Grouping and Cooperating among APs for Energy Efficiency in 5G UUDN

Chunhong Duo, Yongqian Li, Baogang Li, and Yabo Lv

***Abstract*—The user-centric ultra-dense network (UUDN) is considered as a promising technology for 5G. However, the massive deployment of access points (APs) would lead to a considerable increase in energy consumption. Considering user's different service flow and system energy efficiency, we propose a user-centric access algorithm through renewable energy cooperation. Firstly, the access point group (APG), which consists of several APs, is dynamically organized to serve each user in UUDN. Then, to maximize the system energy efficiency, we propose a reinforcement learning approach to cooperate renewable energy. Q neural network which adopts a three-layer BP neural network solves the problem of Q learning in continuous state and discrete action. Meanwhile, by optimizing the resource allocation in a cooperative way, the proposed algorithm compared to the existing algorithms has better performance in satisfying user's demand and improving system energy efficiency.**

***Index Terms*—Energy cooperation, energy harvesting, reinforcement learning, user-centric ultra-dense network (UUDN).**

I. Introduction

With the rapid development of mobile network of 5G, ultra-dense network (UDN) has become hot research field. More and more access points (APs) are needed to satisfy the user's traffic demands, and the number of APs is comparable to the users. User-centric UDN (UUDN) is proposed as an ideal architecture of UDN [1], [2], which is quite different from traditional network-centric cellular architecture. UUDN lets user equipment (UE) feel like a network always following itself. Hence, the network will intelligently recognize the UE's wireless communication environment, and then flexibly organize the required access points group (APG) to serve the UE [3]. To serve each UE seamlessly, authors put forward a maximum data transmission rate oriented dynamic APs grouping scheme, and the APG member will be dynamically refreshed according to the UE's movement or network environment change [4]. Because of the limited radio resources, non-orthogonal multiple access is introduced into UUDN, and an access scheme by grouping multiple APs cooperatively to provide a user-centric access service has been considered [5], [6]. Considering user's different service flow and system energy efficiency, Authors propose a service-driven resource allocation scheme for UUDN and adjust the number of APs dynamically to serve users according to their service flow [7]. Due to heavy CSI feedback overhead, a general MIMO network using a hybrid interference coordination approach is considered, AP load by joint AP grouping and user association is minimized [8].

In order to reduce the operating expenses of these APs, especially minimize energy consumption, energy harvesting (EH) has been considered as a promising energy source [9], [10]. The EH capability of APs increases the network lifetime because APs can use the harvested energy to recharge their batteries [11], [12]. Energy harvesting UUDN (EH-UUDN) and the energy cooperation technology [13], [14] have been developed and widely researched. In EH-UUDN, each AP can harvest energy (e.g., solar and wind power) from the ambient environment. By employing an energy transceiver, each AP can also transmit some energy to other nodes in one time slot and receive energy from others in another time slot, so that the utilization of the available energy over the network could be optimized [15]. In [16], it studies energy cooperation and traffic management using the Lyapunov optimization framework. [15] proposes that users can perform energy cooperation, and the capacity region coincides with that of a traditional *K*-user Gaussian MAC with energy cooperation. In [17], joint renewable energy cooperation and resource allocation for cloud radio access networks with hybrid power supplies (including both the conventional grid and renewable energy sources) is investigated. [18] focuses on resource allocation in energy cooperation enabled two-tier heterogeneous networks with non-orthogonal multiple access, where base stations are powered by both renewable energy sources and the conventional grid. Authors propose a reinforcement learning approach based on Q-learning for the transmitter to learn to cooperation through energy sharing [19], [20], and the reinforcement learning approach is optimized by waterflooding [21]. The Q-network is diverged by combining model-free reinforcement learning algorithms with non-linear function approximators [22], or with off-policy learning [23]. More recently, there has been a revival of interest in combining deep learning with reinforcement learning [24]-[27].

In this paper, a user-centric access service algorithm is proposed which groups multiple APs into one APG for energy cooperation. We first set up AP grouping scheme and select the best servicing AP for one UE in EH-UUDN. Considering as Markov decision process (MDP), energy cooperation problem is then formulated based on Q-learning.

Manuscript received January 4, 2019; revised June 2, 2019.

This work was supported by National Natural Science Foundation of China (No.61775057), Beijing Natural Science Foundation (No. 4164101), Hebei Province Natural Science Foundation (No. F2016502062, F2019502168), and the Fundamental Research Funds for the Central Universities (2015MS125).

The authors are with Department of Electronic and Communications Engineering, North China Electric Power University, Hebei, China (e-mail: duochunhong@163.com, liyongqian1958@163.com, libgbd@163.com, yabolv@163.com).

doi: 10.18178/ijfcc.2019.8.3.543

To solve the problem of reinforcement learning in continuous state and discrete action, we utilize a Q neural network (QNN) which applies a three-layer BP neural network as approximator. The QNN is trained through minimizing a sequence of loss functions that change at each iteration. ε greedy policy can ensure the convergence of the proposed algorithm, and such value iteration algorithm converges to the optimal action-value function, $Q_i \rightarrow Q^*$ as $i \rightarrow \infty$. Simulation results show that the system EE is related to the density of APs and UEs, and the proposed algorithm compared to the existing algorithms achieves a better resource utilization and improves system EE in a cooperative way.

The rest of this paper is organized as follows: In section II, the system model is first introduced. APG scheme for UE in EH-UUDN is described in section III. In section IV, we optimize the problem of energy cooperation through QNN. Simulation results are given in section V. And finally, we draw a conclusion and summarize our work in section VI.

II. System Model

We consider a downlink EH-UUDN where UEs and APs are randomly located. Each AP is equipped with EH unit and rechargeable battery, and is solely powered by renewable energy sources. Time is slotted and time slot (TS) length is T. Assume that each AP is also equipped with an energy transmitting unit for transmitting some of the harvested energy to other APs, and an energy receiving unit for receiving the energy transmitted by other APs. Let $E_i(t)$ denote the amount of energy that AP i harvests in $TS \quad t$, and $B_i(t)$ is the battery capacity of AP i in $TS \quad t$. Suppose the channel state information is $H_i(t)$, which is kept constant in the same TS. In EH-UUDN, it meets that λ_{AP} is comparable to λ_{UE}, where λ_{AP} and λ_{UE} respectively represent the density of APs and UEs [4], as shown in Fig. 1.

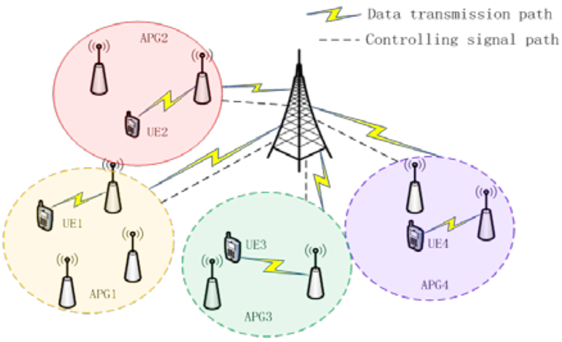


Fig. 1. The system model of UEs and APs in EH-UUDN.

To investigate the performance limit of the network, we consider a subset of the following assumptions (A1-A5).

A1: The energy buffer at each AP is finite, and $B_{\max}$ represents maximum capacity of the battery.

A2: The data buffer at each AP is finite with maximum size $D_{\max}$ to store the incoming data, and the data flow arrives to the data buffer in a stochastic and continuous way.

A3: For each AP, the expectation of the harvested energy in each TS is finite,and the maximum harvested energy is $E_{\max}$. $\{E_i(t), t = 0,1,2,...\}$ is an ergodic, independent and identically distributed sequence.

A4: At any time, AP can transmit energy to other APs or receive energy from other APs. There is at most one selection of charging/discharging energy to/from the battery.

A5: The harvested and transferred energy at $TS \ t$ will be utilized at $TS \quad t+1$.

In EH-UUDN, the coverage ratio is the biggest when all APs are working but apparently it is a waste of energy. As shown in Fig. 2, we consider that AP has three modes: on, sleeping and off. Accordingly, the energy conditions of AP are divided into three situations:

① AP i is qualified to be on when its battery capacity in $TS \quad t$ satisfies $B_i(t) \geq B_{sleep}$, and users can access to it;

② When its current energy satisfies $B_{off} \leq B_i(t) < B_{sleep}$, AP i turns to sleeping mode which can continue to harvest energy and receive energy from other APs, and there is no access to it;

③ No matter in on or sleeping mode AP should be powered off automatically and enter into off mode when its current energy satisfies $B_i(t) < B_{off}$, and wait for the energy replenishment.

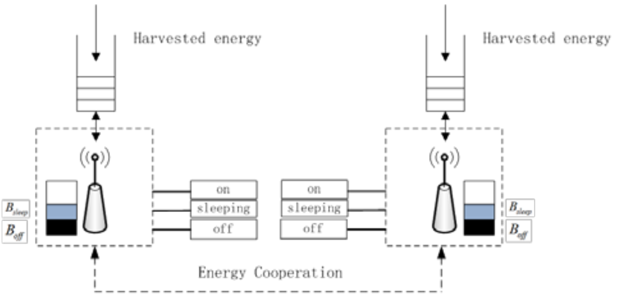


Fig. 2. Three energy mode for Aps.

When the AP in the off mode and the harvested energy from the ambient is zero (for example, the solar energy harvester cannot obtain the energy during the dark night). That is, AP consumes the basic power consumption per time unit while obtains nothing, which is greater than or equal to zero. In this case, AP will meet energy depletion, and the temporal death of the AP will be occurred. In this scenarios, it will get $B_i(t+1) < 0$, which is not correct for practice application. So the battery energy queue length of AP i is as follows ($[x]^+ = \max\{0, x\}$):

$$B_i(t+1) = [B_i(t) - P_i(t) + E_i(t) + \chi c_i(t) - d_i(t)]^+ \qquad (1)$$

The energy of transmitting data at time t is $P_i(t) * (1TS)$ (we omit the implicit multiplication by $1TS$ of $P_i(t)$ when converting between power and energy). The energy charged to the battery is $c_i(t)$, and the energy discharged from the battery is $d_i(t)$. At any time, AP will charge/discharge energy to/from battery. There is at most one of $c_i(t)$ and $d_i(t)$ that is strictly positive, that is, $c_i(t) * d_i(t) = 0$. $\chi \in [0,1]$ is the energy transfer efficiency between two APs.

In summary, the operation of AP i in TS t satisfes the following constraints C1:

$$\begin{cases} 0 \le B_i(t) \le B_{\max} \\ 0 \le E_i(t) \le E_{\max} \\ 0 \le D_i(t) \le D_{\max} \\ c_i(t) \ge 0 \\ d_i(t) \ge 0 \\ c_i(t) * d_i(t) = 0 \\ B_{i+1}(t) = [B_i(t) - P_i(t) + E_i(t) + \chi c_i(t) - d_i(t)]^+ \end{cases} \tag{C1}$$

AP's EE should be considered in detail, which is defined as the sending data ($R_i(t)$) divided by the power consumption of APs. For each AP i, we take two parts into account. P_i^0 is the basic power, and P_i^T is the transmit power.

When the AP is on ($\rho = 1$) and serves some UE, $P_i = P_i^0 + \Delta P_i^T$, where Δ represents the power consumption of feeders and power amplifier of AP i.

When the AP is sleeping ($\rho = 1$) and serves no UE, $P_i = P_i^0$.

When the AP is off ($\rho = 0$) because of low energy, $P_i = \alpha P_i^0, 0 < \alpha < 1$. Actually, AP in off mode consumes approximately one tenth of the basic power for common control signaling.

In conclusion, AP power can be expressed as

$$P_i(\rho) = \Delta P_i^T \rho + (1-\alpha) P_i^0 \rho + \alpha P_i^0 \tag{2}$$

Hence, EE of AP i is represented as

$$EE_i(\rho) = \frac{R_i(\rho)}{\Delta P_i^T \rho + (1-\alpha) P_i^0 \rho + \alpha P_i^0} \tag{3}$$

EE of all APs based on EH-UUDN (I is the AP set in the hot region) can be concluded:

$$EE = \frac{\sum_{i \in I} R_i(\rho)}{\sum_{i \in I} (\Delta P_i^T \rho + (1-\alpha) P_i^0 \rho + \alpha P_i^0)} \tag{4}$$

When ρ=1, there are two situations: ① the on mode of AP: $R_i(\rho)$ is the actual sending data; ②the sleeping mode of AP: $R_i(\rho) = 0$. When $\rho = 0$, $R_i(\rho) = 0$.

III. AP Grouping Scheme in EH-UUDN

A. AP Grouping in EH-UUDN

Each UE u calculates the number of APs in the user centered circle with a radius of r, and store the APs from which the user received signal strength are greater than $P_{u\max} - \eta$, where $P_{u\max}$ is the max received signal strength by UE u from APs, and η is the gap to $P_{u\max}$, which is decided by the traffic demand of UE, so it may be different for different users. The received signal strength of UE u from AP i is $P_{u,i} = P_i H d_{u,i}^{-e}$, where P_i is the transmit power of AP i, e is the path-loss exponent, $d_{u,i}$ is the distance between UE u and AP i, and H is the Rayleigh fading. These APs that satisfy $P_{u,i} \ge P_{u\max} - \eta$ form the potential serving AP group of UE u, which is marked as PG_u. Every AP requires other APs who serve the same UE within distance r to avoid interference.

TABLE I: THE ARRANGEMENT OF CHANNELS

```
Algorithm 1: Grouping APs in EH-UUDN
Input: {The AP set: $I$, the UE set: $U$, the radius: $r$, the gap: $\eta$}
Output: {the serving AP group for each UE: $G_u$}
1) Initialize: mark all APs and UEs, determine the value of $\eta$ and $r$,
   $PG_u = \varnothing$, $G_u = \varnothing$, calculate $P_{u\max}$
2)For $u$=0 to $\mathbb{N}(U)-1$ do
    calculates $S$ (the number of APs in the user centered $r$ circle)
        For s=0 to $S-1$
            if $P_{u,i_s} \ge P_{u\max} - \eta$
            then store the AP $i_s$ in $PG_u$
    End For
    End For
3)For $i$=0 to $\mathbb{N}(PG_u)-1$ do
    if $P_i \ge \frac{1}{2}(\sum_{u \in U'} P_u)$, $U'$ represents the user set who want to access to
   AP $i$
        then $G_u = G_u \cup i$
        else get rid of the AP $i$ from the potential group $PG_u$
    End For
4)For $u$=0 to $\mathbb{N}(U)-1$ do
        if $G_u = \varnothing$
        then select $P_{u,i} = P_{u\max}$, $G_u = G_u \cup i$
    End For
```

For AP i, if its capacity satisfies $P_i > \frac{1}{2}(\sum_{u \in U'} P_u)$, where U' represents the user set who want to access to AP i, then choose all users that want to access to it, else select the top n users from whom the capacity of AP i is greater than half of their sum of traffic demands, that is, $P_i > \frac{1}{2}(\sum_{u \in N} P_u)$, where N represents the top n user set. Each user will decide its serving AP group through getting rid of APs which reject its request from the potential group. If the user has no potential group, then it should access to the strongest AP from those who have remaining capacity. The group for UE u is marked as G_u, and the AP number in G_u is $\mathbb{N}(G_u) = n_u$, where $\mathbb{N}(.)$ represents the number of elements in a set. The algorithm for AP grouping is in Table I.

B. Selecting the Best Serving AP

Suppose there are $|I|$ APs in the group G_u centered by the UE u, which expressed as $G_u = \{b_1, b_2, ..., b_{|I|}\}$ .The SINR of the UE u associated with AP i is

$$SINR_{u,i} = \frac{\rho P_{u,i} H d_{u,i}^{-e}}{\sum_{j \neq i} \rho P_{u,j} H d_{u,j}^{-e} + \sigma^2} \tag{5}$$

where σ^2 is the white noise power, $d_{u,i}^{-e}$ represents the path loss between AP i and UE u.

The potential serving AP for UE u should meet $SINR_{u,i} \geq SINR_{threshold}$, where $SINR_{threshold}$ is the predefined threshold. In this paper, we set $SINR_{threshold} = \frac{SINR_{\max} + SINR_{\min}}{2}$, $SINR_{\max}$ and $SINR_{\min}$ are the maximum and minimum SINR in the serving APG respectively.

According to the Shannon capacity, the achievable data rate of user u associating with AP i is

$$R_i = W_u \log_2(1 + SINR_{u,i}) \tag{6}$$

where W_u is the bandwidth used by AP i for its associated UE u.

AP need to estimate their load beforehand, and the estimation should accurately reflect the actual load. The load of AP i at time t based on history is $L_i(t)$, which is expressed by

$$L_i(t) = \Upsilon(t) L_i(t-1) + (1 - \Upsilon(t)) L_i(t-2) \tag{7}$$

where $\Upsilon(t)$ is the learning rate of the load estimation,and $L_i(0) = 0$.

The remaining renewable energy of AP i is $B_i(t)$, then the AP quality can be estimated by

$$Qua_i = \omega_1 * SINR_{u,i} + \omega_2 * L_i(t) + \omega_3 * B_i(t) \tag{8}$$

where $\omega 1, \omega 2, \omega 3$ is the weight of $SINR, L_i(t), B_i(t)$.

Each UE selects the AP of the biggest Qua value as the best serving AP.

IV. Energy Cooperation Based on AP Grouping

Reinforcement learning has been used for solving various optimization problems, and the traditional Q-learning, which can be assumed to be MDP model, considers quadruples: $(s_t, a_t, p_{a_t}(s_t, s_{t+1}), r_{a_t}(s_t, s_{t+1}))$. s_t belongs to the environment state space; a_t is the system action space; $p_{a_t}(s_t, s_{t+1}) \in [0,1]$ and $r_{a_t}(s_t, s_{t+1})$ respectively represent the state transition probability and the immediate reward of transferring the state from s_t to s_{t+1} by taking action a_t. The system does not need to know other priori information, and the algorithm can converge to the optimal strategy by learning to enhance the discount return value. The Q-value functions can be updated using the following equation:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + l[r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)] \tag{9}$$

where (s_t, a_t) is state-action pair in TS t in MDP. s_{t+1} is the state in TS $t+1$, r_t is the reward in TS t, l ($0 < l < 1$) is the learning factor which controls the convergence speed, and γ ($0<\gamma<1$) is the discount factor. We utilize the ε greedy policy, which enforces sporadic jumps to sub-optimal states for the exploration purposes. Whenever a decision is to be made, the one will be picked at random with the $(1-\varepsilon)$ probability, which is given to the action with the highest Q-value. Such value iteration algorithm converges to the optimal action-value function, $Q_i \to Q^*$ as $i \to \infty$ [28]. The convergence rate increases with the value of δ and the number of learning iterations N_L, and decreases with the number of a, s and γ [29].

Energy cooperation in EH-UUDN can be regarded as a system of multi-agent cooperation. We consider one APG in which APs are not in isolation, but mutual influence and mutual restriction. The traditional Q-learning algorithm uses a table to store the Q-value. However, there is infinite Q-value need to be stored because the state space is continuous. In view of this problem, neural network architectures is adopted to store the Q-value function, which solves the problem of reinforcement learning in continuous state and discrete action.

A neural network function approximator with weights ω is referred as a Q neural network (QNN). QNN can be trained by minimizing a sequence of loss functions that change at each iteration. In QNN, a three-layer BP neural network is utilized to improve the traditional Q-learning algorithm. The input parameters of the network are the state of APs within one APG, and the output is the Q-value for each possible action. The relationship between input and output parameters of the neural network is described as:

$$Q(s, a; \omega) = f_{QNN}(s_1, s_2, ..., s_{\mathbb{N}(APG)}) \tag{10}$$

The update of the value function is expressed as

$$Q(s, a; \omega) = Q(s, a; \omega) + e \tag{11}$$

The direct gradient descent method [30] is utilized to train the parameters of the BP network, and the error is defined as

$$e = r + l \max_{a} Q(s', a; \omega) - Q(s, a; \omega) \tag{12}$$

The loss function is

$$E(t) = \frac{1}{2}(e(t))^2 = \frac{1}{2}[r + l \max_{a} Q(s', a; \omega) - Q(s, a; \omega)]^2 \tag{13}$$

The network weight update rule is

$$\omega(t+1) = \omega(t) - \Delta\omega \tag{14}$$

$$\Delta\omega = -l \frac{\partial E(t)}{\partial \omega(t)} = le(t) \frac{\partial Q(s_t, a_t; \omega)}{\partial \omega(t)} \tag{15}$$

where $E(t) = \frac{1}{2}(e(t))^2$, $\frac{\partial Q(s_t, a_t; \omega)}{\partial \omega(t)}$ is gradient information, and l is the learning rate of network weights.

In the proposed model, the state of AP in TS t is $s_t = [B_{\mathrm{mod}\,e}, E(t), B(t), H(t)]$, which is formed by four components. $B_{\mathrm{mod}\,e}$ represents three AP modes: $[on, sleep, off]$, and the corresponding value is $[2, 1, 0]$,which is shown in Table II.

TABLE II: THREE MODES OF AP

The value of $B_i(t)$	The value of $B_{\mathrm{mod}\,e}$
$B_i(t) \geq B_{sleep}$	$B_{\mathrm{mod}\,e} = 2$
$B_{off} \leq B_i(t) < B_{sleep}$	$B_{\mathrm{mod}\,e} = 1$
$B_i(t) < B_{off}$	$B_{\mathrm{mod}\,e} = 0$

TABLE III: THE ARRANGEMENT OF CHANNELS

```
Algorithm 2: QNN algorithm
Initialize action-value function Q with random weights
Initialize $s_t = [B_{\mathrm{mod}\,e}, E(t), B(t), H(t)]$
repeat
    With probability $\varepsilon$ select a random action $a_t$
    otherwise select $a_t = \max_a Q^*(s_t, a; \omega)$
    Execute action $a_t$ and observe reward $r_t$
    Set $r_j = \begin{cases} r_j & \text{for terminal } s_{j+1} \\ r_j + \gamma \max_{a'} Q(s_{j+1}, a'; \omega) & \text{for non-terminal } s_{j+1} \end{cases}$
    According to (12) and (13)
    Perform a gradient descent step on (14) and (15)
    Set $s_{t+1} = s_t$
Until $s_{t+1}$ is terminal state
```

The joint action is $a_i(t) =< P_i(t), T_{ij}(t) >$, and the set of actions is $A(t) =< a_1(t), a_2(t), ..., a_{\mathbb{N}(APG)}(t) >$, where $\mathbb{N}(APG)$ represents the AP number in one APG. The reward is designed to accomplish the leading of energy cooperation. The main purpose of selecting the appropriate energy allocation strategy is to increase the system throughput of EH-UUDN. Thus, the system reward function is related to the rate of this time slot, which can be defined as

$$r(s_t, a_t) = \sum_{i=1}^{\mathbb{N}(APG)} R_i(t) \tag{16}$$

Energy cooperation through QNN is described in Table III.

V. SIMULATION RESULTS

In this section, the effectiveness of our proposed user centric QNN will be demonstrated. We consider a $1Km * 1Km$ square area in hot spot environment. A large amount of APs and users are modeled as independent homogeneous Poisson point process, and the channel state satisfies Rayleigh distribution, which is kept constant in the same TS. Each TS is $10ms$. At any time, AP either charges energy to battery or discharges energy from battery. The data flow arrives to the data buffer in a stochastic and continuous way. The harvested and transferred energy at TS t can be utilized at TS $t+1$. Set $B_{\max} = 5, D_{\max} = 5, E_{\max} = 2,$ and the basic configuration of the system simulation parameters are shown in the following Table IV and Table V.

TABLE IV: THE PARAMETERS OF AP GROUPING ALGORITHM USED IN SIMULATION

Parameters	Values
The radius of the user centered circle r	$20m$
The path-loss exponent e	4
The bandwidth W_u	$200kHz$
The learning rate of the load estimation $\Upsilon(t)$	0.9

TABLE V: THE PARAMETERS OF ENERGY COOPERATION ALGORITHM USED IN SIMULATION

Parameters	Values
The energy transfer efficiency χ	0.9
The discount factor γ	0.9
The learning rate of network weights l	0.005
The basic power of AP	$20mW$
λ_{AP}	$700users/Km^2$
λ_{UE}	$200users/Km^2$
Battery capacity $B_{\max}, B_{sleep}, B_{off}$	$100kJ, 20kJ, 10kJ$
Maximum harvested energy $E_{\max}$	$30kJ/TS$
Maximum amount of data $D_{\max}$	$1Mbit$

The network structure is a neural network of one hidden layer, and the input layer of the network is the state of AP $s_t = [B_{\mathrm{mod}\,e}, E(t), B(t), H(t)]$. The input layer has 5 neurons, and the hidden layer has 128 neurons, and the output layer has 10 neurons, which are corresponding to 10 discrete actions (transmitting power).

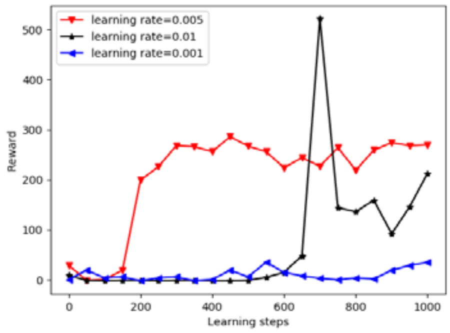


Fig. 3. Reward curve under different learning steps.

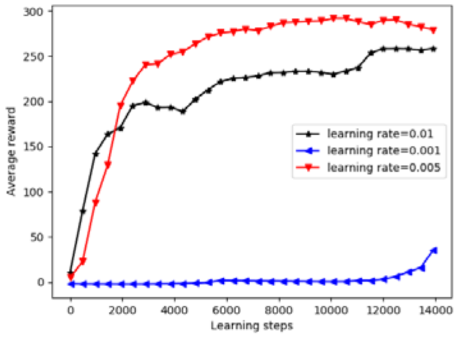


Fig. 4. System average reward curve under different learning steps.

The learning rate controls the loss size added to the parameters in each round of training. It is generally considered that the larger the learning rate is, the faster the algorithm is to reach the optimal value. However, the learning rate is too large to cause concussion near the optimal value, and the learning rate is too small to cause the low learning speed to reach the optimal value, which may not be convergent for a long time. The results are shown in Fig. 3, and the longitudinal axis is the reward for the corresponding learning step. In Fig. 4, $Average\_reward(i) = \frac{1}{T}\sum_{i=1}^{T} r(i)$.

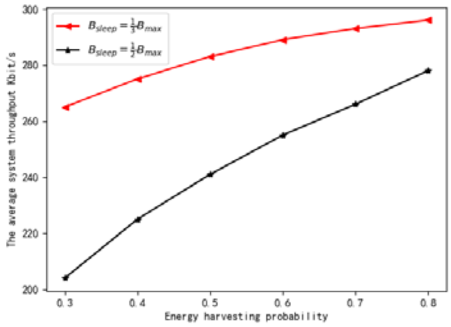


Fig. 5. Average system throughput under different energy harvesting probability.

When $B_{mode} = B_{sleep}$, AP cannot serve any UE and cannot send any data, so the system throughput is decreased as more and more APs turn to sleep mode. As shown in Fig.5, the two lines represent the average system throughput for different energy harvesting probability, and the average system throughput of $B_{sleep} = \frac{1}{3}B_{\max}$ is higher than that of $B_{sleep} = \frac{1}{2}B_{\max}$.

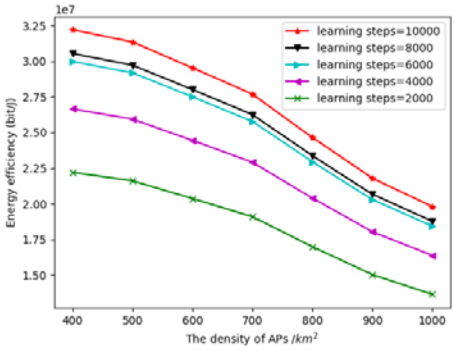


Fig. 6. The system EE versus the density of Aps.

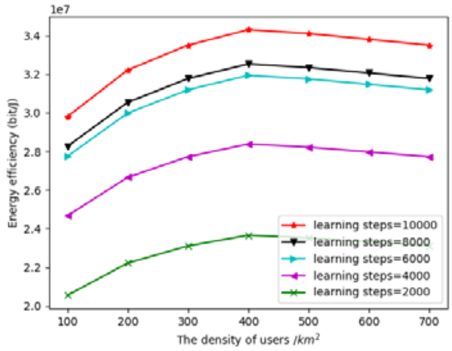


Fig. 7. The system EE versus the density of users.

Fig. 6 and Fig. 7 show the EE performance of the proposed algorithm versus various APs' density and Users' density. As the density of AP increases, on one hand, the inter-group and intra-group interference will lead to a decline in system throughput of the overall network. On the other hand, the more circuit power of APs will be consumed which makes the EE performance decrease gradually. As a result, all the curves go down gradually in Fig. 6. The EE performance firstly go up to a peak then go down gradually with the increase density of users in Fig. 7. When the user density is too small, the proportion of AP circuit power is increased, which results in low energy efficiency. As the number of users increased, the energy efficiency reaches the maximum value. The larger density of users will bring more receiver circuit energy consumption, which may cause EE performance decrease.

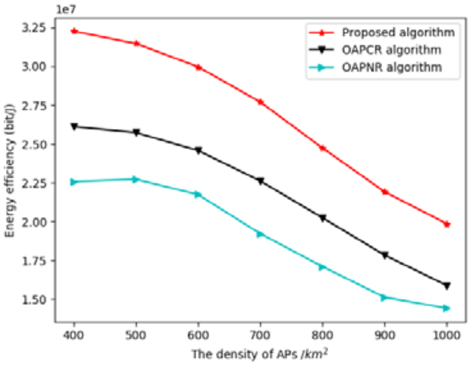


Fig. 8. The system EE of different algorithms versus the density of Aps.

To evaluate the EE performance, the proposed algorithm is compared to two typical access algorithms [5]: 1) an opportunistic APs with cooperative resource (OAPCR); 2) an opportunistic APs with noncooperative resource (OAPNR). Fig. 8 and Fig. 9 show the EE performance of different algorithms versus the various APs' density and users' density respectively. The OAPNR algorithm has the worst EE performance because it assigns the APs opportunistically and without cooperation for optimizing resource allocation, which results in serious interference and low EE. By optimizing resource allocation in a AP grouping and cooperative way, the proposed algorithm can achieve a better resource utilization and EE performance compared with other algorithms.

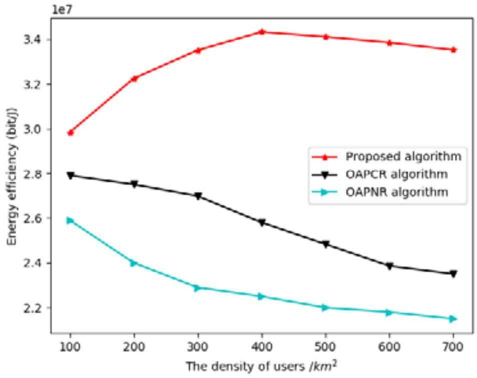


Fig. 9. The system EE of different algorithms versus the density of users.

VI. Conclusions

In this paper, we propose a energy cooperation algorithm based on AP grouping and reinforcement learning in EH-UUDN. Under the proposed algorithm, we formulate the problem as MDP model. Moreover, organizing multiple APs into APG, AP grouping algorithm is proposed to meet user-centric design in UUDN. Then, a reinforcement learning approach based on Q-learning is utilized. The major challenge of reinforcement learning is continuous state and discrete action. To address this problem, we propose QNN which uses a three-layer BP neural network, and is trained by minimizing a sequence of loss functions. Performance evaluation through extensive simulations shows that the system energy efficiency is related to the density of APs and UEs, and the proposed algorithm can achieve better resource utilization and improve system energy efficiency. At last, conclusions and future research directions are presented, which includes: mobility management of users, energy cooperation between clusters, combination of renewable energy and smart grid and etc.

References

[1] S. Chen, F. Qin, B. Hu, *et al.*, "User-centric ultra-dense networks for 5G: Challenges, methodologies, and directions," *IEEE Wireless Communications*, vol. 23, no. 2, pp. 78-85, 2016.

[2] R. A. Aljiznawi, N. H. Alkhazaali, *et al.*, "Quality of service (QoS) for 5G networks," *International Journal of Future Computer and Communication*, vol. 6, no. 1, pp. 27-30, 2017.

[3] H. Zhang and W. Huang, "Tractable mobility model for multi-connectivity in 5G user-centric ultra-dense networks," *IEEE Access*, vol. 6, pp. 43100-43112, 2018.

[4] B. Hu, Y. Wang, and C. Wang, "A maximum data transmission rate oriented dynamic APs grouping scheme in user-centric UDN," in *Proc. International Symposium on Intelligent Signal Processing and Communication Systems*, 2018, pp. 56-61.

[5] Y. Li, X. Li, H. Ji, *et al.*, "A multiple APs cooperation access scheme for energy efficiency in UUDN with NOMA," in *Proc. IEEE Conference on Computer Communications Workshops*, 2017, pp. 892-897.

[6] Y. Liu, X. Li, F. R. Yu, *et al.*, "Grouping and cooperating among access points in user-centric ultra-dense networks with non-orthogonal multiple access," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 10, pp. 2295-2311, 2017.

[7] X. Z. Jiang, X. Li, and H. Ji, "Service-driven resource allocation based on energy efficiency in UUDN," in *Proc. 17th IEEE International Conference on Communication Technology*, 2017, pp. 498-502.

[8] W. Wong and S.-H. G. Chan, "Distributed joint AP grouping and user association for MU-MIMO networks," in *Proc. IEEE Conference on Computer Communications*, 2018, pp. 252-260.

[9] L. Yu, J. Wu, P. Fan, "Minimizing energy consumptions in user-centric ultra-densenetworks," in *Proc. IEEE International Conference on Communications Workshops*, 2018, pp. 1-6.

[10] A. Ammar and D. Reynolds, "Energy harvesting networks: Energy versus data cooperation," *IEEE Communications Letters*, vol. 22, no. 10, pp. 2128-2131, 2018.

[11] Ortiz A, Alshatri H, Weber T, et al. "Multi-Agent Reinforcement Learning for Energy Harvesting Two-Hop Communications with Full Cooperation," *ArXiv*, 2017.

[12] S. Tang and L. Tan, "Reward rate maximization and optimal transmission policy of EH device with temporal death in EH-WSNs," *IEEE Transactions on Wireless Communications*, vol. 16, no. 2, pp. 1157-1167, 2017.

[13] K. Rahbar, C. C. Chai, and R. Zhang. "Energy cooperation optimization in microgrids with renewable energy integration," *IEEE Transactions on Smart Grid*, vol. 9, no. 2, pp. 1482-1493, 2018.

[14] A. Jahid and S. Hossain, "Intelligent energy cooperation framework for green cellular base stations," in *Proc. International Conference on Computer, Communication, Chemical, Material and Electronic Engineering*, 2018, pp. 1-6.

[15] Y. Dong, Z. Chen, and P. Fan, "Capacity region of Gaussian multiple-access channels with energy harvesting and energy cooperation," *IEEE Access*, vol. 5, pp. 1570-1578, 2017.

[16] H. Lee and J. Lee, "Energy cooperation and traffic management in cellular networks with renewable energy," in *Proc. IEEE Global Communications Conference*, 2016.

[17] Y. Li, C. Yin, "Joint energy cooperation and resource allocation in C-RANs with hybrid energy sources," in *Proc. IEEE/CIC International Conference on Communications in China*, 2017.

[18] B. Xu, Y. Chen, J. R. Carrión, *et al.*, "Resource allocation in energy-cooperation enabled two-tier NOMA HetNets towards green 5G," *IEEE Journal on Selected Areas in Communications*, pp. 2758-2770, 2017.

[19] C. Duo, B. Li, Y. Li, *et al.*, "Energy cooperation in ultra-dense network powered by renewable energy based on cluster and learning strategy," *Wireless Communications & Mobile Computing*, vol. 5, pp. 1-10, 2017.

[20] Y. Lv, B. Li, *et al.*, "Energy cooperation in CoMP system based on Q-learning," in *Proc. 11th IEEE International Conference on Anti-counterfeiting, Security, and Identification*, pp. 90-94, 2017.

[21] F. Hourfar, H. J. Bidgoly, B. Moshiri, *et al.*, "A reinforcement learning approach for water flooding optimization in petroleum reservoirs," *Engineering Applications of Artificial Intelligence*, vol. 77, pp. 98-116, 2019.

[22] J. N. Tsitsiklis and B. V. Roy, "An analysis of temporal-difference learning with function approximation," *IEEE Transactions on Automatic Control*, vol. 42, no. 5, pp. 674-690, 2002.

[23] L. Baird, "Residual algorithms: Reinforcement learning with function approximation," *Machine Learning Proceedings*, pp. 30-37, 1995.

[24] V. Mnih, K. Kavukcuoglu, D. Silver, *et al.*, "Playing Atari with deep reinforcement learning," *Computer Science*, 2013.

[25] V. Mnih, K. Kavukcuoglu, D. Silver, *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529-533, 2015.

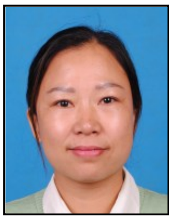
[26] D. Zhang, X. Han, and C. Deng, "Review on the research and practice of deep learning and reinforcement learning in smart grids," *CSEE Journal of Power and Energy Systems*, vol. 4, no. 3, pp. 362-370, 2018.

[27] G. Jeong and H. Y. Kim, "Improving financial trading decisions using deep Q-learning: Predicting the number of shares, action strategies, and transfer learning," *Expert Systems with Applications*, vol. 117, pp. 125-138, 2019.

[28] C. Watkins, "Learning from delayed rewards," Ph.D. dissertation, Royal Holloway, University of London, London, 1989.

[29] E. Evendar, Y. Mansour, "Learning rates for Q-learning," *Journal of Machine Learning Research*, vol. 5, no. 1, pp. 589-604, 2003.

[30] Y. Lv, B. Li, W. Zhao, *et al.*, "Multi-base station energy cooperation based on Nash Q-Learning algorithm," in *Proc. International Conference on 5G for Future Wireless Networks*, 2017, pp. 60-68.



Chunhong Duo has received BE degree in computer science and technology and MS degree in computer application technology at North China Electric Power University, China, in 2005 and 2008, respectively. Now she is studying the PhD degree in communication and information engineering at North China Electric Power University, China. Her research interests include deep reinforcement learning, energy harvesting communication system and Wireless Resource Management.

Yongqian Li received his BE degree in electronic instrument and measurement technology and the MS degree in communication and electronic system from Tianjin University, China, in 1982 and 1988, respectively. And he received his PhD degree at Gunma University, Japan, in 2003. Since 2004, he has been a professor in the Department of Electronics and Communication Engineering, North China Electric Power University. His research interests include optical communication and distributed optical fiber sensing.

Baogang Li received the Ph.D. degree in Beijing University of Posts and Telecommunications, Beijing, China. in 2012. He joined the North China Electric Power University, Baoding, China as an Assistant Professor. During the academic year 2016 to 2017, he was a Visiting Associate Professor with the Department of Electrical Engineering, University of Sydney, Sydney, Australia. His research interests include wireless communication, industry IOT, and smart grid communication.

Yabo Lv is currently working towards MS degree in communication and information engineering at North China Electric Power University, China. He has received B.Eng.degree from Agricultural University of Hebei, China in 2016. His research interests include deep reinforcement learning, energy harvesting communication system and Wireless Resource Management.