Deep Reinforcement Learning for Trajectory Design and Power Allocation in UAV Networks

Nan Zhao* ‡, Yiqiang Cheng*, Yiyang Pei†, Ying-Chang Liang‡, Dusit Niyato§
* Hubei University of Technology, Wuhan 430068, China.
† Singapore Institute of Technology, Singapore.
‡ University of Electronic Science and Technology of China, Chengdu 611731, China.
§ Nanyang Technological University, Singapore.

***Abstract*—Unmanned aerial vehicle (UAV) is considered to be a key component in the next-generation cellular networks. Considering the non-convex characteristic of the trajectory design and power allocation problem, it is difficult to obtain the optimal joint strategy in UAV-assisted cellular networks. In this paper, a reinforcement learning-based approach is proposed to obtain the maximum long-term network utility while meeting with user equipments' quality of service requirement. The Markov decision process (MDP) is formulated with the design of state, action space, and reward function. In order to achieve the joint optimal policy of trajectory design and power allocation, deep reinforcement learning approach is investigated. Due to the continuous action space of the MDP model, deep deterministic policy gradient approach is presented. Simulation results show that the proposed algorithm outperforms other approaches on overall network utility performance with higher system capacity and faster processing speed.**

***Index Terms*—UAV networks, trajectory design, power allocation, deep reinforcement learning.**

I. Introduction

Recently, unmanned aerial vehicles (UAVs) have been regarded as an effective technology in future wireless networks [1]. Due to its fast deployment, flexible configuration, wide coverage, and low cost, UAVs can be used as relays between ground user equipments (UEs) for cooperative communication. Moreover, since UAVs can intelligently change their locations to provide on-demand wireless services for ground UEs, UAVs are also designed as aerial base stations (ABSs) for wireless communication. Thus, UAV-assisted cellular networks have been applied to various applications, such as remote sensing, traffic monitoring, public safety and military [2].

In UAV-assisted cellular networks, there are some technical challenges, including trajectory design, resource allocation, and interference management. By properly designing the trajectory of UAVs, UAVs can move close to target UEs to provide wireless services, which may alleviate the co-channel interference to the unserved UEs. Moreover, the transmit power of UAVs should also be controlled to achieve the trade-off between spectrum efficiency and interference management. Therefore, the problem of trajectory design, power allocation, and interference management should be considered jointly.

Recently, the issue of joint trajectory design and power allocation (JTDPA) has been studied in [3]–[5]. Unfortunately, it may be difficult to find the global optimal method for the JTDPA problem due to the non-convex characteristic. Although some methods have been proposed, i.e., the alternating optimization approach [6], Lagrange dual method [7], and iterative algorithm [8], nearly accurate information, i.e., channel state and UAVs' positions, is still needed to solve the joint optimization problem effectively. Moreover, it may be intractable to find the optimal strategy without the exact information. Therefore, in this work, the reinforcement learning (RL) approach is proposed to solve the JTDPA optimization issue in the UAV-assisted cellular networks.

By interacting with the environment, RL is able to achieve the optimal policy for the intelligent decision problem [9]. Additionally, as one of the online learning methods, RL has been extensively applied in artificial intelligence domain with little prior information [10]. Deep reinforcement learning (DRL) approaches have been also utilized in some UAV-assisted networks, i.e., UAV control [11], trajectory design [12], [13], interference management [14], [15], and multi-user access control [16]. However, most RL works may not be always suitable to deal with the continuous and high-dimensional action spaces in the JTDPA optimization problem. Nonetheless, our prior work has proposed a DRL approach for solving the joint user association and resource allocation issue [17], and solving JTDPA using the DRL approach will complement the work.

In this paper, we investigate a DRL approach for the JTDPA optimization issue in the UAV-assisted cellular networks. The main contributions of this work are as follows.

- *New Solution Technique:* This JTDPA joint optimization issue is formulated to maximize the cumulative discounted reward defined in terms of the transmission rate while considering UEs' quality of service (QoS) requirement. Then, due to the non-convex and combinatorial nature of the problem, the JTDPA optimization is formulated as the Markov decision process (MDP). Then, the DRL solution is investigated for the JTDPA optimization issue.
- *Optimal Algorithm Design:* Due to the continuous action space of the MDP, deep deterministic policy gradient (DDPG) method is applied to find the optimal policy. Target network and experience reply strategies are designed to increase the learning stability.
- *Performance Evaluation:* Numerical results are presented

978-1-7281-5089-5/20/$31.00 ©2020 IEEE

to indicate that the proposed approach has better performance than those of the other optimization strategies with higher computational efficiency.

The rest of this paper is organized as follows. Section II presents system model and problem formulation. The DRL algorithm is proposed to deal with the JTDPA optimization problem in Section III. Section IV describes numerical results to evaluate the performance of the proposed method. The paper is concluded in Section V.

II. System Model and Problem Formulation

A. System Model

In a general UAV-assisted cellular network, N UAVs are deployed as ABSs to provide M UEs with the downlink wireless services in N non-overlapping hotspots. The sets of UEs and UAVs are denoted as $\mathcal{M}$ and $\mathcal{N}$, respectively. The number of UEs in hotspot i is denoted as $M(i)$. For the simplicity of discussion, we assume that UAV i provides service to hotspot i using the same frequency band. Moreover, considering that each UE belongs to only one hotspot, we have $\sum_{i=1}^{N} M(i) = M$. All the UAVs are controlled by a core terrestrial base station (CTBS). At any time t, UEs in the same hotspot are served by the same UAV simultaneously by employing FDMA [18].

Let $v_m = [x_m, y_m]^{\mathrm{T}} \in \mathbb{R}^{2\times 1}, m \in \mathcal{M}$ be the 2D coordinates of UE m, where x_m and y_m are the coordinates of UE m. Then, the horizontal coordinate of UAV i is denoted as $v_i(t) = [x_i(t), y_i(t)]^{\mathrm{T}} \in \mathbb{R}^{2\times 1}, i \in \mathcal{N}$, where $x_i(t)$ and $y_i(t)$ are the X-coordinate and Y-coordinate of UAV i at time t, respectively. Then, we can have the distance between UE m and UAV i in the horizontal dimension $r_{i,m}(t) = \sqrt{[x_i(t) - x_m]^2 + [y_i(t) - y_m]^2}$.

Then, we define the vertical trajectory of UAV i as $h_i(t) \in [H_{min}, H_{max}]$, where H_{min} and H_{max} are the minimum and maximum heights of UAVs, respectively. The distance between UAV i and UE m is denoted as $d_{i,m}(t) = \sqrt{h_i^2(t) + r_{i,m}(t)^2}$.

Considering the limited flight speed of UAVs, the trajectories of UAVs should be subject to the maximum travel distance, that is,

$$\|v_i(t+1) - v_i(t)\| \le V_L T_s, \tag{1}$$

$$\|h_i(t+1) - h_i(t)\| \le V_A T_s, \tag{2}$$

where V_L and V_A are the level-flight and vertical-flight speeds of UAVs in each time slot T_s, respectively.

Moreover, to avoid the collision of any two UAVs, collision avoidance constraints of UAVs should be also considered, that is,

$$\|v_i(t) - v_j(t)\|^2 + \|h_i(t) - h_j(t)\|^2 \ge D_{min}^2, \forall i, j \in \mathcal{N}, i \ne j, \tag{3}$$

where D_{min} is the minimum distance between any two UAVs.

Notice that the time slot T_s should be small enough so that the channel can be treated as approximately constant. Furthermore, considering the collision avoidance between any two UAVs, the following constraint of T_s should be satisfied, $T_s \le T_{max} = \frac{D_{min}}{2\sqrt{V_L^2 + V_A^2}}$. Then, we can have the UAV's maximum horizontal distance $L_{max}^h = V_L T_{max}$ and the maximum vertical distance $L_{max}^v = V_A T_{max}$ in each time slot.

Generally, radio signals emitted from the UAV are composed of Line-of-Sight (LoS) or non-Line-of-Sight (NLoS). The probability of the LoS connection between UE m and UAV i is denoted as [19]

$$P_{i,m}^{LoS}(t) = \frac{1}{1 + a\exp(-b(\frac{180}{\pi}\tan^{-1}(\theta_{i,m}) - a))}, \tag{4}$$

where a and b are the parameters related with the environment, $\theta_{i,m} = \frac{h_i(t)}{r_{i,m}(t)}$ is the angle between UE m and UAV i. Furthermore, the probability of NLoS is $P_{i,m}^{NLoS}(t) = 1 - P_{i,m}^{LoS}(t)$.

Accordingly, at time t, the path loss models of LoS and NLoS in dB can be expressed as [19],

$$L_{i,m}^{LoS}(t) = 20\log\left(\frac{4\pi f_c d_{i,m}(t)}{c}\right) + \eta_{LoS}, \tag{5}$$

$$L_{i,m}^{NLoS}(t) = 20\log\left(\frac{4\pi f_c d_{i,m}(t)}{c}\right) + \eta_{NLoS}, \tag{6}$$

where f_c is the carrier frequency, η_{LoS} and η_{NLoS} are the mean extra losses for LoS and NLoS, respevtively.

Then, the expected mean path loss can be expressed as $L_{i,m}(t) = L_{i,m}^{LoS}(t) \times P_{i,m}^{LoS}(t) + L_{i,m}^{NLoS}(t) \times P_{i,m}^{NLoS}(t)$. Assume that the total available bandwidth B is distributed to each UE equally. Thus, the bandwidth of UE m in hotspots i is expressed as $B_{i,m} = B/M(i)$. Moreover, the transmit power of UAVs is also allocated to each UE uniformly, that is, $p_{i,m}(t) = p_i(t)/M(i)$, where $p_i(t) \in [0, P_{max}]$ is the transmit power of UAV i with the maximum transmit power P_{max}.

Then, for a given transmit power of UAV $p_i(t)$, the received SINR $\Gamma_{i,m}(t)$ of UE m from UAV i is written as

$$\Gamma_{i,m}(t) = \frac{p_{i,m}(t) g_{i,m}(t)}{B_{i,m} N_0 + \sum_{j \ne i} p_{j,m}(t) g_{j,m}(t)}, \forall i, j \in \mathcal{N}, \tag{7}$$

where $g_{i,m}(t)$ is the channel gain between UAV i and UE m, and N_0 is the noise power spectral density.

Accordingly, the achievable rate of UE m from UAV i can be achieved as $r_{i,m}(t) = B_{i,m}\log_2(1 + \Gamma_{i,m}(t))$. Thus, we have the total rate of UAV i, that is,

$$r_i(t) = \sum_{m=1}^{M(i)} r_{i,m}(t) = \sum_{m=1}^{M(i)} B_{i,m}\log_2(1 + \Gamma_{i,m}(t)). \tag{8}$$

B. Problem Formulation

In the UAV-assisted cellular networks, to guarantee that each UE meets the minimum QoS requirement Ω_m from its own UAV, the achievable SINR $\Gamma_{i,m}(t)$ of UE m should be no less than Ω_m, that is,

$$\Gamma_{i,m}(t) \ge \Omega_m. \tag{9}$$

Then, the utility $w_i(t)$ of UAV i is defined as the achievable profit subtracted by the transmit cost, which can be given by,

$$w_i(t) = \rho_i r_i(t) - \lambda_p p_i(t) = \sum_{m=1}^{M(i)} [\rho_i r_{i,m}(t) - \lambda_p p_{i,m}(t)], \tag{10}$$

where ρ_i is the profit of each rate, λ_p is the unit price of UAV's transmit power.

Thus, the JTDPA optimization problem is to maximize the overall network utility by obtaining the joint optimal UAV's trajectory ($v_i(t)$ and $h_i(t)$) and transmit power ($p_i(t)$). The optimization issue is formulated as

$$\begin{aligned} \max_{p_i(t),v_i(t),h_i(t)} \quad & \sum_{i=1}^{N} w_i(t) = \sum_{i=1}^{N} \sum_{m=1}^{M(i)} [\rho_i r_{i,m}(t) - \lambda_p p_{i,m}(t)], \\ s.t. \quad & (1), (2), (3), (9), \\ & H_{min} \le h_i(t) \le H_{max}, \quad 0 \le p_i(t) \le P_{max}. \end{aligned} \tag{11}$$

Due to the non-convex and combinatorial properties, it may be difficult to solve the formulated optimization problem, especially for large networks. Exhaustive search method may obtain the optimal strategy with high computational complexity. Moreover, due to unawareness of UEs' information and channel condition, it is also difficult to achieve the optimal policy with conventional optimization methods. In the following section, a reinforcement learning solution will be proposed to find the optimal JTDPA strategy.

III. DRL for JTDPA Optimization Problem

In order to obtain the maximum network utility, the trajectory and transmit power of UAVs should be determined based on the current network environment. In this case, the above optimization problem is modeled as an MDP. Here, the formulation of the MDP is presented. Then, the DRL method is proposed to deal with the JTDPA optimization problem.

A. MDP Formulation

Generally, an MDP is composed of five elements $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}_{ss'}, \gamma)$, where $\mathcal{S}$ is a state space, $\mathcal{A}$ denotes the action space of UAVs, $\mathcal{R}$ is the reward function, $\mathcal{P}_{ss'}$ means the state transition probability, and $\gamma \in [0, 1)$ is the discount rate related with the future reward's weight.

The state $\mathcal{S}(t)$ is defined to show whether all UEs meet their QoS requirements, that is, $\mathcal{S}(t) = \{s_1(t), s_2(t), \ldots, s_M(t)\}$, where $s_m(t) \in \{0, 1\}$. If the UE m meets its the minimum QoS requirement $\Gamma_{i,m}(t) \ge \Omega_m$, $s_m(t) = 1$, and otherwise $s_m(t) = 0$. Notice that the state space is 2^M, which may be extremely huge for large M.

Moreover, considering that the trajectory and transmit power of UAVs should be obtained, the action space is defined as $\mathcal{A}(t) = \{\boldsymbol{P}(t), \boldsymbol{L}(t), \boldsymbol{\phi}(t), \boldsymbol{H}(t)\}$, where $\boldsymbol{P}(t) = \{p_1(t), p_2(t), \ldots, p_N(t)\}$ is the transmit power of UAVs with $p_i(t) \in \{0, P_{max}\}$. $\boldsymbol{L}(t) = \{l_1(t), l_2(t), \ldots, l_N(t)\}$ is the horizontal distance of UAVs. Considering the horizontal trajectory constraint (1), we set $l_i(t) \in \{0, L^h_{max}\}$. $\boldsymbol{\phi}(t) = \{\phi_1(t), \phi_2(t), \ldots, \phi_N(t)\}$, $\phi_i(t) \in \{0, 2\pi\}$ is the direction angle of UAVs in the horizontal dimension. Moreover, considering the vertical trajectory constraint (2), we define vertical moving distance $\Delta h_i(t) = [h_i(t) - h_i(t-1)] \in \{-L^v_{max}, L^v_{max}\}$. Then, $\boldsymbol{H}(t) = \{\Delta h_1(t), \Delta h_2(t), \ldots, \Delta h_N(t)\}$ is the offset of UAVs in the vertical dimension.

Furthermore, to ensure that all UAVs provide all UEs with the downlink wireless services, the coverage of UEs should be considered in the reward function. If certain UE may not be covered by any UAV, there will be a punishment in the reward function. Then, based on the optimization problem (11) and by incorporating the punishment into the utility function for the collision avoidance constraint (3), the reward function can be given by

$$\begin{aligned} \mathcal{R}(t) = & \sum_{i=1}^{N} \sum_{m=1}^{M'(i)} s_m(t) [\rho_i r_{i,m}(t) - \lambda_p p_{i,m}(t)] \\ & - \zeta_1 (M - \sum_{i=1}^{N} M'(i)) - \sum_{i=1}^{N} \zeta_2^i, \end{aligned} \tag{12}$$

where $M'(i)$ is the number of UEs covered by UAV i, ζ_1 is the punishment coefficient related with UEs' coverage, ζ_2^i is the punishment of UAVs' collision. The first part of (12) is the immediate overall network utility. If UE m meets its QoS demand, $s_m(t) = 1$, and otherwise $s_m(t) = 0$. The second part of (12) is the punishment of UEs' coverage. If all UAVs cover all UEs, this part becomes zero. As for the third part of (12), when the distance between certain two UAVs is less than the minimum distance D_{min}, each UAV will be received a punishment ζ_2^i.

B. DRL Algorithm

In this paper, due to the continuous action space of the MDP, it is challenging to achieve the exact state transition probability $\mathcal{P}_{ss'}$. Although the policy-based learning method can generate continuous action, the learning variance may be high. Moreover, the value-based learning can obtain the optimal policy with the low learning variance, which can only be applied to the discrete action space. Therefore, we resort to DDPG [20] to obtain the optimal policy by combining the process of the policy-based learning (actor network) and value-based learning (critic network) approaches.

In DDPG method, the optimal policy is learned to obtain the maximum expected discounted reward $\Phi(t) = \sum_{t'=t}^{T} \gamma^{t'-t} \mathcal{R}(t')$ over a finite period T. Here, the state-action value function $Q(\mathcal{S}(t), \mathcal{A}(t))$ is defined as the expected reward at state $\mathcal{S}(t)$ with action $\mathcal{A}(t)$, which is written as $Q(\mathcal{S}(t), \mathcal{A}(t)) = E[\Phi(t)|\mathcal{S}(t), \mathcal{A}(t)]$, where $E[\cdot]$ is the expectation operator.

Moreover, based on the actor-critic (AC) framework, the actor and critic networks are implemented using a deep neural network (DNN). Here, the critic network is denoted as $Q(\mathcal{S}(t), \mathcal{A}(t)|\theta_Q)$ with the weight θ_Q, the actor network is denoted as $\mu(o(t)|\theta_\mu)$ with the weight θ_μ, where $o(t)$ is the observation of the network environment.

In order to increase the learning stability, target network strategy is designed in DDPG. Target networks of the DRL

agent are the copy of the actor and critic networks. The weights of target networks are updated by

$$\begin{aligned} \theta_{Q'} &= \tau\theta_Q + (1-\tau)\theta_{Q'}, \\ \theta_{\mu'} &= \tau\theta_\mu + (1-\tau)\theta_{\mu'}, \end{aligned} \tag{13}$$

where τ is the rate of soft updating of target networks' weight, $\theta_{Q'}$ and $\theta_{\mu'}$ are the weights of the corresponding target networks, respectively.

Moreover, due to the model-free characteristic of our method, the experience reply strategy is applied. The transition samples (state $\mathcal{S}(t)$, next state $\mathcal{S}'(t)$, action $\mathcal{A}(t)$, and reward $\mathcal{R}(t)$) are stored in the experience replay memory $\mathcal{D}$. During learning, the actor and critic networks are updated through randomly sampling mini-batches (state s_i, next state s_i', action a_i, and reward r_i) from experience reply memory $\mathcal{D}$.

Then, the actor network's weight is updated with the policy gradient method, which is written as

$$\nabla_{\theta_\mu} J(\theta_\mu) = \frac{1}{M}\sum_{i=1}^{M} \nabla_{\theta_\mu}\mu(o_i|\theta_\mu)\nabla_{a_i}Q(s_i,a_i). \tag{14}$$

where M is the size of mini batches.

Moreover, the critic network is updated by minimizing the loss function $L(\theta_Q)$, which is given by

$$L(\theta_Q) = \frac{1}{M}\sum_{i=1}^{M}\left[y_i - Q(s_i,a_i|\theta_Q)\right]^2, \tag{15}$$

where $y_i = r_i + \gamma Q'(s_{i+1}, a_{i+1}|\theta_{Q'})$ is the target value generated by the target network of critic.

Then, with (14) and (15), the weights of the actor and critic networks can be updated by $\theta_\mu \leftarrow \theta_\mu - \delta_\mu \nabla_{\theta_\mu} J(\theta_\mu)$ and $\theta_Q \leftarrow \theta_Q - \delta_Q \nabla_{\theta_Q} L(\theta_Q)$, respectively, where δ_μ and δ_Q are the corresponding learning rates. Here, we assume that $\delta_\mu = \delta_Q = \delta$.

The DDPG algorithm for the JTDPA optimization problem is summarized in Algorithm 1. The CTBS first initializes the replay memory $\mathcal{D}$, the weights of the actor-critic networks, and the corresponding target networks. The training process has EP episodes, and each episode has T time slots.

In each training episode, the network state $\mathcal{S}(t)$ is initialized firstly. In each time slot of an episode, an action is derived from the current actor network $\mu(o(t)|\theta_\mu)$ with a random noise $\varepsilon\varsigma$, where ς is a random noise with $\varsigma \sim \mathcal{N}(0,1)$ and the cofficient ε decays over time. Then, after the CTBS sends the selected action list to all UAVs, all UAVs set their own trajectories and transit powers accordingly. When certain UAV flies beyond the network area, it will select a random direction angle $\phi_i(t)$. Moreover, if the height of certain UAV $h_i(t)$ is beyond $[H_{min}, H_{max}]$, it will stay at the height of H_{min} or H_{max}. Once certain UAV covers certain hotspot, it stays without making any movement, and just waits for the trainsmit power allocation message from the CTBS.

Then, by the pilot signal, each UE can measure the received power from all UAVs. Based on the maximum received signal powers, UEs are associated with the UAVs. After user association, UEs report their own current states to their associated UAVs. Finally, with the help of the backhaul link, the CTBS can obtain the global network next state $\mathcal{S}'(t)$ and the immediate reward $\mathcal{R}(t)$. Accordingly, the information $(\mathcal{S}(t), \mathcal{A}(t), \mathcal{R}(t), \mathcal{S}'(t))$ is stored in the replay memory $\mathcal{D}$. Then, a mini-batch of transitions is sampled randomly from the replay memory $\mathcal{D}$ to update the actor and critic networks. The weights of the two target networks is slowly updated in (13). Repeat the above process until all UAVs cover all hotspots without overlapping and all UEs' QoS requirements are satisfied.

Algorithm 1 DDPG Algorithm for JTDPA Optimization Problem

- Initialize the replay memory $\mathcal{D}$.
- Initialize the critic network $Q(\mathcal{S}(t), \mathcal{A}(t)|\theta_Q)$ and the actor network $\mu(o(t)|\theta_\mu)$.
- Initialize the corresponding target networks with the weights $\theta_{Q'}$ and $\theta_{\mu'}$.
- **for** $episode = 1, \ldots, EP$
- Initialize the network state $\mathcal{S}(1)$.
- **for** $t = 1, \ldots, T$
- At the state $\mathcal{S}(t)$, the action $\mathcal{A}(t) = \mu(o(t)) + \varepsilon\varsigma$ is selected by the CTBS.
- The CTBS sends the selected action list to all UAVs.
- All UAVs set their own trajectories and transit powers accordingly.
- The CTBS obtains the network state $\mathcal{S}'(t)$ and the immediate reward $\mathcal{R}(t)$.
- The transition $(\mathcal{S}(t), \mathcal{A}(t), \mathcal{R}(t), \mathcal{S}'(t))$ is stored in $\mathcal{D}$.
- Set $\mathcal{S}(t) \leftarrow \mathcal{S}'(t)$.
- Mini-batch of transitions (s_i, a_i, r_i, s_i') is sampled randomly from $\mathcal{D}$.
- Update the weight of the critic network θ_Q by minimizing loss $L(\theta_Q)$ in (15).
- Update the weight of the actor network θ_μ in (14).
- Update the weights of the two target networks in (13).
- **If** all UAVs cover all hotspots without overlapping at the current state $\mathcal{S}(t) = \{1, \ldots, 1\}$, **then**
- break.
- **end If**
- **end for**
- **end for**

IV. Performance Evaluation

Simulation results are presented to validate the performance of the proposed DRL method. The UAV-assisted cellular network with $500m \times 500m$ area is considered. In each episode, several UEs and UAVs are located randomly. The simulation network environment parameters are defined in Table I. In DDPG algorithm, both a two-hidden-layer fully-connected neural networks (64 and 32 neurons) are designed for the actor and critic networks. The RMSPropOptimizer is used as the optimizer of AC framework. The initial value of

ε is 2 with the decay rate 0.9995. With the size of replay memory $\mathcal{D}$ 1000, the mini-batch size M 32, the discount rate γ 0.9, and the number of steps T 200, the proposed DRL model is trained for $EP = 2000$ episodes.

TABLE I
NETWORK ENVIRONMENT PARAMETERS

Parameters	Value
Channel bandwidth B	1 MHz
Downlink carrier frequency f_c	1950 MHz
Maximum transmit power of UAVs P_{max}	30 dBm
Height constraints of UAVs (H_{max}, H_{min})	(300 m, 100 m)
Noise power density N_0	-174 dBm/Hz
Collision punishment cost Ψ_i	20
Coverage cofficient Ψ_i	20
Minimum QoS requirement Ω_m	2 dB
Unit price per transmit power λ_p	2
Punishment cofficient of UEs' coverage ζ_1	120
Punishment of UAVs' collision ζ_2^i	20
Mean excessive pathloss (η_{Los}, η_{NLos})	(1 dB, 20 dB)
Elevation angle $\theta_{i,m}$	42.44°
Speed of UAVs (V_L, V_A)	(20 m/s, 5 m/s)
Minimum distance of UAVs D_{min}	50 m

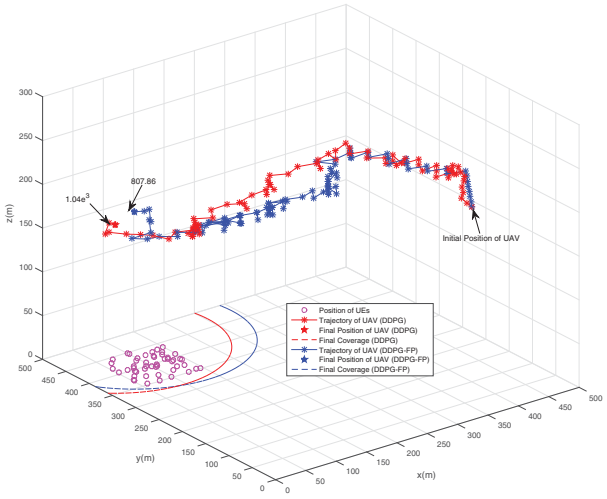

Fig. 1. Positions of the UEs and the UAVs with trajectory design and power allocation strategies with $\Omega_m = 2$.

Figure 1 demonstrates the designed three-dimensional trajectory with the power allocation strategy of one UAV. In each episode, 50 UEs are randomly distributed within the square area of $[50, 150], [350, 450]$ and one UAV starts at a random location. For comparison, the DDPG approach with fixed power allocation ($p_i(t) = P_{max}$) is also considered, which is denoted as DDPG-FP. Compared to the trajectory design with fixed power allocation (DDPG-FPA), the JTDPA strategy demonstrates the different trajectory of UAV with the same flying direction. Moreover, since the tradeoff between spectrum efficiency and interference is considered, the JTDPA optimization algorithm (DDPG) can achieve the higher network utility ($1.04e^3$) than that of DDPG-FP (807.86).

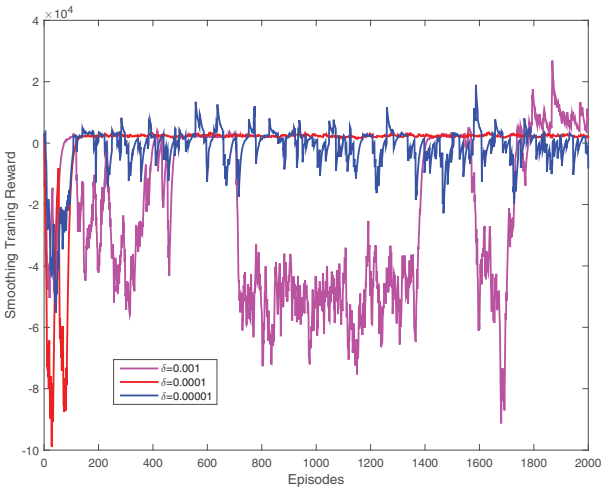

Fig. 2. Smoothing training reward with different learning rates δ.

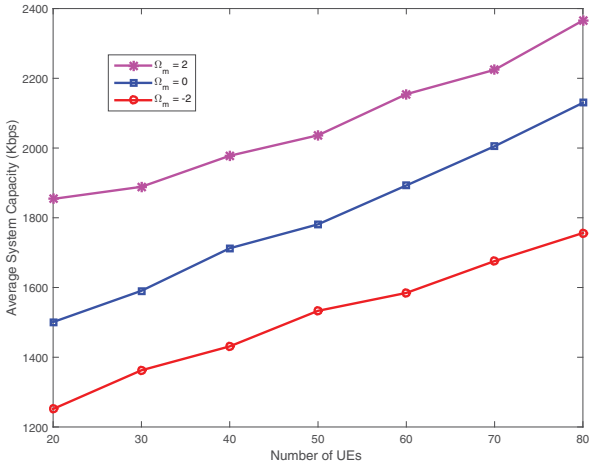

Fig. 3. Average system capacity with different numbers of UEs M.

Figure 2 shows the performance with various learning rates δ. By comparing the three cases of learning rates, we can find that the smoothing training rewards are low at first. Then, as the training episode increases, the smoothing training rewards tend to increase and converge. Furthermore, with the increase of learning rate δ, fewer training episodes are required to satisfy all UEs' QoS requirements. The training speed of $\delta = 10^{-5}$ is slower than that of $\delta = 10^{-4}$. However, if learning rate is too large, the global optimum may tend to the local optimum. We can notice that the training speed of $\delta = 10^{-3}$ is slower than that of $\delta = 10^{-4}$. Therefore, the learning rate $\delta = 10^{-4}$ is regarded to be a good choice by considering the training reward and training speed.

Figure 3 plots the average system capacity with the different numbers of UEs M and QoS requirements Ω_m. Since all UEs' QoS requirements are satisfied, the more number of UEs M is served by the UAV, the higher capacity is obtained. Furthermore, the system capacity with $\Omega_m = -2$ is always less than that of other two cases. The case of $\Omega_m = 2$ obtains

the largest system capacity among the three cases of Ω_m. When the number of UEs Ω_m is small, it may require the few training episodes to satisfy all UEs' QoS requirements, which leads to the small system capacity.

TABLE II
MEAN SYSTEM CAPACITY, NETWORK UTILITY AND COMPUTATIONAL PERFORMANCE ($\Omega_m = 0$ AND $M = 80$)

N	Method	MSC (Kbps)	MNU	CT (s)
$N = 1$	DDPG	**2081.8**	**866.6**	**130.6**
	DDPG-FPA	2061.8	856.8	159.2
	AC	2028.4	845.0	208.2
	Random	2005.6	835.7	415.5
$N = 2$	DDPG	**3502.1**	**1560.8**	**344.1**
	DDPG-FPA	3501.0	1558.5	456.5
	AC	3477.8	1558.8	367.4
	Random	3455.3	1549.2	946.6

Finally, we evaluate the performance of different optimization methods with different numbers of UAVs N. For comparison, the AC method and random algorithm are used as the joint optimization strategies. Mean system capacity (MSC), mean network utility (MNU) and computational time (CT) are shown in Table II. In the case of $N = 1$, UEs are randomly distributed within the square area of $[50, 150], [350, 450]$. As for the case of $N = 2$, UEs are randomly distributed within the two square hotspots ($[50, 150], [350, 450]$ and $[350, 450], [50, 150]$). With the increase number of UAVs N, the values of MSC, MNU and CT increase with all optimization methods. Moreover, in the four optimization strategies, the performance of the random approach is always worst in ASC, ANU and ACT. The reason is that the random approach randomly obtains the policy only based on the current immediate reward. As for the three learning methods, although the AC and DDPG-FPA algorithms have a faster convergence speed than the random approach, the two methods may converge to a local optimal point. Moreover, compared with the other three methods (DDPG-FPA, AC, and random), DDPG approach always achieves the most system capacity and network utility with low computational time cost.

V. CONCLUSION

In this paper, the DRL method is proposed to achieve the optimal JTDPA strategy in the UAV-assisted cellular networks. The optimization problem is investigated to achieve the maximum expected discounted reward while considering UEs' QoS requirement. Moreover, due to the non-convex and combinatorial features of the joint optimization issue, the DRL approach is proposed by considering the trajectory and transmit power of UAVs. Based on the experience replay and target networks, DDPG method can efficiently learn the optimal strategy with fast convergence speed. Numerical results are presented to validate the better performance of the proposed approach compared to other optimization methods.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grants 61631005, the Central Universities under Grant ZYGX2019Z022, and the 111 Project under Grant B20064.

REFERENCES

[1] Y. Zeng, R. Zhang, and T. J. Lim, "Wireless communications with unmanned aerial vehicles: Opportunities and challenges," *IEEE Communications Magazine*, vol. 54, no. 5, pp. 36-42, 2016.

[2] M. Mozaffari, W. Saad, M. Bennis, Y. Nam and M. Debbah, "A tutorial on UAVs for wireless networks: Applications, challenges, and open problems," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2334-2360, 2019.

[3] Q. Wang, Z. Chen, H. Li, and S. Li, "Joint power and trajectory design for physical-layer secrecy in the UAV-aided mobile relaying system," *IEEE Access*, vol. 6, pp. 62849-62855, 2018.

[4] G. Zhang, Q. Wu, M. Cui and R. Zhang, "Securing UAV communications via joint trajectory and power allocation," *IEEE Transactions on Wireless Communications*, vol. 18, no. 2, pp. 1376-1389, 2019.

[5] S. Zhang, H. Zhang, Q. He, K. Bian and L. Song, "Joint trajectory and power optimization for UAV relay networks," *IEEE Communications Letters*, vol. 22, no. 1, pp. 161-164, 2018.

[6] Y. Gao, H. Tang, B. Li, and X. Yuan, "Joint trajectory and power design for UAV-enabled secure communications with no-fly zone constraints," *IEEE Access*, vol. 7, pp. 44459-44470, 2019.

[7] Y. Wu, J. Xu, L. Qiu, and R. Zhang, "Capacity of UAV-enabled multicast channel: Joint trajectory design and power allocation," *in IEEE International Conference on Communications (ICC)*. 2018, pp. 1-7.

[8] G. Yang, R. Dai and Y. Liang, "Energy-efficient UAV backscatter communication with joint trajectory and resource optimization," *in IEEE International Conference on Communications (ICC)*. 2019, pp. 1-6.

[9] R. S. Sutton and A. G. Barto, Reinforcement learning: An introduction. MIT Press Cambridge, 1998.

[10] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Communications Surveys & Tutorials*, doi: 10.1109/COMST.2019.2916583.

[11] C. H. Liu, Z. Chen, J. Tang, J. Xu, and C. Piao, "Energy-efficient UAV control for effective and fair communication coverage: A deep reinforcement learning approach," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 9, pp. 2059-2070, 2018.

[12] J. Hu, H. Zhang, and L. Song, "Reinforcement learning for decentralized trajectory design in cellular UAV networks with sense-and-send protocol," *IEEE Internet of Things Journal*, vol. 6, no. 4, pp. 6177-6189, 2019.

[13] S. Yin, S. Zhao, Y. Zhao and F. R. Yu, "Intelligent trajectory design in UAV-aided communications with reinforcement learning," *IEEE Transactions on Vehicular Technology*. doi: 10.1109/TVT.2019.2923214.

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

[15] X. Liu, Y. Liu, Y. Chen, and L. Hanzo, "Trajectory design and power allocation for multi-UAV assisted wireless networks: A machine learning approach," *IEEE Transactions on Vehicular Technology*, doi: 10.1109/TVT.2019.2920284.

[16] Y. Cao, L. Zhang and Y. Liang, "Deep reinforcement learning for multi-user access control in UAV networks," *in IEEE International Conference on Communications (ICC)*. 2019, pp. 1-6.

[17] N. Zhao, Y. Liang, D. Niyato, Y. Pei, M. Wu, and Y. Jiang, "Deep reinforcement learning for user association and resource allocation in heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, doi: 10.1109/TWC.2019.2933417.

[18] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Unmanned aerial vehicle with underlaid device-to-device communications: Performance and tradeoffs," *IEEE Transactions on Wireless Communications*, vol. 15, no. 6, pp. 3949-3963, 2016.

[19] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal LAP altitude for maximum coverage," *IEEE Wireless Communications Letters*, vol. 3, no. 6, pp. 569-572, 2014.

[20] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Ried miller, "Deterministic policy gradient algorithms," *Proceedings of the 31st International Conference on International Conference on Machine Learning*, vol. 32, pp. 387-395, 2014.