# Multi-UAV Trajectory Design and Power Control Based on Deep Reinforcement Learning

Chiya Zhang, Shiyuan Liang, Chunlong He, Kezhi Wang

***Abstract*—In this paper, multi-unmanned aerial vehicle (multi-UAV) and multi-user system are studied, where UAVs are served as aerial base stations (BS) for ground users in the same frequency band without knowing the locations and channel parameters for the users. We aim to maximize the total throughput for all the users and meet the fairness requirement by optimizing the UAVs' trajectories and transmission power in a centralized way. This problem is non-convex and very difficult to solve, as the locations of the user are unknown to the UAVs. We propose a deep reinforcement learning (DRL)-based solution, i.e., soft actor-critic (SAC) to address it via modeling the problem as a Markov decision process (MDP). We carefully design the reward function that combines sparse with non-sparse reward to achieve the balance between exploitation and exploration. The simulation results show that the proposed SAC has a very good performance in terms of both training and testing.**

***Keywords*—multi-UAV and multi-user wireless system, UAV, power control, trajectory design, throughput maximization, SAC**

Manuscript received Mar. 10, 2022; revised May 01, 2022; accepted May 24, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62101161, in part by Shenzhen Basic Research Program under Grant 20200811192821001 and Grant JCYJ20190808122409660, in part by Guangdong Basic Research Program under Grant 2019A1515110358, Grant 2021A1515012097, Grant 2020ZDZX1037, Grant 2020ZDZX1021, and in part by the open research fund of National Mobile Communications Research Laboratory, Southeast University under Grant 2021D16 and Grant 2022D02. The associate editor coordinating the review of this paper and approving it for publication was J. Xu.

C. Y. Zhang. School of Electronic and Information Engineering, Harbin Institute of Technology, Shenzhen 518055, China. Peng Cheng Laboratory (PCL), Shenzhen 5180523, China (e-mail: chiya.zhang@foxmail.com).

S. Y. Liang, C. L. He. Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen 518060, China (e-mail: 2070436053@email.szu.edu.cn; hclong@szu.edu.cn).

C. Y. Zhang, C. L. He. National Mobile Communications Research Laboratory, Southeast University, Nanjing 211189, China (e-mail: chiya.zhang@foxmail.com; hclong@szu.edu.cn).

K. Z. Wang. School of Computer Sciences and Electrical Engineering, Northumbria University, 5995 Newcastle upon Tyne, Newcastle, United Kingdom of Great Britain and Northern Ireland, UK NE1 8ST, UK (e-mail: kezhi.wang@northumbria.ac.uk).

## I. Introduction

In recent years, unmanned aerial vehicles (UAV) have attracted significant attention in various fields. For example, it can be used for pesticide spraying and crop monitoring in the agricultural and searching for people that are trapped. In particular, UAVs have been extensively investigated in wireless communication serving as aerial base stations (BS)[1-7], mobile relays[8-10], mobile edge computing[11,12] and wireless power transfer[13,14]. UAVs can enhance the probability of line-of-sight (LOS) links and reliably communicate with users by dynamically adjusting their locations[15]. There are mainly two different types of studies called static-UAV and mobile-UAV for wireless communications.

For static-UAV, or quasi-stationary, the altitude or horizontal position of the UAV can be optimized to meet different quality of service (QoS) requirements. In Ref. [16], it maximized the convergence by optimizing the height of UAV given the UAV's horizontal position. In Ref. [2], by fixing the UAV's altitude, it optimized the UAVs' horizontal positions to minimize the number of UAVs which can meet the communication service of a given number of users.

On the mobile-UAV side, the UAV can be more flexibly deployed, e.g., for emergency cases[17]. In the past years, convex optimization algorithms have been used to optimize the trajectory of UAV. In Ref. [8], the studies have considered a mobile relay, which has a more significant throughput gain than traditional static relaying. Ref. [18] has studied maximizing UAV throughput by optimizing UAV trajectory, UAV transmit power, and UAV-to-user scheduling, while considered the flight energy consumption of the UAV. In Ref. [19], the authors considered a multi-user communication system and proposed a novel cyclical time-division-multiple-access (TDMA) protocol. And the authors have also shown that, for delay-tolerant applications, throughput gains can be significantly increased in static-UAV system. Also, the authors in Ref. [19] considered a single-UAV communication system, where one UAV flies with the constant speed and the ground users are assumed to be uniformly located in a one-dimensional (1D)

line. In Ref. [3], the authors considered a multi-UAV and multi-user system, where multiple UAVs are served as aerial BSs for ground users in the same frequency band.

For most of the above works, the traditional algorithms such as block coordinate descent (BCD) and successive convex approximation (SCA), are applied to obtain the optimal or suboptimal solutions, e,g., trajectory design and resource allocation. However, the traditional algorithms require plenty of computational resources and take much time[20]. To address this problem, many studies are proposed to adopt deep reinforcement learning (DRL) to solve joint trajectory design and power allocation (JTDPA) of the UAV[21-28]. The authors in Ref. [21] have considered a single-UAV communication system, which optimizes 3D UAV trajectory and band allocation to maximize the throughput. In Ref. [22], the authors considered a multi-UAV and multi-user communication system and proposed a distributed multi-agent DRL framework, where each agent makes the best decision following its individual policy through a try-and-error learning process. They aimed to reduce the number of user's handover to maximize the total throughput, where the UAV-BSs fly in a pre-designed mobility model. The studies in Ref. [24] optimized UAV trajectory to maximize the energy efficiency, which considers communication coverage, equity, and energy consumption. In Ref. [23], by using multi-agent to solve JTDPA, the authors considered the UAVs communication on the same frequency band and the agent may not know the users' locations and channel parameters in advance.

In this paper, we develop a novel multi-UAV-enabled wireless communication system. For security and practicability, the UAVs don't know the information of users' locations and channel parameters, they only know the achievable rate of users. We assume that the UAVs need to return to initial point and aim to maximize the total throughput for all the users as well as keeping the fairness in communication by optimizing the UAV's trajectory and transmission power in a centralized way. The means of fairness is all ground users can get minimum total throughput. Our contributions are as follows.

(1) We develop a novel wireless communication system and model it as an MDP and propose a DRL-based framework to solve it.

(2) We carefully design the reward function that combines sparse with non-sparse reward to achieve communication fairness. We also compare our solution with other DRL algorithms to show the performance improvement.

The rest of this paper is arranged as follows. In section II, we introduce the multi-UAV communication system. Section III introduces the principle of soft actor-critic (SAC)[29,30] algorithm and the framework to solve the trajectory optimization problem. In section IV, we present the simulation results and then we conclude the paper in section V.

Notations: In this paper, we use italic letters to represent scalars, bold-face to represent vectors. For a vectors $\boldsymbol{S}$, $|\boldsymbol{S}|_{\text{dim}}$ means the dimensional of $\boldsymbol{S}$.

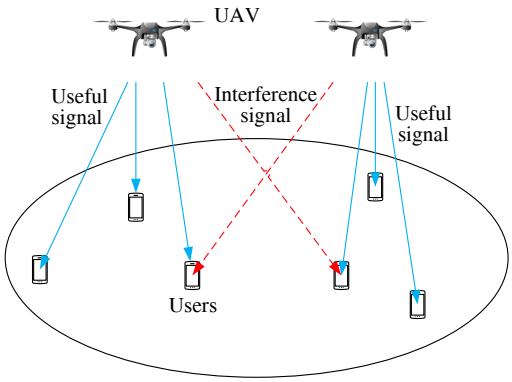



**Fig. 1** The system model

## II. System Model and Problem Formulation

### A. System Model

We consider a wireless communication system contains $M$ UAVs as BSs and $N$ ground users in Fig. 1. The UAV is represented by $d_m \in \mathscr{M}, 1 \leqslant m \leqslant M$ and the user is represented by $u_n \in \mathscr{N}, 1 \leqslant n \leqslant N$. Generally, $N > M$, means that the system is an information broadcast system enabled by UAVs. We assume that the UAVs do not have the location information of the grounds users, but they have the received signal strength indicator (RSSI) which is the total achievable rate of the ground user. In this communication system, the UAVs, on the same frequency band, communicate with ground users over consecutive periods $T > 0$. Then, the ground users communicate with one of the UAVs using TDMA protocol. We consider that the UAV flies at fixing altitude $H$, so that the UAV trajectory can be expressed as $\boldsymbol{q}_m(t) = [x_m(t), y_m(t)]^{\mathrm{T}} \in R^{2\times 1}$, $1 \leqslant m \leqslant M$. we assume that the time interval is $\delta_t = 1$ s, where the UAV completes the flight and communication process in $\delta_t$, and the UAVs need to return the origin point after completing the mission.

### B. Problem Formulation

The coordinate of the $n$th user is represented by $\boldsymbol{u}_n = [x_n, y_n]^{\mathrm{T}} \in R^{2\times 1}, 1 \leqslant n \leqslant N$. Note that $\boldsymbol{u}_n$ may not be accessible for the UAVs. Note that no matter what kind of channel model we adopt, our proposed algorithm only needs RSSI of the ground user. We adopt a typical channel model that the UAV communicates with the group user by using the LoS links, the Doppler shift caused by UAV movement can be compensated at the receivers, and the communication link is only related to the distance between the user and UAV as

$$h_{m,n}[t] = \rho_0 d_{m,n}^{-2}[t] = \frac{\rho_0}{H^2 + \|\boldsymbol{q}_m[t] - \boldsymbol{u}_n\|^2}, \tag{1}$$

where $\rho_0$ denotes the channel power at the reference distance $d_0 = 1$ m. We define the downlink UAV transmission power as $p_m[t]$, then we have,

$$0 \leqslant p_m[t] \leqslant P_{\max}. \tag{2}$$

If the $m$th UAV communicates with $n$th user in time slot $t$, the signal-to-interference-plus-noise ratio (SINR) at $n$th user is

$$\gamma_{m,n}[t] = \frac{p_m[t]h_{m,n}[t]}{\sum_{j=1,j\neq m}^{M} p_j[t]h_{j,n}[t] + \sigma^2}, \tag{3}$$

the denominator of (3) represents the noise $\sigma^2$ which is the power of the additive white Gaussian noise (AWGN) and the interference signal by other UAVs in time slot $t$.

We define a group of binary variables as $a_{m,n}[t]$, where $a_{m,n}[t] = 1$ denotes the $n$th user is served by $m$th UAV in time slot $t$, otherwise, $a_{m,n}[t] = 0$. We assume that each user can communicate with only one UAV and each UAV can only serve one user at any time, which as following

$$\sum_{n=1}^{N} a_{m,n}[t] \leqslant 1, \quad \forall m,t, \tag{4}$$

$$\sum_{m=1}^{M} a_{m,n}[t] \leqslant 1, \quad \forall n,t, \tag{5}$$

$$a_{m,n}[t] \in \{0,1\}, \quad \forall m,n,t, \tag{6}$$

We assume that UAV communicates with the nearest ground user. Moreover, for UAVs trajectory, the maximum flying distance of the UAV in the interval time is $S_{\max}$, and UAV needs flying back to origin point by the end of each period $T$, then we have,

$$\boldsymbol{q}_m[1] = \boldsymbol{q}_m[T], \ \forall m, \tag{7}$$

$$\|\boldsymbol{q}_m[t+1] - \boldsymbol{q}_m[t]\| \leqslant S_{\max}^2, \ t = 1,\cdots,T-1, \quad \forall m, \tag{8}$$

$$\|\boldsymbol{q}_m[t] - \boldsymbol{q}_j[t]\|^2 \geqslant d_{\min}^2, \tag{9}$$

where (9) represents the collision constraints of UAV. We Set the maximum speed of the UAV motion as $V_{\max}$, and $S_{\max} = \delta_t \times V_{\max}$.

Hence, the achievable rate of $n$th user in time slot $t$ can be expressed as

$$R_n[t] = \sum_{m=1}^{M} a_{m,n}[t]\mathrm{lb}(1+\gamma_{m,n}[t]). \tag{10}$$

The UAV-user association, trajectory and transmission power of the UAVs are denoted as $\boldsymbol{A} = \{a_{m,n}[t], \forall m,n,t\}$, $\boldsymbol{Q} = \{\boldsymbol{q}_m[t], \forall m,t\}$, *and* $\boldsymbol{P} = \{p_m[t], \forall m,t\}$, respectively. For fairness, we set the minimum value $R_{\min}$ for each user. We aim to maximize the total throughput for all users by optimizing the UAV-user association $\boldsymbol{A}$, UAV trajectory $\boldsymbol{Q}$, and UAV transmission power $\boldsymbol{P}$ over all time slots. The optimization problem is formulated as

$$\max_{\boldsymbol{A},\boldsymbol{Q},\boldsymbol{P}} \sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{t=1}^{T} a_{m,n}[t]\mathrm{lb}(1+\gamma_{m,n}[t]) \tag{11}$$

$$\text{s.t.} \sum_{t=1}^{\mathrm{T}}\sum_{m=1}^{M} a_{m,n}[t]\mathrm{lb}(1+\gamma_{m,n}[t]) \geqslant R_{\min}, \quad \forall n, \tag{11a}$$

$$\sum_{n=1}^{N} a_{m,n}[t] \leqslant 1, \quad \forall m,t, \tag{11b}$$

$$\sum_{m=1}^{M} a_{m,n}[t] \leqslant 1, \quad \forall n,t, \tag{11c}$$

$$a_{m,n}[t] \in \{0,1\}, \quad \forall m,n,t, \tag{11d}$$

$$\|\boldsymbol{q}_m[t+1] - \boldsymbol{q}_m[t]\| \leqslant S_{\max}^2, \quad t = 1,\cdots,T-1, \forall m, \tag{11e}$$

$$\boldsymbol{q}_m[1] = \boldsymbol{q}_m[T], \quad \forall m, \tag{11f}$$

$$\|\boldsymbol{q}_m[t] - \boldsymbol{q}_j[t]\|^2 \geqslant d_{\min}^2, \quad \forall m,t,j \neq m, \tag{11g}$$

$$0 \leqslant p_m[t] \leqslant P_{\max}, \quad \forall m,t. \tag{11h}$$

Problem (11) cannot be solved by traditional algorithms, as the UAV does not have the users' locations and channel parameters. To solve the problem, we model the communication system as the Markov decision process (MDP) and address it by DRL, where the UAVs act as the agent and learn their optimal policies of trajectory and power.

## III. Soft Actor-Critic

In this section, SAC is adopted to solve the joint trajectory optimization problem. SAC has been proposed by Ref. [29]. SAC belongs to the off-line reinforcement learning, which can transfer the learned model to UAV after learning the previous experience without on-line process. At the same time, compared with other traditional machine-learning-based solutions, e.g., deep deterministic policy gradient (DDPG) and twin delayed DDPG (TD3), it is more stable, where the results are similar for different random seeds. Moreover, the action space is continuous, which leads to a more flexible trajectory.

### A. Preliminaries

In order to facilitate the following description, we give a brief introduction to SAC. Firstly, MDP can be defined by a tuple $(\boldsymbol{S},\boldsymbol{C},\boldsymbol{P},\boldsymbol{Z})$. The state transition probability $p$ satisfies $\boldsymbol{S}\times\boldsymbol{C}\times\boldsymbol{S} \to [0,+\infty)$, which means that the current state $\boldsymbol{s}_t \in \boldsymbol{S}$ and the current action $\boldsymbol{c}_t \in \boldsymbol{C}$ are given, and $\boldsymbol{s}_{t+1} \in \boldsymbol{S}$ is obtained. Define the environment reward as $z : \boldsymbol{S}\times\boldsymbol{C} \to [z_{\min}, z_{\max}]$ on each transition. Also, we define $\rho_\pi(\boldsymbol{s}_t)$ and $\rho_\pi(\boldsymbol{s}_t,\boldsymbol{c}_t)$ respectively as the state and action distribution under the policy $\pi(\boldsymbol{c}_t|\boldsymbol{s}_t)$. Different from the traditional reinforcement learning, SAC adds the maximum in-

formation entropy $\mathscr{H}$ in the objective function to solve the problem of poor exploratory.

$$\pi^* = \arg\max_{\pi} \sum_t E_{(\boldsymbol{s}_t,\boldsymbol{c}_t)\sim\rho_\pi}\left[z(\boldsymbol{s}_t,\boldsymbol{c}_t) + \alpha\mathscr{H}(\pi(\cdot \mid \boldsymbol{s}_t))\right], \tag{12}$$

where $\pi^*$ represents the optimal strategy and $\alpha$ is the temperature parameter, indicating the importance of information entropy to the loss function.

SAC adopts soft iteration which includes soft policy evaluation and soft policy improvement. In the soft policy evaluation, we add the maximum information entropy to Q-values $Q(\boldsymbol{s}_t,\boldsymbol{c}_t)$ which is also called soft Q-values and can be iterated through the modified Bellman backup operator $\mathscr{T}^\pi$, which is as follows

$$\mathscr{T}^\pi Q(\boldsymbol{s}_t,\boldsymbol{c}_t) \triangleq r(\boldsymbol{s}_t,\boldsymbol{c}_t) + \gamma E_{\boldsymbol{s}_{t+1}\sim p}\left[V(\boldsymbol{s}_{t+1})\right], \tag{13}$$

where

$$V(\boldsymbol{s}_t) = E_{\boldsymbol{c}_t\sim\pi}\left[Q(\boldsymbol{s}_t,\boldsymbol{c}_t) - \alpha\log\pi(\boldsymbol{c}_t \mid \boldsymbol{s}_t)\right]. \tag{14}$$

Moreover, $\gamma\in[0,1]$ is the discount factor which makes the sum of expected rewards finite. The larger the discount factor, the greater the influence of previous actions on subsequent decisions. In particular, when $\gamma=1$, the agent can see all the previous information, and $\gamma=0$, otherwise. We define $Q^{k+1} = \mathscr{T}^\pi Q^k$, then the sequence $Q^k$ will converge to the soft Q-funcation of $\pi$ as $k\to\infty$[29].

In the soft policy improvement, we assume that $\pi_{\text{new}}\in\Pi$. For optimizing the $\pi$, we introduce Kullback-Leibler (KL) divergence, which means that making the distribution of policy similar to the distribution of soft Q-values, which is as follows

$$\pi_{\text{new}} = \arg\min_{\pi'\in\Pi} D_{\text{KL}}\left(\pi'(\cdot \mid \boldsymbol{s}_t) \,\|\, \frac{\exp\left(\frac{1}{\alpha}Q^{\pi_{\text{old}}}(\boldsymbol{s}_t,\cdot)\right)}{Z^{\pi_{\text{old}}}(\boldsymbol{s}_t)}\right), \tag{15}$$

where $Z^{\pi_{\text{old}}}(\boldsymbol{s}_t)$ normalizes the distribution, which may be difficult to obtain, but it does not affect the gradient of the policy loss function.

The full soft iteration algorithm alternates between the soft policy evaluation and the soft policy improvement steps. Then, we can get a policy $\pi^*\in\Pi$, where $Q^{\pi^*}(\boldsymbol{s}_t,\boldsymbol{c}_t) \geqslant Q^\pi(\boldsymbol{s}_t,\boldsymbol{c}_t)$ for all $\pi\in\Pi$ and $(\boldsymbol{s}_t,\boldsymbol{c}_t)\in\boldsymbol{S}\times\boldsymbol{C}$, with $|\boldsymbol{C}|<\infty$. The convergence of soft iteration has been proved in Ref. [29].

### B. Problem Definition

As described in section II.B, it is difficult to explicitly formulate the problem as deterministic optimization with the users' locations and channel parameters. Fortunately, SAC (model-free RL algorithms) can train the agent to make the best decision according to environment. To solve (11), we model it as an MDP, with the following information.

*1) State:* We assume that the $m$th UAV can get its own location as

$$\boldsymbol{q}_m(t) = [x_m(t), y_m(t)], \;\; m = 1,\cdots,M. \tag{16}$$

We also assume that the UAV receives the data rate from the ground users as

$$\boldsymbol{R}[t] = \{R_1[t], R_2[t],\cdots,R_n[t]\}, \;\; n = 1,\cdots,N, \tag{17}$$

where each entry $R_n[t]$ represents the RSSI of $n$th user before $t$, where $R_n(t)$ can be calculated by (10). RSSI can help the UAV know which user need communication and get the significant achievable rate. We also add the current slot time $t$ to the state space. The state space needs to be normalized before inputting into the neural network. We use the total throughput of continuous communication $D_{\max}$, which a UAV is directly above a user, as a normalization of the user throughput.

$$R_{\max} = T\text{lb}\left(1 + \frac{P_{\max}\rho_0}{H^2\sigma^2}\right). \tag{18}$$

Therefore, the state space can be defined as

$$\boldsymbol{S}(t) = \left\{ \begin{array}{l} \boldsymbol{Q} = \left\{\left[\frac{x_1(t)}{X_{\max}}, \frac{y_1(t)}{Y_{\max}}\right],\cdots,\left[\frac{x_M(t)}{X_{\max}}, \frac{y_M(t)}{Y_{\max}}\right]\right\}, \\ \boldsymbol{R} = \left\{\frac{R_1(t)}{R_{\max}},\cdots,\frac{R_N(t)}{R_{\max}}\right\}, \frac{t}{T} \end{array} \right\}. \tag{19}$$

We then define the space size of the model as $[X_{\max}, Y_{\max}]$ and the maximum power as $P_{\max}$. At the same time, $|\boldsymbol{S}|_{\text{dim}} = 2M+N+1$.

*2) Action:* According to (11), we set the action space $\boldsymbol{C}$ as the speed, the flight direction and the UAV transmission power

$$\boldsymbol{C}(t) = \left\{ \begin{array}{l} \boldsymbol{V} = \left\{\frac{v_1(t)-V_{\max}/2}{V_{\max}/2},\cdots,\frac{v_M(t)-V_{\max}/2}{V_{\max}/2}\right\}, \\ \boldsymbol{\theta} = \left\{\frac{\theta_1(t)-\pi}{\pi},\cdots,\frac{\theta_M(t)-\pi}{\pi}\right\}, \\ \boldsymbol{P} = \left\{\frac{p_1(t)-P_{\max}/2}{P_{\max}/2},\cdots,\frac{p_M(t)-P_{\max}/2}{P_{\max}/2}\right\} \end{array} \right\}. \tag{20}$$

Where $v_m(t)\in\{0,V_{\max}\}$, $\theta_m(t)\in\{0,2\pi\}$, and $p_m\in\{0,P_{\max}\}$. We aim to maximize the total throughput for all the users, so we set the $a_{m,n}[t]$ as the user with the most achievable rate. For the UAV speed, only the maximum speed $V_{\max}$ and the minimum speed 0 m/s are employed. Also, one has $|\boldsymbol{C}|_{\text{dim}} = 3M$.

*3) Reward:* For solving (11), we consider all the constraints defined in the optimization problem, especially (11a), which reflects the fairness of UAV communication. This brings a great challenge to the design of reward function. We propose a reward function, which not only solves the problem that sparse rewards may not address, but also overcomes the loss of fairness of UAV communication caused by non-sparse

rewards. Five sub-reward parts $\{z_1(t), z_2(t), z_3(t), z_4(t), z_5(t)\}$ are proposed. $z_1(t)$ represents the reward of throughput. $z_2(t)$ represents the reward of UAV returning to the origin for constraint (11f). $z_3(t)$ represents the non-sparse reward corresponding to constraint (11a). $z_4(t)$ represents the sparse reward corresponding to constraint (11a). $z_5(t)$ represents the sparse reward corresponding to constraint (11g). We also define two time nodes $T_1$ and $T_2$, where $T_1$ means that the UAV starts to consider constraints (11a) and $T_2$ means that the UAV begins to consider returning to the origin.

$$z_1(t) = \sum_{m=1}^{M} L_m(t), \tag{21}$$

$$z_2(t) = \sum_{m=1}^{M} F_m(t)\,(t - T_2)\,, t \geqslant T_2, \tag{22}$$

$$z_3(t) = \sum_{n=1}^{N} \left(D_n(t) - R_{\min}\right)(t - T_1)\,, t \geqslant T_1, \tag{23}$$

$$z_4(t) = \sum_{n=1}^{N} \beta_n(t), \tag{24}$$

$$z_5(t) = \sum_{m=1, m\neq j}^{M} \sum_{j=1}^{M} \zeta_{m,j}(t). \tag{25}$$

In the above equation, $F_m(t)$ represents the distance of $m$th UAV between origin at time $t$, $L_m(t)$ represents the throughput of $m$th UAV at time $t$, and one has

$$\beta_n(t) = \begin{cases} 1, & D_n(t) \geqslant R_{\min} \text{ and } \sum_{t=1}^{t} \beta_n(t) = 0, \\ 0, & \text{otherwise}, \end{cases} \tag{26}$$

and

$$\zeta_{m,j}(t) = \begin{cases} 1, & G_{m,j}(t) \leqslant d_{\min}^2, \\ 0, & \text{otherwise}, \end{cases} \tag{27}$$

where $G_{m,j}(t)$ means the distance between $m$th and $j$th UAVs.

In summary, the complete expression of $z(t)$ is as

$$z(t) = \lambda_1 z_1(t) + \lambda_2 z_2(t) + \lambda_3 z_3(t) + \lambda_4 z_4(t) + \lambda_5 z_5(t), \tag{28}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$, and $\lambda_5$ denote the weight coefficients of $z_1(t), z_2(t), z_3(t), z_4(t)$, and $z_5(t)$ separately. Specifically, $\lambda_1$ is a positive constant that is used to adjust the reward of total throughput by $M$ UAV, $\lambda_2$ is a negative constant coefficient to punish non-arrival, $\lambda_3$ and $\lambda_4$ are the positive constant coefficient like $\lambda_1$, $\lambda_5$ is the negative constant coefficient like $\lambda_2$.

### C. SAC

As discussed above, the convergence of SAC is proved in section III.A. Specifically, we use two loss functions to approximate both the soft Q-value and the policy. Then, we adopt stochastic gradient descent (SGD) in training process.

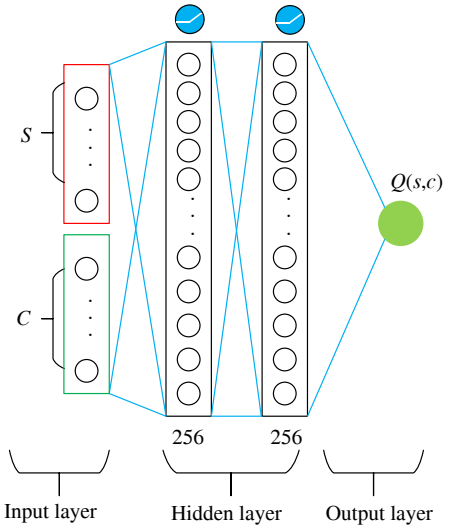



**Fig. 2** The critic network architecture

We use an actor-critic framework which contains a critic network and an actor network to solve the problem with continuous actions.

For the critic network, we define the parameters of the critic-evaluation network and critic-target network as $\theta$ and $\bar{\theta}$, respectively. Then, we optimize the soft Q-values by minimizing the following loss function

$$J_Q(\theta) = E_{(\boldsymbol{S}(t), \boldsymbol{C}(t), z(t)) \sim \mathscr{D}} \left[ \frac{1}{2} \left( Q_\theta\left(\boldsymbol{S}(t), \boldsymbol{C}(t)\right) - \left(z(t) + \gamma E_{\boldsymbol{s}_{t+1} \sim p}\left[V_{\bar{\theta}}\left(\boldsymbol{S}(t+1)\right)\right]\right) \right)^2 \right], \tag{29}$$

where $\mathscr{D}$ means the experience pool and $V_{\bar{\theta}}\left(\boldsymbol{S}(t+1)\right)$ is given by (14).

The critic network takes the state $\boldsymbol{S}(t)$ and action $\boldsymbol{C}(t)$ as input, and outputs the soft Q-value $Q\left(\boldsymbol{S}(t), \boldsymbol{C}(t)\right)$. As shown in Fig. 2, we use the activation function $relu(\cdot)$ in the hidden layers. The role of the critic network is to measure the value of $\boldsymbol{C}(t)$ at the $\boldsymbol{S}(t)$, which means the higher $Q\left(\boldsymbol{S}(t), \boldsymbol{C}(t)\right)$, the more $\boldsymbol{C}(t)$ will be at the $\boldsymbol{S}(t)$.

For the actor network, we define the parameter of the actor network as $\phi$. Because $Z^{\pi_{\text{old}}}(\boldsymbol{s}_t)$ does not contribute to the gradient, we optimize the policy by minimizing the following loss function

$$J_\pi(\phi) = E_{\boldsymbol{S}(t) \sim \mathscr{D}} \Big[ E_{\boldsymbol{C}(t) \sim \pi_\phi} \left[ \alpha \log\left(\pi_\phi\left(\boldsymbol{C}(t) \mid \boldsymbol{S}(t)\right)\right) - Q_\theta\left(\boldsymbol{S}(t), \boldsymbol{C}(t)\right) \right] \Big], \tag{30}$$

To the end, we reparameterize the policy using actor network transformation as

$$\boldsymbol{C}(t) = f_\phi\left(\varepsilon_t; \boldsymbol{S}(t)\right), \tag{31}$$

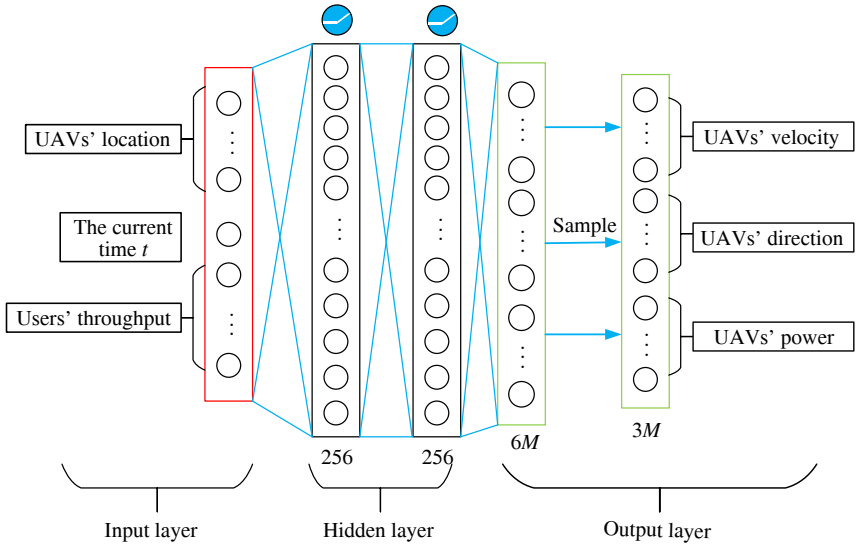



**Fig. 3** The actor network architecture

where $\varepsilon_t$ represents noise, sampled from some fixed distribution, such as the spherical Gaussian distribution. Then, by taking (31) into (30), the loss function can be transformed into the following

$$J_\pi(\phi) = E_{\boldsymbol{S}(t)\sim\mathscr{D},\varepsilon_t\sim\mathscr{N}}\Big[\alpha\log\pi_\phi\left(f_\phi\left(\varepsilon_t;\boldsymbol{S}(t)\right)\mid\boldsymbol{S}(t)\right) - Q_\theta\left(\boldsymbol{S}(t),f_\phi\left(\varepsilon_t;\boldsymbol{S}(t)\right)\right)\Big], \tag{32}$$

where $\pi_\phi$ is the policy of the agent and defined implicitly in terms of $f_\phi$.

The input of actor network contains the state $\boldsymbol{S}(t)$, and the output is the action $\boldsymbol{C}(t)$. As shown in Fig. 3, the hidden layer architecture of actor network is the same as the critic network. Specially, we get the action $\boldsymbol{C}(t)$ by sampling from $6M$ neurons which means the mean and variance of the action $\boldsymbol{C}(t)$, respectively. The role of actor network is to get the action $\boldsymbol{C}(t)$ which maximizes the total reward $\boldsymbol{Z}(t)$ at the state $\boldsymbol{S}(t)$.

The SAC introduces the temperature parameter $\alpha$. Unfortunately, it's usually difficult to set its numerical value. To solve the problem, we propose self-adapting $\alpha$, which optimizes $\alpha$ by SGD. In the SAC, we aim to find the $\pi_t^*$ which maximizes the total reward $\boldsymbol{Z}(t)$ and makes the entropy of action $\boldsymbol{C}(t)$ greater than the expected entropy $\overline{\mathscr{H}}$. To achieve this, its duality problem has been given in Ref. [29], which is as follows

$$\alpha_t^* = \arg\min_{\alpha_t} E_{\boldsymbol{C}(t)\sim\pi_t^*}\left[-\alpha_t\log\pi_t^*\left(\boldsymbol{C}(t)\mid\boldsymbol{S}(t);\alpha_t\right) - \alpha_t\overline{\mathscr{H}}\right], \tag{33}$$

then, we optimize the temperature parameter $\alpha$ by minimizing the following loss function

$$J(\alpha) = E_{\boldsymbol{C}(t)\sim\pi_t}\left[-\alpha\log\pi_t\left(\boldsymbol{C}(t)\mid\boldsymbol{S}(t)\right) - \alpha\overline{\mathscr{H}}\right]. \tag{34}$$

**Algorithm 1** Soft actor-critic

**Require:** $T, M, N, \rho_0, H, \sigma^2, R_{\min}, V_{\max}, P_{\max}$.

Initialize: $\theta_1,\theta_2,\phi$,
$\bar{\theta}_1 \leftarrow \theta_1$ $\bar{\theta}_2 \leftarrow \theta_2$ $\mathscr{D} \leftarrow \emptyset$.
**repeat**
  **repeat**
    get $\boldsymbol{S}(t)$ in (19),
    Select an action $\boldsymbol{C}(t)$ in (20) according to actor network $\pi(\boldsymbol{S}(t))$,
    input $\boldsymbol{C}(t)$ to environment get $\boldsymbol{S}(t+1)$ and $r(\boldsymbol{S}(t),\boldsymbol{C}(t))$,
    $\mathscr{D} \leftarrow \mathscr{D}\cup\{\boldsymbol{S}(t),\boldsymbol{C}(t),z(\boldsymbol{S}(t),\boldsymbol{C}(t)),\boldsymbol{S}(t+1)\}$,
    $\boldsymbol{S}(t) = \boldsymbol{S}(t+1)$,
    update $t = t+1$.
  **until** $t = T$.
  **if** $\mathscr{D} \geqslant \mathscr{D}_{\min}$ and each gradient step **then**
    $\theta_i \leftarrow \theta_i - \lambda\hat{\nabla}_{\theta_i}J_Q(\theta_i)$ for $i\in\{1,2\}$,
    $\phi \leftarrow \phi - \lambda\hat{\nabla}_\phi J_\pi(\phi)$,
    $\alpha \leftarrow \alpha - \lambda\hat{\nabla}_\alpha J(\alpha)$,
    $\bar{\theta}_i \leftarrow \tau\theta_i + (1-\tau)\bar{\theta}_i$ for $i\in\{1,2\}$.
  **end if**
**until**

**Ensure:** $\theta_1,\theta_2,\phi$.

Although the temperature parameter $\alpha$ can be obtained by the methods above, it also introduces a new parameter $\overline{\mathscr{H}}$. Fortunately, we can set $\overline{\mathscr{H}} = -|\boldsymbol{S}|_{\dim}$.

In Algorithm 1, we use two critic-target and two critic-evaluation networks to mitigate positive bias. In particular, we parameterize two critic-target networks with parameters $\bar{\theta}_i$, and train them independently to optimize $J_Q(\bar{\theta}_i)$. As shown in Fig. 4, SAC has 6 networks, including two Q-evaluation (critic) networks, two Q-target (critic) networks, one Actor network, and one $\alpha$ network. Q-target networks adopt soft update, which $\bar{\theta}_i \leftarrow \tau\theta_i + (1-\tau)\bar{\theta}_i$ for $i\in\{1,2\}$, where $\tau$ is the soft update coefficient. The $\mathscr{D}$ represents experience

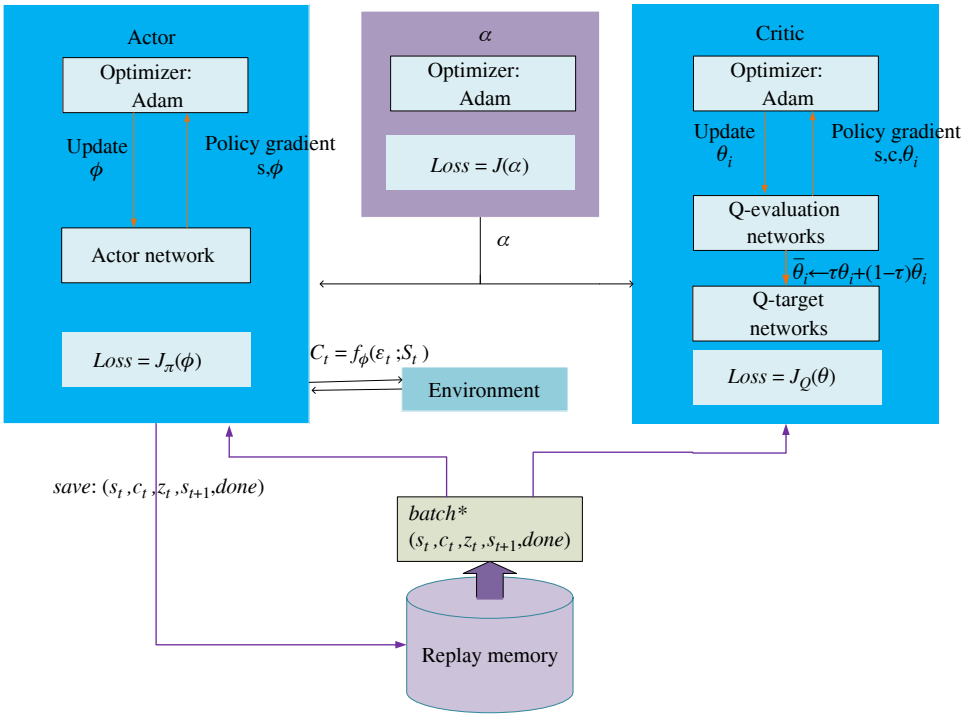



**Fig. 4** The actor network architecture

pool, which can make the states independent of each other. The $\mathscr{D}_{\min}$ means the minimum experiences needed to start the training process.

## IV. Numerical Results

In this section, we provide numerical results with SAC in JTDPA of the multi-UAV system where we consider there are $M = 2$ UAVs and $N = 6$ ground users which are randomly and uniformly distributed within a 2D area of $(3 \times 2)$ km$^2$. The altitude of the UAVs is $H = 100$ m. The noise power at the receiver is $\sigma^2 = -110$ dBm. The channel gain is $\rho_0 = -60$ dB at the reference distance $d_0 = 1$ m. The maximum transmission power, the minimum distance of UAV, and maximum speed of the UAV are $P_{\max} = 0.1$ W, $d_{\min} = 100$ m, and $V_{\max} = 50$ m/s, respectively. The initial locations of the $M$ UAVs are randomly generated in $(3 \times 2)$ km$^2$. In this paper, the initial locations of the two UAVs are set above the 4th user and the 1st user respectively. For hyper parameters of the SAC, we set $\lambda = 3 \times 10^{-4}$, $\gamma = 0.98$, $\tau = 0.001$. For other DRL algorithms, such as DDPG, PPO, TD3, we set the reward function and learning rate the same as the SAC. The specific reward function settings are shown in Tab. 1. All hyper parameters of other DRL algorithms are as following: the DDPG and TD3 belong to the deterministic policy gradient algorithms, which need to add noise. In our simulations, we use a Gaussian action noise. The mean and variance of the noise are set as 0 and 0.1, respectively. Moreover, the variance will reduce with the

**Tab. 1** Reward setting

| Reward parameters | Simulation value |
|---|---|
| $\lambda_1$ | 0.05 |
| $\lambda_2$ | −0.002 |
| $\lambda_3$ | 0.001 |
| $\lambda_4$ | 20 |
| $\lambda_5$ | −50 |

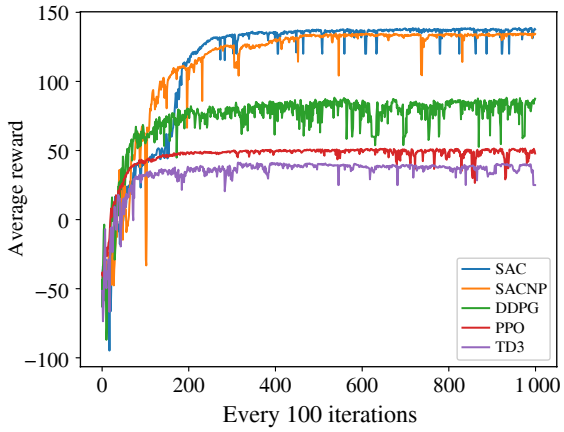



**Fig. 5** Average rewards of Algorithm 1

training. The PPO belongs to on-policy algorithms. We adopt clipped surrogate objective[31], where $\varepsilon = 0.2$.

Fig. 5 demonstrates $10^5$-episode training processes of SAC,

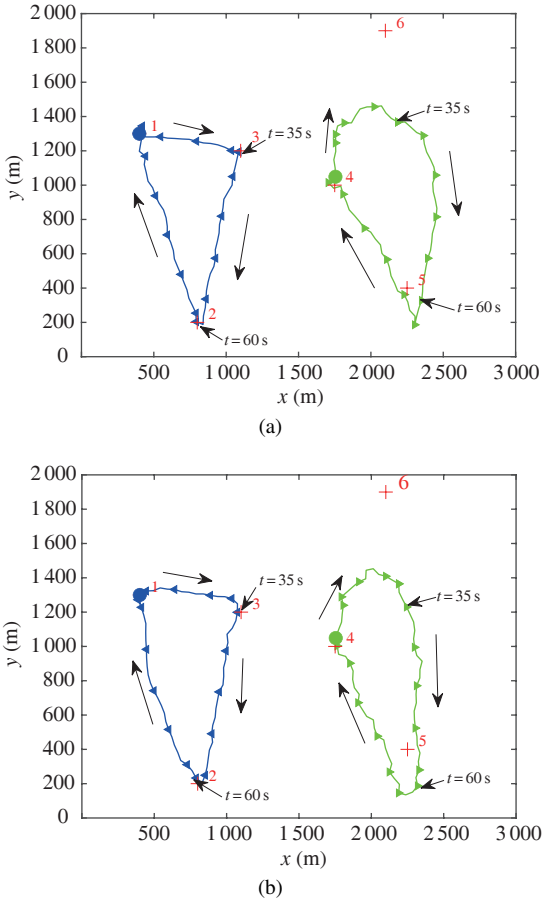



**Fig. 6** The trajectories of two-UAV and six-user communication system under $T = 90$ s and $R_{\min} = 100$ bit/Hz. Blue circle '•' and green cricle '•' represent the inital locations of two UAVs trajectories, respectively. Black arrows represent the directions of UAV. We sample every 5 s on each UAV's trajectory and the sampling points are marked with blue '◀' and green '▶' as their corresponding trajectories: (a) Optimized UAV trajectories by SACNP; (b) Optimized UAV trajectories by SAC

SAC without power control (SACNP) ($p_m(t) = P_{\max}$) and other DRL algorithms. We set $T = 90$ s and $R_{\min} = 100$ bit/Hz. One can see that all the algorithms converge at about $4 \times 10^4$-episode and the SAC, as well as SACNP, have better performance than other DRLs. As for DDPG, TD3, and PPO, their rewards increase quickly at first, but as the number of iterations increase, all of which fall into the local optimal solution. The DDPG and TD3 belong to deterministic policy gradient algorithms, where the actor network outputs a deterministic action, and the PPO is an online algorithm, which cannot solve the problem of smaller experience than off-line algorithm. In contrast, SAC considers the action entropy in Q-value, which encourages the agents to explore, so that SAC may jump out of the local optimal and find the optimal solution as the number of iterations increases. Compared with the SACNP, SAC can change their transmission power to reduce the interference from other UAVs and get more reward, so that the total reward of SAC may be higher than SACNP.

**Tab. 2** The total throughput of ground users

| User/Sum | SAC | SACNP |
|---|---|---|
| 1st | 224.699 85 | 202.585 30 |
| 2nd | 165.751 65 | 150.761 30 |
| 3rd | 162.853 14 | 151.974 15 |
| 4th | 172.619 47 | 160.346 58 |
| 5th | 168.891 44 | 155.841 64 |
| 6th | 106.659 06 | 105.953 67 |
| Sum | 1 001.474 61 | 927.472 64 |

In Fig. 6, we compare the optimized UAVs' trajectories obtained by the SACNP in Fig. 6(a) and SAC in Fig. 6(b) with the time period $T = 90$ s and $R_{\min} = 100$ bit/Hz. The total throughput of ground users are shown in Tab. 2. Note that the user's locations and channel parameters are inaccessible for the UAVs, and SAC outputs action by applying sampling. As a result, the trajectories of the UAV may not be smooth.

It can be observed from Fig. 6(a) that two UAVs tend to keep away from each other as far as possible to reduce co-channel interference from $t = 35$ s to $t = 60$ s. However, UAVs sometimes sacrifice the favourable direct communication links, especially when they serve two users which are close to each other. As a result, the SAC can dynamically adjust the transmission power of the UAVs to reduce the interference from other UAVs.

Both Fig. 6(a) and Fig. 6(b) show that the UAVs will not visit the 6th user. The 6th user is far away from the other users, if the UAV visits the 6th user, the total throughput will reduce. However, the UAV will move to the 6th user to ensure it achieves the minimum RSSI. As a result, the SAC can dynamically adjust the trajectory to maximize the total throughput for all users and ensure all users achieve the minimum RSSI. The initial locations of the two UAVs are set above the 4th user and the 1st user respectively, so the 4th user and 1st user will get more throughput. Meanwhile, the UAVs will fly to other users to ensure them achieve the minimum RSSI.

In contrast, in Fig. 7, the transmission power of the UAVs are complementary at certain time slots, such as 30∼40 s and 15∼25 s, when the two UAVs communicate to two nearby users. Therefore, strong direct links and weak co-channel interference can be achieved at the same time, which may improve the total received data rate in the period $T$. Therefore, without transmission power control, the interference signal by other UAVs can only be mitigated by adjusting the UAV trajectory to keep away from other UAVs, where jointly optimizing transmission power and trajectory of the UAVs provide more flexibility and the higher throughput. In Fig. 8, we compare the maximum throughput of three optimizations for trajectories: 1) the SAC, which uses Algorithm 1; 2) the SACNP, applying Algorithm 1 but without power control; 3) Circular,

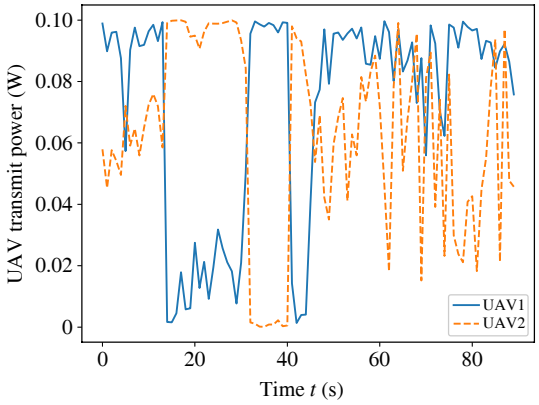



**Fig. 7** UAV transmission power versus time for SAC

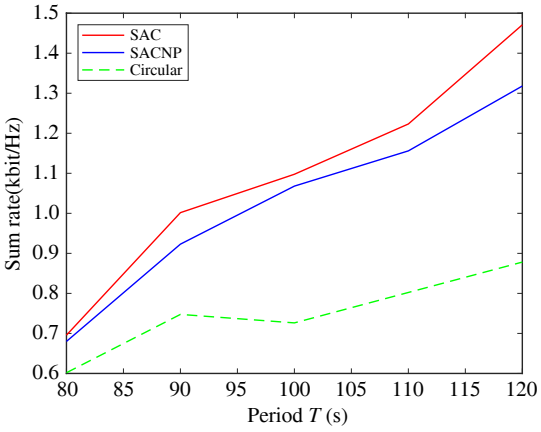



**Fig. 8** Sum rate versus period $T$ with different algorithms

which is obtained by Ref. [32] with $M = 2$ and only optimizes user scheduling and association. Especially, the SAC optimizes the UAVs' trajectories and transmission power at the same time. We can get several important observations from Fig. 8. First, as expected, the total throughput of the three algorithms increases as the period $T$ becomes larger. Second, the performance gap between the SACNP and Circular increases with increasing $T$. Third, comparing the SAC and SACNP, by using power control, it shows that more flexibility can be obtained, result in better throughput.

## V. Conclusion

In this paper, multi-UAV and multi-user system have been studied, where UAVs serve multiple ground users. We aim to maximize the total throughput for all users as well as meeting the fairness by optimizing the UAVs' trajectories and transmission power in a centralized way, without knowing users' locations and channel parameters. Note that no matter what kind of channel model we adopt, our proposed algorithm only needs RSSI of the ground user. To solve this problem, we model it as an MDP and adopt DRL-based SAC to address. Meanwhile, the design of our reward function combines both sparse and non-sparse reward, which not only solves the problem that sparse rewards may not address, but also overcomes the loss of fairness of UAV communication caused by non-sparse rewards. Simulation results have shown that the proposed algorithm, in terms of convergence and performance, is better than other DRL-based solutions.

## References

[1] WU Q, ZENG Y, ZHANG R. Joint trajectory and communication design for UAV-enabled multiple access[C]//IEEE Global Communications Conference. Piscataway: IEEE Press, 2017: 1-6.

[2] LYU J, ZENG Y, ZHANG R, et al. Placement optimization of UAV-mounted mobile base stations[J]. IEEE Communications Letters, 2017, 21(3): 604-607.

[3] WU Q, ZENG Y, ZHANG R. Joint trajectory and communication design for multi-UAV enabled wireless networks[J]. IEEE Transactions on Wireless Communications, 2018, 17(3): 2109-2121.

[4] BOR-YALINIZ R, EL-KEYI A, YANIKOMEROGLU H, et al. Efficient 3-D placement of an aerial base station in next generation cellular networks[C]//IEEE International Conference on Communications (ICC). Piscataway: IEEE Press, 2016: 1-5

[5] LI P, XU J. Fundamental rate limits of UAV-enabled multiple access channel with trajectory optimization[J]. IEEE Transactions on Wireless Communications, 2020, 19(1): 458-474.

[6] FANG S, CHEN G, LI Y Joint optimization for secure intelligent reflecting surface assisted UAV networks[J]. IEEE Wireless Communications Letters, 2021, 10(2): 276-280.

[7] PAN C, REN H, ELKASHLAN M, et al. Weighted sum-rate maximization for the ultra-dense user-centric TDD C-RAN downlink relying on imperfect CSI[J]. IEEE Transactions on Wireless Communications, 2019, 18(2): 1182-1198.

[8] ZENG Y, ZHANG R, LIM T, et al. Throughput maximization for UAV-enabled mobile relaying systems[J]. IEEE Transactions on Communications, 2016, 64(12): 4983-4996.

[9] ZHANG S, ZHANG H, HE Q, et, al. Joint trajectory and power optimization for UAV relay networks[J]. IEEE Communications Letters, 2018, 22(1): 161-164.

[10] PAN C, REN H, DENG Y, et al. Joint blocklength and location optimization for URLLC-enabled UAV relay systems[J]. IEEE Communications Letters, 2019, 23(3): 498-501.

[11] LOKE S. The Internet of flying-things: opportunities and challenges with airborne fog computing and mobile cloud in the clouds[J]. arXiv preprint arXiv:1507.04492v1, 2015.

[12] JEONG S, SIMEONE O, KANG J. Mobile edge computing via a UAV-mounted cloudlet: optimization of bit allocation and path planning[J]. arXiv preprint arXiv:1609.05362, 2017.

[13] XU J, ZENG Y, ZHANG R. UAV-enabled wireless power transfer: trajectory design and energy optimization[J]. IEEE Transactions on Wireless Communications, 2018, 17(8): 5092-5106.

[14] FENG W, TANG J, YU Y, et al. UAV-enabled swipt in IoT networks for emergency communications[J]. IEEE Wireless Communications, 2020, 27(5): 140-147.

[15] ZENG Y, ZHANG R, LIM T. Wireless communications with unmanned aerial vehicles: opportunities and challenges[J]. IEEE Communications Magazine, 2016, 54(5): 36-42.

[16] AL-HOURANI A, KANDEEPAN S, LARDNER S. Optimal lap altitude for maximum coverage[J]. IEEE Wireless Communications Letters, 2014, 3(6): 569-572.

[17] MERWADAY A, GUVENV I. UAV assisted heterogeneous networks

for public safety communications[C]//IEEE Wireless Communications and Networking Conference Workshops. Piscataway: IEEE Press, 2015: 329-334.

[18] AHMED S, CHOWDHURY M, JANG Y. Energy-efficient UAV-to-user scheduling to maximize throughput in wireless networks[J] IEEE Access, 2020, 8(21): 215-225.

[19] LYU J, ZENG Y, ZHANG R. Cyclical multiple access in UAV-aided communications: a throughput-delay tradeoff IEEE Wireless Communications Letters, 2016, 5(6), 600-603.

[20] LUONG N, HOANG D, GONG S, et al. Applications of deep reinforcement learning in communications and networking: a survey[J]. IEEE Communications Surveys Tutorials, 2019, 21(4): 3133-3174.

[21] DING R, GAO F, SHEN X. 3D UAV trajectory design and frequency band allocation for energy-efficient and fair communication: a deep reinforcement learning approach[J]. IEEE Transactions on Wireless Communications, 2020, 19(12): 7796-7809.

[22] CAO Y, ZHANG L, LIANG Y. Deep reinforcement learning for multi-user access control in UAV networks[C]//IEEE International Conference on Communications. Piscataway: IEEE Press, 2019: 1-6.

[23] CUI J, LIU Y, NALLANATHAN A. Multi-agent reinforcement learning-based resource allocation for UAV networks[J]. IEEE Transactions on Wireless Communications, 2020, 19(2): 729-743.

[24] LIU C, CHEN Z, TANG J. Energy-efficient UAV control for effective and fair communication coverage: a deep reinforcement learning approach[J] IEEE Journal on Selected Areas in Communications, 2018, 36(9): 2059-2070.

[25] CUI J, LIU Y, NALLANATHAN A. The application of multi-agent reinforcement learning in UAV networks[C]//IEEE International Conference on Communications Workshops. Piscataway: IEEE Press, 2019: 1-6.

[26] YANG L, YAO H, WANG J, et al. Multi-UAV-enabled load-balance mobile-edge computing for IoT networks[J]. IEEE Internet of Things Journal, 2020, 7(8): 6898-6908.

[27] YIN S, RICHARD YU F. Resource allocation and trajectory design in UAV-aided cellular networks based on multiagent reinforcement learning[J]. IEEE Internet of Things Journal, 2022, 9(4): 2933-2943.

[28] SAXENA V, JALDEN J, KLESSIG H. Optimal UAV base station trajectories using flow-level models for reinforcement learning[J]. IEEE Transactions on Cognitive Communications and Networking, 2019, 5(4): 1101-1112.

[29] HAARNOJA T, ZHOU A, HARTIKAINEN K, et al. Soft actor-critic algorithms and applications[J]. arXiv preprint arXiv:1812.05905, 2018.

[30] HAARNOJA T, ZHOU A, ABBEEL P, et al. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor[J]. arXiv preprint arXiv:1801.01290, 2018.

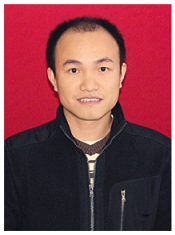
[31] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[J]. arXiv preprint arXiv:1707.06347, 2017.

[32] BERTSEKAS D. Packings of equal circles in fixed-sized containers with maximum packing density. Online available[EB].

## About the Authors

**Chiya Zhang** received the Ph.D. degree in Telecommunication Engineering from the University of New South Wales, Sydney, Australia, in 2019. He is currently an Assistant Professor at Harbin Institute of Technology, Shenzhen, China. His current research interest is AI applications in telecommunication engineering. He received the Exemplary Reviewer Certificates of the IEEE Wireless Communications Letters in 2018 and IEEE ComSoc Asia-Pacific Outstanding Paper Award in 2020. He is serving as an Associate Editor for the IEEE Internet of Things Journal.

**Shiyuan Liang** [corresponding author] received the B.S. degree in Communication Engineering from Yangtzeu University, Jingzhou, China, in 2020. He is a Master Student of Information and Communication Engineering with Guangdong Key Laboratory of Intelligent Information Processing, College of Electronics and Information Engineering, Shenzhen University, Shenzhen, China. His research interests include UAV wireless communications and machine learning.



**Chunlong He** received the M.S. degree in Communication and Information Science from Southwest Jiaotong University, Chengdu, China, in 2010 and the Ph.D. degree from Southeast University, Nanjing, China, in 2014. From September 2012 to September 2014, he was a Visiting Student with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA. Since 2015, he has been with the College of Information Engineering, Shenzhen University, where he is currently an Associate Professor. His research interests include communication and signal processing, green communication systems, channel estimation algorithms, and limited feedback techniques. Dr. He is a member of the Institute of Electronics, Information, and Communication Engineering. He is currently an Associate Editor of IEEE Access.

**Kezhi Wang** received the B.E. and the M.E. degrees in School of Automation from Chongqing University, Chongqing, China, in 2008 and 2011, respectively. He received the Ph.D. degree in Engineering from the University of Warwick, Coventry, U.K. in 2015. He was a Senior Research Officer in University of Essex, Essex, U.K. Currently, he is a Senior Lecturer with Department of Computer and Information Sciences at Northumbria University, Newcastle, U.K. His research interests include wireless communications and machine learning.