Anti-Intelligent UAV Jamming Strategy via Deep Q-Networks

Ning Gao*‡ Zhijin Qin† Xiaojun Jing* Qiang Ni‡
* Beijing University of Posts and Telecommunications, Beijing 100876, China.
† Queen Mary University of London, London E1 4NS, U.K.
‡ Lancaster University, Lancaster LA1 4WA, U.K.
Emails: {ngao, jxiaojun}@bupt.edu.cn, z.qin@qmul.ac.uk, q.ni@lancaster.ac.uk

***Abstract*—The downlink communications are vulnerable to intelligent unmanned aerial vehicle (UAV) jamming attack which can learn the optimal attack strategy in complex communication environments. In this paper, we propose an anti-intelligent UAV jamming strategy, in which the mobile users can learn the optimal defense strategy to prevent the jamming. Specifically, the UAV jammer acts as a leader and the ground users act as followers. The problem is formulated as a stackelberg dynamic game, which includes the leader sub-game and the followers sub-game. As the UAV jammer is only aware of the incomplete channel state information (CSI) of the ground users, we model the leader sub-game as a partially observable Markov decision process (POMDP). The optimal jamming trajectory is obtained via deep recurrent Q-networks (DRQN) in the three-dimension space. For the followers sub-game, we use the Markov decision process (MDP) to model it. Then the optimal communication trajectory can be learned via deep Q-networks (DQN) in the two-dimension space. We prove the existence of the stackelberg equilibrium. The simulations show that the proposed strategy outperforms the benchmark strategies.**

***Index Terms*—Intelligent UAV jamming, game theory, MDP, deep Q-networks.**

I. Introduction

With the demand of high-speed data transmission in wireless communications, various communication technologies have been explored to improve the network capacity. Compared to the conventional technologies, the UAV can provide strong line-of-sight (LoS) links and small path-loss exponents over the air-to-ground communications, which has natural advantages in boosting the network capacity [1]–[3].

UAVs can be exploited as malicious components when considering the communication security issues [3]–[5]. Due to the strong LoS links and small path-loss exponents, the UAV jamming attack can significantly block the data transmission and degrade communication quality of service (QoS), which is more serious than the ground jamming. Therefore, anti-UAV jamming problem in wireless communication is worth investigating. Recently, some meaningful work has been devoted to addressing the malicious UAV jamming problem [6]–[8]. For example, a zero-sum pursuit-evasion game has been formulated to compute optimal strategies to evade the aerial jammer [6]. A smart UAV attacker, who can specify the attack type, such as jamming, eavesdropping and spoofing, has been considered in [7].

However, the above anti-UAV jamming work are based on some ideal assumptions, i.e., the perfect observations. More recent work has considered imperfect observation in anti-ground jamming but few in anti-UAV jamming [8], [9]. For example, with considering the incomplete information constraint and the co-channel mutual interference, the competition between UAVs and jammers have been investigated [8]. Likewise, the impact of observation error of a smart jammer has been evaluated in [9]. As aforementioned, only [8] has considered imperfect observations in anti-UAV jamming problem. Meanwhile, most of the considered UAV jammers are unintelligent [7], [8], [10]. Limited work has considered intelligent UAV jamming, which can easily learn the jamming strategy in complex communication environments, even with the incomplete observation, i.e., incomplete channel state information (CSI). Therefore, investigating the anti-intelligent UAV jamming problem becomes more challenging.

Motivated by the above practical considerations, in this paper, we consider the scenario that both the UAV jammer and the users are intelligent agents. On the one hand, the UAV jammer learns the optimal attack strategy. On the other hand, the ground users learn the optimal defense strategy. To the best of our knowledge, *"How do ground users defend against intelligent UAV jamming using machine learning?"* is still an open problem. Some specific contributions of our work are summarized as follows

- For the first attampt, 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 can defend against this jamming to the greatest extent.
- We propose an anti-intelligent UAV jamming strategy. Specifically, the problem is formulated as a stackelberg dynamic game. The incomplete CSI is considered in the game and the optimal trajectories are learned.
- Some important remarks and insights are obtained from theory and simulations.

The rest of the paper is organized as follows. In Section II, we provide the system model and problem formulation. In Section III, we analyze the proposed game and prove the existence of stackelberg equilibrium. Simulations are presented in Section IV and summaries are given in Section V.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. System Model

We consider the downlink communications between a base station and ground users under the threat of a UAV jammer. 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 in each time slot at constant velocities. 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 adopts a full band barrage jamming. The UAV jammer and the users are considered as intelligent agents, who can learn the optimal strategies to maximize their long-term cumulative rewards, i.e., signal-to-interference-plus-noise ratio (SINR) [11], respectively. Denote $\mathcal{J}$ as the UAV jammer, $\mathcal{B}$ as the base station and $i \in \{1, \cdots, U\}$ as user i. 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 $(x_{\mathcal{J}}, y_{\mathcal{J}}, z_{\mathcal{J}})$. 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 stay, up, down, left, right, forward, backword. Denote the mapping of user action space as

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

which represents stay, left, right, forward, backword.

The channel coefficient from base station to the i-th user 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. Meanwhile, the channel between the UAV jammer and the user i is modeled as an air-to-ground channel model which contains strong LoS, reflected nonline-of-sight (NLoS), and small-scale fading. In general, the probability of having small-scale fading is extremely lower than LoS and NLoS, hence, the small-scale fading is neglected [12]. The path loss between the UAV jammer and user i is denoted as [12]

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

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 the UAV jammer $\mathcal{J}$ and the 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 height of the UAV jammer, which can be represented as

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

Particularly, Φ and Ψ are S-curve parameters, which depend on communication environment, $\theta_i = \frac{180}{\pi} \arcsin(\frac{z_{\mathcal{J}}}{d_{\mathcal{J}i}})$ and the probability of NLoS is $P_{\text{NLoS}} = 1 - P_{\text{LoS}}$. The expectation of the jamming power received at user i is given by [12]

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

where $p_{\mathcal{J}}$ is the power budget of the UAV jammer. Then, the received SINR at the ground 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

We quantize the channel $h_{\mathcal{B}i}$ into a finite state space $\mathcal{S} = \{h_{\mathcal{B}i}^1, \cdots, h_{\mathcal{B}i}^K\}$, and model it as a Markov chain with finite states [13]. In practice, due to the mobility of the UAV jammer, the wireless channel environment is dynamic and unknown. Therefore, the UAV jammer can only obtain the partially observable information which is the location of the users with respect to the distances to the base station, $d_{\mathcal{B}i} = \sqrt{x_i^2 + y_i^2 + H_{\mathcal{B}}^2}, i \in \{1, \cdots, U\}$. Meanwhile, the information observed by the users 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 the action $a_{\mathcal{J}} \in \mathcal{A}_{\mathcal{J}}$ to determine the flying direction, then each user chooses the action $a_i \in \mathcal{A}_i$ to determine the moving direction. 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., $(x_{\mathcal{J}}', y_{\mathcal{J}}', z_{\mathcal{J}}') = (x_{\mathcal{J}}, y_{\mathcal{J}}, z_{\mathcal{J}}) + a_{\mathcal{J}}$. The strategies of the UAV jammer and the users refer to jamming trajectory and communication trajectories, respectively.

The reward of the 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 the 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 the 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

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

station to the user i is $d_{\mathcal{B}i} = \sqrt{x_i'^2 + y_i'^2 + H_{\mathcal{B}}^2}$ 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 reward can be given by

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

where $C_{\mathcal{J}}$ is the unit energy cost of the UAV jammer, i.e., flying and jamming cost per unit distance, and $d_{\mathcal{J}} = \sqrt{(x_{\mathcal{J}}' - x_{\mathcal{J}})^2 + (y_{\mathcal{J}}' - y_{\mathcal{J}})^2 + (z_{\mathcal{J}}' - z_{\mathcal{J}})^2}$ is the flying distance per time slot. The formulated problem is to maximize each user's reward in (5) to find the optimal communication trajectory, within the constraint of the optimal jamming trajectory of the UAV obtained via maximize reward in (6).

III. DQN based Anti-jamming Elude Game

A. The Optimal Jamming Trajectory

We quantize the flying space into L states, then the location 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 space into M states, which is denoted as $\mathcal{S}_i = \{(x_{i,1}, y_{i,1}, 0), \cdots, (x_{i,M}, y_{i,M}, 0\}, i \in \{1, \cdots, U\}$. To simplify the case, we model a virtual user V as a target user, which is a virtual point related to the ground users. The location of the virtual user can be decided by

$$(x_V, y_V, 0) = \left(\frac{\sum_{i=1}^{U} w_i x_i}{\sum_{i=1}^{U} w_i}, \frac{\sum_{i=1}^{U} w_i y_i}{\sum_{i=1}^{U} w_i}, 0 \right), \quad (7)$$

where $w_i = \frac{B_i}{B}$ is the location weight of user i. Then, we denote the motion space of the virtual user as $\mathcal{S}_V$.

Remark. *Since the communication fairness among users, the base station will allocate more bandwidth to the user far away from it. Thus, the value of weights w_i is proportion to the bandwidth allocated to the user i. To simplify analysis, we assume that the bandwidth is assigned the same for each user. Thus, the location weight of the user i is $w_i = \frac{1}{U}$.*

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

$$r_{\mathcal{J}}[\mathscr{T}(a_{\mathcal{J}}), \mathscr{L}(a_V)] = \frac{I_{\mathcal{J}V}}{p_{\mathcal{B}} d_{\mathcal{B}V}^{-\eta} |\tilde{h}_{\mathcal{B}V}|^2 + \sigma^2} - C_{\mathcal{J}} d_{\mathcal{J}}, \quad (8)$$

where $d_{\mathcal{B}V} = \sqrt{x_V'^2 + y_V'^2 + H_{\mathcal{B}}^2}$. Then the optimization problem for the UAV jammer $\mathcal{J}$ is formulated as choosing action $a_{\mathcal{J}}$ to maximize reward (8) under the constraint of unit moving distance per time slot, which can be given by

$$\max_{a_{\mathcal{J}}} \; r_{\mathcal{J}}[\mathscr{T}(a_{\mathcal{J}}), \mathscr{L}(a_V)] \quad \text{s.t.} \;\; |a_{\mathcal{J}}| = 1. \quad (9)$$

However, the complete CSI of the virtual user is not known to the UAV jammer. Considering the dynamic channel environment, we model this process as a partially observable Markov decision process (POMDP) [14]. Define a POMDP as a 6-tuple $\langle \mathcal{S}, \mathcal{A}_{\mathcal{J}}, P, r_{\mathcal{J}}, O, \Omega \rangle$, where

- $\mathcal{S}$ is the 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 choosing in state $s \in \mathcal{S}$;
- $r_{\mathcal{J}}[s, \mathscr{T}(a_{\mathcal{J}})]$ is the 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 space, which is equal to the motion 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

$$b'(s') = \frac{1}{\Theta} \left[\Omega(o'|s', a_{\mathcal{J}}) \sum_{s \in \mathcal{S}} P(s'|s, a_{\mathcal{J}}) b(s) \right], \quad (10)$$

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 $\pi : b \to a_{\mathcal{J}}$. Then solving the POMDP is to find the optimal action selection policy $\pi^* : b^* \to a_{\mathcal{J}}^*$, yields the maximum expected reward for each belief. This maximum expected reward can be obtained by the Bellman equation

$$V_b^* = \max_{a_{\mathcal{J}} \in \mathcal{A}_{\mathcal{J}}} \left[r_{\mathcal{J},b} + \gamma \sum_{o \in O} \Omega(o|b, a_{\mathcal{J}}) V_{b'}^* \right], \quad (11)$$

where $r_{\mathcal{J},b} = \sum_{s \in \mathcal{S}} r_{\mathcal{J},s} b(s)$ represents the expected reward over the belief distribution, and γ is the discount factor.

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 space $\mathcal{M}$ as a new state space instead of the original state space $\mathcal{S}$ [15]. The near-optimal solution to the belief-MDP can be solved by Q-learning [16]. 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 strategy, which directly exploits the sequence of ℓ historical observation-action pairs, $O^t = \{o^{t-\ell}, a_{\mathcal{J}}^{t-\ell}, \cdots, o^{t-1}, a_{\mathcal{J}}^{t-1}\}$ to learn the optimal attack strategy [14]. The deep recurrent Q-networks (DRQN) that combines Q-learning with a recurrent convolutional neural network (CNN), is developed. The framework is shown in Fig 1. 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 θ is the weight parameter set of the Q-network. In time slot t, sequence 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 $e_{\mathcal{J}}^t(\phi^t, a_{\mathcal{J}}^t, r_{\mathcal{J}}^t, \phi^{t+1})$ is

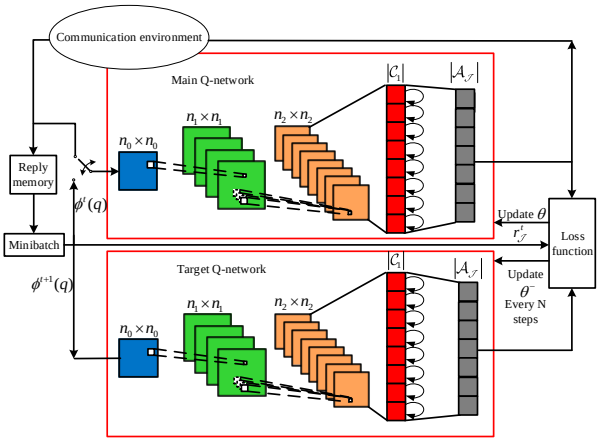

Fig. 1. The developed DRQN framework.

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 by

$$L(\theta) = \mathbb{E}_{\phi,a,r,\phi'}\big[\big(r_{\mathcal{J},\phi} + \gamma \max_{a'_{\mathcal{J}}} Q(\phi', a'_{\mathcal{J}}; \theta^-) - Q(\phi, a_{\mathcal{J}}; \theta)\big)^2\big], \quad (12)$$

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

$$\nabla_\theta L(\theta) = \mathbb{E}_{\phi,a,r,\phi'}\big[\big(r_{\mathcal{J},\phi} + \gamma \max_{a'_{\mathcal{J}}} Q(\phi', a'_{\mathcal{J}}; \theta^-) - Q(\phi, a_{\mathcal{J}}; \theta)\big)\nabla_\theta Q(\phi, a_{\mathcal{J}}; \theta)\big]. \quad (13)$$

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 can be denoted by

$$\mathscr{T}^*(a_{\mathcal{J}}) = (x_{\mathcal{J}0}, y_{\mathcal{J}0}, z_{\mathcal{J}0}) + {a^0_{\mathcal{J}}}^* + {a^1_{\mathcal{J}}}^* + \cdots + {a^t_{\mathcal{J}}}^*, \quad (14)$$

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 optimal action a^*_V based on the observation of the UAV jammer. The optimal communication trajectory $\mathscr{L}^*(a_V)$ can be formulated as

$$\max_{a_V} \; r_V[\mathscr{T}(a_{\mathcal{J}}), \mathscr{L}(a_V)] \quad \text{s.t.} \;\; |a_V| = 1. \quad (15)$$

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 the Theorem 1, optimizing defense strategy problem can be modeled as solving a MDP problem, in which the communication trajectory of the virtual user is determined by the state $\mathcal{S}_{\mathcal{J}}$, 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 state space,
- $\mathcal{A}_V$ is the action space,
- $r_V[s_{\mathcal{J}}, \mathscr{L}(a_V)]$ is the 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

$$\begin{aligned} & P(s^{t+1}_{\mathcal{J}}|s^t_{\mathcal{J}}, a_V) = P(s^t_{\mathcal{J}} + a^{t+1}_{\mathcal{J}}|s^t_{\mathcal{J}}, a_V) \\ = \;& P(a^0_{\mathcal{J}} + \cdots + a^{t+1}_{\mathcal{J}}|a^0_{\mathcal{J}} + \cdots + a^t_{\mathcal{J}}, a_V) \\ = \;& P(a^{t+1}_{\mathcal{J}}|a^t_{\mathcal{J}}, a_V). \end{aligned} \quad (16)$$

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

Considering the state space $\mathcal{S}_{\mathcal{J}}$ is large, we develop the DQN, which is shown in Fig. 2. The Q-value with parameter ξ is estimated by the DQN, which is denoted by $Q(s_{\mathcal{J}}, a_V; \xi)$. Specifically, in time slot t, the sequence of ℓ historical state-action pairs $S^t = \{s^{t-\ell}_{\mathcal{J}}, a^{t-\ell}_V, \cdots, s^{t-1}_{\mathcal{J}}, a^{t-1}_V\}$ 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 $\mathcal{D}_V = \{e^1_V, \cdots, e^t_V\}$. The weight parameter set ξ is updated via the loss function

$$L(\xi) = \mathbb{E}_{\varphi,a,r,\varphi'}\big[\big(r_{V,\varphi} + \gamma \max_{a'_V} Q(\varphi', a'_V; \xi^-) - Q(\varphi, a_V; \xi)\big)^2\big], \quad (17)$$

where the symbol ξ^- is updated from the same Q-network to minimize the loss function in every N steps. The gradient of (17) refers to the weight parameter set ξ is obtained by

$$\nabla_\xi L(\xi) = \mathbb{E}_{\varphi,a,r,\varphi'}\big[\big(r_{V,\varphi} + \gamma \max_{a'_V} Q(\varphi', a'_V; \xi^-) - Q(\varphi, a_V; \xi)\big)\nabla_\xi Q(\varphi, a_V; \xi)\big]. \quad (18)$$

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

$$\mathscr{L}^*(a^t_V) = (x_{V0}, y_{V0}, 0) + {a^0_V}^* + {a^1_V}^* + \cdots + {a^t_V}^*, \quad (19)$$

where $(x_{V0}, y_{V0}, 0)$ is the initial location of the virtual user.

C. 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*

$$\begin{aligned} & r_1(a^*_1, a_2) \ge r_1(a_1, a_2) \\ & r_2(a^*_1, a^*_2) \ge r_2(a^*_1, a_2(a^*_1)), \end{aligned} \quad (20)$$

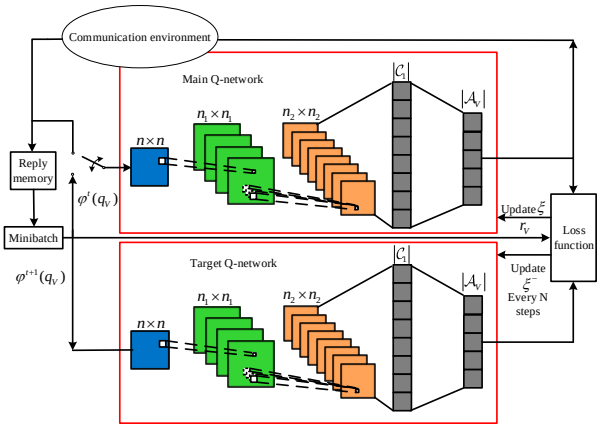

Fig. 2. The developed DQN framework.

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

Remark. *The stackelberg equilibrium with the UAV jammer as leader is the optimal solution for it if the UAV jammer chooses the 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 non-optimal action* $\tilde{a}_V^*$ *to maximize reward* r_V*. In this case, the reward of the UAV jammer will be less than that when the stackelberg equilibrium is used.*

Theorem 2. *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. *The stackelberg equilibrium can be achieved with probability one in each time slot, if the DQN is well trained or via a full greedy strategy. However, to balance the exploration and exploitation with respect to a large state-action space, the stackelberg equilibrium is achieved with probability* $1 - \epsilon$*. In other word, it has a probability* ϵ *that the system cannot obtain the optimal communication trajectory in DQN training.*

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 $p_b = 100$ 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 mobile cost of the virtual user is $C_U = 0.5$ dB≈ 1.12 mW. From [12], 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 location of the virtual user is calculated by (7). 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 $\epsilon = 0.1$.

The long-term cumulative reward of the UAV jammer in 300 time slots is presented in Fig. 3. Specifically, we leverage the greedy strategy and random strategy as benchmark strategies and compare them with the proposed DRQN based intelligent attack strategy. We find that the jamming reward via DRQN can converge to 21.2 dB after 200 time slots. The performance of the proposed attack strategy is already superior to the greedy strategy and random strategy after 25 time slots, for example, the proposed attack strategy can achieve 75% higher reward than greedy reward in the 200-th time slot. We can also find that greedy strategy can achieve a better performance than random strategy.

We present the long-term cumulative reward of the virtual user in Fig. 4. The result suggest that the reward via DQN can converge to 22.3 dB after 100 time slots. After 10 time slots, the DQN based strategy is already get a higher reward than the other two strategies. In summary, these two figures show that both the UAV jammer and the virtual user can obtain the highest long-term cumulative rewards via DRQN and DQN, respectively. Thus, the stackelberg equilibrium exists.

Fig. 5 presents the optimal jamming trajectory and the optimal communication trajectory in one episode. We observe that the communication location of the virtual user starts at (-2 m, 1 m) and ends at (15 m, 18 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). Intuitively, the UAV will always stay close to the virtual user to maximize its reward, but it is not in practice. The reason is that the CSI is dynamic changeable 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 problem as a stackelberg dynamic game, in which the UAV jammer acts as a leader and the users act as followers. With the incomplete channel state information, we have modeled the leader sub-game as a partially observable Markov decision process and have developed the recurrent convolutional neural network to learn the optimal jamming trajectory in the three-dimension space. For the follower, we have modeled the follower sub-game as a Markov decision process. Then, the optimal communication trajectory has been learned in the two-dimension space. Moreover, some remarks and insights have been obtained from theory and simulations.

Appendix A
Proof of Theorem 1

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

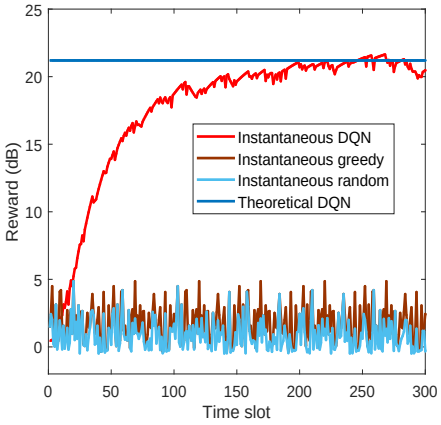

Fig. 3. The rewards of the UAV jammer in DRQN,greedy and random strategies in 300 time slots.

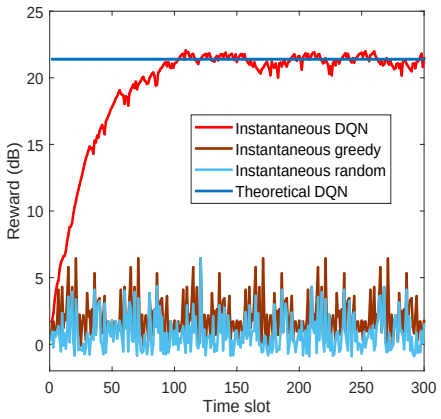

Fig. 4. The rewards of the virtual user in DQN, greedy and random strategies in 300 time slots.

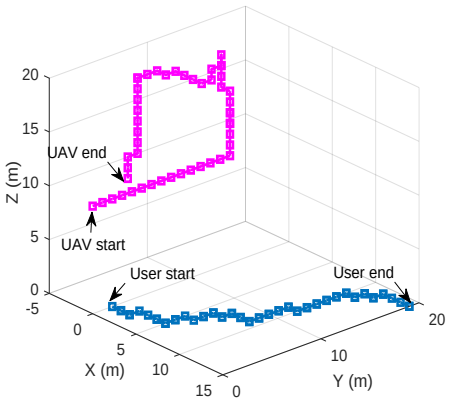

Fig. 5. The optimal trajectories via learning in one episode, the UAV jammer vs. the virtual user.

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

$$\begin{aligned} P({a'_{\mathcal{J}}}^*|a_{\mathcal{J}}) &= P(o', {a'_{\mathcal{J}}}^*|o, a_{\mathcal{J}}) \\ &= P({a'_{\mathcal{J}}}^*)P(o'|o, a_{\mathcal{J}}) \\ &= (1-\epsilon)P(o'|o, a_{\mathcal{J}}), \end{aligned} \tag{21}$$

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

$$\begin{aligned} P({\tilde{a'}_{\mathcal{J}}}^*|a_{\mathcal{J}}) &= P(o', {\tilde{a'}_{\mathcal{J}}}^*|o, a_{\mathcal{J}}) \\ &= P({\tilde{a'}_{\mathcal{J}}}^*)P(o'|o, a_{\mathcal{J}}) \\ &= \epsilon P(o'|o, a_{\mathcal{J}}). \end{aligned} \tag{22}$$

As per (21) (22), 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, the proof is completed.

Appendix B
Proof of Theorem 2

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

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

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 $a^t_V \in \mathcal{A}_V$ to maximize its instantaneous reward $r_V[\mathscr{T}^*(a^t_{\mathcal{J}}), \mathscr{L}^*(a^t_V)]$. For any $a_{-V} \in \mathcal{A}_{-V}$, we have the following

$$r_V[\mathscr{T}^*(a^t_{\mathcal{J}}), \mathscr{L}^*(a^t_V)] \geq r_V[\mathscr{T}^*(a^t_{\mathcal{J}}), \mathscr{L}(a^t_{-V})],$$

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

$$\begin{aligned} &r_{\mathcal{J}}[\mathscr{T}^*(a^t_{\mathcal{J}}), \mathscr{L}^*(a^t_V)] \geq r_{\mathcal{J}}[\mathscr{T}(a^t_{-\mathcal{J}}), \mathscr{L}(a^t_V)] \\ &r_V[\mathscr{T}^*(a^t_{\mathcal{J}}), \mathscr{L}^*(a^t_V)] \geq r_V[\mathscr{T}(a^t_{\mathcal{J}}), \mathscr{L}(a^t_{-V})]. \end{aligned} \tag{23}$$

Based on (20), the proof is completed.

References

[1] 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.

[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] 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.

[4] 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.

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

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

[7] 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.

[8] Y. Xu, G. Ren, J. Chen, Y. Luo, L. Jia, X. Liu, Y. Yang, and Y. Xu, "A one-leader multi-follower bayesian-stackelberg game for anti-jamming transmission in UAV communication networks," *IEEE Access*, vol. 6, pp. 21 697–21 709, Jun. 2018.

[9] 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.

[10] 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.

[11] 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.

[12] A. 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.

[13] 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, Jan. 1995.

[14] M. Hausknecht and P. Stone, "Deep recurrent q-learning for partially observable MDPs," *CoRR, abs/1507.06527*, vol. 7, no. 1, 2015.

[15] N. Meuleau, L. Peshkin, K.-E. Kim, and L. P. Kaelbling, "Learning finite-state controllers for partially observable environments," in *Proc. Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1999, pp. 427–436.

[16] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.