Predictive Deployment of UAV Base Stations in Wireless Networks: Machine Learning Meets Contract Theory

Qianqian Zhang[1], Walid Saad[1], Mehdi Bennis[2], Xing Lu[3], Mérouane Debbah[4,5], and Wangda Zuo[3]

[1]Bradley Department of Electrical and Computer Engineering, Virginia Tech, VA, USA, Emails: {qqz93,walids}@vt.edu
[2]Centre for Wireless Communications, University of Oulu, Finland, Email: mehdi.bennis@oulu.fi
[3]Department of Civil, Environmental and Architectural Engineering, University of Colorado Boulder, CO, USA, Email: {xing.lu-1,wangda.zuo}@colorado.edu
[4]Mathematical and Algorithmic Sciences Lab, Huawei France R&D, Paris, France, Email: merouane.debbah@huawei.com
[5]Large Systems and Networks Group (LANEAS), CentraleSupélec, Université Paris-Saclay, Gif-sur-Yvette, France

***Abstract*—In this paper, a novel framework is proposed to enable a predictive deployment of unmanned aerial vehicles (UAVs) as temporary base stations (BSs) to complement ground cellular systems in face of downlink traffic overload. First, a novel learning approach, based on the weighted expectation maximization (WEM) algorithm, is proposed to estimate the user distribution and the downlink traffic demand. Next, to guarantee a truthful information exchange between the BS and UAVs, using the framework of contract theory, an offload contract is developed, and the sufficient and necessary conditions for having a feasible contract are analytically derived. Subsequently, an optimization problem is formulated to deploy an optimal UAV onto the hotspot area in a way that the utility of the overloaded BS is maximized. Simulation results show that the proposed WEM approach yields a prediction error of around 10%. Compared with the expectation maximization and k-mean approaches, the WEM method shows a significant advantage on the prediction accuracy, as the traffic load in the cellular system becomes spatially uneven. Furthermore, compared with two event-driven deployment schemes based on the closest-distance and maximal-energy metrics, the proposed predictive approach enables UAV operators to provide efficient communication service for hotspot users in terms of the downlink capacity, energy consumption and service delay. Simulation results also show that the proposed method significantly improves the revenues of both the BS and UAV networks, compared with two baseline schemes.**

Index Terms – cellular networks; UAV deployment; traffic prediction; contract theory.

I. INTRODUCTION

The use of unmanned aerial vehicles (UAVs) as flying base stations (BSs) has attracted growing interest in the past few years [1]–[8]. UAVs can be deployed to complement the existing cellular systems, by providing reliable wireless services for ground users, to potentially increase the network capacity, eliminate coverage holes, and cope with the steep surge of communication needs during hotspot events [1]. Compared with the terrestrial BSs that are deployed at a fixed location for a long term, UAVs are more suitable for temporary on-demand service [3]. For instance, UAVs can provide communication service for major events (e.g. sport or musical events) during which the terrestrial network capacity is often strained [4]. Furthermore, UAVs can adjust their positions and establish line-of-sight (LOS) communication links towards ground users, thus improving network performance [5]. Due to their broad range of application domains and low cost, UAVs as flying BSs provide a promising solution to ground users for temporary network connectivity [6].

However, the UAV deployment for on-demand cellular service faces several key challenges. For instance, UAVs are strictly constrained by their on-board energy, which should be efficiently used for communication. However, the on-demand deployment requires UAVs to continuously change their positions to meet instant communication requests. Therefore, most of on-board energy can be consumed by mobility, thus limiting their communication capabilities [1]. Moreover, to effectively alleviate network congestion during a hotspot event, the deployed UAV must have enough on-board power to satisfy the downlink communication demand. In order to allocate a qualified UAV with sufficient energy, the network operator should estimate the required transmit power, based on the real-time traffic load. These challenges, in turn, motivate the need for a comprehensive prediction of cellular traffic, and a predictive approach for UAV deployment [9]. To this end, machine learning (ML) techniques can be applied to estimate the cellular traffic demand within the target system. Given the predicted traffic load, each BS can detect hotspot areas and request suitable UAVs to alleviate network congestion.

Another challenge of the on-demand deployment for aerial wireless service is to incentivize cooperation between the ground BS and the UAV operators under the asymmetric information. As shown in [10], the ground BSs and UAVs can belong to different operators who seek to selfishly maximize their individual benefits. Hence, to request a UAV's assistance, a ground BS must offer an appropriate economic reward to the UAV operator for aerial wireless service. However, given that the BS has no prior knowledge of each UAV, there is no guarantee that the requested UAV is able to provide enough transmit power to satisfy the downlink demand. Therefore, designing an incentive mechanism is necessary to ensure a truthful information exchange between the UAV and BS systems, when the information among different network operators is asymmetric.

A. Related Works

The optimal deployment of UAVs for cellular service has been studied in [11]–[13]. In [11], the authors studied the optimal locations and coverage areas of UAVs that minimizes the transmit power. The work in [12] derived the minimum number of UAVs needed to satisfy the coverage and capacity constraints. In [13], the authors jointly optimized the UAV trajectory and the network resource allocation to maximize the throughput towards ground users. The problem of traffic offloading from an existing wireless network to UAVs has been addressed in [14]–[17]. In [14], the allocation problem of

UAVs to each geographic area was investigated to improve the spectral efficiency and reduce the delay. In [15] and [16], the authors optimized the trajectory of UAVs to provide wireless services to the cell-edge users. In [17], an unsupervised learning approach was presented to solve the deployment of a fleet of UAVs for traffic offloading. However, most of the existing works [11]–[17] assumed that the traffic demand of the cellular users is known a priori, which is challenging to estimate in a practical network. Furthermore, the works [11]–[17] optimized the performance of the cellular network in a centralized approach which assumes all UAVs belong to the same entity. Given the fact that the UAVs can belong to multiple operators, a new framework is needed to consider the individual utility of UAVs in the aerial communication service, while optimizing the performance of the ground cellular networks.

Meanwhile, in [18]–[20], a number of ML approaches are proposed to predict the traffic demands of cellular networks. In [18], a prediction framework is proposed to model the cellular data in the temporal and spatial domains. The authors in [19] predicted the locations of users during daily activities, based on pattern modeling. The work [20] provided surveys that focused on the general use of ML algorithms in cellular networks. Furthermore, the prior art in [21]–[23] studied the use of ML techniques to improve the performance of UAV-aided communications. In [21], an ML framework based on liquid state machine is proposed to optimize the caching content and resource allocation for each UAV. In [22], the authors investigated an ML approach to construct a radio map for autonomous path planning of UAVs. In [23], ML algorithms are applied to detect aerial users from the ground mobile users. However, most of the works in [18]–[23] aim to build an ML model to predict regular traffic patterns, while hotspot events are considered as an anomaly and excluded from these studies. In fact, none of the approaches proposed in [18]–[23] can effectively identify the hotspot areas or accurately predict excessive traffic load during the hotspot event. Thus, results of these prior works cannot enable a predictive UAV deployment for on-demand cellular service to alleviate the traffic congestion.

B. Contributions

The main contribution of this paper is a novel framework for optimally deploying UAVs to assist a ground cellular network in alleviating its downlink traffic congestion during hotspot events. The proposed framework divides the deployment process into four, inter-related and sequential stages: learning stage, association stage, movement stage, and service stage. For each stage, we evaluate the performance of the proposed framework, using an open-source dataset in [24]. Our main contributions include:

- A novel framework, based on the weighted expectation maximization (WEM) approach, is proposed to predict the downlink traffic demand for each cellular system in the learning stage. The proposed WEM method is a general version of the conventional expectation maximization (EM) algorithm, which enables a variable weight at each data point in the distribution modeling. In particular, the proposed approach identifies the user distribution, predicts the cellular data demand, and pinpoints the hotspot areas within the cellular system.
- In the association stage, to employ a UAV with sufficient on-board energy to satisfy the downlink demand, the framework of contract theory [25] is introduced, where each overloaded BS can jointly design the transmit power and unit reward of the target UAV. We analytically derive the sufficient and necessary conditions needed to guarantee a truthful information exchange between the BS and UAV operators. The proposed contract approach yields small communication overhead and a low computational complexity.
- Simulation results show that the mean relative error (MRE) of the proposed ML approach is around 10%. Compared with two baselines, an EM scheme and a k-mean algorithm, the proposed method yields a better prediction accuracy, particularly when the downlink traffic load in the cellular system becomes spatially uneven. Furthermore, simulation results show that the designed contract ensures a non-negative payoff of each UAV, and each UAV will truthfully reveal its communication capability by accepting the contract designed for itself.
- We evaluate the performance of the proposed approach with two event-driven allocation methods, based on the closest-distance and maximal-energy metrics, that deploy a target UAV after the network congestion occurs, without traffic prediction and contract design. Numerical results show that the proposed predictive method enables UAV operators to provide efficient downlink service for hotspot users, in terms of the downlink capacity, energy consumption, and service delay. Moreover, the proposed method significantly improves the economic revenues of both the BS and UAV networks, compared with two baseline schemes.

The rest of this paper is organized as follows. In Section II, we present the system model. The problem formulation is given in Section III. In Section IV, the ML approach is proposed to predict downlink traffic demands. In Section V, the feasible contract is designed with the optimal UAV being employed to offload the cellular traffic. Simulation results are presented in Section VI. Finally, conclusions are drawn in Section VII.

II. System Model

Consider a set $\mathcal{I}$ of I cellular BSs providing downlink wireless service to a group of user equipments (UEs) in a geographical area $\mathcal{A}$. Each BS $i \in \mathcal{I}$ serves an area $\mathcal{A}_i$, such that $\cup_{\forall i \in \mathcal{I}} \mathcal{A}_i = \mathcal{A}$, and $\mathcal{A}_i \cap \mathcal{A}_k = \emptyset$ for any $i \neq k \in \mathcal{I}$. The spatial distribution of the served UEs for each BS i is denoted by $f_i(\boldsymbol{y})$, where $\int_{\boldsymbol{y} \in \mathcal{A}_i} f_i(\boldsymbol{y})\, \mathrm{d}\boldsymbol{y} = 1$. A set $\mathcal{J}$ of J flying UAVs can provide additional cellular service, if the hotspot events happen in the ground cellular network. We assume that the group BSs and UAVs belong to different network operators, and different frequency bands are used for the ground and aerial downlink transmissions, separately.

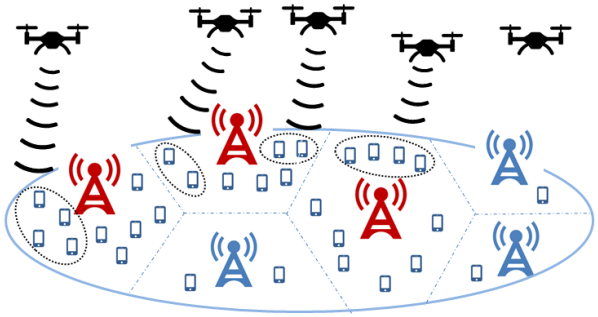


Fig. 1: The red BSs are having excessive traffic load in the downlink, thus each red BS requests a UAV to offload a part of UEs to the aerial cellular system.

A single antenna is equipped at each UE that can receive signals from both the ground BS and the UAV. Initially, a UE will connect to one of the ground BSs. However, as shown in Fig. 1, if a ground BS $i \in \mathcal{I}$ is overloaded in the downlink, BS i can request the assistant of UAVs to offload the service of some UEs. We assume that a UAV only serves the UEs of a single BS at each time, while each BS can employ multiple UAVs, based on the cellular traffic demand. In this regard, if the downlink traffic demand at the level of a given BS is excessive, such that no single UAV is capable to alleviate traffic congestion, then the BS will divide the offloaded UEs into multiple spatially-disjoint sets, and request an individual UAV for each UE set, independently. Meanwhile, each UAV is equipped with a directional antenna array that enables beamforming transmissions [26]. As a result, interference between different UAV networks is negligible.

A. Air-to-ground downlink communications

The path loss of the air-to-ground communication link from a typical UAV located at $\boldsymbol{x} \in \mathbb{R}^3$ to a typical ground UE that is located at $\boldsymbol{y} \in \mathbb{R}^3$ can be given by [27]:

$$h[dB](\boldsymbol{x}, \boldsymbol{y}) = 20 \log \left(\frac{4\pi f_c \left\| \boldsymbol{x} - \boldsymbol{y} \right\|}{c} \right) + \xi(\boldsymbol{x}, \boldsymbol{y}), \quad (1)$$

where f_c is the carrier frequency of UAV downlink communications, $\|\boldsymbol{x} - \boldsymbol{y}\|$ is the UAV-UE distance, c is the speed of light, and $\xi(\boldsymbol{x}, \boldsymbol{y})$ is the additional path loss of the air-to-ground channel, compared with the free space propagation. The value of $\xi(\boldsymbol{x}, \boldsymbol{y})$ can be modeled as a Gaussian distribution with different parameters $(\mu_{\text{LOS}}, \sigma^2_{\text{LOS}})$ and $(\mu_{\text{NLOS}}, \sigma^2_{\text{NLOS}})$ for the LOS and non-line-of-sight (NLOS) links, respectively. Then, the achievable data rate from a UAV $j \in \mathcal{J}$ located at $\boldsymbol{x}_j$ to a UE located at $\boldsymbol{y} \in \mathcal{A}_i$ is

$$r_{ij}(\boldsymbol{x}_j, \boldsymbol{y}, p_j) = w \log_2 \left(1 + \frac{g(\boldsymbol{x}_j, \boldsymbol{y}) p_j}{h(\boldsymbol{x}_j, \boldsymbol{y}) w n_0} \right), \quad (2)$$

where w is the downlink bandwidth of each UAV, $g(\boldsymbol{x}_j, \boldsymbol{y})$ is the antenna gain of UAV j towards the UE located at $\boldsymbol{y}$, p_j is the transmit power of UAV j, $h(\boldsymbol{x}_j, \boldsymbol{y})$ is the path loss in linear scale, and n_0 is the average noise power spectrum density at the UE. The probability of having a LOS link between UAV j located at $\boldsymbol{x}_j$ and the UE located at $\boldsymbol{y}$ is given by [28]:

$$P_{\text{LOS}}(\boldsymbol{x}_j, \boldsymbol{y}) = \frac{1}{1 + a \exp(-b[\frac{180}{\pi} \varphi(\boldsymbol{x}_j, \boldsymbol{y}) - a])}, \quad (3)$$

where a and b are constant values that depend on the communication environment, $\varphi(\boldsymbol{x}_j, \boldsymbol{y}) = \sin^{-1}(\frac{H_j}{\|\boldsymbol{x}_j - \boldsymbol{y}\|})$ is the elevation angle, and H_j is the altitude of UAV j. Consequently, the average downlink rate between a UAV j and a UE at $\boldsymbol{y} \in \mathcal{A}_i$ will be:

$$\begin{aligned} \bar{r}_{ij}(\boldsymbol{x}_j, \boldsymbol{y}, p_j) = {} & P_{\text{LOS}}(\boldsymbol{x}_j, \boldsymbol{y}) \cdot r^{\text{LOS}}_{ij}(\boldsymbol{x}_j, \boldsymbol{y}, p_j) \\ & + (1 - P_{\text{LOS}}(\boldsymbol{x}_j, \boldsymbol{y})) \cdot r^{\text{NLOS}}_{ij}(\boldsymbol{x}_j, \boldsymbol{y}, p_j). \end{aligned} \quad (4)$$

In order to serve multiple downlink UEs, each UAV applies a time-division-multiple-access (TDMA) technique[1] that divides the time resource evenly among all served UEs, and all bandwidth will be allocated to one single UE during each time slot [29]. By using suitable uplink control signals, the UAV-UE channel can be accurately measured, and, thus, the beamforming of UAV's antennas can be properly optimized towards the served UE. Consequently, the average rate that UAV j can provide to the hotspot UEs from BS i will be

$$C_{ij}(\boldsymbol{x}_j, p_j) = \int_{\mathcal{A}^c_i} \bar{r}_{ij}(\boldsymbol{x}_j, \boldsymbol{y}, p_j) f^c_i(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y}, \quad (5)$$

where $\mathcal{A}^c_i \subset \mathcal{A}_i$ is the hotspot area, $f^c_i(\boldsymbol{y})$ is the normalized spatial distribution of UEs within $\mathcal{A}^c_i$, and $\int_{\mathcal{A}^c_i} f^c_i(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} = 1$. When downlink congestion occurs, BS i detects the congested area $\mathcal{A}^c_i$ and offloads the UEs within $\mathcal{A}^c_i$ to the target UAV.

TABLE I: Summary of our notations

Notation	Description
I, J	Number of BSs and number of UAVs
T	Interval of the UAV's offloading service
$\boldsymbol{y}$	Location of a ground user
$\boldsymbol{x}_j$, $\boldsymbol{x}^*_{ij}$	Current location and service point of UAV j associated with BS i
f_i, S_i	User distribution and data demand distribution of BS i
$\mathcal{A}_i$, $\mathcal{A}^c_i$	Service area and hotspot area of BS i
Q_i, Q^c_i	Number of all users and number of hotspot users of BS i
d_i	Data demand of hotspot users within T of BS i
t_{ij}	Movement time of UAV j to the service location of BS i
p_j	Transmit power of UAV j
$\bar{r}_{ij}$	Average rate of UAV j to each hotspot user of BS i
C_{ij}	Average rate of UAV j to all hotspot users of BS i
B_{ij}	Downlink data that UAV j provides to hotspot UEs of BS i within T
u_i	Unit payment of BS i
ρ_i, ρ^c_i	Average rate demand per user/ hotspot user of BS i
U_{ij}	Utility of BS i by employing UAV j
R_{ij}	Utility of UAV j by providing offloading service to BS i
θ_{ij}	Type of UAV j with respect to BS i
$\boldsymbol{\omega}$, $\boldsymbol{\pi}$	Weight vectors in the user and demand distribution models
$\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$	Mean and covariance of Gaussian distribution

B. UAV deployment process

Given the average downlink rate of each UAV in (5), the next step is to deploy suitable UAVs to offload the traffic and alleviate the downlink congestion in the ground cellular network. To facilitate the analysis, we assume that the service interval of each UAV a constant T. As shown in Fig. 2,

[1]The focus of this work is on the deployment stage, and, hence, we do not optimize the multiple access scheme type or operation. Optimizing multiple access can be done post-deployment and will be subject to future work.

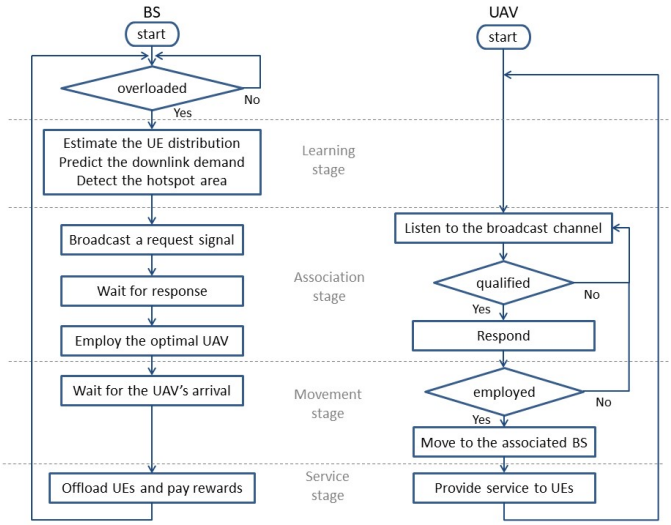


Fig. 2: Flowchart of the proposed UAV predictive deployment process for each BS (left) and each UAV (right).

the deployment process has four sequential stages: learning stage, association stage, movement stage, and service stage. The details of each stage are given as next:

1) Learning stage: For each BS $i \in \mathcal{I}$, once the downlink traffic exceeds its network capacity, a learning stage with a fixed duration τ starts. During τ, BS i collects the transmission record $\mathcal{S}_i = \{(s, \boldsymbol{y}, t) | \boldsymbol{y} \in \mathcal{A}_i, t \in [\Delta t, 2\Delta t, \cdots, \tau]\}$, where s is the data rate that BS i provides to the UE located at $\boldsymbol{y}$ at time t, and Δt is the time slot during which the downlink rate can be considered to be constant. Given that the hotspot area $\mathcal{A}_i^c$ and the UE distribution $f_i(\boldsymbol{y})$ is unknown, a learning stage is necessary for BS i to estimate the spatial distribution of UEs and the traffic demand of the on-going hotspot event. Considering common events, such as sport games and outdoor concerts, where mobile users are often confined to seat or geographically constrained spaces, the mobility of hotspot UEs is scarce. Thus, we assume that the UE distribution $f_i(\boldsymbol{y})$ during one T is time-invariant. Furthermore, to estimate the traffic demand within the congested area, a spatial density function $S_i(\boldsymbol{y})$ is proposed to evaluate the average data rate per UE at each location $\boldsymbol{y} \in \mathcal{A}_i$. The proposed approach for estimating the UE distribution and traffic demand will be discussed in Section IV. Consequently, the total data demand d_i from a hotspot area $\mathcal{A}_i^c$ during a time interval T will be given by:

$$d_i = \int_t^{t+T} \int_{\boldsymbol{y} \in \mathcal{A}_i^c} S_i(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} \, \mathrm{d}t = T \int_{\boldsymbol{y} \in \mathcal{A}_i^c} S_i(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y}. \quad (6)$$

Next, the BS will estimate the necessary number of UAVs to alleviate downlink congestion and calculate the optimal service location of each target UAV. Following from [2, equation (42)] and [11, equations (10) and (11)], given the UE distribution $f_i^c(\boldsymbol{y})$ and the hotspot area $\mathcal{A}_i^c$, the optimal location $\boldsymbol{x}_{ij}^*$ of a target UAV j in serving BS i can be derived in a way to minimize the transmit power $p_{ij}(\boldsymbol{x}_{ij}^*, \rho_i^c)$, while satisfying the average rate requirement ρ_i^c per UE. The average rate per UE is defined by the ratio of the sum data rate within the hotspot area $\mathcal{A}_i^c$ over the total number of hotspot UEs Q_i^c, where $\rho_i^c = \frac{d_i}{TQ_i^c}$. Thus, the optimal service location of the target UAV can be calculated by BS i, prior to the UAV's deployment. We define $p_{\max}$ to be the maximum transmit power of each UAV, which is limited by the antennas' hardware, and $\eta \in (0, 1)$ to be the ratio of efficient transmission time to the service time T, due to the signal overhead and the channel measurement process. If $d_i > \eta T C_{ij}(\boldsymbol{x}_{ij}^*, p_{\max})$, then even though a UAV is located at the optimal service point $\boldsymbol{x}_{ij}^*$ and it applies the maximum transmit power $p_{\max}$, the downlink demand d_i cannot be satisfied. In this case, using a single UAV $j \in \mathcal{J}$ is no longer sufficient to offload the hotspot traffic. Therefore, BS i will evenly divide the hotspot area $\mathcal{A}_i^c$, based on the downlink data demand, into N disjoint areas $\{\mathcal{A}_i^c(n)\}_{n=1,\cdots,N}$, where $\int_{\boldsymbol{y} \in \mathcal{A}_i^c(n)} S_i(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} = \frac{d_i}{N}$, and N is the smallest integer needed to guarantee that, for each subset $n = 1, \cdots, N$, the following requirement holds:

$$d_i(n) = \frac{d_i}{N} < \eta T C_{ij}(\boldsymbol{x}_{ij}^*(n), p_{\max}). \quad (7)$$

For each $n = 1, \cdots, N$, BS i will deploy a UAV onto the service point $\boldsymbol{x}_{ij}^*(n)$ to offload the downlink traffic with the subarea $\mathcal{A}_i(n)$. The requests of multiple UAVs to different subareas are sequential and independent at each round.

2) Association stage : In the association stage, each overloaded BS i requests the assistance of a UAV, by broadcasting a signal with the downlink demand $d_i(n)$ and the service location $\boldsymbol{x}_{ij}^*(n)$ for each subset n. A first-call-first-serve scheme is applied, and each BS $i \in \mathcal{I}$ will listen to the broadcast channel before sending the signal. If the channel is occupied by another BS, then BS i will wait until the on-going association is completed. For each BS i, the goal is to request a UAV that has enough on-board power to meet the downlink demand $d_i(n)$ of the UEs within $\mathcal{A}_i^c(n)$ for each $n = 1, \cdots, N$. The optimal UAV association to each overloaded BS will be studied in Section V.

3) Movement stage: After the association stage, the selected UAV j starts to move from its current location $\boldsymbol{x}_j$ to the service point $\boldsymbol{x}_{ij}^*$ of its target BS i. The duration t_{ij} of the movement stage depends on the distance $\|\boldsymbol{x}_j - \boldsymbol{x}_{ij}^*\|$ and the average speed v_j of UAV j.

4) Service stage: Once it reaches the service point, UAV j will provide downlink communications to its group of associated UEs for a time period $T - t_{ij}$. Note that, during the movement and service stages, the employed UAV is fully dedicated to its associated BS. Thus, the UAV cannot be requested by any other BSs until the end of its current service. Furthermore, to guarantee a sufficient service time, the maximum travel time of UAV j is limited by $t_{ij} \leq \kappa_i T$, where $\kappa_i \in (0, 1)$. If the travel time exceeds $\kappa_i T$, UAV j is not a potential choice for BS i.

After the service stage ends, the BS-UAV association will end. Then, UAV j will listen to the broadcast channel, if its remaining on-board energy E_j can support another service period T; otherwise, the UAV will move to a nearby recharging station. We assume that a number of recharging stations are deployed, such that a UAV can access a recharging station within a short flight time from any location in $\mathcal{A}$. Thus, the movement energy to a recharging station is negligible to effect

the BS-UAV association results. In order to optimally associate UAVs to each overloaded BS, we first define a utility function that each BS aims to maximize when selecting a UAV to offload cellular traffic in Section II-C. Next, the UAV's utility function is given in Section II-D that defines its economic payoff from serving a ground BS.

C. Utility function of a ground BS

In TDMA downlink transmissions, the employed UAV j evenly divides the service time $T - t_{ij}$ to each hotspot UE. Therefore, based on the average downlink rate in (5), the achievable data amount that UAV j can provide to the UEs of BS i is

$$B_{ij}(p_j) = \eta(T - t_{ij})C_{ij}(p_j). \tag{8}$$

Note that, the movement duration t_{ij} and the transmit power p_j are private information for UAV j, and, thus, BS i cannot know their values during the service request process.

Then, the utility of BS i, by employing UAV j to offload the excess cellular traffic, will be:

$$U_{ij}(u_i, p_j, d_i) = \beta B_{ij}(p_j) - u_i d_i, \tag{9}$$

where β is the payment from UEs to BS i (per bit of downlink data), and u_i is the unit payment that BS i gives to UAV j (per bit of aerial data service). Thus, the first term in (9) represents the reward that BS i gets from its UEs by employing UAV j to provide aerial cellular service, and the second term is the total payment that BS i gives to UAV j.

D. Energy model and utility function of a UAV

In the considered problem, the power consumption of each UAV consists of three main components: the transmit power p_j, the propulsion power m, and the hovering power p_h. For tractability and as done in [30], we ignore the acceleration and deceleration stages during the UAV's movement, and the propulsion power m is considered as a constant for a fixed flying speed. Then, the travel time t_{ij} can be uniquely determined based on the moving distance $\|\boldsymbol{x}_j - \boldsymbol{x}_{ij}^*\|$. During the service stage, the maximum available power that UAV j can use for downlink transmissions will be $p_{ij}^{\max} = \frac{E_j - mt_{ij} - p_h(T-t_{ij})}{T-t_{ij}}$, where mt_{ij} is the energy consumed during the UAV's movement, and $p_h(T - t_{ij})$ is the hovering energy during the service stage. Therefore, we have the transmit power $p_j \in [p_{ij}(\boldsymbol{x}_{ij}^*, \rho_i^c), \min\{p_{ij}^{\max}, p_{\max}\}]$, where $p_{ij}(\boldsymbol{x}_{ij}^*, \rho_i^c)$ is the minimum required power to satisfy the downlink data demand, and $p_{\max}$ is the maximum transmit power. Without loss of generality, we assume that $p_{ij}(\boldsymbol{x}_{ij}^*, \rho_i^c) \leq \min\{p_{ij}^{\max}, p_{\max}\}$ holds. Otherwise, UAV j is not a potential option for BS i. Consequently, the utility that a UAV $j \in \mathcal{J}$ can achieve from providing the aerial cellular service to the UEs of BS i will be:

$$R_{ij}(u_i, p_j, d_i) = u_i d_i - \alpha[p_j(T - t_{ij}) + p_h(T - t_{ij}) + mt_{ij}], \tag{10}$$

where α is a unit cost per Joule of UAV's on-board energy. The first term in (10) is the reward that UAV j obtains from BS i, and the second term is the energy cost.

III. Problem formulation

The objective of an overloaded BS is to employ a suitable UAV with sufficient on-board power to offload excessive cellular traffic, while maximizing the utility function in (9). Meanwhile, the goal of each UAV is to optimize its utility in (10). However, by comparing (9) and (10), we realize that $\arg\max_{u_i,p_j} U_{ij} = \arg\min_{u_i,p_j} R_{ij}$ and $\arg\max_{u_i,p_j} R_{ij} = \arg\min_{u_i,p_j} U_{ij}$. Therefore, each BS-UAV pair has conflicting interests. Given that the BSs and UAVs belong to different operators, each will maximize its own utility. The conflict between each BS and each UAV is irreconcilable.

Meanwhile, since the values of the unit payment u_i and the data demand d_i will be broadcast by BS i during the association stage, each UAV j has all necessary information to determine its utility. However, BS i cannot easily acquire some private information of each UAV, such as its current location and onboard energy, which causes the asymmetric information. Since private information of each UAV determines its travel time to a BS and the downlink communication capacity, it is essential for the BS to have accurate information to evaluate the service performance of each UAV. In order to guarantee a truthful information exchange, each BS i can jointly design (u_i, p_j) to ensure mutual benefit for both the BS and UAV operators, so that the conflict of interest can be properly resolved. Therefore, we let $\phi_{ij} = (u_i, p_j)$ be a *traffic offload contract*, which captures the values of p_j and u_i if BS i employs UAV j to offload its hotspot UEs. In order to understand the relationship between the unit payment u_i and the transmit power p_j, we divide both sides of (10) by $\alpha(T - t_{ij})$ and rewrite the utility of UAV j as follows:

$$\begin{aligned}\tilde{R}_{ij}(u_i, p_j, d_i) &= \frac{d_i}{\alpha(T - t_{ij})}u_i - p_j - \frac{mt_{ij}}{T - t_{ij}} - p_h, \\ &= \theta_{ij}u_i - p_j - M_{ij},\end{aligned} \tag{11}$$

where the values of $\theta_{ij} = \frac{d_i}{\alpha(T-t_{ij})}$ and $M_{ij} = \frac{mt_{ij}}{T-t_{ij}} + p_h$ are determined for each BS-UAV pair.

Since θ_{ij} determines the sensitivity of $\tilde{R}_{ij}$ to the increase of u_i and p_j in (11), its value is essential for the joint design of (u_i, p_j). Therefore, we define θ_{ij} as the *type* of UAV j with respect to BS i, where $\theta_{ij} \in \Theta_i = [\frac{d_i}{\alpha T}, \frac{d_i}{\alpha(1-\kappa_i)T}]$. Note that, due to the privacy of t_{ij}, the type θ_{ij} of each UAV $j \in \mathcal{J}$ is unknown for BS i. In order to design the contract without knowing each UAV's type, before broadcasting the request signal, BS i will design a set of contracts $\Phi_i(\Theta_i) = \{\phi_{ij}(\theta_{ij})|\forall\theta_{ij}\} = \{(u_i(\theta_{ij}), p_j(\theta_{ij}))|\forall\theta_{ij}\}$ for all UAV types $\theta_{ij} \in \Theta_i$, where $u_i(\theta_{ij})$ represents the payment that BS i pays to UAV j per bit of data, given that UAV j is of type θ_{ij}, and $p_j(\theta_{ij})$ is the transmit power that UAV j of type θ_{ij} provides to serve BS i. Then, (11) becomes $\tilde{R}(\theta_{ij}) = \theta_{ij}u_i(\theta_{ij}) - p_j(\theta_{ij}) - M_{ij}$.

Meanwhile, to ensure that a UAV will accept the contract of its own type, two constraints, based on contract theory [25], must be considered, which are individual rationality (IR) condition and incentive compatibility (IC) condition.

Definition 1 (Individual Rationality)**.** *A contract designed by BS i satisfies the IR constraint, if a UAV of any type $\theta_{ij} \in \Theta_i$ will receive a non-negative payoff from BS i by accepting the*

contract item for type θ_{ij}, i.e. $\theta_{ij}u_i(\theta_{ij}) - p_j(\theta_{ij}) - M_{ij} \geq 0$, $\forall \theta_{ij} \in \Theta_i$.

A contract satisfying the IR condition guarantees that the reward that each UAV $j \in \mathcal{J}$ can obtain from serving BS i is great than or equal to zero. Compared with the non-employed state in which the payoff is always zero, each UAV is willing to accept the contract from the requesting BS, as long as its contract satisfies the IR condition.

Definition 2 (Incentive Compatibility)**.** *A contract designed by BS i satisfies the IC constraint, if a UAV of type θ_{ij} will get the highest utility from BS i by accepting the contract designed for its own type θ_{ij}, compared with all the other types θ in Θ_i, i.e. $\theta_{ij}u_i(\theta_{ij}) - p_j(\theta_{ij}) - M_{ij} \geq \theta_{ij}u_i(\theta) - p_j(\theta) - M_{ij}$, $\forall \theta \in \Theta_i$.*

A contract satisfying IC condition guarantees that each UAV j will only accept the contract designed for its own type θ_{ij}, since accepting the contract of any other type $\theta \in \Theta_i$ will result in a lower or the same reward. A contract satisfying both IR and IC conditions is called a *feasible* contract, which ensures the UAV will accept and only accept the contract designed for its type.

Consequently, for each overloaded BS $i \in \mathcal{I}$, the objective is to maximize its utility in (9), by estimating the downlink data demand d_i within the hotspot area $\mathcal{A}_i^c$, designing the contract set Φ_i for each UAV of any type in Θ_i, and determining an optimal UAV $j \in \mathcal{J}$ to offload the excessive cellular service. We formulate this predictive UAV deployment problem as follows,

$$\max_{\substack{\{(u_i(\theta_{ij}),p_j(\theta_{ij}))\}_{\forall \theta_{ij}} \\ j \in \mathcal{J}}} \quad U_{ij}(u_i(\theta_{ij}), p_j(\theta_{ij}), d_i), \tag{12a}$$

$$\text{s. t.} \quad R_{ij}(\theta_{ij}) \geq 0, \tag{12b}$$

$$R_{ij}(\theta_{ij}) \geq R_{ij}(\theta), \forall \theta \in \Theta_i, \tag{12c}$$

$$p_{ij}(\boldsymbol{x}_{ij}^*, \rho_i^c) \leq p_j(\theta_{ij}) \leq \min\{p_{ij}^{\max}, p_{\max}\}, \tag{12d}$$

$$t_{ij} \leq \kappa_i T, \tag{12e}$$

$$d_i > 0, u_i(\theta_{ij}) > 0. \tag{12f}$$

The objective function (12a) is the utility that BS i obtains from employing UAV j of type θ_{ij}. (12b) and (12c) are the IR and IC constraints, respectively. (12d) is the constraint on the transmit power, and (12e) limits the maximum travel time. (12f) imposes a positive downlink demand within $\mathcal{A}_i^c$, and a positive unit payment. Note that, (12c) itself is an optimization problem, which must be first addressed to satisfy the IC condition. Since the selection of θ_{ij} will jointly determine the values of the objective function and all constraints in (12), θ_{ij} becomes the key variable to find the optimal association result. To simplify the optimization problem (12), we first derive the necessary and sufficient conditions for IC and IR constraints, based on the UAV type θ_{ij}, which essentially reduces to the problem of designing a feasible contract.

Therefore, in order to solve the predictive UAV deployment problem in (12), first, a learning-based approach is proposed to predict the downlink demand d_i in Section IV. Next, the traffic offload contract Φ_i is developed in Section V, with the optimal UAV being selected to maximize the utility of BS i.

IV. Learning Stage: Estimation of Cellular Traffic Demand

In this section, our goal is to estimate the UE distribution and the downlink data demand during a hotspot event. This estimation is necessary to solve (12) because the data demand d_i is needed to determine the type θ_{ij} of each UAV j with respect to BS i. To enable an accurate modeling, BS i collects the downlink transmission records during the learning stage. For notation simplicity, let N be the total number of records, and $\mathcal{S}_i$ can be rewritten as $\{(s_n, \boldsymbol{y}_n, t_n)|n = 1, \cdots, N\}$. In Section IV-A, we extract the spatial distribution $f_i(\boldsymbol{y})$ of the downlink UEs. Next, in Section IV-B, the downlink rate $S_i(\boldsymbol{y})$ is modeled and the hotspot area $\mathcal{A}_i^c$ is determined. Consequently, the downlink data demand d_i is given by (6).

A. Estimation of the UE distribution

Given $\mathcal{S}_i$, BS i can model the UE distribution, using the location information $\mathcal{Y} = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_N\}$. We assume that each UE's location follows a latent distribution $f_i(\boldsymbol{y})$, and each $\boldsymbol{y}_n$ is an independent sample from this distribution. A Gaussian mixture model (GMM), which is the weighted sum of multiple Gaussian distributions, can model the UE's distribution, as follows:

$$f_i(\boldsymbol{y}) = \sum_{l=1}^{L} \omega_l \mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l), \tag{13}$$

where L is the number of Gaussian distributions, and ω_l, $\boldsymbol{\mu}_l$, and $\boldsymbol{\Sigma}_l$ are the weight, mean and variance of the l-th Gaussian, respectively, with $\omega_l \in (0, 1)$ and $\sum_l \omega_l = 1$. The value of ω_l represents the probability that the data point $\boldsymbol{y}$ is generated by the l-th distribution. GMM has been widely applied in [31]–[33] to model the distribution of a latent variable based the sampled data. Due to its special feature of multiple clusters, GMM is particularly appropriate to model the UE distribution in the congested area, where each hotspot area corresponds to a Gaussian center.

Given the location record $\mathcal{Y}$, the expectation-maximization (EM) algorithm [33] is applied to optimize the parameters $\{\omega_l, \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l\}_{l=1,\cdots,L}$ in (13) via an iterative approach, which maximizes a log-likelihood function $\ln p(\mathcal{Y}|\boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \ln \Pi_{n=1}^{N} \left(\sum_{l=1}^{L} \omega_l \mathcal{N}(\boldsymbol{y}_n|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l) \right)$. After initialization, the EM algorithm alternates between the E and M steps. First, in the E step, the posterior probability that $\boldsymbol{y}_n$ is generated by the l-th Gaussian is calculated by

$$v_{nl} = \frac{\omega_l \mathcal{N}(\boldsymbol{y}_n|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}{\sum_{z=1}^{L} \omega_z \mathcal{N}(\boldsymbol{y}_n|\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)}. \tag{14}$$

Then, in the M step, the parameters are updated using the posterior probability (14) by

$$\begin{aligned} \boldsymbol{\mu}_l &= \frac{\sum_n v_{nl} \boldsymbol{y}_n}{\sum_n v_{nl}}, \\ \boldsymbol{\Sigma_l} &= \frac{\sum_n v_{nl} (\boldsymbol{y}_n - \boldsymbol{\mu}_l)(\boldsymbol{y}_n - \boldsymbol{\mu}_l)^T}{\sum_n v_{nl}}, \\ \omega_l &= \frac{\sum_n v_{nl}}{N}. \end{aligned} \tag{15}$$

After each EM iteration, the updated parameters will result in an increase of the log-likelihood function, and the algorithm is guaranteed to converge to a local optimum [33].

B. Estimation of the downlink data rate

In order to predict the downlink demand d_i, each BS i needs to capture the spatial feature of the cellular traffic. Based on the assumption of the time-invariant data demand, we define the traffic *density* $\bar{S}_i(\boldsymbol{y})$ at each location $\boldsymbol{y} \in \mathcal{A}_i$ as the time-average downlink rate at $\boldsymbol{y}$ during the learning stage, where $\bar{S}_i(\boldsymbol{y}) = \frac{1}{\tau}\sum_{(s_n,\boldsymbol{y},t_n)\in\mathcal{S}_i} s_n \Delta t$. In order to generate a continuous model $S_i(\boldsymbol{y})$ that captures the spatial features of the downlink traffic density, a Gaussian mixture function (GMF) is proposed as follow,

$$S_i(\boldsymbol{y}) = \sum_{k=1}^{K} \pi_k \exp\left(\frac{-(\boldsymbol{y}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_k)}{2}\right), \quad (16)$$

where K is the number of basis functions, and π_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$ are the coefficient, mean and variance of the k-th Gaussian function. Thus, the traffic density at location $\boldsymbol{y}$ is modeled by the sum of K Gaussian functions with coefficient $\{\pi_k\}_{k=1,\cdots,K}$.

Note that, the GMF in (16) is different from the GMM in (13). First, a GMM has a probabilistic interpretation, while a GMF is a deterministic function that calculates the traffic density at each location $\boldsymbol{y}$ by adding the values of K Gaussian functions with different coefficients. Second, the sum of each coefficient π_k in GMF represents the total volume of downlink traffic demand. Thus, it is always greater than one, which make a difference from the unit weight-sum in GMM.

To properly model the downlink traffic density $\bar{S}_i$, the parameters $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1,\cdots,K}$ in (16) need to be optimized. Since the EM method associates the same weight to all data points, it is not suitable to the traffic density modeling, because each data point $\boldsymbol{y}_n$ can have a different traffic density $\bar{S}_i(\boldsymbol{y}_n)$. In order to adapt the weight of each location $\boldsymbol{y}_n$ in determining the parameter values according to the traffic density $\bar{S}_i(\boldsymbol{y}_n)$, as well as to capture the spatial diversity of the traffic load within the cellular network, a weighted expectation maximization (WEM) algorithm is proposed to optimize the parameters in the traffic density model $S_i(\boldsymbol{y})$.

In the proposed WEM method, the initial value of each Gaussian center $\boldsymbol{\mu}_k$ is the location $\boldsymbol{y}_k$ that has the k-th highest traffic density in $\bar{S}_i(\boldsymbol{y})$. The initial variance $\boldsymbol{\Sigma}_k$ equals the identity matrix with the equal weight $\pi_k = \frac{1}{K}\sum_{\boldsymbol{y}} \bar{S}_i(\boldsymbol{y})$. Then, the WEM algorithm updates $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1,\cdots,K}$ via an iterative approach. In the E step, the percentage that each Gaussian function k contributes to the traffic density at location $\boldsymbol{y}_n$ is evaluated via $v_{nk} = \frac{\pi_k \mathcal{N}(\boldsymbol{y}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_{k=1}^{K}\pi_k \mathcal{N}(\boldsymbol{y}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}$. Next, in the M step, the parameters of each Gaussian function will be updated in a weighted approach, where the mean $\boldsymbol{\mu}_k$ is recalculated via

$$\boldsymbol{\mu}_k = \frac{\sum_n v_{nk}\boldsymbol{y}_n \bar{S}_i(\boldsymbol{y}_n)}{\sum_n v_{nk}\bar{S}_i(\boldsymbol{y}_n)}, \quad (17)$$

which is a sum of all locations $\boldsymbol{y}_n \in \mathcal{Y}$, weighted by the posterior probability v_{nk} and the traffic density $\bar{S}_i(\boldsymbol{y}_n)$. Thus, a location $\boldsymbol{y}_n$ with a higher traffic density $\bar{S}_i(\boldsymbol{y}_n)$ will have a higher weight in determining the value of $\boldsymbol{\mu}_k$, and the center of Gaussian k will gradually be driven closer to the high-density locations. Similarly, the variance $\boldsymbol{\Sigma}_k$ and the linear coefficient π_k of each Gaussian function are also updated, with weights $\bar{S}_i(\boldsymbol{y}_n)$, by

$$\begin{aligned}\boldsymbol{\Sigma}_k &= \frac{\sum_n v_{nk}(\boldsymbol{y}_n-\boldsymbol{\mu}_k)(\boldsymbol{y}_n-\boldsymbol{\mu}_k)^T \bar{S}_i(\boldsymbol{y}_n)}{\sum_n v_{nk}\bar{S}_i(\boldsymbol{y}_n)},\\ \pi_k &= \frac{\sum_n v_{nk}\bar{S}_i(\boldsymbol{y}_n)}{\sum_k\sum_n v_{nk}\bar{S}_i(\boldsymbol{y}_n)}.\end{aligned} \quad (18)$$

Furthermore, similar to the EM approach, a WEM method will converge to a local optimum, which maximizes the weighted conditional log-likelihood function [1].

Although the EM and WEM methods have similar mathematical expressions, the physical meaning and iterative process are fundamentally different. First, different from the unit sum-weight in (13), the weight-sum of the WEM method represents the total volume of the downlink data demand, which can be any positive value. Second, when updating the Gaussian parameters, the proposed WEM method considers the traffic density at each location and assigns a higher weight to the location with higher demand in the density model. In contrast, EM method associates each point with an equal weight. Thus, the spatial diversity of the traffic load cannot be properly captured. Therefore, the proposed WEM approach expands the application range of the EM scheme, and can be seen as a general version of EM, which models the distribution with a variable weight at each data point.

The hotspot area $\mathcal{A}_i^c$ is a location set in which the traffic destiny is much higher than other locations in $\mathcal{A}_i$. Given the traffic density model S_i, the average traffic density in $\mathcal{A}_i$ is given by $\bar{s}_i = \frac{1}{|\mathcal{A}_i|}\int_{\boldsymbol{y}\in\mathcal{A}_i} S_i(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}$, where $|\mathcal{A}_i|$ denotes the area of $\mathcal{A}_i$. Then, by calculating the traffic density at each Gaussian center $\{\boldsymbol{\mu}_k\}_{k=1,\cdots,K}$, the mean $\boldsymbol{\mu}_k^*$ with the highest traffic density is chosen, and its neighborhood area, where the traffic density is higher than $\bar{s}_i$ forms the hotspot area $\mathcal{A}_i^c$. The downlink UEs within $\mathcal{A}_i^c$ will be offloaded to the aerial cellular network. Based on the traffic density model $S_i(\boldsymbol{y})$ and the hotspot area $\mathcal{A}_i^c$, the predicted data amount d_i for a time interval T can be calculated based on (6).

Given the downlink traffic demand d_i and the UE distribution $f_i(\boldsymbol{y})$, all variables in (12) have determined values, except for the unit payment u_i and the transmit power p_j. Next, in order to to solve (12), we will jointly decide the value of (u_i, p_j), by designing the feasible contract between an overloaded BS i with each UAV $j \in \mathcal{J}$.

V. Association stage: Contract Design and UAV Allocation

A. Contract design

Given the predicted traffic demand d_i, a BS $i \in \mathcal{I}$ can request UAVs to offload the UEs within the hotspot area $\mathcal{A}_i^c$, so that the future downlink congestion can be alleviated. However, to employ a qualified UAV to meet the downlink demand, each BS needs to carefully design the contract $\Phi_i = \{(u_i(\theta_{ij}), p_j(\theta_{ij}))|\forall\theta_{ij} \in \Theta_i\}$ for UAVs of any type

θ_{ij}. The feasible contract satisfying the IR and IC conditions can guarantee that each UAV will accept the contract designed for its own type and provide the required downlink transmissions. To develop a feasible contract set, we first analyze the sufficient and necessary conditions for a feasible contract.

Proposition 1. *[Necessary Condition] For any $\theta_{ij}, \theta_{ij}^{'} \in \Theta_i$, if $\theta_{ij} > \theta_{ij}^{'}$, then $u_i(\theta_{ij}) \geq u_i(\theta_{ij}^{'})$ and $p_j(\theta_{ij}) \geq p_j(\theta_{ij}^{'})$.*

Proof. See Appendix A. □

Proposition 1 shows that for a typical UAV j, if its type with respect to a typical BS i increases from $\theta_{ij}^{'}$ to θ_{ij}, then it will receive a higher unit payment $u_i(\theta_{ij}) \geq u_i(\theta_{ij}^{'})$, and in return, it should provide a larger transmit power $p_j(\theta_{ij}) \geq p_j(\theta_{ij}^{'})$. Given that $\theta_{ij} = \frac{d_i}{\alpha(T-t_{ij})}$, a higher type θ_{ij} indicates either a higher downlink demand d_i, or a longer travel time t_{ij}. In the first case, if the downlink demand is higher, the employed UAV must increase the transmit power to satisfy the larger traffic needs. Thus, $p_j(\theta_{ij})$ will increase. On the other hand, if UAV j travels for a long time t_{ij}, it consumes more energy on movement, which requires a higher unit payment $u_i(\theta_{ij})$ to compensate for the energy cost. Therefore, a UAV of a higher type is required to provide more transmit power, and will be given a higher unit payment. The conclusion in Proposition 1 will lead to the necessary and sufficient conditions of a feasible contract, as shown next.

Theorem 1. *A contract set $\Phi_i = \{(u_i(\theta_{ij}), p_j(\theta_{ij}))|\forall\theta_{ij}\}$ satisfies IR and IC constraints, if and only if all the following three conditions hold: (a) $\frac{\mathrm{d}p_j(\theta_{ij})}{\mathrm{d}\theta_{ij}} \geq 0$ and $\frac{\mathrm{d}u_i(\theta_{ij})}{\mathrm{d}\theta_{ij}} \geq 0$, (b) $\theta^{min}u_i(\theta^{min}) - p_j(\theta^{min}) - M_{ij} \geq 0$, (c) $\frac{\mathrm{d}p_j(\theta_{ij})}{\mathrm{d}\theta_{ij}} = \theta_{ij} \cdot \frac{\mathrm{d}u_i(\theta_{ij})}{\mathrm{d}\theta_{ij}}$.*

Proof. See Appendix B. □

Theorem 1 gives the necessary and sufficient conditions for a contract set Φ_i to jointly satisfy the IC constraint in (12c) and the IR constraint in (12b). Therefore, each feasible solution of Theorem 1 can guarantee that a UAV only accepts the contract designed for its own type, and provides the required transmit power to meet the downlink demand. Here, we note that Theorem 1 results in a loose solution set. In essence, all of contracts from this solution set meet the necessary and sufficient conditions of the IC and IR requirements, and, thus, they are optimal in the contract-theoretic problem. Meanwhile, Theorem 1 provides each BS with more freedom to choose the feasible contract based on its real-time communication need. In order to minimize the communication overhead in the association stage, we aim to propose a contract with the lowest complexity and the least broadcast overhead. Therefore, to enable an efficient BS-UAV association, we propose the best contract with $\frac{\mathrm{d}u_i(\theta_{ij})}{\mathrm{d}\theta_{ij}} = \gamma_i > 0$. Consequently, the feasible contract that is proposed by BS i is given as follows.

Lemma 1. *Under the condition that $\frac{\mathrm{d}u_i(\theta_{ij})}{\mathrm{d}\theta_{ij}} = \gamma_i$, the feasible contract between BS i and a UAV j of type θ_{ij} is $\phi_{ij} = (u_i, p_j) = (\gamma_i\theta_{ij}, \gamma_i\theta_{ij}^2/2)$, where $\gamma_i = \frac{2\alpha^2T^2p_h}{d_i^2}$.*

Proof. Based on $\frac{\mathrm{d}u_i(\theta_{ij})}{\mathrm{d}\theta_{ij}} = \gamma_i$ and condition (c) of Theorem 1, we have $u_i = \gamma_i\theta_{ij}$, $p_j = \gamma_i\theta_{ij}^2/2$, and condition (a) holds naturally. For BS i, the minimal UAV type is $\theta^{\min} = \frac{d_i}{\alpha T}$, when $t_{ij} = 0$. Therefore, condition (b) becomes $\gamma_i \geq \frac{2M_{ij}}{\theta^{\min 2}} = \frac{2\alpha^2T^2p_h}{d_i^2}$. Therefore, we set $\gamma_i = \frac{2\alpha^2T^2p_h}{d_i^2}$. □

Therefore, for each overloaded BS i, the designed contract is $(u_i, p_j) = (\gamma_i\theta_{ij}, \gamma_i\theta_{ij}^2/2)$ with $\gamma_i = \frac{2\alpha^2T^2p_h}{d_i^2}$ for each UAV in $\mathcal{J}$ with any type θ_{ij}.

Algorithm 1 Proposed process for UAV predictive deployment

For each BS $i \in \mathcal{I}$, once downlink communication exceeds the network capacity, **do**:

1. Learning stage:
- (a) BS i collects $\mathcal{S}_i$ to model the UE distribution $f_i(\boldsymbol{y})$, estimate the downlink traffic density $S_i(\boldsymbol{y})$, and detect the hotspot area $\mathcal{A}_i^c$ based on the WEM approaches proposed in Section IV.
- (b) BS i calculates the downlink demand d_i via (6), estimates the number N of required UAVs through (7), and computes the service point $\boldsymbol{x}_{ij}^*$ for each target UAV j, based on [11].

2. Association stage: for $n = 1, \cdots, N$:
- (a) BS i listens to the broadcast channel. If the channel is occupied, wait; otherwise, BS i broadcasts the request signal with $d_i(n), \boldsymbol{x}_{ij}^*(n)$, κ_i, and $\Phi_i(n) = \{\gamma_i\theta_{ij}, \frac{\gamma_i}{2}\theta_{ij}^2|\forall\theta_{ij}\}$.
- (b) Each UAV $j \in \mathcal{J}$ listens the broadcast channel. After receiving the request from BS i, each UAV calculates the movement time t_{ij}, its UAV type θ_{ij} with respect to BS i, and the available transmit power $p_{ij}^{\max}$ after arriving at $\boldsymbol{x}_{ij}^*$. If $p_{ij}^{\max} \geq \frac{\gamma_i}{2}\theta_{ij}^2$ and $t_{ij} \leq \kappa_i T$, UAV j replies θ_{ij} to BS i; otherwise, ignore.
- (c) BS i identifies the feasible UAV set $\mathcal{J}_i$, and employs the optimal UAV $j^* = \arg\min_{j\in\mathcal{J}_i}\theta_{ij}$.
- (d) If $n = N$, BS i releases the broadcast channel; otherwise, go back to 2(a).

3. Movement stage: The employed UAV j^* starts to move towards the service point of the requesting BS i.

4. Service stage:
- (a) BS i pays $\gamma_i\theta_{ij^*}d_i$ and offloads UEs within $\mathcal{A}_i^c$ to UAV j^*.
- (b) UAV j^* provides the downlink service with a transmit power $p_{j^*} = \frac{\gamma_i}{2}\theta_{ij^*}^2$ for a service time $T - t_{ij^*}$.

End

B. The optimal UAV association under the feasible contract

Given the feasible contract set $\{(\gamma_i\theta_{ij}, \gamma_i\theta_{ij}^2/2)|\forall\theta_{ij}\}$, the utility $R_{ij}(\theta_{ij})$ of each candidate UAV $j \in \mathcal{J}$ and the utility $U_{ij}(\theta_{ij})$ of the requesting BS i can be jointly determined. Then, the optimization problem in (12) becomes

$$\max_{j\in\mathcal{J}} \quad U_{ij}(\theta_{ij}), \tag{19a}$$

$$\text{s. t.} \quad p_{ij}(\boldsymbol{x}_{ij}^*, \rho_i^c) \leq p_j(\theta_{ij}) \leq \min\{p_{ij}^{\max}, p_{\max}\}, \tag{19b}$$

$$t_{ij} \leq \kappa_i T. \tag{19c}$$

Therefore, BS i aims to find a UAV of the optimal type θ_{ij}^* that maximizes its utility in (19a), while satisfying (19b) and (19c). In the association stage, after BS i sends the request signal, each UAV j will respond with its type θ_{ij}. Based on the derivation $\frac{\mathrm{d}U_{ij}(\theta_{ij})}{\mathrm{d}\theta_{ij}} < 0$, the optimal UAV is $j^* = \arg\max_{j\in\mathcal{J}_i} U(\theta_{ij}) = \arg\min_{j\in\mathcal{J}_i}\theta_{ij}$, where $\mathcal{J}_i = \{j|p_{ij}(\boldsymbol{x}_{ij}^*, \rho_i^c) \leq \frac{\gamma_i}{2}\theta_{ij}^2 \leq \min\{p_{ij}^{\max}, p_{\max}\}, t_{ij} \leq \kappa_i T\}$. Thus, the qualified UAV with a smallest type is the optimal solution. The complete process of the predictive UAV deployment is summarized in Algorithm 1.

Compared with conventional UAV deployment, the contract-based optimization has three advantages. First, the proposed method not only reveals that the closest UAV is the optimal solution, but it also optimally determines the amount of the payment that the BS should offer the UAV, such that the utility of the BS can be maximized and the utility of the UAV is non-negative. Second, based on IC constraint, each UAV will receive the highest utility by accepting the contract designed for its real type. Thus, the use of contract theory captures the economic incentive of each UAV, forcing it to truthfully tell the requesting BS with its actual type, which is unknown to the BS a priori. Therefore, the proposed contract approach guarantees a truthful information exchange between the BS and UAV operators, which the traditional optimization method cannot achieve. In the end, the proposed algorithm is more efficient for practical implementation, due to less information exchange in the association stage. In contrast, conventional optimization techniques will require all necessary information from all UAVs for solving the centralized association problem. Hence, compared with traditional optimization methods, our proposed algorithm reduces the communications overhead, exhibits a lower communication overhead, and ensure a truthful information exchange between the BS and UAV operators.

VI. Simulation results and Analysis

A. Simulation parameters

For our simulations, we consider a UAV-assisted wireless network in a dense urban environment, operating at the 2 GHz frequency with a downlink bandwidth of 20 MHz. The parameters in the LOS probability model are $a = 9.6$ and $b = 0.28$ [28]. The Gaussian parameters of the additional air-to-ground path loss are $\mu_{\text{LOS}} = 1.6$ and $\sigma_{\text{LOS}} = 8.41$ for the LOS link while $\mu_{\text{NLOS}} = 23$ and $\sigma_{\text{NLOS}} = 33.78$ for the NLOS case [27]. For the UAV parameters, based on the specifications in [34], we set the mobility power $m = 20$ W with an average moving speed of 5 m/s, and the hovering power is $p_h = 16$ W. The maximal on-board energy of each UAV is 25 Wh, and the battery recharge takes 10 minutes. The maximum downlink transmit power is $p_{\max} = 20$ W, and the unit cost for on-board energy is $\alpha = 1.2$. The noise power spectral density at UE is -174 dBm/Hz, and the data service per bit is $\beta = 10^{-7}$. For the UAV deployment process, we set $\Delta T = 1$ second, the learning duration $\tau = 2$ minutes, and the service time $T = 18$ minutes. The ratio of effective transmission in each time slot is $\eta = 90\%$, and the maximum ratio of the UAV's movement over the time interval is $\kappa_i = 0.1$.

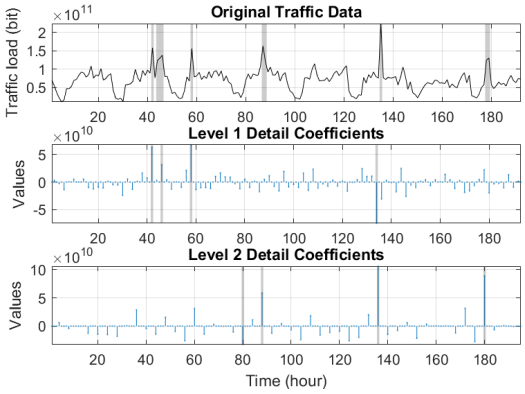
B. Dataset description and preprocessing

An open-source dataset "city-cellular-traffic-map" in [24] is used for the modeling, training, and testing of the proposed UAV deployment framework. The dataset collects HTTP traffic data through the cellular networks during each hour within a middle-sized city of China from August 19 to August 26, 2012. The dataset consist of two parts. One lists the identification number (ID) and the location in longitude and latitude of each BS, and the other collects the number of UEs, packets and traffic data that each BS transmits to downlink UEs during each hour. In order to identify hotspot events in the dataset, we apply the discrete wavelet transform (DWT) to the hourly cellular traffic in the city level. As shown in the upper figure of Fig. 3, the cellular traffic within the city area presents a conspicuously periodic pattern, with several sudden and erratic surges. DWT processes the time-serial data by analyzing both the value and frequency components, where the lower-frequency component defines the long term trend, and the higher-frequency component represents the small-scale rapid variation. A hotspot event usually causes a steep surge in the traffic amount. Therefore, such rapid change can be captured by DWT in the higher frequency domain. As shown in Fig. 3a, a two-level DWT is applied to detect the frequency change of cellular traffic, and the gray bars mark the time points when the traffic amount has a sudden increase. Based on the result, we separate the dataset into the normal traffic data and the potential congested traffic, as given in Fig. 3b. A time window from 42 to 47, which is 18 to 23 p.m. on August 20, shows a continuously high cellular traffic amount, and the hotspot event is highly likely to happen during this period. Therefore, the traffic data from 42 to 47 are used for the predictive UAV deployment in the following analysis.

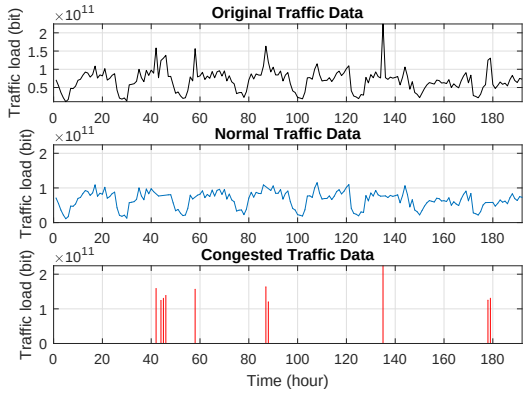
However, the data in [24] does not include the location information of each UE, or the service area of each BS. To identify the UE distribution and the traffic density, the location and time labels are generated and attached to each transmission record via the following approach. First, the service area $\mathcal{A}_i$ of each BS i is partitioned, based on the closest-distance principle. Next, we use the total packet number to denote the number of downlink transmissions. Furthermore, we note that the original time label t in [24] is based on one hour, which is too coarse to enable our analysis. To extract the estimated data with a desired duration, a new label with a finer time grain of one second is randomly generated and attached to each traffic record. Then, given $\tau = 2$ and $T = 18$, we divide each hour evenly into three intervals, such that the cellular data during first two minutes of each interval is used to model the UE distribution and downlink traffic, and data from the following 18 minutes is used to estimate the UAV's transmission performance. Eventually, the location label $\boldsymbol{y}_n$ of each traffic record is generated by a GMM with random parameters to which we add a zero-mean Gaussian noise with a standard deviation of three meters. With additional location and time labels, the dataset is suitable for the studied problem.

C. Performance of the cellular traffic prediction

Fig. 4 shows that over 70% UEs receive, on average, one packet within every two minutes. Thus, the transmission record $\mathcal{S}_i$ that is collected during the learning stage ($\tau = 2$ minutes) is a representative training datase. In this simulation, the proposed WEM approach is applied to predict the data demand d_i, while the actual traffic demand d_i^{actual} is calculated by summing up the real transmission amount within $\mathcal{A}_i^c$. Here, the mean relative error (MRE) is the metric to evaluate the prediction performance, where $\delta_{\text{MRE}} = \mathbb{E}_{i,t}[\frac{|d_i - d_i^{\text{actual}}|}{d_i^{\text{actual}}}]$. Meanwhile, we introduce the EM and k-mean methods as baselines. First, the EM method has been used in Section IV-A for modeling the



(a) Two-level DWT components.



(b) Normal traffic states and potential congestion events.

Fig. 3: Two-level DWT is applied to detect the cellular traffic congestion from a city level.

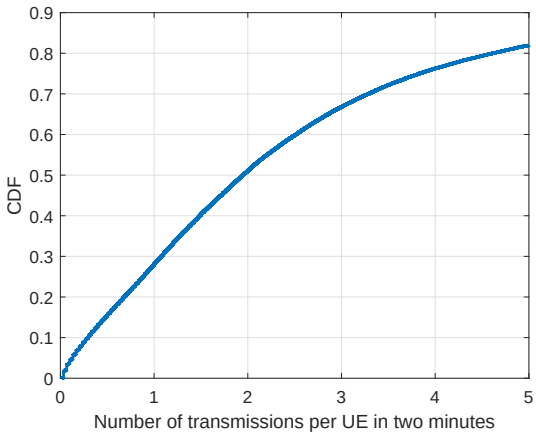


Fig. 4: Number of transmissions per UE per two minutes.

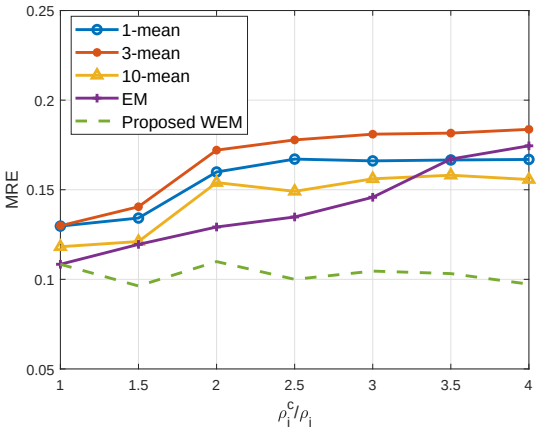


Fig. 5: MRE of the WEM approach and two baselines.

UE distribution $f_i(\boldsymbol{y})$. Here, to predict the traffic demand using the EM method, we have $d_i^{\text{EM}} = T \cdot \mathbb{E}_n(s_n) \cdot \int_{\boldsymbol{y} \in \mathcal{A}_i^c} f_i(\boldsymbol{y})\, \mathrm{d}\boldsymbol{y}$, where $\mathbb{E}_n(s_n) = \frac{\sum_n s_n \Delta t}{\tau}$ is the time-average data rate of all UEs, and $\int_{\boldsymbol{y} \in \mathcal{A}_i^c} f_i(\boldsymbol{y})\, \mathrm{d}\boldsymbol{y}$ is the percentage of UEs within the hotspot area. Note that, this is a commonly-used approach to estimate cellular data demand using the UE distribution and the average rate requirement per UE in the cellular network [6]. The k-mean method predicts the traffic density by averaging the local traffic density from k closest neighbors.

Fig. 5 shows the prediction MRE of the WEM, EM, and k-mean methods, where $k = 1$, 3 and 10, as the average data demand ρ_i^c of the hotspot UEs increases. Note that, $\rho_i^c = \frac{1}{Q_i^c} \int_{\boldsymbol{y} \in \mathcal{A}_i^c} S_i(\boldsymbol{y})\, \mathrm{d}\boldsymbol{y}$ is the average data rate per UE within the hotspot area, and $\rho_i = \frac{1}{Q_i} \int_{\boldsymbol{y} \in \mathcal{A}_i} S_i(\boldsymbol{y})\, \mathrm{d}\boldsymbol{y}$ is the average rate demand of all UEs within the cellular network. When $\frac{\rho_i^c}{\rho_i} = 1$, each hotspot UE will have the same data demand as the other UEs. In this case, the WEM and EM approaches yield a similar prediction accuracy with an MRE of 11%, and the prediction errors of k-mean methods are between 12% and 12.5%. Note that, a prediction error of 11% yields lower than 0.1 W of deviation on the value of $p_{ij}(\boldsymbol{x}_{ij}^*, \rho_i)$. Clearly, this is a very small value compared to the hovering and transmit powers of a typical UAV. When the traffic load within different regions of the cellular network becomes more uneven, the prediction error of WEM remains the same, while the errors of the EM and k-mean methods gradually increase above 15.5%. Clearly, for $\frac{\rho_i^c}{\rho_i} > 1$, the proposed WEM approach outperforms all other baselines.

In the WEM approach, the traffic density $\bar{S}_i(\boldsymbol{y})$ of each location $\boldsymbol{y}$ is considered when optimizing the prediction parameters. Therefore, the spatial feature of downlink transmissions can be accurately captured, and the performance of WEM does not decrease when the traffic load in the cellular system becomes uneven. However, the EM model only considers the location information, but ignores the downlink rate of each transmission. Therefore, when the traffic demand shows distinct patterns in different regions, the EM method fails to capture the spatial diversity, and its prediction error increases significantly. Given that the k-mean method predicts the cellular traffic by averaging data from k closest neighbors, it captures the traffic spatial difference from local information. However, as the cellular traffic becomes more uneven, the local information is more sensitive to noise, and thus, the prediction errors of k-mean methods increase, as ρ_i^c increases. By comparing different k-mean algorithms, we find that 3-mean achieves the worst performance, because information from three neighbors is not sufficient to cancel out the noise.

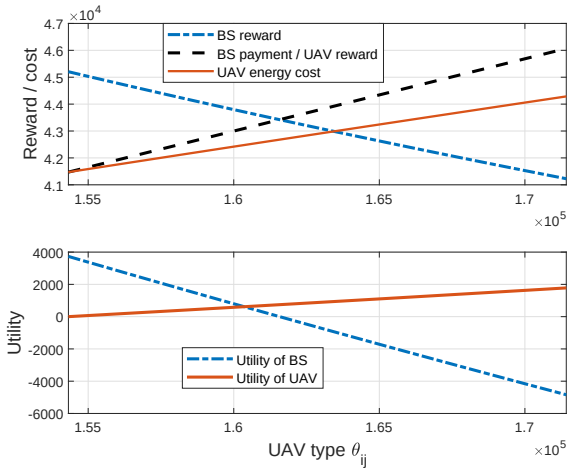


Fig. 6: Costs, rewards and overall utilities of the associated BS and UAV, given different UAV types.

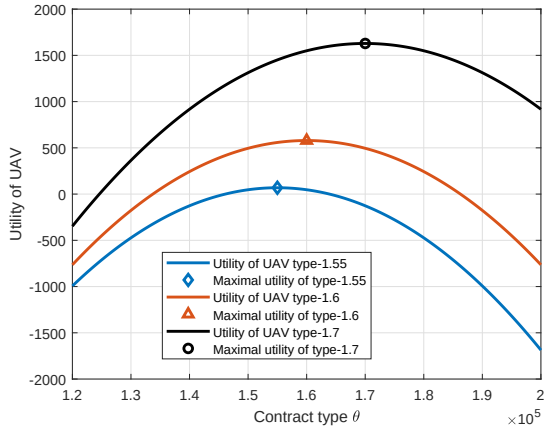


Fig. 7: Utilities of UAVs, given different contract types.

The simulation results also show that 10-mean yields the best performance among all k-mean methods.

D. The Impact of the UAV type on the utilities

In this section, we investigate the impact of the UAV type on the utilities. The contract is designed based on Lemma 1 by a BS with ID 7939, using the data from time 42 in [24]. Fig. 6 shows the relationship between the UAV type and the reward, cost, as well as the overall utilities of the requesting BS and the deployed UAV, respectively. First, as the UAV type θ_{ij} becomes larger, the BS's reward $\beta B_{ij}(p_j)$ from the downlink UEs will decrease. Although the transmit power $p_j(\theta_{ij})(\theta_{ij})$ becomes higher given a larger θ_{ij}, a UAV with a higher type must travel for a longer time t_{ij} before its service. Thus, the downlink data transmission of the UAV becomes lower for a larger θ_{ij}. Meanwhile, based on (8), we have $\frac{\mathrm{d}B_{ij}(\theta_{ij})}{\mathrm{d}\theta_{ij}} < 0$. Therefore, a higher UAV type θ_{ij} leads to a lower BS's reward. Furthermore, a higher UAV type increases the payment $u_i(\theta_{ij})d_i$ from BS i to UAV j, and, thus, the utility of BS i will be lower. For the deployed UAV j, a larger θ_{ij} results in a higher reward $u_i(\theta_{ij})d_i$ from BS i, and the increase of the UAV's reward is faster than the energy cost. Therefore, as shown in Fig. 6, the utility of the deployed UAV will increase, as its type θ_{ij} becomes larger. Meanwhile, Fig. 6 shows that the UAV's utility is always non-negative. Therefore, the IR condition holds in the designed contract.

Fig. 7 investigates the impact of the contract type on the UAV's utility. The utilities of three UAVs, where their actual types are 1.55×10^5 (type-1.55), 1.6×10^5 (type-1.6), and 1.7×10^5 (type-1.7), are given, when they accept different kinds of contracts from BS 7939. As shown in Fig. 7, the maximum utility of each UAV is achieved when the accepted contract is of its own type. Thus, simulation results show that the IC condition holds in the designed contract set.

An interesting observation on the utility function is that the prediction error of d_i does not cause small fluctuations on the utility value of the BS or the employed UAV. Given the transmit power as $p_j = \gamma_i \theta_{ij}^2/2 = \frac{mT^2}{4\alpha(T-t_{ij})^2}$ and the total payment from BS i to UAV j as $u_i d_i = \gamma_i \theta_{ij} d_i = \frac{mT^2}{2(T-t_{ij})}$, d_i no longer appears in the utility formulas, and, thus, an inaccuracy in d_i will not impact the utility functions in (9) and (10). The main effect of d_i in the predictive UAV deployment is to determine the minimum required transmit power $p_{ij}(\boldsymbol{x}_{ij}^*, \rho_i^c)$. If the predicted demand d_i is much lower than the real data demand, then $p_{ij}(\boldsymbol{x}_{ij}^*, \rho_i^c)$ will be smaller. In consequence, some UAVs without enough energy may be inappropriately considered to be a qualified choice, and might be employed. On the other hand, if d_i is much higher than the actual demand, some qualified UAVs with enough power may be excluded from the candidate set $\mathcal{J}_i$. Both cases can lead to a suboptimal solution to (12). However, as long as the error on d_i causes no change to the association result, the utilities of the BS and UAV will always be accurate. Based on this observation, the proposed approach is highly robust to prediction errors.

E. Evaluation of the predictive UAV deployment

In this section, we evaluate the performance of the proposed UAV deployment method with four metrics, which are the downlink capacity, energy consumption, service delay of the employed UAVs, and the utilities of the BS and UAV operators. Meanwhile, for comparison purposes, an event-driven deployment of the closest UAV and an event-driven deployment of a UAV with the maximal on-board energy are introduced as two baselines. In both baseline approaches, the target UAV is requested by the overloaded BS and deployed, after the downlink congestion occurs, without the prediction on traffic demand. The optimal location of the deployed UAV is determined after the UAV arrives at the service area, so as to maximize its downlink transmission rate [11]. Meanwhile, in both baseline approaches, there is no contract design to determine the cost and payment between the BS to its employed UAV. Instead, the employed UAV j provides the downlink service to the best of its power ability, where $p_j = \min\{p_{ij}(\boldsymbol{x}_{ij}^*, \rho_i), p_{ij}^{\max}, p_{\max}\}$, and the unit payment u_i from BS i to the employed UAV is a fixed price β, which equals to the unit payment from the UEs to the BS per bit of data service.

First, we compare the performance of the proposed predictive UAV deployment with two baselines, in terms of the total

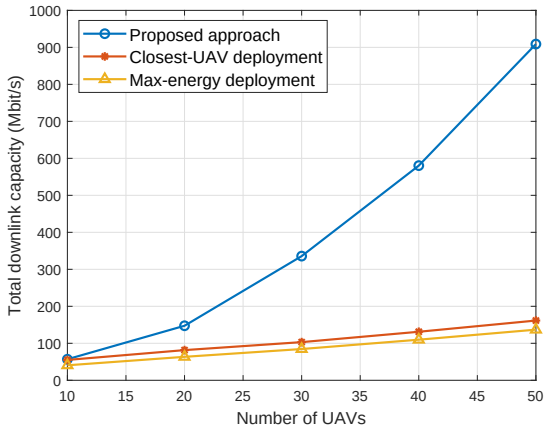


Fig. 8: Total downlink capacity of the UAV downlink service for the proposed predicted UAV deployment and two baselines.

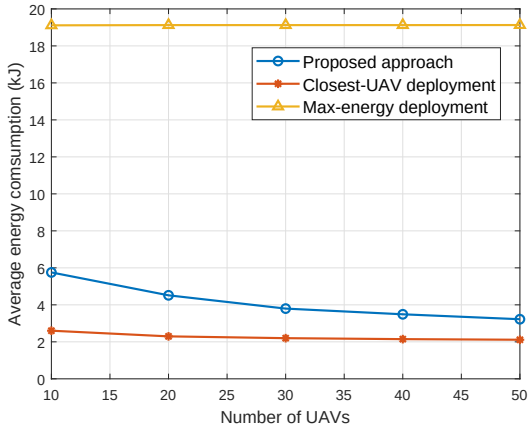


Fig. 9: Average energy consumption per UAV for the proposed predicted UAV deployment and two baselines.

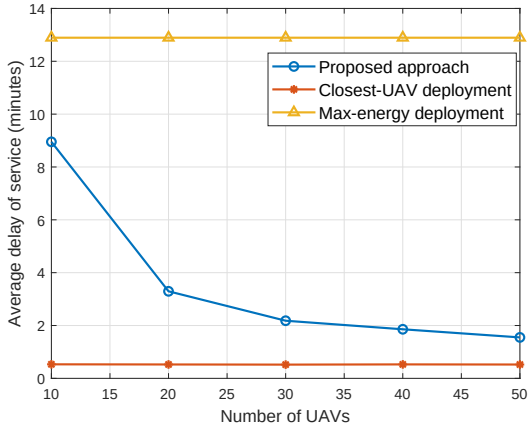


Fig. 10: Average service delay per UAV for the proposed predicted UAV deployment and two baselines.

downlink capacity of the employed UAVs. As shown in Fig. 8, as the number of UAVs within the cellular network increases, the total downlink capacity that the employed UAVs provide to the downlink UEs increases in all three schemes. For a larger number of UAVs, the average movement distance between each overloaded BS and its employed UAV will decrease. Therefore, less energy is consumed during the mobility stage, and more power can be reserved for the downlink transmission service. In consequence, the downlink capacity of all three methods increases. However, in the closest-UAV approach, without the data demand prediction, the deployed closest UAV may not have enough on-board energy to satisfy the downlink data demand. Therefore, the downlink capacity in the closest-UAV baseline is lower than the proposed approach. In the max-energy deployment, the distance between the employed UAV and the service area is usually larger, compared to two other methods. Although the employed UAV has the largest amount of available onboard energy, due to a longer travel distance, most of the onboard energy will be consumed on mobility, and the transmit power may be insufficient. Thus, the max-energy deployment yields the lowest capacity performance among all three schemes. Moreover, the proposed approach improves the downlink capacity by over four-fold and five-fold, compared to the closest-UAV and the max-energy baselines, respectively.

Fig. 9 and Fig. 10 show the average energy consumption and service delay of each employed UAV, respectively, as the number of available UAVs in the network increases. First, we can see that the closest-UAV scheme yields the least energy cost and service delay, due to its shortest movement distance. In the proposed approach, the energy consumption and movement duration are relatively higher, because the selection criteria balances between the distance of the UAV (which determines the movement energy) and the availability of sufficient on-board energy to meet the predicted data demand. Meanwhile, the max-energy deployment results in the highest energy and time cost, due to the largest travel distance during the mobility stage. Next, for a higher number of UAVs, the energy consumption and service delay of the proposed method both drop, while the performance of the baselines remains nearly constant. In particular, as the number of UAVs increases, the performance of the proposed approach improves exponentially, and the gap between the proposed approach and the closest-UAV scheme becomes much smaller. In the proposed method, having more UAVs reduces the average distance between any employed UAV and its service point, and, hence, decreases the energy and time cost. However, in the two baselines, the number of available UAVs does not effect the travel distance during mobility. Thus, the energy consumption and service delay of two baselines remain nearly constant with the increase in the number of UAVs.

In Fig. 11 and Fig. 12, we compared the utilities of the BS and UAV operators in three schemes, respectively. First, in Fig. 11, for a larger number of UAVs, the average utility per BS increases in all three schemes, and the proposed approach yields the highest utility. In the proposed method, by having more UAVs, the average distance between an employed UAV to its service becomes smaller, and, thus, the type θ_{ij} of the employed UAV j with respect to BS i decreases, which yields a higher utility of BS i. For the closest-UAV and max-energy schemes, since the employed UAV cannot always satisfy the data demand of its downlink UEs, the utilities of each BS for both baselines are lower, compared the proposed method.

In Fig. 12, we can see that, as the number of UAVs increases, the total utility of the employed UAVs becomes higher in the proposed approach, while the UAVs' utilities

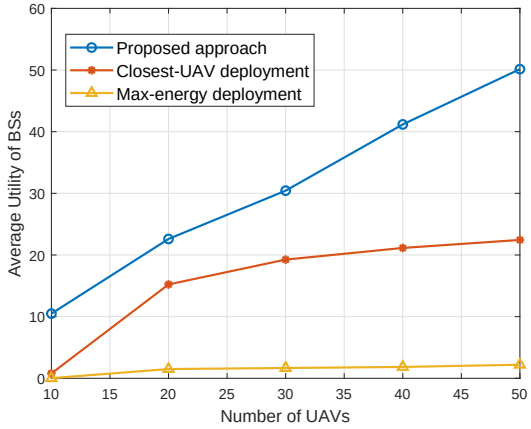


Fig. 11: Average utility of each BS for the proposed predictive UAV deployment and two baselines.

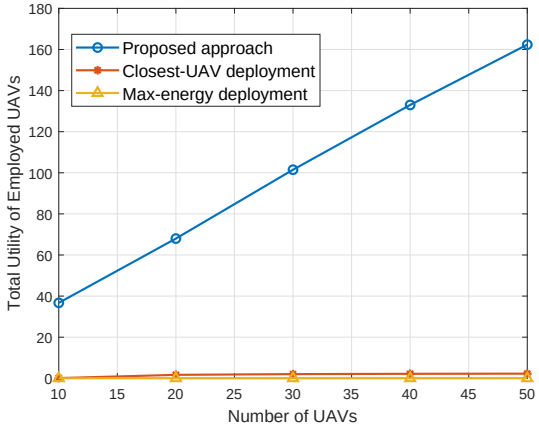


Fig. 12: Total utility of the UAV operators for the proposed predictive UAV deployment and two baselines.

resulting from both baseline schemes are much lower than the proposed method. As shown in Fig. 8 and Fig. 9, by having more UAVs, the average energy cost per UAV resulting from the proposed approach will decrease, while the downlink transmission capacity of the UAV networks increases, which yields a higher income. As a result, the overall utility of the UAV operator in the proposed method will become higher for a larger number of UAVs. For the closest-UAV scheme, its lower energy consumption and shorter service delay yield a smaller deployment cost, compared with the proposed method. However, the lower downlink capacity results in less payment from the BS. Thus, the total utility of the UAV operator in the closest-UAV scheme is less than the proposed method. Moreover, based on Figs. 8, 9, and 10, we can see that the max-energy scheme yields the lowest transmission rate, the highest energy cost, and the longest service delay. Therefore, the utility of the UAV operators in the max-energy scheme is the lowest among all three methods. In consequence, based on Fig. 11 and Fig. 12, we can conclude that the proposed method enables an efficient UAV deployment to alleviate communication congestion in the cellular networks, and shows a significant advantage on the economical revenues of both the BS and UAV operators, compared with two baseline, event-driven approaches.

VII. Conclusion

In this paper, we have proposed a novel approach for predictive deployment of UAVs to complement the ground cellular system in face of the hotspot events. In particular, four interrelated and sequential stages have been proposed to enable the ground BS to optimally employ a UAV to offload the excess traffic. First, a novel framework, based on the EM and WEM methods, has been proposed to estimate the UE distribution and the downlink traffic demand. Next, to guarantee a truthful information exchange between the BS and UAV operators, a traffic offload contract have been developed, and the sufficient and necessary conditions for having a feasible contract have been analytically derived. Then, an optimization problem have been formulated to deploy the optimal UAV onto the hotspot area in a way that the utility of each overloaded ground BS is maximized. Simulation results show that the proposed WEM approach yields a prediction error around 10%, and compared with the EM and k-mean schemes, the WEM algorithm yields a higher prediction accuracy, particularly when the traffic load in the cellular system becomes spatially uneven. Furthermore, compared with two event-driven schemes based on the closest-distance and maximal-energy metrics, the proposed predictive deployment approach enables UAV operators to provide efficient downlink service for hotspot users, and significantly improves the revenues of both the BS and UAV networks.

Appendix A
Proof of Proposition 1

We first use contradiction to prove the proposition that if $\theta_{ij} > \theta_{ij}^{'}$, then $u_i(\theta_{ij}) \geq u_i(\theta_{ij}^{'})$. Suppose that there exists $u_i(\theta_{ij}) < u_i(\theta_{ij}^{'})$, but $\theta_{ij} > \theta_{ij}^{'}$. Then, we have

$$\theta_{ij}u_i(\theta_{ij}^{'}) + \theta_{ij}^{'}u_i(\theta_{ij}) > \theta_{ij}u_i(\theta_{ij}) + \theta_{ij}^{'}u_i(\theta_{ij}^{'}). \quad (20)$$

On the other hand, from IC condition, we have

$$\begin{aligned} \theta_{ij}u_i(\theta_{ij}) - p_j(\theta_{ij}) &\geq \theta_{ij}u_i(\theta_{ij}^{'}) - p_j(\theta_{ij}^{'}), \\ \theta_{ij}^{'}u_i(\theta_{ij}^{'}) - p_j(\theta_{ij}^{'}) &\geq \theta_{ij}^{'}u_i(\theta_{ij}) - p_j(\theta_{ij}). \end{aligned} \quad (21)$$

By adding the inequations in (21), we have $\theta_{ij}u_i(\theta_{ij}) + \theta_{ij}^{'}u_i(\theta_{ij}^{'}) \geq \theta_{ij}u_i(\theta_{ij}^{'}) + \theta_{ij}^{'}u_i(\theta_{ij})$, which contradicts to (20). This completes the first part of the proof.

Next, we prove that if $u_i(\theta_{ij}) \geq u_i(\theta_{ij}^{'})$, $p_j(\theta_{ij}) \geq p_j(\theta_{ij}^{'})$. From the IC condition, we have $\theta_{ij}^{'}u_i(\theta_{ij}^{'}) - p_j(\theta_{ij}^{'}) \geq \theta_{ij}^{'}u_i(\theta_{ij}) - p_j(\theta_{ij})$, i.e. $p_j(\theta_{ij}) - p_j(\theta_{ij}^{'}) \geq \theta_{ij}^{'}\left(u_i(\theta_{ij}) - u_i(\theta_{ij}^{'})\right)$. Since $u_i(\theta_{ij}) > u_i(\theta_{ij}^{'})$, we conclude $p_j(\theta_{ij}) - p_j(\theta_{ij}^{'}) \geq \theta_{ij}^{'}\left(u_i(\theta_{ij}) - u_i(\theta_{ij}^{'})\right) \geq 0$, and thus $p_j(\theta_{ij}) \geq p_j(\theta_{ij}^{'})$. This completes the proof.

Appendix B
Proof of Theorem 1

For notation simplicity, in this section, we denote u_i, p_j, θ_{ij}, M_{ij} as u, P, θ, M respectively.

A. Proof for necessary conditions

Given the IR and IC conditions, we prove Theorem 1 in this section. First, as shown in Proposition 1, for any $\theta, \theta^{'} \in \Theta_i$, once $\theta > \theta^{'}$, then $u(\theta) \geq u(\theta^{'})$ and $P(\theta) \geq P(\theta^{'})$. Therefore, condition (a) of Theorem 1 is proved by Proposition 1. Second, condition (b) of Theorem 1 is supported by the IR condition, where $R_{ij}(\theta) \geq 0$ for all θ in Θ_i, which naturally includes $\theta^{\min}$. Next, we prove condition (c). Let $\Delta\theta = \theta^{'} - \theta$. According to the IC condition, for any $\Delta\theta \in [\theta^{\min} - \theta^{\max}, 0) \cup (0, \theta^{\max} - \theta^{\min}]$, we have: $\theta \cdot u(\theta) - P(\theta) \geq \theta \cdot u(\theta + \Delta\theta) - P(\theta + \Delta\theta)$, i.e., $\theta \cdot [u(\theta) - u(\theta + \Delta\theta)] \geq P(\theta) - P(\theta + \Delta\theta)$. If $\Delta\theta > 0$, then according to Proposition 1, $u(\theta + \Delta\theta) \geq u(\theta)$ and $P(\theta + \Delta\theta) \geq P(\theta)$. Here, we exclude the situation where $u(\theta + \Delta\theta) = u(\theta)$ and $P(\theta + \Delta\theta) = P(\theta)$ in the following discussion of this proof, because condition (c) naturally holds in this case. Therefore, for any $\Delta\theta \in (0, \theta^{\max} - \theta^{\min}]$, we have

$$\theta \leq \frac{P(\theta + \Delta\theta) - P(\theta)}{u(\theta + \Delta\theta) - u(\theta)}. \tag{22}$$

If $\Delta\theta < 0$, then $u(\theta + \Delta\theta) < u(\theta)$ and $P(\theta + \Delta\theta) < P(\theta)$. Thus, for any $\Delta\theta \in [\theta^{\min} - \theta^{\max}, 0)$,

$$\theta \geq \frac{P(\theta + \Delta\theta) - P(\theta)}{u(\theta + \Delta\theta) - u(\theta)}. \tag{23}$$

By combing (22) and (23), we have $\frac{\mathrm{d}P}{\mathrm{d}\theta} / \frac{\mathrm{d}u}{\mathrm{d}\theta} = \lim_{\Delta\theta \to 0} \frac{P(\theta+\Delta\theta)-P(\theta)}{u(\theta+\Delta\theta)-u(\theta)} = \theta$, which proves condition (c) of Theorem 1.

B. Proof for sufficient conditions

From Theorem 1, we will prove the IR and IC conditions in this section. First, we prove the IR condition. According to condition (b) of Theorem 1, $\theta^{\min}$ satisfies the IR condition. Then, we prove that for any $\theta \in (\theta^{\min}, \theta^{\max}]$, the IR condition holds. From condition (c) of Theorem 1, we have the following inequalities, $\frac{P(\theta)-P(\theta^{\min})}{u(\theta)-u(\theta^{\min})} \leq \theta$, i.e.,

$$P(\theta^{\min}) \geq P(\theta) - \theta \cdot [u(\theta) - u(\theta^{\min})]. \tag{24}$$

From condition (b), we have

$$\theta^{\min} \cdot u(\theta^{\min}) - P(\theta^{\min}) - M \geq 0. \tag{25}$$

By combing (24) and (25), we have $\theta \cdot u(\theta) - P(\theta) - M \geq (\theta - \theta^{\min}) \cdot u(\theta^{\min}) \geq 0$. Thus, for any $\theta \in \Theta_i$, the IR condition holds.

In the end, we prove the IC condition. Let $h = \theta \cdot u(\theta) - P(\theta) - M - [\theta \cdot u(\theta^{'}) - P(\theta^{'}) - M]$. And we prove that $h \geq 0$. From condition (c), we have, if $\theta^{'} > \theta$, then $\frac{P(\theta^{'})-P(\theta)}{u(\theta^{'})-u(\theta)} \geq \min\{\theta, \theta^{'}\} = \theta$. i.e., $P(\theta^{'}) - P(\theta) \geq \theta \cdot [u(\theta^{'}) - u(\theta)]$. Therefore, $h = \theta \cdot [u(\theta) - u(\theta^{'})] + P(\theta^{'}) - P(\theta) \geq 0$. On the other hand, if $\theta^{'} < \theta$, then $\frac{P(\theta)-P(\theta^{'})}{u(\theta)-u(\theta^{'})} \leq \max\{\theta, \theta^{'}\} = \theta$. i.e., $P(\theta) - P(\theta^{'}) \leq \theta \cdot [u(\theta) - u(\theta^{'})]$. Therefore, $h \geq 0$. Consequently, the IC condition holds.

REFERENCES

[1] Q. Zhang, M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Machine learning for predictive on-demand deployment of UAVs for wireless communications," in *Proc. of IEEE Global Communications Conference*, Abu Dhabi, UAE, Dec 2018.

[2] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Mobile unmanned aerial vehicles (UAVs) for energy-efficient Internet of Tshings communications," *IEEE Transactions on Wireless Communications*, vol. 16, no. 11, pp. 7574–7589, Sep 2017.

[3] R. I. Bor-Yaliniz, A. El-Keyi, and H. Yanikomeroglu, "Efficient 3-D placement of an aerial base station in next generation cellular networks," in *Proc. of IEEE International Conference on Communications*, Kuala Lumpur, Malaysia, May 2016.

[4] X. Zhang and L. Duan, "Fast deployment of UAV networks for optimal wireless coverage," *IEEE Transactions on Mobile Computing*, vol. 18, no. 3, pp. 588–601, May 2018.

[5] W. Khawaja, I. Guvenc, D. Matolak, U.-C. Fiebig, and N. Schneckenberger, "A survey of air-to-ground propagation channel modeling for unmanned aerial vehicles," vol. 21, no. 3, pp. 2361–2391, May 2019.

[6] M. Mozaffari, W. Saad, M. Bennis, Y.-H. Nam, and M. Debbah, "A tutorial on UAVs for wireless networks: Applications, challenges, and open problems," *IEEE Communications Surveys and Tutorials*, vol. 21, no. 4, pp. 3039–3071, Fourth quarter 2019.

[7] M. Mozaffari, A. T. Z. Kasgari, W. Saad, M. Bennis, and M. Debbah, "Beyond 5G with UAVs: Foundations of a 3D wireless cellular network," *IEEE Transactions on Wireless Communications*, vol. 18, no. 1, pp. 357–372, Jan 2018.

[8] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE network*, vol. 34, no. 3, pp. 134–142, 2019.

[9] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, "Artificial neural networks-based machine learning for wireless networks: A tutorial," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3039–3071, Jul 2019.

[10] Z. Hu, Z. Zheng, L. Song, T. Wang, and X. Li, "UAV offloading: Spectrum trading contract design for UAV assisted cellular networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 9, pp. 6093–6107, July 2018.

[11] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Optimal transport theory for power-efficient deployment of unmanned aerial vehicles," in *Proc. of IEEE International Conference on Communications*, Kuala Lumpur, Malaysia, May 2016.

[12] E. Kalantari, H. Yanikomeroglu, and A. Yongacoglu, "On the number and 3D placement of drone base stations in wireless cellular networks," in *Proc. of IEEE 84th Vehicular Technology Conference*, Montreal, QC, Canada, Sep 2016.

[13] J. Lyu, Y. Zeng, and R. Zhang, "UAV-aided offloading for cellular hotspot," *IEEE Transactions on Wireless Communications*, vol. 17, no. 6, pp. 3988–4001, Mar 2018.

[14] V. Sharma, M. Bennis, and R. Kumar, "UAV-assisted heterogeneous networks for capacity enhancement," *IEEE Communications Letters*, vol. 20, no. 6, pp. 1207–1210, Apr 2016.

[15] J. Lyu, Y. Zeng, and R. Zhang, "Spectrum sharing and cyclical multiple access in UAV-aided cellular offloading," in *Proc. of IEEE Global Communications Conference*, Singapore, Dec 2017.

[16] F. Cheng, S. Zhang, Z. Li, Y. Chen, N. Zhao, R. Yu, and V. C. Leung, "UAV trajectory optimization for data offloading at the edge of multiple cells," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 7, pp. 6732 – 6736, Mar 2018.

[17] S. Sharafeddine and R. Islambouli, "On-demand deployment of multiple aerial base stations for traffic offloading and network recovery," *Computer Networks*, vol. 156, pp. 52–61, June 2019.

[18] R. Li, Z. Zhao, J. Zheng, C. Mei, Y. Cai, and H. Zhang, "The learning and prediction of application-level traffic data in cellular networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 6, pp. 3899–3912, Mar 2017.

[19] C. Yu, Y. Liu, D. Yao, L. T. Yang, H. Jin, H. Chen, and Q. Ding, "Modeling user activity patterns for next-place prediction," *IEEE Systems Journal*, vol. 11, no. 2, pp. 1060–1071, July 2017.

[20] P. Valente Klaine, M. A. Imran, O. Onireti, and R. D. Souza, "A survey of machine learning techniques applied to self organizing cellular networks," *IEEE Communications Surveys and Tutorials*, vol. 19, no. 4, pp. 2392–2431, July 2017.

[21] M. Chen, W. Saad, and C. Yin, "Liquid state machine learning for resource and cache management in LTE-U unmanned aerial vehicle

(UAV) networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 3, pp. 1504 –1517, Jan 2019.

[22] J. Chen, U. Yatnalli, and D. Gesbert, "Learning radio maps for UAV-aided wireless networks: A segmented regression approach," in *Proc. of IEEE International Conference on Communications*, Paris, France, May 2017.

[23] R. Amorim, J. Wigard, H. Nguyen, I. Z. Kovacs, and P. Mogensen, "Machine-learning identification of airborne UAV-UEs based on LTE radio measurements," in *Proc. of IEEE Globecom Workshops*, Singapore, Jan 2017.

[24] "City cellular traffic map," https://github.com/caesar0301/city-cellular-traffic-map, accessed: 2016-10-05.

[25] P. Bolton and M. Dewatripont, *Contract theory*. MIT press, 2005.

[26] L. Zhu, J. Zhang, Z. Xiao, X. Cao, D. O. Wu, and X.-G. Xia, "3D beamforming for flexible coverage in millimeter-wave uav communications," *IEEE Wireless Communications Letters*, vol. 8, no. 3, pp. 837–840, Jan 2019.

[27] A. Al-Hourani, S. Kandeepan, and A. Jamalipour, "Modeling air-to-ground path loss for low altitude platforms in urban environments," in *Proc. of IEEE Global Communications Conference*, Austin, TX, USA, Dec 2014.

[28] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal LAP altitude for maximum coverage," *IEEE Wireless Communications Letters*, vol. 3, no. 6, pp. 569–572, July 2014.

[29] J. Lyu, Y. Zeng, and R. Zhang, "Cyclical multiple access in UAV-aided communications: A throughput-delay tradeoff," *IEEE Wireless Communications Letters*, vol. 5, no. 6, pp. 600–603, Aug 2016.

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

[31] A. T. Z. Kasgari, W. Saad, and M. Debbah, "Human-in-the-loop wireless communications: Machine learning and brain-aware resource management," *IEEE Transactions on Communications*, vol. 67, no. 11, pp. 7727–7743, Jul 2019.

[32] B. Selim, O. Alhussein, S. Muhaidat, G. K. Karagiannidis, and J. Liang, "Modeling and analysis of wireless channels via the mixture of Gaussian distribution," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 10, pp. 8309–8321, 2015.

[33] M. B. Christopher, *Pattern recognition and machine learning*. Springer-Verlag New York, 2016.

[34] "DJI matrice 200 series v2 specifications," https://www.dji.com/downloads/products/matrice-200-series-v2, accessed: 2019.