Proceedings Paper:
Topcu, A., Lawey, A.Q. and Zaidi, S.A.R. (2024) Joint Power and Flexible Numerology Allocation in 5G Networks Using Deep Reinforcement Learning. In: 2024 11th International Conference on Wireless Networks and Mobile Communications (WINCOM). The 11th International Conference on Wireless Networks and Mobile Communications, 23-25 Jul 2024, Leeds, UK. IEEE. ISBN: 979-8-3503-7787-3. ISSN: 2769-9986. EISSN: 2769-9994.

https://doi.org/10.1109/WINCOM62286.2024.10655993

Joint Power and Flexible Numerology Allocation in 5G Networks Using Deep Reinforcement Learning

Alican Topcu, Ahmed Q. Lawey, Syed Ali Raza Zaidi
School of Electronic and Electrical Engineering, University of Leeds, Leeds LS2 9JT, U.K.
alicantopcu88@gmail.com, a.q.lawey@leeds.ac.uk, s.a.zaidi@leeds.ac.uk

***Abstract*—This study presents a novel power allocation optimisation strategy that does not rely on traditional power constraints. Unlike previous works in the literature, we introduce a ratio between the allocated and requested data rates and incorporate this ratio into the reward function of our deep reinforcement learning algorithm. The highest reward of 1 is achieved when the allocated and requested data rates are equal. Additionally, we jointly optimise power and numerology allocation, considering the users' delay and data rate requirements. Any numerology can be allocated to users as long as their requirements are satisfied. This approach enables users to be allocated optimum numerology and transmit power. By addressing the challenge posed by greedy users, our approach enhances the flexibility and performance of the power and numerology allocation process.**

***Index Terms*—Numerology allocation, power allocation, reinforcement learning**

I. Introduction

In the era of 5G and beyond, wireless communication systems are expected to be flexible and efficient. 5G provides this flexibility by supporting services with varying requirements and defining service classes such as enhanced Mobile Broadband (eMBB), Ultra-Reliable and Low-Latency Communication (URLLC), and Massive Machine-Type Communication (mMTC) [1]. To enhance network efficiency in the presence of these diverse services, it is crucial to optimise downlink transmit power among users with different requirements. Classic optimisation techniques, such as Mixed-Integer Linear Programming (MILP) is used in 5G network resource allocation, including transmit power and resource block (RB) allocation [2], subband configuration optimisation for single-base station (BS) scenarios with multiple numerologies [3], and network slicing and numerology allocation in [4]. While effective in some scenarios, MILP uses a snapshot of the network, given the rapidly changing nature of wireless networks, new coefficients emerge, rendering resource allocation in wireless networks an NP-hard problem [5]. Accordingly, the agility of power allocation is of utmost importance.

To address these challenges, several prior studies have employed Reinforcement Learning (RL) techniques in wireless communication systems. RL has been utilised to develop learning-based methods for optimising discrete subband and power allocations [5], as well as dynamic power allocation schemes using CSI [6]. Additionally, in [7], the authors solve the joint user association and resource allocation problem in the downlink direction. Moreover, power allocation problems with multiple cells and shared frequency spectrum are addressed in [8], [9], where multi-agent RL algorithms are employed. Furthermore, Deep Reinforcement Learning (DRL) can effectively tackle various optimisation challenges, including the notorious curse of dimensionality [10]. The authors in [11] introduce a novel DRL method to boost system throughput for URLLC, eMBB, and mMTC users. By removing undesirable resource allocation decisions, they shrink the action space, improving the odds of learning the optimal allocation policy. Similar to our work, the authors in [12], concentrate on flexible numerology allocation among users using DRL. However, they employed puncturing, where one type of traffic transmission (eMBB) is interrupted or stopped to accommodate another type of traffic (URLLC).

DRL has demonstrated its potential for resource allocation tasks; however, some existing research focuses on singular aspects of resource allocation, primarily power control or channel assignment, and some make the problem as hard as MILP optimisation problems. Consequently, to address these challenges, we propose DRL for power allocation and joint power and flexible numerology allocation optimisation.

The conventional reward functions are based on metrics such as SINR, spectral efficiency, or Shannon capacity for power allocation optimisation in wireless communication problems. However, we propose a new approach, where the aim is to maximise the ratio between the requested and allocated data rates. Hence, the objective function implicitly minimises downlink transmit power by maximising the ratio and prevents allocating more than the required data rate at once. Simultaneously, we ensure that each user receives their minimum required data rate. This novel approach simplifies the problem by reducing the number of constraints and enhancing fairness by handling greedy users without over-allocating resources (in terms of power and numerology) beyond their requests while optimising resource allocation in wireless communication systems using DRL techniques. Moreover, we enable flexible numerology allocation among users while considering their delay requirements.

II. System Model

We consider a single cell 5G cellular system with randomly distributed users on the grid, and available numerologies on the network indicated by the sets and $U = \{1, ..., U\}$, $N =$

This work was supported by the Republic of Türkiye Ministry of National Education for Alican Topcu's Ph.D. at the University of Leeds.

$\{1, ..., N\}$, respectively. The downlink transmit power and numerology are determined based on the users' data rate and delay requirements as well as the channel model. To evaluate the performance of our proposed RL model, we compare it with the theoretically derived optimum outputs obtained through mathematical analysis. Specifically, we calculate the path loss considering factors such as distance and frequency then use this path loss to compute the optimal transmit power for each corresponding user, assuming each user is allocated only one RB. These calculations are used to evaluate the effectiveness of the RL model in achieving optimal system performance.

The implicit objective of this problem is to enable the trained RL agent to find an optimal solution quickly. We employ DRL with a Deep Q-Network (DQN) agent, which approximates the optimal result using state, action, and reward functions and leverages deep learning's pattern recognition abilities [13]. Additionally, DQN agents can handle increased complexity in large-scale environments using deep neural networks [13].

We first optimised the power allocation problem and then extended our approach to address the multi-numerology scenario by optimising both power and numerology allocation. We utilised the same DQN agent with a neural network architecture consisting of an input layer, a fully connected layer, a Rectified Linear Unit (ReLU), another fully connected layer, another ReLU, and an output layer for both optimisation problems. Each fully connected layer contains 256 neurons, with the number of neurons in the input and output layers matching the number of users and actions, respectively.

In the DQN framework, the output layer of the neural network represents the Q-values associated with each possible action. In our case, these Q-values are discrete (can be continuous in some cases). We allocate integer numbers as transmit power to the users, and the allocated numerologies are also integer numbers. Therefore, the action space consists of integer numbers. During action selection, the agent chooses the action with the highest Q-value as determined by the neural network output. This action corresponds to the optimal decision according to the learned policy. To create the environment and train the agent, we utilised the Matlab RL environment class and the Matlab RL Designer Toolbox [14].

In the signal propagation model, channel gain is introduced as the multiplication of path loss, large-scale fading (LSF), and small-scale fading (SSF). We consider an NLOS environment where path loss calculation includes free-space path loss (FSPL) and LSF coefficients as introduced in [15]. Rayleigh distribution is used to calculate the SSF coefficients [16], accounting for the absorbing, diffracting, and scattering effects of the environment. Finally, the channel gain is calculated by converting the path loss from dB to the linear scale:

$$g_u = SSF_u \times \left(10^{-(FSPL+10\times m log(k_u)+\sigma)}\right), \quad \forall\ u \in U \tag{1}$$

The parameters $FSPL$, σ, and m have a value of 38.46 dB, 7.7 dB, and 3.1, respectively [15]. d shows the distance between BS and users. We assume block-fading and flat fading for simplicity, similar to the approach the authors took in [6].

III. Power Allocation Optimisation

First, we assess the capability of the power allocation problem, which assumes single numerology, before introducing the joint optimisation problem. Since a single BS scenario is employed, co-channel interference is not considered. Also, the choice of modulation techniques is not our primary concern; therefore, we assume that the allocated data rate equals the Shannon capacity. Moreover, in this section, we focus on the lowest 5G numerology, characterised by a 180 kHz physical RB bandwidth.

In Eq 2, the variables C_u, p_u, n_u, and g_u, represent the allocated data rate, the allocated transmit power, the bandwidth of the allocated numerology and the channel gain for the corresponding user, respectively. Additionally, σ^2 denotes the noise power spectral density, typically expressed as -174 dB/Hz [17]. Within the objective function, the ratio of the requested data rate, d_u, to the allocated data rate is maximised. Eq 4 ensures that each user receives their minimum required data rate, thereby the maximum achievable ratio can reach 1.

$$C_u = n_u \left(log_2\left(1 + \frac{p_u \times g_u}{\sigma^2}\right)\right) \tag{2}$$

$$maximise: \quad \sum_{u\in U} \frac{d_u}{C_u} \tag{3}$$

$$subject\ to: \quad C_u \geq d_u, \quad \forall\ u \in U \tag{4}$$

1) States: The state is represented as a vector of size $|U|$, where each element corresponds to a user's allocated transmit power. The deep Q-learning algorithm operates in discrete state and action spaces; hence, we determined the states as sets of integers. Specifically, the transmit power is represented as integer multiples of 1 mW. This 1 mW change between two states offers reasonable accuracy while maintaining computational efficiency

While a random selection is possible for the initial state, a more effective approach involves leveraging the system's operating conditions. Therefore, at time step 1, the initial state $s(1)$ for each user is determined as the average of the maximum and minimum allowed transmit power, denoted as p_{max} and p_{min} respectively.

$$s(t=1) = \left[\frac{p_{max}+p_{min}}{2}(1) ... \frac{p_{max}+p_{min}}{2}(u)\right] \tag{5}$$

2) Actions: The available actions for each user involve increasing or decreasing transmit power by predefined incremental values, or maintaining the current transmit power. These actions are limited to integer values for finiteness. Regarding the allocated transmit power, denoted as p_u, which also defines the state for each user, the available actions include increasing by a positive integer a, decreasing by a, or maintaining at p_u.

The finite action set, representing the number of combinations of $|a|$ actions taken $|u|$ at a time, is calculated as $a_{\text{comb}} = \binom{|a|}{|u|}$, where $|a|$ and $|u|$ denote the total number of

possible actions and the total number of users, respectively. Additionally, the total number of possible actions can be expressed as $a_{total} = |a|^{|u|}$.

Despite limiting the action space to integers, the number of possible actions is still large enough to enable the system to reach an optimum. Moreover, to maintain each user's transmit power between the p_{min} and p_{max}, the following constraint is employed. Here, the update of each element in the next state $s(t+1)$ is determined by the element-wise addition of the current state $s(t)$ with the action $a(t)$, which can be positive, negative, or zero:

$$s(t+1) = min\left(max\left((s(t)+a(t)), p_{min}\right), p_{max}\right) \tag{6}$$

3) Reward Function: The agent observes the state, applies actions to modify it, and receives a reward based on the result. The reward function operates in discrete time steps: first, computing the channel gain for each user based on the current state; then, calculating the allocated data rate for each user and the ratio of requested to allocated data rate. Individual rewards are assigned based on this ratio at time t defined at Eq 7, and the immediate reward is calculated by summing individual rewards, represented as $R_i(t)$ in Eq 8.

$$r_u(t) = \begin{cases} \dfrac{d_u}{C_u}, & \text{if } (C_u(t) \geq d_u) \\ -1, & \text{otherwise} \end{cases} \tag{7}$$

$$R_i(t) = \sum_{u \in U} r_u(t) \tag{8}$$

The immediate reward $R_i(t)$ and the reward obtained after taking an action $R_i(t+1)$ are compared to calculate the state reward $R(t)$. The range of the state reward $R(t)$ varies, from doubling the allocated data rate for a good action to -1 for an invalid action, when users receive less data rate than requested.

$$R(t) = \begin{cases} 2(R(t+1)), & \text{if } (R_i(t+1) \geq R_i(t)) \ \& \ (C_u(t+1) \geq d_u \\ R(t+1), & \text{if } C_u \geq d_u \\ -1, & \text{if } C_u < d_u \end{cases} \tag{9}$$

A. Simulation Setup

To balance exploration and exploitation, we use the exploration rate (ε), which starts at the maximum value of 1 and gradually decreases to 0.01 after 1000 time steps as the agent gains experience, as illustrated in Figure 1. We initially set γ to 0.99 to increase the future rewards.

In RL, the discount factor (γ) determines the significance of immediate and future rewards [13], which also changes between (0, 1). A low γ aims to increase immediate rewards. The discounted cumulative reward can be represented as:

$$G_t = R_{t+1} + \gamma R_{t+2} + \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{10}$$

In this work, the users' data rate requirements vary between 0.3 Mbps to 2.5 Mbps. Through a series of tests conducted using Matlab, we determined that a step size of 1 mW is an appropriate balance between resolution and the number of steps required to reach optimal solutions.

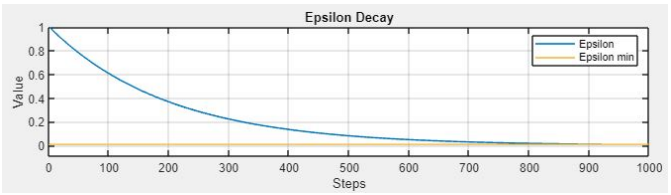

Fig. 1. Epsilon decay of the employed agent.

B. Training

We utilised a single agent, which is able to handle this complex problem with this simplified method. The maximum number of episodes is set to 2000, which provides enough training for the agent. Additionally, the episode length is defined as 50, meaning that the agent takes 50 actions in each episode and receives a summation of those rewards as the reward for the corresponding episode.

Figure 2 illustrates the episode reward achieved during training, where the dark blue line represents the average reward over ten episodes, and the light blue line represents the episode reward. The spikes indicate that the agent is still actively exploring the environment. Figure 2 also shows that the average reward steadily increases over time, indicating that the agent is learning to optimise its actions and make better decisions.

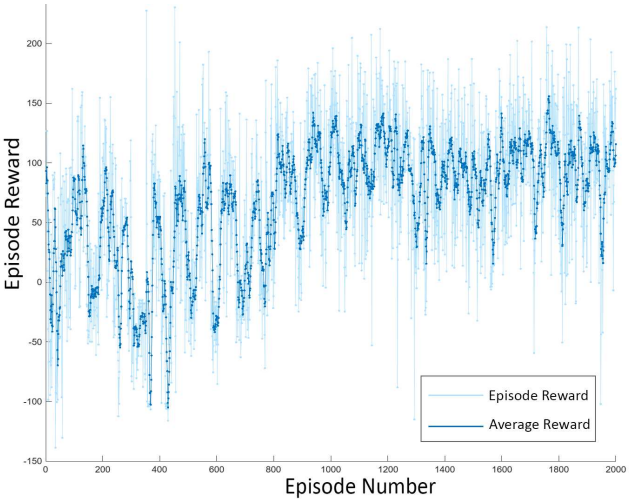

Fig. 2. The episode reward through the training phase.

C. Results

Given that the RL model is not explicitly designed to minimise transmit power, it is reasonable to observe slightly higher allocated data rates than the optimal solution. However, all the allocated data rates are greater than or equal to the requested data rates; four out of six users exceed the requested data rate by less than 0.5% while User2 and User4 exceed by

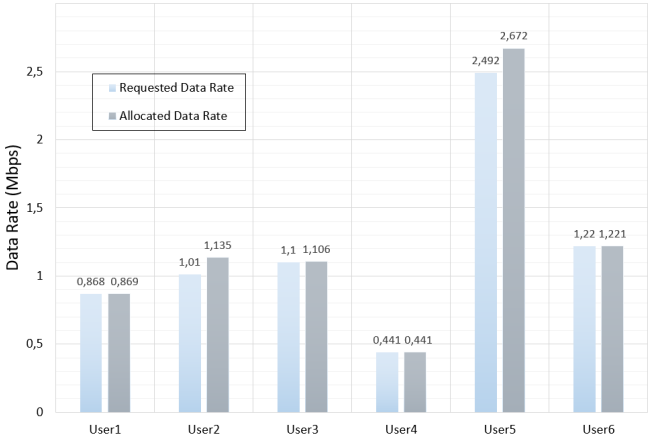

Fig. 3. The requested and allocated data rates for a six-user scenario

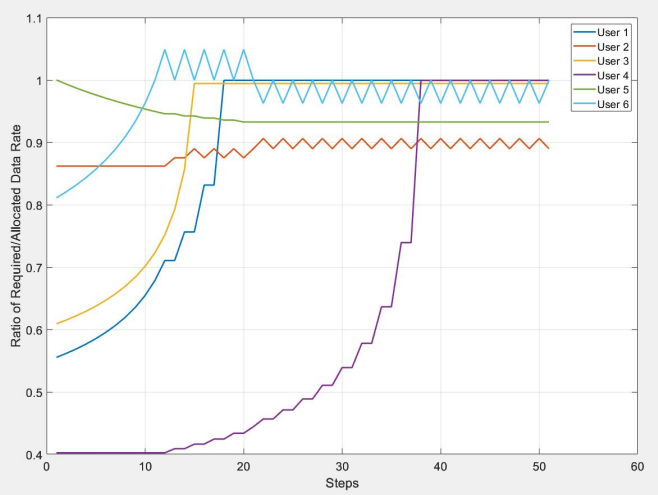

Fig. 4. The individual rewards for each episode.

12.3% and 7.2%, respectively. Figure 3 illustrates a six-user scenario's requested and allocated data rates.

Another crucial aspect in evaluating the performance of our proposed DRL algorithm is its speed in reaching the optimal solution. To evaluate this aspect, we illustrate the steps that are required for the algorithm to achieve near-optimal power allocation in Figure 4. The trained DQN agent could provide an optimal solution within a few steps. The figure shows that after the 37^{th} step, the agent no longer takes significant actions, indicating that the optimal solution has been achieved.

Interestingly, as we obtain the state reward by comparing the current state with the next state, the RL agent continuously changes the power values in each state to maximise the future reward. This necessity for change to achieve a higher reward in each state leads to the observed ripple effect in the performance of users number 2 and 6.

IV. Joint Power & Numerology Allocation Optimisation

In this section, we extend the power allocation problem presented in Section III by incorporating numerology allocation. We assume higher data rate requirements to demonstrate the feasibility of joint power and numerology allocation using DRL, which can be provided by using higher numerologies with higher bandwidths. Moreover, some users have strict delay requirements that can be satisfied by higher numerologies. We assume that any numerology can be allocated if their requirements are met. We maintain the integer power values for the users to prevent further expansion of the action space and to keep most aspects consistent with the previous section.

This problem employs a similar objective function, presented in Eq 11. The reward equals the requested and allocated data rate ratio multiplied by the maximum allowed transmit power, and the employed transmit power is subtracted from the multiplication. The maximum positive outcome can be achieved when the required data rate equals allocated. Conversely, the total transmit power is minimised to encourage the agent to choose the higher numerologies if it is available.

$$maximise: \quad \left(\left(\sum_{u\in U}\frac{d_u}{C_u}\right)\times p_{mx}\right)-\sum_{u\in U}p_u \tag{11}$$

$$subject\ to: \quad C_u \geq d_u, \quad \forall\ u\in U \tag{12}$$

$$subject\ to: \quad t_u \geq t_n, \quad \forall\ u\in U \tag{13}$$

Eq 12, which prevents the model from receiving false positive objectives, is identical to the one presented in Eq 4. Eq 13 guarantees that the allocated numerologies' delay is less than or equal to the delay requirement. In Eq 13, t_u and t_n represent the delay requirement of the users and slot length of the corresponding numerology, respectively. This section assumes that the carrier bandwidth has three numerologies, allowing n_u to take values of 180, 360, and 720 kHz.

1) States: The state is defined by a single vector divided into two parts. The first half represents transmit power allocation for each user in time step t ($p(t)$), while the second half shows the allocated numerology for each user in time step t ($n(t)$). Each vector has a size $|U|$.

$$s_u(t) = [p(t) \quad n(t)] \tag{14}$$

To ensure that the initial system state is a sample of the operating conditions, $p_u(t)$ starts with the average of the maximum and minimum permitted transmit power for each user, as before, and $n(t)$ starts with the maximum numerology for each user.

2) Actions: The agent simultaneously performs transmit power and numerology actions for each user. The action space is the combination of the set of power and numerology action spaces, $a_{\text{comb}} = \binom{|a_p|}{|u|}\binom{|a_n|}{|u|}$, where $|a_p|$, and $|a_n|$ represent the total number of power actions per user and total number

of numerology actions per user, respectively. Therefore, the number of elements in the set of actions increases exponentially. The set of power actions is the same; it involves 1 mW adjustments for each user. The set of numerology actions allows direct switching between numerologies regardless of the previous state for each user. The total number of available actions is expressed as $a_{total} = |a_p|^{|u|} \times |a_n|^{|u|}$.

To ensure transmit power remains within allowable limits after the action is taken, we use the same constraint as in Section III, Eq 6. However, the numerology allocation operation directly modifies the states without any arithmetic operations; hence, an additional constraint is not required.

3) Reward: In this section, the reward function is simplified to be based solely on the current state. The reward function considers the delay and data rate requirements constraints. If any user requirement constraints are not satisfied, the individual reward is set to a negative value of the maximum allowed transmit power. On the other hand, if both constraints are satisfied, the individual reward is computed using the same approach as the objective function. The individual user reward $(r_u(t))$ at time t is defined in Eq 15, and the state reward for the current state at time t is denoted as $R(t)$ in Eq 16.

$$r_u(t) = \begin{cases} \left(\frac{d_u}{C_u} \times p_{mx} - p_u\right), \\ \quad \text{if } (C_u \geq d_u) \ \& \ (t_u \leq t_n), \ \forall \ u \in U \\ -p_{mx}, \\ \quad \text{if } (C_u < d_u) \text{ or } (t_u > t_n), \ \forall \ u \in U \end{cases} \tag{15}$$

$$R(t) = \sum_{u \in U} r_u(t) \tag{16}$$

A. Simulation Setup

The identical ε is employed in this section, as illustrated in Figure 1. Although we keep the ε the same as employed in the power allocation problem, the exploration phase concludes more slowly due to an exponentially expanded action space. Therefore, the agent is expected to take more time to find the optimal policy and improve it with the agent's acquired knowledge. Hence, γ is set to 0.75 to increase the importance of immediate rewards. Thus, the agent prioritises more accurate actions at the beginning to maximise the episode reward.

To demonstrate the feasibility of the proposed model, we implemented it with five users and three numerologies. We have allocated more RBs to 3rd numerology in the available bandwidth to enhance RB performance. While this reduces the overall number of RBs, it is a manageable trade-off given the relatively low network population density. Specifically, we assume a 5 MHz bandwidth, with 1 RB, 1 RB, and 6 RBs allocated for numerologies 1, 2, and 3, with slot lengths of 1 ms, 0.5 ms, and 0.25 ms, respectively.

B. Training

Long episode lengths might be challenging for the agent due to the difficulty of determining the action-reward relationship. Therefore, the episode length is maintained at 50 despite the exponentially increased problem size. Hence, the agent can more easily identify which actions lead to better or worse outcomes in achieving the desired results. However, the maximum number of episodes is increased to 2400. Figure 5 shows the episode reward achieved during training, where the dark blue line represents the average reward over 10 episodes, and the light blue line shows the individual reward.

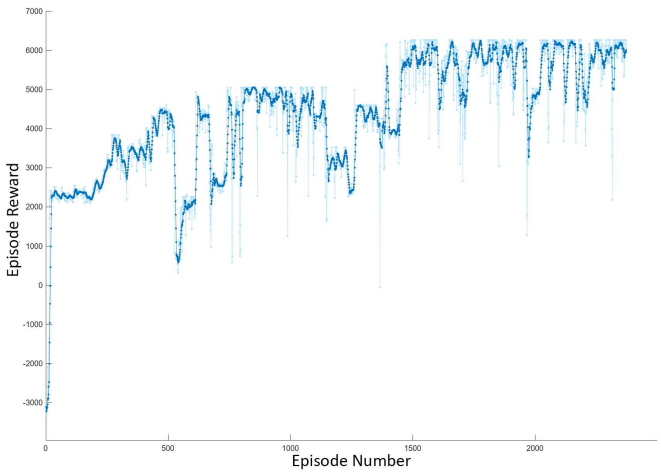

Fig. 5. The episode reward through the training phase.

Compared to the previous section, the increasing trend of the episode reward can be seen more clearly in Figure 5. The presence of spikes in the reward indicates that the agent is still actively exploring the environment till the end of the training. The fluctuations of the spikes, on the other hand, are reduced due to a relative reduction in ε and a reduction in γ.

C. Results

Figure 6 illustrates the requested and allocated data rates, where all data and delay requirements are successfully satisfied for all users. Also, Table I shows the delay requirements, which are also satisfied for all users, corresponding numerology allocations, and the allocated power for all users. Only User3 has a critical delay requirement, necessitating the allocation of the 3^{rd} numerology. Although any numerology can be allocated to the users, the 1^{st} numerology is allocated to User4, to ensure that the allocated data rate exactly meets the requested data rate.

TABLE I
DELAY REQUIREMENTS AND NUMEROLOGY ALLOCATION OF THE USERS

	User1	User2	User3	User4	User5
Delay requirement	> 1 ms	> 1 ms	0.45 ms	> 1 ms	> 1 ms
Allocated numerology	4^{th}	4^{th}	4^{th}	1^{st}	2^{nd}
Allocated power	$1mW$	$5mW$	$1mW$	$1mW$	$1mW$

The proposed DRL algorithm's speed in reaching the optimal solution is evaluated in Figure 7, where the optimal solution is achieved at the 40^{th} step for the first time. The actions of User2 change in a pattern; therefore, the optimum output is achieved more than once in 60 iterations. Also,

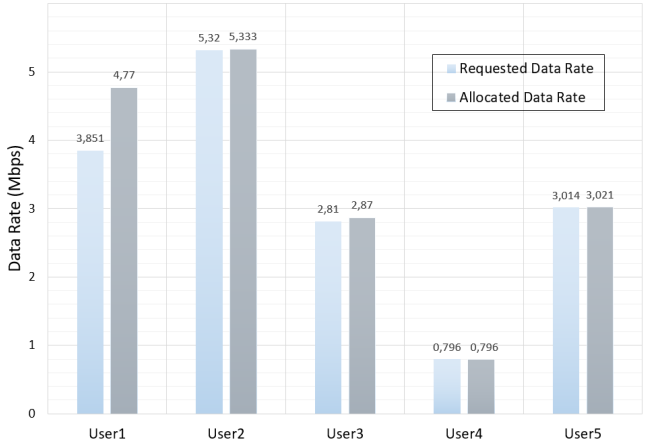

Fig. 6. The requested and allocated data rates for the joint optimisation problem

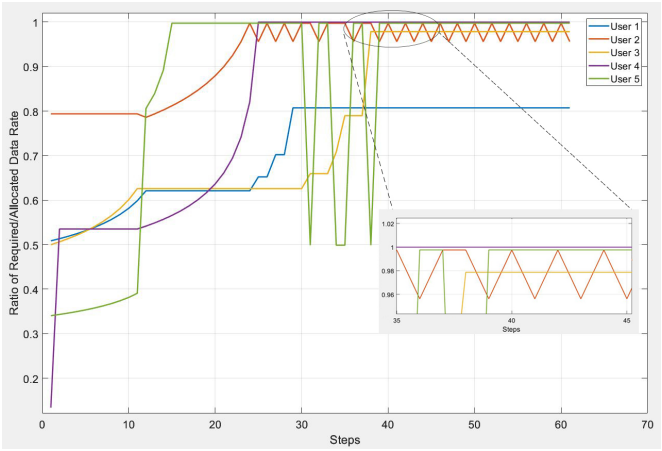

Fig. 7. The requested and allocated data rates for a five-user scenario

it's important to emphasize that using integer numbers for transmit power leads to User1 achieving its maximum ratio of 0.8. The obtained results demonstrate that the proposed DRL algorithm shows promising performance in terms of reaching optimal solutions within a few steps. These findings highlight the efficiency and effectiveness of the proposed RL algorithm.

V. Conclusions

In this paper, we considered a downlink power and numerology allocation for 5G cellular systems. The agent of the DRL algorithm effectively selects appropriate numerologies and power levels for each user based on the reward function and optimises individual user performances within a few steps. We employ a single agent that is responsible for both power and numerology allocation for all users. Introducing additional dimensions exponentially increases the problem's complexity; however, a single DQN agent has successfully solved the problem, demonstrating its feasibility. In achieving optimal performance in the training process, parameters such as ε, γ, and the number of steps in each episode must be adapted; therefore, parameter tuning plays an essential role. As a future work, this DRL-based approach will be extended to address user association in HetNets and data routing.

Acknowledgement

The authors thank Dr Ali M. Hayajneh for his support in demonstrating how to use the Matlab RL Designer Toolbox.

References

[1] 3rd Generation Partnership Project, "Study on scenarios and requirements for next generation access technologies; (release 14)," 3GPP, Tech. Rep. document 38.913, ver 14.3.0, August 2017.

[2] P. K. Korrai, E. Lagunas, A. Bandi, S. K. Sharma, and S. Chatzinotas, "Joint power and resource block allocation for mixed-numerology-based 5g downlink under imperfect csi," *IEEE Open Journal of the Communications Society*, vol. 1, pp. 1583–1601, 2020.

[3] S. Lagen, B. Bojovic, S. Goyal, L. Giupponi, and J. Mangues-Bafalluy, "Subband configuration optimization for multiplexing of numerologies in 5g tdd new radio," in *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. IEEE, 2018, pp. 1–7.

[4] V. N. Ha, T. T. Nguyen, L. B. Le, and J.-F. Frigon, "Admission control and network slicing for multi-numerology 5g wireless networks," *IEEE Networking Letters*, vol. 2, no. 1, pp. 5–9, 2019.

[5] Y. S. Nasir and D. Guo, "Deep reinforcement learning for joint spectrum and power allocation in cellular networks," in *2021 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2021, pp. 1–6.

[6] ——, "Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, pp. 2239–2250, 2019.

[7] N. Zhao, Y.-C. Liang, D. Niyato, Y. Pei, M. Wu, and Y. Jiang, "Deep reinforcement learning for user association and resource allocation in heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 11, pp. 5141–5152, 2019.

[8] E. Ghadimi, F. D. Calabrese, G. Peters, and P. Soldati, "A reinforcement learning approach to power control and rate adaptation in cellular networks," in *2017 IEEE International Conference on Communications (ICC)*. IEEE, 2017, pp. 1–7.

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

[10] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26–38, 2017.

[11] K. Suh, S. Kim, Y. Ahn, S. Kim, H. Ju, and B. Shim, "Deep reinforcement learning-based network slicing for beyond 5g," *IEEE Access*, vol. 10, pp. 7384–7395, 2022.

[12] M. Alsenwi, N. H. Tran, M. Bennis, S. R. Pandey, A. K. Bairagi, and C. S. Hong, "Intelligent resource slicing for embb and urllc coexistence in 5g and beyond: A deep reinforcement learning based approach," *IEEE Transactions on Wireless Communications*, vol. 20, no. 7, pp. 4585–4600, 2021.

[13] Matlab, "Reinforcement Learning Onramp (Overview of Reinforcement Learning)," online, accessed 24/09/23, 2023, https://matlabacademy.mathworks.com/details /reinforcement-learning-onramp/reinforcementlearning.

[14] ——, "Reinforcement Learning Toolbox," online, accessed 24/09/23, 2023, https://uk.mathworks.com/products/reinforcement-learning.html.

[15] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE journal on selected areas in communications*, vol. 32, no. 6, pp. 1164–1179, 2014.

[16] 3rd Generation Partnership Project, "Radio frequency (rf) requirements for lte pico node b (release 13)," 3GPP, Tech. Rep. document TR 36.931 V13.0.0, 2016.

[17] ITU-R, "Guidelines for evaluation of radio interface technologies for imt-2020," ITU, Tech. Rep. ITU-R M. 2412-0, 2017.