Power and Rate Allocation for Energy-Efficient Rate-Splitting Multiple Access via Deep Q-Learning

Maria Diamanti*, Georgios Kapsalis*, Eirini Eleni Tsiropoulou†, and Symeon Papavassiliou*
{mdiamanti@netmode.ntua.gr, georkapsalis@gmail.com, eirini@unm.edu, papavass@mail.ntua.gr}
* Institute of Communication and Computer Systems, School of Electrical and Computer Engineering,
National Technical University of Athens, Athens, Greece
† Dept. of Electrical and Computer Engineering, University of New Mexico, Albuquerque, NM, USA

***Abstract*—Rate-Splitting Multiple Access (RSMA) has recently emerged as an effective technique for increasing network capacity by smartly controlling the tradeoff between decoding and treating interference as noise. In this paper, aligned with the need for sustainable wireless networks, we study the energy-efficient power and rate allocation of the common and private messages in the downlink of a network where rate-splitting is adopted. The corresponding energy efficiency maximization problem is transformed into a multi-agent Deep Reinforcement Learning (DRL) problem, based on which each private stream transmitted in the downlink constitutes a different DRL agent. The formulated DRL problem is solved by using Deep Q-Learning (DQL) algorithm and training a single Deep Q-Network (DQN) from the cumulative experiences gained from the DRL agents and their exploration of the environment, i.e., the exploration of different private-message power allocations. Numerical results obtained via modeling and simulation verify the effectiveness of the proposed DQL algorithm, demonstrating that it concludes solutions that outperform existing approaches from the literature in the achieved energy efficiency.**

***Index Terms*—Energy Efficiency Maximization, Rate-Splitting Multiple Access (RSMA), Deep Reinforcement Learning (DRL).**

I. Introduction

Rate-Splitting Multiple Access (RSMA) is a multiple access technique that has recently gained attention as a promising solution to overcome the limiting factors of its non-orthogonal transmission predecessors regarding the problems of signal decoding complexity and interference mitigation [1]. Considering the downlink, RSMA suggests that a transmitted message intended for multiple users is partitioned into a common and a private message. The common message is decoded by all users involved in the transmission, whereas the private message is intended for each user individually. By partially decoding and partially treating interference as noise, RSMA achieves a good balance between signal processing complexity and interference management. The benefits gained from this tradeoff are directly reflected in the achieved throughput and spectral efficiency of the network in general, which is a topic that has been extensively studied from both theoretical [2] and qualitative viewpoints [3] so far.

This work was partly supported by the European Commission through the Horizon Europe/JU SNS project Hexa-X-II (Grant Agreement no. 101095759).

Targeting sustainability and not sufficing solely on high data rates, a shift is made by modern networks to green communications and energy efficiency. In this context, quantifying the performance gain of the RSMA technique in terms of achieved energy efficiency is crucial though only a handful of research works have investigated it until now, e.g., [4]–[7]. Early work in [4] studied the optimization of beamforming and common-rate allocation in Multiple-Input Single-Output (MISO) broadcast channels, while a similar Single-Input Single Output (SISO) setting was examined in [5]. Both works compared the energy efficiency level of the network against the one achieved by different multiple access techniques, verifying that the energy efficiency region of RSMA is equal or higher. Other more complex network settings were considered by the authors in [6] and [7] that regard a Cloud Radio Access Network (C-RAN) and a Reconfigurable Intelligent Surface (RIS)-assisted RSMA network, respectively, while controlling the power and rate of the common and private transmitted messages. As expected, all initial works relied on convex approximation and heuristic techniques to determine the corresponding power/beamforming and rate allocations.

As networks become even more complex in the number of wireless connections, robust optimization techniques constitute a need rather than a desire. Lately, Deep Reinforcement Learning (DRL) algorithms are broadly considered to effectively manage the burden of data and provide near-optimal solutions to non-convex problems while enabling the network's self-configuration based on the trained model. Different value, policy, or actor-critic-based methods have been implemented and tested over simple Interference Broadcast Channel (ICB) setups (e.g., [8]) to perform power allocation in the downlink. Considering RSMA networks while targeting sum rate maximization, two similar policy-based DRL methods were proposed in [9] and [10] to determine the beamforming of the common and private messages. Another work in [11] designed an actor-critic-based DRL algorithm to perform computation offloading decision, power allocation, and decoding order optimization in the uplink of an RSMA-assisted Mobile Edge Computing (MEC) network, aiming for the minimization of the weighted sum of latency and energy consumed for communication and computing. Apparently, the energy efficiency maximization for RSMA via DRL has been so far overlooked.

In this paper, we aim to fill this gap and design a DRL algorithm for energy-efficient power and rate allocation of the common and private messages transmitted in the downlink of an RSMA network. Owing to its simplicity to implement, robustness, and low complexity [12], the value-based DQL algorithm is employed to solve the corresponding energy efficiency maximization problem after properly designing the states, actions, and rewards to capture the problem's required constraints successfully. Different from the existing works in the literature of RSMA and DRL, i.e., [9]–[11], in this paper, we perform a multi-agent DRL modeling and propose a centralized training and distributed execution approach to the typical DQL algorithm, enabling the latter to scale well and provide accurate solutions as the number of users increases in the network. Each private stream constitutes a different DRL agent that contributes its personal experience from interacting with the environment by trying various actions, i.e., private stream power allocations, to training a common deep network. The latter DRL modeling can be also adopted and tested under different policy-based and actor-critic algorithms, which is part of our current work.

The remainder of the paper 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 algorithm. Section IV presents the numerical evaluation, and Section V concludes the paper.

II. Problem Statement

A. System Model

We consider a single-cell downlink RSMA communication network comprising a set of users $\mathcal{N} = \{1, \ldots, N\}$ served by a base station located at the center of the cell. The users and the base station bear both single antenna receivers and transmitter, respectively, while the data transmissions intended for different users are multiplexed over the same frequency band by adopting the RSMA technique. We denote as W_n the message intended for user n. Each message W_n is split into a common and a private part, i.e., W_n^c and W_n^p, respectively. The common parts intended for all 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 [Watts]. The remaining private messages $W_n^p, \forall n \in \mathcal{N}$ are encoded into different private streams v_n and separately transmitted to the users with downlink transmission powers p_n [Watts], $\forall n \in \mathcal{N}$.

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 respectively 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 the channel gain between user n and the base station is denoted as $G_n^{(t)}$ and $z_n^{(t)} \sim \mathcal{CN}(0, \sigma^2)$ is the corresponding Additive White Gaussian Noise (AWGN). Specifically, we define the channel gain between each user n and the base station as $G_n^{(t)} = |h_n^{(t)}|^2 \beta_n$, where $h_n^{(t)}$ represents the small-scale Rayleigh fading and β_n the large-scale fading that can remain the same over several time slots. To model the time-varying nature of the channel, we adopt Jake's model [13] and express the small-scale Rayleigh fading as a first-order Gaussian-Markov process as $h_n^{(t)} = \rho h_n^{(t-1)} + \sqrt{1-\rho^2}\zeta_n^{(t)}$, 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.

Based on the above, the achievable rate of decoding the common stream $v_0^{(t)}$ transmitted by the base station to each user n is:

$$r_n^{c\,(t)} = \log_2 \left(1 + \frac{G_n^{(t)} p_0^{(t)}}{G_n^{(t)} \sum_{j=1}^{N} p_j^{(t)} + \sigma^2} \right) \text{[bps/Hz]}. \tag{3}$$

Without loss of generality, we consider that the channel gains between the users and the base station are sorted in ascending manner, i.e., $G_1^{(t)} \leq \cdots \leq G_n^{(t)} \leq \cdots \leq G_N^{(t)}$. To ensure that all users $n \in \mathcal{N}$ can successfully decode the common stream $v_0^{(t)}$, the allocated rates $c_n^{(t)}$ of decoding the common stream at their receivers should satisfy the following condition:

$$\sum_{n=1}^{N} c_n^{(t)} \leq \min_{n \in \mathcal{N}} r_n^{c\,(t)}, \tag{4}$$

where $\min_{n \in \mathcal{N}} r_n^{c\,(t)} = r_1^{c\,(t)} = \log_2 \left(1 + \frac{p_0^{(t)}}{\sum_{j=1}^{N} p_j^{(t)} + \sigma^2 / G_1^{(t)}} \right)$, based on the ordering of the channel gains.

Additionally, for the Successive Interference Cancellation (SIC) to be successfully implemented at the receiver of each user n, the following condition should be satisfied:

$$G_n^{(t)} p_0^{(t)} - G_n^{(t)} \sum_{j=1}^{N} p_j^{(t)} \geq p_{tol}, \tag{5}$$

where p_{tol} [Watts] is the corresponding receivers' SIC decoding tolerance/sensitivity, which is assumed as the same for all users. Eq. (5) 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 by each user, the decoding of the corresponding private stream $v_n^{(t)}$ takes place, the achievable rate of which is:

$$r_n^{p\,(t)} = \log_2 \left(1 + \frac{G_n^{(t)} p_n^{(t)}}{G_n^{(t)} \sum_{j=1, j \neq n}^{N} p_j^{(t)} + \sigma^2} \right) \text{[bps/Hz]}. \tag{6}$$

Therefore, the total achievable data rate of user n is:

$$\begin{aligned} R_n^{(t)} &= c_n^{(t)} + r_n^{p\,(t)} \\ &= c_n^{(t)} + \log_2 \left(1 + \frac{G_n^{(t)} p_n^{(t)}}{G_n^{(t)} \sum_{j=1, j \neq n}^{N} p_j^{(t)} + \sigma^2} \right). \end{aligned} \tag{7}$$

B. Problem Formulation

In this paper, we strive to address the energy efficiency maximization problem in a downlink RSMA-based communication network, which is defined as the ratio between the sum of the total achievable data rates of all users $n \in \mathcal{N}$ as defined in Eq. (7), i.e., $\sum_{n=1}^{N} R_n^{(t)}$, and the total consumed power in the downlink by the base station, i.e., $p_0^{(t)} + \sum_{n=1}^{N} p_n^{(t)}$ [Watts]. Towards achieving this goal, we optimize the vectors of allocated common-stream rates $\mathbf{c}^{(t)} = [c_1^{(t)}, \cdots, c_n^{(t)}, \cdots, c_N^{(t)}]^T$ and private-stream transmission powers $\mathbf{p}^{(t)} = [p_1^{(t)}, \cdots, p_n^{(t)}, \cdots, p_N^{(t)}]^T$ by the base station to the users, as well as the common-stream power $p_0^{(t)}$. The corresponding optimization problem to be solved by the base station is formally written as follows:

$$\max_{\mathbf{c}^{(t)}, \mathbf{p}^{(t)}, p_0^{(t)}} \quad EE = \frac{\sum_{n=1}^{N} R_n^{(t)}}{p_0^{(t)} + \sum_{n=1}^{N} p_n^{(t)}} \tag{8a}$$

$$\text{s.t.} \quad \sum_{n=1}^{N} c_n^{(t)} \leq r_1^{c\,(t)}, \tag{8b}$$

$$G_1^{(t)} p_0^{(t)} - G_1^{(t)} \sum_{n=1}^{N} p_n^{(t)} + \sigma^2 \geq p_{tol}, \tag{8c}$$

$$p_0^{(t)} + \sum_{n=1}^{N} p_n^{(t)} \leq p_{max}, \tag{8d}$$

$$c_n^{(t)}, p_n^{(t)} \geq 0, \forall n \quad \text{and} \quad p_0^{(t)} \geq 0. \tag{8e}$$

Eq. (8b) and Eq. (8c) 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. (8d) indicates the base station's maximum power budget p_{max} [Watts], while Eq. (8e) defines the feasible range of values of the different optimization variables.

III. Problem Solution based on Deep Q-Learning

In this section, we first transform the considered energy efficiency maximization problem into a multi-agent DRL problem. The paradigm of centralized training and distributed execution adopted in this paper is analyzed and discussed. Then, the algorithmic framework of DQL is presented to solve the DRL problem.

A. Multi-Agent DRL Model & Architecture

The necessary constituent elements of a typical multi-agent DRL model are the agents, the environment and its state, and the agents' actions and rewards.

Agents: Each private stream $v_n^{(t)}, \forall n \in \mathcal{N}$ of the downlink transmitted signal by the base station to the users is regarded as a different agent in the proposed DRL model. Given that there exists a one-to-one mapping between the users and the private streams, i.e., DRL agents, in the following, we refer to the agents' set as $\mathcal{N} = \{1, \ldots, N\}$ and use index n to indicate a single agent.

State: At each time slot, the agents observe characteristics of the environment and form a corresponding description of what is called state. In the proposed DRL model, the local state $\mathbf{s}_n^{(t)}$ observed by agent n includes information pertinent to the transmission of the corresponding private stream $v_n^{(t)}$. Note that the power of the common and private streams changes at the end of each time slot and remains constant during the next one [13]. Therefore, at the beginning of time slot t, the agents utilize the $\mathbf{p}^{(t-1)}$ and $p_0^{(t-1)}$ power information, whereas at the end of time slot t we refer to $\mathbf{p}^{(t)}$ and $p_0^{(t)}$. The agent's n state $\mathbf{s}_n^{(t)}$ 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., $\sum_{j \in \mathcal{N}, j \neq n} G_j^{(t)} p_j^{(t-1)} + \sigma^2$;
4) the interference sensed from the rest of the private streams at the beginning of time slot $t-1$, i.e., $\sum_{j \in \mathcal{N}, j \neq n} G_j^{(t-1)} p_j^{(t-2)} + \sigma^2$;
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;
8) the data rate $c_n^{(t)}$ of the common stream.

Action: Based on the state $\mathbf{s}_n^{(t)}$, each agent chooses and performs an action $a_n^{(t)} \in \mathcal{A}_n$ from its set of possible actions $\mathcal{A}_n$. Given that p_{max} is the base station's maximum power budget, we define $p_{n,max} = \frac{p_{max}}{N+1}$ the maximum allowable transmission power of the private stream $v_n^{(t)}$ and $p_{n,min}$ is a corresponding minimum allowable power level. Then, the agent's n action space is formally defined as:

$$\mathcal{A}_n = \left\{ 0, p_{n,min}, p_{n,min} \cdot \left(\frac{p_{n,max}}{p_{n,min}} \right)^{\frac{1}{A_n - 2}}, \ldots, p_{n,max} \right\}, \tag{9}$$

where A_n is the cardinality of the set.

Given the chosen actions $a_n^{(t)} \in \mathcal{A}_n$ of all agents, the values of $(\mathbf{c}^{(t)}, p_0^{(t)})$ that maximize the energy efficiency can be obtained by analytically and exhaustively solving the following optimization problem:

$$\max_{\mathbf{c}^{(t)}, p_0^{(t)}} \quad \frac{\sum_{n=1}^{N} c_n^{(t)}}{p_0^{(t)}} \tag{10a}$$

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

$$c_n^{(t)} \geq 0, \forall n \quad \text{and} \quad p_0^{(t)} \in \mathcal{P}_0, \tag{10c}$$

where by $\mathcal{P}_0$ we denote the set of feasible values of $p_0^{(t)}$: $\mathcal{P}_0 = \{ \frac{p_{max} - \sum_{n \in N} a_n^{(t)}}{P_0}, \frac{p_{max} - \sum_{n \in N} a_n^{(t)}}{P_0 - 1}, \ldots, \frac{p_{max} - \sum_{n \in N} a_n^{(t)}}{1} \}$ and P_0 its cardinality. Indeed, the problem in (10) reduces to a linear programming problem for the different values of $p_0^{(t)}$ that can be optimally solved in polynomial time. The obtained values of $(\mathbf{c}^{(t)}, p_0^{(t)})$ satisfy constraints (8b) and (8d), while

the satisfaction of constraint (8c) is guaranteed later by the definition of the agent's reward function.

Reward: As a consequence of the chosen action $a_n^{(t)}$, each agent n moves 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 (8c). Specifically, if constraint (8c) is satisfied, the feedback signal $f_n^{(t+1)}$ is given by:

$$f_n^{(t+1)} = \frac{EE}{N}, \tag{11}$$

otherwise, it is calculated as follows:

$$f_n^{(t+1)} = \frac{EE}{N} \cdot \left(1 + \tanh\left(p_0^{(t)} - \sum_{j=1}^{N} p_j^{(t)} - \frac{p_{tol} + \sigma^2}{G_1^{(t)}}\right)\right). \tag{12}$$

The function $\tanh(x)$ asymptotically reaches -1 for negative values of x. Therefore, by the definition of the feedback signal in Eq. (12), it follows that the latter tends to zero as the violation of constraint (8c) grows, enabling the agent to learn the negative effect of constraint violation.

Concerning the architecture of the proposed DRL framework, a centralized training and distributed execution paradigm is adopted [13]. Following this paradigm, a single general-purpose model is trained centrally and shared among the distributed agents. The agents interact with the environment and employ the learned actions, generating experience samples fed back to the centralized model trainer. In this way, we capitalize on the benefits of multi-agent DRL modeling in terms of reduced action and state spaces, requiring less memory, computational resources, and execution time while maintaining the stability and efficiency of a centralized solution. The design of the reward feedback signal is crucial to effectively optimize the global objective by the agents' distributed decisions and actions. However, given its successful definition, the agents can quickly learn a more general model, benefiting from one another. The convergence of the proposed framework is experimentally shown in Fig. 1 later in Section IV.

B. Deep Q-Learning Algorithm

The DQL algorithm is a value-based DRL model, whose objective is to approximate the optimal Q-function $Q^*(\mathbf{s}, a; \boldsymbol{\theta})$ using a Deep Q-Network (DQN), where $\boldsymbol{\theta}$ is the vector of the neural network's parameters. Specifically, the optimal Q-function yields the maximum expected discounted sum of rewards $Q^{\pi}(\mathbf{s}, a; \boldsymbol{\theta})$ that an agent can take by selecting action a at state $\mathbf{s}$ using some policy π, i.e., $Q^*(\mathbf{s}, a; \boldsymbol{\theta}) = \max_{\pi} Q^{\pi}(\mathbf{s}, a; \boldsymbol{\theta})$. The latter, in turn, is commonly modeled as follows:

$$Q^{\pi}(\mathbf{s}, a; \boldsymbol{\theta}) = \mathbb{E}\left[\sum_{\tau=0}^{\infty} \gamma^{\tau} f^{(t+\tau+1)} \middle| \mathbf{s}^{(t)} = s, a^{(t)} = a\right], \tag{13}$$

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. Note that in the analysis above, the subscripts n referring to the different agents have been dropped for generalization purposes.

To combat potential instability issues of the DQL algorithm due to the high correlation of the successive states observed by an agent, the experience replay mechanism 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 $\mathbf{e}_n^{(t)} = (\mathbf{s}_n^{(t-1)}, a_n^{(t-1)}, f_n^{(t)}, \mathbf{s}_n^{(t)})$. Also, at time slot t, a minibatch $\mathcal{D}^{(t)}$ of size D of experiences is randomly created by a common randomizer, comprising an equal number of experiences from the different agents' queues, eliminating in this way training the DQN over correlated agent experiences.

To approximate the optimal Q-function $Q^*(\mathbf{s}, a; \boldsymbol{\theta})$, the least square error is calculated over minibatch $\mathcal{D}^{(t)}$ as follows:

$$L(\boldsymbol{\theta}^{(t)}) = \sum_{(\mathbf{s}, a, f', \mathbf{s}') \in \mathcal{D}^{(t)}} \left(f' - Q^{\pi}(\mathbf{s}, a; \boldsymbol{\theta}^{(t)})\right)^2, \tag{14}$$

and the DQN's parameters are updated via the gradient descend method with learning rate $\eta \in (0, 1]$:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^{(t)}). \tag{15}$$

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

Algorithm 1 Deep Q-Learning (DQL) Algorithm.

Initialize N_e, N_t, η, ϵ_1, ϵ_{N_e}, M, D.
Randomly initialize DQN's parameters $\boldsymbol{\theta}$.
for $k = 1$ to N_e **do**
  Update ϵ_k based on Eq. (16).
  Derive initial agents' states $\mathbf{s}_n^{(1)}, \forall n$.
  for $t = 1$ to N_t **do**
    for $n = 1$ to N **do**
      if rand() $\leq \epsilon_k$ **then**
        Randomly select action $a_n^{(t)} \in \mathcal{A}_n$.
      else
        Select $a_n^{(t)} = \arg\max_{a_n} Q(s_n^{(t)}, a_n; \boldsymbol{\theta}^{(t)})$.
      end if
    end for
    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 (10).
    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 agent's experience $\mathbf{e}_n^{(t)}$ in the queue.
    Create a minibatch $\mathcal{D}^{(t)}$ and calculate $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^{(t)})$.
    Update DQN's parameters $\boldsymbol{\theta}^{(t+1)}$ based on Eq. (15).
    Set $\mathbf{s}_n^{(t)} \leftarrow \mathbf{s}_n^{(t+1)}, \forall n$.
  end for
end for

selecting an action different from the optimal one $a^* = \arg\max_a Q^\pi(\mathbf{s}, a; \boldsymbol{\theta}^{(t)})$, is given by:

$$\epsilon_k = \epsilon_1 + \frac{k-1}{N_e - 1} \cdot (\epsilon_{N_e} - \epsilon_1),\ k = 1, 2, \ldots, N_e, \qquad (16)$$

where ϵ_1 and ϵ_{N_e} are the initial and final exploration probabilities, accordingly.

IV. Evaluation & Results

In this section, numerical results are presented after proper modeling and simulation to evaluate the performance of the proposed DQL algorithm for energy-efficient power and rate allocation in RSMA networks. In our experiments, we consider a circular area of 500 m radius with $N = 4$ users randomly spatially distributed. 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. The maximum Doppler frequency is $f_d = 10$ Hz and the time slot duration is $T = 20$ ms [13]. The rest of the communication-related parameters are set as $\sigma^2 = -114$ dBm, $p_{tol} = -94$ dBm, $p_{n,min} = 1$ dBm, and $p_{max} = 40$ dBm. Regarding the proposed DQL algorithm, 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. Specifically, the number of power levels of the private and common streams are set as $A_n = 10$ and $P_0 = 100$, unless otherwise explicitly stated. The Rectified Linear Unit (ReLU) is chosen as an activation function, while the rest of the DQN's hyper-parameters are set as $N_e = 1700$, $N_t = 50$, $\eta = 0.01$, $\epsilon_1 = 0.2$, $\epsilon_{N_e} = 0.0001$, $M = 5000$, and $D = 500$.

To assess the achieved energy efficiency optimization level of the proposed DRL framework (referred to as "RS DQL") under both wireless networking and algorithmic perspectives, we consider the following three alternative comparative approaches. The "No-RS DQL" approach employs a similar multi-agent DRL modeling and DQL algorithm to derive the power allocation in an SISO IBC where the messages are not split, and each user decodes only one message intended for them. This comparison allows assessing the performance of the RSMA-based network against an SISO IBC network while at the same time proving that the employed DQL algorithm can generalize well and adapt to fit from less to more complicated network setups that require optimizing a higher number of communication resources. The "RS Algorithm in [5]" regards a similar RSMA network to ours with the difference that the joint power and rate allocation problem is decomposed into sub-problems that are independently solved optimally, and an algorithm is devised to combine the derived solutions. Last, we compare against the Weighted Minimum Mean Square Error (WMMSE) approach, based on which the power allocation is determined for sum-rate maximization, and then, the achieved energy efficiency is calculated, considering that the base station has exhausted its power budget p_{max}.

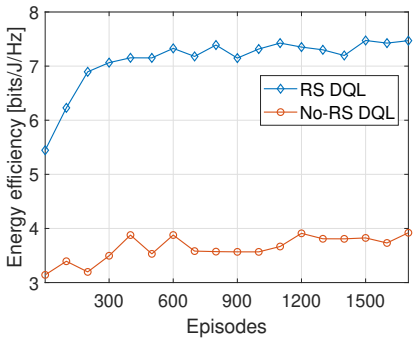

Fig. 1: Achieved energy efficiency of "RS DQL" and "No-RS DQL" over training episodes.

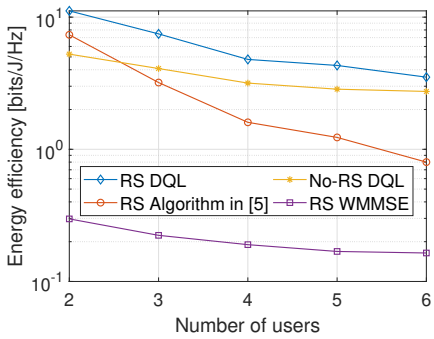

Fig. 2: Achieved energy efficiency of "RS DQL", "No-RS DQL", "RS Algorithm in [5]", and "RS WMMSE" under different numbers of users.

Fig. 1 depicts the variation in the energy efficiency optimization objective during the different training episodes under the proposed (i.e., "RS DQL") and the "NO-RS DQL" approaches. In the first episodes, where the DQL algorithm's exploration probability is high, the optimization objective increases sharply, resulting in continually improved representations of the optimal Q-function - especially under the "RS DQL" approach - until a stabilization energy efficiency level is reached. The minimal variations at the last training episodes are attributed to the time-varying nature of users' channel gains that follow Jake's model. Besides, the superiority of the rate-splitting technique is validated when compared against the "No-RS DQL" approach, achieving almost double energy efficiency at the last episodes of the DQN's training.

In Fig. 2, the "RS DQL", "No-RS DQL", "RS Algorithm in [5]", and "RS WMMSE" approaches are compared in terms of achieved energy efficiency, while accounting for different numbers of users in the cell. Note that the results presented for "RS DQL" and "No-RS DQL" from this point and on correspond to the average energy efficiency concluded by the trained DQN over 50 random episodes of 500 time slots each. A logarithmic scale has been used to better depict the small variations in the values presented. The "RS DQL" and "RS Algorithm in [5]" exhibit a similar decaying trend as the number of users gets higher due to increased interference and total transmission power required in the downlink by

the base station. However, a significant performance gap is shown between them, verifying that DQL is more successful in concluding an energy-efficient power and rate allocation than a heuristic algorithm (i.e., [5]). The "No-RS DQL" approach converges to the mean between the two rate-splitting-based approaches as the number of users increases, since the increased interference limits the benefits of rate-splitting. Regarding the achieved energy efficiency under the "RS WMMSE" approach, this is one order of magnitude lower than the rest approaches due to the WMMSE algorithm seeking to maximize the sum rate, highlighting the need for energy-efficient solutions.

Fig. 3 presents an analysis of the proposed "RS DQL" approach regarding different values of the base station's maximum emitted transmission power p_{max}. It is remarkable that as p_{max} increases, the achieved energy efficiency levels decrease (left vertical axis), which is corroborated by the findings for the achieved sum data rate (right vertical axis). A huge increment in p_{max} results in a small or even no increment in the achieved sum 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, decreases the energy efficiency values reached.

In Fig. 4, we examine the performance of DQL under the proposed "RS DQL" approach when increasing the action space regarding the number of discrete power levels $A_n, \forall n$. It is observed that there exists an "optimal" number of power levels where the tradeoff between exploring different actions and complexity in the exploration is optimal, which in our case is $A_n = 10, \forall n$ as used in the experiments overall.

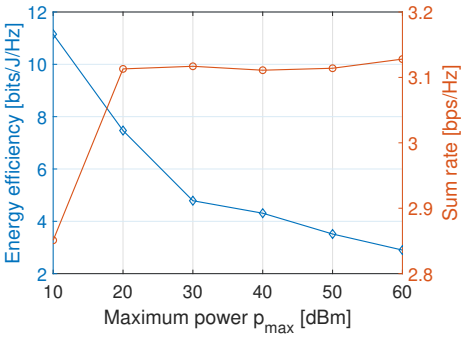

Fig. 3: Achieved energy efficiency and sum rate of "RS DQL" under different values of base station's maximum power.

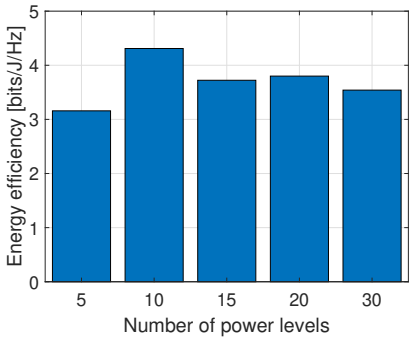

Fig. 4: Achieved energy efficiency of "RS DQL" under different numbers of power levels.

V. Conclusions

In this paper, a multi-agent DRL problem was formulated to address the energy-efficient power and rate allocation of the common and private streams in the downlink of an RSMA network. To solve the problem, a DQL algorithm was designed following the concept of centralized training of the DQN and distributed execution by the DRL agents. The DRL agents were mapped to the private streams that explore the wireless network via their actions, i.e., private stream power allocations, and contribute the experiences gained to training the common DQN. The proposed DQL-based approach achieved at least double energy efficiency levels compared to existing approaches from the literature. Our current work focuses on comparing the proposed DQL algorithm with other policy and actor-critic-based ones, highlighting their strengths and weaknesses related to the considered RSMA network setting.

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, 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, 2020.

[3] 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, 2021.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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, 2020.

[9] 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, 2022.

[10] 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, 2021.

[11] 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. 52 398–52 409, 2022.

[12] A. Feriani and E. Hossain, "Single and multi-agent deep reinforcement learning for ai-enabled wireless networks: A tutorial," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 2, pp. 1226–1252, 2021.

[13] 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, 2019.