EURASIP Journal on Wireless Communications and Networking

REVIEW Open Access

Trajectory optimization for UAV-assisted relay over 5G networks based on reinforcement learning framework

Sara M. M. Abohashish[1*], Rawya Y. Rizk[2] and E. I. Elsedimy[1]

*Correspondence: sara_mohamed@himc.psu.edu.eg

[1] Department of System and Information Technology, Faculty of Management Technology and Information Systems, Port Said University, Port Said, Egypt
[2] Electrical Engineering Department, Port Said University, Port Said, Egypt

Abstract

With the integration of unmanned aerial vehicles (UAVs) into fifth generation (5G) networks, UAVs are used in many applications since they enhance coverage and capacity. To increase wireless communication resources, it is crucial to study the trajectory of UAV-assisted relay. In this paper, an energy-efficient UAV trajectory for uplink communication is studied, where a UAV serves as a mobile relay to maintain the communication between ground user equipment (UE) and a macro base station. This paper proposes a UAV Trajectory Optimization (UAV-TO) scheme for load balancing based on Reinforcement Learning (RL). The proposed scheme utilizes load balancing to maximize energy efficiency for multiple UEs in order to increase network resource utilization. To deal with nonconvex optimization, the RL framework is used to optimize the trajectory UAV. Both model-based and model-free approaches of RL are utilized to solve the optimization problem, considering line of sight and non-line of sight channel models. In addition, the network load distribution is calculated. The simulation results demonstrate the effectiveness of the proposed scheme under different path losses and different flight durations. The results show a significant improvement in performance compared to the existing methods.

Keywords: Reinforcement learning, Sustainable development goals, Trajectory optimization UAVs

1 Introduction

Unmanned aerial vehicles (UAVs) are currently one of the most significant technological developments. Many real-world applications for UAV include entertainment, telecommunications, agriculture, transportation, and infrastructure. The utilization of UAVs necessitates the planning of appropriate vehicle trajectories. UAVs then can overcome the accessibility, speed, and dependability of the terrestrial system [1]. In particular, UAVs are utilized to enhance the communication performance of 5G networks that rely on UAV-assisted communication. UAVs can operate autonomously or through remote pilot control without needing a pilot onboard [2]. As such, UAVs are

utilized in a number of inventive ways to achieve the sustainable development goals (SDGs), including sustainable cities and communities, climate action, industry, innovation and infrastructure, and power saving and clean energy. Climate change has led to increase in the severity of wildfires as well as prolonged heat waves during the drought. UAVs are particularly useful for forest fire prevention. Using Victoria fire frequency statistics over a ten-year period and particle swarm optimization [3], the optimal position of UAV is determined for various fires. UAV-aided communication consists of two types of channels, the UAV-ground channel [4–10] and the UAV-UAV channel [11, 12]. UAV-assisted wireless communication operates in three modes: UAV-aided ubiquitous coverage, UAV-aided relaying and UAV-aided information dissemination and data collection. For instance, in [5, 7, 8, 11], the UAV-aided ubiquitous coverage deploys UAV to provide wireless coverage within an area. Two scenarios in which this can be examined are rapid service recovery and base station (BS) offloading in extremely crowded areas. UAV-aided relaying has been considered in [4, 9] to maximize reliability for user equipment (UE) without direct communication. These efforts aim to enhance the power profile and coverage range of UAVs, thereby reducing energy consumption for the network. Furthermore, UAV-assisted data collection provides an efficient way to collect data from network nodes [8, 10, 13]. Due to their ability to fly, UAV-aided relaying can provide more wireless communication resources, leading to improved coverage.

Recently, there have been several studies on UAV-assisted relay on 5G networks [14–18]. These studies have examined various aspects of UAV-assisted relay, including resource allocation, trajectory optimization, transmit power, load balancing (LB), and UAV channel modeling. Some studies have optimized UAV deployment and trajectory to improve communication performance [19–27]. However, most of these studies have focused on static UAVs, which are not suitable for providing reliable relaying communication. Mobile UAVs can be used as relays to maintain the communication between UE and destination. In addition, previous studies have either considered UAV deployment or traffic load balancing for UAVs, but not both. The LB of UAV wireless communication is a topic of interest in existing literature, with two types of load to consider: the amount of resources connected with each UAV and the amount of resources associated with each UE.

For the UAV-assisted relay network, the existing studies evaluated the LB, which utilizes UAV to improve spectrum allocation [28], data rate [29], wireless latency of users [30], and QoS requirements [31]. Most of the above related works studied one aspect of LB, either allocating resources by UAV or by UE, but ignored the design of an optimal trajectory for the UAV. The load imbalance cannot guarantee efficient distribution in incoming network traffic. Thus, some problems need to be addressed to make an efficient communication environment. The line of sight (LOS) has made a promising solution for energy efficient trajectory. In addition, UAVs offer a better LOS communication between UAV and UE to reduce transmission energy consumption. However, previous works have not considered the design of trajectory and traffic characteristics, especially related to energy consumption.

This paper focuses on an energy-efficient UAV trajectory for uplink communication, where a UAV serves as a mobile relay to balance the load in 5G networks. In particular, the main contributions of this paper are summarized as follows:

1. Our optimization problem proposes a UAV Trajectory Optimization (UAV-TO) scheme for load balancing based on Reinforcement Learning (RL) that maximizes energy efficiency (EE) for uplink communication. The problem considers the Three-Dimensional (3D) flight trajectory of UAV to visualize the aircraft performance and verify the safety and adaptability of the algorithm. The channel model of LOS and non-line of sight (NLOS) are formulated to solve optimization problem while multiple UEs are present on the ground.
2. The proposed scheme utilizes LB to maximize EE for multiple UEs in order to increase network resource utilization. It is based on the amount of resources that the UAV can carry as its load. Specifically, the highest data rate of UE is transmitted first to get the minimum resources. The range of the load takes values between 0 and 1. Therefore, if the load is smaller than 1, the UAV can receive more resources from the UE. Otherwise, the network cannot allocate sufficient resources to UE.
3. The optimization problem is nonconvex, so the RL framework is utilized to build UAV trajectory planning in order to overcome this problem. Two categories of RL, Monte Carlo (MC) and Dynamic Programming (DP) are used to improve resource utilization. The main objective of RL is to provide the optimal solution, which is represented through the interaction between the UAV and its environment. In addition, the network load distribution is calculated.
4. The simulation results demonstrate the performance of the proposed scheme under different path losses and different flight durations. The results show that the proposed scheme outperforms the existing methods under various parameter configurations.

The rest of the paper is organized as follows: Sect. 2 introduces the related work. Section 3 presents the RL. Section 4 presents the system model and channel model. Section 5 describes the problem formulation. Section 6 presents the proposed UAV-TO scheme. The simulation results are provided in Sect. 7. Finally, Sect. 8 concludes the paper.

2 Related work

There have been significant works focused on UAV-aided relay communication for the UAV-ground channel [4–10] and the UAV-UAV channel [11, 12]. The authors highlighted the characteristics of mmWave propagation for 5G in [4]. The Friis Transmission Equation is used to determine UE received power for the relay path in order to investigate the various mmWave propagation characteristics. Furthermore, this study includes the propagation of mmWave channels in a Ray-Tracing simulator, which assesses the relative effectiveness of diffracted, reflected, and scattered paths compared to direct paths. When the height of the UAV increases, the power received by the UEs decreases. At a

height of 30 m, the UAV provides sufficient coverage. At the same time, a throughput maximization is achieved in [5], where multi-UAV is presented through UAV trajectory delay and packet loss enhancement. A graph neural network (GNN) is used to determine the natural order of nodes by graph for transmission and movement properties. In addition, a dynamically reconfigurable topology is described where the information state of aerial nodes is generated and the node parameters are modified. The location of the UAV is modified to reduce packet loss and delay. The energy-efficient for UAV is investigated in [6] based on the LOS and NLOS of communication links in order to optimize the UAV's trajectory path, transmit power, and speed. It is assumed that the UAV's flying speed is fixed. A binary decision variable is assigned to plan the UAV-to-UE connectivity.

By using the estimation of throughput for UAV, the optimal UAV position is proposed in [7], for a multi-rate communication system between two ground nodes. Three modulations on multi-rate in IEEE 802.11b are considered. The estimation of UAV throughput depends on the UAV location and link rates. When the relay is closer to one of the ground nodes than the center position, the maximum throughput is investigated. The uplink channel model is presented in [8], which considers the impact of 3D distance and multi-UAV reflection. Maximizing the EE of the UAV-BS uplink is examined by modifying the UAV's uplink transmit power at different 3D distances.

In [9], an iterative approach is studied to optimize the UAV trajectory and edge user scheduling in a hybrid cellular network. The objective is to maximize the sum rate of edge users while considering the rate requirements of all UEs. The problem is formulated as a mixed-integer nonconvex optimization.

In general, coverage is the most important aspect of UAV-assisted wireless communication. Specifically, in [10], considered the coverage and data rate constraints to identify the minimum number of UAVs and their optimal positions. The mathematical model determines the optimal position and height of UAVs in 3D space. Both decode-and-forward (DF) and amplify-and-forward (AF) are presented in [11] for a cooperative communication system with a single source and receiver. They optimize the UAV's power profile, power-splitting ratio profile, and trajectory to maximize the throughput.

For UAV-assisted relay networks, related works have studied the deployment path for UAV [12–18]. The work of [12] examined the deployment path of UAV and resource allocation in order to ensure user fairness. The aim was to maximize the minimum throughput for all the UEs while taking into account various constraints such as backhaul bandwidth, backhaul information causality, UAV-BS mobility, total bandwidth, and maximum transmit power. As data collection continues to grow exponentially [13], explored energy-efficient data gathering for UAV-assisted WSNs. The proposed method aims to minimize throughput by optimizing UAV deployment and sensor node transmission (SN). The SNs are allowed to choose among three transmission modes, which include waiting, standard sink node transmission, and UAV uploading. In addition, the optimal 3D positioning of the UAVs and the power allocation are

proposed in [14] to improve performance by maintaining secrecy in the presence of eavesdroppers. In [15], the UAV's transmit power is proposed which utilizes the uplink channel model from UAVs to ground BSs. A resources optimization problem in [16] is investigated for UAV-assisted 5G networks. Resource allocation and control of the planning path play an important role in dynamic UAV-assisted 5G networks. Specifically, in [17], multiple UAV-BSs communication is proposed, which utilizes UAV as flying BSs to provide coverage enhancement and QoS requirements in 5G wireless networks. In [18], Ray-tracing simulations are also proposed to enhance the coverage of a UAV operating as aerial BS.

Generally, the trajectory of UAVs plays an active role in improving the performance of UAV-assisted wireless communication. It is worth noting that the related works about UAV-aided relay systems mainly focus on trajectory optimization [19–26]. A survey of UAV characteristics and limitations in the integration of 5G communications into UAV-assisted networks is presented in [19]. An energy-efficient trajectory optimization of UAVs is examined in [20], where the trajectory strategy of the UAV is designed in conjunction with prohibitive power depletion and QoS requirements to optimize data transmission, energy consumption, and coverage fairness. For a multi-hop UAV relaying system, the UAV trajectory and transmit power optimization are utilized in [21] to maximize the end-to-end throughput. The trajectory and transmit power of UAV-BS are optimized to recognize UAV-to-ground (U2G) and ground-to-UAV (G2U) communications [22]. The optimal trajectory of the UAV is presented in [23] to maximize EE for UAV communication. Moreover, when considering the circular trajectory of the UAV, rate maximization and energy minimization are investigated in situations where the UAV trajectory was unconstrained.

Power allocation and trajectory optimization are studied in [24–26] for UAV-assisted relay systems. In particular, [24] investigated the optimal trajectory of UAV to improve the throughput of the communication link between two UEs. UAV is considered as a mobile relay to maximize the minimum transmission rate between transmitter and receiver. To address a nonconvex problem, it converted the main problem to some sub-problems, which jointly optimizes the power allocations and UAV trajectory alternately. It is worth noting that [25] also investigated an outage probability minimization problem by a long-term proactive optimization algorithm. In addition, the closed-form of outage probability is formulated by optimizing UAV's 3D trajectory. Two hop mobile relay of UAV is used in [26] to serve two UEs on the ground. UAV-enabled AF relay network is designed to achieve the maximum end to end throughput. Several works [28–31] have discussed the optimal position of UAVs based on LB. The optimal user association and spectrum allocation schemes are investigated in [28] based on the branch-and-bound method to maximize the sum rate in two adjacent cells. In [29], the LB of UAV and fairness of UEs are investigated which optimizes UAV deployment over multi-UAV networks by diffusion UAV deployment. The LB is utilized to distribute the load balancing across neighbor UAVs. In [30], the deployment of UAVs as flying BS is analyzed to determine suitable

locations to serve more UEs and improve the wireless latency. The goal is to achieve a traffic loading balance that improves the channel quality of UEs in a UAV-assisted fog network. In multi-UAV-aided mobile edge [31], investigated LB of UAV to improve coverage and QoS requirements.

The RL algorithm has been used in many research studies to aid the navigation through unknown environments. The main objective of RL is to provide the optimal solution, which is represented through interaction between the UAV and its environment. Trajectory planning, navigation, and control of UAVs are discussed in [32] using RL. Meanwhile [33], developed trajectory planning of UAV in an uncertain environment using RL.

In recent years, many new approaches for autonomous navigation and path planning of UAVs have emerged [34–36]. Specifically [34], used generative adversarial networks and window functions to develop the spatial resolution of satellite images. A high quality image of UAV is designed in [35] using super-resolution techniques based on Convolutional Neural Network (CNN), while [36] proposed an energy-efficient optimal path of UAV with hybrid ant colony optimization and a variant of A*.

3 Reinforcement learning framework

RL involves an agent interacting with its environment in a cycle, aiming to learn rewards and optimize a policy [37]. The fundamental elements of RL include action, reward, value, policy, transition, and the environment model. Actions describe how agents move from one position to another, while rewards represent the numerical values of the immediate environment state. The policy defines the actions that an agent takes in any given state, and transitions define the probability distribution from the current state to the next. Value represents the expected value of a state, calculated by cumulative discounted rewards. The main objective of RL is to determine an optimal policy to maximize/minimize a certain objective function. It can be defined by a tuple (S, A, P, R, T) where S is a finite set of states and A denotes a finite set of actions and the UAV takes an action $a \epsilon A$ at state $s \epsilon S$. The probability transition function denotes as $P = \Pr\left(s_{t+1} = s' \middle| s_t = s, a_t = a\right)$

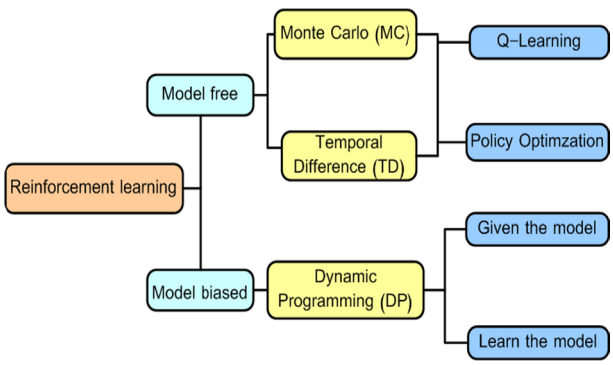

Fig. 1 Classification of RL Algorithm

from state s at time t to state s' at time $t+1$ after executing action a. R is a reward function and T is the set of decision epochs which can be finite or infinite. The policy function represents as π, which is a mapping from a state s to an action a. The RL describes how agents learn the optimal policy (π), where the highest reward value is achieved [33]. RL algorithms are classified into two categories: model-based and model-free as described in Fig. 1.

The model-based RL method involves the agent using the transition probability from the model to determine the next reward and action. This method requires an explicit environment and agent model. DP is an example of a model-based method that requires full observable environmental knowledge. DP is used to identify the optimal solution through value or policy iteration. The agents in model-free RL do not store any information about the environment. Instead, they update their knowledge to determine the quality of a proposed action. The agent's objective is to choose the optimal action, estimating action values based on experience rather than through exploitation. MC and Temporal Difference (TD) are examples of model-free RL algorithms [38]. The MC is a model-free technique that directly learns from episodes of experience [33]. In each episode, the agents move from its current state to its terminal state.

It may be used with sample models without bootstrapping, whereas the TD approach learns from the current value function estimation using bootstrapping. Generally, TD is utilized to predict a quantity that depends on the signal's future values. Q-learning and SARSA are the two major TD-based algorithms [32].

4 System model and channel model

4.1 System model

The following considers uplink communication in a geographical area. The network has one MBS, which is located at the center of the area as depicted in Fig. 2. A UAV flies at a fixed altitude H to serve a group of ground UEs, which serves as a mobile relay. In addition, UAVs can use millimeter waves (mmWave) to provide the back-haul link between UAV and MBS. Total duration to complete a relay communication is T. At each t, $0 \leq t \leq T$, the UAV is deployed as an aerial relay to assist communication between UE and MBS for AF communication. In this paper, the UAV, UE, and MBS are equipped with one antenna. Thus, there is no interference between UE-UAV and UAV-MBS links. UEs are set as $i = (1, 2, \ldots .I)$ and the location of the static UE is assumed as $[W_i, 0]^T$, where $W_i = [x_i, y_i, 0]$ denotes the horizontal coordinate. UEs are distributed randomly in the network. The UAVs are set as $j = (1, 2, \ldots .J)$ and the coordinates of each UAV are assumed as $[q_j(\mathrm{t}), H]^T$ at time t, where the horizontal coordinate of the UAV can be written as $q_j(t) = (t), y_j(t)]^T$ at time t.

This paper proposes UAV-TO flights at fixed altitude H above the ground for a duration of T. Therefore, the location of q_j UAV is considered unchanged within each time slot. After a certain amount of time, the UAVs' positions vary constantly, while the locations of the UEs are assumed to be fixed during each cycle of UAV positioning. The paper studies the channel model based on both LOS and NLOS scenarios.

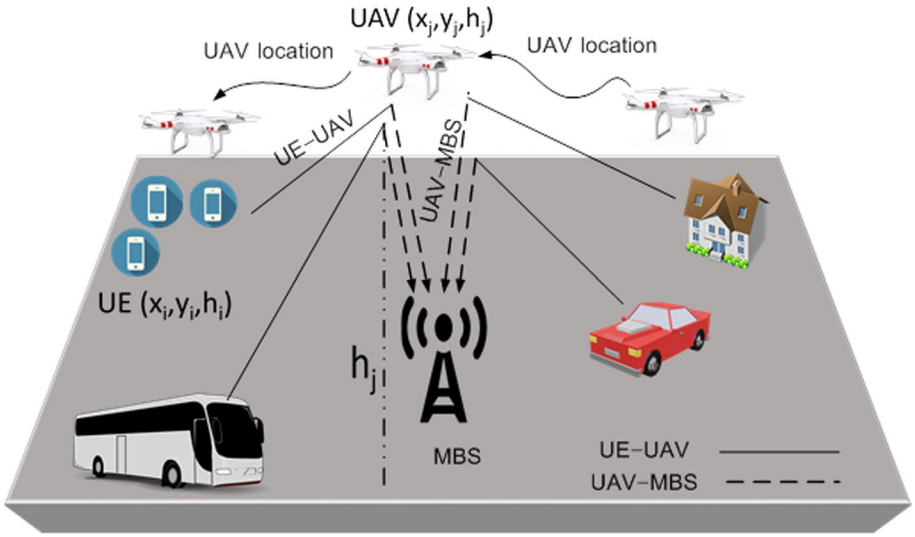

Fig. 2 System model of UAV

4.2 Channel model

4.2.1 The average path loss between UE and UAV

Mobility and the LOS channel are important aspects of 5G networks [19]. This paper aims to improve LOS communication between UAVs and UEs to reduce transmission energy consumption. Hence, the probability of the LOS channel between a UEs and UAV at time t represents as [5]:

$$p_{LOS}(t) = \frac{1}{1 + \alpha e^{-\beta(\Phi_{ij} - \alpha)}} \tag{1}$$

where α and β are the constant values depending on the environment such as Urban and Suburban. Meanwhile, Φ_{ij} is the elevation angle (in degrees) between UAV j and UE i, which can be expressed as $\Phi_{ij} = tan^{-1}\left(\frac{H}{d_{ij}(t)}\right)$, where H is the UAV's altitude and $d_{ij}(t) = \sqrt{q_j(\mathrm{t}) - W_i^2 + H^2}$ is the distance between j UAV and i UE at time t.

So, the probability of NLOS between UAV and UE can be expressed as:

$$p_{NLOS}(t) = 1 - p_{LOS}(t) \tag{2}$$

$¥_{ij}^{los}(t)$ and $¥_{ij}^{Nlos}(t)$ represent the path loss between UEs at location i and UAV j with the LOS and NLOS channels, respectively [39].

$$¥_{ij}^{LOS}(t)(dB) = \zeta^{los} log\left(\frac{4\pi f_c d_{ij}(t)}{c}\right)^{\tau} \tag{3}$$

$$¥_{ij}^{LOS}(t) = \tau\zeta^{LOS}\left(0.5\log\left(q_j(\mathrm{t}) - W_i^2 + H^2\right) + log\left(f_c\frac{4\pi}{c}\right)\right) \tag{4}$$

$$¥_{ij}^{NLOS}(t) = \tau\zeta^{NLOS}\left(0.5\log\left(q_j(\mathrm{t}) - W_i^2 + H^2\right) + log\left(f_c\frac{4\pi}{c}\right)\right) \tag{5}$$

where ζ^{los} and ζ^{Nlos} represent the excessive path loss coefficient for LOS and NLOS channels depending on the urban area. Moreover f_c is carrier frequency,c is the speed of light, and τ is the path loss exponents. The average path loss model for LOS and NLOS is calculated by using Eqs. (4) and (5). Therefore, the average path loss between UE and the UAV can be written as:

$$¥_{ij}^{avg}(t) = p_{LOS}(t)¥_{ij}^{LOS} + p_{NLOS}(t)¥_{ij}^{NLOS} \tag{6}$$

$$¥_{ij}^{avg}(t) = \left(\zeta^{los}p_{LOS}(t) + \zeta^{NLOS}p_{NLOS}(t)\right)log\left(\frac{4\pi f_c d_i(t)}{c}\right)^{\tau} \tag{7}$$

4.2.2 The average path loss between UAV and MBS

Assuming the altitude of the UAV is high and there is no obstacle between UAV and MBS, the backhaul link is assumed to be a LOS link. Therefore, the channel characteristics are unique due to the strong LOS connections. To maximize the EE of UAV-assisted relay in a 5G network, an optimization problem is formulated to jointly determine the optimal trajectory of the UAV. Similarly, the path loss between the UAV and the MBS can be denoted as:

$$¥_{j}^{los}(t) = \tau\zeta^{LOS}\left(\log f_c + \log\sqrt{\left(\mathrm{q(t)}^2 + H^2\right)} + log\frac{4\pi}{c}\right) \tag{8}$$

where $\sqrt{\left(\mathrm{q(t)}^2 + H^2\right)}$ represents the distance between UAV and MBS.

4.3 Data rate

4.3.1 Transmission from UE and UAV

The average of the channel gain from i UE to UAV can be expressed as follows $g_{ij}(\mathrm{t}) = \frac{1}{¥_{ij}(t)}$, where $g_{ij}(\mathrm{t})$ represents the channel gain based on LOS and NLOS communication links. According to [12], the signal to noise ratio (SNR) of the access link can be calculated by:

$$\gamma_{ij}(\mathrm{t}) = \frac{P_i g_{ij}(\mathrm{t})}{\sigma^2} \tag{9}$$

where P_i represents the transmission power of i UE and σ^2 defines the noise power. Therefore, the data rate of the access link of UE i can be modeled as [38]:

$$rate_{ij}(\mathrm{t}) = Blog\left(1 + \gamma_{ij}(\mathrm{t})\right) \tag{10}$$

where B is the bandwidth exclusively used by UAV.

4.3.2 Transmission from UAV and MBS

Here, $g_{jd}(\mathrm{t})$ represents the channel gain from UAV to MBS and P_j as the transmission power of UAV. Thus, the SNR of the backhaul link from UAV to MBS can be calculated as [5]:

$$\gamma_j(\mathrm{t}) = \frac{P_j g_{jd}(\mathrm{t})}{\sigma^2} \tag{11}$$

In AF mode, the time domain for a UE is divided into two slots. In the first half of the time slot, the UE broadcasts its data to both the UAV and MBS. Then, in the second half, the UAV amplifies the received data and forwards it to the MBS. As a result, the achievable data rate *rate*(t) of *i* UE towards MBS via UAV based on AF is given as [12]:

$$rate_i(\mathrm{t}) = \frac{B}{2} log\left(1 + \gamma_i(\mathrm{t}) + \frac{\gamma_{ij}(\mathrm{t})\gamma_{jd}(\mathrm{t})}{1 + \gamma_{ij}(\mathrm{t}) + \gamma_{jd}(\mathrm{t})}\right) \tag{12}$$

5 Problem formula

This paper proposes a UAV-TO scheme for load balancing based on RL in UAV-assisted relay. The proposed scheme formulates LB to maximize EE for multiple UEs to increase network resource utilization. The binary variable x_{ij} is used to indicate whether *i* UE is assigned to *j* UAV or not. $y_j = \sum_{i \in I} x_{ij}$ represents the number of UEs associated to *j* UAV.

$$x_{ij} = \begin{cases} 1 \ user\ i\ served\ by\ j\ UAV \\ 0 \ otherwise \end{cases} \tag{13}$$

5.1 Load definition

There are two types of load, including the amount of resources connected with each UAV and the number of resources associated with each UE [40]. The proposed scheme is based on the amount of resources of each UE as load. Therefore, the total available resources of UE to UAV can be represented as N_i. When *i* UE communicates with *j* UAV, the load caused by transmitting the data of *i* UE to *j* UAV. It can be noted that, the load of UAV can be defined as the ratio of amount of resources allocated of UAV from *i* UE to the total available resources of UEs [41].

$$l_j(t) = \frac{\sum_i \rho_i(t)}{N_i} \tag{14}$$

where $\rho_i(t)$ is the number of required resources of UAV from *i* UE. It can be denoted as follows:

$$\rho_i(t) = min \frac{R_i}{BW\gamma_{ij}(\mathrm{t})} \tag{15}$$

where the data rate required and the bandwidth of the resource block are denoted as R_i and BW, respectively. The bandwidth of resource block is 180 kHz [42]. Specifically, the highest data rate of UE is transmitted first to get the minimum resources. It is important to determine the amount of resources allocated for UAV according to the LB. Note that the range of load takes [0, 1]. Therefore, if the load is smaller than 1, the UAV can receive more than the resources of UE. Otherwise, the network is overloaded because the load of UAV exceeds 1, hence it cannot be sufficient to allocate resources to UE.

Consequently, it can be recognized that the load of UAV cannot exceed 1 for all UEs in the network.

5.2 Energy efficiency

There are many factors that are involved in the energy consumption of a UAV such as communication energy for data transmission, flying energy to keep UAV mobile, and energy due to vertical climb [23]. Flying energy relates to speed and acceleration of the UAV, while the communication energy of the UAVs depends on the data transmission/reception. To avoid complexity, the UAV's takeoff and landing are not considered, hence the energy required for vertical climb is ignored. The energy required for data transmission is given as:

$$E_C(t) = l_d(t)E_i(t) + E_j(t) + l_i(t)E_i(t) \tag{16}$$

where $l_i(t)E_i(t)$ is the energy consumed by i UE which transmitted data to MBS through UAV, $l_d(t)E_i(t)$ is the energy consumed by UE which transmitted data to MBS as direct mode, and $E_j(t)$ is the energy consumed by i UAV j to MBS.

Based on the approach in [23], the energy required for flying is expressed as:

$$E_f(t) = c_1 v_j(t)^3 + \frac{c_2}{v_j(t)}\left(1 + \frac{a_j(t)^4}{\left(d_{ij}(t)\right)^2 g^2}\right) \tag{17}$$

where c_1, c_2 are fixed parameters related to the aircraft's weight, wing area, and air density. Furthermore, $v(t)$ is a velocity of UAV and $a(t)$ is acceleration of UAV. g is the gravitational acceleration. The total energy of UAV can be denoted as E_T which is the combination of communication energy E_c and flying energy E_f. It can be written as:

$$E_T(t) = E_c(t) + E_f(t) \tag{18}$$

Thus, EE of the UAV is defined as the ratio between the data rate and the energy consumption of the UAV. Therefore, the EE can be denoted as

$$EE = \frac{\sum_i \sum_t x_{ij}(t) rate_i(t)}{\sum_i \sum_t \left(E_c(t) + E_f(t)\right)} \tag{19}$$

5.3 Optimization objective

Our objective is to balance the load of the UAV by optimizing its trajectory, utilizing it as a flying relay. Since the UE with the highest data rate is transmitted first, the load of UAV is determined by the number of UEs associated with UAV. The UAV has a time duration T, which can be divided into M time slots with length T/M. These time slots M are used for designing the UAV's trajectory. Therefore, the optimization formula can be written as:

$$\text{P1 Max } EE_j = \frac{\sum_i \sum_t x_{ij}(m) rate_i(m)}{\sum_i \sum_t \left(E_c(m) + E_f(m)\right)} \tag{20}$$

which is subject to:

$$R_{min} \leq rate_i(m) \leq R_i \quad \forall\, i,\ \forall\, m \tag{20a}$$

$$\rho_i(m+1) \leq N_i - \sum_i \rho_i(m) \quad \forall\, i,\ \forall\, m \tag{20b}$$

$$\sum_{i \in I} x_{ij}(m) = y_j \quad \forall\, m \tag{20c}$$

$$x_{ij}(m) \epsilon (0,1) \quad \forall\, m \tag{20d}$$

$$a_j(m) < a_{max} \quad \forall\, m \tag{20e}$$

$$v_j(m) < v_{max} \quad \forall\, m \tag{20f}$$

$$q_j(m+1) - q_j(m) \leq \frac{T}{M} v_{max} \tag{20g}$$

$$q_j[0] = q_0, \quad q_j[T] = q_F \tag{20h}$$

The constraints in the optimization problem are classified into 4 types: user QoS constraints, UAV mechanical constraints, constraints of the load traffic, and UAV trajectory constraints. Constraint (20a) indicates QoS constraints, where R_{min} and R_i denote the minimum data rate required for i UE and the overall data rate required for all i UEs, respectively. Equation (20b) represents a constraint of the load traffic. Constraints (20c) and (20d) provide the total number of UEs that are communicated by the UAV. Equations (20e) and (20f) are the UAV's velocity and acceleration constraints, where v_{max} and a_{max} denote the maximum velocity and maximum acceleration, respectively. Equation (20g) satisfies the constraint of UAV trajectory. Constraint (20h) defines the initial location of UAV $q_j[0]$ and the final location $q_j[T]$ at period T.

It is noted that the optimization problem P1 is a mixed integer non-convex. Furthermore, the various flight constraints and the high dynamic topology of the network increase the complexity of solving the problem. Meanwhile, the Markov decision process (MDP) is a mathematical framework used for describing the environment in RL problems [43]. Two categories of RL, namely MC and DP, are utilized in this paper in order to improve resource utilization.

Q-Learning is a model-free approach in which the optimal policy is learned by using off policy. Q-learning describes an agent that learns the optimal action from an unknown environment. The next action is selected based on the maximum Q-value of the next state, which is a greedy policy. At each time slot m, given state $s(m)$, the agent chooses action $a(m)$ with respect to its policy π. After the action is performed, the agent receives an immediate reward r and transmits to a new state $s(m+1)$. The cumulative reward from the current state up to the terminal state at time M can be calculated by [7]:

$$G_m = \sum_{k=m}^{M} \vartheta^{k-m} r(s_k, a_k) \tag{21}$$

where $\vartheta \epsilon [0, 1]$ is a discount factor balancing between immediate and future rewards. The value of state s under policy π is given by [23]:

$$V^{\pi}(s) = E_{\pi}[G_m | s_t = s] \tag{22}$$

Q-function (state-action-value function) is the expected return when performing action a in state s, which is denoted as [38]:

$$Q^{\pi}(s, a) = E_{\pi}[G_m | s_t = s, a_t = a] \tag{23}$$

$$Q^{\pi}(s, a) = E_{\pi}\left[\sum_{k=m}^{M} \vartheta^{k-m} r(s_k, a_k) \middle| s_t = s, a_t = a\right] \tag{24}$$

According to the Bellman equation in [23], the value function can be written as the following

$$V^{\pi}(s) = r\left(s, a', s'\right) + \vartheta V^{\pi}\left(s'\right) \tag{25}$$

The expected value of state s is defined as the current rewards and values of the next states. The optimal Q-function $Q^{\pi^*}(s, a)$ for each state and action gives the highest expected return that can be obtained from the state when an action is taken. As a result, it can be given as:

$$Q^{\pi^*}(s, a) = \left(r\left(s, a', s'\right) + \vartheta \max\left(Q^{\pi^*}\left(s', a'\right)\right)\right) \tag{26}$$

The optimal value function $V^{\pi^*}\left(s'\right)$ for each state gives the highest expected return that can be obtained from the state. It can be given as:

$$V^{\pi^*}\left(s'\right) = \vartheta Q^{\pi^*}\left(s', a'\right) \tag{27}$$

The Q value can be updated by taking into account the action that has the maximum Q value of the next time slot:

$$Q\left(s', a'\right) = Q(s, a) + \Omega\left(r\left(s', a'\right) + \vartheta \max Q\left(s', a'\right) - Q(s, a)\right) \tag{28}$$

where $Q\left(s', a'\right)$ is a new Q value of the next state, $Q(s, a)$ is the Q value of the previous state, and Ω is a learning rate or step size. Since exploration policy of the agent, $\in$-greedy policy can be used to define the optimal policy π for more information on the Q-function. It is described as:

$$\pi(a|s) = \begin{cases} randomly\ selected\ from\ A & with\ probaility \in \\ arg\ max_{a \epsilon A} Q(s, a) & otherwise \end{cases} \tag{29}$$

It can observed that the optimal policy $\pi^* = argmax_a(Q^*(s, a))$ is a unique value of the solution. Action $a(m)$ is selected randomly from the action space A and otherwise the action $a(m)$ that maximizes the Q-value is selected with probability $\in$. In contrast, the

agents in the model-based learn a representation of the transition function P and reward function r. The optimal Q-function $Q^{\pi^*}(s,a)$ can be written as [38]:

$$Q^{\pi^*}(s,a) = \sum_{s'} P\left(s,\pi(a),s'\right)\left(r\left(s,a',s'\right) + \vartheta \max\left(Q^{\pi^*}\left(s',a'\right)\right)\right) \tag{30}$$

The expected value $V^{\pi}(s)$ of state s is described as the current rewards and values of the next states weighted by their transition probabilities.

$$V^{\pi}(s) = \sum_{s'} P\left(s,\pi(a),s'\right)\left(r\left(s,a',s'\right) + \vartheta V^{\pi}\left(s'\right)\right) \tag{31}$$

In the proposed method, both model-free and model-based RL algorithms are used to plan the UAV trajectory. Model-free algorithms, such as MC, learn the optimal policy by estimating action values from the agent's interactions with the environment, without requiring a model of the environment. In contrast, model-based algorithms, such as DP, require a model of the environment and use the transition probability to estimate the next reward and next action.

6 Proposed UAV-TO scheme

In this section, the RL algorithm is utilized to plan UAV trajectory planning to increase network resource utilization. The system's state depends on the current location of the UAV and the average of the UAV. The state of the space is denoted as $s(m) = \left(s_1(m), \ldots . s_j(m)\right)$, where $s_j(m)$ represents the state of j UAV at time slot m, which includes the current location $q(m)$ and average LB of UAV $\bar{l}(m)$. Therefore, the state element represents as $s_j(m) = \left\{q_j(m), \bar{l}_j(m)\right\}$. The action of the space is defined as $a(m) = \left(a_1(m), \ldots, a_j(m)\right)$, where $a_j(m)$ represents the action of j UAV at time slot m, which provides the movement direction of UAV. Thus, based on the current state (e.g. location and balancing load), agent makes decision and chooses action according to its policy π, which is depicted in Fig. 3.

The main objective of Q-function is to determine the reward function based on the current location $q(m)$ and average LB of UAV $\bar{l}(m)$. A reward function of ending up in state s(m + 1) after executing action $a(m)$ in state s(m) is denoted $r(s(m), a(m), s(m+1))$. The range of loads takes [0, 1]. Assume that $\bar{l}(0) = 0$. The average load of UAV at time m can be written as:

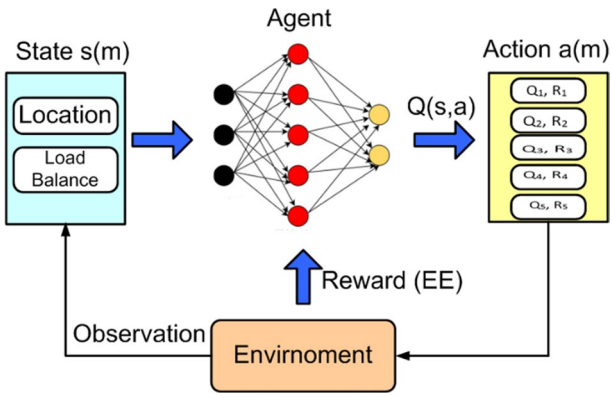

Fig. 3 Proposed scheme based on RL framework

$$\overline{l_j}(m+1) = \overline{l_j}(m) + l_j(m) \tag{32}$$

If $l_j(m) < 0$, it indicates an allowable load of UE and satisfies the LB. Otherwise, the resources of UE could not satisfy the balance of load. The number of resources of UAV $\rho_i(m)$ that are required from i UE is given as follows:

$$\rho_i(m) = \begin{cases} min\frac{R_i}{BW\gamma_{ij}(\mathrm{m})} & \rho_i(m+1) \le N_i - \sum_i \rho_i(m) \\ 0 & \rho_i(m+1) > N_i - \sum_i \rho_i(m) \end{cases} \tag{33}$$

The resources occupied by i UE at time slot (m + 1) may not exceed $\left(\mathrm{N_i} - \sum_{\mathrm{i}} \rho_{\mathrm{i}}(\mathrm{m})\right)$, otherwise $\rho_i(m) = 0$.

Specifically, the highest data rate of UE is transmitted first to get the minimum resources. The action $a_j(m)$ determines the next location of UAV at the next slot $m + 1$. Therefore, the next location can be calculated as:

$$q_j(m+1) = q_j(m) + \beta a_j(m) \tag{34}$$

$$\beta = \begin{cases} 1 & q_j(m+1) - q_j(m)) \le \frac{T}{M} v_{max} \\ 0 & q_j(m+1) - q_j(m)) > \frac{T}{M} v_{max} \end{cases} \tag{35}$$

At each decision point, there are two possible actions for the UAV based on factor β, denoted as 0 and 1. $\beta = 0$ represents the UAV to continue waiting for more UE, while $\beta = 1$ represents the UAV stopping and transferring to another state that provides the best reward.

The reward function P2 can be formula as follows:

$$\text{P2 } r(m) = \frac{\sum_i \sum_t x_{ij}(m) rate_i(m)}{\sum_i \sum_t \left(E_c(m) + E_f(m)\right)} \tag{36}$$

Both model-based and model-free RL approaches are applied to solve the optimization problem. The model-free MC approach is used to design the optimal trajectory of UAV. According to Algorithm 1, the first step is to initialize the Q-value and states. The algorithm starts by resetting time slot m to zero. The optimal Q function determines the reward function (P2) in Eq. (36) and aims to balance the load of the UAV based on the computed value of $\bar{l}(m)$. The objective is to choose the optimal action from an unknown environment. The state space consists of two components: the current location $q(m)$ and average LB of UAV $\bar{l}(m)$. Consequently, the solution of P2 can be obtained by solving Eqs. (32) and (34). The UE with the highest data rate is transmitted first to get the minimum resources. If the load balance is satisfied, i.e., $l_j(m) < 0$, then the resources allocated to the UE are acceptable. However, if $l_j(m) > 0$, then the allocated resources may not be sufficient to satisfy the LB. At each time m, the resources occupied by UE are calculated by using Eq. (33) and should not exceed.the available resources $\left(N_i - \sum_i \rho_i(m)\right)$.

Algorithm 1: the steps to propose UAV-TO scheme for model-free

Inputting: Agent parameters: initial location q_0
System parameter: discount factor ϑ, greedy policy $\in$, step size Ω
Outputting: Trajectory $\pi(a|s)$
Initialize state $s(0)$, Q-value $Q(0)$, $m = 0$
For $m = 1: M$
For $i = 1: I$
Firstly, the highest data rate of UE is transmitted to get the minimum resources
UAV calculates the amount of resources by $\rho_i(m)$

$$\rho_i(m) = \begin{cases} min\ \dfrac{R_i}{BW\gamma_{ij}(m)} & \rho_i(m+1) \le N_i - \sum_i \rho_i(m) \\ 0 & \rho_i(m+1) > N_i - \sum_i \rho_i(m) \end{cases}$$

UAV calculates LB $l_j(m) = \frac{\sum_i \rho_i(m)}{N_i}$
If $\overline{l_j}(m) < 1$ **then**
UAV keeps its current state and waits for more UEs
Else if
UAV calculates reward $r(m)$ for the next action

$$r(m) = \frac{\sum_i \sum_t x_{ij}(m) rate_i(m)}{\sum_i \sum_t \left(E_c(m) + E_f(m)\right)}$$

Update Q value $Q(s,a) \leftarrow Q(s,a) + \Omega\big(r(s',a') + \vartheta \max Q(s',a') - Q(s,a)\big)$

$$a(m) = \begin{cases} randomly\ selected\ from\ A & with\ probaility\ \in \\ arg\ max_{a \epsilon A}\ Q(s,a) & otherwise \end{cases}$$

End If
End for
End for

The agent selects the action that maximizes the Q-value with probability ε according to its policy π, and the chosen action determines the next location of the UAV at the next time slot according to Eq. (34). The factor β in Eq. (35) ensures that UAV should remain within the area for a certain duration T. The Q-value is then updated according to the selected action in order to maximize the Q-value of the next time slot. By repeating these steps, the Algorithm 1 can find the optimal trajectory of the UAV that is suitable for the network environment. On the other hand, the agents in the model-based learn a representation of the transition function P and the reward function r. In the policy iteration method, the optimal solution is based on state-value function in Eq. (31). During the execution, the proposed algorithm constructs greedy action π' that selects actions better than the original policy π. Algorithm 2 is used to find the optimal trajectory of the UAV until a new policy is found that does not improve upon the old policy.

Algorithm 2: the steps to propose UAV-TO scheme for model based

Inputting: Agent parameters: initial location q_0, θ convergence.
System parameter: discount factor ϑ, greedy policy
Outputting: Trajectory $\pi(a)$
Initialize state $s(0)$, Q-value $Q(0)$, $m = 0$, $\Delta=0$
While $\Delta < \theta$
For $m = 1: M$
For $i = 1: I$
Firstly, the highest data rate of UE is transmitted to get the minimum resources
UAV calculates the amount of resources

$$\rho_i(m) = \begin{cases} min\ \frac{R_i}{BW\gamma_{ij}(m)} & \rho_i(m+1) \le N_i - \sum_i \rho_i(m) \\ 0 & \rho_i(m+1) > N_i - \sum_i \rho_i(m) \end{cases}$$

UAV calculates LB $l_j(m) = \frac{\sum_i \rho_i(m)}{N_i}$
If $\overline{l_j}(m) < 1$ **then**
UAV keeps its current state and waits for more UEs
Else if
UAV calculates reward $r(m)$ for the next action

$$r(m) = \frac{\sum_i \sum_t x_{ij}(m) rate_i(m)}{\sum_i \sum_t \left(E_c(m) + E_f(m)\right)}$$

$$V^{\pi}(s) = \sum_{s'} P(s, \pi(a), s')\left(r(s, a', s') + \vartheta V^{\pi}(s')\right)$$

$$\Delta = \max\left(V^{\pi}(s) - V^{\pi}(s)\right)$$

Improve the policy at each state

$$\pi(a) = org\ max_{a \epsilon A} \sum_{s'} P(s, \pi(a), s')\left(r(s, a', s') + \vartheta V^{\pi}(s')\right)$$

End If
End for
End for
End

7 Simulation results

The simulation results evaluate the effectiveness of the proposed scheme UAV-TO, which is optimized to maximize EE under the constraints of LB and the number of resources. Matlab is used to generate the simulation results. It assumes a cell size of (500 m, 500 m) with one MBS located at the center of the cell. The MBS serves UEs located within a distance 500 m and at coordinates (250 m, 250 m). The UAV serves 50 UEs that are distributed randomly between (0 m, 500 m). The UAV starts at position (0, 0, h) and ends at position (500, 500, h). The paper proposes a 3D view of UAV trajectories to visualize the aircraft performance and verify the safety and adaptability of the algorithm. The power consumption for flying and communication is set to

$p_f = 400W$ and $p_c = 40W$, respectively. The maximum speed of the UAV is set as 50 m/s. The flight duration of UAV is T = 120 s.

The simulation parameters are listed in Table 1. The proposed scheme is compared with deployment UAV as Circular scheme [44], Trajectory UAV scheme [21], and Linear scheme [45]. The Circular scheme [44] is defined as the optimal fixed radius being achieved, while the linear scheme is defined as the UAV moving along the linear path. The UAV moved along a linear path from position (0, 0, 100) to position (500, 500, 100) at a constant height of 100 m and a constant velocity of 30 m/s. Figure 4 shows the model of the horizontal path of the UAV. The UAV moved on a horizontal circular path [44] as shown in Fig. 5 with radius 250 m and center at (250, 250, 100) at a constant height of 100 m and a constant velocity of 30 m/s. The UAV completed a single full round starting from position (500, 250, 100).

The efficiency of the UAV-TO scheme is tested using parameters such as EE, LB, flight duration, and number of UEs. The simulation results are divided into three sections; Sect. 1 presents the total EE for various UEs, while Sect. 2 describes the total EE for different heights of UAV. Section 3 includes the LB verse the number of UEs.

Table 1 Simulation parameters

Parameters	Value
σ^2	− 110 dBm
α	0.43
β	4.88
τ	2
ζ^{los}	0.1 dB
ζ^{Nlos}	21 dB
f_c	1 GHz
c	3 × 108 m/s
UAV parameters	
H	100 m
T	120 s
B	1 MHz
P_j	0.1 W
v_{max}	50 m/s
a_{max}	5 m/s^2
c_1	0.001
c_2	2250
g	9.81 m/s^2
UE parameters	
I	50
P_i	0.2 W
R_{min}	0.5 Mb/s
R_i	1 Mb/s
N_i	50

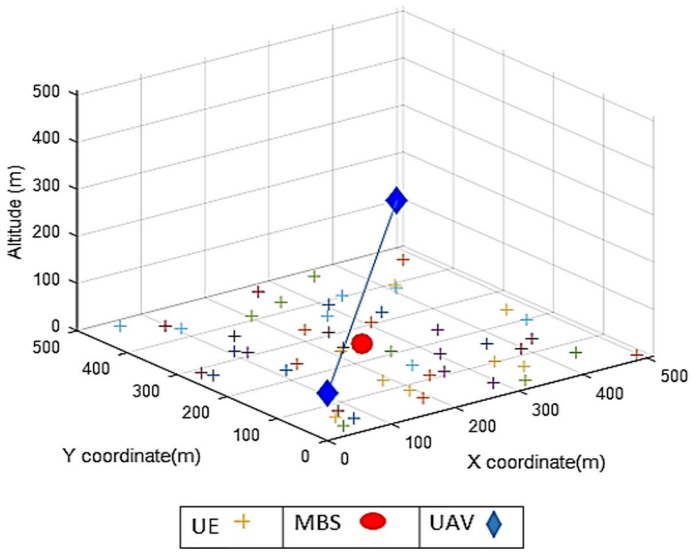

Fig. 4 The horizontal path model of UAV

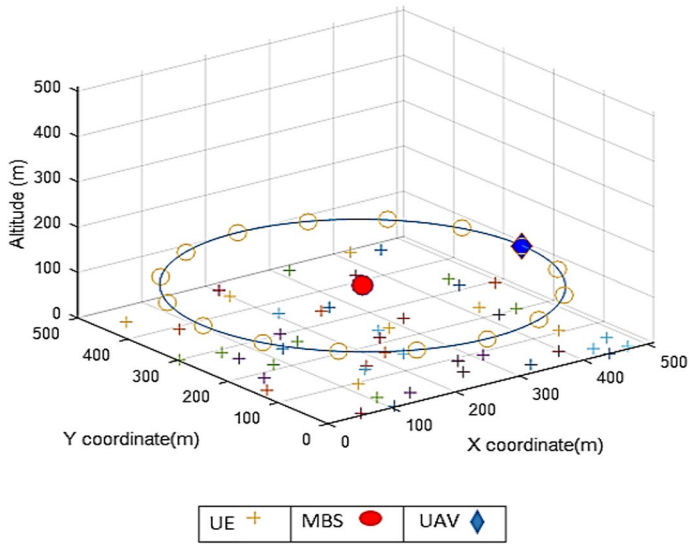

Fig. 5 The circular path model of UAV

7.1 The total EE various different numbers of UE

In this section, two scenarios (MD, DP) of trajectory UAV are proposed for different path loss coefficients such as Suburban, Urban, Dense Urban, and Highrise Urban area. Our goal is to verify the UAV-TO scheme within the cell area to increase network resource utilization based on the RL algorithm. Each UE chooses UAV to balance the load distribution of the cell. In the simulation environment, different path loss coefficients (ζ^{los}, ζ^{Nlos}) pairs (0.1, 21), (1.0,20), (1.6, 23), (2.3, 34) corresponding to Suburban, Urban, Dense Urban, and Highrise Urban respectively [4] (measured in dB). For this purpose, two scenarios of trajectory UAV are tested with five possible states of UAV, as shown in Table 2 and Fig. 6. In this Table, the states and actions of UAV are assumed to

Table 2 Five possible states and actions of UAV

	State 1		State 2		State 3		State 4	
	X	Y	X	Y	X	Y	X	Y
Action 1	150.0000	50.0000	250.0000	50.0000	350.0000	50.0000	450.0000	50.0000
Action 2	142.3880	88.2683	234.7759	126.5367	327.1639	164.8050	419.5518	203.0734
Action 3	120.7107	120.7107	191.4214	191.4214	262.1320	262.1320	332.8427	332.8427
Action 4	88.2683	142.3880	126.5367	234.7759	164.8050	327.1639	203.0734	419.5518
Action 5	50.0000	150.0000	50.0000	250.0000	50.0000	350.0000	50.0000	450.0000

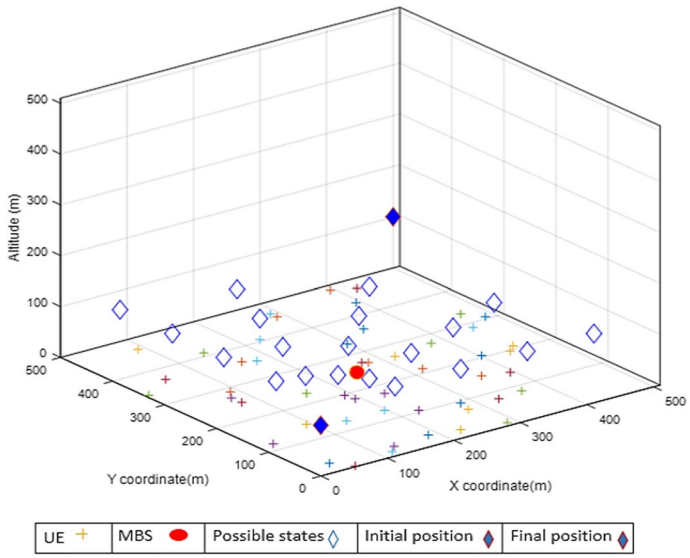

Fig. 6 UAV distributed with five possible states and actions

cover every point in the area. The columns of the Table represent the UAV state, which includes the current location, while the rows represent the possible actions that lead to the next state.

Figure 7 shows the two different scenarios DP and MC of trajectory UAV under various environments when T = 120 s. The initial position of UAV is at (0, 0), while the final position is at (500, 500). Table 3 shows the result of the proposed DP approach for different environments. It indicates a new value function for each state and action. The overall design goal of the reward function is to jointly optimize EE by finding the optimal policy $\pi^*(\mathrm{a}|\mathrm{s})$. From the initial state, the UAV can go to state 1 by finding the value function of five possible actions. As shown in Table 3, the possible actions of state 1 are 2.4851e+07, 2.4905e+07, 2.4772e+07, 2.4848e+07, and 2.4848e+07 for the Suburban environment. The optimal policy $\pi^*(\mathrm{a}|\mathrm{s}) = argmax_a(Q^*(s,a))$ can be obtained by finding action which will lead to the maximum value function. Therefore, the optimal action is action 2 with x = 142.3880 and y = 88.2683 (from Table 2), which is highlighted in green. For example, the optimal action of state 2 is action 1 with x = 250 and y = 50 (from Table 2), which is also highlighted in green. By repeating these steps, the optimal sequence of UAV actions is found to be action 2, action 1, action 3, and action 3 for the Suburban environment.

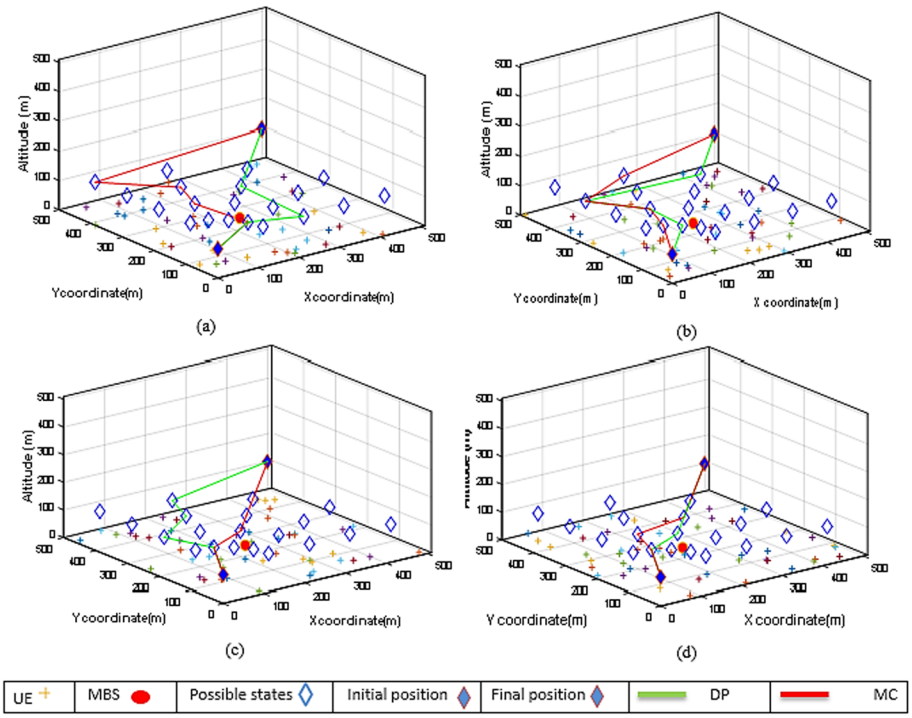

Fig. 7 Trajectory design for DP and MC for different environments. **a** Suburban, **b** Urban, **c** Dense Urban, **d** Highrise

Table 3 The new value function for each state of DP

		State 1	State 2	State 3	State 4
		$V^{\pi}(s)$	$V^{\pi}(s)$	$V^{\pi}(s)$	$V^{\pi}(s)$
Action 1	Suburban	2.4851e+07	**2.2548705e+7**	1.8877e+07	1.1084e+07
	Urban	3.8409e+07	2.898e+07	5.4509e+07	2.1956e+06
	Dense urban	3.8409e+07	1.0313e+07	1.2756e+06	1.5837e+06
	Highrise urban	3.5817e+07	2.2847e+07	2.9118e+07	2.4071e+07
Action 2	Suburban	**2.4905e+07**	1.8784e+07	1.8666e+07	1.495381e+7
	Urban	3.8488e+07	1.4473e+07	1.8257e+07	4.4255e+07
	Dense urban	3.8488e+07	2.0623e+07	1.2865e+07	1.6060e+07
	Highrise urban	3.5058e+07	1.5318e+07	1.9453e+07	3.6423e+07
Action 3	Suburban	2.4772e+07	1.8666e+06	**3.3803e+07**	**9.3204e+06**
	Urban	**3.9521e+07**	1.448e+07	1.8408e+07	**6.6674e+07**
	Dense urban	3.9073e+07	1.0313e+07	3.8757e+07	1.6247e+07
	Highrise Urban	3.4277e+07	**2.2880e+07**	**2.9252e+07**	**3.6677e+07**
Action 4	Suburban	2.4848e+07	1.4975850e+7	2.7950e+07	2.2349e+07
	Urban	3.9073e+07	**4.3534e+07**	3.6638e+07	6.6431e+07
	Dense urban	**3.9521e+07**	2.0568e+07	**3.8993e+07**	**9.7126e+07**
	Highrise urban	**3.5205e+07**	7.6665e+06	9.6957e+06	1.2129e+07
Action 5	Suburban	2.4848e+07	2.2030e+7	1.8901e+07	1.1205e+07
	Urban	3.8769e+07	4.3246e+07	**5.4597e+07**	4.4057e+07
	Dense urban	3.8769e+07	**4.1077e+07**	2.5590e+07	3.1742e+07
	Highrise urban	3.3670e+07	7.644e+06	9.5392e+06	1.2015e+06

The bold defines the optimal action for each state

The optimal actions are highlighted in green, orange, blue, and yellow for the Suburban, Urban, Dense Urban, and Highrise Urban environments, respectively

Table 4 The optimal trajectory path of the UAV for DP

	Optimal trajectory path of UAV			
	State 1	State 2	State3	State 4
Suburban	Action 2	Action 1	Action 3	Action 3
Urban	Action 3	Action 4	Action 5	Action 3
Dense urban	Action 4	Action 5	Action 4	Action 4
Highrise urban	Action 4	Action 3	Action 3	Action 3

The optimal actions are highlighted in green, orange, blue, and yellow for the Suburban, Urban, Dense Urban, and Highrise Urban environments, respectively

Table 5 The optimal trajectory path of the UAV for MC

	Optimal trajectory path of UAV			
	State 1	State 2	State3	State 4
Suburban	Action 2	Action 4	Action 4	Action 5
Urban	Action 4	Action 4	Action 4	Action 5
Dense urban	Action 4	Action 3	Action 3	Action 3
Highrise urban	Action 4	Action 4	Action 3	Action 3

The optimal actions are highlighted in green, orange, blue, and yellow for the Suburban, Urban, Dense Urban, and Highrise Urban environments, respectively

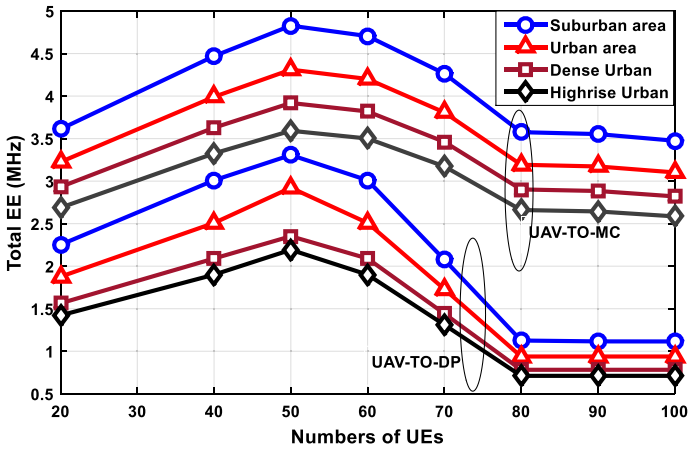

Fig. 8 Total EE versus different number of UEs

Tables 4 and 5 show the optimal trajectory path of the UAV for DP and MC, respectively, which corresponds to Fig. 7. Accordingly, the rows represent the actions that have maximum Q value under various area coefficients. From Table 5, it can be seen that MC must wait until the end of episode to receive the reward.

The proposed scheme is compared with deployment UAV as Circular scheme [44], Trajectory UAV scheme [21], and Linear scheme [45] for balancing the load of trajectory UAV. Figure 8 represents the total EE verse number of UEs for four different environments. As shown in Fig. 8, the EE of the UAV is the largest for the Suburban environment. While EE of the UAV is the smallest for other environments. It can be observed that the proposed UAV-TO for MC (UAV-TO-MC) provides higher EE

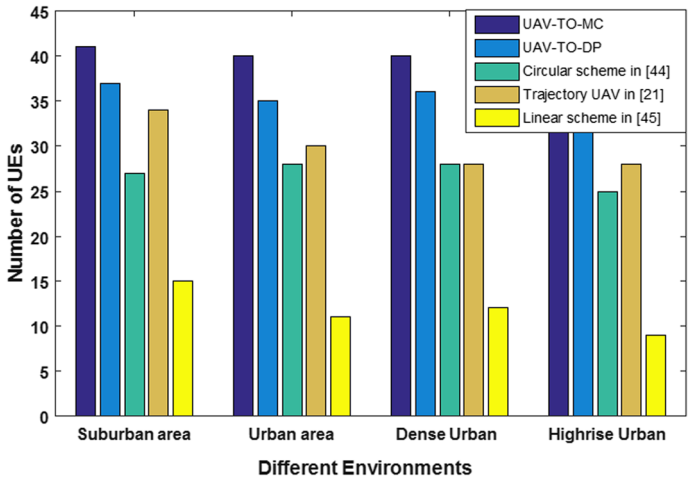

Fig. 9 Number of UEs served by UAV (50UEs). Circular scheme in [44], Trajectory UAV in [21], Linear scheme in [45]

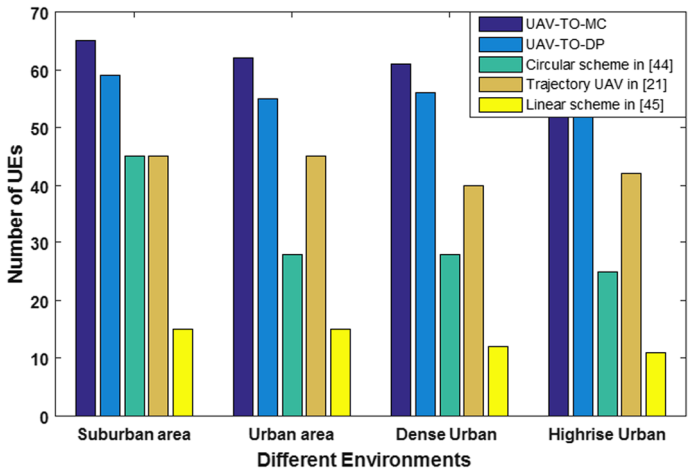

Fig. 10 Number of UEs served by UAV (80UEs). Circular scheme in [44], Trajectory UAV in [21], Linear scheme in [45]

than the proposed UAV-TO for DP (UAV-TO-DP). Figures 9 and 10 are bar graphs show the load performance when UEs are distributed in area for four different environments (Suburban, Urban, Dense Urban, Highrise). It can be observed that the proposed UAV-TO scheme serves more UEs compared to other schemes due to its Circular deployment scheme [44], which has limited coverage area. Additionally, the proposed UAV-TO-MC achieves a better loading balance by serving more UEs through UAV's trajectory. Therefore, the Suburban area is chosen for the remaining results as it has a high EE for all schemes. Figures 11 and 12 show the EE versus flight duration T for different numbers of UE, with UAV having different trajectory designs over different flight durations. The EE of UAV-TO-MC improves as the duration increases since more UEs are allocated to the UAV. In contrast, UAV-TO-DP achieves low EE and fails to utilize the available network resources due to its explicit model of

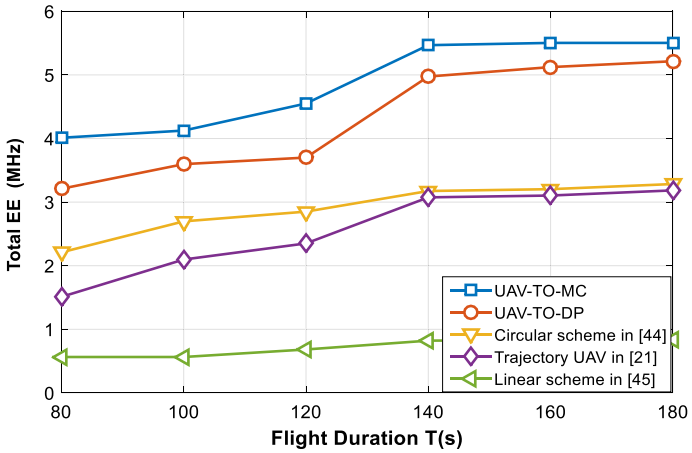

Fig. 11 EE verse flight duration (80UEs). Circular scheme in [44], Trajectory UAV in [21], Linear scheme in [45]

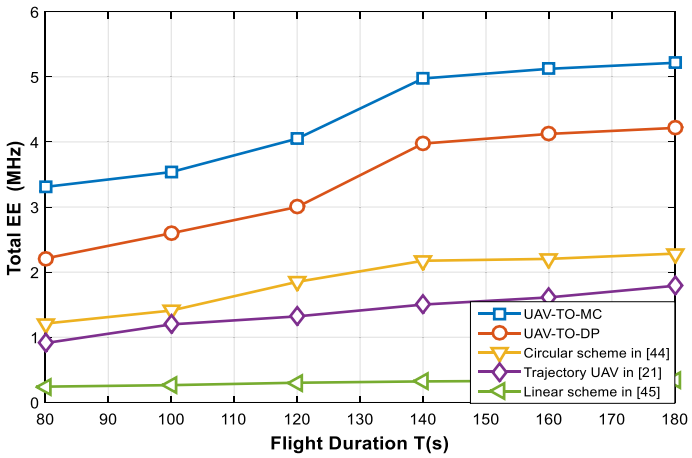

Fig. 12 EE verse flight duration (50UEs). Circular scheme in [44], Trajectory UAV in [21], Linear scheme in [45]

the environment. MC is more efficient in terms of experience, while DP is less efficient in terms of exploitation. As is expected, the proposed scheme UAV-TO-MC achieves the best performance as it optimizes the UAV trajectory for load balancing.

7.2 The total EE for different heights of UAV

In this section, the effect of UAV trajectory design on total EE for different UAV heights is studied. To verify the performance of the proposed UAV-TO, it is compared with other schemes, including Circular scheme [44], Trajectory UAV scheme [21], and Linear scheme [45], to maximize total EE while considering load balancing. Figure 13 shows the total EE as a function of the height of UAV, and it can be seen that MC achieves the maximum of EE. DP requires perfect knowledge of the environment

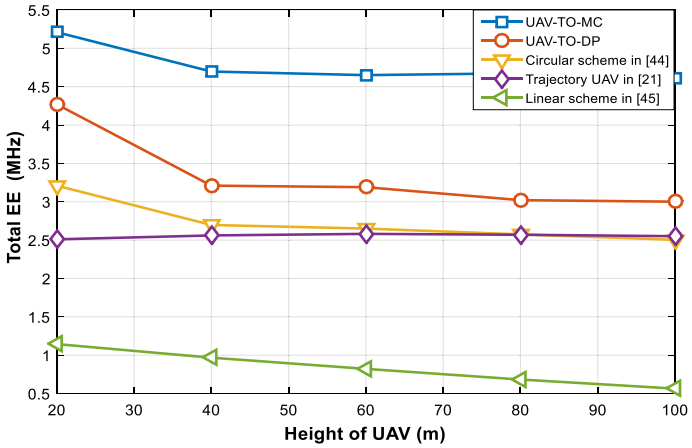

Fig. 13 The total EE verse height of UAV. Circular scheme in [44], Trajectory UAV in [21], Linear scheme in [45]

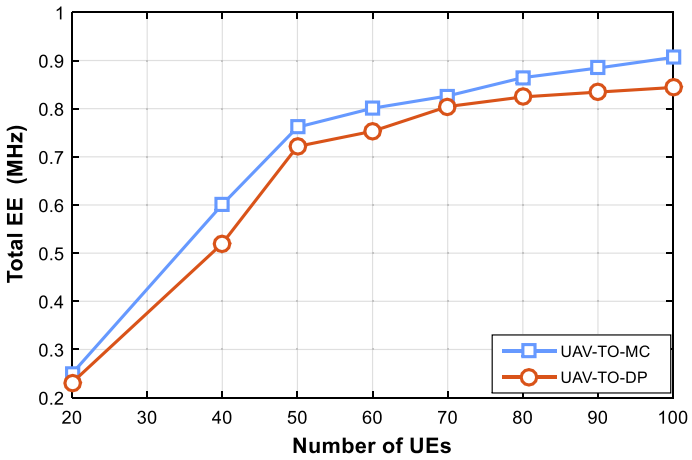

Fig. 14 LB corresponding the number of UEs

and a large amount of memory to store the problem, while MC does not require any prior knowledge and modeling assumptions.

7.3 The LB corresponding to the number of UEs

The LB corresponding to the number of UEs is presented in Fig. 14. It can be observed that both curves increase monotonously with the increase in the number of UEs. It is expected that with increasing UEs, the network will become overloaded to satisfy the requirements for both DP and MC. Obviously, the LB of DP is the biggest, and the LB of MC is the smallest.

8 Conclusion

This paper proposes a UAV-TO scheme for load balancing based on RL. The proposed scheme utilized LB to maximize EE for multiple UEs and improve network resource utilization. It is considered a 3D flight trajectory of UAV to visualize the aircraft performance and verify the safety and adaptability of the algorithm. Since the problem is modeled as nonconvex optimization, RL is utilized for UAV trajectory planning. The proposed scheme was applied for both MC and DP to solve the optimization problem under the LOS and NLOS channel models. Additionally, the network load distribution is calculated. The simulation results demonstrate the performance of the proposed scheme under different path losses and different flight durations. The results show that the proposed scheme outperforms the existing methods under various parameter configurations.

Abbreviations

AF	Amplify-and-forward
BS	Base station
CNN	Convolutional neural network
DF	Decode-and-forward
DP	Dynamic programming
EE	Energy efficiency
GT	Ground terminals
GNN	Graph neural network
G2U	Ground-to-UAV
LB	Load balance
LOS	Line of sight
MBS	Macro base station
MC	Monte Carlo
MDP	Markov decision process
NLOS	Non-line of sight
SDG	Sustainable development goal
SN	Sensor node
SNR	Signal to noise ratio
TD	Temporal difference
UAV	Unmanned aerial vehicle
UE	User equipment
U2G	UAV-to-ground

Author contributions

All Authors contributed equally to this work and approved the final manuscript.

Funding

Open access funding provided by The Science, Technology & Innovation Funding Authority (STDF) in cooperation with The Egyptian Knowledge Bank (EKB). The research received no external funding.

Availability of data and materials

Not applicable.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 2 January 2023 Accepted: 20 June 2023
Published online: 01 July 2023

References

1. S.G. Gupta, M.M. Ghonge, P. Jawandhiya, Review of unmanned aircraft system (UAS). Int. J. Adv. Res. Comput. Eng. Technol. (IJARCET) **2**, 1646–1658 (2013)
2. A.A. Laghari, A.K. Jumani, R.A. Laghari, H. Nawaz, Unmanned aerial vehicles: a review. Cogn. Robot. **3**, 8–22 (2023)

3. J. Yang, J. Qian, H. Gao, Forest wildfire monitoring and communication UAV system based on particle swarm optimization. Journal of Physics: Conference Series, 2021 2nd International Conference on Artificial Intelligence and Information Systems (ICAIIS 2021), vol. 1982 (2021)
4. S.K. Khan, M. Farasat, U. Naseem, F. Ali, Performance evaluation of next-generation wireless (5G) UAV relay. Wirel. Pers. Commun. **113**, 945–960 (2020)
5. S. Ahmed, M.Z. Chowdhury, Y.M. Jang, Energy-efficient UAV-to-user scheduling to maximize throughput in wireless networks. Inst. Electr. Electron. Eng. (IEEE) **8**, 21215–21225 (2020)
6. E. Larsen, L. Landmark, O. Kure, Optimal UAV relay positions in multi-rate networks. Wireless Days Conference, pp. 8–14 (2017)
7. F. Cheng, S. Zhang, Z. Li, Y. Chen, N. Zhao, F. Richard Yu, V.C.M. Leung, UAV trajectory optimization for data offloading at the edge of multiple cells. IEEE Transactions on Vehicular Technology, vol. **67**, pp. 6732–6736 (2018)
8. Z. Rahimi, M.J. Sobouti, R. Ghanbari, S.A.H. Seno, A.H. Mohajerzadeh, H. Ahmadi, H. Yanikomeroglu, An efficient 3D positioning approach to minimize required UAVs for IoT network coverage. IEEE Internet Things J. 558–571 (2021)
9. S. Yin, J. Tan, L. Li, UAV-assisted cooperative communications with wireless information and power transfer. Netw. Internet Archit. 1–31 (2017)
10. D. Huang, M. Cui, G. Zhang, X. Chu, F. Lin, Trajectory optimization and resource allocation for UAV base stations under in-band backhaul constraint. EURASIP J. Wirel. Commun. Netw. **2020**, 1–17 (2020)
11. M.A. Sayeed, R. Kumar, V. Sharma, M.A. Sayeed, Efficient deployment with throughput maximization for UAVs communication networks. Sensors **20**, 1–27 (2020)
12. X. Fu, T. Ding, R. Peng, C. Liu, M. Cheriet, Joint UAV channel modeling and power control for 5G IoT networks. EURASIP J. Wirel. Commun. Netw. **2021**, 1–15 (2021)
13. B. Liu, H. Zhu, Energy-effective data gathering for UAV-aided wireless sensor networks. Sensors. 1–12 (2019)
14. X.A.F. Cabezas, D.P.M. Osorio, M. Latva-aho, Positioning and power optimization for UAV-assisted networks in the presence of eavesdroppers: a multi-armed bandit approach. EURASIP J. Wirel. Commun. Netw. **2022**, 1–24 (2022)
15. X. Fu, T. Ding, R. Peng, C. Liu, M. Cheriet, Joint UAV channel modeling and power control for 5G IoT networks. EURASIP J. Wirel. Commun. Netw. **2021**, 1–15 (2021)
16. S.R. Pandey, K. Kim, M. Alsenwi, Y.K. Tun, Z. Han, C.S. Hong, Latency-sensitive service delivery with UAV-assisted 5G networks. IEEE Wirel. Commun. Lett. **10**, 1518–1522 (2021)
17. H. Yang, J. Zhao, J. Nie, N. Kumar, K.Y. Lam, Z. Xiong, UAV-assisted 5G/6G networks: joint scheduling and resource allocation based on asynchronous reinforcement learning, in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications Workshops* (2021)
18. I. Ahmad, J. Kaur, H.T. Abbas, Q.H. Abbasi, A. Zoha, M.A. Imran, S. Hussain, UAV-assisted 5G networks for optimised coverage under dynamic traffic load. in *2022 IEEE International Symposium on Antennas and Propagation and USNC-URSI Radio Science Meeting (AP-S/URSI)* (2022)
19. R. Shahzadi, M. Ali, H.Z. Khan, M. Naeem, UAV assisted 5G and beyond wireless networks: a survey. J. Netw. Comput. Appl. **189**, 1–20 (2021)
20. L. Zhang, A. Celik, S. Dang, B. Shihada, Energy-efficient trajectory optimization for UAV-assisted IoT networks. IEEE Trans. Mobile Comput. **21**, 4323–4337 (2021)
21. G. Zhang, H. Yan, Y. Zeng, M. Cui, Y. Liu, Trajectory optimization and power allocation for multi-hop UAV relaying communications. IEEE Access. **6**, 48566–48576 (2018)
22. G. Zhang, Q. Wu, M. Cui, R. Zhang, Securing UAV communications via joint trajectory and power control. IEEE Trans. Wirel. Commun. **18**, 1376–1389 (2019)
23. Y. Zeng, R. Zhang, Energy-efficient UAV communication with trajectory optimization. IEEE Trans. Wirel. Commun. **16**, 3747–3760 (2017)
24. S. Nasrollahi, S.M. Mirrezaei, Toward UAV-based communication: improving throughput by optimum trajectory and power allocation. EURASIP J. Wirel. Commun. Netw. **2022** (2022)
25. J. Gu, G. Ding, Y. Xu, H. Wang, Q. Wu, Proactive optimization of transmission power and 3D trajectory in UAV-assisted relay systems with mobile ground users. Chin. J. Aeronaut. **34**(3), 129–144 (2021)
26. X. Jiang, Z. Wu, Z. Yin, Z. Yang, Power and trajectory optimization for UAV-enabled amplify-and-forward relay networks. IEEE Access **4**, 1–9 (2016)
27. A. Salah, H. Abd Elatty, R.Y. Rizk, Joint channel assignment and power allocation based on maximum concurrent multi-commodity flow in cognitive radio networks. Wirel. Commun. Mobile Comput. **2018**, 1-14 (2018)
28. D. Zhai, H. Li, X. Tang, R. Zhang, H. Cao, Joint position optimization, user association, and resource allocation for load balancing in UAV-assisted wireless networks. Digit. Commun. Netw. 1–13 (2022)
29. Z. Luan, H. Jia, P. Wang, R. Jia, B. Chen, Joint UAVs' load balancing and UEs' data rate fairness optimization by diffusion UAV deployment algorithm in multi-UAV networks. Entropy **23**, 1470–1489 (2021)
30. Q. Fan, N. Ansari, Towards traffic load balancing in drone-assisted communications for IoT. IEEE Internet Things J. **6**, 3633–3640 (2019)
31. L. Yang, H. Yao, J. Wang, C. Jiang, A. Benslimane, Y. Liu, Multi-UAV enabled load-balance mobile edge computing for IoT networks. IEEE Internet Things J. **7**, 1–12 (2020)
32. J.-H. Cui, R.-X. Wei, Z.-C. Liu, K. Zhou, UAV motion strategies in uncertain dynamic environments: a path planning method based on Q-learning strategy. Appl. Sci. **8**, 1–16 (2018)
33. A.T. Azar, A. Koubaa, N.A. Mohamed, H.A. Ibrahim, Z.F. Ibrahim, M. Kazim, A. Ammar, B. Benjdira, A.M. Khamis, I.A. Hameed, G. Casalino, Drone deep reinforcement learning: a review. Electronics **10**, 1–30 (2021)
34. K. Karwowska, D. Wierzbicki, Improving spatial resolution of satellite imagery using generative adversarial networks and window functions. Remote Sens. **14**, 1–22 (2022)
35. S.M. Mousavi, Improving quality of images in UAVs navigation using super-resolution techniques based on convolutional neural network with multi-layer mapping. Marine Technol. **4**, 1–11 (2017)
36. E. Balasubramanian, E. Elangovan, P. Tamilarasan, G.R. Kanagachidambaresan, D. Chutia, Optimal energy efficient path planning of UAV using hybrid MACO-MEA* algorithm: theoretical and experimental approach. J. Ambient Intell. Humaniz. Comput. 1–22 (2022)

37. G. Kalnoor, G. Subrahmanyam, A review on applications of Markov decision process model and energy efficiency in wireless sensor networks. Proc. Comput. Sci. **167**, 2308–2317 (2020)
38. M. Abualsheikh, D.T. Hoang, D. Niyato, H.-P. Tan, S. Lin, Markov decision processes with applications in wireless sensor networks: a survey. IEEE Commun. Surv. Tutorials **17**, 1239–1267 (2015)
39. N. Safwat, I.M. Hafez, F. Newagy, UGPL: a MATLAB application for UAV-to-ground path loss calculations. Softw. Impacts. **12** (2022)
40. S.M.M. AboHashish, R.Y. Rizk, F.W. Zaki, Energy efficiency optimization for relay deployment in multi-user LTE-advanced networks. Wirel. Pers. Commun. **108**, 297–323 (2019)
41. B.S. Roh, M.H. Han, J.H. Ham, K.-I. Kim, Q-LBR: Q-learning based load balancing routing for UAV-assisted VANET. Sensors **20**, 1–18 (2020)
42. S.M.M. AboHashish, R.Y. Rizk, F.W. Zaki, Towards energy efficient relay deployment in multi-user LTE-A networks. IET Commun. **13**, 2688–2696 (2019)
43. P.D. Thanh, T.H. Giang, T.N.K. Hoan, I. Koo, Cache-enabled rata rate maximization for solar-powered UAV communication systems. Electronics **9**, 1–28 (2020)
44. O.M. Bushnaq, M.A. Kishk, A. Çelik, M.-S. Alouini, T.Y. Al-Naffor, Optimal deployment of tethered drones for maximum cellular coverage in user clusters. IEEE Trans. Wirel. Commun. **20**, 2092–2108 (2021)
45. Ch. Zhan, Y. Zeng, R. Zhang, Energy-efficient data collection in UAV enabled wireless sensor network. IEEE Wirel. Commun. Lett. **7**, 328–331 (2018)

Publisher's Note