Anti-Intelligent UAV Jamming Strategy via Deep Q-Networks

Ning Gao, *Member, IEEE*, Zhijin Qin, *Member, IEEE*, Xiaojun Jing, *Member, IEEE*, Qiang Ni, *Senior Member, IEEE*, and Shi Jin, *Senior Member, IEEE*

***Abstract*— The downlink communications are vulnerable to intelligent unmanned aerial vehicle (UAV) jamming attack. In this paper, we propose a novel anti-intelligent UAV jamming strategy, in which the ground users can learn the optimal trajectory to elude such jamming. The problem is formulated as a stackelberg dynamic game, where the UAV jammer acts as a leader and the ground users act as followers. First, as the UAV jammer is only aware of the incomplete channel state information (CSI) of the ground users, for the first attempt, we model such leader sub-game as a partially observable Markov decision process (POMDP). Then, we obtain the optimal jamming trajectory via the developed deep recurrent Q-networks (DRQN) in the three-dimension space. Next, for the followers sub-game, we use the Markov decision process (MDP) to model it. Then we obtain the optimal communication trajectory via the developed deep Q-networks (DQN) in the two-dimension space. We prove the existence of the stackelberg equilibrium and derive the closed-form expression for the stackelberg equilibrium in a special case. Moreover, some insightful remarks are obtained and the time complexity of the proposed defense strategy is analyzed. The simulations show that the proposed defense strategy outperforms the benchmark strategies.**

***Index Terms*— UAV, jamming, Markov decision process, deep Q-networks.**

I. Introduction

WITH the urgent demands of high-speed data transmission in wireless communications, various technologies have been explored to improve the network capacity, i.e., massive multiple-input multiple-output (massive-MIMO) and millimeter wave (mmWave) communication. Recently, the unmanned aerial vehicle (UAV) has been adopted to improve the network capacity. For example, compared to the ground communications, UAV can provide strong line-of-sight (LoS) links and small path-loss exponent to the ground users when it is used as the base station. Therefore, by optimizing the UAV trajectory and transmission strategies, the UAVs can be used to boost the network capacity [1]–[4].

When considering the security issues in wireless communication systems, UAVs can be exploited as different components [5]–[13]. As security components, UAVs can be used by the legitimate users. For example, since the friendly jammer can protect the confidential messages by transmitting the artificial noise [14], [15], UAV has been utilized as a friendly jammer to protect the ground users away from the eavesdropper. Specifically, with the assist of an air-to-ground-friendly UAV jammer, the system security can be improved when the location of the eavesdropper is unknown [6]. Then, UAVs can work as relays to forward the message to improve the communication quality [10]. In [11], a reinforcement learning based UAV relay has been studied to against the smart jamming in vehicular ad hoc networks. Additionally, some work has attempted to combine UAV relay and UAV friendly jammer to enhance communication security. For example, a dual-UAV enabled secure communication system has been investigated in [7], in which one UAV can work as a relay to communicate with multiple ground users and another UAV can work as a friendly jammer to jam the ground eavesdropper. As malicious components, UAVs can be exploited by the illegitimate users [12], [13]. The authors in [8] have shown that malicious UAVs equipped with cameras and multi-spectral sensors can eavesdrop the privacy of legitimate users. Due to the LoS links and small path-loss exponent, UAV jamming can significantly block the data transmission and degrade communication quality of service (QoS), which is more serious than ground jamming. Therefore, anti-UAV jamming problem is worth investigating.

Some meaningful work has been developed to address the malicious UAV jamming problem [16]–[19]. Particularly, a zero-sum pursuit-evasion game has been formulated to compute optimal strategies, which aims to evade the attack of an UAV jammer [16]. A smart UAV attacker, who can specify the attack type, such as jamming, eavesdropping, and spoofing, has been considered in [17] and the reinforcement learning based power allocation strategies have been proposed

Manuscript received April 12, 2019; revised August 13, 2019 and September 22, 2019; accepted October 7, 2019. This work was supported in part by the National Key Research and Development Program 2018YFA0701602, in part by the National Natural Science Foundation of China (NSFC) for Distinguished Young Scholars of China under Grant 61625106, and in part by the Royal Society project IEC170324. This article was presented in part at the IEEE International Conference on Communications (ICC), Shanghai, China, May 20-24, 2019. The associate editor coordinating the review of this article and approving it for publication was M. J. Hossain. *(Corresponding author: Shi Jin.)*

N. Gao and S. Jin are with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China (e-mail: ninggao@seu.edu.cn; jinshi@seu.edu.cn).

Z. Qin is with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, U.K. (email: z.qin@qmul.ac.uk).

X. Jing is with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: jxiaojun@bupt.edu.cn).

Q. Ni is with the School of Computing and Communications, Lancaster University, Lancaster LA1 4WA, U.K. (e-mail: q.ni@lancaster.ac.uk).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCOMM.2019.2947918

to defend against such attack. However, the aforementioned anti-UAV jamming work are based on some ideal assumptions, i.e., the perfect observation. More recent work has considered imperfect observation in anti-ground jamming but few in anti-UAV jamming [18]–[24]. For example, with considering the co-channel mutual interference and the incomplete information, i.e., incomplete channel state information (CSI), the competition between UAV users and jammers have been investigated by using a Bayesian stackelberg game [18]. The authors in [19] have designed a secure communication system to deal with the joint impact of UAV smart attack and imperfect channel estimation. The authors in [20] has formulated the jamming game with incomplete information, i.e., the other user's identities, as a Bayesian game and discussed the performance of this game. The prospect theoretic analysis has been used to model anti-jamming communications [21]. Moreover, a Bayesian stackelberg game with incomplete information has been formulated to analyze the jammer in [22], [23]. Likewise, the impact of observation error of a smart jammer has been evaluated in a stackelberg anti-jamming game and the Nash equilibrium has been derived [24]. As aforementioned, only [18], [19] have considered imperfect observations in anti-UAV jamming problem. Meanwhile, only [19] has considered an intelligent UAV attacker with imperfect observations. In other words, limited work has considered intelligent UAV jamming, which can easily learn the optimal attack strategy in complex communication environments, even with imperfect observation, i.e., incomplete CSI.

With the rapid development of artificial intelligence (AI) in communications [25], [26], such an intelligent UAV jamming becomes more reality and more harmful than we have ever considered. One powerful tool is reinforcement learning, by which the intelligent agent can choose jamming action based on the environments and maximize the reward. This reward is called long-term cumulative reward, which is decided by a series of time events. The Q-learning is a model-free reinforcement learning method, which can learn the optimal strategy based on the long-term cumulative reward with an end-to-end approach. Then, to address the curse of high dimensionality in Q-learning, the Deep Q-network (DQN) has been developed by Google DeepMind, which combines Q-learning with convolutional neural network (CNN). It can be used to learn the optimal strategy in a large state space [27]. Whereas, the DQN cannot perform well with the imperfect observations. Then, to learn the optimal strategy with the imperfect observation, the deep recurrent Q-network (DRQN) has been introduced, which is a combination of a long short term memory (LSTM) and a DQN [28]. With AI, some incredible jamming attacks have been realizing, i.e., [17], [29], which makes the anti-UAV jamming problem more challenging.

In this paper, we consider the scenario that both the UAV jammer and the ground users are intelligent agents. On the one hand, the UAV jammer can learn the optimal jamming trajectory via the imperfect observation. On the other hand, the ground users can learn the optimal communication trajectory to elude the UAV jamming. To the best of our knowledge, *"How do ground users defend against intelligent UAV jamming attack using AI?"* is still an open problem. The specific contributions of our work are summarized as follows:

- For the first time, we consider the scenario that both the UAV jammer and the ground users are intelligent agents, in which an UAV jammer can block the data transmission of the ground users and the ground users are capable of defending against the intelligent UAV jamming to the greatest extent.
- For the ground users, we propose a novel anti-intelligent UAV jamming strategy, in which the optimal trajectory of each ground user is obtained. Specifically, the anti-intelligent UAV jamming problem is formulated as a stackelberg dynamic game. The incomplete CSI is considered in the game and the optimal trajectories are learned via DRQN and DQN, respectively.
- Some insightful remarks are obtained from the theory and the simulations: i) we prove that the optimal trajectory of each ground user exists; ii) we prove the existence of the stackelberg equilibrium in the game; iii) to maximize long-term cumulative reward, the action choices of UAV jammer is different from that of maximizing the immediate reward.

The rest of the paper is organized as follows. In Section II, we present the system model and the problem formulation. In Section III, we propose the anti-intelligent UAV jamming strategy and the corresponding discussions. Simulations are presented in Section IV and conclusions are given in Section V.

II. System Model and Problem Formulation

In this section, we first give the system model, then, we formulate the optimization problem. For ease of reference, important symbols are summarized in Table I.

A. System Model

We consider the downlink transmissions between a base station and ground users under the threat of a UAV jammer, which is shown in Fig. 1. In the following, if no confusions occur, the users refer to the ground users. Denote $\mathcal{J}$ as the UAV jammer, $\mathcal{B}$ as the base station and $i \in \{1, \cdots, U\}$ as user i. We assume that the location of the base station is fixed with height $H_{\mathcal{B}}$, while the users and the UAV jammer are mobile at constant velocities in each time slot. Considering the resource-limited devices, all of them are equipped with single antenna and communicate with the base station by adopting frequency division multiple access (FDMA). The total bandwidth is B Hz, and we consider the worst case that the UAV performs barrage jamming, which can jam the full bandwidth of the network [30]. The UAV jammer and the users are considered as intelligent agents, who can learn the optimal actions to maximize their long-term cumulative rewards, i.e., signal-to-interference-plus-noise ratio (SINR) [31], respectively. The locations of base station $\mathcal{B}$, an arbitrary user i, and the UAV jammer $\mathcal{J}$ are denoted as $(0, 0, H_{\mathcal{B}})$, $(x_i, y_i, 0)$, and

TABLE I
SUMMARY OF SYMBOLS

Symbols	Notations
$\mathcal{B}$	Base station
$\mathcal{J}$	UAV jammer
i	User i
$\mathcal{A}_{\mathcal{J}}$	Action space of UAV jammer
$\mathcal{A}_i$	Action space of user i
β_{LoS}	Additional attenuation factor of LoS link
β_{NLoS}	Additional attenuation factor of NLoS link
$I_{\mathcal{J}i}$	Expectation of the jamming power received at user i
Γ_i	Received SINR at user i
$R_{\mathcal{J}}$	Long-term cumulative reward of UAV jammer
R_i	Long-term cumulative reward of user i
r_i	Immediate reward of user i
$r_{\mathcal{J}}$	Immediate reward of UAV jammer
γ	Discount factor
$\mathcal{S}$	Channel state space
$\mathcal{S}_i$	Motion state space of user i
$\mathcal{S}_{\mathcal{J}}$	Flight state space of UAV jammer
$\mathcal{O}$	Observation state space
$\mathcal{M}$	belief state space
$P(\cdot\|\cdot)$	Probability of transition
$\Omega(\cdot\|\cdot)$	Probability of possible observation
b	Belief
$\mathbb{O}$	Sequence of ℓ historical observation-action pairs
$\mathbb{S}$	Sequence of ℓ historical state-action pairs
θ	Weight parameter set of the Q-network of UAV jammer
ξ	Weight parameter set of the Q-network of user
ϵ	Probability that the agent chooses the non-optimal action
$\mathscr{T}^*(a_{\mathcal{J}})$	Optimal jamming trajectory of UAV jammer
$\mathscr{L}^*(a_V)$	Optimal communication trajectory of virtual user
$O(\cdot)$	Time complexity function

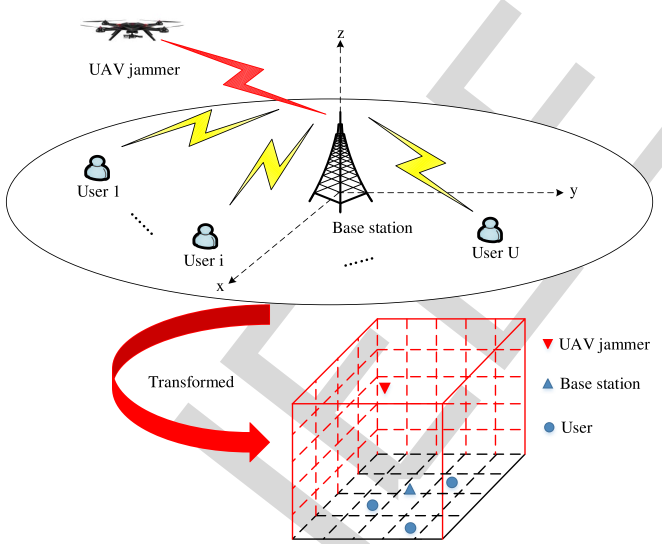


Fig. 1. Schematic diagram. The network includes one base station, U ground users and a UAV jammer, then the network is transformed into a solid figure. The UAV jammer can fly in a three-dimension space, the ground users can move in a two-dimension space, moreover, the base station is in a three-dimension space and deployed at the center of the "x0y" plane.

$(x_{\mathcal{J}}, y_{\mathcal{J}}, z_{\mathcal{J}})$, respectively. Denote the mapping of UAV jammer action space as

$$\mathcal{A}_{\mathcal{J}} = \{(0,0,0),(0,0,1),(0,0,-1),(-1,0,0),(1,0,0),\\ (0,1,0),(0,-1,0)\},$$

which represents moving directions including stay, up, down, left, right, forward, backword. Likewise, we map the user action space as

$$\mathcal{A}_i = \{(0,0,0),(-1,0,0),(1,0,0),(0,1,0),(0,-1,0)\},$$

which represents flight directions including stay, left, right, forward, backword. In time slot t, the UAV jammer $\mathcal{J}$ chooses an action $a_{\mathcal{J}}^t \in \mathcal{A}_{\mathcal{J}}$ to determine the flight direction, and user i chooses an action $a_i^t \in \mathcal{A}_i$ to determine its moving direction.

The channel coefficient from base station $\mathcal{B}$ to user i is denoted as $h_{\mathcal{B}i} = \sqrt{d_{\mathcal{B}i}^{-\eta}}\tilde{h}_{\mathcal{B}i}$, where $d_{\mathcal{B}i}$ represents the distance between base station $\mathcal{B}$ and user i, η is the path loss exponent and $\tilde{h}_{\mathcal{B}i}$ is the small-scale fading, which follows zero-mean complex Gaussian distribution with unit variance. In addition, the communication channel between UAV jammer and user i is modeled as an air-to-ground channel, which contains three parts, including strong LoS, reflected nonline-of-sight (NLoS), and small-scale fading. In general, the influence of small-scale fading is smaller than LoS and NLoS, therefore, the small-scale fading is neglected [32], [33]. The path loss of the air-to-ground channel between UAV jammer and user i is denoted as [34]

$$\text{PL}(\mathcal{J}, i) = \begin{cases} \beta_{\text{LoS}}|d_{\mathcal{J}i}|^{-\alpha}, & \text{for LoS link}, \\ \beta_{\text{NLoS}}|d_{\mathcal{J}i}|^{-\alpha}, & \text{for NLoS link}, \end{cases} \tag{1}$$

where $d_{\mathcal{J}i} = \sqrt{(x_i - x_{\mathcal{J}})^2 + (y_i - y_{\mathcal{J}})^2 + z_{\mathcal{J}}^2}$ is the distance between UAV jammer $\mathcal{J}$ and user i, α is the path-loss exponent for the air-to-ground channel, and β_{LoS} and β_{NLoS} are additional attenuation factors for LoS link and NLoS link, respectively. The probability of LoS connection, P_{LoS}, depends on the elevation angle θ_i between user i and UAV, the communication environment, the surrounding buildings density, and the height of the UAV jammer, $H_{\mathcal{J}}$, which can be represented as

$$P_{\text{LoS}} = \frac{1}{1 + \Phi \exp(-\Psi[\theta_i - \Phi])}. \tag{2}$$

In particular, Φ and Ψ are S-curve parameters, which depend on communication environment, i.e., $\Phi = 150$ and $\Psi = 15$ are the common settings for urban areas, the angle is

$$\theta_i = \frac{180}{\pi}\arcsin\left(\frac{z_{\mathcal{J}}}{d_{\mathcal{J}i}}\right)$$

and the probability of NLoS is $P_{\text{NLoS}} = 1 - P_{\text{LoS}}$. Hence, the expectation of the jamming power received at the user i is given by [32]

$$I_{\mathcal{J}i} = p_{\mathcal{J}} P_{\text{LoS}} \beta_{\text{LoS}} |d_{\mathcal{J}i}|^{-\alpha} + p_{\mathcal{J}} P_{\text{NLoS}} \beta_{\text{NLoS}} |d_{\mathcal{J}i}|^{-\alpha}, \tag{3}$$

where $p_{\mathcal{J}}$ is the power budget of the UAV jammer. Then, the received SINR at user i can be denoted as

$$\Gamma_i = \frac{p_{\mathcal{B}} d_{\mathcal{B}i}^{-\eta} |\tilde{h}_{\mathcal{B}i}|^2}{I_{\mathcal{J}i} + \sigma^2}, \tag{4}$$

where $p_{\mathcal{B}}$ is the power budget of the base station and σ^2 is the noise variance.

B. Problem Formulation

Since the UAV jammer is a malicious user, the UAV jammer cannot obtain the complete observation information of the users, i.e., CSI. The partially observable information that the UAV jammer known is the location of the users, which represents as the distances from the users to the base station, giving by

$$d_{\mathcal{B}i} = \sqrt{x_i^2 + y_i^2 + H_{\mathcal{B}}^2}, \quad i \in \{1, \cdots, U\}.$$

Meanwhile, the information observed by the users continuously is the jamming power received from the UAV.[1] Considering the hierarchical interactions among UAV jammer and the users, we utilize a stackelberg dynamic game $\mathbb{G}\langle\{\mathcal{J}, i\}, \{d_{\mathcal{J}}, d_i\}, \{r_{\mathcal{J}}, r_i\}\rangle$ to formulate the anti-UAV jamming problem, namely, anti-jamming elude game. In the formulated game, we model the foresighted UAV jammer $\mathcal{J}$ as a leader and the myopic users $i \in \{1, \cdots, U\}$ as followers. The UAV jammer first chooses its action $a_{\mathcal{J}} \in \mathcal{A}_{\mathcal{J}}$, then each user chooses its corresponding action $a_i \in \mathcal{A}_i$. We assume that the location of the user i is $(x_i, y_i, 0)$ in the previous time slot and $(x_i', y_i', 0)$ in the current time slot with action a_i, i.e., $(x_i', y_i', 0) = (x_i, y_i, 0) + a_i$. The location of the UAV jammer $\mathcal{J}$ is $(x_{\mathcal{J}}, y_{\mathcal{J}}, z_{\mathcal{J}})$ in the previous time slot and $(x_{\mathcal{J}}', y_{\mathcal{J}}', z_{\mathcal{J}}')$ in the current time slot with action $a_{\mathcal{J}}$, i.e., $(x_{\mathcal{J}}', y_{\mathcal{J}}', z_{\mathcal{J}}') = (x_{\mathcal{J}}, y_{\mathcal{J}}, z_{\mathcal{J}}) + a_{\mathcal{J}}$.

In this case, the immediate reward of user i can be given as

$$r_i[\mathscr{T}(a_{\mathcal{J}}), \mathscr{L}(a_i)] = \frac{p_{\mathcal{B}} d_{\mathcal{B}i}^{-\eta} |\tilde{h}_{\mathcal{B}i}|^2}{I_{\mathcal{J}i} + \sigma^2} - C_U d_i, \tag{5}$$

where $\mathscr{T}(a_{\mathcal{J}}) = (x_{\mathcal{J}}', y_{\mathcal{J}}', z_{\mathcal{J}}')$ denotes the current trajectory of the jammer with action $a_{\mathcal{J}}$, $\mathscr{L}(a_i) = (x_i', y_i', 0)$ denotes the current trajectory of user i with action a_i, C_U is the unit energy cost of the user, i.e., mobility cost per unit distance. The distance between UAV jammer $\mathcal{J}$ and user i is

$$d_{\mathcal{J}i} = \sqrt{(x_{\mathcal{J}}' - x_i')^2 + (y_{\mathcal{J}}' - y_i')^2 + z'^2_{\mathcal{J}}},$$

the distance from the base station to user i is

$$d_{\mathcal{B}i} = \sqrt{x'^2_i + y'^2_i + H_{\mathcal{B}}^2}$$

and the moving distance per time slot is

$$d_i = \sqrt{(x_i' - x_i)^2 + (y_i' - y_i)^2}.$$

The UAV jammer's immediate reward in the current time slot can be given by

$$r_{\mathcal{J}}[\mathscr{T}(a_{\mathcal{J}}), \mathscr{L}(a_i)] = \sum_{i=1}^{U} \frac{I_{\mathcal{J}i}}{p_{\mathcal{B}} d_{\mathcal{B}i}^{-\eta} |\tilde{h}_{\mathcal{B}i}|^2 + \sigma^2} - C_{\mathcal{J}} d_{\mathcal{J}}, \tag{6}$$

where $C_{\mathcal{J}}$ is the unit energy cost of the UAV jammer, i.e., flight cost per unit distance, and the flight distance per time slot can be denoted as

$$d_{\mathcal{J}} = \sqrt{(x_{\mathcal{J}}' - x_{\mathcal{J}})^2 + (y_{\mathcal{J}}' - y_{\mathcal{J}})^2 + (z_{\mathcal{J}}' - z_{\mathcal{J}})^2}.$$

[1]This is a reasonable assumption since that the jamming is continuous and the users can estimate it in each inter frame gap.

The goal of the formulated optimization problem is to maximize the long-term cumulative rewards of UAV jammer and users, respectively. To maximize jammer's long-term cumulative reward $R_{\mathcal{J}}$, we need to find the optimal jamming trajectory for the UAV jammer and then to maximize each user's long-term cumulative reward R_i, we need to find the optimal communication trajectory for each user, with the constraints of flight distance and moving distance per time slot. The formulated optimization problem can be given as

$$\begin{aligned}
\max_{a_{\mathcal{J}}, a_i} \ & R_{\mathcal{J}}[\mathscr{T}(a_{\mathcal{J}}), \mathscr{L}(a_i)], \\
& R_i[\mathscr{T}^*(a_{\mathcal{J}}), \mathscr{L}(a_i)], \\
\text{s.t.} \ & |a_{\mathcal{J}}| \leq 1, \qquad (7) \\
& |a_i| \leq 1, \quad i \in \{1, \cdots, U\}, \qquad (8)
\end{aligned}$$

where $R_{\mathcal{J}} = \sum_{k=0}^{\infty} \gamma^k r_{\mathcal{J}}(k)$ and $R_i = \sum_{k=0}^{\infty} \gamma^k r_i(k)$ denote k steps long-term cumulative rewards of each time slot with discount factor γ, (7) represents the flight distance of UAV jammer per time slot, (8) represents the moving distance of user i per time slot. Due to the mobility of the network, the communication environment is dynamic and complex. The formulated optimization problem faces several challenges, including the need to obtain the complete CSI, the need to obtain the channel state transition probability, as well as the difficulty to obtain the convexity of the problem. Therefore, to solve the formulated optimal problem, we propose the following strategies.

III. Deep Learning Based Optimal Strategy

In this section, we propose a novel anti-intelligent UAV jamming strategy to defend against UAV jammer. Particularly, we analyze the optimal jamming trajectory and the optimal communication trajectory.

A. The Optimal Jamming Trajectory

Since the wireless channel environment is dynamic and complex, we quantize the channel $h_{\mathcal{B}i}$ into a finite channel state space $\mathcal{S} = \{h_{\mathcal{B}i}^1, \cdots, h_{\mathcal{B}i}^K\}, i \in \{1, \cdots, U\}$, and model it as a Markov chain with finite states [35]. Then, by partitioning the flight space of the UAV jammer $\mathcal{J}$ into a finite number of states, i.e., L states, the flight state space of the UAV jammer $\mathcal{J}$ can be denoted as

$$\mathcal{S}_{\mathcal{J}} = \{(x_{\mathcal{J},1}, y_{\mathcal{J},1}, z_{\mathcal{J},1}), \cdots, (x_{\mathcal{J},L}, y_{\mathcal{J},L}, z_{\mathcal{J},L})\}.$$

Again, we quantize the motion state space of the users into M states, which is denoted as

$$\mathcal{S}_i = \{(x_{i,1}, y_{i,1}, 0), \cdots, (x_{i,M}, y_{i,M}, 0)\}, \quad i \in \{1, \cdots, U\}.$$

To simplify the case, we model a virtual user, V, as a target user, which is a virtual point that related to the users in the network. The initial location of the virtual user can be decided by

$$(x_V, y_V, 0) = \left(\frac{\sum_{i=1}^{U} w_i x_i}{\sum_{i=1}^{U} w_i}, \frac{\sum_{i=1}^{U} w_i y_i}{\sum_{i=1}^{U} w_i}, 0 \right), \tag{9}$$

where w_i is the initial location weight of user i. Then, the quantized motion state space of the virtual user can be denoted as

$$\mathcal{S}_V = \{(x_{V,1}, y_{V,1}, 0), \cdots, (x_{V,M}, y_{V,M}, 0)\}.$$

Remark 1: Since the communication fairness among users, the base station will allocate more bandwidth to the user far away from it. Thus, the initial value of the location weights w_i is proportion to the distance between base station and user i, i.e., $w_i \propto d_{\mathcal{B}i}$. As UAV flies at very high altitudes, it can obtain the location of each user, then it can approximately estimate the initial location weights w_i based on the distance between base station and user i, i.e., $w_i = \frac{d_{\mathcal{B}i}}{\sum_{i=1}^{U} d_{\mathcal{B}i}}$. As the users moving, the location weight w_i will be adjusted with the time. Let $\mathbf{A}\mathbf{w} = \mathbf{b}$, where

$$\mathbf{w} = (w_1 \ \ w_2 \ \ \cdots \ \ w_U)^{\dagger},$$

$$\mathbf{A} = \begin{pmatrix} x_1 & x_2 & \cdots & x_U \\ y_1 & y_2 & \cdots & y_U \\ 0 & 0 & \cdots & 0 \end{pmatrix},$$

$$\mathbf{B} = (\mathbf{A}, \mathbf{b}) = \begin{pmatrix} x_1 & x_2 & \cdots & x_U & x_V \\ y_1 & y_2 & \cdots & y_U & y_V \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}.$$

Excepting the special case $\forall i \in \{1, \cdots, U\}, x_i = y_i, x_V \neq y_V$, we can find that the location of the virtual user can be represented by the locations of all the users, linearly. The special case means that all users are on the surface diagonal of the solid figure and the UAV jammer is not. Since the communication environment is complex and the user number is large, the special case above is hard to occur in practice. In the following analysis, we assume that the location relationship between virtual user and users are always linear.

The UAV jammer's immediate reward in (6) can be transformed to

$$r_{\mathcal{J}}[\mathscr{T}(a_{\mathcal{J}}), \mathscr{L}(a_V)] = \frac{I_{\mathcal{J}V}}{p_{\mathcal{B}} d_{\mathcal{B}V}^{-\eta} |\tilde{h}_{\mathcal{B}V}|^2 + \sigma^2} - C_{\mathcal{J}} d_{\mathcal{J}}, \quad (10)$$

where the distance

$$d_{\mathcal{B}V} = \sqrt{{x'}_V^2 + {y'}_V^2 + H_{\mathcal{B}}^2}.$$

Then the optimization problem for the UAV jammer $\mathcal{J}$ is formulated as choosing action $a_{\mathcal{J}}$ to maximize UAV jammer's long-term cumulative reward under the constraint of moving distance per time slot, which can be given by

$$\begin{aligned} &\max_{a_{\mathcal{J}}} \ R_{\mathcal{J}}[\mathscr{T}(a_{\mathcal{J}}), \mathscr{L}(a_V)], \\ &\text{s.t.} \ \ |a_{\mathcal{J}}| \leq 1. \end{aligned} \quad (11)$$

However, the complete CSI of the virtual user is not known to the UAV jammer. Considering the dynamic channel environments, we model this process as a partially observable Markov decision process (POMDP) [28]. Define a POMDP as a 6-tuple $\langle \mathcal{S}, \mathcal{A}_{\mathcal{J}}, P, r_{\mathcal{J}}, \mathcal{O}, \Omega \rangle$, where

- $\mathcal{S}$ is the channel state space;
- $\mathcal{A}_{\mathcal{J}}$ is the action space;
- $P(\cdot|s, a_{\mathcal{J}})$ is the transition probability of the next state, conditioned on action $a_{\mathcal{J}}$ being chosen in state $s \in \mathcal{S}$;
- $r_{\mathcal{J}}[s, \mathscr{T}(a_{\mathcal{J}})]$ is the immediate reward obtained when action $a_{\mathcal{J}}$ is taken in state s, and the symbol $r_{\mathcal{J}}[s, \mathscr{T}(a_{\mathcal{J}})]$ is omitted to $r_{\mathcal{J},s}$ if no confusion occurs;
- $\mathcal{O}$ is the observation state space, which is equal to the motion state space $\mathcal{S}_V$;
- $\Omega(\cdot|s, a_{\mathcal{J}})$ is the probability of the possible observation, conditioned on action $a_{\mathcal{J}}$ being taken to reach state s.

According to the observation o, the probability of being in state s is defined by the belief b, which can be updated by

$$b'(s') = \frac{1}{\Theta} \left[\Omega(o'|s', a_{\mathcal{J}}) \sum_{s \in \mathcal{S}} P(s'|s, a_{\mathcal{J}}) b(s) \right], \quad (12)$$

where

$$\Theta = \sum_{s' \in \mathcal{S}} \Omega(o'|s', a_{\mathcal{J}}) \sum_{s \in \mathcal{S}} P(s'|s, a_{\mathcal{J}}) b(s)$$

is the normalization function of the belief and the belief is initialized at $b^0 = P_0$, i.e., $P_0 = 0.1$. Define the action selection policy as $\pi : b \rightarrow a_{\mathcal{J}}$. Then, solving the POMDP is to find the optimal action selection policy $\pi^* : b^* \rightarrow a_{\mathcal{J}}^*$, yields the maximum expected reward for each belief. This maximum expected reward can be obtained by the Bellman equation

$$V_b^* = \max_{a_{\mathcal{J}} \in \mathcal{A}_{\mathcal{J}}} \left[r_{\mathcal{J},b} + \gamma \sum_{o \in \mathcal{O}} \Omega(o|b, a_{\mathcal{J}}) V_{b'}^* \right], \quad (13)$$

where

$$r_{\mathcal{J},b} = \sum_{s \in \mathcal{S}} r_{\mathcal{J},s} b(s)$$

represents the expected reward over the belief distribution.

For any partially observable with known state transition probability $P(\cdot|s, a_{\mathcal{J}})$, the problem can be reformulated as a belief-MDP, which uses belief state space $\mathcal{M}$ as a new state space instead of the original channel state space $\mathcal{S}$ [36]. The near-optimal solution to the belief-MDP can be solved by Q-learning [37]. By storing and updating a Q-value function for each belief in the system, the optimal action $a_{\mathcal{J}}^*$ with respect to the maximum Q-value is obtained. However, in practice, the belief space is large and the state transition probability is unknown, the Q-learning is impossible to store and update the Q-value function. Therefore, we use the model-free approach to learn the trajectory, which directly exploits the sequence of ℓ historical observation-action pairs, $\mathbb{O}^t = \{o^{t-\ell}, a_{\mathcal{J}}^{t-\ell}, \cdots, o^{t-1}, a_{\mathcal{J}}^{t-1}\}$ to learn the optimal jamming trajectory [28]. The DRQN that combines Q-learning with a recurrent convolutional neural network (CNN), is developed. The framework is shown in Fig. 2. In each Q-network, the neural network consists of two convolutional layers, one long short-term memory (LSTM) layer, and one fully connected (FC) layer. The first convolutional layer convolves $\mathcal{F}_1$ filters of $n_1 \times n_1$ with stride 1, and the second convolutional layer convolves $\mathcal{F}_2$ filters of $n_2 \times n_2$ with stride 1. The LSTM layer consists of $\mathcal{C}_1$ rectifier unites and FC layer includes $|\mathcal{A}_{\mathcal{J}}|$ rectifier unites.

Solving the formulated POMDP problem via the developed DRQN, the Q-values are parameterized by $Q(\phi, a_{\mathcal{J}}; \theta)$, where

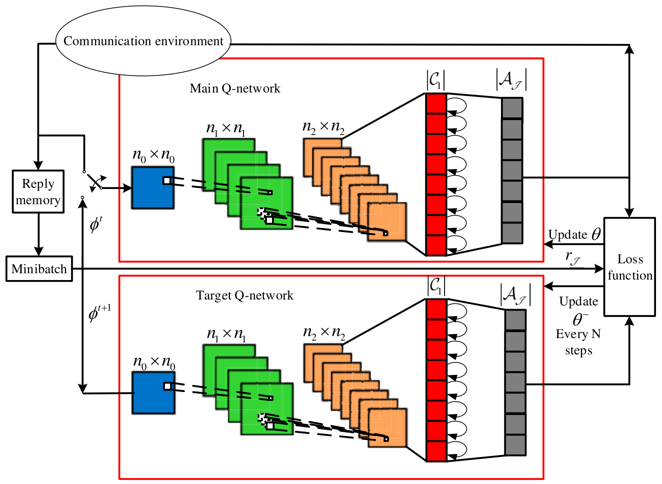


Fig. 2. The developed DRQN framework, which includes one main Q-network and one target Q-network. Each Q-network consists of one input layer, two convolutional layers, one LSTM layer, and one FC layer.

θ is the weight parameter set of the Q-network. In time slot t, sequence $\mathbb{O}^t$ can be preprocessed to an $n_0 \times n_0$ matrix ϕ^t, then input this matrix to the recurrent CNN to calculate $Q(\phi^t, a_{\mathcal{J}}; \theta)$. Once θ is learned, the Q-values are determined. Then, the UAV jammer's experience $e^t_{\mathcal{J}}(\phi^t, a^t_{\mathcal{J}}, r^t_{\mathcal{J}}, \phi^{t+1})$ is stored in the replay memory $\mathcal{D}_{\mathcal{J}} = \{e^1_{\mathcal{J}}, \cdots, e^t_{\mathcal{J}}\}$. When training the DRQN, mini-batches of experience $e^g_{\mathcal{J}}, 1 \le g \le t$ from the pool of the reply memory is randomly chosen to update the weight parameter set θ via a stochastic gradient descent (SGD). The weight parameter set θ is updated via the loss function

$$L(\theta) = \mathbb{E}_{\phi,a,r,\phi'}\big[\big(r_{\mathcal{J},\phi} + \gamma \max_{a'_{\mathcal{J}}} Q(\phi', a'_{\mathcal{J}}; \theta^-) - Q(\phi, a_{\mathcal{J}}; \theta)\big)^2\big], \quad (14)$$

where the symbol θ^- is only updated with θ every N steps from the same Q-network. The gradient of loss function with respect to the weight parameter set θ is obtained by

$$\nabla_\theta L(\theta) = \mathbb{E}_{\phi,a,r,\phi'}\big[\big(r_{\mathcal{J},\phi} + \gamma \max_{a'_{\mathcal{J}}} Q(\phi', a'_{\mathcal{J}}; \theta^-) - Q(\phi, a_{\mathcal{J}}; \theta)\big)\nabla_\theta Q(\phi, a_{\mathcal{J}}; \theta)\big]. \quad (15)$$

To balance the exploration and exploitation, we utilize the ϵ-greedy policy $\pi_{\mathcal{J}}$ to select the action with greedy probability $P(a_{\mathcal{J}} = a^*_{\mathcal{J}}) = 1 - \epsilon$, where $\epsilon \in (0, 1)$ is a small positive value, i.e., $\epsilon = 0.01$. Then, the optimal jamming trajectory at time t can be denoted by

$$\mathscr{T}^*(a^t_{\mathcal{J}}) = (x_{\mathcal{J}0}, y_{\mathcal{J}0}, z_{\mathcal{J}0}) + {a^0_{\mathcal{J}}}^* + {a^1_{\mathcal{J}}}^* + \cdots + {a^t_{\mathcal{J}}}^*, \quad (16)$$

where $(x_{\mathcal{J}0}, y_{\mathcal{J}0}, z_{\mathcal{J}0})$ is the initial location of the UAV jammer.

B. The Optimal Communication Trajectory

In the follower sub-game, the virtual user V chooses the optimal action $a^*_V \in \mathcal{A}_V$ based on the observation of the UAV jammer, and obtains the optimal communication trajectory $\mathscr{L}^*(a_V)$ by solving

$$\max_{a_V} \; R_V[\mathscr{T}^*(a_{\mathcal{J}}), \mathscr{L}(a_V)], \quad \text{s.t. } |a_V| \le 1. \quad (17)$$

Since the optimal action a^*_V of the virtual user depends on the observation of the UAV jammer, we can derive the insightful property between action a_V and action $a_{\mathcal{J}}$, which is given by the following theorem.

Theorem 1: The communication trajectory is decided by the observation-action transition of the UAV jammer, and the action transition probability $P(a_{\mathcal{J}}|a'_{\mathcal{J}})$ follows an independent and identically distribution finite state Markov chain. ■

Proof: Please see Appendix A.

From Theorem 1, the optimizing communication trajectory problem can be modeled as solving a MDP problem, in which the communication trajectory of the virtual user is determined by the state $\mathcal{S}_{\mathcal{J}}$ with respect to the action of the UAV jammer, i.e., $s'_{\mathcal{J}} = s_{\mathcal{J}} + a'_{\mathcal{J}}$. The MDP can be denoted as a 4-tuple $\langle \mathcal{S}_{\mathcal{J}}, \mathcal{A}_V, r_V, P(\cdot|s_{\mathcal{J}}, a_V)\rangle$, where

- $\mathcal{S}_{\mathcal{J}}$ is the flight state space,
- $\mathcal{A}_V$ is the action space,
- $r_V[s_{\mathcal{J}}, \mathscr{L}(a_V)]$ is the immediate reward obtained when action a_V is taken in state $s_{\mathcal{J}}$, and the symbol $r_V[s_{\mathcal{J}}, \mathscr{L}(a_V)]$ is omitted to $r_{V,s_{\mathcal{J}}}$ if no confusion occurs.
- $P(\cdot|s_{\mathcal{J}}, a_V)$ is the transition probability of the next state, conditioned on action a_V being chosen in state $s_{\mathcal{J}} \in \mathcal{S}_{\mathcal{J}}$.

We have

$$\begin{aligned} P(s^{t+1}_{\mathcal{J}}|s^t_{\mathcal{J}}, a_V) &= P(s^t_{\mathcal{J}} + a^{t+1}_{\mathcal{J}}|s^t_{\mathcal{J}}, a_V) \\ &= P(a^0_{\mathcal{J}} + \cdots + a^{t+1}_{\mathcal{J}}|a^0_{\mathcal{J}} + \cdots + a^t_{\mathcal{J}}, a_V) \\ &= P(a^{t+1}_{\mathcal{J}}|a^t_{\mathcal{J}}, a_V). \end{aligned} \quad (18)$$

Then, we apply the Q-learning to derive the optimal communication trajectory of virtual user $\mathscr{L}^*(a_V)$ with the observation of the UAV jammer.

Considering the state space $\mathcal{S}_{\mathcal{J}}$ is large, we develop the CNN to approximate the Q-value function. Then, we utilize the DQN to estimate the Q-value with the weight parameter ξ [27]. The developed DQN framework is shown in Fig. 3, including the main Q-network and the target Q-network. Specifically, in time slot t, the sequence of ℓ historical state-action pairs $\mathbb{S}^t = \{s^{t-\ell}_{\mathcal{J}}, a^{t-\ell}_V, \cdots, s^{t-1}_{\mathcal{J}}, a^{t-1}_V\}$ is preprocessed to an $n \times n$ matrix φ^t as the input to the CNN. The experience of the user $e^t_V(\varphi^t, a^t_V, r^t_V, \varphi^{t+1})$ is stored in the replay memory $\mathcal{D}_V = \{e^1_V, \cdots, e^t_V\}$. When training the DQN, mini-batches of experience $e^g_V, 1 \le g \le t$ from the pool of the replay memory is randomly chosen to update weight parameter set ξ via a SGD. The weight parameter set ξ is updated via the following loss function

$$L(\xi) = \mathbb{E}_{\varphi,a,r,\varphi'}\big[\big(r_{V,s_{\mathcal{J}}} + \gamma \max_{a'_V} Q(\varphi', a'_V; \xi^-) - Q(\varphi, a_V; \xi)\big)^2\big],$$

where the symbol ξ^- is updated from the same Q-network to minimize the loss function in every N steps. The gradient

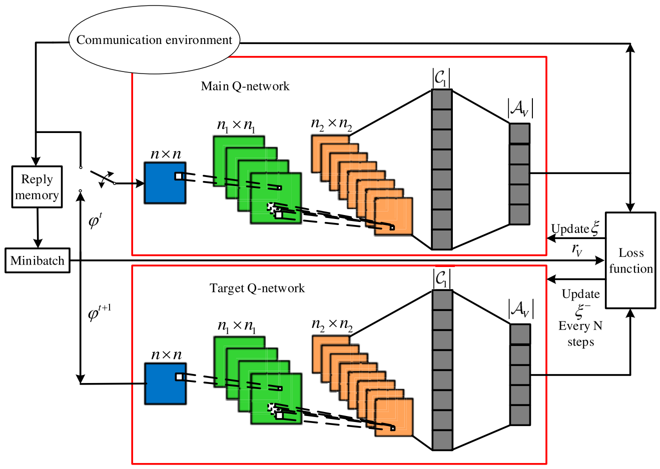


Fig. 3. The developed DQN framework, which includes one main Q-network and one target Q-network. Each Q-network consists of one input layer, two convolutional layers and two FC layers.

of loss function with respect to the weight parameter set ξ is obtained by

$$\nabla_\xi L(\xi) = \mathbb{E}_{\varphi,a,r,\varphi'}\big[\big(r_{V,s_{\mathcal{J}}} + \gamma \max_{a'_V} Q(\varphi', a'_V; \xi^-) \\ - Q(\varphi, a_V; \xi)\big)\nabla_\xi Q(\varphi, a_V; \xi)\big]. \quad (19)$$

The optimal action in ϵ-greedy policy π_V with greedy probability $P(a_V = a_V^*) = 1 - \epsilon$ is given by

$$a_V^* = \arg \max_{a_{\mathcal{J}} \in \mathcal{A}_{\mathcal{J}}} Q(\varphi, a_V; \xi). \quad (20)$$

The optimal communication trajectory of virtual user $\mathscr{L}^*(a_V)$ in time slot t is given by

$$\mathscr{L}^*(a_V^t) = (x_{V0}, y_{V0}, 0) + {a_V^0}^* + {a_V^1}^* + \cdots + {a_V^t}^*, \quad (21)$$

where $(x_{V0}, y_{V0}, 0)$ is the initial location of the virtual user. However, the optimal communication trajectory of virtual user is an equivalent solution, as described in (9). Actually, we have to prove the existence of the optimal communication trajectory for each user after using the DQN, thus, we derive the following lemma and theorem.

Lemma 1: For any multivariate function $f(\mathbf{c}_1, \cdots, \mathbf{c}_U) = f_1(\mathbf{c}_1) + \cdots + f_U(\mathbf{c}_U)$, if

$$\frac{\partial^2\ f_i(\mathbf{c}_i)}{\partial^2 \mathbf{c}_i} \geqslant 0, \quad \forall i \in 1, \cdots, U \quad (22)$$

then, the optimal solution that satisfies $f^*(\mathbf{c}_1, \cdots, \mathbf{c}_U) = f_1^*(\mathbf{c}_1) + \cdots + f_U^*(\mathbf{c}_U)$.

Proof: Please see Appendix B. ■

Theorem 2: For the optimal communication trajectory of virtual user in each time slot, denoted as $\mathscr{L}_V^*$, the optimal communication trajectory $\mathscr{L}_i^*, i \in 1, \cdots, U$ that maximizes the long-term cumulative reward for each user is existent.

Proof: Please see Appendix C. ■

Remark 2: The relationship between optimal communication trajectory of virtual user and optimal communication trajectories of users are linear. In addition, we can further derive that if the optimal communication trajectory of virtual user exists, then the optimal communication trajectory of each user is existent but not unique, which can be proved as follow:

Based on the non-homogeneous linear equations, we can rewrite (33) as $(\mathbf{A}^*\mathbf{w})^\dagger = \mathbf{b}^{*\dagger}$, where

$$\mathbf{A}^* = \begin{pmatrix} \mathbf{a}_1^* \\ \mathbf{a}_2^* \\ \mathbf{a}_3^* \end{pmatrix} = \begin{pmatrix} x_1^* & x_2^* & \cdots & x_U^* \\ y_1^* & y_2^* & \cdots & y_U^* \\ 0 & 0 & \cdots & 0 \end{pmatrix},$$

$$\mathbf{w} = (w_1 \quad w_2 \quad \cdots \quad w_U)^\dagger,$$

$$\mathbf{b}^* = (b_1, b_2, b_3)^\dagger = (x_V^* \ y_V^* \ 0)^\dagger.$$

Let $(\mathbf{a}_j^*\mathbf{w})^\dagger = b_j, \mathbf{p}_j = (\mathbf{w}^\dagger, b_j),\ j \in \{1, 2, 3\}$, then for given $\mathbf{w}$ and $\forall\ j \in \{1, 2, 3\}$, we have $Rank(\mathbf{w}^\dagger) = Rank(\mathbf{p}_j) = 1 < U$, the solutions of $x_i^*, y_i^*,\ i \in \{1, \cdots, U\}$ are existent but not unique.

As per Theorem 2, the optimal communication trajectory of each user in time slot t is given by

$$\mathscr{L}^*(a_i^t) = (x_{i0}, y_{i0}, 0) + {a_i^0}^* + {a_i^1}^* + \cdots + {a_i^t}^*, i \in 1, \cdots, U \quad (23)$$

where $(x_{i0}, y_{i0}, 0)$ is the initial location of user i.

C. Discussions

Here, we prove the existence of stackelberg equilibrium in the game, and then we analyze the time complexity of the proposed defense strategy.

1) Stackelberg Equilibrium:

Definition 1: Given a two-player stackelberg game, where player 1 as a leader wants to maximize a reward function $r_1(a_1, a_2)$ and player 2 as a follower wants to maximize a reward function $r_2(a_1, a_2)$ by choosing a_1, a_2 from action space $\mathcal{A}_1$ and $\mathcal{A}_2$, respectively. Then the pair (a_1^*, a_2^*) is called a stackelberg equilibrium if for any a_1 belonging to $\mathcal{A}_1$ and a_2 belonging to $\mathcal{A}_2$, satisfies

$$r_1(a_1^*, a_2) \geq r_1(a_1, a_2) \\ r_2(a_1^*, a_2^*) \geq r_2(a_1^*, a_2(a_1^*)), \quad (24)$$

where the reward $r_2(a_1^*, a_2^*) = \max_{a_2} r_2(a_1^*, a_2(a_1^*))$ [38].

Remark 3: We note that the stackelberg equilibrium with the UAV jammer as a leader is the optimal solution for it if the UAV jammer chooses its action $a_{\mathcal{J}}^*$ first, and if the goal of the virtual user is to maximize R_V, while that of the UAV jammer is to maximize $R_{\mathcal{J}}$. If the leader chooses any other action $a_{\mathcal{J}}$, then the follower will choose an action $\tilde{a}_V^*$ to maximize R_V. In this case, the reward of the UAV jammer will be less than that when the stackelberg equilibrium with UAV jammer is used.

Theorem 3: In the proposed game with one UAV jammer $\mathcal{J}$ and one virtual user V, the DQN based optimal trajectory pairs $[\mathscr{T}^*(a_{\mathcal{J}}), \mathscr{L}^*(a_V)]$ is a stackelberg equilibrium.

Proof: Please see Appendix B. ■

Remark 4: Theoretically, a stackelberg equilibrium can be achieved with probability one, if the DQN is well trained. To balance the exploration and exploitation with respect to a large state-action space, it has a probability $2\epsilon - \epsilon^2$ that the system cannot obtain the optimal communication trajectory with respect to a stackelberg equilibrium in DQN training.

TABLE II
THE TIME COMPLEX OF THE PROPOSED DEFENSE STRATEGY

The testing time complexity	The training time complexity
$O\Big(\mathcal{F}_1\big(n_1^2(n-n_1+1)^2+\mathcal{F}_2n_2^2(n-n_1-n_2+2)^2\big)\Big)$	$O\Big(3\mathcal{F}_1\big(n_1^2(n-n_1+1)^2+\mathcal{F}_2n_2^2(n-n_1-n_2+2)^2\big)\Big)$

Since $\epsilon \in \{0,1\}$ is a small positive value, the probability event $2\epsilon-\epsilon^2$ is extremely small, i.e., $\epsilon = 0.05, 2\epsilon-\epsilon^2 = 0.0975$. Such occasional small probability event can help to fully explored and exploited the large state-action space and help to obtain the global optimal solution, then, the DQN can be well trained.

Corollary 1: If the initial location of the UAV jammer and the virtual user satisfies $x_{\mathcal{J}0} = y_{\mathcal{J}0}$ and $x_{V0} = y_{V0}$, and the channel is quasi-static block fading, then the anti-jamming elude game has a stackelberg equilibrium $[\mathscr{T}^*(a_{\mathcal{J}}), \mathscr{L}^*(a_V)]$, which is given by

$$\mathscr{T}^*(a_{\mathcal{J}}) = \Big(\frac{x_{\mathcal{J}0}-x_{V0}+x_{V0}z_{\mathcal{J}0}}{z_{\mathcal{J}0}}, \frac{y_{\mathcal{J}0}-y_{V0}+y_{V0}z_{\mathcal{J}0}}{z_{\mathcal{J}0}}, 1\Big),$$
$$\mathscr{L}^*(a_V) = (1,1,0).$$

Proof: Please see Appendix E. ■

Remark 5: In the above case, we note that the stakelberg equilibrium of the system is independent of the initial flight height $z_{\mathcal{J}0}$, and the optimal flight height $z_{\mathcal{J}}^*$ is a constant. The optimal communication trajectory of the virtual user satisfies $\{(x_V^*, y_V^*, 0)|(x_V^*, y_V^*, 0) \in \mathcal{S}_i, x_V^* = y_V^*\}$. In particular, $\mathscr{L}^*(a_V) = (0,0,0)$ has no physical meaning in practice, and $\mathscr{L}^*(a_V) = (1,1,0)$ is a special case.

2) Time Complexity Analysis: The total time complexity of anti-intelligent UAV jamming strategy mainly depends on the all convolutional layers, which can be defined as [39]

$$O\Big(\sum_{m=1}^{2}\mathcal{F}_{m-1}n_m^2\mathcal{F}_m\mu_m^2\Big), \tag{25}$$

where m is the index of convolutional layer, the symbol $\mathcal{F}_{m-1}$ is the number input channels of the m-th layer, i.e., $\mathcal{F}_0 = 1$, the symbol μ_m is the spatial size of the output feature map of the m-th convolutional layer.

In our developed CNN, the number of the convolutional layer $m = 2$. Thus, with regard to the first convolutional layer, each filter has size $n_1 \times n_1$ with stride 1, it inputs a $n \times n$ matrix, then outputs a feature map with size $(n-n_1+1)$. With each filter size $n_2 \times n_2$ and stride 1, the second convolutional layer inputs a $(n-n_1+1)$ matrix and outputs a feature map with size $(n-n_1-n_2+2)$. The total testing time complexity of the proposed strategy can be obtained via (25). Meanwhile, since the CNN training includes one forward propagation and two backward propagation, the training time complexity is roughly three times of the testing time complexity [39]. Therefore, the time complex of the proposed defense strategy is given in table II.

IV. SIMULATION RESULTS

In this section, we evaluate the performance of the anti-jamming elude game via simulations. In the simulations, the transmit power of the base station is $p_{\mathcal{B}} = 100$ mW, the jamming power of the UAV is $p_{\mathcal{J}} = 30$ mW, the noise power is $\sigma^2 = 1$ mW, the unit energy cost of the UAV jammer is $C_j = 0.9$ dB ≈ 1.23 mW and the unit energy cost of the virtual user is $C_U = 0.5$ dB ≈ 1.12 mW. From [32], we set the path-loss exponents for air-to-ground channel $\alpha = 3$, ground-to-ground channel $\eta = 2$, and the additional attenuation factors $\beta_{\text{LoS}} = 1$ dB, $\beta_{\text{NLoS}} = 20$ dB, respectively. The location of the base station is $(0,0,0)$ and the initial location of the virtual user is calculated by (9). The virtual user can move in a square area with size $X \times Y \times 1$, and the UAV jammer can move in a cube area with size $X \times Y \times Z$, where $X \in [-30\text{ m}, 30\text{ m}]$, $Y \in [-30\text{ m}, 30\text{ m}]$, and $Z \in [0\text{ m}, 30\text{ m}]$. To simplify simulation, the CSI is set to be real number, which changes in each time slot, and the size of state $\mathcal{S}$ is set to be 50. Likewise, the size of state $\mathcal{S}_{\mathcal{J}}$ is also set to be 50. The neural network consists of 2 hidden layers with the discount factor $\gamma = 0.95$, and greedy rate $\epsilon = 0.1$.

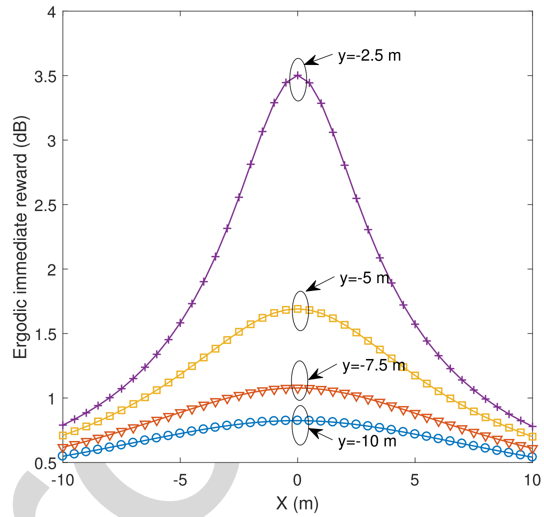


Fig. 4. The ergodic immediate reward of the virtual user at different location. The UAV is at $(-10$ m, 20 m, 50 m) and state changes 1000 times.

As the channel environment is dynamic, it is difficult to directly analyze the immediate reward. Thus, we analyze the immediate reward based on the ergodic immediate reward. Fig. 4. shows the tangent plane of ergodic immediate reward of virtual user in different location, corresponding to the location of the UAV is $(-10$ m, 20 m, 50 m). Some interesting insights are obtained. For instance, with the distance between virtual user and base station decreases, the immediate reward received by the virtual user increases. In particular, such increasing trend is non-linear and the ergodic immediate reward of the virtual user is maximum at (0 m, 0 m, 0 m). For example, when coordinate $x = 0$ m is fixed, the coordinate y changes from -10 m to -7.5 m which increases 0.25 dB ergodic immediate reward, and from -7.5 m to -5 m which increases 0.65 dB ergodic immediate reward.

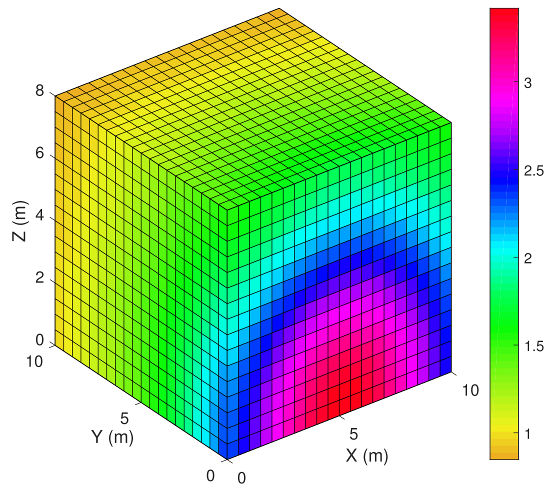


Fig. 5. The ergodic immediate reward of the UAV jammer at different location. The virtual user is at (5 m, -5 m) and the state changes 1000 times.

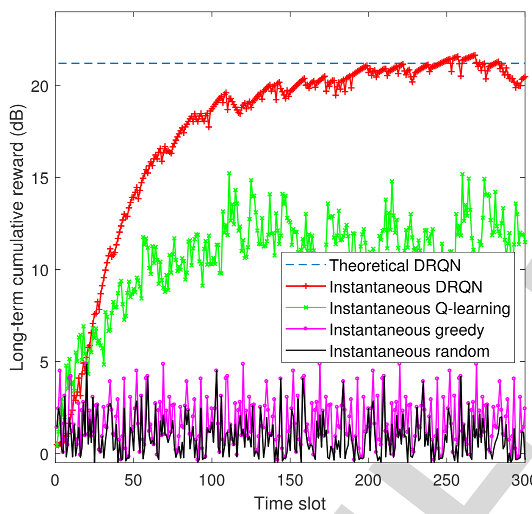


Fig. 6. The long-term cumulative rewards of the UAV jammer in DRQN, greedy, random and Q-learning strategy in 300 time slots.

When the location of the virtual user is (5 m, -5 m, 0 m), making state s change 1000 times, the ergodic immediate reward of the UAV jammer is shown in Fig. 5. We find that the tangent plane of the ergodic immediate reward can be approximated to a hemisphere. It shows that the closer the distance between virtual user and UAV jammer is, the higher the ergodic immediate reward will be. In addition, we observe that the ergodic immediate reward decreases with the increasing flight height $z_{\mathcal{J}}$ and it decreases rapidly when the coordinate y is greater than 2 m. The reason is that the gradient of the edge is large, which leads to the immediate reward decreases rapidly. The result suggests that if the attacker only launches jamming in one time slot, the UAV jammer should stay close to the virtual user as soon as possible to obtain a high ergodic immediate reward. Furthermore, one interesting observation is that the ergodic immediate reward is symmetric about $x = 5$ under the parameters setting above.

The long-term cumulative rewards of the UAV jammer in 300 time slots is presented in Fig. 6. We leverage the greedy strategy, random strategy and Q-learning strategy as

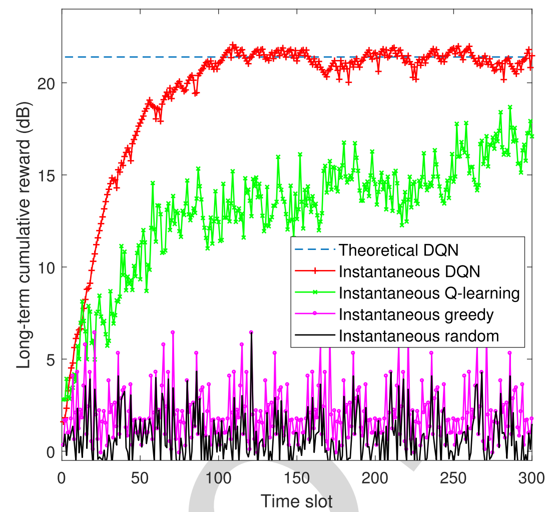


Fig. 7. The long-term cumulative rewards of the virtual user in DQN, greedy, random and Q-learning strategy in 300 time slots.

benchmark methods and compare them with the proposed DRQN based intelligent jamming strategy. Since the greedy strategy and the random strategy do not consider a series of time events, for these two strategies, the long-term cumulative rewards are equal to immediate rewards. We find that the long-term cumulative reward via DRQN can converge to 21.2 dB after 200 time slots. However, due to the state spaces are large, the Q-learning strategy cannot update the Q-table effectively. Thus, the convergence speed of Q-learning is slower than DRQN based strategy. And, even after 300 time slots, the Q-learning based strategy cannot converge to a fixed value. The performance of the proposed strategy is already superior to the greedy strategy and the random strategy after 25 time slots. For example, the proposed strategy can achieve 75% higher long-term cumulative reward than the greedy reward in the 200-th time slot. In benchmark methods, we also find that the greedy strategy can achieve a better performance than the random strategy, and the Q-learning based strategy is the best of the three.

We obtain the long-term cumulative rewards of the virtual user in Fig. 7. The result suggests that the long-term cumulative reward via DQN can converge to 22.3 dB after 100 time slots. After 10 time slots, the DQN based strategy has already get a higher long-term cumulative reward than random and greedy strategies. Then, after 20 time slots, the proposed strategy is better than the Q-learning base strategy. In summary, these two figures show that both the UAV jammer and the virtual user can obtain the highest long-term cumulative rewards via the proposed strategy, respectively. That is, the stackelberg equilibrium exists after the long-term cumulative reward converges.

Fig. 8. presents the optimal jamming trajectory of the UAV and the optimal communication trajectory of the virtual user in one episode. We observe that the communication location of the virtual user starts at (-2 m, 1 m, 0 m) and ends at (15 m, 18 m, 0 m) and the jamming location of the UAV starts at (0 m, 0 m, 10 m) and ends at (15 m, 15 m, 0 m). To obtain the maximum long-term cumulative reward, the UAV jammer

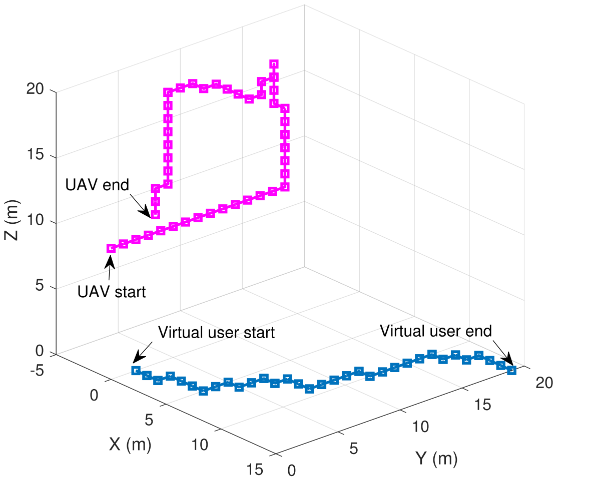


Fig. 8. The optimal trajectories via learning in one episode, the UAV jammer via DRQN vs. the virtual user via DQN.

will not prefer to stay close to the virtual user in each time slot as analyzed in Fig. 5. The reason is that the CSI is time varying in each time slot, the UAV jammer will consider the CSI transition probability to maximize long-term cumulative reward rather than considering the instantaneous CSI only.

V. Conclusions

In this paper, we have proposed the anti-intelligent UAV jamming strategy via deep Q-networks. Specifically, we have formulated the anti-UAV jamming problem as a stackelberg dynamic game, in which the UAV jammer acts as a leader and the users act as followers. We have modeled the leader sub-game as a partially observable Markov decision process and have learned the optimal jamming trajectory via deep recurrent Q-networks in the three-dimension space. Then, we have modeled the follower sub-game as a Markov decision process. The optimal communication trajectory has been learned via deep Q-networks in the two-dimension space. The time complexity of the defense strategy has been analyzed via theory and the performance of the proposed defense strategy has been evaluated by simulations. Some insightful remarks have been obtained: 1) If the optimal trajectory of virtual user exists, the optimal communication trajectory of each user is existent but is not unique. 2) In quasi-static block fading, the stakelberg equilibrium of the system is independent of the initial flight height, and the optimal flight height is a constant. 3) To maximize long-term cumulative reward, the action choices of UAV jammer is different from that of maximizing the immediate reward.

Appendix A
Proof of Theorem 1

The action transition probability of UAV jammer can be divided into two cases based on ϵ-greedy policy $\pi_{\mathcal{J}}$.

Case 1: If the UAV jammer chooses the optimal action ${a'_{\mathcal{J}}}^*$ in the next time slot, then

$$\begin{aligned} P({a'_{\mathcal{J}}}^*|a_{\mathcal{J}}) &= P(o', {a'_{\mathcal{J}}}^*|o, a_{\mathcal{J}}) \\ &= P({a'_{\mathcal{J}}}^*)P(o'|o, a_{\mathcal{J}}) \\ &= (1-\epsilon)P(o'|o, a_{\mathcal{J}}). \end{aligned} \tag{26}$$

Case 2: If the UAV jammer chooses the non-optimal action $\tilde{a'_{\mathcal{J}}}^*$ in the next time slot, then

$$\begin{aligned} P(\tilde{a'_{\mathcal{J}}}^*|a_{\mathcal{J}}) &= P(o', \tilde{a'_{\mathcal{J}}}^*|o, a_{\mathcal{J}}) \\ &= P(\tilde{a'_{\mathcal{J}}}^*)P(o'|o, a_{\mathcal{J}}) \\ &= \epsilon P(o'|o, a_{\mathcal{J}}), \end{aligned} \tag{27}$$

where the action $a_{\mathcal{J}} \in \{a^*_{\mathcal{J}}, \tilde{a_{\mathcal{J}}}^*\}$. As per (26) (27), we have the action transition probability $P(a'_{\mathcal{J}}|a_{\mathcal{J}}) = P(o'|o, a_{\mathcal{J}})$. Given current action $a_{\mathcal{J}}$, we note that the next action $a'_{\mathcal{J}}$ is independent of the previous action, which has a Markov property. Then proof is completed.

Appendix B
Proof of Lemma 1

Taking the second derivative of function $f(\mathbf{c}_1, \cdots, \mathbf{c}_U)$, we can get the Hessian matrix

$$\frac{\partial^2\ f(\mathbf{c}_1, \cdots, \mathbf{c}_U)}{\partial^2 \mathbf{c}_1, \cdots, \mathbf{c}_U} = \begin{bmatrix} \frac{\partial^2 f}{\partial \mathbf{c}_1^2} & \frac{\partial^2 f}{\partial \mathbf{c}_1 \partial \mathbf{c}_2} & \cdots & \frac{\partial^2 f}{\partial \mathbf{c}_1 \partial \mathbf{c}_U} \\ \frac{\partial^2 f}{\partial \mathbf{c}_2 \partial \mathbf{c}_1} & \frac{\partial^2 f}{\partial \mathbf{c}_2^2} & \cdots & \frac{\partial^2 f}{\partial \mathbf{c}_2 \partial \mathbf{c}_U} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial \mathbf{c}_U \partial \mathbf{c}_1} & \frac{\partial^2 f}{\partial \mathbf{c}_U \partial \mathbf{c}_2} & \cdots & \frac{\partial^2 f}{\partial \mathbf{c}_U^2} \end{bmatrix}. \tag{28}$$

According to (22), we can obtain

$$\begin{aligned} &\frac{\partial^2 f}{\partial \mathbf{c}_i \partial \mathbf{c}_j} = 0, \quad i, j \in \{1, \cdots U\},\ i \neq j, \\ &\frac{\partial^2\ f_1(\mathbf{c}_1)}{\partial^2 \mathbf{c}_1} \geqslant 0, \\ &\vdots \\ &\frac{\partial^2\ f_1(\mathbf{c}_1)}{\partial^2 \mathbf{c}_1} + \cdots + \frac{\partial^2\ f_1(\mathbf{c}_U)}{\partial^2 \mathbf{c}_U} \geqslant 0, \end{aligned} \tag{29}$$

and deduce that the Hessian matrix is positive definite. The result indicates that $f(\mathbf{c}_1, \cdots, \mathbf{c}_U)$ is a convex function, therefore, there is an optimal solution that satisfies $f^*(\mathbf{c}_1, \cdots, \mathbf{c}_U) = f_1^*(\mathbf{c}_1) + \cdots + f_U^*(\mathbf{c}_U)$, and the proof is completed.

Appendix C
Proof of Theorem 2

Substituting w_i into (9), we can obtain the linear representation among users, which are

$$\begin{aligned} (x_V, y_V, 0) &= (w_1 x_1 + \cdots + w_U x_U, w_1 y_1 + \cdots + w_U y_U, 0) \\ &= w_1(x_1, y_1, 0) + \cdots + w_U(x_U, y_U, 0). \end{aligned} \tag{30}$$

Since the Q-values with respect to the locations of the users, we can get

$$Q(\varphi, a_V; \xi) \propto Q(\varphi_1, a_i; \xi_1) + \cdots + Q(\varphi_U, a_U; \xi_U), \tag{31}$$

where $\varphi_i, \xi_i, \in \{1, \cdots, U\}$ is the DQN parameter of each user. According to Lemma 1, we have

$$Q^*(\varphi, a_V; \xi) \propto Q^*(\varphi_1, a_i; \xi_1) + \cdots + Q^*(\varphi_U, a_U; \xi_U). \quad (32)$$

Then, we can get

$$(x_V, y_V, 0)^* = w_1(x_1, y_1, 0)^* + \cdots + w_U(x_U, y_U, 0)^*, \quad (33)$$

which shows that all users have effectively learned the optimal communication trajectory to maximum its long-term cumulative reward, if and only if the virtual user obtains the optimal communication trajectory $\mathscr{L}_V^*$. Then proof is completed.

Appendix D
Proof of Theorem 3

As the leader, the UAV jammer first chooses the action $a_{\mathcal{J}}^t \in \mathcal{A}_{\mathcal{J}}$ to maximize its long-term cumulative reward in each time slot t. For any $a_{-\mathcal{J}} \in \mathcal{A}_{-\mathcal{J}}$, we have the following

$$R_{\mathcal{J}}[\mathscr{T}^*(a_{\mathcal{J}}^t), \mathscr{L}(a_V^{t-1})] \geq R_{\mathcal{J}}[\mathscr{T}(a_{-\mathcal{J}}^t), \mathscr{L}(a_V^{t-1})],$$

where $\mathcal{A}_{-\mathcal{J}}$ is the action space except the action $a_{\mathcal{J}}$. Then, as the follower, the virtual user observes the action of the leader and chooses the action $a_V^t \in \mathcal{A}_V$ to maximize its long-term cumulative reward $R_V[\mathscr{T}^*(a_{\mathcal{J}}^t), \mathscr{L}^*(a_V^t)]$. For any $a_{-V} \in \mathcal{A}_{-V}$, we have the following

$$R_V[\mathscr{T}^*(a_{\mathcal{J}}^t), \mathscr{L}^*(a_V^t)] \geq R_V[\mathscr{T}^*(a_{\mathcal{J}}^t), \mathscr{L}(a_{-V}^t)],$$

where $\mathcal{A}_{-V}$ is the action space except the action a_V. For any $a_{-\mathcal{J}} \in \mathcal{A}_{-\mathcal{J}}$ and $a_{-V} \in \mathcal{A}_{-V}$, we can obtain

$$\begin{aligned} R_{\mathcal{J}}[\mathscr{T}^*(a_{\mathcal{J}}^t), \mathscr{L}^*(a_V^t)] &\geq R_{\mathcal{J}}[\mathscr{T}(a_{-\mathcal{J}}^t), \mathscr{L}(a_V^t)], \\ R_V[\mathscr{T}^*(a_{\mathcal{J}}^t), \mathscr{L}^*(a_V^t)] &\geq R_V[\mathscr{T}(a_{\mathcal{J}}^t), \mathscr{L}(a_{-V}^t)]. \end{aligned} \quad (34)$$

Based on (24), the proof is completed.

Appendix E
Proof of Corollary 1

Substituting (3) into (10) and defining $\mathcal{K} + \mathcal{J} = p_{\mathcal{J}} P_{\mathrm{LoS}} \beta_{\mathrm{LoS}} + p_{\mathcal{J}} P_{\mathrm{NLoS}} \beta_{\mathrm{NLoS}}$, we can get immediate reward in (35), which is shown at the bottom of this page.

According to Lagrange multiplier

$$\mathscr{F}(x_{\mathcal{J}}, y_{\mathcal{J}}, z_{\mathcal{J}}, \lambda_{\mathcal{J}}) = r_{\mathcal{J}}[\mathscr{T}(a_{\mathcal{J}}), \mathscr{L}(a_V)] + \lambda_{\mathcal{J}}(|a_{\mathcal{J}}| - 1) \quad (35)$$

and sufficient Karush-Kuhn-Tucker (KKT) conditions,

$$\begin{aligned} \frac{\partial \mathscr{F}(x_{\mathcal{J}}, y_{\mathcal{J}}, z_{\mathcal{J}}, \lambda_{\mathcal{J}})}{\partial x_{\mathcal{J}}} &= 0 \\ \frac{\partial \mathscr{F}(x_{\mathcal{J}}, y_{\mathcal{J}}, z_{\mathcal{J}}, \lambda_{\mathcal{J}})}{\partial y_{\mathcal{J}}} &= 0 \\ \frac{\partial \mathscr{F}(x_{\mathcal{J}}, y_{\mathcal{J}}, z_{\mathcal{J}}, \lambda_{\mathcal{J}})}{\partial z_{\mathcal{J}}} &= 0 \\ \lambda_{\mathcal{J}}(|a_{\mathcal{J}}| - 1) &= 0 \\ \lambda_{\mathcal{J}} &\geq 0, \end{aligned} \quad (36)$$

we obtain

$$\mathscr{T}^*(a_{\mathcal{J}}) = \left(\frac{x_{\mathcal{J}0} - x_{V0} + x_{V0} z_{\mathcal{J}0}}{z_{\mathcal{J}0}}, \frac{y_{\mathcal{J}0} - y_{V0} + y_{V0} z_{\mathcal{J}0}}{z_{\mathcal{J}0}}, 1\right).$$

Defining

$$\begin{aligned} &(x_{\mathcal{J}}^*, y_{\mathcal{J}}^*, z_{\mathcal{J}}^*) \\ &= \left(\frac{x_{\mathcal{J}0} - x_{V0} + x_{V0} z_{\mathcal{J}0}}{z_{\mathcal{J}0}}, \frac{y_{\mathcal{J}0} - y_{V0} + y_{V0} z_{\mathcal{J}0}}{z_{\mathcal{J}0}}, 1\right) \end{aligned} \quad (37)$$

and substituting (3) into (5), we can get immediate reward in (39), which is presented at the bottom of this page.

Similarly, if the initial location of the UAV jammer and the virtual user satisfies $x_{\mathcal{J}0} = y_{\mathcal{J}0}$ and $x_{V0} = y_{V0}$, using Lagrange multiplier and KKT conditions,

$$\mathscr{F}(x_V, y_V, 0, \lambda_V) = r_V[\mathscr{T}^*(a_{\mathcal{J}}), \mathscr{L}(a_V)] + \lambda_V(|a_V| - 1) \quad (40)$$

$$\begin{aligned} \frac{\partial \mathscr{F}(x_V, y_V, 0, \lambda_V)}{\partial x_V} &= 0 \\ \frac{\partial \mathscr{F}(x_V, y_V, 0, \lambda_V)}{\partial y_V} &= 0 \\ \lambda_V(|a_V| - 1) &= 0 \\ \lambda_V &\geq 0, \end{aligned} \quad (41)$$

we have $x_V^* = y_V^*$. Then, we derive that $\mathscr{L}^*(a_V) = (1, 1, 0)$ is one of the optimal solution for the virtual user in this special case.

$$r_{\mathcal{J}}[\mathscr{T}(a_{\mathcal{J}}), \mathscr{L}(a_V)] = \frac{(\mathcal{K}+\mathcal{J})\left(\sqrt{(x_{\mathcal{J}} - x_{V0})^2 + (y_{\mathcal{J}} - y_{V0})^2 + z_{\mathcal{J}}^2}\right)^{-\alpha}}{p_b\left(\left(\sqrt{x_{V0}^2 + y_{V0}^2 + H_{\mathcal{B}}^2}\right)^{-\eta}|\tilde{h}_{\mathcal{B}V}|^2 + \sigma^2\right)} - C_{\mathcal{J}}\sqrt{(x_{\mathcal{J}} - x_{\mathcal{J}0})^2 + (y_{\mathcal{J}} - y_{\mathcal{J}0})^2 + (z_{\mathcal{J}} - z_{\mathcal{J}0})^2} \quad (35)$$

$$r_V[\mathscr{T}^*(a_{\mathcal{J}}), \mathscr{L}(a_V)] = \frac{p_{\mathcal{B}}\left(\sqrt{x_V^2 + y_V^2 + H_{\mathcal{B}}^2}\right)^{-\eta}|\tilde{h}_{\mathcal{B}V}|^2}{(\mathcal{K}+\mathcal{J})\left(\sqrt{(x_V - x_{\mathcal{J}}^*)^2 + (y_V - y_{\mathcal{J}}^*)^2 + z_{\mathcal{J}}^{*2}}\right)^{-\alpha}} - C_V\sqrt{(x_V - x_{V0})^2 + (y_V - y_{V0})^2} \quad (39)$$

References

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

[2] Y. Liu, Z. Qin, Y. Cai, Y. Gao, G. Y. Li, and A. Nallanathan, "UAV communications based on non-orthogonal multiple access," *IEEE Wireless Commun.*, vol. 26, no. 1, pp. 52–57, Feb. 2019.

[3] J. Xu, Y. Zeng, and R. Zhang, "UAV-enabled wireless power transfer: Trajectory design and energy optimization," *IEEE Trans. Wireless Commun.*, vol. 17, no. 8, pp. 5092–5106, Aug. 2018.

[4] S. Zhang, Y. Zeng, and R. Zhang, "Cellular-enabled UAV communication: A connectivity-constrained trajectory optimization perspective," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2580–2604, Mar. 2019.

[5] S. A. R. Naqvi, S. A. Hassan, H. Pervaiz, and Q. Ni, "Drone-aided communication as a key enabler for 5G and resilient public safety networks," *IEEE Commun. Mag.*, vol. 56, no. 1, pp. 36–42, Jan. 2018.

[6] Y. Zhou *et al.*, "Improving physical layer security via a UAV friendly jammer for unknown eavesdropper location," *IEEE Trans. Veh. Technol.*, vol. 67, no. 11, pp. 11280–11284, Nov. 2018.

[7] Y. Cai, F. Cui, Q. Shi, M. Zhao, and G. Y. Li, "Dual-UAV-enabled secure communications: Joint trajectory design and user scheduling," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 9, pp. 1972–1985, Sep. 2018.

[8] D. He, S. Chan, and M. Guizani, "Communication security of unmanned aerial vehicles," *IEEE Wireless Commun.*, vol. 24, no. 4, pp. 134–139, Aug. 2017.

[9] U. Challita, A. Ferdowsi, M. Chen, and W. Saad, "Machine learning for wireless connectivity and security of cellular-connected UAVs," *IEEE Wireless Commun.*, vol. 26, no. 1, pp. 28–35, Feb. 2019.

[10] Q. Wang, Z. Chen, W. Mei, and J. Fang, "Improving physical layer security using UAV-enabled mobile relaying," *IEEE Wireless Commun. Lett.*, vol. 6, no. 3, pp. 310–313, Jun. 2017.

[11] L. Xiao, X. Lu, D. Xu, Y. Tang, L. Wang, and W. Zhuang, "UAV relay in VANETs against smart jamming with reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4087–4097, May 2018.

[12] T. Humphreys, "Statement on the security threat posed by unmanned aerial systems and possible countermeasures," Oversight Manage. Efficiency Subcommittee, Homeland Secur. Committee, Washington, DC, USA, Tech. Rep., 2015.

[13] M. Min, L. Xiao, D. Xu, L. Huang, and M. Peng, "Learning-based defense against malicious unmanned aerial vehicles," in *Proc. IEEE VTC-Spring*, Jun. 2018, pp. 1–5.

[14] D. Wang, P. Ren, J. Cheng, and Y. Wang, "Achieving full secrecy rate with energy-efficient transmission control," *IEEE Trans. Commun.*, vol. 65, no. 12, pp. 5386–5400, Dec. 2017.

[15] D. Wang, P. Ren, Q. Du, L. Sun, and Y. Wang, "Security provisioning for MISO vehicular relay networks via cooperative jamming and signal superposition," *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 10732–10747, Dec. 2017.

[16] S. Bhattacharya and T. Başar, "Game-theoretic analysis of an aerial jamming attack on a UAV communication network," in *Proc. Amer. Control Conf.*, Jun./Jul. 2010, pp. 818–823.

[17] L. Xiao, C. Xie, M. Min, and W. Zhuang, "User-centric view of unmanned aerial vehicle transmission against smart attacks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3420–3430, Apr. 2018.

[18] Y. Xu *et al.*, "A one-leader multi-follower Bayesian-Stackelberg game for anti-jamming transmission in UAV communication networks," *IEEE Access*, vol. 6, pp. 21697–21709, 2018.

[19] C. Li, Y. Xu, J. Xia, and J. Zhao, "Protecting secure communication under UAV smart attack with imperfect channel estimation," *IEEE Access*, vol. 6, pp. 76395–76401, 2018.

[20] Y. E. Sagduyu, R. A. Berry, and A. Ephremides, "Jamming games in wireless networks with incomplete information," *IEEE Commun. Mag.*, vol. 49, no. 8, pp. 112–118, Aug. 2011.

[21] L. Xiao, J. Liu, Y. Li, N. B. Mandayam, and H. V. Poor, "Prospect theoretic analysis of anti-jamming communications in cognitive radio networks," in *Proc. IEEE Globecom*, Dec. 2014, pp. 746–751.

[22] L. Jia, F. Yao, Y. Sun, Y. Niu, and Y. Zhu, "Bayesian Stackelberg game for antijamming transmission with incomplete information," *IEEE Commun. Lett.*, vol. 20, no. 10, pp. 1991–1994, Oct. 2016.

[23] L. Jia, F. Yao, Y. Sun, Y. Xu, S. Feng, and A. Anpalagan, "A hierarchical learning solution for anti-jamming Stackelberg game with discrete power strategies," *IEEE Wireless Commun. Lett.*, vol. 6, no. 6, pp. 818–821, Dec. 2017.

[24] L. Xiao, T. Chen, J. Liu, and H. Dai, "Anti-jamming transmission Stackelberg game with observation errors," *IEEE Commun. Lett.*, vol. 19, no. 6, pp. 949–952, Jun. 2015.

[25] Z. Qin, H. Ye, G. Y. Li, and B.-H. F. Juang, "Deep learning in physical layer communications," *IEEE Wireless Commun.*, vol. 26, no. 2, pp. 93–99, Apr. 2019.

[26] X. Gao, S. Jin, C.-K. Wen, and G. Y. Li, "ComNet: Combination of deep learning and expert knowledge in OFDM receivers," *IEEE Commun. Lett.*, vol. 22, no. 12, pp. 2627–2630, Dec. 2018.

[27] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–533, Feb. 2015.

[28] M. Hausknecht and P. Stone, "Deep recurrent Q-learning for partially observable MDPs," in *Proc. AAAI-SDMIA*, Nov. 2015, pp. 1–9.

[29] L. Xiao, D. Jiang, D. Xu, H. Zhu, Y. Zhang, and H. V. Poor, "Two-dimensional antijamming mobile communication based on reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9499–9512, Oct. 2018.

[30] A. Mpitziopoulos, D. Gavalas, C. Konstantopoulos, and G. Pantziou, "A survey on jamming attacks and countermeasures in WSNs," *IEEE Commun. Surveys Tuts.*, vol. 11, no. 4, pp. 42–56, 4th Quart., 2009.

[31] E. Altman, K. Avrachenkov, and A. Garnaev, "Jamming in wireless networks under uncertainty," *Mobile Netw. Appl.*, vol. 16, no. 2, pp. 246–254, Apr. 2011.

[32] A. Al-Hourani, S. Kandeepan, and A. Jamalipour, "Modeling air-to-ground path loss for low altitude platforms in urban environments," in *Proc. IEEE Globecom*, Dec. 2014, pp. 2898–2904.

[33] Q. Feng, J. McGeehan, E. K. Tameh, and A. R. Nix, "Path loss models for air-to-ground radio channels in urban environments," in *Proc. IEEE VTC-Spring*, May 2006, pp. 2901–2905.

[34] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal LAP altitude for maximum coverage," *IEEE Wireless Commun. Lett.*, vol. 3, no. 6, pp. 569–572, Dec. 2014.

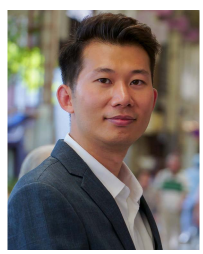
[35] H. S. Wang and N. Moayeri, "Finite-state Markov channel—A useful model for radio communication channels," *IEEE Trans. Veh. Technol.*, vol. 44, no. 1, pp. 163–171, Feb. 1995.

[36] N. Meuleau, L. Peshkin, K.-E. Kim, and L. P. Kaelbling, "Learning finite-state controllers for partially observable environments," in *Proc. Conf. Uncertainty Artif. Intell.* San Francisco, CA, USA: Morgan Kaufmann Publishers, 1999, pp. 427–436.

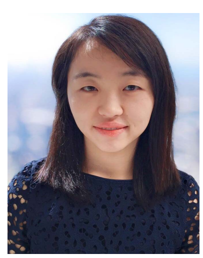
[37] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, 1992.

[38] Z. Han, D. Niyato, W. Saad, T. Baar, and A. Hjrungnes, *Game Theory in Wireless and Communication Networks*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[39] K. He and J. Sun, "Convolutional neural networks at constrained time cost," in *Proc. IEEE CVPR*, Jun. 2015, pp. 5353–5360.



Ning Gao received the Ph.D. degree in information and communications engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2019. From 2017 to 2018, he was a Visiting Ph.D. Student with the School of Computing and Communications, Lancaster University, Lancaster, U.K. He is currently a Research Fellow with the National Mobile Communications Research Laboratory, Southeast University. His research interests include wireless eavesdropping and spoofing, intelligent communications, and UAV communications.



Zhijin Qin (S'13–M'16) received the B.S. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2012, and the Ph.D. degree in electronic engineering from the Queen Mary University of London (QMUL), London, U.K., in 2016. She was a Research Associate with Imperial College London from 2016 to 2017 and a Lecturer with Lancaster University from 2017 to 2018. Since 2018, she has been a Lecturer with the School of Electronic Engineering and Computer Science, QMUL. She has also been an Honorary Research Fellow with Imperial College London since 2018. Her research interests include machine learning and compressive sensing in wireless signal processing and IoT networks. She is an Associate Editor for IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE COMMUNICATIONS LETTERS, and IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING. She was a recipient of the 2017 IEEE GLOBECOM Best Paper Award and the 2018 IEEE Signal Processing Society Young Author Best Paper Award.

Xiaojun Jing received the B.S. degree from Beijing Normal University and the M.S. and Ph.D. degrees from the National University of Defense Technology in 1999 and 1995, respectively. From 2000 to 2002, he was a Post-Doctoral Researcher with the Beijing University of Posts and Telecommunications (BUPT), where he is currently a Full Professor with the School of Information and Communication Engineering. His research interests include information security and fusion, wireless communication, and image processing.

Qiang Ni (M'04–SM'08) received the B.Sc., M.Sc., and Ph.D. degrees from the Huazhong University of Science and Technology, China, all in engineering. He is currently a Professor and the Head of the Communication Systems Group, School of Computing and Communications, Lancaster University, Lancaster, U.K. His research interests include the area of future generation communications and networking, including green communications and networking, millimeter-wave wireless communications, cognitive radio network systems, non-orthogonal multiple access (NOMA), heterogeneous networks, 5G and 6G, SDN, cloud networks, energy harvesting, wireless information and power transfer, IoTs, cyber physical systems, AI and machine learning, big data analytics, and vehicular networks. He has authored or coauthored over 200 articles in these areas. He was an IEEE 802.11 Wireless Standard Working Group Voting Member and a contributor to the IEEE Wireless Standards.

Shi Jin (S'06–M'07–SM'17) received the B.S. degree in communications engineering from the Guilin University of Electronic Technology, Guilin, China, in 1996, the M.S. degree from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2003, and the Ph.D. degree in communications and information systems from Southeast University, Nanjing, China, in 2007.

From June 2007 to October 2009, he was a Research Fellow with the Adastral Park Research Campus, University College London, London, U.K. He is currently with the Faculty of the National Mobile Communications Research Laboratory, Southeast University. His research interests include space time wireless communications, random matrix theory, and information theory.

Dr. Jin and his coauthors received the 2011 IEEE Communications Society Stephen O. Rice Prize Paper Award in the field of communication theory and the 2010 Young Author Best Paper Award by the IEEE Signal Processing Society. He serves as an Associate Editor for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE COMMUNICATIONS LETTERS, and *IET Communications*.