Energy-Efficient NOMA for 5G Heterogeneous Services: A Joint Optimization and Deep Reinforcement Learning Approach

Duc-Dung Tran, *Member, IEEE,* Vu Nguyen Ha, *Member, IEEE,* Shree Krishna Sharma, *Senior Member, IEEE,* Ti Ti Nguyen, *Member, IEEE,* Symeon Chatzinotas, *Fellow, IEEE,* Petar Popovski, *Fellow, IEEE*

***Abstract*—The escalating number of wireless users requiring different services, such as enhanced mobile broadband (eMBB), massive machine-type communications (mMTC), and ultra-reliable low-latency communications (URLLC), has led to exploring non-orthogonal multiplexing methods like heterogeneous non-orthogonal multiple access (H-NOMA). This method allows users demanding divergent services to share the same resources. However, implementing the H-NOMA scheme faces major resource management challenges due to unpredictable interference caused by the random access mechanism of mMTC users. To address this issue, this paper proposes a joint optimization and cooperative multi-agent (MA) deep reinforcement learning-based resource allocation mechanism, aimed at maximizing the energy efficiency (EE) of H-NOMA-based networks. Specifically, this work initially establishes an optimization framework capable of determining the optimal power allocation for any specific sub-channel assignment (SA) setting for all users. Based on that, a cooperative MA double deep Q network (CMADDQN) scheme is carefully designed at the base station to conduct SA among users. In addition, a distributed full learning-based approach using MADDQN for both SA and power allocation is also designed for comparison purposes. Simulation results show that the proposed joint optimization and machine learning method outperforms the solely-learning-based approach and other benchmark schemes in terms of convergence rate and EE performance.**

***Index Terms*—Joint Optimization and Machine Learning, DDQN, eMBB, Heterogeneous NOMA, mMTC, URLLC.**

I. Introduction

THE future wireless networks are anticipated to support a tremendous number of devices requiring heterogeneous services, e.g., enhanced mobile broadband (eMBB), massive machine type communications (mMTC), and ultra-reliable and low-latency communications (URLLC), together with different quality-of-service (QoS) demands [2]. Specifically, the eMBB service aims to bring a significant increase in user data rate; the mMTC service supports the connectivity for a huge number of devices; and the URLLC is expected to provide a service with unprecedented high reliability and low latency [3]. Due to the high data rate demand, eMBB communication is designed under the assumption of infinite blocklength (iFBL) to target Shannon's channel capacity utilizing long data packets. In contrast, the demands of high connectivity density in mMTC, and ultra-reliable and low latency in URLLC require a new transmission method since mMTC and URLLC packets are generally short. In this regard, short-packet communications (SPC) have been applied for mMTC and URLLC transmissions to meet their requirements [4].

To meet the diverse requirements arising from heterogeneous services, non-orthogonal multiple access (NOMA) technology is considered a promising solution [5], [6]. Specifically, numerous studies in the literature have considered employing NOMA to efficiently manage the transmission in systems where heterogeneous services coexist [6], [7]. Thanks to the NOMA mechanism, users of heterogeneous services can simultaneously communicate with the base station (BS) using the same time-frequency resource block (RB). This is achieved through various methods such as power domain [5], [8], rate splitting [9], or codebook/pilot sequences [10]. Recently, NOMA technologies have been empowered by the introduction of the semi-grant-free (semi-GF) strategy, also known as semi-GF NOMA [11]. Following this novel integrated strategy, users having stringent QoS requirements (e.g., eMBB or URLLC users) are scheduled orthogonally by the system controllers (e.g., BS, access point, etc.) using grant-based (GB) access to fulfill their demands. Meanwhile, other users, such as mMTC users, can access the opened RBs freely according to a grant-free (GF) access mechanism. This approach can significantly increase connectivity opportunities in dense networks. On top of the semi-GF scheme, the NOMA transmission becomes particularly advantageous when more than one user accesses a specific RB.

A. Related Works

Recently, the applications of NOMA to the systems with multiplexed diverse services have been investigated [6], [11]–[15]. In particular, the authors in [6] explored a heterogeneous NOMA (H-NOMA)-based network slicing scheme for wireless communication systems supporting the eMBB, URLLC, and mMTC services. In this work, H-NOMA is defined as a novel approach to non-orthogonal sharing of the RBs for various services, distinct from the conventional NOMA which caters to homogeneous demands. In particular, the employment of this H-NOMA network slicing scheme enables users requiring

This work was supported in part by the FNR-funded project CORE 5G-Sky (Grant C19/IS/13713801), ERC-funded project Agnostic (Grant 742648), and the Villum Investigator Grant "WATER" from the Velux Foundations, Denmark. An earlier version of this manuscript was presented in part at the IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Toronto, Canada, September 2023 [1].

D.-D. Tran, Vu Nguyen Ha, S. K. Sharma, and S. Chatzinotas are with Interdisciplinary Center for Security, Reliability and Trust (SnT), University of Luxembourg, 4365 Esch-sur-Alzette, Luxembourg (e-mail: duc.tran, vu-nguyen.ha@uni.lu; shree.sharma@ieee.org; symeon.chatzinotas@uni.lu).

Ti Ti Nguyen is with Université du Québec, Montréal, Canada (e-mail: titi.nguyen@emt.inrs.ca).

Petar Popovski is with the Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark (e-mail: petarp@es.aau.dk).

various services to continuously utilize the same specific frequency radio resources over time. This approach leads to a significant improvement in spectrum efficiency. This work also showed that the H-NOMA-based slicing scheme can outperform heterogeneous orthogonal multiple access (H-OMA) in meeting diverse requirements under certain considered scenarios. In [11], [12], the advantage of the semi-GF NOMA scheme serving both GB and GF users has been investigated in different circumstances. Furthermore, the closed-form expressions of outage probability and ergodic rate were derived in these works to analyze system performance. In [13], the authors have developed an efficient NOMA-based slicing solution for eMBB and URLLC coexisting networks to minimize the total power consumption. In [14], the coexistence of eMBB and URLLC in multiple-input multiple-output (MIMO) NOMA systems was investigated by maximizing eMBB users' data rate while fulfilling the latency demands of URLLC users. In [15], a network slicing method for eMBB, URLLC, and mMTC based on a rate-splitting MA (RSMA) scheme was proposed. This work showed that RSMA-based network slicing can achieve better performance in terms of sum-rate in some investigated regions as compared to conventional OMA-based and NOMA-based ones.

The GF strategy can reduce the overhead time associated with setting up transmission links; however, it also introduces complex issues related to interference management. Specifically, for massive access scenarios, the random access nature of GF or mMTC users can result in severe interference when a large number of users attempt to access a limited number of RBs. This can make users' heterogeneous QoS requirements unsatisfied, leading to significant performance degradation. Furthermore, in the context of the wireless channel varying unpredictably over time, developing dynamic resource allocation (RA) mechanisms addressing the above congestion problem and fulfilling the various QoS requirements from different services becomes more challenging. In recent years, the reinforcement learning (RL) method has been applied to intelligently resolve the RA problem in communications [16]–[18]. Its application to the coexistence of heterogeneous services has been investigated in [19]–[22].

Specifically, the authors in [19] developed an intelligent resource-slicing approach for downlink eMBB-URLLC coexisting orthogonal frequency-division MA (OFDMA) systems by exploiting the well-known deep RL (DRL) tools. Considering uplink transmission, the authors in [20] proposed a multi-agent (MA) DRL (MADRL) RA framework using deep Q network (DQN) and transfer learning for OFDMA-based uplink systems serving multiple users with different QoS requirements, such as high reliability, low latency, and high data rate. In this study, the power quantization (PQ) method is utilized to discretize the continuous range of possible transmission powers into a limited set of transmission power levels (TPLs), thereby facilitating the learning process. However, this approach may lead to performance loss if the discretized power levels are unable to closely approximate the optimal points. In [21], Fayaz *et al.* developed a DRL-based sub-channel (SC) and power allocation mechanism for semi-GF NOMA-based uplink HetNets, which aims to maximize the sum rate while meeting different rate demands of both GB and GF users. In this work, the PQ method is only employed for GF users' power allocation while the transmission power of the GB users is fixed. In [22], the authors investigated a two-hop NOMA-based uplink HetNet, where each GF user first transmits its message to a selected GB user, the chosen GB users then forward the information to BS. The GF users in this scheme were considered as DQN agents to select SC and transmission power from finite sets of all SCs and pre-discretized TPLs. In addition, a GB user is designated as the head of the GF-user cluster. Meanwhile, the power allocation for GB users is centrally defined at the BS using another DRL algorithm, namely proximal policy optimization (PPO).

B. Contributions

This paper develops a joint optimization and MADRL-based method for H-NOMA-based uplink systems serving eMBB, mMTC, and URLLC users requesting heterogeneous QoS requirements, not only to speed up the learning process but also to achieve the optimal RA solution. Specifically, the main contributions of this paper are summarized as follows:

- We investigate the coexistence of eMBB, mMTC, and URLLC in a H-NOMA uplink system, where eMBB and URLLC users are assigned orthogonally to a number of SCs to fulfill their stringent QoS requirements on high reliability, low latency, and high data rate. Meanwhile, mMTC users can access any SCs freely and quickly without any admission approval from BS to improve the spectrum access efficiency and connectivity density.
- We formulate an energy efficiency (EE) maximization problem for the considered system. The objective is to maximize the long-term average EE under constraints on various QoS requirements of users.
- We design a novel learning-based RA strategy to address the proposed problem. In particular, we propose a joint optimization and cooperative MA DDQN (JOCDDQN) method to optimize the RA policy as well as significantly improve learning performance. The JOCDDQN method utilizes a cooperative MA DDQN (CMADDQN) scheme centralized at the BS for SC assignment based on which a dynamic power allocation solution is developed to obtain optimal transmission power for users. We also design a distributed full learning solution based on MADDQN, namely FDDQN, where all users are considered as learning agents to find the best SC and power level selection policy in order to resolve the investigated problem. It is noteworthy that the PQ method is applied in the proposed FDDQN method similar to the existing works [20], [21].
- We carry out the performance comparison between our proposed methods and other benchmark schemes to evaluate the efficiency of the former in terms of convergence property and EE performance. Additionally, we provide numerical results to analyze the effects of different system parameters, such as the number of SCs, number of transmission power levels (TPLs), number of users, maximum transmission power, and divergent QoS requirements, on the system performance.

TABLE I: Main Notations and Symbols

Notation	Description
$a_z(t)$	The action of agent z at TS t
α	The learning rate
$\mathcal{CN}\left(0,\sigma^2\right)$	A scalar complex Gaussian distribution with zero mean and variance σ^2
$c_z^{(k)}(t)$	Binary SC allocation variable for user z over SC k in time-slot (TS) t
D_z	The transmission latency
$\mathbb{E}\left[\cdot\right]$	The expectation operator
ε_z	The decoding error probability of user z
γ	The discount factor
$\gamma_z^{(k)}(t)$	The received signal-to-interference-plus-noise ratio (SINR) according to user z over SC k in TS t
$h_z^{(k)}(t)$	The channel coefficient of the link from user z to the BS over SC k in TS t
$\mathcal{K}$	The set of SCs
L	The number of TPLs
$\mathcal{M}_{\mathsf{E}}$	eMBB user set
$\mathcal{M}_{\mathsf{M}}$	mMTC user set
$\mathcal{M}_{\mathsf{U}}$	URLLC user set
n_b	The packet size
ν_{E}	eMBB numerology indices
ν_{U}	URLLC numerology indices
$P_z^{(k)}(t)$	The Tx-power of user z over SC k in TS t
$P_z^{\max}$	The maximum Tx-power of user z
$\hat{P}_l$	The l-th TPL
$Q\left(x\right)$	The Gaussian Q-function
$Q^{-1}\left(x\right)$	The inverse of the Gaussian Q-function
$r(t)$	The reward function at TS t
r_c	The cell radius
$R_z^{(k)}(t)$	The achievable rate of user z over SC k in TS t
$s_z(t)$	The network state of agent (user) z at TS t
W_{E}	SC bandwidth according to eMBB users
W_{U}	SC bandwidth according to URLLC users
$x_z^{(k)}(t)$	The message of user z transmitted on SC k in TS t
$\zeta(t)$	EE in TS t
$\left\|\cdot\right\|$	The absolute value

The remainder of the paper is organized as follows. Section II presents the system model, uplink transmission strategy for H-NOMA-based systems, the achievable rate of users, and the EE maximization (EE-Max) problem. Section III describes the proposed JOCDDQN solution to address the EE optimization problem. Section IV shows the proposed distributed FDDQN method to resolve the considered problem. Section V provides the simulation results and discussions. Finally, Section VI concludes this paper. For clarity, a summary of the main notations and symbols is provided in Table I.

II. System Model

As shown in Fig. 1, we investigate an H-NOMA-based uplink system that consists of one BS located at the center of the cell with a radius of r_c (m) and a number of users randomly distributed in this cell requiring different services such as eMBB, mMTC, and URLLC. Let $\mathcal{M}_{\mathsf{U}}$, $\mathcal{M}_{\mathsf{E}}$ and $\mathcal{M}_{\mathsf{M}}$ be the sets of URLLC, eMBB, and mMTC users, whose cardinalities are M_{U}, M_{E} and M_{M}, respectively. For convenience, we also denote the set of all users as $\mathcal{M} = \mathcal{M}_{\mathsf{U}} \cup \mathcal{M}_{\mathsf{E}} \cup \mathcal{M}_{\mathsf{M}}$ and $M = M_{\mathsf{U}} + M_{\mathsf{E}} + M_{\mathsf{M}}$. To serve these users, a total bandwidth of W (Hz) is assumed in the system, which is divided into K SCs. Let $\mathcal{K}$ be the set of all K SCs. Furthermore, due to high requirements of eMBB (data rate) and URLLC (reliability and latency) services, one assumes that each of the eMBB and URLLC users is preassigned several orthogonal SCs for its transmissions. Meanwhile, the mMTC users are assumed to be able to access any available SCs freely to improve the connection density due to the massive access requirement of the mMTC service. Therefore, the mMTC users can use the SCs granted to the eMBB and URLLC users. In this case, when more than one user occupies the same SC, the power-domain H-NOMA scheme is applied for multi-user communication. In practice, the number of active mMTC and URLLC users is random resulting in a dynamic traffic load, which can be described by Poisson distribution [7]. Here, we consider the worst scenario where all M_{M} mMTC users and M_{U} URLLC users have packets to transmit in each time-slot (TS), leading to the highest co-channel interference.

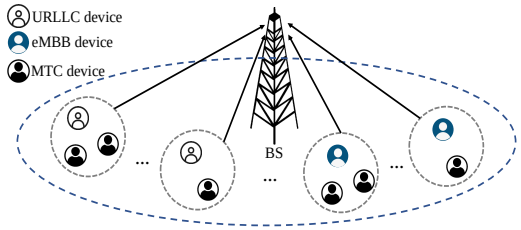


Fig. 1: Illustration of a H-NOMA-based uplink system.

A. Uplink Transmission Strategy for H-NOMA

1) 5G New Radio (NR) Numerologies: 5G NR standard introduces various physical-resource-block (PRB) or subchannel (SC) types in order to support different communication requirements and use-cases, which is referred to as *"numerology"*. In particular, the bandwidth of SC in 5G NR schemes is defined as 2^{ν} times of SC's bandwidth in 4G systems (i.e., 180 kHz), where $\nu \in \{0;1;2;3;4\}$ is the numerology index [2]. PRBs with high SC spacing are arranged for URLLC services while traffic flows from the eMBB service can adopt a numerology with the smaller SC spacing [2]. Therefore, this paper focuses on an SC setting that the whole bandwidth is divided into two sets of SCs, $\mathcal{K}_{\mathsf{U}}$ and $\mathcal{K}_{\mathsf{E}}$. Particularly, $\mathcal{K}_{\mathsf{U}}$ represents the set of SCs serving URLLC users with numerology ν_{U} while $\mathcal{K}_{\mathsf{E}}$ is the set of eMMB-service SCs with numerology ν_{E}. Herein, $\mathcal{K}_{\mathsf{U}} \cup \mathcal{K}_{\mathsf{E}} = \mathcal{K}$. One assumes that $\nu_{\mathsf{E}} < \nu_{\mathsf{U}}$ and denotes $W_{\mathsf{E}} = 2^{\nu_{\mathsf{E}}} \times 180$ (kHz) and $W_{\mathsf{U}} = 2^{\nu_{\mathsf{U}}} \times 180$ (kHz) as the bandwidth of SCs corresponding to eMBB and URLLC services, respectively.

2) Uplink Transmission Mechanism: Considering the transmission over SC k ($k \in \mathcal{K}$), we denote $c_z^{(k)}(t)$ ($z \in \mathcal{M}$) as a binary SC allocation variable at TS t, where $c_z^{(k)}(t) = 1$ if user z occupies SC k and $c_z^{(k)}(t) = 0$ otherwise. As mentioned earlier, each of eMBB and URLLC users is assigned a set of orthogonal SCs to guarantee its strict requirements. In addition, we assume a one-SC freely access strategy for mMTC users where each mMTC user can select only one arbitrary SC for its transmission. These assumptions yield

$$(C1): \quad \sum_{z\in\mathcal{M}_{\mathsf{E}}\cup\mathcal{M}_{\mathsf{U}}} c_z^{(k)}(t) \le 1, \quad \forall k \in \mathcal{K}. \tag{1}$$

$$(C2): \quad \sum_{k\in\mathcal{K}} c_z^{(k)}(t) = 1, \quad \forall z \in \mathcal{M}_{\mathsf{M}}. \tag{2}$$

Thus, many mMTC users can access the same SC and they can use the SCs granted to the eMBB and URLLC users. To enable a multi-user data stream over the same SC, the power-domain H-NOMA scheme is employed. Following H-NOMA

principle, many users requiring different services can occupy the same SC for their transmissions. In this regard, the received signal over SC k at the BS in TS t can be expressed as

$$y_k(t) = \sum_{z \in \mathcal{M}} c_z^{(k)}(t) \sqrt{P_z^{(k)}(t)} h_z^{(k)}(t) x_z^{(k)} + w_k(t), \quad (3)$$

where $w_k(t) \sim \mathcal{CN}\left(0, \sigma_k^2\right)$ is the additive white Gaussian noise (AWGN) over SC k at the BS; $P_z^{(k)}(t)$, $h_z^{(k)}(t)$, and $x_z^{(k)}$ denote the transmission power (Tx-power), channel coefficient, and transmitted symbol of user z over SC k, respectively. It is worth noting that the Tx-power is defined as $P_z^{(k)}(t) = 0$ if $c_z^{(k)}(t) = 0$ and $P_z^{(k)}(t) \neq 0$, otherwise. From (3), the BS can decode the received multi-user data systematically through the use of the successive interference cancellation (SIC) technique [23]. In uplink NOMA, the decoding order of the multi-user data stream is affected by various different factors. Specifically, a decoding order can be formulated based on channel gain conditions [12], received power levels [24], or QoS constraints of users [23]. In this paper, the messages of the users over each SC can be decoded at the BS as follows:

- Due to strict requirements on reliability and latency, the URLLC users' messages need to be decoded first. However, as long as their requirements are guaranteed, the SCs granted to them can still be used by the mMTC users to improve the spectrum efficiency.
- The symbols belonging to eMBB and mMTC users will be decoded in the order of the corresponding channel gains. In particular, the user having the higher channel gain will be decoded earlier at the BS.
- After decoding the message of a user based on the decoding order mentioned above, the BS removes this component from its observation to decode the remaining users' messages by using the successive interference cancellation (SIC) technique.

Given the above context, it is noteworthy that an SC selection of mMTC users needs to guarantee the minimum co-channel interference to optimize network performance while satisfying the QoS requirements of both granted users (i.e., URLLC and eMBB) and themselves. Otherwise, an alternative SC allocation for mMTC users is necessary.

Without loss of generality, one assumes there are Z^k users accessing SC k in TS t, then they are arranged in the decoding order discussed above as $\mathcal{Z}^{(k)}(t) = \left\{z_1^{(k)}, ..., z_{Z^k}^{(k)}\right\}$. Accordingly, the received signal-to-interference-plus-noise ratio (SINR) of user $z_\ell^{(k)}$ $(1 \leq \ell \leq Z^k)$ is expressed as

$$\gamma_{z_\ell^{(k)}}^{(k)}(t) = \mathcal{Y}_{z_\ell^{(k)}}^{(k)}(t) / \left[\sum_{j=\ell+1}^{Z^k} \mathcal{Y}_{z_j^{(k)}}^{(k)}(t) + \sigma_k^2\right], \quad (4)$$

where $\mathcal{Y}_z^{(k)}(t) = P_z^{(k)}(t) g_z^{(k)}(t)$ is the power of signal due to user z's data over SC k, $g_z^{(k)}(t) = \left|h_z^{(k)}(t)\right|^2$ denote the corresponding channel gain, and $\sigma_k^2 = F N_0 W_k$ represents the noise power over SC k. Herein, F is the noise figure in dB, N_0 is the noise power spectral density (PSD) in dBm/Hz, W_k denotes the bandwidth of SC k, $W_k = W_{\mathsf{E}}$ if $k \in \mathcal{K}_{\mathsf{E}}$ and $W_k = W_{\mathsf{U}}$ if $k \in \mathcal{K}_{\mathsf{U}}$. It is noted that we consider a perfect SIC scenario similar to [22], [24], [25], where the interference from stronger users $\left\{z_1^{(k)}, \ldots, z_{\ell-1}^{(k)}\right\}$ is perfectly removed before decoding the message of user $z_\ell^{(k)}$. Meanwhile, the power noise σ_k^2 is still considered in the power estimation process.

B. Achievable Rate of Users

1) URLLC Communication: Regarding the transmission of URLLC user u over SC k in $\mathcal{K}_{\mathsf{U}}$, which happens when $c_u^{(k)} = 1$. Based on the NOMA transmission mechanism given in Section II-A2, one must have $u \equiv z_1^{(k)}$. Moreover, the SINR of URLLC device u over SC k is expressed as

$$\gamma_u^{(k)}(t) = \mathcal{Y}_u^{(k)}(t) / \left[\mathcal{I}_u^{(k)}(t) + \sigma_u^2\right], \quad (5)$$

where $\mathcal{I}_u^{(k)}(t) = \sum_{j=2}^{Z^k} \mathcal{Y}_{z_j^{(k)}}^{(k)}(t)$ represents the interference caused by mMTC users over SC k. Furthermore, bandwidth of SC k in $\mathcal{K}_{\mathsf{U}}$ is W_{U} and $\sigma_u^2 = F N_0 W_{\mathsf{U}}$. In URLLC communication, short packet communication (SPC) in finite blocklength (FBL) regime is implemented to meet the strict URLLC requirements. Consequently, Shannon's capacity formula cannot be applied for URLLC communication model to capture the transmission data rate and decoding error probability effectively since it is designed under the assumption of infinite blocklength (iFBL). According to [26], [27], the achievable rate of URLLC user u over SC k in FBL regime for a quasi-static flat fading channel can be approximated as

$$R_u^{(k)}(t) = W_{\mathsf{U}}[\log_2(1 + \gamma_u^{(k)}(t)) - \Phi_u^{(k)}(t)], \quad (6)$$

where $\Phi_u^{(k)}(t) = \sqrt{\frac{V_u^{(k)}(t)}{D_u W_{\mathsf{U}}}} \frac{Q^{-1}(\varepsilon_u)}{\ln 2}$, ε_u is the decoding error probability (DEP) which can be used to evaluate the transmission reliability, D_u is the transmission latency threshold, $Q^{-1}(x)$ is the inverse of the Gaussian Q-function, and $V_u^{(k)}(t)$ is the channel dispersion which is given by

$$V_u^{(k)}(t) = 1 - \left[1 + \gamma_u^{(k)}(t)\right]^{-2} \approx 1. \quad (7)$$

Here, the approximation is very tight when $\gamma_u^k(t) \geq 5$ dB [27], [28]. In contrast, $V_u^k(t)$ decreases $(V_u^{(k)}(t) < 1)$ as $\gamma_u^{(k)}(t)$ gets lower. Thus, by using the approximation $V_u^k(t) \approx 1$ in low SNR regime ($\gamma_u^{(k)}(t) < 5$ dB), we can achieve the lower bound of the achievable rate in (6). Users' QoS requirements can be satisfied if the lower bound is employed for RA optimization. Therefore, the approximation in (7) can be used for further analysis. Note that the channel dispersion definition in (7) is achieved based on the assumption that each user has its perfect channel state information[1] (CSI), such that the packet error occurs due to the noise instance only [4], [28].

Remark 1. *For short-packet transmission, the packet size can influence the latency and the decoding error probability (reliability). Considering the achievable rate in [packets/s], the packet size-latency and packet size-reliability relations can be, respectively, mathematically determined based on* (6) *as*

$$D_u = \left\{Q^{-1}(\varepsilon_u) / \left[\sqrt{W_{\mathbf{U}}}\left(C_u(t) - n_u R_u^{(k)}(t)/W_{\mathbf{U}}\right) \ln 2\right]\right\}^2, \quad (8)$$

[1]CSI-related signaling exchange between BS and users may yield a latency increase. The effect of this scenario can be analyzed by addressing the problem of latency minimization which is beyond the scope of this paper.

$$\varepsilon_u = Q\left(\left(C_u(t) - n_u R_u^{(k)}(t)/W_{\mathbf{U}}\right)\ln 2\sqrt{D_u W_{\mathbf{U}}}\right), \quad (9)$$

where $C_u(t) = \log_2(1+\gamma_u^{(k)}(t))$ *and* n_u *denotes the packet size. One can observe from* (8) *and* (9) *that the increase in the value of the packet size leads to higher latency* D_u *and lower reliability (i.e., higher decoding error probability* ε_u*).*

For URLLC service, we assume that each URLLC user tries to upload one packet over one SC in each TS. To analyze the URLLC requirements, we transform them into an SNR constraint, as discussed in [27], where a target SNR threshold is defined based on the transmission latency threshold D_u and the decoding error probability ε_u. Thus, the target SNR threshold for URLLC user u that satisfies the URLLC requirements (i.e., D_u and ε_u) can be derived based on (6) as [27]

$$\gamma_u^{\mathsf{tar}} = 2^{\frac{n_u}{D_u W_{\mathsf{U}}} + \frac{Q^{-1}(\varepsilon_u)}{\ln 2\sqrt{D_u W_{\mathsf{U}}}}} - 1, \quad (10)$$

where n_u is the packet size. This demand yields

$$(C3): \quad c_u^{(k)}(t)\gamma_u^{(k)}(t) \geq \gamma_u^{\mathsf{tar}}, \quad \forall k \in \mathcal{K}. \quad (11)$$

2) eMBB Communication: Regarding the transmission of eMBB user e over SC k in $\mathcal{K}_{\mathsf{E}}$, e.g., $c_e^{(k)}(t) = 1$. Due to its order in the NOMA-based decoding process, its SINR denoted as $\gamma_e^{(k)}(t)$, as in (4) with noting that $\sigma_k^2 = FN_0W_{\mathsf{E}}$. Then, the achievable rate of eMBB user e is given by

$$R_e^{(k)}(t) = W_{\mathsf{E}}\log_2\left[1+\gamma_e^{(k)}(t)\right]. \quad (12)$$

Herein, one addresses a predetermined target transmission rate, R_e^{tar}, for each eMBB user e in every TS as

$$(C4): \quad \sum_{k\in\mathcal{K}} c_e^{(k)}(t)R_e^{(k)}(t) \geq R_e^{\mathsf{tar}}, \quad \forall e \in \mathcal{M}_{\mathsf{E}}. \quad (13)$$

3) mMTC Communication: Based on the NOMA transmission strategy mentioned earlier in Section II-A2, the mMTC users can select a free SC or the one occupied by either eMBB or URLLC user. When $c_m^{(k)}(t) = 1$, mMTC user m utilizes SC k in TS t. In such case, the SINR of this user, denoted as $\gamma_m^{(k)}(t)$, can be calculated as in (4) with noting that $\sigma_k^2 = FN_0W_k$. In mMTC communication, SPC is utilized to exchange small packets. Consequently, the achievable rate of mMTC user m over SC k in FBL regime for a quasi-static flat fading channel can be approximated as [26], [27]

$$R_m^{(k)}(t) = W_k\left[\log_2(1+\gamma_m^{(k)}(t)) - \Phi_m^{(k)}(t)\right], \quad (14)$$

whose parameters are defined similarly in (6). From (14), we define a target SINR threshold for mMTC user m to satisfy its DEP and latency requirements (i.e., ε_m and D_m) when transmitting one packet over one SC in each TS as [27]

$$\gamma_m^{\mathsf{tar}} = 2^{\frac{n_m}{D_m W_k} + \frac{Q^{-1}(\varepsilon_m)}{\ln 2\sqrt{D_m W_k}}} - 1, \quad (15)$$

where n_m is the packet size. This demand yields

$$(C5): \quad c_m^{(k)}(t)\gamma_m^{(k)}(t) \geq \gamma_m^{\mathsf{tar}}, \quad \forall k \in \mathcal{K}. \quad (16)$$

It is noteworthy that the successful packet transmission of mMTC users is considered based on SINR condition in (16) and the NOMA decoding order mentioned in Section II-A2.

C. Energy Efficiency Maximization Problem

In this paper, we aim to develop an effective SC and power allocation (PA) strategy based on users' channel conditions and requirements to minimize the co-channel interference to maximize the long-term average EE while ensuring the specific QoS requirements from different services [2]. To guarantee the transmission rate requirement while reducing the power consumption for the system, we define an EE factor as

$$\zeta(t) = R^{\mathsf{tot}}(t)/\left(P^{\mathsf{Tx}}(t) + MP_{\mathsf{c}}\right), \quad (17)$$

where $R^{\mathsf{tot}}(t) = \sum_{k\in\mathcal{K}}\sum_{z\in\mathcal{M}} c_z^{(k)}(t)R_z^{(k)}(t)$, $P^{\mathsf{Tx}}(t) = \sum_{k\in\mathcal{K}}\sum_{z\in\mathcal{M}} P_z^{(k)}(t)$, and P_{c} denotes the circuit power consumption. Then, the EE-Max problem can be stated as

$$\max_{\mathbf{c},\mathbf{P}} \quad \mathbb{E}_t\left[\zeta(t)\right] \quad (18a)$$

$$\text{s.t.} \quad (C1)-(C5), \quad (18b)$$

$$(C6): \sum_{k\in\mathcal{K}} P_z^{(k)}(t) \leq P_z^{\mathsf{max}}, \ \forall(z,t), \quad (18c)$$

where $\mathbf{c}$ and $\mathbf{P}$ denote the SC assignment and power control variables and $(C6)$ stands for the user power budget.

Remark 2. *Problem* (18) *is a mixed-integer non-linear programming (MINLP), well-known as NP-hard, which is difficult to solve. In particular, the challenges of solving this problem include the coupling between binary variables* $\mathbf{c}$ *and continuous ones* $\mathbf{P}$*. Moreover, the complicating H-NOMA-based SINR formula of users as given in* (4) *has raised another extremely critical issue to define the solution of this problem.*

III. Proposed Joint Optimization and Cooperative MADDQN Method

In this section, we present the proposed joint optimization and MADRL method to solve (18), where a cooperative MADDQN (CMADDQN) scheme is built for SC assignment based on which a dynamic PA (DPA) for every SC setting is proposed to maximize the EE in (18). To do this, one assumes that the perfect CSI is available at the BS.

A. Power Allocation for Given SC Assignment

Before presenting the CMADDQN-based SC assignment, we first introduce our proposed PA solution for a given SC assignment in this section. In particular, for a fixed SC allocation, problem (18) in TS t can be expressed as

$$\max_{\mathbf{P}} \ \zeta(t) \ \text{s.t.} \ (C1)-(C6), \quad (19)$$

where $\mathbf{P}(t) = \{P_z^{(k)}(t)\}_{\forall(z,k)}$. To tackle the fraction form of $\zeta(t)$ in (19), an efficient method well-known as the Dinkelbach algorithm [29], [30] can be employed. Following this, an iterative approach can be developed to obtain the optimal solution of (19) by dealing with a sequence of parameterized subtracting-form problems. Let us state the parameterized problem for a given value of ζ as follows:

$$\max_{\mathbf{P}} \ R^{\mathsf{tot}}(t) - \zeta P^{\mathsf{Tx}}(t) \ \text{s.t.} \ C1)-(C6). \quad (20)$$

[2]To further analyze the distinct objectives of the URLLC and mMTC services, more specific optimization problems such as minimizing latency or decoding error probability, and maximizing successful packet rate or number of successfully served mMTC users can be investigated, which are interesting for future works.

Then, **Theorem 1** in [30] suggests to iteratively solve problem (20) for a certain value of ζ and adjusting ζ until an optimal $\zeta^\star \geq 0$ satisfying $R^{\mathsf{tot}}(t) = \zeta^\star \left(P^{\mathsf{Tx}}(t) + MP_{\mathsf{c}}\right)$ is found.

To address the problem (20), we first provide the following valuable remark based on (4) and the uplink NOMA transmission mechanism discussed in Section II-A2.

Remark 3. *The SINR in* (4) *demonstrates that there is no interference suffering the decoding process due to user $z^{(k)}_{Z^k}$. Moreover, once the power of all users in set $\left\{z^{(k)}_{\ell+1}, ..., z^{(k)}_{Z^k}\right\}$ is defined, the Tx-power of user $z^{(k)}_{\ell}$, i.e., $P^{(k)}_{z^{(k)}_{\ell}}$, can be optimized without coupling to other users. Hence, the Tx-power can be determined in the reverse order of the coding sequence.*

Thanks to Remark 3, an efficient approach to solving the problem (20) will be proposed in the following. The concept of this solution is to decompose problem (20) into a number of sub-problems each of which aims to obtain the Tx-power of a user separately. Then, the sub-problems are solved in the order suggested in Remark 3. Particularly, the PA strategy for all types of users is described as follows:

- Each mMTC user will have its Tx-power optimized only when the power of all other users accessing the same SC with weaker channel are determined.
- For eMBB users, the Tx-power of an eMBB user over all SCs assigned to it will be optimized jointly when all mMTC users using the same SCs with weaker channel gains have their power defined.
- For URLLC users, they will be the last ones having their power optimized over all SCs that they are assigned.

Next, one presents the sub-problems and their solution corresponding to eMBB, mMTC, and URLLC services.

1) Power Allocation for mMTC Users: Considering the user set accessing SC k, $\mathcal{Z}^{(k)}(t)$ defined in Section II-A2, one assumes that user $m \equiv z^{(k)}_{\ell}$ is an mMTC user. When $\{P^{(k)}_{z^{(k)}_i}\}^{Z^k}_{i=\ell+1}$ are determined, this user's Tx-power, i.e., $P^{(k)}_m$, will be optimized by solving the following sub-problem,

$$\max_{p} \quad W_k \left[\log_2\left(1 + A^{(k)}_m p\right) - \Phi^{(k)}_m\right] - \zeta p \tag{21a}$$

$$\text{s.t.} \quad P^{\mathsf{tar}}_m \leq p \leq P^{\mathsf{max}}_m, \tag{21b}$$

where $P^{\mathsf{tar}}_m = \gamma^{\mathsf{tar}}_m / A^{(k)}_m$, p is the Tx-power variable, $A^{(k)}_m$ represents the NOMA-based channel gain over interference and noise ratio (NOMA-CINR) of user m, i.e.,

$$A^{(k)}_m = \left|h^{(k)}_m(t)\right|^2 / \Big[\sum_{j=\ell+1}^{Z^k} \left|h^{(k)}_{z^{(k)}_j}(t)\right|^2 P^{(k)}_{z^{(k)}_j} + \sigma^2_k\Big]. \tag{22}$$

Note that constraint $P^{\mathsf{tar}}_m \leq p$ is equivalent to $(C5)$ for this user. The solution of the problem (21) is determined in the following proposition.

Proposition 1. *The Tx-power of mMTC user m over SC k is the solution of* (21) *which is given as*

$$P^{(k)\star}_m = \min\left(\max\left(\frac{W_{\mathsf{k}}}{\zeta \ln 2} - \frac{1}{A^{(k)}_m}, P^{\mathsf{tar}}_m\right), P^{\mathsf{max}}_m\right). \tag{23}$$

Proof: The proof can be described simply as follows. Let $y^{(k)}_m(p) = W_k \left[\log_2\left(1 + A^{(k)}_m p\right) - \Phi_m(k)\right] - \zeta p$. Then, problem (21) is convex since $y^{(k)}_m(p)$ is concave and the feasible set is convex. With feasible-set regardless, we first derive the derivative of $y^{(k)}_m(p)$ concerning the variable p as

$$\partial y^{(k)}_m(p)/\partial p = W_k A^{(k)}_m / \Big[\Big(1 + A^{(k)}_m p\Big) \ln 2\Big] - \zeta. \tag{24}$$

From (24), the maximum point of the objective function can be defined by resolving the equation $\partial y^{(k)}_m(p)/\partial p = 0$ as

$$\hat{p} = W_k/(\zeta \ln 2) - 1/A^{(k)}_m. \tag{25}$$

Then, by taking the feasible set into account, the optimal solution of problem (21) can be defined as given in (23). ■

2) Power Allocation for eMBB Users: Considering the transmission of eMBB user $e \in \mathcal{M}_{\mathsf{E}}$. Assume that user e is assigned n SCs named as $\{k^e_1, ..., k^e_n\}$, and it is denoted as user $z^{(k^e_j)}_{\ell} \in \mathcal{Z}^{(k^e_j)}(t)$ over SC k^e_j ($1 \leq j \leq n$ and $1 \leq \ell \leq Z^{k^e_j}$), where $\mathcal{Z}^{(k^e_j)}(t) = \left\{z^{(k^e_j)}_1, ..., z^{(k^e_j)}_{Z^{k^e_j}}\right\}$ denotes the user set accessing SC k^e_j at TS t arranged in the decoding order explained in Section II-A2. Then, if all mMTC users with weaker channel gains on $\{k^e_1, ..., k^e_n\}$ have their power defined, the Tx-power over all SCs of eMBB user e can be determined as follows. Denote A^e_j be the NOMA-CINR of eMBB user e over SC k^e_j which is defined similarly as in (22). Then, the decomposed part of the problem (20) according to eMBB user e can be stated as

$$\max_{\mathbf{P}^e} \quad \sum_{\forall j} \left(W_{\mathsf{E}} \log_2\left(1 + A^e_j p^e_j\right) - \zeta p^e_j\right) \tag{26a}$$

$$\text{s.t.} \quad \sum_{\forall j} W_{\mathsf{E}} \log_2\left(1 + A^e_j p^e_j\right) \geq R^{\mathsf{tar}}_e, \tag{26b}$$

$$\sum_{\forall j} p^e_j \leq P^{\mathsf{max}}_e, \tag{26c}$$

where $\mathbf{P}^e = [p^e_1, ..., p^e_n]$ and p^e_j denotes the Tx-power variable corresponding to eMBB user e over SC k^e_j. As can be observed, this problem is convex and hence its optimal solution can be obtained by employing the duality method. In particular, the solution approach can begin with describing the Lagrangian of (26) as

$$\mathcal{L}(\mathbf{P}^e, \kappa, \lambda) = \sum_{\forall j} \left[(1+\kappa) W_{\mathsf{E}} \log_2\left(1 + A^e_j p^e_j\right) - (\zeta + \lambda) p^e_j\right] - \kappa R^{\mathsf{tar}}_e + \lambda P^{\mathsf{max}}_e, \tag{27}$$

where κ and λ are the Lagrangian multipliers of (26b) and (26c), respectively. Then, the dual function can be defined as

$$\mathrm{g}(\kappa, \lambda) = \max_{\mathbf{P}^b} \mathcal{L}(\mathbf{P}^b, \kappa, \lambda). \tag{28}$$

Proposition 2. *The solution of the right-hand-side (RHS) of* (28) *is defined as*

$$p^b_j = \max\Big(\frac{(1+\kappa) W_{\mathsf{E}}}{(\lambda + \zeta) \ln 2} - \frac{1}{A^b_j}, 0\Big). \tag{29}$$

Proof: The proof of this proposition can be obtained easily by solving the equation $\partial \mathcal{L}(\mathbf{P}^b, \kappa, \lambda)/\partial p^b_j = 0$. ■

To determine κ and λ, the dual problem can be written as

$$\max_{\kappa, \lambda} \mathrm{g}(\kappa, \lambda) \quad \text{s.t. } \kappa, \lambda \geq 0. \tag{30}$$

Since (26) is convex, the dual-gap corresponding to this problem is zero [31]. In the following, one will describe a searching approach to define the optimal solution of the dual

problem by using the standard sub-gradient method, where the dual variable κ and λ can be iteratively updated as

$$\kappa^{(v+1)}=\Big[\kappa^{(v)}-\delta^{(v)}\big(\sum_{\forall j} W_{\mathsf{E}}\log_2(1+A_j^b p_j^b)-R_e^{\mathsf{tar}}\big)\Big]^+, \quad (31)$$

$$\text{and } \lambda^{(v+1)}=\Big[\lambda^{(v)}+\delta^{(v)}\big(\sum_{\forall j} p_j^e-P_e^{\mathsf{max}}\big)\Big]^+, \quad (32)$$

where the suffix (v) represents the iteration index, $\delta^{(v)}$ is the step size, and $[x]^+=\max(0,x)$. This sub-gradient method guarantees the convergence if the step-size $\delta^{(v)}$ is chosen appropriately so that $\delta^{(v)}\overset{v\to\infty}{\longrightarrow}0$, e.g., $\delta^{(v)}=1/\sqrt{v}$ [31], [32].

3) Power Allocation for URLLC Users: Similarly, one assumes that there are l SCs assigned to URLLC user u denoted as $\{k_1^u,...,k_l^u\}$. If the power of all mMTC users on SCs $\{k_1^u,...,k_l^u\}$ are determined, the Tx-power over all SCs can be optimized by solving the following

$$\max_{\mathbf{P}^u} \quad \sum_{j=1}^{l}\left(W_{\mathsf{U}}\left(\log_2\left(1+A_j^u p_j^u\right)-\Phi_j^u\right)-\zeta p_j^u\right) \quad (33a)$$

$$\text{s.t.} \quad p_j^u \geq \gamma_u^{\mathsf{tar}}/A_j^u, \quad \forall j \quad (33b)$$

$$\sum_{j=1}^{l} p_j^u \leq P_u^{\mathsf{max}}, \quad (33c)$$

where γ_u^{tar} is defined in (10), $\Phi_j^u=\sqrt{\frac{V_j^u}{D_u W_{\mathsf{U}}}}\frac{Q^{-1}\left(\varepsilon_j^u\right)}{\ln 2}$, $V_j^u=1-\left(1+A_j^u p_j^u\right)^{-2}\approx 1$ [33], $\mathbf{P}^u=[p_1^u,...,p_l^e]$ and p_j^u denotes the Tx-power of URLLC user u over SC k_j^u. Similar to the results obtained in solving problem (26), the Tx-power of URLLC user u over SCs $\{k_1^u,...,k_l^u\}$, can be defined as

$$p_j^u=\max\left(W_{\mathsf{U}}/\left[(\psi+\zeta)\ln 2\right]-1/A_j^u,\gamma_u^{\mathsf{tar}}/A_j^u\right),\forall j, \quad (34)$$

where ψ can be iteratively updated as follows:

$$\psi^{[v+1]}=\Big[\psi^{[v]}+\delta^{[v]}\big(\sum_{\forall j} p_j^u-P_u^{\mathsf{max}}\big)\Big]^+. \quad (35)$$

4) Proposed Power Allocation Algorithm for Given SC Assignment: Thanks to the Dinkelbach solution approach and the PA mechanism given above, one proposes an energy-efficiency power control algorithm which is summarized in Algorithm 1. In particular, once the SC Assignment is defined, Algorithm 1 can be implemented at the BS to determine the Tx-power of all users in the network. For the user set employing the same SC k, arranged in the decoding order as $\mathcal{Z}^{(k)}(t)=\left\{z_1^{(k)},\ldots,z_{Z^k}^{(k)}\right\}$, their Tx-power is dictated by the observation detailed in Remark 3. Specifically, considering that user $z_{Z^k}^{(k)}$, the final one to have its message decoded, operates without interference, its Tx-power is initially calculated using (23) if $z_{Z^k}^{(k)}\in\mathcal{M}_{\mathsf{M}}$, (29) if $z_{Z^k}^{(k)}\in\mathcal{M}_{\mathsf{E}}$, or (34) if $z_{Z^k}^{(k)}\in\mathcal{M}_{\mathsf{U}}$. Subsequently, the PA for the next user $z_{Z^k-1}^{(k)}$ follows a similar procedure, where the messages of users with defined power, i.e., $z_{Z^k}^{(k)}$, are treated as noise. This process iterates until all users have their power levels determined.

Algorithm 1 Energy-Efficiency Power Allocation Algorithm

Initialize $v=0$, $\zeta^{(0)}=0$, step size δ, the Lagrangian multipliers κ, λ, and ψ, $\chi_i=1$ $(1\leq i\leq 4)$, and choose predetermined tolerate ϕ.
repeat
  for $k=1,\ldots,K$ **do**
    Determine the number of users accessing SC k, i.e., Z^k.
    for $z=Z^k,\ldots,1$ **do**
      if $z\in\mathcal{M}_{\mathsf{M}}$ **then**
        Determine $P_z^{(k)}$ as in (23).
      end if
      if $z\in\mathcal{M}_{\mathsf{E}}$ **then**
        Determine $P_z^{(k)}$ as in (29).
      end if
      if $z\in\mathcal{M}_{\mathsf{U}}$ **then**
        Determine $P_z^{(k)}$ as in (34).
      end if
    end for
  end for
  Update $\kappa^{(v+1)}$, $\lambda^{(v+1)}$, and $\psi^{(v+1)}$ as in (31), (32), and (35), respectively.
  Update $\zeta^{(v+1)}=\frac{R^{\mathsf{tot}}(t)}{P^{\mathsf{Tx}}(t)+MP_c}$.
  Update $\chi_1=|\kappa^{(v+1)}-\kappa^{(v)}|$, $\chi_2=|\lambda^{(v+1)}-\lambda^{(v)}|$, $\chi_3=|\psi^{(v+1)}-\psi^{(v)}|$, and $\chi_4=|\zeta^{(v+1)}-\zeta^{(v)}|$.
  Set $v:=v+1$.
until $\chi_i\leq\phi$.

B. CMADDQN-based SC Assignment Strategy

We assume that each eMBB or URLLC user is preassigned several SCs for its transmissions to guarantee its high requirements. Meanwhile, mMTC users can use SCs freely to reduce access latency and increase the number of active users [16]. However, this scenario can lead to severe co-channel interference when multiple users occupy the same SC. To overcome this issue, we investigate a CMADDQN-based DRL method to help mMTC users quickly select the best SCs for interference mitigation. By using the CMADDQN scheme, each mMTC user is mapped to a learning agent and all M_{M} agents are centralized at the BS to exploit full information on users available at the BS [34]. This facilitates the learning process and helps users select the most appropriate SC for their transmissions to maximize the overall network EE.

1) Optimization-based DRL Approach: We denote $\mathcal{S}$, $\mathcal{A}$, and $\mathcal{R}$ as the set of **states**, **actions**, and **rewards**, respectively. At the beginning of each TS t, an agent observes the current state $s(t)\in\mathcal{S}$ to take an action $a(t)\in\mathcal{A}$. Performing the action $a(t)$ can return a reward/penalty from the environment and discover the next state $s(t+1)$. Thanks to the feedback from the environment, the agent can update/strengthen its decision policy. The process can be operated continuously until an optimal policy can be obtained at the agent. Given this context, the definitions of state, action, and reward according to each agent $m\in\mathcal{M}_{\mathsf{M}}$ are described as follows.

- **State**: The state of agent m in TS t, $s_m(t)\in\mathcal{S}_m$ with cardinality of $K(1+M_{\mathsf{M}})$, is defined as the combination of its current channel gains over K SCs and the previous SC selection status of all M_{M} mMTC users, i.e.,

$$s_m(t)=\{\mathbf{g}_m(t),\mathbf{c}(t-1)\}, \quad (36)$$

  where $\mathbf{g}_m(t)=\{g_m^{(1)}(t),\ldots,g_m^{(K)}(t)\}$, $\mathbf{c}(t-1)=\{\mathbf{c}_1(t-1),\ldots,\mathbf{c}_{M_{\mathsf{M}}}(t-1)\}$, and $\mathbf{c}_m(t-1)=\{c_m^{(1)}(t-1),\ldots,c_m^{(K)}(t-1)\}$.
- **Action**: Since each mMTC user m can use only one SC every TS, the action of agent m in TS t, $a_m(t)\in\mathcal{A}_m$, is defined as its current SC selection which is given as

$$a_m(t)\in\mathcal{A}_m=\{1,\ldots,K\}. \quad (37)$$

  As can be observed, the action space size of agent m is determined as $|\mathcal{A}_m|=K$. For the action selection

strategy, the ϵ-greedy policy can be exploited, where the random action is taken with the probability of ϵ and the action with the highest Q-value is employed for the remaining probability. In particular, the action $a_m(t)$ is selected based on the ϵ-greedy policy can be mathematically expressed as

$$a_m(t) = \begin{cases} \text{random action}, & \text{with prob. of } \epsilon, \\ a_m^{\max}, & \text{with prob. of } 1-\epsilon, \end{cases} \quad (38)$$

where, $a_m^{\max} = \arg\max_{a \in \mathcal{A}_m} \{Q(s_m(t), a; \boldsymbol{\theta}_m)\}$, $Q(s_m(t), a_m(t); \boldsymbol{\theta}_m)$ is the Q-value of $(a_m(t), s_m(t))$.

- **Reward**: In MADRL frameworks, both centralized and decentralized reward structures can be employed for building learning models. Specifically, MADRL methods with centralized rewards provide a common reward for all agents, whereas each agent can receive a distinct reward in MADRL methods with decentralized rewards [35]. However, using decentralized rewards can lead to the selfish behavior of agents, where they may compete with each other to maximize their own rewards. This degrades the global performance. To acreess this, the same reward can be allocated to all agents to achieve a common objective [36]. This paper aims to optimize the average network EE while fulfilling the heterogeneous QoS requirements of all users. Therefore, we design a CMADDQN algorithm for SC assignment with centralized rewards to optimize the above common objective. Specifically, we use the achieved EE in (17) to define the immediate common reward for all agents. Thus, the reward function in TS t, denoted by $r(t)$, is defined as

$$r(t) = \begin{cases} \zeta(t), & \text{if all constraints are satisfied,} \\ 0, & \text{otherwise.} \end{cases} \quad (39)$$

Based on the actions and rewards obtained from trials, each agent m builds its own DDQN model consisting of two deep neural networks (DNNs), namely online and target networks corresponding to weight matrices $\boldsymbol{\theta}_m$ and $\boldsymbol{\theta}'_m$, respectively. Herein, the online network is used to select an action. Meanwhile, the target network is applied to evaluate the online-network-based action. Thus, the objective is to reduce the loss function which is formulated as [25]

$$L(\boldsymbol{\theta}_m) = \left[y_m(t) - Q_m(s_m(t), a_m(t); \boldsymbol{\theta}_m)\right]^2, \quad (40)$$

where $y_m(t)$ denotes the target Q-value determined as

$$y_m(t) = r(t) + \gamma Q_m\big(s_m(t+1), \arg\max_{a \in \mathcal{A}_m} \mathcal{Q}_m(t+1); \boldsymbol{\theta}'_m\big), \quad (41)$$

where γ denotes the discount factor, $\mathcal{Q}_m(t+1) = Q_m(s_m(t+1), a; \boldsymbol{\theta}_m)$. It is noteworthy from (41) that in DDQN model, the action selection, i.e., $\arg\max_{a \in \mathcal{A}_m} \mathcal{Q}_m(t+1)$, and action evaluation, i.e., $Q_m(s_m(t+1), \arg\max_{a \in \mathcal{A}_m} \mathcal{Q}_m(t+1); \boldsymbol{\theta}'_m)$, are decoupled using two different Q-value function to avoid the conventional overestimation [37].

2) Joint Optimization and Cooperative MADDQN Algorithm: Based on the above discussions, we propose a joint optimization and cooperative MADDQN method as in Algorithm 2, so-called JOCDDQN, to address the problem (16). The proposed method is the combination of the CMADDQN-based SC assignment solution presented in Section III-B and the dynamic PA approach proposed in Algorithm 1. Specifically, the JOCDDQN method with the integration of the CMADDQN-based SC assignment and PA algorithms, shown in Algorithm 2, can be depicted through the following main steps:

Algorithm 2 JOCDDQN ALGORITHM FOR MAXIMIZING EE

```
Initialize the weight matrices of the online and target networks, i.e., θ_m and θ'_m, ∀m ∈ M_M.
for i = 1, ..., E_p do
    Initialize the state s_m(t), ∀m.
    for t = 1, ..., T do
        All agents take an action (SC selection) a_m(t) (∀m) following the ε-greedy policy in (38).
        Run Algorithm 1 to achieve the optimal Tx-power for all users.
        The BS broadcasts the SC selection and the PA to users.
        All agents observe the reward r(t) in (39) and move to the next states s_m(t+1) (∀m).
        for m = 1, ..., M_M do
            Store an experience tuple (s_m(t), a_m(t), r(t), s_m(t+1)) to the memory of agent m.
            Randomly sample a mini-batch of experiences from the memory to train the online network.
            Update θ_m by using gradient descent to minimize the loss function in (40).
            if t%F = 0 then
                Set θ'_m = θ_m.
            end if
        end for
    end for
end for
```

- **Step 1 - Input collection:** The input information (state) on channel gains and SCs status are collected.
- **Step 2 - CMADDQN-based SC Assignment:** Based on the input information, the CMADDQN is applied for SC allocation, where each agent m mapping to mMTC user m observes its current state $s_m(t) \in \mathcal{S}_m$ and takes an action of SC selection $a_m(t) \in \mathcal{A}_m$. Thus, the actions of all agents can be expressed as a vector $\boldsymbol{a}(t) = \{a_1(t), \ldots, a_m(t), \ldots, a_{M_\mathbf{M}}(t)\}$.
- **Step 3 - Input PA algorithm:** The output of the CMADDQN algorithm, i.e., $\boldsymbol{a}(t)$, and the input information are then fed into the PA algorithm in Algorithm 1.
- **Step 4 - PA decision:** Using the input provided, the PA method in Algorithm 1 is then utilized to achieve optimal Tx-power for all users. The output power vector for all users is indicated as $\boldsymbol{P}(t) = \left\{\boldsymbol{P}^u(t), \boldsymbol{P}^e(t), P_m^{a_m(t)}(t)\right\}$ $\forall (u \in \mathcal{M}_\mathbf{F}, e \in \mathcal{M}_\mathbf{E}, m \in \mathcal{M}_\mathbf{M})$. Herein, $P_m^{a_m(t)}(t)$ is the power of mMTC user m over the SC identified from the selected action $a_m(t)$; $\boldsymbol{P}^u(t) = \{P_1^u(t), \ldots, P_l^u(t)\}$ and $\boldsymbol{P}^e(t) = \{P_1^e(t), \ldots, P_n^e(t)\}$ denote the power vector allocated to the URLLC user u and the eMBB user e over preassigned SCs, respectively, as defined in Section III-A.
- **Step 5 - Feedback:** Based on the agents' selected actions and optimized power allocated to users, the EE is determined using (17) and agent m observes the environment to receive a reward $r(t)$ and moves to a new state $s_m(t+1)$. Thus, the output of the PA algorithm is utilized to identify the reward of the learning model in

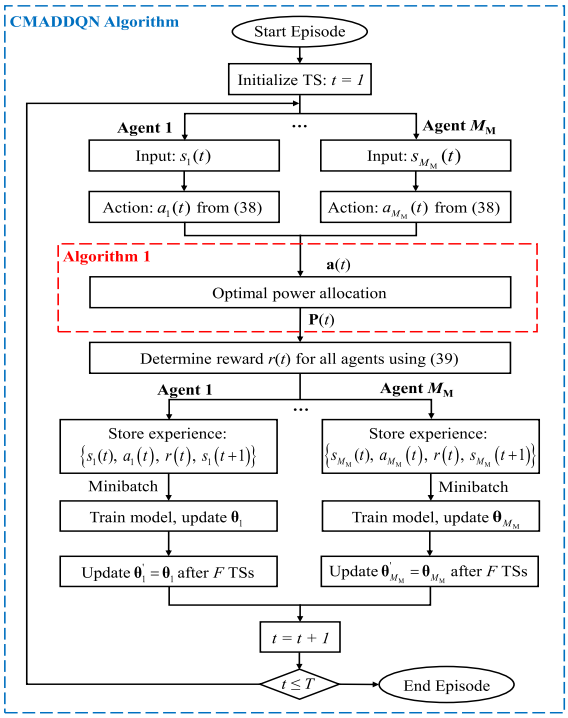


Fig. 2: Flowchart of Algorithm 2.

the CMADDQN algorithm.

- **Step 6 - Train and update the learning model of CMADDQN scheme:** Agent m 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 weight matrix of the online network $\boldsymbol{\theta}_m$ is then updated to minimize the loss function in (40) by using the stochastic gradient method. After a predetermined number of TSs (F), the weight matrix of the target network $\boldsymbol{\theta}'_m$ is updated by copying $\boldsymbol{\theta}_m$.
- **Step 7 - Iteration:** At the next TSs, steps 1-6 are repeated until reaching a predefined number of episodes (NoE) guaranteeing the algorithm's convergence.

The single-episode flowchart of Algorithm 2 is provided in Fig. 2 showing the implementation steps and the integration of the CMADDQN and PA algorithms into this algorithm.

IV. Proposed Full Distributed Reinforcement Learning Method and Further Discussion

In this section, we present another DRL-based solution to address (18) by developing a full MADDQN method, namely FDDQN, in a distributed manner to reduce the information exchange between the BS and users. The FDDQN method is designed to conduct both SC assignment and PA.

A. FDDQN Method

Employing FDDQN method, all users are considered as learning agents to find the optimal policies for selecting both SC and Tx-power. In addition, the multi-level quantization strategy is exploited to deal with the continuous characteristic of power variables in the similar approach introduced in [24], [25]. Specifically, we investigate a scenario where mMTC users build their own DDQN model to learn how to choose the best SC and power level for their transmissions from the available SCs and TPLs sets. Furthermore, eMBB and URLLC users are also learning agents to select suitable TPLs for their communication over preassigned orthogonal SCs.

1) MADDQN-based DRL Model: Given the above context, the states, actions, and rewards of agents according to eMBB, mMTC, and URLLC users are defined as follows:

- State: Considering a TS t, the state definitions of agents according to user $z \in \mathcal{M}_{\mathsf{E}} \cup \mathcal{M}_{\mathsf{U}}$ and user $m \in \mathcal{M}_{\mathsf{M}}$ are, respectively, expressed as

$$s_z(t) = \{\mathbf{g}_z(t), \mathbf{p}_z(t-1)\}, \tag{42}$$

$$s_m(t) = \{\mathbf{g}_m(t), \mathbf{c}_m(t-1), \mathbf{p}_m(t-1)\}, \tag{43}$$

  where $\mathbf{g}_z(t) = \left\{g_z^{(k_1^z)}(t), \ldots, g_z^{(k_b^z)}(t)\right\}$ ($b = n$ if $z \in \mathcal{M}_{\mathsf{E}}$ and $b = l$ if $z \in \mathcal{M}_{\mathsf{U}}$) is the channel gain vector of user z over assigned SCs, $\mathbf{p}_z(t-1) = \left\{P_z^{(k_1^z)}(t-1), \ldots, P_z^{(k_b^z)}(t-1)\right\}$ denotes the TPLs selection of user z over assigned SCs, $\mathbf{g}_m(t)$ and $\mathbf{c}_m(t-1)$ are defined in (36), and $\mathbf{p}_m(t-1) = \left\{P_m^{(1)}(t-1), \ldots, P_m^{(K)}(t-1)\right\}$.
- Action: In TS t, the actions of agents according to user $z \in \mathcal{M}_{\mathsf{E}} \cup \mathcal{M}_{\mathsf{U}}$, $a_z(t)$, and user $m \in \mathcal{M}_{\mathsf{M}}$, $a_m(t)$, are defined as the power selection of user z over n granted SCs, and the SC and power selection of mMTC user m, respectively. Thus, the actions $a_z(t)$ and $a_m(t)$ are, respectively, written by

$$a_z(t) = \left\{P_z^{(k_1^z)}(t), \ldots, P_z^{(k_b^z)}(t)\right\} \in \hat{\mathcal{A}}_z, \tag{44}$$

$$a_m(t) \in \hat{\mathcal{A}}_m = \{1, \ldots, kl, \ldots, KL\}, \tag{45}$$

  where $P_z^{(k)}(t) \in \mathcal{P}$, $\mathcal{P} = \{P_1, \ldots, P_L\}$ is the available L TPLs set, $a_m(t) = kl$ indicates that mMTC user m selects SC k and TPL l in TS t. Thus, the action space size of agents z and m are $\left|\hat{\mathcal{A}}_z\right| = L^b$ and $\left|\hat{\mathcal{A}}_m\right| = KL$.
- Reward: Similar to the JOCDDQN method, the reward function in TS t for the FDDQN approach is also determined based on the achieved EE as in (39).

2) Full MA Double Deep Q-Network Algorithm: The FDDQN method requires each agent to create its own DDQN model following the same process as described in Section III-B. The details of the proposed FDDQN algorithm are summarized in Algorithm 3. Particularly, at TS t, every agent $z \in \mathcal{M}$ assesses its present state $s_z(t)$ and makes actions of both SC and power level selections, denoted as $a_z(t)$. Subsequently, the agent observes the environment, receives a reward $r(t)$, and transitions to a new state $s_z(t+1)$. It then records this experience, trains its online network, and adjusts the weight matrices of both the online and target networks following a procedure akin to Algorithm 2. This process repeats at TS $t+1$, involving another round of SC and power level selection. Iteration continues until reaching a predetermined number of episodes, ensuring the convergence of the algorithm.

Algorithm 3 FDDQN ALGORITHM FOR MAXIMIZING EE

Initialize the weight matrices of the online and target networks, i.e., $\boldsymbol{\theta}_z$ and $\boldsymbol{\theta}'_z$, $\forall z \in \mathcal{M}$.
for $i = 1, \ldots, E_p$ **do**
  Initialize the state $s_z(t)$, $\forall z$.
  for $t = 1, \ldots, T$ **do**
    All agents take an action $a_z(t)$, $\forall z$, following the ε-greedy policy in (38), where $a_z(t)$ is defined in (44) if $z \in \mathcal{M}_{\mathsf{E}} \cup \mathcal{M}_{\mathsf{U}}$ and in (45) if $z \in \mathcal{M}_{\mathsf{M}}$.
    All agents observe the reward $r(t)$ in (39) and move to the next states $s_z(t+1)$ ($\forall z$).
    for $z = 1, \ldots, M$ **do**
      Perform steps of storing experience, training the online network, and updating $\boldsymbol{\theta}_z$ and $\boldsymbol{\theta}'_z$, according to agent z similar to steps $10 - 15$ in Algorithm 2.
    end for
  end for
end for

B. Complexity Analysis

1) FDDQN Algorithm: Let H, N_h, and I_s be the number of training layers, the number of neurons in layer h, and the size of the input layer. For each TS, the computational complexity of FDDQN algorithm in algorithm 3 can be calculated by $C_{\mathsf{TS}} = \mathcal{O}(X)$, 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 complexity of the algorithm is given by

$$C_{\mathsf{FDDQN}} = \mathcal{O}(METX). \tag{46}$$

2) JOCDDQN Algorithm: The complexity of the JOCDDQN algorithm in Algorithm 2 consists of the complexity of the DPA algorithm in Algorithm 1 and the CMADDQN scheme. For the DPA algorithm, its complexity is given by

$$C_{\mathsf{DPA}} = \mathcal{O}(I(M_{\mathsf{M}} + nM_{\mathsf{E}} + lM_{\mathsf{U}})), \tag{47}$$

where I denotes the number of iterations to get convergence. For CMADDQN scheme, its complexity is determined similarly to the one of the FDDQN algorithm as

$$C_{\mathsf{CMADDQN}} = C_{\mathsf{FDDQN}} = \mathcal{O}(METX). \tag{48}$$

Thus, the JOCDDQN-Alg. complexity can be calculated as

$$C_{JOCDDQN} = C_{\mathsf{DPA}} C_{\mathsf{CDDQN}}. \tag{49}$$

C. Signaling Overhead Analysis

Signaling overhead (SO) refers to the information-bit number required to transmit channel status data, subcarrier indicators (SCI), and Tx-power of a specific user over an SC [40]. Factors such as the total number of users, SCs, and the exchange of states and rewards between agents and the environment also influence the SO. Higher SO leads to greater computational complexity and processing time. According to [40], transmitting a continuous value for channel status, data rate, and reward requires 16 bits, while 4 bits are used for SCI, Tx-power, and other related parameters. Table II compares the SO of the proposed methods with several related works [21], [38], [39], where $M_{\mathsf{UE}} = M_{\mathsf{U}} + M_{\mathsf{E}}$. In particular, the related works have high SOs due to the need for feedback, including data rates, rewards, power sets, pilot signals, CSI, and interference thresholds (IT). The proposed JOCDDQN method's SO stems from the pilot signals used for channel estimation and feedback of the SCI and Tx-powers from the BS to users. For the proposed FDDQN method, the overhead is based on feedback related to CSI, rewards, and power sets from the BS to users. Unlike the works [21], [39], which assume fixed Tx-power for high-demand users (i.e., URLLC and eMBB users), we consider dynamic Tx-power for all users to optimize network EE, leading to additional SO.

D. Convergence Discussion

This part provides a discussion regarding the convergence of the proposed learning methods. In particular, the learning mechanisms in the JOCDDQN and FDDQN methods are built based on the DDQN model which combines Q-learning and neural networks (NNs). Therefore, their convergence properties can be described by considering the convergence conditions of the Q-learning and NNs' possibility of effectively approximating the non-linear Q-values [41]. Firstly, the convergence constraints of the Q-learning algorithm are expressed as [42], [43]

$$\lim_{T\to\infty} \sum_{t=1}^{T} \alpha_t = +\infty \quad \& \quad \lim_{T\to\infty} \sum_{t=1}^{T} \alpha_t^2 < +\infty, \tag{50}$$

where $0 \leq \alpha \leq 1$ denotes the learning rate. The conditions in (50) indicate that it is necessary to progressively reduce the learning rate during the training process to ensure the algorithm's convergence. Secondly, the works in [43], [44] showed that NNs can approximate any non-linear continuous functions. Given the above discussions, the proposed methods can achieve the convergence status.

E. Resource Allocation under Imperfect CSI and SIC

This section delves into adapting our proposed SC and PA strategies for scenarios involving imperfect CSI (iCSI) and imperfect SIC (iSIC).

1) Imperfect CSI Implementation: Under iCSI, the channel coefficient $h_z^{(k)}(t)$ of user $z \in \mathcal{M}$ can be estimated as

$$\hat{h}_z^{(k)}(t) = h_z^{(k)}(t) + \eta, \tag{51}$$

where $\hat{h}_z^{(k)}(t)$ and $h_z^{(k)}(t)$ denote the estimated and true channel coefficients, respectively. In addition, $\eta \sim \mathcal{CN}(0, \sigma_e^2)$ indicates the estimated error, which is a random variable approximated by a Gaussian distribution with zero mean and variance σ_e^2. To assess the effect of iCSI, we consider a scenario as follows.

- For the JOCDDQN method with iCSI, the BS employs the estimated channel based on iCSI to formulate the state $s_m(t)$ in (36) and conduct the PA for all users in Algorithm 1. The users then use the allocated Tx-power to transmit their messages over the true channel. Subsequently, the BS calculates the obtained EE to determine the reward for the learning process.
- For the FDDQN method with iCSI, a similar approach can be applied. In particular, the iCSI-based estimated channel is used for creating the agent state in (42) and (43) and actions of SC and power level selection. After that, users employ the selected Tx-power to communicate with the BS over the true channel and receive feedback.

TABLE II: Quantitative Signaling Overhead Comparison

References	Optimization Method	Signalling Overhead
[21]	Decentralized MADRL	$\underbrace{16M_{\mathrm{M}}}_{\text{Data rates}} + \underbrace{16}_{\text{Reward}} + \underbrace{16}_{\text{PS}}$
[38]	Conventional	$\underbrace{2}_{\text{Pilot}} + \underbrace{4M_{\mathsf{UE}}}_{\text{Power}} + \underbrace{16M_{\mathsf{UE}}}_{\text{CSI}} + \underbrace{4}_{\text{IT}}$
[39]	Decentralized MADRL	$\underbrace{16KM_{\mathbf{M}}}_{\text{CSI}} + \underbrace{16}_{\text{Reward}}$
Proposed JOCDDQN	Centralized joint optimization and MADRL	$\underbrace{4M}_{\text{Pilot}} + \underbrace{4M}_{\text{SCI}} + \underbrace{4M}_{\text{Power}}$
Proposed FDDQN	Decentralized MADRL	$\underbrace{16KM}_{\text{CSI}} + \underbrace{16}_{\text{Reward}} + \underbrace{16}_{\text{PS}}$

2) Imperfect SIC Solution: Given this context, the SINR in (4) is rewritten as

$$\gamma_{z_\ell^{(k)}}^{(k)}(t) = \frac{\mathcal{Y}_{z_\ell^{(k)}}^{(k)}(t)}{\sum_{j=1}^{\ell-1} \psi \mathcal{Y}_{z_j^{(k)}}^{(k)}(t) + \sum_{j=\ell+1}^{Z^k} \mathcal{Y}_{z_j^{(k)}}^{(k)}(t) + \sigma_k^2}, \tag{52}$$

where $0 \leq \psi \leq 1$ denotes the imperfect SIC factor, where higher values of ψ implies larger interference; and $\mathcal{Y}_z^{(k)}(t) = P_z^{(k)}(t) \left| h_z^{(k)}(t) \right|^2$. In Section III-A, we propose a PA method under perfect SIC (pSIC), where the residual interference $\sum_{j=1}^{\ell-1} \psi \mathcal{Y}_{z_j^{(k)}}^{(k)}(t)$ is perfectly removed to decode the message of user $z_\ell^{(k)}$. However, the presence of this component in the case of iSIC makes the problem (18) become more challenging to solve due to the non-convex log term in the achievable rate. To overcome this issue, we introduce a new PA approach using a loop method as follows.

To tackle the residual interference component in (52), we define an initial power value set $\left\{ P_{z_1^{(k)}}^{(0)}, \ldots, P_{z_{\ell-1}^{(k)}}^{(0)} \right\}$ for this term and treat it as noise to determine the Tx-power for user $z_\ell^{(k)}$. After that, the PA approach presented in Section III-A can be utilized to calculate the Tx-power $P_{z_\ell^{(k)}}^{(1)}$ for all users. The power set is then updated as $\left\{ P_{z_1^{(k)}}^{(0)}, \ldots, P_{z_{\ell-1}^{(k)}}^{(0)} \right\} = \left\{ P_{z_1^{(k)}}^{(1)}, \ldots, P_{z_{\ell-1}^{(k)}}^{(1)} \right\}$ to initiate another loop. This process continues until reaching a predetermined number of loops guaranteeing optimal PA for all users. To assess the convergence of this method, simulation results across varying numbers of loops are provided in Fig. 11 of Section V.

V. Simulation Results

This section provides the simulation results to evaluate our proposed algorithms' performance. The DDQN model consists of three fully-connected hidden layers including 256, 128, and 64 neurons. The experimental parameters are provided in Table III. To investigate the channel conditions, we generate different Rician channel realizations randomly for all users at each time-slot during the training and testing phases. Here, we investigate the behavior and the performance of our proposed algorithms (i.e., JOCDDQN and FDDQN) as well as the benchmark methods which are described as follows:

- Fixed Tx-power (FTP): In this scheme, each user transmits their messages with a predetermined Tx-power [11]. Thus, the MADDQN algorithm is applied only for SC assignment in this case.
- Fixed RA (FRA): This scheme is considered in [21], where SC and PA for high-demand users (i.e., URLLC and eMBB users) is fixed, whereas mMTC users can find the best RA policy based on MADDQN scheme.
- Random SC selection (RSS): The SC assignment is carried out randomly in this RSS method. Based on this SC setting, Algorithm 1 is implemented for optimal PA.

Since the hyper-parameters significantly affect the learning process in the DRL algorithms, we first evaluate the effect of the learning rate (α) on the convergence performance in terms of reward in Fig. 3a. Note that this figure only investigates the convergence behaviour of the FDDQN method versus different values of α. This figure shows that varying the values of α lead to distinct convergence behaviours. Specifically, smaller values of α (i.e., $\alpha = 0.0001$ and $\alpha = 0.00001$) cause the algorithm to converge to lower rewards than the cases $\alpha = 0.01$ and $\alpha = 0.001$. Additionally, to further clarify the influence of α on the system performance, we plot the achieved EE in the testing phase versus the learning rate across different action space scenarios (i.e., variations in the number of SCs (K) and TPLs (L)) in Fig. 3b. The figure indicates that the case of $\alpha = 0.001$ provides better EE performance than other cases. Moreover, we can observe from Figs. 3a and 3b that a lower learning rate can make the agents learn a poor policy leading to low performance in terms of reward and EE. Conversely, a higher learning rate accelerates learning but risks rapid convergence to sub-optimal solutions. Therefore, careful selection of the learning rate is necessary. Based on the results observed from Fig. 3, we set the learning rate value as 0.001.

Fig. 4 depicts the convergence behavior of the proposed approaches (i.e., JOCDDQN and FDDQN) by plotting the achieved reward versus number of episodes. Furthermore, different values of the number of TPLs (L) are considered in this figure. One can observe from Fig. 4 that both JOCDDQN and FDDQN methods can achieve the convergence status after a number of learning episodes. More specifically, the JOCDDQN method obtains higher reward and converges faster than the FDDQN method. This is because the JOCDDQN method utilizes the CMADDQN scheme to select only SCs for users, resulting in a small action space; and applies the DPA algorithm in Algorithm 1 to attain an optimal PA, leading to a performance improvement in terms of the achieved reward and convergence. In contrast, the FDDQN method is designed by using the power quantization (PQ) method [21], [24] to

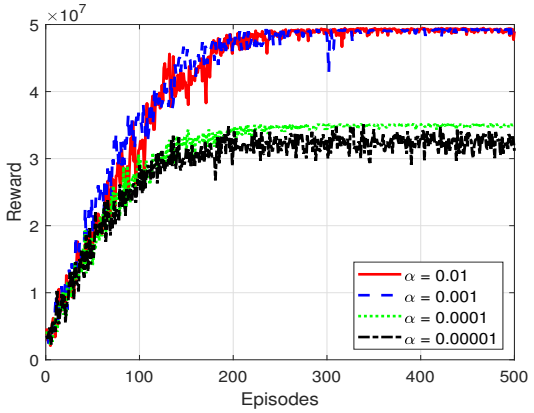
TABLE III: Experimental Parameters

Parameters	Value
Cell radius (r)	500 m
Channel model	Rician
eMBB and URLLC numerology indices (ν_{E} & ν_{U})	1 & 2
Number of SCs (K)	4
Number of URLLC users (M_{U})	1
Number of eMBB users (M_{E})	1
Number of mMTC users (M_{M})	6
Number of SCs assigned to a URLLC/eMBB user	2
eMBB data-rate demand $\left(R_e^{\mathsf{tar}}\right)$	2 bps/Hz
URLLC & mMTC latency threshold (D_u & D_m)	1 & 2 ms
URLLC & mMTC reliability threshold (ε_u & ε_m)	10^{-5} & 10^{-3}
Maximum Tx-power	23 dBm
Circuit power consumption (P_c)	0.05 W
Noise figure & PSD (F & N_0)	6 dB & -174 dBm/Hz
Packet length (m_b)	32 bytes
Number of hidden layers	3
Number of neurons per hidden layers	{256, 128, 64}
Learning rate (α)	0.001
Discount factor (γ)	0.9
Optimizer	Adam

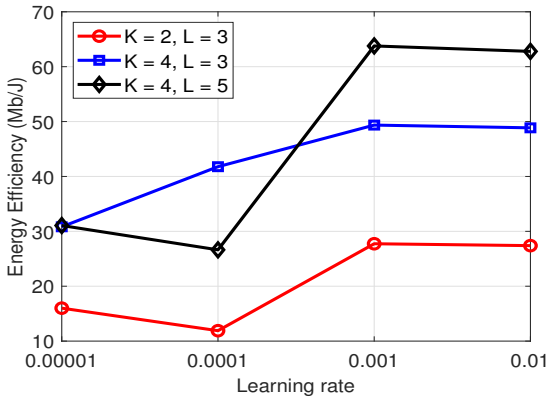
split the Tx-power into multiple discrete power levels, and the MADDQN scheme is then applied for SC and power level selection. This significantly increases the action space requiring agents to spend more time (episodes) exploring the environment. Additionally, the PQ method makes it challenging to obtain an optimal PA policy. Therefore, the FDDQN method demonstrates inferior learning performance compared to the JOCDDQN method. Furthermore, Fig. 4 shows that the FDDQN method can get a higher reward towards the one achieved by the JOCDDQN method when increasing L. However, the value of L should be selected carefully since increasing L makes the action space larger leading to a longer learning process. Note that the JOCDDQN is not influenced by the values of L since it does not employ the PQ method.

Fig. 5 shows the variation of EE versus the number of mMTC users (M_{M}) of different approaches including JOCDDQN, FDDQN, RSS, FRA, and FTP. This figure indicates that lower EE performance for all considered methods can be observed with the increase in the value of M_{M} due to higher interference. Moreover, the proposed JOCDDQN and FDDQN methods outperform other investigated methods in terms of EE performance when M_{M} increases. In particular, the proposed JOCDDQN method achieves the best EE performance thanks to its joint optimization and DRL scheme, whereas the RSS method gives the worst performance due to the random SC selection strategy applied in this method. Among the remaining approaches (i.e., FDDQN, FRA, and FTP), the proposed FDDQN method attains higher EE performance as compared to others. It can be explained by the fact that the FRA and FTP methods are built based on some ideal assumptions leading to their performance degradation. Specifically, the FRA method fixes the Tx-power of high-demand users (i.e., URLLC and eMBB users), whereas the FTP method considers a communication protocol with fixed Tx-power for all users.

In addition, Fig. 5 also provides the results obtained using another advanced DRL technique, known as duelling double deep Q network (D3QN) [45], for further comparative analysis. The D3QN is expected to improve learning efficiency by applying a duelling architecture to expedite the identification of important states and optimal actions, as discussed in [21], [45]. Within Fig. 5, we explore two different applications of



(a) Convergence performance with different learning rate values.



(b) Effect of learning rate on the EE performance in the testing phase with different number of SCs (K) and TPLs (L).

Fig. 3: Learning rate analysis.

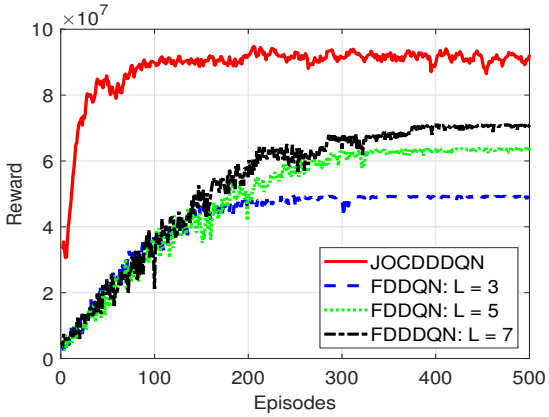


Fig. 4: Effect of TPLs (L) on convergence performance.

the D3QN method in the studied context: (i) Joint optimization and cooperative MA D3QN (JOCD3QN): This scheme utilizes the D3QN for SC assignment while employing the methodology outlined in Algorithm 1 for dynamic PA, similar to the method proposed in Section III; and (ii) Decentralized full MA D3QN (FD3QN): The distributed D3QN method is applied for both SC and PA, mirroring the methodology introduced in Section IV. Fig. 5 shows that the JOCD3QN and FD3QN methods yield EE performances comparable to the proposed JOCDDQN and FDDQN approaches, respectively. However, it is noted that the former methods exhibit higher complexity due to the implementation of the duelling architecture.

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2024.3476083

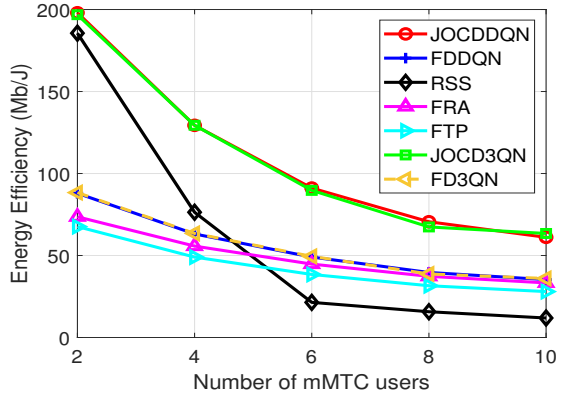


Fig. 5: The EE performance vesrus number of mMTC users.

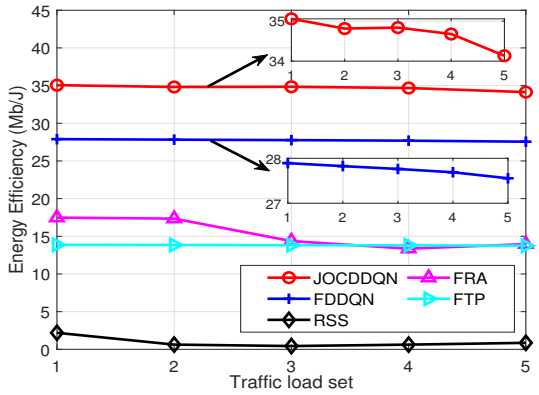


Fig. 8: Effect of traffic load on the EE performance, where $K = 2$.

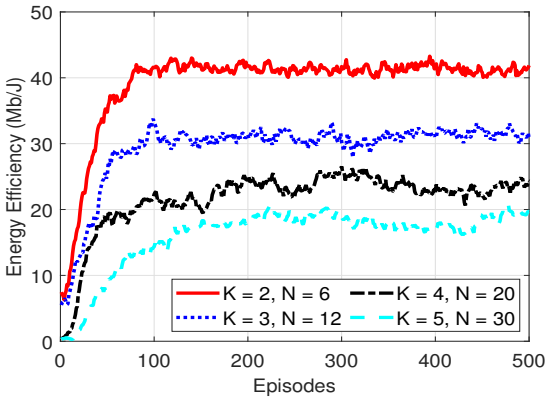


Fig. 6: JOCDDQN performance with various network sizes.

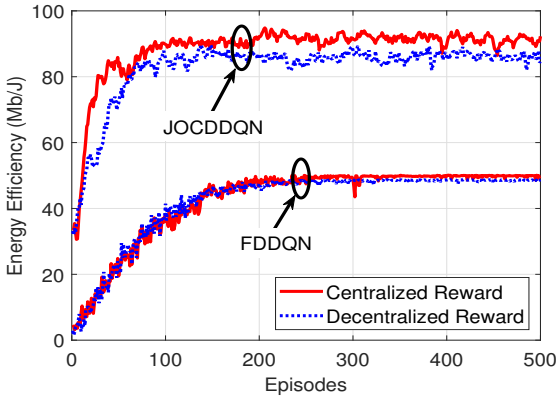


Fig. 9: Effect of centralized and decentralized rewards.

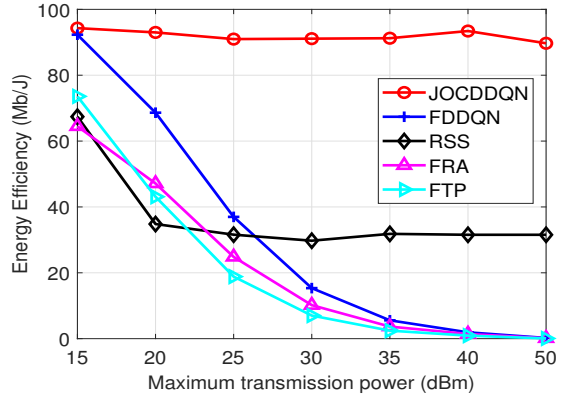


Fig. 7: Effect of maximum transmission power.

Fig. 6 provides the EE performance validation across different network sizes with the JOCDDQN method. In this context, the number of SCs is defined as $K = \{2, 3, 4, 5\}$, while the number of agents (users) is set to $N = \{6, 12, 20, 30\}$, corresponding to the user density (UD) values of UD $= N/K = \{3, 4, 5, 6\}$. Fig. 6 shows that the EE performance decreases as the UD increases. This degradation occurs because a larger UD implies that N is much greater than K, leading to stronger interference among users sharing the same SCs. Moreover, this figure highlights the scalability of our proposed method, demonstrating its ability to converge and guarantee users' requirements even in highly overloaded scenarios where the number of users far exceeds the available SCs. Additionally, new agents can be seamlessly integrated into the system by replicating the parameters of existing trained agents, facilitating the generalization of our method. Thus, our proposed method is well-suited for H-NOMA systems with a large user base and limited network resources.

Fig. 7 depicts the effect of the maximum Tx-power ($P_{\max}$) on the achieved EE performance of different approaches. Herein, we consider a wide range of $P_{\max}$ varying from 15 dBm to 50 dBm. Furthermore, for the power quantization (PQ) approach utilized in the FDDQN method, we design a Tx-power level (TPL) set for users' selection including L levels. Each power level (P_i) within this set is calculated as $P_i = i \times P_{\max}/L$ for $1 \leq i \leq L$. In Fig. 7, the value of L is set to 3. It is noteworthy that during the training process, the total Tx-power allocated to users in a time slot (TS) can exceed the set range of $P_{\max}$, as users (URLLC and eMBB) can utilize multiple subcarriers (SCs). This scenario provides valuable learning experience for the agents to develop optimal policies while respecting users' available power budgets. Fig. 7 demonstrates that the proposed JOCDDQN method can still bring the best EE performance when $P_{\max}$ gets larger. In contrast, the EE performance achieved by the FDDQN, FRA, and FTP methods is significantly reduced with the increase in $P_{\max}$. This phenomenon occurs since the PQ approach utilized in these methods leads to higher Tx-power when $P_{\max}$ scales up, causing the EE performance loss. Meanwhile, the RSS method can get higher EE performance than the FDDQN, FRA, and FTP methods when $P_{\max}$ becomes much larger.

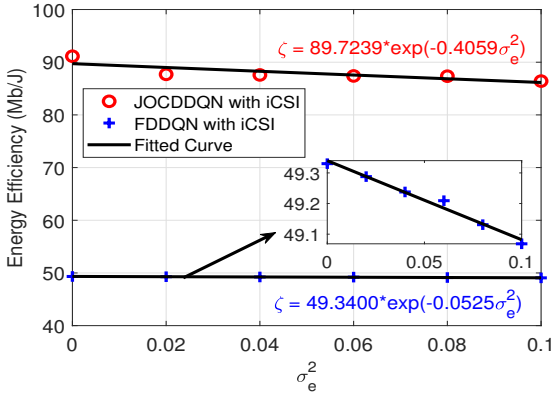


Fig. 10: Effect of iCSI on the EE performance.

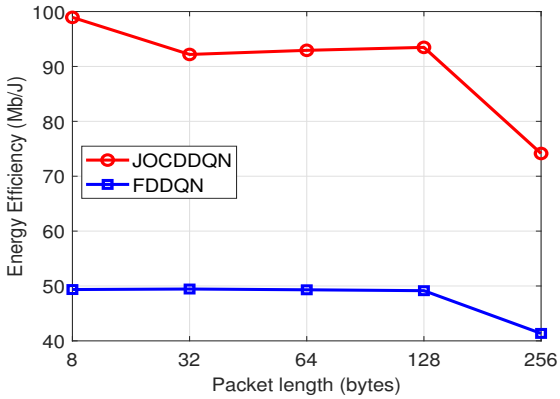


Fig. 11: EE of different approaches versus the packet length.

However, its random SC selection makes it difficult to obtain an optimal solution.

Fig. 8 evaluates the impact of traffic load on the achieved EE performance of different approaches. The traffic load reflects users' SINR or data rate requirements, which increase with higher demands. In this context, we define the traffic load sets through the users' requirements on latency D_u, reliability ε_u, and data rate R_e^{tar} as $(D_u(i), \varepsilon_u(i), R_e^{\text{tar}}(i))$, where $1 \leq i \leq 5$, $D_u = \{4, 3, 2, 1, 0.5\}$ (ms), $\varepsilon_u = \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}\}$, and $R_e^{\text{tar}} = \{2, 4, 6, 8, 10\}$ (bps/Hz). As observed from Fig. 8, when the traffic load gets higher, the worse EE performance can be observed. In addition, this figure also indicates that the proposed JOCDDQN method outperforms other schemes, i.e., the FDDQN, RSS, FRA, and FTP methods across various traffic load scenarios.

The computational complexity of an algorithm has crucial practical implications that can impact system design and performance. In particular, high complexity can increase processing time, leading to greater latency. To clarify this concern, Table IV provides a convergence time comparison of various methods. Among the two proposed methods, the JOCDDQN scheme converges to a larger EE value with fewer episodes but requires a higher average time per episode than the FDDQN, due to its smaller action space and power optimization process. The JOCD3QN method converges to the similar EE value to the JOCDDQN at the same number of episodes but needs more average time per episode. In contrast, the FRA and FTP methods converge faster than the others but

TABLE IV: Convergence Time Comparison

Parameters	JOCDDQN	FDDQN	JOCD3QN	FRA	FTP
Avg. time per episode (s)	4.94	4.28	5.79	3.25	3.19
No. of episode to converge	**95**	190	95	101	29
Convergence time (s)	**469.3**	813.2	550.05	328.25	92.51
EE (Mb/J)	**91.67**	48.97	91.01	44.43	38.37

with significantly lower EE. Overall, the JOCDDQN method demonstrates superiority, offering the highest EE performance with a moderate convergence time.

Fig. 9 presents a comparison of EE performance between methods utilizing centralized and decentralized rewards. The centralized reward, as defined in equation (39), contrasts with decentralized rewards where each agent's reward depends on its individual transmission outcome. Specifically, in pursuit of maximizing EE while guaranteeing the heterogeneous requirements of all users, an agent z's decentralized reward is determined as the ratio of its achievable rate to its Tx-power, i.e., $r_z(t) = \sum_{k\in\mathcal{K}} c_z^{(k)}(t) R_z^{(k)}(t) / \sum_{k\in\mathcal{K}} P_z^{(k)}(t)$, if a successful transmission is achieved, and is zero, i.e., $r_z(t) = 0$, otherwise. Fig. 9 illustrates that EE performance achieved with decentralized rewards is inferior to that achieved with centralized rewards. This difference arises because decentralized rewards can encourage selfish behavior among agents, leading them to prioritize individual objectives over the collective goal of maximizing overall EE while accommodating diverse user requirements. Consequently, a degradation in global EE performance under decentralized reward can be observed.

In practice, acquiring pCSI is challenging. Fig. 10 is provided to assess the effect of iCSI on the EE performance by plotting the achieved EE against different values of the estimated error variance σ_e^2 for the JOCDDQN and FDDQN methods. The considered iCSI scenario is presented in Section IV-E1. This figure shows that the EE achieved under iCSI (i.e., $\sigma_e^2 > 0$) is lower than that with pCSI (i.e., $\sigma_e^2 = 0$). This is because when $\sigma_e^2 > 0$, the Tx-PA for all users is determined based on iCSI, not optimally suited for the true channel based on pCSI, degrading the system EE. In particular, for the JOCDDQN method, the EE drops by 5.18%, from 91.1358 Mb/J with pCSI to 86.4150 Mb/J at error variance value $\sigma_e^2 = 0.1$. Meanwhile, the FDDQN method shows a decrease of 0.53%, from 49.3302 Mb/J to 49.0685 Mb/J. The relationship between the EE and σ_e^2 can be further quantitatively analyzed by developing a relation function using Matlab's curve fitting tool [46]. Given the Gaussian distribution of the estimated error, we utilize the exponential function for fitting distribution and model the relationship as $\zeta = \hat{\alpha} e^{-\hat{\beta}\sigma_e^2}$, where ζ stands for the EE, $\hat{\alpha}$ and $\hat{\beta}$ denote the empirical parameters obtained through the fitting process. This yields the relation functions $\zeta = 89.7239 e^{-0.4059\sigma_e^2}$ for the JOCDDQN method and $\zeta = 49.3400 e^{-0.0525\sigma_e^2}$ for the FDDQN method. These functions can provide insights into how different levels of channel estimation accuracy affect the EE.

Fig. 11 shows the effect of the packet length on the EE performance when employing the proposed JOCDDQN and FDDQN methods. As can be seen from this figure, the variation in the EE performance is small as the packet length varies

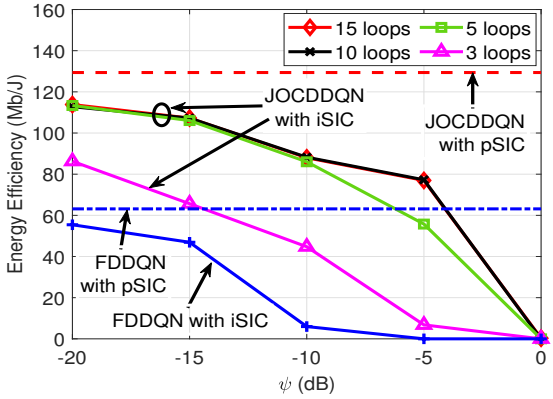


Fig. 12: Effect of iSIC on EE performance, where $M_m = 4$.

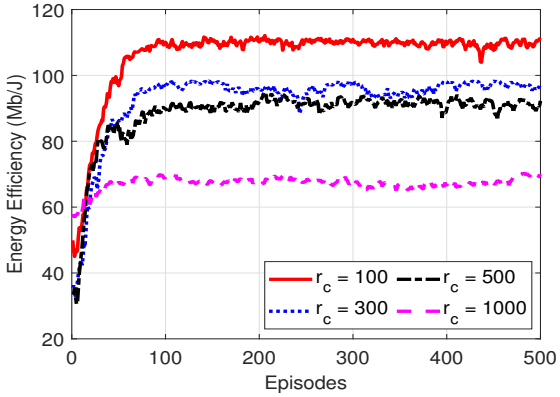


Fig. 13: JOCDDQN performance with various channel conditions.

from 8 (bytes) to 128 (bytes), particularly evident in the case of using the FDDQN method. However, as the packet length gets much higher, i.e., 256 (bytes), a significant degradation in the EE performance can be observed with both JOCDDQN and FDDQN methods, attributed to stricter QoS requirements imposed on users under this scenario.

Fig. 12 provides the EE evaluation achieved by the proposed JOCDDQN and FDDQN methods under the iSIC scenario by plotting the variation of the achieved EE versus different values of the iSIC factor ψ. Here, we also provide the EE attained with pSIC case for comparison purposes. In addition, using the loop method presented in Section IV-E2, we show the EE results of the JOCDDQN method with iSIC across different numbers of loops to validate its convergence. Particularly, Fig. 12 demonstrates that under iSIC context, the JOCDDQN method outperforms the FDDQN method in terms of the EE performance. Furthermore, both methods approach their respective EE results under pSIC as ψ gets smaller. Meanwhile, the increase in the value of ψ leads to a significant degradation in the EE performance due to larger interference. Additionally, as the number of loops scales up, the JOCDDQN method with iSIC exhibits higher EE performance, with a convergence of results observed between cases with 10 and 15 loops.

Fig. 13 illustrates the effect of channel conditions on the EE performance with the JOCDDQN method. The channel conditions vary according to the distance from users to the BS. To represent this, we adjust the cell radius r_c, which indicates the maximum user-BS distance, with users uniformly distributed within this radius. Fig. 13 shows that stronger channel conditions, indicated by shorter user-BS distance (smaller r_c), result in improved EE performance. Conversely, increasing r_c leads to weaker channel conditions and a degradation in EE performance due to higher path loss.

Thanks to the results, we can conclude that the proposed JOCDDQN method demonstrates superior performance over other considered approaches in terms of EE and convergence rate thanks to its joint optimization and cooperative MADDQN framework. However, this approach comes with a higher complexity. Meanwhile, although the proposed FDDQN method outperforms other benchmarks, it shows lower EE performance and convergence characteristics than the proposed JOCDDQN method. Notably, it allows for a distributed implementation with reduced complexity as compared to the JOCDDQN method. In addition, while the integration of optimization and DRL in the JOCDDQN method can expedite the learning process of the proposed algorithm, it encounters extended learning time when scaling up to support larger numbers of users and SCs typical of real-world networks. This is primarily due to the increased complexity and enlarged discrete action space. Exploring user-SC grouping strategies (such as partitioning users and SCs into smaller groups for independent algorithm execution) and investigating continuous DRL techniques tailored to both continuous and extensive discrete action spaces could potentially mitigate these challenges for real-world network applications [47], [48]. In addition, offline training through digital twins and online operation through transfer learning to adapt an offline policy to the online reality could be another promising solution [49]. Further research is thus required to delve into these potential approaches.

VI. Conclusion and Future Work

We have investigated the H-NOMA method for the coexistence of eMBB, mMTC, and URLLC services. To analyze the system performance, we have formulated the EE-Max problem subject to divergent QoS constraints of various users. We then have proposed two MADRL-based RA solutions, i.e., JOCDDQN and FDDQN, to address the considered problem. In particular, the JOCDDQN method utilizes a CMADDQN scheme for SC assignment among users while the power corresponding to a given SC setting is optimized effectively using the proposed dynamic PA algorithm. Meanwhile, the FDDQN method implements a MADDQN scheme for both SC and PA, where the continuous power variable is split into multiple discrete power levels to facilitate the learning process. Simulation results have shown that the JOCDDQN method can achieve higher EE and converge faster than the FDDQN method and other benchmark schemes. To improve the scalability of the proposed methods towards real-world networks, offline training and online execution leveraging the benefits of digital twins, transfer learning, and continuous DRL techniques could be an interesting future research direction.

References

[1] D.-D. Tran, V. N. Ha, S. Chatzinotas, and T. T. Nguyen, "A hybrid optimization and deep RL approach for resource allocation in semi-GF NOMA networks," in *Proc. IEEE Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, 2023.

[2] S.-Y. Lien *et al.*, "5G new radio: Waveform, frame structure, multiple access, and initial access," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 64–71, 2017.

[3] G. J. Sutton *et al.*, "Enabling technologies for ultra-reliable and low latency communications: From PHY and MAC layer perspectives," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2488–2524, 2019.

[4] 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.

[5] 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.

[6] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55 765–55 779, 2018.

[7] 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 *IEEE Global Commun. Conf. (GLOBECOM)*, Madrid, Spain, Dec. 2021.

[8] 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.*, pp. 1–16, Mar. 2021.

[9] Y. Mao *et al.*, "Rate-splitting multiple access: Fundamentals, survey, and future research trends," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 4, pp. 2073–2126, 2022.

[10] 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, 2017.

[11] Z. Ding, R. Schober, P. Fan, and H. V. Poor, "Simple semi-grant-free transmission strategies assisted by non-orthogonal multiple access," *IEEE Trans. Commun.*, vol. 67, no. 6, pp. 4464–4478, 2019.

[12] Z. Yang *et al.*, "Adaptive power allocation for uplink non-orthogonal multiple access with semi-grant-free transmission," *IEEE Wireless Commun. Lett.*, vol. 9, no. 10, pp. 1725–1729, 2020.

[13] F. Saggese, M. Moretti, and P. Popovski, "NOMA power minimization of downlink spectrum slicing for eMBB and URLLC users," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2022, pp. 1725–1730.

[14] Q. Chen, J. Wu, J. Wang, and H. Jiang, "Coexistence of URLLC and eMBB services in MIMO-NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 72, no. 1, pp. 839–851, 2023.

[15] Y. Liu, B. Clerckx, and P. Popovski, "Network slicing for eMBB, URLLC, and mMTC: An uplink rate-splitting multiple access approach," *IEEE Trans. Wireless Commun.*, pp. 1–1, 2023.

[16] 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, 2020.

[17] 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.

[18] D.-D. Tran, S. K. Sharma, V. N. Ha, S. Chatzinotas, and I. Woungang, "Multi-agent DRL approach for energy-efficient resource allocation in URLLC-enabled grant-free NOMA systems," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 1470–1486, 2023.

[19] M. Alsenwi *et al.*, "Intelligent resource slicing for eMBB and URLLC coexistence in 5G and beyond: A deep reinforcement learning based approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4585–4600, 2021.

[20] H. Yang *et al.*, "Deep reinforcement learning based massive access management for ultra-reliable low-latency communications," *IEEE Trans. Wireless Commun.*, vol. 20, no. 5, pp. 2977–2990, 2021.

[21] M. Fayaz, W. Yi, Y. Liu, and A. Nallanathan, "A power-pool-based power control in semi-grant-free NOMA transmission," *arXiv preprint arXiv:2106.11190v2*, pp. 1–14, 2022.

[22] A. Alajmi, M. Fayaz, W. Ahsan, and A. Nallanathan, "Semi-centralized optimization for energy efficiency in IoT networks with NOMA," *IEEE Wireless Commun. Lett.*, vol. 12, no. 2, pp. 366–370, 2023.

[23] Z. Ding, R. Schober, and H. V. Poor, "Unveiling the importance of SIC in NOMA systems-part 1: State of the art and recent findings," *IEEE Commun. Lett.*, vol. 24, no. 11, pp. 2373–2377, 2020.

[24] 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, 2020.

[25] 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, 2021.

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

[27] C. Sun *et al.*, "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.

[28] S. Schiessl, J. Gross, and H. Al-Zubaidy, "Delay analysis for wireless fading channels with finite blocklength channel coding," in *Proc. 18th ACM Int. Conf. Model., Anal. Simul. Wireless Mobile Syst.*, Nov. 2015, pp. 13–22.

[29] W. Dinkelbach, "On Nonlinear Fractional Programming," *Management Science*, vol. 13, no. 7, pp. 492–498, March 1967. [Online]. Available: https://ideas.repec.org/a/inm/ormnsc/v13y1967i7p492-498.html

[30] V. N. Ha, D. H. N. Nguyen, and J.-F. Frigon, "System energy-efficient hybrid beamforming for mmwave multi-user systems," *IEEE Trans. Green Commun. Netw.*, vol. 4, no. 4, pp. 1010–1023, 2020.

[31] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[32] V. N. Ha and L. B. Le, "End-to-end network slicing in virtualized OFDMA-based cloud radio access networks," *IEEE Access*, vol. 5, pp. 18 675–18 691, 2017.

[33] X. Xi *et al.*, "Network resource allocation for eMBB payload and URLLC control information communication multiplexing in a multi-UAV relay network," *IEEE Trans. Commun.*, vol. 69, no. 3, pp. 1802–1817, 2021.

[34] Y. Liu, Y. Deng, M. Elkashlan, A. Nallanathan, and G. K. Karagiannidis, "Optimization of grant-free NOMA with multiple configured-grants for mURLLC," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 4, pp. 1222–1236, 2022.

[35] D. Lee, N. He, P. Kamalaruban, and V. Cevher, "Optimization for reinforcement learning: From a single agent to cooperative agents," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 123–135, 2020.

[36] 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, 2019.

[37] H. V. Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," *arXiv preprint arXiv:1509.06461*, 2015.

[38] H. Liu *et al.*, "Rate splitting multiple access for semi-grant-free transmissions," *arXiv preprint arXiv:2110.02127*, 2021.

[39] J. Chen, L. Guo, J. Jia, J. Shang, and X. Wang, "Resource allocation for IRS assisted SGF NOMA transmission: A MADRL approach," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 4, pp. 1302–1316, 2022.

[40] A. Nouruzi *et al.*, "Toward a smart resource allocation policy via artificial intelligence in 6G networks: Centralized or decentralized?" *arXiv preprint arXiv:2202.09093*, pp. 1–8, 2022.

[41] 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, 2020.

[42] F. S. Melo, "Convergence of Q-learning: A simple proof," Inst. Syst. Robot., Tech. Rep., 2001.

[43] X. Liu, Y. Liu, Y. Chen, and L. Hanzo, "Trajectory design and power control for multi-UAV assisted wireless networks: A machine learning approach," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 7957–7969, 2019.

[44] A. Sannai, Y. Takai, and M. Cordonnier, "Universal approximations of permutation invariant/equivariant functions by deep neural networks," *arXiv:1903.01939*, pp. 1–17, 2019.

[45] Z. Wang *et al.*, "Dueling network architectures for deep reinforcement learning," *arXiv preprint arXiv:1511.06581*, 2016.

[46] J. Yu *et al.*, "Channel measurement and modeling of the small-scale fading characteristics for urban inland river environment," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3376–3389, 2020.

[47] 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.

[48] G. Dulac-Arnold *et al.*, "Deep reinforcement learning in large discrete action spaces," *arXiv preprint arXiv:1512.07679*, 2015.

[49] Y. Yuan *et al.*, "Adapting to dynamic LEO-B5G systems: Meta-critic learning based efficient resource scheduling," *IEEE Trans. Wireless Commun.*, vol. 21, no. 11, pp. 9582–9595, 2022.