Multiobjective Reinforcement Learning Based Energy Consumption in C-RAN Enabled Massive MIMO

Shruti SHARMA, Wonsik YOON

Dept. of Electrical and Computer Engineering, Ajou University, Suwon, South Korea

{wsyoon, shruti}@ajou.ac.kr

Submitted March 26, 2021 / Accepted, July 26, 2021

Abstract. *Multiobjective optimization has become a suitable method to resolve conflicting objectives and enhance the performance evaluation of wireless networks. In this study, we consider a multiobjective reinforcement learning (MORL) approach for the resource allocation and energy consumption in C-RANs. We propose the MORL method with two conflicting objectives. Herein, we define the state and action spaces, and reward for the MORL agent. Furthermore, we develop a Q-learning algorithm that controls the ON-OFF action of remote radio heads (RRHs) depending on the position and nearby users with goal of selecting the best single policy that optimizes the trade-off between EE and QoS. We analyze the performance of our Q-learning algorithm by comparing it with simple ON-OFF scheme and heuristic algorithm. The simulation results demonstrated that normalized ECs of simple ON-OFF, heuristic and Q-learning algorithm were 0.99, 0.85, and 0.8, respectively. Our proposed MORL-based Q-learning algorithm achieves superior EE performance compared with simple ON-OFF scheme and heuristic algorithms.*

Keywords

Convergence, energy consumption, reinforcement learning, reward, optimization

1. Introduction

Energy consumption (EC) in communication networks is a serious concern for wireless devices because they are mostly battery operated. As their batteries have limited capacities, the lifetimes of such networks are limited. Therefore, a considerable amount of attention is being given to energy harvesting in current generation of wireless networks. The introduction of multi-input multi- output (MIMO) technology in such networks provides superior spectral energy (SE) and energy efficiency (EE). However, there is a gradual increase in the EC of massive MIMO networks owing to associated massive antenna arrays [1]. To overcome these issues, heterogeneous cloud-radio access networks (C-RANs) have generated significant interest in communication technology, where base-band units (BBUs) and remote radio heads (RRHs) are parted [2], [3]. Here, processing units are relocated into a centralized BBU in the form of a cloud and distributed RRHs handle the radio signals received from users to BBUs through radio links. In a heterogeneous C–RAN, the inter-tier interference is mitigated to improve SE and EE. In addition, coupling MIMO networks with C-RAN architecture is an appealing solution to achieve energy savings while maintaining the QoS requirements. This lowers the overall manufacturing and operational cost for possible large-scale high-density network deployments because these networks cannot solely achieve fifth-generation (5G) targets [3], [4]. Using this centralized MIMO enabled C-RAN strategy, it is easy to analyze and calculate the statistical data, as it encourages the use of independent schemes for network energy management. Numerous works have explored resource utilization in wireless networks using multi-consisting of multivariable environment with conflicting objectives.

In wireless communication networks, resource allocation involves multiobjective optimization to achieve proper utilization of limited physical resources (e.g., power, energy, user scheduling, QoS, EE, SE, etc.) and enhance system performance [5]. Specifically, reinforcement learning (RL) is a valuable technique in ML where the agent learns a good resource management policy from the environment, and rapid decisions are made once the policy is learned [6], [7]. Therefore, to overcome the aforementioned challenges, ML techniques including model-free RL and neural networks could be employed. To collect information, the agent must completely explore the environment to select the best action. Gradually, the agent approaches and maps the behavior of the observed environment. In most RL techniques, a single agent is used [7], [8]. However, the wireless network environment is complex for an RL agent. Particularly, in a multiobjective environment, an agent hardly fulfills the requirement of analyzing multiple decision making process. The popular Markov decision processes (MDPs) of sequential decision-making problems in RL have received widespread attention for resource allocation and application predictions in several wireless communication networks [9].

DOI: 10.13164/re.2022.0155

When the system dynamics are unidentified and system observations are not precise, RL methods can be employed to study the system dynamics and identify the unknown environment. By adapting and responding to these dynamic changes, RL methods show significant promise for MIMO enabled C-RAN networks [9], [10]. RL techniques are best to switch RRHs on or off at defined time steps to achieve low EC in RRHs and satisfy the user QoS requirements under varying traffic demand and network densities. This study focuses on RL methods to solve conflicting objectives problems to maximize EE and reduce network EC [10]. Compared to traditional RL methods, MORL considers optimization of more than one objective simultaneously by the learning agent. Traditional RL methods use various iterative algorithms and greedy search algorithms to determine the system requirements. In contrast, MORL agents examine the whole network environment with respect to each and every possible state [11]. In this work, we have used multiple rewards for dissimilar objectives, and a trade-off among conflicting objectives based on their significance and optimization methods can be achieved. We have used sum total reward function to optimize the combined objective problems as a weighted reward sum.

2. Related Work and Motivation

Power control methods have been studied to mitigate EC and maximize EE. There are several studies on improving EE in wireless communication systems. Several authors have extensively focused on energy-efficient wireless communications, and accepted that EE is a key challenge of 5G networks [3]. In addition, joint rate allocation, scheduling, power control and channel assignment problems were studied with the aim of maximizing throughput [12]. Shi et al. investigated RRH selection and resource allocation with power minimization and sparse beamforming for C-RANs [13]. The effect of optimizing compression on EE was studied in a C-RAN [2, 3, 9]. The authors minimized the total EC in a wireless network and showed that a higher EE depends on the user sum rate. To improve the EE of the network, the authors minimized the inter-cell or inter-tier interference. Luo et al. jointly investigated uplink and downlink mobile users for a C-RAN and proposed an algorithm for access point association, beamforming strategy to minimize interference, and EC [5]. The authors suggested that an improved scheme for the optimization of EE by joint power allocation and resource block assignment for mitigation of interference in heterogeneous C-RAN [3]. The authors proposed a multi-objective evolutionary algorithm for power assignment to control the transmission power of sensor nodes in a wireless sensor network. Through a multi-objective two-nested genetic algorithm in [9], the authors clustered a homogeneous wireless sensor network.

Joint optimization through cell activation, user association, with sub-carrier allocation was studied under the constraints of maintaining EE and QoS [14], [15]. Recently, researchers used MORL methods. Natarajan et al. studied an alternative approach called linear secularization method, where a user specifies the weight of objectives [16]. A single objective reward can be expressed as the calculated weighted sum of each objective rewards. The developments in wireless communication network design have increased the system capacity. Keeping this in mind, the following two aspects motivate C-RAN enabled MIMO implementations. As the EC of a wireless system is reduced, the operational expenses for the wireless network reduce. These reports led to the introduction of "Green Radio," dedicated to creating novel methods to limit the EC of wireless networks, specifically improving the operational design of base stations.

3. MORL Agent Environment

We transform multiobjective function into a single objective [4], specifically, through the weighted-sum scalarization [12], [17]. On-line MORL processes are based on linear scalarization methods with Q-learning. In MORL, Q-values are changed to Q-vectors to categorize the Pareto front of policies [18], [19]. MORL has different instructions for Q-vectors and the next action, compared to a single objective MDP. For example, reward vectors are scalarized, Q-vectors are updated in each step, and the next action is selected from a list generated using greedy exploration-based methods [12]. The advantage of using linear scalarization functions is the ease in showing the convergence to the true Pareto front.

Within this framework, we consider the EC and user association problem in a massive MIMO empowered C-RAN architecture based on MORL. The agent can be gradually trained and kept informed of the learned data to know the state of each RRHs to implement continuous control. The main aim is to improve the resource allocation scheme, which reduces the power consumption and maintains the QoS requirement. The proposed framework is divided into two steps: First, we identify the active and inactive RRHs using the MORL algorithm. Second, based on the active RRHs set, we develop resource allocation module by optimizing power. In the "on-line training" framework, the agent progressively learns the network environment in real time with each episode.

Based on related studies [8], where EC is optimized over the current frame, we present cell activation MORL-based approach that makes a sequential resource allocation decisions to minimize the total EC during the operational period. To solve this complex problem, we first used a Q-learning method to simplify the user resource allocation problem as a convex optimization. EC, user satisfaction, and delay are very important constraints that affect the EE and QoS requirements of wireless networks. This study examines the trade-off between the EC and the delay using a MORL technique model. We determine the solution by showing the EE-QoS trade-offs using the Pareto optimality [19].

In a massive MIMO enabled C-RAN, users are randomly uniform distributed in the cell and have different access positions to RRHs. Therefore, some RRHs are significant to users, whereas others are not. Accordingly, for each user, instead of being equally served by all RRHs, only a few nearby RRHs must be chosen to serve the user. Moreover, if all RRHs are participating in the communication, the power consumed by the network and antennas could be large, which subsequently degrades the EE [4]. Based on these facts, selection of RRHs for an individual user becomes complex, considering the complexity of user association for the network. Various user association algorithms consisting of different performance metrics have been formulated using MORL [20], [21]. In this work, EE-based user association problem is solved by minimizing the EC while satisfying the QoS requirements and maximizing the EE. We aim to balance the EC and user satisfaction to satisfy the requirements of infrastructure providers via flexible power and bandwidth control or allocation, which is decoupled from cell activation techniques. The main contribution of this study based on MORL, where agents focus on resource management problems in C-RANs, is that we propose a novel MORL approach for energy management of C-RANs with two objectives: minimize EC of the network and maximize user throughput so that QoS is satisfied. In MORL approach, an action in the environment leads to multiple rewards.

- Agents perform on-line training. A policy of channels for C-RAN groups is defined that maximizes user throughput performance.
- A simultaneous multiple objective networks to maximize user throughput and energy savings is defined.
- Rewards that maximize user throughput performance with respect to energy savings are characterized.

4. System Model

Here, we briefly describe the architecture of massive MIMO enabled C-RANs, and the motivation for employing a MORL method in this network. Here, we adopt a switching OFF and ON strategy for RRHs to reduce the power consumption as shown in Fig. 1.

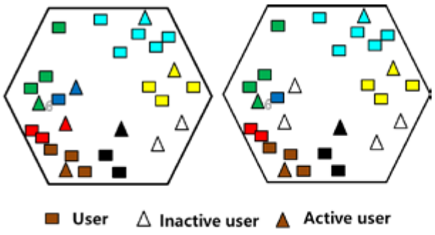


Fig. 1. MIMO enabled C-RAN architecture. Illustration of RRHs in active and sleep mode.

The idea of RRH on-off switching in applications is simple and can be summarized as follows.

Depending on users' positions, the number of active RRHs is minimized, whereas the other active RRHs present the radio coverage (required QoS) for users within the network.

The architecture of a cell in the C-RAN comprises of a set of N RRHs = {1, 2… N} and a set of K users = {1, 2… K}. The network reports the associated user requirements and their present states of RRH (in active or sleep mode). We suppose that all RRHs and users have a single antenna connected to BBU pool. Therefore, the received signal of the user k can be expressed according to the beamforming model [2],

$$y_{kn} = \sqrt{p_{kn}}\mathbf{g}_{kn}^{\mathrm{H}} s_{kn} + \sum_{i=1}^{K} q_{kn} s_{in} + \mathbf{z}_0 \tag{1}$$

where $\mathbf{g}_{kn} = \mathbf{H}_{kn}\mathbf{I}_{kn}$ is the channel gain between the nth RRH and the kth user and $\mathbf{z}_0$ is the receiver noise power. The p_{kn} represents the transmitted signal from the kth user and the nth RRH, s_{in} is the transmit power sent from the ith user, H_{kn} and $\mathbf{I}_{kn}$ are the fast fading coefficient and large fading coefficient between the nth RRH and the kth user respectively, and $(.)^{\mathrm{H}}$ denotes the transpose matrix. For the resource allocation, the signal-to-interference-plus-noise-ratio given by each user K associated with an RRH N is given as follows:

$$\gamma_{kn} = \frac{\mathbf{g}_{kn} P_{kn}^{\mathrm{trans}}}{\sum\limits_{j \neq n} \mathbf{g}_{kn} P_{jn}^{\mathrm{trans}} + \sigma}. \tag{2}$$

We define two power consumption states for each RRH, sleep and active states. The active state combines the power consumption of transmit power and RRH power. Power consumed by RRHs in the sleep state is almost zero. Therefore, the total power model for each RRH is given by:

$$P_k^{\mathrm{total}} = \begin{cases} P_k^{\mathrm{active}} + P_k^{\mathrm{trans}} \\ P_k^{\mathrm{sleep}} \end{cases}. \tag{3}$$

where P_k^{active} denotes the essential power consumed by the nth RRH in the active state needed for performing the basic RRH operations. Here, P_k^{active} is the used transmit power of the RRH, which ensures the data transmission of the user. In case the RRH is not used for transmission, it enters the sleep mode. Given at time t = {1, 2, 3, … T}, a set of active RRHs and a set of sleep mode RRHs, the total EC of RRHs in a complete period is expressed as:

$$E = \sum_{t=1}^{T} \left(\sum\nolimits_{n \in N_A} P_n^{\mathrm{active}} + \sum\nolimits_{n \in N_A} P_n^{\mathrm{trans}} + \sum\nolimits_{n \in N_S} P_n^{\mathrm{sleep}} \right). \tag{4}$$

In this architecture, we place the inactive RRHs into sleep modes to save power. The proposed Q-learning-based framework provides the control for managing energy consumption.

5. MORL Formulation

The MORL problem is a collection of multiple criteria decision-making problems. These problems are modeled as mathematical optimization problems. Therefore, problems with more than one objective function must be optimized simultaneously. In MORL, multiple Pareto-optimal solutions exist. The circumstances measured so far consisted of a single objective, which was to maximize the user throughput performances. In contrast, this section defines the use of a MORL agent to assist the C-RAN with multiple objectives. Importantly, in the case of energy saving in the network, the agent has dual goals to meet: maximize the average throughput of network by minimizing EC and maximize the user satisfaction. A separate area of research on RL is devoted to studying this class of problem. Based on the energy models, the average throughput of the *k*th user with the *n*th RRH is represented as

$$R_{kn} = B\log_2\left(1+\gamma_{kn}\right), \tag{5}$$

$$\text{s.t}\quad \text{C1}: \sum_{k=1}^{K} P_{kn}^{\text{trans}} = P_{n}^{\text{trans}}, \forall j \in \{1,2,...N\}, \tag{6}$$

$$\text{C2}: \sum_{n=1}^{N} R_{kn} > d_k, \tag{7}$$

$$\text{C3}: \sum_{n=1}^{N} \left|x_{kn}\right|_0 = 1, \quad \forall k \in \{1,2,...K\} \tag{8}$$

where B is the bandwidth. C1 indicates the transmit power allocated to the *k*th user to the *n*th RRH, constraint C2 states that maximizing throughput of the *k*th user should be greater than their minimum QoS. The constraint C3 states that one user is associated to one RRH at a time. Based on M/M/1 queuing theory, the average delay is expressed as

$$D_{kn} = \frac{1}{R_{kn}^{*} - x_{kn}\lambda_{kn}^{i}} \tag{9}$$

where x_{kn} denotes an association indicator between user k and RRH n. If $x_{kn} = 0$, there is no association between the users and RRH; otherwise, $x_{kn} = 1$. λ^{i}_{kn} denotes packet arrival rate. The satisfaction on delay is used to define when the maximum delay requirement of a user is not exceeded [22], [23]. The satisfaction on delay is given

$$\chi(D) = \frac{1}{1+\mathrm{e}^{-\varepsilon(D_k^{\max}-D_k)}} \tag{10}$$

where $D_k^{\max}$ is the maximum tolerant packet delay, which is required to satisfy the upper bound delay for user k and ε is a steepness constant which specifies the shape of the satisfactory curve. $\chi(D)$ of each user k is scaled between 0 and 1.

An ideal traffic profile is designed based on the trapezoidal traffic shape [14]. The traffic function is defined by the angular coefficient y which is given by

$$f(t) = \begin{cases} 1 - yt; & (0 \le t \le \frac{1}{y}) \\ 0; & \frac{1}{y} \le t \le T - \frac{1}{y} \\ 1 + y(t-T); & (T - \frac{1}{y} \le t \le T) \end{cases} \tag{11}$$

If y is equal to 1/10, then we move the $f(t)$ to $f(t) + 4$, which is close to the real measurements [23]. Here, T is time period of 24 hour and value of $f(t)$ is between 0 to 1.

6. Proposed Method

In this proposed resource allocation scheme, our objective is to minimize the EC and maximize user satisfaction. Our MORL problem is framed as a MDP with unknown transition probabilities. The MDP includes a set of state, a set of action, a reward and state transition probabilities. In the end, the resource allocation scheme based on MORL Q-learning is described. The system model is shown in Fig. 2. Based on the above optimization problem in (5), we define the state, action and reward in the MDP framework as follows:

State (s): As stated, the goal of MORL is to minimize the number of active RRHs ensuring QoS requirements of users. Based on this, the state space needs to reflect the on-off states of RRHs and their bandwidth occupancy. Therefore, we define the state space of an agent to include the on-off state of RRHs and the proportion of bandwidth resources occupied by the current RRHs. Each RRH has two states: the active and sleep state.

Environment: The massive MIMO enabled C-RAN functions in the environment where the learning agent acts. The environment shown as an MDP, with a state transition function, provides the probability defined as $p(s' \mid s', a')$ of shifting from state s to a when action a is performed. The agent receives a real-valued award for deciding action a in state s.

Episode: A sequence or an arrangement of actions passed by the agent and, the respective states and rewards gained through the environment is the episode of this framework. An episode ends when certain number of actions is completed.

Action (a): In this study, the objective of Q-learning-based energy consumption of active RRHs is based on the switching action of some RRHs when few RRHs satisfy the demand of the user. Here, the action is the agent turning the RRH on/off. In other words, the agent makes a corresponding switching action of RRHs according to its current state. Each RRH has two actions, switching on or off. The two actions are denoted as $N_j \in \{0,1\}$, N_j=0 represents switching off RRH j to sleep, and N_j denotes switching on RRH j to make it active. We express the MORL Q-value of state-action pair (s,a) in terms of the next state s' using the Bellman equation as follows:

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left[r + \gamma \max_{a'} Q(s',a') - Q(s,a)\right] \tag{12}$$

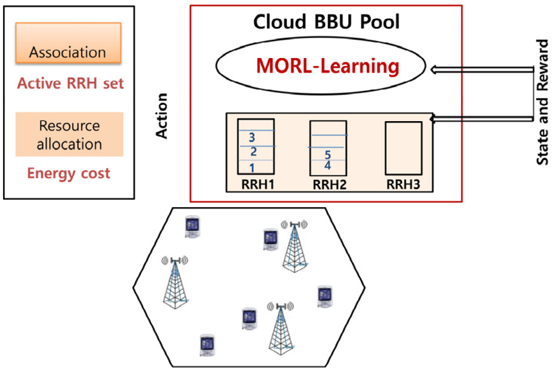


Fig. 2. System model.

where α is the learning rate, denoting the method chosen by the algorithm to reach a new reward value, γ denotes the discount factor reflecting the weights given to future rewards. s' is the new state after performing an action a in the state s. Let the combined MORL function be

$$\mathbf{MQ}^{\pi}(s,a)=[Q_1^{\pi}(s,a),Q_2^{\pi}(s,a)...Q_N^{\pi}(s,a)] \quad (13)$$

where $\mathbf{MQ}^{\pi}$ is a vector that satisfies the Bellman equation (12). The optimal state-action function can be expressed as

$$MQ^{*}(s,a)=\max_{\pi} MQ_{\pi}(s,a). \quad (14)$$

Policy: The actions of an agent are directed by a mapping function π, called the policy. It chooses the probability of the agent selecting an action $a'=a$, when the environment is in the state $s'=s$.

The letter Q is derived from the word "quality," as the Q-function represents the quality score for performing an action in a certain state. The ideal policy π_s for an agent to maximize the future (discounted) reward from state s is to always select the action with the highest Q-value as follows:

$$\pi_{(s)}=\arg\max_{a\in A_s} Q(s,a). \quad (15)$$

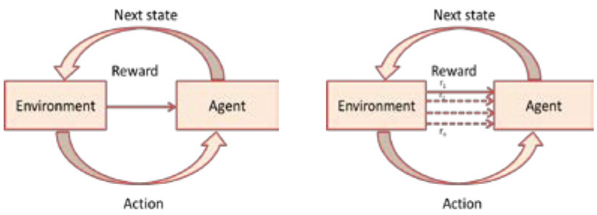
The optimal policy $\pi*$ is obtained as follows:

$$\pi^{*}(s)=\arg\max_{a} MQ^{*}(s,a). \quad (16)$$

Reward (r): Here, the reward function can be equated to the user throughput as:

$$r(s,a)=\sum_{k=1}^{K} R_{kn}. \quad (17)$$

Our target is to choose the best policy that maximizes the reward r. Owing to two conflicting objectives, we express the reward function as a combination of energy gain (E_{G}) and on delay (D) [11], [23]. We calculated energy gain as energy consumption reduction (%) that shows the energy savings. In MORL, the update of the state-action pair depends upon the reward. Thus, the agent learns the trade-off between these conflicting objectives by optimizing the reward function. In this case we define r for a given episode as the weighted sum given by:



(a) (b)

Fig. 3. (a) Basic agent–environment interaction model. (b) MORL agent–environment interaction model.

$$r(s,a)=-\omega\chi(D)+(1-\omega)\mathrm{E}_{\mathrm{G}}. \quad (18)$$

The main logic of the MORL algorithm is to achieve an optimal policy by exploring the environment. Through MORL method and updating situations to improve a numerical consequence, we obtain results that are negative or positive values. If the result is negative, it is considered as a cost or penalty, and if the result is positive, then it is considered as a reward. Here, because there are more than one conflicting objectives, this problem can be classified as an MORL problem. MORL is different from traditional RL. Here, the suitability of the performed action is checked by the decision maker. As the environment evolves with time, the agent must adapt and continuously learn.

By working and exploring the environment, the agent gains information that is useful when making future decisions. The agent-environment interaction model scheme is shown in Fig. 3(a,b). In Q-learning, agent takes an action a, then changes to a new state s' while getting a reward r (Fig. 3a). In addition, the agent selects new actions to find the optimal solution and better future rewards as shown in Fig. 3b.

A common solution to such MORL problems is the greedy algorithm defined below.

Algorithm 1 Single policy approach

```
1.  Initialize $Q(s, a) = 0$;  , learning rate $\alpha$
2.  Set the weight ,discount factor for each training episode
3.  Scheduled user and randomly select action
4.  For each step do
5.  Select the RRH based on the action policy
6.  Else, $a'$= argmax $Q(s, a)$; \\ where $Q(\bullet)$ is estimated by the
      energy consumption\\
7.  Activate action $a$ and select the set of active RRH
8.  Calculate user throughput and user satisfaction using (10).
9.  Compute the reward
10. Obtain current reward and the next state from environment.
11. Update the scheduled user set $K$.
12. Obtain the energy consumption based on set of active RRHs.
13. Set the next state as the current state.
14. Observe next state $s'$.
15. Update the MQ-value $Q(s; a)$
16. End procedure.
```

For resource allocation, we update Q-function and select the best improved policy. The detailed execution is as follows. In the first steps, we initialize the Q-table by using an initial value of 0, as agents do not have the knowledge of the environment. In second step, the user association between user and RRHs follow. In the third step, we update the MQ-function and obtain the new state based on the determined state transition sequence (*s,a,r,s′*). In this case, Q-function is updated when the new Q-function is more than the previous Q-function; else, Q-function is unaffected. Here, actions are selected randomly on the basis of Q function. At last we achieve an EC based on the set of active RRHs by solving the resource allocation models. Such an update criterion appropriately reduces the computational complexity.

7. Numerical Analysis

In this section, we investigate the performance of our proposed MORL algorithm. We consider three RRHs, which are connected to a BBU in the network. In each group, the BBU takes over the on-off action using the Q-table generated from the Q-learning agent. The number of RRHs is based on the use-case setup in the experiment. The two use cases defined in this study assume a fixed number of RRHs from the performance evaluation.

In addition, we assume the maximum number of RRHs for a BBU cloud to be 3 and 18 for the whole system. To be more realistic, we set the system bandwidth of RRHs at 20 MHz. The coverage of the RRH in the network is 200 m. The path loss in (1) is given at a carrier frequency of 2 GHz assuming that the base station height is 15 m [7]. The number of user range between 4 and 100 according to the traffic model [13]. The user demand of 1 Mbps (data traffic rate in current 5G network) is equal for all users. The packet arriving rate λ is taken as 160 packets per second and the steepness constant ε is taken as 1 [23]. Each user is assumed to have one interface. The EC largely depends on transceiver power settings, traffic loads and the active RRH duration. The traffic pattern was considered for a period of 24 h in our network. The active users and traffic demands vary over this period which increases the complexity of the traffic mode. The discipline of the traffic is given by a daily traffic model in a day based on the real on-site measurements from the EARTH project [14]. To track the EC with higher accuracy, system time is allocated into 24 time slots in a typical day (24 h).

For evaluating the proposed model and algorithms, we define four different schemes. Scheme I is a simple on-off cell activation with the simple nearest-RRH association, named the simple on-off scheme. Scheme II comprises cell activation with the load ordering-based heuristic scheduling algorithm. Scheme III is based on Q-learning. Finally, we use Q-learning to make a cell activation decision, but the resource allocation is solved using the CVX solver; this scheme is named as Q-learning with CVX (Q-learning-CVX). We compare our proposed algorithm with the simple on-off scheme and the heuristic scheduling algorithm because they are model-free and consider the dynamics of traffic distribution. However, the simple on-off scheme is a baseline algorithm that prefers the nearest-association of user and RRHs. If no user is present near an RRH, the RRH is switched off and vice versa. The difference between heuristic scheduling-based algorithm and proposed Q-learning algorithm is that the heuristic algorithm is based on static policy, i.e., there is no feedback to the former after scheduling. The learning agent in the Q-learning-based algorithm receives feedback in the form of a reward. As the traffic distribution changes, the learning agent selects an optimal solution to the problem. The simulation parameters are listed in Tab. I.

The objective of this study is to optimize the wireless network energy consumption and radio resource occupancy while satisfying the QoS requirement of users. The simulation results are classified into the following metrics: the number of active RRHs, transmit power cost, accumulated total EC and average user QoS satisfaction. The normalized number of active RRHs is used to evaluate the effect of cell activation. For the Q-learning scheme, we assume that the used bandwidth proportion is equal to transmit power cost proportion. The transmit power cost can be used to evaluate this effect.

Simulator Parameter	Value
Maximum transmit power per RRH	1 W
Power consumption (active RRH)	6.8 W
Power consumption (sleep mode RRH)	4.3 W
Carrier frequency	2 GHz
Bandwidth	20 MHz
Path loss (*d* in km)	[128.1 + 37.6 $\log_{10}(d)$] dB
Noise power density	−174 dB/Hz
Number of users	4–100
Number of RRH	3–18
Weighted factor in reward function	100

Tab. 1. Simulation parameters.

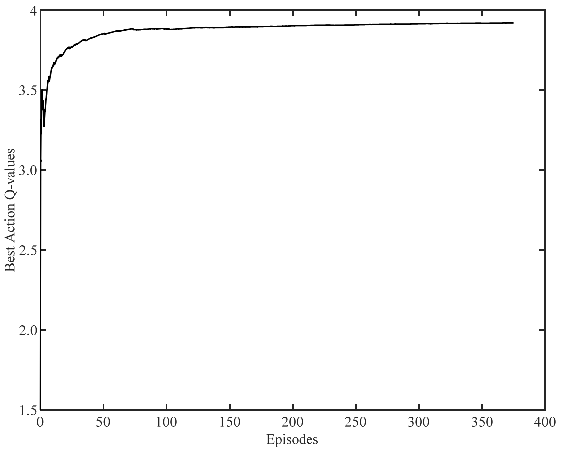


Fig. 4. Convergence of the Q-learning algorithm.

Figure 4 shows the convergence evolution of the best action in the learning phase. For a sufficiently high number of iterations (approx. 104), the maximum deviation is almost zero for all state-action pairs. As shown in Fig. 4, the proposed algorithm converges for all RRHs and terminates for all state-action pairs (*s, a*) after 100 episodes. When the Q-learning algorithm converges, we stop the training and obtained policies. Then, the best action policy is retained in a look-up table, and this stored value is used during the exploitation process.

Figure 5 shows the traffic load during 24 hours. The traffic model based on the real Earth project shows that traffic demands vary over the clock time in a day. Here, the user demands do not change but the total user demand changes over 24 hours in a day leading to the states of RRHs changing between active and sleep. As shown, traffic load (25 Mbps) is low from 6 to 15 hours. Our proposed Q-learning algorithm will activate the nearest RRHs to handle user demands. Similarly, for high traffic loads (70–90 Mbps) during 21 to 24 hours, all RRHs can be activated causing more energy consumption.

Figure 6 shows average throughput per user during day time. Since throughput depends on the number of resources available, it is seen that when traffic load increases, the throughput decreases. It is also depicted that simple on-off and heuristic algorithm causes only few RRHs to be active to meet user demands. Our proposed Q-learning algorithm achieves higher throughput even at low traffic loads.

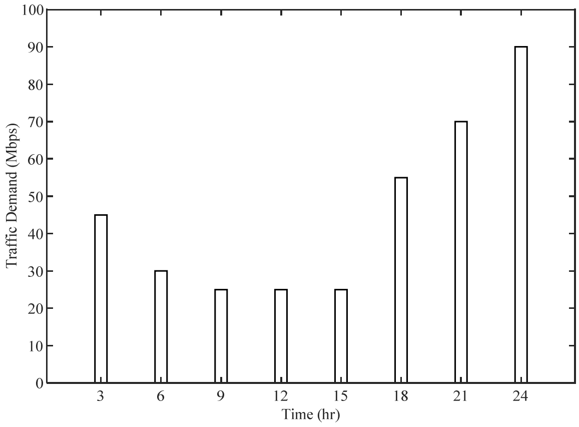


Fig. 5. Traffic load during 24 hours.

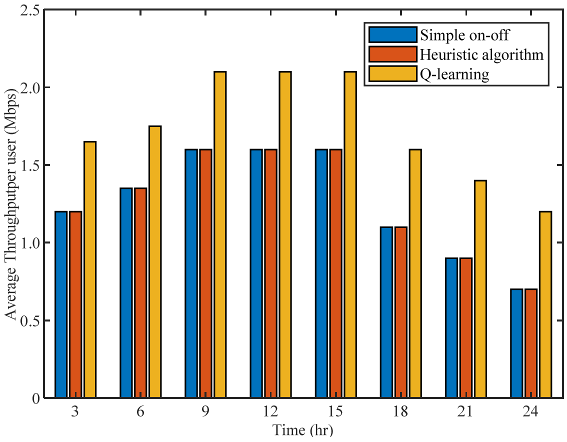


Fig. 6. Average throughput per user over daytime.

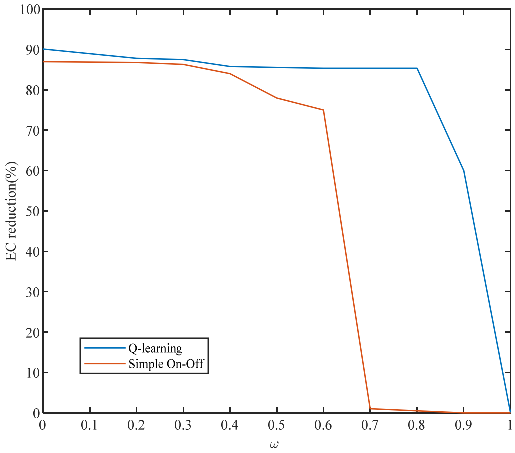


Fig. 7. Energy savings comparison between different algorithms.

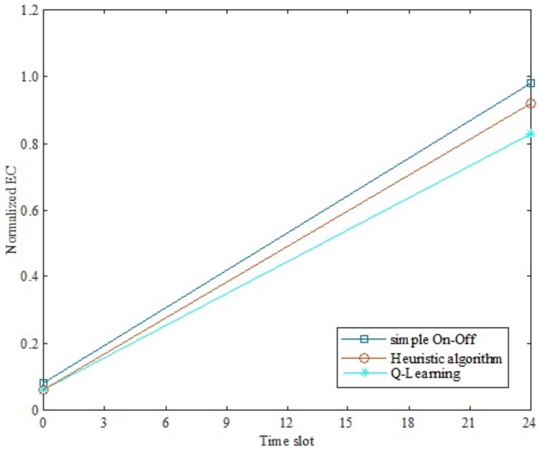


Fig. 8. Total normalized energy consumption.

Figure 7 presents the balance between the energy consumption of the average network and the delay with different values of ω. The number of users is 20. We can manage the energy consumption when the RRH is turned off. Here, significant energy savings up to 90% are observed. The proposed Q-learning algorithm outperforms simple on-off method.

In Fig. 8, we demonstrate the normalized EC of different policies with respect to the time slot numbers. The performances of the Q-learning-based policies show suboptimal EC. This is because the Q-learning-based methods exploit the causal information of channel gains, which helps avoid energy wastage. This confirms that Q-learning-based policies improve the EC over the long process and can make real-time adjustments to the dynamic environment.

In Fig. 8, the normalized EC of the simple switching scheme is approximately 0.99, and it is approximately 0.8 in the Q-learning-based schemes. The proposed Q-learning

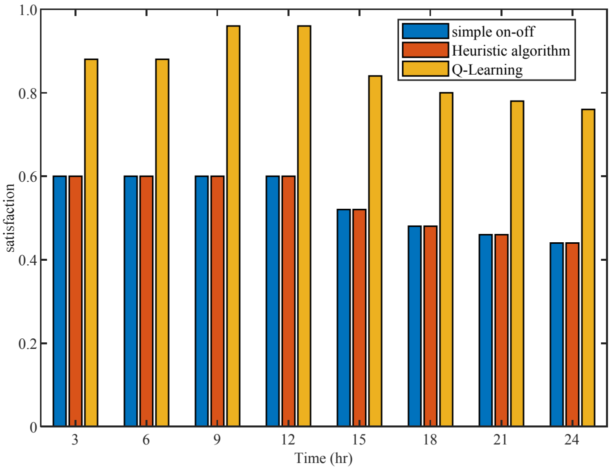


Fig. 9. User satisfaction comparison between different algorithms.

algorithm outperforms other algorithms, having the least total normalized EC, which is close to that obtained via heuristic approach.

Figure 9 presents the user throughput satisfaction during 24 hours. The satisfaction depends on number of active RRH. For Q-learning algorithm, user satisfaction is almost 100% and remains over 80% even at higher loads during daytime satisfying the QoS requirement. In case of simple on-off and heuristic approach, user demands are satisfied at low loads. The satisfaction is 60% during 3 to 12 hours. However, when load increases to peak hours from 19 to 24 hours, the satisfaction decreases down to as low as 40%.

We observed the performance of the three schemes in terms of total energy consumption, user throughput and satisfaction that satisfy the QoS of the network. Based on the above discussion, we conclude that our proposed algorithm works better than the other schemes even while changing traffic demands during a clock time.

8. Conclusion

In this paper, we have defined a MORL method to solve the resource allocation problem in MIMO enabled C-RANs. Our scheme is designed to increase the system throughput. This study explored the problem of energy consumption and delay that are linked with the QoS requirements of 5G networks. Although minimizing the energy consumption of these wireless communication networks is critical, an algorithm based on the Q-learning model is required to control the state of RRH for achieving a trade-off between energy consumption and user satisfaction. We focused on EE, user satisfaction, and delay to optimize the trade-off and QoS requirements associated with multiobjective environment using the MORL method. In the model, we maximized the user throughput depending on the user position according to the turn-on and turn-off of the C-RAN RRHs. For selecting the best policy, we proposed a MORL algorithm that showed better performance than a heuristic algorithm. The normalized ECs of simple ON-OFF scheme, heuristic and Q-learning algorithm were 0.99, 0.85, and 0.8, respectively. This indicates that our Q-learning MORL algorithm outperformed the traditional simple on-off method in an ordinary system and showed better energy saving ability while maintaining a good QoS. Furthermore, it allows the operator to manage the trade-off according to the 5G network usage.

Acknowledgments

This research was funded by the National Research Foundation of Korea (NRF), Ministry of Education, Science and Technology (Grant No. 2016R1A2B4012752).

References

[1] BJÖRNSON, E., HOYDIS, J., SANGUINETTI, L. *Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency*, *Foundations and Trends in Signal Processing*, 2017, vol. 11, no. 3–4, p. 154–655. DOI: 10.1561/2000000093

[2] LUO, S., ZHANG, R., LIM, T. J. Downlink and uplink energy minimization through user association and beam forming in C-RAN. *IEEE Transactions on Wireless Communications*, 2015, vol. 14, no. 1, p. 494–508. DOI: 10.1109/TWC.2014.2352619

[3] PARK, S., CHAE, C.-B., BAHK, S. Large-scale antenna operation in heterogeneous cloud radio access networks: A partial centralization approach. *IEEE Wireless Communications*, 2015, vol. 22, no. 3, p. 32–40. DOI: 10.1109/MWC.2015.7143324

[4] SHARMA, S., YOON, W. Multiobjective optimization for energy efficiency in cloud radio access. *International Journal of Engineering Research and Technology*, 2019, vol. 12, no. 5, p. 607–610. ISSN: 0974-3154

[5] PENG, M., ZHANG, K., JIANG, J., et al. Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks. *IEEE Transactions on Vehicular Technology*, 2015, vol. 64, no. 11, p. 5275–5287. DOI: 10.1109/TVT.2014.2379922

[6] SUTTON, R. S., BARTO, A. G. *Reinforcement Learning: An Introduction*. 2nd ed. Cambridge (MA): MIT Press, 2018. ISBN: 978-0262039246

[7] EL-AMINE, A., ITURRALDE, M., HASSAN, H. A. H., et al. A distributed Q-Learning approach for adaptive sleep modes in 5G networks. In *2019 IEEE Wireless Communications and Networking Conference (WCNC)*. Marrakesh (Morocco), 2019, p. 1–6. DOI: 10.1109/WCNC.2019.8885818

[8] ORTIZ, A., AL-SHATRI, H., LI, X., et al. Reinforcement learning for energy harvesting point-to-point communications. In *Proceedings of the IEEE International Conference on Communications*. Kuala Lumpur (Malaysia), 2016, p. 1–6. DOI: 10.1109/ICC.2016.7511405

[9] VAN OTTERLO, M., WIERING, M. Reinforcement learning and Markov decision processes. In Wiering, M., van Otterlo, M. (eds) *Reinforcement Learning. Adaptation, Learning, and Optimization*. Berlin, Heidelberg (Germany): Springer, 2012, vol. 12. DOI: 10.1007/978-3-642-27645-3_1

[10] CHEN, X., LI, N., WANG, J., et al. A dynamic clustering algorithm design for C-RAN based on multi-objective optimization theory. In *IEEE 79th Vehicular Technology Conference (VTC*

Spring). Seoul (South Korea), 2014, p. 1–5. DOI: 10.1109/VTCSpring.2014.7022775

[11] SALEM, F. E., ALTMAN, Z., GATI, A., et al. Reinforcement learning approach for advanced sleep modes management in 5G networks. In *2018 IEEE Vehicular Technology Conference (VTC-Fall)*. Chicago (IL, USA), July 2018, p. 1–5. DOI: 10.1109/VTCFall.2018.8690555

[12] LIU, C., XU, X., HU, D. Multiobjective reinforcement learning: A comprehensive overview. *IEEE Transactions on Systems, Man, and Cybernetics: Systems,* 2014, vol. 45, no. 3, p. 385–398. DOI: 10.1109/TSMC.2014.2358639

[13] SHI, Y., ZHANG, J., LETAIEF, K. B. Group sparse beamforming for green cloud radio access networks. In *Proceedings of Global Communications Conference (Globecom)*. Atlanta (GA, USA), 2013, p. 4662–4667. DOI: 10.1109/GLOCOMW.2013.6855687

[14] SUN, G., ADDO, P. C., WANG, G., et al. Energy efficient cell management by flow scheduling in ultra-dense networks. *KSII Transactions on Internet and Information Systems,* 2016, vol. 10, no. 9, p. 4108–4122. DOI: 10.3837/tiis.2016.09.005

[15] ZHUANG, B., GUO, D., HONIG, M. L. Energy-efficient cell activation, user association, and spectrum allocation in heterogeneous networks, *IEEE Journal of Selected Areas in Communication,* 2016, vol. 34, no. 4, p. 823–831. DOI: 10.1109/JSAC.2016.2544478

[16] NATARAJAN, S., TADEPALLI, P. Dynamic preferences in multi-criteria reinforcement learning. In *Proceedings of the 22nd International Conference on Machine Learning*. Bonn (Germany), 2005, p. 601–608. DOI: 10.1145/1102351.1102427

[17] SHARMA, S., YOON, W. Multi-objective energy efficient resource allocation for WPCN. *International Journal of Engineering Research and Technology*, 2018, vol. 11, no. 12, p. 2035–2043. ISSN: 0974-3154

[18] ROIJERS, D. M, VAMPLEW, P., WHITESON, S., et al. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 2013, vol. 48, p. 67–113. DOI: 10.1613/jair.3987

[19] YANG, R., SUN, X., NARASIMHAN, K. A generalized algorithm for multi-objective reinforcement learning and policy adaptation. In *Proceedings of 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*. Vancouver (Canada), 2019, p. 1–12.

[20] HAO, Y., NI, Q., LI, H., et al. Robust multi-objective optimization for EE-SE tradeoff in D2D communications underlaying heterogeneous networks. *IEEE Transaction on Communication*, 2018, vol. 66, no. 10, p. 4936–4949. DOI: 10.1109/TCOMM.2018.2834920

[21] SAKULKAR, P., KRISHNAMACHARI, B. Online learning schemes for power allocation in energy harvesting communications. *IEEE Transactions on Information Theory*, 2018, vol. 64, no. 6, 4610–4628. DOI: 10.1109/TIT.2017.2773526

[22] SHENG, M., XU, C., WANG, X., et al. Utility-based resource allocation for multi-channel decentralized networks. *IEEE Transactions on Communications,* 2014, vol. 62, no. 10, p. 3610–3620. DOI: 10.1109/TCOMM.2014.2357028

[23] SUN, G., ZHANG, T., OWUSU BOATENG, G., et al. Revised reinforcement learning based on anchor graph hashing for autonomous cell activation in cloud-RANs. *Future Generation Computer Systems*, 2020, vol. 104, p. 60–73. DOI: 10.1016/j.future.2019.09.044

About the Authors ...

Shruti SHARMA received her M.Sc. degree in Electronics from Pt. Ravishankar Shukla University, Raipur, C.G., India in 2008. Further, she completed her Diploma in Embedded System Design form C-DAC, Kolkata, India and Post Graduate Diploma in Computer Applications form Guru Ghasidas University, Bilaspur. From 2011 to 2014, she served in GENPACT multinational firm. She also worked as Assistant Professor in CMD College, Bilaspur, C.G. India from 2014 to 2015. Currently, she is pursuing her Ph.D. degree in Electrical and Computer Engineering from Ajou University, South Korea. Her research interests include embedded design, machine learning, 5G communication, and massive MIMO.

Wonsik YOON graduated with his B.S. degree in Control and Instrumentation Engineering from Seoul National University in 1984. He received his M.S. and Ph.D. degrees from KAIST in 1986 and 1991, respectively. From 1986 to 1994, he worked with Goldstar Electrical Company and LG Innotek, South Korea. From 2000 to 2001, he worked as the CTO at Contela Inc., South Korea. Since 1994, he has been with the Department of Electrical Engineering at Ajou University, South Korea, where he is a Professor. His research interests include wireless communications and networks, massive MIMO, and network coding. He is a senior member of IEEE.