RIS-Assisted UAV-D2D Communications Exploiting Deep Reinforcement Learning

YOU Qian[1], XU Qian[1], YANG Xin[1], ZHANG Tao[2], CHEN Ming[3]
(1. School of electronics and information, Northwestern Polytechnical University, Xi'an 710072, China;
2. China Academy of Launch Vehicle Technology, Beijing 100076, China;
3. Hangzhou Hikvision Digital Technology Co., Ltd., Hangzhou 310051, China)

DOI: 10.12142/ZTECOM.202302009
https://kns.cnki.net/kcms/detail/34.1294.TN.20230505.1523.002.html, published online May 6, 2023
Manuscript received: 2023-02-21

Abstract: Device-to-device (D2D) communications underlying cellular networks enabled by unmanned aerial vehicles (UAV) have been regarded as promising techniques for next-generation communications. To mitigate the strong interference caused by the line-of-sight (LoS) air-to-ground channels, we deploy a reconfigurable intelligent surface (RIS) to rebuild the wireless channels. A joint optimization problem of the transmit power of UAV, the transmit power of D2D users and the RIS phase configuration are investigated to maximize the achievable rate of D2D users while satisfying the quality of service (QoS) requirement of cellular users. Due to the high channel dynamics and the coupling among cellular users, the RIS, and the D2D users, it is challenging to find a proper solution. Thus, a RIS softmax deep double deterministic (RIS-SD3) policy gradient method is proposed, which can smooth the optimization space as well as reduce the number of local optimizations. Specifically, the SD3 algorithm maximizes the reward of the agent by training the agent to maximize the value function after the softmax operator is introduced. Simulation results show that the proposed RIS-SD3 algorithm can significantly improve the rate of the D2D users while controlling the interference to the cellular user. Moreover, the proposed RIS-SD3 algorithm has better robustness than the twin delayed deep deterministic (TD3) policy gradient algorithm in a dynamic environment.

Keywords: device-to-device communications; reconfigurable intelligent surface; deep reinforcement learning; softmax deep double deterministic policy gradient

Citation (Format 1): YOU Q, XU Q, YANG X, et al. RIS-assisted UAV-D2D communications exploiting deep reinforcement learning [J]. ZTE Communications, 2023, 21(2): 61 - 69. DOI: 10.12142/ZTECOM.202302009
Citation (Format 2): Q. You, Q. Xu, X. Yang, et al., "RIS-assisted UAV-D2D communications exploiting deep reinforcement learning," *ZTE Communications*, vol. 21, no. 2, pp. 61 - 69., Jun. 2023. doi: 10.12142/ZTECOM.202302009.

1 Introduction

Current communication systems and applications are pursuing higher and higher transmission rates, which brings greater challenges to the scarce spectrum resources. Thus, spectrum-efficient communications become increasingly important, which promotes the development of the next-generation cellular networks. Among the various spectrum-efficient techniques, the device-to-device (D2D) communication underlying the cellular network has been considered a promising technique for boosting the communication rates between two neighbor nodes, since it allows the two users to transmit signals directly without passing through a base station (BS)[1]. To maximize the performance of the D2D and cellular network, the location of the BS usually needs to be optimized, which is difficult to realize for the traditional terrestrial cellular network. Fortunately, unmanned aerial vehicles (UAVs) have played a critical role in 6G networks due to their flexibility. For instance, UAVs can work as the aerial BS to improve the network capacity and expand the coverage area, and thus help overcome the limitations of the terrestrial wireless communication at the physical layer[2].

With the UAV aerial BS, the dominant links are usually line-of-sight (LoS) links that benefit the intended receivers while causes strong interference to the unintended users. In this case, reconfigurable intelligent surface (RIS) can be employed to reconstruct the transmission environment and thus reach a compromise between the performances of the intended and other users[3-5]. RIS consists of many low-cost passive reflection elements, where each element can adaptively adjust its reflection amplitude and/or phase to control the intensity and the direction of the electromagnetic wave. In this way, RIS can enhance and/or weaken the strength of the reflected signal for different users[3]. For the D2D communication system with plenty of low-power terminal devices,

This work is supported by the National Natural Science Foundation of China under Grant Nos. 62201462 and 62271412.

RIS can be deployed to improve the quality of the communication links for cellular user equipment (CUE) and mitigate the co-channel interference between CUE and D2D users[6]. There have been some studies on RIS-assisted D2D communications. For example, the authors in Ref. [7] proposed a relaxation-based algorithm called Riemannian manifold based alternating direction multiplier (RM-ADMM) to optimize the system configuration, which is a quadratic constraint quadratic optimization (QCQP) problem. This kind of proposal adopts traditional optimization methods, which may converge to the local optima and cause system performance loss.

Recently, artificial intelligence (AI) has been regarded as a powerful tool to solve complicated non-linear optimization problems. In Ref. [8], a deep learning-based method is proposed for the effective online configuration of the smart surface, where the proposed deep neural network (DNN) model maps the target user's information and the optimal phase matrix to maximize the user's received signal strength by calculating the measurement coordinates. It is worth noting that the deep learning method requires large-scale data sets, which is impractical for some applications. To overcome the limitations of deep learning, deep reinforcement learning (DRL), which combines deep learning and reinforcement learning, has been widely used in wireless communication systems. In Ref. [9], the non-convex optimization problem consisting of beamforming design, power control, and interference coordination is jointly optimized by DRL. In Ref. [10], the authors investigated the simultaneous wireless information and power transfer network where the UAV and the RIS are deployed. By exploiting the DRL to optimize the RIS passive beamforming, the total harvested energy is maximized while meeting the quality of service (QoS) requirements for communications. Ref. [11] is a very early attempt to develop a framework for integrating DRL techniques into optimization designs with no need to understand explicit models or specific mathematical formulas of the wireless environment to solve large-dimensional optimization problems.

At present, the commonly used algorithms for processing continuous action space in DRL are deep deterministic policy gradient (DDPG) and its improved version, the twin delayed deep deterministic (TD3) policy gradient. But the introduction of the underestimation bias by the TD3 algorithm will affect the performance. Studies have shown that softmax's smoothing effect can help learn and reduce the number of local optima[12]. Thus, the authors in Ref. [13] proposed a softmax deep double deterministic (SD3) policy gradients algorithm. The analyses show that the error between the value function and the optimal value under the softmax operator is bounded.

To overcome the complex problem of traditional algorithm calculation, we exploit the SD3 algorithm to jointly design the transmit power of the UAV, the transmit power of the D2D users, and the RIS phase configuration. The main contributions of this paper are summarized as follows:

1) Firstly, we formulate a RIS-assisted UAV-D2D communication system model. In our considered system, the UAV is used as an aerial BS to overcome the limitations of conventional terrestrial BSs. Besides, to investigate the impact of the time-varying channels on the system performance, the motion state of the UAV moving from the CUE to the D2D users is taken into consideration.

2) Secondly, we propose a RIS-SD3 algorithm to solve the complex optimization problem involved in the RIS-assisted UAV-D2D communication system. Unlike the TD3 algorithm, SD3 merges the softmax operator into the action key of continuous control, which makes the optimization environment smoother and thus is conducive to empirical learning.

3) Finally, unlike previous studies that exploit alternating methods to optimize the transmit power and the RIS phase, the proposed algorithm optimizes the transmit power and the phase of the RIS simultaneously. To be more specific, the sum rate of the D2D users is adopted as an immediate reward for training the RIS-SD3 algorithm. The sum rate is gradually maximized by iteratively adjusting the parameters of the RIS-SD3 according to the reward.

The remainder of this paper is organized as follows. The system model is described in Section 2. In Section 3, the RIS-SD3 algorithm is introduced to optimize the phase shift and the transmit power. In Section 4, simulation results are presented to evaluate the performance of the proposed algorithm. The conclusions are given in Section 5.

2 System Model

We consider a practical RIS-assisted UAV-D2D communication network. For example, in a dense urban environment with tall buildings, the primary user, like CUE, is close to the RIS, while the D2D user is located at the edge of the cell. The detailed system description is as follows.

2.1 System Descriptions

The system model is depicted in Fig. 1. We consider a downlink cellular transmission assisted by UAV and RIS. The system consists of one UAV serving as the BS, one RIS, K CUE, and D D2D pairs. To simplify the following analysis, only one CUE is considered in this paper, and a scenario with multiple CUE will be studied in future work. The BS, CUE, D2D transmitter (DT), and the associated D2D receiver (DR) are all single antenna devices. Besides, the RIS is equipped with M reflecting elements and the reflection coefficient matrix $\boldsymbol{\Theta}$ can be described as $\boldsymbol{\Theta} = \mathrm{diag}\left(\beta_1 e^{j\theta_1}, \beta_2 e^{j\theta_2}, \cdots, \beta_M e^{j\theta_M}\right)$.

The CUE receives the desired signals including the signals sent by the BS and the signals reflected from the RIS. In addition, it will receive the interference signals from all the D2D pairs. Therefore, the signal received by the CUE can be

written as:

$$y_c = \left(\boldsymbol{h}_{r,c}^H \boldsymbol{\Theta} \boldsymbol{h}_r + h_c\right)\sqrt{P_c}\, s + \sum_{d=1}^{D} h_{d,c}\sqrt{P_d}\,\mu_d + n_c, \tag{1}$$

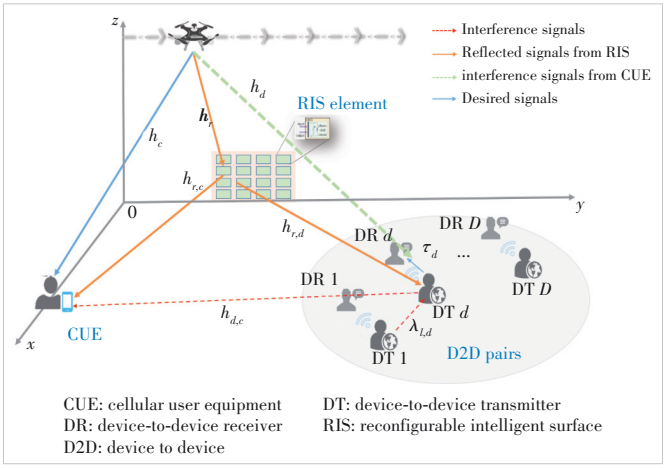
where $\boldsymbol{h}_r \in \mathbb{C}^{M\times 1}$, $\boldsymbol{h}_{r,c} \in \mathbb{C}^{M\times 1}$, $h_c \in \mathbb{C}$, and $h_{d,c} \in \mathbb{C}$ represent the channel gains of UAV-RIS, RIS-CUE, UAV-CUE, and the d-th DT to CUE, respectively; $P_c \in \mathbb{R}$ and $s \in \mathbb{C}$ denote the transmit power and the transmit signal of the BS-CUE link, respectively; $P_d \in \mathbb{R}$ and $\mu_d \in \mathbb{C}$ are the transmit power of the d-th DT and the data transmitted to the d-th DR, respectively; $n_c \sim \mathcal{CN}\left(0,\sigma_c^2\right)$ is the additive Gaussian white noise at the CUE.

The wireless transmission link between the user and the UAV can be either LoS or NLoS. Thus, the received signal power at each user's location is given by Ref. [14].

$$P_r = \begin{cases} P_c d^{-\alpha_0}, & \text{LoS} \\ \eta P_c d^{-\alpha_0}, & \text{NLoS}, \end{cases} \tag{2}$$

where d is the distance between the user and the UAV, α_0 is the path loss exponent over the user-UAV link, and η is an additional factor related to the NLoS link. The LoS probability can be expressed as $P_{\text{LoS}} = \dfrac{1}{1 + A\exp\left(-B(\theta - A)\right)}$, where A and B are constant values that depend on the environment. In this paper, we set $A = 9.6, B = 0.15$, and $\eta = 20$ dB; $\theta = \dfrac{180}{\pi}\sin^{-1}\left(\dfrac{h}{d}\right)$ is the elevation angle where h is the altitude between the user and UAV. The probability of NLoS is $P_{\text{NLoS}} = 1 - P_{\text{LoS}}$[14].

For the terrestrial links, we assume that they follow the Rayleigh distribution where the path-loss is given by $\rho\left(\dfrac{d}{d'}\right)^{-v}$, where ρ, d and v represent the path loss at the reference distance of $d' = 1$, the individual link distance, and the corresponding path loss exponent, respectively.



▲ **Figure 1. System model of a practical RIS-assisted unmanned aerial vehicle (UAV)-D2D communication network**

Note that the m-th element of the diagonal matrix can be written as $\phi_m = \beta_m e^{j\theta_m}$, where $\theta_m \in [0, 2\pi)$ is the phase shift. Generally speaking, phase-shift control achieves better passive beamforming performance than amplitude control, so we assume ideal reflection by the RIS so that the signal power is lossless from each reflection element, e.g., the amplitude reflection coefficient $\beta_m = 1$[15].

The Signal to Interference plus Noise Ratio (SINR) for the received signal of CUE can be calculated as:

$$\text{SINR}_c = \frac{\left|\boldsymbol{h}_{r,c}^H \boldsymbol{\Theta} \boldsymbol{h}_r + h_c\right|^2 P_c}{\sum_{d=1}^{D}\left|h_{d,c}\right|^2 P_d + \sigma_c^2}. \tag{3}$$

Thus, the achievable rate of CUE is:

$$R_c = \log_2\left(1 + \text{SINR}_c\right) = \log_2\left(1 + \frac{\left|\boldsymbol{h}_{r,c}^H \boldsymbol{\Theta} \boldsymbol{h}_r + h_c\right|^2 P_c}{\sum_{d=1}^{D}\left|h_{d,c}\right|^2 P_d + \sigma_c^2}\right). \tag{4}$$

The signals received at the d-th DR consists of the desired signal received from the d-th DT, the interference signal from the UAV, and the reflected signal from the RIS, in addition to the interference signal received from the other D2D pairs. Thus, the signal received at the d-th DR is given by:

$$y_d = \tau_d\sqrt{P_d}\,\mu_d + \left(\boldsymbol{h}_{r,d}^H \boldsymbol{\Theta} \boldsymbol{h}_r + h_d\right)\sqrt{P_c}\, s + \sum_{l\neq d}^{D} \lambda_{l,d}\sqrt{P_l}\,\mu_l + n_d, \tag{5}$$

where $\boldsymbol{h}_{r,d} \in \mathbb{C}^{M\times 1}$, $h_d \in \mathbb{C}$, $\tau_d \in \mathbb{C}$, and $\lambda_{l,d} \in \mathbb{C}$ denote the channel gains of RIS-DR d, UAV-DR d, DT d-DR d, and DT l-DR d, respectively; P_l and μ_l are the transmit power of the l-th DT and the transmit data of D2D to the l-th DR, respectively; $n_d \sim \mathcal{CN}\left(0,\sigma_d^2\right)$ denotes the additive Gaussian white noise at the d-th DR.

Similarly, the received SINR for the d-th DR is given by:

$$\text{SINR}_d = \frac{\left|\tau_d\right|^2 P_d}{\sum_{l\neq d}^{D}\left|\lambda_{l,d}\right|^2 P_l + \left|\boldsymbol{h}_{r,d}^H \boldsymbol{\Theta} \boldsymbol{h}_r + h_d\right|^2 P_c + \sigma_d^2}. \tag{6}$$

The achievable rate of the d-th DR is

$$R_d = \log_2\left(1 + \text{SINR}_d\right) = \log_2\left(1 + \frac{\left|\tau_d\right|^2 P_d}{\sum_{l\neq d}^{D}\left|\lambda_{l,d}\right|^2 P_l + \left|\boldsymbol{h}_{r,d}^H \boldsymbol{\Theta} \boldsymbol{h}_r + h_d\right|^2 P_c + \sigma_d^2}\right). \tag{7}$$

Accordingly, the sum rate of all the D2D pairs is

$$R_{\text{total}} = \sum_{d=1}^{D} R_d. \tag{8}$$

2.2 Problem Formulation

In order to increase the sum rate of the D2D pairs while limiting the amount of interference to the CUE, the problem is formulated as a non-convex optimization problem as follows

$$\max_{\boldsymbol{p},\Theta,p_c} \sum_{D}^{d=1} \log_2\left(1 + \text{SINR}_d\right) \tag{9}$$

$$\text{s.t.} \sum_{D}^{d=1} \left|h_{d,c}\right|^2 P_d \leqslant I_T, \tag{9a}$$

$$0 \leqslant P_d \leqslant P_t, \forall d \in \{1,2,\cdots,D\}, \tag{9b}$$

$$0 \leqslant P_c \leqslant P_{\max}, \tag{9c}$$

$$\text{SINR}_c \geqslant \text{SINR}_{\text{thr}}, \tag{9d}$$

$$R_d \geqslant R_{d\_\text{thr}}, \forall d \in \{1,2,\cdots,D\}, \tag{9e}$$

$$\theta_m \in [0, 2\pi), \forall m \in \{1,2,\cdots,M\}, \tag{9f}$$

$$\left|\phi_m\right| = 1, \forall m \in \{1,2,\cdots,M\}, \tag{9g}$$

where $\boldsymbol{P} = \{P_1, P_2, \cdots, P_D\}$ is the transmit power vector for D2D pairs; P_t is the maximum transmit power of DT and $P_{\max}$ is the maximum transmit power of UAV . I_T in Constraint (9a) indicates the maximum allowable interference to the cellular transmission. Constraints (9b) and (9c) denote the transmit power limit for each DT and the maximum power limit for the UAV BS. Constraints (9d) and (9e) denote the QoS requirements for CUE and D2D pairs. Constraints (9f) and (9g) specify the phase shift and the amplitude constraint of the RIS.

Due to the non-convexity of the objective function and the performance loss of the traditional successive convex approximation (SCA) method[16], we propose a DRL-based framework to solve the non-convex optimization problem.

3 Proposed RIS-SD3 Algorithm

3.1 Description of SD3

SD3 is the abbreviation for the deep double deterministic policy gradient algorithm, which enables a better value estimation by reducing the overestimation bias in DDPG and smoothing the optimized environment, thus contributing to experiential learning[13].

The process of the SD3 algorithm is shown in Fig. 2. SD3 includes an actor network $\mu(\cdot)$ and a critic network $Q(\cdot)$. The actor network consists of two online and two target policy networks with the different parameters $\theta_i^{\mu}, \theta_i^{\mu'} (i = 1,2)$. Similarly, the critic network consists of two online and two target Q-networks with different parameters $\theta_i^{Q}, \theta_i^{Q'} (i = 1,2)$.

According to Fig. 2, we can see that at the time step t, the agent selects an action a_t based on the actor network $\mu\left(s_t; \theta^{\mu}\right)$. Meanwhile, a random noise $N_t \sim \mathcal{N}(0,\sigma)$ is added to interact with the environment to more fully explore the policy. Thus, the action a_t can be written as

$$a_t = \mu\left(s_t; \theta^{\mu}\right) + N_t. \tag{10}$$

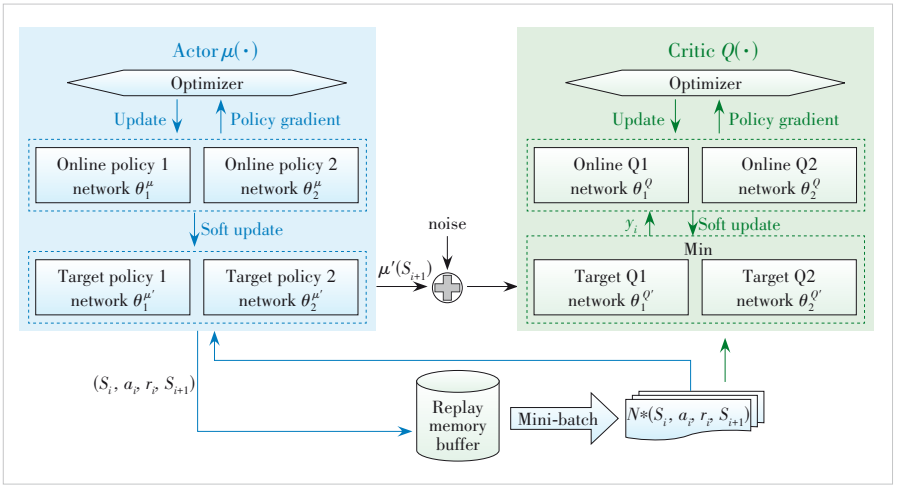
After the agent executes the action, it will return a reward defined as below

$$Q_{t+1}(s,a) = r_t(s,|a) + \gamma \mathbb{E}_{s' \sim p(\cdot|s,a)}\left[V_t(s')\right], \tag{11}$$

where $V_{t+1}(s) = \text{softmax}_{\beta}\left(Q_{t+1}(s,\cdot)\right)$ is the softmax operator, which is used to update the value function $Q_{t+1}(s,a)$ iteratively. Since the softmax operator itself involves integrals and thus is difficult to handle in continuous action space, we use the following unbiased estimation to replace the term $V_t(s')$ as in Ref. [13]:

$$\mathbb{E}_{a' \sim p}\left[\frac{\exp\left(\beta \hat{Q}(s',a')\right)\hat{Q}(s',a')}{p(a')}\right] \Big/ \mathbb{E}_{a' \sim p}\left[\frac{\exp\left(\beta \hat{Q}(s',a')\right)}{p(a')}\right], \tag{12}$$

where $p(a')$ is the probability density function of Gaussian distribution.



▲Figure 2. Workflow of the deep double deterministic (SD3) policy gradient algorithm

For ease of representation, we introduce

$$\hat{Q}_i(s',a') = \min\left(Q_i\left(s',a';\theta_i^{Q'}\right),Q_{-i}\left(s',a';\theta_{-i}^{Q'}\right)\right), \tag{13}$$

and

$$Y_{SD_3}^{-i}(s') = \text{soft}\max_{\beta}\left(\hat{Q}_i(s',\cdot)\right). \tag{14}$$

With Eqs. (13) and (14), the target values for the critic network in Fig. 2 can be estimated as:

$$y_i = r + \gamma Y_{\text{SD}_3}^{-i}(s'), i = 1,2. \tag{15}$$

Then, the critic network optimizes its parameters θ_i^Q by minimizing the loss function given by:

$$L = \frac{1}{N_B}\left(y_i - Q_i\left(s,a;\theta_i^Q\right)\right)^2. \tag{16}$$

After the critic network updates its parameters, the actor network is updated by θ_i^{μ} following the applying the chain rule.

$$\nabla_{\theta_i^{\mu}} = \frac{1}{N_B}\nabla_a Q_i\left(s,a;\theta_i^{Q}\right)\Big|_{s=s_t,a=\mu\left(s_i;\theta_i^{\mu}\right)}\nabla_{\theta^{\mu}}\mu\left(s;\theta_i^{\mu}\right)\Big|_{s=s_t}, \tag{17}$$

To make the learning process more stable, the SD3 also uses a soft target update approach:

$$\begin{aligned}\theta_i^{Q'} &\leftarrow \tau\theta_i^{Q} + (1-\tau)\theta_i^{Q'},\\ \theta_i^{\mu'} &\leftarrow \tau\theta_i^{\mu} + (1-\tau)\theta_i^{\mu'},\end{aligned} \tag{18}$$

where τ is the learning rate for updating the target critic network and the target actor network.

Algorithm 1: Learning algorithm of RIS-SD3

Input: $\boldsymbol{h}_{r,c}, \boldsymbol{h}_r, \boldsymbol{h}_{r,d}, h_c, h_{d,c}, h_d, \tau_d, \lambda_{l,d}$
Output: the optimal action
$a = \left\{p_1^{\text{opt}}, p_2^{\text{opt}}, \theta_1^{\text{opt}}, \theta_2^{\text{opt}}, \cdots, \theta_M^{\text{opt}}, p_c^{\text{opt}}\right\}$

```
Initialize actor networks μ1, μ2 and critic networks Q1, Q2
with random parameters θ1^μ, θ2^μ, θ1^Q, θ2^Q;
Initialize the size of experience replay N_R, the size of mini-batches N_B and replay buffer R;
for t = 1,⋯,T do
    Select an action with exploration noise N_t ~ N(0, σ) based on executing action a, obtained reward r, new state s' and done;
    Store transition tuple (s, a, r, s', done) in R;
    for i = 1, 2 do
        Sample a random minibatch of N from R;
        Sample K noises N_t ~ N(0, σ);
        Set â' = μ_i'(s_{i+1}) + clip(N_t, −c, c);
        Set Q̂(s', â') = min_{j=1,2}(Q_j(s', â'; θ_j^{Q'}));
        Set softmax_β(Q̂(s', â')) as Eq. (12)
        Update critic net via minimizing Eq. (16);
        Update actor net by policy gradient in Eq. (17);
        Update the target networks
        θ_i^{Q'} ← τθ_i^Q + (1 − τ)θ_i^{Q'};
        θ_i^{μ'} ← τθ_i^μ + (1 − τ)θ_i^{μ'}
    end
end
```

3.2 Details of RIS-SD3

In this paper, the environment depends on our proposed system model. At the time step t, the agent can collect the current channel information, and combined with the current state, the agent selects the action and calculates the reward according to the current policy. There are E episodes in the whole training process, and each episode is iterated by T times. The detailed workflow of the proposed RIS-SD3 algorithm is shown in Algorithm 1. The state space, action space and reward function are given as follows.

1) State: The state s_t at the t-th time step is constructed by the received signal of CUE, the UAV's location at the t-th time step, and the SINR of D2D pairs. So the total number of the state is $D+M+K+1$.

2) Action: The action is constructed by the transmit power vector $\boldsymbol{P} = \left\{P_1, P_2, \cdots, P_D\right\}$, the transmit power of BS P_c and the phase $\theta_i (i = 1,2,\cdots,M)$ of RIS. In order to reduce the complexity of the action space, we convert both phase and power into one-dimensional vectors, i. e., action = $\left\{P_1, P_2, \cdots P_D, \theta_1, \theta_2, \cdots \theta_M, P_c\right\}$. So the total number of the action is $D+M+1$.

3) Reward: In the proposed RIS-SD3 algorithm, the sum rate of the D2D pairs is taken as the reward. Furthermore, in order to satisfy the minimum signal-to-noise ratio and the maximum interference requirements for CUE users and the QoS requirements for D2D users, the reward can therefore be set as:

$$R_t = \begin{cases} & \text{if } \sum_{d=1}^{D}\left|h_{d,c}\right|^2 P_d \leqslant I_T \\ R_{\text{total}}, & \text{SINR}_c \geqslant \text{SINR}_{\text{thr}} \\ & R_d \geqslant R_{d\_\text{thr}} \\ 0, & \text{else} \end{cases}. \tag{19}$$

The reward for each episode is:

$$R = \sum_{t=1}^{T} R_t. \tag{20}$$

4 Numerical Results

In order to facilitate the analysis, we consider D=2, and the other parameters used in the algorithm are shown in

Table 1, and the establishment of the coordinate system is shown in Fig. 1. After continuous training tests, we then find the training work the best when the main hyper-parameters in the RIS-SD3 are set as follows: E = 10 000, T = 61, γ = 0.99, and τ = 0.005.

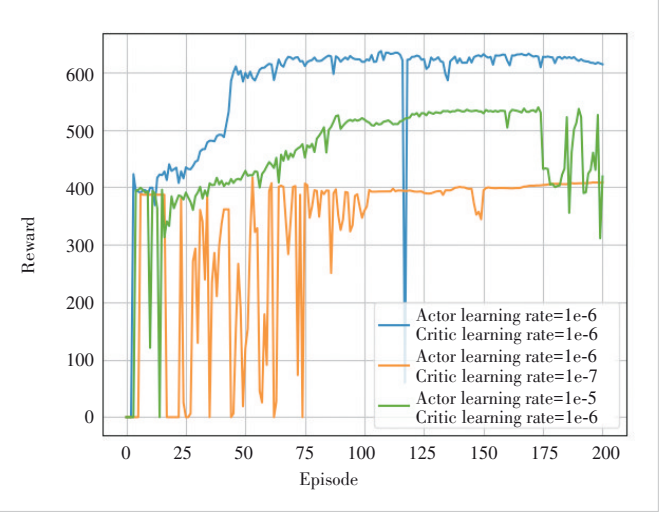
4.1 Impact of Parameters Settings of RIS-SD3

In our proposed RIS-SD3 algorithm, we use a constant learning rate and batch size for all networks to investigate their effects on the performance and convergence speed for the DRL-based approach. Fig. 3 demonstrates the average rewards versus time episodes at different learning rates. It can be seen that different learning rates have a great impact on the performance of the proposed RIS-SD3 algorithm. As shown in Fig. 3, RIS-SD3 with actor and critic learning rates of 1e-6 performs best. Specifically, when the learning rate is too large, the algorithm will be unstable and even cannot converge. On the contrary, when the learning rate is too small, the convergence rate will be slow or even incapable to learn, and thus the training time is wasted.

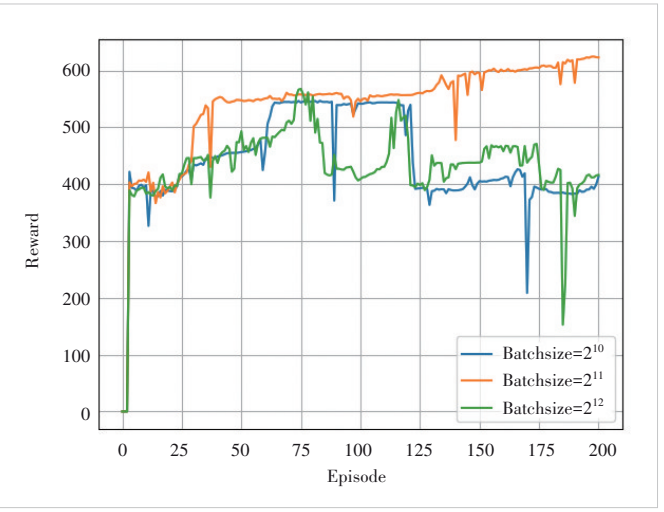
▼Table 1. Parameters of the proposed system

Parameter		Value
Location	UAV	From (0, 0, 1) m to (0, 60, 1) m
	RIS	(0, 10, 2) m
	CUE	(20, 0, 1) m
	DT1	(20, 60, 1) m
	Distance of D2D	5 m
	Size area of D2D	10 m
$SINR_{thr}$	Minimum SINR of CUE	12 dB
R_{d_thr}	Minimum achievable rate of D2D	2 dB
I_T	Maximum interference of CUE	−30 dB
P_{max}	Max transmit power of UAV	30 W
P_t	Max transmit power of DT	10 W, 20 W, 30 W
β	Path loss coefficient	−30 dB
α_0	Path loss exponent over the user-UAV link	3
ν	Path loss exponent	2.5
ρ	The path loss at the reference distance	0.01

CUE: cellular user equipment
D2D: device-to-device
DT: D2D transmitter
RIS: reconfigurable intelligent surface
SINR: Signal to Interference plus Noise Ratio
UAV: unmanned aerial vehicle



▲Figure 3. Effect of the learning rate



▲Figure 4. Effect of batchsize on the training model

Batchsize is the number of data used for each update when using the optimizer. In short, it is how many data we want to put into the model at a time to train. This value is between 1 and the total number of training samples.

As shown in Fig. 4, we explore the impact of batchsize on the training model. If the batchsize is too small, time-consuming and training efficiency is low, the training data will be very difficult to converge, resulting in a state of under-fitting. In a certain range, generally speaking, the larger the batchsize, the more accurate the determined descending direction, and the smaller the training shock. The batchsize increases to a certain extent, and its determined decline direction has basically not changed. Therefore, the larger the batch size is, the more stable the gradient will be, while the smaller the batch size is, the higher the randomness of the gradient will be. However, if the batch size is too large, the the demand for memory will be higher, and it is not conducive to the network jumping out of the local minimum. We can see that batchsize = 2^{11} is the best, so this value is used in the following simulations.

4.2 Comparisons with Benchmarks

To further demonstrate the performance and the time complexity of the proposed RIS-SD3 algorithm, we consider the following baseline schemes. Firstly, we use the exhaustive searching approach to find the approximate optimal value, where the transmit power and the phase are limited to ten equally spaced values. Then, for the weighted minimum

mean-square error (WMMSE)-power" baseline scheme, we use the WMMSE algorithm in Ref. [17] to optimize the transmit power of D2D. For the "max-power and random-phase" baseline scheme, we assume that the RIS configures the phase shifts in a random manner with the maximum D2D transmit power. For the "without RIS" baseline scheme, we assume that the D2D transmit power is random without the deployment of the RIS. Moreover, the TD3 algorithm is also introduced. Unless otherwise specified, the learning rate for the RIS-SD3 algorithm is set as 1e-6.

As shown in Fig. 5, compared with the TD3 algorithm for continuous actions, the proposed RIS-SD3 algorithm is more robust under dynamic channel conditions because it refers to the soft operator in Ref. [13]. Compared with the more accurate exhaustive searching results and the WMMSE algorithm, the proposed RIS-SD3 algorithm can obtain a larger sum rate. In addition, it can be seen from the figure that the proposed algorithm and the exhaustive searching algorithm are more robust to the position change of the UAV. Finally, by comparing the results of the proposed algorithm with the "without RIS" scheme, we improve the system performance by introducing RIS, since the RIS provides the additional degrees of freedom (DoF) to improve the sum rate. In addition, since exhaustive search only considers partially discrete values, its effect is slightly lower than that of the RIS-SD3 algorithm that considers continuous values.

Moreover, it can be seen from Fig. 5 that the sum rate fluctuates as the position of UAVs changes, especially for the TD3 and the WMMSE algorithms. Actually, due to the introduction of RIS, the system performance is not that sensitive to the position of the UAV. The up and down phenomena indicate that the performance of the TD3 algorithm is poor for the considered scenario, which motivates us to propose the RIS-SD3 algorithm. As for the WMMSE algorithm, the reason for the fluctuation is that this algorithm only optimizes the D2D user's transmit power, while the phase is random.

To evaluate the time complexity of the proposed method, the time consumption of the proposed scheme and the baseline schemes are shown in Table 2, where the device we use is NVIDIA GPU RTX 3090. It can be observed that the time consumption of the proposed algorithm is less than most of the baselines, but a little bit more than the TD3 algorithm.

However, it should be noted that the TD3 cannot adapt to the change of the UAV location, as observed in Fig. 5.

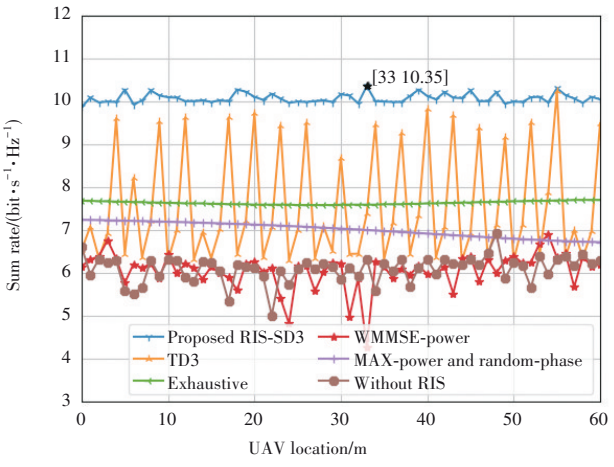
4.3 Impact of Parameter Settings on System

To get a better understanding of the RIS-SD3 method, we investigate the impact of the max power of DT. When more transmitting power is allocated to D2D users, the proposed RIS-SD3 algorithm can obtain a higher sum rate. This observation is consistent with the results in the traditional multi-input single-output (MISO) system. Through the joint design of transmit beamforming and phase shift, the common channel interference of multi-user MISO systems can be effectively reduced, thereby improving performance.

▼Table 2. Time consumption comparison

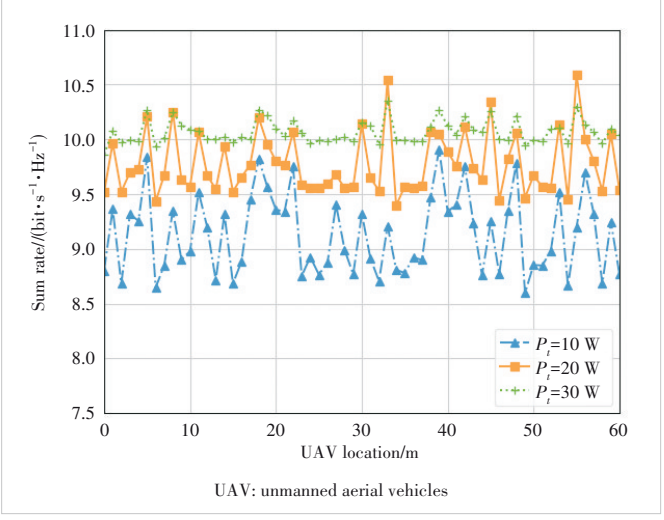
Scheme	Time Consumption/s
Proposed RIS-SD3	1.74
TD3	1.08
Exhaustive	3.79e+05
WMMSE-power	1.68
Max-power and random-phase	3.67
Without RIS	3.98

RIS: Reconfigurable intelligent surface
SD3: softmax deep double deterministic
TD3: delayed deep deterministic
WMMSE: weighted minimum mean-square error



RIS: reconfigurable intelligent surface
SD3: deep double deterministic
TD3: delayed deep deterministic
WMMSE: weighted minimum mean-square error
UAV: unmanned aerial vehicles

▲Figure 5. RIS-SD3 in comparison with other baseline schemes



UAV: unmanned aerial vehicles

▲Figure 6. Sum Rate under different P_t

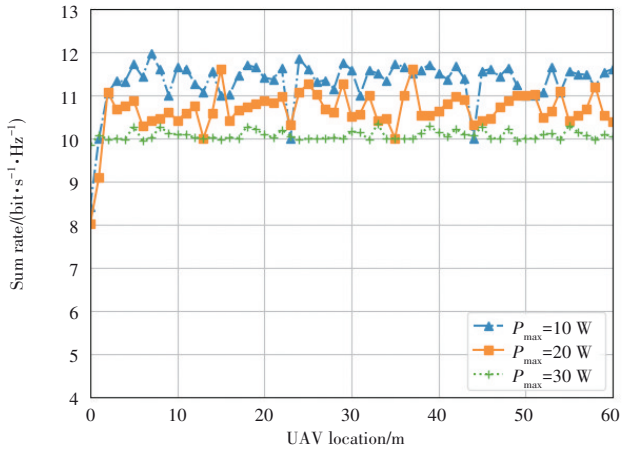
As can be seen from Fig. 6, during the movement of the UAV from (0, 0, 30) m to (0, 60, 30) m, there are also certain fluctuations in the system about the sum rate. In fact, the UAV only changes continuously by 1 m, while the receiver may not have time to feel this change, and then the UAV moves to the next position. Therefore, it can be seen from Fig. 6 that the undulating change is random.

In addition, we simulate the effect of the maximum transmitting power of UAV $P_{\max}$ on the D2D sum rate. It can be seen from Fig. 7 that, as the maximum transmitting power of the UAV increases, the D2D sum rate decreases. This is because with the transmit power of the UAV increases, the interference of the cellular user to the D2D user increases, so the D2D sum rate decreases.

It can also be seen from Fig. 7 that in the process of the drone moving from (0,0,30) m to (0,60,30) m, the sum rate has certain ups and downs. The specific reason may be that the UAV changes less, so the fluctuations are more random, but the overall change is not very significant.

5 Conclusions

Based on the latest progress in DRL for continuous action space, a RIS-SD3 optimization algorithm is proposed to solve the joint power allocation and phase optimization problem in a dynamic RIS-assisted UAV-D2D communication network. With the RIS-SD3 algorithm, the sum rate of the D2D users is maximized while meeting the QoS requirement for the cellular user. Specifically, by introducing softmax operators, the proposed algorithm learns about the environment more efficiently, and thus has better robustness to the change of the environment. Simulation results show that the proposed RIS-SD3 method can learn from the environment by observing the instantaneous reward got from the time-varying wireless channels, and then gradually improves its behavior to the optimal result. Compared with the baseline schemes, the proposed scheme can increase the sum rate as well as improve the robustness of the transmission environment.



UAV: unmanned aerial vehicles

▲**Figure 7. Sum rate under different $P_{\max}$**

References

[1] DANG S P, CHEN G J, COON J P. Multicarrier relay selection for full-duplex relay-assisted OFDM D2D systems [J]. IEEE transactions on vehicular technology, 2018, 67(8): 7204 – 7218. DOI: 10.1109/TVT.2018.2829401
[2] WU Q 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. DOI: 10.1109/TWC.2017.2789293
[3] WU Q Q, ZHANG R. Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network [J]. IEEE communications magazine, 2020, 58(1): 106 – 112. DOI: 10.1109/MCOM.001.1900107
[4] WU Q Q, ZHANG R. Intelligent reflecting surface enhanced wireless network: Joint active and passive beamforming design [C]//IEEE Global Communications Conference (GLOBECOM). IEEE, 2018: 1 – 6. DOI: 10.1109/GLOCOM.2018.8647620
[5] HUANG C W, ZAPPONE A, ALEXANDROPOULOS G C, et al. Reconfigurable intelligent surfaces for energy efficiency in wireless communication [J]. IEEE transactions on wireless communications, 2019, 18(8): 4157 – 4170. DOI: 10.1109/TWC.2019.2922609
[6] YANG G, LIAO Y, LIANG Y C, et al. Reconfigurable intelligent surface empowered device-to-device communication underlaying cellular networks [J]. IEEE transactions on communications, 2021, 69(11): 7790 – 7805. doi: 10.1109/TCOMM.2021.3102640
[7] CAO Y S, LV T J, NI W, et al. Sum-rate maximization for multi-reconfigurable intelligent surface-assisted device-to-device communications [EB/OL]. [2023-01-20]. https://arxiv.org/abs/2108.07091
[8] HUANG C W, ALEXANDROPOULOS G C, YUEN C, et al. Indoor signal focusing with deep learning designed reconfigurable intelligent surfaces [C]// IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC). IEEE, 2019: 1 – 5. DOI: 10.1109/SPAWC.2019.8815412
[9] SHAFIN R, CHEN H, NAM Y H, et al. Self-tuning sectorization: Deep reinforcement learning meets broadcast beam optimization [J]. IEEE transactions on wireless communications, 2020, 19(6): 4038 – 4053. DOI: 10.1109/TWC.2020.2979446
[10] PENG H R, WANG L C, YE L G, et al. Long-lasting UAV-aided RIS communications based on SWIPT [C]//2022 IEEE Wireless Communications and Networking Conference (WCNC). IEEE, 2022: 1844 – 1849. DOI: 10.1109/WCNC51071.2022.9771999
[11] HUANG C W, MO R H, YUEN C. Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning [J]. IEEE Journal on selected areas in communications, 2020, 38(8): 1839 – 1850. DOI:10.1109/JSAC.2020.3000835
[12] CESA-BIANCHI N, GENTILE C, LUGOSI G, et al. Boltzmann exploration done right [C]//The 31st International Conference on Neural Information Processing Systems. NIPS, 2017: 6287 – 6296
[13] PAN L, CAI Q P, HUANG L B. Softmax deep double deterministic policy gradients [C]//The 34th International Conference on Neural Information Processing Systems. NIPS, 2020: 11767 – 11777
[14] AL-HOURANI A, KANDEEPAN S, LARDNER S. Optimal LAP altitude for maximum coverage [J]. IEEE wireless communications letters, 2014, 3(6): 569 – 572. doi: 10.1109/LWC.2014.2342736
[15] WU Q Q, ZHANG S W, ZHENG B X, et al. Intelligent reflecting surface-aided wireless communications: A tutorial [J]. IEEE transactions on communications, 2021, 69(5): 3313 – 3351. DOI: 10.1109/TCOMM.2021.3051897
[16] WANG W H, YANG L, MENG A Q, et al. Resource allocation for IRS-aided JP-CoMP downlink cellular networks with underlaying D2D communications [J]. IEEE transactions on wireless communications, 2022, 21(6): 4295 – 4309. DOI: 10.1109/TWC.2021.3128711
[17] SHI Q J, RAZAVIYAYN M, LUO Z Q, et al. An iteratively weighted MMSE

approach to distributed sum-utility maximization for a MIMO interfering broadcast channel [C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2011: 3060 – 3063. DOI: 10.1109/ICASSP.2011.5946304

Biographies

YOU Qian received her BS degree from Yangzhou University, China in 2021. She is currently pursuing her MS degree at the school of electronics and information, Northwestern Polytechnical University, China. Her research interests include machine learning for communications, IRS-assisted communications, and UAV-assisted Communications.

XU Qian (qianxu@nwpu.edu.cn) received her BS and PhD degrees both from Xi'an Jiaotong University, China. She is currently an associate professor at the school of electronics and information, Northwestern Polytechnical University, China. Her current research interests include mobile wireless communications with emphasis on physical layer security and QoS provisioning. She has published more than 20 technical papers.

YANG Xin received his BS degree in communication engineering and MS degree in electronics and communication engineering from Xidian University, China in 2011 and 2014, respectively, and PhD degree in information and communication engineering from Northwestern Polytechnical University, China in 2018. He is an associate professor with school of electronics and information, Northwestern Polytechnical University. His research interests include wireless communications and ad hoc networks.

ZHANG Tao received his MS degree from Nankai University, China. He is currently a senior engineer in China Academy of Launch Vehicle Technology. His current research interests mainly focus on wireless communications.

CHEN Ming received his BS degree from Xidian University, China. He is currently a senior engineer in Hangzhou Hikvision Digital Technology Co., Ltd. His current research interests mainly focus on intelligent signal processing.