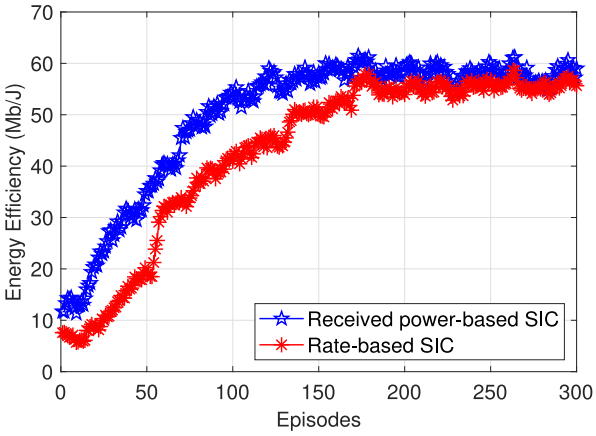


FIGURE 14. EE performance of different SIC methods, where $M = 4$, $K = 2$, and $L = 7$.

in the rate-based SIC approach, which depends on the users' channel conditions and TPL selection. This can help the users find the most appropriate SC and TPL for their transmissions to optimize the global EE performance and satisfy the different rate demands of all users. In contrast, the decoding order is fixed in the rate-based SIC method due to the predetermined rate demand of the users. It is, therefore, difficult for users to find the best learning policy, especially in time-varying and strong-interference environments, leading to performance degradation.

From the results achieved above, it can be concluded that the proposed URLLC-GF-NOMA methods based on MA3DQN, MA2DQN, and MADQN can obtain similar performance and outperform other benchmark schemes in terms of EE, convergence rate, and signaling overhead. However, the methods based on MA2DQN and MADQN exhibit lower complexity compared to the MA3DQN-based method as indicated in Section III-C1, thereby reducing the power consumption and processing latency for the URLLC users. This benefit makes them better suited for the considered URLLC-GF-NOMA system.

V. CONCLUSION

In this paper, we have investigated a resource allocation problem in an uplink URLLC-GF-NOMA system where the users aim to maximize energy efficiency while satisfying their URLLC requirements. To achieve this, we have proposed three MADRL-based URLLC-GF-NOMA approaches (MA3DQN, MA2DQN, and MADQN) for the users to learn how to select the most suitable sub-channel and transmission power for their transmissions. In particular, we have designed an MADRL framework that guarantees a rapid convergence and small signaling overhead to maximize energy efficiency and satisfy users' URLLC requirements. Our simulation results have shown that the proposed URLLC-GF-NOMA methods based on MA3DQN, MA2DQN, and MADQN can achieve similar performance, but MA2DQN and MADQN are more appropriate for the investigated URLLC-GF-NOMA system due to their lower complexity compared to MA3DQN. Moreover, our proposed methods outperform existing benchmark schemes in terms of energy efficiency performance, convergence property, and signaling overhead to guarantee the URLLC requirements of energy-limited users.

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] P. Popovski et al., "Wireless access for ultra-reliable low-latency communication: Principles and building blocks," *IEEE Netw.*, vol. 32, no. 2, pp. 16–23, Mar./Apr. 2018.

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

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

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

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

[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] 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.

[9] C. Xiao et al., "Downlink MIMO-NOMA for ultra-reliable low-latency communications," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 780–794, Apr. 2019.

[10] Z. Wang, T. Lv, Z. Lin, J. Zeng, and P. T. Mathiopoulos, "Outage performance of URLLC NOMA systems with wireless power transfer," *IEEE Wireless Commun. Lett.*, vol. 9, no. 3, pp. 380–384, Mar. 2020.

[11] 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, Apr. 2021.

[12] *5G NR, Physical Layer Procedures for Data, V15.9.0*, 3GPP Standard TS 38.214, Mar. 2020.

[13] C. Jiang, H. Zhang, Y. Ren, Z. Han, K.-C. Chen, and L. Hanzo, "Machine learning paradigms for next-generation wireless networks," *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 98–105, Apr. 2017.

[14] H. Ye, G. Y. Li, and B. Juang, "Power of deep learning for channel estimation and signal detection in OFDM systems," *IEEE Wireless Commun. Lett.*, vol. 7, no. 1, pp. 114–117, Feb. 2018.

[15] H. Huang, W. Xia, J. Xiong, J. Yang, G. Zheng, and X. Zhu, "Unsupervised learning-based fast beamforming design for downlink MIMO," *IEEE Access*, vol. 7, pp. 7599–7605, 2019.

[16] N. C. 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, 4th Quart., 2019.

[17] J. Yu and L. Chen, "Stability analysis of frame slotted aloha protocol," *IEEE Trans. Mobile Comput.*, vol. 16, no. 5, pp. 1462–1474, Jul. 2016.

[18] H. Cao and J. Cai, "Distributed opportunistic spectrum access in an unknown and dynamic environment: A stochastic learning approach," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4454–4465, Jan. 2018.

[19] 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, Oct. 2017.

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

[21] M. V. da Silva, R. D. Souza, H. Alves, and T. 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.

[22] D.-D. Tran, S. K. Sharma, and S. Chatzinotas, "BLER-based adaptive Q-learning for efficient random access in NOMA-based mMTC networks," in *Proc. IEEE Veh. Technol. Conf. (VTC)*, Helsinki, Finland, Apr. 2021, pp. 1–5.
[23] D.-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 *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Madrid, Spain, Dec. 2021, pp. 1–6.
[24] 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 Veh. Technol. Conf. (VTC)*, 2022, pp. 1–5.
[25] Y. Hua, R. Li, Z. Zhao, X. Chen, and H. Zhang, "GAN-powered deep distributional reinforcement learning for resource management in network slicing," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 2, pp. 334–349, Feb. 2020.
[26] A. T. Z. Kasgari, W. Saad, M. Mozaffari, and H. V. Poor, "Experienced deep reinforcement learning with generative adversarial networks (GANs) for model-free ultra reliable low latency communication," *IEEE Trans. Commun.*, vol. 69, no. 2, pp. 884–899, Feb. 2021.
[27] R. Huang, V. W. S. Wong, and R. Schober, "Throughput optimization for grant-free multiple access with multiagent deep reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 228–242, Jan. 2021.
[28] J. Zhang, X. Tao, H. Wu, N. Zhang, and X. Zhang, "Deep reinforcement learning for throughput improvement of the uplink grant-free NOMA system," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6369–6379, Jul. 2020.
[29] M. Fayaz, W. Yi, Y. Liu, and A. Nallanathan, "Transmit power pool design for grant-free NOMA-IoT networks via deep reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 11, pp. 7626–7641, Nov. 2021.
[30] 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)*, 2022, pp. 1–6.
[31] Y. Liu, Y. Deng, H. Zhou, M. Elkashlan, and A. Nallanathan, "Deep reinforcement learning-based grant-free NOMA optimization for mURLLC," *IEEE Trans. Commun.*, vol. 71, no. 3, pp. 1475–1490, Mar. 2023.
[32] S.-Y. Lien, S.-L. Shieh, Y. Huang, B. Su, Y.-L. Hsu, and H.-Y. Wei, "5G new radio: Waveform, frame structure, multiple access, and initial access," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 64–71, Jun. 2017.
[33] 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 Netw. Lett.*, vol. 2, no. 1, pp. 5–9, Mar. 2020.
[34] H. Liu, N. I. Miridakis, T. A. Tsiftsis, K. J. Kim, and K. S. Kwak, "Coordinated uplink transmission for cooperative NOMA systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Abu Dhabi, UAE, Dec. 2018, pp. 1–6.
[35] C. Sun, C. She, C. Yang, T. Q. S. Quek, Y. Li, and B. Vucetic, "Optimizing resource allocation in the short blocklength regime for ultra-reliable and low-latency communications," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 402–415, Jan. 2019.
[36] Z. Sheng, D. Tian, and V. C. M. Leung, "Toward an energy and resource efficient Internet of Things: A design principle combining computation, communications, and protocols," *IEEE Commun. Mag.*, vol. 56, no. 7, pp. 89–95, Jul. 2018.
[37] L. Liang, H. Ye, and G. Y. Li, "Spectrum sharing in vehicular networks based on multi-agent reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2282–2292, Oct. 2019.
[38] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.
[39] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
[40] H. V. Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double *Q*-learning," 2015, *arXiv:1509.06461*.
[41] M. H. H. M. L. Z. Wang, T. Schaul and N. Freitas, "Dueling network architectures for deep reinforcement learning," 2016, *arXiv:1511.06581*.
[42] X. Liu, Y. Liu, Y. Chen, and L. Hanzo, "Enhancing the fuel-economy of V2I-assisted autonomous driving: A reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 69, no. 8, pp. 8329–8342, Aug. 2020.
[43] X. Liu, Y. Liu, and Y. Chen, "Machine learning empowered trajectory and passive beamforming design in UAV-sRIS wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 2042–2055, Jul. 2021.
[44] A. Sannai, Y. Takai, and M. Cordonnier, "Universal approximations of permutation invariant/equivariant functions by deep neural networks," 2019, *arXiv:1903.01939*.
[45] F. Wu, H. Zhang, J. Wu, and L. Song, "Cellular UAV-to-device communications: Trajectory design and mode selection by multi-agent deep reinforcement learning," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4175–4189, Jul. 2020.
[46] A. Nouruzi et al., "Toward a smart resource allocation policy via artificial intelligence in 6G networks: Centralized or decentralized?" 2022, *arXiv:2202.09093*.
[47] M. Zeng, X. Li, G. Li, W. Hao, and O. A. Dobre, "Sum rate maximization for IRS-assisted uplink NOMA," *IEEE Commun. Lett.*, vol. 25, no. 1, pp. 234–238, Jan. 2021.

DUC-DUNG TRAN (Member, IEEE) received the B.E. degree in electronics and telecommunications from the Hue University of Sciences, Vietnam, in 2013, and the M.Sc. degree in computer sciences from Duy Tan University, Vietnam, in 2016. He is currently pursuing the Ph.D. degree with the Interdisciplinary Center for Security, Reliability and Trust, University of Luxembourg, Luxembourg. From 2014 to 2019, he was with the Faculty of Electrical and Electronics Engineering, Duy Tan University. His current research interests include 5G and beyond wireless networks, machine learning, URLLC, and multiple access techniques.

SHREE KRISHNA SHARMA (Senior Member, IEEE) received the Ph.D. degree in wireless communications from the University of Luxembourg in 2014. He held various research and academic positions with the Interdisciplinary Center for Security, Reliability and Trust, University of Luxembourg; Western University, Canada; and Ryerson University, Canada. He is a Lead Editor of two IET books on *Satellite Communications in the 5G Era* and *Communications Technologies for Networked Smart Cities*. He has published more than 100 technical papers in scholarly journals, international conferences, and book chapters, and has over 6000 Google Scholar citations with an H-index of 36.

VU NGUYEN HA (Member, IEEE) received the B.Eng. degree (Hons.) from the French Training Program for Excellent Engineers in Vietnam, Ho Chi Minh City University of Technology, Vietnam, the Addendum degree from the École Nationale Supérieure des Télécommunications de Bretagne-Groupe des École des Télécommunications, Bretagne, France, in 2007, and the Ph.D. degree (Hons.) from the Institut National de la Recherche Scientifique-Énergie, Matériaux et Télécommunications, Université du Québec, Montreal, QC, Canada, in 2017. From 2016 to 2021, he worked as a Postdoctoral Fellow with the Ecole Polytechnique de Montreal, and then the Resilient Machine Learning Institute, École de Technologie Supérieure, University of Québec. He is currently a Research Scientist with the Interdisciplinary Centre for Security, Reliability, and Trust, University of Luxembourg. His research interests include applying/developing optimization and machine-learning-based solution for RRM problems in MAC/PHY layers of several wireless communication systems, including SATCOM, 5G/beyond-5G, HetNets, Cloud RAN, massive MIMO, mobile-edge computing, and 802.11ax WiFi. He received the Innovation Award for his Ph.D. degree. He was a recipient of the FRQNT Postdoctoral Fellowship for International Researcher (PBEEE) awarded by the Québec Ministry of Education, Canada, in 2018 and 2019. In 2021 and 2022, he was also awarded the Certificate for Exemplary Reviews by the IEEE WIRELESS COMMUNICATIONS LETTERS.

SYMEON CHATZINOTAS (Fellow, IEEE) received the M.Eng. degree in telecommunications from the Aristotle University of Thessaloniki, Greece, in 2003, and the M.Sc. and Ph.D. degrees in electronic engineering from the University of Surrey, U.K., in 2006 and 2009, respectively. He is currently a Full Professor/Chief Scientist I and the Head of the Research Group SIGCOM, Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg. In parallel, he is an Adjunct Professor with the Department of Electronic Systems, Norwegian University of Science and Technology and a Collaborating Scholar with the Institute of Informatics and Telecommunications, National Center for Scientific Research "Demokritos." In the past, he has lectured as a Visiting Professor with the University of Parma, Italy and contributed in numerous research and development projects for the Institute of Telematics and Informatics, Center of Research and Technology Hellas and the Mobile Communications Research Group, Center of Communication Systems Research, University of Surrey. He has authored more than 700 technical papers in refereed international journals, conferences, and scientific books and has received numerous awards and recognitions, including the IEEE Fellowship and an IEEE Distinguished Contributions Award. He is currently on the editorial board of the IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE OPEN JOURNAL OF VEHICULAR TECHNOLOGY, and the *International Journal of Satellite Communications and Networking*.

ISAAC WOUNGANG (Senior Member, IEEE) received the Ph.D. degree in mathematics from the University of the South, Toulon-Var, France, in 1994. From 1999 to 2002, he worked as a Software Engineer with Nortel Networks Corporation, Ottawa, ON, Canada. Since 2002, he has been with Ryerson University, Toronto, ON, Canada, where he is currently a Professor of Computer Science. He has published eight books and over 90 refereed technical papers in scholarly international journals and proceedings of international conferences. His current research interests include radio resource management in next-generation wireless networks, big data, Internet of Things, and cloud computing. He has served as the Chair of the Computer Chapter, IEEE Toronto Section, from 2012 to 2018. He has guest edited several special issues with various reputed journals, such as *Computer Communications* (Elsevier) and *Telecommunication Systems* (Springer).