



A novel approach to efficient resource allocation in load-balanced cellular networks using hierarchical DRL

A. Mirzaei Rahimi¹ · Amir Ziaeddini² · Shu Gonglee³

Received: 1 February 2020 / Accepted: 25 March 2021 / Published online: 13 April 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract

5G wireless networks require highly spectral-efficient multiple access techniques, which play an important role in determining the performance of mobile communication systems. Multiple access techniques can be classified into orthogonal and nonorthogonal based on the way the resources are allocated to the users. This paper investigates the joint user association and resource allocation problem in an uplink multicast NOMA system to maximize the power efficiency with guaranteeing the quality-of-experience of all subscribers. We also introduce an adaptive load balancing approach that aspires to obtain “almost optimal” fairness among servers from the quality of service (QoS) perspective in which learning automata (LA) has been used to find the optimal solution for this dynamic problem. This approach contains a sophisticated learning automata which consists of time-separation and the “artificial” ergodic paradigms. Different from conventional model-based resource allocation methods, this paper suggested a hierarchical reinforcement learning based frameworks to solve this non-convex and dynamic power optimization problem, referred to as hierarchical deep learning-based resource allocation framework. The entire resource allocation policies of this framework are adjusted by updating the weights of their neural networks according to feedback of the system. The presented learning automata find the ϵ -optimal solution for the problem by resorting to a two-time scale-based SLA paradigm. Numerical results show that the suggested hierarchical resource allocation framework in combination with the load balancing approach, can significantly improve the energy efficiency of the whole NOMA system compared with other approaches.

Keywords Hierarchical resource allocation · Fair load balancing · Network function virtualization · Next generation mobile networks

1 Introduction

The ever-increasing traffic demand in mobile communications has motivated research activities to design the next generation (5G) wireless networks that can offer significant improvements in coverage and user experience (Li et al. 2020). 5G wireless networks require highly spectral-efficient multiple access techniques, which play an important role in determining the performance of mobile

communication systems. Multiple access techniques can be classified into orthogonal and nonorthogonal based on the way the resources are allocated to the users (Liu et al. 2018). Recently, related work has emerged to investigate the resource allocation problem in NOMA systems to optimize the system SE. In Khan et al. (2020), game theory is applied to allocate power among users in single-carrier NOMA (SC-NOMA) systems to maximize the revenue of BS. Song et al. (2018) proposed a suboptimal power and channel allocation algorithm combining Lagrangian duality and dynamic programming to maximize the weighted sum rate in a multi-carrier NOMA (MC-NOMA) system. In Wei et al. (2018), for a two-user multiple-input multiple-output (MIMO) NOMA system, the authors propose two power allocation algorithms to maximize the sum capacity under the total power constraint and the minimum rate requirement of weak users.

✉ Shu Gonglee
shu.g.gust@gmail.com

¹ Department of Computer Engineering, Ardabil Branch, Islamic Azad University, Ardabil, Iran

² Computer Networks Lab, Communications and Information Technology Institute, Tehran, Iran

³ Department of Computer Science, Chengdu University of Technology, Sichuan 610059, China

Many model-based resource allocation algorithms have been proposed to increase EE or other objectives in NOMA systems. The power allocation problems were studied by (Zeng et al. 2017, 2019; Liu et al. 2019a, b; Baidas et al. 2019), and the joint scheduling and power allocation problems were studied by (Cao et al. 2018; Praveenchandar and Tamilarasi 2020; Zhai et al. 2018; Maimó et al. 2019; Fu et al. 2019). However, considering the dynamics and uncertainty that are inherent in wireless communication systems, it is generally hard or even unavailable to obtain the complete knowledge or mathematical model that are required in these conventional resource allocation approaches in practice (Abedin et al. 2018 and Ye and Li 2018). Besides, due to the high computational complexity of these algorithms, they are inefficient or even inapplicable for future communication networks.

Some studies have used deep learning (DL) as a model-free and data-driven approach to reduce the computational complexity with available training inputs and outputs (Celdrán et al. 2019; Huang et al. 2019). As one main branch of machine learning (ML), DL has been used to solve resource allocation problems (Liu et al. 2019a, b; Ye et al. 2019) by training the neural networks offline with simulated data first, then outputting results with the well-trained networks during the online process. However, the correct data set or optimal solutions used for training can be difficult to obtain, and the training process itself is usually time-consuming. Given the above issues, reinforcement learning (RL) (Xu et al. 2018), as another main branch of ML, can be a feasible option for real-time decision-making tasks (e.g., dynamic resource allocation) since in RL the requirements of system model and priori data is widely relaxed. Besides, instead of optimizing current benefits only, RL can generate almost optimal decision policy which maximizes the long-term performance of systems through constant interactions. However, conventional RL algorithms suffer from slow convergence speed and become less efficient for problems with large state and action spaces. Therefore, deep reinforcement learning (DRL), which combines DL with RL, has been proposed to overcome these issues. One famous algorithm of DRL named deep Q-learning (Teng et al. 2020), uses a deep Q network (DQN) which applies deep neural networks as function approximators to conventional RL, and has already been used in many aspects such as power control in NOMA system (Zhao et al. 2019), resource allocation in heterogeneous network (Liu et al. 2018) and internet of things (IoT) (Verhelst and Moons 2017).

However, the main drawback of DQN is that the output decision can only be discrete, which brings quantization error for continuous action tasks (e.g., power allocation). Besides, the output dimension of DQN increases exponentially for multi-action and joint optimization tasks. Fortunately, the recently proposed deep deterministic policy gradient (DDPG) (Qiu et al. 2019) is able to solve these issues. DDPG is an enhanced version of the deterministic policy

gradient (DPG) algorithm (Xie and Zhong 2020) based on the actor-critic architecture, which uses an actor network to generate a deterministic action and a critic network to evaluate the action. DDPG also takes the advantages of experience replay and target network strategies from DQN to improve learning stability, which makes DDPG more efficient for dynamic resource allocation problems.

This research considers the load distribution criterion in hierarchical heterogeneous networks which is very important in the next-generation wireless networks and tries to propose a prominent load balancing approach to be adaptive to multi-layer configured networks. We also investigate the hierarchical resource allocation problem in an uplink NOMA system. Motivated by the aforementioned considerations, we present a dynamic framework to improve the long-term energy efficiency of the NOMA system while ensuring a minimum data rate of all subscribers. The first part of the framework is fully based on the dynamic model of DQN, while the second part of the framework combines the advantages of DQN and DDPG considering the load balancing constraints. Based on the DRL methods they use, we refer to this framework as the continuous DRL-based resource allocation, the continuous DRL based resource allocation (CDRA) framework.

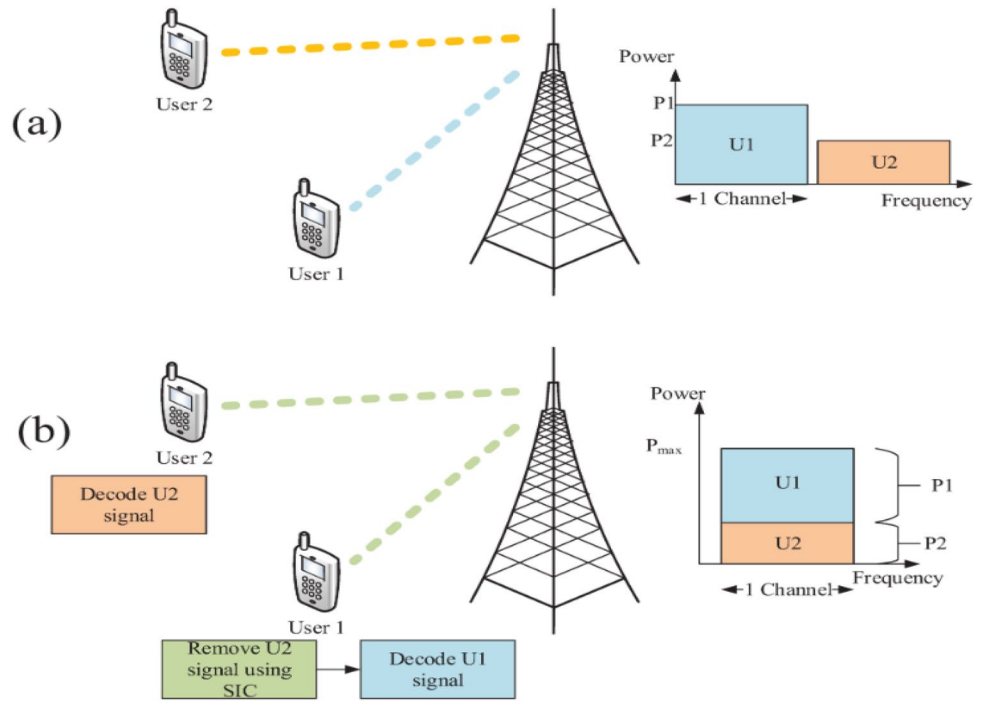
The main idea of this paper is based on a claim which the performance of NOMA resource allocation schemes can significantly increase joining with stochastic-based load balancing approaches. This research mainly investigates about the functionality of a prominent resource allocation scheme empowered with an effective load balancing approach in NOMA heterogeneous wireless networks. In this regard, we first interpret a multi agent-based resource allocation scheme using DRL, which has been presented by Li et al. After presenting an adaptive load balancing approach, we prove that the performance of the integrated scheme will significantly increase from the total network energy consumption, forced termination probability, and resource utilization rate perspectives.

The rest of this paper is organized as follows. Section 2 introduces the system model and problem formulation of the uplink NOMA system. Section 3 is relevant to the discrete and continues resource allocation framework. And in Sect. 4, the load balancing approach for dynamic resource allocation is proposed. The simulation results are given in Sect. 5, and final conclusions in Sect. 6.

2 System model and assumptions

As shown in Fig. 1, we consider a user association and resource allocation system considering NOMA channels in which, one base station (BS) considered at the center point of the macro cell and M users scattered randomly in the covered area. The UEs have randomly movement with a velocity of v within the macro cell covered radius. The total provided

Fig. 1 User association and resource allocation system considering NOMA channels



bandwidth will be shared equally to K orthogonal carriers, so the interference among carriers is harmful. The set of users and carriers are demonstrated as $\mathbf{m} = \{1, 2, \dots, M\}$ and $\mathcal{K} = \{1, 2, \dots, K\}$. Suppose each carrier can support two UEs with the highest level of channel quality at a same time, $b_{k,m}(t)$ represents the carrier allocation factor on time-slot (TS) t , in which $b_{k,m}(t) = 1$ illustrates that carrier k is allocated to UE m on time slot t , otherwise $b_{k,m}(t) = 0$. Hence, the superposition coded signal transmitted on subchannel k is as Eq. (1).

From the performance assessment perspective, based on the indicated system model, the system functionality can be represented as the following:

$$S = \{(i, j, k, m, n) \mid i \geq 0, j \geq 0, i + j + k \leq C, 0 \leq m \leq M, 0 \leq n \leq N\}$$

All of the steady states can be categorized in some different classes based on the queues' condition and the available resources. In this stage, each state's manner should be investigated and the equivalent formulation is obtained. In this scenario, we apply the presented queuing method in (Zhai et al. 2018; Maimó et al. 2019) in which we need Boolean indexes in order to study interaction between different states.

All of the probable states we can consider for the scenario are distinct but to simplify the deployment of the scenario, we consider only four states S_i ($i = 1, 2, 3, 4$) so as $S = S_1 \cup S_2 \cup S_3 \cup S_4$ we can denote the four mentioned states as below cases:

$$S_1 = \{(i, j, k, m, n) \mid i = 0; j = 0; k = 0; m = 0; n = 0\}$$

$$S_2 = \{(i, j, k, m, n) \mid 0 < (i + j + k) < C; m = 0; n = 0\}$$

$$S_3 = \{(i, j, k, m, n) \mid 0 < (i + j + k) = C; m = 0; n = 0\}$$

$$S_4 = \{(i, j, k, m, n) \mid C < (i + j + k + n + m) \leq (C + M + N); 0 \leq m \leq M; 0 \leq n \leq N\}$$

in the following, we should interpret these four steady states and try to find their equivalent formulations.

$$x_k = \sum_{m=1}^M b_{k,m}(t) \sqrt{P_{k,m}(t)} d_{k,m}(t) + \beta_{k,m}(t) \tag{1}$$

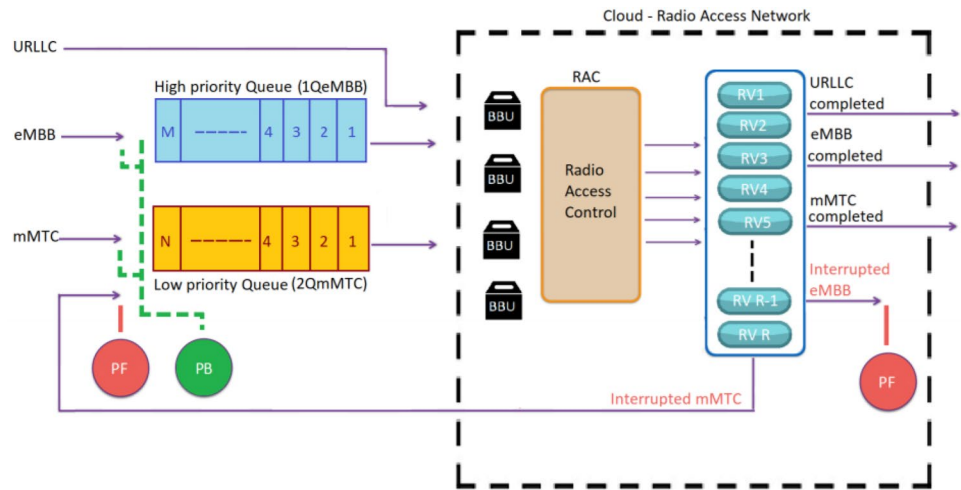
where $d_{k,m}(t)$ and $P_{k,m}(t)$ illustrate the data symbol and transmission power of UE m on carrier k , and $d_{k,m}(t)$ meets the constraint of $\mathbb{E}[|d_{k,m}(t)|^2] = 1$ in which $\mathbb{E}[\cdot]$ denotes the expectation and $|\cdot|$ shows the absolute degree operation. Note that we have applied queuing model for different QoS-class of data as shown in Fig. 2.

It should be noted that according to the described framework, overall normalized balancing formulation will be achievable through equality (12)

$$\pi Q = 0; \text{ and } \sum_{\forall S} \pi_{i,j,k,m,n} = 1$$

In this formulation, $g_{k,m}(t)$ demonstrate the fast Rayleigh fading between base station and UE m on carrier k , which can be calculated as $g_{k,m}(t) = \sqrt{\beta_{k,m}(t)} h_{k,m}(t)$, in which $\beta_{k,m}(t)$ is the fast Rayleigh fading factor supposed to be fixed during one specific time slot but varies over different time slots, and $h_{k,m}(t)$ represents the small scale fading with a normal Gaussian distribution $h_{k,m}(t) \sim \mathcal{CN}(0,1)$. So, the

Fig. 2 Applied queuing model for different QoS-class of data



corresponding signal received at the base station in time slot t can be mentioned as

$$y_{k,m}(t) = b_{k,m}(t)g_{k,m}(t)\sqrt{p_{k,m}(t)}d_{k,m}(t) + \sum_{i=1, i \neq m}^M b_{k,i}(t)\sqrt{p_{k,i}(t)}d_{k,i}(t) + \beta_{k,m}(t)z_{k,m}(t) \tag{2}$$

The first part of the formulation (2) indicates the actual sent signal of UE m on carrier k , and the second section of this equation shows received signal related to other UEs on the particular carrier. The last part of the equation also indicates complete gaussian noise with the distribution pattern of $z_{k,m}(t) \sim \mathcal{CN}(0, \sigma_z^2)$.

In the assessment stage, in accordance with the defined probability matrix π we will be able to determine the performance assessment indexes in order to compare enhanced mobile broadband and massive and machine-type communications from the quality-of-service point of view. During the performance evaluation process, resource utilization process is considered in both enhanced mobile broadband and massive machine-type communications services as the following. At the destination, successive interference cancellation is applied to decode multiple arrival packets concurrently (Pan et al. 2018), so the interference caused by carrier sharing is prohibited. First, the destination decodes the UE with the highest channel quality level, extracts it from the total received signal and then decodes the remaining signal which contains weaker signals. Hence, the SINR of UE m on carrier k can be represented as

$$\Gamma_{k,m}(t) = \frac{b_{k,m}(t)p_{k,m}(t)|g_{k,m}(t)|^2}{\sum_{i=1, |g_{k,m}(t)|^2 < |g_{k,m}(t)|^2}^M b_{k,i}(t)\beta_{k,m}(t)p_{k,i}(t)|g_{k,i}(t)|^2 + \sigma_n^2} \tag{3}$$

With normalized bandwidth, the corresponding data rate is

$$\xi_{k,m}(t) = \log(1 + \Gamma_{k,m}(t)) \tag{4}$$

In the fully-utilized system which the resource is not enough for the congested network, enhanced mobile broadband should send a demand sequence to release some of the basebands' resource blocks and dedicate the released resources to the coming ultra-reliable and low-latency communications. In this condition, the forced termination probability will be one of the primary assessment indexes applied in the paper. Because limitation of the uplink transmission rate of each UEs, the non-orthogonal multiple access system determines a user level uplink energy efficiency on carrier k which is computed as ratio of throughput to the total power consumption.

$$E_{k,m}(t) = \frac{\xi_{k,m}(t)}{\beta_{k,m}(t)p_{k,m}(t) + \mathcal{P}_m}, \tag{5}$$

where \mathcal{P}_m is a certain amount of power consumed by the device of user m itself (such as baseband signal processing, digital to analog converter and transmit filter), and the larger $E_{k,m}(t)$ represents the idea of having more transmit rate while consuming less power. We aim to optimize the EE performance of the whole NOMA system, so the resource allocation problem is formulated as maximizing the sum EE of all users through selecting the subchannel assignment index $\{b_{k,m}(t)\}$ and the allocated power $\{p_{k,m}(t)\}$ of every TS t , that is

$$P1 : \max_{\{b_{k,m}(t), p_{k,m}(t)\}} \sum_{k=1}^K \sum_{m=1}^M \beta_{k,m}(t) E_{k,m}(t) \tag{6}$$

$$S.t.C1 : \sum_{k=1}^k b_{k,m}(t) \beta_{k,m}(t) p_{k,m}(t) \leq P_{max}, \forall m \in M,$$

$$C2 : \sum_{k=1}^k b_{k,m}(t) R_{k,m}(t) \geq R_{min}, \forall m \in M,$$

$$C3 : p_{k,m}(t) \geq 0, \forall k \in \mathcal{K}, m \in M,$$

$$C4 : b_{k,m}(t) \in \{0,1\}, \forall k \in \mathcal{K}, m \in M,$$

$$C5 : \sum_{k=1}^K b_{k,m}(t) \leq 1, \forall m \in M,$$

$$C6 : \sum_{m=1}^M b_{k,m}(t) \leq C, \forall k \in \mathcal{K},$$

Applying the balancing formulation, the number of enhanced mobile broadband resource blocks which should be released can be calculated by

$$\sum_{\substack{vs \\ i+j+k=C \text{ and } k=0}} \frac{\lambda_1 j}{C-j} \pi_{i,j,k,m,n}$$

where P_{max} is the maximum available power of a device for signal transmission. Constraint C1 indicates that the transmit power of any user cannot exceed the power limitation P_{max} , and C2 is the minimum data rate requirement R_{min} for all users to ensure their quality-of-service (QoS). C3 and C4 are the inherent constraints of $p_{k,m}(t)$ and $b_{k,m}(t)$. C5 and C6 suggest that one subchannel can serve no more than C users, while each user can only access to one subchannel at the same time.

The service completion index of enhanced mobile broadband and massive and machine-type communications which can be calculated as final number of completed services is achievable as

$$\begin{aligned} S_1 &= \{(i, j, k, m, n) \mid i = 0; j = 0; k = 0; m = 0; n = 0\} \\ S_2 &= \{(i, j, k, m, n) \mid 0 < (i + j + k) < C; m = 0; n = 0\} \\ S_3 &= \{(i, j, k, m, n) \mid 0 < (i + j + k) = C; m = 0; n = 0\} \\ S_4 &= \{(i, j, k, m, n) \mid C < (i + j + k + n + m) \leq (C + M + N); 0 \leq m \leq M; 0 \leq n \leq N\} \end{aligned}$$

$$SCR_1 = \sum_{vs} j \mu_2 \pi_{i,j,k,m,n}$$

$$SCR_2 = \sum_{vs} j \mu_3 \pi_{i,j,k,m,n}$$

With deploying the scenario, it's obvious that for each steady state (i, j, k, m, n) with C number of available discrete resource, $(i + j + k)$ base band units are utilized by ultra-reliable and low-latency communications, enhanced mobile broadband or massive machine-type communications services. So, the utilization rate of enhanced mobile broadband or massive machine-type communications demands can be calculated as:

$$U = \sum_{vs} \frac{i + j + k}{C} \pi_{i,j,k,m,n}$$

The optimization problem is non-convex and NP-hard, and the global optimal solution is usually difficult to obtain in practice due to the high computational complexity and the randomly evolving channel conditions. More importantly, the conventional model-based approaches can hardly satisfy the requirements of future wireless communication services. Thus, we present the DRL-based frameworks in the following sections to deal with these problems.

From the performance assessment perspective, based on the indicated system model, the system functionality can be represented as the following:

$$S = \{(i, j, k, m, n) \mid i \geq 0, j \geq 0, i + j + k \leq C, 0 \leq m \leq M, 0 \leq n \leq N\} \tag{7}$$

All of the steady states can be categorized in some different classes based on the queues' condition and the available resources. In this stage, each state's manner should be investigated and the equivalent formulation is obtained. In this scenario, we apply the presented queueing method of (Zhai et al. 2018; Maimó et al. 2019) in which we need Boolean indexes in order to study interaction between different states.

All of the probable states we can consider for the scenario are distinct but to simplify the deployment of the scenario, we consider only four states S_i ($i = 1, 2, 3, 4$) so as $S = S_1 \cup S_2 \cup S_3 \cup S_4$ we can denote the four mentioned states as below cases:

Now we should interpret these four steady states and try to find their equivalent formulations.

The first state (S1): this state is relevant to conditions in which the system is completely empty that in this state we suppose T equal to 0. In this state, none of the base band units are in service and there is no any traffic in C-radio access network. This state can be defined as the initiation state of the system. So, we can define this state as the following:

$$S_1 = \{(i, j, k, m, n) \mid i = 0; j = 0; k = 0; m = 0; n = 0\}$$

Which all of the parameters are 0 and the equivalent equation is as Eq. (8).

$$(\lambda_1 + \lambda_2 + \lambda_3)\pi_{0,0,0,0,0} = \pi_{1,0,0,0,0} + \pi_{0,1,0,0,0} + \pi_{0,0,1,0,0} \quad (8)$$

$$\lambda_{S_2R_2} = \dots = \lambda_{S_NR_N} = \lambda/2 \text{ and } \lambda_{R_1D_1} = \lambda_{R_2D_2} = \dots = \lambda_{R_ND_N} = \lambda/2$$

$$\lambda_{R_0R_1} = \lambda_{R_1R_2} = \dots = \lambda_{R_NR_{N-1}} = \lambda, \lambda = 1, \text{ and } \gamma_{th}^j = \gamma_{th}^1, j \in [1, N].$$

The second state 2 (S2): in this condition, we have some requests in C-RAN with the amount of more than zero and less than a predefined threshold (C). In this state the system isn't face to any congestion and any delay for requests due to queueing. So, consequently the total accessible resource blocks are more than zero and less than the indicated fixed value (C) and all of the queues are completely empty. The equation of the second state can be demonstrated as:

$$S_2 = \{(i, j, k, m, n) \mid 0 < (i + j + k) < C; m = 0; n = 0\}$$

The more detail of Eq. (9) shown in the following.

$$(i\mu_1 + \lambda_1 + j\mu_2 + \lambda_2 + k\mu_3 + \lambda_3)\pi_{i,j,k,0,0} = \lambda_1\pi_{i-1,j,k,0,0} + (i+1)\mu_1\pi_{i+1,j,k,0,0} + \lambda_2\pi_{i,j-1,k,0,0} + (j+1)\mu_2\pi_{i,j+1,k,0,0} + \lambda_3\pi_{i,j,k-1,0,0} + (k+1)\mu_3\pi_{i,j,k+1,0,0} \quad (9)$$

The third state (S3): this steady state represents the situation that $i + j + k = 0$ and $m = n = 0$ in which all of the queues are fully empty but the maximum achievable resources applied in radio access network due to cover the requests. This situation can be formulized as Eq. (10).

$$(i\mu_1 + \alpha_1\lambda_1 + j\mu_2 + \lambda_2 + k\mu_3 + \lambda_3)\pi_{i,j,k,0,0} = \alpha_3\lambda_1\pi_{i-1,j,k,0,0} + (j\mu_2 + k\mu_3)\pi_{i,j,k,1,0} + \lambda_2\pi_{i,j-1,k,0,0} + k\mu_3\pi_{i,j,k,0,1} + \lambda_3\pi_{i,j,k-1,0,0} \quad (10)$$

The fourth state (S4): this state is dedicated to all of the remaining possible conditions so as to $0 < T > C, 0 < m \leq M, 0 < n \leq N$. In this state, most parts of the resource blocks are utilized and the requests execution is done based on the queueing process. So, we can formulate this state as Eq. (11).

$$\left(\begin{array}{l} \alpha_6 i \mu_1 + \alpha_1 \lambda_1 + \alpha_7 (1 - \alpha_6) i \mu_1 + \\ \alpha_6 j \mu_2 + \alpha_7 (1 - \alpha_6) j \mu_2 + \alpha_8 k \lambda_1 + \\ \alpha_6 k \mu_3 + \alpha_7 (1 - \alpha_6) k \mu_3 + \alpha_9 \lambda_3 \end{array} \right) \pi_{i,j,k,m,n} = \alpha_3 \lambda_1 \pi_{i-1,j+1,k,m-1,n} + \alpha_3 \lambda_1 \pi_{i-1,j+1,k,m,n-1} + \alpha_3 \lambda_1 \pi_{i,j,k,m-1,n} + \alpha_3 \lambda_1 \pi_{i,j,k,m,n-1} \quad (11)$$

Considering these four defined states and the T-matrix, the below condition should be quarantined.

$$\sum_{\forall S} \pi = 1$$

$$0 < \pi_{i,j,k,m,n} < 1$$

According to the described framework, overall normalized balancing formulation will be achievable through equality (12)

$$\pi Q = 0; \text{ and } \sum_{\forall S} \pi_{i,j,k,m,n=1} \quad (12)$$

In the assessment stage, in accordance with the defined probability matrix π we will be able to determine the performance assessment indexes in order to compare enhanced mobile broadband and massive and machine-type communications from the quality-of-service point of view. During the performance evaluation process, resource utilization process is considered in both enhanced mobile broadband and massive machine-type communications services as the following.

In the fully-utilized system which the resource is not enough for the congested network, enhanced mobile broadband should send a demand sequence to release some of the basebands' resource blocks and dedicate the released resources to the coming ultra-reliable and low-latency communications. In this condition, the forced termination probability which is represented as PF1 can be calculated as

$$P_{F_1} = \frac{\text{the average rate of terminated eMBB}}{\text{the eMBB admitted rate}}$$

Applying the balancing formulation, the number of enhanced mobile broadband resource blocks which should be released can be calculated by

$$\sum_{\forall S} \frac{\lambda_{1j}}{C-j} \pi_{i,j,k,m,n}$$

$$i+j+k=C \text{ and } k=0$$

The final enhanced mobile broadband rate is equal to $\lambda_2(1 - PB_2)$. Therefore, the forced termination probability obtains via Eqs. (13) and (14).

$$PF_1 = \sum_{\forall S} \frac{\lambda_{1j}}{(C-j)\lambda_2(1 - PB_2)} \pi_{i,j,k,m,n} \quad (13)$$

$$i+j+k=C \text{ and } k=0$$

$$PF_2 = \sum_{\substack{VS \\ i+j+k=C \text{ and } n=N}} \frac{\lambda_1 k}{(C-j)\lambda_3(1-PB_2)} \pi_{i,j,k,m,n} \quad (14)$$

The service completion index of enhanced mobile broadband and massive and machine-type communications which

$$L_2 = \sum_{0 \leq n \leq N} \sum_{VS} n \pi_{i,j,k,m,n} \text{ and } R_{\text{Interrupt}} = \sum_{\substack{VS \\ i+j+k=C \\ v < C}} \frac{\lambda_1 k}{C-v} \pi_{i,j,k,m,n}$$

can be calculated as final number of completed services is achievable through Eqs. (15) and (16) consequently.

$$SCR_1 = \sum_{VS} j \mu_2 \pi_{i,j,k,m,n} \quad (15)$$

$$SCR_2 = \sum_{VS} j \mu_3 \pi_{i,j,k,m,n} \quad (16)$$

With deploying the scenario, it's obvious that for each steady state (i, j, k, m, n) with C number of available discrete resource, $(i + j + k)$ base band units are utilized by ultra-reliable and low-latency communications, enhanced mobile broadband or massive machine-type communications services. So, the utilization rate of enhanced mobile broadband or massive machine-type communications demands can be calculated as:

$$U = \sum_{VS} \frac{i+j+k}{C} \pi_{i,j,k,m,n} \quad (17)$$

One of the measurement indexes we select to the assessment of our model is average transmission delay for both enhanced mobile broadband or massive machine-type communications services. As mentioned before, each demand has own distinct queue process based on its quality-of-service thresholds and having different delays for different demands is inevitable. In this scenario, we assume queue length of enhanced mobile broadband or massive machine-type communications as L_1 and L_2 and the mean value of enhanced mobile broadband service (D_1) and the mean value of massive machine-type communications (D_2) is achievable as:

$$D_1 = \frac{L_1}{\lambda_1} \quad (18)$$

which

$$L_1 = \sum_{\substack{VS \\ 0 \leq m \leq M}} m \pi_{i,j,k,m,n}$$

It should be noted that in calculation of delay for massive machine-type communications demands, all of the interrupts

must be calculated in addition to the usual transmission delay. So D_2 changes to:

$$D_2 = \frac{L_2}{\lambda_3 + R_{\text{Interrupt}}} \quad (19)$$

In which,

3 Hierarchical resource allocation

This part of the paper allocated to present the structure of the hierarchical framework called CDRA, in which we applied the same carrier allocation DQN unit based on the introduces framework in the former section to select the best carrier, and subsequently application of a DDPG unit to result in the transmission power of all UEs as shown in Fig. 3. First, we briefly introduce the framework of DDPG. It is not possible for us to use (19) formulation to select actions because the actions considered infinite and consequently Q will also be infinite. In order to find the optimal solution for this problem, the policy gradient approach directly formulates the rule π to determine actions, that makes policy gradient the most appropriate method for stablishing continuous reinforcement learning tasks. The formulated framework can be described as Eq. (20) but it should be noted that virtue of the $M/M/1$ queue, the mean-response time at server i is:

$$MRT_i(t) = \frac{1}{\mu_i - \beta_{k,m}(t)\lambda_i(t)},$$

where $\lambda_i(t)$ is the average arrival rate at server i . If the $\{p_i\}$ are constant or vary slowly over time, then $\lambda_i(t)$ can be approximated using $\lambda_i(t) = p_i(t)\lambda$, which is a consequence of the $M/M/1$ queue model (Zhai et al. 2018).

$$\pi_{\emptyset}(s, a) = P_r\{a|s;\emptyset\}, \quad (20)$$

In which \emptyset represents the policy index. This formulation represents the probability of selecting action a considering state s and \emptyset . To evaluate the appropriateness of policy π_{\emptyset} , we can utilize the expectation of the cumulative discounted reward to demonstrate the target, which is shown as Eq. (21)

$$J(\pi_{\emptyset}) = \int_S \rho^{\pi}(s) \int_A \pi_{\emptyset}(s, a) r(s, a) d_a d_s. \quad (21)$$

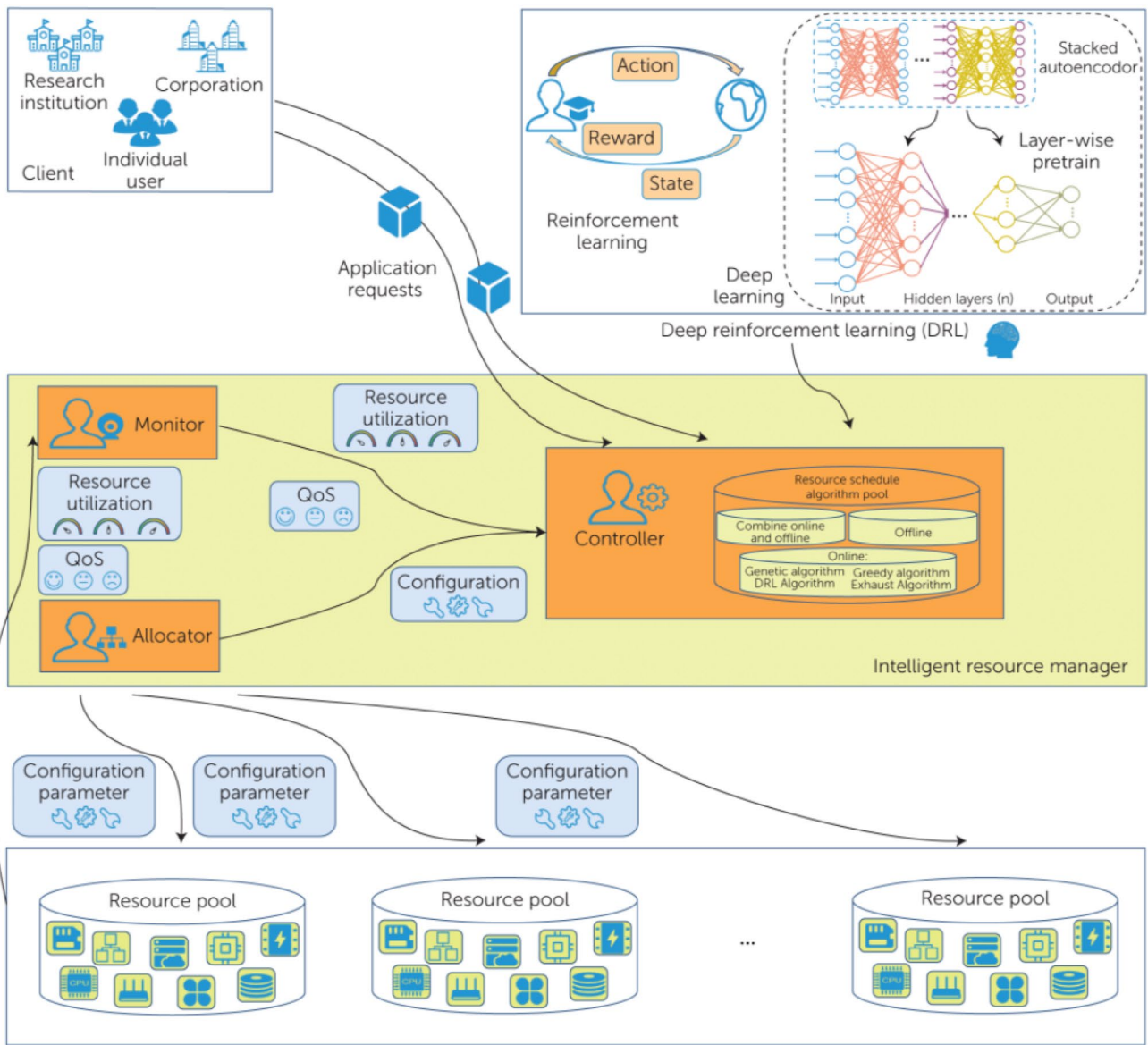


Fig. 3 Configuration of the dynamic DRL based resource allocation

In which, $\rho^\pi(s)$ is the distribution of states. Therefore, the goal of PG is to maximize (21) by adjusting the parameters \varnothing in the direction of the performance gradient defined as the following. To proceed with the formulation, let $\alpha(t)$ be the index of the chosen action at time instant t . Then, the value of $p_i(t)$ is updated as per the following simple rule (the rules for other values of $p_j(t)$, $j \neq i$, are analogous):

$$p_i(t + 1) \leftarrow p_i(t) + \theta(p_{max} - p_i(t))$$

when $\alpha(t) = i$ and $v_i = 1$

$$p_i(t + 1) \leftarrow p_i(t) + \theta(p_{min} - p_i(t))$$

when $\alpha(t) = j, j \neq i$ and $v_i = 1$,

where θ is a user-defined parameter $0 < \theta < 1$, typically close to zero. Further, v_i is a reward function indicator defined by:

- $v_i = 1$, reward, if the instantaneous response of the chosen server is under the running moving average of the mean response time, $s_i(t) \leq \frac{1}{r} \sum_{k=1}^r \hat{s}_k(t)$.

- $v_i = 0$, penalty, if the instantaneous response of the chosen server exceeds the running moving average of the mean response time, $\hat{s}_i(t) > \frac{1}{r} \sum_{k=1}^r \hat{s}_k(t)$.

$$\nabla_{\varnothing} J(\pi_{\varnothing}) = \int_S \beta_{k,m}(t) p^{\pi}(s) \int_{A \varnothing} \nabla \pi_{\varnothing}(s,a) Q^{\pi}(s,a) d_a d_s \tag{22}$$

and Eq. (9) is referred to as the policy gradient theorem (Zhang et al. 2020). However, two main drawbacks are existed in PG methods. First, PG algorithms typically converge to a local optimum rather than the global optimum. Besides, evaluating a policy π_{\varnothing} is generally inefficient and has high variance, because it requires integration over both state and action spaces (Weisz et al. 2018). To address these problems, the deterministic policy gradient theorem (Wu et al. 2020) is proposed, where an action is deterministically generated by a parameterized policy instead of the probability of actions (20). That is, $a = \pi_{\varnothing}(s)$, and Eq. (25) is rewritten as

$$J(\pi_{\varnothing}) = \int_S \beta_{k,m}(t) p^{\pi}(s) r(s, \pi_{\varnothing}(s)) ds, \tag{23}$$

and the corresponding gradient is

$$\begin{aligned} \nabla_{\varnothing}(s, a) &= \int_S \beta_{k,m}(t) p^{\pi}(s) \nabla_{\varnothing} \pi_{\varnothing}(s) \nabla_a Q^{\pi}(s, a) |_{a=\pi_{\varnothing}(s)} ds \\ &= E_{s \sim p^{\pi}} \left[\nabla_{\varnothing} \pi_{\varnothing}(s) \Delta_a Q^{\pi}(s, a) |_{a=\pi_{\varnothing}(s)} \right] + \beta_{k,m}(t) \end{aligned} \tag{24}$$

where the deterministic policy gradient (DPG) (28) is a special case of PG (26) when the variance of Eq. (24) reaches 0.

The delay for massive machine-type communications demands, all of the interrupts must be calculated in addition to the usual transmission delay. So D_2 changes to:

$$D_2 = \frac{L_2}{\lambda_3 + R_{Interrupt}}$$

In which,

$$\begin{aligned} L_2 &= \sum_{\forall s} n \pi_{i,j,k,m,n} \text{ and } R_{Interrupt} \\ &= \sum_{\substack{0 \leq n \leq N \\ i+j+k=C \\ v < C}} \frac{\lambda_1 k}{C-v} \pi_{i,j,k,m,n} \end{aligned}$$

Algorithm 1: Hierarchical Resource Allocation Algorithm

- 1: **Initialize** the replay coefficient λ with delay D .
- 2: **Initialize** the carrier allocation DQN unit $Q(s, a; \omega)$ with weights ω .
- 3: **Initialize** the resource allocation DDPG unit including the actor system $\pi(s; \mu)$ and the critic system $Q(s, a; \theta)$ with weights μ and θ .
- 4: **Initialize** the state $S1, S = \{(i, j, k, m, n) \mid i \geq 0, j \geq 0, i + j + k \leq C, 0 \leq m \leq M, 0 \leq n \leq N\}$.
- 5: **Initialize** the matrix T for DDPG unit, and the weights update process with interval size K .
- 6: The DRL controller at the base station receives the first channel information of all UEs as the initial state s_1 . Calculate PF1 as
$$\sum_{\substack{\forall s \\ i+j+k=C \text{ and } k=0}} \frac{\lambda_1 j}{C-j} \pi_{i,j,k,m,n}$$
- 7: **for** $t = 1, 2, \dots, T_{max}$ **do**
- 8: The carrier allocation DQN unit chooses action $a_t^1 \in \mathcal{A}_1$ following the ϵ -greedy policy.
- 9: Calculate of the forced termination probability obtains via
$$PF_1 = \sum_{\substack{\forall s \\ i+j+k=C \text{ and } k=0}} \frac{\lambda_1 j}{(C-j)\lambda_2(1-PB_2)} \pi_{i,j,k,m,n}$$

$$PF_2 = \sum_{\substack{\forall s \\ i+j+k=C \text{ and } n=N}} \frac{\lambda_1 k}{(C-j)\lambda_3(1-PB_2)} \pi_{i,j,k,m,n}$$
- 10: The resource allocation DDPG unit chooses the action $a_t^2 \in \mathcal{A}_2$ according to (29)
- 11: determination of Service Completion Index $SCR_i = \sum_{\forall s} j \mu_k \pi_{i,j,k,m,n}$
- 12: If μ_k can satisfy the data rate threshold, controller receives current energy efficiency as reward r_t . Otherwise, it doesn't receive any reward.
- 13: Considering length of the queues Class n
$$L_n \sum_{\substack{\forall s \\ 0 \leq m \leq M}} m \pi_{i,j,k,m,n}$$
- 14: Store tuple (s_t, a_t, r_t, s_{t+1}) in the experience replay memory \mathcal{D} .
- 15: Sample a random mini-batch of N tuples (s_i, a_i, r_i, s_{i+1}) from \mathcal{D} .
- 16: Final Transmission Delay $D_n = \frac{L_n}{\lambda_n + R_{Interrupt}}$
- 17: Calculate R Interrupt $R_{Interrupt} = \sum_{\substack{\forall s \\ i+j+k=C \\ v < C}} \frac{\lambda_1 k}{C-v} \pi_{i,j,k,m,n}$
- 18: Update goal Q system weights ω^- of the DQN unit by copying ω in every w time slots.
- 19: Update goal actor network weights θ^- and goal critic system weights μ^- of the DDPG unit based on (21)
- 20: **end for**

4 DRL-based load balancing model

In this section we try to apply an effective Load Balancing approach presented by H. Ismail and A. Yazidi which is based on two-time scale LA with barriers into the presented resource allocation approach to empower the suggested approach from the load distribution and load sharing point of views. So, we interpret the archived results of this combination approach after a brief presentation of their load balancing model.

A. Model

In this scenario, r servers considered each of which organized as an $M/M/1$ queue so that entrance probability is modeled as a poisson distribution statistics considering λ_i (Shu and Zhu 2020). The service rate is equal to μ_i and the service time has an exponential statistical distribution.

In this framework, the learning automata will be responsible for dispatching the demand. The learning automata submits the demand to server i which its probability is illustrated as $p_i(t)$. In the following of this section we will describe the update process of these learning automata formulations. We can define the update process considering the advantage of $M/M/1$ queue modelling. Based on this approach, the equation of mean-feedback time relevant to server i will be equal to:

$$MRT_i(t) = \frac{1}{\beta_{k,m}(t)\mu_i - \lambda_i(t)}, \tag{25}$$

In which, $\lambda_i(t)$ demonstrates the mean entrance rate of server i . it should be noted that if the $\{p_i\}$ considered as a fixed value or we can ignore its changes, we can consider $\lambda_i(t) = p_i(t)\lambda$, which is as a result of utilizing $M/M/1$ as a model for the queue (Zhai et al. 2018). We suppose that $s_i(t)$ shows the immediate response time at the exact time t at server i . considering α as the learning index, to calculate the mean feedback time of server i , we have to apply the *exponential – moving – average* method so that $\hat{s}_i(t)$ shows the approximate mean feedback time.

The estimation parameter $\hat{s}_i(t + 1)$ will be updated right after despatching a demand toward server i . this action is done utilizing a dynamic estimator called *exponential – moving – average* as the following:

$$\hat{s}_i(t + 1) = \hat{s}_i(t) + \alpha(s_i(t) - \hat{s}_i(t)). \tag{26}$$

The mean feedback time for each other servers remain without change during this process. Here,

$$\hat{s}_i(t + 1) = \hat{s}_i(t) \text{ for } j \neq i, j \in [1, n]. \tag{27}$$

The process of penalty and reward determination for each action is clear. If action i is selected, we have the following instructions relevant to penalty/reward process.

- Reward if $\hat{s}_i \leq \frac{1}{r} \sum_{k=1}^r \hat{s}_k$.
- Penalty if $\hat{s}_i > \frac{1}{r} \sum_{k=1}^r \hat{s}_k$.

Based on this framework, as a back-drop, the proposed approach can be described step-by-step as algorithm 2.

B. Two-time scale learning automata

The beginning stage of the preposed approach is related to modification of Markov process from the

absorbing form to the ergodic form. Instead of applying the actual constraints of the probability space equal to 0 or 1, we assume that none of the probability factors can be less than a predefined lower bound threshold which we indicated it by p_{min} or more than a upper bound threshold denoted as p_{max} (Colonius and Rasmussen 2021).

After definition of these two thresholds, the action-selection stochastic parameter in the learning resolution method will be changed and the optimal value in this scheme will be moving to the determind levels p_{min} and p_{max} and such a little adjustments can make the framework ergodic and the results will also be completely different compared to other current schemes.

To obtain such goal, we try to define a minimum level p_{min} as the probability factor, in which $0 < p_{min} < 1$ for all selection probability x_i , $1 \leq i \leq r$ and r illustrates the quantity of actions. The optimum value for any selection probability p_i , $1 \leq i \leq r$, is obtained as $p_{max} = 1 - (r - 1)p_{min}$ which it can be considered as a primary target. It will be happened if the other $r - 1$ actions are equal to their minimal degree p_{min} , nevertheless, the action with the highest probability will be equal to p_{max} . Subsequently, p_i , for $1 \leq i \leq r$, will take values more than p_{min} and less than p_{max} .

In this framework, $\alpha(t)$ denotes the parameter of selected action at relevant to time t . To proceed with the formulation, let $\alpha(t)$ be the index of the chosen action at time instant t . Hence, $p_i(t)$ will be renewed based on the following formulation (the relevant rules for other degrees of $p_j(t)$, $j \neq i$, will be similar):

$$p_i(t + 1) \leftarrow p_i(t) + \theta(p_{max} - p_i(t))$$

when $\alpha(t) = i$ and $v_i = 1$

$$p_i(t + 1) \leftarrow p_i(t) + \theta(p_{min} - p_i(t))$$

when $\alpha(t) = j, j \neq i$ and $v_i = 1$,

In this formulation, θ demonstrates user level variable $0 < \theta < 1$, which is normally close to 0. Also, v_i indicates a reward function described via:

- $v_i = 1$, known as the reward, if the imediate feedback of the selected server is less than the *moving – average* of the average feedback time $s_i(t) \leq \frac{1}{r} \sum_{k=1}^r \hat{s}_k(t)$.
- $v_i = 0$, is known as the penalty, if the imediate feedback of the selected server more than the *moving – average* of the average feedback time, $\hat{s}_i(t) > \frac{1}{r} \sum_{k=1}^r \hat{s}_k(t)$.

This procedure has been summarized in Algorithm 2. As per our description, $\hat{s}(t)$ indicates the mean immedi-

ate feedback times relevant to all the entities during time duration t is exhibited as the following equation:

$$\hat{s}(t) \leq \frac{1}{r} \sum_{k=1}^r \hat{s}_k(t) \tag{28}$$

Algorithm 2: Load Balancing with the hierarchical resource allocation

Loop

1: Run the discrete resource allocation algorithm. The DRL controller at the base station receives the first carrier information of all UEs as the initial state s_1 . compute PF1 as

$$\sum_{\substack{vs \\ i+j+k=C \text{ and } k=0}} \frac{\lambda_1 j}{C-j} \pi_{i,j,k,m,n}$$

2: Poll an action in time duration $[0, t]$ in accordance with the stochastic vector, $[p_1, p_2, \dots, p_r]$.

3: determination of Service Completion Index $SCR_t = \sum_{vs} j \mu_k \pi_{i,j,k,m,n}$

4: Updating the response time estimates based on D_t .

5: Determination of forced termination probability indexes

6: Calculate R Interrupt

$$R_{Interrupt} = \sum_{\substack{vs \\ i+j+k=C \\ v < C}} \frac{\lambda_1 k}{C-v} \pi_{i,j,k,m,n}$$

7: Final Transmission Delay $D_n = \frac{L_n}{\lambda_n + R_{Interrupt}}$

- Update the feedback time relevant to the selected action:

$$\hat{s}_i(t+1) = \hat{s}_i(t) + \alpha (s_i(t) - \hat{s}_i(t)).$$

- Considering queues' length L_i The estimation of feedback for the actions $j, j \neq i$ will be without change, and hence,

$$L_n \sum_{\substack{vs \\ 0 \leq m \leq M}} m \pi_{i,j,k,m,n}$$

$$\hat{s}_j(t+1) = \hat{s}_j(t) \text{ for } j \neq i, j \in [1, r].$$

8: Response Determination: Reward/Penalty.

$$v_i = 1: (\text{Reward}) \text{ if: } \hat{s}_i \leq \frac{1}{r} \sum_{k=1}^r \hat{s}_k, \\ \text{Otherwise, } v_i = 0 (\text{Penalty}),$$

9: Let $\alpha(t)$ as the variable of the selected action during the time duration $[0, t]$. $p_i(t)$ will be renewed based on the following rule:

$$p_i(t+1) \leftarrow p_i(t) + \theta(p_{max} - p_i(t)) \\ \text{when } \alpha(t) = i \text{ and } v_i = 1,$$

$$p_i(t+1) \leftarrow p_i(t) + \theta(p_{min} - p_i(t)) \\ \text{when } \alpha(t) = j, j \neq i \text{ and } v_i = 1,$$

We explained all aspects of the proposed scheme so we should evaluate the performance of this approach compared to the other state of the art schemes during the next part. In the assessment stage, it will be exhibited that as $p_i(t)$ enhances, $Prob(s_i(t) > \hat{s}(t))$, decreases which this results can be very exciting.

5 Simulation results

In this part of the paper, we tried to evaluate the performance of the presented model compared to some other schemes in a same and identical scenario. The results are simulated in an

uplink multi-user NOMA system, where the BS is located at the center of the cell and four users are randomly distributed in the cell with a radius of 400 m. The path loss model is $114 + 38 \lg(d) + 9.2$, where $d(\text{km})$ is the distance between users and BS. The minimum data rate requirement $R_{min} = 15$ bps/Hz, $\sigma_z^2 = -174$ dBm and C is set to be 4. If not specified, parameters are set as follows. $P_{max} = 5$ W, $P_0 = 25$ W, $T_{max} = 1000$ and all users perform random moving with a speed of $v = 2.5$ m/s. For the DDRA framework, each network of all DQN units has three consecutive layers with 64 neurons per layer. The learning rate is 0.015, $\gamma = 0.98$, $\epsilon = 0.9$, memory capacity $D = 450$, weights update interval $W = 50$ and batch size $N = 64$. For the CDRA framework, the subchannel assignment DQN unit is the same as the one in the DDRA framework.

In this scenario we consider a single C- radio access network with 45 accessible base band units (i.e., $C = 45$). We also have assumed three distinct quality of service class for different arrival rate of services which all of the service streams are categorized in one of the following cases:

- Ultra-reliable and low-latency communications.
- Enhanced mobile broadband.
- Massive machine-type communications.

In which we have $\lambda_1 = 3, \lambda_2 = 3, \lambda_3 = 6$ for the three services respectively. And data rate relevant to each service is $\mu_1 = 0.8, \mu_2 = 1.0, \text{ and } \mu_3 = 0.7$. The queue length for two distinct applied queues is 10 ($L = 6$) and 2 ($L = 2$) respectively.

In comparison with DL-based approaches in deep learning for effective non-orthogonal multiple access networks, we have compared the performance of the proposed scheme with (Liu et al. 2017; Gui et al. 2018) which Liu et al. (2017) applied a long-short-term memory (LSTM) to detect the channel characteristics automatically. As we mentioned before, evaluation of the resource allocation approach is done both individually (CDRA) and combined with the introduced load balancing approach (CDRA + LB). We can also consider LSTM individually or in combination with RTN method.

In the first round of the assessment, we have deployed three different algorithms in a same scenario in which the assessment index was rate of resource utilization per service rae. As shown in Fig. 4, the combination of CDRA with the suggested load balancing model achieved the best result with the significant difference to other two schemes. In this scenario the effectiveness of the stochastic load balancing method is completely obvious. It has been clear that without empowerment of CDRA, this approach has no any excellence compared to LSTM.

Fig. 4 Resource utilization of ultra-reliable and low-latency communications

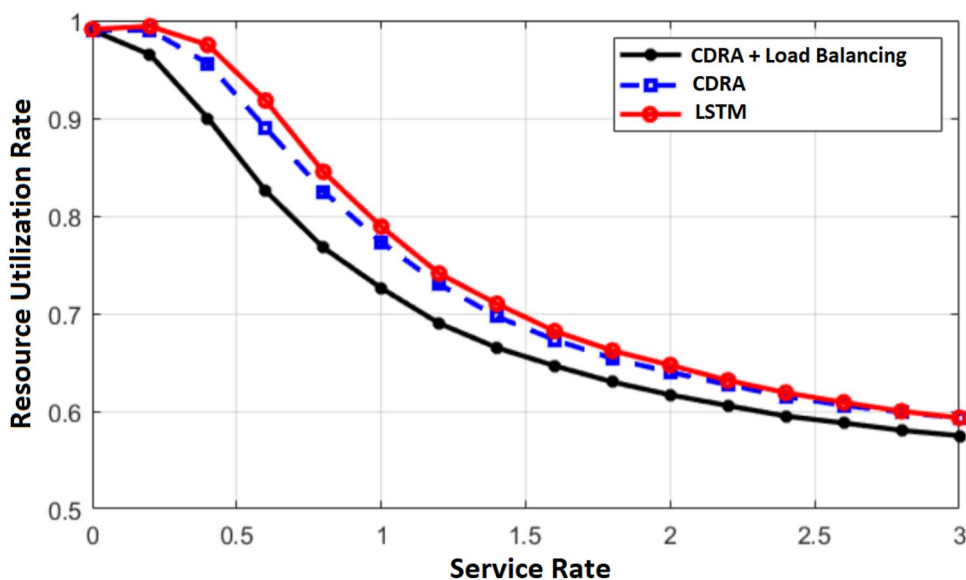
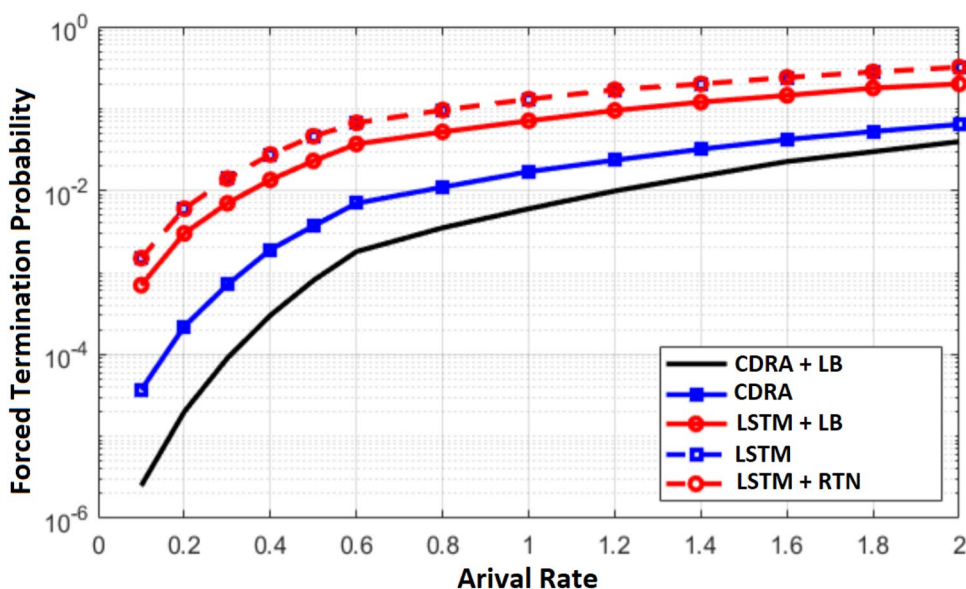


Fig. 5 Forced termination probability per arrival rate for massive machine-type communications



In Fig. 5, the evaluation index is the amount of forced termination probability versus arrival rate. In this scenario we considered five different algorithms in which CDRA and LSTM are the main algorithms with/without additional methods like as load balancing and RTN. The major target of this plot is to interpret the functionality of each algorithm, when the arrival service rate of massive machine-type communications request with the best QoS class increases. As we mentioned before, in such scenario, evaluation of the queues' length or queue delays cannot be enough individually and the more effective index which should be investigated in force terminated probability of each services. Based on the obtained results, it is obvious that the combination of CDRA is chosen as the best solution for having least force termination probability. The performance of CDRA

can also be better when this algorithm empowered by the load balancing approach.

Figure 6 demonstrated the queuing delay of massive machine-type communications service for various approaches. It shown that CDRA outperforms that of LSTM generally although in some cases when the service arrival rate increases, the combination of LSTM and RTN has excellence compared to CDRA. It's why that when any interrupted requests relevant to the massive machine-type communications join back to its queue, we will face to queue length increasing and in this condition increasing the queuing delay is inevitable.

Actually there is a tradeoff between the average queuing delay and the force termination probability but the effect of using the load balancing approach can completely change

Fig. 6 Average queue delay of enhanced mobile broadband arrival rate for different approaches

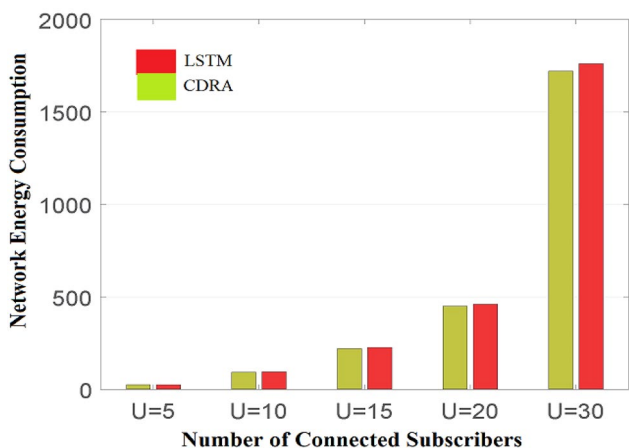
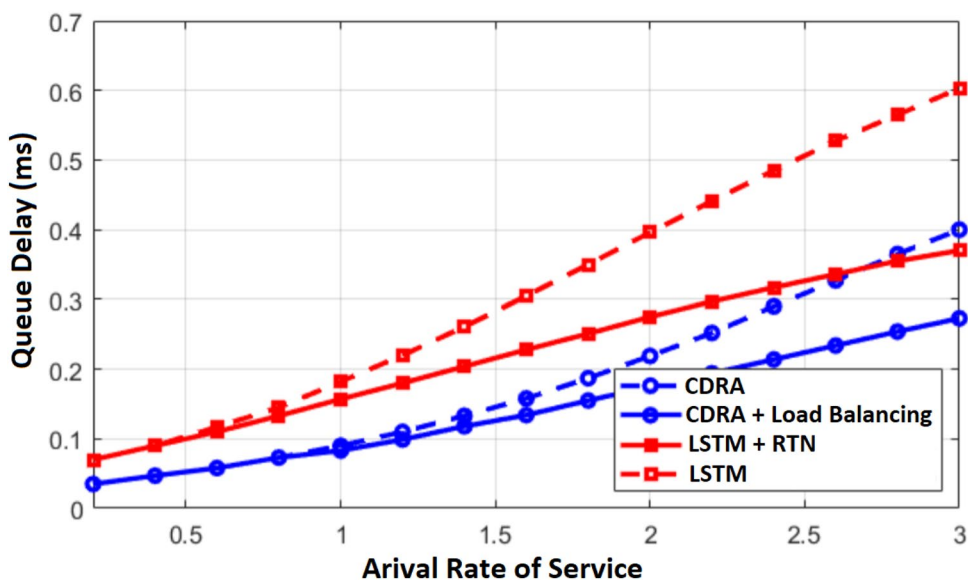


Fig. 7 Total network energy consumption (mj) versus the number of connected users (#)

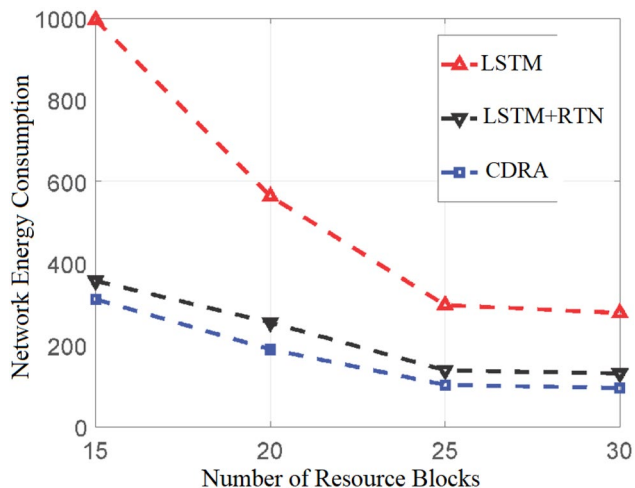


Fig. 8 Total energy consumption (mj) versus the number of resource block (#)

this process. However, the best solution for this problem is finding the optimal queue length which be able to balance this tradeoff’s parameters.

In Fig. 7, the performance of the proposed CDRA approach and LSTM have been compared from the perspective of total consumption energy for different number of connected subscribers. As it is obvious in this figure, the total consumed energy of LSTM is completely close to CDRA.

Figure 8 demonstrates the effect of number of resource block on the energy consumption. This result proves that we can decrease the consumed energy by increasing the number of resource blockes. It has been also exhibited that CDRA

has better performance than other two DL-based NOMA resource allocation approaches.

6 Conclusion

This paper investigates the joint user association and resource allocation problem in an uplink multicast NOMA system to maximize the power efficiency with guaranteeing the quality-of-experience of all subscribers. We also introduced an adaptive load balancing approach that aspires to obtain “almost optimal” fairness among servers from the quality of service (QoS) perspective in which learning

automata (LA) has been used to find the optimal solution for this dynamic problem. Different from conventional model-based resource allocation methods, this paper suggested a hierarchical reinforcement learning-based framework to solve this non-convex and dynamic power optimization problem. The presented learning automata find the ϵ -optimal solution for the problem by resorting to a two-time scale-based SLA paradigm. Numerical results show that the suggested hierarchical resource allocation framework in combination with the load balancing approach, can significantly improve the energy efficiency of the whole NOMA system compared with other approaches. For the DDRA framework, a discrete and distributed multi-DQN strategy is proposed to reduce the output dimension. And for the CDRA framework, we combine the advantages of DQN and DDPG with the load balancing approach to design a joint resource allocation network. Both frameworks are trained jointly to find the optimal subchannel assignment and power allocation policy through constant interaction with the NOMA system. Compared with other alternative approaches, this framework is able to provide better EE under different transmit power limitations, and are applicable in various moving speed conditions by adjusting the parameters of networks, which proves the effectiveness of our proposed frameworks.

References

- Abedin SF, Alam MGR, Kazmi SA, Tran NH, Niyato D, Hong CS (2018) Resource allocation for ultra-reliable and enhanced mobile broadband IoT applications in fog network. *IEEE Trans Commun* 67(1):489–502
- Baidas MW, Alsusa E, Hamdi KA (2019) Joint relay selection and energy-efficient power allocation strategies in energy-harvesting cooperative NOMA networks. *Trans Emerg Telecommun Technol* 30(7):e3593
- Cao X, Ma R, Liu L, Shi H, Cheng Y, Sun C (2018) A machine learning-based algorithm for joint scheduling and power control in wireless networks. *IEEE Internet Things J* 5(6):4308–4318
- Celdrán AH, Pérez MG, Clemente FJG, Pérez GM (2019) Towards the autonomous provision of self-protection capabilities in 5G networks. *J Ambient Intell Humaniz Comput* 10(12):4707–4720
- Colonius F, Rasmussen M (2021) Quasi-ergodic limits for finite absorbing Markov chains. *Linear Algebra Appl* 609:253–288
- Fu S, Fang F, Zhao L, Ding Z, Jian X (2019) Joint transmission scheduling and power allocation in non-orthogonal multiple access. *IEEE Trans Commun* 67(11):8137–8150
- Gui G, Huang H, Song Y, Sari H (2018) Deep learning for an effective nonorthogonal multiple access scheme. *IEEE Trans Veh Technol* 67(9):8440–8450
- Huang H, Peng Y, Yang J, Xia W, Gui G (2019) Fast beamforming design via deep learning. *IEEE Trans Veh Technol* 69(1):1065–1069
- Khan WU, Liu J, Jameel F, Sharma V, Jantti R, Han Z (2020) Spectral efficiency optimization for next generation NOMA-enabled IoT networks. *IEEE Trans Veh Technol* 69(12):15284–15297
- Li J, Bao X, Zhang W, Bao N (2020) QoE probability coverage model of indoor visible light communication network. *IEEE Access* 8:45390–45399
- Liu C, Jin Z, Gu J, Qiu C (2017) Short-term load forecasting using a long short-term memory network. In: 2017 IEEE PES innovative smart grid technologies conference Europe (ISGT-Europe), pp 1–6
- Liu M, Song T, Gui G (2018) Deep cognitive perspective: resource allocation for NOMA-based heterogeneous IoT with imperfect SIC. *IEEE Internet Things J* 6(2):2885–2894
- Liu P, Li Y, Cheng W, Zhang W, Gao X (2019a) Energy-efficient power allocation for millimeter wave beamspace MIMO-NOMA systems. *IEEE Access* 7:114582–114592
- Liu Y, Yu H, Xie S, Zhang Y (2019b) Deep reinforcement learning for offloading and resource allocation in vehicle edge computing and networks. *IEEE Trans Veh Technol* 68(11):11158–11168
- Maimó LF, Celdrán AH, Pérez MG, Clemente FJG, Pérez GM (2019) Dynamic management of a deep learning-based anomaly detection system for 5G networks. *J Ambient Intell Humaniz Comput* 10(8):3083–3097
- Pan H, Liew SC, Liang J, Shao Y, Lu L (2018) Network-coded multiple access on unmanned aerial vehicle. *IEEE J Sel Areas Commun* 36(9):2071–2086
- Praveenchandar J, Tamilarasi A (2020) Dynamic resource allocation with optimized task scheduling and improved power management in cloud computing. *J Ambient Intell Humaniz Comput*. <https://doi.org/10.1007/s12652-020-01794-6>
- Qiu C, Hu Y, Chen Y, Zeng B (2019) Deep deterministic policy gradient (DDPG)-based energy harvesting wireless communications. *IEEE Internet Things J* 6(5):8577–8588
- Shu Y, Zhu F (2020) An edge computing offloading mechanism for mobile peer sensing and network load weak balancing in 5G network. *J Ambient Intell Humaniz Comput* 11(2):503–510
- Song Z, Ni Q, Sun X (2018) Spectrum and energy efficient resource allocation with QoS requirements for hybrid MC-NOMA 5G systems. *IEEE Access* 6:37055–37069
- Teng Z, Zhang B, Fan J (2020) Three-step action search networks with deep q-learning for real-time object tracking. *Pattern Recogn* 101:107188
- Verhelst M, Moons B (2017) Embedded deep neural network processing: algorithmic and processor techniques bring deep learning to IOT and edge devices. *IEEE Solid State Circuits Mag* 9(4):55–65
- Wei Z, Zhao L, Guo J, Ng DWK, Yuan J (2018) Multi-beam NOMA for hybrid mmWave systems. *IEEE Trans Commun* 67(2):1705–1719
- Weisz G, Budzianowski P, Su PH, Gašić M (2018) Sample efficient deep reinforcement learning for dialogue systems with large action spaces. *IEEE ACM Trans Audio Speech Lang Process* 26(11):2083–2097
- Wu D, Dong X, Shen J, Hoi SC (2020) Reducing estimation bias via triplet-average deep deterministic policy gradient. *IEEE Trans Neural Netw Learn Syst* 31(11):4933–4945
- Xie D, Zhong X (2020) Semicentralized deep deterministic policy gradient in cooperative StarCraft games. *IEEE Trans Neural Netw Learn Syst*. <https://doi.org/10.1109/TNNLS.2020.3042943>
- Xu X, Zuo L, Li X, Qian L, Ren J, Sun Z (2018) A reinforcement learning approach to autonomous decision making of intelligent vehicles on highways. *IEEE Trans Syst Man Cybern Syst* 50(10):3884–3897
- Ye H, Li GY (2018) Deep reinforcement learning for resource allocation in V2V communications. In: 2018 IEEE international conference on communications (ICC), pp 1–6
- Ye H, Li GY, Juang BHF (2019) Deep reinforcement learning based resource allocation for V2V communications. *IEEE Trans Veh Technol* 68(4):3163–3173
- Zeng M, Yadav A, Dobre OA, Tsiropoulos GI, Poor HV (2017) Capacity comparison between MIMO-NOMA and MIMO-OMA with multiple users in a cluster. *IEEE J Sel Areas Commun* 35(10):2413–2424

- Zeng M, Hao W, Dobre OA, Poor HV (2019) Energy-efficient power allocation in uplink mmWave massive MIMO with NOMA. *IEEE Trans Veh Technol* 68(3):3000–3004
- Zhai D, Zhang R, Cai L, Li B, Jiang Y (2018) Energy-efficient user scheduling and power allocation for NOMA-based wireless networks with massive IoT devices. *IEEE Internet Things J* 5(3):1857–1868
- Zhang K, Koppel A, Zhu H, Basar T (2020) Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM J Control Optim* 58(6):3586–3612
- Zhao N, Liang YC, Niyato D, Pei Y, Wu M, Jiang Y (2019) Deep reinforcement learning for user association and resource allocation in heterogeneous cellular networks. *IEEE Trans Wirel Commun* 18(11):5141–5152

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.