Multi-UAV Trajectory and Power Optimization for Cached UAV Wireless Networks With Energy and Content Recharging-Demand Driven Deep Learning Approach

Shuqi Chai, *Associate Member, IEEE*, and Vincent K. N. Lau, *Fellow, IEEE*

***Abstract*— In this paper, we propose a novel joint trajectory and communication scheduling scheme for multiple unmanned aerial vehicles (UAVs) enabled wireless caching networks. To exploit the favorable propagation of air-to-ground channels, we consider an ultra dense UAVs enabled content-centric wireless transmission network, where massive UAVs are deployed to transmit cached contents to a group of random distributed ground users. We formulate the problem as an infinite horizon ergodic stochastic differential game (SDG) for optimizing the users' quality-of-experience (QoE). In particular, stochastic dynamics of channel states, UAVs' mobility, energy queues and content request queues are modeled in this game. To deal with the state coupling between the UAVs, we consider a limiting problem for large number of UAV based on mean field analysis. A reduced-complexity decentralized solution can be obtained through mean-field equilibrium analysis. To further reduce the solution complexity on each UAV, we propose a model-specific deep neural network (DNN) to learn the optimal control solution in an online manner. The DNN is not arbitrarily generated but tailored to the structural properties of the value function and stationary distribution based on the homotopy perturbation method analysis. Finally, simulation results are provided to show that the proposed solution can achieve significant gain over the existing baselines.**

***Index Terms*— UAV caching networks, multi-UAV trajectory design, radio resource control, mean-field game, deep learning.**

I. Introduction

IT IS widely anticipated that the capacity in 5G wireless networks will continue to increase to meet the requirements of high-data-rate demand applications, such as multimedia streaming. Currently, there is a surge of interest in utilizing unmanned aerial vehicles (UAVs) as moving base stations in wireless networks to improve the network capacity and content-delivery performance. Using UAVs-assisted wireless network has several motivations. First, traditional terrestrial base stations (BSs) are comparatively inefficient to support mobile users with popularly requested content due to the lack of line-of-sight (LOS). Hence, they have to use the lower frequency spectrum which is heavily congested. On the other hand, UAVs can assist the BSs with a higher data transmission rate by exploiting the inherit LOS propagation, and this can allow the usage of the high frequency spectrum to serve hotspot demands. Furthermore, the agility and controllability of UAV-aided wireless networks can achieve larger wireless coverage, making them ideal to support reliable transmission for special events, such as large-scale ad-hoc activities or sports events.

Many research works have analyzed UAV-aided wireless networks in recent years. In [1]–[4], the authors studied UAVs' transmission technology and discussed various open research problems of UAVs-aided networks. The authors in [5]–[10] optimized UAVs' aerial positions to assist the ground static networks by deploying UAVs as the moving relays. In addition, many recent works have deployed UAVs as aerial mobile BSs. The works [10]–[14] analyzed the UAV-based aerial BS area coverage problem by optimizing the UAV placement. In [15], [16], the authors proposed a novel framework by jointly optimizing the UAV trajectory and radio resources to achieve UAV-aided network throughput maximization and secure communication. However, the scenario with multiple UAVs and multiple users is of more practical and appealing in the content-centric network design. In [17], the authors considered the multi-UAV trajectory and deployment optimization problem, and both the cases of static and mobile ground terminals were studied. The authors in [18] exploited multiple UAVs to form a C-RAN to assist ground BSs and designed the UAVs' travel path to optimize the spectral efficiency. In [19], the authors expanded their previous work to the multi-UAV case by jointly optimizing UAVs' trajectory, power and communication scheduling. However, these multi-UAV control framework are not cross-layer design, in the sense that only physical layer metrics are considered. In addition, the computation complexity of the multi-UAV optimization algorithms used in above works are very high. Thus the implementation of such algorithm is quite challenging for UAV online control.

The deployment of UAVs in wireless networks may induce a tremendous backhaul burden, and the BS to UAV link may be the performance bottleneck. Cache-enabled UAVs

Manuscript received October 28, 2020; revised February 24, 2021; accepted April 12, 2021. Date of publication June 14, 2021; date of current version September 16, 2021. This work was supported by the Research Grants Council, Hong Kong, under Project 16204018. *(Corresponding author: Shuqi Chai.)*

The authors are with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong (e-mail: schai@ust.hk; eeknlau@ust.hk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSAC.2021.3088694.

Digital Object Identifier 10.1109/JSAC.2021.3088694

content-centric networking [20]–[22][1] has thus emerged as a promising solution to alleviate the bottlenecks. In such networking, UAVs are proactively loaded with cached content and deployed to serve the users according to the instantaneous user demands, with the transmitted packets coming from the local cache instead of the BS-UAV link. There are a few research works on UAV caching communications. In [23], a deep learning-based UAV caching design was proposed. However, only optimized UAV transmission and cache content scheduling was considered, but not the trajectory design. In [24], the authors studied UAV-assisted caching in small-cell networks, while the authors in [21] proposed a UAV-based content caching and position design scheme. However, the above multi-UAV control designs suffer from exponential complexity w.r.t. the number of UAVs. Furthermore, in these works, the solutions have ignored the limited battery life as well as the limited storage capacity of the UAV. In [25], the authors proposed a cache-enabled UAV transmission and trajectory scheme taking into account the UAV endurance issue. In our recent paper [26], we proposed an online reinforcement learning algorithm to control UAV trajectory and radio resource allocation for UAV-assisted wireless networks. The solution considers a novel energy and content recharging scheme, which can effectively address the finite battery life and finite storage capacity issue of the UAV. However, only the single UAV case is considered in [25] and [26].

In this paper, we extend the work in [26] to a multi-UAV system. Specifically, we consider demand-driven dynamic control of the trajectories, the decision to return to the power and content charging station or not, and transmission power control of the multi-UAV system subject to the constraints of limited battery and limited cache storage capacity. The extension to multiple UAVs is highly non-trivial due to the strong coupling among the states of the UAVs induced by the anti-collision constraint. The following summarizes the key challenges and our contributions.

- **Multi-UAV Trajectory Optimization with Content and Energy Charging:** In most existing works, the finite battery life of the UAV as well as the finite storage have been ignored. In this paper, we propose a novel energy and content recharging scheme which enables continuous operation of the cache-assisted UAV network as well as adaptation to the drift of content popularity in the requests distribution of the user. As such, the decision space of the UAV system has to include decisions as to whether a particular UAV has to continue to serve the next demands or to fly back to the charging station to replenish the energy and cache content in addition to the trajectory control and transmission power control. The proposed solution is demand-driven in the sense that it is adaptive to both the channel state information (CSI) between the UAVs and the users as well as the request queue state information (QSI) of the demands from various users.
- **Complex Dynamic State Coupling in the Multi-UAV System:** One of the key challenges in the extension to a multi-UAV system is the complexity issue induced by complex dynamic state coupling. Specifically, the states of the multiple UAVs are tightly coupled together due to the *inter-UAV collision avoidance constraint*. The inter-UAV collision avoidance constraint is to prevent UAVs from crashing into each other due to the wind disturbance or mis-coordinated trajectory control. Such complex state coupling induce exponential complexity in the brute-force optimization approach.
- **Reduced-Complexity Mean-Field Game (MFG) Solution:** To reduce the complex coupling issues, we first consider an ergodic stochastic optimization formulation with the state evolution of the multi-UAV system modeled as an N-dimensional continuous time stochastic differential equations (SDEs) [27], [28]. To address the inter-UAV collision issue, we define a new anti-collision loss function in the optimization objective that penalizes the collision avoidance violation. To deal with the coupling induced by the inter-collision, we exploit mean-field analysis [28]–[30] to construct a *mean-field equilibrium* which captures a limiting behavior with the number of UAVs tending to infinity. Using *propagation of chaos* in statistical mechanics, we show that the original ergodic control problem for the multi-UAV is asymptotically equivalent to a reduced-complexity mean-field game for a large number of UAVs. Such mean-field approximation enables decomposition of the original coupled HJB equations into the MFG equations.
- **Deep-learning Design and Online Training:** Solving the derived MFG equations using classical approach is still computational intractable. There are several research works on the learning based UAV trajectory and transmission control. In [31], the authors design a CNN based deep supervised learning to enable UAV-aided edge caching. In [32], the authors propose a hierarchical deep reinforcement learning based UAV trajectory and resource control. In contrast, we exploit problem-specific structures to design an online deep learning solution to learn the solution. The motivation of using the DNN method is on the fast inferencing to compute the control policy given a state realization. To enhance the expressibility of the DNN, we tailored the datapath of the DNN according to the structural properties of the optimal control and the value functions based on the *homopty perturbation method* [33], [34]. As such, the DNN can approximate the value function more efficiently with less weights. Based on this, we develop an *online training algorithm* which does not require *labeled data*. In other words, the online training algorithm autonomously adjust the weights of the DNN to converge to the best approximator of the solution of the MFG equations [35]–[37]. We have also obtained the error bound of the DNN and showed that the complexity is independent of the number of UAVs.

The paper is organized as follows. In Section II, we introduce the architecture of the UAV-assisted wireless caching system and communication model, as well as the UAV control and energy consumption model. In Section III, we formulate the UAV trajectory and radio resource control

[1]Over 70% of the new capacity demands for 5G wireless networks come from high quality video streaming applications. As such, these capacity demands are mostly content-centric. These demands have huge spatial and temporal correlations between content requests by users in the network.

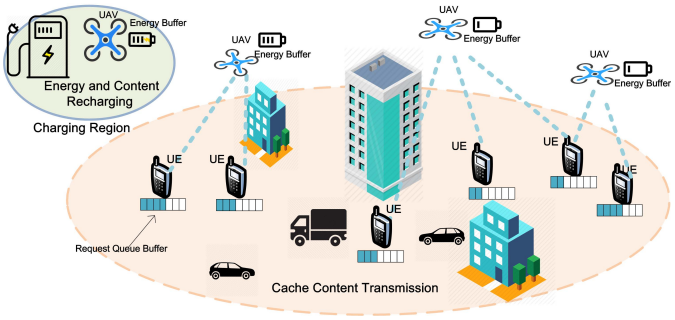


Fig. 1. Architecture of the multi-UAV wireless caching network with energy and content recharging.

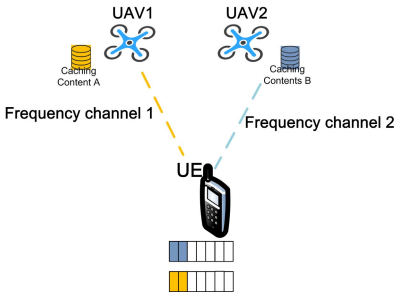


Fig. 2. Illustration of UAV transmission model with the case of 2 UAV and single user.

problem as a stochastic differential game. In Section IV, using the mean-field game approach, we derive a reduced-complexity solution based on mean-field equilibrium analysis. In Section V, we propose a novel model-specific deep neural network (DNN) to learn the partial differential equations (PDEs) numerically by exploiting an HPM method. Section VI illustrates the numerical simulation results and presents discussion. Finally, we summarize our main results in Section VII.

II. System Model

In this section, we introduce the architecture of the multi-UAV wireless caching network, with the basic stochastic dynamic models of the UAV trajectory, user content requests queues and UAV energy/content recharging discussed in detail.

A. UAV-Assisted Wireless Caching Network Architecture

As illustrated in Fig.1, we consider a cache-enabled wireless network where N UAVs are deployed to provide data transmission to M users by using unicast. Assume that there are a total of NJ content files in the content server, and each UAV will be responsible for the content files set $|\mathcal{J}^n| = J$ independently. Denote the multi-UAV service groups as $\mathcal{N} = \{1, \ldots, N\}$, and each UAV n can only be cached with a single content file $j \in \mathcal{J}^n$ to serve the users and is subject to a finite energy and storage buffer[2]. Specifically, let $\mathcal{M} = \{1, \ldots, M\}$ denote the set of users served by all UAVs where M denotes the number of users. To support continuous operation of the UAV network, we assume there is a charging station at position $\boldsymbol{X}^{ch} = [x^{ch}, y^{ch}, z^{ch}]^T$. The charging station can support energy and content recharging for UAVs.

B. UAV Communication Model

The 3D coordinates of the n-th UAV are given by $\boldsymbol{X}_n = [x_n, y_n, z_n]^T$. Let $\bar{\boldsymbol{X}}_m = [\bar{x}_m, \bar{y}_m, \bar{z}_m]^T$ denote the position of user m. The distance between the UAV n and user m can be written as

$$D_{nm} = ||\boldsymbol{X}_n - \bar{\boldsymbol{X}}_m||_2. \tag{1}$$

Based on this, the large-scale path gain from UAV n to user m at time t can be modeled as [3]

$$L_{nm}(t) = \rho_0 D_{nm}^{-2}(t) = \frac{\rho_0}{||\boldsymbol{X}_n(t) - \bar{\boldsymbol{X}}_m||_2^2}, \tag{2}$$

[2]The solution in this paper can be extended easily to include caching multiple files. In Sec VI, we have compare the performance with the extension.

[3]We assume that all users positions are static for simplicity in this paper.

where ρ_0 denotes the reference path gain at distance $d_0 = 1$m.

The instantaneous channel power gain between UAV n and UE m can be modeled as

$$|H_{nm}(t)|^2 = L_{nm}(t)|h_{nm}(t)|^2, \tag{3}$$

where $L_{nm}(t)$ is the large-scale path gain as defined in (2), and $h_{nm}(t)$ accounts for the short-term channel with dynamics given by the following stochastic differential equation:

$$dh_{nm}(t) = -\frac{1}{2}a_H h_{nm}(t)dt + \sqrt{a_H}\mathrm{d}W_{nm}^h(t), \tag{4}$$

where $a_H > 0$ determines the temporal correlation (depending on the mobility of the user) and $W_{nm}^h(t)$ is the standard independent Wiener process for the short-term fading process [38]. According to Ito's formula, the channel power gain process $G_{nm}(t) = |h_{nm}|^2$ will have the dynamic

$$\mathrm{d}G_{nm}(t) = (a_H - a_H G_{nm}(t))dt + 2\sqrt{a_H G_{nm}(t)}\mathrm{d}W_{nm}^h(t), \tag{5}$$

We assume each UAV has a dedicated channel and hence, can serve one user at a time. However, the users are equipped with multiple radios and they are allowed to receive data streams from different UAVs simultaneously. Figure 2 illustrates an example.

As each UAV has a dedicated channel to serve users, and thus there's no frequency interference over dierent UAV-users links. As a result, the received signal at the m-th user from the n-th UAV can be expressed as $Y_{nm} = \sqrt{P_n L_{nm}(t)}h_{nm}(t)X_n + Z_{nm}$, where X_n denotes the transmitted symbol, and P_n is the transmit power of the n-th UAV, and $Z_{nm} \sim \mathcal{CN}(0, 1)$ is the channel noise from UAV n to user m. Let $u_{nm} \in \{0, 1\}$ denote the user selection index for the m-th user at the n-th UAV with the scheduling constraint $\sum_{m=1}^{M} u_{nm} = 1$. The achievable data rate (bits/sec) from the n-th UAV to the m-th user at time t is given by

$$R_{nm}(t) = u_{nm}(t)W\log_2(1 + |H_{nm}(t)|^2 P_n(t)), \tag{6}$$

where W is the bandwidth allocated to the UAV.

We assume that the $n-$th UAV can only be preloaded with one content file $j \in \mathcal{J}^n$ in the cache. Denote the caching state as $\boldsymbol{s}_n = \{s_n^j, \forall j \in \mathcal{J}^n\}_{n\in\mathcal{N}}$, where $s_n^j \in \{0, 1\}$ is the j-th content file caching placement state at UAV n. Specifically, $s_n^j = 1$ if the j-th content file is cached in UAV n, and $s_n^j = 0$, otherwise. The cache states at each UAV should satisfy $\sum_{j=1}^{J} s_n^j = 1, \forall n \in \mathcal{N}$.

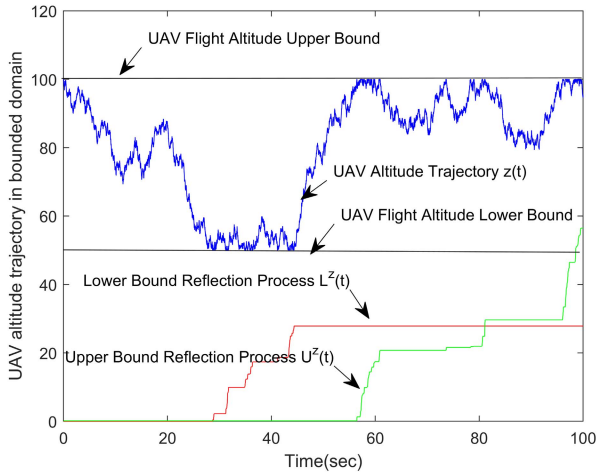


Fig. 3. Example of UAV altitude trajectory in bounded region and reflection processes with upper bound at 100m, lower bound at 50 m.

C. UAV Trajectory Control Model

Given the UAV flight trajectory control $\boldsymbol{v}_n = [\boldsymbol{v}_n^x, \boldsymbol{v}_n^y, \boldsymbol{v}_n^z]^T$, the nth UAV trajectory follows the dynamic:

$$\mathrm{d}\boldsymbol{X}_n(t) = \boldsymbol{v}_n(t)\mathrm{d}t + \boldsymbol{\sigma_X}\mathrm{d}\boldsymbol{W}_n^X(t) + \mathrm{d}\boldsymbol{L}_n^X(\mathrm{t}) - \mathrm{d}\boldsymbol{U}_n^X(t), \quad (7)$$

where $\boldsymbol{\sigma_X} = \mathrm{diag}(\sigma_X^x, \sigma_X^y, \sigma_X^z)$ denotes the volatility of the process due to the wind disturbance, and $\boldsymbol{W}_n^X(t)$ is a three dimensional independent standard Brownian motion process. $\boldsymbol{L}_n^X(t) = [L_n^x(t), L_n^y(t), L_n^z(t)]^T$ and $\boldsymbol{U}_n^X(t) = [U_n^x(t), U_n^y(t), U_n^z(t)]^T$ are the non-decreasing and continuous reflection processes associated with the UAVs flight space boundaries $x(t) = L_x$, $x(t) = U_x$, $y(t) = L_y$, $y(t) = U_y$ and $z(t) = L_z$, $z(t) = U_z$ respectively, where L_x, L_y, L_z and U_x, U_y, U_z are the boundary value. As a result, $\boldsymbol{X}_n(t)$ will always be within $[L_x, U_x] \times [L_y, U_y] \times [L_z, U_z]$. The reflection process associated with the UAV trajectory can be uniquely determined by the following equations:

$$\begin{aligned}&\int_0^t 1_{\{x_n(t)>L_x\}}\mathrm{d}L_n^x(t) = \int_0^t 1_{\{x_n(t)<U_x\}}\mathrm{d}U_n^x(t) \quad (8)\\ &= \int_0^t 1_{\{y_n(t)>L_y\}}\mathrm{d}L_n^y(t) = \int_0^t 1_{\{y_n(t)<U_y\}}\mathrm{d}U_n^y(t)\\ &= \int_0^t 1_{\{z_n(t)>L_z\}}\mathrm{d}L_n^z(t) = \int_0^t 1_{\{z_n(t)<U_z\}}\mathrm{d}U_n^z(t) = 0\end{aligned}$$

Figure 3 illustrates an example of the UAV altitude trajectory with reflective boundaries $L_z = 60m$ and $U_z = 100m$.

D. Content Request Queue Model

All users in the service group will send content request to each UAV for the specific content file, and we adopt the *request queue model* (in bits) to characterize the requests of the users. Assume that each content has the same file size. Let $Q_{nm}^j(t)$ denote the number of bits of the j-th content ($j \in \mathcal{J}^n$) requested by the m-th user in the n-th UAV. $Q_{nm}^j(t)$ evolves according to the dynamic equation [22]:

$$\mathrm{d}Q_{nm}^j(t) = [A_{nm}^j(t) - s_n^j(t)R_{nm}(t)]\mathrm{d}t + \mathrm{d}\boldsymbol{L}^{Q_{nm}^j}(t), \quad (9)$$

TABLE I
UAV ENERGY PROPULSION CONSUMPTION PARAMETERS

Parameter	Physical Meaning	Parameter	Physical Meaning
δ_e	Blade drag coefficient 0.012	s	Rotor solidity 0.05
Ω_e	Blade angular velocity 400 radians/sec	g	Gravity acceleration 9.8
R_e	Rotor radius 0.38	A	Rotor disc area 0.79
ρ	Air density 1.225	M_{UAV}	UAV mass 4 kg
κ_p	Induced power factor 0.1	$\boldsymbol{v}$	UAV velocity

where $A_{nm}^j(t)$ is the request arrival rate (bits/sec) with $\mathbb{E}[A_{nm}^j] = \bar{A}, \forall n, m, j$. $R_{nm}(t)$ is the achievable data rate (bits/sec) of the n-th UAV to the user m as defined in (6). $\boldsymbol{L}^{Q_{nm}^j}(t)$ denotes the reflection process associated with the lower request queue boundary $Q_{nm}^j(t) = 0$ such that the request queue $Q_{nm}^j(t)$ will not go below zero. The reflection process $\boldsymbol{L}^{Q_{nm}^j}(t)$ can be uniquely determined by

$$\int_0^t 1_{\{Q_{nm}^j(t)<0\}}\mathrm{d}\boldsymbol{L}^{Q_{nm}^j}(t) = 0. \quad (10)$$

Note that the request queue $Q_{nm}^j(t)$ is a virtual queue and hence, there is no upper bound on $Q_{nm}^j(t)$.

E. UAV Energy Consumption and Recharging Model

In this paper, all UAVs will be able to recharge the energy and cache content when it hovers over the charging station within the distance of D_0. Let $E_n(t)$ denote the remaining energy (in Joule) in the battery of the n-th UAV. It follows a dynamic equation given by:

$$\begin{aligned}\mathrm{d}E_n(t) = &\left(-P_n(t) - E_n^p(t) + E_n^{ch}(t)\right)\mathrm{d}t + \mathrm{d}L_n^E(t)\\ &-\mathrm{d}\boldsymbol{U}_n^E(t). \quad (11)\end{aligned}$$

$E_n^p(t) = (c_1 + c_2\left\|\boldsymbol{v}_n(t)\right\|_2^2)$ denotes the UAV propulsion consumption rate (Joule/sec) which depends on the UAV velocity with two terms: profile power and induced power [39]–[42], where $c_1 = \frac{\delta_e}{8}\rho s A \Omega_e^3 R_e^3 + \kappa_p \frac{(M_{UAV}g)^{3/2}}{\sqrt{2\rho A}}$ and $c_2 = \frac{3\delta_e}{8\Omega_e^2 R_e^2}\rho s A \Omega_e^3 R_e^3$. The detailed parameters of the UAV propulsion consumption are listed in Table I.

Given a certain energy level $E_n(t)$ in the battery, the n-th UAV has an option to continue it's journey to serve user requests or to return to the charging station to replenish the battery. We assume there is an efficient power transfer mechanism from the charging station to the UAV such as wireless power transfer [43], [44] or laser power transfer [45]. However, the detailed mechanism is outside the scope of this paper. The UAV energy charging rate in (11) is given by:

$$E_n^{ch}(t) = a\mathbf{1}(D_n^{ch}(t) \le D_0), \quad (12)$$

where a is the energy transfer rate (Joule/sec) from the charging station to the UAV, and $D_n^{ch}(t)$ denotes the distance from UAV n to the charging region. In addition, we assume that the energy charging rate is larger than the consumption rate in the charging region: $a \ge c_1 + c_2 v_c^2$.

$\boldsymbol{L}_n^E(t)$ and $\boldsymbol{U}_n^E(t)$ in (11) denote the reflection processes associated with the lower and upper energy queue boundary $E_n(t)=0$ and $E_n(t)=N_E$ such that when the UAV energy buffer is empty, no energy can be consumed and when the UAV energy buffer is full, additional energy cannot be replenished. The reflection processes $\boldsymbol{L}_n^E(t)$ and $\boldsymbol{U}_n^E(t)$ together with the admissible control constraint on the available energy defined in Section III ensures the energy queue length $E_n(t)$ lies in the domain $[0, N_E]$. The energy queue reflection process can be uniquely determined by

$$\int_0^t 1_{\{E_n(t)<0\}} \mathrm{d}\boldsymbol{L}_n^E(t) = 0. \tag{13}$$

$$\int_0^t 1_{\{E_n(t)>N_E\}} \mathrm{d}\boldsymbol{U}_n^E(t) = 0. \tag{14}$$

F. UAV Cache Content Recharging Model

Beside the energy charging, the charging station can also update the content in the UAV cache buffer. When the n-th UAV gets to the charging region, the cache content states are determined by the following equation:

$$s_n^j(t) = \begin{cases} 1, & \text{if } j = \arg\max \sum_m Q_{nm}^j(t) \\ 0, & \text{otherwise} \end{cases}, \forall j \in \mathcal{J}^n.$$

Specifically, UAVs will be charged with the most popular content file into the caching buffer. Instantaneous popularity at time t can be characterized by the instantaneous request queue length $\sum_m Q_{nm}^j(t)$. The content with the longest request queue $j^* = \arg\max \sum_m Q_{nm}^j(t)$ will be replenishment at the n-th UAV buffer.

Remark 1: We assume that the content recharging can be completed within a limited time using mmWave beamforming when the UAV is in close proximity with the charging station. In practice, the energy charging will take longer time compared with content charging. Furthermore, the UAV will always be triggered by insufficient energy to return to the charging station. Hence, the content charging will always be completed well before the energy charging.

III. UAV Trajectory and Transmission Control Dynamic Optimization

In this section, we shall design a dynamic trajectory and radio resource control policy for the multi-UAV system by considering an infinite horizon stochastic differential game. The dynamic policy is adaptive to the instantaneous request queue length (QSI) $Q_n(t)$, the instantaneous energy queue length (ESI) $E_n(t)$, the instantaneous channel condition (CSI) $G_n(t)$ as well as the current location of the UAV $\boldsymbol{X}_n(t)$. Adaptation to the CSI reveals **good transmission opportunities** induced by time-varying channel fading between the UAV and the user. Adaptation to the QSI reveals the **dynamic urgency of individual content flows**. Adaptation to the ESI reveals the **dynamic urgency induced by the battery life**. We first define the admissible control policy which considers multi-UAVs' safe operation conditions such as the finite energy constraint. We then formulate the problem as an N-player stochastic differential game and derive the associated optimality condition (Nash equilibrium).

A. UAV Trajectory and Radio Resource Control Admissible Policy

Given the system model in Section III, we now introduce the UAV control policy for content-centric caching network. Let $\boldsymbol{S}_n = \{\boldsymbol{X}_n, \boldsymbol{Q}_{nm}, E_n, G_{nm}\} \in \mathcal{S}_n$ denotes the local system state at the n-th UAV and $\boldsymbol{S}^N = \{\boldsymbol{S}_1, \ldots, \boldsymbol{S}_N\} \in \mathcal{S}^N := \mathcal{S}_1 \times \ldots \times \mathcal{S}_N$ denotes the global system state. Based on the system states, each UAV n implements an admissible control policy to determine the transmission power P_n^N, the user to serve $\{u_{nm}^N\}_{m\in\mathcal{M}}$ and the trajectory control $\boldsymbol{v}_n^N = [\boldsymbol{v}_n^{Nx}, \boldsymbol{v}_n^{Ny}, \boldsymbol{v}_n^{Nz}]^T$. We define the stationary admissible control policy Ω^N as below.

Definition 1 Admissible UAV Control Policy: Let $\mathfrak{U}$ denote the set of admissible control policies. The admissible control policy of the n-th UAV $\Omega_n^N \in \mathfrak{U}$ is a mapping from the system state $\boldsymbol{S}^N(t)$ to the trajectory and transmission actions $\{P_n^N, \{u_{nm}^N\}_{m\in\mathcal{M}}, \boldsymbol{v}_n^N\}$. In addition, the actions should satisfy the following constraints:

- $\left\|\boldsymbol{v}_n^N(t)\right\|_2 \le v_c$, where v_c denotes the UAV maximum speed in m/s;
- Under the admissible policy, the energy state and the control of the n-th UAV should satisfy

$$\begin{aligned} E_n(s) - \int_s^t (P_n(\tau) + c_1 + c_2 \left\|\boldsymbol{v}_n(\tau)\right\|_2^2 - E_n^{ch}(\tau))\mathrm{d}\tau - \boldsymbol{U}_n^E(t) \\ \ge \frac{(c_1 + c_2 v_c^2)}{v_c} (\Big\| \boldsymbol{X}(s) + \int_s^t \boldsymbol{v}_n(\tau)\mathrm{d}\tau + \int_s^t \boldsymbol{\sigma}_{\boldsymbol{X}} \mathrm{d}\boldsymbol{W}_n^X(\tau) \\ + \boldsymbol{L}_n^X(\mathrm{t}) - \boldsymbol{U}_n^X(t) - \boldsymbol{X}^{ch} \Big\|_2 - D_0) \end{aligned} \tag{15}$$

.

- The control policy is a stationary unichain policy. The request queuing system $Q_{nm}^j(t), \forall n \in \mathcal{N}, m \in \mathcal{M}, j \in J^n$ under the policies $\Omega_1^N \times \ldots \times \Omega_n^N$ is stable with $\lim_{T\to\infty} \frac{1}{T} \int_0^T \sum_{n,m,j} \mathbb{E}^{\Omega^N}[Q_{nm}^j(t)]\mathrm{d}t < \infty$.

The first condition is due to the constraint on the maximum speed of a UAV. The second condition is the safety constraint, which ensures that each UAV must have enough energy to get to the charging station with the highest speed. In the third condition, we restrict our attention to stationary unichain policies, i.e., the controlled stochastic process under Ω has a single recurrent class (and possibly some transient states) [27]. Note that such assumptions are reasonable as the optimal policy is also stationary[4].

B. Infinite Horizon Ergodic Differential Game Formulation

We formulate the problem as an infinite horizon ergodic stochastic differential game, where the UAV control policies $\{\Omega_1^N, \ldots, \Omega_n^N\}$ are jointly designed to minimize its average power consumption cost, its average collision cost and the average request delay cost. Specifically, each UAV is a player targeted to optimize the following ergodic cost:

[4]In this paper, we focus on the set of the addmissible control policies which give irreducible and ergodic markov process with well-defined steady state distribution.

Problem 1: Ergodic Stochastic Differential Game for multi-UAV Trajectory and Transmission Control

$$\min_{\Omega_n^N \in \mathfrak{U}} J^n(\Omega_n^N, \Omega_{-n}^N) \tag{16}$$

$$= \limsup_{T\to\infty} \mathbb{E}^{\Omega^N} \left[\frac{1}{T} \Big[\int_0^T \left[c_n(\boldsymbol{S}_n(t), \boldsymbol{S}_{-n}(t), \Omega_n^N) \right] dt + \varrho_1 (L_n^x(T) + L_n^y(T) + L_n^z(T)) + \varrho_2 (\boldsymbol{U}_n^x(T) + \boldsymbol{U}_n^y(T) + \boldsymbol{U}_n^z(T)) \Big] + \rho_3 L_n^E(T) \right]$$

$$s.t. \quad d\boldsymbol{S}_n(t) = F_n(\boldsymbol{S}_n(t), \boldsymbol{S}_{-n}(t), \Omega_n^N) dt + \boldsymbol{D}_n(\boldsymbol{S}_n(t)) d\boldsymbol{W}_n(t) + d\boldsymbol{L}_n(t) - d\boldsymbol{U}_n(t). \tag{17}$$

The per-stage cost function $c_n(\boldsymbol{S}_n, \boldsymbol{S}_{-n}, \Omega_n^N)$ *is defined as:* $c_n(\boldsymbol{S}_n, \boldsymbol{S}_{-n}, \Omega_n^N) = P_n^N + \alpha c_n^{delay}(\boldsymbol{S}_n) + \frac{1}{N}\sum_{k\neq n} \frac{\beta}{\gamma + \|\boldsymbol{X}_n - \boldsymbol{X}_k\|_2^2}$, [5]*where the first term in the cost function denotes the transmission power cost, and the second term is the users request delay cost* $c_n^{delay}(\boldsymbol{S}_n) = \sum_{m\in\mathcal{M}, j\in\mathcal{J}^n} Q_{nm}^j$ *with cost weighting* $\alpha > 0$*. The third term refers to the normalized UAV anti-collision cost and aims to penalize the UAVs collision events*[6] *with cost weighting* $\beta > 0$ *and* $\gamma > 0$*.* $\boldsymbol{S}_n(t) = [E_n(t), Q_{nm}^j(t), \boldsymbol{X}_n(t), G_{nm}(t)]_{m\in\mathcal{M}, j\in\mathcal{J}^n}^T$ *is the local state of the* n*-th UAV and* $S_{-n}(t)$ *is the states of all other UAVs. Note that the last two terms in the ergodic control cost (16) refer to the penalty associated with the UAVs displacement at the lower and upper boundary of the 3D flight region, where* ϱ_1 *and* ϱ_2 *are the penalty rates. From (7), (9) and (11), the dynamic drift can be written as* $F_n(\boldsymbol{S}_n, \boldsymbol{S}_{-n}, \Omega_n^N) = [(-P_n^N - E_n^p + E_n^{ch}), A_{nm}^j - \boldsymbol{s}_n^j R_{nm}(\Omega_n^N), \boldsymbol{v}_n^{Nx}, \boldsymbol{v}_n^{Ny}, \boldsymbol{v}_n^{Nz}, -\frac{1}{2} a_H G_{nm}]_{m\in\mathcal{M}, j\in\mathcal{J}^n}^T$*, and the diffusion term* $\boldsymbol{D}_n = diag(\boldsymbol{0}_{1+MJ}, \sigma_X^x, \sigma_X^y, \sigma_X^z, 2\sqrt{a_H G_{n1}}, \ldots, 2\sqrt{a_H G_{nM}})$*. The reflection process* $\boldsymbol{L}_n(t)$ *and* $\boldsymbol{U}_n(t)$ *of the* $n-$*th UAV are defined in (8)-(14).*

C. Optimality Condition for Multi-UAV Control Problem

The optimal solution for the stochastic differential game (SDG) in (16)-(17) is defined as the Nash equilibrium (NE):

Definition 2 (Nash Equilibrium of the Stochastic Differential Game): The set of controls $\Omega_n^{N*}(\boldsymbol{S}^N) = \{P_n^{N*}, \{u_{nm}^{N*}\}_{m\in\mathcal{M}}, \boldsymbol{v}_n^{N*}\}$*,* $\forall n \in \mathcal{N}$ *constitute the Nash equilibrium for the* N*-UAVs SDG Problem (16)-*(17) *if*

$$J^n(\Omega_n^{N*}, \Omega_{-n}^{N*}) \leq J^n(\Omega_n^{N'}, \Omega_{-n}^{N*}), \forall n \in \mathcal{N}, \tag{18}$$

for any other alternative control $\{\Omega_1^{N'}, \ldots, \Omega_N^{N'}\}$*.*

Given the problem formulation in (16), the NE policy Ω^{N*} defined in (18) can be obtained by solving the following sufficient condition for NE:

[5]In this paper, the per-stage transmission power cost accounts for the penalty of using a large transmission power and can be denoted as the soft power constraint. The weight α and β can be interpreted as the corresponding Lagrange Multipliers which determine the Pareto trade off between the power, delay and the UAV collision cost. The precise operating point on the Pareto boundary is determined by the application scenario and is out of the scope of the paper.

[6]In practice, there are local anti-collision mechanisms in each UAV and the onboard controller will avoid the crashing of the UAV. However, the cost term is to penalize the triggering of such on-board anti-collision events.

Theorem 1 Sufficient Condition for the NE : Assume there exists functions $V_n^N(\boldsymbol{S}_n) \in \mathcal{C}^2(\mathcal{S}_n)$ *and* θ_n^N*,* $\forall n$ *that solves the following average cost HJB optimality equation*:

$$\min_{\Omega_n^N \in \mathfrak{U}} \Bigg[P_n^N + \sum_{m=1}^M \sum_{j=1}^J \alpha Q_{nm}^j + \frac{1}{N} \sum_{k\neq n} \frac{\beta}{\gamma + \|\boldsymbol{X}_n - \boldsymbol{X}_k\|_2^2} + \sum_{n=1}^N F_n(\boldsymbol{S}_n, \boldsymbol{S}_{-n}, \Omega_n^N)^T \cdot \nabla_{\boldsymbol{S}_n} V_n^N(\boldsymbol{S}_n) + \sum_{n=1}^N \frac{1}{2} tr\left(\boldsymbol{D}_n \boldsymbol{D}_n^T \nabla_{\boldsymbol{S}_n}^2 V_n^N(\boldsymbol{S}_n) \right) \Bigg] = \theta_n^N, \forall n \in \mathcal{N} \tag{19}$$

where $\nabla_{\boldsymbol{S}_n} V_n^N(\boldsymbol{S}_n)$ *is the gradient of* $V_n^N(\boldsymbol{S}_n)$*, and* $\nabla_{\boldsymbol{S}_n}^2 V_n^N(\boldsymbol{S}_n)$ *is the Hessian of* $V_n^N(\boldsymbol{S}_n)$*. Then we have the following results:*

1. θ_n^N *is the optimal cost for UAV* n *in Problem (16), and* $V_n^N(\boldsymbol{S}_n)$ *is the optimal value function, with the boundary conditions*

$$\nabla_{x_n} V_n^N(\boldsymbol{S}_n)\big|_{x_n = L_x} = \nabla_{y_n} V_n^N(\boldsymbol{S}_n)\big|_{y_n = L_y} = \nabla_{z_n} V_n^N(\boldsymbol{S}_n)\big|_{z_n = L_z} = -\varrho_1; \tag{20}$$

$$\nabla_{x_n} V_n^N(\boldsymbol{S}_n)_{x_n = U_x} = \nabla_{y_n} V_n^N(\boldsymbol{S}_n)\big|_{y_n = U_y} = \nabla_{z_n} V_n^N(\boldsymbol{S}_n)\big|_{z_n = U_z} = \varrho_2; \tag{21}$$

$$\nabla_{Q_{nm}^j} V_n^N(\boldsymbol{S}_n)\Big|_{Q_{nm}^j = 0} = 0, \forall j \in \mathcal{J}, m \in \mathcal{M}, \tag{22}$$

$$\nabla_{E_n} V_n^N(\boldsymbol{S}_n)\big|_{E_n = N_E} = 0; \tag{23}$$

$$\nabla_{E_n} V_n^N(\boldsymbol{S}_n)\big|_{E_n = 0} = -\rho_3; \tag{24}$$

2. If admissible control $\{P_n^{N*}, \{u_{nm}^{N*}\}_{m\in\mathcal{M}}, \boldsymbol{v}_n^{N*}\}$ *attains the minimum of the L.H.S. of (19), then* $\Omega_n^{N*}(\boldsymbol{S}) = \{P_n^{N*}, \{u_{nm}^{N*}\}_{m\in\mathcal{M}}, \boldsymbol{v}_n^{N*}\}$ *is the optimal control policy for Problem (16).*

Proof: Please see Appendix A. ■

Based on this, each UAV player n will focus on optimizing its long-term cost $J^n(\Omega_n^N, \Omega_{-n}^N)$ in a selfish manner and adapt its trajectory and power control to the system state. However, the problem is rather challenging as each UAV player should have access to the knowledge of its own state $\boldsymbol{S}_n(t)$ and other UAV players state feedback $\boldsymbol{S}_{-n}(t)$. Furthermore, in order to build the Nash equilibrium for this N-UAV SDG, the system of N coupled HJB equations in (19) should be jointly solved. When $N \geq 2$, the computational complexity grows exponentially.

Challenge 1: Solving the N-tuple of HJB equations (19) either analytically or numerically is highly non-trivial due to the huge dimension of the state and mutual coupling via the anti-collision cost.
Challenge 2 (Centralized Solution): Brute force optimization of (19) results in a centralized control policy which requires global real-time state observations $\{\boldsymbol{S}_1, \ldots, \boldsymbol{S}_N\}$ and they are very difficult to obtain.

IV. Decoupling by Mean-Field Analysis

In this section, we shall overcome the aforementioned challenges using mean-field limit analysis [28]–[30]. Specifically, the coupling in the states can be fully characterized by a mean-field equilibrium, which is specified by a pair of backward Hamiltonian-Jacobi-Bellman (HJB) and forward Fokker-Planck-Kolmogorov (FPK) equations. As such, conditioned on the mean-field equilibrium, the N-dim HJB equations in (19) can be decoupled and this can substantially reduce the complexity. Furthermore, the stationary policies in (19) can be reduced to decentralized policies in which the actions of the n-th UAV will be adaptive to the local states S_n only. We will then solve the mean-field equilibrium equation and the HJB equation by deriving a reduced-complexity optimality equation and analyze the structural properties of the mean-field optimal solution.

A. Mean-Field Analysis

We first define the following indistinguishable property to characterize the individual behavior in the mean-field equilibrium:

Definition 2: Indistinguishable Property of multi-UAV SDG Problem The multi-UAV wireless communication system states $\boldsymbol{S}^N(t)=\{\boldsymbol{S}_1(t),\ldots,\boldsymbol{S}_N(t)\}$ are said to be indistinguishable under the policy Ω^N, if the law of $\boldsymbol{S}^N(t)$, $\mathcal{P}(\cdot)$ is invariant by any permutation of N elements:

$$\mathcal{P}(\boldsymbol{S}_1(t),\ldots,\boldsymbol{S}_N(t))=\mathcal{P}(\boldsymbol{S}_{\pi(1)}(t),\ldots,\boldsymbol{S}_{\pi(N)}(t)), \quad (25)$$

where $\pi(\cdot)$ denotes the permutation operation over $\{1,\ldots,N\}$.

Remark 2: Note that to achieve the convergence of mean field equilibrium, indistinguishable condition defined in Def.2 should be satisfied, meaning that the mean field game comprises of generic players.

The system of coupled HJB equations in (19) can be rewritten in the form of:

$$\sum_{n=1}^{N}\mathcal{L}_n^{\Omega_n^{N*}}V_n^N(\boldsymbol{S}_n)+c_n(\boldsymbol{S}_n,\boldsymbol{S}_{-n},\Omega_n^{N*})=\theta_n^N,\forall n\in\mathcal{N}, \quad (26)$$

where $\mathcal{L}_n^{\Omega_n^{N*}}$ is defined as the n-th UAV dynamic generator with

$$\mathcal{L}_n^{\Omega_n^{N*}}V_n^N(\boldsymbol{S}_n)\overset{def}{=}F_n(\boldsymbol{S}_n,\boldsymbol{S}_{-n},\Omega_n^{N*})^T\cdot\nabla_{\boldsymbol{S}_n}V_n^N(\boldsymbol{S}_n)+\frac{1}{2}\mathrm{tr}\left(\boldsymbol{D}_n\boldsymbol{D}_n^T\nabla^2_{\boldsymbol{S}_n}V_n^N(\boldsymbol{S}_n)\right) \quad (27)$$

We then have the following results on the optimal admissible policy in (26):

Lemma 1: The system states $\boldsymbol{S}^N(t)=\{\boldsymbol{S}_1(t),\ldots,\boldsymbol{S}_N(t)\}$ under the optimal admissible policy in (26) are indistinguishable.

Proof: Please see Appendix B. ∎

Based on this, each UAV dynamic $\boldsymbol{S}_n(t)$ in the multi-UAV system has a unique ergodic invariant distribution with $\mu_n^N(\boldsymbol{S}_n)$, which satisfies the following Fokker-Planck-Kolmogorov (FPK) equation:

$$\mathcal{L}_n^{*\Omega_n^{N*}}\mu_n^N=0,\int_{\mathcal{S}_n}\mu_n^N(\boldsymbol{S}_n)\mathrm{d}\boldsymbol{S}_n=1,\forall n\in\mathcal{N}, \quad (28)$$

where $\mathcal{L}_n^{*\Omega_n^{N*}}$ is the formal adjoint operator of $\mathcal{L}_n^{\Omega_n^{N*}}$ defined in (26). Hence, all UAVs become generic and the problem P1 can be approximated by the following mean-field problem P2:

Problem 2 Mean-field Problem:

$$\min_{\Omega\in\mathfrak{U}}J(\Omega,\mu) \quad (29)$$

$$=\limsup_{T\to\infty}\big[\mathbb{E}^{\Omega}\big[\frac{1}{T}(\int_0^T c^{\Omega}(\boldsymbol{S}(t),\mu(t))dt+\varrho_1c_L(\boldsymbol{S}(T))+\varrho_2c_U(\boldsymbol{S}(T)))+\rho_3c_L^E(\boldsymbol{S}(T))\big]\big]$$

$$s.t.\ \ d\boldsymbol{S}(t)=F^{\Omega}(\boldsymbol{S}(t))dt+\boldsymbol{D}(\boldsymbol{S}(t))d\boldsymbol{W}(t)+d\boldsymbol{L}(t)-d\boldsymbol{U}(t), \quad (30)$$

where $J(\Omega,\mu)$ is optimized over the admissible control policy set $\Omega\in\mathfrak{U}$. $c^{\Omega}(\boldsymbol{S},\mu)=P+\alpha\sum_{m\in\mathcal{M},j\in\mathcal{J}}Q_m^j+\int_{S'}\frac{\beta}{\gamma+\|\boldsymbol{X}-\boldsymbol{X}'\|_2^2}\mu(S')dS'$, $c_L(\boldsymbol{S})=L^x(T)+L^y(T)+L^z(T)$, $c_U(\boldsymbol{S})=U^x(T)+U^y(T)+U^z(T)$, $c_L^E(\boldsymbol{S}(T))=L^E(T)$, and $F^{\Omega}(\boldsymbol{S})=[(-P-E^p+E^{ch}),A_m-\boldsymbol{s}^jR_m,\boldsymbol{v}^x,\boldsymbol{v}^y,\boldsymbol{v}^z,-\frac{1}{2}a_HG_m]^T_{m\in\mathcal{M},j\in\mathcal{J}}$, and $\boldsymbol{D}(\boldsymbol{S})=diag(\boldsymbol{0}_{1+M},\sigma_X^x,\sigma_X^y,\sigma_X^z,2\sqrt{a_HG_m},\ldots,2\sqrt{a_HG_M})$. $\mu(t)$ denotes the stationary distribution of the process (30). The reflection processes $\boldsymbol{L}(t)$ and $\boldsymbol{U}(t)$ are defined similarly to (8)-(14), which satisfies $\int_0^t1_{\{x(t)<L_x\}}dL^x(t)=\int_0^t1_{\{x(t)>U_x\}}dU^x(t)=\int_0^t1_{\{y(t)<L_y\}}dL^y(t)=\int_0^t1_{\{y(t)>U_y\}}dU^y(t)=\int_0^t1_{\{z(t)<L_z\}}dL^z(t)=\int_0^t1_{\{z(t)>U_z\}}dU^z(t)=0$ and $\int_0^t1_{\{E(t)>N_E\}}d\boldsymbol{U}^E(t)=0$, $\int_0^t1_{\{E(t)<0\}}dL^E(t)$.

Note that the original state couplings $\frac{1}{N}\sum_{k\neq n}\frac{\beta}{\gamma+\|\boldsymbol{X}_n-\boldsymbol{X}_k\|_2^2}=\frac{1}{N}\sum_{k\neq n}\int_{\boldsymbol{S}'}\frac{\beta}{\gamma+\|\boldsymbol{X}'_n-\boldsymbol{X}'_k\|_2^2}\delta_{\boldsymbol{S}_k}\mathrm{d}S'$ are completely characterized by mean-field limiting term $\int_{S'}\frac{\beta}{\gamma+\|\boldsymbol{X}-\boldsymbol{X}'\|_2^2}\mu(S')\mathrm{d}S'$ as the empirical distribution produced by all members $\frac{1}{N}\sum_n\delta_{\boldsymbol{S}_n}(\mathrm{d}\boldsymbol{S})$ can be approximated by single UAV player's statistical property $\mu(S)$. In addition, the optimal solution for the mean-field problem in (29)-(30) is defined as the mean-field equilibrium (MFE):

Definition 3 Mean-Field Equilibrium for P2 : The controls $\Omega^*(\boldsymbol{S})=\{P^*,\{u_m^*\}_{m\in\mathcal{M}},\boldsymbol{v}^*\}$ achieves the mean-field equilibrium of the mean-field problem P2 if for all alternative policy Ω'

$$J(\Omega^*,\mu^*)\leq J(\Omega',\mu^*), \quad (31)$$

where μ^* denotes the stationary distribution induced by Ω^*.

The MFE policy $\Omega^*=\{P^*,\{u_m^*\}_{m\in\mathcal{M}},\boldsymbol{v}^*\}$ and stationary distribution defined in (31) can be obtained by solving the following sufficient condition for mean field equilibrium (HJB-FPK equations):

$$\min_{\Omega(\boldsymbol{S})}\left(\mathcal{L}^{\Omega}V(\boldsymbol{S})+c(\boldsymbol{S},\Omega(\boldsymbol{S}),\mu)\right)=\theta \quad (32)$$

$$\mathcal{L}^{*\Omega}\mu(\boldsymbol{S})=0,\int_{\mathcal{S}}\mu(\boldsymbol{S})\mathrm{d}\boldsymbol{S}=1, \quad (33)$$

where $\mathcal{L}^{\Omega}V(\boldsymbol{S})=F^{\Omega}(\boldsymbol{S})\cdot\nabla_{\boldsymbol{S}}V(\boldsymbol{S})+\frac{1}{2}\mathrm{tr}\left(\boldsymbol{D}\boldsymbol{D}^T\nabla^2_{\boldsymbol{S}}V(\boldsymbol{S})\right)$, and $\mathcal{L}^{*\Omega}$ is the formal adjoint operator of $\mathcal{L}^{\Omega}$.

Next, we shall establish that as $N\to\infty$, the original SDG problem P1 is asymptotically equivalent to the mean-field problem P2 in the following sense.

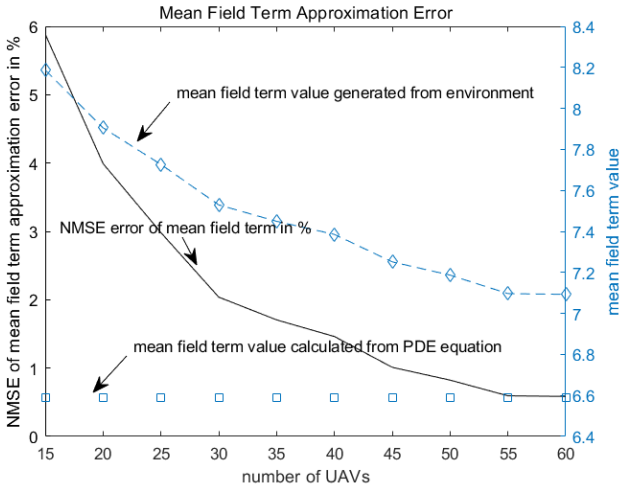


Fig. 4. Normalized mean field term approximation error with $M = 60$, $J = 10$.

Theorem 2 Asymptotic Equivalence of P1 and P2 : *Let* $(\theta_1^N, V_1^N), .., (\theta_1^N, V_1^N)$ *be the solution of (19) in P1 and* $\mu_1^N, \ldots, \mu_N^N$ *be the stationary distributions of the states induced by the optimal policy* $\{\Omega_1^{N*}, \ldots, \Omega_N^{N*},\}$ *of P1. We have the following results:*

1) $\sup_{n,k} \left(\left|\theta_n^N - \theta_k^N\right| + \left\|V_n^N - V_k^N\right\|_{\mathcal{C}^2(\mathcal{S})} + \left\|\mu_n^N - \mu_k^N\right\|_\infty \right) \to 0$ as $N \to \infty$;
2) any limit point (θ, V, μ) of $\{(\theta_1^N, V_1^N, \mu_1^N), \ldots, (\theta_N^N, V_N^N, \mu_N^N)\}$ satisfies the mean-field equilibrium in (32)-(33).

Proof: Please see the Appendix C.

As such, we can focus on solving the mean-field problem P2, which involves one HJB equation in (32) to determine the optimal value θ and value function $V(\cdot)$, and one FPK equation in (33) to determine stationary distribution $\mu(\cdot)$. Figure 4 illustrates the comparison of the actual mean field term $\frac{1}{N}\sum_{k\neq n}\int_{\boldsymbol{S}'} \frac{\beta}{\gamma + \left\|\boldsymbol{X}_n' - \boldsymbol{X}_k'\right\|_2^2} \delta_{\boldsymbol{S}_k} \mathrm{d}S'$ and the mean field term computed from (32)-(33). It can be seen that the proposed mean field approximation solution is quite accurate for number of UAV $N \to 50$.

B. Structural Properties of the Mean-Field Optimal Solution

According to equation (33), we shall derive the optimal solution based on the value function and analyze some key policy structural properties for the multi-UAV mean-field control. One challenge to solve (32)-(33) is the safety constraints in the admissible action space. Specifically, inequality (15) is a compulsory state-action coupled constraint limited on the power and trajectory control, which has no closed-form solution.

Challenge 3: There is no closed-form solution for the action given the value function due to the complex state-action coupled constraints (15).

In the following, we shall transform the state-action constraint into a solely state-dependent *inward pointing boundary condition* for the HJB equation so as to ensure that the system state will not leave the safety region. As Fig. 5 illustrates, the admissible policy should guarantee the state trajectory in the safety domain instead of reaching outwards to the boundary. These are formally stated in the following lemma.

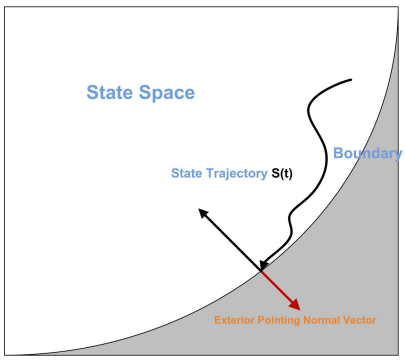


Fig. 5. Illustration of the inward pointing condition when state trajectory reaches the boundary of the state space.

Lemma 2: Transformation of the Instantaneous Safety Constraint: The mean-field equilibrium optimization in (32) is equivalent to the following equations:

$$\begin{aligned}
&\min_{\{P,\boldsymbol{u},\boldsymbol{v}\}} \Big[P + \sum_{m=1}^{M}\sum_{j=1}^{J} \alpha Q_m^j + \int_{S'} \frac{\beta}{\gamma + \left\|\boldsymbol{X} - \boldsymbol{X}'\right\|} \mu(S') dS' \\
&+ \sum_{m=1}^{M}\sum_{j=1}^{J} \nabla_{Q_m^j} V(\boldsymbol{S}) (A_m^j - \boldsymbol{s}^j u_m W \log_2(1 + |H_m|^2 P)) \\
&+ \nabla_E V(\boldsymbol{S})(E^{ch} - P - E^p) + \boldsymbol{v}^T \cdot \nabla_{\boldsymbol{X}} V(\boldsymbol{S}) \\
&- \frac{1}{2} a_H \sum_{m=1}^{M} G_m \nabla_{G_m} V(\boldsymbol{S}) \\
&+ \frac{1}{2} tr\left(\boldsymbol{\sigma}_X \boldsymbol{\sigma}_X^T \nabla_{\boldsymbol{X}}^2 V(\boldsymbol{S})\right) + a_H \sum_{m=1}^{M} G_m \nabla_{G_m}^2 V(\boldsymbol{S}) \Big] \\
&= \theta, \boldsymbol{S} \in \mathcal{S}; \qquad (34)
\end{aligned}$$

with boundary conditions

$$\begin{aligned}
F^{\{P^*,\boldsymbol{u}^*,\boldsymbol{v}^*\}}(\boldsymbol{S}) \cdot \eta(\boldsymbol{S})\Big|_{E=\frac{(c_1+c_2 v_c^2)}{v_c}(D^{ch}(\boldsymbol{X})-D_0)} &\leq 0, \qquad (35)\\
\nabla_{\boldsymbol{X}^x} V(\boldsymbol{S})|_{\boldsymbol{X}^x = L_x} &= \nabla_{\boldsymbol{X}^y} V(\boldsymbol{S})|_{\boldsymbol{X}^y = L_y} \qquad (36)\\
&= \nabla_{\boldsymbol{X}^z} V(\boldsymbol{S})|_{\boldsymbol{X}^z = L_z} = -\varrho_1;\\
\nabla_{\boldsymbol{X}^x} V(\boldsymbol{S})_{\boldsymbol{X}^x = U_x} &= \nabla_{\boldsymbol{X}^y} V(\boldsymbol{S})|_{\boldsymbol{X}^y = U_y} \qquad (37)\\
&= \nabla_{\boldsymbol{X}^z} V(\boldsymbol{S})|_{\boldsymbol{X}^z = U_z} = \varrho_2;\\
\nabla_{Q_m^j} V(\boldsymbol{S})\Big|_{Q_m^j = 0} &= 0, \forall j \in \mathcal{J}, m \in \mathcal{M}, \qquad (38)\\
\nabla_E V(\boldsymbol{S})|_{E=N_E} &= 0; \qquad (39)\\
\nabla_E V(\boldsymbol{S})|_{E=0} &= -\rho_3; \qquad (40)
\end{aligned}$$

where $\eta(\boldsymbol{S})$ *denotes the exterior pointing normal vector defined on the boundary* $E = \frac{(c_1+c_2 v_c^2)}{v_c}(D^{ch}(\boldsymbol{X}) - D_0)$.

Proof: Please refer to Appendix D.

Based on this, the admissible mean-field optimal control actions can be obtained given the value function as follows: ■

Corollary 1 Mean-Field Optimal Control Actions: The optimal power control policy has the following water-filling

structure (41), as shown at the bottom of this page, where u_m^* is the optimal user scheduling control as given by:

$$u_m^* = \begin{cases} 1, & m = m^*, \\ 0, & otherwise. \end{cases} \tag{42}$$

$$m^* = \arg\max_{n\in[1,M]} \sum_{j=1}^{J} s^j \frac{\partial V(\boldsymbol{S})}{\partial Q_m^j} \left(W \log_2 (\frac{\frac{W}{ln2}\frac{\partial V(\boldsymbol{S})}{\partial Q_m^j}\frac{\sigma^2}{H_m}}{1-\frac{\partial V(\boldsymbol{S})}{\partial E}}) \right)^+. \tag{43}$$

The optimal UAV velocity policy has the following structure (44), as shown at the bottom of the next page, where $V(S)$ and $\mu(S)$ are the joint solutions of HJB-FPK equations (33).

Proof: Please refer to Appendix E. ∎

V. Data-Driven DNN Solution for Solving HJB-FPK Equations

The remaining task is to solve the HJB-FPK PDEs (32)-(33). In this section, we will focus on developing a data-driven deep learning algorithm to learn the value function $V(\boldsymbol{S})$, distribution $\mu(\boldsymbol{S})$ and control policy $\Omega(\boldsymbol{S})$. Specifically, we construct a model-specific DNN by exploiting the HPM method [33], [34]. The learning agent has system state realizations CSI, QSI, UAV energy queue and UAV position as the real-time input, and produces transmission and trajectory control actions as the real-time output. Compared with generic DNN structures, the proposed structure is more $expressive$ in the sense that it can represent the $V(\boldsymbol{S})$ and $\mu(\boldsymbol{S})$ more efficiently with less number of parameters.

A. Problem-Specific DNN Architecture Based on Homotopy Perturbation Method

We will first construct the DNN architecture for the PDE learning. Consider the I layers DNN parameterized approximation structure for the value function and stationary distribution as follows:

$$\hat{V}^I(\boldsymbol{S};\boldsymbol{w}_H) = \sum_{i=0}^{I} V_i(\boldsymbol{S}), \tag{45}$$

$$\hat{\mu}^I(\boldsymbol{S};\boldsymbol{w}_F) = \sum_{i=0}^{I} \mu_i(\boldsymbol{S}), \tag{46}$$

where $\boldsymbol{w}_H$ and $\boldsymbol{w}_F$ denote the neural network tunable weighting parameters for HJB and FPK equations respectively. $\{V_i(S), \mu_i(S)\}$ denotes the pair of the ith homotopy deformation layer approximation of value function and distribution $\{V(\boldsymbol{S}), \mu(\boldsymbol{S})\}$.

Based on the HPM method [33], [34], with the proper selection of the convergence control parameters $\{\boldsymbol{w}_H, \boldsymbol{w}_F\}$, the homotopy-series solution $\{\hat{V}^I(x;\boldsymbol{w}_H), \hat{\mu}^I(x;\boldsymbol{w}_F)\}$ can be constructed to approximate the solution of the nonlinear PDE system (32)-(33) accurately

$$\hat{V}^I(\boldsymbol{S};\boldsymbol{w}_H) \approx V(\boldsymbol{S}),\ \hat{\mu}^I(\boldsymbol{S};\boldsymbol{w}_F) \approx \mu(\boldsymbol{S}).$$

The homotopy solutions $\{\hat{V}^I(x;\boldsymbol{w}_H), \hat{\mu}^I(x;\boldsymbol{w}_F)\}$ are governed by the following HPM deformation equations:

$$\begin{aligned}&(1-\alpha_H(q;\boldsymbol{w}_H))\mathcal{I}^H[\hat{V}^I(\boldsymbol{S};\boldsymbol{w}_H) - V_0(\boldsymbol{S})] \\ &\quad = \beta_H(q;\boldsymbol{w}_H)\left(\mathcal{B}^H[\hat{V}^I(\boldsymbol{S};\boldsymbol{w}_H), \hat{\mu}^I(\boldsymbol{S};\boldsymbol{w}_F)]\right)\end{aligned} \tag{47}$$

$$\begin{aligned}&(1-\alpha_F(q;\boldsymbol{w}_F))\mathcal{I}^F[\hat{\mu}^I(\boldsymbol{S};\boldsymbol{w}_F) - \mu_0(\boldsymbol{S})] \\ &\quad = \beta_F(q;\boldsymbol{w}_F)q\left(\mathcal{B}^F[\hat{V}^I(\boldsymbol{S};\boldsymbol{w}_H), \hat{\mu}^I(\boldsymbol{S};\boldsymbol{w}_F)]\right)\end{aligned} \tag{48}$$

where $\alpha_H(q;\boldsymbol{w}_H)$, $\beta_H(q;\boldsymbol{w}_H)$, $\alpha_F(q;\boldsymbol{w}_F)$ and $\beta_F(q;\boldsymbol{w}_F)$ are the deformation functions with $\alpha_H(0;\boldsymbol{w}_H) = \alpha_F(0;\boldsymbol{w}_F) = \beta_H(0;\boldsymbol{w}_H) = \beta_F(0;\boldsymbol{w}_F) = 0$ and $\alpha_H(1;\boldsymbol{w}_H) = \alpha_F(1;\boldsymbol{w}_F) = \beta_H(1;\boldsymbol{w}_H) = \beta_F(1;\boldsymbol{w}_F) = 1$. $V_0(\boldsymbol{S})$ and $\mu_0(\boldsymbol{S})$ are the initial approximation of the solution to the PDE system. $q \in [0,1]$ denotes the embedding perturbation parameter. $\mathcal{I}^H$ and $\mathcal{I}^F$ are the identity operators with $\mathcal{I}^H[V] = V$, $\mathcal{I}^F[\mu] = \mu$. $\mathcal{B}^H$ and $\mathcal{B}^F$ are the nonlinear differential operators of the HJB and FPK equations respectively given by:

$$\begin{aligned}&\mathcal{B}^H[V,\mu] \\ &= P^* + \sum_{m=1}^{M}\sum_{j=1}^{J} \alpha Q_m^j + \int_{S'} \frac{\beta}{\gamma + \left\|\boldsymbol{X} - \boldsymbol{X}'\right\|} \mu(S')\mathrm{d}S' \\ &\quad + \sum_{m=1}^{M}\sum_{j=1}^{J} \nabla_{Q_m^j} V(A_m^j - \boldsymbol{s}^j u_m^* W \log_2(1+|H_m|^2 P^*)) \\ &\quad + \nabla_E V(E^{ch} - P^* - E^p) + \boldsymbol{v}^{*T}\cdot\nabla_{\boldsymbol{X}} V \\ &\quad + \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\sigma}_X \boldsymbol{\sigma}_X^T \nabla_{\boldsymbol{X}}^2 V\right) + \frac{1}{2} a_H \sum_{m=1}^{M} G_m \nabla_{G_m} V \\ &\quad + a_H \sum_{m=1}^{M} G_m \nabla_{G_m}^2 V - \theta\end{aligned} \tag{49}$$

$$\begin{aligned}&\mathcal{B}^F[V,\mu] \\ &= \sum_{m=1}^{M}\sum_{j=1}^{J} \nabla_{Q_m^j}\left((A_m^j - \boldsymbol{s}^j u_m^* W \log_2(1+|H_m|^2 P^*))\mu\right) \\ &\quad + \nabla_E\left((-P^* - E^p + E^{ch})\mu\right) + \nabla_{\boldsymbol{X}}\left(\boldsymbol{v}^{*T}\cdot\mu\right) \\ &\quad - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\sigma}_X \boldsymbol{\sigma}_X^T \nabla_{\boldsymbol{X}}^2 \mu\right) + \frac{1}{2} a_H \sum_{m=1}^{M} \nabla_{G_m}\left(G_m \mu\right) \\ &\quad - a_H \sum_{m=1}^{M} \nabla_{G_m}^2 \left(G_m \mu\right)\end{aligned} \tag{50}$$

As the embedding parameter q varies from 0 to 1, the approximated solutions $\{\hat{V}^I(S;\boldsymbol{w}_H), \hat{\mu}^I(S;\boldsymbol{w}_F)\}$

$$P^* = \begin{cases} \sum_{j,m=1}^{J,M} s^j u_m^* (\frac{\frac{W}{ln2}\frac{\partial V(\boldsymbol{S})}{\partial Q_m^j}}{1-\frac{\partial V(\boldsymbol{S})}{\partial E}} - \frac{\sigma^2}{H_m})^+, & \text{if} \\ & E > D^{ch}(\boldsymbol{X})(\frac{c_1}{v_c} + c_2 v_c), \\ 0 & \text{otherwise}, \end{cases} \tag{41}$$

vary from the initial guess $\{V_0(\boldsymbol{S}), \mu_0(\boldsymbol{S})\}$ to the exact solutions $\{V(\boldsymbol{S}), \mu(\boldsymbol{S})\}$. The approximated solutions $\{\hat{V}^I(S;\boldsymbol{w}_H), \hat{\mu}^I(S;\boldsymbol{w}_F)\}$ are subject to the boundary conditions and the normalization condition [7]

$$\begin{aligned}
&\nabla_{\boldsymbol{X}^x}\hat{V}^I(\boldsymbol{S};\boldsymbol{w}_H)\Big|_{\boldsymbol{X}^x=L_x}\\
&= \nabla_{\boldsymbol{X}^y}\hat{V}^I(\boldsymbol{S};\boldsymbol{w}_H)\Big|_{\boldsymbol{X}^y=L_y}, \qquad (51)\\
&= \nabla_{\boldsymbol{X}^z}\hat{V}^I(\boldsymbol{S};\boldsymbol{w}_H)\Big|_{\boldsymbol{X}^z=L_z} = -\varrho_1;\\
&\nabla_{\boldsymbol{X}^x}\hat{V}^I(\boldsymbol{S};\boldsymbol{w}_H)\Big|_{\boldsymbol{X}^x=U_x}\\
&= \nabla_{\boldsymbol{X}^y}\hat{V}^I(\boldsymbol{S};\boldsymbol{w}_H)\Big|_{\boldsymbol{X}^y=U_y},\\
&= \nabla_{\boldsymbol{X}^z}\hat{V}^I(\boldsymbol{S};\boldsymbol{w}_H)\Big|_{\boldsymbol{X}^z=U_z} = \varrho_2; \qquad (52)\\
&\nabla_{Q_m^j}\hat{V}^I(\boldsymbol{S};\boldsymbol{w}_H)\Big|_{Q_m^j=0} = 0, \forall j \in \mathcal{J}, m \in \mathcal{M}; \qquad (53)\\
&\nabla_E\hat{V}^I(\boldsymbol{S};\boldsymbol{w}_H)\Big|_{E=N_E} = 0; \qquad (54)\\
&\nabla_E\hat{V}^I(\boldsymbol{S};\boldsymbol{w}_H)\Big|_{E=0} = -\rho_3; \qquad (55)\\
&\nabla_{\boldsymbol{X}^x}\hat{\mu}^I(\boldsymbol{S};\boldsymbol{w}_F)\big|_{\boldsymbol{X}^x=L_x,U_x} = \nabla_{\boldsymbol{X}^y}\hat{\mu}^I(\boldsymbol{S};\boldsymbol{w}_F)\big|_{\boldsymbol{X}^y=L_y,U_y}\\
&= \nabla_{\boldsymbol{X}^z}\hat{\mu}^I(\boldsymbol{S};\boldsymbol{w}_F)\big|_{\boldsymbol{X}^z=L_z,U_z} = 0; \qquad (56)\\
&\hat{\mu}^I(\boldsymbol{S};\boldsymbol{w}_F)\big|_{Q_m^j=0} = 0, \forall j \in \mathcal{J}, m \in \mathcal{M}; \qquad (57)\\
&\hat{\mu}^I(\boldsymbol{S};\boldsymbol{w}_F)\big|_{E=N_E} = 0; \qquad (58)\\
&\hat{\mu}^I(\boldsymbol{S};\boldsymbol{w}_F)\big|_{E=0} = 0;\\
&\int_{\boldsymbol{S}}\hat{\mu}^I(\boldsymbol{S};\boldsymbol{w}_F)\mathrm{d}\mathcal{S} = 1 \qquad (59)
\end{aligned}$$

By substituting the Ith order homotopy-series expansion w.r.t. embedding perturbation parameter $\{\sum_{i=0}^I V_i(\boldsymbol{S})q^i, \sum_{i=0}^I \mu_i(S)q^i\}$ into the (47)-(48), and equating the like-power of q, we have the deformation equation to construct the recursion relationship of the series expansion. By using the neural network to represent the unrolled recursion, the output of the ith layer can be represented as:

$$\begin{aligned}
V_i(\boldsymbol{S}) &= \chi_i\sum_{j=1}^{i-1}\boldsymbol{w}_H^{i,j}V_j(\boldsymbol{S}) \qquad (60)\\
&\quad +\boldsymbol{w}_H^{i,i}\mathcal{B}^H[V_{i-1}(\boldsymbol{S}), \mu_{i-1}(\boldsymbol{S})],\\
\mu_i(\boldsymbol{S}) &= \chi_i\sum_{j=1}^{i-1}\boldsymbol{w}_F^{i,j}\mu_j(\boldsymbol{S}) \qquad (61)\\
&\quad +\boldsymbol{w}_F^{i,i}\mathcal{B}^F[V_{i-1}(\boldsymbol{S}), \mu_{i-1}(\boldsymbol{S})],
\end{aligned}$$

[7]The boundary conditions of the value function and stationary distribution are derived by the reflecting boundary on the bounded state space.

where

$$\chi_i = \begin{cases}1, & i > 1,\\ 0, & i \le 1,\end{cases} \qquad (62)$$

The weightings $\boldsymbol{w}_H = \{\boldsymbol{w}_H^{i,j}, 1 \le j \le i \le I\}$ and $\boldsymbol{w}_F = \{\boldsymbol{w}_F^{i,j}, 1 \le j \le i \le I\}$ are the DNN tunable parameters to control the convergence during the learning. Fig.6 illustrates an example of the $I = 3$-layer DNN architecture, where each hidden layer is constructed to approximate the recursion relationship in (60)- (61). The final layer of the DNN generates the control actions output based on the real time state input and the approximated value function and stationary distribution.

B. Online Training Algorithm

Based on the problem-specific DNN architecture, we now describe our DNN based data-driven learning algorithm. Denote $\{\boldsymbol{S}^i, i \in \{1,\ldots,d\}\}$ and $\{D_v^i, D_\mu^i, i \in \{1,\ldots,d\}\}$ as the finite set of points and boundary operators on the value function and stationary distribution at boundaries according to (51)-(58). The proposed DNN aims to find appropriate DNN weightings $\{\boldsymbol{w}_H, \boldsymbol{w}_F\}$ to minimize the loss function as follows:

$$\begin{aligned}
&\min_{\{\boldsymbol{w}_H,\boldsymbol{w}_F\}} L(\boldsymbol{w}_H, \boldsymbol{w}_F)\\
&= E^{\hat{\Omega}^*}\Bigg[\Big|\mathcal{B}^H[\hat{V}^I(\boldsymbol{S};\boldsymbol{w}_H), \hat{\mu}^I(\boldsymbol{S};\boldsymbol{w}_F)]\Big|^2\\
&\quad +\Big|\mathcal{B}^F[\hat{V}^I(\boldsymbol{S};\boldsymbol{w}_H), \hat{\mu}^I(\boldsymbol{S};\boldsymbol{w}_F)]\Big|^2\\
&\quad +\sum_{i=1}^d\Big(\Big|D_v^i\left(\hat{V}^I(\boldsymbol{S}^i;\boldsymbol{w}_H)\right)\Big|^2\\
&\quad +\left|D_\mu^i\left(\hat{\mu}^I(\boldsymbol{S}^i;\boldsymbol{w}_F)\right)\right|^2\Big) + \Big|\int_{\boldsymbol{S}}\hat{\mu}^I(\boldsymbol{S};\boldsymbol{w}_F)\mathrm{d}\mathcal{S} - 1\Big|^2\Bigg]
\end{aligned} \qquad (63)$$

where $\hat{V}^I(\boldsymbol{S};\boldsymbol{w}_H)$ and $\hat{\mu}^I(\boldsymbol{S};\boldsymbol{w}_F)$ are the I layers DNN approximated solution of the HJB-FPK PDE equations. The first two terms in the loss function measure how closely the DNN approximated solution satisfies the PDE operator, and the second two terms $\sum_{i=1}^d\left(\left|D_v^i\left(\hat{V}^I\right)\right|^2 + \left|D_\mu^i\left(\hat{\mu}^I\right)\right|^2\right)$ denote the approximation error of the boundary conditions (51)-(58). The last term $\left|\int_{\boldsymbol{S}}\hat{\mu}^I(\boldsymbol{S};\boldsymbol{w}_F)\mathrm{d}\mathcal{S} - 1\right|^2$ denotes the approximation error of the distribution normalization condition (59). $\hat{\Omega}^*$ denotes the derived optimal control solution based on $\hat{V}^I(\boldsymbol{S};\boldsymbol{w}_H)$ and $\hat{\mu}^I(\boldsymbol{S};\boldsymbol{w}_F)$ according to (41)-(44). Note that the loss function depends on the online sampled realizations of system state only and is independent of the target optimal action. As such, the learning process is completely online and autonomous without requiring knowledge of the

$$\boldsymbol{v}^* = \begin{cases}\frac{-\nabla_{\boldsymbol{X}}V(\boldsymbol{S})}{\|\nabla_{\boldsymbol{X}}V(\boldsymbol{S})\|_2}\min(v_c, \left\|\frac{\nabla_{\boldsymbol{X}}V(\boldsymbol{S})}{2c_2\frac{\partial V(\boldsymbol{S})}{\partial E}}\right\|_2), & \text{if}\\ & E > D^{ch}(\boldsymbol{X})(\frac{c_1}{v_c} + c_2v_c),\\ \frac{-v_c(\boldsymbol{X}-\boldsymbol{X}^{ch})}{\|\boldsymbol{X}-\boldsymbol{X}^{ch}\|_2}, & \text{otherwise},\end{cases} \qquad (44)$$

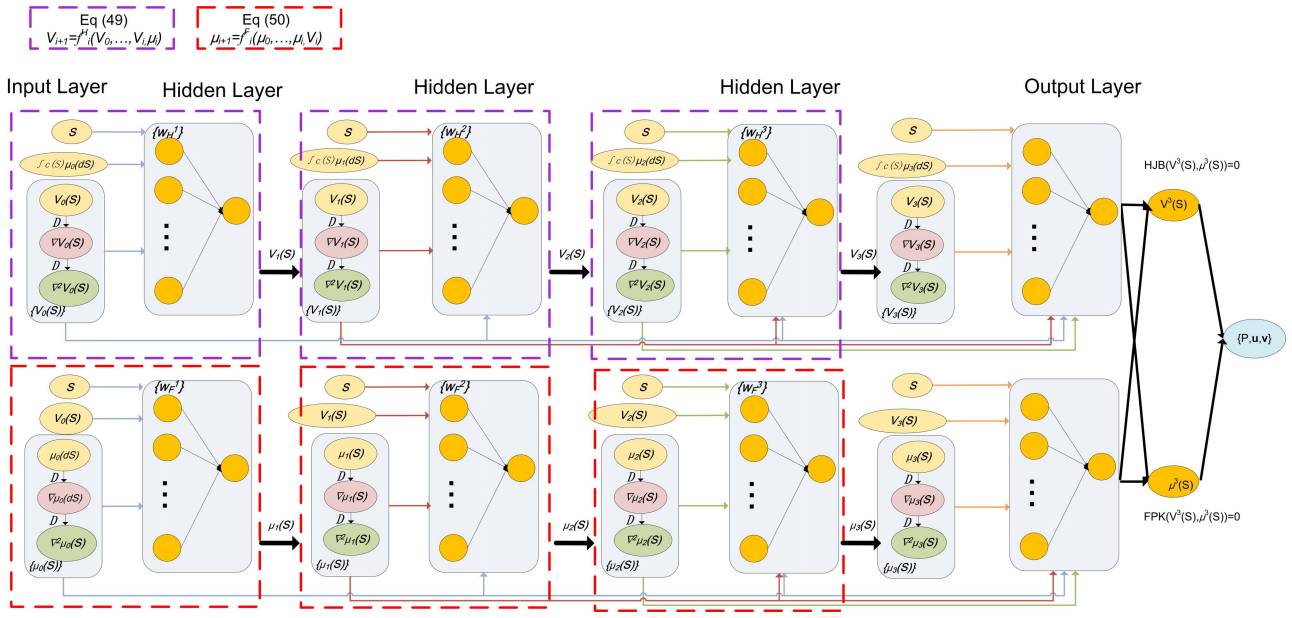


Fig. 6. Problem-specific deep neural network architecture based on (60)-(61) with the approximation order $I = 3$.

target optimal action to train the DNN. Based on this, we will apply a stochastic gradient descent approach to update the DNN parameters $\{\boldsymbol{w}_H, \boldsymbol{w}_F\}$ using the online sampled data. The detailed algorithm is summarized in Fig.7.

C. Convergence and Complexity Analysis

We now establish the convergence condition and the approximation power of the proposed DNN learning algorithm for the nonlinear PDEs by proving the approximation error bounds.

Theorem 3: Convergence of the Proposed Learning Algorithm: Assume that the initial approximation $V_0(\boldsymbol{S})$ and $\mu_0(\boldsymbol{S})$ satisfy the boundary conditions (51)-(58). Then we have $\lim_{I\to\infty} \hat{V}^I(\boldsymbol{S}) = V(\boldsymbol{S})$, $\lim_{I\to\infty} \hat{\mu}^I(\boldsymbol{S}) = \mu(\boldsymbol{S})$, and

$$\left\|\hat{V}^I(\boldsymbol{S}) - V(\boldsymbol{S})\right\| \leq \frac{\gamma_H^{I+1}}{1-\gamma_H} \left\|V_0(\boldsymbol{S})\right\|, \tag{64}$$

$$\left\|\hat{\mu}^I(\boldsymbol{S}) - \mu(\boldsymbol{S})\right\| \leq \frac{\gamma_F^{I+1}}{1-\gamma_F} \left\|\mu_0(\boldsymbol{S})\right\|, \tag{65}$$

where $0 < \gamma_H < 1$ and $0 < \gamma_F < 1$.

Proof: Please refer to Appendix F. ■

Remark 3: Note that the proposed DNN learning structure is capable of approximating the high dimensional HJB-FPK PDE solutions in the sense that the approximation error bound (convergence rate) is independent of the number of UAVs N, the number of users M and number of content files J. As a result, the complexity and training time of the proposed solution is free from the curse of dimensionality.

VI. Results and Discussions

In this section, we shall evaluate the performance of the proposed online data-driven learning scheme for the multi-UAV control through simulation results. As will be described later, the corresponding system setup for the simulation is summarized in Table II, and several benchmark schemes are considered for the performance comparison.

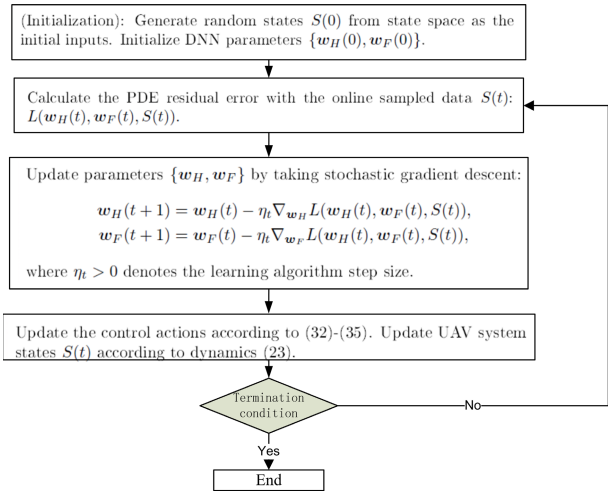


Fig. 7. Algorithm flowchart of the proposed data-driven learning algorithm.

A. Simulation Setup

In our simulation, the UAVs are constrained in the cube target region with a size of 200 × 200 × 50 m, and the initial point is set at the center of the region. We consider a multi-users scenario that the M users are grouped into several user clusters/hotspots, where each cluster is of the same size and uniformly distributed in the target region. The heights of the charging station and each user are assumed to be $z^{ch} = 50$ m and $\bar{z}_m = 1$ m, $\forall m \in \mathcal{M}$, respectively. We apply $12V$ Lipo battery for the UAV energy charging with charging current $5A$. The overall system setup for the simulation is given in Table II.

In the simulation, the continuous dynamic is discretized by the Euler-Maruyama method with the time slot $\tau = 5$ms. We compare the proposed learning scheme with the following baselines using numerical simulations:

TABLE II
SYSTEM SETUP FOR SIMULATIONS

Parameter	Value	Parameter	Value
W	20 MHz	D_0	20 m
$[L_x, L_y, L_z]$	$[-100, -100, 50]$	ρ_0	1
$[U_x, U_y, U_z]$	$[100, 100, 100]$	$A^j_{nm}, \forall n, m, j$	5Mbps
$\sigma^x_X, \sigma^y_X, \sigma^z_X$	0.5	$P_{\max}$	10 dBW
a_H	1	a	60W
$\boldsymbol{v}_c$	50m/s	N_E	37000 Joule
M_{UAV}	4 kg	α, β, γ	1
c_1	272.6	c_2	0.0337
ϱ_1	50	ϱ_2	50

- **(1) Throughput maximization offline design [19], [46]–[48]:** In this scheme, the UAV trajectory, power and user scheduling control are jointly optimized by maximizing the overall throughput: $\max_{\boldsymbol{X}(t), 0 \le P_n(t) \le P_{max}, \sum_m u_{nm}(t)=1} \sum_{t,m,j} s^j_n(t) R_{nm}(t)$, s.t. $\|\boldsymbol{X}_n(t+1) - \boldsymbol{X}_n(t)\|_2 \le \boldsymbol{v}_c, \|\boldsymbol{X}_n(t) - \boldsymbol{X}_m(t)\|_2 \ge d_{\min}, \forall n, m$, where P_{max} denotes the maximum power consumption at each time slot, and $d_{\min}$ denotes the collision avoidance distance between different UAVs.
- **(2) Energy-efficient UAV trajectory offline design [10], [49]:** In this scheme, the UAV trajectory and power control are optimized by solving the following energy-efficient maximization problem: $\max_{\boldsymbol{X}(t), 0 \le P_n(t) \le P_{max}, \sum_m u_{nm}(t)=1} \frac{\sum_{t,m,j} s^j_n(t) R_{nm}(t)}{\sum_{t,m,j} E^p_n(t)}$, s.t. $\|\boldsymbol{X}_n(t+1) - \boldsymbol{X}_n(t)\|_2 \le \boldsymbol{v}_c, \|\boldsymbol{X}_n(t) - \boldsymbol{X}_m(t)\|_2 \ge d_{\min}, \forall n, m$ where $E^p_n(t)$ denotes the per-stage UAV propulsion energy consumption.
- **(3) Queue weighted design:** In this scheme, the UAV moves sequentially and selects the user with the highest request queue length for transmission scheduling. The transmission power is adaptive to both CSI and QSI by solving the following optimization problem: $\min_{P_n} P_n(t) - \alpha \sum_{m,j} s^j_n(t) Q_{nm}(t) R_{nm}(t)$.

For baseline (1) and (2), we apply block SCA based iterative algorithm to solve the non-convex optimization problem. For proposed HPM-based learning scheme, we first choose

$$V_0(\boldsymbol{S}) = 1 - \left[1 - \frac{\rho_3}{2N_E} E^2 + \rho_3 E\right] \prod_{j,m} \left[1 - (Q^j_m)^2\right] \times \left[1 + \frac{1}{2} \frac{\varrho_1 + \varrho_2}{L_x - U_x} (\boldsymbol{X}^x)^2 - \frac{L_x \varrho_2 + U_x \varrho_1}{L_x - U_x} \boldsymbol{X}^x\right] \times \left[1 + \frac{1}{2} \frac{\varrho_1 + \varrho_2}{L_y - U_y} (\boldsymbol{X}^y)^2 - \frac{L_y \varrho_2 + U_y \varrho_1}{L_y - U_y} \boldsymbol{X}^y\right] \times \left[1 + \frac{1}{2} \frac{\varrho_1 + \varrho_2}{L_z - U_z} (\boldsymbol{X}^z)^2 - \frac{L_z \varrho_2 + U_z \varrho_1}{L_z - U_z} \boldsymbol{X}^z\right]$$

$$\mu_0(\boldsymbol{S}) = \prod_{j,m} \left(1 - e^{-Q^j_m}\right) \left(1 - e^{E - N_E}\right) \left(1 - e^{E}\right)$$

as the initial guess of $V(\boldsymbol{S})$ and $\mu(\boldsymbol{S})$ to satisfy the conditions in (51)-(58), where $\varrho_1, \varrho_2 > 0$ are the parameters specifying boundary conditions in (51)-(52)

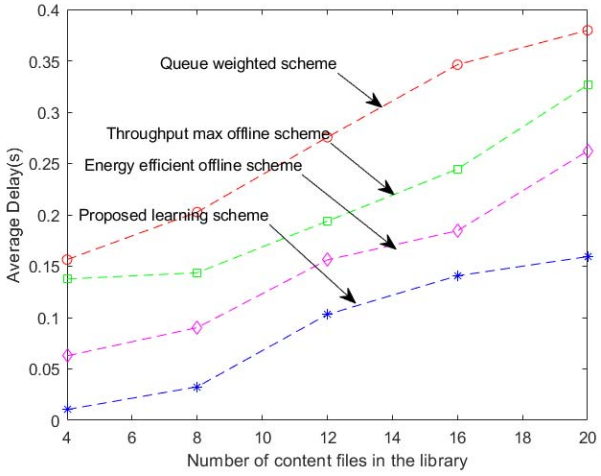


Fig. 8. Long-term average delay performance comparison vs number of content files in the content library with $M = 60$, $N = 10$ with single caching capacity in UAV.

B. Multi-UAV Control Performance

Fig.8 shows the average delay cost (seconds) versus number of content files in the library. As number of content files in the library increases, the average delay increases because each UAV can only serve single content at each time. In Fig. 9, we extended the simulation to consider multiple content files per UAV cache. In both cases, the proposed solution has significant gain over the baselines..

In Fig. 10, we compare the performance of proposed solution applied on heterogeneous and homogeneous UAVs respectively, where in heterogeneous scenario, UAVs are grouped into 3 types with different flight height region and power cost consumption. It can be seen that the proposed solution is quite robust for the case of heterogeneous UAVs.

Fig.11 illustrates the average delay cost per user versus the UAV numbers. It can be noted that as the density of the multi-UAV system increases, the proposed scheme will be more advantageous when compared to the offline baseline schemes. This is because the adopted mean-field approach benefits multi-UAV system with high density, while the baseline schemes cannot handle the UAVs trajectory and transmission scheduling well in a denser system.

Fig.12 shows the optimized UAV trajectory obtained by using the proposed online learning scheme with UAV number $N = 5$. To be specific, there are 5 user clusters (in the red circles), and each user cluster has 12 users. We assume that users in the different clusters request for the different content files cached in the specific UAVs. For instance, UAV1 only serve the users in Cluster 1, 2 and 5, while UAV2 serve the users in cluster 2,3 and 4. Each UAV under the proposed learning scheme can control radio resources and trajectory that are adaptive to its energy consumption state and its serving group QoS state, and won't violate its safety constraint due to the boundary conditions in (35)-(39). Note that the queue states carry a higher weight in the value function compared with other state components. Hence, UAV1 flies around from cluster 1 to cluster 5 rather that directly flying to cluster

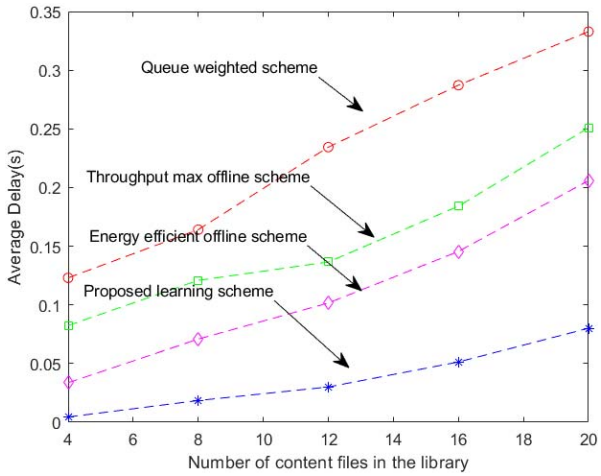


Fig. 9. Long-term average delay cost vs number of content files in the content library with $M = 60$, $N = 10$, each UAV can cache 10 files.

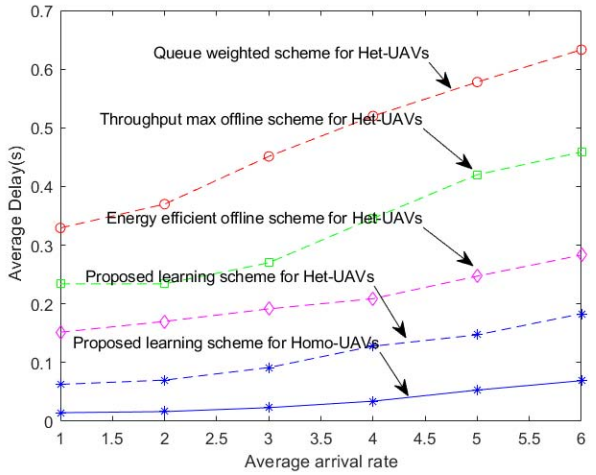


Fig. 10. Performance comparison of proposed solution and baselines applied on heterogeneous and homogeneous UAVs with $M = 60$, $N = 10$, $J = 10$.

5 because the request queue state of the users in cluster 1 is much larger than the other clusters.

C. Convergence and Complexity Analysis

Figure 13 illustrates the approximation errors versus the number of neurons in the DNN. As illustrated, we can achieve similar approximation quality with much smaller size of weight vectors, illustrating the superiority in the expressive power of the proposed neural network. Fig.14 illustrates convergence curve of the proposed learning algorithm. It can be found that the performance converges to the steady point quite fast. Fig.15 shows the performance gap of average delay vs average transmission power between brute-force optimal, proposed learning solution and baselines. It can be seen that the performance of the proposed scheme is very close to the brute-force solution performance, thus the quality of value function $V(\boldsymbol{S})$ and stationary distribution $\mu(\boldsymbol{S})$ approximation is good w.r.t. the optimal one under the proposed deep learning algorithm. Furthermore, Table III shows the computation time comparison of the baselines, brute-force optimal and proposed scheme. We can see that the computation time of the queue

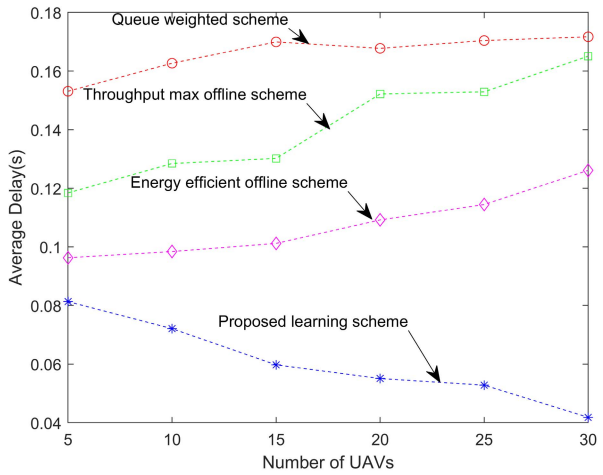


Fig. 11. Average delay vs UAV numbers with $M = 60$, $J = 10$.

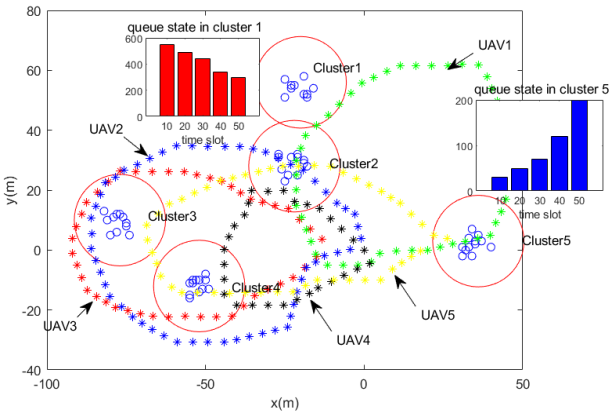


Fig. 12. UAV trajectory under the proposed scheme with number of UAVs $N = 5$, number of users $M = 60$.

TABLE III
COMPARISON OF AVERAGE COMPUTATION TIME TO COMPUTE THE CONTROL ACTION IN ONE SLOT

	$M = 30$ $N = 5$	$M = 60$ $N = 5$	$M = 60$ $N = 10$
Queue Weighted	15.2ms	16.2ms	18.4ms
Throughput Max	5.621s	5.778s	10.271s
Energy Efficient	5.74s	5.733s	10.134s
Proposed Scheme	11.232ms	11.26ms	11.4ms
B-F Algorithm	$> 10^2$ s	$> 10^2$ s	$> 10^2$ s

weighted baseline is smaller, but it has the worst performance than the other schemes. The brute-force numerical solution has a close performance to that of the proposed scheme, but with the largest computational complexity. While the offline schemes are of high time-consuming and worse than the proposed scheme.

VII. CONCLUSION

In this paper, we propose an online data-driven multi-UAV trajectory and transmission control scheme. We formulate the optimization problem as an ergodic stochastic differential game to optimize the users' quality-of-experience. By using mean-field game analysis, we derive a reduced-complexity optimality condition, and analyze the corresponding feasible

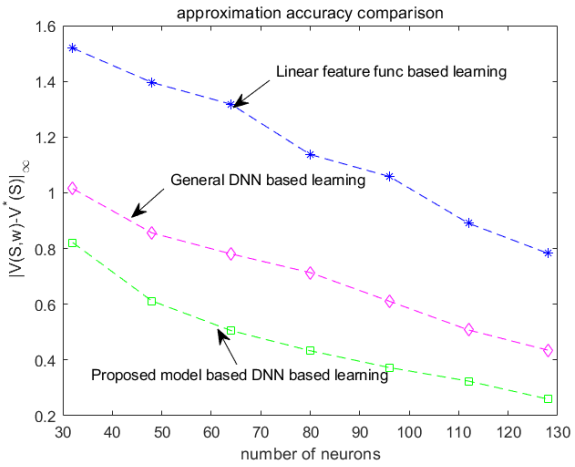


Fig. 13. Approximation accuracy comparison of generic DNN and proposed problem specific DNN illustrating the proposed DNN architecture is more expressive than generic DNN structure.

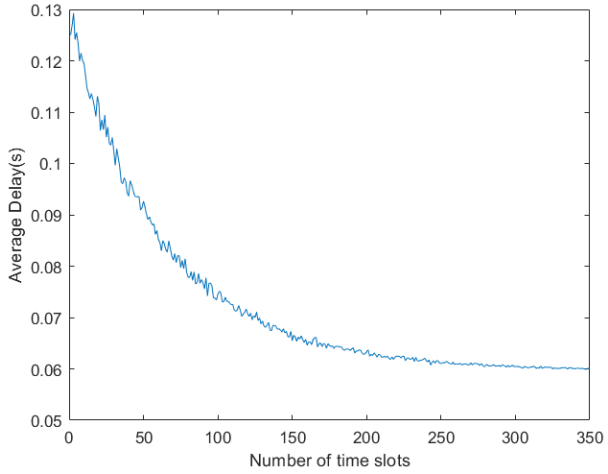


Fig. 14. Convergence of the proposed learning algorithm.

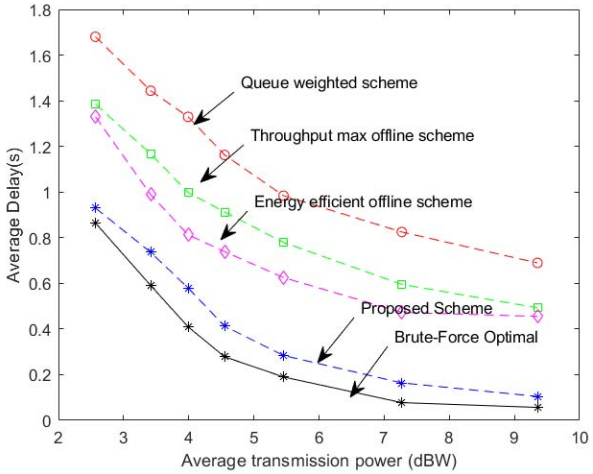


Fig. 15. Average delay performance comparison of proposed learning algorithm and brute-force optimal solution.

control solution under a complex coupled state-action constraint. As solving such a high dimensional PDE is still highly challenging numerically and analytically, we apply a novel data-driven online learning approach to resolve the problem. Numerical results show that the proposed scheme has much better performance than the existing baselines.

VIII. Appendix A

We have the differential equations as $\mathrm{d}V_n^N(\boldsymbol{S}_n) = \sum_l \frac{\mathrm{d}V_n^N(\boldsymbol{S}_n)}{\mathrm{d}\boldsymbol{S}_n^l}\mathrm{d}\boldsymbol{S}_n^l + \frac{1}{2}\sum_l\sum_k \frac{\mathrm{d}^2V_n^N(\boldsymbol{S}_n)}{\mathrm{d}\boldsymbol{S}_n^l\mathrm{d}\boldsymbol{S}_n^k}\mathrm{d}\boldsymbol{S}_n^l\mathrm{d}\boldsymbol{S}_n^k$, where $\boldsymbol{S}_n^l$ and $\boldsymbol{S}_n^k$ denote the state elements in $\boldsymbol{S}_n$. By substituting the differential dynamics into this equation with SDE properties, we have

$$
\begin{aligned}
&\mathrm{d}V_n^N(\boldsymbol{S}_n(t)) \\
&= \sum_{j,m}\Bigg[[A_{nm}^j(t) - s_n^j(t)R_{nm}(t)]\frac{\mathrm{d}V_n^N(\boldsymbol{S}_n)}{\mathrm{d}Q_{nm}^j}\mathrm{d}t \\
&\quad + \frac{\mathrm{d}V_n^N(\boldsymbol{S}_n)}{\mathrm{d}Q_{nm}^j}\mathrm{d}\boldsymbol{L}^{Q_{nm}^j}(t)\Bigg] + (-P_n(t) - E_n^p(t) + \\
&\quad + E_n^{ch}(t))\frac{\mathrm{d}V_n^N(\boldsymbol{S}_n)}{\mathrm{d}E_n}\mathrm{d}t - \frac{\mathrm{d}V_n^N(\boldsymbol{S}_n)}{\mathrm{d}E_n}\mathrm{d}\boldsymbol{U}_n^E(t) \\
&\quad + \frac{\mathrm{d}V_n^N(\boldsymbol{S}_n)}{\mathrm{d}E_n}\mathrm{d}\boldsymbol{L}_n^E(t) \\
&\quad + \boldsymbol{v}_n^T(t)\nabla_{\boldsymbol{X}_n}V_n^N(\boldsymbol{S}_n)\mathrm{d}t + \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\sigma}_X\boldsymbol{\sigma}_X^T\nabla_{\boldsymbol{X}_n}^2V_n^N(\boldsymbol{S}_n)\right)\mathrm{d}t \\
&\quad + \nabla_{\boldsymbol{X}_n}^TV_n^N(\boldsymbol{S}_n)\mathrm{d}\boldsymbol{L}_n^X(\mathrm{t}) - \nabla_{\boldsymbol{X}_n}^TV_n^N(\boldsymbol{S}_n)\mathrm{d}\boldsymbol{U}_n^X(t) \\
&\quad + \sum_m((a_H - a_HG_{nm}(t))\frac{\mathrm{d}V_n^N(\boldsymbol{S}_n)}{\mathrm{d}G_{nm}} \\
&\quad + a_HG_{nm}(t)\frac{\mathrm{d}^2V_n^N(\boldsymbol{S}_n)}{\mathrm{d}G_{nm}^2})\mathrm{d}t
\end{aligned}
\tag{66}
$$

Integrating this equation over the interval $(0,T)$ under the given policy Ω_n^N, dividing T and subtracting θ_n^N on both sides, we can obtain

$$
\begin{aligned}
&\frac{\mathbb{E}^{\Omega_n^N}\left[V_n^N(\boldsymbol{S}_n(T))\right] - V_n^N(\boldsymbol{S}_n(0))}{T} - \theta_n^N \\
&\quad + \mathbb{E}^{\Omega_n^N}[\frac{1}{T}\int_0^T[c_n(\boldsymbol{S}_n(t),\boldsymbol{S}_{-n}(t),\Omega_n^N)\mathrm{d}t]] \\
&= -\theta_n^N + \mathbb{E}^{\Omega_n^N}[\frac{1}{T}\int_0^T[\sum_i\mathcal{L}_i^{\Omega_i^N}(V(\boldsymbol{S}_n(t)))\mathrm{d}t]) \\
&\quad + \mathbb{E}^{\Omega_n^N}[\frac{1}{T}\int_0^T[c_n(\boldsymbol{S}_n(t),\boldsymbol{S}_{-n}(t),\Omega_n^N)\mathrm{d}t] \\
&\quad + \mathbb{E}^{\Omega_n^N}[\frac{1}{T}[\int_0^T\sum_{j,m}\frac{\mathrm{d}V_n^N(\boldsymbol{S}_n)}{\mathrm{d}Q_{nm}^j}\Big|_{Q_{nm}^j=0}\mathrm{d}\boldsymbol{L}^{Q_{nm}^j}(t) \\
&\quad - \int_0^T\frac{\mathrm{d}V_n^N(\boldsymbol{S}_n)}{\mathrm{d}E_n}\Big|_{E_n=N_E}\mathrm{d}\boldsymbol{U}_n^E(t) \\
&\quad + \int_0^T\frac{\mathrm{d}V_n^N(\boldsymbol{S}_n)}{\mathrm{d}E_n}\Big|_{E_n=0}\mathrm{d}\boldsymbol{L}_n^E(t) \\
&\quad + \int_0^T\nabla_{\boldsymbol{X}_n}^TV_n^N(\boldsymbol{S}_n)\Big|_{\boldsymbol{X}_n=[L_x,L_y,L_z]^T}\mathrm{d}\boldsymbol{L}_n^X(\mathrm{t}) \\
&\quad - \int_0^T\nabla_{\boldsymbol{X}_n}^TV_n^N(\boldsymbol{S}_n)\Big|_{\boldsymbol{X}_n=[U_x,U_y,U_z]^T}\mathrm{d}\boldsymbol{U}_n^X(\mathrm{t})],
\end{aligned}
$$

where $\mathcal{L}_i^{\Omega_i^N}$ is defined as the i-th UAV dynamic generator with (27). If (19) and the boundary conditions in (20)-(23) are satisfied, for any admissible control policy, we have (after rearranging) $\frac{\mathbb{E}^{\Omega_n^N}\left[V_n^N(\boldsymbol{S}_n(T))\right]-V_n^N(\boldsymbol{S}_n(0))}{T}+\mathbb{E}^{\Omega_n^N}[\frac{1}{T}\int_0^T[c_n(\boldsymbol{S}_n(t),\boldsymbol{S}_{-n}(t),\Omega_n^N)\mathrm{d}t]+\mathbb{E}^{\Omega_n^N}[\frac{1}{T}\varrho_1(L_n^x(T)+L_n^y(T)+L_n^z(T))+\frac{1}{T}\varrho_2(\boldsymbol{U}_n^x(T)+\boldsymbol{U}_n^y(T)+\boldsymbol{U}_n^z(T))+\frac{1}{T}\varrho_3(\boldsymbol{L}_n^E(T)]\geq\theta_n^N$. After taking $\limsup$ as $T\to\infty$ on both sides, we have $\limsup_{T\to\infty}\frac{\mathbb{E}^{\Omega_n^N}\left[V_n^N(\boldsymbol{S}_n(T))\right]-V_n^N(\boldsymbol{S}_n(0))}{T}=0$. Therefore, $\limsup_{T\to\infty}\mathbb{E}^{\Omega^N}\Big[\frac{1}{T}[\int_0^T\left[c_n(\boldsymbol{S}_n(t),\boldsymbol{S}_{-n}(t),\Omega_n^N)\right]\mathrm{d}t+\varrho_1(L_n^x(T)+L_n^y(T)+L_n^z(T))+\varrho_2(\boldsymbol{U}_n^x(T)+\boldsymbol{U}_n^y(T)+\boldsymbol{U}_n^z(T))+\rho_3\boldsymbol{L}_n^E(T)]\Big]\geq\theta_n^N$, where the equality can be realized if there exists an admissible policy Ω_n^{N*} such that it attains the minimum of the L.H.S. of (19). Hence, policy $\Omega_n^{N*}(\boldsymbol{S})=\{P_n^{N*},\{u_{nm}^{N*}\}_{m\in\mathcal{M}},\boldsymbol{v}_n^{N*}\}$ is the optimal control policy for Problem 2.

IX. Appendix B

Based on HJB equations (19) and the admissible control policy, all UAV individuals have the symmetric cost structure and dynamic construction, e.g., state boundary, diffusion coefficients, and it can be noted that the obtained optimal feasible UAV controls are homogeneous. Hence we have the result that the states of UAVs are indistinguishable (exchangeable) [29], [30]: when the number of UAV players tends to infinity, all UAVs exhibit identical behavior and the game can be approximated by mean-field game.

X. Appendix C

(1) We first rewrite the per stage cost as $c(\boldsymbol{S}_n,\boldsymbol{S}_{-n},\Omega_n^N)=c^1(\boldsymbol{S}_n,\Omega_n^N)+c^2(\boldsymbol{S}_n,\frac{1}{N}\sum_{k\neq n}^N\delta_{\boldsymbol{S}_k})$, where $c^1(\boldsymbol{S}_n,\Omega_n^N)=\alpha\sum_{m=1}^M\sum_{j=1}^J Q_{nm}^j+P_n^N$, and $c^2(\boldsymbol{S}_n,\frac{1}{N}\sum_{k\neq n}^N\delta_{\boldsymbol{S}_k})=\frac{1}{N}\sum_{k\neq n}^N\int_{\boldsymbol{S}'}\frac{\beta}{\gamma+\|\boldsymbol{X}_n'-\boldsymbol{X}_k'\|_2^2}\delta_{\boldsymbol{S}_k}\mathrm{d}S'$ with $\delta_{\boldsymbol{S}_k}=\delta(\boldsymbol{S}'=\boldsymbol{S}_k)$. Then based on the property of N independent UAV dynamics, we have the stationary distributions of N UAVs should be decoupled with $\mu_n^N(\mathrm{d}\boldsymbol{S}_n)$, $\forall n\in\mathcal{N}$.

Define that $\hat{c}_n^{\boldsymbol{\mu}}(\boldsymbol{S}_n)=\int_{\mathcal{S}^N}c^2(\boldsymbol{S}_n,\frac{1}{N}\sum_{k\neq n}^N\delta_{\boldsymbol{S}_k})\prod_{k\neq n}^N\mu_k^N(\mathrm{d}\boldsymbol{S}_k)$, we have

$$\min_{\Omega_n^N=\{P_n^N,\{u_{nm}^N\}_{m\in\mathcal{M}},\boldsymbol{v}_n^N\}}\left(\mathcal{L}_n^{\Omega_n^N}V_n^N+c^1(\boldsymbol{S},\Omega_n^N)\right)+\hat{c}_n^{\boldsymbol{\mu}}(\boldsymbol{S})=\theta_n^N,\forall n\in\mathcal{N}$$

as $\int\mathcal{L}_n^{\Omega_n^N}f(\boldsymbol{S})\mu_n^N(\mathrm{d}\boldsymbol{S})=0,\ for\ all\ f\in\mathcal{C}^2(\mathcal{S}),\ \forall n\in\mathcal{N}$. Then consider the following equation with the solution $(\check{V}^N,\check{\theta}^N,\mu^N)$:

$$\min_{\Omega=\{P,\{u_m\}_{m\in\mathcal{M}},\boldsymbol{v}\}}\left(\mathcal{L}^{\Omega}\check{V}^N+c^1(\boldsymbol{S},\Omega)\right)+\check{c}^{\boldsymbol{\mu}}(\boldsymbol{S})=\check{\theta}^N,\forall n\in\mathcal{N},$$

$$\mathcal{L}^{*\Omega}\mu^N(\boldsymbol{S})=0,\int_{\mathcal{S}}\mu^N(\boldsymbol{S})\mathrm{d}\boldsymbol{S}=1,$$

where $\check{c}^{\boldsymbol{\mu}}(\boldsymbol{S})=\int_{\mathcal{S}^N}c^2(\boldsymbol{S},\frac{1}{N}\sum_{k=1}^N\delta_{\boldsymbol{S}_k})\prod_{k=1}^N\mu_k^N(\mathrm{d}\boldsymbol{S}_k)$. It can be deduced that

$$\left\|\check{c}_n^{\boldsymbol{\mu}}-\hat{c}^{\boldsymbol{\mu}}\right\|_\infty=\sup\Big|\int_{\mathcal{S}^N}\Big(c^2(\boldsymbol{S}_n,\frac{1}{N}\sum_{k\neq n}^N\delta_{\boldsymbol{S}_k})-c^2(\boldsymbol{S}_n,\frac{1}{N}\sum_{k=1}^N\delta_{\boldsymbol{S}_k})\Big)\prod_{k=1}^N\mu_k^N(\mathrm{d}\boldsymbol{S}_k)\Big|\leq\frac{\epsilon_1}{N}$$

with constant ϵ_1 independent of N as $\left|c^2(\boldsymbol{S}_n,\frac{1}{N}\sum_{k\neq n}^N\delta_{\boldsymbol{S}_k})-c^2(\boldsymbol{S}_n,\frac{1}{N}\sum_{k=1}^N\delta_{\boldsymbol{S}_k})\right|=\frac{\beta/\gamma}{N}$. Then we have $\left\|\check{c}_n^{\boldsymbol{\mu}}-\hat{c}^{\boldsymbol{\mu}}\right\|_\infty\to0$ as $N\to\infty$. Based on this, by considering the continuous dependence of HJB equation and FPK equation, we can conclude that as $N\to\infty$

$$\sup_n\left(\left|\theta_n^N-\check{\theta}^N\right|+\left\|V_n^N-\check{V}^N\right\|_{\mathcal{C}^2(\mathcal{S})}+\left\|\mu_n^N-\mu^N\right\|_\infty\right)\to0.$$

Then the convergence in Theorem (2) (1) can be established.

(2) We first have the following result from [50]

$$\lim_{N\to\infty}\int_{\mathcal{S}^N}c^2(\boldsymbol{S}_n,\frac{1}{N}\sum_{k=1}^N\delta_{\boldsymbol{S}_k})\prod_{k=1}^N\mu(\mathrm{d}\boldsymbol{S}_k)=c^2(\boldsymbol{S},\mu).\quad(67)$$

Since $\lim_{N\to\infty}\left|\mu_n^N-\mu\right|_\infty=0$ according to the Lemma 3.2.5 in [51], by using the result in (1) and triangle inequality, we can conclude that

$$\lim_{N\to\infty}\Big\|\int_{\mathcal{S}^N}c^2(\boldsymbol{S}_n,\frac{1}{N}\sum_{k=1}^N\delta_{\boldsymbol{S}_k})\prod_{k=1}^N\mu(\mathrm{d}\boldsymbol{S}_k)\quad(68)$$

$$-\int_{\mathcal{S}^N}c^2(\boldsymbol{S}_n,\frac{1}{N}\sum_{k=1}^N\delta_{\boldsymbol{S}_k})\prod_{k=1}^N\mu^N(\mathrm{d}\boldsymbol{S}_k)\Big\|_\infty=0.\quad(69)$$

Then we have

$$\lim_{N\to\infty}\left\|\int_{\mathcal{S}^N}c^2(\boldsymbol{S}_n,\frac{1}{N}\sum_{k=1}^N\delta_{\boldsymbol{S}_k})\prod_{k=1}^N\mu^N(\mathrm{d}\boldsymbol{S}_k)-c^2(\boldsymbol{S},\mu)\right\|_\infty=0.$$

Finally, it is sufficed to conclude that the limit point of (26)-(28) is the solution of

$$\min_{\Omega=\{P,\{u_m\}_{m\in\mathcal{M}},\boldsymbol{v}\}}\left(\mathcal{L}^{\Omega}V(\boldsymbol{S})+c^1(\boldsymbol{S},\Omega)+c^2(\boldsymbol{S},\mu)\right)=\theta$$

$$\mathcal{L}^{*\Omega}\mu(\boldsymbol{S})=0,\int_{\mathcal{S}}\mu(\boldsymbol{S})\mathrm{d}\boldsymbol{S}=1.$$

XI. Appendix D

The coupled state-action constraint in (15) can be transformed to a state only constraint w.r.t. the UAV energy state and position state: $E\geq\frac{(c_1+c_2v_c^2)}{v_c}(D^{ch}(\boldsymbol{X})-D_0)$. Based on this, ever since the system state reaches the safety boundary $E=\frac{(c_1+c_2v_c^2)}{v_c}(D^{ch}(\boldsymbol{X})-D_0)$, the control action should guarantee the energy and position state pointing inward to the safety region [52], [53], thus we have the equation (35). The value function boundary conditions (36)-(39) come from reflecting boundary property in (20)-(23).

XII. Appendix E

According to the HJB-FPK equation in (33), by solving $\min_{\Omega(\boldsymbol{S})=\{P,\boldsymbol{u},\boldsymbol{v}\}}\left(P+\mathcal{L}^{\Omega}V(\boldsymbol{S})\right)$, we can obtain the optimal solution in the following form:

$$P^*=\sum_{j,m=1}^{J,M}s^j u_m^*\left(\frac{\frac{W}{ln2}\frac{\partial V(\boldsymbol{S})}{\partial Q_m^j}}{1-\frac{\partial V(\boldsymbol{S})}{\partial E}}-\frac{\sigma^2}{H_m}\right)^+,$$

$$u_m^*=\begin{cases}1, & m=m^*,\\ 0, & otherwise,\end{cases} \tag{70}$$

where m^* is defined as (43).

$$\boldsymbol{v}^*=\frac{v_c\nabla_{\boldsymbol{X}}V(\boldsymbol{S})/(2c_2\frac{\partial V(\boldsymbol{S})}{\partial E})}{\left\|\nabla_{\boldsymbol{X}}V(\boldsymbol{S})/(2c_2\frac{\partial V(\boldsymbol{S})}{\partial E})\right\|_2}, \tag{71}$$

Based on the boundary condition in (35), we impose that $\boldsymbol{v}^*=\frac{v_c(\boldsymbol{X}-\boldsymbol{X}^{ch})}{\left\|\boldsymbol{X}-\boldsymbol{X}^{ch}\right\|_2}, P^*=0$ at the state space boundary $E=\frac{(c_1+c_2v_c^2)}{v_c}(D^{ch}-D_0)$. Specifically, the transmission is suspended, and energy are all utilized for the UAV flight such that it can be guaranteed to get to the recharge station in case of running out of on-board energy.

XIII. Appendix F

We first show that the series constructed by DNN in (60)-(61) converges to the PDE solutions $V(\boldsymbol{S})$, $\mu(\boldsymbol{S})$. The proof is based on the Banach's fixed point theorem. According to the recursion equations (60)-(61), we have $\left\|\hat{V}^{I+1}(\boldsymbol{S})-\hat{V}^{I}(\boldsymbol{S})\right\|=V_{I+1}(\boldsymbol{S})$ and $\left\|\hat{\mu}^{I+1}(\boldsymbol{S})-\hat{\mu}^{I}(\boldsymbol{S})\right\|=\mu_{I+1}(\boldsymbol{S})$, and we need to show that the series $\hat{V}^{I+1}(\boldsymbol{S})$ and $\hat{\mu}^{I}(\boldsymbol{S})$ are the Cauchy sequences in the Banach space. When the auxiliary convergence parameters $\{\boldsymbol{w}_H,\boldsymbol{w}_F\}$ are properly chosen at each recursion update, it can be shown that the series generated by (60)-(61) satisfy

$$\begin{aligned}\left\|\hat{V}^{I+1}(\boldsymbol{S})-\hat{V}^{I}(\boldsymbol{S})\right\|&=\left\|V_{I+1}(\boldsymbol{S})\right\|\le\gamma_H\left\|V_I(\boldsymbol{S})\right\| \qquad (72)\\ &\le\gamma_H^2\left\|V_{I+1}(\boldsymbol{S})\right\|\le\dots\\ &\le\gamma_H^{I+1}\left\|V_0(\boldsymbol{S})\right\|,\end{aligned}$$

$$\begin{aligned}\left\|\hat{\mu}^{I+1}(\boldsymbol{S})-\hat{\mu}^{I}(\boldsymbol{S})\right\|&=\left\|\mu_{I+1}(\boldsymbol{S})\right\|\le\gamma_F\left\|\mu_I(\boldsymbol{S})\right\| \qquad (73)\\ &\le\gamma_F^2\left\|\mu_{I+1}(\boldsymbol{S})\right\|\le\dots\\ &\le\gamma_F^{I+1}\left\|\mu_0(\boldsymbol{S})\right\|,\end{aligned}$$

where $0\le\gamma_H<1$, $0\le\gamma_F<1$ are the constant independent of the state dimensions. Then we have that $\lim_{I,J\to\infty}\left\|\hat{V}^{I}(\boldsymbol{S})-\hat{V}^{J}(\boldsymbol{S})\right\|=0$, $\lim_{I,J\to\infty}\left\|\hat{\mu}^{I}(\boldsymbol{S})-\hat{\mu}^{J}(\boldsymbol{S})\right\|=0$ such that $\hat{V}^{I}(\boldsymbol{S})$ and $\hat{\mu}^{I}(\boldsymbol{S})$ are the Cauchy sequences. As discussed of homotopy analysis method in [34], if the homotopy series solutions $\hat{V}^{I}(\boldsymbol{S})$ and $\hat{\mu}^{I}(\boldsymbol{S})$ are convergent, and the initial guess $V_0(\boldsymbol{S})$, $\mu_0(\boldsymbol{S})$ satisfy the boundary conditions (51)-(58), then the solutions $\lim_{I\to\infty}\hat{V}^{I}(\boldsymbol{S})$ and $\lim_{I\to\infty}\hat{\mu}^{I}(\boldsymbol{S})$ must be the exact solutions of nonlinear PDE equations (32)-(33), and satisfy the boundary conditions.

Next we will show the approximation error bound between the proposed solutions and the true solutions in (32)-(33). According to the triangle inequality and the equations (67)-(68), we can derive that $\forall I,J\in\mathbb{N}$, $J\ge I>0$,

$$\left\|\hat{V}^{J}(\boldsymbol{S})-\hat{V}^{I}(\boldsymbol{S})\right\|\le\frac{1-\gamma_H^{J-I}}{1-\gamma_H}\gamma_H^{I+1}\left\|V_0(\boldsymbol{S})\right\|$$

$$\left\|\hat{\mu}^{J}(\boldsymbol{S})-\hat{\mu}^{I}(\boldsymbol{S})\right\|\le\frac{1-\gamma_F^{J-I}}{1-\gamma_F}\gamma_F^{I+1}\left\|\mu_0(\boldsymbol{S})\right\|.$$

Then based on that the series generated by the homotopy perturbation converge to the exact PDE solutions as discussed above: $\lim_{I\to\infty}\hat{V}^{I}(\boldsymbol{S})=V(\boldsymbol{S})$ and $\lim_{I\to\infty}\hat{\mu}^{I}(\boldsymbol{S})=\mu(\boldsymbol{S})$, we can have $\left\|V(\boldsymbol{S})-\hat{V}^{I}(\boldsymbol{S})\right\|\le\frac{1}{1-\gamma_H}\gamma_H^{I+1}\left\|V_0(\boldsymbol{S})\right\|$ and $\left\|\mu(\boldsymbol{S})-\hat{\mu}^{I}(\boldsymbol{S})\right\|\le\frac{1}{1-\gamma_F}\gamma_F^{I+1}\left\|\mu_0(\boldsymbol{S})\right\|$. This completes the theorem (3) proof.

References

[1] H. Wang, G. Ding, F. Gao, J. Chen, J. Wang, and L. Wang, "Power control in UAV-supported ultra dense networks: Communications, caching, and energy transfer," *IEEE Commun. Mag.*, vol. 56, no. 6, pp. 28–34, Jun. 2018.

[2] M. Mozaffari, W. Saad, M. Bennis, Y.-H. Nam, and M. Debbah, "A tutorial on UAVs for wireless networks: Applications, challenges, and open problems," 2018, *arXiv:1803.00680*. [Online]. Available: http://arxiv.org/abs/1803.00680

[3] Y. Zeng, Q. Wu, and R. Zhang, "Accessing from the sky: A tutorial on UAV communications for 5G and beyond," 2019, *arXiv:1903.05289*. [Online]. Available: http://arxiv.org/abs/1903.05289

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

[5] S. Zhang, H. Zhang, Q. He, K. Bian, and L. Song, "Joint trajectory and power optimization for UAV relay networks," *IEEE Commun. Lett.*, vol. 22, no. 1, pp. 161–164, Jan. 2018.

[6] F. Ono, H. Ochiai, and R. Miura, "A wireless relay network based on unmanned aircraft system with rate optimization," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7699–7708, Nov. 2016.

[7] U. Challita and W. Saad, "Network formation in the sky: Unmanned aerial vehicles for multi-hop wireless backhauling," 2017, *arXiv:1707.09132*. [Online]. Available: http://arxiv.org/abs/1707.09132

[8] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Mobile unmanned aerial vehicles (UAVs) for energy-efficient Internet of Things communications," *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, pp. 7574–7589, Nov. 2017.

[9] M. Jiang, Y. Li, Q. Zhang, and J. Qin, "Joint position and time allocation optimization of UAV enabled time allocation optimization networks," *IEEE Trans. Commun.*, vol. 67, no. 5, pp. 3806–3816, May 2019.

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

[11] M. M. Azari, F. Rosas, K.-C. Chen, and S. Pollin, "Ultra reliable UAV communication using altitude and cooperation diversity," *IEEE Trans. Commun.*, vol. 66, no. 1, pp. 330–344, Jan. 2018.

[12] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Efficient deployment of multiple unmanned aerial vehicles for optimal wireless coverage," *IEEE Commun. Lett.*, vol. 20, no. 8, pp. 1647–1650, Aug. 2016.

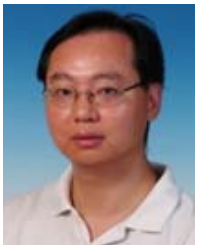
[13] J. Lyu, Y. Zeng, R. Zhang, and T. J. Lim, "Placement optimization of UAV-mounted mobile base stations," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 604–607, Mar. 2017.

[14] Y. Chen, W. Feng, and G. Zheng, "Optimum placement of UAV as relays," *IEEE Commun. Lett.*, vol. 22, no. 2, pp. 248–251, Feb. 2018.

[15] S. Jeong, O. Simeone, and J. Kang, "Mobile edge computing via a UAV-mounted cloudlet: Optimization of bit allocation and path planning," *IEEE Trans. Veh. Technol.*, vol. 67, no. 3, pp. 2049–2063, Mar. 2018.

[16] 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.
[17] E. Koyuncu, M. Shabanighazikelayeh, and H. Seferoglu, "Deployment and trajectory optimization of UAVs: A quantization theory approach," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 8531–8546, Dec. 2018.
[18] S. Roth, A. Kariminezhad, and A. Sezgin, "Base-stations up in the air: Multi-UAV trajectory control for min-rate maximization in uplink C-RAN," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–6.
[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] W. Han, A. Liu, and V. K. N. Lau, "PHY-caching in 5G wireless networks: Design and analysis," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 30–36, Aug. 2016.
[21] M. Chen, M. Mozaffari, W. Saad, C. Yin, M. Debbah, and C. S. Hong, "Caching in the sky: Proactive deployment of cache-enabled unmanned aerial vehicles for optimized quality-of-experience," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 5, pp. 1046–1061, May 2017.
[22] A. Liu, V. Lau, W. Ding, and E. Yeh, "Mixed-timescale online PHY caching for dual-mode MIMO cooperative networks," 2018, *arXiv:1810.07503*. [Online]. Available: http://arxiv.org/abs/1810.07503
[23] M. Chen, W. Saad, and C. Yin, "Echo-liquid state deep learning for 360° content transmission and caching in wireless VR networks with cellular-connected UAVs," 2018, *arXiv:1804.03284*. [Online]. Available: https://arxiv.org/abs/1804.03284
[24] N. Zhao *et al.*, "Caching unmanned aerial vehicle-enabled small-cell networks: Employing energy-efficient methods that store and retrieve popular content," *IEEE Veh. Technol. Mag.*, vol. 14, no. 1, pp. 71–79, Mar. 2019.
[25] X. Xu, Y. Zeng, Y. L. Guan, and R. Zhang, "Overcoming endurance issue: UAV-enabled communications with proactive caching," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 6, pp. 1231–1244, Jun. 2018.
[26] S. Chai and V. K. N. Lau, "Online trajectory and radio resource optimization of cache-enabled UAV wireless networks with content and energy recharging," *IEEE Trans. Signal Process.*, vol. 68, pp. 1286–1299, 2020.
[27] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, vol. 2, no. 3. Belmont, MA, USA: Athena scientific, 2011.
[28] O. Guéant, J.-M. Lasry, and P.-L. Lions, "Mean field games and applications," in *Paris-Princeton Lectures on Mathematical Finance 2010*. Berlin, Germany: Springer, 2011, pp. 205–266.
[29] H. Tembine and M. Huang, "Mean field difference games: McKean-Vlasov dynamics," in *Proc. IEEE Conf. Decis. Control Eur. Control Conf.*, Dec. 2011, pp. 1006–1011.
[30] K. Hamidouche, W. Saad, M. Debbah, and H. V. Poor, "Mean-field games for distributed caching in ultra-dense small cell networks," in *Proc. Amer. Control Conf. (ACC)*, Jul. 2016, pp. 4699–4704.
[31] H. Wu, F. Lyu, C. Zhou, J. Chen, L. Wang, and X. Shen, "Optimal UAV caching and trajectory in aerial-assisted vehicular networks: A learning-based approach," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 12, pp. 2783–2797, Dec. 2020.
[32] W. Shi, J. Li, H. Wu, C. Zhou, N. Cheng, and X. Shen, "Drone-cell trajectory planning and resource allocation for highly mobile networks: A hierarchical DRL approach," *IEEE Internet Things J.*, vol. 8, no. 12, pp. 9800–9813, Jun. 2021.
[33] J.-H. He, "Homotopy perturbation method for solving boundary value problems," *Phys. Lett. A*, vol. 350, nos. 1–2, pp. 87–88, Jan. 2006.
[34] S. Liao, *Beyond Perturbation: Introduction to the Homotopy Analysis Method*. Boca Raton, FL, USA: CRC Press, 2003.
[35] P. Grohs, F. Hornung, A. Jentzen, and P. von Wurstemberger, "A proof that artificial neural networks overcome the curse of dimensionality in the numerical approximation of black-scholes partial differential equations," 2018, *arXiv:1809.02362*. [Online]. Available: http://arxiv.org/abs/1809.02362
[36] Y. Khoo, J. Lu, and L. Ying, "Solving parametric PDE problems with artificial neural networks," 2017, *arXiv:1707.03351*. [Online]. Available: http://arxiv.org/abs/1707.03351
[37] I. E. Lagaris, A. Likas, and D. I. Fotiadis, "Artificial neural networks for solving ordinary and partial differential equations," *IEEE Trans. Neural Netw.*, vol. 9, no. 5, pp. 987–1000, Sep. 1998.
[38] T. Feng, T. R. Field, and S. Haykin, "Stochastic differential equation theory applied to wireless channels," *IEEE Trans. Commun.*, vol. 55, no. 8, pp. 1478–1483, Aug. 2007.
[39] Y. Zeng, J. Xu, and R. Zhang, "Energy minimization for wireless communication with rotary-wing UAV," 2018, *arXiv:1804.02238*. [Online]. Available: http://arxiv.org/abs/1804.02238
[40] C. Di Franco and G. Buttazzo, "Energy-aware coverage path planning of UAVs," in *Proc. IEEE Int. Conf. Auto. Robot Syst. Competitions*, Apr. 2015, pp. 111–117.
[41] A. Filippone, *Flight Performance of Fixed and Rotary Wing Aircraft*. Amsterdam, The Netherlands: Elsevier, 2006.
[42] A. R. S. Bramwell, D. Balmford, and G. Done, *Bramwell's Helicopter Dynamics*. Amsterdam, The Netherlands: Elsevier, 2001.
[43] M. Simic, C. Bil, and V. Vojisavljevic, "Investigation in wireless power transmission for UAV charging," *Procedia Comput. Sci.*, vol. 60, pp. 1846–1855, 2015.
[44] B. Zhang, C. H. Liu, J. Tang, Z. Xu, J. Ma, and W. Wang, "Learning-based energy-efficient data collection by unmanned vehicles in smart cities," *IEEE Trans. Ind. Informat.*, vol. 14, no. 4, pp. 1666–1676, Apr. 2018.
[45] T. J. Nugent, Jr. and J. T. Kare, "Laser power beaming for defense and security applications," *Proc. SPIE*, vol. 8045, May 2011, Art. no. 804514.
[46] F. Cheng *et al.*, "UAV trajectory optimization for data offloading at the edge of multiple cells," *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 6732–6736, Jul. 2018.
[47] Y. Zeng, R. Zhang, and T. J. Lim, "Throughput maximization for UAV-enabled mobile relaying systems," *IEEE Trans. Commun.*, vol. 64, no. 12, pp. 4983–4996, Dec. 2016.
[48] A. Liu and V. K. N. Lau, "Optimization of multi-UAV-aided wireless networking over a ray-tracing channel model," *IEEE Trans. Wireless Commun.*, vol. 18, no. 9, pp. 4518–4530, Sep. 2019.
[49] J. Zhang, Y. Zeng, and R. Zhang, "Spectrum and energy efficiency maximization in UAV-enabled mobile relaying," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.
[50] E. Hewitt and L. J. Savage, "Symmetric measures on Cartesian products," *Trans. Amer. Math. Soc.*, vol. 80, no. 2, pp. 470–501, 1955.
[51] A. Arapostathis, V. S. Borkar, and M. K. Ghosh, *Ergodic Control of Diffusion Processes*, no. 143. Cambridge, U.K.: Cambridge Univ. Press, 2012.
[52] H. M. Soner, "Optimal control with state-space constraint I," *SIAM J. Control Optim.*, vol. 24, no. 3, pp. 552–561, May 1986.
[53] I. Capuzzo-Dolcetta and P.-L. Lions, "Hamilton-Jacobi equations with state constraints," *Trans. Amer. Math. Soc.*, vol. 318, no. 2, pp. 643–683, Feb. 1990.

Shuqi Chai (Associate Member, IEEE) received the B.Eng. degree in electronic and information engineering from Xi'an Jiaotong University, Xi'an, China, in 2014. She is currently pursuing the Ph.D. degree with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology. Her research interests include dynamic programming in wireless communication systems, reinforcement learning, and UAV communication.



Vincent K. N. Lau (Fellow, IEEE) received the B.Eng. degree (Hons.) from The University of Hong Kong, Hong Kong, in 1992, and the Ph.D. degree from Cambridge University, Cambridge, U.K., in 1997. From 1997 to 2004, he was with Bell Labs. In 2004, he was with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology (HKUST), Hong Kong. He is currently a Chair Professor and the Founding Director of the Huawei-HKUST Joint Innovation Laboratory, HKUST. His research interests include robust and delay-optimal cross layer optimization for MIMO/OFDM wireless systems, interference mitigation techniques for wireless networks, massive MIMO, M2M, and network control systems.