

# UAV Trajectory, User Association and Power Control for Multi-UAV Enabled Energy Harvesting Communications: Offline Design and Online Reinforcement Learning

Chien-Wei Fu, *Student Member, IEEE*, Meng-Lin Ku, *Senior Member, IEEE*, Yu-Jia Chen, *Senior Member, IEEE*, and Tony Q. S. Quek, *Fellow, IEEE*

**Abstract**—In this paper, we consider multiple solar-powered wireless nodes which utilize the harvested solar energy to transmit collected data to multiple unmanned aerial vehicles (UAVs) in the uplink. In this context, we jointly design UAV flight trajectories, UAV-node communication associations, and uplink power control to effectively utilize the harvested energy and manage co-channel interference within a finite time horizon. To ensure the fairness of wireless nodes, the design goal is to maximize the worst user rate. The joint design problem is highly non-convex and requires causal (future) knowledge of the instantaneous energy state information (ESI) and channel state information (CSI), which are difficult to predict in reality. To overcome these challenges, we propose an offline method based on convex optimization that only utilizes the average ESI and CSI. The problem is solved by three convex subproblems with successive convex approximation (SCA) and alternative optimization. We further design an online convex-assisted reinforcement learning (CARL) method to improve the system performance based on real-time environmental information. An idea of multi-UAV regulated flight corridors, based on the optimal offline UAV trajectories, is proposed to avoid unnecessary flight exploration by UAVs and enables us to improve the learning efficiency and system performance, as compared with the conventional reinforcement learning (RL) method. Computer simulations are used to verify the effectiveness of the proposed methods. The proposed CARL method provides 25% and 12% improvement on the worst user rate over the offline and conventional RL methods.

**Index Terms**—Unmanned aerial vehicle (UAV) communication, energy harvesting (EH), UAV trajectory, communication association, power control, convex optimization, reinforcement learning (RL)

## I. INTRODUCTION

In the era of Internet of Things (IoT), many wireless communication nodes are deployed over wide areas for applications toward sustainable development, e.g., environmental monitoring. Traditional ways to powering up electrical devices by connecting the power grids or non-rechargeable batteries

with frequent replacement incur high deployment and maintenance cost. In recent years, energy harvesting (EH) technology has been recognized as an effective means to meet the energy requirement of electrical devices and prolong the lifetime of wireless networks. Due to its ubiquitous abundance, solar power is one of the most preferable EH sources and is capable of providing nearly-permanent power supply. However, renewable energy sources are often affected by the environment resulting in stochastic EH, and therefore the power control design of communication nodes becomes an important issue for self-sustaining wireless services [1].

More recently, unmanned aerial vehicle (UAV) communications have received tremendous attention in various applications such as data collection [2], wireless power transfer [3]–[7], relaying [8], and mobile edge computing [9], due to the potential to improve the transmission quality with the significant advantages of high mobility, flexible deployment, and low cost [10]. In contrast to the wireless networks with terrestrial infrastructures, the UAV can be flexibly deployed to collect sensing data from widely distributed nodes by dynamically adjusting its location. This can dramatically improve the energy efficiency and facilitate the successful deployment of EH communications. Multi-UAV can form multiple mobile base stations to improve the network performance, but interference management becomes a critical challenge when multiple nodes communicate with multiple UAVs concurrently. Since the flight of UAVs is limited by the battery power, it is crucial to investigate wireless resource allocation in UAV-enabled communications.

When multi-UAV-enabled communications are designed for serving multi-EH wireless nodes (WNs), several design challenges need to be carefully addressed. First, the communication association between the UAVs and the WNs is crucial for mitigating the interference from the concurrently served multi-EH WNs to the multi-UAVs, either through appropriate scheduling of alternate transmissions or by selecting WNs with the least interference impact to other users. The communication association is especially important for EH WNs with limited energy. Second, the power control of EH WNs is subject to energy causality constraints, where the energy consumed by power control at any time should not exceed

Chien-Wei Fu, Meng-Lin Ku and Yu-Jia Chen are with the Department of Communication Engineering, National Central University, Zhongli 32001, Taiwan (E-mail: fuchienwei8666@gmail.com, mlku@ce.ncu.edu.tw, yjchen@ce.ncu.edu.tw). Tony Q. S. Quek is with the Information Systems Technology and Design (ISTD) Pillar, Singapore University of Technology and Design, Singapore 487372 (e-mail: tonyquek@sutd.edu.sg).

Corresponding author: Meng-Lin Ku

the total amount of energy collected from the past to the present, and battery storage constraints, where energy charging at any time is limited to the maximum battery capacity [11]. Third, the UAV flight has initial and final flight position constraints, maximum flight speed constraints, flight altitude constraints, and safety distances between multiple UAVs for collision avoidance [12]-[14]. In particular, the flight paths of the multiple UAVs will interact with each other, which has a huge impact on the system performance. In this regard, the joint design of the transmit power control of EH WNs, the communication association between UAVs and EH WNs, and the multi-UAV flight trajectories is a problem worthy of study, as they are closely related to each other. The joint design over the time horizon typically relies on the channel state information (CSI) and energy state information (ESI)<sup>1</sup> from the current time to the future time. Unfortunately, it is very difficult to predict and obtain accurate non-causal CSI and ESI in reality due to the random nature of harvested energy and the susceptibility of wireless channels. While some dynamic programming methods, such as reinforcement learning (RL), can learn an optimal policy based only on the current system states by repeatedly interacting with the environment, such methods suffer from the curse of dimensionality when the cardinality of the system state and action space is high. Therefore, we are motivated to design offline methods which only require the statistical/average or instantaneous CSI and ESI, called *offline methods*, and the developed offline strategies can provide some useful insights to improve RL-based *online methods*.

#### A. Literature Survey

Many researchers have contributed greatly to the topics of UAV deployment and resource management for UAV communications. There are some early works discussing the deployment of a single UAV. In [15], successive convex approximation (SCA) is used to optimize the UAV trajectory, transmit power, and subcarrier allocation for throughput maximization. In [2], a maximum-minimum data collection rate problem is solved for UAV wireless sensor networks in urban areas, where the three-dimensional UAV trajectory and transmission scheduling of sensors are jointly optimized by convex approximation. In [16], the UAV trajectory is optimized via Q-learning to improve energy efficiency while ensuring QoS of ground users. For multiple UAVs, the interference problem becomes a key issue dominating wireless communication performance and requires new solutions to exploit its inherent spatial degree of freedom. The literature [17] maximizes the uplink transmission rate by designing multi-UAV cooperative transmission, sub-channel allocation and UAV speed. In [18], UAV flight trajectories are investigated through a dueling deep Q-network to maximize downlink channel capacity under the line-of-sight (LOS) channel probability and user coverage constraint. The joint design of communication scheduling, power control, and UAV trajectory is proposed to maximize the minimum

throughput in [19] and to minimize the weighted sum of aerial and ground costs in [20] via alternative optimization and SCA, while the authors in [21] focus on the UAV flight speed, altitude, and power control problems. To simplify the design, most studies, such as [3][19], only consider LOS channels and ignore small-scale fading. To make the research more realistic for UAV applications, the LOS/NLOS channels are emphasized in [2][17], and the effect of small-scale fading is evaluated in [15][18][20][21].

Another line of work in the existing literature is to exploit EH in UAV communications for prolonging the lifetime of wireless devices. In [3], UAVs with radio-frequency (RF) EH capability are utilized to assist mobile edge computing, which can provide users with edge computing services and energy through wireless charging. In [4], a UAV with EH is considered for uninterrupted service, where harvesting and charging time, flight trajectory and speed, and UAV's transmit power allocation are jointly optimized by block coordinate descent and SCA methods. A dynamic fly-hover-transmission scheme is studied in [5][6] for UAV-assisted wireless energy and information transfer in cognitive radio networks, where a constrained Markov decision process (MDP) problem is cast to design the UAV transmission and trajectory based on causal system information for throughput maximization, while [6] proposes an efficient suboptimal but low-complexity transmission policy. In [7], a Q-learning method is used to find the flight policy for UAVs as power stations to maximize mission duration. However, the main drawback of wireless charging is the low charging efficiency due to channel path loss. Also, in some hazardous areas, RF signals may not be readily available or dense enough. As an alternative, EH nodes with the self-sustainability from renewable energy are fascinating in UAV communications and enable UAVs to focus more on data collection and improve system performance [22][23]. In [22], a single UAV with a fixed flight path is dispatched to collect data from multiple sensing nodes equipped with solar cells, which can extend network lifetime without human intervention. In [23], the UAV placement and resource allocation are investigated through deep learning in the renewable energy paradigm.

#### B. Contributions

Although there has been prior work on the study of UAV trajectory and resource allocation for single-UAV scenarios with solar EH nodes [23], multi-UAV scenarios with solar EH nodes have not been investigated in the literature. In addition, most existing multi-UAV design frameworks only consider the LOS channels and neglect the small-scale fading effect. Although LOS/NLOS channels are considered in [2][17] and small-scale fading is considered in [15][18][20][21], the multi-UAV communications with EH nodes remain unexplored. Besides, in the context of power control and UAV trajectory, the MDP approaches [5][6] necessitate the state transition probability in advance, while the convex approaches typically achieve better performance [2][4][15][17][19]-[21] under the assumption that the future CSI and ESI are perfectly known.

<sup>1</sup>The ESI refers to the amount of harvested solar energy.

While the RL can maximize long-term utility without knowing the exact stochastic transition dynamics as in [7][16][18][23], it notably suffers from the curse of dimensionality for systems with large-scale state and action sets, which can lead to ineffective convergence of Q-values. To fill these gaps, this paper presents a framework for jointly designing multi-UAV flight trajectories, communication association between UAVs and nodes, and uplink transmit power control of multi-UAV communication networks with solar-powered ground nodes. The main contributions of this paper are stated as follows.

- To the best of our knowledge, this is the first work to investigate a multi-UAV communication network with multiple solar-powered ground nodes, with a focus on the resource management strategies and multi-UAV flight paths planning to minimize the worst-case user rate.

- A real solar power harvesting dataset and a composite channel model, including LOS/NLOS and small-scale fading, are considered in this design. The joint design problem is non-convex. Based on a series of SCAs, we propose an offline method for jointly optimizing multi-UAV flight trajectories, UAV-node communication associations and transmit power control which only requires average CSI and ESI. The offline design provides valuable insight into the flight path planning of multiple UAVs, especially when future instantaneous CSI and ESI are unpredictable.

- The online RL is then proposed by further taking the current instantaneous CSI and ESI into account. With the help of multi-UAV trajectories in the offline design, a new concept of UAV flight corridor is presented in the online RL design, called convex-assisted RL (CARL), to effectively guide the executed actions and state exploration. By arranging flight corridors ahead together with the offline UAV trajectories, the proposed online method can not only avoid exploring the state space randomly but also effectively respond to the current instantaneous CSI and ESI for performance improvement.

- Computer simulations are performed to demonstrate the effectiveness of the proposed offline and online methods. The proposed offline method performs better than other baseline schemes with fixed UAV paths or fixed uplink transmit power. Furthermore, the proposed CARL method significantly improves the performance, compared to the conventional RL and the proposed offline methods.

The rest of this paper is organized as follows. In Section II, we present the system model and the joint design problem of the UAV flight trajectory, communication association, and power control of solar EH nodes, along with the MDP formulation. In Section III, an offline convex optimization method that utilizes only the average CSI and ESI is proposed. An online CARL method based on the proposed offline design is given in Section IV. Simulation results are provided in Section V, and Section VI concludes the paper.

## II. SYSTEM MODEL AND PROBLEM FORMULATION

In this section, we first present the system model for multi-UAV communications with multiple solar-powered ground nodes, in which the ground wireless nodes harvest energy from

solar and transmit data to the UAVs in the uplink channels over the same frequency band. Afterwards, the joint design problem of multi-UAV flight trajectory, communication association between UAVs and ground nodes, and power control is formulated for multiple UAVs serving multiple uplink solar-powered nodes.

### A. System Model

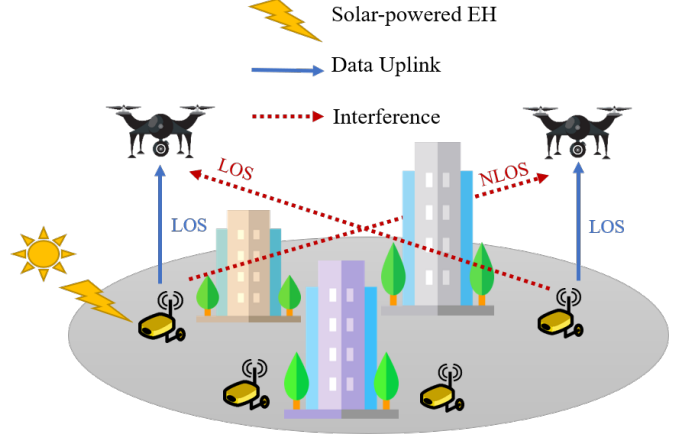


Fig. 1: Multi-UAV communication networks with solar-powered ground nodes for uplink transmissions ( $M=2$  and  $K=4$ ).

Fig. 1 shows the multi-UAV communication networks with solar-powered ground nodes for uplink transmissions. Consider a service area consisting of a group of  $K$  wireless nodes (WNs) which utilize the harvested solar power for data transmission. A group of  $M$  UAVs flies over the area to collect the data from the  $K$  wireless nodes. Both the UAVs and the WNs are only equipped with a single antenna. A time-slotted model is adopted, and we assume that the entire task period  $T_s$  is divided into  $N$  time slots, where there are  $N + 1$  discrete time instants ( $n = 0, 1, \dots, N$ ) and the time interval  $\delta_D$  is defined as

$$\delta_D = \frac{T_s}{N}. \quad (1)$$

We assume that the positions of the WNs are unchanged, and the horizontal two-dimensional coordinate of the  $k$ th WN is given by

$$\mathbf{g}_k = [\bar{x}_k, \bar{y}_k]^T \in \mathbb{R}^2, \forall k \in \{1, \dots, K\}, \quad (2)$$

where  $[\cdot]^T$  takes the vector transpose. It is assumed that the UAVs fly at a fixed altitude  $H$ , and the horizontal two-dimensional coordinate of the  $m$ th UAV at the time instant  $n$  is given as

$$\mathbf{q}_m[n] = [x_m[n], y_m[n]]^T \in \mathbb{R}^2, \quad \forall m \in \{1, \dots, M\}, \forall n \in \{0, \dots, N\}. \quad (3)$$

The trajectory coordinate of the UAVs is subject to the following constraints:

$$\mathbf{q}_m[0] = \mathbf{q}_m[N] = \mathbf{q}_m^{ini}, \forall m \in \{1, \dots, M\}; \quad (4)$$

$$\frac{1}{\delta_D} \|\mathbf{q}_m[n+1] - \mathbf{q}_m[n]\|_2 \leq V_{max},$$

$$\forall m \in \{1, \dots, M\}, \forall n \in \{0, \dots, N-1\}; \quad (5)$$

$$\|\mathbf{q}_m[n] - \mathbf{q}_j[n]\|_2 \geq D_{min},$$

$$\forall n \in \{1, \dots, N-1\}, \forall m, j \in \{1, \dots, M\}, m \neq j, \quad (6)$$

where  $\mathbf{q}_m^{ini}$  is the initial position of the UAVs,  $V_{max}$  is the maximum flight speed of the UAVs, and  $D_{min}$  is the minimum safe distance for any two UAVs. We assume that the UAVs are equipped with finite batteries, and the constraint (4) stipulates that each UAV is required to fly back to the initial point  $\mathbf{q}_m^{ini}$  at the end of the task period. The constraint (5) indicates that the flight speed of each UAV is limited to the maximum flight speed  $V_{max}$ . Moreover, the constraint (6) represents that the minimum safe distance  $D_{min}$  must be maintained between any two UAVs for avoiding collision.

For the channel model, both the LOS and NLOS channels are taken into consideration in the large-scale fading, and the path loss (in decibel) between the  $m$ th UAV and the  $k$ th WN at the time instant  $n$  can be expressed as [24]:

$$L_{\epsilon, m, k}[n] = 20 \log_{10} \left( \frac{4\pi f_c d_{m, k}[n]}{c} \right) + \eta_\epsilon + S,$$

$$\forall m \in \{1, \dots, M\}, \forall k \in \{1, \dots, K\}, n \in \{0, \dots, N\},$$

$$\forall \epsilon \in \{LOS, NLOS\}, \quad (7)$$

where  $f_c$  is the carrier frequency (Hz),  $c$  is the speed of light (m/s),  $\eta_\epsilon$  is an LOS and NLOS environment-related parameter, and  $S$  represents the shadowing effect. Furthermore, the term  $d_{m, k}[n]$  represents the distance between the  $m$ th UAV and the  $k$ th WN at the time instant  $n$ , given as

$$d_{m, k}[n] = \sqrt{\|\mathbf{q}_m[n] - \mathbf{g}_k\|_2^2 + H^2}. \quad (8)$$

According to [24], the probability of occurring the LOS channel between the  $m$ th UAV and the  $k$ th WN at the time instant  $n$  can be modelled as

$$\rho_{LOS, m, k}[n] = \frac{1}{1 + A \exp(-B(\frac{180}{\pi} \tan^{-1}(\frac{H}{\|\mathbf{q}_m[n] - \mathbf{g}_k\|_2}) - A))},$$

$$\forall m \in \{1, \dots, M\}, \forall k \in \{1, \dots, K\}, \forall n \in \{0, \dots, N\}, \quad (9)$$

where the coefficients  $A$  and  $B$  depend on the operating environments [25], and the probability of occurring the NLOS channel can be computed as  $\rho_{NLOS, m, k}[n] = 1 - \rho_{LOS, m, k}[n]$ . By combining the effect of large-scale and small-scale channel fading, we can obtain the channel gain between the  $m$ th UAV and the  $k$ th WN at the time instant  $n$  as follows:

$$H_{m, k}[n] = \check{L}_{\epsilon, m, k} |\chi_{m, k}[n]|^2, \quad (10)$$

where  $\check{L}_{\epsilon, m, k}[n]$  represents the linear scale of  $L_{\epsilon, m, k}[n]$ ,  $\chi_{m, k}[n] \sim CN(0, 1)$  is a zero-mean complex Gaussian random variable with unit variance to describe the Rayleigh fading.

Next, we define the communication association variable between the  $m$ th UAV and the  $k$ th WN at the time instant  $n$  as  $a_{m, k}[n]$ . The communication association variables are subject to the following three constraints:

$$a_{m, k}[n] \in \{0, 1\}, \forall m \in \{1, \dots, M\}, \forall k \in \{1, \dots, K\},$$

$$\forall n \in \{0, \dots, N-1\}; \quad (11)$$

$$\sum_{k=1}^K a_{m, k}[n] \leq 1, \forall m \in \{1, \dots, M\}, \forall n \in \{0, \dots, N-1\};$$

$$(12)$$

$$\sum_{m=1}^M a_{m, k}[n] \leq 1, \forall k \in \{1, \dots, K\}, \forall n \in \{0, \dots, N-1\},$$

$$(13)$$

where the constraint (11) means that the communication association variables take binary values. If  $a_{m, k}[n] = 1$ , the  $k$ th WN is associated with the  $m$ th UAV at the time instant  $n$ . The constraint (12) confines that each UAV can serve at most one WN at each time instant, while each active WN can be only served by one UAV under the constraint (13). Besides, if the  $m$ th UAV associates with one WN at the time instant  $n$  for data transmissions, its position remains unchanged during the time interval  $\delta_D$ , leading to the following constraint:

$$a_{m, k}[n] \cdot \|\mathbf{q}_m[n+1] - \mathbf{q}_m[n]\|_2 = 0,$$

$$\forall m \in \{1, \dots, M\}, \forall n \in \{0, \dots, N-1\}. \quad (14)$$

Since the WNs utilize the same frequency band for uplink communications, each UAV suffers from the uplink multiuser interference problem. From (10), the signal-to-interference plus noise power ratio (SINR) of the  $k$ th WN at the  $m$ th UAV and the  $n$ th time instant can be expressed as

$$\Gamma_{m, k}[n] = \frac{P_k[n] H_{m, k}[n]}{\sum_{i=1, i \neq k}^K P_i[n] H_{m, i}[n] + \sigma_n^2},$$

$$\forall m \in \{1, \dots, M\}, \forall k \in \{1, \dots, K\}, \forall n \in \{0, \dots, N-1\}, \quad (15)$$

where  $\sigma_n^2$  is the power of additive white Gaussian noise, and  $P_k[n]$  is the uplink transmit power of the  $k$ th WN at the time slot  $n$ . Accordingly, the achievable sum rate of the  $k$ th WN over the whole UAV task period can be calculated as

$$R_k = \sum_{n=0}^{N-1} \underbrace{\sum_{m=1}^M a_{m, k}[n] W \log_2(1 + \Gamma_{m, k}[n])}_{\triangleq R_{k, n}},$$

$$\forall k \in \{1, \dots, K\}, \quad (16)$$

where  $W$  is the system bandwidth, and  $R_{k, n}$  is the data rate of the  $k$ th WN at the time instant  $n$ .

Since each WN relies on the harvested solar energy in a finite-capacity battery for uplink transmission, the transmit power of a WN is constrained by its harvested energy and battery capacity. Let  $E_k[n]$  and  $B_k[n]$  represent the harvested energy and battery level of the  $k$ th WN at the time instant  $n$ . For simplicity, we assume that the battery level of WN is zero

at the time instant  $n = 0$ . Therefore, the residual energy in the battery of the  $k$ th WN at the time instant  $n$  can be described as  $B_k[n] = \sum_{l=0}^n E_k[l] - \delta_D \sum_{l=0}^n P_k[l]$ , and the battery power evolution from the time  $n$  to  $(n+1)$  can be described as  $B_k[n+1] = B_k[n] + E_k[n+1]$ . Since the battery power is limited by  $0 \leq B_k[n] \leq B_{max}$ , it then yields two constraints about the uplink transmit power, harvested solar energy, and battery power, namely energy causality and battery storage constraints:

$$\delta_D \sum_{l=0}^n P_k[l] \leq \sum_{l=0}^n E_k[l], \forall k \in \{1, \dots, K\},$$

$$\forall n \in \{0, \dots, N-1\}; \quad (17)$$

$$\sum_{l=0}^{n+1} E_k[l] - \delta_D \sum_{l=0}^n P_k[l] \leq B_{max}, \forall k \in \{1, \dots, K\},$$

$$\forall n \in \{0, \dots, N-1\}, \quad (18)$$

where  $B_{max}$  is the maximum battery storage capacity of WNs. Here, the energy causality constraint (17) mandates that the harvested energy at each WN cannot be used until it arrives. The battery storage constraint (18) indicates that for each WN, the total amount of harvested energy minus the energy expenditure for data transmission cannot exceed the maximum battery storage capacity at each time instant.

### B. Problem Formulation

In order to achieve a fair communication service for multiple WNs during the entire UAV task period, the design goal of this paper is to maximize the worst sum rate among  $K$  WNs. The joint design problem of the trajectories of UAVs, communication association between UAVs and WNs, and transmit power of WNs with solar EH can be formulated as follows:

$$(P1) \quad \max_{\{\mathbf{q}_m[n], P_k[n], a_{m,k}[n], \forall m, k, n\}} \min_{k=1, \dots, K} R_k$$

$$s.t. \text{ (4), (5), (6), (11), (12), (13), (14), (17), (18).}$$

It is worth noting that if the  $k$ th WN is not associated with any UAV, the optimal solution to the uplink transmit power  $P_k[n]$  of the  $k$ th WN must be equal to zero; otherwise, it causes the interference problem and degrades the overall sum rate performance of the network.

In this optimization problem, the UAV needs to know the CSI  $H_{m,k}[n]$  and ESI  $E_k[n]$  not only in the current time instant but also in the future time instants. However, in real applications, it is impractical to acquire the full (past and future) information for carrying out the optimization problem. In response to this design challenge, a RL method can be utilized to perform dynamic optimization for the UAV flight direction, UAV and WN association, and uplink transmit power, solely based on the current battery state information (BSI) of WNs and the current CSI between UAVs and WNs.

As compared with the convex optimization approach, the conventional RL is exempt from the future instantaneous ESIs and CSIs but only depends on the current state information.

However, it requires the UAV to repeatedly take various actions in each state of the environment, which results in a long exploration learning time and a slow update of Q-values. The learning time and convergence performance become even worse for multi-UAV communication networks with multi-EH WNs, since the number of system states and actions increases exponentially and many inefficient actions may be executed during the learning process. For this reason, in this paper, we will first investigate an offline convex optimization method to find the best offline flight trajectory for the UAVs by only applying the ‘‘average’’ LOS/NLOS channel gain and EH profile. The obtained solution is then served as the reference solution for the online RL. The main idea is to mark out a flight corridor based on the reference solution to guide the UAV flight actions in the online learning for reducing the learning time of the conventional RL while improving the system performance of the offline convex optimization approach.

### III. OFFLINE CONVEX OPTIMIZATION DESIGN OF MULTI-UAV SYSTEMS WITH MULTIPLE EH WNs

In this section, we develop an offline convex optimization approach by using the average LOS/NLOS channel gain value on the UAV trajectory  $\bar{H}_{m,k}[n] = \mathbb{E}[H_{m,k}[n]]$  and the average value of the past solar EH time series  $\bar{E}_k[n] = \mathbb{E}[E_k[n]]$  to replace the instantaneous information  $H_{m,k}[n]$  and  $E_k[n]$  in the optimization problem (P1), respectively. It is worth noting that although the statistical average of solar EH profiles is difficult to obtain, it can be acquired by numerically averaging real solar energy historical data. Besides, from (7)–(10), the average LOS/NLOS channel gain  $\bar{H}_{m,i}[n]$  can be derived in terms of the UAV position  $\mathbf{q}_m[n]$  as follows:

$$\begin{aligned} \bar{H}_{m,i}[n] &= \rho_{LOS,m,i}[n] \bar{L}_{LOS,m,i}[n] + \rho_{NLOS,m,i}[n] \bar{L}_{NLOS,m,i}[n] \\ &= \left( \frac{c}{4\pi f_c d_{m,i}[n]} \right)^2 \times 10^{\frac{-S}{10}} \\ &\quad \times \left( \rho_{LOS,m,i}[n] \times 10^{\frac{-\eta_{LOS}}{10}} + \rho_{NLOS,m,i}[n] \times 10^{\frac{-\eta_{NLOS}}{10}} \right) \\ &= C_1 \times C_2 \times d_{m,i}^{-2}[n] \times \rho_{LOS,m,i}[n] + C_3 \times d_{m,i}^{-2}[n] \\ &= C_1 \times C_2 \times \frac{1}{\|\mathbf{q}_m[n] - \mathbf{g}_i\|_2^2 + H^2} \\ &\quad \times \frac{1}{1 + A \exp\left(-B \left(\frac{180}{\pi} \tan^{-1}\left(\frac{H}{\|\mathbf{q}_m[n] - \mathbf{g}_i\|_2}\right) - A\right)\right)} \\ &\quad + C_3 \times \frac{1}{\|\mathbf{q}_m[n] - \mathbf{g}_i\|_2^2 + H^2}, \end{aligned} \quad (19)$$

where  $C_1$ ,  $C_2$ , and  $C_3$  are constants and all greater than zero:

$$C_1 = \left( \frac{c}{4\pi f_c} \right)^2 \times 10^{\frac{-S}{10}} > 0; \quad (20)$$

$$C_2 = \left( 10^{\frac{-\eta_{LOS}}{10}} - 10^{\frac{-\eta_{NLOS}}{10}} \right) > 0; \quad (21)$$

$$C_3 = \left( \frac{c}{4\pi f_c} \right)^2 \times 10^{\frac{-S}{10}} \times 10^{\frac{-\eta_{NLOS}}{10}} > 0. \quad (22)$$

Therefore, the objective function considered in the offline optimization problem is given by

$$\bar{R}_k = \sum_{n=0}^{N-1} \sum_{m=1}^M a_{m,k}[n] W \log_2(1 + \bar{\Gamma}_{m,k}[n]), \quad \forall k \in \{1, \dots, K\}, \quad (23)$$

where  $\bar{\Gamma}_{m,k}[n]$  is the SINR calculated throughout the average channel gain:

$$\bar{\Gamma}_{m,k}[n] = \frac{P_k[n] \bar{H}_{m,k}[n]}{\sum_{i=1, i \neq k}^K P_i[n] \bar{H}_{m,i}[n] + \sigma_n^2}, \quad \forall m \in \{1, \dots, M\}, \quad \forall k \in \{1, \dots, K\}, \quad \forall n \in \{0, \dots, N-1\}. \quad (24)$$

From (P1), the offline joint design problem is formulated as

$$\begin{aligned} (\text{P2}) \quad & \max_{\{\mathbf{q}_m[n], P_k[n], a_{m,k}[n], \forall m, k, n\}} \min_{k=1, \dots, K} \bar{R}_k \\ \text{s.t.} \quad & (4), (5), (6), (11), (12), (13), (14), (17), (18). \end{aligned}$$

Since the objective function of the joint design problem (P2) and the constraints (6), (11) and (14) are non-convex for the three design variables  $\mathbf{q}_m[n]$ ,  $a_{m,k}[n]$  and  $P_k[n]$ . To overcome this issue, we adopt the alternative optimization to decompose the joint optimization problem into three subproblems to optimize the UAV flight trajectory  $\{\mathbf{q}_m[n], \forall m, n\}$ , communication association  $\{a_{m,k}[n], \forall m, k, n\}$ , or power control  $\{P_k[n], \forall k, n\}$  under the fixed values of other variables. Nevertheless, the three subproblems are still non-convex: (i) the communication association subproblem is a non-convex mixed integer programming, and we will relax the integer constraint (11), i.e.,  $\{a_{m,k}[n] \in \{0, 1\}, \forall m, k, n\}$  to  $\{0 \leq a_{m,k}[n] \leq 1, \forall m, k, n\}$  in order to convert the problem into a linear programming problem; (ii) the UAV flight trajectory subproblem and the power control subproblem are also non-convex due to the objective function and the constraint (6), and we will use SCA methods to transform these two non-convex subproblems into convex ones.

#### A. UAV-WN Communication Association Subproblem

Given the transmit power control  $P_k[n]$  of the WNs and the UAV flight trajectory  $\mathbf{q}_m[n]$  in (P2), the optimization can be performed for the communication association. By relaxing the integer constraint (11) and introducing an auxiliary variable  $\zeta_a$ , the subproblem becomes an equivalent epigraph form:

$$\begin{aligned} (\text{P3}) \quad & \max_{\{\zeta_a, a_{m,k}[n], \forall m, k, n\}} \zeta_a \\ \text{s.t.} \quad & \sum_{n=0}^{N-1} \sum_{m=1}^M a_{m,k}[n] W \log_2(1 + \bar{\Gamma}_{m,k}[n]) \geq \zeta_a, \\ & \forall k \in \{1, \dots, K\}, \quad (25) \\ & 0 \leq a_{m,k}[n] \leq 1, \forall m, \forall n \in \{0, \dots, N-1\}, \quad (26) \\ & (12), (13). \end{aligned}$$

The subproblem (P3) now becomes linear programming which can be directly optimized by using off-the-shelf optimization software (CVX) [26]. Since the obtained solution  $a_{m,k}^*[n]$  in (P3) may take a value between 0 and 1, it is quantized to 1, if  $a_{m,k}^*[n] \geq 0.5$ ; otherwise, we set  $a_{m,k}^*[n]$  to 0.

#### B. UAV Flight Trajectory Subproblem

Given the communication association  $a_{m,k}[n]$  and the power control  $P_k[n]$  of WNs, the UAV flight trajectory subproblem can be obtained from the optimization problem (P2) and represented as an epigraph form with the introduction of an auxiliary variable  $\zeta_q$ :

$$\begin{aligned} (\text{P4}) \quad & \max_{\{\zeta_q, \mathbf{q}_m[n], \forall m, k, n\}} \zeta_q \\ \text{s.t.} \quad & \bar{R}_k \geq \zeta_q, \forall k \in \{1, \dots, K\}, \quad (27) \\ & (4), (5), (6), (14). \end{aligned}$$

Observing the constraint (27), we know that the SINR  $\bar{\Gamma}_{m,k}[n]$  in (24) contains  $\bar{H}_{m,k}[n]$  which is related to the multiple UAV flight trajectory variables  $\mathbf{q}_m[n]$ . This makes the constraint (27) non-convex. In addition, the minimum safe distance constraint (6) is also non-convex. As a result, the subproblem (P4) is a non-convex optimization problem, which cannot be directly solved by convex optimization tools. In the following, we resort to the SCA method to convexify the non-convex constraints considering LOS/NLOS channels for multiple UAVs. Note that the SCA method was applied in [2][19], whereas these two works merely considered the design in LOS channels or single-UAV environments and cannot be directly applied to our work.

First, we expand the left-hand side of the constraint (27) into the difference of two logarithmic functions, as shown in (28) at the top of the next page, where we define  $\check{R}_{1m}[n] \triangleq \log_2 \left( \sum_{i=1}^K P_i[n] \bar{H}_{m,i}[n] + \sigma_n^2 \right)$  and  $\check{R}_{2m,k}[n] \triangleq -\log_2 \left( \sum_{i=1, i \neq k}^K P_i[n] \bar{H}_{m,i}[n] + \sigma_n^2 \right)$ . Our goal is to find a lower bound concave function for  $\bar{R}_k$  in (28) in order to transform the constraint (27) into solvable convex constraints. Below we elaborate on the ways to find concave lower bounds for  $\check{R}_{1m}[n]$  and  $\check{R}_{2m,k}[n]$  in terms of  $\mathbf{q}_m[n]$ , followed by the transformation of the subproblem (P4) into a solvable convex optimization problem through these derived lower bounds.

1) A concave lower bound for  $\check{R}_{1m}[n]$ : We define two variables  $X_{m,i}$  and  $Y_{m,i}$ , given by

$$\begin{aligned} X_{m,i}[n] &= \|\mathbf{q}_m[n] - \mathbf{g}_i\|_2^2 + H^2, \quad \forall m \in \{1, \dots, M\}, \\ & \quad \forall i \in \{1, \dots, K\}, \quad \forall n \in \{0, \dots, N-1\}. \quad (29) \\ Y_{m,i}[n] &= 1 + A e^{-B \left( \frac{180}{\pi} \tan^{-1} \left( \frac{H}{\|\mathbf{q}_m[n] - \mathbf{g}_i\|_2} \right) - A \right)}, \\ & \quad \forall m \in \{1, \dots, M\}, \quad \forall i \in \{1, \dots, K\}, \quad \forall n \in \{0, \dots, N-1\}. \quad (30) \end{aligned}$$

The following lemma is provided.

**Lemma 1.** The function  $\check{R}_{1m}[n] = \log_2 \left( \sum_{i=1}^K P_i[n] \bar{H}_{m,i}[n] + \sigma_n^2 \right)$  is convex in  $X_{m,i}[n]$  and  $Y_{m,i}[n]$  for all  $i$ .

*Proof.* See Appendix A for the detailed proof.  $\square$

Since our purpose is to find a lower bound concave function for  $\check{R}_{1m}[n]$  in (28), we use the property that the first-order Taylor expansion of a convex function at any point is always a lower bound of that convex function. By using Lemma 1, a

$$\bar{R}_k = \sum_{n=0}^{N-1} \sum_{m=1}^M a_{m,k}[n] W \left( \underbrace{\log_2 \left( \sum_{i=1}^K P_i[n] \bar{H}_{m,i}[n] + \sigma_n^2 \right)}_{\triangleq \check{R}_{1m}[n]} - \underbrace{\log_2 \left( \sum_{i=1, i \neq k}^K P_i[n] \bar{H}_{m,i}[n] + \sigma_n^2 \right)}_{\triangleq \check{R}_{2m,k}[n]} \right) \quad (28)$$

theorem for the first-order Taylor expansion of  $\check{R}_{1m}[n]$  is then provided as follows.

**Theorem 1.** *Given any  $\mathbf{q}_m[n] = \mathbf{q}_m^r[n]$ , the first-order Taylor expansion of  $\check{R}_{1m}[n]$  can be derived and served as a lower bound:*

$$\begin{aligned} \check{R}_{1m}[n] &\triangleq \log_2 \left( \sum_{i=1}^K P_i[n] \bar{H}_{m,i}[n] + \sigma_n^2 \right) \\ &\geq \log_2 \left( \sum_{i=1}^K P_i[n] \left( \frac{C_1 C_2}{X_{m,i}^r[n] Y_{m,i}^r[n]} + \frac{C_3}{X_{m,i}^r[n]} \right) + \sigma_n^2 \right) \\ &\quad + \sum_{i=1}^K O_{m,i}[n] (X_{m,i}[n] - X_{m,i}^r[n]) \\ &\quad + \sum_{i=1}^K G_{m,i}[n] (Y_{m,i}[n] - Y_{m,i}^r[n]) \\ &\triangleq \check{R}_{1m}^{lb}[n], \forall m \in \{1, \dots, M\}, \forall n \in \{0, \dots, N-1\}, \end{aligned} \quad (31)$$

where we define

$$X_{m,i}^r[n] = \|\mathbf{q}_m^r[n] - \mathbf{g}_i\|^2 + H^2; \quad (32)$$

$$Y_{m,i}^r[n] = 1 + Ae^{-B \left( \frac{180}{\pi} \tan^{-1} \left( \frac{H}{\|\mathbf{q}_m^r[n] - \mathbf{g}_i\|_2} \right) - A \right)}; \quad (33)$$

$$O_{m,i}[n] = \frac{-P_i[n] \left( \frac{C_1 C_2 + C_3 Y_{m,i}^r[n]}{(X_{m,i}^r[n])^2 Y_{m,i}^r[n]} \right)}{\left( \sum_{j=1}^K P_j[n] \left( \frac{C_1 C_2}{Y_{m,i}^r[n] X_{m,i}^r[n]} + C_3 \right) + \sigma_n^2 \right) \ln(2)}; \quad (34)$$

$$G_{m,i}[n] = \frac{-P_i[n] \left( \frac{C_1 C_2}{X_{m,i}^r[n] (Y_{m,i}^r[n])^2} \right)}{\left( \sum_{j=1}^K P_j[n] \left( \frac{C_1 C_2}{Y_{m,i}^r[n] X_{m,i}^r[n]} + C_3 \right) + \sigma_n^2 \right) \ln(2)}. \quad (35)$$

*Proof.* See Appendix B for the detailed proof.  $\square$

It can be observed from Theorem 1 that the term  $O_{m,i}[n] (X_{m,i}[n] - X_{m,i}^r[n])$  is concave in  $\mathbf{q}_m[n]$ , since  $O_{m,i}[n] \leq 0$  and  $X_{m,i}[n] = \|\mathbf{q}_m[n] - \mathbf{g}_i\|_2^2 + H^2$  is a square norm function of  $\mathbf{q}_m[n]$ . However, the term  $G_{m,i}[n] (Y_{m,i}[n] - Y_{m,i}^r[n])$  is neither concave nor convex, and we further provide the following theorem to transform this term into a concave function.

**Theorem 2.** *Given any  $\mathbf{q}_m[n] = \mathbf{q}_m^r[n]$ ,  $Y_{m,i}[n]$  can be upper bounded by*

$$\begin{aligned} Y_{m,i}[n] &\leq 1 + Ae^{-B \left( \frac{180}{\pi} \left( \tan^{-1} \left( \frac{H}{\sqrt{U_{m,i}^r[n]}} \right) - \frac{H (\|\mathbf{q}_m[n] - \mathbf{g}_i\|_2^2 - U_{m,i}^r[n])}{2\sqrt{U_{m,i}^r[n]} (H^2 + U_{m,i}^r[n])} \right) - A \right)} \\ &\triangleq Y_{m,i}^{up}[n], \forall m \in \{1, \dots, M\}, \forall i \in \{1, \dots, K\}, \\ &\quad \forall n \in \{0, \dots, N-1\}, \end{aligned} \quad (36)$$

where  $U_{m,i}^r[n] = \|\mathbf{q}_m^r[n] - \mathbf{g}_i\|_2^2$  and  $Y_{m,i}^{up}[n]$  is a convex function in terms of  $\mathbf{q}_m[n]$ .

*Proof.* See Appendix C for the detailed proof.  $\square$

By applying Theorem 2, we replace  $Y_{m,i}[n]$  in  $\check{R}_{1m}^{lb}[n]$  by  $Y_{m,i}^{up}[n]$ . Since  $G_{m,i}[n] \leq 0$ , the term  $G_{m,i}[n] (Y_{m,i}^{up}[n] - Y_{m,i}^r[n])$  is a concave lower bound in  $\mathbf{q}_m[n]$ . Hence, the concave lower bound of  $\check{R}_{1m}^{lb}[n]$  in (31) can be derived as follows:

$$\begin{aligned} \check{R}_{1m}^{lb}[n] &\geq \log_2 \left( \sum_{i=1}^K P_i[n] \left( \frac{C_1 C_2}{X_{m,i}^r[n] Y_{m,i}^r[n]} + \frac{C_3}{X_{m,i}^r[n]} \right) + \sigma_n^2 \right) \\ &\quad + \sum_{i=1}^K O_{m,i}[n] (X_{m,i}[n] - X_{m,i}^r[n]) \\ &\quad + \sum_{i=1}^K G_{m,i}[n] (Y_{m,i}^{up}[n] - Y_{m,i}^r[n]) \\ &\triangleq \check{R}_{1m}^{llb}[n], \forall m \in \{1, \dots, M\}, \forall n \in \{0, \dots, N-1\}. \end{aligned} \quad (37)$$

2) *A concave lower bound for  $\check{R}_{2m,k}[n]$ :* In order to find a concave lower bound for  $\check{R}_{2m,k}[n]$ , we introduce two auxiliary variables  $\tilde{X}_{m,i}[n]$  and  $\tilde{Y}_{m,i}[n]$ , which satisfy the following constraints:

$$\begin{aligned} \tilde{X}_{m,i}[n] &\leq \|\mathbf{q}_m[n] - \mathbf{g}_i\|_2^2 + H^2, \forall m \in \{1, \dots, M\}, \\ &\quad \forall i \in \{1, \dots, K\}, \forall n \in \{0, \dots, N-1\}. \end{aligned} \quad (38)$$

$$\begin{aligned} \tilde{Y}_{m,i}[n] &\leq 1 + Ae^{-B \left( \frac{180}{\pi} \tan^{-1} \left( \frac{H}{\|\mathbf{q}_m[n] - \mathbf{g}_i\|_2} \right) - A \right)}, \\ &\quad \forall m \in \{1, \dots, M\}, \forall i \in \{1, \dots, K\}, \forall n \in \{0, \dots, N-1\}. \end{aligned} \quad (39)$$

Thus, an upper bound  $\bar{H}_{m,i}^{up}[n]$  for  $\bar{H}_{m,i}[n]$  in (28) can be derived as

$$\begin{aligned}\bar{H}_{m,i}[n] &\leq C_1 \times C_2 \times \frac{1}{\tilde{X}_{m,i}[n]} \times \frac{1}{\tilde{Y}_{m,i}[n]} + C_3 \times \frac{1}{\tilde{X}_{m,i}[n]} \\ &\triangleq \bar{H}_{m,i}^{up}[n], \forall m \in \{1, \dots, M\}, \forall i \in \{1, \dots, K\}, \\ &\quad \forall n \in \{0, \dots, N-1\},\end{aligned}\quad (40)$$

where  $C_1$ ,  $C_2$  and  $C_3$  are constants defined in (20), (21), and (22), respectively.

**Lemma 2.** *The upper bound  $\bar{H}_{m,i}^{up}[n]$  is a convex function for the two auxiliary variables  $\tilde{X}_{m,i}[n]$  and  $\tilde{Y}_{m,i}[n]$ ,  $\forall m \in \{1, \dots, M\}$ ,  $\forall i \in \{1, \dots, K\}$ ,  $\forall n \in \{0, \dots, N-1\}$ .*

*Proof.* According to the proof in Lemma 1, we know that the function  $\ln(\frac{C_1 C_2}{xy} + \frac{C_3}{x})$  is convex in  $(x, y)$  if  $C_1, C_2, C_3 > 0$ . By using the fact that  $e^{g(x)}$  is convex if  $g(x)$  is convex [27], it can be shown that  $\frac{C_1 C_2}{xy} + \frac{C_3}{x} = e^{\ln(\frac{C_1 C_2}{xy} + \frac{C_3}{x})}$  is also convex. Hence, the proof is completed.  $\square$

**Theorem 3.** *With (38) and (39), the function  $\check{R}_{2m,k}[n] = -\log_2(\sum_{i=1, i \neq k}^K P_i[n] \bar{H}_{m,i}[n] + \sigma_n^2)$  can be lower bounded by*

$$\begin{aligned}\check{R}_{2m,k}[n] &\geq -\log_2 \left( \sum_{i=1, i \neq k}^K P_i[n] \bar{H}_{m,i}^{up}[n] + \sigma_n^2 \right) \\ &\triangleq \check{R}_{2m,k}^{lb}[n], \forall m \in \{1, \dots, M\}, \forall n \in \{0, \dots, N-1\},\end{aligned}\quad (41)$$

where  $\check{R}_{2m,k}^{lb}[n]$  is a concave function in terms of  $\tilde{X}_{m,i}[n]$  and  $\tilde{Y}_{m,i}[n]$ .

*Proof.* With (38) and (39), we have  $\bar{H}_{m,i}[n] \leq \bar{H}_{m,i}^{up}[n]$  in (40). By replacing  $\bar{H}_{m,i}[n]$  in  $\check{R}_{2m,k}[n]$  with the upper bound  $\bar{H}_{m,i}^{up}[n]$ , it then yields  $\check{R}_{2m,k}[n] \geq -\log_2(\sum_{i=1, i \neq k}^K P_i[n] \bar{H}_{m,i}^{up}[n] + \sigma_n^2)$ . Moreover, it can be shown that  $\check{R}_{2m,k}^{lb}[n]$  is a concave function in terms of  $\tilde{X}_{m,i}[n]$  and  $\tilde{Y}_{m,i}[n]$  by applying the similar proof in Lemma 1.  $\square$

3) *Transformation of subproblem (P4) into a convex problem:* By adopting the derived lower bounds in Theorem 1, Theorem 3 and (37), a concave lower bound for  $\bar{R}_k$  can be computed as

$$\bar{R}_k \geq \sum_{n=0}^{N-1} \sum_{m=1}^M a_{m,k}[n] W(\check{R}_{1m}^{lb}[n] + \check{R}_{2m,k}^{lb}[n]). \quad (42)$$

We then transform the subproblem (P4) into a convex one by replacing  $\bar{R}_k$  in the constraint (27) with its lower bound (42) and inserting the two imposed constraints (38) and (39):

$$\begin{aligned}(\text{P5}) \quad &\max_{\{\zeta_q, \mathbf{q}_m[n], \tilde{X}_{m,i}[n], \tilde{Y}_{m,i}[n], \forall m, i, n\}} \zeta_q \\ \text{s.t.} \quad &\sum_{n=0}^{N-1} \sum_{m=1}^M a_{m,k}[n] W(\check{R}_{1m}^{lb}[n] + \check{R}_{2m,k}^{lb}[n]) \geq \zeta_q, \\ &\quad \forall k \in \{1, \dots, K\}, \\ &(4), (5), (6), (14), (38), (39).\end{aligned}\quad (43)$$

In (P5), the constraint (43) now becomes convex, since the left-hand side of (43) is a concave function in terms of the variables  $\mathbf{q}_m[n]$ ,  $\tilde{X}_{m,i}[n]$  and  $\tilde{Y}_{m,i}[n]$  according to (42). The constraints (38) and (39), which are due to the introduction of the auxiliary variables  $\tilde{X}_{m,i}[n]$  and  $\tilde{Y}_{m,i}[n]$  in (40), are also included in (P5) for guaranteeing the lower bound relationship of (42) and thus the legitimacy of the constraint (43). Due to the lower bound relationship in (42), one can also conclude that the feasible set of the subproblem (P5) is a subset of the feasible set of the original UAV flight trajectory subproblem (P4).

However, the subproblem (P5) is still not a convex optimization problem, since the constraints (6), (38) and (39) are non-convex. We in turn apply the first-order Taylor expansion to deal with these constraints and transform them into convex ones. To make the constraint (6) easier to handle, we first square the both sides of the constraint (6):

$$\begin{aligned}\|\mathbf{q}_m[n] - \mathbf{q}_j[n]\|_2^2 &\geq D_{min}^2, \\ \forall n \in \{1, \dots, N-1\}, \forall m, j \in \{1, \dots, M\}, m \neq j.\end{aligned}\quad (44)$$

Since  $\|\mathbf{q}_m[n] - \mathbf{q}_j[n]\|_2^2$  is a convex function with respect to  $\mathbf{q}_m[n]$  and  $\mathbf{q}_j[n]$ , the first-order Taylor expansion of  $\|\mathbf{q}_m[n] - \mathbf{q}_j[n]\|_2^2$  at given points  $\mathbf{q}_m^r[n]$  and  $\mathbf{q}_j^r[n]$  can be derived and served as a lower bound:

$$\begin{aligned}\|\mathbf{q}_m[n] - \mathbf{q}_j[n]\|_2^2 &\geq \|\mathbf{q}_m^r[n] - \mathbf{q}_j^r[n]\|_2^2 \\ &\quad + 2(\mathbf{q}_m^r[n] - \mathbf{q}_j^r[n])^T (\mathbf{q}_m[n] - \mathbf{q}_m^r[n]) \\ &\quad - 2(\mathbf{q}_m^r[n] - \mathbf{q}_j^r[n])^T (\mathbf{q}_j[n] - \mathbf{q}_j^r[n]).\end{aligned}\quad (45)$$

From (45), the constraint (44) is then replaced by the following new constraint:

$$\begin{aligned}D_{min}^2 &\leq \|\mathbf{q}_m^r[n] - \mathbf{q}_j^r[n]\|_2^2 \\ &\quad + 2(\mathbf{q}_m^r[n] - \mathbf{q}_j^r[n])^T (\mathbf{q}_m[n] - \mathbf{q}_m^r[n]) \\ &\quad - 2(\mathbf{q}_m^r[n] - \mathbf{q}_j^r[n])^T (\mathbf{q}_j[n] - \mathbf{q}_j^r[n]), \\ \forall n \in \{1, \dots, N-1\}, \forall m, j \in \{1, \dots, M\}, m \neq j.\end{aligned}\quad (46)$$

Note that the solutions that satisfy the constraint (46) must also satisfy the constraint (6) due to the lower bound relationship in (45).

Similarly, the function  $\|\mathbf{q}_m[n] - \mathbf{g}_i\|_2^2 + H^2$  in the constraint (38) is a convex function with respect to the trajectory variable  $\mathbf{q}_m[n]$ , and its lower bound is obtained by the first-order Taylor expansion at a given point  $\mathbf{q}_m^r[n]$ , as follows:

$$\begin{aligned}\|\mathbf{q}_m[n] - \mathbf{g}_i\|_2^2 + H^2 &\geq \|\mathbf{q}_m^r[n] - \mathbf{g}_i\|_2^2 + H^2 \\ &\quad + 2(\mathbf{q}_m^r[n] - \mathbf{g}_i)^T (\mathbf{q}_m[n] - \mathbf{q}_m^r[n]).\end{aligned}\quad (47)$$

By using (47), the constraint (38) is then replaced by the following new constraint:

$$\begin{aligned}\tilde{X}_{m,i}[n] &\leq \|\mathbf{q}_m^r[n] - \mathbf{g}_i\|_2^2 + H^2 \\ &\quad + 2(\mathbf{q}_m^r[n] - \mathbf{g}_i)^T (\mathbf{q}_m[n] - \mathbf{q}_m^r[n]), \forall n \in \{1, \dots, N-1\}, \\ &\quad \forall m \in \{1, \dots, M\}, \forall i \in \{1, \dots, K\}.\end{aligned}\quad (48)$$

The next step is to convexify the constraint (39) because the right-hand side of (39) is neither convex nor concave with



respect to the trajectory variable  $\mathbf{q}_m[n]$ . To achieve this, we first define an angle elevation variable  $\theta_{m,i}[n]$  which satisfies the following imposed constraint:

$$\theta_{m,i}[n] \geq \frac{180}{\pi} \tan^{-1} \left( \frac{H}{\|\mathbf{q}_m[n] - \mathbf{g}_i\|_2} \right). \quad (49)$$

This implies that

$$Ae^{-B(\theta_{m,i}[n]-A)} \leq Ae^{-B\left(\frac{180}{\pi} \tan^{-1} \left( \frac{H}{\|\mathbf{q}_m[n] - \mathbf{g}_i\|_2} \right) - A\right)}, \quad (50)$$

where  $A$  and  $B$  are the non-negative coefficients defined in (9). In addition, it is worth mentioning that the function  $A \exp(-B(\theta_{m,i}[n] - A))$  is convex with respect to  $\theta_{m,i}[n]$ , and its first-order Taylor expansion at the point  $\theta_{m,i}^r[n] = \frac{180}{\pi} \tan^{-1} \left( \frac{H}{\|\mathbf{q}_m^r[n] - \mathbf{g}_i\|_2} \right)$  can be derived as

$$Ae^{-B(\theta_{m,i}[n]-A)} \geq Ae^{-B(\theta_{m,i}^r[n]-A)} + \left( -ABe^{B(A-\theta_{m,i}^r[n])} \right) (\theta_{m,i}[n] - \theta_{m,i}^r[n]). \quad (51)$$

Applying the first-order expression of (51) into (39) yields a convex constraint:

$$\begin{aligned} \tilde{Y}_{m,i}[n] &\leq 1 + Ae^{-B(\theta_{m,i}^r[n]-A)} + \left( -ABe^{B(A-\theta_{m,i}^r[n])} \right) \\ &\times (\theta_{m,i}[n] - \theta_{m,i}^r[n]), \forall n \in \{1, \dots, N-1\}, \\ &\forall m \in \{1, \dots, M\}, \forall i \in \{1, \dots, K\}. \end{aligned} \quad (52)$$

This new convex constraint (52) ensures the satisfaction of the constraint (39) according to the lower bound relationship in (49) and (51).

However, the angle elevation angle constraint (49) is non-convex in terms of the flight trajectory variable  $\mathbf{q}_m[n]$ . In the following, we attempt to convexify this non-convex constraint. According to the fact that  $\tan^{-1} \left( \frac{1}{\sqrt{x}} \right)$  is a convex function when  $x > 0$ , we first define a variable  $\tilde{U}_{m,i}[n]$  which is enforced to satisfy the following constraint:

$$\tilde{U}_{m,i}[n] \leq \|\mathbf{q}_m[n] - \mathbf{g}_i\|_2^2. \quad (53)$$

Thus, we can derive

$$\frac{180}{\pi} \tan^{-1} \left( \frac{H}{\sqrt{\tilde{U}_{m,i}[n]}} \right) \geq \frac{180}{\pi} \tan^{-1} \left( \frac{H}{\|\mathbf{q}_m[n] - \mathbf{g}_i\|_2} \right). \quad (54)$$

By using (54), the angle elevation angle constraint (49) can be replaced by an upper bound constraint:

$$\begin{aligned} \theta_{m,i}[n] &\geq \frac{180}{\pi} \tan^{-1} \left( \frac{H}{\sqrt{\tilde{U}_{m,i}[n]}} \right), \forall n \in \{1, \dots, N-1\}, \\ &\forall m \in \{1, \dots, M\}, \forall i \in \{1, \dots, K\}. \end{aligned} \quad (55)$$

Moreover, the constraint (53) is non-convex in terms of  $\mathbf{q}_m[n]$ , and we again apply the first-order Taylor expansion to find a lower bound for  $\|\mathbf{q}_m[n] - \mathbf{g}_i\|_2^2$ :

$$\begin{aligned} \|\mathbf{q}_m[n] - \mathbf{g}_i\|_2^2 &\geq \|\mathbf{q}_m^r[n] - \mathbf{g}_i\|_2^2 \\ &+ 2(\mathbf{q}_m^r[n] - \mathbf{g}_i)^T (\mathbf{q}_m[n] - \mathbf{q}_m^r[n]), \end{aligned} \quad (56)$$

where  $\mathbf{q}_m^r[n]$  is a given reference point. Hence, the constraint (53) can be convexified through the lower bound relationship in (56), given by

$$\begin{aligned} \tilde{U}_{m,i}[n] &\leq \|\mathbf{q}_m^r[n] - \mathbf{g}_i\|_2^2 + 2(\mathbf{q}_m^r[n] - \mathbf{g}_i)^T (\mathbf{q}_m[n] - \mathbf{q}_m^r[n]), \\ &\forall n \in \{1, \dots, N-1\}, \forall m \in \{1, \dots, M\}, \forall i \in \{1, \dots, K\}, \end{aligned} \quad (57)$$

where the solutions that satisfy this constraint are also in the feasible set of the constraint (53).

In summary, the convex constraints (46) and (48) convexify the constraints (6) and (38) in the subproblem (P5), respectively, while the convex constraints (52), (55) and (57) are able to convexify the constraint (39). Accordingly, the UAV flight trajectory subproblem (P5) can be rewritten as

$$\begin{aligned} \text{(P6)} \quad &\max_{\{\zeta_q, \mathbf{q}_m[n], \tilde{X}_{m,i}[n], \tilde{Y}_{m,i}[n], \theta_{m,i}[n], U_{m,i}[n], \forall m, k, n\}} \zeta_q \\ \text{s.t.} \quad &(4), (5), (14), (43), (46), (48), (52), (55), (57). \end{aligned}$$

The UAV flight trajectory subproblem (P6) now becomes a solvable convex optimization problem under a given reference point  $\mathbf{q}_m^r[n]$ , and the optimal value of the UAV trajectory can be iteratively obtained by using the existing solution tool (CVX) [26] with SCA methods [28]. Note that the obtained optimal solution of the subproblem (P6) is a lower bound of the subproblem (P4), since the feasible set of (P4) contains that of (P6).

### C. Transmit Power Control Subproblem

Given the communication association  $a_{m,k}[n]$  and the UAV flight trajectory  $\mathbf{q}_m[n]$ , the transmit power control subproblem for WNs can be obtained from the original optimization problem (P2) by using the epigraph form and introducing an auxiliary variable  $\zeta_p$ , as follows:

$$\begin{aligned} \text{(P7)} \quad &\max_{\{\zeta_p, P_k[n], \forall k, n\}} \zeta_p \\ \text{s.t.} \quad &\bar{R}_k \geq \zeta_p, \forall k \in \{1, \dots, K\}, \\ &(17), (18). \end{aligned} \quad (58)$$

In this subproblem, the objective function and the constraints (17) and (18) are affine and thus convex in  $P_k[n]$ . However, from (28), it can be found that the user rate  $\bar{R}_k$  is the sum of concave functions and convex functions, and thus  $\bar{R}_k$  in the constraint (58) is neither convex nor concave in terms of the variables  $P_i[n]$ . To solve this problem, we convexify the constraint (58) by the first-order Taylor expansion. From (28), it is known that  $\tilde{R}_{2m,k}[n]$  is a convex function in  $P_i[n]$ , and we can get the following lower bound relationship:

$$\begin{aligned} \tilde{R}_{2m,k}[n] &\geq -\log_2 \left( \sum_{i=1, i \neq k}^K P_i^r[n] \bar{H}_{m,i}[n] + \sigma_n^2 \right) \\ &- \sum_{i=1, i \neq k}^K \frac{\bar{H}_{m,i}[n]}{\ln(2) \left( \sum_{j=1, j \neq k}^K P_j^r[n] \bar{H}_{m,j}[n] + \sigma_n^2 \right)} \\ &\times (P_i[n] - P_i^r[n]) \triangleq \tilde{R}_{m,k}^{lb}[n], \forall n \in \{1, \dots, N-1\}, \\ &\forall m \in \{1, \dots, M\}, \forall k \in \{1, \dots, K\}, \end{aligned} \quad (59)$$

where  $P_i^r[n]$  is a reference point for the Taylor expansion. Accordingly, the user rate  $\bar{R}_k$  in (28), i.e., the left-hand side of the constraint (58), can be lower bounded by a concave function, as follows:

$$\bar{R}_k \geq \sum_{n=0}^{N-1} \sum_{m=1}^M a_{m,k}[n] W \left( \check{R}_{1m,k}[n] + \dot{R}_{m,k}^{lb}[n] \right), \quad \forall k \in \{1, \dots, K\}. \quad (60)$$

Thus, the subproblem (P7) can be transformed into a convex one by replacing the constraint (58) with the lower bound (60):

$$\begin{aligned} \text{(P8)} \quad & \max_{\{\zeta_p, P_k[n], \forall k, n\}} \zeta_p \\ \text{s.t.} \quad & \sum_{n=0}^{N-1} \sum_{m=1}^M a_{m,k}[n] W \left( \check{R}_{1m,k}[n] + \dot{R}_{m,k}^{lb}[n] \right) \geq \zeta_p, \\ & \forall k \in \{1, \dots, K\}, \end{aligned} \quad (61)$$

(17), (18).

Note that the optimization result of (P8) is a lower bound of (P7). Given a reference point  $P_i^r[n]$ , the transmit power control subproblem (P8) can be iteratively solved by the optimization tool, e.g., CVX, with SCA methods.

After the above transformation, the offline joint design problem (P2) can be alternatively solved by the three convex subproblems (P3), (P6) and (P8), and the proposed offline algorithm is summarized in Algorithm 1. For convenience, we assume that the superscript  $r$  of the reference points  $\mathbf{q}_m^r[n]$ ,  $P_k^r[n]$ ,  $a_{m,k}^r[n]$  also refers to the iteration index in the successive convex optimization. We first initialize  $r = 0$ , and initialize  $\mathbf{q}_m^r[n]$ ,  $P_k^r[n]$ , and  $a_{m,k}^r[n]$ . Let  $\epsilon > 0$  be a threshold for the stopping criterion. In the communication association subproblem (P3), the optimal solution to the communication association  $a_{m,k}^*[n]$  is obtained under the given values of UAV trajectory  $\mathbf{q}_m^r[n]$  and power control  $P_k^r[n]$ , and then we update  $a_{m,k}^{r+1}[n] = a_{m,k}^*[n]$ . Afterwards, the UAV flight trajectory subproblem (P6) is solved under the given values of communication association  $a_{m,k}^{r+1}[n]$  and power control  $P_k^r[n]$  through the successive convex optimization. In the inner loop of the successive convex optimization, we update  $\mathbf{q}_m^r[n]$  with the last result of  $\mathbf{q}_m^*[n]$ , and subsequently renew the related constraints that contain  $\mathbf{q}_m^r[n]$ . Now the new optimal solution  $\mathbf{q}_m^*[n]$  can be obtained by solving the subproblem (P6). The steps in the loop are repeated until the convergence is achieved, and we update  $\mathbf{q}_m^{r+1}[n] = \mathbf{q}_m^*[n]$ . Finally, the power control subproblem (P8) is solved under the given values of UAV trajectory  $\mathbf{q}_m^{r+1}[n]$  and communication association  $a_{m,k}^{r+1}[n]$ , where we update  $P_k^r[n]$  and (61) with the last result of  $P_k^*[n]$  in the inner loop of the successive convex optimization. We repeat these steps in the inner loop until the convergence is attained, and update  $P_k^{r+1}[n] = P_k^*[n]$ . We then update the iteration number of the outer loop for the next round. The outer loop is stopped when the increase of the objective value is smaller than the preset threshold  $\epsilon$ .

---

**Algorithm 1** Offline Alternative Optimization Algorithm for Problem (P2)

---

- 1: Initialize  $\{a_{m,k}^r[n], \mathbf{q}_m^r[n], P_k^r[n], \forall m, k, n\}$ ,  $r = 0$
  - 2: Set  $\epsilon > 0$
  - 3: **while** Increase of the objective value  $< \epsilon$  **do**
  - 4:   For given  $\{\mathbf{q}_m^r[n], P_k^r[n], \forall m, k, n\}$ , find the optimal communication association solution of the problem (P3) as  $a_{m,k}^{r+1}[n]$ .
  - 5:   For given  $\{a_{m,k}^{r+1}[n], \mathbf{q}_m^r[n], P_k^r[n], \forall m, k, n\}$ .
  - 6:   **repeat** (SCA for solving problem (P6))
  - 7:     Update  $\mathbf{q}_m^r[n]$  using the last result  $\mathbf{q}_m^*[n]$ .
  - 8:     Update (43), (46), (48), (52) and (57) using  $\mathbf{q}_m^r[n]$ .
  - 9:     Find the new optimal solution  $\mathbf{q}_m^*[n]$  by solving the problem (P6).
  - 10:   **until** convergence to the optimal UAV flight trajectory solution  $\mathbf{q}_m^*[n]$
  - 11:   Set  $\mathbf{q}_m^{r+1}[n] \leftarrow \mathbf{q}_m^*[n]$ .
  - 12:   For given  $\{a_{m,k}^{r+1}[n], \mathbf{q}_m^{r+1}[n], P_k^r[n], \forall m, k, n\}$ .
  - 13:   **repeat** (SCA for solving problem (P8))
  - 14:     Update  $P_k^r[n]$  using the last result  $P_k^*[n]$ .
  - 15:     Update (61) using  $P_k^r[n]$ .
  - 16:     Find the new optimal solution  $P_k^*[n]$  by solving the problem (P8).
  - 17:   **until** convergence to the optimal power control strategy solution  $P_k^*[n]$
  - 18:   Set  $P_k^{r+1}[n] \leftarrow P_k^*[n]$ .
  - 19:   Update  $r \leftarrow r + 1$ .
  - 20: **end while**
- 

#### D. Convergence of the Offline Algorithm

The convergence of the proposed offline algorithm for UAV flight trajectory, WN communication association and power control is analyzed as follows. Let  $f(a_{m,k}[n], \mathbf{q}_m[n], P_k[n])$  be the objective function of the original problem (P2). First, in the communication association subproblem (P3), by fixing  $\{\mathbf{q}_m^r[n], P_k^r[n]\}$  to obtain the optimal solution of the communication association  $a_{m,k}^{r+1}[n]$ , it results in

$$f(a_{m,k}^r[n], \mathbf{q}_m^r[n], P_k^r[n]) \leq f(a_{m,k}^{r+1}[n], \mathbf{q}_m^r[n], P_k^r[n]). \quad (62)$$

Next we discuss the convergence of the inner loop of the UAV flight trajectory subproblem (P6). Given  $\{a_{m,k}^{r+1}[n], \mathbf{q}_m^r[n], P_k^r[n]\}$ , the subproblem is optimized by the successive convex optimization to obtain the trajectory  $\mathbf{q}_m^{r+1}[n]$ . In the  $i$ th inner loop of the successive convex optimization, we assume that  $\mathbf{q}_m^{(i)*}[n]$  is the obtained solution with respect to the given first-order Taylor expansion reference point  $\mathbf{q}_m^{(i)}[n]$ . Further, denote  $\zeta_{\mathbf{q}_m^{(i)}[n]}(a_{m,k}^{r+1}[n], \mathbf{q}_m^{(i)*}[n], P_k^r[n])$  as the worst sum rate for the obtained solution  $\mathbf{q}_m^{(i)*}[n]$  at the reference point  $\mathbf{q}_m^{(i)}[n]$  in

the subproblem (P6). Then we have

$$\begin{aligned} & \zeta_{\mathbf{q}_m^{(i)}[n]} \left( a_{m,k}^{r+1}[n], \mathbf{q}_m^{(i)*}[n], P_k^r[n] \right) \\ & \leq f \left( a_{m,k}^{r+1}[n], \mathbf{q}_m^{(i)*}[n], P_k^r[n] \right) \\ & = \zeta_{\mathbf{q}_m^{(i)*}[n]} \left( a_{m,k}^{r+1}[n], \mathbf{q}_m^{(i)*}[n], P_k^r[n] \right) \\ & \leq \zeta_{\mathbf{q}_m^{(i+1)}[n]} \left( a_{m,k}^{r+1}[n], \mathbf{q}_m^{(i+1)*}[n], P_k^r[n] \right), \end{aligned} \quad (63)$$

where the first inequality comes from the lower bound relationship in (42), and the second equality is because the equality of the lower bound relationship holds at the tangent point. Moreover, the third inequality is due to the fact that we set  $\mathbf{q}_m^{(i+1)}[n] = \mathbf{q}_m^{(i)*}[n]$  to update the reference point for the next inner iteration according to the SCA and obtain the corresponding optimal solution  $\mathbf{q}_m^{(i+1)*}[n]$ . Hence it concludes that the worst sum rate performance can be monotonically increased in the inner loop of the UAV flight trajectory subproblem (P6), and we can get the following relationship for UAV flight trajectory in the  $r$ th outer loop:

$$f \left( a_{m,k}^{r+1}[n], \mathbf{q}_m^r[n], P_k^r[n] \right) \leq f \left( a_{m,k}^{r+1}[n], \mathbf{q}_m^{r+1}[n], P_k^r[n] \right), \quad (64)$$

Likewise, we can apply the similar derivation of (63) to prove that the sum rate performance can be monotonically increased in the inner loop of the SCA for the power control subproblem (P8). Thus, under the given  $\{a_{m,k}^{r+1}[n], \mathbf{q}_m^{r+1}[n], P_k^r[n]\}$ , it implies that

$$\begin{aligned} & \zeta_{P_k^{(i)}[n]} \left( a_{m,k}^{r+1}[n], \mathbf{q}_m^{r+1}[n], P_k^{(i)*}[n] \right) \\ & \leq \zeta_{P_k^{(i+1)}[n]} \left( a_{m,k}^{r+1}[n], \mathbf{q}_m^{r+1}[n], P_k^{(i+1)*}[n] \right), \end{aligned} \quad (65)$$

where  $\zeta_{P_k^{(i)}[n]} \left( a_{m,k}^{r+1}[n], \mathbf{q}_m^{r+1}[n], P_k^{(i)*}[n] \right)$  is referred to as the worst sum rate for the obtained solution  $P_k^{(i)*}[n]$  at the reference point  $P_k^{(i)}[n]$  in the  $i$ th inner loop of the power control subproblem (P8). As a result, it implies that

$$\begin{aligned} & f \left( a_{m,k}^{r+1}[n], \mathbf{q}_m^{r+1}[n], P_k^r[n] \right) \leq \\ & f \left( a_{m,k}^{r+1}[n], \mathbf{q}_m^{r+1}[n], P_k^{r+1}[n] \right). \end{aligned} \quad (66)$$

Due to the alternative optimization for the three subproblems, it can be concluded from (62), (64) and (66) that the performance of the proposed algorithm can be monotonically increased and the local optimal solution of the original problem (P2) can be found until the algorithm is converged.

#### IV. CONVEX-ASSISTED REINFORCEMENT LEARNING (CARL)

In this section, we propose a CARL approach, in which a flight corridor is marked out, based on the proposed offline design in Algorithm 1, to guide the UAV flight actions in an online learning fashion. Besides, the actions of the RL agents can be restricted and the number of system states can be efficiently reduced through the assistance of offline design. The system states, actions and real-time rewards are designed and presented in the following.

##### A. System States

Let  $\mathbf{S} = \mathbf{L} \times \mathbf{H}$  be a two-tuple state space, where  $\times$  denotes the Cartesian product,  $\mathbf{L}$  represents a UAV location state set, and  $\mathbf{H}$  is a channel state set. Moreover, we define a random variable  $\mathbf{s} = (\mathbf{l}, \mathbf{h}) \in \mathbf{S}$  as the system stochastic state of the Markov decision process. It is assumed that the UAV location and channel remain steady during the time interval  $\delta_D$ . The detailed definition of each state is specified in the following.

• **UAV location state:** Assume that a square UAV coverage region is quantized into  $N_L$  lattice points with a scale of  $\Delta$  (the minimum distance between two adjacent horizontal/vertical lattice points), and the UAV location state space is defined as  $\mathbf{L} = \mathbf{L}_x \times \mathbf{L}_y$ , where  $\mathbf{L}_x = \mathbf{L}_y = \{0, 1, \dots, \sqrt{N_L} - 1\}$ . When the location state of the  $m$ th UAV at the time instant  $n$  is given as  $\mathbf{l}_m^n = [\tilde{x}_m^n, \tilde{y}_m^n] \in \mathbf{L}$ , it means that the horizontal coordinate of the  $m$ th UAV at the time instant  $n$  is

$$\mathbf{q}_m[n] = \left[ \Delta \times \tilde{x}_m^n + \frac{\Delta}{2}, \Delta \times \tilde{y}_m^n + \frac{\Delta}{2} \right]^T. \quad (67)$$

Let  $\mathbf{q}_m^*[n]$  be the offline horizontal flight path of the  $m$ th UAV at the time instant  $n$ , obtained by the offline convex optimization. By using the offline UAV flight trajectory  $\mathbf{q}_m^*[n]$ , the offline trajectory-assisted location state of  $m$ th UAV at the time instant  $n$  is given as

$$\mathbf{l}_m^n \in \hat{\mathbf{L}}_m = \{\mathbf{l}_m^n \in \mathbf{L} \mid \|\mathbf{q}_m[n] - \mathbf{q}_m^*[n]\|_2 \leq D_F, \forall n\}, \quad (68)$$

where  $D_F$  is the flight corridor width, and the distance between the real UAV location  $\mathbf{q}_m[n]$  and the offline UAV trajectory  $\mathbf{q}_m^*[n]$  at the time instant  $n$  is smaller than the preset corridor width  $D_F$ . Then the location state of all UAVs at the time instant  $n$  can be expressed as

$$\mathbf{l}^n = [\mathbf{l}_1^n, \dots, \mathbf{l}_M^n] \in \hat{\mathbf{L}}_1 \times \hat{\mathbf{L}}_2 \times \dots \times \hat{\mathbf{L}}_M. \quad (69)$$

• **Channel state:** With the assistance of the offline UAV flight trajectory results, the average LOS/NLOS channel strength  $\bar{H}_{m,k}[n]$  between the  $k$ th WN and the  $m$ th UAV along the flight path at each time can be calculated by (19) to simplify the quantization of channel states. Let  $\epsilon_H > 0$  be a threshold value for channel quantization. Considering the fact that the channel strength in the vicinity of the flight corridor is related to the channel strength of the offline flight path, the channel state of the  $m$ th UAV can be defined as:

$$h_{m,k}^n = \begin{cases} 0, & H_{m,k}[n] < \bar{H}_{m,k}[n] - \epsilon_H; \\ 1, & \bar{H}_{m,k}[n] - \epsilon_H \leq H_{m,k}[n] \leq \bar{H}_{m,k}[n] + \epsilon_H; \\ 2, & H_{m,k}[n] > \bar{H}_{m,k}[n] + \epsilon_H. \end{cases} \quad (70)$$

By using this relative quantization method, the changes in channel strength can be better described than the direct quantization of channel strength, and the number of quantization levels can be greatly reduced. As such, the channel state of all UAVs and nodes at the time instant  $n$  is defined as:

$$\mathbf{h}^n = [h_{1,1}^n, \dots, h_{1,K}^n, \dots, h_{M,1}^n, \dots, h_{M,K}^n] \in \hat{\mathbf{H}}^{M \times K}, \quad (71)$$

where  $\hat{\mathbf{H}} = \{0, 1, 2\}$ .

### B. System Actions

Based on the states of UAV locations and channels, we can decide the UAV flight directions, the association between UAVs and WNs, and uplink transmit power. Define  $\mathbf{A} = \mathbf{A}_F \times \mathbf{A}_C$  as the action space, where  $\mathbf{A}_F = \{0, 1, 2, 3, 4\}$  represents the set of five flight direction actions,  $\mathbf{A}_C = \{0, 1, \dots, N_p K\}$  is the set of communication (including association and transmission) actions, and  $N_p$  is the number of transmit power levels available for each WN. Denote  $a_{m,F}^n$  and  $a_{m,C}^n$  as the UAV flight direction action and the communication action of the  $m$ th UAV at the time instant  $n$ , respectively. Hence, the concatenated action of all UAVs that can be chosen at the time instant  $n$  is given as  $\mathbf{a}^n = [\mathbf{a}_F^n, \mathbf{a}_C^n]$ , where  $\mathbf{a}_F^n = [a_{1,F}^n, \dots, a_{M,F}^n]$  and  $\mathbf{a}_C^n = [a_{1,C}^n, \dots, a_{M,C}^n]$ . The details of the actions are specified below. For the flight direction action  $\mathbf{a}_F^n \in \mathbf{A}_F^M$ , the action  $a_{m,F}^n = 0$  means that the  $m$ th UAV is hovering at the time instant  $n$ , while the action  $a_{m,F}^n = 1, 2, 3, 4$  means that the  $m$ th UAV moves to the left, right, forward or backward with a predefined distance  $\Delta$ , respectively. By flight corridor restrictions, for a given system state  $\mathbf{s}^n = [\mathbf{l}^n, \mathbf{h}^n]$  at the time instant  $n$ , the UAV flight action is partially constrained by  $\mathbf{a}_F^n \in \hat{\mathbf{A}}_F^n = \{\mathbf{a}_F^n \in \mathbf{A}_F^M \mid \|\mathbf{q}_m[n+1] - \mathbf{q}_m^*[n+1]\|_2 \leq D_F, \forall m, \text{ and } \mathbf{q}_m[n+1] \neq \mathbf{q}_j[n+1], \forall m \neq j\}$ . In this setting, the actions that can ensure that the UAVs' position at the next time being within the range of the flight path and avoid the collision among the multiple UAVs are regarded as legal actions.

Let  $\bar{\mathbf{A}}_C^n$  be the affordable communication action set that satisfies the battery constraints and the communication association constraints at the time instant  $n$ , i.e.,  $\mathbf{a}_C^n \in \bar{\mathbf{A}}_C^n \subseteq \mathbf{A}_C^M$ , and it can be constructed as follows. Assume that  $p_k^n$  and  $b_k^n$  are the transmit power level and battery power of the  $k$ th WN at the time instant  $n$ , respectively, and  $p_k^n \in \mathbf{P} = \{1, 2, \dots, N_p\}$ . For  $p_k^n$ , it means that the energy expenditure of the  $k$ th WN for uplink data transmission is  $p_k^n E_p$  at the time instant  $n$ , for  $k = 1, \dots, K$  and  $p_k^n = 1, \dots, N_p$ , where  $E_p$  is the basic energy unit with respect to the transmit power level  $p_k^n = 1$ . If the  $k$ th WN is associated with the  $m$ th UAV with the uplink transmit power level  $p_k^n$  at the time instant  $n$ , the action  $a_{m,C}^n = (k-1)N_p + p_k^n$  is performed. On the other hand, the action  $a_{m,C}^n = 0$  indicates that the  $m$ th UAV is not connected to any WNs at the time  $n$ . In addition, the communication action  $a_{m,C}^n = (k-1)N_p + p_k^n$  is constrained by the available battery power of the WNs and the UAV-WN association. For the battery power constraint, the action  $a_{m,C}^n = (k-1)N_p + p_k^n$  is eligible to be performed only if the energy expenditure  $p_k^n E_p$  is less than the battery power  $b_k^n$  of the  $k$ th WN, i.e.,  $p_k^n E_p \leq b_k^n$ . For the UAV-WN association, if  $a_{m,C}^n \neq 0$ , the actions of the other UAVs are limited by  $a_{j,C}^n \neq a_{m,C}^n$ , for any  $j \neq m$ . This is because each WN can be only served by one UAV during a time interval.

### C. State Transition

After all UAVs perform an action  $\mathbf{a}^n$  at the time  $n$ , the current system state  $\mathbf{s}^n$  is transited to the next system state  $\mathbf{s}^{n+1}$ , and the state transitions are elaborated as follows.

• **UAV location state:** Assume the current UAV location state is  $\mathbf{l}_m^n = [\tilde{x}_m^n, \tilde{y}_m^n]$ , and the action  $\mathbf{a}_m^n = [a_{m,F}^n, a_{m,C}^n]$  is performed. The next UAV location state becomes

$$\mathbf{l}_m^{n+1} = [\tilde{x}_m^{n+1}, \tilde{y}_m^{n+1}] = \begin{cases} [\tilde{x}_m^n, \tilde{y}_m^n], & a_{m,F}^n = 0; \\ [\tilde{x}_m^n - 1, \tilde{y}_m^n], & a_{m,F}^n = 1; \\ [\tilde{x}_m^n + 1, \tilde{y}_m^n], & a_{m,F}^n = 2; \\ [\tilde{x}_m^n, \tilde{y}_m^n + 1], & a_{m,F}^n = 3; \\ [\tilde{x}_m^n, \tilde{y}_m^n - 1], & a_{m,F}^n = 4. \end{cases} \quad (72)$$

From (3) and (67), the corresponding change of the  $m$ th UAV coordinates  $\mathbf{q}_m[n]$  with respect to the action  $\mathbf{a}_m^n = [a_{m,F}^n, a_{m,C}^n]$  is thus given by

$$\mathbf{q}_m[n+1] = \begin{cases} [x_m[n], y_m[n]]^T, & a_{m,F}^n = 0; \\ [x_m[n] - \Delta, y_m[n]]^T, & a_{m,F}^n = 1; \\ [x_m[n] + \Delta, y_m[n]]^T, & a_{m,F}^n = 2; \\ [x_m[n], y_m[n] + \Delta]^T, & a_{m,F}^n = 3; \\ [x_m[n], y_m[n] - \Delta]^T, & a_{m,F}^n = 4. \end{cases} \quad (73)$$

Note that if the UAV selects the communication action, its coordinates and location state remain unchanged, i.e.,  $a_{m,F}^n = 0$  if  $a_{m,C}^n \neq 0$ .

• **Channel state :** The channel state  $h_{m,k}^{n+1}$  between the  $m$ th UAV and the  $k$ th WN at the time instant  $(n+1)$  is decided by quantizing the observed channel gain  $H_{m,k}[n+1]$  according to (70), where the instantaneous channel  $H_{m,k}[n+1]$  depends on the positions of the UAV and the WN. Based on the channel model in (7)–(10), the large-scale channel components of  $H_{m,k}[n+1]$  and  $H_{m,k}[n]$  are correlated with each other.

### D. Reward Design

During the entire mission time  $n = 0, \dots, N-1$ , the system receives a negative reward  $C_P < 0$  as a penalty if the UAV flight action set  $\hat{\mathbf{A}}_F^n$  constrained by the offline UAV trajectory is empty, i.e., the next position of any UAV is beyond the flight corridor. On the contrary, if the UAV flight action set  $\hat{\mathbf{A}}_F^n$  is non-empty, we discuss the reward in two cases: 1)  $n = 0, \dots, N-2$  and 2)  $n = N-1$ . For  $n = 0, \dots, N-2$ , the system can get a reward  $\rho(\mathbf{s}^n, \mathbf{a}^n)$  at a state  $\mathbf{s}^n$  with respect to a performed action  $\mathbf{a}^n \in \hat{\mathbf{A}}_F^n$ . At  $n = N-1$ , all the UAVs are required to return to the starting point when the action is performed. Hence, the system gets a negative reward  $C_P < 0$  if any UAV does not comply with this constraint, i.e.,  $\mathbf{l}_m^N \neq \mathbf{l}_m^0$  for any  $m$ ; otherwise, the system can get a reward  $\rho(\mathbf{s}^n, \mathbf{a}^n)$  when all the UAVs can successfully flight back to the starting point. The system reward can be summarized as

$$R^n(\mathbf{s}^n, \mathbf{a}^n) = \begin{cases} C_P, n = \{0, \dots, N-1\}, \text{ and } \hat{\mathbf{A}}_F^n = \emptyset; \\ \rho(\mathbf{s}^n, \mathbf{a}^n), n = \{0, \dots, N-2\}, \text{ and } \hat{\mathbf{A}}_F^n \neq \emptyset; \\ C_P, n = N-1, \hat{\mathbf{A}}_F^n \neq \emptyset, \text{ and } \mathbf{l}_m^N \neq \mathbf{l}_m^0 \text{ for any } m; \\ \rho(\mathbf{s}^n, \mathbf{a}^n), n = N-1, \hat{\mathbf{A}}_F^n \neq \emptyset, \text{ and } \mathbf{l}_m^N = \mathbf{l}_m^0, \forall m. \end{cases} \quad (74)$$

The design of the reward  $\rho(\mathbf{s}^n, \mathbf{a}^n)$  is related to the goal of the original design problem (P1) which attempts to maximize the worst sum rate of WNs. Here, we consider three kinds of reward designs for  $\rho(\mathbf{s}^n, \mathbf{a}^n)$ , namely worst accumulated sum rate among users (WASR), difference of worst accumulated sum rate among users at two adjacent time slots (DWASR) [29], and instantaneous average sum rate of users (ISR), which are in turn defined as follows:

$$\rho_{WASR}(\mathbf{s}^n, \mathbf{a}^n) = \min_{k=1, \dots, K} \left( Z_k^{n-1} + R_{k,n} \right); \quad (75)$$

$$\rho_{DWASR}(\mathbf{s}^n, \mathbf{a}^n) = \min_{k=1, \dots, K} \left( Z_k^{n-1} + R_{k,n} \right) - \min_{k=1, \dots, K} Z_k^{n-1}; \quad (76)$$

$$\rho_{ISR}(\mathbf{s}^n, \mathbf{a}^n) = \frac{1}{K} \sum_{k=1}^K R_{k,n}, \quad (77)$$

where  $Z_k^{n-1}$  represents the accumulated sum data rate for the  $k$ th WN up to the  $(n-1)$ th time, and  $R_{k,n}$  is calculated by (16) with respect to the action  $\mathbf{a}^n$  and the state  $\mathbf{s}^n$  at the time instant  $n$ . The idea of the WASR method is to simply use the worst accumulated sum rate of the WNs as the reward according to the objective function of the optimization problem (P1). The DWASR method is conceptualized in accordance with [29], for which the difference of the worst accumulated sum rate at two adjacent time slots  $n$  and  $n-1$  is computed as the reward. The ISR method is proposed in this paper to take the instantaneous average sum rate of all WNs as the reward.

In Algorithm 2, we summarize the procedures of the proposed CARL algorithm. The algorithm starts from the offline UAV flight trajectory  $\mathbf{q}_m^*[n]$  obtained by using Algorithm 1. Let  $N_e$  be the number of episodes. We develop the flight corridor based on  $\mathbf{q}_m^*[n]$  and find the legal UAV flight actions  $\hat{\mathbf{A}}_F^n$  under the flight corridor constraint and the legal communication actions  $\bar{\mathbf{A}}_C^n$  under the node battery constraint. If the set of legal UAV flight actions  $\hat{\mathbf{A}}_F^n$  is non-empty, action selection is performed based on a decaying  $\epsilon$ -greedy approach [30], which means that  $\epsilon$  decreases slowly as training proceeds. After performing an action, the corresponding reward is received according to (74), which is used to update the Q-table. Note that if the set of legal UAV flight actions  $\hat{\mathbf{A}}_F^{n+1}$  at the next time instant  $n+1$  is empty or the mission reaches to the final time instant  $n = N-1$ , the current episode round will be terminated after performing the action, and the above procedures are repeated for the next episode round.

## V. NUMERICAL SIMULATION

### A. Simulation Settings

For simulation settings, we consider two UAVs flying over a  $600 \text{ m} \times 600 \text{ m}$  area at an altitude of 150 m with a mission period time of  $T_s = 100$  minutes, where the total number of time slots is set to 100. The initial positions of the first and the second UAVs are  $[0, 300, 150]$  m and  $[600, 300, 0]$  m, respectively. The maximum speed limit of the UAVs is

### Algorithm 2 Convex-assisted Reinforcement Learning (CARL) Algorithm

---

```

1: Obtain the UAVs' trajectory  $\mathbf{q}_m^*[n]$  through the offline
   convex optimization in Algorithm 1.
2: Initialize Q-value  $Q(\mathbf{s}, \mathbf{a}) = 0$  for all  $\mathbf{s}$  and  $\mathbf{a}$ .
3: for  $j = 0$  to  $N_e$  ( $N_e$  is the number of episodes) do
4:   Set  $n = 0$ 
5:   while  $n < N$  do
6:     Find available action sets  $\hat{\mathbf{A}}_F^n$  and  $\bar{\mathbf{A}}_C^n$  that meet
       the flight corridor limits as well as the battery
       constraints.
7:     if  $\hat{\mathbf{A}}_F^n \neq \emptyset$  then
8:       Draw a random number  $\epsilon_0$  between 0 and 1.
9:       if  $\epsilon_0 < \epsilon$  then
10:        Perform an action  $\mathbf{a}^n = (\mathbf{a}_F^n, \mathbf{a}_C^n) \in \hat{\mathbf{A}}_F^n \times \bar{\mathbf{A}}_C^n$ 
          randomly.
11:       else
12:        Perform the action  $\mathbf{a}^n$  with the highest Q-
          value, i.e.,  $\mathbf{a}^n = \arg \max_{\mathbf{a}} Q(\mathbf{s}^n, \mathbf{a})$ .
13:       end if
14:     end if
15:     if ( $\hat{\mathbf{A}}_F^{n+1} = \emptyset$  or  $n = N-1$ ) then
16:       Update Q-table as  $Q(\mathbf{s}^n, \mathbf{a}^n) \leftarrow (1-a) \cdot$ 
          $Q(\mathbf{s}^n, \mathbf{a}^n) + a \cdot R^n(\mathbf{s}^n, \mathbf{a}^n)$ 
17:       break;
18:     else
19:       Update Q-table as
          $Q(\mathbf{s}^n, \mathbf{a}^n) \leftarrow (1-a) \cdot Q(\mathbf{s}^n, \mathbf{a}^n) + a \cdot$ 
          $(R^n(\mathbf{s}^n, \mathbf{a}^n) + \gamma \max_{\mathbf{a}} Q(\mathbf{s}^{n+1}, \mathbf{a}))$ 
20:       Update  $n \leftarrow n+1$ 
21:     end if
22:   end while
23: end for

```

---

1 m/sec, and the safety distance between any two UAVs is 100 m. The system has a carrier frequency of 2.4 GHz with a transmission bandwidth of 5 MHz. The solar panel size of the wireless node is  $10 \text{ cm} \times 10 \text{ cm}$ . The solar EH data is obtained from NREL website [31], and we use 16 years of monitoring data at Elizabeth State University from 1997 to 2012, in which the average solar irradiance profiles over days in the morning (from 10 : 40 am to 12 : 20 pm and afternoon (from 12 : 20 pm to 14 : 00 pm) are depicted in Fig. 2. The environmental parameters for the channel models are set to  $A = 9.61$ ,  $B = 0.1592$ ,  $\eta_{LOS} = 1$  dB, and  $\eta_{NLOS} = 20$  dB [25]. The stopping criterion in the proposed offline method is given as  $\epsilon = 10^{-4}$ . For the CARL, the parameters for the quantization of system states are given by  $\Delta = 60$  m and  $\epsilon_H = 5$  dB. Moreover, we set the discount factor  $\gamma = 0.5$  and the penalty  $C_P = -10^3$ , and the decaying learning rate and decaying  $\epsilon$ -greedy are adopted with  $[a_{\max}, a_{\min}] = [0.9, 0.3]$  and  $[\epsilon_{\max}, \epsilon_{\min}] = [0.9, 0.1]^2$ . The width of the flight corridor

<sup>2</sup>The learning rate is given as  $a = (a_{\max} - a_{\min}) \times \max(\frac{N_e - n_{step}}{N_e}, 0) + a_{\min}$ , where  $n_{step}$  is the number of learning steps so far. The same method is applied for the decaying  $\epsilon$ -greedy.

is set to  $D_F = 90$  m in the morning and  $D_F = 105$  m in the afternoon. The values of the battery capacity  $B_{\max}$ , noise power  $\sigma_n^2$ , training number  $N_e$  are set to 1500 J,  $-80$  dBm and  $2 \times 10^6$ , respectively. The above parameters are used as default settings, except as otherwise stated.

For the offline designs, some heuristic methods are considered for performance comparison. An exhaustive power control method [32] is applied, in which each WN exhausts the available energy at each time slot for transmit power control. For the UAV trajectory, three flight plans are considered for the two UAVs: (1) uncrossed circles (UC), (2) crossed circles (CC), (3) straight lines and circles (SLC), as shown in Fig. 3. For the communication association, the nearest user association method is applied [33], and the color of the line in Fig. 3 represents the closest association node.

For the online design, a conventional RL method is included for performance comparison, in which the UAV location state is defined as in (67) and the channel state is quantized into three states with the threshold  $[-100, -90]$  (dB). For the conventional RL, the system receives a reward  $R^n(\mathbf{s}^n, \mathbf{a}^n)$ :

$$R^n(\mathbf{s}^n, \mathbf{a}^n) = \begin{cases} f(R_{k,n}) - \mu_n \left( \sum_{m=1}^M \|\mathbf{l}_m^{n+1} - \mathbf{l}_m^0\|_2 \right), n = \{0, \dots, N-1\}; \\ C_P, n = N-1, \mathbf{l}_m^N \neq \mathbf{l}_m^0 \text{ for any } m, \end{cases} \quad (78)$$

where  $f(R_{k,n})$  is a function of user rates  $R_{k,n}$ , the coefficient  $\mu_n = 10^{-4}n$  is positive and increased with the time index  $n$ , and it is applied before the total distance metric to force the UAVs to learn to return the starting position at the end of the mission. Note that at  $n = N-1$ , all the UAVs must comply with the constraint of returning to the starting point at the end of the mission; otherwise, the system receives a negative penalty  $C_P = -10^3$ .

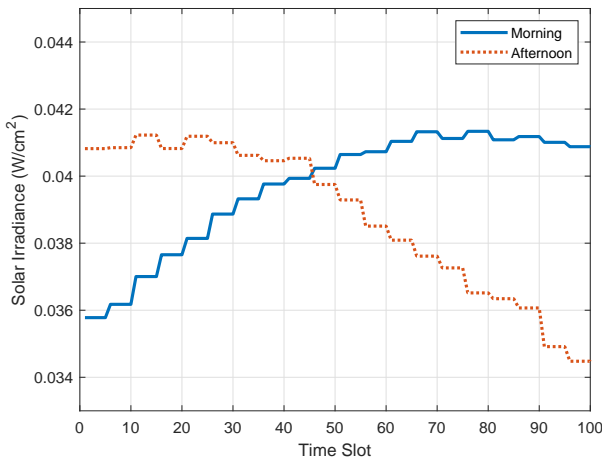


Fig. 2: Average solar irradiance profiles in the morning and afternoon at the solar site of Elizabeth State University from 1997 to 2012.

## B. Performance of Offline Designs

Fig. 4 shows the convergence of the proposed offline method for different numbers of WNs in the afternoon. The performance can be monotonically improved until convergence, which validates our analysis in Sec. III-D. Moreover, the required number of outer iterations for convergence increases with the number of WNs, and the performance is converged within four and eleven iterations for  $K = 2$  and  $K = 5$ , respectively.

To evaluate the influence of different combinations of design factors on the performance, the following offline methods are compared in Fig. 5 by only optimizing some of the design factors: (1) only communication association (OA), (2) only communication association and UAV flight trajectory (AFT), (3) only communication association and transmit power control (APC). In case that no optimization is applied, the exhaustive power control method and the UC flight trajectory in Fig. 3 are used. Here, the solar irradiance in the afternoon is adopted, and  $K = 3$ . From this figure, the proposed offline method with fully joint optimization performs much better than the other methods that optimize only one or two design factors. Comparing AFT and APC shows that optimizing the transmit power can improve the performance more than optimizing the UAV trajectory.

With the proposed offline method, Fig. 6 shows the optimal flight trajectory and communication association of the two UAVs during different EH periods when  $K = 3$ . The colors of the UAV flight paths represent the user association results. We can find that in this deployment, WN 3 is the farthest node from the two UAVs, and both UAVs serve WN 3 in the morning and afternoon to improve the worst user rate. As can be seen, in the early stage of the mission, UAV 1 decides to serve WN 3 because of the abundant energy harvested in the afternoon. On the other hand, in the morning, UAV 1 serves its closest node WN 1 in the early stage while serving WN 3 in the middle stage, for which WN 3 collects enough battery energy and gets closer to UAV 1.

Fig. 7 shows the optimal UAV flight trajectory and communication association of the proposed offline method for different numbers of WNs. Obviously, the placement of WNs affects the optimal flight paths and communication association. For example, when  $K = 2$ , the distances from the two WNs to the two UAVs are equal, and the two UAVs fly clockwise and counter-clockwise to avoid the interference problem. Similar observations can be found in the other cases. Taking another example of  $K = 4$ , each UAV tends to serve two closer WNs.

Fig. 8 compares three heuristic methods with the proposed offline method for different numbers of WNs<sup>3</sup>. For the three heuristic methods, the nearest user association and exhaustive power control methods are applied with the three different flight plans in Fig. 3. As can be seen, the proposed offline method is superior to the three compared methods, and the performance gap expands with the increase of  $K$ .

<sup>3</sup>The positions of WNs for  $K = 6$  are  $\mathbf{g}_1 = [200, 200]^T$ ,  $\mathbf{g}_2 = [200, 400]^T$ ,  $\mathbf{g}_3 = [400, 200]^T$ ,  $\mathbf{g}_4 = [400, 400]^T$ ,  $\mathbf{g}_5 = [200, 300]^T$  and  $\mathbf{g}_6 = [300, 300]^T$ .

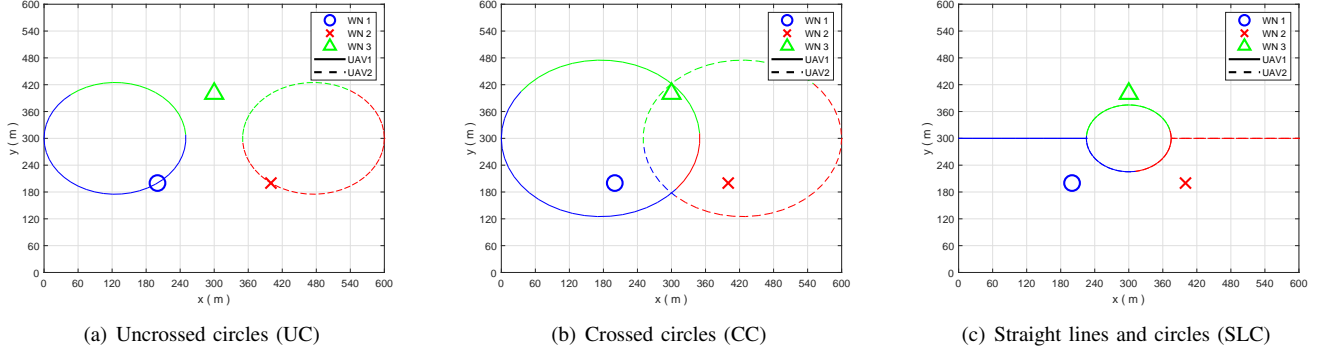


Fig. 3: The UAV flight plans and the nearest node association for the three heuristic schemes.

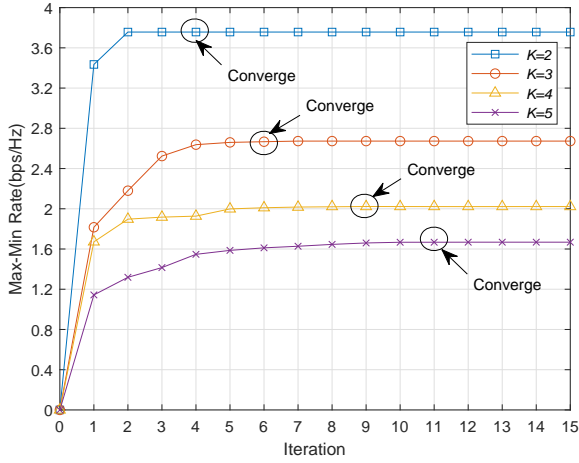


Fig. 4: Convergence of the proposed offline method for different numbers of WNs in the afternoon.

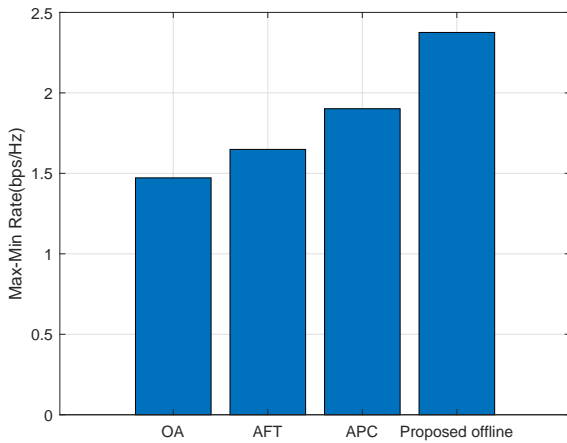


Fig. 5: Comparison of different combinations of design factors for offline optimization in the afternoon ( $K = 3$ ).

Fig. 9 shows the performance of the proposed offline method in the afternoon under different battery capacities. A larger battery capacity can improve the worst user rate performance. For a larger number of WNs, this improvement becomes more significant as each WN has less time to access the channel and is allowed to store more energy when waiting for data transmission.

### C. Performance of Online Designs

Fig. 10 shows the performance of the CARL method with the three different reward designs under various noise power values. It is clearly seen that the ISR method can achieve the best performance, in terms of the worst user rate and the probability of successful mission, i.e., all the UAVs successfully fly back to the initial point, among the three reward designs. This implicitly suggests that it is more appropriate to adopt the instantaneous average sum rate of all WNs, rather than the worst accumulated sum rate, as a reward in the online learning for balancing user rates. Therefore, the ISR reward is used for the CARL and the conventional RL (i.e., the reward in (78) is set as  $f(R_{k,n}) = \frac{1}{K} \sum_{k=1}^K R_{k,n}$ ) in the following simulations.

In Fig. 11, we compare the max-min user rates of the proposed offline and CARL methods for  $K = 3$  under various noise power values in the afternoon. The performance of the CARL method that utilizes a non-optimal UC trajectory is also simulated for comparison. It shows that the CARL can further enhance the performance of the proposed offline method and outperform the CARL method with a non-optimal UC trajectory. Regarding the probability of successful mission, the CARL with the optimal offline trajectory can attain a probability of 0.935, which is slightly better than the CARL with the UC trajectory.

Fig. 12 shows the performance of the CARL method for different widths of the flight corridor  $D_F$  in the morning. We can see that the flight corridor width performs best at  $D_F = 60$  m when  $N_e$  is small, while the flight corridor width that achieves the best performance becomes  $D_F = 90$  m when  $N_e$  increases. This is because the number of exploration states is fewer when  $D_F$  is small, and the Q-table converges quickly even if  $N_e$  is small. Moreover, as  $N_e$  increases, the performance improvement for  $D_F = 60$  m saturates,

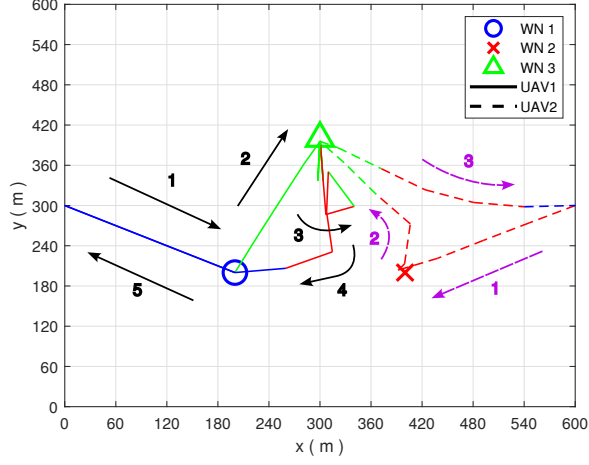
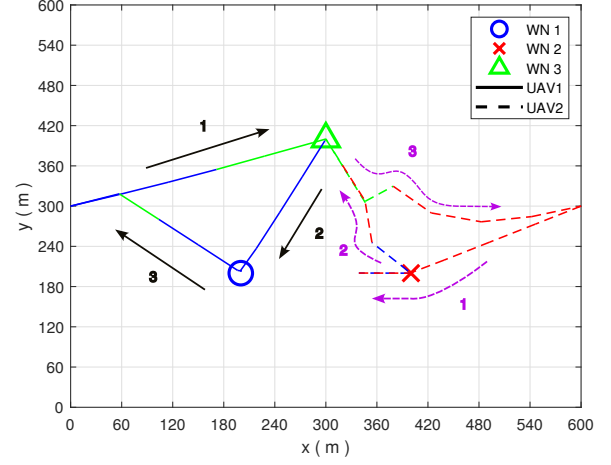
(a)  $K = 3$  in the morning(b)  $K = 3$  in the afternoon

Fig. 6: The optimal UAV flight trajectory and communication association of the proposed offline method in the morning and afternoon.

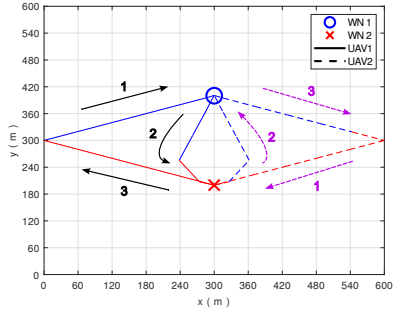
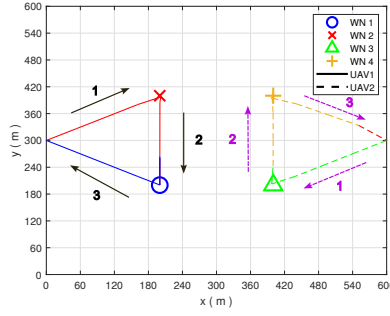
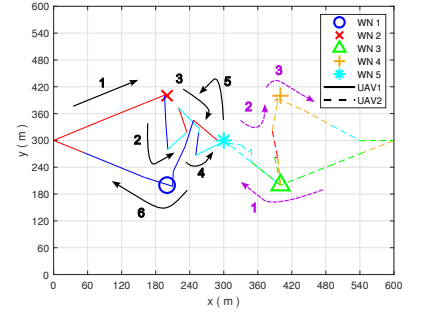
(a)  $K = 2$ (b)  $K = 4$ (c)  $K = 5$ 

Fig. 7: The optimal UAV flight trajectory and communication association of the proposed offline method in the afternoon for various numbers of WNs.

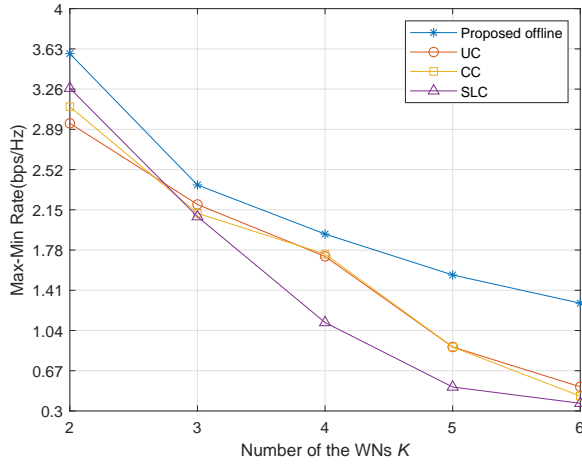
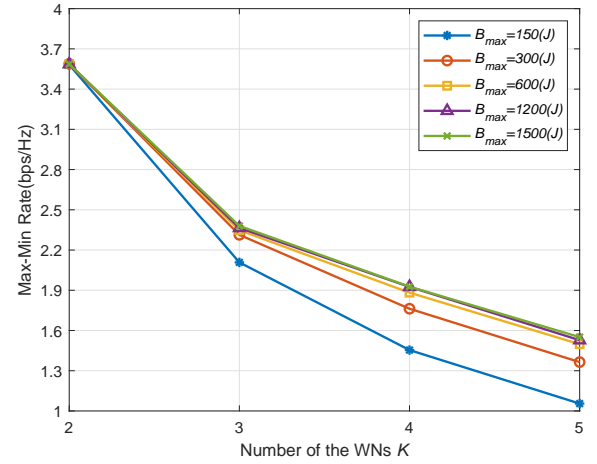


Fig. 8: Performance comparison of three heuristic schemes and the proposed offline method for various numbers of WNs in the afternoon.

Fig. 9: Performance of the proposed offline method in the afternoon under different battery capacities  $B_{max}$ .



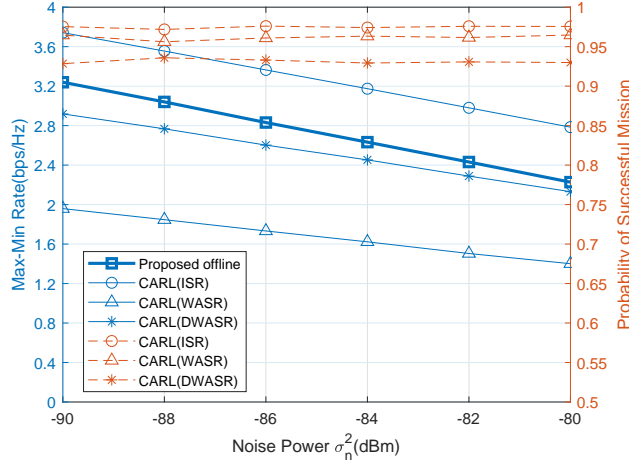


Fig. 10: Performance of the CARL method with the three different reward designs in the morning ( $K = 3$ ).

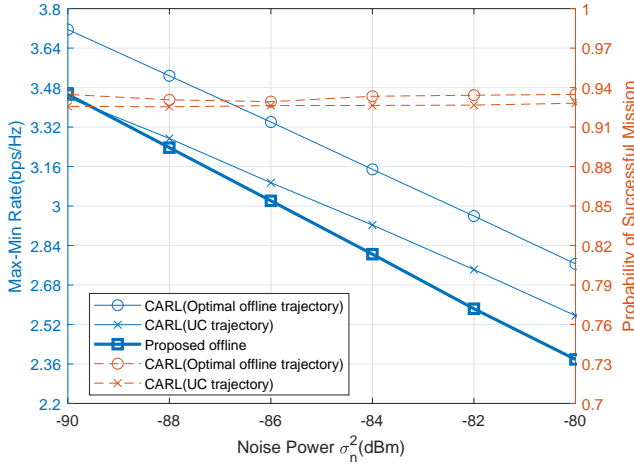


Fig. 11: Performance of the CARL, conventional RL, and offline methods under various noise power values in the afternoon ( $K = 3$ ).

whereas the other larger flight corridor widths can still offer continuous improvement. Hence, a larger flight corridor width can potentially improve the performance at the expense of more training epochs  $N_e$ .

In Fig. 13, we compare the performance of the CARL, conventional RL and proposed offline methods under different training epochs  $N_e$ . We can see that the max-min rate of the CARL method increases with  $N_e$ , whereas there is no significant performance improvement for the conventional RL method. This is because the performance of the conventional RL saturates quickly, while the performance of the CARL can be efficiently improved under the guidance of optimal offline trajectories. In addition, the probability of successful mission of the conventional RL is higher than that of the CARL when  $N_e$  is small, but the probability of the CARL eventually surpasses that of the conventional RL as  $N_e$  exceeds  $50 \times 10^4$ . This is because the reward design in the conventional RL forces the UAVs to fly back to the initial point with less

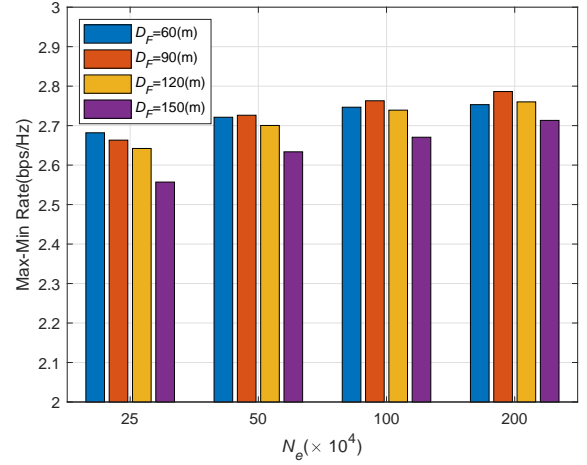


Fig. 12: Performance of the CARL method under different widths of the flight corridor  $D_F$  in the morning ( $K = 3$ ).

exploration, while the CARL can learn to fly back to the initial point by following the flight corridor if the training number is sufficiently large. Also, both online methods outperform the offline method due to the dynamic responses to the changes of channel and battery conditions.

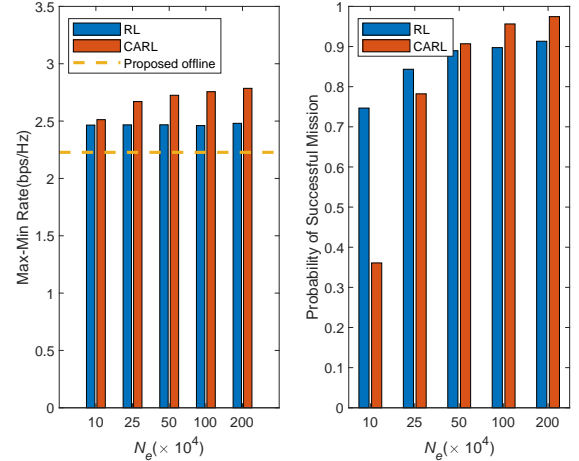


Fig. 13: Performance of the CARL, conventional RL, and offline methods under various numbers of training periods  $N_e$  in the morning ( $K = 3$ ).

## VI. CONCLUSION

In this paper, we investigate the joint design problem of multi-UAV flight trajectories, communication association between UAVs and ground nodes, and uplink power control for a multi-UAV network with multiple uplink EH nodes. Under a mixed channel model with LOS/NLOS and small-scale fading, a series of SCAs are performed to deal with this non-convex joint design problem, and an offline method that does not require causal knowledge of instantaneous CSI and ESI is proposed. An online CARL method is proposed to further improve the performance by exploiting preset UAV flight corridors

according to the optimal offline UAV trajectories. Through efficient flight guidance and reducing the number of state spaces to be explored, the CARL can improve the learning convergence and performance compared to the conventional RL without the assistance from the offline UAV trajectories.

#### APPENDIX A PROOF OF LEMMA 1

We first consider a function  $\phi(x, y) = \ln(\frac{C_1 C_2}{xy} + \frac{C_3}{x})$  with two variables  $x > 0$  and  $y > 0$ , where  $C_1, C_2, C_3 > 0$ . The Hessian matrix of  $\phi(x, y)$  is

$$\nabla^2 \phi(x, y) = \begin{bmatrix} \frac{\partial^2 \phi(x, y)}{\partial x^2} & \frac{\partial^2 \phi(x, y)}{\partial x \partial y} \\ \frac{\partial^2 \phi(x, y)}{\partial x \partial y} & \frac{\partial^2 \phi(x, y)}{\partial y^2} \end{bmatrix}, \quad (\text{A.1})$$

where each term can be derived as

$$\frac{\partial^2 \phi(x, y)}{\partial x^2} = \frac{1}{x^2}; \quad (\text{A.2})$$

$$\frac{\partial^2 \phi(x, y)}{\partial y^2} = \frac{C_1 C_2 (C_1 C_2 + 2C_3 y)}{(C_1 C_2 y + C_3 y^2)^2}; \quad (\text{A.3})$$

$$\frac{\partial^2 \phi(x, y)}{\partial x \partial y} = \frac{\partial^2 \phi(x, y)}{\partial y \partial x} = 0. \quad (\text{A.4})$$

Since  $\det(\nabla^2 \phi(x, y)) \geq 0$  and  $\text{tr}(\nabla^2 \phi(x, y)) \geq 0$ , the Hessian matrix  $\nabla^2 \phi(x, y)$  is positive semi-definite. Thus, the function  $\phi(x, y)$  is convex in  $(x, y)$ .

From (19), (29) and (30), we rewrite  $\tilde{R}_{1m}[n]$  as a log-sum-exp form:

$$\tilde{R}_{1m}[n] = \log_2 \left( \sum_{i=1}^K e^{\tilde{\phi}_i(X_{m,i}[n], Y_{m,i}[n])} + \sigma_n^2 \right), \quad (\text{A.5})$$

where  $\tilde{\phi}_i(X_{m,i}[n], Y_{m,i}[n]) = \ln \left( \frac{P_i[n] C_1 C_2}{X_{m,i}[n] Y_{m,i}[n]} + \frac{P_i[n] C_3}{X_{m,i}[n]} \right)$ . By using the fact that a function  $\log_2 \left( \sum_{i=1}^K e^{g_i(x)} \right)$  is convex whenever  $g_i(x)$  is convex for all  $i$  [27], it can be shown that  $\log_2 \left( \sum_{i=1}^K \exp \{ \tilde{\phi}_i(X_{m,i}[n], Y_{m,i}[n]) \} \right)$  is convex in  $X_{m,i}[n]$  and  $Y_{m,i}[n]$  for all  $i$ , since  $\tilde{\phi}_i(X_{m,i}[n], Y_{m,i}[n])$  is convex in  $(X_{m,i}[n], Y_{m,i}[n])$ . As a result,  $\tilde{R}_{1m}[n]$  is convex in  $X_{m,i}[n]$  and  $Y_{m,i}[n]$  for all  $i$ .

#### APPENDIX B PROOF OF THEOREM 1

We first consider a convex function  $\varphi(\mathbf{x}, \mathbf{y})$  with two variables  $\mathbf{x}$  and  $\mathbf{y}$ , and a lower bound can be obtained by using the first-order Taylor expansion at  $(\mathbf{x}_0, \mathbf{y}_0)$ :

$$\begin{aligned} \varphi(\mathbf{x}, \mathbf{y}) &\geq \varphi(\mathbf{x}_0, \mathbf{y}_0) + \nabla_{\mathbf{x}} \varphi(\mathbf{x}, \mathbf{y}) \Big|_{(\mathbf{x}, \mathbf{y})=(\mathbf{x}_0, \mathbf{y}_0)} (\mathbf{x} - \mathbf{x}_0) \\ &\quad + \nabla_{\mathbf{y}} \varphi(\mathbf{x}, \mathbf{y}) \Big|_{(\mathbf{x}, \mathbf{y})=(\mathbf{x}_0, \mathbf{y}_0)} (\mathbf{y} - \mathbf{y}_0). \end{aligned} \quad (\text{B.1})$$

For a given  $\mathbf{q}_m[n] = \mathbf{q}_m^r[n]$  and from (29) and (30), we can calculate  $X_{m,i}^r[n]$  and  $Y_{m,i}^r[n]$  as in (32) and (33). From Lemma 1, the function  $\tilde{R}_{1m}[n] = \log_2 \left( \sum_{i=1}^K P_i[n] \left( \frac{C_1 C_2}{X_{m,i}[n] Y_{m,i}[n]} + \frac{C_3}{X_{m,i}[n]} \right) + \sigma_n^2 \right)$  is a convex function with respect to the variables  $X_{m,i}[n]$  and  $Y_{m,i}[n]$  for all  $i$ . By using (B.1), we can derive the first-order

Taylor expansion of  $\tilde{R}_{1m}[n]$  at  $(X_{m,i}[n], Y_{m,i}[n]) = (X_{m,i}^r[n], Y_{m,i}^r[n])$ , given by  $\tilde{R}_{1m}^{lb}[n]$  in (31). Hence, we can get  $\tilde{R}_{1m}[n] \geq \tilde{R}_{1m}^{lb}[n]$ .

#### APPENDIX C PROOF OF THEOREM 2

Define a variable  $U_{m,i}[n] = \|\mathbf{q}_m[n] - \mathbf{g}_i\|_2^2$ . By applying the change of variables into (30), we can get

$$Y_{m,i}[n] = 1 + Ae^{-B \left( \frac{180}{\pi} \tan^{-1} \left( \frac{H}{\sqrt{U_{m,i}[n]}} \right) - A \right)}. \quad (\text{C.1})$$

Since  $\tan^{-1} \left( \frac{1}{\sqrt{x}} \right)$  is a convex function when  $x > 0$ , the first-order Taylor expansion of  $\tan^{-1} \left( \frac{H}{\sqrt{U_{m,i}[n]}} \right)$  at the point  $U_{m,i}[n] = \|\mathbf{q}_m^r[n] - \mathbf{g}_i\|_2^2 \triangleq U_{m,i}^r[n]$  can be derived as

$$\begin{aligned} \tan^{-1} \left( \frac{H}{\sqrt{U_{m,i}[n]}} \right) &\geq \tan^{-1} \left( \frac{H}{\sqrt{U_{m,i}^r[n]}} \right) \\ &\quad - \frac{H}{2\sqrt{U_{m,i}^r[n]} (H^2 + U_{m,i}^r[n])} \times (U_{m,i}[n] - U_{m,i}^r[n]). \end{aligned} \quad (\text{C.2})$$

By simply applying (C.2) in (C.1), an upper bound  $Y_{m,i}^{up}[n]$  can be obtained for  $Y_{m,i}[n]$ , as shown in (36). In addition, it reveals in (36) that  $Y_{m,i}^{up}[n]$  is convex in  $\mathbf{q}_m[n]$ , since the term  $\|\mathbf{q}_m[n] - \mathbf{g}_i\|_2^2$  is a square norm function of  $\mathbf{q}_m[n]$ .

#### REFERENCES

- [1] J. Lu, H. Okada, T. Itoh, R. Maeda and T. Harada, "Towards the world smallest wireless sensor nodes with low power consumption for 'Green' sensor networks," in *Proc. IEEE ICSENS*, pp. 1-4, 2013.
- [2] C. You and R. Zhang, "Hybrid offline-online design for UAV-enabled data harvesting in probabilistic LoS channels," *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 3753-3768, Jun. 2020.
- [3] R. Chen, X. Li, Y. Sun, S. Li, and Z. Sun, "Multi-UAV coverage scheme for average capacity maximization," *IEEE Commun. Lett.*, vol. 24, no. 3, pp. 653-657, Mar. 2020.
- [4] M. A. Ali and A. Jamalipour, "Dynamic aerial wireless power transfer optimization," *IEEE Trans. Veh. Technol.*, vol. 71, no. 4, pp. 4010-4022, Apr. 2022.
- [5] Y. Che, Y. Lai, S. Luo, K. Wu and L. Duan, "UAV-aided information and energy transmissions for cognitive and sustainable 5G networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1668-1683, Mar. 2021.
- [6] Y. Lai, Y. L. Che, S. Luo and K. Wu, "Optimal wireless information and energy transmissions for UAV-enabled cognitive communication systems," in *Proc. IEEE Int. Conf. Commun. Syst. (ICCS)*, pp. 168-172, Dec. 2018.
- [7] S. A. Hoseini, J. Hassan, A. Bokani and S. S. Kanhere, "Trajectory optimization of flying energy sources using Q-learning to recharge hotspot UAVs," in *IEEE INFOCOM Workshops*, pp. 683-688, 2020.
- [8] Y. Hu, F. Zhang, T. Tian, and D. Ma, "Placement optimization method for multi-UAV relay communication," *IET Commun.*, vol. 14, no. 6, pp. 1005-1015, Apr. 2020.
- [9] X. Xu, Y. Zhao, L. Tao and Z. Xu, "Resource allocation strategy for dual UAVs-assisted MEC system with hybrid solar and RF energy harvesting," in *Proc. 3rd IEEE Int. Conf. Comput. Commun. Internet (ICCCI)*, pp. 52-57, 2021.
- [10] Y. Zeng, Q. Wu, and R. Zhang, "Accessing from the sky: a tutorial on UAV communications for 5G and beyond," *Proc. IEEE*, vol. 107, no. 12, pp. 2327-2375, Dec. 2019.

- [11] K. Singh, M. Ku, J. Lin and T. Ratnarajah, "Toward optimal power control and transfer for energy harvesting amplify-and-forward relay networks," *IEEE Trans. Wireless Commun.*, Vol. 17, No. 8, pp. 4971-4986, Aug. 2018.
- [12] G. Zhang, Q. Wu, M. Cui, and R. Zhang, "Securing UAV communications via joint trajectory and power control," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 1376-1389, Feb. 2019.
- [13] Q. Wu, Y. Zeng, and R. Zhang, "Joint trajectory and communication design for UAV enabled multiple access," in *Proc. IEEE Global Commun. Conf.*, pp. 1-6, 2017.
- [14] 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.
- [15] Y. Sun, D. Xu, D. W. K. Ng, L. Dai, and R. Schober, "Optimal 3D-trajectory design and resource allocation for solar-powered UAV communication systems," *IEEE Trans. Commun.*, vol. 67, no. 6, pp. 4281-4298, Jun. 2019.
- [16] S. Salehi, J. Hassan, A. Bokani, S. A. Hoseini and S. S. Kanhere, "Poster abstract: a QoS-aware, energy-efficient trajectory optimization for UAV base stations using Q-learning," in *19th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pp. 329-330, 2020.
- [17] S. Zhang, H. Zhang, B. Di, and L. Song, "Cellular UAV-to-X communications: design and optimization for multi-UAV networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 1346-1359, Feb. 2019.
- [18] Q. Wang, W. Zhang, Y. Liu, and Y. Liu, "Multi-UAV dynamic wireless networking with deep reinforcement learning," *IEEE Commun. Lett.*, vol. 23, no. 12, pp. 2243-2246, Dec. 2019.
- [19] Q. Wu, Y. Zeng, and R. Zhang, "Joint trajectory and communication design for multi-UAV enabled wireless networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 2109-2121, Mar. 2018.
- [20] C. Zhan and Y. Zeng, "Aerial-ground cost tradeoff for multi-UAV enabled data collection in wireless sensor networks," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1937-1950, Mar. 2020.
- [21] C. Shen, T. Chang, J. Gong, Y. Zeng, and R. Zhang, "Multi-UAV interference coordination via joint trajectory and power control," *IEEE Trans. Signal Process.*, vol. 68, pp. 843-858, 2020.
- [22] J. Bowman, J. Brooks, C. Lopez, A. Marcos-Martinez and A. Salman, "Secure data collection using autonomous unmanned aerial vehicles," in *Proc. Syst. Inf. Eng. Design Symp. (SIEDS)*, pp. 1-6, 2020.
- [23] Q. V. Do, Q. -V. Pham and W. -J. Hwang, "Deep reinforcement learning for energy-efficient federated learning in UAV-enabled wireless powered networks," *IEEE Commun. Lett.*, vol. 26, no. 1, pp. 99-103, Jan. 2022.
- [24] 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.
- [25] J. Holis and P. Pechac, "Elevation dependent shadowing model for mobile communications via high altitude platforms in built-up areas," *IEEE Trans. Antennas Propag.*, vol. 56, no. 4, pp. 1078-1084, 2008.
- [26] M. Grant and S. Boyd. (2016).CVX: *MATLAB software for disciplined convex programming*. [Online]. Available: <http://cvxr.com/cvx>
- [27] S. Boyd and L. Vandenberghe, "Convex optimization," *Cambridge University Press*, 2004.
- [28] M. Razaviyayn, "Successive convex approximation: analysis and applications," 2014.
- [29] U. F. Siddiqi, S. M. Sait and M. Uysal, "Deep reinforcement based power allocation for the max-min optimization in non-orthogonal multiple access," *IEEE Access*, vol. 8, pp. 211235-211247, 2020.
- [30] H. Afifi and H. Karl, "Reinforcement learning for virtual network embedding in wireless sensor networks," in *Proc. 16th Int. Conf. Wireless Mobile Comput. Netw. Commun. (WiMob)*, Thessaloniki, Greece, 2020, pp. 123-128.
- [31] N. R. E. Laboratory. (2012, Feb.) Solar radiation resource information. [Online]. Available: <http://www.nrel.gov/tredc/>.
- [32] M. Ku, Y. Chen and K. J. R. Liu, "Data-Driven stochastic models and policies for energy harvesting sensor communications," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 8, pp. 1505-1520, Aug. 2015.
- [33] 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.