Deployment and Association of Multiple UAVs in UAV-Assisted Cellular Networks With the Knowledge of Statistical User Position

Leiyu Wang, *Student Member, IEEE*, Haixia Zhang, *Senior Member, IEEE*, Shuaishuai Guo, *Member, IEEE*, and Dongfeng Yuan, *Senior Member, IEEE*

***Abstract*— Exploiting unmanned aerial vehicles (UAVs) as flying relays is becoming an indispensable strategy to assist terrestrial cellular networks to enhance coverage. One challenging problem for UAV-integrated cellular networks is how to design their deployment and association schemes to provide on-demand coverage with minimum network power consumption. In this paper, the uplink transmission in a UAV-assisted cellular network is studied with the objective of minimizing the transmit power consumption of users and UAVs through designing proper UAV deployment and association schemes. To avoid the computational complexity caused by the estimation of instantaneous position of users, we investigate UAV deployment and association schemes based on the statistical user position. By discretizing the space where UAV can be located, we build a centralized multi-agent Q-learning algorithm, with which multiple UAVs update their positions in a joint manner. In the training process of Q-learning algorithm, a reward function is built based on the optimal association scheme and its corresponding power consumption. By adopting the optimal transport theory, the existence of the unique optimal association scheme for given statistical user distribution and UAVs' state is proved. Simulation results demonstrate that the proposed designs considerably outperform the similar existing algorithms. Comparisons with the benchmark scheme show that the proposed scheme can bring about 85% energy efficiency improvement under the same simulation setups.**

***Index Terms*— UAVs communication, power minimization, deployment and association, optimal transport theory, multi-agent Q-Learning.**

Manuscript received 31 January 2021; revised 3 August 2021 and 13 December 2021; accepted 31 January 2022. Date of publication 16 February 2022; date of current version 12 August 2022. This work was supported in part by the Project of International Cooperation and Exchanges, National Natural Science Foundation of China (NSFC), under Grant 61860206005; in part by the NSFC under Grant 62171262; and in part by the Key Research and Development (Major Scientific and Technological Innovation) Project of Shandong Province under Grant 2020CXGC010108. The associate editor coordinating the review of this article and approving it for publication was M. C. Vuran. *(Corresponding author: Haixia Zhang.)*

Leiyu Wang and Dongfeng Yuan are with the Shandong Key Laboratory of Wireless Communication Technologies, Jinan, Shandong 250061, China, and also with the School of Information Science and Engineering, Shandong University, Qingdao, Shandong 266237, China (e-mail: leiyu_wang@mail.sdu.edu.cn; dfyuan@sdu.edu.cn).

Haixia Zhang and Shuaishuai Guo are with the Shandong Key Laboratory of Wireless Communication Technologies, Jinan, Shandong 250061, China, and also with the School of Control Science and Engineering, Shandong University, Jinan, Shandong, 250061, China (e-mail: haixia.zhang@sdu.edu.cn; shuaishuai_guo@sdu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2022.3150429.

Digital Object Identifier 10.1109/TWC.2022.3150429

I. Introduction

UNMANNED aerial vehicle (UAV)-assisted cellular networks, where multiple UAVs are integrated as relays to help the terrestrial base stations improve ground users' quality of service (QoS), have attracted increasing attention in recent years [1]–[4]. The integration of UAVs into cellular networks can bring extra spatial degrees of freedom, which has been shown to be an effective way to against unfavorable wireless environments. In UAV-assisted cellular networks, one crucial issue that must be addressed with top priority is to reduce power consumption, especially that of the UAVs and ground users. This is because all UAVs and ground users are typically powered by capacity-, size- and weight-limited batteries. Reducing their power consumption can greatly prolong the working life of the integrated networks and improve users' quality of experience. In this paper, we focus on minimizing the power consumption for uplink data transmission, including user-UAV transmission and UAV base station transmission. The power consumption of these two parts depends highly on the deployment of UAVs. Placing UAVs far from base stations will inevitably increase UAVs' transmit power, while placing UAVs far from ground users will increase the users' transmit power. Apart from that, the power consumption of the two parties is also highly related to the association schemes. Specifically, inappropriate association in the integrated network will impose unbalanced loads for UAVs. Those UAVs with heavy loads will consume much more power and their working time will be greatly shortened. From the user side, if the association area is too large, the edge users will have to spend more power to transmit data to their associated UAVs. Inspirited by all of those, we pay our attention to developing an energy-efficient UAVs' deployment and association schemes with the objective of minimizing the transmit power consumption of the UAVs and ground users in this work.

A. Prior Works

Researchers have done a lot of works in developing energy-efficient deployment and association schemes for UAV-integrated communication networks. Based on the assumption on user position information, prior works on this topic can be classified into two categories:

instantaneous-user-position-based schemes and statistical-user-position-based schemes.

1) Instantaneous-User-Position-Based Deployment and Association: Instantaneous-user-position-based schemes assume that the locations of the ground users are known. With such prior information, lots of deployment and association schemes of UAVs were widely investigated. For instance, from the point of view of reducing the power consumption of UAVs, [5] and [6] developed energy-efficient placement optimization algorithms to determine the optimal hover position for a single UAV based on the deterministic position of users. For multiple UAVs cases, Chen *et al.* [7] formulated and solved UAVs' deployment and association optimization problem to guarantee the requirement of each user while minimizing the transmit power consumption of all UAVs. Wang *et al.* [8] came up with a deep learning framework for dynamically optimizing the location and association of UAVs to reduce their transmit power consumption. Meanwhile, extensive research has also shown that the movement of UAVs will cost huge propulsion energy. All these works mentioned above have overlooked the power consumed when UAVs fly to the desired positions.

Taking the moving process into account, there was a body of literature developing various energy-efficient trajectory optimization approaches [9]–[17]. Specifically, in [9], UAV was dispatched to serve cell edge users and offload data from base station. In order to reduce the required propulsion-related power consumption, an energy-efficient dynamic placement and user partition strategy was designed. Later on, the dynamic deployment for the propulsion energy-constrained UAVs in the data uploading process was investigated in [10]. Besides, Zeng *et al.* proposed a mathematical model for propulsion power consumption of fixed-wing UAVs [13] and rotary-wing UAVs [15], respectively. Then, they explored energy-efficient dynamic trajectory design accordingly. For the purpose of balancing the performance of ground users and the power consumption of UAVs, Wu *et al.* discussed several trade-offs in trajectory design for UAVs [17], [18]. Recently, UAV deployment problem has been regarded as a dynamic game and solved in reinforcement learning framework [19]–[25]. For instance, by using distributed multi-agent reinforcement learning algorithm, the transmit power of UAVs was optimized implicitly through determining the deployment positions of UAVs to improve the quality of experience of ground users [25]. The above works only paid their main attention to UAVs' power consumption minimization. In UAV-assisted cellular networks, the power consumption of ground users is also highly related to UAVs' deployment and association. Taking that into consideration, [26] and [27] investigated dynamic UAV deployment schemes to minimize the maximum transmit power of the ground devices in Internet-of-Things (IoT) networks.

2) Statistical-User-Position-Based Deployment and Association: Designing deployment and association schemes based on statistical user position information can be seen as a more practical and efficient method in reality, compared with those works based on instantaneous user position information. The reasons are as follows. In practical systems, the ground users normally have high mobility, which forces UAVs to track users' positions constantly and adjust their deployment positions and association policies frequently. This will considerably increase their power consumption. Differently, designing policies based on statistical user position information will enable UAVs to hold stable deployment and association schemes for a long period. In such designs, UAVs only need to know the long-term position state information instead of instantaneous position information, thus the overhead for real-time user position estimation and feedback can be saved. Moreover, practical network data analysis indicates that the statistical user position follows a periodic pattern in cellular networks [28]. This characteristic makes deployment and association schemes based on statistical user position information more valuable. Realizing this fact, an initial attempt for designing energy-efficient deployment and association for UAV networks has been done in [29], where Mozaffari *et al.* optimized the hover locations of multiple UAVs as well as the association scheme to minimize the transmit power of UAVs. Later on, by predicting the number and statistical position of users, Zhang *et al.* [30] designed UAVs' deployment and association schemes to minimize consumed transmit power of UAVs. It is noteworthy that all of these works assumed the drones operated independently as flying base stations. In UAV-assisted cellular networks, where UAVs work as relays, the results in [29] and [30] can not be directly adopted. This is because the backhaul communications between flying UAVs and base stations were not considered. As far as we know, designing the energy-efficient deployment and association scheme for UAV-assisted cellular networks based on statistical user position information is still an open problem.

B. Our Work and Contributions

As discussed above, the power consumption of UAVs and ground users is one of the key concerns in UAV-assisted cellular networks, which depends highly on UAVs' deployment and association schemes. Most of the prior works developed corresponding schemes based on the assumption that instantaneous user positions. Only a few works are done based on statistical user position. These works assume that UAV acts as the base station, and try to minimize the power consumption of UAVs. Against this background, we take both ground users-UAVs and UAVs-base station transmissions into account, and try to minimize the transmit power consumption of UAV-assisted cellular networks through UAV deployment and association scheme design in this work. The main contributions of our work can be summarized as follows:

- Based on the statistical user position information, the transmit power consumption minimization problem for both the ground users and UAVs is formulated. How to minimize power consumption through deployment and association design is mathematically described.
- To solve the formulated problem, we discretize the space where UAVs can be located. After that, we transform the UAVs deployment and association as a Markov decision process and use a centralized multi-agent Q-learning algorithm to determine the positions of UAVs.

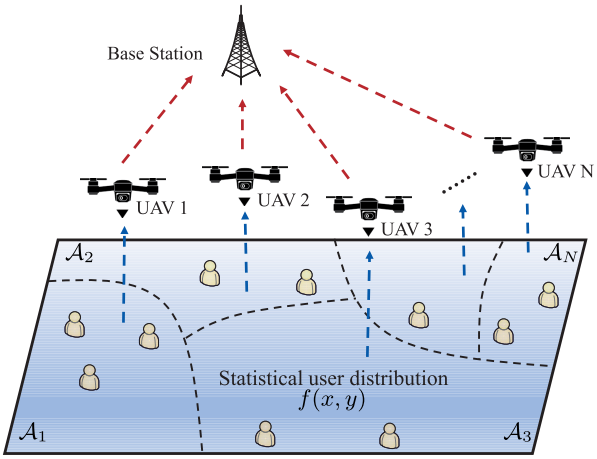


Fig. 1. UAV-assisted cellular network system model.

- In the training process of the Q-learning, the reward for a state with fixed UAV positions is computed by developing the optimal association scheme and then computing its corresponding power consumption. Specifically, we introduce the optimal transport theory to find out the optimal association. The uniqueness of the optimal association is proved and an iterative algorithm is adopted to approach the optimal association.
- It has been shown that the deployment and association scheme obtained based on the proposed algorithms can match any given statistical user position and minimize the transmit power consumption accordingly. The convergence and complexity of the proposed algorithms are analyzed. Simulations are done and results show the superiority of the proposed scheme.

C. Paper Organization

The remainder of the paper is organized as follows. Section II gives a detailed description of the model for UAV-assisted cellular networks. The power minimization problem is also formulated in this section. In Section III, the deployment and association algorithms for the integrated networks are designed elaborately. Besides, the convergence and complexity of the proposed algorithms are provided. Section IV presents simulation results and analysis. Finally, Section V concludes the whole paper.

II. System Model and Problem Statement

In this work, we consider the uplink transmission in a UAV-assisted cellular network as illustrated in Fig. 1. In this network, due to the malfunction or overload for the nearby base stations, a group of N UAVs denoted by $\mathcal{N} = \{1, 2, \ldots, N\}$ are adopted to act as relays to help users deliver their data to a remote base station. They serve M ground users following a given statistical probability density distribution $f(x, y)$ in a target area $\mathcal{A}$. Let $l_i = (x_i, y_i, h)$ be the coordinates of UAV i, where (x_i, y_i) are the horizontal plane coordinates of UAV i, and h is the altitude of UAVs.[1]

[1]We assume that all UAVs fly at the same height in this paper. Therefore, only the horizontal deployment is investigated in this work.

The base station is placed at point (x_0, y_0, h_b), which is far from all users. In order to fully cover all the ground users, the target area $\mathcal{A}$ is divided into N non-overlapping districts, i.e., $\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2 \cup \cdots \cup \mathcal{A}_N$ and $\mathcal{A}_i \cap \mathcal{A}_j = \emptyset$, $i \neq j, i, j \in \mathcal{N}$. All through this work, we assume that $\mathcal{A}_i$ and UAV i have a one-to-one mapping relationship. It means that users in $\mathcal{A}_i$ can only associate with and get service from UAV i, and in the same time UAV i only serves the users within area $\mathcal{A}_i$. We define the set $\Omega = \{\mathcal{A}_1, \mathcal{A}_2, \cdots, \mathcal{A}_N\}$ and let it represent an association scheme for all UAVs. Besides, the directional antenna technology and the orthogonal frequency allocation are used to reduce the interference among users and UAVs as much as possible [31].

A. Path Loss Model

We assume that the system works in the urban environment. Therefore, the air-to-ground signal propagation is affected by the obstacles and buildings. To model the channel fading between an air platform and a terrestrial terminal in urban environment (e.g., suburban, urban), the probabilistic path loss model is adopted [32], [33], the average path loss between a ground user at point (x, y) and UAV i can be denoted as

$$\Lambda_i(x,y) = P_i^{\mathrm{LoS}} \underbrace{K_o d_i^2(x,y)\mu_{\mathrm{LoS}}}_{\text{LoS link path loss}} + P_i^{\mathrm{NLoS}} \underbrace{K_o d_i^2(x,y)\mu_{\mathrm{NLoS}}}_{\text{NLoS link path loss}}, \quad (1)$$

where $d_i(x,y) = \sqrt{(x-x_i)^2 + (y-y_i)^2 + h^2}$ denotes the distance between UAV i at (x_i, y_i, h) and the ground user at (x, y), $K_o = \left(\frac{4\pi f_c}{c}\right)^2$ and f_c is carrier frequency, c denotes the speed of light, μ_{LoS} and μ_{NLoS} are the additional path loss for LoS link and NLoS link that incurred by the shadowing effect, P_i^{LoS} and P_i^{NLoS} are occurrence probability that ground user transmits data to UAV i in a line-of-sight (LoS) manner or non-line-of-sight (NLoS) manner, respectively. P_i^{LoS} and P_i^{NLoS} are calculated by

$$P_i^{\mathrm{LoS}} = \frac{1}{1 + b_1 \exp\left(-b_2\left(\frac{180}{\pi}\theta_i - b_1\right)\right)}, \quad (2)$$

and

$$P_i^{\mathrm{NLoS}} = 1 - P_i^{\mathrm{LoS}}, \quad (3)$$

where b_1 and b_2 are constants related to given wireless propagation environment, θ_i is the elevation angle, which can be calculated by $\theta_i = \arcsin\left(\frac{h}{d_i(x,y)}\right)$.

For the UAV-base station transmission part, due to the fact that the base station is typically deployed in open area, UAV-base station transmission can effectively avoid the obstruction caused by buildings or trees. Therefore, we assume the channel gain between UAV i and the base station at (x_0, y_0, h_b) follows the free space path loss [34], i.e.,

$$\Lambda_{i0} = K_o d_{i0}^2, \quad (4)$$

where $d_{i0} = \sqrt{(x_0 - x_i)^2 + (y_0 - y_i)^2 + (h - h_b)^2}$ is the distance between UAV i and the base station.

B. Problem Formulation

As our objective is to minimize the transmit power of UAV-assisted cellular network through UAVs' deployment location and association optimizing based on the knowledge of statistical user position. To ensure that the data of all users can reach the base station, the required total transmit power is decomposed into two parts, i.e., the transmit power for the ground user to transmit its data to the associated UAV (blue dashed line) and the transmit power for the associated UAV to relay the ground user's data to the cellular base station (red dashed line), as shown in Fig. 1. To proceed, how deployment and association affect the power consumption of the ground users and UAVs should be described firstly.

To model the transmit power consumption from ground users to their associated UAVs, we assume that the minimum transmission rate of ground users should be guaranteed, i.e., the received signal-to-noise ratio (SNR) should be greater than the given threshold β. With such an assumption, the required transmit power of a user in $\mathcal{A}_i$ that is associated with UAV i can be expressed as

$$\begin{aligned} p_u^i(x,y,l_i) &= \beta\sigma_u^2\Lambda_i(x,y) \\ &= \beta K_o d_i^2(x,y)\sigma_u^2[P_i^{\mathrm{LoS}}\mu_{\mathrm{LoS}} + P_i^{\mathrm{NLoS}}\mu_{\mathrm{NLoS}}], \end{aligned} \tag{5}$$

where the user is assumed to be located at (x, y) and σ_u^2 is the noise power which can be denoted by κB_u, with κ is the noise power spectral density, B_u denotes the bandwidth of user u. From (5), it can be observed that $p_u^i(x, y, l_i)$ is mainly determined by $d_i^2(x, y)$. Recall that $d_i(x, y) = \sqrt{(x - x_i)^2 + (y - y_i)^2 + h^2}$, it can be concluded that the required transmit power for ground user depends on which UAV it is associated to and where the UAV is. Therefore, the transmit power of the ground users will be optimized through designing UAVs' deployment and association in the following section.

For UAV-base station transmission, UAV i at location l_i will collect all the ground users' data in cell $\mathcal{A}_i$ and forward them to the remote base station at (x_0, y_0, h_b). Let $a_i = \int_{\mathcal{A}_i} f(x, y)dxdy$ represent the user ratio in area $\mathcal{A}_i$ and $r_u = B_u \log_2(1 + \beta)$ stand for the data rate of user u in $\mathcal{A}_i$, the required relaying transmission rate of UAV i writes

$$R(a_i) = Ma_i r_u. \tag{6}$$

To obtain such a rate with available bandwidth B_v, the required SNR can be given by

$$\beta_v = 2^{R(a_i)/B_v} - 1. \tag{7}$$

Taking the path loss into consideration, the required transmit power for UAV i should be

$$p_i^0(a_i, l_i) = \beta_v \sigma_v^2 \Lambda_{i0}. \tag{8}$$

where σ_v^2 represents the noise power of the link from UAV i to the base station and can be denoted by κB_v. Thus, the transmit power for UAV i to deliver one user's data to the base station can be averagely computed as

$$p_{i,u}^0(a_i, l_i) = \frac{p_i^0(a_i, l_i)}{Ma_i}. \tag{9}$$

Substituting (4) and (6)-(8) into (9) yields

$$p_{i,u}^0(a_i, l_i) = \frac{\left(2^{Ma_i r_u/B_v} - 1\right)\sigma_v^2 K_o d_{i0}^2}{Ma_i}. \tag{10}$$

From (10), it is observed that the required transmit power for each UAV to relay data of one user depends not only on the UAV's position l_i (related to given deployment scheme) but also the user ratio a_i (related to association scheme). It can also be seen that the transmit power consumed by a UAV will increase exponentially with the data rate requirement $R(a_i) = Ma_i r_u$. This indicates that unbalanced loads for UAVs will lead to even bigger unbalance in power consumption. Those UAVs with heavy transmission loads will consume much more power, and their working time will be greatly shortened accordingly.

By now, we have obtained the required transmit power of a ground user (5) and that of its associated UAV (10) for the uplink data transmission. Aggregating all the transmit power in the UAV-assisted cellular network together yields

$$P = \sum_{i=1}^{N} \int_{\mathcal{A}_i} \left[p_u^i(x, y, l_i) + p_{i,u}^0(a_i, l_i)\right] f(x, y)dxdy, \tag{11}$$

which is obtained by taking the integral over the distribution on all areas $\mathcal{A}_i, \forall i \in \mathcal{N}$. Based on the above modeling, the optimization problem finding UAVs' deployment location and association to minimize the transmit power in the UAV-assisted cellular network can be formulated as

$$\begin{aligned} &\mathcal{P}1: \text{Given}: \ \beta, N, M, f(x,y), \mathcal{A}, b_1, b_2, K_o, B_u, B_v, \sigma_u^2, \sigma_v^2 \\ &\text{Find}: \ \Omega, l_i, \forall i \in \mathcal{N} \\ &\text{Minimize}: \ P \\ &\text{Subject to}: \ \mathcal{A}_i \cap \mathcal{A}_j = \emptyset, \quad \forall\, i \neq j \in \mathcal{N}, && \text{(12a)} \\ &\qquad \bigcup_{i \in \mathcal{N}} \mathcal{A}_i = \mathcal{A}, && \text{(12b)} \\ &\qquad a_i = \int_{\mathcal{A}_i} f(x, y)dxdy \in [0, 1], && \text{(12c)} \\ &\qquad \sum_{i=1}^{N} a_i = 1. && \text{(12d)} \end{aligned}$$

where (12a) represents that multiple UAVs cover mutually non-intersect areas, (12b) represents that all N UAVs cover all areas, (12c) denotes the user ration contained in an area to be in-between 0 and 1, and (12d) means all users to locate in the given area $\mathcal{A}$.

III. Multi-Agent Q-Learning Based UAV Deployment and OTT-Based Association Algorithms

In Subsection II-B, we jointly consider the transmit power of UAVs and ground users and formulate the power consumption minimization problem $\mathcal{P}1$ based on the given statistical user position information. To solve the formulated problem, UAV deployment and association schemes should be elaborately designed. However, it can be seen that obtaining the desired deployment and association schemes for multiple UAVs is a nontrivial problem. The reasons are as follows. On the one

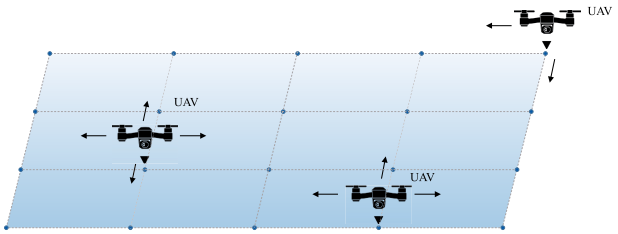


Fig. 2. Illustration on the set of available state and action of UAVs.

hand, the deployment and association variables are coupled with each other. On the other hand, UAVs are deployed into a continuous space which leads to innumerable deployment solutions. Apparently, exhaustive search for solutions is prohibitive. As far as we know, there is no standard solution for solving the formulated problem. To solve it, we first discretize the space where UAVs can be located, as illustrated in Fig. 2. After that, we adopt a centralized multi-agent Q-learning algorithm, in which multiple UAVs will be regarded as the agents to determine their deployment positions. Meanwhile, in the training process of the multi-agent, we propose an association algorithm based on the optimal transport theory to determine the reward at each training state. The details of the proposed algorithms are shown in Subsections III-A and III-B.

A. Multi-Agent Q-Learning Based Multiple UAVs Deployment

As we discussed above, jointly deploying multiple UAVs is challenging. After discretization as shown in Fig. 2, the joint deployment can be done following the steps below. First, we initiate the UAV position randomly. Then, we let the UAVs move following the rewards obtained from the multi-agent Q-learning algorithm. To ensure the optimality of the deployment, a centralized training strategy is adopted. After several large-reward movements, the algorithm will finally converge to the desired positions and the UAV deployment is obtained. It is noteworthy that the learning process is totally off-line and does not need the real movement of UAVs. Nothing but the position obtained after convergence matters for a given statistical user position. The result can be applied on-line once the training process is ended, i.e., the controller will inform the base station to adjust UAVs' deployment policy and following which all the UAVs fly to their newly generated optimal deployment positions. The state set, action set and reward function for multi-agent Q-learning algorithm can be described as follows.

- **State set**: The state space of individual UAV i is defined as $\mathbf{S}_i = \{s_i^1, \cdots, s_i^m\}$, where m is the total number of discrete states in $\mathbf{S}_i$, $s_i^1 = (x_i^1, y_i^1)$, $s_i^m = (x_i^m, y_i^m)$. The state space of each UAV is not overlapped and the joint state space $\mathbf{S} = \mathbf{S}_1 \times \mathbf{S}_2 \times \cdots \mathbf{S}_N$. For each episode, UAV i will stay in one state, i.e., $s_i \in \{s_i^1, \cdots, s_i^m\}$, thus, the joint state at episode k for all UAVs can be written as $\mathbf{s}_k = \{s_1, \cdots, s_N\}$.
- **Action set**: The action set $\mathbf{U}_i$ of UAV i consists of five movement directions as shown in Fig. 2, including moving forward, moving backward, turning left, turning right and keeping still. UAV i will select one action u_i at each episode from set $\{u_i^1, \cdots, u_i^5\}$. The joint action at episode k for all UAVs forms a vector as $\mathbf{u}_k = \{u_1, \cdots, u_N\}$.
- **Reward function**: Since the target of our design is to find out the optimal joint deployment, all UAVs share the same reward function. We introduce $r(\mathbf{s}_k, \mathbf{u}_k)$ to represent the reward of UAVs at state $\mathbf{s}_k$ while taking joint action $\mathbf{u}_k$ at the kth episode. The reward $r(\mathbf{s}_k, \mathbf{u}_k)$ is to evaluate the quality of the transition from current state $\mathbf{s}_k$ to the next state $\mathbf{s}_{k+1}$ accurately. It is a function of the required minimum transmit power consumption at state $\mathbf{s}_{k+1}$. Specifically, the less power consumption, the more reward will be got. In our work, we simply employ the inverse of the power consumption as the reward,[2] which is given by

$$r(\mathbf{s}_k, \mathbf{u}_k) = -P(\mathbf{s}_{k+1}, \Omega_{k+1}^*), \tag{13}$$

  where $P(\mathbf{s}_{k+1}, \Omega_{k+1}^*)$ represents the transmit power consumption at state $\mathbf{s}_{k+1}$ with the optimal association scheme Ω_{k+1}^*. Since the state $\mathbf{s}_{k+1}$ is known by taking action $\mathbf{u}_k$ at state $\mathbf{s}_k$, the minimization of the transmit power consumption only depends on Ω_{k+1}^*. Finding the optimal association scheme Ω_{k+1}^* at state $\mathbf{s}_{k+1}$ becomes the key issue.

Fortunately, since all agents (i.e., UAVs) are coordinated and controlled by the base station, the multi-agent system can be seen as a single-agent system which can be described by the Markov decision process, in which the next state of all UAVs only depends on the current state and all UAVs' action. It can be solved through learning the joint action using Q-learning [25], [35]. The Q-value in the Q-table can be computed by

$$\begin{aligned} Q_{k+1}(\mathbf{s}_k, \boldsymbol{u}_k) = Q_k(\mathbf{s}_k, \boldsymbol{u}_k) + \alpha[r(\mathbf{s}_k, \mathbf{u}_k) \\ +\gamma \max_{\boldsymbol{u}'} Q_k(\mathbf{s}_{k+1}, \boldsymbol{u}') - Q_k(\mathbf{s}_k, \boldsymbol{u}_k)], \end{aligned} \tag{14}$$

where $\gamma \in [0, 1]$ is the discount factor representing the weight of future rewards, $Q_k(\mathbf{s}_k, \boldsymbol{u}_k)$ is defined as the accumulated discounted rewards when UAVs take joint actions $\mathbf{u}_k$ at joint state $\mathbf{s}_k$, and α is the learning rate.

B. OTT-Based Association

As defined in (13), the reward $r(\mathbf{s}_k, \mathbf{u}_k)$ at state $\mathbf{s}_k$ is computed according to the transmit power consumption of state $\mathbf{s}_{k+1}$. Finding the optimal association scheme Ω_{k+1}^* that minimizes the transmit power consumption at state $\mathbf{s}_{k+1}$ can be formulated as an optimization problem as

$$\Omega_{k+1}^* = \arg\min_{\Omega_{k+1}} P(\mathbf{s}_{k+1}, \Omega_{k+1}), \tag{15}$$

where $\Omega_{k+1} = \{\tilde{\mathcal{A}}_1, \tilde{\mathcal{A}}_2, \cdots, \tilde{\mathcal{A}}_N\}$ represents a legitimate association scheme at state $\mathbf{s}_{k+1}$ satisfying $\tilde{\mathcal{A}}_i \cap \tilde{\mathcal{A}}_j = \emptyset$ and $\bigcup_{i \in \mathcal{N}} \tilde{\mathcal{A}}_i = \mathcal{A}$, $\Omega_{k+1}^* = \{\tilde{\mathcal{A}}_1^*, \tilde{\mathcal{A}}_2^*, \cdots, \tilde{\mathcal{A}}_N^*\}$ represents the

[2]In this paper, it is assumed that the reward can be negative, which can be explained as a penalty on the power consumption.

optimal association scheme at state $\mathbf{s}_{k+1}$. In the optimization problem, the objective function $P(\mathbf{s}_{k+1}, \Omega_{k+1})$ can be extended similarly as that in (11) by

$$P(\mathbf{s}_{k+1}, \Omega_{k+1}) = \sum_{i=1}^{N} \int_{\mathcal{A}_i} [p_u^i(x, y, l_i) + p_{i,u}^0(a_i, l_i)] f(x, y) dx dy. \quad (16)$$

With the optimal transport theory, the unique optimal solution of (15) can be guarantteed. According to the existing works [36], [37], the form of the unique optimal solution is written as

Theorem 1: Given a UAV deployment state $\mathbf{s}_{k+1}$, the unique optimal association cell boundary $\tilde{\mathcal{A}}_i^*$ for UAV i and its covered area a_i have the following relationship

$$\tilde{\mathcal{A}}_i^* = \Bigg\{ (x, y) : p_u^i(x, y, l_i) + p_{i,u}^0(a_i, l_i) + a_i \frac{\partial p_{i,u}^0(a_i, l_i)}{\partial a_i} \leq p_u^j(x, y, l_j) + p_{j,u}^0(a_j, l_j) + a_j \frac{\partial p_{j,u}^0(a_j, l_j)}{\partial a_j}, \forall j \neq i \in \mathcal{N} \Bigg\}, \quad (17)$$

where

$$a_i = \int_{\tilde{\mathcal{A}_i}^*} f(x, y) dx dy. \quad (18)$$

In (17), $p_u^i(x, y, l_i)$ represents the transmit power consumed by the user at point (x, y) if it is associated to UAV i, $p_{i,u}^0(a_i, l_i)$ represents the transmit power consumed by UAV i when relaying the data of a user to the base station, $\frac{\partial p_{i,u}^0(a_i, l_i)}{\partial a_i}$ stands for the partial derivative of $p_{i,u}^0(a_i, l_i)$ with respect to a_i, $p_u^j(x, y, l_j)$ represents the transmit power consumed by the user at point (x, y) if it is associated to UAV j, $p_{j,u}^0(a_j, l_j)$ represents the transmit power consumed by UAV j when relaying the data of a user to the base station, and $\frac{\partial p_{j,u}^0(a_j, l_j)}{\partial a_j}$ stands for the partial derivative of $p_{j,u}^0(a_j, l_j)$ with respect to a_i.

Proof: See Appendix A. ∎

Theorem 1 gives the cell boundary information of any UAV. But, the solution is not in a closed form. In the solution, $\tilde{\mathcal{A}_i}^*$ and a_i are coupled with each other. Based on **Theorem 1**, we adopt an iterative algorithm to approach the optimal association.

1) Initialization: First, we initialize the iteration indicator $t = 1$. We set a random initial association $\Omega_{k+1}^{(t)} = \{\tilde{\mathcal{A}}_1^{(t)}, \tilde{\mathcal{A}}_2^{(t)}, \cdots, \tilde{\mathcal{A}}_N^{(t)}\}$ and calculate its transmit power consumption $P(\mathbf{s}_{k+1}, \Omega_{k+1}^{(t)})$. At the beginning, we define a new variable $\phi_i^{(t)}(x, y), \forall i \in \mathcal{N}$, where $\phi_i^{(t)}(x, y) = 0$ represents that users at point (x, y) will associate with UAV i, and $\phi_i^{(t)}(x, y) = 1$ represents that users at point (x, y) will not associate with UAV i. To start, we make the initialization that $\phi_i^{(t)}(x, y) = 0, \forall i \in \mathcal{N}$ at $t = 1$.

2) Step 1: As $\phi_i^{(t)}(x, y)$ indicates the relationship between users at point (x, y) and UAV i, we can determine the user association area $\tilde{\mathcal{A}}_i^*$ by finding the optimal value of $\phi_i^{(t)}(x, y)$. To this end, we update each $\phi_i(x, y)$ in an iterative manner. In detail, based on an initialization association, the value of $\phi_i^{(t+1)}(x, y)$ can be calculated, i.e., if point (x, y) is located in $\tilde{\mathcal{A}}_i^{(t)}$, then $\phi_i^{(t+1)}(x, y)$ will be set to be $\rho\phi_i^{(t)}(x, y)$, if point (x, y) is not located in $\tilde{\mathcal{A}}_i^{(t)}$, then $\phi_i^{(t+1)}(x, y)$ will be set to be $1 - \rho(1 - \phi_i^{(t)}(x, y))$, where $\rho = 1 - 1/t$ is a prudence parameter defined in [38] expressing the "resistance" that the user feels against changing his association.

Algorithm 1 Iterative Algorithm for Near Optimal Association Scheme

Input: $\mathbf{s}_{k+1}, N, M, f(x, y), \mathcal{A}, b_1, b_2, K_o, B_u, B_v, \sigma_u^2, \sigma_v^2$ and the halting criterion τ

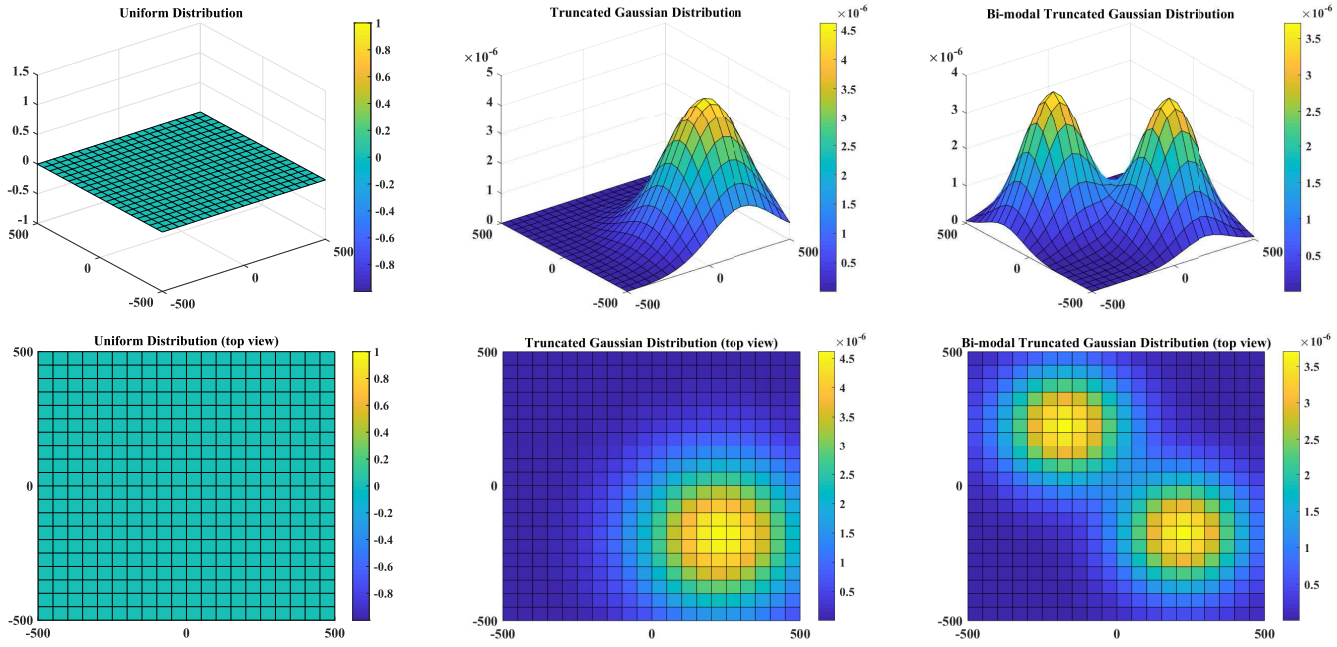
Output: Ω_{k+1}^*.

1: Initialize a random association $\Omega_{k+1}^{(t)}$, compute the initial transmit power $P(\mathbf{s}_{k+1}, \Omega_{k+1}^{(t)})$ using (16), and set $\phi_i^{(t)}(x, y) = 0, \forall i \in \mathcal{N}$.
2: **repeat**
3: $\rho = 1 - 1/t$.
4: **if** $(x, y) \in \tilde{\mathcal{A}}_i^{(t)}$ **then**
5: $\phi_i^{(t+1)}(x, y) = \rho\phi_i^{(t)}(x, y)$.
6: **else**
7: $\phi_i^{(t+1)}(x, y) = 1 - \rho\left(1 - \phi_i^{(t)}(x, y)\right)$.
8: **end if**
9: Calculate $a_i = \int_{\mathcal{A}} \left(1 - \phi_i^{(t+1)}(x, y)\right) f(x, y) \mathrm{d}x\mathrm{d}y$.
10: Update $t = t + 1$.
11: Update the $\tilde{\mathcal{A}}_i^{(t)}, \forall i \in \mathcal{N}$ through (17).
12: Calculate the transmit power $P_{tot}(t)$ using (16).
13: **until** $|P(\mathbf{s}_{k+1}, \Omega_{k+1}^{(t)}) - P(\mathbf{s}_{k+1}, \Omega_{k+1}^{(t-1)})| < \tau$
14: $\tilde{\mathcal{A}}_i^* = \tilde{\mathcal{A}}_i^{(t)}, \forall i \in \mathcal{N}$.
15: $\Omega_{k+1}^* = \{\tilde{\mathcal{A}}_1^*, \tilde{\mathcal{A}}_2^*, \cdots, \tilde{\mathcal{A}}_N^*\}$.

3) Step 2: According to $\phi_i^{(t+1)}(x, y)$, we can calculate the average user ratio a_i for UAV i and determine the association scheme at $t+1$, as shown in line 9 - line 11. With the increase of the iteration number, variable $\phi_i^{(t+1)}(x, y)$ will approach to its optimal value, then the corresponding association scheme can also be determined. The convergence is achieved when $|P(\mathbf{s}_{k+1}, \Omega_{k+1}^{(t)}) - P(\mathbf{s}_{k+1}, \Omega_{k+1}^{(t-1)})| < \tau$ is satisfied, where τ is the halting criterion. The pseudo code of the iteration process can be seen in **Algorithm 1**. Through **Algorithm 1**, we can acquire a near optimal solution of the association scheme Ω_{k+1}^* at state $\mathbf{s}_{k+1}$. The proof of the convergence of **Algorithm 1** can be found in [38].

C. Joint Deployment and Association for Multiple UAVs

When $\mathbf{s}_{k+1}$ and Ω_{k+1}^* are determined, the reward can also be calculated through (13). According to the reward, the Q-value in the multi-agent Q-learning algorithm can be further updated by (14). The detail of the proposed multi-agent Q-learning algorithm is given in **Algorithm 2**. To be specific, the Q-value in the Q-table is initialized to be zero and the deployment is initialized randomly. Then, all UAVs will select their action according to the ε-greedy policy [25], [35], where ε represents the probability to select random action. **Algorithm 1** is adopted to calculate the reward, based on which Q table can be updated. To accelerate the convergence



(a) Scenario 1 (b) Scenario 2 (c) Scenario 3

Fig. 3. The distribution model of the ground users in given area. Scenario 1: Uniform Distribution (left); Scenario 2: Truncated Gaussian Distribution (middle); Scenario 3: Bi-modal Truncated Gaussian Distribution (right).

Algorithm 2 Centralized Multi-Agent Q-Learning Algorithm for UAVs Deployment and Association

1: Set $k = 0$ and $Q(\mathbf{s}_0, \boldsymbol{u}_0) = 0$ and initialize a random deployment.
2: **while** $k < K$ **do**
3: Select joint action $\mathbf{u}_k$ according to the ε-greedy policy at joint state $\mathbf{s}_k$.
4: **if rand** $< \varepsilon$ **then**
5: Randomly choose a joint action $\mathbf{u}_k$ to next state $\mathbf{s}_{k+1}$.
6: **else**
7: Choose the joint action $\mathbf{u}_k$ according to $Q_k(\mathbf{s}_k, \mathbf{u_k})$ to the next state $\mathbf{s}_{k+1}$.
8: **end if**
9: Compute Ω^*_{k+1} according to **Algorithm 1** at state $\mathbf{s}_{k+1}$.
10: Calculate the reward according to (13).
11: Update $Q_{k+1}(\mathbf{s}_k, \boldsymbol{u}_k)$ according to (14).
12: **end while**

speed of the learning process, the learning rate is set to be decreasing with time, which is given by [34], [39]

$$\alpha = \frac{1}{(t + c_\alpha)^{\varphi_\alpha}}, \tag{19}$$

where $c_\alpha > 0$ and $\frac{1}{2} < \varphi_\alpha \leq 1$. The proof of the convergence of the Q-learning and the optimality for the convergence point can be found in [40]. The convergence conditions are

- The state and action spaces of system are finite.
- $\sum_t \alpha^{(t)} = \infty$ and $\sum_t \left(\alpha^{(t)}\right)^2 < \infty$.
- The variance of $r(\mathbf{s}_k, \mathbf{u}_k)$ is bounded.

It is obviously that the first and the third conditions can be satisfied. To make the second condition satisfied, we set the learning rate in this work to be (19) as defined in [34]. So, we conclude the Q-learning algorithm is convergent. Once the deployment policy is acquired, UAVs can fly to their optimal discrete positions according to this policy accurately. It should be pointed out that, the output of Q-learning is not the optimal solution to the original problem since we discretize the original space in which UAVs can stay. Apparently, the fine-grained discretization scheme can make UAVs deploy closer to the optimal positions.

The complexity of the proposed multi-agent Q-learning algorithm is composed of space complexity and time complexity. In the multi-agent Q-learning algorithm, assuming that the individual state space $|\mathbf{S}_1| = \cdots = |\mathbf{S}_N| = S$, the individual action space $|\mathbf{U}_1| = \cdots = |\mathbf{U}_2| = U$, then the total update space in Q-table is $\mathcal{O}(S^N U^N)$. In terms of computational time, the complexity for Q-learning is $\mathcal{O}(K)$, where K is the total number of episodes in the training process [41]. In each episode, **Algorithm 1** is performed to compute the reward. The above process can be implemented through the Monte Carlo method which has a complexity $\mathcal{O}(TL)$, where T represents the number of iterations required for **Algorithm 1** to converge and L is the number of discrete sampling point in Monte Carlo simulation. Thus, the total complexity of **Algorithm 2** is $\mathcal{O}(TLK)$.

IV. Simulation Results and Analysis

In order to show the superiority of the proposed deployment and association scheme, intensive simulations are done in this section.

TABLE I
SIMULATION PARAMETERS

Parameters	Description	Value
f_c	Carrier frequency for users	2GHz
f_b	Carrier frequency for UAVs	2GHz
κ	The noise power spectral density	-174 dBm/Hz
B_u	Bandwidth for each user	128 kHz
B_v	Bandwidth for each UAV	[15,20] MHz
h	The altitude of UAVs	210 m
h_b	The altitude of base station	10 m
b_1	The LoS probability constant	9.61 [32]
b_2	The LoS probability constant	0.16 [32]
c	The light speed	3×10^8 m/s
β	The SNR threshold for users	20 dB
μ_{LoS}	The additional path loss for LoS	3 dB
μ_{NLoS}	The additional path loss for NLoS	23 dB
c_α	Learning rate parameter	0.5
γ	Discount factor	[0.5, 0.8]
K	The training episode in Q-learning	2×10^5

A. Simulation Parameters

Through all the simulations, the scenarios with three classical user distributions (uniform user distribution, uni-modal user distribution and bi-modal user distribution) are considered. It is assumed that all the users are located within a target rectangular area $[-L_x, L_x] \times [-L_y, L_y]$, where $L_x = 500$ meters, $L_y = 500$ meters, as illustrated in Fig. 3. Particularly, the probability density function for user uniform distribution can be expressed as

$$f(x, y) = \frac{1}{|\mathcal{A}|}, \tag{20}$$

where $|\mathcal{A}|$ represents the total area of the network. The uni-modal user probability density function can be described by the two-dimensional truncated Gaussian distribution, i.e.,

$$f(x, y) = \frac{1}{\eta} \exp\left[-\left(\frac{x-\mu_x}{\sqrt{2}\sigma_x}\right)^2\right] \exp\left[-\left(\frac{y-\mu_y}{\sqrt{2}\sigma_y}\right)^2\right], \tag{21}$$

where $\eta = 2\pi\sigma_x\sigma_y \operatorname{erf}\left(\frac{L_x-\mu_x}{\sqrt{2}\sigma_x}\right) \operatorname{erf}\left(\frac{L_y-\mu_y}{\sqrt{2}\sigma_y}\right)$, the parameters μ_x, μ_y, σ_x, σ_y are the mean and standard deviation values of L_x and L_y, respectively, and $\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2}$ is the Gauss error function. (μ_x, μ_y) represents the center of hot spot. As shown in Fig. 3, the closer to the hot spot center, the higher the density of users will be. Similarly, the bi-modal user distribution can be seen as the sum of two truncated Gaussian distributions, whose probability density function can be written as

$$f(x, y) = \lambda f_1(x, y) + (1 - \lambda) f_2(x, y), \tag{22}$$

where $0 \leq \lambda \leq 1$ represents a weight factor. $f_1(x, y)$ and $f_2(x, y)$ represent two truncated Gaussian distribution models. This model corresponds to multiple hot spots scenarios such as residential areas. The simulation parameters are listed in Table I.

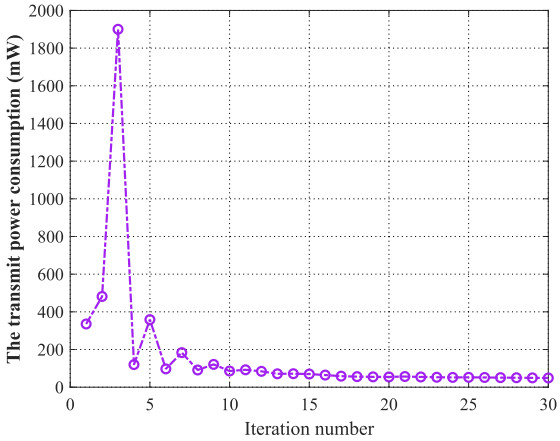


Fig. 4. The convergence versus the iteration number.

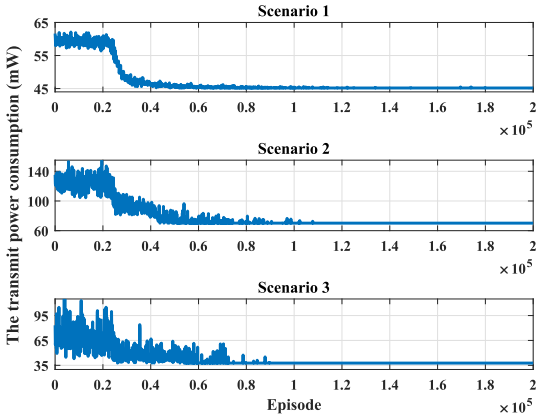


Fig. 5. The multi-agent Q-learning convergence speed versus the training episodes.

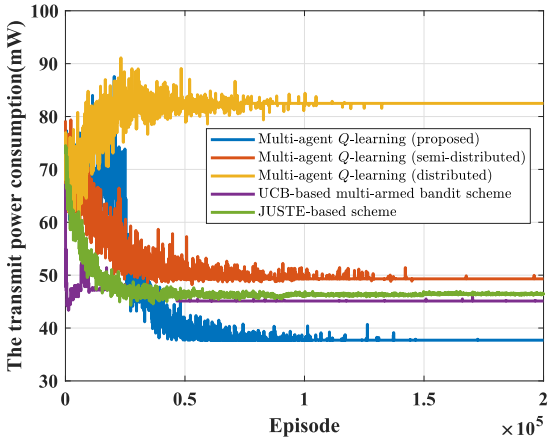


Fig. 6. The multi-agent Q-learning algorithms comparison.

B. The Convergence for the Proposed Algorithm

The convergence of **Algorithm 1** and **Algorithm 2** are first investigated. We show the relationship between the transmit power consumption and the number of iteration of **Algorithm 1** in Fig. 4. Apparently, after convergence, the transmit power consumption is minimized (corresponding to the optimal association scheme). To investigate the convergence property of **Algorithm 2**, we set $M = 700$ and $N = 4$,

and the simulations are done for three different scenarios, and results are illustrated in Fig. 5. It can be observed that the transmits power fluctuates dramatically at the beginning of the training process. The Q-table is updated step by step along with episode. In this way, the optimal action policy at each state will be learned, and finally the system performance will attain a stable value, i.e. it converges to an optimal state. Besides, it should be noted that the convergence speed in each training process is highly dependent on the initiated deployment of UAVs. It means that, if the initiated positions of UAVs are close to the optimal deployment, then the algorithm will converge to the stable state quickly and vice versa.

Besides, the distributed multi-agent reinforcement learning algorithm and the semi-distributed (or semi-central) multi-agent reinforcement learning algorithm are simulated as baselines. In the distributed learning scheme, each UAV is an independent agent. The power consumption of each UAV and its associated users are regarded as the individual reward in the training process. For the semi-distributed or semi-central scheme, the global transmit power consumption is regarded as the reward for each independent UAV. It should be pointed out that the distributed multi-agent reinforcement learning method is a kind of selfish training strategy since each UAV (agent) only pays attention to its own interest, which will lead to the worst power consumption performance, as shown in Fig. 6. For the semi-distributed learning algorithm, it can be seen as the revised version of distributed reinforcement learning. Each UAV (agent) considers the team's interest (i.e., the global power consumption) when it takes action. Therefore, the power consumption performance can be greatly improved compared with the distributed scheme. Nevertheless, it is hard to ensure the optimality for the semi-distributed learning method since it lacks an action coordinator among the UAVs to avoid short-sighted solutions. Meanwhile, the joint utility and strategy estimation (JUSTE) based scheme [42] and the upper-confidence-bound (UCB) based scheme [43] are also simulated for comparison purpose. In performing JUSTE-RL scheme, an estimation of the expected utility for each of the actions is built. Such utility estimations are then used in the same iteration to finally build an action strategy. Since each UAV only holds an estimation of its expected strategy utility, the system is expected to achieve some steady states or equilibrium instead of the global optimal solution. While with the UBC-based multi-armed bandit scheme, UAVs calculate an upper confidence bound of the mean reward of each action and choose the action with the highest estimated bound. From the results in Fig. 6, it is observed that the proposed algorithm can help determine the optimal discrete UAVs deployment positions and realize the minimum transmit power consumption.

To investigate the impact of discretization on the proposed algorithm using Q-learning, we simulate the discretization schemes from 8 points to 36 points, as shown in Fig. 7. There is no doubt that the more discrete points mean the less discretization error the UAV network has. It is obvious that discretization with more points will make UAVs deploy closer to the real optimal positions. However, the fine-grained discretization scheme will make it hard to find the optimal deployment policy, since it will result in a huge state-action matrix and a very higher computational complexity in the training process. On the contrary, the coarse discretization scheme (i.e., sparse discrete points) will contribute to the convergence of Q-learning but at the cost of UAV deployment accuracy and performance.

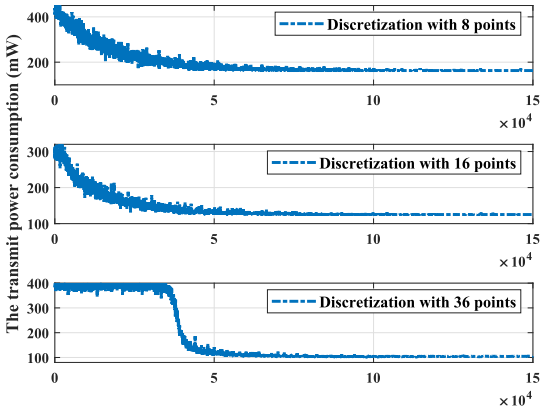
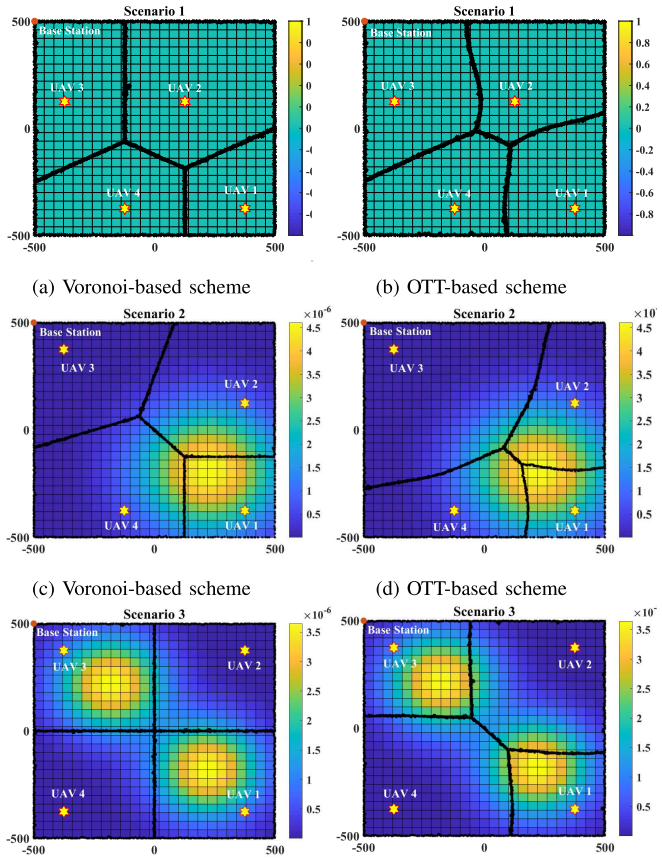


Fig. 7. The convergence comparison in three discretization schemes (N = 3).

C. Superiority of the Proposed OTT-Based Association

In this subsection, the proposed OTT-based association scheme and its corresponding communication loads distribution in three different scenarios are studied. To demonstrate the superiority of the proposed scheme, the Voronoi-based association scheme is chosen as the baseline. Voronoi approach is a classical cell partition method and is widely used in the planning of wireless networks. In this approach, all users associate base stations following the shortest distance rule. Besides, in all three scenarios, UAVs (marked by yellow stars) are deployed randomly and the base station is settled at the northwest corner of the target area (i.e., the point $(-500, 500)$). By setting $M = 700$ and $N = 4$, the simulations are done and the results are shown in Fig. 8 and Fig. 9, respectively. Specifically, figures (a), (c) and (e) of Fig. 8 show the performance of the baseline scheme and the rest show the performance of the proposed OTT-based scheme. It can be seen from Fig. 8 (a) that, in scenario 1, the area is divided into four parts and each cell boundary is determined by finding UAVs' perpendicular bisector. Such an association scheme relies on the distance among the UAVs and ignores the communication load of each UAV. Just as shown in the subfigure on the top of Fig. 9, about 40% of users are serviced by UAV 2 alone and the rest of users are serviced by UAVs 1, 3 and 4 jointly. On the contrary, the proposed OTT-based association scheme offers more balanced communication loads for UAVs. As shown in Fig. 8 (b), the cell boundary of each UAV depends on the UAVs' locations and user distribution jointly. Therefore, the ground users serviced by UAV 2 when employing the OTT-based association decreases dramatically from 40% to 27%, and the proportion of the ground users covered by UAV 1 increases from about 15% to 20%. In Scenario 2, i.e., Fig. 8 (c) and Fig. 8 (d),



(a) Voronoi-based scheme (b) OTT-based scheme

(c) Voronoi-based scheme (d) OTT-based scheme

(e) Voronoi-based scheme (f) OTT-based scheme

Fig. 8. OTT-based association schemes in three scenarios ($M = 700$, random deployment).

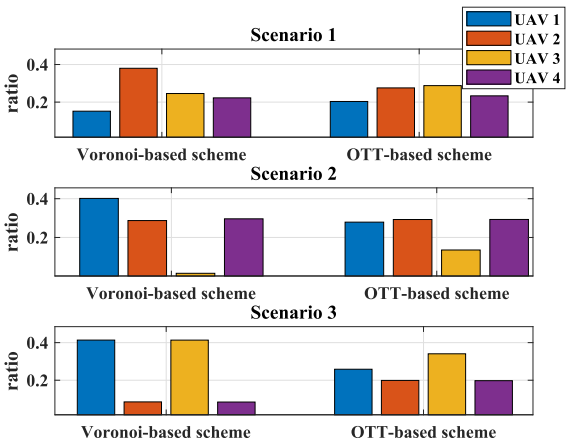


Fig. 9. The user ratio comparison for UAVs in three scenarios.

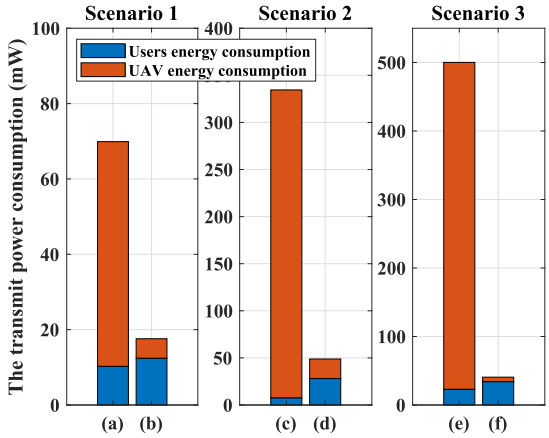


Fig. 10. The transmit power consumption comparison in three scenarios.

the ground users are clustered in a hot spot area. In this case, for the Voronoi-based scheme, only a small percentage (about 1.5%, as shown in the middle subfigure of Fig. 9) of users are covered by UAV 3. Most of the ground users have been allocated to UAV 1 (about 40%), UAV 2 (about 28.7%) and UAV 4 (about 29.6%), which leads to extreme unbalanced communication loads. Fig. 8 (d) offers much more balanced loads, this is because of the capability of user distribution perceiving of proposed OTT-based scheme. It can be seen that the cell covered by UAV 3 increases dramatically to carry more communication loads and to alleviate possible network congestion. In detail, the proportion of the ground users served by UAV 3 increases from 1.5% to 13.5%, the amount of the ground users served by UAV 1 decreases from about 40% to 27%. However, UAV 3 can not cover even larger areas because it is far away from the hot spot area. Since if so, the edge users served by UAV 3 will consume more power to fulfill its data transmission. In Scenario 3, the ground users are mainly distributed in two regions. It can be seen from Fig. 8 (e), most of the ground users (about 82%) will associate with UAVs 1 and 3. When employing OTT-based association scheme, UAVs 2 and 4 expand their coverage to alleviate the transmission pressure of UAVs 1 and 3 that are near to ground users. It can be concluded that the proposed scheme can greatly balance the load among all the UAVs compared with the baseline scheme.

We also simulated the power consumption when employing both the Voronoi-based scheme and the proposed OTT-based association scheme under the same system setups, and the simulation results are shown in Fig. 10. As discussed above, due to the unbalanced loads brought by the Voronoi-based association scheme, the transmit power consumed by the UAVs with heavier communication loads will increase dramatically. For instance, in Scenario 1, the power consumption when employing the Voronoi-based scheme reaches 70 milliwatts and about 85% transmit power is consumed by UAVs, as shown in the left subfigure (a). Similarly, the results obtained in scenarios 2 and 3 confirm the conclusion. We can see more power consumed by UAVs to cope with the unbalanced user association. Comparisons in Fig. 10 show that the proposed OTT-based association can avoid extremely high power consumption since it considers the transmit power consumption of users and UAVs simultaneously when designing the user association scheme. Thanks to the load balancing advantage of the proposed scheme, the transmit power consumption can be reduced and minimized. Specifically, in Scenario 2, compared with the Voronoi-based scheme, the proposed OTT-based association scheme can reduce the transmit power from about 330 milliwatts to about 50 milliwatts, with 84% transmit power consumption reduction.

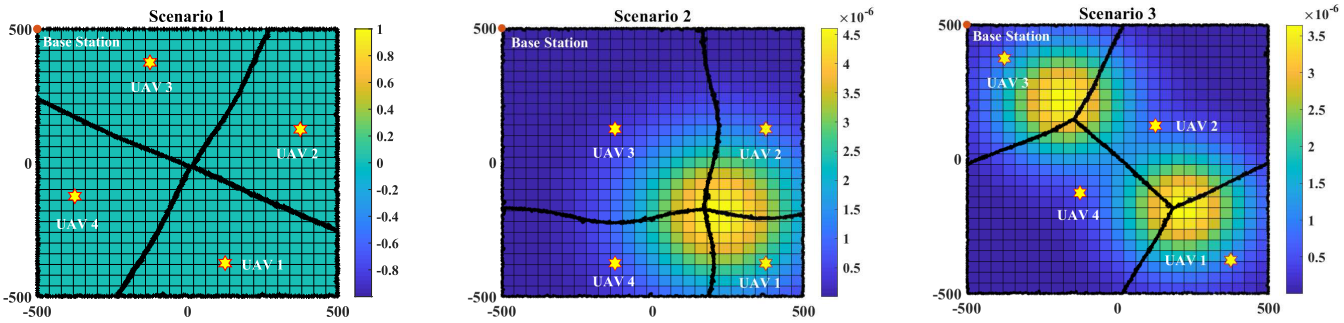


Fig. 11. The multi-agent Q-learning based optimal deployment and OTT-based association scheme ($M = 700, N = 4$).

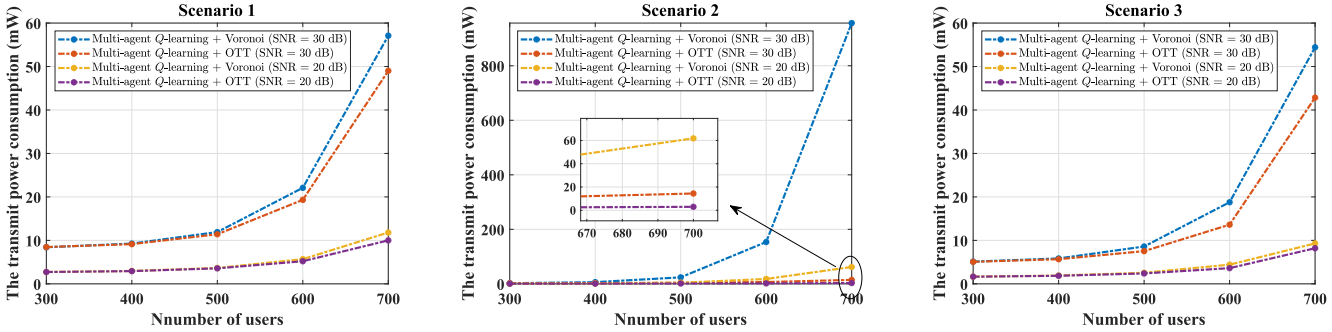


Fig. 12. The transmit power consumption versus the number of users ($M = 700, N = 4$).

D. Superiority of the Proposed Joint Deployment and Association

The system performance of the proposed joint deployment and association are also investigated in three different scenarios described in the last subsection. In doing so, we set $M = 700$ and $N = 4$ and the obtained results are shown in Fig. 11. It is obvious that all UAVs are deployed close to the hot spot area in Scenarios 2 and 3, but not at the center of hot spot area. This is because deployment and association are jointly designed to minimize the transmit power consumption. In addition, as discussed above, the optimal deployment and association scheme guarantees balanced communication loads, which has been confirmed in scenario 1.

To verify the validity of the proposed joint deployment and association scheme, we adopt the multi-agent Q-learning (MAQL)-based deployment + Voronoi-based association scheme as benchmark, and investigate the transmit power consumption. The results are illustrated in Fig. 12. Clearly, with the increase of users, the consumed transmit power increases and the proposed schemes outperform the benchmark schemes in all scenarios. For example, the consumed transmit power of the benchmark scheme increases sharply to about 950 milliwatts (SNR $=$ 30 dB) as the user number goes to 700 in Scenario 2. On the contrary, the proposed scheme only consumes about 18 milliwatts (SNR $=$ 30 dB) and 8 milliwatts (SNR $=$ 20 dB). This is because although the benchmark schemes adopt multi-agent Q-learning deployment scheme, the inappropriate association still leads to unbalanced communication loads. This phenomenon confirms the fact that accumulated communication loads will cost large energy overhead. In scenario 1 and scenario 3, the proposed scheme also brings about 16% (SNR $=$ 30 dB) and 27% (SNR $=$ 30 dB) performance gain, respectively. In our simulation, users only transmit a small part of communication data, therefore the advantage of our proposed scheme will become more significant as the communication loads become heavier. The area energy efficiency of the proposed scheme has also been investigated at different receive SNR in the described three scenarios. The results are illustrated in Fig. 13. It can be seen that, in the UAV-assisted cellular network, as receive SNR increases, the increment of the consumed transmit power is far beyond the increment of the number of the transmitted data, thus the total energy efficiency is decreased. In this case, in order to achieve higher energy efficiency, the deployment and association schemes should be designed elaborately to balance the communication load among multiple UAVs. The proposed scheme can sense the distribution of the ground users, with which UAVs can find the hot spot area and decide where to hover to avoid the transmission congestion. Therefore, the communication load can be well balanced. Simulation results confirm that the proposed sensing-based scheme has higher energy efficiency than the non-sensing scheme in all the three scenarios, e.g., in Scenario 2, the proposed scheme (MAQL-based deployment + OTT-based association, $N = 4$) achieves about 85% energy efficiency performance gain when SNR $=$ 15dB, compared with the non-sensing scheme (MAQL-based deployment + Voronoi-based association, $N = 4$). Besides, compared with the benchmark scheme (random deployment + OTT-based association, $N = 4$), our scheme still has obvious advantages in terms of the area energy efficiency (about 64% performance

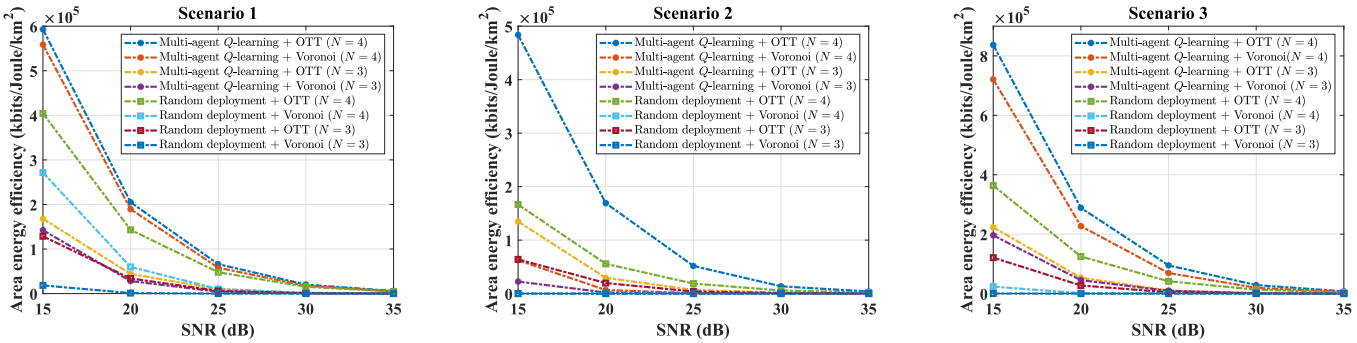
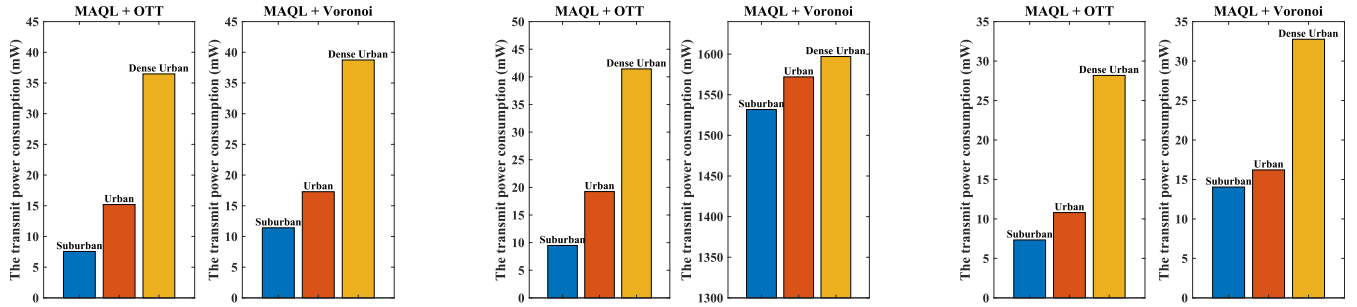


Fig. 13. The area energy efficiency versus the SNR in three scenarios ($M = 600$).



(a) Scenario 1 (b) Scenario 2 (c) Scenario 3

Fig. 14. The power consumption comparison in multiple urban environment: Subruban, Urban, Dense Urban.

improvement). It should be pointed out that, due to the heavier communication loads, the schemes with fewer drones ($N = 3$) consume more transmit power and have lower energy efficiency.

To be thorough, we also simulate the performance of the proposed method in other propagation environments, i.e., suburban ($b_1 = 4.88$, $b_2 = 0.43$), urban ($b_1 = 9.61$, $b_2 = 0.16$) and dense urban $b_1 = 12.08$, $b_2 = 0.11$ [32]. All the obtained results are inluded in Fig. 14. It can be seen that in all the three scenarios, the proposed scheme can greatly outperforms the benchmarks, and similar conclusions can be drawn as in urban scenarios.

V. Conclusion and Discussion

This paper investigated the transmit power minimization problem in UAV-assisted cellular uplink networks. The UAV deployment and association scheme were designed based on the knowledge of statistical user position information. In doing so, the deployment of multiple UAVs was solved by adopting centralized multi-agent Q-learning algorithm. In the training process of the Q-learning, the reward for a state was calculated by developing the optimal association scheme of the next state and the corresponding power consumption. The existence of the unique optimal association scheme for given statistical user distribution and UAVs state was proved by adopting the optimal transport theory, and an iteration algorithm was developed to approach the optimal solution. By the proposed algorithms, the communication load for each UAV in any scenario can be balanced and the transmit power consumption in the UAV-assisted cellular networks can be minimized. Intensive simulations were done, and the results verified the analysis and superiority of the proposed algorithm.

This initial study opens up the opportunity to model and investigate UAV deployment and association problems with the knowledge of statistical user distribution information in the mobile relay systems. The proposed method can be extended to improve the other network performance, such as the communication delay minimization and the throughput maximization. Besides, the proposed method also can be adopted to other research areas, such as city planning. The location of the point of interest (e.g., hospital or shopping mall) can be optimized by using the long-term statistical user distribution information.

Appendix A
Proof of Theorem 1

As we can see, the formulated problem in (15) is a continuous function over the compact set $\Omega_{k+1} = \{\tilde{\mathcal{A}}_1, \tilde{\mathcal{A}}_2, \cdots, \tilde{\mathcal{A}}_N\}$. Therefore, there must exist an association scheme which can minimize the transmit power consumption. In order to show the uniqueness of the optimal association, the convexity of the optimization function has to be proved. However, it is difficult to determine whether the optimization function is convex or not, since it has a complex and highly non-linear structure. To tackle this problem, using the optimal transport theory, we can convert the formulated problem into an equivalent form. To be more specific, we rewrite the optimization problem as follows.

$$\min \; P(\mathbf{s}_{k+1}, \Omega_{k+1}) = \min_{\tilde{\mathcal{A}}_i} \sum_{i=1}^{N} \int_{\tilde{\mathcal{A}}_i} \left[p_u^i(x, y, l_i) + p_{i,u}^0(a_i, l_i)\right] f(x, y) dxdy. \quad (23)$$

Recalling that $a_i = \int_{\tilde{\mathcal{A}}_i} f(x, y) dxdy$, we can define the unit simplex in $\mathbb{R}^N$, i.e., $\mathbf{a} = \{a_1, a_2, \ldots, a_N\}$, then (23) can be

written as

$$\min_{\tilde{\mathcal{A}}_i} \sum_{i=1}^{N} \left\{ \int_{\tilde{\mathcal{A}}_i} p_u^i(x,y,l_i) f(x,y) dxdy + \sum_{i=1}^{N} a_i p_{i,u}^0(a_i,l_i) \right\}. \tag{24}$$

Apparently, in the considered system, the ground users are distributed within the area according to the continuous distribution function μ with the probability density function $f(x,y)$. For UAVs, they can be seen as discrete distribution and we have $\nu = \sum_{i\in\mathbb{N}} a_i \delta_{s_i}$, where δ_{s_i} denotes the Dirac function at point s_i, i.e., the individual state of UAV i. According to the dual formulation in the optimal transport theory, (24) can be converted into an equivalent form,

$$\begin{aligned} &\min_{\tilde{\mathcal{A}}_i} \sum_{i=1}^{N} \left\{ \int_{\tilde{\mathcal{A}}_i} p_u^i(x,y,l_i) f(x,y) dxdy + \sum_{i=1}^{N} a_i p_{i,u}^0(a_i,l_i) \right\} \\ &\qquad = \min_{\mathbf{a}} \left\{ K_p^d(\mu,\nu) + \sum_{i=1}^{N} a_i p_{i,u}^0(a_i,l_i) \right\}, \end{aligned} \tag{25}$$

where $K_p^d(\mu,\nu)$ is the dual formulation which can described as

$$\begin{aligned} &K_p^d(\mu,\nu) \\ &\quad = \max_{\varphi,\psi} \left\{ \int_{\mathcal{A}} \varphi(x,y) \mathrm{d}\mu(x,y) + \int_{\mathcal{A}'} \psi(x,y) \mathrm{d}\nu(x,y) \right\} \\ &\quad = \max_{\varphi,\psi} \left\{ \int_{\mathcal{A}} \varphi(w) \mathrm{d}\mu(w) + \int_{\mathcal{A}'} \psi(v) \mathrm{d}\nu(v) \right\} \\ &\quad = \min_{\Gamma\in\Pi(\mu,\nu)} \int_{\mathcal{A}\times\mathcal{A}'} p_u^i(w,v,l_i) \mathrm{d}\Gamma(w,v) \end{aligned} \tag{26}$$

where φ and ψ are the functions satisfying that $\varphi(w)+\psi(v) \le p_u^i(w,v)$, $\mathcal{A}$ and $\mathcal{A}'$ are the space where users and UAVs located in, respectively, Γ represents the transport plan which is the distribution on the joint space $\mathcal{A}\times\mathcal{A}'$ with marginals μ and ν, $\Pi(\mu,\nu)$ denotes the set of transport plans π.

Then, we only need to show the convexity of the equivalent form. In order to show the convexity of $K_p^d(\mu,\nu)$, we select $\{a_1,a_2,\ldots,a_N\}$ and $\{a_1',a_2',\ldots,a_N'\}$ randomly from $\mathbf{a}$ and let $t\in[0,1]$. In addition, we define function F, and let it be

$$F(a_1,a_2,\ldots,a_N) = K_p^d(\mu,\sum_{i\in\mathbb{N}} a_i\delta_{s_i}). \tag{27}$$

Then, considering the two optimal transport plans, i.e.,

$$\begin{aligned} \Gamma &\in \Pi(\mu,\sum_{i\in\mathbb{N}} a_i\delta_{s_i}). \\ \Gamma' &\in \Pi(\mu,\sum_{i\in\mathbb{N}} a_i'\delta_{s_i}). \end{aligned} \tag{28}$$

That means

$$t\Gamma + (1-t)\Gamma' \in \Pi(\mu,\sum_{i\in\mathbb{N}}(ta_i+(1-t)a_i')\delta_{s_i}). \tag{29}$$

Hence, we have

$$\begin{aligned} &F\left(t(a_1,a_2,\ldots,a_N)+(1-t)(a_1',a_2',\ldots,a_N')\right) \\ &\quad = K_p^d(\mu,\sum_{i\in\mathbb{N}}(ta_i+(1-t)a_i')\delta_{s_i}) \\ &\quad \le \int_{\mathcal{A}\times\mathcal{A}'} p_u^i(w,v)\mathrm{d}(t\Gamma+(1-t)\Gamma')(w,v) \\ &\quad = t\left(\int_{\mathcal{A}\times\mathcal{A}'} p_u^i(w,v)\mathrm{d}\Gamma(w,v)\right) \\ &\qquad +(1-t)\left(\int_{\mathcal{A}\times\mathcal{A}'} p_u^i(w,v)\mathrm{d}\Gamma'(w,v)\right) \\ &\quad = tK_p^d(\mu,\sum_{i\in\mathbb{N}} a_i\delta_{s_i}) + (1-t)K_p^d(\mu,\sum_{i\in\mathbb{N}} a_i'\delta_{s_i}) \\ &\quad = tF(a_1,a_2,\ldots,a_N)+(1-t)F(a_1',a_2',\ldots,a_N'). \end{aligned} \tag{30}$$

Thus, $K_p^d(\mu,\nu)$ is convex over $\mathbf{a}$. Apparently, $\sum_{i=1}^{N} a_i p_{i,u}^0(a_i,l_i)$ is also a convex function over $\mathbf{a}$, since it is the sum of multiple exponential functions. The convexity of $K_p^d(\mu,\nu)$ together with the convexity of $\sum_{i=1}^{N} a_i p_{i,u}^0(a_i,l_i)$ implies the uniqueness of the optimal value $\mathbf{a}^* = \{a_1^*,a_2^*,\ldots,a_N^*\}$ of (25). Since users are distributed continuously, UAVs are distributed following discrete distribution $\nu = \sum_{i\in\mathbb{N}} a_i^*\delta_{(s_i)}$, according to the optimal transport theory, we have an unique transport map T such that

$$T(x,y) = \sum_{i\in\mathcal{N}} s_i \mathbf{1}_{\tilde{\mathcal{A}}_i^*}(x,y) \quad \text{and} \quad \mu\left(\tilde{\mathcal{A}}_i^*\right) = a_i^*. \tag{31}$$

That is to say there is only one optimal association scheme.

We denote the optimal association scheme by $(\tilde{\mathcal{A}}_i^*)_{i=1,\cdots,N}$. Now, we conceive another association scheme $(\widehat{\mathcal{A}}_i)_{i=1,\cdots,N}$ which can be described as

$$\begin{cases} \widehat{\mathcal{A}}_m = \tilde{\mathcal{A}}_m^* \cup B_\varepsilon(\boldsymbol{v}_*) \\ \widehat{\mathcal{A}}_n = \tilde{\mathcal{A}}_n^* \backslash B_\varepsilon(\boldsymbol{v}_*) \\ \widehat{\mathcal{A}}_i = \tilde{\mathcal{A}}_i^*, \quad i\ne m,n \end{cases} \tag{32}$$

where point $\boldsymbol{v}_* = (x_*,y_*) \in \tilde{\mathcal{A}}_i^*$. $B_\varepsilon(\boldsymbol{v}_*)$ is an area with a center point $\boldsymbol{v}_*$ and its radius $\varepsilon > 0$. Thus, it can be derived that

$$\begin{aligned} &\sum_{i=1}^{N} \int_{\tilde{\mathcal{A}}_i^*} \left[p_u^i(x,y,l_i) + p_{i,u}^0(a_i,l_i)\right] f(x,y) dxdy \\ &\le \sum_{i\ne m,n} \int_{\widehat{\mathcal{A}}_i} \left[p_u^i(x,y,l_i) + p_{i,u}^0(a_i,l_i)\right] f(x,y) dxdy \\ &\quad + \int_{\widehat{\mathcal{A}}_m} \left[p_u^m(x,y,l_m) + p_{m,u}^0(a_m+a_\varepsilon,l_m)\right] f(x,y) dxdy \\ &\quad + \int_{\widehat{\mathcal{A}}_n} \left[p_u^n(x,y,l_n) + p_{n,u}^0(a_n-a_\varepsilon,l_n)\right] f(x,y) dxdy. \end{aligned} \tag{33}$$

This leads to

$$\begin{aligned}
&\int_{\tilde{\mathcal{A}}_m^*} p_u^m(x,y,l_m)f(x,y)dxdy + a_m p_{m,u}^0(a_m,l_m) \\
&\quad + \int_{\tilde{\mathcal{A}}_n^*} p_u^n(x,y,l_n)f(x,y)dxdy + a_n p_{n,u}^0(a_n,l_n) \\
&\le \int_{\widehat{\mathcal{A}}_m} p_u^m(x,y,l_m)f(x,y)dxdy \\
&\quad +(a_m+a_\varepsilon)p_{m,u}^0(a_m+a_\varepsilon,l_m) \\
&\quad +\int_{\widehat{\mathcal{A}}_n} p_u^n(x,y,l_n)f(x,y)dxdy \\
&\quad +(a_n-a_\varepsilon)p_{n,u}^0(a_n-a_\varepsilon,l_n).
\end{aligned}$$

Then, we have

$$\begin{aligned}
&\int_{B_\varepsilon(\boldsymbol{v}_*)} p_u^n(x,y,l_n)f(x,y)dxdy + a_\varepsilon p_{n,u}^0(a_n-a_\varepsilon,l_n) \\
&\quad +a_n\left(p_{n,u}^0(a_n,l_n) - p_{n,u}^0(a_m-a_\varepsilon,l_n)\right) \\
&\le \int_{B_\varepsilon(\boldsymbol{v}_*)} p_u^m(x,y,l_m)f(x,y)dxdy + a_\varepsilon p_{m,u}^0(a_m+a_\varepsilon,l_m) \\
&\quad +a_m\left(p_{m,u}^0(a_m+a_\varepsilon,l_m) - p_{m,u}^0(a_m,l_m)\right). \qquad (34)
\end{aligned}$$

Now, we divide a_ε at both sides of equation (34) and take the limit for $\varepsilon \to 0$, we have

$$\begin{aligned}
&p_u^n(x_*,y_*,l_n) + p_{n,u}^0(a_n,l_n) + a_n \cdot \frac{\partial p_{n,u}^0(a_n,l_n)}{\partial a_n} \\
&\le p_u^m(x_*,y_*,l_m) + p_{m,u}^0(a_m,l_m) + a_m \cdot \frac{\partial p_{m,u}^0(a_m,l_m)}{\partial a_m}. \qquad (35)
\end{aligned}$$

This means that (x_*,y_*) in $\widehat{\mathcal{A}}_m$ will cost more power consumption than that in $\widehat{\mathcal{A}}_n$. Thus, the optimal area partition can be described as

$$\begin{aligned}
\tilde{\mathcal{A}}_i^* = \Bigg\{ &(x,y) : p_u^i(x,y,l_i) + p_{i,u}^0(a_i,l_i) + a_i \cdot \frac{\partial p_{i,u}^0(a_i,l_i)}{\partial a_i} \\
&\le p_u^j(x,y,l_j) + p_{j,u}^0(a_l,l_j) + a_j \cdot \frac{\partial p_{i,u}^0(a_j,l_j)}{\partial a_j} \Bigg\}, \\
&\forall i \ne j \in \mathcal{N} \qquad (36)
\end{aligned}$$

Thus, Theorem 1 is proved.

REFERENCES

[1] Y. Zeng, J. Lyu, and R. Zhang, "Cellular-connected UAV: Potential, challenges, and promising technologies," *IEEE Wireless Commun.*, vol. 26, no. 1, pp. 120–127, Feb. 2019.

[2] R. Amer, W. Saad, and N. Marchetti, "Mobility in the sky: Performance and mobility analysis for cellular-connected UAVs," *IEEE Trans. Commun.*, vol. 68, no. 5, pp. 3229–3246, May 2020.

[3] S. Zhang, Y. Zeng, and R. Zhang, "Cellular-enabled UAV communication: A connectivity-constrained trajectory optimization perspective," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2580–2604, Mar. 2019.

[4] W. Mei, Q. Wu, and R. Zhang, "Cellular-connected UAV: Uplink association, power control and interference coordination," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5380–5393, Nov. 2019.

[5] M. Alzenad, A. El-Keyi, F. Lagum, and H. Yanikomeroglu, "3-D Placement of an unmanned aerial vehicle base station (UAV-BS) for energy-efficient maximal coverage," *IEEE Wireless Commun. Lett.*, vol. 6, no. 4, pp. 434–437, Aug. 2017.

[6] L. Wang, B. Hu, and S. Chen, "Energy efficient placement of a drone base station for minimum required transmit power," *IEEE Wireless Commun. Lett.*, vol. 9, no. 12, pp. 2010–2014, Dec. 2020.

[7] M. Chen, M. Mozaffari, W. Saad, C. Yin, M. Debbah, and C. S. Hong, "Caching in the sky: Proactive deployment of cache-enabled unmanned aerial vehicles for optimized quality-of-experience," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 5, pp. 1046–1061, May 2017.

[8] Y. Wang, M. Chen, Z. Yang, T. Luo, and W. Saad, "Deep learning for optimal deployment of UAVs with visible light communications," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7049–7063, Nov. 2020.

[9] M. Hua, Y. Wang, C. Li, Y. Huang, and L. Yang, "Energy-efficient optimization for UAV-aided cellular offloading," *IEEE Wireless Commun. Lett.*, vol. 8, no. 3, pp. 769–772, Jun. 2019.

[10] C. Zhan and Y. Zeng, "Energy-efficient data uploading for cellular-connected UAV systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7279–7292, Nov. 2020.

[11] G. Yang, R. Dai, and Y.-C. Liang, "Energy-efficient UAV backscatter communication with joint trajectory design and resource optimization," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 926–941, Feb. 2021.

[12] J. Zhang, Y. Zeng, and R. Zhang, "Spectrum and energy efficiency maximization in UAV-enabled mobile relaying," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.

[13] Y. Zeng and R. Zhang, "Energy-efficient UAV communication with trajectory optimization," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 3747–3760, Jun. 2017.

[14] M. Hua, Y. Wang, Z. Zhang, C. Li, Y. Huang, and L. Yang, "Power-efficient communication in UAV-aided wireless sensor networks," *IEEE Commun. Lett.*, vol. 22, no. 6, pp. 1264–1267, Jun. 2018.

[15] Y. Zeng, J. Xu, and R. Zhang, "Energy minimization for wireless communication with rotary-wing UAV," *IEEE Trans. Wireless Commun.*, vol. 18, no. 4, pp. 2329–2345, Mar. 2019.

[16] H. D. Tran, T. X. Vu, S. Chatzinotas, S. Shahbazpanahi, and B. Ottersten, "Coarse trajectory design for energy minimization in UAV-enabled wireless communications with latency constraints," *IEEE Trans. Veh. Technol.*, vol. 69, no. 9, pp. 9483–9496, Jun. 2020.

[17] Q. Wu, L. Liu, and R. Zhang, "Fundamental trade-offs in communication and trajectory design for UAV-enabled wireless network," *IEEE Wireless Commun.*, vol. 26, no. 1, pp. 36–44, Feb. 2019.

[18] D. Yang, Q. Wu, Y. Zeng, and R. Zhang, "Energy tradeoff in ground-to-UAV communication via trajectory design," *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 6721–6726, Jul. 2018.

[19] P. Yang, X. Cao, X. Xi, W. Du, Z. Xiao, and D. Wu, "Three-dimensional continuous movement control of drone cells for energy-efficient communication coverage," *IEEE Trans. Veh. Technol.*, vol. 68, no. 7, pp. 6535–6546, Jul. 2019.

[20] 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.

[21] C. H. Liu, Z. Chen, and Y. Zhan, "Energy-efficient distributed mobile crowd sensing: A deep learning approach," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1262–1276, Jun. 2019.

[22] X. Liu, Y. Liu, Y. Chen, and L. Hanzo, "Trajectory design and power control for multi-UAV assisted wireless networks: A machine learning approach," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 7957–7969, Aug. 2019.

[23] C. H. Liu, X. Ma, X. Gao, and J. Tang, "Distributed energy-efficient multi-UAV navigation for long-term communication coverage by deep reinforcement learning," *IEEE Trans. Mobile Comput.*, vol. 19, no. 6, pp. 1274–1285, Jun. 2020.

[24] S. F. Abedin, M. S. Munir, N. H. Tran, Z. Han, and C. S. Hong, "Data freshness and energy-efficient UAV navigation optimization: A deep reinforcement learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 9, pp. 5994–6006, Sep. 2020.

[25] X. Liu, Y. Liu, and Y. Chen, "Reinforcement learning in multiple-UAV networks: Deployment and movement design," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 8036–8049, Aug. 2019.

[26] C. Zhan, Y. Zeng, and R. Zhang, "Energy-efficient data collection in UAV enabled wireless sensor network," *IEEE Wireless Commun. Lett.*, vol. 7, no. 3, pp. 328–331, Jun. 2018.

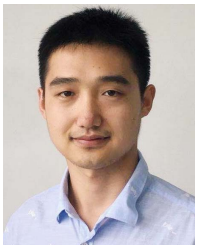
[27] C. Zhan and H. Lai, "Energy minimization in Internet-of-Things system based on rotary-wing UAV," *IEEE Wireless Commun. Lett.*, vol. 8, no. 5, pp. 1341–1344, Oct. 2019.

[28] C. Zhang, H. Zhang, J. Qiao, D. Yuan, and M. Zhang, "Deep transfer learning for intelligent cellular traffic prediction based on cross-domain big data," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1389–1401, Jun. 2019.

[29] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Optimal transport theory for power-efficient deployment of unmanned aerial vehicles," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–6.
[30] Q. Zhang, M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Machine learning for predictive on-demand deployment of UAVs for wireless communications," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–6.
[31] Z. Yang, C. Pan, K. Wang, and M. Shikh-Bahaei, "Energy efficient resource allocation in UAV-enabled mobile edge computing networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 9, pp. 4576–4589, Sep. 2019.
[32] 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.
[33] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Mobile unmanned aerial vehicles (UAVs) for energy-efficient Internet of Things communications," *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, pp. 7574–7589, Nov. 2017.
[34] J. Cui, Y. Liu, and A. Nallanathan, "Multi-agent reinforcement learning-based resource allocation for UAV networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 729–743, Feb. 2020.
[35] L. Busoniu, R. Babuska, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 38, no. 2, pp. 156–172, Mar. 2008.
[36] A. Silva, H. Tembine, E. Altman, and M. Debbah, "Optimum and equilibrium in assignment problems with congestion: Mobile terminals association to base stations," *IEEE Trans. Autom. Control*, vol. 58, no. 8, pp. 2018–2031, Aug. 2013.
[37] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Wireless communication using unmanned aerial vehicles (UAVs): Optimal transport theory for hover time optimization," *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 8052–8066, Dec. 2017.
[38] G. Crippa, C. Jimenez, and A. Pratelli, "Optimum and equilibrium in a transport problem with queue penalization effect," *Adv. Calculus Variat.*, vol. 2, no. 3, pp. 207–246, 2009.
[39] C. Claus and C. Boutilier, "The dynamics of reinforcement learning in cooperative multiagent systems," *AAAI/IAAI*, vol. 1998, nos. 746–752, p. 2, Jul. 1998.
[40] F. S. Melo, "Convergence of Q-learning: A simple proof," Inst. Syst. Robot., Zurich, Switzerland, Tech. Rep, 2001.
[41] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan, "Is Q-learning provably efficient?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 4863–4873.
[42] S. M. Perlaza, H. Tembine, and S. Lasaulce, "How can ignorant but patient cognitive terminals learn their strategy and utility?" in *Proc. IEEE 11th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jun. 2010, pp. 1–5.
[43] S. Maghsudi and E. Hossain, "Distributed user association in energy harvesting dense small cell networks: A mean-field multi-armed bandit approach," *IEEE Access*, vol. 5, pp. 3513–3523, 2017.

Leiyu Wang (Student Member, IEEE) received the B.E. degree in electronics and information engineering from Qufu Normal University, China, in 2014, and the M.Eng. degree in electronics and communication engineering from Inner Mongolia University, China, in 2016. He is currently pursuing the Ph.D. degree with the School of Information Science and Engineering, Shandong University, China. His research interests include unmanned aerial vehicle (UAV) communications, mobile edge computing, and machine learning.

Haixia Zhang (Senior Member, IEEE) received the B.E. degree from the Department of Communication and Information Engineering, Guilin University of Electronic Technology, China, in 2001, and the M.Eng. and Ph.D. degrees in communication and information systems from the School of Information Science and Engineering, Shandong University, China, in 2004 and 2008, respectively. From 2006 to 2008, she was an Academic Assistant with the Institute for Circuit and Signal Processing, Munich University of Technology. From 2016 to 2017, she was a Visiting Professor with the University of Florida, USA. She is currently a Full Professor with Shandong University. She has been actively participating in many professional services. Her current research interests include the industrial Internet of Things (IIoT), resource management, mobile edge computing, and smart communication technologies. She has been serving as a TPC chair, a symposium chair, a TPC member, a keynote speaker, and an invited speaker for conferences. She serves on the Editorial Boards for the IEEE INTERNET OF THINGS JOURNAL, the IEEE WIRELESS COMMUNICATION LETTERS, and *China Communications*.



Shuaishuai Guo (Member, IEEE) received the B.E. and Ph.D. degrees in communication and information systems from the School of Information Science and Engineering, Shandong University, Jinan, China, in 2011 and 2017, respectively. He visited the University of Tennessee at Chattanooga (UTC), USA, from 2016 to 2017. He worked as a Post-Doctoral Research Fellow at the King Abdullah University of Science and Technology (KAUST), Saudi Arabia, from 2017 to 2019. He is currently working as a Full Professor with Shandong University. His research interests include advanced modulation techniques, reconfigurable intelligent surface (RIS), and semantic communications.

Dongfeng Yuan (Senior Member, IEEE) received the M.S. degree from the Department of Electrical Engineering, Shandong University, China, in 1988, and the Ph.D. degree from the Department of Electrical Engineering, Tsinghua University, China, in January 2000. From 1993 to 1994, he was with the Electrical and Computer Department, University of Calgary, Calgary, AB, Canada. He was with the Department of Electrical Engineering, University of Erlangen, Germany, from 1998 to 1999; with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA, from 2001 to 2002; with the Department of Electrical Engineering, Munich University of Technology, Germany, in 2005; and with the Department of Electrical Engineering, Heriot-Watt University, U.K., in 2006. He is currently a Full Professor with the School of Information Science and Engineering, Shandong University. His current research interests include intelligent communication systems, mobile edge computing and cloud computing, AI, and big data processing for communications.