Enhanced Machine Learning Scheme for Energy Efficient Resource Allocation in 5G Heterogeneous Cloud Radio Access Networks

Ismail AlQerm and Basem Shihada
CEMSE Division, King Abdullah University of Science and Technology, Saudi Arabia,
{ismail.qerm, basem.shihada}@kaust.edu.sa

***Abstract*—Heterogeneous cloud radio access networks (H-CRAN) is a new trend of 5G that aims to leverage the heterogeneous and cloud radio access networks advantages. Low power remote radio heads (RRHs) are exploited to provide high data rates for users with high quality of service requirements (QoS), while high power macro base stations (BSs) are deployed for coverage maintenance and low QoS users support. However, the inter-tier interference between the macro BS and RRHs and energy efficiency are critical challenges that accompany resource allocation in H-CRAN. Therefore, we propose a centralized resource allocation scheme using online learning, which guarantees interference mitigation and maximizes energy efficiency while maintaining QoS requirements for all users. To foster the performance of such scheme with a model-free learning, we consider users' priority in resource blocks (RBs) allocation and compact state representation based learning methodology to enhance the learning process. Simulation results confirm that the proposed resource allocation solution can mitigate interference, increase energy and spectral efficiencies significantly, and maintain users' QoS requirements.**

***Index Terms*—Resource allocation, energy efficiency, online learning, H-CRAN**

I. Introduction

The wireless industry is witnessing an avalanche of mobile traffic fueled by the increasing interest in high data rate services such as video conferencing, online high definition video streaming, and the increasing popularity of handheld devices. 5G is envisioned as the potential solution to handle this extraordinary data rate demand [1]. Cloud radio access networks (CRAN) are recognized to reduce operating expenditures, manage inter-cell interference, and provide high data rates with considerable energy efficiency performance [2]. It consists of remote radio heads (RRHs), which act as relays that forwards the users (UEs) data to the centralized baseband unit (BBU) pool for processing through wired/wireless fronthaul links. On the other hand, high power macro base stations (BSs) in heterogeneous networks support coverage and guarantee backward compatibility with traditional cellular networks since small cells focus only on boosting the data rate in special zones [3]. To realize the 5G vision, heterogeneous networks with massive densification of small cells and CRAN are combined in one network structure called heterogeneous cloud radio access network (H-CRAN) to improve spectral efficiency, resource management, and energy efficiency [4]. H-CRAN aims at exploiting the heterogeneous networks advantages to overcome the limited fronthaul links capacity problem in CRAN. However, resource allocation in H-CRAN considering the heterogeneity of the network, interference between network tiers, and users QoS constraints in an energy efficient manner is a challenging problem.

In this paper, we propose a green resource allocation scheme for the downlink of H-CRAN between RRHs and their associated UEs that aims at maximizing energy efficiency while satisfying the UE QoS requirements and mitigate the inter-tier interference. Allocation includes resource blocks (RB) assignment and power allocation. The scheme is developed using enhanced online learning where the allocation is performed at a designated controller integrated with the BBU, thanks to the macro BS, which is interfaced to the BBU pool for coordinating the inter-tier interference and exchange resource allocation control signals. This alleviates the capacity and time delay constraints on the fronthaul links and supports the burst traffic efficiently. The proposed learning methodology functions with support of an enhanced spectrum partitioning that classifies the available RBs according to the users traffic priority and location. This partitioning will assist the online learning in RBs allocation as the information about UE location and requirements can be used as prior knowledge for learning. Moreover, compact state representation is employed to handle the curse of dimensionality and expedite the learning process. To the best of our knowledge, there is no solutions in the literature for resource allocation in H-CRAN with consideration of energy efficiency using machine learning techniques. The key contributions of this paper are summarized as follows,

- We develop an enhanced spectrum partitioning based system model that divides the available spectrum into two RBs sets where each set is dedicated for certain group of UEs according to their location and QoS requirements.
- A centralized joint RBs and power allocation scheme for RRH and their associated UEs is proposed, which relies on a single controller integrated to the BBU. This controller acquires the network state information through its interface with the macro BSs and consequently, utilizes this knowledge to select the most appropriate actions that enhance energy efficiency and guarantee UEs QoS

requirements from different tiers.

- The proposed online learning model incorporates compact state representation to reduce the size of the state space, handle the curse of dimensionality, and augment the algorithm convergence.

The paper is organized as follows, section II presents the related work. The system model and problem formulation are described in section III. The learning model for resource allocation and the centralized resource allocation scheme are presented in section IV and V respectively. Section VI shows the simulation results and the paper concludes in section VII.

II. Related Work

Energy efficiency problem in the context of CRAN has been studied in the literature while designing resource allocation schemes. The authors in [5] formulated joint RRH selection and power consumption minimization, subjected to user QoS requirements and RRH power budget, as a group sparse beamforming problem. However, the proposed scheme ignored the fact that the fronthaul links have a limited capacity. The work in [6] aimed at optimizing the end-to-end TCP throughput performance of Mobile Cloud Computing (MCC) users in a CRAN network through topology configuration and rate allocation. One limitation of this work is that it did not constrain the capacity consumption of individual links. The authors in [7] minimized the total network power consumption in a CRANs subject to users' QoS for both secure communication and efficient wireless power transfer, limited backhaul capacity, and power budget constraints. However, this proposal considers single tier network only. Consequently, it is necessary to decouple data and control signals in order to alleviate the influence of fronthaul links on energy efficiency, dedicate RRHs to provide high data rates, and utilize macro BSs to convey control signals [8]. Despite the fact that heterogeneous networks are able to improve the coverage and the capacity, inter-tier interference and the cumulative power consumption of the small cells are critical challenges that must be considered [9]. The energy efficiency oriented resource allocation for heterogeneous networks attracts the researchers' attention in the literature. In [10], distributed power allocation for multi-cell OFDMA networks taking both energy efficiency and intercell interference mitigation into account was investigated where bi-objective problem was formulated and solved using multi-objective optimization. The power allocation, RBs allocation and relay selection were optimized in [11] with the goal of maximizing energy efficiency. An optimal power allocation algorithm using equivalent conversion is proposed in [12] to maximize energy efficiency under interference constraints. The authors in [13] explored a system framework of cooperative green heterogeneous networks for 5G wireless communication systems. Regrading resource allocation in H-CRAN, authors in [14] proposed a joint optimization for RBs and power allocation subjected to interference constraints in H-CRAN. The work in [15] formulated a joint optimization problem of relay selection, power allocation and network selection to maximize the energy efficiency in H-CRAN. However, the proposed schemes are limited to the specified optimization model, which is not practical in such dynamic environment.

III. System Model and Problem Formulation

The architecture of the considered H-CRAN system is presented in Fig. 1, in which a two-tier cellular network is shown, the macro tier and the small cells tier. The macro BSs denoted by $u \in \mathbf{U}$ are underlaid with small cells (RRHs) denoted by $s \in \mathbf{S}$, and the BBU pool performs all baseband processing functionalities. The resource allocation process is executed at the controller integrated with the BBU pool. The network UEs in this model are classified into two categories: MUEs denoted by index $m \in \mathbf{M}$, which are the users associated with the macro BS, while RUEs with index n are the users associated with RRHs. The backhaul interface is utilized to link the macro BSs with the BBU for control exchange. All the RRHs are connected to the BBU pool by the fronthaul links to facilitate data processing, transmissions, and cloud computing. The system RBs are donated by $k \in \mathbf{K}$ with total bandwidth B. To maximize the spectral efficiency and assist the mitigation of the inter-tier interference, the available spectrum is partitioned into two sets of RBs: the first set denoted by Γ_1 incorporates RBs dedicated to RUEs with high QoS requirements or located at their corresponding cell edge. The second set of RBs denoted by Γ_2 comprises RBs to be shared between RUEs with low QoS or located at the cell center and MUEs. This specific partitioning maximizes the spectrum efficiency and limits the inter-tier interference compared to the traditional schemes and avoids random exploration of machine learning in RBs allocation, which expedites the learning speed of convergence. The QoS requirement is defined as the minimum data rate of the RUEs. The sets $\mathbf{N} = \{1, 2, ..., N\}$ and $\mathbf{q} = \{N+1, ...N+q\}$ represent the users associated with RRH and occupying the sets Γ_1 and Γ_2 respectively. Let us assume that a_n^k is the RB allocation indicator, $a_n^k = 1$, if the RB k is allocated to RUE n and 0 otherwise. For RUE n served by RRH s on RB k, the received signal can be written as,

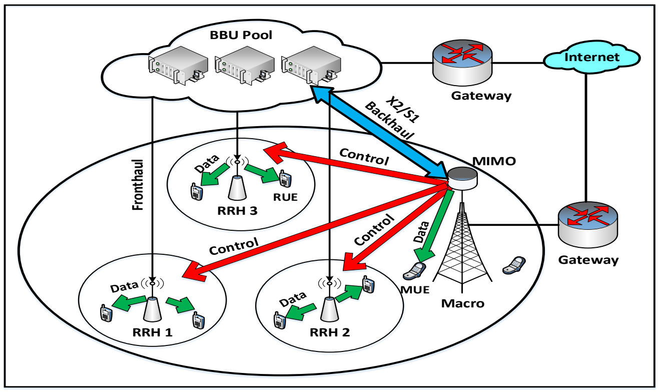


Fig. 1. H-CRAN Architecture

$$\eta_{s,n}^k = P_{s,n}^k g_{s,n}^k \rho_n^k + I_n^k + N_0 \qquad (1)$$

where $P_{s,n}^k$ is the transmission power of RRH s allocated to RUE n on RB k, ρ_n^k is the transmitted information symbol for RUE n on RB k, N_0 is the noise power, and $g_{s,n}^k = H_{s,n}^k l_{s,n}$ is the channel gain between RRH s and RUE n on RB k. Note that $H_{s,n}^k$ and $l_{s,n}$ refer to the fast fading coefficient and the path loss between RRH s and RUE n respectively. The encountered interference denoted by I_n^k is calculated according to the category of the RUE n. Therefore, I_n^k is determined as follows,

$$I_n^k = \begin{cases} \sum_{i=1,i\neq n}^{N}(\sum_{r\in\mathbf{S},r\neq s} P_{r,i}^k g_{r,i}^k) \; k\in\Gamma_1 \\ \sum_{m=1}^{M} P_{u,m}^k g_{u,m}^k + \sum_{i=N,i\neq n}^{N+q}(\sum_{r\in\mathbf{S},r\neq s} P_{r,i}^k g_{r,i}^k) \\ k\in\Gamma_2 \end{cases} \tag{2}$$

where $P_{u,m}^k$ is the transmission power of the macro BS u allocated to MUE m on RB k, and $g_{u,m}^k = H_{u,m}^k l_{u,m}$ is the channel gain between macro BS u and MUE m on RB k. Note that the index i represents the other RUEs served by other RRH r. The signal to interference and noise ratio (SINR) for nth RUE associated with sth RRH and occupying kth RB is given by,

$$\gamma_{s,n}^k = \frac{P_{s,n}^k g_{s,n}^k}{I_n^k + N_0} \tag{3}$$

Similarly, the received signal by an MUE m served by macro BS u is written as follows,

$$\eta_{u,m}^k = P_{u,m}^k g_{u,m}^k \rho_m^k + I_m^k + N_0 \tag{4}$$

where ρ_m^k is the received information symbol and I_m^k is the encountered interference by the MUE m allocated to RB k and it is evluated as follows,

$$I_m^k = \sum_{i=N}^{N+q}(\sum_{s\in\mathbf{S}} P_{s,i}^k g_{s,i}^k) \tag{5}$$

The SINR achieved by MUE m utilizing RB k and associated with macro BS u is found as follows,

$$\gamma_{u,m}^k = \frac{P_{u,m}^k g_{u,m}^k}{I_m^k + N_0} \tag{6}$$

The capacity for all RUEs associated with RRH s is expressed as follows,

$$C_s = \sum_{n=1}^{N+q}\sum_{k=1}^{K} a_{s,n}^k B \log_2(1+\gamma_{s,n}^k) \tag{7}$$

where $n \in \mathbf{N}$ represents the RUE allocated to RB k in Γ_1, $n \in \mathbf{q}$ is the RUE allocated to RB in Γ_2. The power consumption of RRH s is evaluated as follows,

$$PC_s = P_{ct} + P_f + \sum_{n=1}^{N+q}\sum_{k=1}^{K} P_{s,n}^k a_{s,n}^k \tag{8}$$

where P_{ct} is the circuit power, P_f is the power consumption of the fronthaul links. Similarly, the capacity of all MUEs associated with the macro BS u is found as follows,

$$C_u = \sum_{m=1}^{M}\sum_{k=1}^{K} a_{u,m}^k B \log_2(1+\gamma_{u,m}^k) \tag{9}$$

where $a_{u,m}^k$ is the RB resource allocation indicator similar to $a_{s,n}^k$. The energy efficiency (EE) of RRHs in the considered model is defined as follows,

$$EE = \frac{C_s}{PC_s} \tag{10}$$

The allocation of RBs $a_{s,n}^k$ and transmission power $P_{s,n}^k$ with objective of EE maximization in the downlink of RRHs in H-CRAN subjected to QoS requirements is formulated as follows,

$$\max_{P_{s,n}^k, a_{s,n}^k} EE \quad s.t \tag{11}$$

$$\text{C.1: } \sum_{k=1}^{K} a_{s,n}^k \leq z_n,\; a_{s,n}^k \in \{0,1\}\; \forall n \in \mathbf{N}\cup\mathbf{q}$$

$$\text{C.2: } C_s \geq \theta,\; \forall n \in \mathbf{N},\; C_s \geq \theta^*,\; \forall n \in \mathbf{q}$$

$$\text{C.3: } C_u \geq \delta,\; \forall m \in \mathbf{M}$$

$$\text{C.4: } P_{s,n}^k \leq a_{s,n}^k P_{s,max},\; \forall s \in \mathbf{S}, \forall n \in \mathbf{N}\cup\mathbf{q}, \forall k \in \mathbf{K}$$

$$\text{C.5: } \sum_{n=1}^{N+q} \phi(\sum_{k=1}^{K} P_{s,n}^k) \leq \chi_{max},\; \forall s \in \mathbf{S}$$

The constraint C.1 limits the number of allocated RBs to RUE n to z_n RBs which prevents the cloud from greedily allocating all the available RBs to its RUEs and $a_{s,n}^k$ is a binary variable. The capacity constraint C.3 ensures that the achieved capacity for RUEs with high QoS requirements and low QoS requirements RUEs are above the thresholds θ and θ^* respectively. C.3 is the constraint to guarantee QoS for the MUE m, which limits the interference introduced by the transmissions of RRHs on RB k. The transmit power on an unallocated RB is enforced to be zero and the maximum allowed power for each RRH is indicated in C.4 as $P_{s,max}$. Finally, C.5 is a fronthaul constraint which limits the number of baseband signals transmitted on the fronthaul link between the cloud and RRH s to χ_{max}, where ϕ is a step function that takes value 1 if $\sum_{k=1}^{K} P_{s,n}^k > 0$ and 0 otherwise. The resource allocation problem in (11) is tackled using enhanced online learning to reach ultimate allocation of RB and transmission power to guarantee capacity in the designated sub-bands and inter-tier interference mitigation.

IV. Online Learning Model for Resource Allocation

As the spectrum partitioning based RB allocation considered in this work is not only location dependent but also considers QoS requirements, it is not practical to predefine the state transition model for solving the optimization problem. Therefore, we choose online Q-learning [16] for resource allocation in the specified model. The basic concept of online learning is described as follows, when the network is in state x^t at time step t, a finite number of possible actions y^t, which are elements of the action space Y can be selected. As a result, a reward is received which is the network feedback for the action selected at state x^t. In this section, we describe the learning model employed to achieve resource allocation that maximizes the network energy efficiency. The considered learning model for resource allocation is defined as $\zeta = (\mathbf{N}, \mathbf{q}, \gamma_{s,n}^{k}, \gamma_{u,m}^{k}, a_{s,n}^{k}, P_{s,n}^{k}, EE)$.

The online learning parameters are defined as follows:

- **State:** the environment state at certain time slot t is defined as, x^t=(n, RUE location, θ, θ^*, δ, $\gamma_{s,n}^{k}$, $\gamma_{u,m}^{k}$). The state information is acquired from the BBU using its interface with the macro BSs.
- **Action:** the action $y^t = (a_{s,n}^{k}, P_{s,n}^{k})$ is defined as the allocation of RB and transmission power, which RRH allocates to its associated RUEs.
- **Reward:** the reward function is defined as the energy efficiency as,
$$R(x,y) = EE(x,y) \tag{12}$$
The reward is achieved if the conditions in C.1 to C.5 are satisfied.
- **Transition Function:** for a given resource allocation strategy $\pi \in \Pi$, the state transition probability is defined as follows,
$$T_{x,x'}(y) = Pr(x(t+1) = x'|^t = x, y^t = y) \tag{13}$$
Basically, the strategy π is defined as the probability of selection of action y at state x.

The optimal Q-value of the online learning model is defined as the current expected reward plus a future discounted reward as follows,

$$Q^*(x,y) = E[EE(x,y)] + \beta \sum_{s' \in X} T_{x,x'}(y) \max_{y' \in Y} Q^*(x', y') \tag{14}$$

where β is the discount factor. The optimal Q-value $Q^*(x,y)$ is learned by updating the Q-value function on the transition from state x to state x' under the action y in time slot t as follows,

$$\begin{aligned} Q^{t+1}(x,y) = (1-\alpha^t)Q^t(x,y) + \alpha^t[EE(x,y) \\ +\beta max_{y' \in Y} Q^t(x', y')] \end{aligned} \tag{15}$$

where $\alpha^t \in (0,1]$ is the learning rate. The initial Q-value for all (x,y) is arbitrary. The considered online learning model here is a stochastic approximation method that solves the Bellman's optimality equation associated with the discrete time markovian decision process (DTMDP). Online learning does not require explicit state transition probability model and it converges with probability one to an optimal solution if $\sum_{t=1}^{\infty} \alpha^t$ is infinite, $\sum_{t=1}^{\infty} (\alpha^t)^2$ is finite, and all state action pairs are visited infinitely often [17]. Balancing exploration and exploitation is an essential issue in the stochastic learning process. Exploration aims to try new allocation strategies while exploitation is the process of using well-established strategies. The most common technique to achieve this balance is to use the ϵ-greedy selection [18]. However, this approach selects equally among the available actions i.e. (the worst action is likely to be chosen as the best one). In order to overcome this drawback, the action selection probabilities are varied as a graded function of the Q-value. The best power level is given the highest selection probability while all other levels are ranked according to their Q-values. The learning algorithm exploits Boltzmann probability distribution to determine the probability of the resource allocation action that fulfill the energy efficiency maximization constraints in C.1 to C.5. Thus, the action y in state x is selected at t with the following probability,

$$\pi_n^t(x,y) = \frac{e^{Q^t(x,y)/\tau}}{\sum_{y' \in Y} e^{Q^t(x,y')/\tau}} \tag{16}$$

where τ is a positive integer that controls the selection probability. With high value of τ, the action probabilities become nearly equal. However, low value of τ causes big difference in selection probabilities for actions with different Q-values. One issue to report is that the 5G H-CRAN system has a large space. Therefore, the curse of dimensionality increases the required computations and makes it unfeasible to use the typical online learning methodology to maintain the Q-value for each state/action pair, which slows the system convergence.

V. Centralized Approximated Online Learning Resource Allocation Scheme

The resource allocation process is performed at a dedicated controller that is integrated with the BBU pool and the macro BSs act as brokers between the controller and the RRHs for control exchange. Resource allocation and data processing signals from the controller to the macro BSs are sent through X1 and S1 interfaces respectively, which are obtained from definitions of the Third-Generation Partnership Project (3GPP) standards. All the agents in the network including RUEs, MUEs, RRHs report their channel information to the macro BS that they operate under its coverage in a hierarchical manner. The channel information includes path loss and channel gains from the serving RRH and the macro BS to the RUEs and MUEs. All the macro BSs provide this information to the controller through the control exchange interface. The controller's exploitation of the reported information and QoS requirements as prior knowledge to enhance online learning, facilitates the proper selection of RBs and transmission power. The allocation decision made by the controller is sent to the RRH through the macro BS. Note that SINR is the state information exploited in action selection and it is determined according to the channel information reported and the allocated transmission power in previous learning processes.

The computational complexity of the system increases along with the size of the states and action spaces. The simple look-up table where separate Q-value is maintained for each state/action pair is not feasible in large space with massive number of states like our system. Therefore, we propose a brief representation for the Q-values in which they are approximated as a function of much smaller set of variables to account for the curse of dimensionality. The brief representation of Q-value focuses on a countable state space X^* using the function $Q' : X^* \times Y$ which is referred as a function approximator. The parameter vector $\xi = \{\xi_z\}_{z=1}^{Z}$ is adopted to approximate the Q-value by minimizing the metric of difference between $Q^*(x,y)$ and $Q'(x,y,\xi)$ for all $(x,y) \in X^* \times Y$. Thus, the approximated Q' value is formulated as follows,

$$Q'(x,y,\xi) = \sum_{z=1}^{Z} \xi_z \psi_z(x,y) = \xi\psi^T(x,y) \tag{17}$$

where T denotes the transpose operator and the vector $\psi(x,y) = [\psi_z(x,y)_{z=1}^{Z}]$ with a scalar function $\psi_z(x,y)$ defined as the basis function (BF) over $X^* \times Y$, and $\xi_z(z = 1, ..., Z)$ are the associated weights. A gradient function $\psi(x,y)$, which is a vector of partial derivative with respect to the elements of ξ^t, is used to combine the typical online learning model with the linearly parametrized approximated online learning proposed.

The Q-value update rule in (15) is reconstructed to include the parameter vector updates as follows,

$$\xi^{t+1}\psi^T(x,y) = \big\{(1-\alpha^t)\xi^t\psi^T(x,y) + \alpha^t\big[EE(x,y) + \beta \max_{y' \in Y} \xi^t\psi^T(x',y')\big]\big\}\psi(x,y) \tag{18}$$

The probability of selecting certain action presented in (16) is updated with the Q-value approximation as follows,

$$\pi^t(x,y) = \frac{e^{\xi^t\psi^T(x,y)/\tau}}{\sum_{y' \in Y} e^{\xi^t\psi^T(x,y)/\tau}} \tag{19}$$

The online learning process with approximated Q-value is illustrated in Algorithm 1. The algorithm takes the QoS requirements for RUEs and MUEs as input to check the quality of the strategies selected and they are compared to the capacity achieved by different BSs. The algorithm select actions strategies according to (19) and the achieved capacities are checked. If the conditions C.1 to C.5 are satisfied, then, the reward is achieved. Finally, the Q-value and resource allocation strategy are updated according to (18) and (19) respectively, and the new state is observed.

Algorithm 1 Centralized approximated online learning algorithm for resource allocation

```
Data: $\pi^t(x,y)$, $t = 1$, $\delta$, $\theta^*$, $\theta^*$, $\gamma_{s,n}^k$, $\gamma_{u,m}^k$
Result: RB and $P_{s,n}^k$ allocation for RUE
initialization of Learning
 for each($x$, $y \in Y'$) do
    initialize resource allocation strategy $\pi^t(x,y)$;
    initialize approximated Q-value $\xi^t\psi^T(x,y)$;
end
while (true) do
    evaluate the state $x = x^t$
    Select action $y$ according to $\pi^t(x,y)$ in (19);
    Check $C_s$ and $C_u$
    if (C.1 to C.5 are satisfied ) then
        $R(x,y)$ is achieved
    else
        $R(x,y) = 0$
    end
    Update $\xi_i^{t+1}\psi^T(x,y)$ according to (18)
    Update $\pi_i^{t+1}(x,y)$ according to (19)
    $x = x^{t+1}$
    $t = t + 1$
end
```

To demonstrate Algorithm 1 convergence, we found the necessary conditions for convergence of the proposed approximated learning resource allocation. To start the proof, we introduce the following definition and assumptions.

Definition 1: Let $\Psi = E[\psi^T(x,y)\psi(x,y)]$. For the parameter vector ξ and a particular network state $x \in X^*$, we define a vector $\psi(x,\xi) = [\psi_z(x,y)]$ for $z = 1 \rightarrow Z$ where $y \in \{y = arg\ \max_{y' \in Y} \xi\psi^T(x,y')\}$ is the set of optimal joint resource allocation actions for x. We define the following a ξ-dependent matrix:

$$\Psi' = E[\psi^T(x,\xi)\psi(x,\xi)] \tag{20}$$

Assumption 1: The basis functions $\psi_z(x,y)$ are linearly independent for all (x,y) and all the properties of $Q^t(x,y)$ in previous discussion are applicable to the dot product for the vectors $\xi^t\psi^T(x,y)$.

Assumption 2: For every $z = (1, 2....Z)$, $\psi_z(x,y)$ is bounded, which means $E\{\psi_z^2(x,y)\} < \infty$ and the reward function satisfies $E\{R^2(x,y)\} < \infty$.

Assumption 3: The learning rate satisfies $\sum_{t=1}^{\infty} \alpha^t = \infty$ and $\sum_{t=1}^{\infty} (\alpha^t)^2 < \infty$.

Proposition 1: With the *assumptions 1-3* and *Definition 1*, the approximated online learning algorithm converges with probability (w.p) 1, if

$$\Psi' < \Psi,\ \forall\xi \tag{21}$$

Proof:

The proof of convergence requires finding stable fixed points of the ordinary deferential equations (ODE) associated with the update rule in (18), which can be written as.

$$\xi^{\cdot t} = E[EE(x,y) + \beta\xi^t\psi^T(x',\xi^t) - \xi^t\psi^T(x,y))\psi(x,y)] \tag{22}$$

where $\xi^{\cdot t} = \frac{\partial\xi}{\partial t}$ as $\alpha \rightarrow 0$. We define two trajectories of the ODE ξ_1^t and ξ_2^t that have different initial conditions and satisfies $\xi_0^t = \xi_1^t - \xi_2^t$. Then, we have

$$\frac{\partial\|\xi_0^t\|^2}{\partial t} = 2(\xi_1^{\cdot t} - \xi_2^{\cdot t})(\xi_0^t)^T = 2\beta E\big[\xi_1^t\psi^T(x',\xi_1^t)\psi(x,y)(\xi_0^t)^T - \xi_2^t\psi^T(x',\xi_2^t)\psi(x,y)(\xi_0^t)^T\big] - 2\xi_0^t\Psi(\xi_0^t)^T \tag{23}$$

From *Definition 1*, we can deduce the following two inequalities,

$$\xi_1^t\psi^T(x',\xi_1^t) \leq \xi_1^t\psi^T(x',\xi_2^t) \tag{24}$$

$$\xi_2^t\psi^T(x',\xi_2^t) \leq \xi_2^t\psi^T(x',\xi_1^t) \tag{25}$$

As the expectation E in (23) is taken over different states and different actions, we can define two sets $\Lambda_+ = \{(x,y) \in X \times Y | \xi_0^t\psi^T(x,y) > 0\}$ and $\Lambda_- \in X \times Y - \Lambda_+$. If we combine (24) and (25) in (23), we get,

$$\frac{\partial\|\xi_0^t\|^2}{\partial t} \leq 2\beta\big(E\big[\xi_0^t\psi^T(x',\xi_2^t))\psi(x,y)(\xi_0^t)^T|\Lambda_+\big]$$

$$+E\big[\xi_0^t\psi^T(x',\xi_1^t))\psi(x,y)(\xi_0^t)^T\big|\Lambda_-\big]\big) - 2\xi_0^t\Psi(\xi_0^t)^T \quad (26)$$

After the application of Holder's inequality [19] to the expectation in (26), we get,

$$\frac{\partial\|\xi_0^t\|^2}{\partial t} \le 2\beta\Bigg(\sqrt{E\big[\big(\xi_0^t\psi^T(x',\xi_2^t)\big)^2\big|\Lambda_+\big]}\times \sqrt{E\big[\big(\psi(x,y)(\xi_0^t)^T\big)^2\big|\Lambda_+\big]} + \sqrt{E\big[\big(\xi_0^t\psi^T(x',\xi_1^t)\big)^2\big|\Lambda_-\big]} \times\sqrt{E\big[\big(\psi(x,y)(\xi_0^t)^T\big)^2\big|\Lambda_-\big]}\Bigg) - 2\xi_0^t\Psi(\xi_0^t)^T$$

$$\le 2\beta\Bigg(\sqrt{E\big[\big(\xi_0^t\psi^T(x',\xi_2^t)\big)^2\big]} \times \sqrt{E\big[\big(\psi(x,y)(\xi_0^t)^T\big)^2\big|\Lambda_+\big]} + \sqrt{E\big[\big(\xi_0^t\psi^T(x',\xi_1^t)\big)^2\big]} \times \sqrt{E\big[\big(\psi(x,y)(\xi_0^t)^T\big)^2\big|\Lambda_-\big]}\Bigg) -2\xi_0^t\Psi(\xi_0^t)^T$$

If we apply the definition of Ψ' in *Definition 1*, we get,

$$\le 2\beta\sqrt{max\big[\xi_0^t\Psi_1'(\xi_0^t)^T, \xi_0^t\Psi_2'(\xi_0^t)^T\big]} \times\sqrt{E\big[\big(\psi(x,y)(\xi_0^t)^T\big)^2\big]} + -2\xi_0^t\Psi(\xi_0^t)^T \quad (27)$$

According to the condition in (21), we can state that,

$$\frac{\partial\|\xi_0^t\|^2}{\partial t} < 2\beta\xi_0^t\Psi(\xi_0^t)^T - 2\xi_0^t\Psi(\xi_0^t)^T = (2-\beta)\xi_0^t\Psi(\xi_0^t)^T < 0 \quad (28)$$

which means that ξ_0^t converges to the origin and this confirms that there exists a stable point of the ODE in (22). Thus, the proposed online learning with Q approximation converges w.p 1. ■

Consequently, the stable point $\xi^*$of the ODE in (22) indicates that,

$$0 = E[EE(x,y)+\beta\xi^t\psi^T(x',\xi^t)-\xi^t\psi^T(x,y))\psi(x,y)] \quad (29)$$

and ξ^* can be found as follows,

$$\xi^* = E[EE(x,y)+\beta\xi^*\psi^T(x',\xi^*))\psi(x,y)]\,\Psi^{-1} \quad (30)$$

As a result, the approximated online Q-function is stated as follows,

$$Q'(x,y,\xi^*) = \xi^*\psi(x,y) \quad (31)$$

VI. Numerical Results

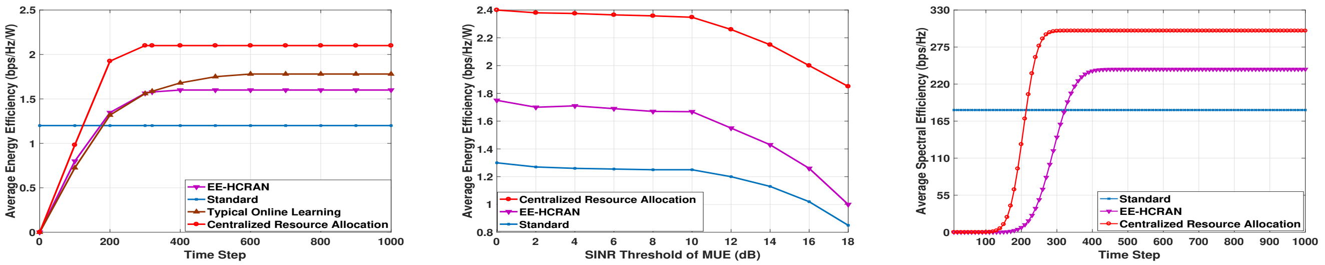
In this section, we verify the performance of the proposed scheme in terms of energy efficiency, spectral efficiency, and QoS guarantee. The evaluation environment consists of three macro BSs with 21 MUEs, 15 RRHs with 47 RUEs accessing Γ_1 and 28 RUEs sharing the spectrum with MUEs in Γ_2. It is assumed that the path-loss model is expressed as $31.5 + 40 * log_{10}(d)$ for RRH to RUE link and $31.5 + 35 * log_{10}(d)$ for macro BS to RUE and RRH to MUE links where d is the distance between the transmitter and receiver. Fast-fading coefficients are all generated as independent and identically distributed Rayleigh random variables with unit variances. The data rate thresholds per both types of RUEs and MUEs are assumed to be 2 Mbps, 512 Kbps, and 4 Mbps, respectively. The rest of the simulation parameters are presented in Table 1.

Parameter	Value
User distribution	uniform
Number of RBs	100
Total bandwidth	20 MHz
Thermal noise power	-112 dBm
Macro BS transmission power	43 dBm
Back-haul power consumption P_{bh}	23 dBm
Antenna gain for macro/RRH	17/6 dB
RRH maximum transmission power	25 dBm
Trials per experiment	1000

TABLE I
System Parameters

The performance of our scheme is compared to two schemes including the standard with fixed power allocation for all RBs and the scheme proposed in [14], which aims at tackling the energy efficiency problem in H-CRAN denoted by (EE-HCRAN). Moreover, we include online learning resource allocation without compact-state representation to demonstrate the advantage of our enhanced online learning in the speed of convergence. The evaluation of the proposed scheme includes energy efficiency, spectral efficiency, and users achieved data rate. First, we plot energy efficiency as function of the number of time steps in Fig. 2 (a). We noticed that our scheme converges after 300 iterations faster than typical online learning and EE-HCRAN and achieved the highest energy efficiency. The second energy efficiency evaluation focuses on the system with variable SINR threshold of MUEs. Fig. 2 (b) presents the energy efficiency achieved versus the data rate threshold of MUEs accessing Γ_2 with $P_{max} = 25$ dBm. We notice that the proposed scheme outperforms both EE-HCRAN and the standard schemes. Fig. 2 (b) also reveals that when the threshold is not large, the energy efficiency is stable with the increasing threshold because the inter-tier interference is not severe thanks to the proposed RB allocation strategy, which considers both location and QoS of the RUEs. Moreover, appropriate power allocation has a considerable contribution to the achieved energy efficiency.

In addition, we evaluate the proposed scheme performance in terms of spectral efficiency and QoS represented by the data rate achieved by both RUEs accessing Γ_1 and Γ_2 respectively and MUE. Thus, we plot the average system spectral efficiency against the time steps Fig. 2 (c). The figure emphasizes the speed of convergence achieved by our scheme compared to others where our scheme is the fastest with the highest spectral efficiency. QoS requirements stated in C.2 and C.3 are investigated in this evaluation. Fig. 3 (a), (b) and (c) present the commutative distribution function (CDF) of the data rate for both RUEs accessing sub-band Γ_1 and Γ_2, and the CDF of the data rate achieved by MUEs respectively. We noticed that our scheme is the only scheme that managed to have more than 97 % of the users above the specified thresholds in the figure compared to EE-HCRAN that records 78 % above threshold, and standard with 65 %. This evaluation demonstrates the capability of the online learning scheme to allocate RB and transmission power efficiently while maintaining QoS of users at the maximum level.

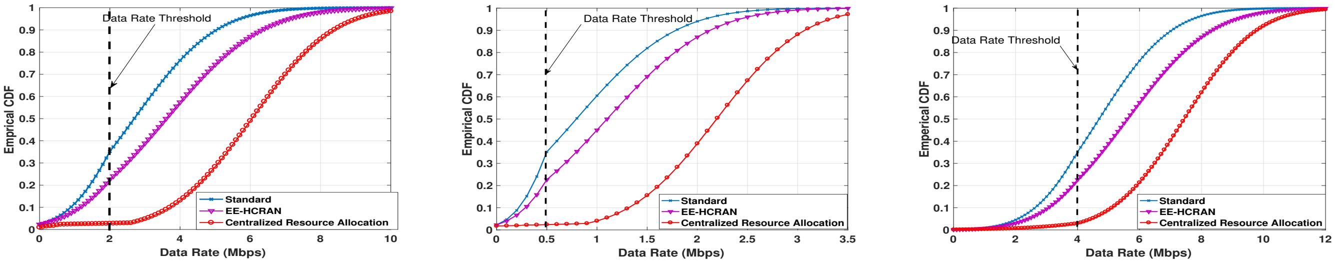


(a) Energy efficiency convergence

(b) Energy efficiency with variable SINR threshold for MUEs

(c) Spectral efficiency convergence

Fig. 2. Energy efficiency and spectral efficiency evaluation results



(a) Data rate CDF for RUEs accessing Γ_1

(b) Data rate CDF for RUEs accessing Γ_2

(c) Data rate CDF for MUEs accessing Γ_2

Fig. 3. Data rate evaluation results

VII. Conclusion

In this paper, a green resource allocation scheme in H-CRAN network is proposed. RBs and transmission power are allocated subjected to the inter-tier interference and capacity constraints. Sophisticated frequency partitioning is proposed to account for the inter-interference and support better online learning exploration. In addition, approximated online learning methodology is exploited to perform joint allocation for RB and transmission power. Simulation results including energy efficiency, spectral efficiency and data rate have demonstrated the capability of the proposed scheme to allocate resources in green and suppressed interference fashion. Moreover, the approximation in the proposed online learning algorithm increases the convergence speed compared to the standard online learning and other schemes.

References

[1] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5g wireless networks: A comprehensive survey", *IEEE Communications Surveys Tutorials*, vol. 18, no. 3, pp. 1617–1655, thirdquarter 2016.

[2] C. L. I, C. Rowell, S. Han, Z. Xu, G. Li, and Z. Pan, "Toward green and soft: a 5g perspective", *IEEE Communications Magazine*, vol. 52, no. 2, pp. 66–73, February 2014.

[3] M. Peng, D. Liang, Y. Wei, J. Li, and H. H. Chen, "Self-configuration and self-optimization in lte-advanced heterogeneous networks", *IEEE Communications Magazine*, vol. 51, no. 5, pp. 36–45, May 2013.

[4] M. Peng, Y. Li, J. Jiang, J. Li, and C. Wang, "Heterogeneous cloud radio access networks: a new perspective for enhancing spectral and energy efficiencies", *IEEE Wireless Communications*, vol. 21, no. 6, pp. 126–135, December 2014.

[5] Y. Shi, J. Zhang, and K. B. Letaief, "Group sparse beamforming for green cloud-ran", *IEEE Transactions on Wireless Communications*, vol. 13, no. 5, pp. 2809–2823, May 2014.

[6] Y. Cai, F. R. Yu, and S. Bu, "Dynamic operations of cloud radio access networks (c-ran) for mobile cloud computing systems", *IEEE Transactions on Vehicular Technology*, vol. 65, no. 3, pp. 1536–1548, March 2016.

[7] D. W. K. Ng and R. Schober, "Secure and green swipt in distributed antenna networks with limited backhaul capacity", *IEEE Transactions on Wireless Communications*, vol. 14, no. 9, pp. 5082–5097, Sept 2015.

[8] M. Peng, Y. Li, T. Q. S. Quek, and C. Wang, "Device-to-device underlaid cellular networks under rician fading channels", *IEEE Transactions on Wireless Communications*, vol. 13, no. 8, pp. 4247–4259, Aug 2014.

[9] Y. Dong, Z. Chen, P. Fan, and K. B. Letaief, "Mobility-aware uplink interference model for 5g heterogeneous networks", *IEEE Transactions on Wireless Communications*, vol. 15, no. 3, pp. 2231–2244, March 2016.

[10] Z. Fei, C. Xing, N. Li, and J. Kuang, "Adaptive multiobjective optimisation for energy efficient interference coordination in multicell networks", *IET Communications*, vol. 8, no. 8, pp. 1374–1383, May 2014.

[11] C. Y. Ho and C. Y. Huang, "Energy efficient subcarrier-power allocation and relay selection scheme for ofdma-based cooperative relay networks", in *2011 IEEE International Conference on Communications (ICC)*, June 2011, pp. 1–6.

[12] Y. Wang, W. Xu, K. Yang, and J. Lin, "Optimal energy-efficient power allocation for ofdm-based cognitive radio networks", *IEEE Communications Letters*, vol. 16, no. 9, pp. 1420–1423, September 2012.

[13] R. Q. Hu and Y. Qian, "An energy efficient and spectrum efficient wireless heterogeneous network framework for 5g systems", *IEEE Communications Magazine*, vol. 52, no. 5, pp. 94–101, May 2014.

[14] M. Peng, K. Zhang, J. Jiang, J. Wang, and W. Wang, "Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks", *IEEE Transactions on Vehicular Technology*, vol. 64, no. 11, pp. 5275–5287, Nov 2015.

[15] Y. Zhang, Y. Wang, and W. Zhang, "Energy efficient resource allocation for heterogeneous cloud radio access networks with user cooperation and qos guarantees", in *2016 IEEE Wireless Communications and Networking Conference*, April 2016, pp. 1–6.

[16] Richard S. Sutton and Andrew G. Barto, *Introduction to Reinforcement Learning*, MIT Press, Cambridge, MA, USA, 1st edition, 1998.

[17] C. J. C. H. Watkins and P. Dayan, "Q-learning", *Mach. Learn*, vol. 8, no. 3, pp. 279–292, 1992.

[18] Eduardo Rodrigues Gomes and Ryszard Kowalczyk, "Dynamic analysis of multiagent q-learning with -greedy exploration", in *Proceedings of the 26th Annual International Conference on Machine Learning*, New York, NY, USA, 2009, ICML '09, pp. 369–376, ACM.

[19] "On nonlinear fractional programming", *Management Science*, vol. 13, no. 7, pp. 492–498, 1967.