IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 18, 2022, accepted April 5, 2022, date of publication April 11, 2022, date of current version April 18, 2022.

*Digital Object Identifier 10.1109/ACCESS.2022.3166179*

# Optimal Frequency Reuse and Power Control in Multi-UAV Wireless Networks: Hierarchical Multi-Agent Reinforcement Learning Perspective

**SEUNGMIN LEE[1,2], (Student Member, IEEE), SUHYEON LIM[1,2], (Student Member, IEEE), SEONG HO CHAE[3], (Member, IEEE), BANG CHUL JUNG[4], (Senior Member, IEEE), CHAN YI PARK[5], (Member, IEEE), AND HOWON LEE[1,2], (Member, IEEE)**

[1]School of Electronic and Electrical Engineering, Hankyong National University, Anseong 17579, South Korea
[2]Institute for IT Convergence (IITC), Hankyong National University, Anseong 17579, South Korea
[3]Department of Electronics Engineering, Tech University of Korea, Siheung-si 15073, South Korea
[4]Department of Electronics Engineering, Chungnam National University, Daejeon 34134, South Korea
[5]Agency for Defense Development, Daejeon 34186, South Korea

Corresponding authors: Howon Lee (hwlee@hknu.ac.kr) and Bang Chul Jung (bcjung@cnu.ac.kr)

This work was supported by the Agency for Defense Development, Republic of Korea.

**ABSTRACT** To overcome the problems caused by the limited battery lifetime in multiple-unmanned aerial vehicle (UAV) wireless networks, we propose a hierarchical multi-agent reinforcement learning (RL) framework to maximize the energy efficiency (EE) of UAVs by finding the optimal frequency reuse factor and transmit power. The proposed algorithm consists of distributed inner-loop RL for transmit power control of the UAV terminal (UT) and centralized outer-loop RL for finding the optimal frequency reuse factor. Specifically, the proposed algorithm adjusts these two factors jointly to effectively mitigate intercell interference and reduce undesired transmit power consumption in multi-UAV wireless networks. We show that, for this reason, the proposed algorithm outperforms conventional algorithms, such as a random action algorithm with a fixed frequency reuse factor and a hierarchical multi-agent Q-learning algorithm with binary transmit power controls. Furthermore, even in the environment where UTs are continuously moving based on the mixed mobility model, we show that the proposed algorithm can find the best reward when compared to conventional algorithms.

**INDEX TERMS** Unmanned aerial vehicle, optimal frequency reuse, transmit power control, energy efficiency, hierarchical multi-agent Q-learning, multi-UAV wireless network.

## I. INTRODUCTION

The utilization of unmanned aerial vehicles (UAVs) is one of the promising characteristics for future sixth-generation (6G) wireless networks because the key objective of 6G is to provide three-dimensional (3D) wireless connectivity [1]. There are many challenges to achieve this goal, e.g., 3D channel modeling, multilayered network architecture design, seamless 3D handover, and network lifetime maximization [1], [2]. In particular, the limited battery lifetime of UAVs shortens the time UAVs can operate [3], [4]. Considering this battery problem, many studies have been conducted to improve UAV's energy efficiency (EE) [5], [6] [7]. Specifically, in [5], the authors proposed a multi-agent reinforcement learning (RL)-based UAV deployment and power control algorithm for maximizing EE in multi-UAV wireless networks. Additionally, the authors of [6] proposed an online random access protocol by adjusting the packet transmission opportunities based on the residual energy of drones in S-ALOHA-based swarming drone networks.

In addition, finding optimal frequency reuse is a crucial enabling technologies to simultaneously maximize network-wide resource utilization efficiency and EE in practical wireless communication networks. First, several frequency reuse techniques have been introduced to mitigate intracell and intercell interferences when sharing frequency resources in wireless networks [8]–[10]. The authors of [8] compared two representative 4G standards, IEEE 802.16m (WiMAX) and 3GPP-LTE, which provide adaptive fractional frequency reuse (FFR) techniques based on hard FFR shutting off

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

**TABLE 1. Summary of notations and symbols.**

| Symbol | Description | Symbol | Description |
|---|---|---|---|
| $N_g$ | Number of GCSs | $N_u$ | Number of UTs |
| $N_c$ | Number of frequency resources per cell | $N_f$ | Number of frequency resources per cluster |
| $B_{tot}$ | Total bandwidth size of each cluster | $B_\mu$ | Bandwidth size of frequency resource according to $\mu$ |
| $P_{ij}^{\mathrm{TX}}$ | Transmission power of UT $j$ from GCS $i$ | $P_{ij}^{\mathrm{RX}}$ | Received power of GCS $i$ from UT $j$ |
| $R_u$ | Cell radius | $D_g$ | Inter-GCS distance |
| $\mu$ | Frequency reuse factor | $(x_j,y_j)$ | Cartesian coordinate point of UT $j$ |
| $\mu_n$ | Maximum frequency reuse factor | $h_j$ | Altitude of UT $j$ |
| $\widetilde{\mu_n}$ | Number of divisors of $\mu_n$ | $h_j^W$ | Way point of UT $j$ |
| $v_j^H$ | Horizontal movement velocity of UT $j$ | $v_j^V$ | Vertical movement velocity of UT $j$ |
| $\phi_j^H$ | Movement azimuth of UT $j$ | $T_j^d$ | Dwell time of UT $j$ |
| $p_j^s$ | Flight maintenance probability of UT $j$ | $P_{ij}^{LoS}/P_{ij}^{NLoS}$ | LoS/NLoS probability between GCS $i$ and UT $j$ |
| $d_{ij}$ | LoS distance between GCS $i$ and UT $j$ | $L_{ij}^{LoS}/L_{ij}^{NLoS}$ | LoS/NLoS path loss between GCS $i$ and UT $j$ |
| $\Delta_p^{\mathrm{TX}}$ | Step size of power increment | $L_{ij}^{LoS,ins}/L_{ij}^{NLoS,ins}$ | LoS/NLoS instantaneous path loss between GCS $i$ and UT $j$ |
| $L_{ij}^{avg}$ | Average path loss between GCS $i$ and UT $j$ | $\zeta^{LoS}/\zeta^{NLoS}$ | LoS/NLoS excessive path loss |
| $S_t^{\mathrm{out}}$ | Outer-loop state at time step $t$ | $S_t^{\mathrm{IN},(i,\mu)}$ | Inner-loop state of GCS $i$ when frequency reuse factor is $\mu$ at time step $t$ |
| $A_t^{\mathrm{out}}$ | Outer-loop action at time step $t$ | $A_t^{\mathrm{IN},(i,\mu)}$ | Inner-loop action of GCS $i$ when frequency reuse factor is $\mu$ at time step $t$ |
| $R_t$ | Reward of inner- and outer-loop RL | $P_j^{\mathrm{CRT}}$ | Fixed circuit power consumption of UT $j$ |
| $\theta_{ij}$ | Elevation angle between GCS $i$ and UT $j$ | $\psi$ | Exploration parameter |
| $T_{\mathrm{outer}}$ | Outer-loop period | $(a,b)$ | A2G channel parameters |
| $\alpha$ | Learning rate | $\beta$ | Discount factor |
| $\Gamma_{ij}$ | SINR between GCS $i$ and UT $j$ | $\eta_t^\mu(i)$ | Average energy efficiency of GCS $i$ |

interfering BSs and soft FFR limiting the transmit power of interfering BSs. Soft FFR can be beneficial because lower-power frequency resources can be exploited to support additional cell center users compared to hard FFR. In [9], strict FFR was used to partition a cell area into spatial regions with different frequency reuse factors, and soft frequency reuse (SFR) was used to divide a cell area into two regions: an inner region where the entire frequency resources were available and an outer region where a small fraction of the resources were available. SFR can be more bandwidth-efficient than strict FFR but results in more intercell interference for both cell-interior users and edge users. That is, strict FFR has the advantage of reducing interference between cell-interior users and cell-edge users, as it does not share any frequencies. To accommodate flying BSs, the authors of [10] proposed a flexible SFR (F-SFR) technique to assign a frequency resource plan that considered the dynamic network topology and maximizes inter-BS distance among cells with the same resource plan by assigning different SFR levels in each cell. Since this technique aimed at supporting aerial BSs and ground users, it is not easy to apply directly to a wireless network consisting of ground BSs and aerial terminals.

Several hierarchical reinforcement approaches have been proposed to improve the performance of multi-UAV wireless networks [11]–[13]. In [11], to resolve the problem of limited data collection coverage of the backscatter sensor nodes, the hierarchical deep reinforcement learning (DRL) framework was proposed to extend the data collection coverage and minimize the total flight time of the rechargeable UAVs when performing data collecting missions. The authors of [12] proposed a hierarchical deep Q-network (h-DQN) model for dynamic spectrum access. The proposed h-DQN shows faster convergence, higher performance, and higher channel utilization than Q-learning for dynamic sensing (QADS) [14] or deep reinforcement learning for dynamic access (DRLDA) [15]. Additional, in [13], a hierarchical scheduling architecture with top-layer scheduling for satellite selection and foundation-layer precise scheduling for urgent tasks was introduced to solve the real-time earth observation satellite (EOS) scheduling problem. Here, Q-learning with an adaptive action selection strategy was proposed to solve the Markov decision process model more efficiently. Furthermore, it is expected to realize real-time task scheduling of agile satellites. However, in complicated three-dimensional network environments where UAVs are continuously moving, it is still difficult to utilize conventional algorithms in practice due to their huge computational complexity.

In this paper, we assume multicell network environments in which multiple UAV terminals (UTs) perform their own missions, and each UT transmits its information to ground control systems (GCSs). The goal of this paper is to find the optimal frequency reuse factor and transmit power for maximizing EE in multi-UAV wireless networks by using a hierarchical multi-agent reinforcement learning algorithm. The main contributions of this paper are as follows:

- To maximize network-wide EE while reducing the computational complexity of multi-agent reinforcement learning, we adopt a hierarchical approach. Specifically,

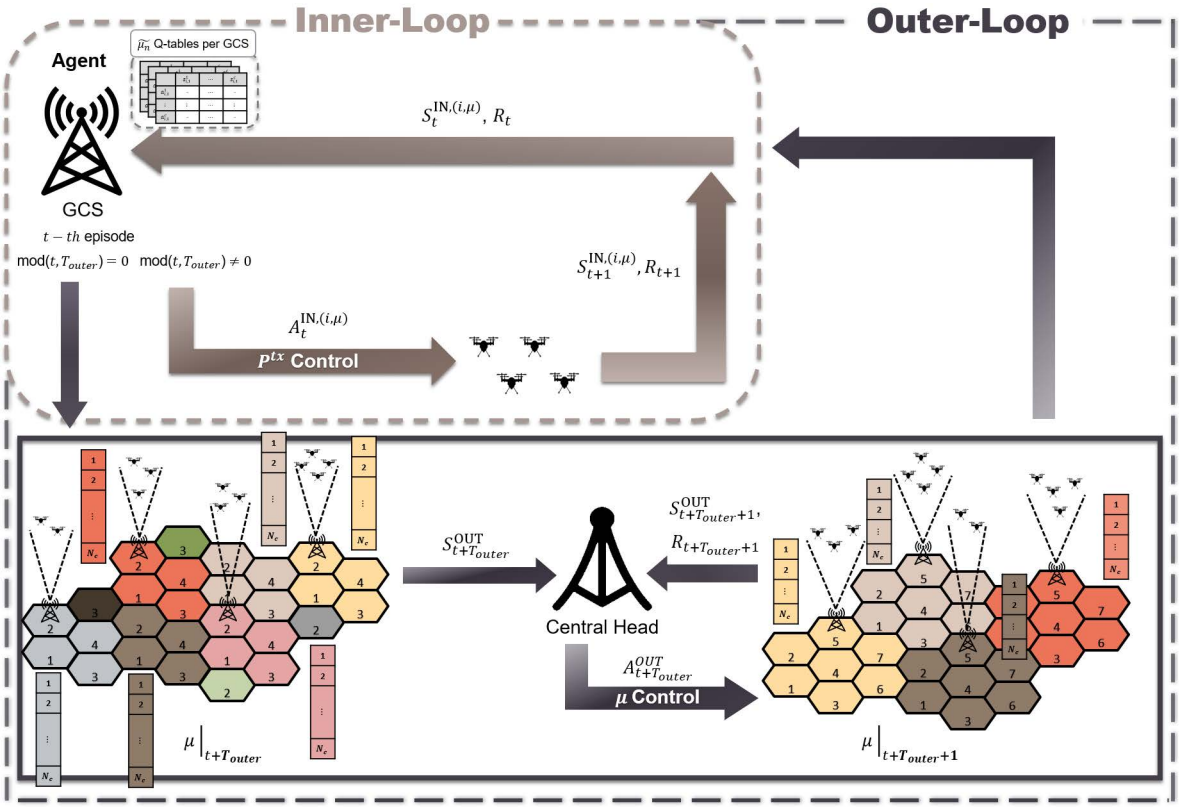



**FIGURE 1.** System model of proposed hierarchical multi-agent Q-learning framework for optimal frequency reuse and transmit power control in multi-UAV wireless networks.

the proposed algorithm consists of distributed multi-agent inner-loop RL for finding the optimal transmit power of UTs and centralized outer-loop RL for determining the optimal frequency reuse factor. The simultaneous adjustment of these factors is very challenging and complicated.

- To reduce the complexity of inner-loop RL, we propose a distributed multi-agent approach. In the inner-loop RL, each agent only considers its own state but shares its reward with others. That is, after gathering the separate reward from each agent, the central head redistributes this shared reward at each time step so that the overall reward of the hierarchical reinforcement learning can be maximized.
- Even in the hexagonal prism-shaped three-dimensional environment where UTs are continuously moving, the proposed algorithm can well-converge to the optimal solution obtained by the exhaustive search algorithm. This demonstrates the practicality and scalability of our proposed algorithm.

The rest of this paper is organized as follows. In Section II, we describe the air-to-ground (A2G) channel model and UAV mobility model. Additionally, we propose a hierarchical multi-agent Q-learning-based optimal frequency reuse and power control algorithm in Section III. In Section IV, we demonstrate that the proposed algorithm outperforms the conventional algorithms with respect to network-wide EE. Finally, the conclusions are drawn in Section V. The notations and symbols used in this paper are summarized in Table 1.

## II. SYSTEM MODEL

Fig. 1 represents the system model of the proposed hierarchical multi-agent Q-learning framework for optimal frequency reuse and transmit power control in uplink multi-UAV wireless networks. We consider a hexagonal prism-shaped three-dimensional cell architecture with $N_g$ GCSs and $N_u$ UTs, and the number of cells per cluster is determined by the frequency reuse factor ($\mu$).

### A. AIR-TO-GROUND (A2G) CHANNEL MODEL

In the A2G channel model, the line-of-sight (LoS) signal between UTs and GCSs may be occasionally blocked by ground buildings and obstacles so that LoS and non-LoS (NLoS) propagations should be separately considered. Thus, we herein utilize the elevation angle-dependent probabilistic LoS model as the A2G channel model [16]. The LoS path loss and NLoS path loss between GCS $i$ and UT $j$ can be represented as

$$L_{ij}^{LoS} = 20\log_{10}\left(\frac{4\pi f_c d_{ij}}{v_l}\right) + \zeta^{LoS}, \tag{1}$$

$$L_{ij}^{NLoS} = 20\log_{10}\left(\frac{4\pi f_c d_{ij}}{v_l}\right) + \zeta^{NLoS}, \tag{2}$$

where $v_l$ is the speed of light, $f_c$ is the carrier frequency, and $d_{ij}$ denotes the distance between GCS $i$ and UT $j$. In

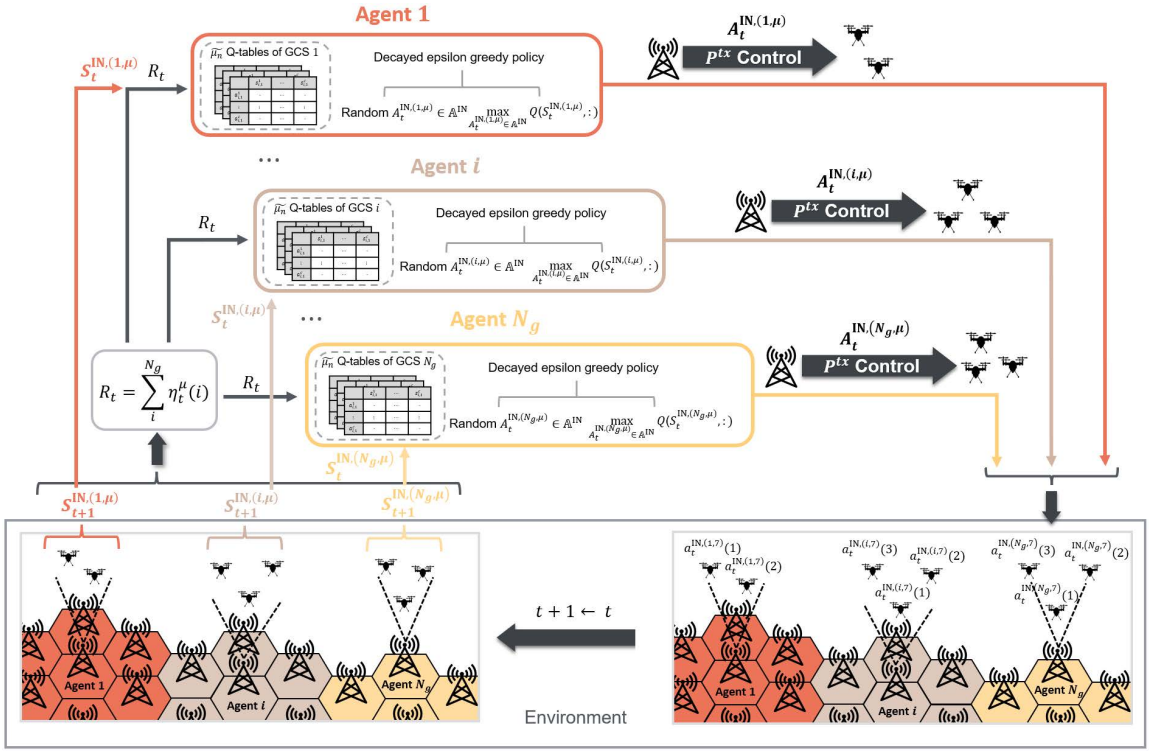



**FIGURE 2.** Detailed operation of distributed multi-agent Q-learning inner-loop RL.

Equations (1) and (2), the free space path loss is common and can be calculated as '$20\log\left(\frac{4\pi f_c d_{ij}}{v_l}\right)$'. Additionally, $\zeta^{LoS}$ and $\zeta^{NLoS}$ denote the excessive path loss caused by artificial obstacles (e.g., skyscrapers) in LoS and NLoS paths, respectively. The excessive path loss varies depending on the urban environmental deployment models proposed by the International Telecommunication Union - Radio communication sector (ITU-R) [17]. Even though the A2G channel model used in this paper does not consider small-scale fading directly, it reflects the A2G channel characteristics of various environmental deployments (suburban, urban, dense urban, and highrise urban) by using these statistical parameters.

In this A2G channel model, the LoS probability between GCS $i$ and UT $j$ can be calculated as

$$P_{ij}^{LoS}(\theta_{ij}) = \frac{1}{1 + a \times \exp(-b \times (\theta_{ij} - a))}, \tag{6}$$

where $a$ and $b$ are other statistical parameters representing the A2G channel characteristics of four urban environmental deployments. $\theta_{ij}$ is the elevation angle between GCS $i$ and UT $j$, and can be calculated as $\theta_{ij} = \arcsin\frac{h_j}{d_{ij}}$, where $h_j$ denotes the altitude of UT $j$. From Equation (6), the NLoS probability between GCS $i$ and UT $j$ can be obtained as $P_{ij}^{NLoS} = 1 - P_{ij}^{LoS}$. Using Equations (1)–(6), the average path loss of the A2G link between GCS $i$ and UT $j$ considering the LoS and NLoS probabilities can be described as

$$L_{ij}^{avg} = 10^{\left(P_{ij}^{LoS}/20\right)} \times L_{ij}^{LoS} + 10^{\left(P_{ij}^{NLoS}/20\right)} \times L_{ij}^{NLoS}. \tag{7}$$

From Equation (7), the received power of GCS $i$ from UT $j$ ($P_{ij}^{\mathrm{RX}}$) can be represented as

$$P_{ij}^{\mathrm{RX}} = \frac{P_{ij}^{\mathrm{TX}}}{10^{\left(P_{ij}^{LoS}/20\right)} \times L_{ij}^{LoS} + 10^{\left(P_{ij}^{NLoS}/20\right)} \times L_{ij}^{NLoS}}, \tag{8}$$

where $P_{ij}^{\mathrm{TX}}$ indicates the transmit power of UT $j$ associated with GCS $i$.

As mentioned above, instead of instantaneous small-scale fading, excessive path loss depending on the LoS and NLoS paths, $\zeta^{LoS}$ and $\zeta^{NLoS}$, is included in the path loss model. If the small-scale fading effect is included, the empirical mean of the signal-to-interference-plus-noise-ratio (SINR) between GCS $i$ and UT $j$ is given by Equations (3)–(5), as shown at the bottom of the next page. Here, $\mathbb{E}\{\cdot\}$ is an expectation, and $\sigma^2$ is the thermal noise power. In addition, $L_{ij}^{LoS,\mathrm{ins}}$ and $L_{ij}^{NLoS,\mathrm{ins}}$ are instantaneous path losses caused by small-scale fading of LoS and NLOS paths, respectively. Here, Jensen's inequality is used in both inequalities (4) and (5). For a simple performance evaluation, the upper bound in (5) is adopted as a performance measure instead of the empirical mean of the SINR. Furthermore, we verify that the upper bound is sufficiently tight to be used in the proposed algorithm. Accordingly, SINR between GCS $i$ and UT $j$ can be represented as

$$\Gamma_{ij} = \sum_{k=1}^{N_c} \left( \frac{P_{ij}^{\mathrm{RX}} \cdot I_t^{ij}(k)}{\sigma^2 + \sum_{m=1, m\neq j}^{N_u} \left(P_{im}^{\mathrm{RX}} \cdot \sum_{n=1, n\neq i}^{N_g} I_t^{nm}(k)\right)} \right), \tag{9}$$

$$I_t^{ij}(k) = \begin{cases} 1, & \text{if } k\text{-th resource is assigned to UT } j, \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

In Equation (10), $I_t^{ij}(k)$ is an indicator function that determines whether the $k$-th frequency resource of GCS $i$ is assigned to UT $j$.

### B. UAV MOBILITY MODEL

UTs' movement is modeled as a mixed mobility (MM) model, which is a combination of the random waypoint and random walk models [18]. The detailed operation of the mixed mobility model is described as follows.

1) When the current way point is $h_j^W(t)$, UT $j$ randomly selects the next way point $h_j^W(t+1)$ between $[h_{\min}, h_{\max}]$. Additionally, UT $j$ randomly selects a horizontal movement speed $v_j^H(t)$ between $[v_{\min}^H, v_{\max}^H]$ and an azimuthal $\phi_j^H(t)$ between $[180°, -180°]$.
2) UT $j$ ascends or descends to the next way point $h_j^W(t+1)$ with a vertical movement speed $v_j^V(t)$ between $[v_{\min}^V, v_{\max}^V]$.
3) At the $(x_j(t), y_j(t))$ position, UT $j$ randomly chooses the dwell time $T_j^d$ between $[T_{\min}^d, T_{\max}^d]$.
4) According to the flight maintenance probability $p_j^s$, UT $j$ holds or alters its position during $T_j^d$ as follows.

$$x_j(t+1) = \begin{cases} x_j(t), & 1-p_j^s, \\ x_j(t) + v_j^H(t)\cos(\phi_j^H(t)) \times T_j^d, & p_j^s. \end{cases} \tag{11}$$

$$y_j(t+1) = \begin{cases} y_j(t), & 1-p_j^s, \\ y_j(t) + v_j^H(t)\sin(\phi_j^H(t)) \times T_j^d, & p_j^s. \end{cases} \tag{12}$$

5) The process is repeated until the learning is complete.

## III. HIERARCHICAL MULTI-AGENT Q-LEARNING FOR OPTIMAL FREQUENCY REUSE AND POWER CONTROL

To maximize network-wide EE in multi-UAV wireless networks while reducing the computational complexity of RL, we propose a hierarchical multi-agent Q-learning framework. As shown in Fig. 1, the proposed RL framework consists of distributed multi-agent inner-loop RL for finding UTs' optimal transmit power and centralized outer-loop RL for determining the optimal frequency reuse factor. The change in the frequency reuse factor requires a large control overhead in terms of the entire network so that it needs to be adjusted intermittently through the outer-loop RL. In contrast, UTs' transmit power can be controlled every time step through the inner-loop RL to maximize network-wide EE. The detailed operations of the inner-loop RL and the outer-loop RL are as follows.

### A. INNER-LOOP RL

The goal of the inner-loop RL is to find the optimal transmit power of UTs for maximizing EE. Fig. 2 shows the detailed operation of the distributed multi-agent inner-loop RL. In the inner-loop RL, each GCS performs the role of an agent. According to the frequency reuse factor ($\mu$) the number of frequency resources available in each agent is determined. Additionally, each agent should manage $\widetilde{\mu_n}$ different Q-tables, where $\widetilde{\mu_n}$ is the number of divisors of $\mu_n$. Here, $\mu_n$ denotes the maximum value of $\mu$, and $\mu$ is determined by the outer-loop RL. Namely, $\mu$ is one of the divisors of $\mu_n$. Additionally, the agents share their rewards with the central head, and the integrated reward is redistributed to the agents in each time step. In the proposed inner-loop RL algorithm, because each agent considers only its action set and state set, the computational complexity of the proposed algorithm is significantly reduced compared to the centralized Q-learning algorithm. Detailed descriptions for the state, action, and reward of the inner-loop RL are expressed as follows.

$$\Gamma_{ij} = \left[ P_{ij}^{LoS} \cdot \mathbb{E}\left\{ \sum_{k=1}^{N_c} \left( \frac{P_{ij}^{\mathrm{TX}} \times I_t^{ij}(k)}{L_{ij}^{LoS,ins}(\sigma^2 + \sum_{m=1,m\neq j}^{N_u}(P_{im}^{\mathrm{RX}} \times \sum_{n=1,n\neq i}^{N_g} I_t^{nm}(k)))} \right) \right\} \right.$$
$$\left. + P_{ij}^{NLoS} \cdot \mathbb{E}\left\{ \sum_{k=1}^{N_c} \left( \frac{P_{ij}^{\mathrm{TX}} \times I_t^{ij}(k)}{L_{ij}^{NLoS,ins}(\sigma^2 + \sum_{m=1,m\neq j}^{N_u}(P_{im}^{\mathrm{RX}} \times \sum_{n=1,n\neq i}^{N_g} I_t^{nm}(k)))} \right) \right\} \right], \tag{3}$$

$$\leq \left[ P_{ij}^{LoS} \cdot \sum_{k=1}^{N_c} \left( \frac{P_{ij}^{\mathrm{TX}}}{L_{ij}^{LoS}} \times \frac{I_t^{ij}(k)}{\sigma^2 + \sum_{m=1,m\neq j}^{N_u}(P_{im}^{\mathrm{RX}} \times \sum_{n=1,n\neq i}^{N_g} I_t^{nm}(k))} \right) \right.$$
$$\left. + P_{ij}^{NLoS} \cdot \sum_{k=1}^{N_c} \left( \frac{P_{ij}^{\mathrm{TX}}}{L_{ij}^{NLoS}} \times \frac{I_t^{ij}(k)}{\sigma^2 + \sum_{m=1,m\neq j}^{N_u}(P_{im}^{\mathrm{RX}} \times \sum_{n=1,n\neq i}^{N_g} I_t^{nm}(k))} \right) \right], \tag{4}$$

$$\leq \sum_{k=1}^{N_c} \left( \frac{P_{ij}^{\mathrm{TX}}}{L_{ij}^{avg}} \times \frac{I_t^{ij}(k)}{\sigma^2 + \sum_{m=1,m\neq j}^{N_u}(P_{im}^{\mathrm{RX}} \times \sum_{n=1,n\neq i}^{N_g} I_t^{nm}(k))} \right). \tag{5}$$

**Algorithm 1:** Detailed Procedure of Proposed Hierarchical Multi-Agent Q-Learning Based Optimal Frequency Reuse and Power Control in Multi-UAV Wireless Networks

Place GCSs according to the inter-GCS spacing distance, $D_g$.
Determine the altitude of UTs randomly between $h_{\min}$ and $h_{\max}$ and place UTs within $R_u$ on the basis of the horizontal position of each GCS.
Initialize Q-tables of inner-loop and outer-loop agents, and the frequency reuse factor $\mu$.
Partition GCSs into ceil($\frac{N_g}{\mu}$) clusters.
**for** *Every episode* **do**
  **for** *Every iteration* **do**
    Calculate SINR between GCS $i$ and UT $j$ ($\Gamma_{ij}$), for all $i$ and $j$.
    Allocate frequency band $k$ of GCS $i$, which provides the greatest SINR, to UT $j$, for all $j$.
    **if** mod $(t, T_{outer}) == 0$ **then**
      Choose the outer-loop action $A_t^{\mathrm{OUT}}$ by decayed $\varepsilon$-greedy policy.

$$A_t^{\mathrm{OUT}} = \begin{cases} \text{Random Action}, & \text{with probability } \varepsilon, \\ \arg\max_{A_t^{\mathrm{OUT}}} Q(S_t^{\mathrm{OUT}}, :), & \text{with probability } 1-\varepsilon. \end{cases}$$

      The central header adjusts the frequency reuse factor $\mu$ according to $A_t^{\mathrm{OUT}}$ and re-partition GCSs.
      Re-calculate $\Gamma_{ij}(t)$ and $\eta_t^{\mu}(i)$.
      Move on to the next state $S_{t+1}^{\mathrm{OUT}}$, calculate $R_{t+1}$, and update the Q-value of outer-loop RL.
      $Q(S_t^{\mathrm{OUT}}, A_t^{\mathrm{OUT}}) \leftarrow (1-\alpha)Q(S_t^{\mathrm{OUT}}, A_t^{\mathrm{OUT}}) + \alpha \times (R_{t+1} + \beta \times \max_{A_{t+1}^{\mathrm{OUT}}} Q(S_t^{\mathrm{OUT}}, :)$.
    **else**
      Choose the inner-loop action $A_t^{\mathrm{IN},(i,\mu)}$ by decayed $\varepsilon$-greedy policy.

$$A_t^{\mathrm{IN},(i,\mu)} = \begin{cases} \text{Random Action}, & \text{with probability } \varepsilon, \\ \arg\max_{A_t^{\mathrm{IN},(i,\mu)}} Q(S_t^{\mathrm{IN},(i,\mu)}, :), & \text{with probability } 1-\varepsilon. \end{cases}$$

      GCS $i$ adjusts the transmit power of UT $j$ according to $A_t^{\mathrm{IN},(i,\mu)}(j)$, for all $i$ and $j$.
      Re-calculate $\Gamma_{ij}(t)$ and $\eta_t^{\mu}(i)$.
      Move on to the next state $S_{t+1}^{\mathrm{IN},(i,\mu)}$, calculate $R_{t+1}$, and update the Q-value of inner-loop RL.
      $Q(S_t^{\mathrm{IN},(i,\mu)}, A_t^{\mathrm{IN},(i,\mu)}) \leftarrow (1-\alpha)Q(S_t^{\mathrm{IN},(i,\mu)}, A_t^{\mathrm{IN},(i,\mu)}) + \alpha \times (R_{t+1} + \beta \times \max_{A_{t+1}^{\mathrm{IN},(i,\mu)}} Q(S_{t+1}^{\mathrm{IN},(i,\mu)}, :)$.
    **end**
  **end**
  Update the three-dimensional positions of UTs based on the mixed UAV mobility model.
**end**

- **Inner-loop RL state:** When the frequency reuse factor is $\mu$ at time step $t$, the inner-loop state of GCS $i$ ($S_t^{\mathrm{IN},(i,\mu)}$) is defined as

$$S_t^{\mathrm{IN},(i,\mu)} = [s_t^{\mathrm{IN},(i,\mu)}(1), \cdots, s_t^{\mathrm{IN},(i,\mu)}(N_c)], \quad (13)$$
$$s_t^{\mathrm{IN},(i,\mu)}(k) \in \{P_{\min}^{\mathrm{TX}}, \cdots, P_{\max}^{\mathrm{TX}}, \mathrm{None}\}. \quad (14)$$

  Here, $s_t^{\mathrm{IN},(i,\mu)}(k)$ denotes the amount of transmit power of the $k$-th frequency resource when the frequency reuse factor is $\mu$ at time step $t$ and $|S_t^{\mathrm{IN},(i,\mu)}| = N_c$. In Equation (14), "None" means that the $k$-th frequency resource is not assigned to any UTs. Additionally, $P_{\min}^{\mathrm{TX}}$ and $P_{\max}^{\mathrm{TX}}$ indicate the minimum and maximum transmission power of UTs, respectively.
- **Inner-loop RL action:** When the frequency reuse factor is $\mu$ at time step $t$, GCS $i$ adjusts UTs' transmit power associated with it as follows.

$$A_t^{\mathrm{IN},(i,\mu)} = [a_t^{\mathrm{IN},(i,\mu)}(1), \cdots, a_t^{\mathrm{IN},(i,\mu)}(N_c)]. \quad (15)$$
$$a_t^{\mathrm{IN},(i,\mu)}(k) \in \{\Delta_p^{\mathrm{TX}}, -\Delta_p^{\mathrm{TX}}, 0\}. \quad (16)$$

  In Equation (15) and (16), $A_t^{\mathrm{IN},(i,\mu)}$ indicates the inner loop RL action of GCS $i$, and $a_t^{\mathrm{IN},(i,\mu)}(k)$ is the element of $A_t^{\mathrm{IN},(i,\mu)}$. Additionally, $\Delta_p^{\mathrm{TX}}$, $-\Delta_p^{\mathrm{TX}}$, and "0" represent "transmit power up", "transmit power down" and "maintain the current transmit power", respectively.
- **Inner-loop RL reward:** The objective of the proposed hierarchical multi-agent Q-learning is maximizing the EE of the multi-UAV wireless networks. Accordingly, we define EE ($\eta_t^{\mu}(i)$) as

$$\eta_t^{\mu}(i) = \sum_{j}^{N_u} \sum_{k}^{N_c} \left( \frac{B_{\mu} \log_2(1+\Gamma_{ij}(t))}{P_j^{\mathrm{CRT}} + P_{ij}^{\mathrm{TX}}} \times I_t^{ij}(k) \right). \quad (17)$$

Here, $\eta_t^{\mu}(i)$ is the EE of GCS $i$ when the frequency reuse factor is $\mu$ at time step $t$. $B_\mu$ is the bandwidth size of each frequency resource when the frequency reuse factor is $\mu$, and $B_\mu = B_{tot}/N_f$ where $B_{tot}$ is the total bandwidth of each cluster and $N_f$ is the total number of frequency resources. Additionally, $P_j^{\mathrm{CRT}}$ is the fixed circuit power consumption of UT $j$. Using Equation (17), the reward of the proposed hierarchical multi-agent Q-learning algorithm can be expressed as

$$\begin{aligned} R_t &= \sum_{i}^{N_g} \eta_t^{\mu}(i), \\ &= \sum_{i}^{N_g} \sum_{j}^{N_u} \sum_{k}^{N_c} \left( \frac{B_\mu \log_2(1+\Gamma_{ij}(t))}{P_j^{\mathrm{CRT}} + P_{ij}^{\mathrm{TX}}} \times I_t^{ij}(k) \right). \end{aligned} \tag{18}$$

### B. OUTER-LOOP RL

To find the optimal frequency reuse factor $\mu$, we consider a central header that manages all GCSs as the agent of outer-loop RL. The total number of frequency resources ($N_f$) that can be used in the entire network is fixed. Therefore, according to the variation in $\mu$, the number of resources available in each GCS ($N_c$) is different, $N_c = N_f/\mu$. As $\mu$ increases, the number of resources available in each GCS reduces, but intercell interference also decrease. Therefore, finding the optimal $\mu$ is crucial for maximizing network-wide EE. Detailed descriptions for the state, action, and reward of the outer-loop RL are described as follows.

- **Outer-loop RL state:** At time step $t$, the outer-loop state ($S_t^{\mathrm{OUT}}$) is defined as

$$S_t^{\mathrm{OUT}} = \mu. \tag{19}$$

  In practical wireless networks, the number of divisors of $\mu_n$ is not large so that the Q-table size of the centralized outer-loop RL does not become large. That is, the Q-table size can be calculated as $|\mathbb{S}_t^{\mathrm{OUT}}| \times |\mathbb{A}_t^{\mathrm{OUT}}|$ where $\mathbb{S}_t^{\mathrm{OUT}}$ and $\mathbb{A}_t^{\mathrm{OUT}}$ denote the state set and action set of the outer-loop RL, respectively.
- **Outer-loop RL action:** In outer-loop RL, the agent adjusts the frequency reuse factor $\mu$ as follows.

$$A_t^{\mathrm{OUT}} \in \{\Delta_\mu, -\Delta_\mu, 0\}, \tag{20}$$

  where $\Delta_\mu$, $-\Delta_\mu$, and '0' denote "increase in $\mu$", "decrease in $\mu$", and "maintain the current $\mu$", respectively.
- **Outer-loop RL reward:** The reward of outer-loop RL is the same as the reward of inner-loop RL.

### C. POLICY

We adopt the decayed $\varepsilon$-greedy model to adaptively control the ratio between exploitation and exploration [19]. The policy utilized in this paper is as follows.

$$a_t = \begin{cases} \text{Random Action}, & \text{with probability } \varepsilon, \\ \arg\max\limits_{a_t \in \mathbb{A}} Q(s_t, a_t), & \text{with probability } 1-\varepsilon. \end{cases} \tag{21}$$

**TABLE 2. Simulation parameters.**

| Parameters | Value |
|---|---|
| Frequency reuse factor | 1, 2, 3, 6 |
| Number of cells ($N_g$) | 12 |
| Number of UTs ($N_u$) | 72 |
| Number of frequency resources ($N_f$) | 6 |
| Carrier frequency ($f_c$) | 1 (GHz) |
| Frequency bandwidth ($B_\mu$) | 200 (kHz) |
| Excessive path loss ($\zeta_{\mathrm{LoS}}, \zeta_{\mathrm{NLoS}}$) | 1, 20 (dB) |
| Urban deployment parameters ($a, b$) | 9.6117, 0.1581 |
| Thermal noise power ($\sigma^2$) | -120 (dBm) |
| Vertical way point interval ($[h_{\min}, h_{\max}]$) | [120, 150] (m) |
| Dwell time interval ($[T_{\min}^d, T_{\max}^d]$) | [2, 4] (s) |
| Vertical movement speed ($[v_{\min}^V, v_{\max}^V]$) | [3, 7] (m/s) |
| Horizontal movement speed ($[v_{\min}^H, v_{\max}^H]$) | [5, 10] (m/s) |
| Flight maintenance probability ($p^s$) | 0.5 |
| Episode time duration | 0.5 (ms) |
| Max. and min. transmit power ($P_{\min}^{\mathrm{TX}}, P_{\max}^{\mathrm{TX}}$) | 0.5, 2.0 (W) |
| Fixed circuit power consumption ($P^{\mathrm{CRT}}$) | 20 (W) |
| Step size of power increment ($\Delta_p^{\mathrm{TX}}$) | 0.75 (W) |
| Outer-loop RL period ($T_{\mathrm{outer}}$) | 20 (episodes) |
| Learning rate ($\alpha$) | 0.1 |
| Discount factor ($\beta$) | 0.9 |
| Exploration parameter ($\psi$) | 130 |

Here, $\varepsilon = \varepsilon_{\mathrm{init}} \times (1-\varepsilon_{\mathrm{init}})^{\frac{E}{\psi \times |\mathbb{A}|}}$. $\varepsilon_{\mathrm{init}}$ and $E$ represent the initial value of $\varepsilon$ and the current episode index, respectively. Additionally, $|\mathbb{A}|$ represents the cardinality of $\mathbb{A}$ and $\psi$ is an exploration parameter to adjust the attenuation rate of $\varepsilon$. It is noteworthy that the importance of exploration depends on the number of actions. In this policy, the agent acts randomly with $\varepsilon$ and chooses the optimal action maximizing the reward obtained from the Q-value with $1-\varepsilon$. According to the variation in $\varepsilon$, we can adaptively adjust the ratio between exploration and exploitation to find the optimal solution quickly and accurately.

### D. Q-TABLE UPDATE

The Q-table of the proposed hierarchical multi-agent Q-learning is updated as follows.

$$\begin{aligned} Q(S_t, A_t) \leftarrow (1-\alpha) \times Q(S_t, A_t) \\ + \alpha \times (R_{t+1} + \beta \times \max_{A_{t+1} \in \mathbb{A}} Q(S_{t+1}, A_{t+1}), \end{aligned} \tag{22}$$

where $\alpha$ and $\beta$ are the learning rate and discount factor, respectively. With $\alpha$, we can control the speed of the Q-value update, and the ratio between the current reward and the future expected reward is adjusted by using $\beta$.

In the proposed hierarchical multi-agent Q-learning algorithm, first, GCSs are placed with a constant spacing $D_g$ and UTs are randomly distributed in a three-dimensional

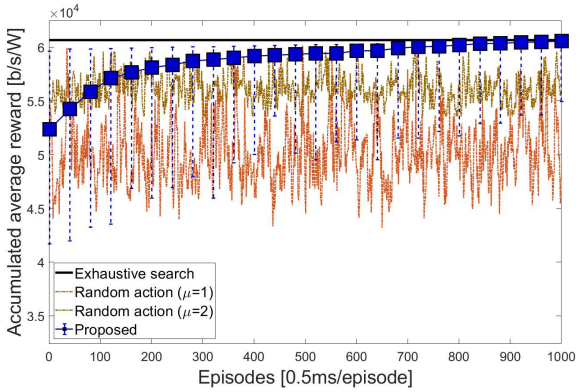



**FIGURE 3.** Accumulated average reward for exhaustive search, random action, and proposed algorithms when $N_g = 4$, $N_u = 8$, $N_f = 2$, $D_g = 100$, $R_u = 50$, and $\mu_n = 2$.

network area based on each GCS. Because each inner-loop agent needs to learn $\widetilde{\mu_n}$ Q-tables, relatively more learning opportunities should be given to the inner-loop agent than the outer-loop agent. Therefore, outer-loop RL performed intermittently, every $T_{\text{outer}}$. Additionally, in each episode, the 3D positions of UTs are updated according to the MM model. The detailed procedure of the proposed hierarchical multi-agent Q-learning based optimal frequency reuse and power control algorithm is summarized in Algorithm 1.

Finally, the proposed hierarchical multi-agent Q-learning-based optimal frequency reuse and power control problem for maximizing network-wide EE can be formulated as

$$\mathcal{P}_0: \quad \max_{\mathbb{P}_t^{\text{TX}}, \mu_t} \; R_t = \sum_{i}^{N_g} \eta_t^{\mu}(i) \tag{23}$$

$$s.t.\, \mathcal{C}_1 : P_{\min}^{\text{TX}} \le P_{ij}^{\text{TX}} \le P_{\max}^{\text{TX}}, \quad \forall i \in \mathbb{N}_g,\ j \in \mathbb{N}_u, \tag{24}$$

$$\mathcal{C}_2 : \sum_{i}^{N_g} \sum_{k}^{N_c} I_t^{ij}(k) \le 1, \quad \forall j \in \mathbb{N}_u, \tag{25}$$

$$\mathcal{C}_3 : \sum_{t=0}^{\infty} \sum_{i^* \ne i}^{N_g} \sum_{k}^{N_c} I_t^{i^* j}(k) = 0, \quad \forall j \in \mathbb{N}_u, \tag{26}$$

$$\mathcal{C}_4 : \sum_{k}^{N_c} \sum_{j}^{N_u} I_t^{ij}(k) \le N_c, \quad \forall i \in \mathbb{N}_g. \tag{27}$$

Here, $\mathcal{C}_1$ is the constraint of the transmit power for the UTs, and $\mathcal{C}_2$ describes the constraint that only one channel of the serving GCS can be allocated to a UT. Additionally, $\mathcal{C}_3$ defines the constraint that the serving GCS is not changed during the end of the learning and $\mathcal{C}_4$ means that GCS $i$ has $N_c$ frequency resources.

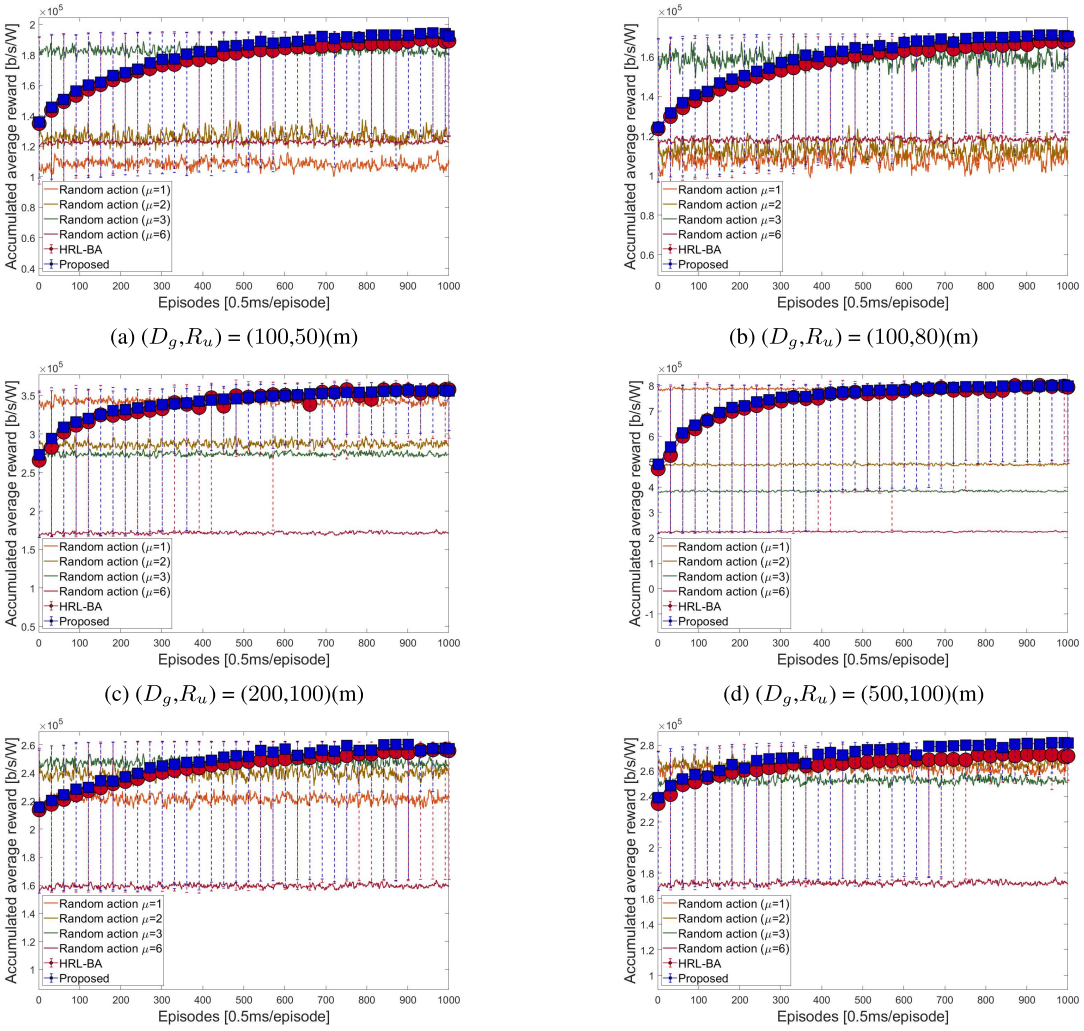
## IV. RESULTS AND DISCUSSION

We show the performance results according to the variations in the inter-GCS distance ($D_g$) and cell radius ($R_u$). We conducted simulations considering the following ($D_g$,$R_u$) combinations: (100, 50)(m), (100, 80)(m), (200, 100)(m), (500, 100)(m), (200, 150)(m), and (500, 300)(m). Initially, users of each cell were randomly distributed within the cell radius of $R_u$. However, not all GCSs served the same number of UTs because each UT associated with the GCS that provided the highest received signal power. In addition, when $D_g < 2R_u$, the number of UTs in a cell boundary or overlapping area increases, and the intercell interference becomes severe. Conversely, when $D_g \ge 2R_u$, GCSs might receive relatively low intercell interference. Other simulation parameters are summarized in Table 2. Furthermore, a random action (RA) algorithm and a hierarchical RL-based binary action algorithm (HRL-BA) were considered benchmarks to compare the performance of the proposed algorithm in terms of network-wide EE. A detailed description of these benchmark algorithms is as follows.

- **Random Action (RA) Algorithm with Fixed $\mu$**: Each UT randomly chooses its transmit power assuming that $\mu$ is fixed. As the optimal solution cannot be obtained in complicated network environments, the convergence of the proposed hierarchical multi-agent Q-learning algorithm to the optimal solution can be roughly demonstrated through the random action algorithm. That is, because of the extremely high computational complexity required for obtaining the optimal solution based on the exhaustive search, we compared the performance of the proposed algorithm with the random action algorithm for each $\mu$.
- **Hierarchical RL-based Binary Action (HRL-BA) Algorithm**: Similar to the proposed algorithm, HRL-BA exploits a hierarchical multi-agent Q-learning framework. However, HRL-BA has binary actions when agents adjust the transmit power of UTs. HRL-BA has a relatively small computational complexity compared to the proposed algorithm.

Fig. 3 shows the accumulated average reward versus episode for exhaustive search, random action, and proposed algorithms when $N_g = 4$, $N_u = 8$, $N_f = 2$, $T_{\text{outer}} = 20$, $D_g = 100$, $R_u = 50$, and $\mu_n = 2$. To obtain these results, we performed 1, 000 episodes, where each episode had 5, 000 iterations. As shown in this figure, we find that the proposed hierarchical multi-agent RL algorithm well converges to the optimal solution obtained by the exhaustive search algorithm.

Figs. 4a-4f show the accumulated average reward versus episode for the proposed, HRL-BA, and random action algorithms under $N_g = 12$, $N_u = 72$, and $N_f = 6$ according to combinations of ($D_g$,$R_u$). Figs. 4a, 4b, 4c, 4d, 4e, and 4f show the results under ($D_g$, $R_u$) = (100, 50)(m), (100, 80)(m), (200, 100)(m), (500, 100)(m), (200, 150)(m), (500, 300)(m), respectively. To obtain these results, we performed 1, 000 episodes, where each episode had 50, 000 iterations. Each random action algorithm had a fixed frequency reuse factor. Thus, this algorithm cannot obtain the performance improvement according to the change in the frequency reuse factor. That is, the random




(a) $(D_g,R_u)$ = (100,50)(m) (b) $(D_g,R_u)$ = (100,80)(m)

(c) $(D_g,R_u)$ = (200,100)(m) (d) $(D_g,R_u)$ = (500,100)(m)

(e) $(D_g,R_u)$ = (200,150)(m) (f) $(D_g,R_u)$ = (500,300)(m)

**FIGURE 4.** Accumulated average EE for proposed and random action algorithms under $N_g = 12$, $N_u = 72$, and $N_f = 6$ according to variation in $(D_g,R_u)$.

**TABLE 3.** Maximum EE and throughput for proposed and random action algorithms $N_g = 12$, $N_u = 72$, and $N_f = 6$ according to variation in $(D_g,R_u)$.

| Algorithm | Energy Efficiency (×100Kb/s/Watt) | | | | | | Throughput (×1Mb/s) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (100,50) | (100,80) | (200,100) | (200,150) | (500,100) | (500,300) | (100,50) | (100,80) | (200,100) | (200,150) | (500,100) | (500,300) |
| Proposed | 1.9279 | 1.7075 | 3.6092 | 8.0385 | 2.6390 | 2.8129 | 3.9879 | 3.6035 | 7.6856 | 16.9048 | 5.4287 | 5.8623 |
| RA ($\mu$=1) | 1.1747 | 1.1656 | 3.5310 | 7.9918 | 2.2907 | 2.7206 | 2.5298 | 2.5170 | 7.5630 | 17.1432 | 4.9201 | 5.8706 |
| RA ($\mu$=2) | 1.3744 | 1.2482 | 2.9721 | 5.0260 | 2.4808 | 2.7393 | 2.9585 | 2.7039 | 6.3410 | 10.5771 | 5.3268 | 5.8377 |
| RA ($\mu$=3) | 1.8635 | 1.6727 | 2.8055 | 3.9095 | 2.5519 | 2.5871 | 3.9610 | 3.5480 | 5.9683 | 8.2809 | 5.4399 | 5.4964 |
| RA ($\mu$=6) | 1.2456 | 1.1536 | 1.7576 | 2.2944 | 1.6398 | 1.7655 | 2.5908 | 2.4068 | 3.6411 | 4.7501 | 3.4005 | 3.6532 |

action algorithm can obtain only performance improvement according to power control under the fixed frequency reuse factor. If the transmit power of a UT becomes larger, the received signal strength increases. At the same time, the intercell interference signal could increase, making it very important and difficult to obtain the optimal transmission power.

By comparing the combinations of (Fig. 4a, Fig. 4b) and (Fig. 4c, Fig. 4e), when $R_u$ increases for the same $D_g$, the EE relatively decreases due to the increase in the amount of intercell interference. In contrast, when $D_g$ increases for the same $R_u$, EE can increase because the intercell interference decreases. Additionally, since each GCS can use more frequency resources when $\mu$ is low, the average EE

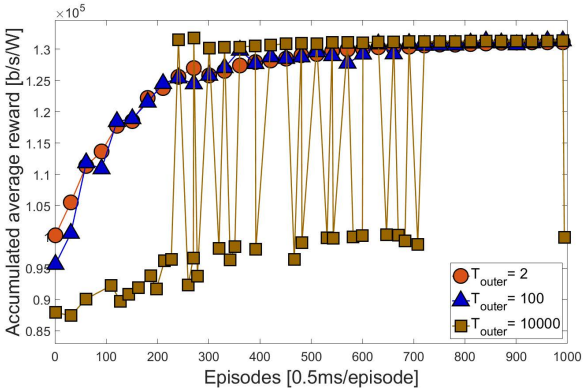



**FIGURE 5.** Accumulated average EE for proposed algorithms according to variation in $T_{\text{outer}}$ when $N_g = 8$, $N_u = 32$, $N_f = 4$, $D_g = 100$, $R_u = 50$, $\mu_n = 4$.

value considering the relatively large number of users may be smaller than when $\mu$ is high. Furthermore, as shown in these figures, HRL-BA has a slightly lower EE than the proposed algorithm. Because there are only two actions to choose from, performance degradation in terms of EE might occur in HRL-BA.

Because obtaining the optimal solution by using the exhaustive search is impossible owing to the computational complexity, we find that the proposed algorithm can roughly converge to the optimal solution for all simulation scenarios through the error bars in Figs. 4a-4f. That is, by adjusting UT's transmit power in inner-loop RL and finding the optimal frequency reuse factor in outer-loop RL, the proposed hierarchical multi-agent Q-learning can increase the chances of finding the optimal solution.

Moreover, the performance behavior of the proposed algorithm according to the variation in $T_{\text{outer}}$ is shown in Fig. 5. When $T_{\text{outer}} = 2$, very frequent changes of $\mu$ are needed, and thus it will be a significant burden to network operators. In contrast, when $T_{\text{outer}}$ becomes significantly larger, it is difficult to find the optimal number of frequency resources in each cell, and thus large reward oscillations occur as the episode progresses. That is, GCSs should find the optimal network-wide EE by performing transmit power control only. Therefore, it is very important to set an appropriate $T_{\text{outer}}$ value in consideration of the characteristics of the network environment.

Table 3 summarizes the maximum EE and throughput for the proposed and random action algorithm under $N_g = 12$, $N_u = 72$, and $N_f = 6$ according to the variation in $(D_g, R_u)$. As shown in this table, EE and throughput are significantly affected by the variations in $D_g$ and $R_u$. We find that the largest $D_g$ and the smallest $R_u$ give the greatest EE result, e.g., (500, 100). As mentioned before, when $D_g \geq 2R_u$, GCSs might receive relatively lower intercell interference leading to an increase in EE and throughput. Conversely, when $R_u$ increases for the same $D_g$, the intercell interference becomes severe because the number of UTs in a cell boundary or overlapping area increases. Consequently, EE and throughput can become worse.

## V. CONCLUSION

In this paper, we propose a hierarchical multi-agent Q-learning-based optimal frequency reuse and power control algorithm to maximize network-wide EE in uplink multi-UAV wireless networks. First, to mitigate an intercell interference problem, we focused on obtaining the optimal frequency reuse factor with centralized outer-loop RL. Additionally, UTs' transmit power was optimally adjusted by using distributed inner-loop RL. Because the simultaneous adjustment of these factors is very challenging and complicated in practice, it is almost impossible to propose an optimal algorithm working in real-time. Nevertheless, in this paper, we obtained the best EE results with the hierarchical multi-agent Q-learning algorithm compared to the random action algorithms using the fixed frequency reuse factor. To evaluate performance in various network environments, we considered many $(D_g, R_u)$ combinations. Even in the case when the number of UTs in a cell boundary or overlapping area increases, we showed that the proposed algorithm outperformed conventional algorithms and converged. For further work, we will investigate the joint optimization of power consumption at a transceiver and a propulsion system for maximizing network-wide EE in multi-UAV wireless networks.

## ACKNOWLEDGMENT

*Seungmin Lee and Suhyeon Lim contributed equally to this work.*

## REFERENCES

[1] G. Flagship, *Key Drivers and Research Challenges for 6G Ubiquitous Wireless Intelligence*, document 6G Research Visions 1, Sep. 2019, pp. 1–36.
[2] *The Next Hyper-Connected Experience for All*, Samsung, Suwon-si, South Korea, Jul. 2020, pp. 1–46.
[3] 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.
[4] Z. Xiao, H. Dong, L. Bai, D. O. Wu, and X.-G. Xia, "Unmanned aerial vehicle base station (UAV-BS) deployment with millimeter-wave beamforming," *IEEE Internet Things J.*, vol. 7, no. 2, pp. 1336–1349, Feb. 2020.
[5] 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.
[6] S. Lim, S. H. Chae, and H. Lee, "RE-ORA: Residual energy-aware online random access for improving the lifetime of slotted ALOHA-based swarming drone networks," *IEEE Access*, vol. 9, pp. 45504–45511, 2021.
[7] C. H. Liu, Z. Chen, J. Tang, J. Xu, and C. Piao, "Energy-efficient UAV control for effective and fair communication coverage: A deep reinforcement learning approach," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 9, pp. 2059–2070, Sep. 2018.
[8] N. Himayat, S. Talwar, A. Rao, and R. Soni, "Interference management for 4G cellular standards [WIMAX/LTE UPDATE]," *IEEE Commun. Mag.*, vol. 48, no. 8, pp. 86–92, Aug. 2010.
[9] T. Novlan, R. Ganti, A. Ghosh, and J. Andrews, "Analytical evaluation of fractional frequency reuse for OFDMA cellular networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 12, pp. 4294–4305, Dec. 2011.
[10] M. Hossain and Z. Becvar, "Flexible soft frequency reuse for interference management in the networks with flying base stations," in *Proc. IEEE VTC-Spring*, May 2020, pp. 1–7.
[11] Y. Zhang, Z. Mou, F. Gao, L. Xing, J. Jiang, and Z. Han, "Hierarchical deep reinforcement learning for backscattering data collection with multiple UAVs," *IEEE Internet Things J.*, vol. 8, no. 5, pp. 3786–3800, Mar. 2021.

[12] S. Liu, J. Wu, and J. He, "Dynamic multichannel sensing in cognitive radio: Hierarchical reinforcement learning," *IEEE Access*, vol. 9, pp. 25473–25481, 2021.
[13] L. Ren, X. Ning, and J. Li, "Hierarchical reinforcement-learning for real-time scheduling of agile satellites," *IEEE Access*, vol. 8, pp. 220523–220532, 2020.
[14] Y. Zhang, Q. Zhang, B. Cao, and P. Chen, "Model free dynamic sensing order selection for imperfect sensing multichannel cognitive radio networks: A Q-learning approach," in *Proc. IEEE Int. Conf. Commun. Syst.*, Nov. 2014, pp. 364–368.
[15] S. Wang, H. Liu, P. H. Gomes, and B. Krishnamachari, "Deep reinforcement learning for dynamic multichannel access in wireless networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 4, no. 2, pp. 257–265, Jun. 2018.
[16] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal LAP altitude for maximum coverage," *IEEE Wireless Commun. Lett.*, vol. 3, no. 6, pp. 569–572, Dec. 2014.
[17] *Propagation Data and Prediction Methods Required for the Design of Terrestrial Broadband Radio Access Systems Operating in a Frequency Range From 3 to 60 GHz*, document ITU-R, Recommendation P.1410-5, Feb. 2012, pp. 1–14.
[18] P. K. Sharma and D. I. Kim, "Random 3D mobile UAV networks: Mobility modeling and coverage probability," *IEEE Trans. Wireless Commun.*, vol. 18, no. 5, pp. 2527–2538, May 2019.
[19] 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.

**SEUNGMIN LEE** (Student Member, IEEE) received the B.S. degree from the School of Electronic and Electrical Engineering, Hankyong National University, Anseong, South Korea, in 2021. His current research interests include B5G/6G wireless communications, ultra-dense distributed networks, reinforcement learning for UAV networks, unsupervised learning for wireless communication networks, and the Internet of Things. He was a recipient of Best Paper Award at KICS 2020 Winter Conference, in 2020.

**SUHYEON LIM** (Student Member, IEEE) received the B.S. degree from the School of Electronic and Electrical Engineering, Hankyong National University, Anseong, South Korea, in 2021. Her current research interests include B5G/6G wireless communications, ultra-dense distributed networks, reinforcement learning for UAV networks, unsupervised learning for wireless communication networks, and the Internet of Things.

**SEONG HO CHAE** (Member, IEEE) received the B.S. degree from the School of Electronic and Electrical Engineering, Sungkyunkwan University, Suwon, South Korea, in 2010, and the M.S. and Ph.D. degrees from the School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2012 and 2016, respectively. He was a Postdoctoral Research Fellow with KAIST, in 2016, and a Senior Researcher with the Agency for Defense Development, from 2016 to 2018, where he was involved in research and development for military communication system and frequency management software. He is currently an Assistant Professor with Tech University of Korea. His research interests include edge computing and caching, UAV communications, and the Internet of Things (IoT).

**BANG CHUL JUNG** (Senior Member, IEEE) received the B.S. degree in electronics engineering from Ajou University, Suwon, South Korea, in 2002, and the 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 2004 and 2008, respectively. He was a Senior Researcher/a Research Professor with the KAIST Institute for Information Technology Convergence, Daejeon, from 2009 to 2010. From 2010 to 2015, he was a Faculty Member with Gyeongsang National University. He is currently an Associate Professor with the Department of Electronics Engineering, Chungnam National University, Daejeon. His research interests include 5G mobile communication systems, statistical signal processing, opportunistic communications, compressed sensing, interference management, interference alignment, random access, relaying techniques, device-to-device networks, in-network computation, and network coding. He was a recipient of the Fifth IEEE Communication Society Asia-Pacific Outstanding Young Researcher Award, in 2011. He was also a recipient of the Bronze Prize of Intel Student Paper Contest, in 2005, the First Prize of KAIST's Invention Idea Contest, in 2008, the Bronze Prize of Samsung Humantech Paper Contest, in 2009, and the Outstanding Paper Award in Spring Conference of the Korea Institute of Information and Communication Engineering, in 2015. He received the Haedong Young Scholar Award, in 2015, which is sponsored by the Haedong Foundation and given by the Korea Institute of Communications and Information Science.

**CHAN YI PARK** (Member, IEEE) received the B.S. degree in computer engineering from Pusan National University, Pusan, South Korea, in 1998, and the M.S. degree in computer science and engineering from the Pohang University of Science and Technology (POSTECH), Pohang, South Korea, in 2000. He is currently a Principal Researcher for the Agency for Defense Development (ADD), Daejeon, South Korea. His research interests include operational concept and radio resource management for UAV networks.

**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/a 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, ultra-dense distributed networks, in-network computations for 3D images, cross-layer radio resource management, reinforcement learning for UAV networks, unsupervised learning for wireless communication networks, and the Internet of Things.

He was a recipient of the Joint Conference on Communications and Information (JCCI) 2006 Best Paper Award and of 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, and the Best Paper Award at the KICS Winter Conference, in 2020. He received the Minister's Commendation by the Minister of Science and ICT, in 2017.