Learning in the Sky: Towards Efficient 3D Placement of UAVs

Atefeh Hajijamali Arani*, M. Mahdi Azari‡, William Melek§, and Safieddin Safavi-Naeini*  
*Department of Electrical and Computer Engineering, University of Waterloo, Canada  
‡Communication Systems Division, CTTC, Spain  
§Department of Mechanical and Mechatronics Engineering, University of Waterloo, Canada

Abstract—Deployment of unmanned aerial vehicles (UAVs) as aerial base stations to support cellular networks can deliver a fast and flexible solution for serving high and varying traffic demand. In order to adequately leverage the benefit of UAVs deployment, their efficient placement is of utmost importance, and requires to intelligently adapt to the environment changes. In this paper, we propose novel learning-based mechanisms for the three-dimensional deployment of UAVs assisting terrestrial networks in the downlink for overloaded situations. The problem is modeled as a game among UAVs. To solve the game, we utilize tools from reinforcement learning, and develop low complexity algorithms based on the multi-armed bandit and satisfaction methods to learn UAVs’ locations. Simulation results reveal that the proposed satisfaction based UAV placement algorithm can yield significant performance gains up to about 50% and 41% in terms of throughput and the number of outage users, respectively, compared to a learning based benchmark algorithm.

Index Terms—Game Theory, Reinforcement Learning, Multi-armed Bandit, Satisfaction, Unmanned Aerial Vehicles.

I. INTRODUCTION

Due to the novel types of services and ongoing advances in unmanned aerial vehicle (UAV) technologies, there is a growing consensus on integrating UAVs into cellular networks. It is expected that UAVs will play a prominent role for traffic offloading, capacity enhancement, and disaster recovery to assist terrestrial networks [1]–[3]. From an economic perspective, deploying small cell base stations (BSs) and/or advanced fifth generation (5G) components, such as massive multiple-input and multiple-output (MIMO), may not be cost-effective for big temporary events. In this regard, deployment of UAVs can be considered as an alternative or complement solution to compensate outage and alleviate overload conditions in cellular networks. UAVs are able to establish line-of-sight (LoS) communication links with high probability for ground users resulting in increased coverage, enhanced reliability, and agility [4].

Mobility of UAVs and their flexibility in adjusting their locations significantly impact the LoS probability and the network performance. Most research efforts have addressed this issue from a non-learning perspective [4]–[7]. In [4], the optimum altitude of a single UAV was obtained in order to minimize the outage probability and maximize the coverage region. In [5], the location of a UAV in one-dimensional and two-dimensional (2D) networks was investigated to maximize the average throughput. In [6], the authors developed an algorithm for the three-dimensional (3D) placement of a UAV to maximize the number of covered users in the network. For the sake of simplicity, the problem was decoupled to the placement of the UAV in the horizontal and vertical dimensions. In [7], the 3D placement and trajectory designs for a UAV to improve rate were investigated.

However, there has been a growing attention recently devoted to the use of learning algorithms for the UAV deployment problem [8]–[11]. In [8] and [9], learning based approaches to find the 2D trajectory of UAVs flying with fixed altitudes were proposed. In [10], a learning based approach for 3D placement of a single UAV is developed to maximize network throughput. However, these works do not address the 3D deployment of multiple UAVs integrated into already existing terrestrial networks for the purpose of network overload relief. In [11], the problem of 3D placement was dealt as a two separate optimization problems in horizontal 2D plane and altitude of UAVs, where a prior knowledge of the users locations is required. However, providing this information for UAVs in real time can be challenging. Moreover, the proposed algorithm is not able to adapt to the environment changes. Finally, to our best knowledge, the channel models in the previous learning based reports do not capture the dependency of path loss exponents to the height of UAVs which might have a significant influence on the performance of UAVs [4], [12].

In this paper, we address the 3D placement of multiple UAVs aiming at assisting the existing terrestrial cellular networks to improve the throughput in overloaded and outage scenarios. We formulate an optimization problem using frameworks from game theory, and leverage reinforcement learning algorithms to develop two novel dis-
A. Network Topology

We consider the downlink of a cellular network consisting of a set of terrestrial BSs $\mathcal{M}$, and a set of UAVs $\mathcal{U}$ as the aerial BSs, in a particular area $\mathcal{R} \subseteq \mathbb{R}^2$. The UAVs are deployed to support the cellular network in overload and outage situations. The set of total users and the set of users associated to BS $b \in \mathcal{B}$ at time $t$ are denoted by $\mathcal{K}$ and $\mathcal{K}_b(t)$, respectively, where $\mathcal{B} = \mathcal{M} \cup \mathcal{U}$ indicates the total BSs in the system. We assume that all the BSs transmit over the same channel (i.e. co-channel deployment), and the time horizon is discretized into $N$ equally spaced time instants with duration $T_s$. Let $\mathbf{a}_b(t) = (x_b(t), y_b(t), h_b(t))$ be the location of BS $b$ at time $t \in \{1, \ldots, N\}$, where $(x_b(t), y_b(t))$ and $h_b(t)$ are the location of BS $b$ in the horizontal dimension and its altitude at time $t$, respectively. Moreover, we assume that all the UAVs fly with a constant speed of $V$. An illustration of the system model is given in Fig. 1.

B. Radio Propagation and Signal Quality

We consider a channel model, in which the link between each user $k \in \mathcal{K}$ and each BS $b \in \mathcal{B}$ comprises LoS and non-LoS propagation conditions. Let $\Pr_{b,k}^\text{LoS}$ denote the probability of having a LoS link between user $k$ and BS $b$ which is determined as follows [12]:

$$\Pr_{b,k}^\text{LoS} = \prod_{n=0}^{\left\lfloor \frac{z_{b,k}(t)}{1000} \right\rfloor - 1} \left[ 1 - \exp\left( \frac{h_b(t) - (\alpha + \beta)(h_b(t) - h_k(t))}{2\sigma^2} \right) \right],$$

where $\alpha$, $\beta$ and $\xi$ represent statistical environment-dependent parameters. Here, $h_b(t)$ and $h_k(t)$ are the altitude of BS $b$ and the altitude of user $k$ at time $t$, respectively. The horizontal location of user $k$, and its horizontal distance to BS $b$ at time $t$ are denoted by $(x_k(t), y_k(t))$ and $r_{b,k}(t) = \sqrt{(x_b(t) - x_k(t))^2 + (y_b(t) - y_k(t))^2}$, respectively. Consequently, the non-LoS probability can be determined as $\Pr_{b,k}^\text{NLoS} = 1 - \Pr_{b,k}^\text{LoS}$.

Let $d_{b,k}(t) = \sqrt{r_{b,k}^2(t) + (h_b(t) - h_k(t))^2}$ denote the 3D distance between BS $b$ and user $k$ at time $t$. The path loss between BS $b$ and user $k$ can be expressed as [12]:

$$L_{b,k}^z(t) = A_b^z(d_{b,k}(t))^\delta_{b,k}^z,$$

where superscript $z \in \{\text{LoS}, \text{NLoS}\}$ denotes LoS and non-LoS components on the link. Here, $A_b^z$ and $\delta_{b,k}^z$ are the reference path loss and the path loss exponent, respectively. For small scale fading, we use Nakagami-$m$ fading which is a generalized model to characterize various fading environments. Therefore, the channel fading gains $\mathbf{Z}_z$ undergoing Nakagami fading is Gamma distributed, i.e. $\mathbf{Z}_z \sim \text{Gamma}(m_z, \frac{1}{m_z})$, where $z \in \{\text{LoS}, \text{NLoS}\}$. Here, $m_{\text{LoS}}$ and $m_{\text{NLoS}}$ denote the fading parameters for the LoS and non-LoS links, respectively.
Therefore, the probability density function (PDF) of the Gamma-distributed channel fading gain $Z$ is determined as follows:

$$f_{Z}(z) = \frac{m_z^{m_z} z^{m_z-1}}{\Gamma(m_z)} \exp(-m_z z), \quad z \in \{\text{LoS}, \text{NLoS}\}$$

where $\Gamma(.)$ is the Gamma function. We consider $m_{\text{NLoS}} = 1$ that provides Rayleigh distribution, and $m_{\text{LoS}} > 1$ that can approximate the Ricean fading distribution. Moreover, the case of no fading can be modeled by setting $m_{\text{LoS}} \to \infty$.

The signal to interference plus noise ratio (SINR) experienced by user $k$ associated with BS $b$ can be formulated as:

$$\gamma_{b,k}(t) = \frac{p_b g_{b,k}(t)}{\sum_{b' \in B \setminus \{b\}} p_{b'} g_{b',k}(t) + \sigma^2},$$

where $p_b$ is the transmit power of BS $b$, and $g_{b,k}(t)$ is the channel gain between BS $b$ and user $k$ at time $t$ which is determined as follows [12], [13]:

$$g_{b,k}(t) = x_{b,k}, \frac{Z_{\text{LoS}}}{L_{b,k}(t)} + (1 - x_{b,k}), \frac{Z_{\text{NLoS}}}{L_{b,k}(t)},$$

where $x_{b,k}$ is a Bernoulli random variable with parameter $p_{b,\text{LoS}}$ which characterizes the case of LoS or non-LoS propagation occurrence. Parameter $\sigma^2$ represents the additive white Gaussian noise (AWGN) power. Therefore, the achievable data rate provided by BS $b$ to user $k$ using Shannon’s capacity formula is given by:

$$C_{b,k}(t) = \omega \log_2(1 + \gamma_{b,k}(t)),$$

where $\omega$ is the total bandwidth.

### C. Load and User Association Policy

Let $v_k$ denote the traffic influx rate of user $k$, in which each user can have a different quality of service (QoS) requirement (i.e. heterogeneous users). The fraction of time BS $b$ requires to serve the traffic $v_k$ to the location of user $k$ is defined as $\frac{v_k}{C_{b,k}(t)}$. Therefore, the load of BS $b$ at time $t$ is given by [14]:

$$p_b(t) = \sum_{k \in K_b(t)} \frac{v_k}{C_{b,k}(t)}.$$  

Since the limited resources are available in the network, the loads of BSs cannot exceed one. Thus, if the load of a BS exceeds one, some of its users experience drops in their rates [15].

To associate the users to the BSs, we consider a user association rule based on the best signal strength. Accordingly, each user $k$ connects to the BS $b_k(t)$ which offers the highest received signal strength, as follows:

$$b_k(t) = \arg\max_{b \in B} \{p_b g_{b,k}(t)\}.$$  

### III. Learning Based Placement Algorithms

In this section, we aim at optimizing the 3D locations of the UAVs for maximizing the throughput, in a distributed manner. Let $A(t) = (a_1(t), \ldots, a_{|U|}(t))$ denote the locations of the UAVs at time $t$. Therefore, the optimization problem is formulated as follows:

$$\max_{A(t)} \sum_{t \in N} \sum_{b \in B} \sum_{k \in K_b(t)} C_{b,k}(t)$$

subject to:

$$h_u(t) \in [h_{\text{min}}, h_{\text{max}}], \quad \forall u \in U,$$

$$0 \leq \rho_b(t) \leq 1, \quad \forall b \in B,$$

where parameters $h_{\text{min}}$ and $h_{\text{max}}$ denote the minimum and maximum altitude of the UAVs, respectively. Here, $d_{u,u'}(t)$ and $d_{\text{min}}$ represent the 3D distance between UAV $u$ and UAV $u'$ at time $t$, and a certain minimum distance, respectively. For the optimization problem (9), the constraints in (9b)-(9e) define the feasible 3D space for the locations of the UAVs. The constraint in (9d) ensures the safety of UAVs and collision avoidance. The constraint in (9e) corresponds to the definition of load, in which it avoids outages and ensures service for the users [15].

The problem of finding optimum 3D locations of the UAVs is complex mainly due to the mobility of the UAVs and temporal traffic statistics. Therefore, we leverage the tools of machine learning to solve the problem. Hereby, each UAV determines its position based on the proposed learning approaches which are presented in Section III-A and III-B.

#### A. Multi-Armed Bandit Based UAV Placement Approach

In a MAB problem, each bandit (player) chooses an arm from a set of arms (set of actions). After choosing an arm, the bandit observes the reward associated to the arm [16]. The bandit has no initial knowledge about the rewards of the arms. Therefore, it is allowed to explore the different arms to improve the estimated reward, while exploits the arm that can maximize the immediate reward. Accordingly, the MAB approach requires to effectively balance the tradeoff between exploration and exploitation.

We model our problem as a MAB problem. The set of players, actions, and reward function are defined as follows:

- **Players**: the set of UAVs $U$ is considered as the set of players.
- **Actions**: the set of actions is denoted by $S_u$ which is defined as the movement in different directions, i.e. $S_u = \{\text{up, down, left, right, forward, backward, no change}\}$. 
- **Rewards**: the reward of the arm is defined as the throughput obtained by the arm. Therefore, the reward of the arm is given by:

$$r_u(t) = \sum_{k \in K_u(t)} C_{b_k(t),k}(t),$$

where $C_{b_k(t),k}(t)$ is the data rate provided by the BS $b_k(t)$ to user $k$ at time $t$. Therefore, the reward of the arm is given by:

$$r_u(t) = \sum_{k \in K_u(t)} C_{b_k(t),k}(t).$$

In a MAB problem, each bandit (player) chooses an arm from a set of arms (set of actions). After choosing an arm, the bandit observes the reward associated to the arm [16]. The bandit has no initial knowledge about the rewards of the arms. Therefore, it is allowed to explore the different arms to improve the estimated reward, while exploits the arm that can maximize the immediate reward. Accordingly, the MAB approach requires to effectively balance the tradeoff between exploration and exploitation.

We model our problem as a MAB problem. The set of players, actions, and reward function are defined as follows:

- **Players**: the set of UAVs $U$ is considered as the set of players.
- **Actions**: the set of actions is denoted by $S_u$ which is defined as the movement in different directions, i.e. $S_u = \{\text{up, down, left, right, forward, backward, no change}\}$. 
- **Rewards**: the reward of the arm is defined as the throughput obtained by the arm. Therefore, the reward of the arm is given by:

$$r_u(t) = \sum_{k \in K_u(t)} C_{b_k(t),k}(t),$$

where $C_{b_k(t),k}(t)$ is the data rate provided by the BS $b_k(t)$ to user $k$ at time $t$. Therefore, the reward of the arm is given by:

$$r_u(t) = \sum_{k \in K_u(t)} C_{b_k(t),k}(t).$$
• **Reward:** the reward of UAV $u$ can be expressed as follows [16] and [17]:

$$R_u(t) = \frac{\sum_{k \in \mathcal{K}_u(t)} C_{u,k}(t)}{\sum_{k \in \mathcal{K}_u(t)} w_k}$$  \hspace{1cm} (10)

One of the widely used solution method for MAB problems is upper confidence bound (UCB) algorithm. In this method, each player first selects each arm once. Then, for each $t > |\mathcal{S}_u|$ it selects the action $s_{\text{UCB}}^u(t)$ as follows [16]:

$$s_{\text{UCB}}^u(t) = \arg \max_{s_{u,i} \in \mathcal{S}_u} \left\{ \bar{R}_{u,i}(t) + c \sqrt{\frac{2 \ln t}{n_{u,i}(t)}} \right\},$$  \hspace{1cm} (11)

where $\bar{R}_{u,i}(t)$ denotes the mean reward of UAV $u$ for playing action $s_{u,i} \in \mathcal{S}_u$ at time $t$, and $n_{u,i}(t)$ is the number of times that action $s_{u,i}$ has been selected by UAV $u$ until time $t$. Parameter $c > 0$ balances the tradeoff between exploration and exploitation, in which high values of $c$ lead to a high level of exploration. The pseudocode for the UCB based approach is presented in Algorithm 1.

**B. Satisfaction Based UAV Placement Approach**

The 3D deployment problem of the UAVs can be formulated as a noncooperative game in satisfaction form which comprises a multi-step process. In a satisfaction-form game, each UAV is interested in the satisfaction of its constraints. In the first phase, UAVs try to be satisfied with the high satisfaction threshold. After a predefined time interval, unsatisfied players can decide to reduce their thresholds, and play the game with the redefined thresholds. A game in satisfaction form can be described by the following triplet [14], [18]:

$$\mathcal{G} = (\mathcal{U}, \{\mathcal{S}_u\}_{u \in \mathcal{U}}, \{c_u\}_{u \in \mathcal{U}}),$$  \hspace{1cm} (12)

where $\mathcal{U}$ and $\mathcal{S}_u$ are, respectively, the set of players and the set of actions similar to the definitions in Section III-A. The correspondence $c_u(s_{-u}) \subseteq \mathcal{S}_u$ denotes the set of actions that can satisfy the constraints of UAV $u$ given the actions played by all other UAVs, where $s_{-u}$ is the actions of all UAVs except UAV $u$. Therefore, the correspondence $c_u(s_{-u})$ can be defined as follows [18]:

$$c_u(s_{-u}) = \{s_{u,i} \in \mathcal{S}_u : R_u(t) \geq \kappa_u \},$$  \hspace{1cm} (13)

where $\kappa_u$ is the satisfaction threshold for UAV $u$. According to the observed utility, each UAV $u \in \mathcal{U}$ updates a satisfaction indicator $\vartheta_u(t)$ at time $t$ as follows:

$$\vartheta_u(t) = \begin{cases} 1, & \text{if } s_u(t) \in c_u(s_{-u}) \\ 0, & \text{otherwise} \end{cases},$$  \hspace{1cm} (14)

where $s_u(t)$ is the action played by UAV $u$ at time $t$. For a game in satisfaction form, an important outcome is called satisfaction equilibrium where all players are satisfied, and $c_u(s_{-u})$ is not empty for each player $u$. The notion of satisfaction equilibrium can be formulated as a fixed point as follows [18]:

**Definition 1 (Satisfaction Equilibrium):** An action profile $s^* = (s_u^*, s_{-u}^*)$ is a satisfaction equilibrium if $\forall u \in \mathcal{U}$, $s_u^* \in c_u(s_{-u}^*)$.

However, for a given satisfaction threshold, some UAVs may be faced with the situations where they can not be satisfied. Therefore, a satisfaction equilibrium may not always exist. In this context, we can reduce the satisfaction threshold $\kappa_u$ for unsatisfied UAVs after a certain time interval $\tau$. Therefore, we use an adaptive threshold approach. In this approach, after each time interval $\tau$, an unsatisfied UAV $u'$ decreases its satisfaction threshold $\kappa_u'$ to $\kappa_u'(1 - \delta)$, where $0 < \delta < 1$ denotes a constant coefficient used to decrease the satisfaction threshold.

To solve the game in a distributed manner, we use a learning algorithm, and assume that each UAV $u$ can observe its obtained utility. If the UAV is satisfied with its current utility (i.e. $\vartheta_u(t) = 1$), the UAV has no incentive to change its action. Otherwise, it may change its action according to the probability distribution $\pi_u(t) = (\pi_{u,1}(t), \ldots, \pi_{u,|\mathcal{S}_u|}(t))$, where $\pi_{u,i}(t)$ is the probability assigned to action $s_{u,i} \in \mathcal{S}_u$ with $i \in \{1, \ldots, |\mathcal{S}_u|\}$. Therefore, each UAV $u$ updates the probability assigned to each action $s_{u,i} \in \mathcal{S}_u$ as follows [14]:

---

**Algorithm 1: MAB based learning algorithm for the 3D deployment of UAVs**

1. **Initialization:** $\bar{R}_{u,i}(t) = 0, n_{u,i}(t) = 0$ for $t = 0, \forall u \in \mathcal{U}, \forall s_{u,i} \in \mathcal{S}_u$ and $i \in \{1, \ldots, |\mathcal{S}_u|\}$
2. while $t < N$ do
3. $t \leftarrow t + 1$
4. for each $u \in \mathcal{U}$ do
5. if $\exists s_{u,i} \in \mathcal{S}_u$ s.t. $n_{u,i}(t) = 0$ then
6. Select arm $s_{\text{UCB}}^u(t) = s_{u,i}$
7. else
8. Select arm $s_{\text{UCB}}^u(t)$ according to (11)
9. end if
10. Update the location $a_u(t)$ based on $s_{\text{UCB}}^u(t)$
11. Calculate $R_u(t)$ according to (10)
12. for $\forall s_{u,i} \in \mathcal{S}_u$ do
13. Update $n_{u,i}(t)$ as:
14. $n_{u,i}(t) = n_{u,i}(t-1) + 1_{\{s_{u,i} = s_{\text{UCB}}^u(t)\}}$
15. Update $\bar{R}_{u,i}(t)$ as:
16. $\bar{R}_{u,i}(t) = \frac{n_{u,i}(t-1) + 1_{\{s_{u,i} = s_{\text{UCB}}^u(t)\}} R_u(t)}{n_{u,i}(t)}$
17. end for
18. end for
19. end while
\[
\pi_{u,i}(t + 1) = \begin{cases} 
\pi_{u,i}(t), & \text{if } \vartheta_u(t) = 1 \\
L_u(\pi_{u,i}(t)), & \text{otherwise.}
\end{cases}
\] (15)

Here,

\[
L_u(\pi_{u,i}(t)) = \pi_{u,i}(t) + \mu_u(t)q_u(t)(1 - \pi_{u,i}(t)),
\] (16)

where \(\mu_u(t) = \frac{1}{100.0.t+1}\) is the learning rate of UAV \(u\). The parameter \(q_u(t)\) is computed as \(q_u(t) = R_{\text{max}} + R_{\text{max}}(1 - \pi_{u,i})\), where \(R_{\text{max}}\) is the maximum utility that UAV \(u\) can achieve.

The pseudocode for the proposed satisfaction approach is presented in Algorithm 2.

**Proposition 1:** The procedure described in Algorithm 2 converges to an equilibrium of the game \(G\) in finite time.

**Proof.** Since the set of players and the set of actions are finite, the convergence of the algorithm in finite time is ensured. Furthermore, due to the fact that for all \(u \in U\) and for all \(s_{u,i} \in S_u\), \(\pi_{u,i} > 0\), thus each action profile will be played at least once, and when an equilibrium is played by the players, the convergence is observed. \(\blacksquare\)

### IV. SIMULATION RESULTS

To evaluate the performance of our proposed approaches, we consider a hexagonal layout with radius 250 m, and one terrestrial BS located in the center of the area. The set of users are uniformly distributed in the area, and their locations are fixed during one simulation.

**Algorithm 2:** Satisfaction based learning algorithm for the 3D deployment of UAVs

1: **Initialization:** \(t = 0\), \(\pi_{u,i}(t) = \frac{1}{|S_u|}\), \(\vartheta_u(0) = 0\), \(\forall u \in U, \forall s_{u,i} \in S_u\) and \(i \in \{1, \ldots, |S_u|\}\)
2: **while** \(t < N\) **do**
3: \(t \leftarrow t + 1\)
4: **for each** \(u \in U\) **do**
5: **if** \(\vartheta_u(t - 1) = 1\) **then**
6: \(s_u(t) = s_u(t - 1)\)
7: **else**
8: Select a action \(s_u(t)\) based on \(\pi_u(t)\)
9: **if** \(\text{mod}(t, \tau) = 0\) **then**
10: \(\kappa_u \leftarrow \kappa_u(1 - \delta)\)
11: **end if**
12: **end if**
13: Update the location \(\alpha_u(t)\) based on \(s_u(t)\)
14: Calculate utility \(R_u(t)\) according to (10)
15: Calculate \(\vartheta_u(t)\) according to (14)
16: Update \(\pi_u(t)\) according to (15)
17: **end for**
18: **end while**

| Parameter | Value |
|-----------|-------|
| Carrier frequency \((f_c)\) / Channel bandwidth | 2 GHz / 10 MHz |
| Noise power spectral density | \(-174\) dBm/Hz |
| \(v_k\) | 1800 Kbps |
| \(f_b\) | 4 |
| \(d_{\text{min}}\) | 20 m |
| \(R_{\text{max}}\) | 22.5 m, 150 m |
| Altitude of users | 1.5 m |
| Altitude of terrestrial BS | 25 m |
| \(\alpha, \beta, \xi\) | 0.3, 500, 20 |
| \(R_{\text{max}}, \kappa_u\) | 1, 1 |
| \(\delta\) | 0.2 |
| \(c\) | 1 |
| \(T_c\) | 1 sec |
| \(N\) | 2700 |
| \(\tau\) | 100 |
| \(\nu\) | 10 m/s |

**Table I:** System-Level Simulation Parameters [12], [14]

In our simulations, we assume that the number of UAVs is 4, unless we mention it is set to be others values. We conduct multiple simulations for various configurations, and average over the independent runs. The simulation parameters are summarized in Table I. Furthermore, we demonstrate the performance gain of our proposed learning based schemes over the following benchmark references:

- **Blind placement approach:** the 3D locations of the UAVs are chosen randomly.
- **Q-learning based placement approach:** the UAVs update their locations based on the Q-learning approach proposed in [8]. It is assumed that the altitude of the UAVs is 100 m.

In the proposed approaches and Q-learning approach, the initial locations of the UAVs are chosen according to a heuristic approach, in which the horizontal location of a new UAV is determined from a predefined horizontal locations to achieve the furthest distances from the BSs in the system similar to [19].

Fig. 2 depicts the convergence behavior of the MAB and satisfaction based approaches in terms of average throughout during the simulations for a network with 500 users. As can be observed, the satisfaction based learning approach needs more iterations to converge...
compared to the MAB based learning approach. It shows that for the MAB based approach and the satisfaction based approach, the average number of iterations for convergence reaches up to about 450 and 800 iterations, respectively. Furthermore, the proposed satisfaction approach yields better performance compared to the MAB based algorithm. As a result, there is a tradeoff between convergence speed and performance.

Fig. 3 illustrates the average rate per user. We can observe that as the number of users increases, average rate per user decreases due to increasing load and the availability of limited resource in the network. Since the satisfaction based approach optimizes the 3D locations of UAVs in terms of satisfying UAV’s throughput, it improves the average rate of users compared to the benchmark algorithms. For instance, for a network with 400 users, the satisfaction based approach yields 26.25%, 49.28%, and 97.29% average user’s rate improvement compared to the MAB based algorithm, Q-learning, and blind approaches, respectively.

In Fig. 4, we show the average number of outage users (i.e. the users that their rates are less than \(v_k\)) per BS. The figure shows that the average number of outage users increases with an increase in the number of users. This is mainly due to the fact that, a limited resource is available in the network. Therefore, with increasing the number of users, some of them may experience reductions in their rates due to the overloaded BSs. Since the proposed satisfaction approach improves the average rate compared to the other approaches, it yields better performance in terms of average number of outage users. The performance gain of the satisfaction based approach compared to the MAB, Q-learning and blind approaches is up to 33.1%, 41.2%, and 48.9%, respectively, for a network with 250 users. Accordingly, the satisfaction approach is more resistant to the higher traffic demand.

Fig. 5 illustrates the impact of the number of UAVs on the performance of the system for the proposed and benchmark approaches. As can be seen, satisfaction based approach outperforms others, though the advantage over the MAB decreases as the number of UAVs increases. Moreover, one can see that for very...
low number of UAVs the Q-learning slightly performs better than MAB. This figure, in general, reveals that the number of required UAVs for a target performance decreases using learning methods, and hence the UAVs networking is even more cost-effective. For instance, to reduce the outage users to 100 with learning approach 2 UAVs are needed as compared to 3 UAVs when adopting a blind approach. Such benefit of learning improves with less number of tolerated outage users.

V. Conclusion

In this paper, we have proposed two novel algorithms for the placement of UAVs integrated into terrestrial cellular networks to compensate outage and overload. The proposed approaches, which are based on the MAB and satisfaction methods, leverage the tools from game theory and machine learning to learn the 3D locations of the UAVs. Our results have shown that the proposed approaches can significantly outperform the benchmark algorithms in terms of both throughput and outage users. Furthermore, our proposed approaches do not require full information of the network (e.g. the locations of the users), and they do not impose significant amount of information exchange in the network. Therefore, they can be executed in a distributed manner.

REFERENCES

[1] M. Mozaffari, W. Saad, M. Benni, Y. Nam, and M. Debbah, “A tutorial on UAVs for wireless networks: Applications, challenges, and open problems,” IEEE Commun. Surveys Tuts., vol. 21, no. 3, pp. 2334–2360, 2019.
[2] E. Vinogradov, H. Sallouha, S. D. Bast, M. M. Azari, and S. Pollin, “Tutorial on UAV: A blue sky view on wireless communication,” Journal of Mobile Multimedia, 2019. [Online]. Available: http://arxiv.org/abs/1901.02306
[3] Y. Zeng, Q. Wu, and R. Zhang, “Accessing from the sky: A tutorial on UAV communications for 5G and beyond,” Proceedings of the IEEE, vol. 107, no. 12, pp. 2327–2375, 2019.
[4] M. M. Azari, F. Rosas, K. Chen, and S. Pollin, “Ultra reliable UAV communication using altitude and cooperation diversity,” IEEE Trans. Commun., vol. 66, no. 1, pp. 330–344, 2018.
[5] Z. Wang, L. Duan, and R. Zhang, “Adaptive deployment for UAV-aided communication networks,” IEEE Trans. Wireless Commun., vol. 18, no. 9, pp. 4531–4543, Sep. 2019.
[6] M. Alzenad, A. El-Keyi, and H. Yanikomeroglu, “3-D placement of an unmanned aerial vehicle base station for maximum coverage of users with different QoS requirements,” IEEE Wireless Commun. Lett., vol. 7, no. 1, pp. 38–41, 2018.
[7] Y. Huang, W. Mei, J. Xu, L. Qu, and R. Zhang, “Cognitive UAV communication via joint maneuver and power control,” IEEE Trans. Commun., vol. 67, no. 11, pp. 7872–7888, 2019.
[8] B. Khamidehi and E. S. Sousa, “Reinforcement learning-based trajectory design for the aerial base stations,” in 2019 IEEE 30th Annual International Symposium on PIMRC, 2019, pp. 1–6.
[9] H. Dai, H. Zhang, M. Hua, C. Li, Y. Huang, and B. Wang, “How to deploy multiple UAVs for providing communication service in an unknown region?,” IEEE Wireless Commun. Lett., vol. 8, no. 4, pp. 1276–1279, Aug 2019.
[10] R. Ghanavi, E. Kalantari, M. Sabbughian, H. Yanikomeroglu, and A. Yongacoglu, “Efficient 3D aerial base station placement considering users mobility by reinforcement learning,” in 2018 IEEE WCNC, April 2018, pp. 1–6.
[11] H. El Hamrouni, M. Benjillali, B. Shihada, and M. Alouini, “Learn-as-you-fly: A distributed algorithm for joint 3D placement and user association in multi-UAVs networks,” IEEE Trans. Wireless Commun., vol. 18, no. 12, pp. 5831–5844, Dec 2019.
[12] M. M. Azari, G. Geraci, A. Garcia-Rodriguez, and S. Pollin, “UAV-to-UAV communications in cellular networks,” IEEE Trans. Wireless Commun., pp. 1–1, 2020.
[13] J. Lyu and R. Zhang, “Network-connected UAV: 3-D system modeling and coverage performance analysis,” IEEE Internet Things J., vol. 6, no. 4, pp. 7048–7060, 2019.
[14] A. Hajijamali Arani, M. J. Omidi, A. Mehbdoumi, and F. Adachi, “Minimizing base stations’ ON/OFF switchings in self-organizing heterogeneous networks: A distributed satisfactory framework,” IEEE Access, vol. 5, pp. 26267–26278, 2017.
[15] R. L. G. Cavelcante, S. Stojcic, J. Zhang, and H. Zhuang, “Low complexity iterative algorithms for power estimation in ultra-dense load coupled networks,” IEEE Trans. Signal Process., vol. 64, no. 22, pp. 6058–6070, 2016.
[16] S. Ali, A. Ferdowsi, W. Saad, and N. Rajatheva, “Sleeping multi-armed bandits for fast uplink grant allocation in machine type communications,” in 2018 IEEE Globecom Workshops (GC Wkshps), 2018, pp. 1–6.
[17] A. Fotouhi, M. Ding, L. Galati Giordano, M. Hassan, J. Li, and Z. Lin, “Joint optimization of access and backhaul links for UAVs based on reinforcement learning,” in 2019 IEEE Globecom Workshops (GC Wkshps), 2019, pp. 1–6.
[18] S. M. Perlaza, H. Tembine, S. Lasaulce, and M. Debbah, “Satisfaction equilibrium: A general framework for QoS provisioning in self-configuring networks,” in 2010 IEEE GLOBECOM 2010, Dec 2010, pp. 1–5.
[19] F. Lagum, I. Bor-Yaliniz, and H. Yanikomeroglu, “Strategic densification with UAV-BSs in cellular networks,” IEEE Wireless Commun. Lett, vol. 7, no. 3, pp. 384–387, 2018.