Received 19 September 2023; accepted 27 September 2023. Date of publication 6 October 2023; date of current version 20 October 2023.

Digital Object Identifier 10.1109/OJCOMS.2023.3322047

Energy-Efficient Rate-Splitting Multiple Access: A Deep Reinforcement Learning-Based Framework

MARIA DIAMANTI[1] (Member, IEEE), GEORGIOS KAPSALIS[1], EIRINI ELENI TSIROPOULOU[2] (Senior Member, IEEE), AND SYMEON PAPAVASSILIOU[1] (Senior Member, IEEE)

[1]Institute of Communication and Computer Systems, School of Electrical and Computer Engineering, National Technical University of Athens, 15780 Zografou, Greece
[2]Department of Electrical and Computer Engineering, University of New Mexico, Albuquerque, NM 87131, USA

CORRESPONDING AUTHOR: S. PAPAVASSILIOU (e-mail: papavass@mail.ntua.gr)

This work was supported in part by the European Commission through the Horizon Europe/JU SNS Project Hexa-X-II under Grant 101095759.

ABSTRACT Rate-Splitting Multiple Access (RSMA) has been recognized as an effective technique to reconcile the tradeoff between decoding interference and treating interference as noise in 6G and beyond networks. In this paper, in line with the need for network sustainability, we study the energy-efficient power and rate allocation of the common and private messages transmitted in the downlink of a single-cell single-antenna RSMA network. Contrary to the literature that resorts to heuristic approaches to deal with the joint problem, we transform the formulated energy efficiency maximization problem into a multi-agent Deep Reinforcement Learning (DRL) problem, based on which each transmitted private message represents a different DRL agent. Each agent explores its own state-action space, the size of which is fixed and independent of the number of agents, and shares its gained experience by exploration with a common neural network. Two DRL algorithms, namely the value-based Deep Q-Learning (DQL) and the policy-based REINFORCE, are properly configured and utilized to solve it. The adaptation of the proposed DRL framework is also demonstrated for the treatment of the considered network's sum-rate maximization objective. Numerical results obtained via modeling and simulation verify the effectiveness of the proposed DRL framework to conclude a solution to the joint problem under both optimization objectives, outperforming existing heuristic approaches and algorithms from the literature.

INDEX TERMS Energy efficiency maximization, rate-splitting multiple access (RSMA), deep reinforcement learning (DRL).

I. INTRODUCTION

6G AND beyond communication networks must deal with the ever more challenging issue of multi-user interference, given the requirements for massive connectivity to be supported over the same physical resources. In this context, Rate-Splitting Multiple Access (RSMA) has been recognized as a promising technique to transcend the immense controversy between decoding interference and treating interference as noise in such multi-user communication systems [1]. The rate-splitting lies in splitting a message into two or more parts that can be flexibly decoded at one or more receivers, respectively. The common message – as it is called – is intended for and decoded by all the involved users in the transmission, contrary to the private message intended for each user separately. As a result, when decoding the private message, the interference originating from the other users' private messages is treated as noise. By smartly controlling the split among the common and private messages, an acceptable tradeoff between efficient spectrum usage, multi-user interference management, and signal processing complexity at the receivers is achieved [2].

In light of elucidating the performance limits of the RSMA technique, systematic attempts have focused on resource optimization in RSMA-based wireless networks. Accordingly, the power control, precoder design, and rate allocation should be jointly studied in single or multi-antenna

systems, resulting in highly non-convex and combinatorial optimization problems that are difficult to solve optimally using conventional optimization techniques [3]. Moreover, the network complexity in the number of wireless connections, calls for robust optimization techniques that can scale well and dynamically adapt to the environment. Deep Reinforcement Learning (DRL) has been broadly considered in communications and networking to handle the complexity, scalability, and autonomicity issues therein [4]. Leveraging the power of deep neural networks, DRL algorithms explore a vast state-action space and conclude near-optimal solutions to non-convex problems while allowing the network's self-adaptation based on the trained model.

In this article, we target energy efficiency maximization in a single-antenna RSMA-based wireless network. To the best of the authors' knowledge, this is the first time in the literature to design and propose a DRL-based framework for energy-efficient power and rate allocation of the common and private messages transmitted in the downlink. The optimization problem is transformed into a multi-agent DRL problem, such that each agent autonomously explores its own state-action space and contributes its gained experience to a commonly trained neural network. Two different DRL algorithms are properly configured and utilized to solve it, namely the value-based Deep Q-Learning (DQL) and the policy-based REINFORCE algorithm. The algorithms are evaluated in terms of effectiveness in determining a solution to the problem by comparison against other existing heuristic approaches from the literature. Complementary to this and for better revealing the benefits and tradeoffs of the obtained solution when aiming at energy efficiency, we also analyze and assess the proposed framework under the objective of sum-rate maximization considering the same network setting, which is again a problem that has not been similarly targeted in the literature so far.

A. RELATED WORK

RSMA provides a generalization of several existing orthogonal and non-orthogonal multiple access techniques, leading to superior performance in terms of achieved throughput and spectral efficiency as has been theoretically proved for two-user Single-Input Single-Output (SISO) [5] and Multiple-Input Single-Output (MISO) [2] broadcast channels. The existence of such theoretical analyses provoked active research around RSMA lately, with an emphasis on resource allocation under various network settings. In [6] and [7], the sum-rate and weighted sum-rate maximization in the downlink of multi-user SISO and MISO systems are targeted, respectively, by jointly performing power control/precoder design and rate allocation. Other works, e.g., [8], [9], are devoted to achieving a tradeoff between energy and spectral efficiency in downlink single-cell and multi-cell MISO systems. The aforementioned tradeoff is formulated as a multi-objective optimization problem that is either approximated by the weighted sum of the two contradicting objectives [8] or decomposed into two subproblems solved iteratively [9]. Subsequently, the method of Successive Convex Approximation (SCA) is used to convexify the resulting problems and obtain a solution.

Toward accounting for sustainability and not restricting the resource allocation procedure to achieving high data rates, a different line of research pursues the maximization of the studied system's energy efficiency while potentially ensuring some minimum rate requirements, e.g., [10], [11], [12], [13], [14]. Similar to the above, the joint power control/beamforming and common-rate allocation constitute fundamental problems studied in SISO [10] and MISO [11] broadcast channels under the energy efficiency optimization objective. Both [10] and [11] conclude with suboptimal solutions contrariwise to [12] that under a similar MISO setting with [11] manages to obtain a globally optimal solution based on Successive Incumbent Transcending (SIT) Branch and Bound (BB) algorithm. Continuing with more complex network settings, the authors in [13] and [14] investigate the application of the RSMA technique in a Cloud Radio Access Network (C-RAN) and a Reconfigurable Intelligent Surface (RIS)-assisted network, accordingly. In the former, the typical power control and rate allocation problem is addressed toward energy efficiency maximization subject to the additional per-base station's transmission power and common fronthaul links' capacity constraints, whereas, in the latter, the RIS's phase-shift optimization is considered along.

The overwhelming majority of research works in the field of RSMA network optimization has relied on model-oriented and heuristic algorithms that (i) conclude suboptimal solutions, (ii) are characterized by high computational complexity as the network scales, and (iii) prohibit adaptability to the network's unpredictable changes. To tackle these challenges, the application of DRL algorithms is becoming increasingly popular. The works in [15], [16], [17], [18] provide representative examples of DRL algorithms successfully implemented to solve optimization problems in various communication environments. In [15] and [16], the power control toward sum-rate maximization is modeled as a multi-agent DRL problem, according to which the transmitter of each wireless link, i.e., agent, autonomously executes its action in selecting an appropriate transmission power level based on a commonly trained neural network, which is a paradigm referred to as "centralized training and distributed execution" in the literature. Value-based DQL, policy-based REINFORCE, and actor-critic Deep Deterministic Policy Gradient (DDPG) algorithms are then implemented and tested in this context. In [17], the DQL algorithm is used to derive the user pairing in the downlink of a Non-Orthogonal Multiple Access (NOMA) network, while the joint channel selection and power control problem is treated in [18] under both value-based and actor-critic-based DRL algorithm implementations. Both works in [17], [18] consider the sum-rate maximization objective.

Regarding the application of DRL algorithms for resource optimization in RSMA networks, only a handful of research

works can be found in the literature, i.e., [19], [20], [21], [22], [23]. In [19] and [20], two similar policy-based DRL algorithms are proposed to determine the beamforming in the downlink of an RSMA network, targeting the system's sum-rate maximization. Under the same optimization objective, the joint problem of uplink-downlink user association and beamforming is tackled in [21] for a multiple Unmanned Aerial Vehicle (UAV)-assisted RSMA network using an actor-critic DRL algorithm. In [22], an actor-critic DRL algorithm is introduced to perform computation offloading decision-making, power allocation, and decoding order optimization in the uplink of an RSMA-assisted Mobile Edge Computing (MEC) network while aiming for the minimization of the weighted sum of latency and consumed energy. Last, accounting for communications powered by energy harvesting, the authors in [23] design a DRL framework to perform harvested power allocation from a UAV to end-user devices, and then the beamforming in the RSMA network is determined using the Minimum Mean Square Error (MMSE) technique. It should be noted that none of the aforementioned works in [19], [20], [21], [22], [23] has inherited the paradigm of centralized training and distributed execution by following a multi-agent DRL modeling, while both continuous [19], [21], [22], [23] and discrete [20], [21] action spaces have been scrutinized. In the meantime, the energy efficiency maximization in RSMA-based networks via DRL algorithms has been significantly overlooked, creating a research gap.

B. CONTRIBUTIONS & OUTLINE

In this article, a DRL framework for energy-efficient power and rate allocation of the common and private messages transmitted in the downlink of a single-antenna RSMA-based network is proposed for the first time in the literature. Different from the existing works in the intersection of RSMA and DRL, multi-agent DRL modeling is adopted, according to which each private stream plays the role of a different DRL agent that contributes its personal experience from interacting with the environment toward training a common neural network. Two different DRL algorithms are then utilized to solve the formulated DRL problem, namely the value-based DQL and the policy-based REINFORCE. The key contributions of this article are summarized as follows.

1) The non-convex energy efficiency maximization problem is converted into a multi-agent DRL problem by properly designing the states, actions, and rewards to capture the problem's objective and constraints and ultimately obtain the joint power and rate solution sought while modeling each private stream as a different DRL agent.
2) The multi-agent DRL modeling, the adoption of the centralized training and distributed execution paradigm, and the appropriate discretization of the action space – for DRL algorithms' application purposes – result in a computationally scalable, though robust, DRL framework that is independent of the number of users in the network.
3) The applicability and adaptation of the proposed DRL framework are also demonstrated for the treatment of the system's sum-rate maximization, which serves as a basis for highlighting the benefits and tradeoffs of the obtained solution when targeting energy efficiency.
4) The overall DRL framework's performance is evaluated via modeling and simulation and numerical results are presented that verify its superiority under both optimization objectives when compared against existing heuristic approaches from the literature.

The remainder of this article is organized as follows. Section II presents the system model and the energy efficiency maximization problem formulation. In Section III, the multi-agent DRL modeling and distributed DRL architecture are discussed along with the description of the DQL and REINFORCE algorithms. In Section IV, the sum-rate maximization benchmark problem's formulation and solution are analyzed. Section V presents the numerical evaluation, and Section VI concludes the paper.

II. PROBLEM STATEMENT

A. SYSTEM MODEL

We consider a single-cell single-antenna wireless network consisting of a set of users $\mathcal{N} = \{1, \ldots, N\}$ served by a base station positioned at the center of the cell. The multiplexing of data transmissions for different users in the downlink is performed over the same frequency band by employing the RSMA technique. The message intended for user n is denoted as W_n, which is further divided into two parts: a common part W_n^c and a private part W_n^p. The common parts intended for the different users, i.e., $W_1^c, \ldots, W_n^c, \ldots, W_N^c$, are combined and encoded into a single common stream v_0 that is transmitted to all users with downlink transmission power p_0 [Watt]. On the other hand, the remaining private messages $W_n^p, \forall n \in \mathcal{N}$ are encoded into separate private streams v_n and transmitted individually with power p_n [Watt], $\forall n \in \mathcal{N}$. Given that the system operates on a per-time slot basis, the transmitted signal by the base station at time slot t is:

$$x^{(t)} = \sqrt{p_0^{(t)}} v_0^{(t)} + \sum_{n=1}^{N} \sqrt{p_n^{(t)}} v_n^{(t)}. \tag{1}$$

The received signal by each user n is:

$$y_n^{(t)} = \sqrt{G_n^{(t)} p_0^{(t)}} v_0^{(t)} + \sum_{j=1}^{N} \sqrt{G_n^{(t)} p_j^{(t)}} v_j^{(t)} + z_n^{(t)}, \tag{2}$$

where $G_n^{(t)}$ denotes the channel gain from the base station to user n and $z_n^{(t)} \sim \mathcal{CN}(0, \sigma^2)$ is the corresponding Additive White Gaussian Noise (AWGN). An overview of a simplified two-user RSMA-based network is presented in Fig. 1.

With reference to the channel gain modeling, in this article, block fading is adopted, such that:

$$G_n^{(t)} = |h_n^{(t)}|^2 \beta_n, \tag{3}$$

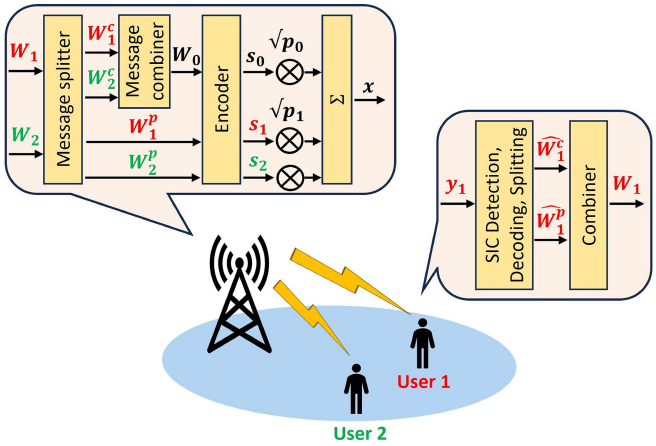

FIGURE 1. Overview of simplified two-user RSMA-based network.

where β_n is the large-scale fading that can remain the same over several time slots, whereas the term $h_n^{(t)}$ represents the small-scale Rayleigh fading. To model the time-varying nature of the channel, Jake's model [24] is used and the small-scale Rayleigh fading is expressed as a first-order Gaussian-Markov process:

$$h_n^{(t)} = \rho h_n^{(t-1)} + \sqrt{1-\rho^2}\zeta_n^{(t)}, \tag{4}$$

where $\zeta_n^{(t)} \sim \mathcal{CN}(0, 1-\rho^2)$ is an independent and identically distributed random variable. The correlation parameter ρ is $\rho = J_0(2\pi f_d T)$, where J_0 is the zero-order Bessel function, f_d is the maximum Doppler frequency, and T is the time slot over which the correlated channel variation occurs.

Following the above, the achievable rate for decoding the common stream $v_0^{(t)}$ transmitted by the base station to user n is calculated as:

$$r_n^{c\,(t)} = \log_2\left(1 + \frac{G_n^{(t)} p_0^{(t)}}{G_n^{(t)} \sum_{j=1}^{N} p_j^{(t)} + \sigma^2}\right)\left[bps/Hz\right]. \tag{5}$$

To guarantee the successful decoding of the common stream $v_0^{(t)}$ by all users $n \in \mathcal{N}$, the allocated decoding rates $c_n^{(t)}$ must adhere to the following condition:

$$\sum_{n=1}^{N} c_n^{(t)} \leq \min_{n\in\mathcal{N}} r_n^{c\,(t)}, \tag{6}$$

where $\min_{n\in\mathcal{N}} r_n^{c\,(t)} = r_1^{c\,(t)}$, given the channel gains sorted as $G_1^{(t)} \leq \cdots \leq G_n^{(t)} \leq \cdots \leq G_N^{(t)}$.

Furthermore, to ensure the successful implementation of the Successive Interference Cancellation (SIC) technique at the receiver of each user n, the following condition must be met:

$$G_n^{(t)} p_0^{(t)} - G_n^{(t)} \sum_{j=1}^{N} p_j^{(t)} \geq p_{tol}, \tag{7}$$

with p_{tol} [Watt] indicating the receivers' SIC decoding tolerance/sensitivity that is assumed to be the same for all users. Eq. (7) is rewritten as $G_1^{(t)} p_0^{(t)} - G_1^{(t)} \sum_{n=1}^{N} p_n^{(t)} \geq p_{tol}$, based on the ordering of the channel gains.

After decoding the common stream, the decoding of the corresponding private stream $v_n^{(t)}$ takes place at the receiver of each user, the achievable rate of which is:

$$r_n^{p\,(t)} = \log_2\left(1 + \frac{G_n^{(t)} p_n^{(t)}}{G_n^{(t)} \sum_{j=1, j\neq n}^{N} p_j^{(t)} + \sigma^2}\right)\left[bps/Hz\right]. \tag{8}$$

As a result, the total achievable data rate of a user n in the downlink of an RSMA-based network is:

$$\begin{aligned} R_n^{(t)} &= c_n^{(t)} + r_n^{p\,(t)} \\ &= c_n^{(t)} + \log_2\left(1 + \frac{G_n^{(t)} p_n^{(t)}}{G_n^{(t)} \sum_{j=1, j\neq n}^{N} p_j^{(t)} + \sigma^2}\right). \end{aligned} \tag{9}$$

B. PROBLEM FORMULATION

In this article, the energy efficiency maximization is targeted in the downlink of a single-antenna RSMA-based wireless network that is defined as the ratio between the sum of the total achievable data rates of all users in the system, i.e., $\sum_{n=1}^{N} R_n^{(t)}$, and the total consumed power by the base station, i.e., $p_0^{(t)} + \sum_{n=1}^{N} p_n^{(t)}$. Toward achieving this objective, the allocated by the base station common-stream rates $\mathbf{c}^{(t)} = [c_1^{(t)}, \ldots, c_n^{(t)}, \ldots, c_N^{(t)}]^T$, private-stream powers $\mathbf{p}^{(t)} = [p_1^{(t)}, \ldots, p_n^{(t)}, \ldots, p_N^{(t)}]^T$, and common-stream power $p_0^{(t)}$ to the users, are optimized. Specifically, the corresponding optimization problem to be solved by the base station is formally written as follows:

$$\max_{\mathbf{c}^{(t)}, \mathbf{p}^{(t)}, p_0^{(t)}} \quad EE = \frac{\sum_{n=1}^{N} R_n^{(t)}}{p_0^{(t)} + \sum_{n=1}^{N} p_n^{(t)}} \tag{10a}$$

$$\text{s.t.} \quad \sum_{n=1}^{N} c_n^{(t)} \leq r_1^{c\,(t)}, \tag{10b}$$

$$G_1^{(t)} p_0^{(t)} - G_1^{(t)} \sum_{n=1}^{N} p_n^{(t)} + \sigma^2 \geq p_{tol}, \tag{10c}$$

$$p_0^{(t)} + \sum_{n=1}^{N} p_n^{(t)} \leq p_{max}, \tag{10d}$$

$$c_n^{(t)}, p_n^{(t)} \geq 0, \forall n \quad \text{and} \quad p_0^{(t)} \geq 0. \tag{10e}$$

Eq. (10b) and Eq. (10c) represent the required constraints over the allocated common-stream rates and powers, respectively, for the successful decoding and implementation of the SIC technique at the receivers of the users, as described earlier in Section II-A. Eq. (10d) indicates the base station's maximum power budget p_{max} [Watt], while Eq. (10e) defines the feasible range of values of the different optimization variables.

III. PROBLEM SOLUTION

In this section, the formulated energy efficiency maximization problem is equivalently transformed into a multi-agent DRL problem to capitalize on the architectural

paradigm of centralized training and distributed execution. Subsequently, the application of the value-based DQL and policy-based REINFORCE algorithms is analyzed and discussed to solve the multi-agent DRL problem.

A. MULTI-AGENT DRL MODEL & ARCHITECTURE

A typical multi-agent DRL problem is characterized by the set of agents, the environment's state space, the agents' action spaces, and the reward function. The definition of the aforementioned constituent elements in the context of the studied optimization problem is as follows.

Agents: Each private stream $v_n^{(t)}$ from the downlink transmitted signal by the base station to the users is regarded as a distinct agent in the considered transformation. Given that there exists a one-to-one correspondence between the users and the private streams, which are henceforth termed DRL agents, we denote the set of agents as $\mathcal{N} = \{1, \ldots, N\}$ and use index n to refer to a particular agent.

State: At each time slot, the agents observe specific characteristics of the environment and create a corresponding representation known as the state. In more detail, the local state $\mathbf{s}_n^{(t)}$ observed by agent n encompasses relevant information to the transmission of its corresponding private stream $v_n^{(t)}$. Given that the power levels of the common and private streams undergo changes at the end of each time slot and remain constant during the subsequent slot [15], the agent's n state $\mathbf{s}_n^{(t)}$ at the beginning of time slot t is a tuple of the following eight components:

1) the channel gain $G_n^{(t)}$ at time slot t;
2) the channel gain $G_n^{(t-1)}$ at time slot $t-1$;
3) the interference sensed from the rest private streams at the beginning of time slot t, i.e., $G_n^{(t)} \sum_{j\in\mathcal{N}, j\neq n} p_j^{(t-1)} + \sigma^2$;
4) the interference sensed from the rest of the private streams at the beginning of time slot $t-1$, i.e., $G_n^{(t-1)} \sum_{j\in\mathcal{N}, j\neq n} p_j^{(t-2)} + \sigma^2$;
5) the power $p_n^{(t-1)}$ of the private stream;
6) the power $p_0^{(t-1)}$ of the common stream;
7) the data rate $r_n^{p\,(t)}$ of the private stream at the beginning of time slot t, calculated considering $p_n^{(t-1)}, \forall n \in \mathcal{N}$;
8) the data rate $c_n^{(t)}$ of the common stream.

Action: Each agent chooses and performs an action $a_n^{(t)} \in \mathcal{A}_n$ from its set of possible actions $\mathcal{A}_n$ following some policy $\pi(a_n^{(t)}|\mathbf{s}_n^{(t)})$ conditioned on the current state $\mathbf{s}_n^{(t)}$. Specifically, the agent's n action space is formally defined as:

$$\mathcal{A}_n = \left\{0, p_{n,min}, p_{n,min} \cdot \left(\frac{p_{n,max}}{p_{n,min}}\right)^{\frac{1}{A_n-2}}, \ldots, p_{n,max}\right\}, \quad (11)$$

where $p_{n,max} = \frac{p_{max}}{N+1}$ is the maximum allowable transmission power of the private stream $v_n^{(t)}$, with p_{max} denoting the base station's maximum power budget, and $p_{n,min}$ is a corresponding minimum allowable power level. Also, A_n indicates the cardinality of the set $\mathcal{A}_n$.

After determining the selected actions $a_n^{(t)} \in \mathcal{A}_n$ of all agents at time slot t, the optimal values of $(\mathbf{c}^{(t)}, p_0^{(t)})$ that maximize the system's energy efficiency can be obtained through analytically and exhaustively solving the following optimization problem:

$$\max_{\mathbf{c}^{(t)}, p_0^{(t)}} \quad \frac{\sum_{n=1}^{N} c_n^{(t)}}{p_0^{(t)}} \quad (12a)$$

$$\text{s.t.} \quad \sum_{n=1}^{N} c_n^{(t)} \leq r_1^{c\,(t)}, \quad (12b)$$

$$c_n^{(t)} \geq 0, \forall n \quad \text{and} \quad p_0^{(t)} \in \mathcal{P}_0, \quad (12c)$$

where by $\mathcal{P}_0$ we denote the set of feasible values of $p_0^{(t)}$: $\mathcal{P}_0 = \{\frac{p_{max}-\sum_{n\in\mathcal{N}} a_n^{(t)}}{P_0}, \frac{p_{max}-\sum_{n\in\mathcal{N}} a_n^{(t)}}{P_0-1}, \ldots, \frac{p_{max}-\sum_{n\in\mathcal{N}} a_n^{(t)}}{1}\}$ and P_0 represents its cardinality. The problem in (12) reduces to a linear programming problem for the different values of $p_0^{(t)}$ that can be, in turn, optimally solved in polynomial time. It is remarkable that the obtained solution for $(\mathbf{c}^{(t)}, p_0^{(t)})$ satisfies constraints (10b) and (10d) owing to the proper definition of problem (12). The satisfaction of the remaining constraint in Eq. (10c) is guaranteed later by the definition of the DRL problem's reward function.

Reward: As a consequence of the chosen action $a_n^{(t)}$, each agent n transitions to a new state $\mathbf{s}_n^{(t+1)}$ and receives a scalar reward feedback signal $f_n^{(t+1)}$. Aiming to maximize the energy efficiency of the system, the agent's feedback signal increases with an increase in the normalized energy efficiency $\frac{EE}{N}$, while it decreases with the level of violation of constraint (10c). Specifically, if constraint (10c) is satisfied, the reward $f_n^{(t+1)}$ is given by:

$$f_n^{(t+1)} = \frac{EE}{N}, \quad (13)$$

otherwise, it is calculated as follows:

$$f_n^{(t+1)} = \frac{EE}{N} \cdot \left(1 + \tanh\left(p_0^{(t)} - \sum_{j=1}^{N} p_j^{(t)} - \frac{p_{tol} + \sigma^2}{G_1^{(t)}}\right)\right). \quad (14)$$

The function $\tanh(x)$ approaches -1 as x tends to negative values. Hence, considering the definition of the reward in Eq. (14), it follows that the latter tends to zero as the violation of constraint (10c) grows. This behavior allows the agent to learn the negative impact of constraint violation.

Based on the proposed multi-agent DRL problem modeling described above, the centralized training and distributed execution architectural paradigm can be adopted [15], [25]. Following this paradigm, a single general-purpose model is trained centrally and shared among the distributed agents. The agents interact with their environment and utilize the learned actions (or policies depending on the employed DRL algorithm), generating experience samples that are then provided as feedback to the centralized model trainer (see Fig. 2). This approach allows leveraging the advantages of multi-agent DRL modeling in terms

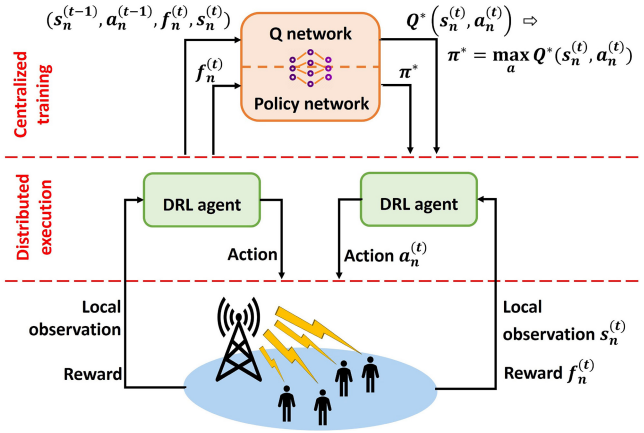

FIGURE 2. Overview of proposed multi-agent DRL architecture.

of reduced action and state spaces that require less memory, computational resources, and execution time while maintaining the stability and efficiency of a centralized solution. Each agent explores its own state-action space, which in our case consists of eight components that describe the state of the agent and A_n power levels, i.e., actions, that are independent of the number of users existing in the network, combating the curse of dimensionality issue of discrete state-action space modeling in DRL frameworks. Undoubtedly, the design of the reward feedback signal is crucial to effectively optimize the global objective by the agents' distributed decisions and actions. However, upon its successful definition, the agents can quickly learn a more general model, benefiting from one another. The centralized model training can also be performed offline using data from a simulated wireless environment and be further fine-tuned in real scenarios. In this way, the burden of online training from the inherent large volumes of data is eliminated.

B. DEEP Q-LEARNING: A VALUE-BASED ALGORITHM

DQL is a value-based algorithm that approximates the Q-function $Q^{\pi}(\mathbf{s}, a)$, i.e., the expected reward when choosing an action a in state $\mathbf{s}$ according to some policy π. The definition of the Q-function $Q^{\pi}(\mathbf{s}, a)$ is given as follows:

$$Q^{\pi}(\mathbf{s}, a) = \mathbb{E}\left[\sum_{\tau=0}^{\infty} \gamma^{\tau} f^{(t+\tau+1)} \middle| \mathbf{s}^{(t)} = s, a^{(t)} = a\right], \quad (15)$$

where γ is the discounted rate that determines the importance of future rewards, with $\gamma \in [0, 1]$. In the special case that $\gamma = 0$, only the instantaneous reward is considered.

The Q-function satisfies the recursive Bellman equation:

$$Q^{\pi}(\mathbf{s}, a) = \mathbb{E}\left[f^{(t)} + \gamma\, Q^{\pi}\left(\mathbf{s}', a'\right) \middle| \mathbf{s}^{(t)} = s, a^{(t)} = a\right], \quad (16)$$

describing the relationship of the value in state $\mathbf{s}$ with the values in all states $\mathbf{s}'$ that are likely to follow in the next time slots. By solving Eq. (16), the optimal state-action value $Q^*(\mathbf{s}, a) = \max_a Q^{\pi}(\mathbf{s}, a)$ can be determined, implying the optimal policy $\pi^* = \arg\max_a Q^*(\mathbf{s}, a)$. In the preceding definitions, the subscripts n referring to the different agents have been dropped for notation convenience.

To approximate the optimal Q-function $Q^*(\mathbf{s}, a)$, a neural network with parameter vector $\boldsymbol{\theta}_q$ is used, referred to as Deep Q-Network (DQN). Consequently, solving the DRL problem reduces to determining the optimal parameter vector $\boldsymbol{\theta}_q$, regardless of the dimensions of the state-action space. The DQN is trained from the experiences gained by the distributed agents interacting with the environment. Specifically, to combat potential instability issues of the DQL algorithm due to the high correlation of the successive states observed by a particular agent, the experience replay mechanism [26] is used. Based on this mechanism, N different First In First Out (FIFO) queues of size M are used, in which each agent n separately stores the experience acquired at time step t of training, represented by the tuple $\mathbf{e}_n^{(t)} = (\mathbf{s}_n^{(t-1)}, a_n^{(t-1)}, f_n^{(t)}, \mathbf{s}_n^{(t)})$. A minibatch $\mathcal{D}^{(t)}$ of size D of experiences is randomly created at time slot t by a common randomizer, comprising an equal number of experiences from the different agents' queues, to eliminate training the DQN over correlated agent experiences.

Given a minibatch $\mathcal{D}^{(t)}$, the least-square error of the trained DQN with parameters $\boldsymbol{\theta}_q$ is calculated as:

$$L\left(\boldsymbol{\theta}_q^{(t)}\right) = \sum_{(\mathbf{s}, a, f', \mathbf{s}') \in \mathcal{D}^{(t)}} \left(y_{DQN}^{(t)} - Q^{\pi}\left(\mathbf{s}, a; \boldsymbol{\theta}_q^{(t)}\right)\right)^2. \quad (17)$$

The target state-action value $y_{DQN}^{(t)}$ is given by:

$$y_{DQN}^{(t)} = f' + \gamma \max_{a'} Q^{\pi}\left(\mathbf{s}', a'; \mathbf{w}^{(t)}\right), \quad (18)$$

where $\mathbf{w}^{(t)}$ is the parameter vector of a second "target" DQN – as it is called – that is updated to be equal to the trained DQN, i.e., $\mathbf{w}^{(t)} = \boldsymbol{\theta}_q^{(t)}$, once every T_u time slots. The idea behind creating a second instance of the DQN that is sporadically updated serves the purpose of eliminating the correlation between the trained and the targeted state-action value. In the special case that $\gamma = 0$, the target state-action value coincides with the agent's immediate reward f' and, thus, there is no need to keep a target DQN instance.

To progressively derive a better approximation of the Q-function, the trained DQN's parameters $\boldsymbol{\theta}_q$ are updated via the gradient descend method with learning rate $\eta_q \in (0, 1]$:

$$\boldsymbol{\theta}_q^{(t+1)} = \boldsymbol{\theta}_q^{(t)} - \eta_q \nabla_{\boldsymbol{\theta}_q} L\left(\boldsymbol{\theta}_q^{(t)}\right). \quad (19)$$

Given the updated DQN's parameters and the agent's state, the optimal action that is selected at each time slot t of the designed DQL algorithm follows a dynamic ϵ-greedy policy. Let N_e denote the number of episodes, each comprising N_t time slots, then the exploration probability of randomly selecting an action different from the optimal one $a^* = \arg\max_a Q^{\pi}(\mathbf{s}, a; \boldsymbol{\theta}_q^{(t)})$, is given by:

$$\epsilon_k = e^{-\lambda k}, \; k = 1, 2, \ldots, N_e, \quad (20)$$

where $\lambda \in [0, 1]$ is the exploration probability. The proposed DQL algorithm is summarized in Algorithm 1.

Algorithm 1 Deep Q-Learning Algorithm

Initialize N_e, N_t, η_q, λ, M, D.
Randomly initialize DQN's parameters $\boldsymbol{\theta}_q$.
for $k = 1$ to N_e **do**
  Update ϵ_k based on Eq. (18).
  Derive initial agents' states $\mathbf{s}_n^{(1)}, \forall n$.
  for $t = 1$ to N_t **do**
    if rand() $\leq \epsilon_k$ **then**
      Randomly select action $a_n^{(t)} \in \mathcal{A}_n, \forall n$.
    else
      Select $a_n^{(t)} = \arg\max_{a_n} Q^{\pi}(s_n^{(t)}, a_n; \boldsymbol{\theta}_q^{(t)}), \forall n$.
    end if
    Set $\mathbf{p}^{(t)} = [a_1^{(t)}, \ldots, a_n^{(t)}, \ldots, a_N^{(t)}]$ and calculate $(\mathbf{c}^{(t)}, p_0^{(t)})$ by solving problem (12).
    Assign $(\mathbf{p}^{(t)}, \mathbf{c}^{(t)}, p_0^{(t)})$ solution to the base station and observe new states $\mathbf{s}_n^{(t+1)}$ and rewards $f_n^{(t+1)}, \forall n$.
    Obtain and store experience $\mathbf{e}_n^{(t)}, \forall n$ in the corresponding agent's n queue.
    Create a minibatch $\mathcal{D}^{(t)}$ and calculate $\nabla_{\boldsymbol{\theta}_q} L(\boldsymbol{\theta}_q^{(t)})$.
    Update DQN's parameters $\boldsymbol{\theta}_q^{(t+1)}$ based on Eq. (19).
    Set $\mathbf{s}_n^{(t)} \leftarrow \mathbf{s}_n^{(t+1)}, \forall n$.
  end for
end for

C. REINFORCE: A POLICY-BASED ALGORITHM

REINFORCE is a policy-based algorithm that directly generates the stochastic policy $\pi(a|\mathbf{s})$ using a Deep Policy Network (DPN) with $\boldsymbol{\theta}_\pi$ being the corresponding parameter vector. Therefore, the goal at each time slot t is to derive the parameter vector $\boldsymbol{\theta}_\pi^{(t)}$ that maximizes the agents' expected mean immediate reward defined as:

$$J(\boldsymbol{\theta}_\pi) = \mathbb{E}\left[\frac{\sum_{n=1}^{N} f_n^{(t)}}{N}\right]. \tag{21}$$

Then, the optimal policy $\pi^*(\mathbf{s}, a; \boldsymbol{\theta}_\pi) = \arg\max_\pi J^*(\boldsymbol{\theta}_\pi)$ is derived that is applied by each agent to determine its action $a^{(t+1)}$ at the next time slot.

To progressively conclude the parameters $\boldsymbol{\theta}_\pi$ that maximize J, the gradient ascend method is used, such that

$$\boldsymbol{\theta}_\pi^{(t+1)} = \boldsymbol{\theta}_\pi^{(t)} + \eta_\pi \nabla_{\boldsymbol{\theta}_\pi} J\left(\boldsymbol{\theta}_\pi^{(t)}\right), \tag{22}$$

where $\eta_\pi \in (0, 1]$ is the corresponding learning rate.

Due to the exploration of the algorithm in the state-action space during the training phase, there is a high probability that the values of the mean immediate rewards $J(\boldsymbol{\theta}_\pi)$ obtained between sequential time slots diverge significantly between each other. This behavior affects the algorithm's performance, resulting in its instability. To circumvent this issue, each agent's reward is normalized:

$$\hat{f}_n^{(t)} = \frac{f_n^{(t)} - \mu_f^{(t)}}{\sigma_f^{(t)}}, \tag{23}$$

Algorithm 2 REINFORCE Algorithm

Initialize N_e, N_t, η_π.
Randomly initialize DPN's parameters $\boldsymbol{\theta}_\pi$.
for $k = 1$ to N_e **do**
  Derive initial agents' states $\mathbf{s}_n^{(1)}, \forall n$.
  for $t = 1$ to N_t **do**
    Select action $a_n^{(t)} \in \mathcal{A}_n, \forall n$ based on $\pi(a_n|\mathbf{s}_n; \boldsymbol{\theta}_\pi)$.
    Set $\mathbf{p}^{(t)} = [a_1^{(t)}, \ldots, a_n^{(t)}, \ldots, a_N^{(t)}]$ and calculate $(\mathbf{c}^{(t)}, p_0^{(t)})$ by solving problem (12).
    Assign $(\mathbf{p}^{(t)}, \mathbf{c}^{(t)}, p_0^{(t)})$ solution to the base station and observe new states $\mathbf{s}_n^{(t+1)}$ and rewards $f_n^{(t+1)}, \forall n$.
    Calculate $\mu_f^{(t)}$, $\sigma_f^{(t)}$, and $\hat{f}_n^{(t)}, \forall n$ based on Eq. (23).
    Calculate $\nabla_{\boldsymbol{\theta}_\pi} J(\boldsymbol{\theta}_\pi^{(t)})$ using $\hat{f}_n^{(t)}, \forall n$.
    Update DPN's parameters $\boldsymbol{\theta}_\pi^{(t+1)}$ based on Eq. (22).
    Set $\mathbf{s}_n^{(t)} \leftarrow \mathbf{s}_n^{(t+1)}, \forall n$.
  end for
end for

where $\mu_f^{(t)} = \frac{\sum_{i=1}^{N} f_i^{(t)}}{N}$ and $\sigma_f^{(t)} = \sqrt{\frac{\sum_{i=1}^{N} (f_i^{(t)} - \mu_f^{(t)})^2}{N}}$ represent the mean value and the dispersion of the agents' rewards at time slot t. The proposed REINFORCE algorithm is outlined in Algorithm 2.

IV. SUM-RATE MAXIMIZATION BENCHMARK

In this section, we extend our proposed DRL framework analyzed in detail in Section III to account for an alternative objective, namely the sum-rate maximization in the considered downlink RSMA-based communication network. On the one hand, we aim to corroborate the applicability, effectiveness, and efficiency of the devised DRL framework under different optimization objectives, given that the problem of sum-rate maximization has not been treated similarly by the literature so far. On the other hand, we seek to macroscopically identify and promote the significance of targeting energy efficiency, resulting in a better trade-off between resource utilization, system performance, and algorithmic complexity.

The formal representation of the corresponding sum-rate maximization problem toward optimizing the vectors of allocated common-stream rates $\mathbf{c}^{(t)} = [c_1^{(t)}, \ldots, c_n^{(t)}, \ldots, c_n^{(t)}]^T$, and the private and common-stream transmission powers $\mathbf{p}^{(t)} = [p_1^{(t)}, \ldots, p_n^{(t)}, \ldots, p_N^{(t)}]^T$ and $p_0^{(t)}$ is as follows:

$$\max_{\mathbf{c}^{(t)}, \mathbf{p}^{(t)}, p_0^{(t)}} \quad \sum_{n=1}^{N} R_n^{(t)} \tag{24a}$$

$$\text{s.t.} \quad \sum_{n=1}^{N} c_n^{(t)} \leq r_1^{c\,(t)}, \tag{24b}$$

$$G_1^{(t)} p_0^{(t)} - G_1^{(t)} \sum_{n=1}^{N} p_n^{(t)} + \sigma^2 \geq p_{tol}, \tag{24c}$$

$$p_0^{(t)} + \sum_{n=1}^{N} p_n^{(t)} \leq p_{max}, \tag{24d}$$

$$c_n^{(t)}, p_n^{(t)} \geq 0, \forall n \quad \text{and} \quad p_0^{(t)} \geq 0. \tag{24e}$$

The definition of problem (24) is in accordance with its energy efficiency counterpart and a similar approach with Section III-A can be followed for its transformation into a multi-agent DRL scenario. Each private stream $v_n^{(t)}$ of the downlink transmitted signal constitutes a different agent whose description of the local state $\mathbf{s}_n^{(t)}$ comprises the eight components analyzed in Section III-A. Each agent autonomously chooses an action $a_n^{(t)} \in \mathcal{A}_n$ from the set of possible actions $\mathcal{A}_n$ in Eq. (11) after evaluating its state. Based on the agents' chosen actions, the values of $(\mathbf{c}^{(t)}, p_0^{(t)})$ that maximize the sum rate can be obtained by setting $p_0^{(t)} = p_{max} - \sum_{n=1}^{N} p_n^{(t)}$ and solving the following linear programming problem:

$$\max_{c_n^{(t)} \geq 0, \forall n} \quad \sum_{n=1}^{N} c_n^{(t)} \tag{25a}$$

$$\text{s.t.} \quad \sum_{n=1}^{N} c_n^{(t)} \leq r_1^{c\,(t)}. \tag{25b}$$

It should be noted that the common stream does not interfere with the private streams and, thus, the allocation of all available power, i.e., $p_{max} - \sum_{n=1}^{N} p_n^{(t)}$, to the common stream maximizes the sum rate [6]. This observation can be easily derived by closely examining Eq. (5) and (6).

Last, to target the system's sum-rate maximization, the reward feedback signals provided to the agents should be redefined accordingly. Following a similar rationale with the one in Section III-A, if constraint (10c) is satisfied, the reward $f_n^{(t+1)}$ provided to agent n at time slot $t+1$ about the action $a_n^{(t)}$ chosen at the previous time slot t is captured by its normalized achieved data rate, i.e.,

$$f_n^{(t+1)} = \frac{R_n^t}{N}, \tag{26}$$

whereas, in case of the constraint violation, the reward is:

$$f_n^{(t+1)} = \frac{R_n^t}{N} \cdot \left(1 + \tanh\left(p_0^{(t)} - \sum_{j=1}^{N} p_j^{(t)} - \frac{p_{tol} + \sigma^2}{G_1^{(t)}}\right)\right). \tag{27}$$

The physical meaning and interpretation of the designed reward are identical with Eq. (13) and (14) described earlier.

Subsequently, the proposed DRL framework based on the value-based DQL algorithm or policy-based REINFORCE alternative can be directly applied to render a solution to the sum-rate maximization problem.

V. EVALUATION & RESULTS

In this section, the performance of the proposed DRL framework for energy-efficient power and rate allocation in the downlink of single-cell single-antenna RSMA networks is evaluated via modeling and simulation. Throughout our experiments, we consider $N = 4$ users randomly spatially distributed with minimum and maximum distance from the base station set as 10 m and 500 m, respectively. The channel gain between the users and the base station is calculated considering the log-distance path loss model $PL = 120.9 + 37.6\log(d)$ with d measured in km and log-normal shadowing standard deviation equal to 8 dB [6]. The maximum Doppler frequency is $f_d = 10$ Hz and the time slot duration is $T = 20$ ms [15]. The rest of the communication-related parameters are summarized in Table 1.

TABLE 1. Simulation parameters.

Parameters	Values
Noise power σ^2	-114 dBm
Receiver decoding sensitivity p_{tol}	-94 dBm
Total maximum power p_{max}	25 dBm
Minimum power $p_{n,min}$	1 dBm
Episodes N_e	6000 for DQL 200 for REINFORCE
Time slots per episode N_t	50
Discounted rate γ	0
Learning rate η_q	10^{-2}
Learning rate η_π	10^{-3}
FIFO queues size M	5000
Minibatch size D	500
Exploration probability λ	$4 \cdot 10^{-3}$

Considering the definition of the action space in the multi-agent DRL problem, a number of $A_n = 10, \forall n$ and $P_0 = 100$ discrete power levels for the private and common streams is considered unless otherwise explicitly stated. The structure of the neural networks used as part of the DQL and REINFORCE algorithms is similar and is as follows. A feedforward neural network with 3 hidden layers is chosen, having 200, 100, and 40 neurons, respectively. The input layer has 8 neurons, i.e., one neuron for each state feature, while the output layer has A_n neurons equal to the number of power levels of the private streams. The Rectified Linear Unit (ReLU) is chosen as an activation function, while the specific values used for the DQN and REINFORCE algorithms' hyper-parameters are listed in Table 1. A comprehensive numerical analysis is included in the following, justifying the selection of the latter values.

To characterize the effectiveness of the proposed DRL algorithms in concluding a solution under both optimization objectives, two heuristic approaches from the literature are also considered and simulated. First, a heuristic algorithm to solve the energy-efficient power and rate allocation is used as a benchmark, where the decoupling of the joint problem into distinct subproblems is performed. The respective algorithm is presented in [10] and is referred to as "Heuristic" henceforth. Furthermore, regarding the sum-rate maximization objective, a modified version of the Weighted Minimum-Mean Square Error (WMMSE) [27] algorithm is used to solve the power allocation problem and, then,

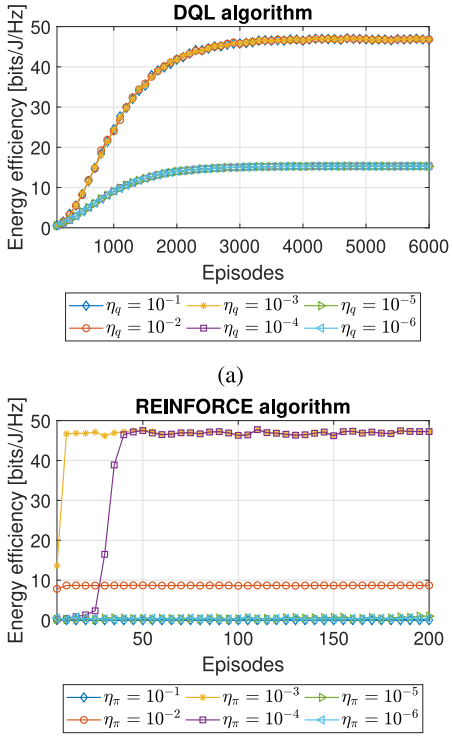

FIGURE 3. Average energy efficiency per user under the (a) DQL and (b) REINFORCE algorithms for different values of the learning rate when targeting energy efficiency maximization.

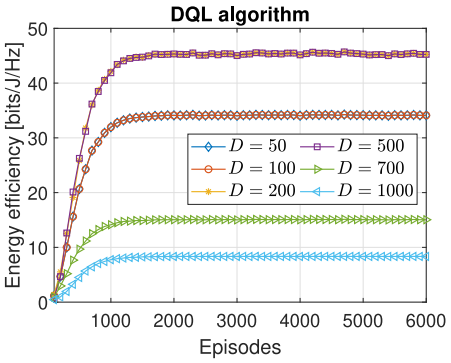

FIGURE 4. Average energy efficiency per user under the DQL algorithm for different values of the minibatch size when targeting energy efficiency maximization.

determine the rate splitting for the RSMA network. The latter benchmarking heuristic is denoted as "WMMSE".

In the sequel, the plotted values of the energy efficiency and sum rate metrics have been normalized with the number of users in the system to capture the average achieved energy efficiency and rate per user. This representation serves the purpose of accurately reflecting the performance of the system under the specific number of served users. To ensure reasonable system performance, we consider as successful and valid those network and algorithm settings that allow each user to achieve at least 1 Mbps/Hz downlink data rate [6].

A. DRL ALGORITHMS' HYPER-PARAMETER ANALYSIS

First, we perform a numerical analysis over different values of the DRL algorithms' hyper-parameters, reflecting their impact on the algorithms' behavior over the training episodes. The obtained results are indicatively presented for the energy efficiency optimization objective, while similar observations can be rendered considering the sum-rate maximization of the system. In Fig. 3(a) and 3(b), the achieved energy efficiency is illustrated as a function of the training episodes for different values of the DQL and REINFORCE algorithms' learning rates η_q and η_π, respectively. In more detail, the learning rate controls the adjustment level of the parameter vector, i.e., the neural network's weights, in response to the estimated error at each time slot. As a consequence, small values of the learning rate, i.e., $\eta_q = \eta_\pi = 10^{-1}$, result in suboptimal solutions, whereas larger values, i.e., $\eta_q = \eta_\pi = 10^{-5}$, 10^{-6}, may prevent optimization and cause the algorithms' training to get stuck. There is a turning point where optimal performance in the achieved energy efficiency can be achieved for both DQL and REINFORCE algorithms. The DQL algorithm performs best for $\eta_q = 10^{-2}$, 10^{-3}, 10^{-4}, whereas $\eta_\pi = 10^{-3}$, 10^{-4} are the values of the learning rate parameter yielding best performance for REINFORCE algorithm. Under these particular values that training is performed successfully, both algorithms present stable performance and reach almost identical energy efficiency levels. However, the REINFORCE algorithm requires fewer episodes to conclude, exhibiting stable performance from the very beginning. Concluding, based on the results of Fig. 3(a) and 3(b), the learning rate parameters are set equal to $\eta_q = 10^{-2}$ and $\eta_\pi = 10^{-3}$ for the rest of the simulation experiments.

Especially with reference to the DQL algorithm, the hyper-parameter related to the size of the minibatch of experiences used as input to the DQN should be additionally configured. For this purpose, different values of the minibatch size D are scrutinized, and the performance of the DQL algorithm in the achieved energy efficiency is observed over the training episodes. The results are presented in Fig. 4, where a similar tradeoff between small and large values for the minibatch size hyper-parameter is depicted. A minibatch with inadequate experience samples, i.e., $D = 50, 100$, may cause the trained model to converge to a local maximum, whereas a large size of the minibatch, i.e., $D = 700, 1000$, may have the opposite effect and result in the DQN's overtraining during the very first episodes. This prohibits the DQN from learning actions by experiences gained at later episodes, yielding solutions of lower achieved energy efficiency compared to the optimal hyper-parameter setting. The latter optimal setting is found for a minibatch size of $D = 500$ experience samples officially selected for our experiments.

Apart from properly configuring the DRL algorithms, the design of the state-action space is crucial for the solution outcome. In this context, controlling the size of the agents' action space is also performed numerically to strike a

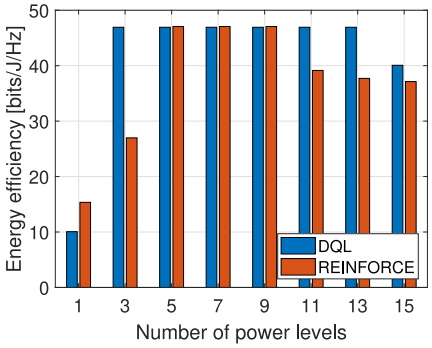

FIGURE 5. Average energy efficiency per user under the DQL and REINFORCE algorithms for different numbers of power levels when targeting energy efficiency maximization.

balance between achieved energy efficiency and algorithmic complexity. Fig. 5 illustrates the achieved energy efficiency under different numbers of power levels A_n for the private streams for both DQL and REINFORCE algorithms. The results reveal that there exists an "optimal" number of power levels where the tradeoff between exploring different actions and complexity in the exploration is optimal for both algorithms, which in our case is $A_n = 10, \forall n$ as used in the experiments overall.

Concluding, the trained DRL models follow the configuration that resulted from the hyper-parameter analysis so far. The results presented for both DQL and REINFORCE algorithms from this point and on correspond to the average energy efficiency (and rate accordingly) given as output from the trained deep model over $N_e = 100$ randomly simulated episodes, comprising $N_t = 500$ time slots each.

B. SCALABILITY ANALYSIS

Subsequently, we conduct a scalability analysis considering an increasing number of users in the cell, aiming to evaluate the performance of the proposed DRL framework as the network size increases while comparing at the same time against the "Heuristic" and "WMMSE" approaches. The range considered regarding the number of users is $N = [2, 7]$ in alignment with good practices followed in the existing literature of RSMA, e.g., [6], [8]. It should be noted that the common and private streams transmitted by the base station to the users are multiplexed over the same frequency resources, resulting in interference between them, as expressed in Eq. (5) and Eq. (8), respectively. Therefore, for the interference not to become unbearable, an upper bound in the number of users sharing the same frequency band is considered in the literature, equal to $N = 7$. In case more users should be considered in the simulation topology, then the same problem with the proposed one will be solved independently for different clusters of users that operate over a different frequency band.

Fig. 6(a) demonstrates the achieved energy efficiency per user for different numbers of users in the horizontal axis when targeting the energy efficiency maximization of the system. As expected, the results present a decaying trend

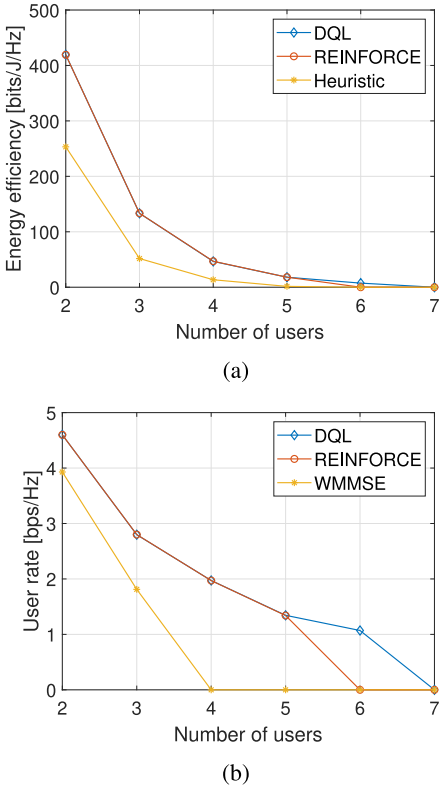

FIGURE 6. Average (a) energy efficiency and (b) rate per user under the DQL, REINFORCE, "Heuristic", and "WMMSE" approaches for different numbers of users when targeting (a) energy efficiency and (b) sum-rate maximization.

under all approaches and algorithms as the number of users gets higher due to the increased interference and the total transmission power required in the downlink by the base station. A significant gap is shown between the DRL-based algorithms and the "Heuristic" approach for a small number of users transmitting over the same frequency band, i.e., $N = 2, 3$. For larger values of N, when the system is congested and constrained, the DRL algorithms and the "Heuristic" perform closely. Especially for $N = 6, 7$, the majority (if not all) of the comparative scenarios are unable to conclude a solution that provides at least 1 Mbps per user. For this reason, their achieved energy efficiency value is set equal to 0. The latter justifies that the number of users sharing the same frequency resource cannot be arbitrarily increased. Fig. 6(b) depicts the achieved average rate per user for different numbers of users when seeking the sum-rate maximization. In this simulation case, it is remarkable that the "WMMSE" approach fails to conclude a solution that secures a data rate higher than 1 Mbps for each user for $N \geq 4$ contrariwise to the proposed DRL algorithms that can provide an effective resource allocation solution for at least five users under the same frequency band. In this way, the power of DRL to explore a vast state-action space is further demonstrated.

The outcome of the scalability analysis so far is that DRL is more successful in deriving an energy-efficient power and rate allocation in RSMA networks than a heuristic approach

TABLE 2. Resulting testing time under the DQL, REINFORCE, "Heuristic", and "WMMSE" approaches for different numbers of users.

N	Testing time [sec]					
	Energy efficiency maximization			Sum-rate maximization		
	DQL	REIN.	Heuristic	DQL	REIN.	WMMSE
2	49.09	54.60	93.60	46.90	54.42	10.09
3	46.86	55.22	71.40	49.17	56.24	10.64
4	42.80	63.72	94.80	39.04	45.89	-
5	47.20	55.03	124.80	45.91	53.62	-
6	47.47	-	-	47.73	-	-
7	-	-	-	-	-	-

under both optimization objectives. In the following, we also measure the resulting testing time, i.e., the execution time of the resource allocation procedure based on the pre-trained deep neural network over the testing dataset that includes simulated channel gain distributions of the users that are different from the ones used during pre-training. The obtained numerical results are listed in Table 2. The results reveal that the two DRL algorithms behave similarly in the resulting testing time. However, both of them outperform the "Heuristic" approach, whose mean execution time is 96.15 sec under the energy efficiency optimization objective. On the other hand, although the "WMMSE" approach proves to be significantly faster than the DRL algorithms during their testing, its ability to conclude a solution is limited and restricted to a very small number of users. Note that the cells missing numerical values refer to the specific simulation cases with $N = 6, 7$, where a minimum acceptable rate of 1 Mbps for each user cannot be secured by some of the different comparative algorithms and approaches.

Our scalability analysis is complemented by a comparison against the well-known Q-Learning algorithm [4], which allows for further justifying the need for solutions based on deep neural networks to tackle optimization problems of the scale and complexity of the examined one. Based on the Q-Learning algorithm, the optimal Q-function is derived after exhaustive exploration and calculation of its value for the different state-action pairs, contrary to the proposed DQL algorithm that employs a deep neural network to perform function approximation. The calculated value of the Q-function for each state-action pair is stored in a lookup table, i.e., the Q-table. For the implementation of the Q-Learning algorithm, the modeling of the reward function and the discrete action space in Section III-A are kept unchanged, while the only differentiation lies in the design of the state space that is discretized to facilitate the construction of the Q-table. Directly discretizing the state space of our proposed DRL framework that comprises eight distinct components (see Section III-A) leads to the creation of a huge Q-table. For this reason, inspired by the majority of Q-learning applications in wireless networks from the literature, we consider that an agent's n state is completely captured by its channel gain $G_n^{(t)}$ at a particular time slot t, i.e., $s_n^{(t)} = G_n^{(t)}$ [4]. The agent's state, i.e., channel gain, is further quantized into 10 value ranges, each of which creates a separate row in the Q-table while the discrete actions form different columns. Table 3 includes the obtained numerical results regarding the achieved energy efficiency and resulting training time. The training time of the DQL, REINFORCE, and Q-Learning algorithms has been measured considering 4000, 200, and 100 episodes, respectively, where convergence is reached. Also, a small number of users N has been considered owing to the inherent difficulty of constructing a Q-table of all combinations of state-action pairs for all users in the system. Despite the small scale of the simulated system, the Q-Learning algorithm still concludes a resource allocation solution of notably low energy efficiency, i.e., approximately 49 times lower when $N = 2$ and 26 times lower when $N = 3$ compared to the DRL algorithms.

TABLE 3. Average energy efficiency per user and resulting training time under the DQL, REINFORCE, and Q-Learning approaches for different numbers of users.

N	Energy efficiency [bits/J/Hz]			Training time [sec]		
	DQL	REIN.	Q-Learn.	DQL	REIN.	Q-Learn.
2	419.40	419.25	8.64	196.36	109.20	23.14
3	133.30	133.44	5.18	187.44	110.44	23.92

C. NETWORK OPTIMIZATION OBJECTIVES ANALYSIS

To gain more insight into the impact of energy efficiency optimization on the overall network's performance, we proceed to a comparative examination between the performance of the proposed DRL framework under (a) energy efficiency and (b) sum-rate maximization objectives. In particular, Fig. 7(a) and 7(b) demonstrate the achieved values under both metrics when (a) energy efficiency and (b) sum-rate maximization is targeted, respectively. To render this comparison even more plausible, we also account for different values of the base station's maximum power budget within the range $p_{max} = [20, 40]$ dBm, characterizing its total maximum emitted transmission power in the downlink at each simulation scenario. Note that for each 5 dBm-increment of p_{max}, we increase the number of power levels A_n concluded from Fig. 5 by five to fairly maintain the sensitivity of exploration within the action space $\mathcal{A}_n, \forall n$. The number of users considered in this simulation case is $N = 4$.

Under the energy efficiency objective, both DRL algorithms "stick" to the pursued minimum data rate requirement for each user (see right part of Fig. 7(a)) and target to maximize the achieved energy efficiency without necessarily spending the total amount of power p_{max} available. Apparently, there exists a turning point regarding the available maximum power budget and the resulting number of power levels, where the DRL algorithms find the best solution to the problem. Specifically, both algorithms manage to achieve a maximum energy efficiency level approximately equal to 46.5 bits/J/Hz, as shown in the left part of Fig. 7(a) and coincide in that this is found for $p_{max} =$

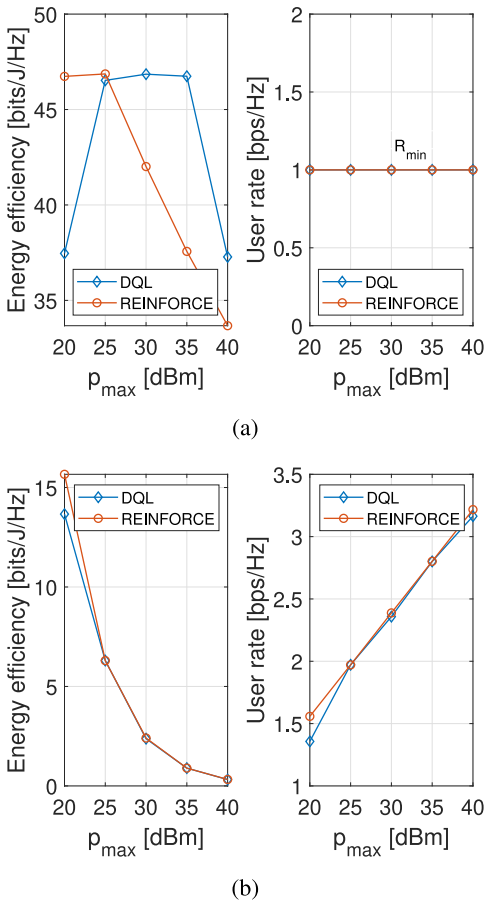

FIGURE 7. Average energy efficiency and rate per user under the DQL and REINFORCE algorithms for different values of the base station's maximum power budget p_{max} when targeting (a) energy efficiency and (b) sum-rate maximization.

25 dBm. Regarding the sum-rate maximization, the two designed DRL algorithms exhibit identical performance (see Fig. 7(b)). Higher values of the parameter p_{max} allow for achieving higher user data rates (right part of Fig. 7(b)) while decreasing the corresponding energy efficiency of the system (left part of Fig. 7(b)). To be more specific, when p_{max} gets doubled from 20 dBm to 40 dBm, a small increment of two times is observed in the user data rate due to higher interference sensed by the users, which in conjunction with the higher sum of transmission powers in the denominator of the energy efficiency function, rapidly decreases the energy efficiency by almost 15 times. Furthermore, closely inspecting the right parts of Fig. 7(a) and 7(b), it can be easily seen that for an average rate equal to 1.5 bps/Hz per user, the concluded energy efficiency under the sum-rate maximization objective is 15 bits/J/Hz, whereas a value close to 46.5 bits/J/Hz could be achieved if pursuing the energy efficiency maximization, following the results of Fig. 7(a). Interestingly, this comes with the cost of 31 times lower achieved energy efficiency when myopically targeting the system's sum-rate maximization, highlighting the need to focus on energy-efficient resource allocation approaches.

VI. CONCLUSION AND FUTURE WORK

In this paper, the problem of energy efficiency maximization was investigated in a single-cell single-antenna RSMA network. Specifically, the joint power and rate allocation of the common and private messages transmitted in the downlink of the RSMA network was designed to maximize the system's energy efficiency. To manage such a combinatorial problem, a multi-agent DRL modeling was proposed, according to which the DRL agents were mapped to the private streams that explore the wireless network via their actions, i.e., private stream power allocations. The DRL agents contribute their experiences gained to training a common neural network, at which point, two different DRL algorithms were properly configured and utilized. The first DRL algorithm regarded the value-based DQL, while the second corresponded to the policy-based REINFORCE. The output of the respective DRL algorithm, which is the optimal private-stream power allocations of the DRL agents, was then used as input to a linear programming problem that directly derived the common-stream power and rate allocations for the considered network setting. The same multi-agent DRL modeling, architecture, and algorithms were also evaluated under a different network optimization objective, namely the sum-rate maximization of the considered RSMA network. The proposed DRL framework showed to perfectly adapt to both optimization settings and conclude solutions that are closer to optimal when compared against existing approaches and algorithms from the literature.

Our current and future work focuses on the design and testing of actor-critic-based algorithms over the same network setup. Furthermore, the extension of the networking setting to account for multiple antenna transmissions will be targeted by adapting both the multi-agent DRL modeling and architecture, as well as the employed DRL algorithms.

REFERENCES

[1] H. Joudeh and B. Clerckx, "Robust transmission in downlink multiuser MISO systems: A rate-splitting approach," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6227–6242, Dec. 2016.

[2] B. Clerckx, Y. Mao, R. Schober, and H. V. Poor, "Rate-splitting unifying SDMA, OMA, NOMA, and multicasting in MISO broadcast channel: A simple two-user rate analysis," *IEEE Wireless Commun. Lett.*, vol. 9, no. 3, pp. 349–353, Mar. 2020.

[3] Y. Mao, O. Dizdar, B. Clerckx, R. Schober, P. Popovski, and H. V. Poor, "Rate-splitting multiple access: Fundamentals, survey, and future research trends," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 4, pp. 2073–2126, 4th Quart., 2022.

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

[5] T. Han and K. Kobayashi, "A new achievable rate region for the interference channel," *IEEE Trans. Inf. Theory*, vol. 27, no. 1, pp. 49–60, Jan. 1981.

[6] Z. Yang, M. Chen, W. Saad, and M. Shikh-Bahaei, "Optimization of rate allocation and power control for rate splitting multiple access (RSMA)," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 5988–6002, Sep. 2021.

[7] H. Xia, Y. Mao, B. Clerckx, X. Zhou, S. Han, and C. Li, "Weighted sum-rate maximization for rate-splitting multiple access based secure communication," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2022, pp. 19–24.

[8] G. Zhou, Y. Mao, and B. Clerckx, "Rate-splitting multiple access for multi-antenna downlink communication systems: Spectral and energy efficiency tradeoff," *IEEE Trans. Wireless Commun.*, vol. 21, no. 7, pp. 4816–4828, Jul. 2022.

[9] J. Zhang, J. Zhang, Y. Zhou, H. Ji, J. Sun, and N. Al-Dhahir, "Energy and spectral efficiency tradeoff via rate splitting and common beamforming coordination in multicell networks," *IEEE Trans. Commun.*, vol. 68, no. 12, pp. 7719–7731, Dec. 2020.

[10] W. De Souza Junior, V. Croisfelt, and T. Abrão, "On the energy efficiency of one-layer SISO rate-splitting multiple access," in *Proc. IEEE URUCON*, 2021, pp. 42–46.

[11] Y. Mao, B. Clerckx, and V. O. Li, "Energy efficiency of rate-splitting multiple access, and performance benefits over SDMA and NOMA," in *Proc. 15th Int. Symp. Wireless Commun. Syst. (ISWCS)*, 2018, pp. 1–5.

[12] B. Matthiesen, Y. Mao, A. Dekorsy, P. Popovski, and B. Clerckx, "Globally optimal spectrum- and energy-efficient beamforming for rate splitting multiple access," *IEEE Trans. Signal Process.*, vol. 70, pp. 5025–5040, Oct. 2022, doi: 10.1109/TSP.2022.3214376.

[13] A. A. Ahmad, B. Matthiesen, A. Sezgin, and E. Jorswieck, "Energy efficiency in C-RAN using rate splitting and common message decoding," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, 2020, pp. 1–6.

[14] Z. Yang, J. Shi, Z. Li, M. Chen, W. Xu, and M. Shikh-Bahaei, "Energy efficient rate splitting multiple access (RSMA) with reconfigurable intelligent surface," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, 2020, pp. 1–6.

[15] Y. S. Nasir and D. Guo, "Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2239–2250, Oct. 2019.

[16] F. Meng, P. Chen, L. Wu, and J. Cheng, "Power allocation in multi-user cellular networks: Deep reinforcement learning approaches," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6255–6267, Oct. 2020.

[17] F. Jiang, Z. Gu, C. Sun, and R. Ma, "Dynamic user pairing and power allocation for NOMA with deep reinforcement learning," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2021, pp. 1–6.

[18] Z. Lu, C. Zhong, and M. C. Gursoy, "Dynamic channel access and power control in wireless interference networks via multi-agent deep reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 71, no. 2, pp. 1588–1601, Feb. 2022.

[19] J. Huang, Y. Yang, L. Yin, D. He, and Q. Yan, "Deep reinforcement learning-based power allocation for rate-splitting multiple access in 6G LEO satellite communication system," *IEEE Wireless Commun. Lett.*, vol. 11, no. 10, pp. 2185–2189, Oct. 2022.

[20] N. Q. Hieu, D. T. Hoang, D. Niyato, and D. I. Kim, "Optimal power allocation for rate splitting communications with deep reinforcement learning," *IEEE Wireless Commun. Lett.*, vol. 10, no. 12, pp. 2820–2823, Dec. 2021.

[21] J. Ji, L. Cai, K. Zhu, and D. Niyato, "Decoupled association with rate splitting multiple access in UAV-assisted cellular networks using multi-agent deep reinforcement learning," *IEEE Trans. Mobile Comput.*, early access, Mar. 15, 2023, doi: 10.1109/TMC.2023.3256404.

[22] T. P. Truong, N.-N. Dao, and S. Cho, "HAMEC-RSMA: Enhanced aerial computing systems with rate splitting multiple access," *IEEE Access*, vol. 10, pp. 52398–52409, 2022.

[23] J. Seong, M. Toka, and W. Shin, "Sum-rate Maximization of RSMA-based aerial communications with energy harvesting: A reinforcement learning approach," *IEEE Wireless Commun. Lett.*, vol. 12, no. 10, pp. 1741–1745, Oct. 2023.

[24] P. Dent, G. E. Bottomley, and T. Croft, "Jakes fading model revisited," *Electron. Lett.*, vol. 13, no. 29, pp. 1162–1163, 1993.

[25] F. D. Calabrese, L. Wang, E. Ghadimi, G. Peters, L. Hanzo, and P. Soldati, "Learning radio resource management in RANs: Framework, opportunities, and challenges," *IEEE Commun. Mag.*, vol. 56, no. 9, pp. 138–145, Sep. 2018.

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

[27] S. S. Christensen, R. Agarwal, E. De Carvalho, and J. M. Cioffi, "Weighted sum-rate maximization using weighted MMSE for MIMO-BC beamforming design," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 4792–4799, Dec. 2008.

MARIA DIAMANTI (Member, IEEE) received the Diploma degree in electrical and computer engineering from the Aristotle University of Thessaloniki in 2018. She is currently pursuing the Ph.D. degree with the School of Electrical and Computer Engineering, National Technical University of Athens, where she is also a Research Assistant. Her research interests lie in the areas of 5G/6G wireless networks, resource management and optimization, game theory, contract theory, and reinforcement learning.

GEORGIOS KAPSALIS received the Diploma degree in electrical and computer engineering from the National Technical University of Athens in 2022. His Diploma thesis focused on the topic of resource allocation in rate-splitting multiple access networks with the use of optimization and reinforcement learning techniques. His overall research interests lie in the broader area of resource optimization in 5G/6G wireless communications systems.

EIRINI ELENI TSIROPOULOU (Senior Member, IEEE) is currently an Associate Professor with the Department of Electrical and Computer Engineering, University of New Mexico. Her main research interests lie in the area of cyber–physical social systems and wireless heterogeneous networks, with emphasis on network modeling and optimization, resource orchestration in interdependent systems, reinforcement learning, game theory, network economics, and Internet of Things. Four of her papers received the Best Paper Award at IEEE WCNC in 2012, ADHOCNETS in 2015, IEEE/IFIP WMNC 2019, and INFOCOM 2019 by the IEEE ComSoc Technical Committee on Communications Systems Integration and Modeling. She was selected by the IEEE Communication Society—N2Women—as one of the top ten Rising Stars of 2017 in the communications and networking field. She received the NSF CRII Award in 2019 and the Early Career Award by the IEEE Communications Society Internet Technical Committee in 2019.

SYMEON PAPAVASSILIOU (Senior Member, IEEE) is currently a Professor with the School of ECE, National Technical University of Athens. From 1995 to 1999, he was a Senior Technical Staff Member with AT&T Laboratories, Middletown, NJ, USA. In August 1999, he joined the ECE Department, New Jersey Institute of Technology, USA, where he was an Associate Professor until 2004. He has an established record of publications in his field of expertise, with more than 400 technical journal and conference published papers. His main research interests lie in the area of computer communication networks, with emphasis on the analysis, optimization, and performance evaluation of mobile and distributed systems, wireless networks, and complex systems. He received the Best Paper Award in IEEE INFOCOM 94, the AT&T Division Recognition and Achievement Award in 1997, the U.S. National Science Foundation Career Award in 2003, the Best Paper Award in IEEE WCNC 2012, the Excellence in Research Grant in Greece in 2012, the Best Paper Awards in ADHOCNETS 2015, ICT 2016 and IEEE/IFIP WMNC 2019, IEEE Globecom 2022, as well as the 2019 IEEE ComSoc Technical Committee on Communications Systems Integration and Modeling Best Paper Award (for his INFOCOM 2019 paper).