IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 6, 2022, accepted February 10, 2022, date of publication February 16, 2022, date of current version February 28, 2022.

*Digital Object Identifier 10.1109/ACCESS.2022.3152162*

# Optimal Resource Allocation Considering Non-Uniform Spatial Traffic Distribution in Ultra-Dense Networks: A Multi-Agent Reinforcement Learning Approach

**EUNJIN KIM[1], (Student Member, IEEE), HYUN-HO CHOI[2], (Senior Member, IEEE), HYUNGSUB KIM[3], JEEHYEON NA[3], (Member, IEEE), AND HOWON LEE[1], (Member, IEEE)**

[1]School of Electronic and Electrical Engineering and IITC, Hankyong National University, Anseong 17579, Republic of Korea
[2]School of ICT, Robotic, and Mechanical Engineering and IITC, Hankyong National University, Anseong 17579, Republic of Korea
[3]Electronics and Telecommunications Research Institute, Daejeon 34129, Republic of Korea

Corresponding authors: Hyun-Ho Choi (hhchoi@hknu.ac.kr) and Howon Lee (hwlee@hknu.ac.kr)

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) funded by the Korean Government (MSIT) through the 5G Open Intelligence-Defined RAN (ID-RAN) Technique Based on 5G New Radio under Grant 2018-0-01659.

**ABSTRACT** Recently, the demand for small cell base stations (SBSs) has been exploding to accommodate the explosive increase in mobile data traffic. In ultra-dense small cell networks (UDSCNs), because the spatial and temporal traffic distributions are significantly disproportionate, the efficient management of the energy consumption of SBSs is crucial. Therefore, we herein propose a multi-agent distributed Q-learning algorithm that maximizes energy efficiency (EE) while minimizing the number of outage users. Through intensive simulations, we demonstrate that the proposed algorithm outperforms conventional algorithms in terms of EE and the number of outage users. Even though the proposed reinforcement learning algorithm has significantly lower computational complexity than the centralized approach, it is shown that it can converge to the optimal solution.

**INDEX TERMS** Multi-agent Q-learning, spatial traffic distribution, energy efficiency, user outage, ultra-dense small cell network.

## I. INTRODUCTION

The massive amount of data traffic generated by the many different types of mobile services has led to a rapid increase in the number of base stations (BSs) deployed within the same network region [1]–[3]. This gradually accelerates network densification [4]–[6]. In addition, cellular networks have tended to use a higher frequency (e.g., a frequency in the terahertz range), which decreases the cell radius because of the larger attenuation of transmit power and requires more BSs to be deployed in the same network area [4], [5]. This network densification has resulted in the proliferation of ultra-dense small-cell networks (UDSCNs). In UDSCNs, because the average inter-site distance between small cell BSs (SBSs) and users has been decreasing considerably, the link quality can be improved. However, this may cause severe interference between neighboring SBSs and vastly increase the energy consumption of the entire network [7]. In this regard, it is worth noting that 80% of the energy in mobile networks is consumed by radio access networks (RANs), and most of the energy in current cellular networks is consumed by BSs, which is approximately 58% of the total power consumption [8], [9]. Therefore, in UDSCNs, maximizing the energy efficiency (EE) of SBSs is one of the most critical research challenges facing next-generation communication networks.

Recently, many researchers have been actively conducting research on minimizing the network energy consumption of UDSCNs. In [10], the impact of the idle-mode operation of BSs, transmit power control, user density, and user distribution on network energy efficiency was considered to find potential gains and limitations of ultra-dense networks (UDNs). The authors of [11] proposed a joint optimization framework for energy-efficient switching on/off strategy and user association policy for UDNs with partial

The associate editor coordinating the review of this manuscript and approving it for publication was Xujie Li.

conventional BSs. In addition, an energy-aware user association and power allocation algorithm was proposed for ultra-dense networks with energy-harvesting BSs based on millimeter waves (mmWaves), in [12].

Moreover, various techniques have been proposed to effectively utilize radio resources based on Q-learning in UDN environments. In [13], a Q-learning based solution for a small-scale cooperative coded caching system was proposed to maximize the long-term expected cumulative traffic load served by SBSs without accessing macro cell BSs (MBSs). The authors of [14] proposed a Q-learning based dynamic load adjustment algorithm to reduce energy consumption and adjust the traffic load. It has been proven that the algorithm can save energy consumption compared to the existing on/off algorithm and other conventional algorithms. Furthermore, a Q-learning based downlink transmit power control algorithm was proposed in [15]. A transfer learning method called hotbooting was applied to accelerate the learning speed and reduce the energy consumption based on the estimated user density without any information about the network and channel model of the other small cells.

The traffic generated by actual UDSCNs is geographically disproportionate. According to [16], half of the network sites carry only 15% of the total traffic, whereas 5% of the sites carry 20% of the traffic. Therefore, the network operator should efficiently manage and control network energy consumption by considering the dynamics of the spatial network traffic. Many studies have been conducted to improve these spatial and temporal traffic dynamics. In [17], the authors presented a load balancing scheme based on deep-reinforcement learning (DRL) to solve global and local traffic variations in irregular dense small cell networks. The authors of [18] proposed unmanned aerial vehicle (UAV)-assisted cell-edge mobile user offloading in non-uniform heterogeneous cellular networks. Here, cell-edge mobile users are periodically scheduled between coordinated ground base stations and flying UAVs. In addition, [19] solved the user association problem using resource and handover management based on the deep deterministic policy gradient (DDPG) method for mmWave networks. They showed that intelligent load-balancing handover could effectively associate users in the case of a high-load situation. In [20], the authors proposed cluster-based resource allocation and user association via efficient co-channel interference management in mmWave dense femtocell networks. This study altered the binary optimization problem into a continuous problem using deductive penalty functions and solved it by computing the difference of two convex functions. Furthermore, the authors of [21] proposed a load-aware cell selection scheme for multi-connectivity in intra-frequency 5G ultra-dense networks to efficiently utilize available idle resources and reduce the probability of radio link failure.

Previous methods using optimization and deep reinforcement learning frameworks are computationally extremely complex and require intensive iteration. In particular, because tabular Q-learning does not exploit deep neural networks for functional approximation, it can significantly reduce the computational overhead caused when performing neural network training in the conventional approaches. This motivated us to propose a SBS control algorithm based on multi-agent distributed Q-learning to maximize the EE while simultaneously minimizing the number of outage users in UDSCNs. The proposed algorithm, which considers the spatial traffic dynamics, can efficiently control the transmit power of SBSs based on multi-agent Q-learning. The main contributions of this study are as follows.

- Two types of network dynamics are considered for proposing a reinforcement learning algorithm that maximizes EE in UDSCNs: uniform/non-uniform spatial traffic distributions and random user mobility.
- Regardless of uneven spatial traffic distribution and unpredictable user movements, we demonstrate that the proposed multi-agent Q-learning algorithm can converge to the optimal solution obtained by exhaustive search.
- Even in ultra-dense network environments, the proposed algorithm outperforms the conventional algorithm in terms of EE and the number of outage users. However, achieving these two objectives may not be feasible in a conventional optimization framework.
- The proposed algorithm can significantly reduce the computational complexity by allowing the agent to consider only its own state.

The remainder of this paper is organized as follows: In Section II, the system model of the proposed algorithm is presented. The proposed multi-agent Q-learning algorithm for maximizing EE while minimizing the number of outage users is proposed in Section III. Section IV presents the effectiveness of the proposed algorithm verified through intensive simulations with respect to the EE and the number of outage users. Finally, conclusions are presented in Section V.

## II. SYSTEM MODEL

Herein, we describe the system model for the proposed algorithm and the assumptions used in this study. Consider a downlink communication for UDSCNs configured with several MBSs ($\mathbf{M}$), SBSs ($\mathbf{N}$), and users ($\mathbf{U}$), as shown in Fig. 1. The MBSs are considered as interferers in this network and the SBSs adjust their transmit power to maximize the system performance.

### A. SINR CALCULATION

The channel quality of users received from SBSs is measured by the reference signal received power (RSRP), which is commonly used as a channel quality metric between users and BSs in cellular networks. RSRP between user $i$ and SBS $j$ ($P_r(i,j)$) is expressed as $P_r(i,j) = \frac{P_t(j)}{d(i,j)^{\rho}}$, where $P_t(j)$ is the transmit power of SBS $j$, $d(i,j)$ is the distance between user $i$ and SBS $j$, and $\rho$ is the path loss exponent in UDSCNs. Using $P_r(i,j)$, the signal-to-interference-plus-noise ratio (SINR) of

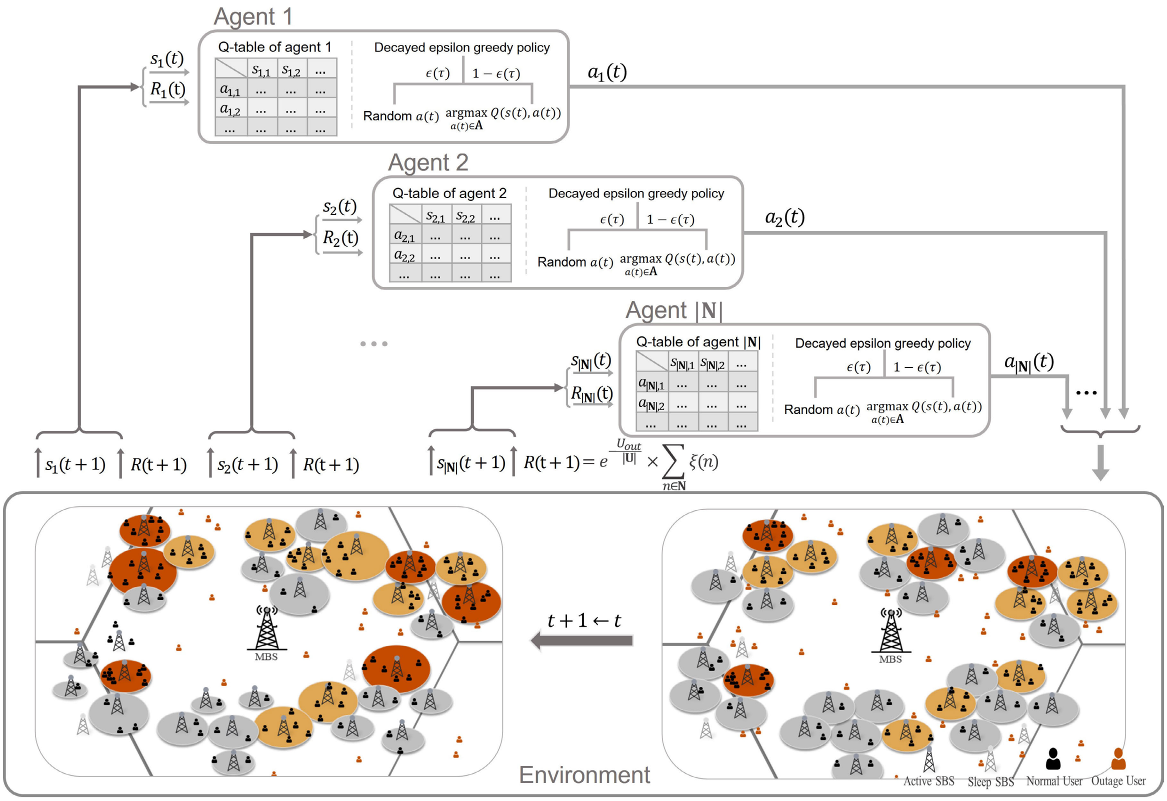


**FIGURE 1.** System model of proposed multi-agent distributed Q-learning framework maximizing EE while minimizing number of outage users in UDSCNs.

user $i$ for SBS $j$ can be calculated as

$$\gamma(i,j) = \frac{P_r(i,j)}{\sum_{n\neq j,n\in\mathbf{N}} P_r(i,n) + \sum_{m\in\mathbf{M}} P_r(i,m) + \sigma_i^2}, \tag{1}$$

where $\sigma_i^2$ is the thermal noise power. As mentioned before, all other SBSs, except the serving SBS and MBSs, are considered as interferers. When $\gamma(i,j) < \gamma_{th}, \forall j \in \mathbf{N}$, user $i$ is considered as an outage user. Here, $\gamma_{th}$ denotes the SINR outage threshold.

### B. EE CALCULATION CONSIDERING SBS POWER CONSUMPTION

From equation (1), the achievable data rate of user $i$ for SBS $j$ ($\zeta(i,j)$) can be obtained as

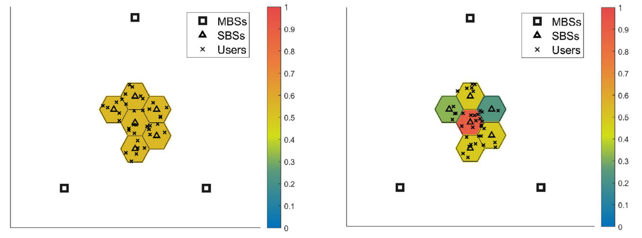
$$\zeta(i,j) = \frac{1}{|\mathbf{U}(j)|} \cdot W_j \cdot \log_2(1+\gamma(i,j)), \tag{2}$$

Here, $W_j$ is the system bandwidth of SBS $j$ and $|\mathbf{U}(j)|$ is the number of users associated with SBS $j$, and using equation (2), the EE of SBS $j$ ($\xi(j)$) can be calculated as

$$\xi(j) = \frac{\sum_{i\in\mathbf{U}(j)} \zeta(i,j)}{P_{tot}(j)}, \tag{3}$$

where $P_{tot}(j)$ is the total power consumption of SBS $j$, which can be represented as

$$P_{tot}(j) = P_c(j) + P_p(j) + \frac{1}{\delta} \cdot P_t(j), \tag{4}$$



(a) Uniform spatial traffic distribution

(b) Non-uniform spatial traffic distribution

**FIGURE 2.** Network deployment with uniform and non-uniform spatial traffic distribution when $|\mathbf{M}| = 3$, $|\mathbf{N}| = 6$, and $|\mathbf{U}| = 36$.

Here, $\delta$ is the power amplifier efficiency, and $P_c(j)$, $P_p(j)$, and $P_t(j)$ are the fixed circuit power consumption, radio frequency (RF) power amplifier power consumption, and transmit power consumption, respectively. In particular, $P_p(j)$ describes the power consumption according to the variation in the cell load. Accordingly, $P_p(j)$ can be obtained as

$$P_p(j) = P_{pa} \times \frac{|\mathbf{U}(j)|}{\kappa}, \tag{5}$$

where $\kappa$ is the maximum number of users supportable per RF power amplifier and $P_{pa}$ is the power consumed by each RF power amplifier.

**Algorithm 1** Proposed Multi-Agent Q-Learning Algorithm for Maximizing EE in UDSCNs

```
Initialize: Place |M| MBSs, |N| SBSs, and |U| users in network, and initialize Q-tables of all agents
for $\tau = 1 : \tau_{\max}$ do
  Calculate $\epsilon(\tau) = \epsilon_{\text{init}} \times (1 - \epsilon_{\text{init}})^{\frac{\tau}{\chi \times |\mathbf{A}|}}$.
  for $i = 1 : |\mathbf{U}|$ do
    Move the position of user $i$ corresponding to the random walk model with $v_i(\tau)$ and $\theta_i(\tau)$.
  end for
  for $t = 1 : t_{\max}$ do
    for $j = 1 : |\mathbf{N}|$ do
      Chooses action of agent $j$ ($a_j(t)$) according to the decayed epsilon greedy policy with $\epsilon(\tau)$.
          $a_j(t) = \begin{cases} \text{random action, with } \epsilon(\tau), \\ \arg\max_{a \in \mathbf{A}}(Q_j(s_j(t), a)), \text{ with } 1 - \epsilon(\tau). \end{cases}$
    end for
    for $i = 1 : |\mathbf{U}|$ do
      Calculate SINR of user $i$ for its serving SBS ($\gamma(i, j)$) and achievable data rate of user $i$ for SBS $j$ ($\zeta(i, j)$).
      if $\gamma(i, n) < \gamma_{th}, \forall n \in \mathbf{N}$ then
        $U_{\text{out}} = U_{\text{out}} + 1$
      else
        $|\mathbf{U}(j)| = |\mathbf{U}(j)| + 1$
      end if
    end for
    for $j = 1 : |\mathbf{N}|$ do
      Calculate $\xi(j)$ by using $\zeta(i, j)$.
    end for
    Calculate $R(s(t + 1), a(t))$ and update Q-values for all agents using equation (10).
  end for
end for
```

### C. USER MOBILITY MODEL

We apply a random walk model to emulate users' unpredictable movements [22]. At each episode, the position of each user is altered according to the random walk model. The speed of user $i$ at episode $\tau$ ($v_i(\tau)$) is randomly determined within $[0, v_{\max}]$. In addition, the moving direction of user $i$ ($\theta_i(\tau)$) is randomly chosen within $[0, 2\pi]$. Consequently, user $i$ moves with the velocity vector ($\mathbf{v_i}(\tau)$) at episode $\tau$ as follows:

$$\mathbf{v_i}(\tau) = \{v_i(\tau) \cos \theta_i(\tau), v_i(\tau) \sin \theta_i(\tau)\}. \tag{6}$$

### D. SPATIAL TRAFFIC DISTRIBUTION IN UDSCN

An important characteristic of actual UDSCNs is the geographically disproportionate network traffic. Accordingly, we assume that a non-uniform spatial traffic distribution is generated according to the constraints described in [16]. This spatial traffic distribution is based on real-world measurements. From [16], half of the network cells carry only 15% of the total network traffic, whereas 5% of the cells carry 20% of the traffic. Unfortunately, spatial traffic growth would be expected to increase most in cells that already have high loads. For instance, Figs. 2a and 2b show the network deployment results considering uniform and non-uniform spatial traffic distributions for the 3 MBSs, 6 SBSs, and 36 users. The vertical axis denotes relative traffic density in each cell. Specifically, when $|\mathbf{U}(i)| = \frac{|\mathbf{U}|}{|\mathbf{N}|}$, we assumed the traffic density of SBS $i$ as 0.5. Also, this value was used as a criterion for determining the traffic density of other SBSs.

## III. PROPOSED MULTI-AGENT Q-LEARNING ALGORITHM FOR MAXIMIZING EE IN UDSCN

We herein propose an EE maximization algorithm based on multi-agent Q-learning for UDSCNs with small cell clusters, as shown in Fig. 1. In our multi-agent distributed reinforcement framework, agents, states, actions, a reward, a Q-function, and a policy are defined as follows.

### A. AGENT

Consider that each SBS is an agent of the proposed multi-agent reinforcement learning framework in UDSCNs. In a centralized reinforcement framework, a single agent can manage all state information of the SBSs, but it generates a large amount of overhead as a result of the computational complexity. Thus, we consider a multi-agent distributed Q-learning framework.

### B. STATE

In this study, the agent does not share its state information with other agents. Thus, each agent considers only its transmit power. The state of the agent can be defined as

$$\mathbf{S} = \{P_{t,\mathbf{N}}^{\min} : \Delta_{P_t} : P_{t,\mathbf{N}}^{\max}\}, \tag{7}$$

where $P_{t,\mathbf{N}}^{\min}$ and $P_{t,\mathbf{N}}^{\max}$ are the minimum and maximum transmit power of the SBS, respectively. In addition, the $\Delta_{P_t}$ is the step size of the power increases. Here, $\{a{:}b{:}c\}$ represents a set of values from $a$ to $c$ with a step size of $b$.

### *C. ACTION*

To maximize the EE of UDSCNs, the agent can choose one of three actions ($\mathbf{A}$): "transmit power up ($\Delta_{P_t}$)", "transmit power down ($\Delta_{P_t}$)", and "keep current transmit power ($\Delta_0$)" as follows:

$$\mathbf{A} = \{-\Delta_{P_t}, \Delta_0, \Delta_{P_t}\}, \tag{8}$$

### *D. REWARD*

Assume that each agent shares its reward information with each other agent to maximize the EE of the entire network. In addition, because minimizing the number of outage users is essential when maximizing the EE, we design an outage-aware reward in the proposed reinforcement framework. Accordingly, the reward of agent $j$ ($R_j^c$) is represented as

$$R_j^c = e^{-\frac{U_{\text{out}}}{|\mathbf{U}|}} \times \sum_{n\in\mathbf{N}} \xi(n), \tag{9}$$

where $|\mathbf{U}|$ and $U_{\text{out}}$ are the total number of users and the number of outage users in the entire network, respectively.

### *E. Q-FUNCTION UPDATE*

A Q-function is a state-action value function that externally implies a value to the action in a specific state of the agent, and internally implies the expected reward when the action is performed. In other words, it describes the benefit of an agent performing a particular action in a state with a specific policy. In this study, the Q-function ($Q_j(s_j(t), a_j(t))$) is expressed as:

$$\begin{aligned} Q_j(s_j(t), a_j(t)) = (1-\alpha)\cdot Q_j(s_j(t), a_j(t)) + \alpha\cdot [R_j(s_j(t+1),\\ \times a_j(t)) + \eta \cdot \max_{a_j'\in\mathbf{A}} Q_j(s_j(t+1), a_j')], \end{aligned} \tag{10}$$

Here, $\alpha$ is the learning rate and $\eta$ is a discount factor.

### *F. POLICY*

We adopt the decayed $\epsilon$-greedy policy for extensive exploration in early episodes [23], [24]. According to $\epsilon(\tau)$, each agent chooses a random action with a probability of $\epsilon(\tau)$, and the optimal action with a probability of $1-\epsilon(\tau)$. As the number of episodes increases, the value of $\epsilon(\tau)$ decreases; therefore, in the latter part of the learning, the agent exploits more than it explores. $\epsilon(\tau)$ can be described as

$$\epsilon(\tau) = \epsilon_{\text{init}} \times (1-\epsilon_{\text{init}})^{\frac{\tau}{\chi\times|\mathbf{A}|}}, \tag{11}$$

where $\epsilon_{\text{init}}$ is the initial epsilon value, $\chi$ is the decay parameter, and $|\mathbf{A}|$ is the size of the action set of each agent. The detailed operational procedure of the proposed multi-agent distributed Q-learning algorithm is described in Algorithm 1.

**TABLE 1. Simulation parameters.**

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| $\lvert\mathbf{M}\rvert$ | $\{3, 5\}$ | $\lvert\mathbf{N}\rvert$ | $\{6, 100\}$ |
| $\lvert\mathbf{U}\rvert$ | $\{36, 600\}$ | $r$ | $100\text{m} \sim 150\text{m}$ |
| $P_c$ | 6.7W | $P_{pa}$ | 2W |
| $P_{t,\mathbf{M}}$ | 20W | $P_{t,\mathbf{N}}^{\max}$ | 2W |
| $P_{t,\mathbf{N}}^{\min}$ | 0W | $\Delta_{P_t}$ | 0.5W |
| $\mathbf{A}$ | $\{-0.5, 0, 0.5\}$W | $\delta$ | 1 |
| $W_j$ | 10MHz | $\gamma_{th}$ | 0dB |
| $\kappa$ | 5 | $\sigma^2$ | −174dBm/Hz |
| $\alpha$ | 0.1 | $\eta$ | 0.9 |
| $\epsilon_{\text{init}}$ | 0.99 | $\chi$ | 330 |

## IV. SIMULATION RESULTS AND DISCUSSION

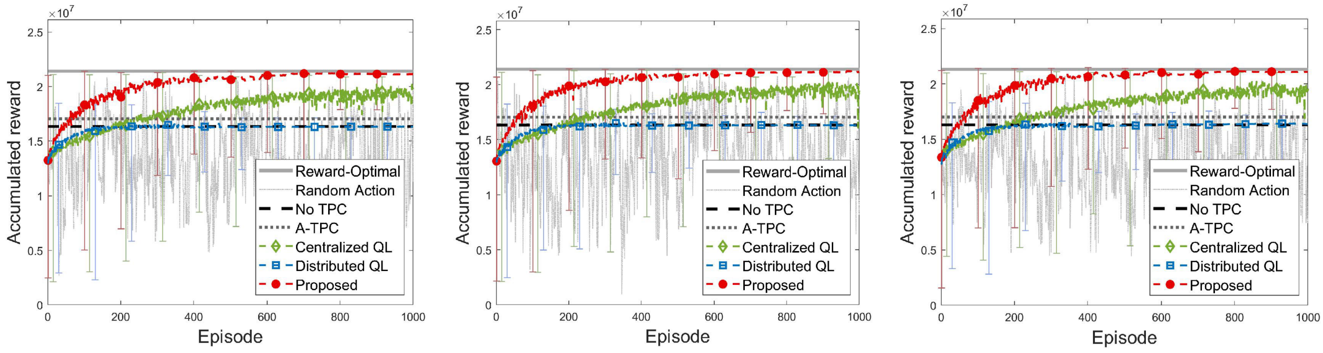
### *A. SIMULATION ENVIRONMENTS*

We considered two types of network deployments: [3 MBSs, 6 SBSs, 36 users] and [5 MBSs, 100 SBSs, 600 users]. To obtain the results, we performed 1,000 episodes where each episode involved 3,000 iterations. The simulation parameters are listed in Table 1. In addition, we compared the performance of the proposed algorithm with several benchmark algorithms: "Reward-Optimal," "No TPC," "A-TPC," "Random Action," "Centralized QL," and "Distributed QL". Details of these conventional algorithms are as follows:

- **Reward-Optimal:** The reward-optimal solution is obtained using the exhaustive search algorithm. This algorithm enumerates and checks all possible states of the agents. In the case of complicated network environments, it is difficult to obtain a reward-optimal solution owing to its high computational complexity.
- **No Transmit Power Control (No TPC)**: None of the agents control their transmit power. That is, each agent always sends its signal using maximum transmit power.
- **Adaptive Transmit Power Control (A-TPC):** The transmit power of each agent is calculated by the number of associated users. In this study, user association was determined by measuring the SINR in the initial network deployment.
- **Random Action:** This algorithm randomly chooses an action in each episode. We can use the random action algorithm to roughly prove that the proposed algorithm to the optimal solution because the exhaustive search-based optimal solution cannot be obtained in simulation scenarios with extremely high computational complexity.
- **EE Maximization based on Centralized Q-Learning (C-QL):** This algorithm is based on Q-learning and considers the overall state information of the agents. However, because all possible cases explorable by the agent should be considered, the size of the Q-table increases

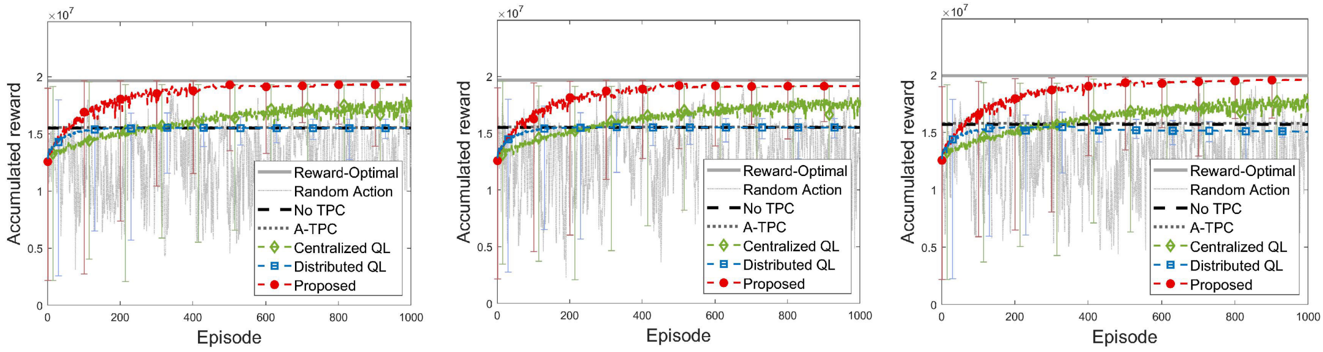
**TABLE 2. Computational complexity analysis of conventional and proposed algorithms.**

| Algorithm | Reward-Optimal | C-QL Based EE Max. | D-QL Based EE Max. | Proposed Multi-Agent QL |
|---|---|---|---|---|
| Computational complexity ($\mathcal{O}(\cdot)$) | $\mathcal{O}(\lvert\mathbf{S}\rvert^{\mathbf{N}}\lvert\mathbf{A}\rvert^{\mathbf{N}})$ | $\mathcal{O}(\lvert\mathbf{S}\rvert^{\mathbf{N}}\lvert\mathbf{A}\rvert^{\mathbf{N}})$ | $\mathcal{O}(\lvert\mathbf{S}\rvert\lvert\mathbf{A}\rvert)$ | $\mathcal{O}(\lvert\mathbf{S}\rvert\lvert\mathbf{A}\rvert)$ |



(a) $v_{\max} = 0$ m/s (b) $v_{\max} = 0.01$ m/s (c) $v_{\max} = 0.1$ m/s

**FIGURE 3. Accumulated average reward vs. episode when |M| = 3, |N| = 6, and |U| = 36 with uniform spatial traffic distribution.**



(a) $v_{\max} = 0$ m/s (b) $v_{\max} = 0.01$ m/s (c) $v_{\max} = 0.1$ m/s

**FIGURE 4. Accumulated average reward vs. episode when |M| = 3, |N| = 6, and |U| = 36 with non-uniform spatial traffic distribution.**

exponentially according to the number of agents and the sizes of the state and action sets. Because of its complexity, we cannot apply this algorithm to complicated network environments.

- **EE Maximization based on Distributed Q-learning (D-QL):** The basic operation of this algorithm is almost similar to that of the proposed algorithm. However, the only difference is that reward sharing is not considered in this distributed Q-learning algorithm. That is, each agent only considers its own reward before choosing the action to perform. Hence, the reward ($R_j^d$) can be described as

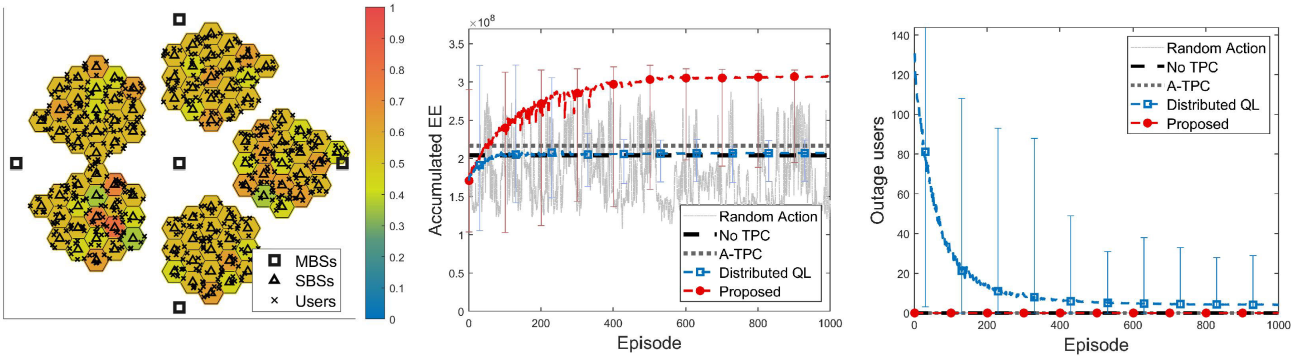
$$R_j^d = e^{-\frac{U_{\text{out}}(j)}{|\mathbf{U}(j)|}} \times \xi(j), \tag{12}$$

where $|\mathbf{U}(j)|$ and $U_{\text{out}}(j)$ are the number of users associated with SBS $j$ and the number of outage users of SBS $j$, respectively.

The computational complexity of the reward-optimal algorithm, centralized Q-learning algorithm, distributed Q-learning algorithm, and the proposed algorithm are summarized in Table 2. Because the reward-optimal and centralized Q-learning algorithms are designed to consider all cases that could occur in networks, their computational complexity is significantly larger than that of other algorithms. The proposed multi-agent distributed Q-learning algorithm can greatly reduce the computational complexity by allowing the agent to consider only its own state.

### B. RESULTS AND DISCUSSION

Figs. 3a–3c show the accumulated average rewards based on the progress of the episode for each algorithm when $|\mathbf{M}| = 3$, $|\mathbf{N}| = 6$, and $|\mathbf{U}| = 36$ under a network deployment with a uniform spatial traffic distribution, as shown in Fig. 2a. Here, the users were randomly distributed within 100m of each SBS. Furthermore, the results in Figs. 3a, 3b, and 3c were obtained for $v_{\max} = 0$ m/s, 0.01 m/s, and 0.1 m/s, respectively. The results in these figures show that the proposed algorithm converges to the optimal solution even if the user's mobility increases. Because A-TPC controls the transmit power of

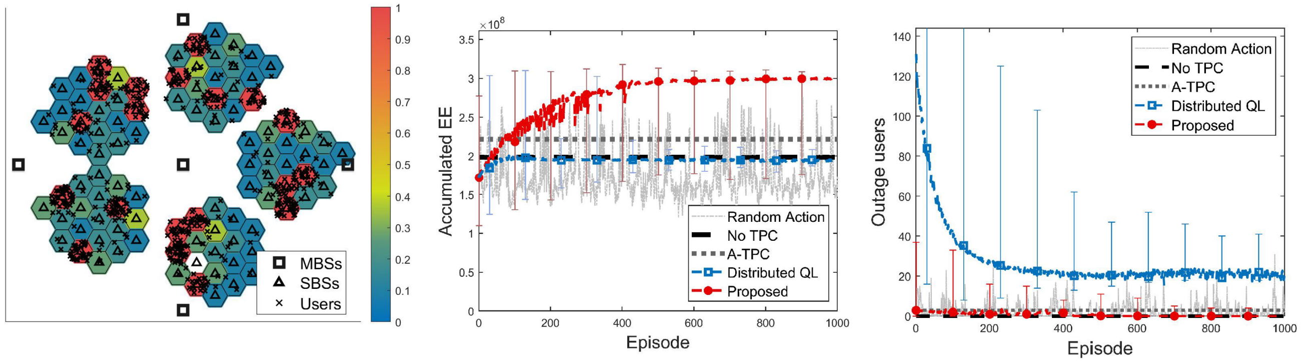


(a) Network deployment with uniform traffic distribution

(b) Accumulated average energy efficiency

(c) Accumulated average number of outage users

**FIGURE 5.** Network deployment with uniform spatial traffic distribution, accumulated EE vs. episode, and number of outage users vs. episode when $|\mathbf{M}| = 5$, $|\mathbf{N}| = 100$, $|\mathbf{U}| = 600$, and $v_{\max} = 0.01$ m/s.



(a) Network deployment with non-uniform traffic distribution

(b) Accumulated average energy efficiency

(c) Accumulated average number of outage users

**FIGURE 6.** Network deployment with non-uniform spatial traffic distribution, accumulated EE vs. episode, and number of outage users vs. episode when $|\mathbf{M}| = 5$, $|\mathbf{N}| = 100$, $|\mathbf{U}| = 600$, and $v_{\max} = 0.01$ m/s.

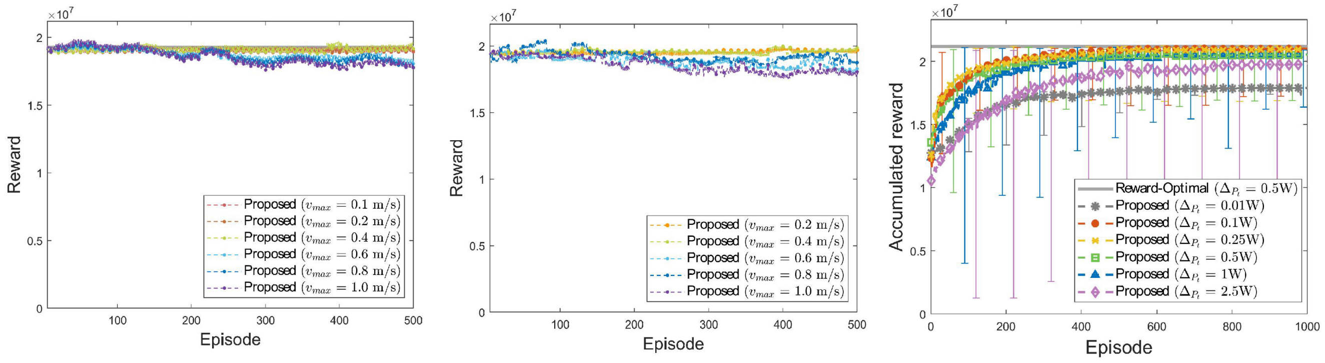
each SBS according to the initial association results, A-TPC delivers performance results superior to those of the No TPC algorithm. In the case of the centralized Q-learning algorithm, because this algorithm needs to consider the states and reward information of all the SBSs, the convergence speed of this algorithm is relatively slow compared to that of the proposed algorithm. In addition, each agent in the distributed Q-learning algorithm tries to maximize its reward without considering the status and rewards of other agents. As a result, the transmit power of each agent gradually increases to reach the maximum power, and the result finally converges to that of the No TPC algorithm.

Figs. 4a–4c demonstrate that the proposed algorithm can converge to the optimal solution even for a network deployment with a non-uniform spatial traffic distribution, as shown in Fig. 2b. Similar to the results for the uniform traffic distribution, our proposed algorithm outperforms the conventional algorithms with respect to the accumulated average reward regardless of the increase in user mobility. The overall performance behavior is clearly similar to the case of uniform spatial traffic distribution, but the accumulated reward value is relatively smaller than that of the uniform distribution owing to the regionally biased traffic. Moreover, as learning progressed, the length of the error bars gradually became shorter, which shows that the learning progressed well.

To prove the operational flexibility of the proposed multi-agent distributed Q-learning algorithm, we considered ultra-dense network environments. Figs. 5a and 6a show the network deployment results considering uniform and non-uniform spatial traffic distributions when $|\mathbf{M}| = 5$, $|\mathbf{N}| = 100$, and $|\mathbf{U}| = 600$. Here, users were randomly distributed within 150m of the SBS and were moving in correspondence to the random walk mobility model with $v_{\max} = 0.01$ m/s. Figs. 5b and 5c show the accumulated average EE and the average number of outage users based on the progress of the episode for each algorithm under the network deployment with the uniform spatial traffic distribution in Fig. 5a. In addition, Figs. 6b and 6c show the accumulated average EE and the average number of outage users based on the progress of the episode for each algorithm under the network

**TABLE 3. Comparison of the fairness of SBSs in conventional and proposed algorithms.**

| Algorithm | \|N\| = 6 | | | | | | \|N\| = 100 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Uniform distribution | | | Non-uniform distribution | | | Uniform distribution | | | Non-uniform distribution | | |
| | $v_{\max}$=0 | $v_{\max}$=0.01 | $v_{\max}$=0.1 | $v_{\max}$=0 | $v_{\max}$=0.01 | $v_{\max}$=0.1 | $v_{\max}$=0 | $v_{\max}$=0.01 | $v_{\max}$=0.1 | $v_{\max}$=0 | $v_{\max}$=0.01 | $v_{\max}$=0.1 |
| Reward-Optimal | 0.771 | 0.771 | 0.772 | 0.716 | 0.716 | 0.712 | - | - | - | - | - | - |
| Random Action | 0.801 | 0.628 | 0.629 | 0.694 | 0.745 | 0.817 | 0.731 | 0.771 | 0.716 | 0.595 | 0.643 | 0.642 |
| No TPC | 0.959 | 0.959 | 0.96 | 0.91 | 0.909 | 0.904 | 0.899 | 0.899 | 0.898 | 0.786 | 0.786 | 0.784 |
| A-TPC | 0.961 | 0.961 | 0.962 | 0.924 | 0.924 | 0.917 | 0.894 | 0.894 | 0.893 | 0.744 | 0.744 | 0.731 |
| Centralized QL | 0.747 | 0.805 | 0.774 | 0.783 | 0.781 | 0.695 | - | - | - | - | - | - |
| Distributed QL | 0.959 | 0.959 | 0.957 | 0.91 | 0.91 | 0.944 | 0.899 | 0.899 | 0.896 | 0.744 | 0.761 | 0.741 |
| Proposed | 0.772 | 0.78 | 0.775 | 0.714 | 0.739 | 0.713 | 0.855 | 0.867 | 0.876 | 0.702 | 0.726 | 0.716 |



(a) $v_{\max}$ used in training = 0.1 m/s (b) $v_{\max}$ used in training = 0.2 m/s (c) $P_t^{\max}$ = 10W

**FIGURE 7. Reward vs. episode in proposed algorithm obtained under network deployment with non-uniform spatial traffic distribution when |M| = 3, |N| = 6, and |U| = 36.**

deployment with the non-uniform spatial traffic distribution in Fig. 6a. Even in ultra-dense network environments, the proposed algorithm outperforms the conventional algorithms in terms of the accumulated EE and number of outage users. Because of the extremely high computational complexity of the reward-optimal and centralized Q-learning algorithms, these algorithms did not produce results in this simulation scenario. However, the error bars of the proposed algorithm and the random action algorithm provide a rough indication that the proposed algorithm converges to the optimal solution. Furthermore, in the case of non-uniform spatial traffic distribution, distributed Q-learning yields several outage users because this algorithm only considers its state and reward. However, we can show that the average number of outage users of the proposed algorithm converges to zero.

Table 3 compares the fairness of EE between SBSs for each algorithm once learning is complete. We used Jain's fairness index to obtain the fairness results [25]. Jain's fairness index can be represented as

$$\vartheta = \frac{(\sum_{j=1}^{\mathbf{N}} \xi(j))^2}{|\mathbf{N}| \cdot \sum_{j=1}^{\mathbf{N}} \xi(j)^2}, \tag{13}$$

With the No TPC, A-TPC, and distributed QL methods, because all SBSs transmit at almost maximum power, the difference in the energy efficiency between SBSs is smaller than that in other algorithms. As a result, these algorithms produce superior SBS fairness results compared to the reward-optimal, random action, centralized QL, and proposed algorithms. Moreover, as expected, the SBS fairness results for the uniform spatial traffic distribution are larger than those for the non-uniform spatial traffic distribution. As the learning progresses, the SBSs with higher traffic density use relatively larger transmit power than those with lower traffic density. Consequently, the EE results for each SBS could be gradually different, and accordingly, the SBS fairness in the non-uniform traffic distribution becomes smaller than that in the uniform traffic distribution.

To show the performance behavior against delayed learning information, we obtained Figs. 7a and 7b under network deployment with non-uniform spatial traffic distribution when $|\mathbf{M}| = 3$, $|\mathbf{N}| = 6$, and $|\mathbf{U}| = 36$. In Fig. 7a, while training the Q-values of each agent, we assumed $v_{\max}$ as 0.1 m/s. However, in the test environments, the users moved to larger $v_{\max}$ values to reflect the effect of the delayed learning information. Similarly, in Fig. 7b, while training the Q-values of each agent, we assumed $v_{\max}$ as 0.2 m/s. Also, the tests were performed against larger $v_{\max}$ values. It can be seen that the greater the difference in $v_{\max}$ between the learning environment and the test environment, the larger the performance degradation. In addition, in the case of Fig. 7a, since $v_{\max}$ is smaller than that in Fig. 7b, it has a chance to perform learning for more diverse positions. As a result, the performance degradation might be smaller.

Fig. 7c shows how $\Delta_{P_t}$ affects the system performance. To obtain this figure, we set $P_t^{\max}$ as 5W, and the reward-optimal solution is obtained using the exhaustive search algorithm under $\Delta_{P_t} = 0.5$W. When $\Delta_{P_t} = 0.01$W, because the action set size of each agent becomes too large, it needs to take a long time to achieve the optimal solution. In contrast, when $\Delta_{P_t} = 2.5$W, the agent rarely finds the optimal solution because its action set size is too small. Therefore, network operators should determine $\Delta_{P_t}$ very carefully considering the network environment and spatial traffic distribution.

## V. CONCLUSION

In this paper, we proposed a SBS power control algorithm based on multi-agent distributed Q-learning to maximize the network EE while reducing the number of outage users in UDSCNs. To consider practical network environments, we utilized uniform and non-uniform spatial traffic distributions based on real-world measurements. Even in the non-uniform distribution, we showed that the proposed algorithm converges well to the optimal solution obtained by the exhaustive search algorithm. In addition, to demonstrate the performance in ultra-dense network environments, we considered 100 SBSs and 600 users in a network with five small cell clusters. In this network environment, we demonstrated that the proposed algorithm outperforms conventional algorithms such as random action, No TPC, A-TPC, and distributed Q-learning algorithms regardless of the increase in user mobility. Furthermore, by allowing the agent to consider only its own state, the computational complexity of the proposed algorithm can be significantly reduced compared to that of the centralized Q-learning algorithm.

## REFERENCES

[1] *Key Drivers and Research Challenges for 6G Ubiquitous Wireless Intelligence*, 6G Res. Vis. 1, 6G Flagship, Univ. Oulu, Oulu, Finland, Sep. 2019, pp. 1–36.

[2] H. Yu, H. Lee, and H. Jeon, "What is 5G? Emerging 5G mobile services and network requirements," *Sustainability*, vol. 9, pp. 1–22, Oct. 2017.

[3] W. Lee, B. C. Jung, and H. Lee, "DeCoNet: Density clustering-based base station control for energy-efficient cellular IoT networks," *IEEE Access*, vol. 8, pp. 120881–120891, 2020.

[4] G. Yu, R. Liu, Q. Chen, and Z. Tang, "A hierarchical SDN architecture for ultra-dense millimeter-wave cellular networks," *IEEE Commun. Mag.*, vol. 56, no. 6, pp. 79–85, Jun. 2018.

[5] X. Cai, A. Chen, L. Chen, and Z. Tang, "Joint optimal multi-connectivity enabled user association and power allocation in mmWave networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2021, pp. 1–6.

[6] P. Liu, Q. Hu, K. Jin, G. Yu, and Z. Tang, "Toward the energy-saving optimization of WLAN deployment in real 3-D environment: A hybrid swarm intelligent method," *IEEE Syst. J.*, early access, Apr. 5, 2021, doi: 10.1109/JSYST.2021.3065434.

[7] Z. Hasan, H. Boostanimehr, and V. K. Bhargava, "Green cellular networks: A survey, some research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 13, no. 4, pp. 524–540, 4th Quart., 2011.

[8] C. Han, T. Harrold, S. Armour, I. Krikidis, S. Videv, P. M. Grant, H. Haas, J. S. Thompson, I. Ku, C.-X. Wang, T. A. Le, M. R. Nakhai, J. Zhang, and L. Hanzo, "Green radio: Radio techniques to enable energy-efficient wireless networks," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 46–54, Jun. 2011.

[9] J. Wu, Y. Zhang, M. Zukerman, and E. K. N. Yung, "Energy-efficient base-stations sleep-mode techniques in green cellular networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 803–826, 2nd Quart., 2015.

[10] D. López-Pérez, M. Ding, H. Claussen, and A. H. Jafari, "Towards 1 Gbps/UE in cellular systems: Understanding ultra-dense small cell deployments," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2078–2101, 4th Quart., 2015.

[11] Z. Jian, W. Muqing, and Z. Min, "Energy-efficient switching ON/OFF strategies analysis for dense cellular networks with partial conventional base-stations," *IEEE Access*, vol. 8, pp. 9133–9145, 2020.

[12] H. Zhang, S. Huang, C. Jiang, K. Long, V. C. M. Leung, and H. V. Poor, "Energy efficient user association and power allocation in millimeter-wave-based ultra dense networks with energy harvesting base stations," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1936–1947, Sep. 2017.

[13] S. Gao, P. Dong, Z. Pan, and G. Y. Li, "Reinforcement learning based cooperative coded caching under dynamic popularities in ultra-dense networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 5, pp. 5442–5456, May 2020.

[14] Q. Zhang, X. Xu, J. Zhang, X. Tao, and C. Liu, "Dynamic load adjustments for small cells in heterogeneous ultra-dense networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, May 2020, pp. 1–6.

[15] H. Zhang, M. Min, L. Xiao, S. Liu, P. Cheng, and M. Peng, "Reinforcement learning-based interference control for ultra-dense small cells," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–6.

[16] P. Frenger, C. Friberg, Y. Jading, M. Olsson, and O. Persson, "Radio network energy performance: Shifting focus from power to precision," *Ericsson Rev.*, pp. 1–9, Feb. 2014. [Online]. Available: https://www.ericsson.com/49868a/assets/local/reports-papers/ericsson-technology-review/docs/2014/er-radio-network-energy-performance.pdf

[17] Y. Xu, W. Xu, Z. Wang, J. Lin, and S. Cui, "Load balancing for ultradense networks: A deep reinforcement learning-based approach," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 9399–9412, Dec. 2019.

[18] H. Wu, Z. Wei, Y. Hou, N. Zhang, and X. Tao, "Cell-edge user offloading via flying UAV in non-uniform heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2411–2426, Apr. 2020.

[19] S. Khosravi, H. S. Ghadikolaei, and M. Petrova, "Learning-based load balancing handover in mobile millimeter wave networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2020, pp. 1–7.

[20] B. Soleimani and M. Sabbaghian, "Cluster-based resource allocation and user association in mmWave femtocell networks," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1746–1759, Mar. 2020.

[21] X. Ba and Y. Wang, "Load-aware cell select scheme for multi-connectivity in intra-frequency 5G ultra dense network," *IEEE Commun. Lett.*, vol. 23, no. 2, pp. 354–357, Feb. 2019.

[22] F. Bai and A. Helmy, "A survey of mobility modeling and analysis in wireless ad hoc networks," in *Wireless Ad Hoc and Sensor Networks*. Norwell, MA, USA: Kluwer, Jun. 2004.

[23] S. Lee, H. Yu, and H. Lee, "Multi-agent Q-learning based multi-UAV wireless networks for maximizing energy efficiency: Deployment and power control strategy design," *IEEE Internet Things J.*, early access, Sep. 16, 2021, doi: 10.1109/JIOT.2021.3113128.

[24] M. Srinivasan, V. J. Kotagi, and C. S. R. Murthy, "A Q-learning framework for user QoE enhanced self-organizing spectrally efficient network using a novel inter-operator proximal spectrum sharing," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 11, pp. 2887–2901, Nov. 2016.

[25] K. Lee, H. Lee, and D. Cho, "On the low-complexity resource allocation for self-healing with reduced message passing in indoor wireless communication systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 2080–2089, Mar. 2016.

**EUNJIN KIM** (Student Member, IEEE) received the B.S. degree in electronic and electrical engineering from Hankyong National University (HKNU), Anseong, South Korea, in 2021. Since 2021, she has been with the School of Electronic and Electrical Engineering, HKNU. Her current research interests include B5G/6G wireless communications, ultra-dense distributed networks, reinforcement learning for UDN, and the Internet of Things (IoT).

**HYUN-HO CHOI** (Senior Member, IEEE) received the B.S. *(cum laude)*, M.S., and Ph.D. degrees *(summa cum laude)* from the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2001, 2003, and 2007, respectively.

From February 2007 to February 2011, he was a Senior Engineer with the Communication Laboratory, Samsung Advanced Institute of Technology (SAIT), Suwon, South Korea. Since March 2011, he has been a Professor with Hankyong National University, Anseong, South Korea, where he is currently a Professor with the School of ICT, Robotics and Mechanical Engineering. He has also experienced as a Visiting Researcher with TeleCIS Wireless Inc., Mountain View, CA, USA, in 2006, and a Visiting Scholar with Stanford University, Stanford, CA, USA, in 2008, and the University of California at Irvine, Irvine, CA, USA, in 2017. He has published 77 international journal articles and 37 international conference papers. His current research interests include bio-inspired networking, distributed optimization, machine learning, wireless energy harvesting, mobile *ad-hoc* networks, and next-generation wireless communication.

Prof. Choi is currently a Life Member of KICS and KIICE. He received the Excellent Paper Award at ICUFN 2012, the Best Paper Award at ICN 2014, the Best Paper Award at Qshine 2016, the Excellent Paper Award in *The Journal of Korean Institute of Communications and Information Sciences* (KICS) 2018, and the Hankyong Academic awards, in 2014, 2016, and 2017. He was a co-recipient of the SAIT Patent Award, in 2010, and the Paper Award at Samsung Conference, in 2010.

**HYUNGSUB KIM** received the B.S. and M.S. degrees from the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2002 and 2004, respectively. Since 2004, he has been with the Intelligent Ultra Dense Small Cell Research Section, Electronics and Telecommunications Research Institute (ETRI), Daejeon. His research interests include 4G LTE, 5G small cells, EPC, RRC, and interface protocols for mobile communication networks.

**JEEHYEON NA** (Member, IEEE) received the B.S. degree from the Department of Computational Statistics, Chonnam National University, in 1989, and the M.S. and Ph.D. degrees from the Department of Computer Science, Chungnam National University, in 2000 and 2008, respectively. She has been with the Electronics and Telecommunications Research Institute (ETRI), since 1989, where she is currently the Director of the Intelligent Ultra Dense Small Cell Research Section. Her research interests include 4G, 5G small cells, self-organizing networks (SON), and location management and paging for mobile communication networks. She is a member of IEICE Communication Part.

**HOWON LEE** (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical and computer engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2003, 2005, and 2009, respectively. From 2009 to 2012, he was a Senior Research Staff/the Team Leader of the Knowledge Convergence Team, KAIST Institute for Information Technology Convergence (KI-ITC). Since 2012, he has been with the School of Electronic and Electrical Engineering and the Institute for IT Convergence (IITC), Hankyong National University (HKNU), Anseong, South Korea. He has also experienced as a Visiting Scholar with the University of California at San Diego (UCSD), La Jolla, CA, USA, in 2018. His current research interests include B5G/6G wireless communications, hyper-connected 3D networks, in-network computations for 3D images, cross-layer radio resource management, reinforcement learning for wireless communication networks, and autonomous Internet of Things.

He was a recipient of the Joint Conference on Communications and Information (JCCI) 2006 Best Paper Award and the Bronze Prize at Intel Student Paper Contest, in 2006. He was also a recipient of the Telecommunications Technology Association (TTA) Paper Contest Encouragement Award, in 2011, the Best Paper Award at the Korean Institute of Communications and Information Sciences (KICS) Summer Conference, in 2015, the Best Paper Award at the KICS Fall Conference, in 2015, the Honorable Achievement Award from 5G Forum Korea, in 2016, the Best Paper Award at the KICS Summer Conference, in 2017, the Best Paper Award at the KICS Winter Conference, in 2018, the Best Paper Award at the KICS Summer Conference, in 2018, the Best Paper Award at the KICS Winter Conference, in 2020, and the Best Paper Award at the Institute of Electronics and Information Engineers (IEIE) Academic Symposium, in 2021. He received the Minister's Commendation by the Minister of Science and ICT, in 2017.