IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 September 2022, accepted 2 October 2022, date of publication 5 October 2022, date of current version 12 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3212149

RESEARCH ARTICLE

Intelligent Drone-Base-Station Placement for Improved Revenue in B5G/6G Systems Under Uncertain Fluctuated Demands

HAYTHEM BANY SALAMEH[1,2], (Senior Member, IEEE), ALA'EDDIN MASADEH[3], (Member, IEEE), AND GHALEB EL REFAE[4]

[1]Department of Network and Communications Engineering, Al Ain University, Al Ain, United Arab Emirates
[2]Department of Telecommunications Engineering, Yarmouk University, Irbid 21163, Jordan
[3]Electrical Engineering Department, Al-Huson University College, Al-Balqa Applied University, Al-Salt 19110, Jordan
[4]College of Business, Al Ain University, Al Ain, United Arab Emirates

Corresponding author: Haythem Bany Salameh (haythem@email.arizona.edu)

This work was supported by the ASPIRE Award for Research Excellence (AARE-2021), Abu Dhabi, United Arab Emirates, under Grant AARE20-161.

ABSTRACT In this paper, the drone base-station (DBS) dispatching problem in a multi-cell B5G/6G network is investigated. The main objective is to achieve the highest system profit by serving the largest possible number of users with the least possible cost while considering the uncertain time-dependent fluctuated user's (service) demand in the different cells, the cost of dispatched drones, and the possible profit loss due to un-served users. The problem is formulated as a profit-maximization discount return problem. Due to the uncertainty in the demand (users) in each cell, the problem cannot be solved using conventional optimization methods. Hence, the problem is reformulated as a Markov decision problem (MDP). Due to the exponential complexity of finding the solution and the unavailability of statistical knowledge about user availability (demand) in the considered regions for such optimization, we adopt a reinforcement learning (RL) approach based on the state-action-reward-state-action (SARSA) algorithm to efficiently solve the MDP. Simulation results reveal that our RL-based approach significantly increases the overall operator profit by continuously adapting its DBS dispatching strategy based on the learned users' behavior in the network, which enables serving a larger number of users (highest revenue) with least number of DBSs (least cost).

INDEX TERMS Reinforcement learning, revenue, on-demand dispatching, uncertain demand, drone base-station.

I. INTRODUCTION

Beyond 5G (B5G)/6G networks are expected to support massive number of interconnected devices (orders of magnitude of today's number of devices), offer high-speed transmissions, enable real-time applications/services and extend network coverage [1], [2], [3]. In B5G/6G networks, the installation of a large-number of permanent terrestrial infrastructure to support user requirements in temporary crowded areas (or terrestrial infrastructure in rural areas) is economically infeasible from the operator's perspective due to the incurred high operational cost and/or the volatile and sophisticated environments. To this end, drone base stations (DBSs) have been envisioned as an integral low-cost part of the B5G/6G network architecture that can dynamically extend network coverage, support massive the Internet-of-Things (IoT) networking, and provide high-speed cellular services [3], [4], [5], [6]. DBSs provisions flexible, easy-to-deploy and low-cost networking by providing dynamic on-demand high-speed B5G/6G cellular coverage in temporary crowded areas (hot-spots) and limited-infrastructure areas with low cellular-coverage (e.g., rural areas, hard-to-reach areas, areas with emergency situations). Hence, a key design challenge in deploying DBSs in B5G/6G systems

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

is how to dynamically (on-demand) dispatch the available DBSs to the different regions in the network such that the overall profit made by the operator is maximized while considering the uncertain fluctuated demand over time. Specifically, the operation expenditure of the cellular operators can be minimized by avoid sending unnecessarily drones to the regions with low demands. On the other hand, the operator's profit in the crowded regions (hotspots) can be maximized by dispatching higher number of DBSs based on the prevailing time-dependent high traffic demand in these regions (reduce the profit-loss due to the un-served users). This paper investigates the DBS dispatching problem in a multi-cell B5G/6G networks with the objective of maximizing the overall operator's profit while considering the uncertain service demand fluctuations across the different service areas. This problem is analytically modeled as a profit-maximization dispatching problem. The uncertainty in the service demand in each cell makes the conventional optimization methods not applicable in solving this problem. Thus, our dispatching problem is redefined as a Markov decision problem (MDP). Due to the exponential complexity of our MDP problem and the availability of only causal knowledge about the availability of users in the considered regions, a reinforcement learning (RL) method is adopted to intelligently obtain efficient dispatching decisions. These decisions allows the operator to continuously adapt its DBS dispatching strategy based on the learned users' behaviour in the different regions. Accordingly, the cellular operator can significantly enhance its profit by intelligently form on-demand data-coverage hot-spots across the network based on users' demand (on-demand dispatching the needed number of DBSs per region). This allows serving larger number of users (achieving higher revenue) with least number of dispatched DBSs (least operating cost). We conduct simulation experiments to investigate the performance of the proposed RL-based dispatching optimization. The results indicate that significantly improvement in the overall system profit is demonstrated compared to reference dispatching algorithms.

The rest of this paper is organized as follows. Section II overviews the related work on using DBSs in B5G/6G networks. In Section III, the DBS-based cellular network model is described. The problem statement, formulation and RL-based solution are presented in Section IV. Section V discusses the performance evaluation of the proposed RL-based dispatching method. Finally, Section VI provides concluding remarks.

II. RELATED WORK

Several approaches and mechanisms have been proposed to enable efficient DBSs deployment and operation in cellular networks in terms of finding the best DBS placement, reducing DBS energy consumption (traveling time), air-to-ground channel modeling, optimizing path trajectory/transmit power-control, maximizing data rate, and DBS communications [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13]. Very few attempts have been done to optimize the overall revenue of using DBSs from the operator perspective. In [4], the authors proposed integrating the in-band full-duplex technology and DBSs in cellular networks for improving spectrum efficiency. They addressed the problem of DBS placement with joint bandwidth/power allocation, in which two heuristic methods were proposed. The authors in [5] developed a UAV-based cellular system based on tethered UAVs (tUAVs) with the objective of extending the drone flight time. They proposed connecting the UAVs with a ground station (GS) through a tether such that the GS can provide the tUAV with data and energy, which allows the DBS to fly for days. In [3], the authors investigated the on-demand use of drones for extending the broadband connectivity. They discussed the implementation challenges of drones in future networks including mobility, handover, energy constraint, optimal positioning and drone localization. The authors in [8] investigated the problem of finding the optimal interference-aware UAV trajectory planning in both single- and multi-cell wireless networks. In [9], the authors investigated the DBS placement with the objective of maximizing the number of covered users by each DBS. The paper in [10] investigated the DBS placement problem with the objective of providing energy efficient wireless coverage based on the successive convex approximation. In [11], the authors developed a delay-aware

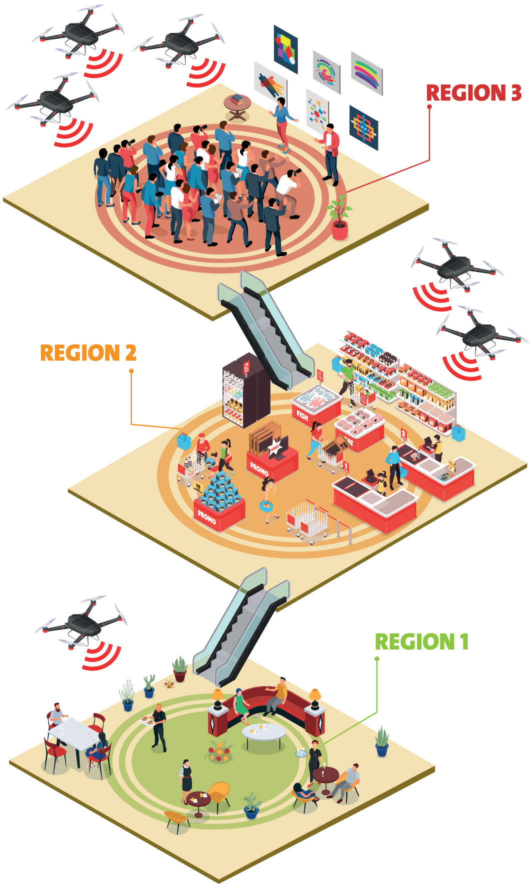


FIGURE 1. Network model of a flying network with 3 hotspots and 6 DBSs.

DBS placement algorithm that jointly determines the location of each DBS and its associated coverage area such that the total average latency is minimized subject to DBS battery constraint. The paper in [12] proposed a machine learning (ML)-based framework that enables a predictive DBS deployment to accompaniment the cellular infrastructure coverage. The ML approach implements the weighted-expectation maximization that estimates users' distribution and downlink traffic. Their main goal was improving the hotspot users' downlink speed and delay while reducing the DBSs energy consumption. However, their framework did not consider the time-varying uncertainty of the user demand and traffic distribution across the different regions in the network. Specifically, their work was based on an available historical dataset, which is used to model, train, and test their ML prediction algorithm. This significantly limits the adaptability of their drone deployment framework to any uncertain change in user's demand. In [13], the authors exploited the advantage of DBSs' mobility and Q-learning to detect autonomously the low-coverage regions (holes) in the network. Accordingly, they developed a centralized mechanism to find the optimal DBS 3D-placement that maximizes the number of associated users to the DBSs under wireless-backhaul/core-network limitations.

In summary, none of the previous works related to drone placement have optimized the overall cost of using the DBSs and their impact on the achieved operator's profit under uncertain users' demand. Our proposed RL-based DBS dispatching mechanism allows for continuously learning by interacting with the operating environment with the goal of maximizing the achieved operator's profit through continuously learning and updating its adopted dispatching policy to the one that results in a better (at least equal) the current achieved profit.

III. NETWORK MODEL

We consider a flying network that consists of N DBSs and a dispatching charging station. The service area consists of M sub-regions (i.e., $Z_1, Z_2, \ldots Z_M$). The DBSs are dynamically deployed in the service area to form small cells. Each DBS can serve up to $U_{\max}$ users. Fig. 1 shows an example of the considered network model, in which we consider 6 DBSs serving three in-door regions in a shopping mall. The users in each sub-region Z_i are served by the dispatched DBSs to that sub-region. The number of users (service demand) in each region ($\widetilde{U}_i$) is dynamically changing with uncertain behaviour due to user's uncertain mobility and fluctuated demands. The possible user demand in each region is discretized into K demand ranges $D_1 = [0-U_{\max}], D_2 = [(U_{\max}+1)-2U_{\max}]$, …, and $D_K = [(K-1)U_{\max} + 1 - KU_{\max}]$ users. The transition probability of demand ranges in each cell follows a Markov process according to a transition probability matrix. Figure 2 shows an example of a transition diagram of the demand variation (along with the transition probabilities) for 3-demand ranges in a given region Z_i. Let the number of dispatched DBSs for region Z_i during a given dispatching flight time slot t is denoted by $N_D^{(i,t)}$. The DBSs covering the same area use different frequency bands such that the co-channel interference is limited. The backhaul connectivity to the DBSs can be realized through wireless connection to any operating core network entity (e.g., satellite connectivity, neighboring macro-ground BS, mobile-based BS vehicles) [14]. The maximum number of served users in a given region Z_i is determined by the number of dispatched DBSs to that region $N_D^{(i,t)}$, in which the maximum possible served users is limited by $N_D^{(i,t)} \times U_{\max}$. Each user in a given region Z_i pays a subscription service fee (P) to the cellular operator if it gets served by the dispatched DBSs during the flight period. The cost of dispatching the DBSs for the region Z_i consists of a total fixed-cost (FC) and a total variable cost (VC). Specifically, the total cost (TC_i^t) in Z_i during the flight time t can be computed as [15], [16]:

$$TC_i^t = FC + VC(N_D^{(i,t)}) = FC + C^D N_D^{(i,t)} \tag{1}$$

where C^D represents the operating cost for each dispatched DBS. The total revenue (TR_i^t), given that the subscription service fee is P, can be computed as:

$$TR_i^t = U_{served}^{(i,t)} \times P \tag{2}$$

where $U_{served}^{(i,t)} = \min\{N_D^{(i,t)} \times U_{\max}, \widetilde{U}_i^t\}$ denotes the number of served users in Z_i during flight time t, and $\widetilde{U}_i^t$) is the actual number of users (demand) in cells i during flight time t.

Based on (1) and (2), the total achieved profit in each cell during one DBS flight period t can be computed as:

$$\begin{aligned}\Pi_i^t &= TR_i^t - TC_i^t \\ &= U_{served}^{(i,t)} \times P - (FC + C^D N_D^{(i,t)})\end{aligned} \tag{3}$$

Consequently, the over all achieved profit in the system during the flight period t can be determined as:

$$\begin{aligned}\Pi_t &= \sum_{i=1}^{M} \Pi_i^t \\ &= \sum_{i=1}^{M} (U_{served}^{(i,t)} P - (FC + C^D N_D^{(i,t)})) \\ &= \sum_{i=1}^{M} (\min\{N_D^{(i,t)} U_{\max}, \widetilde{U}_i^t\})P - (FC + C^D N_D^{(i,t)})\end{aligned} \tag{4}$$

The overall profit during each flight period in the system depends heavily on the number of dispatched DBSs for each region as well as the uncertain number of users (fluctuated demand) in each sub-region.

IV. DRONE-BS DISPATCHING PROFIT MAXIMIZATION PROBLEM

A. PROBLEM STATEMENT AND FORMULATION

The main goal of our formulation is to dispatch DBSs to the different cells in the network such that the achieved system profit over-time is maximized subject to DBS availability and coverage constraints under uncertain fluctuated service demands across the different cells. This can be done by finding the optimal policy of DBS dispatching that minimizes the total cost while increasing the overall revenue by serving the largest possible number of users across the network. This DBS dispatching decision problem can be formulated as a cumulative discounted profit maximization problem with the availability of only causal knowledge of the number of users in each cell (actual demand per cell). Due to the unavailability of future information about the demand in each cell, we define the objective function of our optimization as the maximization of the expected discounted profit (return), i.e., the expected cumulative discounted dispatching profit. The discounted return after the flight time t, G_t, can be written as:

$$G_t = \sum_{j=t}^{T-1} \gamma^{j-t} \Pi_{j+1} \tag{5}$$

where t is the start point for accumulating the consecutive rewards, T is a last time slot of an episode ($T - t$ represents one DBS flight period), Π_{j+1} is the immediate reward (profit), achieved in all regions at time slot $j+1$ resulting from taking action A_j, which is defined as the number of DBSs that are to be dispatched to each region Z_i at time j ($N_D^{(i,j)}$, $\forall Z_i$). The term $0 \leq \gamma \leq 1$ represents the discount factor that prioritizes the importance of rewards over time. The old received rewards contribute more in the future cumulative reward. The number of available DBS constraint indicates that, at any given time j, no more than N DBSs can be dispatched to the different regions/cells, which can be written as:

$$\sum_{i=1}^{M} N_D^{(i,j)} \leq N, \quad j = t, \ldots, T-1. \tag{6}$$

The DBS capacity constraint limits the maximum number of users a DBs can serve to $U_{\max}$, in which the total number of served user in a given region Z_i at time j ($U_{served}^{(i,j)}$) is limited to $N_D^{(i,j)} \times U_{\max}$.

To maintain coverage in each region Z_i, $N_D^{(i,j)}$ should satisfy the following constraint, in which at least one DBS should be dispatched to each region i:

$$N_D^{(i,j)} \geq 1, \quad j = t, \ldots, T-1, \ \forall i \in \{1, 2, \ldots M\} \tag{7}$$

Our optimization problem attempts to determine the number of dispatched DBSs $N_D^{(i,j)}$, $\forall i$ (i.e., the action A_j) that results in maximizing the expected discounted profit over an infinite horizon, which can be expressed as:

$$\begin{aligned} &\max_{\{A_j\}} \lim_{T\to\infty} \mathbb{E}[G_t] \\ &\text{such that for } j = t, \ldots, T-1, \\ &\sum_{i=1}^{M} N_D^{(i,j)} \leq N, \\ &N_D^{(i,j)} \geq 1, \quad \forall i \in \{1, 2, \ldots M\} \\ &U_{served}^{(i,j)} \leq N_D^{(i,j)} \times U_{\max}, \quad \forall i \in \{1, 2, \ldots M\}. \end{aligned} \tag{8}$$

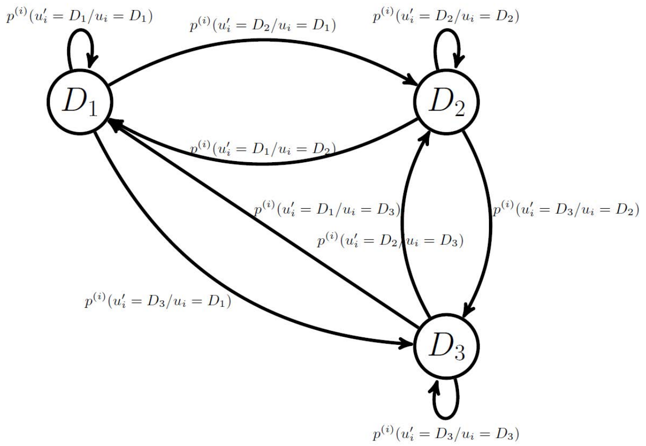

$p^{(i)}(u'_i = D_1/u_i = D_1)$
$p^{(i)}(u'_i = D_2/u_i = D_1)$
$p^{(i)}(u'_i = D_2/u_i = D_2)$
D_1
D_2
$p^{(i)}(u'_i = D_1/u_i = D_2)$
$p^{(i)}(u'_i = D_1/u_i = D_3)$
$p^{(i)}(u'_i = D_3/u_i = D_2)$
$p^{(i)}(u'_i = D_2/u_i = D_3)$
$p^{(i)}(u'_i = D_3/u_i = D_1)$
D_3
$p^{(i)}(u'_i = D_3/u_i = D_3)$

FIGURE 2. Transition diagram of the user-demand in cell Z_i for 3 demand ranges.

B. THE MDP REFORMULATION

Since the exact number of users (demand $\widetilde{U}_i^t$) in the different cells is unknown in the previous time slot $t - 1$, the formulated Profit-maximization problem cannot be solved using traditional optimization techniques. Since our optimization involves decision making (dispatching actions) under uncertainty, the dispatching actions, decided by the cellular operator, can be transformed to an MDP [17]. The MDP model is analytically defined using: (1) set of states $\mathcal{S}$, (2) set of actions $\mathcal{A}$, (3) state-transition probability $p(s'|s, a)$, (4) immediate rewards $r(s, a, s')$ and (5) the forgotten (discount) factor γ. The set of actions is defined as $\mathcal{A} \triangleq \{a_1, \ldots, a_k, \ldots a_L\}$, where L is the total number of feasible actions. The term $p(s'|s, a)$ presents the probability of transiting from state $s \in \mathcal{S}$ to another state $s' \in \mathcal{S}$ given that action $a \in \mathcal{A}$ is decided at state s, whereas $r(s, a, s')$ quantifies the rewards granted when transiting from s to s' when action a is taken [17]. Accordingly, the MDP formulation associated with our optimization problem can be defined using the following states, action and rewards:

It is important to note the difference between the notations used to represent different variables in time domain and in the MDP formulation, in which capital letters are used in time domain and small letters are used in the formulated underlying MDP.

1) POSSIBLE STATES

The state $s \in \mathcal{S}$ is defined as a group of $2M$ elements representing the number of dispatched DBSs ($n_D^{(i)}$) and the user-demand range ($D^{(i)} \in \mathcal{D}$) in each region $Z_i \in \{1, 2, \ldots M\}$, where $\mathcal{D} = \{D_1, D_2, \cdot\cdot\cdot, D_K\}$ is the set of all possible demand ranges. We note that $D^{(i)}$ depends on the actual number of users in region Z_i. Specifically, the state s can be expressed as $s = (n_D^{(1)}, n_D^{(2)}, \ldots n_D^{(M)}, D^{(1)}, D^{(2)}, \ldots D^{(M)})$.

2) POSSIBLE ACTIONS

The action a is defined as the number of dispatched DBSs for each region at each state. The set of available actions to the cellular operator is given by $\mathcal{A} = \{a_1, \dots, a_k, \dots a_L\}$, where L is the total number of feasible combinations of DBS dispatching to the M regions, and action $a \in \mathcal{A}$ is an M-dimensional vector representing the taken action of dispatching DBSs to each region Z_i at state s, for $i = \{1, 2, \dots M\}$ (e.g., $a_1 = (1, 1, \dots 1)$ represents the action that only one DBS is sent to each region). We note that $\sum_{i=1}^{M} n_D^{(i)} \leq N$, which indicates that the total number of dispatched DBSs during one transition step/flight period does not exceed the available number of DBSs N.

3) STATE TRANSITION PROBABILITY

The demand (number of users) in the region Z_i (u_i, $\forall i$), in general, evolves according to an unknown Markov process with transition probability $p^{(i)}(u_i'|u_i)$, where u_i and u_i' denote the current and evolved number of users in region Z_i during one step transition, respectively. The user-demand range in the region Z_i at a given state not time, denoted as $D^{(i)}$, is mapped into one of the defined demand ranges in $\mathcal{D}$. Hence, the transition probability from demand range D to D' in region Z_i is given as $p^{(i)}(u_i' \in D'|u_i \in D)$.

The number of dispatched DBSs evolves according to:

$$(n'_{D^{(1)}}, n'_{D^{(2)}}, \dots n'_{D^{(M)}}) = a \tag{9}$$

where a is the taken action at the current state s. Given the independent user-demand in the different regions, the transition probability from state s with demand range $D^{(i)} = D$ to state s' with user-demand range D', $\forall i$, given an action a is taken, can be written as:

$$p(s'|s, a) = \begin{cases} \Pi_{i=1}^{M} p^{(i)}(u_i' \in D'|u_i \in D), & \text{if (9) is satisfied} \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

4) IMMEDIATE REWARDS

The immediate reward (i.e., the reward resulting from taking action a at s and transiting to s') is given as:

$$r(s, a, s') = \begin{cases} r, & \text{if all users are served without extra DBSs} \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

where $r = \sum_{i=1}^{M} u_{served}^{(i)} P - (FC + C^D n_D^{*(i)})$, and $n_D^{*(i)}$ is the least number of DBSs needed to serve all the users in each region Z_i. Mathematically, $n_D^{*(i)}$ is the one that satisfies the following inequality: $U_{\max}(n_D^{(i)} - 1) + 1 \leq u_i \leq U_{\max} n_D^{(i)}, \forall i$, where $i = 1, 2, \dots M$. The lower inequality ensures that the number of dispatched DBSs to each region Z_i can serve the current demand in that region, while the upper inequality ensures that no unnecessary DBSs are assigned to region Z_i.

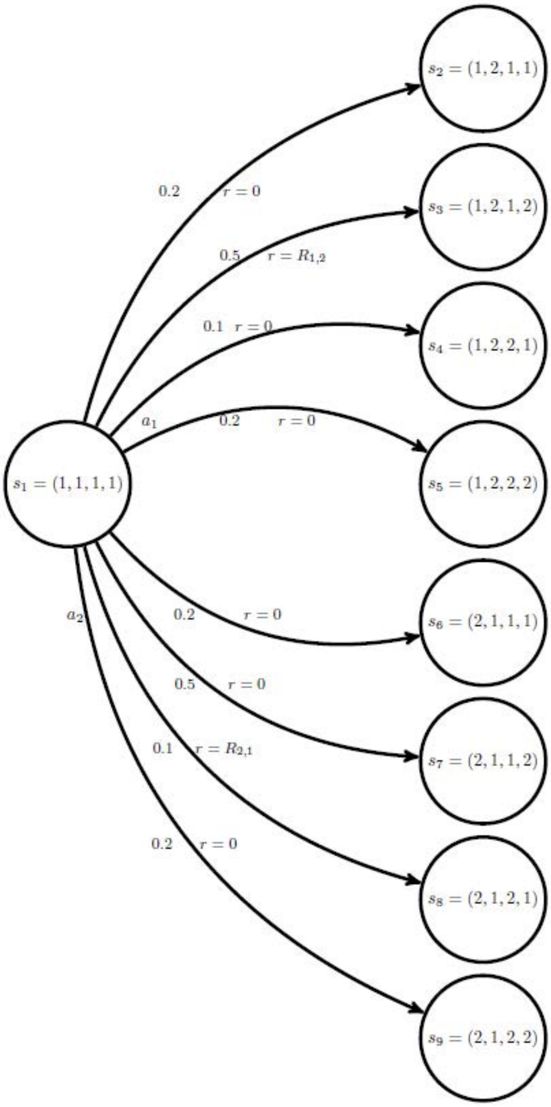
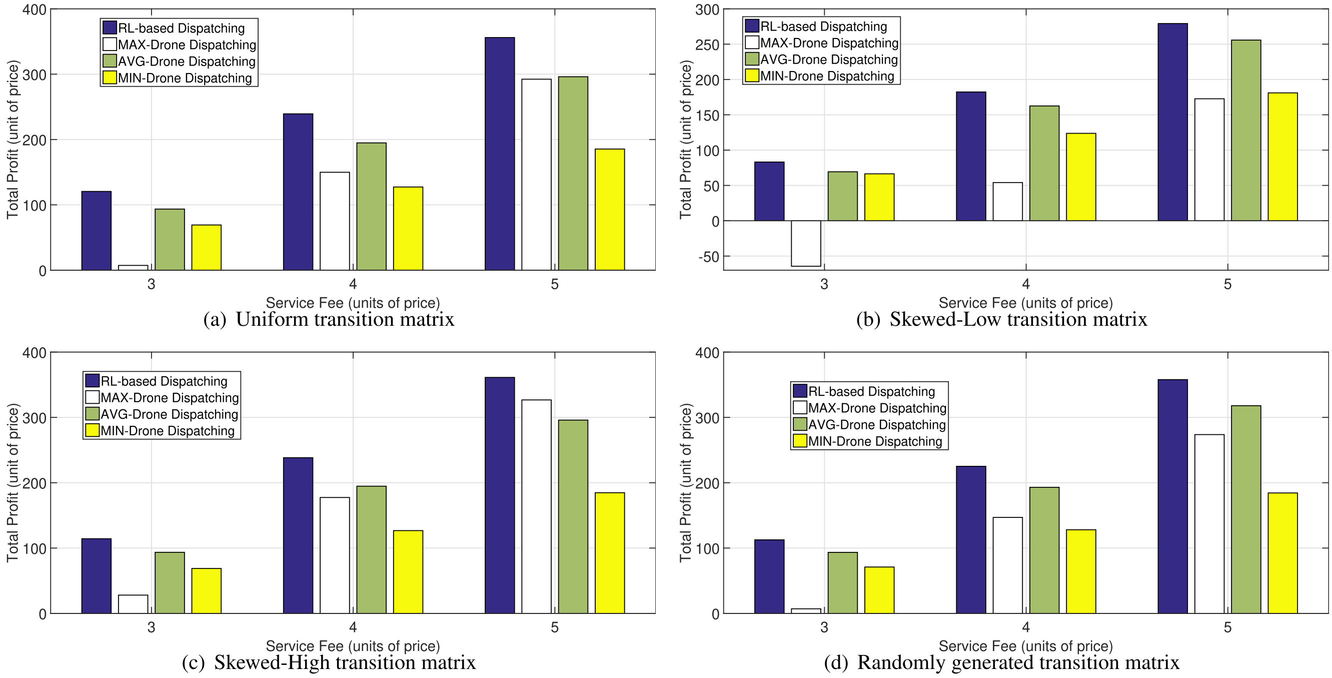


FIGURE 3. An MDP showing all possible next states for the actions $a_1 = \{1, 2\}$ and $a_2 = \{2, 1\}$ for a network with 2-region and 2-demand-range.

5) POLICY

A deterministic policy π is used to map states into the dispatching action selected at each state, $\pi(\cdot) : s \rightarrow a, \forall s$. The main goal is to maximize our objective function (achieved profit) by finding the optimal policy (π^*), where π^* results in a better (or at least equal) action-value function (i.e., $q_{\pi^*}(s, a) \geq q_\pi(s, a), \quad \forall s \in \mathcal{S}$) [18].

Figure 3 shows an illustrative example for an MDP that represents the considered DBS system. The service area consists of 2 cells/regions (i.e., Z_1, and Z_2). The set of user demands consists of 2 demand ranges (i.e., $\mathcal{D} = \{D_1, D_2\}$). This example shows all possible next states s' for the actions $a_1 = \{1, 2\}$ and $a_2 = \{2, 1\}$ given that the current state is $s = (1, 1, 1, 1)$ with demand range D_1 in both cells. This figure also shows the transition probabilities from state s to all other states s'.



(a) Uniform transition matrix

(b) Skewed-Low transition matrix

(c) Skewed-High transition matrix

(d) Randomly generated transition matrix

FIGURE 4. The operator's profit versus *P* for 3 regions under uniform, skewed and random demand transition matrices.

C. THE PROPOSED RL-LEARNING APPROACH

This part discusses the proposed solution for the considered problem using RL, where RL has the capability of handling the challenge of knowledge unavailability about the user demand and mobility. In the proposed RL approach, state-action-reward-state-action (SARSA) learning method is used as a prediction algorithm, while the convergence-based algorithm is used as an exploration algorithm.

1) THE PREDICTION METHOD

SARSA learning is a temporal-difference (TD) method [17], which is used to predict and evaluate the values of different actions taken by the DBS operator at different states. The values of actions at the feasible states are predicted and evaluated using SARSA as follows. Let the current state be s. The exploration algorithm (e.g., the convergence-based exploration algorithm) is used to select an action a (i.e., dispatching a number of DBSs for each region) at s. This results in a transition from s to a new state s' and obtaining a reward $r(s, a, s')$. After transiting to s', the used exploration algorithm is used to select an action at s'. Using the information collected from the previous steps, the value of the current state-action pair (i.e., $Q(s, a)$) is updated as [17]:

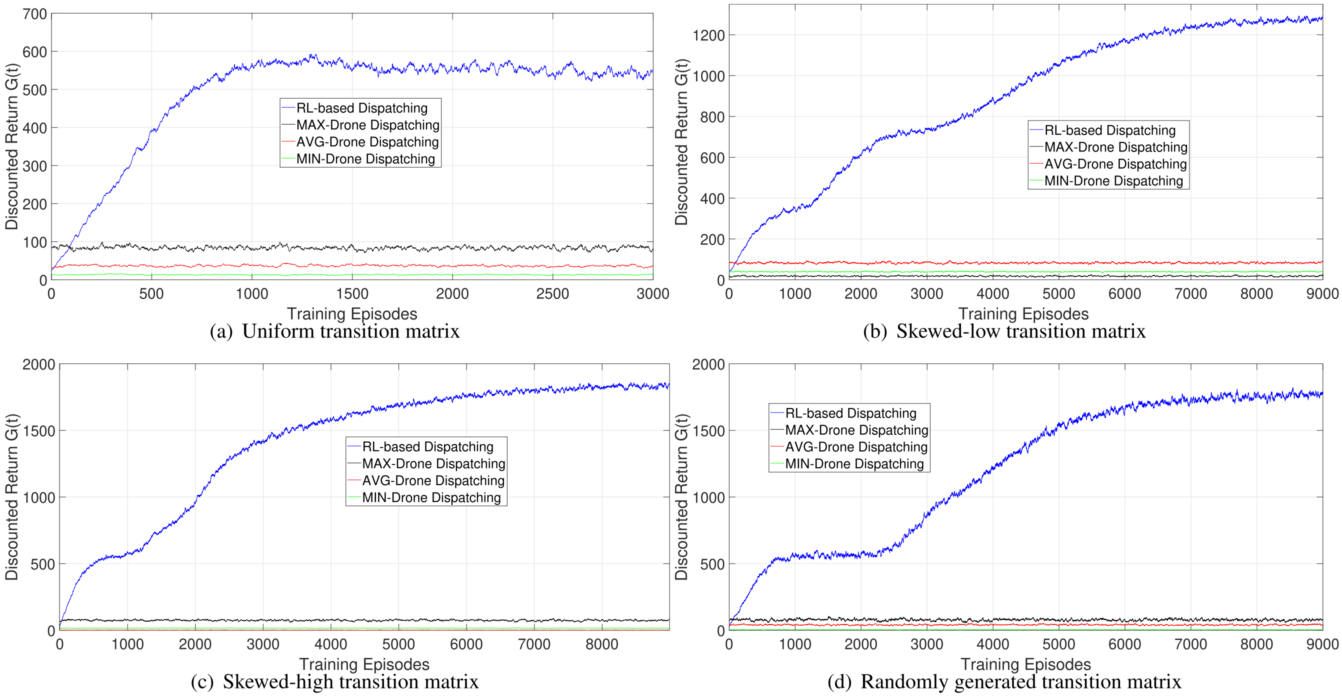
$$Q(s, a) \leftarrow Q(s, a) + \alpha[r(s, a, s') + \gamma Q(s', a') - Q(s, a)] \tag{12}$$

where α is the learning rate that determines how much the newly acquired data contributes to the older information.

2) THE EXPLORATION ALGORITHM

This part discusses the exploration algorithm implemented in our proposed RL-based algorithm, which is called the convergence-based exploration algorithm. The exploration algorithm handles the challenge of having only causal knowledge about the dynamic change in users' demand in the network. It is responsible for finding a balance between the exploration and exploitation modes during the learning process. In the exploration mode, the agents gather more information about the underlying model by trying new policies in hope of finding a policy that performs better than the current best one. On the other hand, in the exploitation mode, the agents make their decisions by following the best available policy according to the current available information, while it is possible that there is an unexplored policy that might perform better than the current best policy.

The convergence-based exploration algorithm aims to balance exploitation and exploration modes by employing two parameters, called exploration time τ, and the action-value function convergence error ζ. In the considered task, τ denotes the maximum time that DBSs are allowed to explore different policies. After this time, DBSs are obligated to use the best policy as per the current information. ζ specifies the approximate accuracy of estimate value of a state-action pair, which is defined as the maximum error allowed in estimating the value of a state-action pair during the exploration mode. The same action continues to be used at a state when it is visited until the value of the state-action pair is approximated to a value with an error less than or equal to ζ [19], [20]. The convergence-based exploration algorithm is implemented by



(a) Uniform transition matrix

(b) Skewed-low transition matrix

(c) Skewed-high transition matrix

(d) Randomly generated transition matrix

FIGURE 5. The operator profit for 3 regions and $P = 3$ under different demand transition matrices (similar behaviour has been observed for other values of P).

conducting the following steps: before learning, actions of dispatching DBSs are randomly assigned to all available states. At each visited state, DBS operator continues to use the same dispatching actions according to the current policy until the approximate error of the estimate value of the state-action pair becomes equal or less than ζ as long as the available time is less than τ. Once the estimated value of the state-action pair converges to a value with an approximate error that is less than or equal to ζ, a random unexplored action is assigned to this state and the policy is updated. The DBS operator keeps following the aforementioned steps unless one of these two conditions is met. The first one is met if all dispatching actions available at a state have been evaluated before reaching τ. At this time, the best learned action at this state will be used during the remaining time. The second condition occurs if the remaining time is τ. At this time, DBS operator terminates the exploration mode, and turns to the exploitation mode, where the best learned policy is followed during the remaining time.

3) COMPLEXITY AND CONVERGENCE

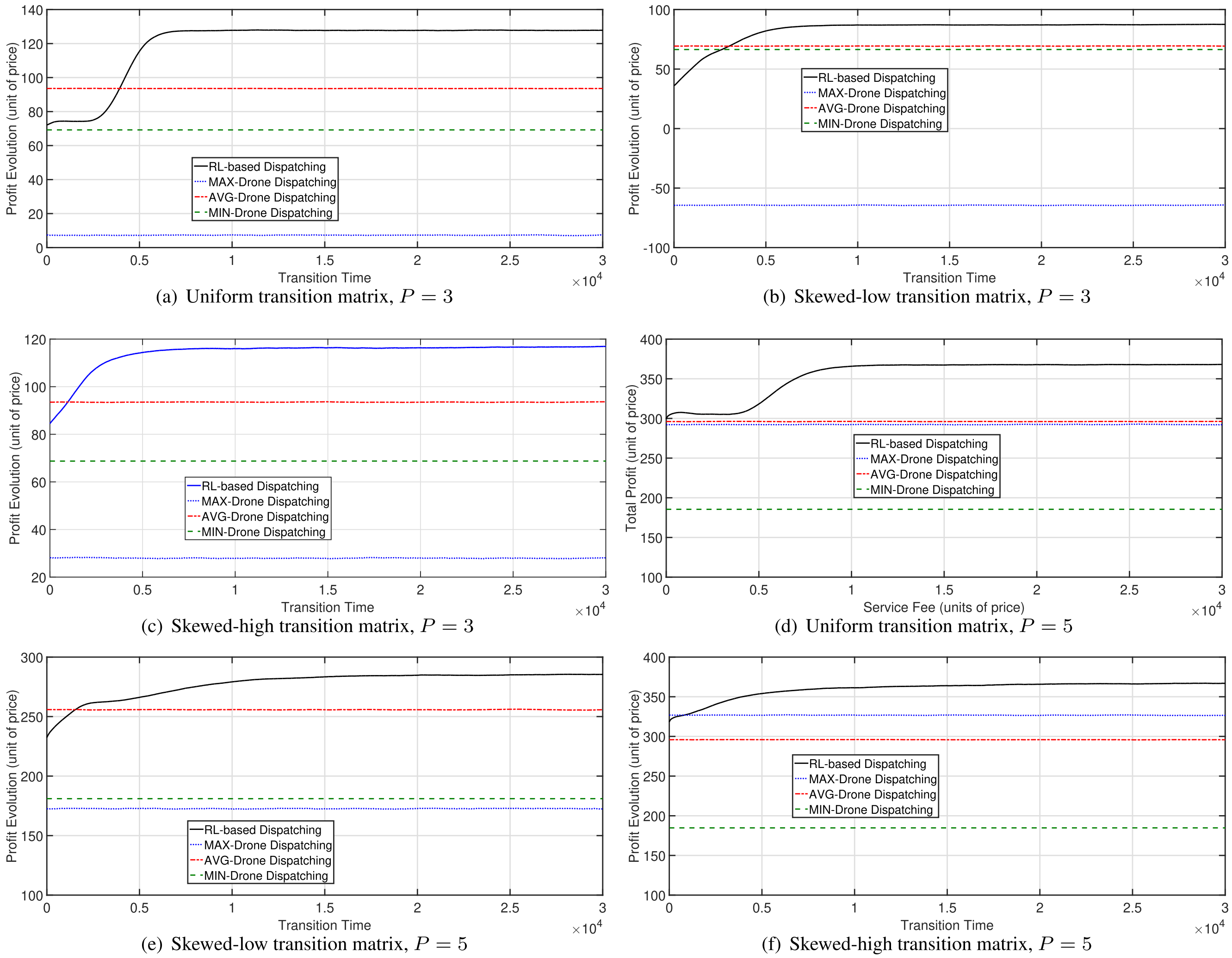
The adopted convergence-based exploration algorithm along with SARSA prediction algorithm aims at achieving an efficient learning performance. This RL-based algorithm tries to accurately predict the values of the different state-action pairs, and then, utilize the best resulting policy. The work in [20] has proven that the complexity of this RL-based algorithm is $\mathcal{O}(|T|)$, where T is the final time step of an episode.

V. PERFORMANCE EVALUATION

Simulation experiments using MATLAB programs are conducted to analyze the effectiveness of our proposed RL-based dispatching algorithm, referred to as RL-based Dispatching. The performance of our dispatching algorithm is compared with that of the MAX-Drone Dispatching (4 DBSs for each region), AVG-Drone Dispatching (2 DBSs for each region) and MIN-Drone Dispatching (1 DBSs for each region) algorithms. The simulation time is divided into time slots, each presenting a flight time supported by each DBS's battery. During each flight-period (time slot), the cellular operator dispatches DBSs to the different regions according to the learned policy. The discount factor and learning rate used in our RL-based experiments are set to $\gamma = 0.95$ and 0.1, respectively. The reported results are averaged of 1000 independent episodes (learning sessions), each with 30000 time slots. The main performance metrics are the overall achieved operator's profit and the cumulative discounted return over the learning sessions.

A. SIMULATION SETUP

A network of 3 hotspots and 12 DBSs is simulated in a cellular coverage area. The user demand in each hotspot varies according to an unknown time-varying behavior. We set the operating drone cost per DBS to $C^D = 35$ units of price, and the fixed operating cost to 10 units. Four different user-demand scenarios are simulated. The four scenarios differ in the values of the transition probabilities of the demand range indicator (i.e., the probability of transiting from one demand range to another). Specifically, uniform,



(a) Uniform transition matrix, $P = 3$

(b) Skewed-low transition matrix, $P = 3$

(c) Skewed-high transition matrix, $P = 3$

(d) Uniform transition matrix, $P = 5$

(e) Skewed-low transition matrix, $P = 5$

(f) Skewed-high transition matrix, $P = 5$

FIGURE 6. The operator profit for 3 regions and $P = 3$ and 5 under different demand transition matrices.

skewed high, skewed low, and randomly generated transition matrices are considered. For the uniform demand scenario, the transition probability for user demand in any region Z_i ($p(u_i' \in D_j|u_i \in D_k)$) are assigned uniformly for the 4 demand ranges as follows $p(u_i' \in D_j|u_i \in D_k) = 0.25$, $\forall j, k \in \{1, 2, 3, 4\}$. For the skewed-high and skewed-low-demand behaviour in each hotspot, the transition probabilities between the different demand ranges are randomly assigned for each of the learning session such the total probability of all possible transitions from one user-demand range to all other ranges equals to 1. For the skewed high-demand scenarios, the transition probabilities from any demand range j to the high demand ranges (D_3 and D_4) are randomly selected such that $p(u_i' \in D_3|u_i \in D_k) + p(u_i' \in D_4|u_i \in D_k)= 0.7$, where $k = 1, 2, 3, 4$. For the skewed low-demand scenarios, the transition probabilities to the low demand ranges (D_1 and D_2) in each region Z_i are randomly selected such that $p(u_i' \in D_1|u_i \in D_k) + p(u_i' \in D_2|u_i \in D_k) = 0.7$. For the randomly generated matrix, the transition probabilities are randomly selected from the range (0, 1) such that the sum of all transition probabilities from a given demand range to all others (including staying in the same demand range) is 1.

B. SIMULATION RESULTS

Fig. 4(a)-(d) plots the total operator's profit for different subscription service fee (P) values under uniform, skewed-high, skewed-low and randomly generated transition probability matrices, respectively. Fig. 4 indicates that our proposed RL-based Dispatching algorithm significantly enhances the operator's profit compared to the other algorithms, irrespective of P and user-demand scenarios. For example, Fig. 4(a) reveals that under uniform user-demand and $P = 3$, our algorithm outperforms the MAX, AVG- and MIN-Drone Dispatching algorithms by up to 130%, 27% and 87%, respectively. Fig. 4 also indicates that as P increases the MAX-Drone Dispatching outperforms the AVG- and MIN-Drone Dispatching algorithms. This is because the total revenue increases by increasing P, which compensates for the profit loss due to the higher number of un-necessarily dispatched DBSs in MAX-Drone Dispatching. On the other hand, MAX-Drone Dispatching performs the worst under low values of P as the profit loss due to the high number of un-necessarily dispatched DBSs becomes dominant, resulting in reduced profit.

Fig. 5 plots the discounted return G_t under uniform, skewed-high, skewed-low and randomly generated transition probability matrices for $P = 3$. The discounted return, known as cumulative discounted received reward, is defined as the cumulative valuable rewards received from a DBS dispatching process during a given time. This figure illustrates that the RL-based Dispatching algorithm outperforms the other algorithms in terms of the overall achieved operator's profit. For the RL-based dispatching algorithm, the discounted return G_t significantly increases with more collected data (i.e., learning experience). Fig. 5 shows that the performance of the proposed RL-based dispatching algorithm saturates as time elapses. This is due to the fact that the system has reached a learned policy that cannot be enhanced further. This figure also indicates that the performance of the MAX-, AVG- and MIN-Drone Dispatching algorithms does not change with time (no performance enhancement is observed with time). This is expected as each one of the three algorithms adopts a fixed DBSs dispatching policy that does not change with time. Similar behaviour are observed for other values of P.

Finally, Fig. 6 plots the operator profit under uniform, skewed-high, and skewed-low transition probability matrices for $P = 3$ and $P = 5$. This Figure reveals that our propose RL-based algorithm outperforms the other algorithms in terms of the achieved profit, irrespective of P and the user's demand behaviour. It is clear that the achieved profit of our algorithm significantly increases with time due to the continuous learning and more collected data. This figure also shows that as P increases the overall profit increases. For $P = 3$, Figures 6(a) and(c) show that the AVG-Drone Dispatching outperforms the MIN- and MAX-Drone Dispatching algorithms under uniform and skewed-high demand scenarios. For skewed-low scenarios, Fig. 6(b) indicates that AVG- and MIN-Drone Dispatching provide comparable performance that outperforms the achieved performance of the MAX-Drone Dispatching algorithm. This is because the MAX-Drone algorithm dispatches higher number of DBSs than required, resulting in extra cost and reduced profit. For high price $P = 5$, Fig. 6(d) and (f) reveal that MAX-Drone Dispatching outperforms the MIN- and AVG-Drone Dispatching algorithms under uniform and skewed-high scenarios. This is because the total revenue of serving more users increases as P increases, which compensates for the extra cost incurred by dispatching higher number of DBSs. Under skewed-low scenarios (Fig. 6(e)), the MAX-Drone Dispatching performs the worse as it incurs higher operating cost with least revenue. We note that under higher price P, AVG-Drone Dispatching outperforms the MIN-Drone Dispatching algorithm. This is because the possible revenue gain that can be made by dispatching larger number of DBSs is higher than their incurred operating cost.

VI. CONCLUSION

This paper investigated the profit-maximization DBS dispatching problem in a multi-cell B5G/6G network while being aware of the time-varying uncertain fluctuated user demand across the different cells in the network. The problem was modeled as a profit-maximization discount-return problem with the goal of improving the overall operator's profit by minimizing the cost of the dispatched DBSs and reducing the possible profit loss due to the unavailability of DBSs to achieve user demand. To deal with the uncertainty in user demand, the profit-maximization was mapped to an MDP problem, for which a SARSA-based RL algorithm was used to solve the formulated MDP. The proposed RL-based dispatching algorithm continuously adapts its dispatching policy based on a continuous learning provided by the online interaction with the operating environment. Simulation results showed that significant operator profit improvement can be achieved, compared to reference static dispatching algorithms, by considering the demand uncertainty across the network when performing on-demand per-cell DBS dispatching.

REFERENCES

[1] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan, and M. Zorzi, "Toward 6G networks: Use cases and technologies," *IEEE Commun. Mag.*, vol. 58, no. 3, pp. 55–61, Mar. 2020.

[2] Z. Zhang, Y. Xiao, Z. Ma, M. Xiao, Z. Ding, X. Lei, G. K. Karagiannidis, and P. Fan, "6G wireless networks: Vision, requirements, architecture, and key technologies," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 28–41, Sep. 2019.

[3] G. Amponis, T. Lagkas, M. Zevgara, G. Katsikas, T. Xirofotos, I. Moscholios, and P. Sarigiannidis, "Drones in B5G/6G networks as flying base stations," *Drones*, vol. 6, no. 2, p. 39, Feb. 2022.

[4] L. Zhang, Q. Fan, and N. Ansari, "3-D drone-base-station placement with in-band full-duplex communications," *IEEE Commun. Lett.*, vol. 22, no. 9, pp. 1902–1905, Sep. 2018.

[5] M. Kishk, A. Bader, and M.-S. Alouini, "Aerial base station deployment in 6G cellular networks using tethered drones: The mobility and endurance tradeoff," *IEEE Veh. Technol. Mag.*, vol. 15, no. 4, pp. 103–111, Dec. 2020.

[6] M. Giordani and M. Zorzi, "Non-terrestrial networks in the 6G era: Challenges and opportunities," *IEEE Netw.*, vol. 35, no. 2, pp. 244–251, Mar./Apr. 2021.

[7] B. Li, Z. Fei, and Y. Zhang, "UAV communications for 5G and beyond: Recent advances and future trends," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2241–2263, Apr. 2019.

[8] J. Lee and V. Friderikos, "Interference-aware path planning optimization for multiple UAVs in beyond 5G networks," *J. Commun. Netw.*, *J. Commun. Netw.*, vol. 24, no. 2, pp. 1–14, 2022.

[9] E. Kalantari, M. Z. Shakir, H. Yanikomeroglu, and A. Yongacoglu, "Backhaul-aware robust 3D drone placement in 5G+ wireless networks," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, May 2017, pp. 109–114.

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

[11] X. Sun and N. Ansari, "Latency aware drone base station placement in heterogeneous networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2017, pp. 1–6.

[12] Q. Zhang, W. Saad, M. Bennis, X. Lu, M. Debbah, and W. Zuo, "Predictive deployment of UAV base stations in wireless networks: Machine learning meets contract theory," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 637–652, Jan. 2021.

[13] S. A. Al-Ahmed, M. Z. Shakir, and S. A. R. Zaidi, ''Optimal 3D UAV base station placement by considering autonomous coverage hole detection, wireless backhaul and user demand,'' *J. Commun. Netw.*, vol. 22, no. 6, pp. 467–475, Dec. 2020.

[14] X. Li, ''Deployment of drone base stations for cellular communication without apriori user distribution information,'' in *Proc. 37th Chin. Control Conf. (CCC)*, Jul. 2018, pp. 7274–7281.

[15] M. Baye and J. Prince, *Managerial Economics and Business Strategy*, 9th ed. New York, NY, USA: McGraw-Hill, 2017.

[16] D. Besanko, D. Dranove, M. Shanley, and S. Schaefer, *Economics of Strategy*. Hoboken, NJ, USA: Wiley, 2009.

[17] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[18] P. Blasco, D. Gündüz, and M. Dohler, "A learning theoretic approach to energy harvesting communication system optimization," *IEEE Trans. Wireless Commun.*, vol. 12, no. 4, pp. 1872–1882, Apr. 2013.

[19] A. Masadeh, Z. Wang, and A. E. Kamal, ''Convergence-based exploration algorithm for reinforcement learning,'' *Electr. Comput. Eng. Tech. Rep. White Papers*, vol. 1, pp. 1–13, Jan. 2018.

[20] A. Masadeh, Z. Wang, and A. E. Kamal, "Look-ahead and learning approaches for energy harvesting communications systems," *IEEE Trans. Green Commun. Netw.*, vol. 4, no. 1, pp. 289–300, Mar. 2020.

HAYTHEM BANY SALAMEH (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from The University of Arizona, Tucson, AZ, USA, in 2009. He is currently a Professor of communications and networking engineering and the Dean of Scientific Research and Graduate Studies, Al Ain University, Al Ain, United Arab Emirates (on leave from Yarmouk University (YU), Irbid, Jordan). He also holds a Visiting Professor position with Staffordshire University, Stoke-on-Trent, U.K. His research interests include wireless networking, with emphasis on dynamic spectrum access, cognitive radio networking, the Internet of Things, security, and distributed protocol design. He was a recipient of the Jordan Science Research Support Foundation (SRSF) Prestigious Award for Distinguished Research in ICT, in 2015, the Best Researcher Award for Scientific Colleges in YU, in 2016, and the SRSF Award for Creativity and Technological Scientific Innovation, in 2017. He has served and continues to serve on the Technical Program Committee of many international conferences.

ALA'EDDIN MASADEH (Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from the Jordan University of Science and Technology, Irbid, Jordan, in 2010 and 2013, respectively, and the Ph.D. degree in electrical engineering and computer engineering from Iowa State University, Ames, IA, USA, in 2019. He is currently an Assistant Professor with the Electrical Engineering Department, Al-Huson University College, Al-Balqa Applied University, Jordan. His research interests include wireless networks, energy harvesting communications, unmanned aerial vehicles, reinforcement learning, machine learning, and artificial intelligence.

GHALEB EL REFAE is currently a Professor in financial economics with expertise in higher education management, risk management in higher education institutions, and university corporate governance. Since July 2011, he has been the President of Al Ain University (AAU). Before joining Al Ain University, he led the Faculty of Economics and Administrative Science, Al Zaytoonah University of Jordan (ZUJ), for 16 years as the Dean.