Inter-cell Interference Mitigation for Cellular-connected UAVs using MOSDS-DQN

Liyana Adilla binti Burhanuddin, *Student, IEEE*, Xiaonan Liu, *Student, IEEE,*, Yansha Deng, *Member, IEEE*, Maged Elkashlan, *Member, IEEE,* Arumugam Nallanathan, *Fellow, IEEE*.

***Abstract*—In 5G and beyond, UAVs are integrated into cellular networks as new aerial mobile users to support many applications and provide higher probability of line-of-sight (LoS) transmission to base stations (BSs). Nevertheless, due to limited frequency bandwidth and spectrum resource reuse when BSs serving terrestrial users (TUEs) and UAVs, it causes severe downlink interference to TUEs, especially when the network has a heavy load. Thus, in this paper, we study the performance of radio connectivity of UAVs and TUEs in an urban area and introduce a downlink inter-cell interference coordination mechanism. Then, we propose adaptive cell muting interference and resource allocation scheduling schemes. A value function approximation solution (VFA), Tabular-Q, and Deep-Q Network (DQN) are proposed to maximize the long-term network throughput of TUEs while guaranteeing the data rate requirements of UAVs. With increasing number of UAVs and TUEs and dynamic wireless environment, we further propose a Muting Optimization Scheme and Dynamic time-frequency Scheduling (MOSDS) algorithm to increase throughput and satisfactory level for both UAVs and TUEs. Simulation results show that the proposed algorithms achieve 80% performance improvement of throughput of UAV and TUE networks and mitigate the interference among them. Also, the proposed MOSDS-DQN shows 18% improvement compared to the DQN algorithm.**

***Index Terms*—Interference Management; UAV; Cellular Networks; Dynamic scheduling; Deep Reinforcement Learning.**

I. Introduction

Unmanned Aerial Vehicle (UAV) is becoming an important solution in the future cellular-connected networks to increase coverage of base stations (BSs) and support many applications, e.g., video streaming [1, 2]. Studies show that in 2029, the worldwide commercial UAV market will achieve 14 billion dollars [3] and these will lead to traffic congestion of communication between UAVs and cellular-connected ground users. Compared to ground users, the flying UAVs have higher altitudes and the channels between UAVs and BSs are usually line-of-sight (LoS) channels [1].

Current regulations in most countries limit UAV operations to the case in which there is Visual-Line-of-Sight (VLOS) between a UAV and its pilot. However, it is expected that Beyond-Visual-Line-of-Sight (BVLOS) operations will be allowed for extended range, if there is a reliable Command and Control (C2) link to the UAV. The C2 link is critical to safe operations for UAVs. Moreover, in cellular networks, the BS's inter-site distance (ISD) is designed according to ground level channel models and the density of Terrestrial users (TUEs) [2]. However, ISD is not optimized for UAVs in different propagation environments. As a result, the transmission performance of UAVs and TUEs is severely affected by interference among them, when BSs serving them in the same frequency simultaneously [4]. Using the model in [5], the study in [6] showed that highly loaded scenarios decreased UAV coverage due to high interference. In addition, the authors in [7] gave theoretical interference analysis of cellular-connected UAV networks with TUEs based on radio characteristics, including UAVs' heights, ISD and signal-to-interference ratio level.

Consequently, authors in [4, 8–11] considered interference mitigation schemes between TUEs and UAVs, by considering power control [8–11], reducing UAV height [11], and antenna beam selection [4]. Although decreasing power allocation, reducing UAV height, and selecting proper antenna beams can mitigate interference and improve throughput, they can result in low coverage of UAVs and increase outage probability when BSs serving UAVs and TUEs simultaneously. To address this issue, authors in [3, 12] designed a cooperative beamforming technique to effectively suppress inter-cell interference (ICI) to the UAV, and authors in [2] deployed a muting scheme to mute the cells with high interference to decrease interference between UAVs and TUEs.

Furthermore, when large number of moving TUEs and UAVs exist in 5G networks, there will be high interference when BSs serving them. In [13–15], the authors considered cell muting and traditional optimization methods to mitigate the ICI. Specifically, authors in [13] optimized UAV resource allocation based on their cell association to maximize throughput performances of TUEs and UAVs, and considered inter-cell interference coordination (ICIC) based on Release-10/11 to mitigate strong interference to TUEs. However, only a single UAV was considered in [13] and the UAV could only access to the resource block (RB) that had not been occupied by any TUEs, thus, the approach in [13] could not be adapted to the scenario with multiple UAVs. Authors in [14, 15] used the cell range expansion (CRE), enhanced inter-cell interference coordination (eICIC), and further-enhanced ICIC (feICIC) schemes to improve the overall spectral efficiency. However, the optimization methods in [14] and [15] aimed at optimal solutions in each time slot with high computation

L. A. B. Burhanuddin and Y. Deng are with Department of Engineering, King's College London, London, UK (e-mail: liyana.burhanuddin@kcl.ac.uk, yansha.deng@kcl.ac.uk).

X. Liu, M. Elkashlan and A. Nallanathanare with School of Electronic Engineering and Computer Science, Queen Mary University of London, London, U.K (e-mail: x.l.liu@qmul.ac.uk; maged.elkashlan@qmul.ac.uk; a.nallanathan@qmul.ac.uk).

This work was supported in part by Engineering and Physical Sciences Research Council (EPSRC), U.K., under Grant EP/W004348/1.

Corresponding author: Yansha Deng

complexity and were not designed for long-term optimization problem.

To solve the problems in [13–15], authors in [16] proposed an interference-aware path planning scheme for cellular-connected UAV. Through deep reinforcement learning (DRL), each UAV is required to make a trade-off between maximizing energy efficiency and minimizing wireless latency and interference. With increasing number of devices in future terrestrial networks, the interference problem between UAVs and TUEs becomes more complicated. Therefore, efficient ICIC designs are required for enabling efficient spectrum sharing between UAVs and TUEs in future cellular-connected networks, in which, the resource allocation can be designed to mitigate interference and improve throughput of UAVs and TUEs. Based on the previous RB allocation and traffic patterns, authors in [17] proposed a deep Q-network (DQN) to select proper RBs for UAVs and TUEs to perform transmission with low interference. Authors in [8] deployed DRL algorithms, including DQN and actor-critic (AC), to co-design the video resolution, movement, and power control of UAV-BS and UAV-UEs to maximize the quality of experience (QoE) of real-time video streaming.

Although 5G helps improve data rate performance, it has some drawbacks. Therefore, more 5G BSs are built to support 5G connectivity in multiple areas and brought BSs closer to users [18]. With the increase in the number of 5G BSs and TUEs, the interference among them increases, and with the increase in transmission opportunity, the situation becomes more complex to reduce interference while guaranteeing high quality of service (QoS) of UAVs and TUEs. To mitigate interference in complex scenarios, the DRL algorithm is considered.

To the best of our knowledge, none of these studies investigated multiple UAVs and TUEs' coordination in the cellular network and deployed dynamic RB scheduling to maximize long-term throughput performance. In practice, the control signal reception of UAVs is not only affected by the link quality of the communication channel, but also susceptible to interference. Thus, the control links between the BS and TUEs and UAVs are important, especially when the spectrum resources are constrained. To effectively solve the aforementioned problems, UAVs and TUEs require high-level coordination to ensure all users meet their minimum requirements and optimize their data-rate performance, especially in a highly dynamic environment. Therefore, in this paper, we propose a dynamic muting and RB allocation scheme to maximize the throughput of TUEs via DRL algorithms. The interference is decreased by muting the cells with the strongest interference and RBs are properly shared and scheduled to UAVs and TUEs, with the aim to satisfy high QoS requirements of UAVs and TUEs. The main contribution of this paper is that we propose dynamic cell muting by using DQN algorithm to maximize long-term reward of downlink transmission with moving UAVs and TUEs in a downlink scenario of a connected-cellular UAV network. Our results show that our proposed MOSDS-DQN improves the throughput performance of TUEs and reduce overall interference, compared with the downlink inter-cell interference coordination mechanism proposed in [2]. The contributions of this paper are summarized as follows:

- We propose a dynamic muting scheme for moving UAVs and terrestrial users (TUEs) in a downlink scenario of a cellular network. The UAVs and TUEs are uniformly distributed in the communication environment, and the dynamic requests from them follow Poisson process in each time slot.
- To guarantee excellent service among TUEs in a dynamic network, we formulate a long-term problem to mitigate the interference level of each UAV by muting cells, which can satisfy QoS requirements of TUEs and UAVs over time and maximize sum-rate of TUEs.
- To further increase the throughput of downlink transmission based on cell muting technique, we propose a dynamic muting and time-frequency scheduling algorithm. The muting scheme mutes proper number of interfering cells, and the time-frequency scheduling scheme allocates proper physical resource blocks (PRBs) to TUEs and UAVs.
- To solve the aforementioned problem, we deploy value function approximation solution (VFA), Tabular-Q, Deep Q Network (DQN), and MOSDS-DQN. Learning algorithms help the agent to select actions to maximize the long-term throughput of downlink transmission. The linear muting scheme from [2] is set as a benchmark as it using traditional optimization muting scheme to mitigate the inter-cell interference. Simulation results show that our proposed DQN approach outperforms the linear muting scheme in terms of higher throughput and lower interference. Furthermore, the proposed MOSDS-DQN guarantees the throughput performance of TUEs with increasing number of UAVs.

The rest of this paper is organized as follows. The system model and problem formulation are given in Section II. The optimization problem via deep reinforcement learning is presented in Section III. Simulation results and conclusions are presented in Sections IV and V, respectively.

II. System Model

We assume that C base stations (BSs) with M antennas are deployed at the centre of C cells, using Orthogonal Frequency Division Multiple (OFDM) to serve their associated users, as shown in Fig. 1. OFDM has been used for over a decade and proved its robustness in multi-carrier technologies. In OFDM, the available frequency band is divided into multiple subcarriers, each of which is assigned to a specific TUE or UAV in a specific time slot. This allows multiple users to transmit or receive data simultaneously over different subcarriers without interfering with each other. OFDM uses multiple smaller subcarriers to avoid the Inter-Channel Interference and Inter-Symbol Interference over wireless networks, and adds a Cyclic Prefix (CP) to demodulate the signal effectively on the receiver side [19].

According to [20–22], as shown in Fig. 2, the antenna element pattern $A(\theta, \phi)$ for the mth antenna array is given

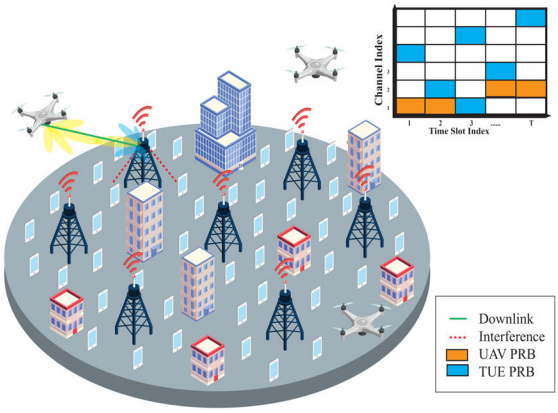

Fig. 1. Illustration of UAV-cellular network model and resource block scheduling. The solid green lines denote the signal links between BSs and their associated UEs, while the dashed red lines denote (strong) interference links between adjacent UAV/UEs.

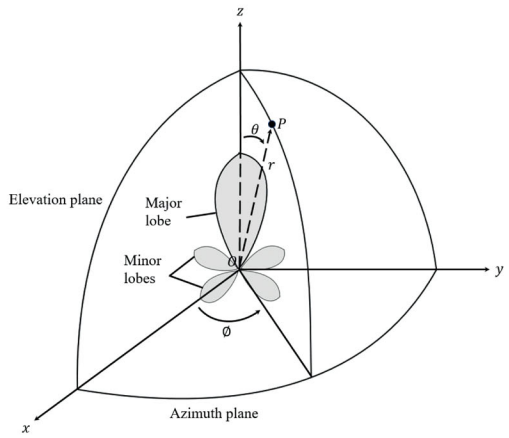

Fig. 2. Illustration of antenna pattern.

by

$$A(\theta,\phi) = -\min\left\{-\left[A_{E,V}(\theta)+A_{E,H}(\phi)\right], A_m\right\}, \tag{1}$$

where $A_{E,V}(\theta)$ and $A_{E,H}(\phi)$ are vertical and horizontal radiation patterns of antenna elements, respectively. $A_{E,V}(\theta)$ is denoted as

$$A_{E,V}(\theta) = -\min\left\{12\left(\frac{\theta-90^o}{\theta_{3dB}}\right)^2, SLA_V\right\}. \tag{2}$$

CIn Eq. (1) and (2), θ is the vertical angle, θ_{3dB} is the vertical 3dB beamwidth [20], SLA_V is the side-lobe level limit, and $A_{E,H}(\phi)$ is denoted as

$$A_{E,H}(\phi) = -\min\left\{12\left(\frac{\phi}{\phi_{3dB}}\right)^2, A_m\right\}. \tag{3}$$

In (3), ϕ is the horizontal angle, ϕ_{3dB} is the horizontal 3dB beamwidth, and A_m is the front-back ratio [20]. Based on (2) and (3), the 3D antenna element gain for each pair of angles (θ,ϕ) is calculated as

$$A_G(\theta,\phi) = G_{max} - \min\left\{-\left[A_{E,V}(\theta)+A_{E,H}(\phi)\right], A_m\right\}, \tag{4}$$

where G_{max} is the maximum directional gain of the antenna element [20, 22, 23]. The above equations (1) - (4) provide the dB gain experienced by a ray with angle pair (θ,ϕ) based on the effect of the element radiation pattern. The cth BS ($c \in \mathcal{C}$) operates in a single-user mode serving either a terrestrial UE (TUE) with DL data or an UAV with Command and Control (C2) data. Both TUEs and UAVs are assumed to be equipped with a single antenna. Each cell consists of I uniformly distributed TUEs, while the total number of UAVs is J and they are uniformly distributed over the entire network with radius R_{NW}. The UAV in a cell is prioritized and assigned with PRBs, as it requires critical C2 data transmitted in a required data rate [21]. The distribution of TUEs is modelled as Poisson Process and the remaining available PRBs should be allocated to all TUEs.

A. Mobility Model

In the 3D environment, we assume that the UAV flies at a fixed height with a fixed speed. Thus, the 3D environment for the UAV is mapped into a 2D image with $W \times W$ grids. We assume that the length of the side of each grid is a, and the UAV moves along the centre of each grid, which creates a finite set of possible paths. Also, the moving latency between two grids is the same because of the fixed speed. The UAV moves with four directions, right, left, forward, and backward.

B. Channel Model

We adopt two different 3GPP standards to model the channels for TUEs and UAVs, respectively [21] [22]. Here, the UAV is assumed to be flying at a height where a line-of-sight (LoS) link is ensured. The small-scale channel gain of the UAV is modelled as Rician channel, while for TUEs, they are modelled as Rayleigh channels [16, 22, 24]. The pathloss from the uth UE to the BS is written as

$$PL^t_{\text{LoS,u}} = \begin{cases} 15.3 + 37.6\log_{10}(d_{3D}), & 1.5m \le h^t_u \le 22.5m \\ 28.0 + 22\log_{10}(d_{3D}) + 20\log_{10}(f_c), & 22.5m < h^t_u \le 300m \end{cases} \tag{5}$$

where f_c is the carrier frequency, and $d^u_{3D}(t)$ is the distance between the uth user and the BS. Next, we assume that each BS uses the same transmit power and each user has perfect knowledge of its channel state information (CSI), so that the signal to interference plus noise ratio (SINR) $\gamma_{c,j}$ between the cth BS and the jth UAV is written as

$$\gamma_{c,j} = \frac{P_c \|\boldsymbol{h}^H_{c,j} \cdot \mathbf{v}_{c,j}\|^2}{N_{c,j} + \sum_{k \in C \setminus c} P_c \|\boldsymbol{h}^H_{k,u} \cdot \mathbf{v}_{k,u}\|^2}, \tag{6}$$

where $P_c = \frac{P_{DL}}{10^{PL_{LOS,k}/10}} \times A_G(\theta,\phi)$, and P_{DL} is the downlink transmit power per PRB. In (6), $\boldsymbol{h}_{c,j} \in \mathbb{C}^{M\times 1}$ denotes the channel vector between the cth BS and the jth UAV, and $\boldsymbol{h}_{k,u}$ is the channel between the kth BS and the uth user in the kth cell. The channel model $\mathbf{h}$ includes both the small-scale fading and large-scale fading calculated by $h = g.\beta^{1/2}$, where g and β are small-scale fading and large-scale fading parameters, respectively. In (6), $\mathbf{v}_{k,u} = \left(\mathbf{g}_{k,u}\right)^{\mathrm{H}} \left(\mathbf{g}_{k,u}\left(\mathbf{g}_{k,u}\right)^{\mathrm{H}}\right)^{-1}$ represents the transmit zero-forcing precoding vector of the uth user in the kth BS [25], and $\boldsymbol{g}_{k,u} \in \mathbb{C}^{M\times 1}$ is the channel vector between the kth BS and the uth user in the kth cell. In addition, $N_{c,j}$ is the additive white Gaussian noise at the jth user.

C. User Association

According to [6], we consider the maximum Reference Signal Receive Power (RSRP) in the user association policy. The RSRP is the average power of Resource Elements (RE) that carries cell specific Reference Signals (RS) over the entire bandwidth [2, 25], thus, the RSRP is only measured in the symbols carrying RS and is denoted as

$$RSRP_{c,u} = P_c - PL_{LoS,k}. \tag{7}$$

In the maximum RSRP-based user association, the UAV is connected to the BS that provides the maximum RSRP. Specifically, the associated BS is chosen by the UAV via

$$u_j = \left\{u \mid \max RSRP_{c,u}, \forall j \in J\right\}. \tag{8}$$

Based on (6), the achievable rate of the jth UAV is calculated as

$$R_{c,j}^{UAV} = B_c \log_2(1+\gamma_{c,j}), \tag{9}$$

where B_c is the bandwidth of the cth BS.

D. Inter-Cell Interference Coordination (ICIC) for Macrocell Muting

To improve received SINR of the UAV, the BSs coordinate PRBs among TUEs and UAVs. The interfering BS $c \in C'$ leaves the PRBs blank/muted, allowing the UAV-serving BS to schedule its transmission within the same frame shared with TUEs. Therefore, Eq. (6) is rewritten as

$$\gamma_{c,j} = \frac{P_c \|\boldsymbol{h}_{c,j} \cdot \mathbf{v}_{c,j}\|^2}{N_{c,j} + \sum_{k\in C'\setminus c, k\notin C'} P_c \|\boldsymbol{h}_{k,u}^H \cdot \mathbf{v}_{k,u}\|^2}, \tag{10}$$

where $C' \subseteq \mathcal{C}$ is the set of BSs being muted, and the second term in the denominator is the total interference from other BSs. The SINR between the cth BS and TUE is given by

$$\gamma_{c,u} = \frac{P_c \|\boldsymbol{h}_{c,u}^H \cdot \mathbf{v}_{c,u}\|^2,}{N_{c,u} + \sum_{k\in C'\setminus c, k\notin C'} P_c \|\boldsymbol{h}_{k,u}^H \cdot \mathbf{v}_{k,u}\|^2}. \tag{11}$$

In (11), $\boldsymbol{h}_{c,u} \in \mathbb{C}^{M\times 1}$ denotes the channel vector between the cth BS and the uth user. Based on Eq. (10) and (11), the date rate of the uth TUE is defined as

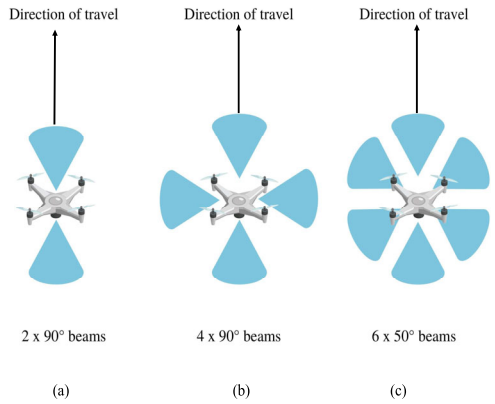

Fig. 3. Modelled antenna beam configurations for the UAV.

$$R_{c,u}^{TUE} = B_c \log_2(1+\gamma_{c,u}), \tag{12}$$

where B_c is the bandwidth of the cth BS and the data rate of the jth UAV is defined as

$$R_{c,j}^{UAV} = B_{c,j} \log_2(1+\gamma_{c,j}). \tag{13}$$

In (13), $B_{c,j}$ is the fraction of bandwidth allocated to user j at cth BS, with respect to the available bandwidth.

E. Antenna Beam Selection

When assuming the antenna elements are mounted on the UAV at the right spacing and angle/orientation, antenna selection with two or more directional antenna elements is equivalent to a simple beam selection [2]. For example, the UAV rotates its fuselage in the azimuth plane while keeping the right direction, then 1 or 2 antenna elements are sufficient to generate a 'beam' towards the serving BS. If degrees of freedom of the UAV are more restricted, at least 4 antenna elements need to be mounted to provide four beams in the azimuth plane. Thus, we assume that antenna beam selection of the UAV is applied only in the azimuth plane, and an omnidirectional elevation radiation pattern is considered. An antenna beam radiation pattern is modelled as a $sinc()^2$ function, with -3 dB beam-widths of approximately 90°, or 50° in the azimuth plane with six beams [2]. The modelled beam patterns provide +6.6 dBi gain in the main direction and -3 dB gain in the front-to-side lobe attenuation according to [2], which can be used to compensate for the non-ideal orientation and shape of beams [2]. As shown in Fig. 3, a simple setup with a grid of 2, 4 or 6 fixed beams is used (fixed relative to the UAV fuselage) to emulate a practical antenna selection mechanism.

F. Downlink Resource Block Scheduler

LTE transmission is segmented into frames, each one consists of 10 subframes, and each subframe is further divided into two slots. Each slot is 0.5 ms, so that the total time for each frame is 10 ms. Each time slot on the LTE downlink system

consists of 7 OFDM symbols. The flexible spectrum allows the LTE system to use bandwidths ranging from 1.4 MHz to 20 MHz, where higher bandwidths are used for higher LTE data rates. The physical resources of the LTE downlink can be illustrated as a frequency-time resource grid, as shown in Fig. 4. A Resource Block (RB) has a duration of 0.5 ms (one slot) and a bandwidth of 180 kHz (12 sub-carriers). Each RB has 84 resource elements in the case of a normal cyclic prefix and 72 resource elements for extended cyclic prefix.

In the RB scheduler technique, there are several types of scheduling algorithms, such as Round Robin (RR) [26] and proportional fairness (PF) [27]. RR scheduling is a non-aware scheduling scheme that allows users to take turns in using the shared resources (time-RBs), without taking the instantaneous channel conditions into account. The radio resources in RR are assigned equally among all users, which compromises the throughput performance of the system. While PF is defined as the ratio of the average data rate to all users to maintain the equality of fairness [27]. To solve the problem, dynamic scheduling is introduced to schedule the available data for each Transmission Time Interval (TTI), which maximizes the scheduling gains. As shown in Fig. 4, PRBs are allocated to sub-bands according to their channel and resource allocation models. However, the challenge is that it requires frequent coordination for exchanging control signals between cells, which increases the overhead among cells.

G. Problem Formulation

The objective is to maximize the throughput of TUE networks by selecting optimal actions in A^t subject to the UAV's QoS requirements (i.e., reliability). Thus, the optimization problem is formulated as

$$(P1): \max_{\pi(A^t|S^t)} \sum_{i=t}^{\infty} \sum_{c=1}^{C} \sum_{u=1}^{U} \beta^{(i-t)} R_{c,u}^{TUE}(i,f) \tag{14}$$

$$\text{s.t.} \sum_{f=1}^{F} p_f \leq P_c, \quad p_f \geq 0, \tag{15}$$

$$R_{c,j}^{UAV}(i,f) \geq R_{Th}^{UAV}, \quad \forall j \in \mathcal{J} \tag{16}$$

$$R_{c,u}^{TUE}(i,f) \geq R_{Th}^{TUE}, \quad \forall u \in \mathcal{U} \tag{17}$$

$$\beta^i \in [0,1]. \tag{18}$$

where $\beta^i \in [0,1)$ is the discount factor determining the performance accumulated in the future reward. When $\beta^i = 0$, the agent only concerns about the immediate reward. Eq. (15) guarantees the maximum transmit power threshold at the BS in each (i,f), where f is the selected sub-frames of F sub-frames and Eq. (16) and Eq. (17) guarantee the transmission rate threshold for UAVs and TUEs, respectively. The optimization problem aims at maximizing the total long-term reward in continuous time slots with respect to the policy π that maps the current state information s_t to the probabilities of selecting possible actions in A^t. The state S^t contains the set of instantaneous and cumulative data rates of both UAVs and TUEs, and the agent selects a specific action $A^t \in \mathcal{A}(S^t)$ that determines the index of cell(s) being muted.

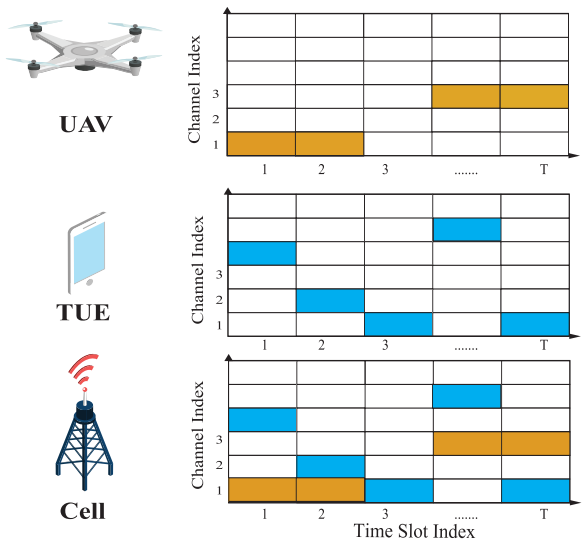

Fig. 4. Dynamic PRB scheduling for UAVs and TUEs.

To solve the problem, we consider muting schemes using reinforcement learning (RL), which are introduced in detail in the next section.

III. Muting Optimization Scheme using Reinforcement Learning

Since the channel and locations of UAVs and TUEs change over time, different muting and dynamic scheduling schemes are required in continuous time slots. Therefore, the problem in P1 is a long-term problem and cannot be solved by a traditional optimization method.Therefore, in this section, we design several Reinforcement Learning (RL) algorithms to solve the problem in P1. The RL agent learns the optimal mapping from the input states to select the resource allocation action to maximize the long-term throughput.

A. Tabular Q-Learning

Consider a Q-agent deployed at the central unit to optimize the service provision for both UAVs and TUEs. To optimize the long-term reward, the agent first explores the environment. Let $s \in \mathcal{S}$, $a \in \mathcal{A}$, and $r \in \mathcal{R}$ denote the state, action, and reward, respectively.

1) State Representation: The current state S^t corresponds to a set of current observations. The state of the system is denoted as $\mathcal{S} = [\sum R_{TUE}, \sum R_{UAV}]$, where R_{TUE} is a set of data rate of TUEs and R_{UAV} is a set of instantaneous rate of UAVs.

2) Action Space: Q-agent selects action A from set $\mathcal{A}$. The action is denoted as $A^t = \{N_m\}$, where N_m is the index of muting cells. To ensure the balance of exploration and exploitation actions of the agent, ϵ-greedy ($0 < \epsilon \leq 1$) exploration is deployed. In the tth TTI, the agent randomly generates a probability p_ϵ^t to compare with ϵ. If the probability $p_\epsilon^t < \epsilon$, the algorithm randomly selects an action from the feasible actions to improve the value of the non-greedy

action. However, if $p_\epsilon^t \geq \epsilon$, the algorithm exploits the current knowledge of the Q-value table to choose the action that maximizes the expected reward.

3) Rewards: At the beginning of each TTI, the Q-agent observes the current state S^t and selects a specific action $A^t \in \mathcal{A}$. After performing the selected action A^t, the agent receives a reward R^{t+1} and observes a new state S^{t+1}. The optimization goal is to maximize the long-term throughput of TUEs while guaranteeing the quality of service (QoS) of UAVs, which is defined as:

$$Reward_t^{p1} = \sum_{u=1}^{U} R_u \cdot \mathbb{1}\left[R_{u,j} \geq R_{Th}\right], \tag{19}$$

where,

$$\mathbb{1}\left[R_{u,j} \geq R_{Th}\right] = \mathbb{1}\left[R_u \geq R_{Th}\right] \cap \mathbb{1}\left[R_j \geq R_{Th}\right]. \tag{20}$$

In Eq. (20), $\mathbb{1}[x]$ is the indicator function, $\mathbb{1}[x] = 1$ when x is true, otherwise, $\mathbb{1}[x] = 0$, and $\cap$ is a logical and operation function. In the logical and operation function, $\mathbb{1}[x] \cap \mathbb{1}[y] = 1$ as x and y are true, otherwise, $\mathbb{1}[x] \cap \mathbb{1}[y] = 0$ [28].

In tabular Q, the state to action mapping is learned through value function $Q(s,a)$, which consists of a scalar value for all state and action spaces. The action that has the maximum value is selected from $\mathcal{A}$. To dynamically optimize the number of muted cells, the function learns the optimal policy π^* and optimizes the Q-table. The agent updates its Q-table using the immediate reward $\mathcal{R}^{t+1}$ and the next state-action value $Q(S^{t+1}, a)$, which is given by

$$\begin{aligned} &Q(S^t, A^t) = Q(S^t, A^t) + \\ &\alpha\left[R^{t+1} + \gamma \max_{a \in \mathcal{A}} Q(S^{t+1}, a) - Q(S^t, A^t)\right]. \end{aligned} \tag{21}$$

In Eq. (21), $\alpha \in (0,1)$ is the learning rate, and $\gamma \in [0,1)$ is the discount rate that determines how much the current reward affects the future value. In each TTI, the agent selects the action with the highest probability with probability $p_e^t \geq \epsilon$, or vice versa. The learning rate α, most importantly, is set to be a small constant to guarantee stable convergence, as the reward can be biased due to unknown and unpredictable distribution of the observed states. The implementation of cell muting using tabular-Q method is shown in Algorithm 1.

Algorithm 1 : Tabular Q-Learning/Linear VFA to optimize cumulative terrestrial users' throughput

Algorithm hyperparameters: $\alpha \in (0,1], \gamma \in [0,1), \epsilon \in (0,1]$
Tab-Q: Initialize Q-table $Q(s, a)$ **VFA:** Initialize $\mathbf{w}$
for *Iteration* $\leftarrow 1$ to I **do**
  Initialize s^1 by executing a random action A^0;
  UAVs identify the BS with the highest RSRP and associate with it.
  for $t \leftarrow 1$ to T **do**
    if $p_\epsilon < \epsilon$
      Randomly select an action A^t from $\mathcal{A}$;
    else
      Tab-Q: select $A^t = \underset{A \in \mathcal{A}}{\text{argmax}} Q\left(S^t, A\right)$;
      VFA: select $A^t = \underset{A \in \mathcal{A}}{\text{argmax}} Q\left(S^t, A, , \mathbf{w}\right)$;
    The agent performs A^t and mutes the selected cells.
    The agent observes S^{t+1} and calculates R^{t+1} using Eq. (19).
    Tab-Q: Update $Q(S, A)$ according to Eq. (21).
    VFA: Update $\mathbf{w}$ according to Eq. (29).
  end for
  Determine all active UEs (TUEs and UAVs) using Bernoulli process.
  Determine associate cell UEs and active UEs matrices.
  Update the queue matrices.
  Calculate SINR and transmission rate.
end for

B. Linear Value Function Approximation

However, the Tabular-Q requires large space to store state-action value, and needs to update each parameter to achieve convergence. To address these issues, we consider a linear value function approximation (VFA) method. VFA uses a 'Value Function' approximator to obtain a sub-optimal policy, but its efficiency depends on the deployed approximation function, such as Linear Approximator (LA), Deep Q-Learning (DQN), and decision trees.

LA approximates the value function $Q(S^t, A^t)$ by

$$Q(S^t, A^t) \approx \hat{Q}(S^t, A^t, \mathbf{w}^t), \tag{22}$$

where $\mathbf{w}^t$ is the weight vector. The objective is to minimize the mean-squared error between these two values, given by

$$J(\mathrm{w}^t) = \mathbb{E}_\pi[(Q(S^t, A^t) - \hat{Q}(S^t, A^t, \mathbf{w}^t))^2]. \tag{23}$$

To obtain the optimal policy, $\mathbf{w}^t$ is updated by stochastic gradient descent (SGD), which is calculated as

$$\begin{aligned} -\frac{1}{2}\nabla_{\mathbf{w}} J(\mathbf{w}^t) = &[Q(S^t, A^t) - \hat{Q}(S^t, A^t, \mathbf{w}^t)] \\ &\nabla_{\mathbf{w}} \hat{Q}(S^t, A^t, \mathbf{w}^t), \end{aligned} \tag{24}$$

and

$$\nabla \mathbf{w}^t = \alpha[Q(S^t, A^t) - \hat{Q}(S^t, A^t, \mathbf{w}^t)] \nabla_{\mathbf{w}} \hat{Q}(S^t, A^t, \mathbf{w}^t). \tag{25}$$

In LA, $\hat{Q}(S^t, A^t, \mathbf{w}^t)$ is represented as a dot product of feature vector $\mathbf{x}(S^t, A^t)$ and weight vector $\mathbf{w}^t$, which is denoted as

$$\hat{Q}(S^t, A^t, \mathbf{w}^t) = \mathbf{x}^T(S^t, A^t)\mathbf{w}^t = \sum_{k=1}^{K} \mathrm{x}_k(S^t, A^t)\mathrm{w}_k^t, \tag{26}$$

and

$$\mathbf{x}(S^t, A^t) = \begin{bmatrix} \mathrm{x}_1(S^t, A^t) \\ \mathrm{x}_2(S^t, A^t) \\ \vdots \\ \mathrm{x}_K(S^t, A^t) \end{bmatrix}, \quad (27)$$

where $\mathbf{x}(S^t, A^t)$ corresponds to the entire state-action space. The current action is selected from the vector $\hat{Q}(S^t, A^t, \mathbf{w}^t)$ in Eq. (26), following the ϵ-greedy policy, which is the same as that of tabular Q-learning. The gradient descent in Eq. (25) is calculated as

$$\nabla_{\mathbf{w}} \hat{Q}(S^t, A^t, \mathbf{w}^t) = \nabla_{\mathbf{w}} [\mathbf{x}^T(S^t, A^t)\mathbf{w}^t] = \mathbf{x}(S^t, A^t), \quad (28)$$

and

$$\nabla \mathbf{w} = \alpha [Q(S^t, A^t) - \hat{Q}(S^t, A^t, \mathbf{w}^t)]\mathbf{x}(S^t, A^t). \quad (29)$$

The implementation of cell muting using the VFA method is shown in Algorithm 1. However, the basic linear tabular-Q is not suitable, as the state-action space is so large and are increasing with the number of cells, and also function approximation technique is unable to train and get the optimal solution. Therefore, we consider DQN in our scenario.

C. Deep Q-Network

When large number of cells, TUEs, and UAVs exist in the network, the state-action space increases exponentially. To address this issue, DQN is used to update the network's weights. Just like LA, it also changes the value function $Q(S, A)$ into $Q(S, A, \boldsymbol{\theta})$, and $\boldsymbol{\theta}$ is the weight matrix of the multi-layer Deep Neural Network (DNN). DNN is used to approximate the state-action value function [29]. S^t is the state observed by the agent and acts as an input to DNN. The outputs are selected actions in $\mathcal{A}$. Furthermore, the intermediate layer contains multiple hidden layers and is connected to Rectifier Linear Units (ReLu) via using a $f(x) + max(0, x)$ function, and the output layer performs the linear activation to select actions from $\mathcal{A}$. In the tth time slot, the weight vector is updated by SGD and Adam Optimizer, which is expressed as

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^t - \lambda_{\mathrm{ADAM}} \cdot \nabla \mathcal{L}(\boldsymbol{\theta}^t), \quad (30)$$

where λ_{ADAM} is the Adam learning rate, and $\lambda_{\mathrm{ADAM}} \cdot \nabla \mathcal{L}(\boldsymbol{\theta}^t)$ is the gradient of the loss function $\mathcal{L}(\boldsymbol{\theta}^t)$. In (30), $\nabla \mathcal{L}(\boldsymbol{\theta}^t)$ is denoted as

$$\nabla \mathcal{L}(\boldsymbol{\theta}^t) = \mathbb{E}_{S^i, A^i, Re^{i+1}, S^{i+1}} \Big[\Big(Q_{\mathrm{tar}} - Q(S^i, A^i; \boldsymbol{\theta}^t) \cdot \nabla Q(S^i, A^i; \boldsymbol{\theta}^t) \Big]. \quad (31)$$

In (31), the expectation is calculated with respect to a so-called minibatch, which is randomly selected in previous samples $(S^i, A^i, Re^{i+1}, S^{i+1})$ for some $i \in \{t - M_r, t - M_r + 1, \ldots, t\}$, with M_r being the replay memory. The minibatch sampling is able to improve the convergence reliability of the updated value function [30]. In addition, the target Q-value Q_{tar} is denoted as

$$Q_{\mathrm{tar}} = re^{i+1} + \gamma \max_{a \in \mathcal{A}} Q(S^{i+1}, a; \bar{\boldsymbol{\theta}}^t), \quad (32)$$

Algorithm 2 : DQN to optimize cumulative terrestrial users' throughput

Algorithm hyper-parameters: $\alpha \in (0, 1], \gamma \in [0, 1), \epsilon \in (0, 1]$
Initialize replay memory M, primary Q-network $\boldsymbol{\theta}$, and target Q-network $\bar{\theta}$
for $e \leftarrow 1$ to I **do**
  Initialize s^1 by executing a random action a^0;
  UAVs identify the BS with the highest RSRP and associate with it.
  for $t \leftarrow 1$ to T **do**
    If $p_\epsilon < \epsilon$: Randomly select action a^t from $\mathcal{A}$;
    else select $a^t = \underset{a \in \mathcal{A}}{\mathrm{argmax}} Q\left(S^t, a, \theta\right)$;
    Agent performs the selected A and mutes cells.
    Agent observes S^{t+1} and calculates R^{t+1}.
    Store transitions (S^t, A^t, R^t, S^{t+1}) in replay memory, and sample random minibatch of transitions (S^t, A^t, R^t, S^{t+1}) from M.
    Calculate $\hat{Q}(S^{t+1}, a, \boldsymbol{\theta})$ according to Eq. (32).
    Calculate gradient descent using Eq. (29).
    Update $\bar{\boldsymbol{\theta}}$ every K steps.
  end for
  Determine all active UEs (TUEs and UAVs) using Bernoulli process.
  Determine associate cell UEs and active UE matrices.
  Update the queue matrices.
  Calculate SINR and transmission rate.
end for

where $\bar{\boldsymbol{\theta}}^t$ is the weight vector of the target Q-network to estimate the future value of the Q-function for the next state-action pair, and it is updated as $\bar{\theta} \leftarrow \theta$. This parameter is periodically updated from the current value $\boldsymbol{\theta}^t$ and kept fixed for some episodes. The DQN algorithm is a value-based algorithm, which can obtain an optimal strategy. This is due to the experience replay mechanism and randomly sampling in DQN, and DQN uses the training samples efficiently to smooth the training distribution over previous behaviours. Not only does this massively reduce the amount of interactions needed with the environment, but also reduce the variance of learning updates. The DQN algorithm creates a sequence of policies whose corresponding value functions converge to the optimal value function, when both the sample size and the number of iterations tend to infinity. The DQN algorithm is presented in Algorithm 2.

D. Muting Optimization Scheme and Dynamic time-frequency PRB Scheduling (MOSDS)

In this section, solutions on solving interference among TUEs and UAVs while maximizing the TUEs' capacity is proposed. Dynamic requests from both TUEs and UAVs can cause higher interference, especially in a high dense urban area. To deal with this issue, we consider a MOSDS-DQN to maximize the total capacity of TUEs, mitigate the interference, and mute the cell causing high interference.

The effect of blank subframes is modelled by assuming that the downlink transmission from the corresponding cells is

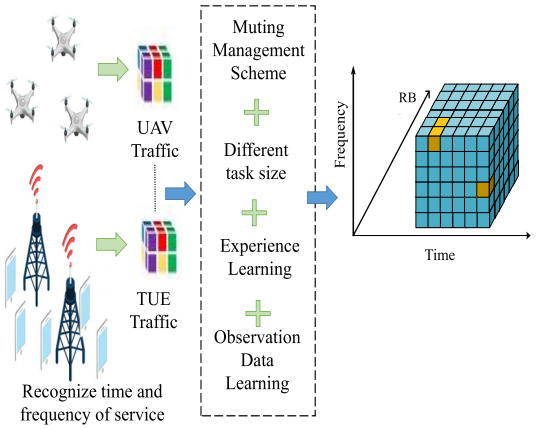

Fig. 5. The dynamic scheduling design for MOSDS-DQN.

muted in the corresponding frequencies. The main component is to suppress the blank sub-frames of the interfering cell, and the Almost Blank Sub-frames (ABS) scheme is applied according to ICIC Release 10 [14]. However, to reduce interferences, further-enhanced ICIC (feICIC) solutions are implemented in the system, which pre-allocate the packet in frequency and time domain for UAVs and TUEs as visualized in Fig. 1 and Fig. 4.

In this model, the available frequency bandwidth for the DL transmissions is divided into F sub-frames indexed by $f = 1, 2, \ldots, F$ and the time interval is slotted into transmission time intervals (TTIs) indexed by $i = 1, 2, \ldots, N$ as shown in Fig. 4. The time-frequency resource grid consists of $F \times N$ RBs. Therefore, the date rate of the uth TUE is defined as

$$R_{c,u}^{TUE} = B_{i,f}^{c} \log_2(1 + \gamma_{c,u}), \tag{33}$$

where $B_{i,f}^{c}$ is the bandwidth of the RB (i, f) at the cth cell and the data rate of the jth UAV is defined as

$$R_{c,j}^{UAV} = B_{i,f}^{c} \log_2(1 + \gamma_{c,j}). \tag{34}$$

For dynamic resource scheduling, we mainly consider efficient dynamic scheduling, where different data sizes and requirements are considered in this scenario. As proposed in [21], the UAV data rate and latency requirements need to satisfy 60-100Kbps and 50ms. Specifically, for the resource allocation problems with different time and frequency requirements, quantized time-frequency resource block allocation scheme is considered, as shown in Fig. 5. First, the controller classifies different services with the specific QoS requirements according to the service characteristics and the current network congestion. Second, according to the admission control policy, the resource block of each scheduled UAVs and TUEs are continuously mapped to the specific time and frequency domain. Finally, based on the current muting scheme, data sizes, and previous learning experience, dynamic resource scheduling for UAVs and TUEs are considered to reduce interference. Thus, to maximize TUE throughput and guarantee the reliability and latency of UAVs, optimizing both of the scheduling policies and cell muting selection are considered.

In addition, the omnidirectional antenna [2] is utilized in the algorithm to help mitigate the interference efficiently while maximizing the capacity of both TUEs and UAVs. It is assumed that each UAV transmits 1250B every 100ms [2]. Authors in [2] showed that TUEs could achieve the lowest capacity loss when the UAVs were scheduled to send information at every 10th and 50th TTI. However, the results are different when they have different number of UAVs in different scenarios, i.e., high load scenario. In real scenarios, it is difficult to predict the number of users. Thus, the main focus in this section is to jointly optimize the number of muting cells and UAVs' scheduling schemes, and the optimization problem is formulated as

$$(P2): \max_{\pi(A_t|S_t)} \sum_{i=t}^{\infty} \sum_{c=1}^{C} \sum_{u=1}^{U} \beta^{(i-t)} R_{c,u}^{TUE}(i) \tag{35}$$

$$\text{s.t.} \mathcal{N}_{B_u} \cap \mathcal{N}_{B_j} = \varnothing, \quad \forall u \neq j, \tag{36}$$

$$\sum_{l=1}^{L_u} \mathcal{N}_{B_u} \leq Q, \quad \forall (u, l), \tag{37}$$

$$R_{c,j}^{UAV}(i, f) \geq R_{Th}^{UAV}, \quad \forall j \in \mathcal{J}, \tag{38}$$

$$R_{c,u}^{TUE}(i, f) \geq R_{Th}^{TUE}, \quad \forall u \in \mathcal{U}, \tag{39}$$

$$\beta^{i} \in [0, 1]. \tag{40}$$

where $\beta^{i} \in [0, 1)$ is the discount factor determining the performance accumulated in the future reward. If $\beta^{i} = 0$, it means that the agent only concerns the immediate reward. Eq. (36) shows a RB should always be allocated to one user. The scheduler length $\mathcal{N}_{B_u}$ in Eq. (37) should allocate no more than the maximum queue length Q. Next, Eq. (38) and Eq. (39) ensure a good service rate for UAVs and TUEs, respectively. As the arrival of UAVs cause a trade-off between available PRBs and interferences among all users, it is important to consider an optimal trade-off among the RSRP, the group of UAV's RB, and muting scheme, which further motivates us to use the learning algorithms to jointly optimize long-term throughput of all users. The DRL agent then learns the optimal mapping from the input states to select the resource allocation action.

1) State Representation: The current state of the system includes commutative throughput of TUEs and UAVs, given by $S_2^t = \{\sum_{j=1}^{N_{TUE}} R_{j,t}^{TUE}, \sum R_t^{UAV}\}$, where R_{TUE} is a set of data rates of TUEs and R_{UAV} is the instantaneous rate of UAVs.

2) Action Space: Q-agent will choose an action a from set $\mathcal{A}$. The dimension of the action set is calculated as $\mathcal{A} = N_m \cdot t_s$. The action is denoted as $A_{t+1}^{P2} = \{N_m, t_s\}$, where N_m is the number of muting cells and t_s is the slice time allocation for UAVs to transmit data.

3) Rewards: After performing the selected actions, the accumulated reward function is given as

$$Reward_t^{P2} = \sum_{u=1}^{U} R_u \cdot \mathbb{1}\left[R_{u,j} \geq R_{Th}\right]. \tag{41}$$

The MOSDS-DQN algorithm is shown in Algorithm 3.

Algorithm 3 : MOSDS

Initialization α, ϵ, M, $\boldsymbol{\theta}$, and $\bar{\theta}$.
UAV identifies the highest RSRP and associate with it.
Receive muting cell ID from Algorithm 2 and the packet scheduling.
Determine all active UEs (TUEs & UAVs) using Bernoulli process and time packet scheduling.
Determine associate cell UEs & active UEs matrices.
for $e \leftarrow 1$ to I **do**
  Initialize s^1 by executing a random action a^0;
  for $t \leftarrow 1$ to T **do**
    If $p_\epsilon < \epsilon$: Randomly select action a^t from $\mathcal{A}$;
    else select $a^t = \underset{a \in \mathcal{A}}{\text{argmax}} Q\left(S^t, a, \theta\right)$;
    for $PRB_i \leftarrow 1$ to I **do**
      for $activecell_c \leftarrow 1$ to C **do**
        Update the queue matrices.
        Calculate pathloss.
        Calculate Antenna Gain.
        Calculate received power.
        Calculate the channel states.
        Calculate SINR and transmission rate.
      end for
    end for
  end for
end for

Fig. 6 shows the proposed network architecture, where the current state is input into the neural network for the DQN algorithm. Next, an RNN-based GRU network is used to approximate the value function of the DRL algorithm. The GRU can capture the correlation between the state and action over time, and helps DRL to select the optimal actions.

E. Computational Complexity Analysis

In this section, we evaluate the computational complexity of one iteration of our proposed algorithm with respect to the size of the network, namely, the number of UEs and available resources. The computational complexity of the DQN algorithm, including DQN learning architecture, the action selection of the agent, and the downlink transmission, is given by $O(mlogn + 2^A + N_i N_k)$, where m is the number of layers, n is the number of units per learning layer, and A is the number of actions [8, 31].

IV. Numerical Results and Evaluation

In this section, we examine the effectiveness of our proposed muting optimization schemes using DQN algorithm. The network consists of 7 cells covering 1500m x 1500m area. In the simulation, the UAVs are distributed with a fixed flying height. The height of all TUEs is 1.5m, and the height of all UAVs is assumed to be of 120 m following UK regulations [32]. Both TUEs and UAVs are assumed to equipped with a single antenna. The TUEs and UAVs in each cell are uniformly distributed and the maximum number of UAVs in the entire network is 10. We assume that all users move within their corresponding cells. When a TUE reaches the boundary, it turns back and moves in a random direction. The network

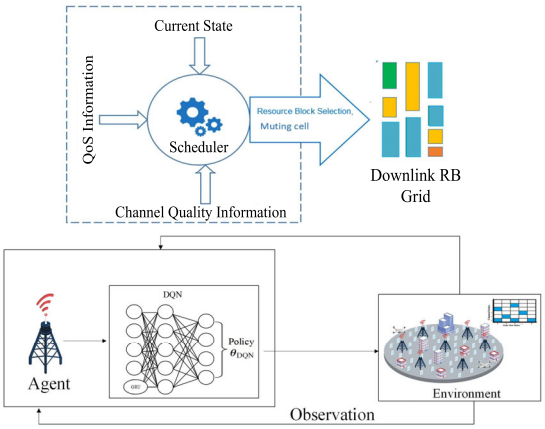

Fig. 6. The learning network architecture for MOSDS-DQN

TABLE I
Simulation Parameter

Parameters	Value
Transmission power, P_c	30 dBm
Bandwidth, B	3 MHz
Noise power $N_{c,u}$	-142.39 dBm
Center frequency, f_c	2 GHz
θ_{3dB}	65^o [20]
ϕ_{3dB}	65^o [20]
SLA_V	$30dB$
A_m	$30dB$
β	3.4 [2]
Antenna Gain, G_{max}	8dbi
UAV Threshold, R_{Th}^{UAV}	1Mbps [35]
UAV Threshold, R_{Th}^{TUE}	20 bps [36]
Alpha, α	0.001
Gamma, γ	0.999
Learning Rate	0.1, 0.01
Discount Rate	0.8
Replay memory	1000

parameters for the system are shown in Table I, and follow the 3GPP specifications in [21], [33], and [34]. All results are obtained by averaging over 100 episodes, with each episode containing 100 TTIs.

In the downlink, the TUE traffic pattern is modeled as File Transfer Protocol (FTP) sessions [2], where both packet size and arrival time follow Poisson distribution. The downlink scheduler prioritizes the UAV transmission and C2 traffic over the FTP traffic, meaning that the BS schedules the UAV transmission first, and then the remaining TUEs and resources are divided equally among the connected TUEs that have FTP data to receive. If there is no downlink data to be transmitted, users are assumed to be in an idle mode. Otherwise, the user switches from the idle mode to a connected mode. Once the data buffer is clear, the user returns to the idle mode.

A. Muting Optimization Scheme using Deep Q-Learning

Fig. 7 shows the reward of a dynamic muting scheme for different learning algorithms. From the simulation results, it

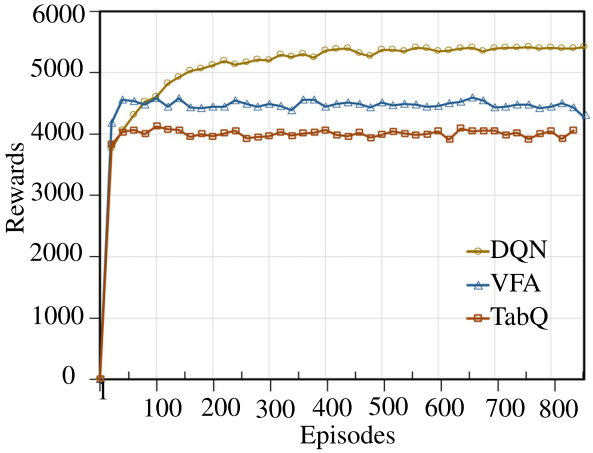

Fig. 7. Rewards performance comparison between different learning algorithms.

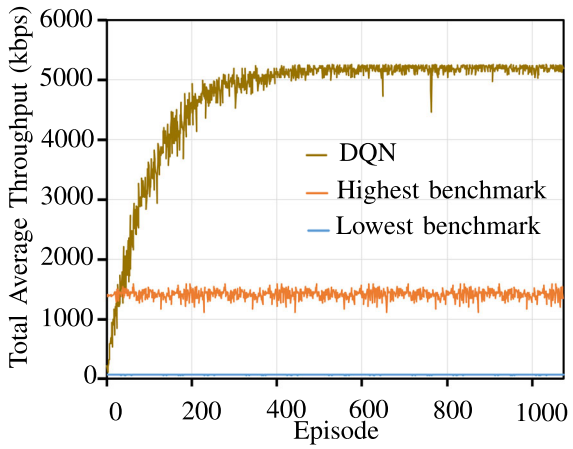

Fig. 9. Average TUEs' throughput comparison between DQN-based muting scheme and linear muting.

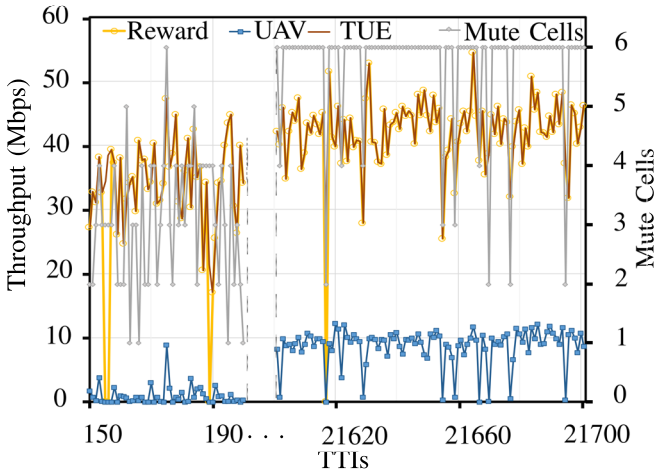

Fig. 8. Dynamic actions influence the rate for all TUEs and UAVs over time.

is clearly shown that DQN outperforms both VFA and Tab-Q. The convergence of both VFA and Tab-Q are slightly faster than DQN, but unable to obtain the maximum reward as that of DQN. Tab-Q fails to perform exploitation of each action in continuous time slots as it is fixed in a suboptimal strategy [37]. In addition, the VFA's target network might not fully works due to features and high number of state-action, which causes VFA cannot perform better exploration over time. DQN can explore and exploit actions, which enable it to obtain the maximum state-action value. Moreover, the convergence analysis of the reinforcement learning algorithms has been proven in [28], [38], so that the agent of the Q-learning algorithm can converge to the optimal Q value.

Fig. 8 shows how dynamic muting actions affect the total throughput for all TUEs and UAVs via DQN muting scheme over time. In Fig. 8, "Reward" represents the cumulative reward, "UAV" represents the total throughput for all UAVs, "TUE" represents the total throughput for all TUEs, and "Mute Cells" represents the number of muting cell in each time slot. At the early stage of learning, DQN learns to be adapted to the environment based on the observations, and the reward continues increasing. When $t = 155, 156, 189,$ and 21617, the rewards drop to zero due to the difficulties in choosing the correct muting number that suits the current environment, which leads to the transmission rate of UAV not satisfying the threshold in Eq. (16). As time passes, DQN can predict and learn how to maximize the reward. However, the system can become unstable when the epsilon-greedy parameter is less than the threshold, namely, $p_e < \epsilon$, as it directly selects a random action and decreases the performances of TUE. When the algorithm converges, the performance of all TUEs and UAVs maintain at the maximum value with an optimized number of muting cell.

Fig. 9 plots the convergence performance of DQN with different mitigation schemes [2]. For simplicity, "Highest benchmark" represents the linear muting scheme with 3 strongest interfering neighboring cells muted to allow UAVs to transmit their data without interference from TUEs, and "Lowest benchmark" shows the performance of the linear muting scheme when the system mutes a single neighboring cell with the highest interference, which can mitigate the interference between UAVs and TUEs. The "Highest benchmark" and "Lowest benchmark" use linear mitigation schemes in [2]. In [2], the "Highest benchmark" muted a maximum of 3 strongest RSRP interference signals to cancel the interference following the 3GPP Release-13 model [2, 39]. The DQN-based muting scheme shows 48% improvement compared to the "Highest benchmark". It is proved that the DQN scheme can adequately select the correct number of muting cells to reduce interference, even though the proposed system changes dynamically. In addition, DQN is able to perform in a dynamic scenario with a varying number of UAVs and TUEs, and select proper actions for the agent to maintain a higher data rate of TUEs. Compare to the lowest benchmark scheme, they are limited by the fixed set of rules it using and mute the fixed number of cells over time. In addition, the benchmark algorithm may not be able to find the optimal solution if the environment or condition is changing and DQN shows 80% improvement in overall data rate.

Fig. 10 plots the interference comparison analysis for DQN-based muting and linear muting schemes in [2]. It can be

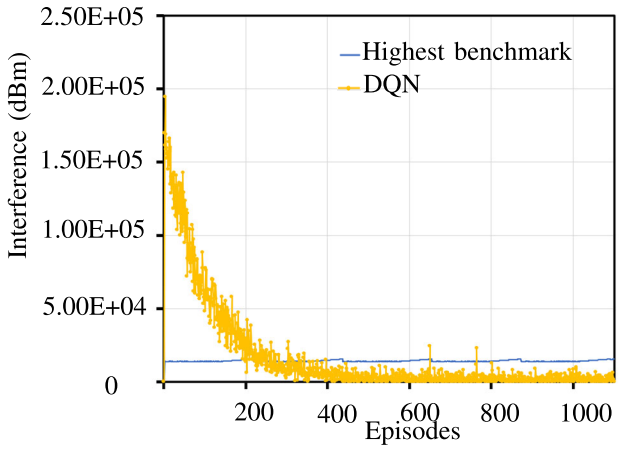

Fig. 10. Comparison of interference analysis between DQN-based muting scheme and linear muting.

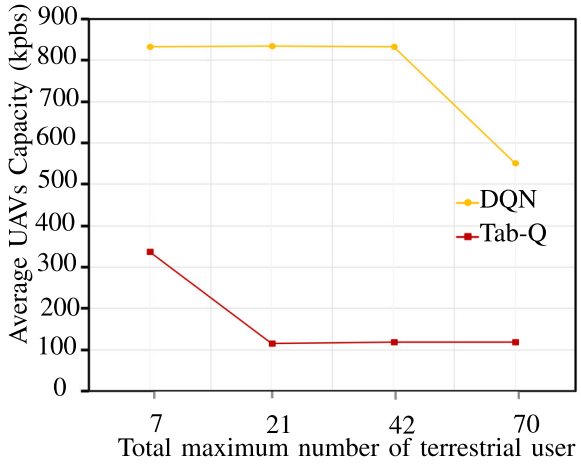

Fig. 12. Average capacity rate for UAV based on different number of TUEs.

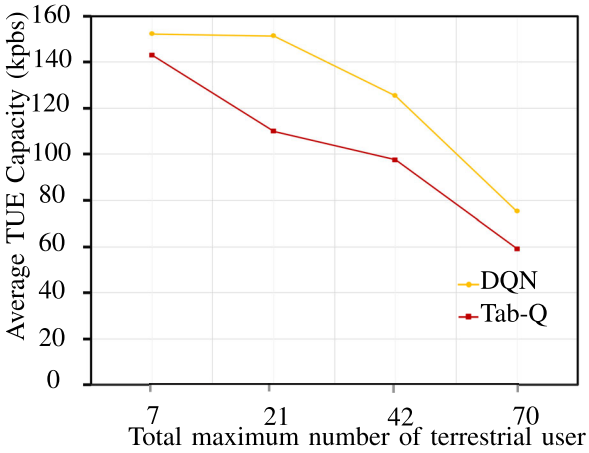

Fig. 11. Average capacity rate for TUE based on different number of TUEs.

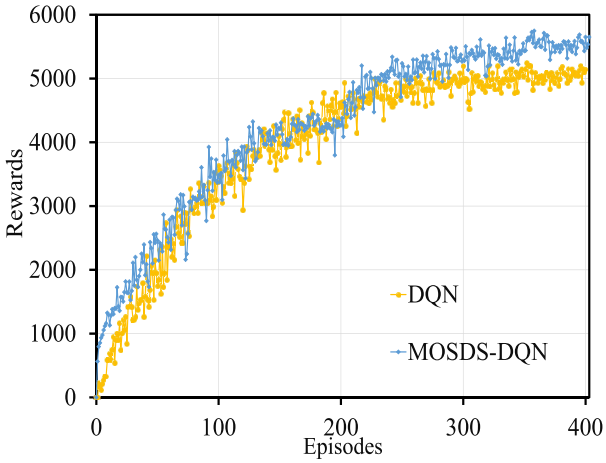

Fig. 13. Rewards performance comparison between different schemes.

seen that the proposed DQN muting scheme outperforms the Highest benchmark scheme. The result proves that the DQN muting scheme can accurately choose the cell muting index to reduce the interference in dynamic environments, and further maximize the average throughput.

Fig. 11 and Fig. 12 show the throughput performance of TUEs and UAVs in different situations, respectively. From Fig. 11, we observe that when the number of TUEs increases, the interference increases, and more number of TUEs and UAVs cannot satisfy their minimum transmission requirements. From Fig. 12, we can obtain that when the number of TUEs is small, UAVs achieve high throughput. However, when the number of TUEs increases up to 70, the average capacity of UAVs decreases because of high interference, and more UAVs cannot satisfy their minimum transmission requirements. It is because the muting schemes try to decrease the number of muting cells to let a high number of TUEs transmit their data, which leads to the UAVs being unable to satisfy their minimum requirements of transmission rate. Also, high number of TUEs causes less bandwidth allocated to UAVs, which further leads to lower throughput of UAVs. In addition, the performance of the UAV with the Tab-Q algorithm decreases dramatically when the number of TUEs increases. This is because Tab-Q with high dimensional state space requires large memory, and has difficulty in selecting proper actions to achieve optimal results.

B. Muting Optimization Scheme and Dynamic PRB Scheduling (MOSDS-DQN)

This section evaluates the proposed muting optimization scheme and dynamic PRB scheduling with MOSDS-DQN algorithm. Fig. 13 shows the convergence performance of the MOSDS-DQN muting scheme. For instant, "MOSDS-DQN" represents DQN muting optimization scheme and dynamic PRB scheduling. It is observed that the MOSDS-DQN algorithm performs better than the DQN muting scheme. However, the MOSDS-DQN scheme shows a lower convergence speed. It is because MOSDS-DQN muting scheme has larger state and action space, and needs to select more proper actions to mute cells and allocate PRBs to UAVs and TUEs.

Fig. 14 shows the throughput performance of dynamic actions over time. In Fig. 14, "Reward" represents the cumulative reward, "UAV" represents the total throughput of all UAVs, and "Mute Cells" represents the number of muting cells each time slot. Fig. 14a shows the muting cell action selected to

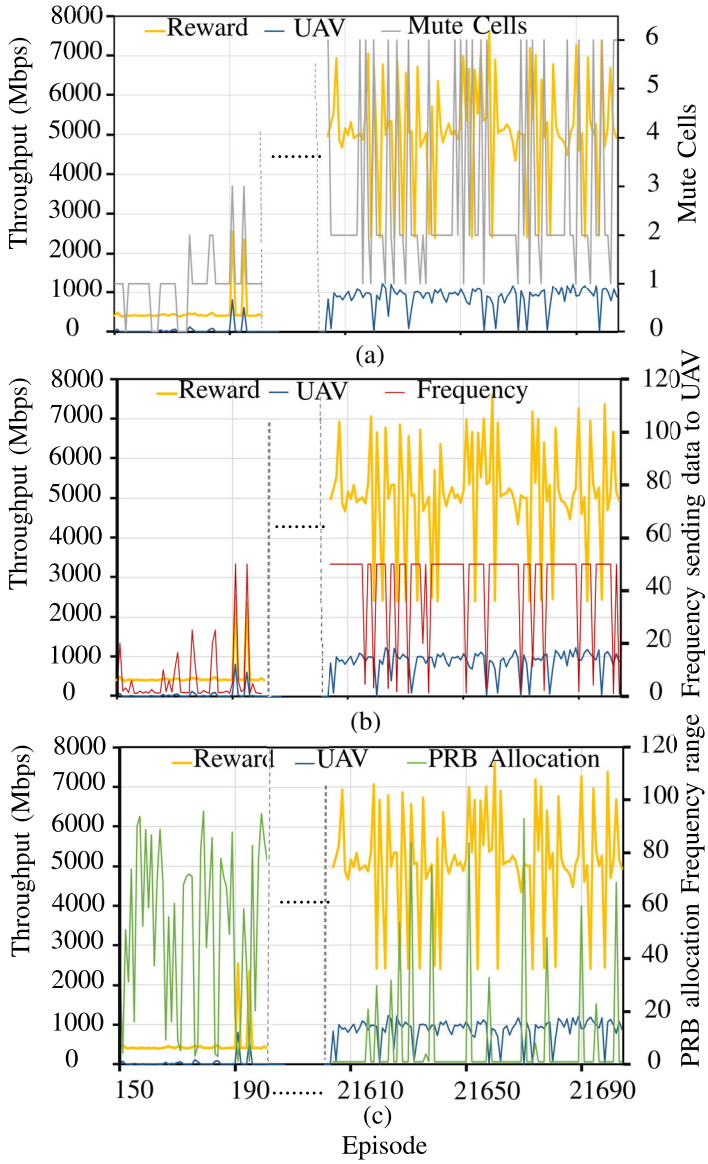

Fig. 14. Dynamic actions influence the reward for all UAVs in each episode.

maximize the overall throughput. At the beginning of the learning process, the reward continues increasing, it is because the algorithm is learning the environment based on previous experience. When the learning algorithms converge,F proper number of muting cells is selected to decrease the interference and improve the throughput. However, in some time slots, the performance of the UAV severely decreases and cannot satisfy the minimum QoS because of some factors, such as the data size of UAV, time allocation, and bandwidth allocation.

Fig. 14b shows how the BS sends data to the UAV in 100ms. The network environment condition and the location of UAV play important roles in MOSDS-DQN to plan the number of data pack of UAV transmission. For example, if the UAV is far from the cell, the frequency of sending UAV's data pack should be reduced to decrease the transmission failure. Thus, less data pack of the UAV is transmitted, and less PRB is allocated to the UAV, which decreases the interference and improve the throughput performance of TUEs.

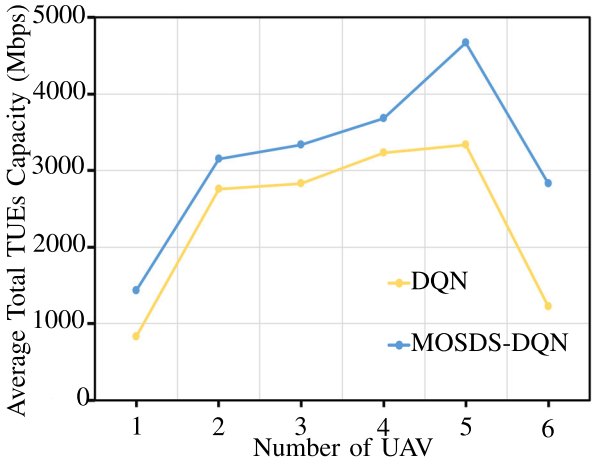

Fig. 15. Comparison of average capacity rate for all TUEs based on different number of UAVs.

In the early phase of Fig. 14c, the high-frequency range is used frequently, and it causes the performance of both UAVs and TUEs to decrease because of high interference. When the learning algorithms converge, proper frequency range for each PRB is selected to satisfy the QoS requirement. Therefore, the MOSDS-DQN algorithm is able to provide an effective way to select proper number of cells to mute and allocate proper PRBs to TUEs and UAVs, especially in different scenarios.

Fig. 15 plots the average capacity rate for all TUEs with different number of UAVs based on muting schemes via DQN and MOSDS-DQN. It is observed that when the number of UAVs is less than 4, both algorithms increase average capacity rate. However, when the number of UAVs increases from 4 to 5, the DQN algorithm is unable to increase average capacity. It is because it cannot select proper actions in a large action space. When the scenario becomes more complex, the algorithms need to balance the performance between UAVs and TUEs. In the simulation, UAVs have higher priority, thus, the performance of TUEs decreases tremendously when a high number of UAVs exist. Furthermore, when the number of UAVs increases, the performance of MOSDS-DQN is higher than that of DQN.

V. Conclusion

In this paper, a downlink inter-cell interference coordination mechanism was developed to mitigate the interference between BSs and TUEs while satisfying the rate requirements of UAVs. Then, adaptive muting optimization scheme and dynamic scheduling of PRBs were proposed to maximize the throughput of all users, and mitigate the interference by muting the cell(s) that caused high interference. Simulation results showed that our proposed learning-based schemes achieved 80% and 48% performance improvement of throughput compared to the lowest and highest linear muting algorithms, respectively. Furthermore, the proposed MOSDS-DQN also showed 18% improvement compared to DQN algorithm. In addition, the coordination of multiple agents in the MOSDS algorithm should be considered in the future work.

REFERENCES

[1] M. Mozaffari, X. Lin, and S. Hayes, "Toward 6G with connected sky: UAVs and beyond," *IEEE Commun. Mag.*, vol. 59, no. 12, pp. 74–80, 2021.

[2] H. C. Nguyen *et al.*, "How to ensure reliable connectivity for aerial vehicles over cellular networks," *IEEE Access*, vol. 6, pp. 12 304–12 317, 2018.

[3] W. Mei and R. Zhang, "Aerial-ground interference mitigation for cellular-connected UAV," *IEEE Wirel. Commun.*, vol. 28, no. 1, pp. 167–173, 2021.

[4] M. M. Azari, F. Rosas, A. Chiumento, and S. Pollin, "Coexistence of terrestrial and aerial users in cellular networks," in *2017 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2017, pp. 1–6.

[5] R. Amorim *et al.*, "Radio channel modeling for UAV communication over cellular networks," *IEEE Wirel. Commun. Lett.*, vol. 6, no. 4, pp. 514–517, 2017.

[6] H. C. Nguyen *et al.*, "Using LTE networks for UAV command and control link: A rural-area coverage analysis," in *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*. IEEE, 2017, pp. 1–6.

[7] I. Kovacs *et al.*, "Interference analysis for UAV connectivity over LTE using aerial radio measurements," in *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*. IEEE, 2017, pp. 1–6.

[8] L. A. b. Burhanuddin *et al.*, "QoE optimization for live video streaming in UAV-to-UAV communications via deep reinforcement learning," *IEEE Trans. Veh. Technol.*, pp. 1–14, 2022.

[9] V. Yajnanarayana *et al.*, "Interference mitigation methods for unmanned aerial vehicles served by cellular networks," in *2018 IEEE 5G World Forum (5GWF)*, Jul. 2018, pp. 118–122.

[10] A. Azari, M. Ozger, and C. Cavdar, "Risk-aware resource allocation for URLLC: Challenges and strategies with machine learning," *IEEE Commun. Mag.*, vol. 57, no. 3, pp. 42–48, 2019.

[11] A. S. Abdalla, K. Powell, V. Marojevic, and G. Geraci, "UAV-assisted attack prevention, detection, and recovery of 5G networks," *IEEE Wirel. Commun.*, vol. 27, no. 4, pp. 40–47, 2020.

[12] W. Mei and R. Zhang, "Cooperative downlink interference transmission and cancellation for cellular-connected UAV: A divide-and-conquer approach," *IEEE Trans. on Commun.*, vol. 68, no. 2, pp. 1297–1311, 2019.

[13] W. Mei, Q. Wu, and R. Zhang, "Cellular-connected UAV: Uplink association, power control and interference coordination," *IEEE Trans. Wirel. Commun.*, vol. 18, no. 11, pp. 5380–5393, 2019.

[14] A. Kumbhar, Guvenç, S. Singh, and A. Tuncer, "Exploiting LTE-advanced HetNets and FeICIC for UAV-assisted public safety communications," *IEEE Access*, vol. 6, pp. 783–796, 2018.

[15] A. Kumbhar, H. Binol, I. Guvenc, and K. Akkaya, "Interference coordination for aerial and terrestrial nodes in three-tier LTE-advanced HetNet," in *Proc IEEE Radio Wirel Symp*. IEEE, 2019, pp. 1–4.

[16] U. Challita, W. Saad, and C. Bettstetter, "Interference management for cellular-connected UAVs: A deep reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 18, no. 4, pp. 2125–2140, 2019.

[17] F. Tang, Y. Zhou, and N. Kato, "Deep reinforcement learning for dynamic uplink/downlink resource allocation in high mobility 5G hetnet," *IEEE Journal on Selected Areas in Commun.*, vol. 38, no. 12, pp. 2773–2782, 2020.

[18] J. An, K. Yang, J. Wu, N. Ye, S. Guo, and Z. Liao, "Achieving sustainable ultra-dense heterogeneous networks for 5g," *IEEE Communications Magazine*, vol. 55, no. 12, pp. 84–90, 2017.

[19] A. Vora and K.-D. Kang, "Effective 5G wireless downlink scheduling and resource allocation in cyber-physical systems," *Technologies*, vol. 6, no. 4, p. 105, 2018.

[20] M. Rebato, L. Resteghini, C. Mazzucco, and M. Zorzi, "Study of realistic antenna patterns in 5G mmWave cellular scenarios," in *Proc. 2018 IEEE Int. Commun. Conf. (ICC)*. IEEE, Jul. 2018, pp. 1–7.

[21] "Study on enhanced LTE support for aerial vehicles," 3GPP, TR 36.777, Dec. 2017, V15.0.0.

[22] "Study on channel model for frequencies from 0.5 to 100 GHz," 3GPP, TR 38.901, Jun. 2018, V15.0.0.

[23] C. A. Balanis, *Antenna theory: analysis and design*. John wiley & sons, 2015.

[24] "Technical specification group (TSG) RAN WG4; RF system scenarios," 3GPP, TR 25.942, Jun. 2001, V15.0.0.

[25] P. V. Klaine, J. P. Nadas, R. D. Souza, and M. A. Imran, "Distributed drone base station positioning for emergency cellular networks using reinforcement learning," *Cognitive computation*, vol. 10, no. 5, pp. 790–804, 2018.

[26] C. Zhan *et al.*, "Unmanned aircraft system aided adaptive video streaming: A joint optimization approach," *IEEE Trans. Multimedia*, vol. 22, no. 3, pp. 795–807, 2020.

[27] I. Budhiraja, N. Kumar, and S. Tyagi, "Deep-reinforcement-learning-based proportional fair scheduling control scheme for underlay d2d communication," *IEEE Internet of Things J.*, vol. 8, no. 5, pp. 3143–3156, 2021.

[28] J. Hu, H. Zhang, and L. Song, "Reinforcement learning for decentralized trajectory design in cellular UAV networks with sense-and-send protocol," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 6177–6189, 2018.

[29] Y. Zeng, X. Xu, S. Jin, and R. Zhang, "Simultaneous navigation and radio mapping for cellular-connected UAV with deep reinforcement learning," *IEEE Trans. on Wireless Commun.*, 2021.

[30] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, Feb. 2015.

[31] X. Liu and Y. Deng, "Learning-based prediction, rendering and association optimization for mec-enabled wireless virtual reality (vr) networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 10, pp. 6356–6370, 2021.

[32] "Drones: how to fly them safely and legally," Sep 2017. [Online]. Available: https://www.gov.uk/government/news/drones-are-you-flying-yours-safely-and-legally

[33] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal LAP altitude for maximum coverage," *IEEE Commun. Lett.*, vol. 3, no. 6, pp. 569–572, Dec. 2014.

[34] S. Zhang, H. Zhang, B. Di, and L. Song, "Cellular UAV-to-X communications: Design and optimization for multi-UAV networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 1346–1359, Feb. 2019.

[35] Y. Zeng, X. Xu, and R. Zhang, "Trajectory design for completion time minimization in UAV-enabled multicasting," *IEEE Trans. on Wireless Commun.*, vol. 17, no. 4, pp. 2233–2246, 2018.

[36] Y. Li, H. Zhang, K. Long, C. Jiang, and M. Guizani, "Joint resource allocation and trajectory optimization with QoS in UAV-based NOMA wireless networks," *IEEE Trans. on Wireless Commun.*, vol. 20, no. 10, pp. 6343–6355, 2021.

[37] N. Jiang, Y. Deng, A. Nallanathan, and J. A. Chambers, "Reinforcement learning for real-time optimization in NB-IoT networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1424–1440, Jun. 2019.

[38] T. Jaakkola, M. I. Jordan, and S. P. Singh, "On the convergence of stochastic iterative dynamic programming algorithms," *Neural computation*, vol. 6, no. 6, pp. 1185–1201, 1994.

[39] H. Holma, A. Toskala, and J. Reunanen, *LTE small cell optimization: 3GPP evolution to Release 13*. John Wiley & Sons, 2016.