Learning-based joint UAV trajectory and power allocation optimization for secure IoT networks

Dan Deng, Xingwang Li, Varun Menon, Md. Jalil Piran, Hui Chen, Mian Ahmad Jan

PII: S2352-8648(21)00048-1

DOI: https://doi.org/10.1016/j.dcan.2021.07.007

Reference: DCAN 296

To appear in: *Digital Communications and Networks*

Received Date: 18 August 2020

Revised Date: 14 July 2021

Accepted Date: 19 July 2021

Please cite this article as: D. Deng, X. Li, V. Menon, M.J. Piran, H. Chen, M.A. Jan, Learning-based joint UAV trajectory and power allocation optimization for secure IoT networks, *Digital Communications and Networks*, https://doi.org/10.1016/j.dcan.2021.07.007.

Learning-based joint UAV trajectory and power allocation optimization for secure IoT networks

Dan Deng[a], Xingwang Li*[b], Varun Menon[c], Md. Jalil Piran*[d], Hui Chen[e] and Mian Ahmad Jan[f]

[a]School of Information Engineering, Guangzhou Panyu Polytechnic, Guangzhou, 410630, China.(dengdan@ustc.edu).
[b]School of Physics and Electronic Information Engineering, Henan Polytechnic University, Jiaozuo, 454150, China. (lixingwang@hpu.edu.cn).
[c]Department of Computer Science and Engineering, SCMS School of Engineering and Technology, India. (varunmenon@ieee.org).
[d]Department of Computer Science and Engineering, Sejong University, South Korea, (email:piran@sejong.ac.kr).
[e]School of Physics and Electronic Information Engineering, Henan Polytechnic University, Jiaozuo, 454150, China. (hchen@hpu.edu.cn).
[f]Informetrics Research Group, Ton Duc Thang University, Ho Chi Minh City, Vietnam. (mianjan@tdtu.edu.vn).

Abstract

Non-Orthogonal Multiplex Access (NOMA) can be deployed in Unmanned Aerial Vehicle (UAV) networks to improve the spectrum efficiency. Due to the broadcasting feature of NOMA-UAV networks, it is essential to focus on the security of the wireless system. This paper focuses on maximizing the secrecy sum-rate under the constraint of the achievable rate of the legitimate channels. To tackle the non-convexity optimization problem, a reinforcement learning-based alternative optimization algorithm is proposed. Firstly, with the help of successive convex approximation, the optimal power allocation scheme with a given UAV trajectory is obtained by using convex optimization tools. Afterwards, through plenty of explorations on the wireless environment, the Q-learning networks approach the optimal location transition strategy of the UAV, even without the wireless channel state information.

Keywords:
Unmanned aerial vehicle (UAV); NOMA; reinforcement learning; secure communications; deep Q-learning.

1. Introduction

In the past decades, the throughput of wireless big data increases exponentially [1–3], especially when the Internet of Things (IoT) is widely adopted. Numbers of emerging wireless technologies are proposed to enhance the quality of service both in academic and industry [4–7]. Unmanned Aerial Vehicles (UAV), due to their economical and quick deployment, has been considered as an important research area for the next generation of wireless IoT networks. Thanks to the mobility, the UAV base station can hover in the air or fly to an arbitrary location to enlarge the wireless coverage area, which can help to accommodate massive connections.

Meanwhile, Non-Orthogonal Multiplex Access (NOMA), which can transmit multiple data streams simultaneously to different users, can be employed in UAV networks to improve the spectrum efficiency [8–11]. By using successive interference cancellation with suitable decoding order, NOMA outperforms classical Orthogonal Multiple Access (OMA) in terms of the achievable sum data rate of both downlink and uplink [12, 13]. In addition, the optimal user pairing scheme is studied in [14], where the user with the worst channel gain pairs with the one with the best gain. Through carefully design of the trajectory and power control [15], as well as the precoding scheme [16], the NOMA-assisted UAV networks can greatly improve its sum rate. As an extended work, the authors in [17] investigated the power allocation optimization with circular trajectory for NOMA-UAV networks under secure constrains.

On the other hand, machine learning (ML) has aroused great interest in the optimization of wireless

[*] *The corresponding authors are Xingwang Li (lixingwang@hpu.edu.cn) and Md. Jalil Piran. (piran@sejong.ac.kr).

networks in recent years [18]. Due to the model-free feature, Reinforcement Learning (RL) is capable of tackling the optimization problems for NOMA-UAV networks [19], where it is difficult to obtain plenty of training data. Specifically, Q-learning networks, which are modeled as a Markov Decision Process (MDP), have been widely used in reinforcement learning algorithms. Generally, Q-learning networks consist of a set of states, a set of actions, the reward function as well as the Q-table [20]. In each exploration of Q-learning, the Q-table is updated according to the Bellman equation, where the Q-table depends on the current state, the next action, as well as the reward. Through plenty of explorations or exploitations of the current Q-table, the optimal state transition function can be obtained. For example, a RL-based approach is proposed in [21] to optimize the trajectory of the UAV base station, and simulation results show that this algorithm can achieve 3 times of average user throughput gain. Focusing on the uplink sum rate of UAV networks, the RL method is used in [22] to track the group mobile users with acceptable performance loss. By using a stochastic game, authors in [23, 24] extended single UAV to multiple UAV scenarios, where a multi-agent RL method is introduced to the joint optimization problem on user, power allocation and sub-channel selection scheme.

Due to the broadcasting feature of NOMA-UAV networks, it is essential to focus on the security of the wireless system. To this end, the aim of this paper is to maximize the secrecy sum-rate under the constraint of the achievable rate of the legitimate channels. Due to the non-convexity of the objective function, it is hard to solve the optimization problem directly. To tackle the issue, a reinforcement learning-based alternative optimization algorithm is proposed. Firstly, with the help of successive convex approximation, the optimal power allocation scheme with a given UAV trajectory is obtained by using convex optimization tools. Afterwards, through plenty of explorations on the wireless environment, the Q-learning networks approach the optimal location transition strategy of the UAV, even without the wireless channel state information. Simulation results are provided to verify the converge and the effectiveness of the proposed algorithm.

The main contributions of this paper can be summarized as follows:

- A deep RL-based optimization algorithm is proposed to maximize the secrecy sum-rate under the constraint of the achievable rate of the legitimate channels, where the trajectory and the power allocation scheme for UAV-NOMA are jointly optimized through the two-stage Q-learning networks.
- We provide deep insight into the effects of the system parameters, such as the predefined rate for the legitimate channels, the altitude of the UAV as well as the transmission power, on the secrecy sum rate through simulation results.

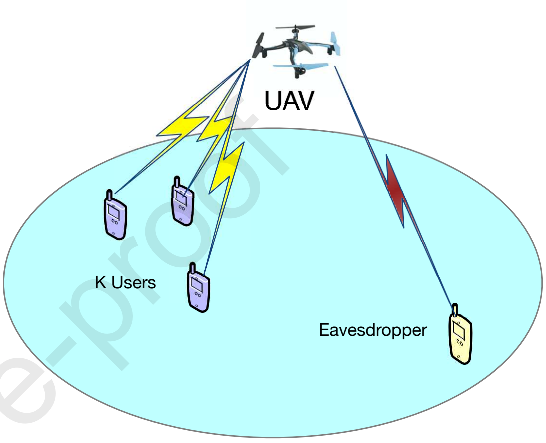

Figure 1: System model of joint UAV trajectory and power allocation optimization for secure IoT networks.

2. System model

As depicted in Fig.1, there is one UAV node, K associated users, as well as a passive eavesdropper. All nodes are assumed to be equipped with a single antenna, and all links experience Line-of-Sight (LoS) propagation. Specifically, the locations of users are denoted as $\mathbf{L}_i = [x_i, y_i]^T, i \in [1, K]$, while the eavesdropper's location is L^e. Moreover, the UAV flies with fixed altitude H, and the horizontal trajectory of the UAV at the n-th time slot is defined as $\mathbf{W} = \{\mathbf{w}[n] = [x[n], y[n]]^T, n = 1, 2, ..., N.\}$. The period from the first location to the last location is denoted as T, and the time interval between adjacent locations is given as $\delta = T/N$.

The channel power gain from the UAV to the i-th user can be given as

$$\begin{aligned} g_i[n] &= \rho_o d_i^{-\alpha}[n] \\ &= \rho_o(H^2 + ||\mathbf{w}[\mathbf{n}] - L_i||^2)^{-\alpha/2}, \forall i \in \Omega, n \in [1, N], \end{aligned} \quad (1)$$

where ρ_o is the reference power gain at the distance $d_i = 1m$, and $\alpha \geq 2$.

Note that the Doppler effect is negligible when the moving velocity of the UAV is small enough. Thus, in this case, the wireless channels from the UAV to the users is dominated by the LOS component. Therefore, LOS model is mostly adopted in the current research literature on UAV [14, 15, 17, 25]. Furthermore, our research results can be easily extended to NLOS or composite propagation models [26].

Similarly, the channel power gain from the UAV to the eavesdropper at the n-th time slot can be given as

$$\begin{aligned} g^e[n] &= \rho_o d_e^{-\alpha}[n] \\ &= \rho_o(H^2 + \|\mathbf{w}[\mathbf{n}] - \mathbf{L}^e[n]\|^2)^{-\alpha/2}, \forall n \in [1, N]. \end{aligned} \tag{2}$$

Since NOMA protocol is adopted at the UAV, the downlink transmission signal is a linear superposition of K data streams with different power allocation factors. Consequently, the downlink NOMA signal can be expressed as can be expressed as

$$x = \sum_{i=1}^{K} \sqrt{\xi_i P} x_i, \tag{3}$$

where P is the transmission power of the UAV, x_i is the original message transmitted from the UAV to the i-th user, and ξ_i is the power allocation factor for x_i.

In addition, we use $\xi[n]$ to denote the power allocation scheme for n-th location, i.e., $\xi[n] = [\xi_1[n], \xi_2[n], ..., \xi_K[n]]^T$. We use $\mathbf{\Xi}$ to denote the power allocation scheme for all locations, i.e., $\mathbf{\Xi} = \{\xi[n], n \in [1, N]\}$.

Considering the total power constrain, we have the following equation

$$\sum_{i\in\Omega} \xi_i[n] = 1, \forall n \in [1, N]. \tag{4}$$

Thus, the received signal at the user can be given as

$$y_i = \sqrt{g_i} x + n_i = \sqrt{g_i} \sum_{i\in\Omega} \sqrt{\xi_i P} x_i + n_i, \tag{5}$$

where $n_i \sim \mathcal{CN}(0, \sigma^2)$ is the additive white Gauss noise.

Without loss of generality, it is assumed that the channel power gains are in an ascending order with respect to the index, i.e., $g_1 \leq g_1 \leq ... \leq g_K$. According the NOMA protocol, the power allocation factors are in a descending order with respect to the index, i.e., $\xi_1 \geq \xi_2 \geq ... \geq \xi_K$.

3. Problem formulation

According to the NOMA protocol, the perfect Successive Interference Cancellation (SIC) receiver is adopted at all users based on the power allocation factors. That is, the k-th user decodes the first $k - 1$ data streams with larger power before decoding its own message, while the residual data streams with smaller power will be treated as interference. In addition, the SNR of k-th data stream at the i-th user at the n-th time slot can be given as

$$\begin{aligned} \gamma_{i,k}[n] &= \frac{\xi_k[n]}{I_k[n] + \frac{\sigma^2}{g_i P}} \\ &= \frac{\xi_k[n]}{I_k[n] + \frac{\sigma^2}{\rho_o P}(H^2 + \|\mathbf{w}[n] - \mathbf{L}_i\|^2)^{\alpha/2}}, \end{aligned} \tag{6}$$

with

$$I_k[n] = \sum_{j=k+1}^{K} \xi_j[n], \forall n \in [1, N]. \tag{7}$$

Then, the achievable transmission rate of k-th data stream at the k-th user at the n-th time slot can be given as

$$R_{k,k}[n] = \log_2(1 + \gamma_{k,k}[n]), \forall k \in [1, K], n \in [1, N]. \tag{8}$$

On the other hand, it is assumed that the eavesdropper has no prior information on the decoding order, and all other data streams are treated as interference. The corresponding SINR of k-th data stream at the eavesdropper at the n-th time slot can be given as

$$\begin{aligned} \gamma_k^e[n] &= \frac{\xi_k[n]}{I_k^e[n] + \frac{\sigma^2}{g^e P}} \\ &= \frac{\xi_k[n]}{I_k^e[n] + \frac{\sigma^2}{\rho_o P}(H^2 + \|\mathbf{w}[n] - \mathbf{L}^e[n]\|^2)^{\alpha/2}}, \end{aligned} \tag{9}$$

with

$$I_k^e[n] = \sum_{j\neq k} \xi_j[n], \forall n \in [1, N]. \tag{10}$$

Then the achievable rate of the k-th data stream at the eavesdropper can be calculated as

$$R_k^e[n] = \log_2(1 + \gamma_k^e[n]), \forall k \in [1, K], n \in [1, N]. \tag{11}$$

Thus, according to the secure communication model, the secrecy rate of the k-th user at the n-th time slot is given as

$$\begin{aligned} r_k[n] &= (R_{k,k}[n] - R_k^e[n])^+ \\ &= \Big[\log_2 \frac{(1+\gamma_{k,k}[n])}{(1+\gamma_k^e[n])}\Big]^+ = \Big[\log_2 \frac{1+\frac{\xi_k[n]}{I_k[n]+\frac{\sigma^2}{g_k[n]P}}}{1+\frac{\xi_k[n]}{I_k^e[n]+\frac{\sigma^2}{g^e[n]P}}}\Big]^+ \\ &= \Big[\log_2(1+\frac{\xi_k[n]}{I_k[n]+c_k[n]}) + \log_2(I_k^e[n]+q[n]) \\ &\quad - \log_2(1+q[n])\Big]^+, \end{aligned} \tag{12}$$

with

$$c_k[n] = \frac{\sigma^2}{g_k[n]P}, q[n] = \frac{\sigma^2}{g^e[n]P} \tag{13}$$

The goal of this paper is to maximize the toal secrecy rate of all users by jointly optimizing power allocation scheme and UAV trajectory, with a predefined threshold rate R_j^{th} for the all users, i.e.,

$$R_{k,k}[n] \geq R_k^{th}, \forall k \in [1,K], n \in [1,N]. \tag{14}$$

The optimization problem can be formulated as

$$(\text{P0}): \max_{\mathbf{W},\Xi,I} \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} r_k[n] \tag{15}$$

$$s.t. \quad 0 \leq \xi_i[n] \leq 1, \forall n \in [1,N], \tag{16}$$

$$\sum_{i\in\Omega} \xi_i[n] = 1, \forall n \in [1,N], \tag{17}$$

$$\|\mathbf{w}[n] - \mathbf{w}[n+1]\|^2 \leq (v_m\delta)^2, \forall n \in [1,N-1], \tag{18}$$

$$R_{k,k}[n] \geq R_k^{th}, \forall k \in [1,K], n \in [1,N]. \tag{19}$$

where v_m is the maximum horizontal velocity of the UAV.

4. Optimization of power allocation factors

Since it is difficult to solve the problem in (15) directly, a two-stage joint trajectory and power allocation alternating optimization algorithm are adopted in this paper.

Firstly, with a given trajectory $\mathbf{w}[n]$ at n-th time slot, the optimal power allocation factors $\xi[n]$ is obtained through successive convex optimization technique. Consequently, the UAV trajectory optimization problem is solved by deep reinforcement learning-based iterative algorithm. With the help of deep Q-learning networks, the UAV can learn the environment of the sum secrecy rate, and move to the best location.

With a given UAV trajectory $\mathbf{w}[n]$ at n-th time slot, the optimization problem of sum-rate of $r_k[n]$ with respect to $\xi[n]$ can be reformulated as

$$\begin{aligned} (\text{P1}): \max_{\xi[n],I[n]} \sum_{k=1}^{K} r_k'[n], \\ s.t. \quad (14),(16),(17). \end{aligned} \tag{20}$$

where

$$\begin{aligned} r_k'[n] &= \log_2(1+\frac{\xi_k[n]}{I_k[n]+c_k[n]}) \\ &\quad + \log_2(I_k^e[n]+q[n]) - \log_2(1+q[n]) \end{aligned} \tag{21}$$

We turn to consider the rate constraint in (14), which can be rewritten as

$$\gamma_{k,k}[n] \geq \gamma_k^{th}, \forall k \in [1,K] \tag{22}$$

where $\gamma_k^{th} = 2^{R_k^{th}} - 1$.

By substituting (6) into (22), we have

$$\xi_k[n] \geq \gamma_k^{th}(I_k[n]+c_k[n]), \forall k \in [1,K]. \tag{23}$$

Considering the objective function in (20), it is easy to prove that $r_k'[n]$ is concave with respect to $\xi_k[n]$ and $I_k^e[n]$, convex with respect to $I_k[n]$. Thus, the objective function in (20) is non-convex and intractable to be solved. To tackle the problem in (20), the successive convex optimization technique is used to give a tight approximation solution. The main idea of successive convex optimization is to convert the non-convex function to a linear function by using first-order Taylor expansion at the prior feasible solution.

To perform the successive convex optimization, a set of auxiliary variables $\eta_k[n]$ is introduced to reformulate (20) as

$$(\text{P2}): \max_{\xi[n],I[n],\eta[n]} \sum_{k=1}^{K} \hat{r}_k[n], \tag{24}$$

$$\begin{aligned} s.t. \quad & I_k[n]+c_k[n] \leq \eta_k[n], \forall k \in [1,K], \\ & (23),(16),(17). \end{aligned} \tag{25}$$

where

$$\begin{aligned} \hat{r}_k[n] &= \log_2(I_k[n]+c_k[n]+\xi_k[n]) \\ &\quad + \log_2(I_k^e[n]+q[n]) \\ &\quad - \log_2\eta_k[n] - \log_2(1+q[n]) \end{aligned} \tag{26}$$

Since the objective function in (24) is convex with respect to $\eta_k[n]$, the following lower-bound can be obtained given r-th iteration results $\eta_k^r[n]$.

$$\begin{aligned}\hat{r}_k[n] \geq& \log_2(I_k[n] + c_k[n] + \xi_k[n]) + \log_2(I_k^e[n] + q[n])\\ &- \log_2(1 + q[n]) - \log_2 \eta_k^r[n]\\ &- \frac{1}{\ln 2 * \eta_k^r[n]}(\eta_k[n] - \eta_k^r[n])\\ &= r_k^{lb}[n].\end{aligned} \quad (27)$$

Thus, the optimization problem in (24) can be reformulated as

$$\text{(P3)}: \max_{\xi[n], I[n], \eta[n]} \sum_{k=1}^{K} r_k^{lb}[n] \quad (28)$$
$$s.t. \quad (25), (23), (16), (17).$$

We can observe that the objective function in (28) is concave with respect to $\xi_k[n], I_k[n]$ and $I_k^e[n]$, and affine with respect to $\eta_k[n]$. Also, all the constraints of (P3) are affine. Thus, the optimization problem in (P3) is convex and can be efficiently solved by classical convex optimization tools, such as CVX [27, 28].

The details of the iteration optimization algorithm for (P3) is given as in **Algorithm 1**.

Algorithm 1 Iteration optimization algorithm for (P3)

```
Initialize $\xi^r[n], I^r[n], I^{e,r}[n], \eta^r[n]$, and set iteration index $r = 0$.
repeat
    With $\eta^r[n]$, use CVX to solve problem (P3) and obtain the optimal solution $\xi^{r+1}[n], I^{r+1}[n], I^{e,r+1}[n], \eta^{r+1}[n]$.
    Update iteration index: $r = r + 1$.
until $r$ reaches the maximum iteration number, or the increase of the objective function in (P3) is smaller than the predefined threshold $\tau$.
```

5. Q-learning algorithm on UAV movement

In this section, given power allocation factors $\xi[n]$ at the n-th time slot, our objective is to obtain the *best* movement to maximize the sum secrecy rate for the next UAV location. Since the optimization problem is hard to be dealt with, an effective algorithm is required to solve the horizontal trajectory optimization. As the sum secrecy rate is related to the distance between the UAV to all user, as well as the distance between the UAV to the eavesdropper, the RL-based algorithm is proposed to solve the optimization problem.

Q-learning algorithm is a powerful model-free reinforcement learning method, which is widely used for resource allocation and channel estimation in wireless communication networks. The main advantage of Q-learning algorithm is that it is not necessary to obtain the channel state information or the state transition probability. To this end, deep Q-learning network is adopted in this paper to assist the learning of the environment of secure NOMA communications.

In our Q-learning model, the UAV acts as an intelligent agent, and the Q-learning model consists of four elements, i.e., the state space S, the action space A, the reward function R_a as well as the Q-table Q. In each step of the Q-learning network, the agent explores the environment from the initial state, calculates the reward of the selected action, and updates the Q-table step by step.

In order to simplify the problem, without loss of generality, it is assumed that actions for each step are the finite set of coordinates, which are used to denote 5 directions of UAV nodes, i.e., $A = \{0, 1, 2, 3, 4\}$. We use $\lambda = v_m\delta$ to denote the length of each movement. Specifically, $a = 0$ indicates that the UAV holds statically, and $a = 1, 2, 3, 4$ means the UAV moves forward, backward, left turn and right turn with length λ, respectively. Note that any number of directions can be used in our proposed algorithm, while The number of directions are a tradeoff between the computational complexity and the approximation accuracy. That is, the larger number of directions leads to higher complexity and higher accuracy.

Afterwards, all user are randomly distributed within a square area with the size $M\lambda \times M\lambda$. And the coordinates of the UAV measured based on λ are used as the states of the agent, i.e.,

$$S = \{s_n = [x[n], y[n]]^T, n \in [1, N]\} \quad (29)$$

with

$$x[n] \in [0, M-1], y[n] \in [0, M-1]. \quad (30)$$

Moreover, we set the initial state as coordinate origin, i.e., $s[1] = [0, 0]^T$.

The reward function Ra is dependent on the current state s as well as the selected action a. According to the optimization problem in (15), R_a can be modeled as

$$R_a = \sum_{k=1}^{K} r_k[n]. \quad (31)$$

$$Q_{n+1}(s_n, a_n) = (1-\theta)Q_n(s_n, a_n) + \theta\Big[R_n + \beta \max_{a\in A} Q_n(s_{n+1}, a)\Big], \quad (32)$$

where $\theta \in (0, 1]$ denotes the learning rate that is also the weight of the current reward. The larger the θ, the faster learning speed. $\beta \in [0, 1]$ represents the discount factor, which indicates the importance of the future earnings, i.e., the larger the β, the greater the forward returns.

In each step, ϵ-greedy policy is adopted for the UAV to select an action, which is also a tradeoff between the exploration of environment and the exploitation of the current Q-table. Specifically, according to the ϵ-greedy policy, the UAV exploits the optimal action based on the current Q-table with probability ϵ, and explores the other actions by randomly selecting an action with probability $(1-\epsilon)$. Generally, ϵ may be large enough to guarantee that the current optimal action is hit with high probability.

The details of the deep Q-learning algorithm for (15) is given as in **Algorithm 2**.

Algorithm 2 Deep Q-learning Algorithm for (P0)

```
Initialize $\xi$, $s$, initialize $Q(s, a)$ with arbitrary value, and set iteration index $n = 1$.
repeat
    For each step of iterations
    Employ $\epsilon$-greedy policy to select action $a_n$, updated state $s_{n+1}$.
    Given UAV location $s_{n+1}$, calculate the optimal power allocation factors $\xi[n+1]$ as in Algorithm 1 .
    Observe the reward $R_a$ according to (31).
    Update Q-table as in (32).
    Update iteration index: $n = n + 1$.
until $n$ reaches the maximum iteration number $N$.
```

6. Simulation results

In this section, simulation results are provided to validate the converge and the effectiveness of the proposed algorithm. In all simulations, unless otherwise specified, we set the UAV's altitude as $H = 20m$, the maximum horizontal velocity of the UAV as $v_m = 10m/s$, and the interval of each time slot is set as $\delta = 0.2s$. The square area of the UAV's trajectory is a $100 \times 100m$, and the action space is a set with five different directions, i.e., $A = \{0, 1, 2, 3, 4\}$. The number of users is set as $K = 9$, and the locations of K users are given as $\mathbf{L}_k = [2k, 2k]', \forall k \in [1, K]$. The initial location of the UAV is given as $[90, 0]'m$, while that of the Eavesdropper is $[100, 100]'m$.

In **Algorithm 1**, the predefined increase threshold is set as $\tau = 0.001$, and the maximum iteration number is 30. As to **Algorithm 2**, the key parameters are given as $\epsilon = 0.95$ and $\theta = 1, \beta = 0.4$, while the number of time slots is set as $N = 10000$. The transmission power of the UAV is $P = 30dBm$, the reference power gain is set as $\rho_0 = -30dB$, the noise power is given as $\sigma^2 = -74dBm$. While the path loss exponent is $\alpha = 2$, and the threshold rate is $R_k^{th} = 0.5bps/Hz, \forall k \in [1, K]$.

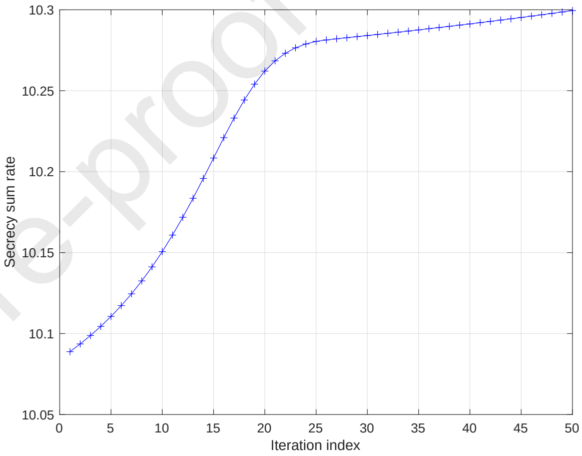

Figure 2: The converge of **Algorithm 1** with respect to the iteration number r.

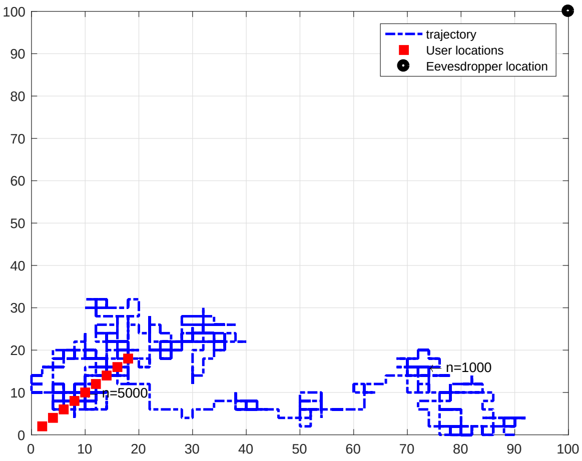

Figure 3: The trajectory of the UAV in **Algorithm 2** .

Firstly, the converge of **Algorithm 1** is presented in Fig.2, where there are $K = 9$ legitimate users in total,

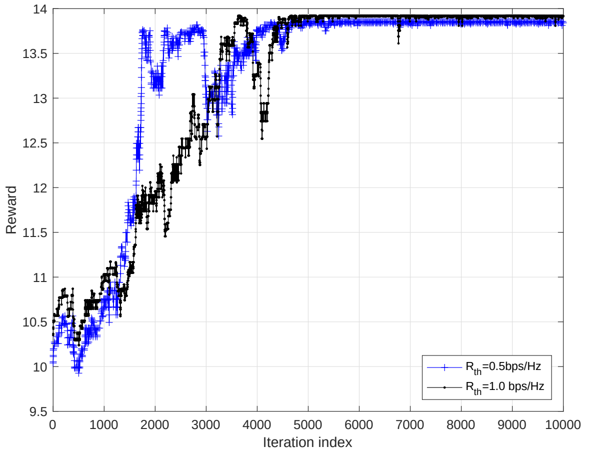

Figure 4: The reward of **Algorithm 2** with respect to the iteration number.

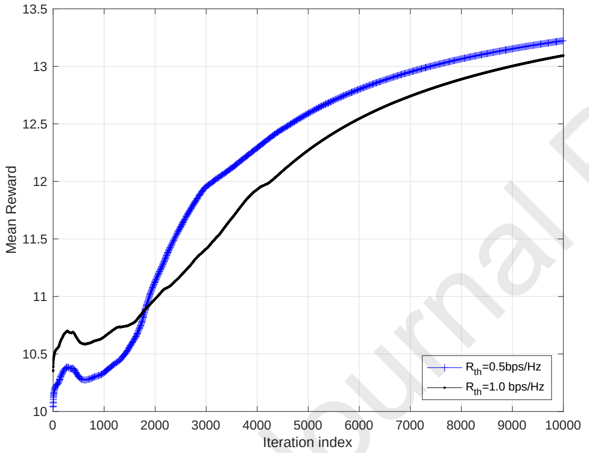

Figure 5: The mean reward of **Algorithm 2** with respect to the iteration number.

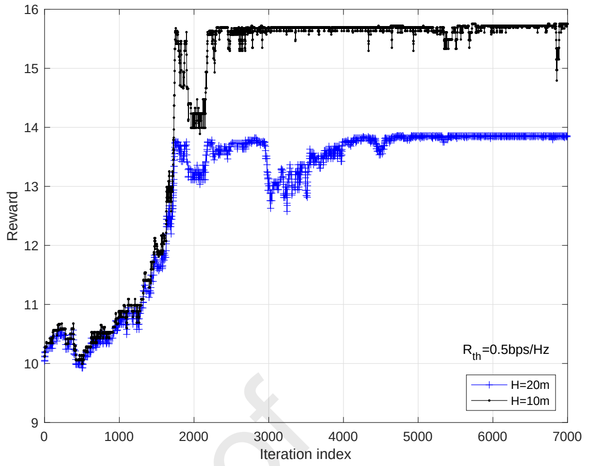

Figure 6: The reward function curves with respect to the altitude of UAV.

and the eavesdropper is located as $[100, 100]'m$. We can observe that, when the iteration number reaches 30, the secrecy sum rate begins to converge, which validates the effectiveness of the proposed algorithm. Thus, in the following simulations, we set the maximum iteration number of **Algorithm 1** as 30.

As to the trajectory optimization in **Algorithm 2**, an example is presented in Fig.3. We can see from this figure that the UAV flies to the *best* location with the help of deep Q-learning networks. Specifically, when the iteration number reaches 5000, the locations of the UAV remain stable near the origin. The reason is that, with the positions around the origin, the UAV is the farthest from the eavesdropping node and the closest to the legitimate users. Thus, it is the optimal position for the secure UAV-NOMA networks.

On the other hand, the reward for each step is shown in Fig.4, where in the first 1800 explorations the rewards increase sharply, while become stationary in the following phase. When the iteration number is larger than 4000, the reward of the UAV-NOMA becomes converge to 13.8 bps/Hz. To give an insight into the effects of the predefined threshold rate constrains for the legitimate channels R_th on the system performance, we compare the rewards curve with $R_{th} = 1.0bps/Hz$ with that of $R_{th} = 0.5bps/Hz$ in Fig.4. It is observed that, the larger the rate constraints are, the slower the converging speed of the reward is. The reason is that, when the rate constraints are stronger, it is more difficult for the UAV to obtain the optimal power allocation scheme for the NOMA protocol. As such, the growth rate of the reward is slowing down. Moreover, the mean reward with respect to iteration number is given in Fig.5, with $R_{th} = 1.0bps/Hz$ and $R_{th} = 0.5bps/Hz$, respectively. We can see from this figure, with weak rate constraints for the legitimate channels, larger secrecy sum rate can be obtained. Specifically, about $0.2bps/Hz$ gain can be obtained when the rate constraints change from $R_{th} = 1.0bps/Hz$ to $R_{th} = 0.5bps/Hz$. All results validate the convergence and the effectiveness of the proposed two-stage algorithm.

The effects of the UAV's altitude H are depicted in Fig.6 and Fig.7. We can see from Fig.6 that, the reward of UAV-NOMA is sensitive with respect to H. When H changes from $20m$ to $10m$, the stationary reward changes from 13.8 *bps/Hz* to 15.7 *bps/Hz*. The

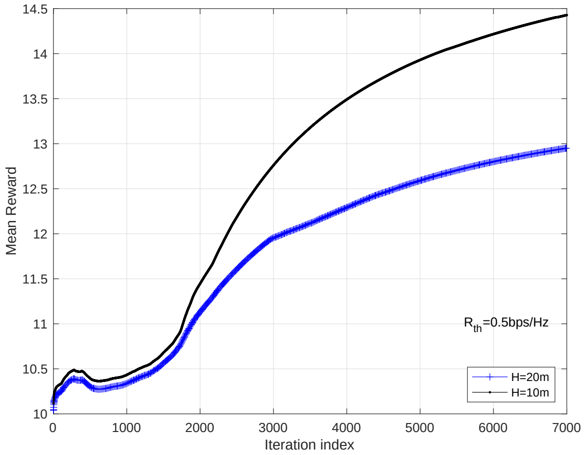

Figure 7: The mean reward function curves with respect to the altitude of UAV.

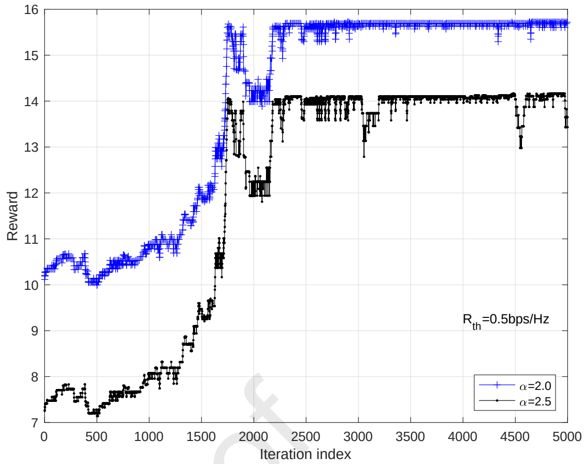

Figure 8: The reward function curves with respect to the path loss exponent α.

reason for the system performance gain is that, when the UAV's altitude H becomes lower, the wireless channels of the legitimate users become stronger. Thus, the secrecy sum rate can be improved. Similar results can be observed in Fig.7, where the mean reward is compared with different H.

Also, the effects of the path loss exponent α on the system performance are presented in Fig.8 and Fig.9, where the UAV's altitude H is set as $H = 10m$, and the path loss exponent α is set as $\alpha = 2.0$ and $\alpha = 2.5$, respectively. One can see from Fig.8 that the stationary reward values of the system changes from 15.7 to 13.8 bps/Hz, when α changes from 2.0 to 2.5. The reason is that when the path loss exponent α grows larger, both the capacity of the legitimate channels and the eavesdropping channel decrease sharply. As such, the secrecy rate sum becomes smaller. Also, we can find that there exists a performance degradation when the iteration number reaches 3000. The reason is that we adopt the ϵ-greedy policy to select the optimal action in each iteration. Specifically, $\epsilon = 0.95$. According to the ϵ-greedy policy, the UAV exploits the optimal action based on the current Q-table with probability ϵ, and explores other actions by randomly selecting an action a probability of 0.05. Generally, ϵ may be large enough to guarantee that the current optimal action is hit with high probability. Thus, there exists the probability of 0.05 that a random action is selected. Then, there will be a performance degradation. However, when the iteration number reaches large enough, the performance will converge to the optimal value. Similar results can be observed in Fig.9, where the mean reward value is considered.

7. Conclusion

NOMA can be deployed in UAV networks to improve the spectrum efficiency. Due to the broadcasting feature of NOMA-UAV networks, it is essential to focus on the security of the wireless system. This paper focuses on maximizing the secrecy sum rate under the constraints of the achievable rate of the legitimate channels. To tackle the non-convexity optimization problem, a RL-based alternative optimization algorithm is proposed. In the first step of the proposed two-stage method, a Q-learning network is adopted to obtain the optimal action with given the location of the UAV node. Through a lot of exploration of the wireless environment, the Q-learning networks approach the optimal location transfer strategy of the UAV, even without the wireless channel state information. Afterwards, with the help of successive convex approximation, the optimal power allocation scheme for the updated trajectory is obtained by using convex optimization tools. Simulation results verify the convergence and effectiveness of the proposed algorithm. In future works, we will combine NOMA-UAV with other emerging technologies, such as Reconfigurable Intelligent Surface (RIS) and mmWave, and discuss the effects of system parameters on the secure communications.

Acknowledgements

This work was partly supported by the Natural Science Foundation of Guangdong Province with grant number 2018A030313736, the Scientific Research Project of Education Department of Guangdong with

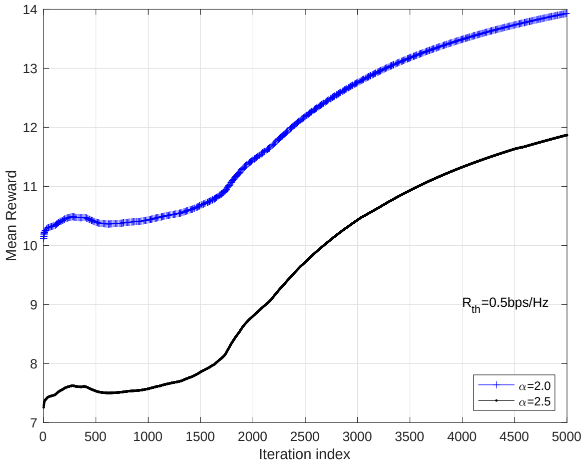

Figure 9: The mean reward function curves with respect to the path loss exponent α.

grant number 2019GZDXM002, the Project of Technology Development Foundation of Guangdong with grant number 706049150203, the Application Technology Collaborative Innovation Center of GZPYP with grant number 2020ZX01, the Yangcheng scholar, scientific research project of Guangzhou Education Bureau with grant number 202032761. the Guangzhou Municipal Education Bureau of China with grant number 201831785, the Guangzhou Municipal Science and Technology Bureau of China with grant number 202002030133, the Key Scientific Research Projects of Higher Education Institutions in Henan Province under Grant 20A510007, the Outstanding Youth Science Foundation of Henan Polytechnic University under Grant J2019-4, the the Fundamental Research Funds for the Universities of Henan Province under Grant NSFRF180309, the National Natural Science Foundation of China under Grant 62001320, the Key Research and Development Program of Shanxi under Grant 201903D121117, the Natural Science Foundation of China under Grant 61901367, the Natural Science Foundation of Shaanxi Province under Grant 2020JQ-8.

Data Availability

The data of this work can be available through the request on the corresponding author by e-mail.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

[1] S. Jacob, V. G. Menon, S. P. G, F. S. KS, B. Mahapatra, S. Joseph, Bidirectional multi-tier cognitive swarm drone 5g network, in: IEEE INFOCOM 2020, Toronto, Canada, 2020, 2020, pp. 1219–1224. doi:10.1109/ICC.2019.8761681.

[2] J. Zhu, M. Zhao, S. Zhang, W. Zhou, Exploring the road to 6g: Abc foundation for intelligent mobile networks, China Communications 17 (6) (2020) 51–67.

[3] J. Zhu, C. Gong, S. Zhang, M. Zhao, W. Zhou, Foundation study on wireless big data: Concept, mining, learning and practices, China Ccommunications 15 (12) (2018) 1–15.

[4] S. J. et al, A novel spectrum sharing scheme using dynamic long short-term memory with cp-ofdma in 5g networks, IEEE Transactions on Cognitive Communications and Networking pp (99) (2020) 1–1. doi:10.1109/TCCN.2020.2970697.

[5] D. Deng, L. Fan, X. Lei, W. Tan, D. Xie, Joint user and relay selection for cooperative NOMA networks, IEEE Access 5 (2017) 20220–20227.

[6] Z. Ming, S. Zhou, W. Zhou, J. Zhu, An improved uplink sparse coded multiple access, IEEE Communications Letters 21 (1) (2017) 176–179.

[7] Y. Wang, M. Zhao, D. Deng, S. Zhou, W. Zhou, Fractional sparse code multiple access and its optimization, IEEE Wireless Communications Letters 7 (6) (2018) 990–993.

[8] Y. Huang, W. Mei, J. Xu, L. Qiu, R. Zhang, Cognitive uav communication via joint maneuver and power control, IEEE Transactions on Communications 67 (11) (2019) 7872–7888.

[9] X. Li, M. Huang, Y. Liu, V. G. Menon, I/q imbalance aware nonlinear wireless-powered relaying of b5g networks: Security and reliability analysis, arXiv preprint pp (99) (2020) 1–1. doi:https://arxiv.org/abs/2006.03902.

[10] X. Li, M. Zhao, Y. Liu, L. Li, Z. Ding, A. Nallanathan, Secrecy analysis of ambient backscatter noma systems under i/q imbalance, IEEE Transactions on Vehicular Technology pp (99) (2020) 1–1.

[11] X. Li, Q. Wang, Y. Liu, T. A. Tsiftsis, Z. Ding, A. Nallanathan, Uav-aided multi-way noma networks with residual hardware impairments, IEEE Wireless Communications Letters pp (99) (2020) 1–1.

[12] Z. Yang, W. Xu, C. Pan, Y. Pan, M. Chen, On the optimality of power allocation for noma downlinks with individual qos constraints, IEEE Communications Letters 21 (7) (2017) 1649–1652.

[13] N. Zhang, J. Wang, G. Kang, Y. Liu, Uplink nonorthogonal multiple access in 5g systems, IEEE Communications Letters 20 (3) (2016) 458–461.

[14] L. Zhu, J. Zhang, Z. Xiao, X. Cao, D. O. Wu, Optimal user pairing for downlink non-orthogonal multiple access (noma), IEEE Wireless Communications Letters 8 (2) (2019) 328–331.

[15] N. Zhao, X. Pang, Z. Li, Y. Chen, F. Li, Z. Ding, M. Alouini, Joint trajectory and precoding optimization for uav-assisted noma networks, IEEE Transactions on Communications 67 (5) (2019) 3723–3735.

[16] W. Wang, J. Tang, N. Zhao, X. Liu, Y. Qian, Joint precoding optimization for secure swipt in uav-aided noma networks, IEEE Transactions on Communications PP (99) (2020).

[17] X. Chen, Z. Yang, N. Zhao, Y. Chen, J. Wang, Z. Ding, R. Yu, Secure transmission via power allocation in noma-uav networks with circular trajectory, IEEE Transactions on Vehicular Technology (2020) 1–1.

[18] Y. Lin, M. Wang, X. Zhou, G. Ding, S. Mao, Dynamic spectrum interaction of uav flight formation communication with priority: A deep reinforcement learning approach, IEEE Transactions on Cognitive Communications and Networking (2020) 1–1.

[19] W. Zhang, K. Song, X. Rong, Y. Li, Coarse-to-fine uav target tracking with deep reinforcement learning, IEEE Transactions

on Automation Science and Engineering 16 (4) (2019) 1522–1530.

[20] Z. Yang, Y. Liu, Y. Chen, L. Jiao, Learning automata based q-learning for content placement in cooperative caching, IEEE Transactions on Communications 68 (6) (2020) 3667–3680.

[21] V. Saxena, J. Jaldn, H. Klessig, Optimal uav base station trajectories using flow-level models for reinforcement learning, IEEE Transactions on Cognitive Communications and Networking 5 (4) (2019) 1101–1112.

[22] S. Yin, S. Zhao, Y. Zhao, F. R. Yu, Intelligent trajectory design in uav-aided communications with reinforcement learning, IEEE Transactions on Vehicular Technology 68 (8) (2019) 8227–8231.

[23] J. Cui, Y. Liu, A. Nallanathan, Multi-agent reinforcement learning-based resource allocation for uav networks, IEEE Transactions on Wireless Communications 19 (2) (2020) 729–743.

[24] X. Liu, Y. Liu, Y. Chen, Reinforcement learning in multiple-uav networks: Deployment and movement design, IEEE Transactions on Vehicular Technology 68 (8) (2019) 8036–8049.

[25] Q. Wu, J. Xu, R. Zhang, Capacity characterization of uav-enabled two-user broadcast channel, IEEE Journal on Selected Areas in Communications 36 (9) (2018) 1955–1971. doi:10.1109/JSAC.2018.2864421.

[26] P. S. Bithas, A. G. Kanatas, D. B. Da Costa, P. K. Upadhyay, U. S. Dias, On the double-generalized gamma statistics and their application to the performance analysis of v2v communications, IEEE Transactions on Communications (2017) 1–12.

[27] M. Grant, S. Boyd, CVX: Matlab software for disciplined convex programming, version 2.1, http://cvxr.com/cvx (Mar. 2014).

[28] M. Grant, S. Boyd, Graph implementations for nonsmooth convex programs, Lecture Notes in Control and Information Sciences, Springer-Verlag Limited, 2008.