Adversarial Game Against Hybrid Attacks in UAV Communications With Partial Information

Chaoqiong Fan, Huayi Liu, Bin Li, *Member, IEEE*, Chenglin Zhao, and Shiwen Mao, *Fellow, IEEE*

***Abstract*—Unmanned aerial vehicle (UAV) communications are vulnerable to smart attacks, where the attacker can change the attack mode (e.g., eavesdropping and jamming) via smart radio devices. To ensure secure transmissions against hybrid attacks, UAVs may transmit confidential information and misleading information alternately. In this paper, we consider a dynamic anti-hybrid attack framework with trajectory optimization, where both a UAV and an attacker attempt to find their optimal trajectories without knowing the information type and the attack mode of each other. Given the major challenge due to incomplete knowledge (i.e., each agent knows only its own information), we establish an adversarial game with partial-observation feature to formulate an optimization problem, and propose a counterfactual regret minimization learning scheme to achieve the correlated equilibrium for both the UAV and attacker. Simulation results validate the superiority of our scheme over a benchmark in UAV communication scenarios with partial information.**

***Index Terms*—Unmanned aerial vehicle (UAV), eavesdropping, jamming, adversarial game, counterfactual regret minimization.**

I. Introduction

Due to the advantages of high mobility and flexible deployment, unmanned aerial vehicles (UAVs) have found increasingly myriad applications, and are becoming an important component of 5 G and beyond networks [1]. However, guaranteeing the secrecy of UAV communications is a surely important yet extremely challenging problem. On one hand, due to the broadcast and line-of-sight (LoS) dominated nature of air-to-ground wireless channel, UAV communications are prone to be intercepted or jammed by potential attackers than conventional terrestrial communication systems [2]. On the other hand, by leveraging smart and programmable radio devices, illegitimate nodes become more intelligent, which are able to flexibly change their attack mode to create more severe damages [3]. Compared with the traditional fixed-mode passive attackers, smart attackers with multi-mode operation would be more harmful to UAV secure communications. To unlock the potential of UAV services, considerable prior works have been conducted to ensure secure UAV communications [4]–[8]. For cellular-connected UAV networks, the authors in [4] examined the security challenges and introduced a machine learning scheme [9]. To combat jamming attacks, the authors in [5] proposed a deep reinforcement learning algorithm for UAV-aided cellular communications. While [6] and [7] were focused on dealing with eavesdropping threats via trajectory design and power control of UAVs. When multiple attacks modes (e.g. spoofing, jamming, and eavesdropping) were considered, Xiao *et al.* in [8] formulated a prospect theory game from a user-centric perspective to optimize the attack policy of the attacker and the power allocation of the UAV.

Motivated by the work in [8],[1]in this paper we present a novel paradigm for the anti-hybrid attack optimization problem, where the smart attacker operates in three modes (i.e., eavesdropping, jamming, and sleep) to destroy UAV's secure transmissions, while the legitimate UAV can transmit both confidential information and misleading information dynamically to combat against the malicious activities. Each of the two intelligent agents aims to optimize the trajectories to maximize their long-term utilities. Taking the multiple attack modes of the attacker and the multiple types of information of the UAV into consideration, the network state, which is represented by the combination of attack mode and information type, would be highly dynamic. Since these information cannot be fully observed by the other agent, the complete knowledge of the network state is not available to either the UAV or the attacker. To address the challenges of dynamic state and incomplete information, we adopt an adversarial game with partial observation to formulate the trajectory optimization problem, and propose a counterfactual regret minimization (CRM) scheme to obtain the expected equilibrium solutions to the adversarial game. Numerical results show that our CRM scheme attains a 20% performance improvement over the Q-learning method in the presence of incomplete information.

The rest of this paper is structured as follows. We present the system model and problem formulation in Section II. The adversarial game with partial observation and the CRM scheme are presented in Section III. We evaluate the proposed scheme and demonstrate its advantages by numerical results in Section IV. The conclusions are drawn in Section V. The key notations used in this paper are summarized in Table I.

II. System Model and Problem Formulation

A. System Model

As shown in Fig. 1, we consider secure transmissions in a UAV network, where a legitimate UAV U convey information to the ground base station (GBS) G through a UAV-to-GBS (U2G) link in the presence of a smart attacker A. The malicious attacker is capable of eavesdropping and jamming and works in a half-duplex mode [10]. Specifically, when the eavesdropping mode is adopted, the attacker turns on its receiver and attempts to tap the legitimate transmissions of the U2G link; When the attacker is in the jamming mode, it sends a jamming signal to block the reception of UAV signal at the GBS. Apart from these two operation modes, the attacker may occasionally switch to a sleep mode to avoid being discovered and reduce its energy consumption. For the UAV, to combat the attack, it dynamically transmits confidential information and misleading information. In practice, the attack mode of the smart attacker and the information type of the UAV are unknown to each other, which is the main challenge to be addressed in this work.

We define a three-dimensional Cartesian coordinate system $\mathcal{S} \triangleq \{(x,y,z)|x,y,z \in \mathbb{R}\}$, where the locations of the GBS G, the UAV U, and the attacker A are given by $\mathbf{l}_{\mathrm{G}} = (0,0,H_{\mathrm{G}})$, $\mathbf{l}_{\mathrm{U}} = (x_{\mathrm{U}}, y_{\mathrm{U}}, z_{\mathrm{U}})$,

Manuscript received August 29, 2021; revised November 6, 2021; accepted December 2, 2021. Date of publication December 7, 2021; date of current version February 14, 2022. This work was supported by the Natural Science Foundation of China (NSFC) under Grants 62101051 and U1805262. The work of S. Mao was supported by NSF under Grants ECCS-1923163 and CNS-2107190. The review of this article was coordinated by Dr. Haijun Zhang. *(Corresponding author: Bin Li.)*

Chaoqiong Fan is with the School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China (e-mail: fcq@bnu.edu.cn).

Huayi Liu is with the Department of Electronic Technology, Ministry of Industry and Information Technology, Beijing 100804, China (e-mail: liuhuayi@sohu.com).

Bin Li and Chenglin Zhao are with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: stonebupt@gmail.com; clzhao@bupt.edu.cn).

Shiwen Mao is with the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849-5201 USA (e-mail: smao@ieee.org).

Digital Object Identifier 10.1109/TVT.2021.3132934

[1] Note that the formulated optimization problems, the adopted game models, and the proposed learning algorithms of [8] and our work are all different.

TABLE I
NOTATION

Symbol	Description
G, U, A	Notations of the GBS, UAV, and attacker
$\mathbf{l}_{\rm G}$, $\mathbf{l}_{\rm U}$, $\mathbf{l}_{\rm A}$	Locations of the GBS, UAV, and attacker
$h_{\rm UG}$, $h_{\rm UA}$, $h_{\rm AG}$	Channel power gains of the three links
$d_{\rm UG}$, $d_{\rm UA}$, $d_{\rm AG}$	Ranges of the three links
n, $P_{\rm U}$	Information type and transmit power of the UAV
m, $P_{\rm A}$	Attack mode and jamming power of the attacker
$R_{\rm S}^{mn}$	Secrecy rate of the U2G communication
$D_{\rm A}^{mn}$	Damage caused by the attacker
$V_{\rm U}$, $c_{\rm U}$, $\Delta d_{\rm U}$	Flight speed, flight cost per unit distance, and flight distance per unit slot of the UAV
$V_{\rm A}$, $c_{\rm A}$, $\Delta d_{\rm A}$	Moving speed, moving cost per unit distance, and moving distance per unit slot of the attacker
$w_{\rm U}$, $\mathcal{A}_{\rm U}$, $a_{\rm U}(t)$, $\mathcal{L}_{\rm U}$, $W_{\rm U}$	Instantaneous utility, action space, selected action, trajectory, and long-term utility of the UAV
$w_{\rm A}$, $\mathcal{A}_{\rm A}$, $a_{\rm A}(t)$, $\mathcal{L}_{\rm A}$, $W_{\rm A}$	Instantaneous utility, action space, selected action, trajectory, and long-term utility of the attacker
$\mathbf{p}_n$, $p_{n,a_{\rm U}}$	Probability vector of the UAV with observation n and probability of taking action $a_{\rm U}$ among $\mathcal{A}_{\rm U}$
$\mathbf{q}_m$, $q_{m,a_{\rm A}}$	Probability vector of the attacker with observation m and probability of taking action $a_{\rm A}$ among $\mathcal{A}_{\rm A}$

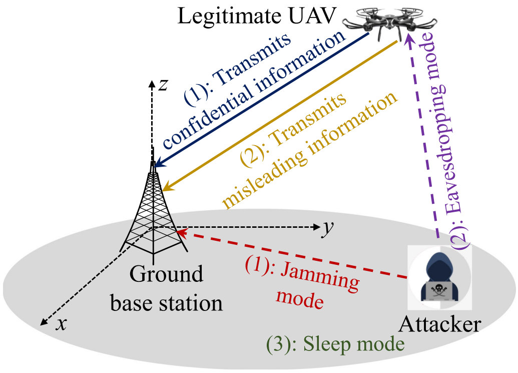


Fig. 1. Illustration of a UAV communication network in the presence of a smart, multi-mode attacker.

and $\mathbf{l}_{\rm A} = (x_{\rm A}, y_{\rm A}, 0)$, respectively. Correspondingly, the distance $d_{\rm XY}$ between any two nodes X and Y, $\rm X, Y \in \{U, G, A\}$, is $d_{\rm XY} = f(\mathbf{l}_{\rm X}, \mathbf{l}_{\rm Y})$, where $f(\mathbf{a}, \mathbf{b}) \triangleq \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2 + (z_a - z_b)^2}$ with $\mathbf{a} \triangleq (x_a, y_a, z_a) \in \mathcal{S}$ and $\mathbf{b} \triangleq (x_b, y_b, z_b) \in \mathcal{S}$.

Denote the channel gain between any pair of nodes X and Y as $h_{\rm XY}$. The link between the UAV and a ground node (i.e., G or A) are modeled as an air-to-ground channel, which could be either LoS or non-LoS (NLoS). Thus, the channel gains $h_{\rm UG}$ and $h_{\rm UA}$ are respectively given by:

$$h_{\rm UG} = \begin{cases} \beta_{\rm LoS}(d_{\rm UG})^{-\alpha} \\ \beta_{\rm NLoS}(d_{\rm UG})^{-\alpha} \end{cases} h_{\rm UA} = \begin{cases} \beta_{\rm LoS}(d_{\rm UA})^{-\alpha} \\ \beta_{\rm NLoS}(d_{\rm UA})^{-\alpha}, \end{cases} \quad (1)$$

where α is the path-loss exponent for the air-to-ground channel, and $\beta_{\rm LoS}$ and $\beta_{\rm NLoS}$ are the additional attenuation factors for the LoS link and NLoS link, respectively. The terrestrial channel experiences quasi-static independent Rayleigh fading [11]. Therefore, the channel gain $h_{\rm AG}$ of the link between the attacker and the GBS is given by:

$$h_{\rm AG} = \rho_0 (d_{\rm AG})^{-\kappa} v_{\rm AG}, \quad (2)$$

where ρ_0 represents the channel gain at a reference distance of 1 m, κ is the path-loss exponent, and $v_{\rm AG}$ is an exponentially distributed random variable with unit mean.

B. Problem Formulation

As stated, the smart attacker operates in three modes and the legitimate UAV can transmit two types of information. To characterize the attack mode of the attacker and the information type of the UAV, two indicators $m \in \{1, 2, 3\}$ and $n \in \{1, 2\}$ are introduced to represent the current attack mode and information type, respectively. Specifically, for the UAV, if confidential information is transmitted, we have $n = 1$; and if misleading information is transmitted, we have $n = 2$. If the attacker operates in the eavesdropping mode, we have $m = 1$; if the jamming mode is adopted, we have $m = 2$; and if the sleep mode is adopted, we have $m = 3$. Clearly, the secrecy rate of the UAV and the damage caused by the attacker are different for different network state $\{n, m\}$, which are elaborated as follows.

1) When the UAV Transmits Confidential Information ($n = 1$): Assume the transmit power of the UAV U remains constant during the flight period, denoted as $P_{\rm U}$. Then the communication rate $R_{\rm U}$ over the U2G link in bps/Hz is $R_{\rm U} = \log_2(1 + P_{\rm U} h_{\rm UG}/\sigma^2)$, where σ^2 is the power of additive white Gaussian noise. When the attacker operates in the eavesdropping mode, the secrecy rate $R_{\rm S}^{mn}$ of U2G communication is defined as the difference between the communication rate $R_{\rm U}$ and the eavesdropping rate, i.e.,

$$R_{\rm S}^{m1} = \left[R_{\rm U} - \log_2 \left(1 + \frac{P_{\rm U} h_{\rm UA}}{\sigma^2} \right) \right]^+, \text{if } m = 1, \quad (3)$$

where $[x]^+$ is defined as $[x]^+ \triangleq \max\{x, 0\}$. When the attacker enters the jamming mode, the secrecy rate $R_{\rm S}^{mn}$ is given by:

$$R_{\rm S}^{m1} = \log_2 \left(1 + \frac{P_{\rm U} h_{\rm UG}}{P_{\rm A} h_{\rm AG} + \sigma^2} \right), \text{if } m = 2, \quad (4)$$

where $P_{\rm A}$ is the jamming power of the attacker. If the attacker is in the sleep mode, it does not cause any damage and the secrecy rate $R_{\rm S}^{mn}$ is equal to the communication rate $R_{\rm U}$, i.e., $R_{\rm S}^{m1} = R_{\rm U}$ when $m = 3$. Moreover, the damage $D_{\rm A}^{mn}$ caused by the attacker is defined as the difference between the communication rate $R_{\rm U}$ and the secrecy rate $R_{\rm S}^{mn}$ of the U2G link, i.e., $D_{\rm A}^{m1} \triangleq R_{\rm U} - R_{\rm S}^{m1}$, for $m \in \{1, 2, 3\}$.

2) When the UAV Transmits Misleading Information ($n = 2$): Since no useful information is received by the GBS in this case, the secrecy rate $R_{\rm S}^{mn}$ for the U2G communication is zero, i.e., $R_{\rm S}^{m2} = 0$ for $m \in \{1, 2, 3\}$. For the malicious attacker, operating in the eavesdropping mode or jamming mode not only brings no gain, but also increases the risk of being misled or detected. Taking this into consideration, the damage caused by the attacker, $D_{\rm A}^{mn}$, under the two modes (i.e., eavesdropping and jamming) is defined as a decreasing function of the signal-to-noise ratio of the corresponding links. In addition, when the sleep mode is adopted, the damage caused by the attacker will be zero. Therefore, we have,

$$D_{\rm A}^{m2} \triangleq \begin{cases} -P_{\rm U} h_{\rm UA}/\sigma^2, & \text{if } m = 1 \\ -P_{\rm A} h_{\rm AG}/\sigma^2, & \text{if } m = 2 \\ 0, & \text{if } m = 3. \end{cases} \quad (5)$$

Assume the UAV and the smart attacker are both mobile with constant speeds $V_{\rm U}$ and $V_{\rm A}$, respectively. As intelligent agents, they can learn the optimal trajectories to maximize their own utility, which are defined as the difference between the reward and the cost, i.e.,

$$\begin{cases} w_{\rm U} \triangleq (R_{\rm S}^{mn} - c_{\rm U} \Delta d_{\rm U}) + C_1 \\ w_{\rm A} \triangleq (D_{\rm A}^{mn} - c_{\rm A} \Delta d_{\rm A}) + C_2, \end{cases} \quad (6)$$

where c_{U} and $\Delta d_{\mathrm{U}} = V_{\mathrm{U}} \times \Delta t$ are the flight cost per unit distance and the flight distance of the UAV, c_{A} and $\Delta d_{\mathrm{A}} = V_{\mathrm{A}} \times \Delta t$ are the moving cost per unit distance and the moving distance of the attacker, C_1 and C_2 are constants to guarantee that the utilities are nonnegative.

To facilitate the analysis, we assume that the UAV and the attacker are moving parallel to the coordinate axis directions [12], and discretize the trajectories by their per unit distance in each time slot. Then the action space of the UAV is denoted as:

$$\begin{aligned}\mathcal{A}_{\mathrm{U}} = \{&(0,0,0), (-V_{\mathrm{U}},0,0), (V_{\mathrm{U}},0,0), (0,V_{\mathrm{U}},0),\\ &(0,-V_{\mathrm{U}},0), (0,0,V_{\mathrm{U}}), (0,0,-V_{\mathrm{U}})\},\end{aligned} \tag{7}$$

which represents its flight directions including stay, left, right, forward, backward, up, and down in the three-dimensional space. Likewise, the action space of the attacker is given by:

$$\begin{aligned}\mathcal{A}_{\mathrm{A}} = \{&(0,0,0), (-V_{\mathrm{A}},0,0), (V_{\mathrm{A}},0,0),\\ &(0,V_{\mathrm{A}},0), (0,-V_{\mathrm{A}},0)\},\end{aligned} \tag{8}$$

which represents its moving directions including stay, left, right, forward, and backward in the two-dimensional space.

In time slot t, the UAV U chooses an action $a_{\mathrm{U}}(t)$ from its action space $\mathcal{A}_{\mathrm{U}}$, and the attacker A chooses an action $a_{\mathrm{A}}(t)$ from its action space $\mathcal{A}_{\mathrm{A}}$. The trajectories of the UAV U and the attacker A from start to time slot T are given by:

$$\begin{cases}\mathcal{L}_{\mathrm{U}} = [\mathbf{l}_{\mathrm{U}}(t), t=0,\ldots,T|\mathbf{l}_{\mathrm{U}}(t) = \mathbf{l}_{\mathrm{U}}(t-1) + a_{\mathrm{U}}(t), \mathbf{l}_{\mathrm{U}}(0) = \mathbf{l}_{\mathrm{U}}],\\ \mathcal{L}_{\mathrm{A}} = [\mathbf{l}_{\mathrm{A}}(t), t=0,\ldots,T|\mathbf{l}_{\mathrm{A}}(t) = \mathbf{l}_{\mathrm{A}}(t-1) + a_{\mathrm{A}}(t), \mathbf{l}_{\mathrm{A}}(0) = \mathbf{l}_{\mathrm{A}}].\end{cases} \tag{9}$$

Denote the partial observation vectors of the UAV and the attacker as $\mathbf{n} = [n(0), n(1), \ldots, n(T)]$ and $\mathbf{m} = [m(0), m(1), \ldots, m(T)]$, respectively. Then, the long-term utilities of the UAV and the attacker are given by $W_{\mathrm{U}}[\mathcal{L}_{\mathrm{U}}, \mathcal{L}_{\mathrm{A}}|\mathbf{n}] = \sum_{t=0}^{T} w_{\mathrm{U}}(t)$ and $W_{\mathrm{A}}[\mathcal{L}_{\mathrm{U}}, \mathcal{L}_{\mathrm{A}}|\mathbf{m}] = \sum_{t=0}^{T} w_{\mathrm{A}}(t)$, respectively. The goals of the UAV and the attacker are to maximize their long-term cumulative utilities via optimizing the trajectories with partial information. Therefore, the optimization problem can be formulated as:

$$\begin{cases}\max_{a_{\mathrm{U}}(t)\in\mathcal{A}_{\mathrm{U}}} W_{\mathrm{U}}[\mathcal{L}_{\mathrm{U}}, \mathcal{L}_{\mathrm{A}}|\mathbf{n}],\\ \max_{a_{\mathrm{A}}(t)\in\mathcal{A}_{\mathrm{A}}} W_{\mathrm{A}}[\mathcal{L}_{\mathrm{U}}, \mathcal{L}_{\mathrm{A}}|\mathbf{m}],\end{cases} t = 0, 1, \ldots, T. \tag{10}$$

III. Adversarial Game Based Scheme

Due to the multiple attack modes and information types, the network state is dynamic. Since the information pattern of the UAV and the mode of the attacker are private information, the complete knowledge of the network state remains partially observable for the two agents. Facing these practical challenges, an adversarial game based optimization scheme for coping with the dynamical state and incomplete information is proposed in this section.

A. Adversarial Game With Partial Observation

Considering the adversarial interactions between the UAV and the smart attacker, and the incomplete knowledge of the two agents about the network state, we adopt an adversarial game $\mathcal{G}$ with partial observation characteristics to reformulate the anti-hybrid attack optimization problem in the UAV communication network. Formally, we have $\mathcal{G} \triangleq \langle\{\mathrm{U}, \mathrm{A}\}, \{n, m\}, \{\mathcal{A}_{\mathrm{U}}, \mathcal{A}_{\mathrm{A}}\}, \{\mathbf{p}_n, \mathbf{q}_m\}, \{w_{\mathrm{U}}, w_{\mathrm{A}}\}\rangle$, where n and m are partial observations of the two agents U and A, respectively; $\mathbf{p}_n = [p_{n,a_{\mathrm{U}}}]_{1\times|\mathcal{A}_{\mathrm{U}}|}$ is the probability distribution over the action space $\mathcal{A}_{\mathrm{U}}$ with observation n, and $p_{n,a_{\mathrm{U}}}$ is the probability of taking action a_{U} among $\mathcal{A}_{\mathrm{U}}$; similarly, $\mathbf{q}_m = [q_{m,a_{\mathrm{A}}}]_{1\times|\mathcal{A}_{\mathrm{A}}|}$ is the probability distribution over the action space $\mathcal{A}_{\mathrm{A}}$ with observation m, and $q_{m,a_{\mathrm{A}}}$ is the probability of taking action a_{A} among $\mathcal{A}_{\mathrm{A}}$; and $|\mathcal{X}|$ represents the number of elements in set $\mathcal{X}$.

Although Nash equilibrium (NE) is a leading notation for non-cooperative games, it retains some weaknesses in practice: (i) computing an NE could be intractable; (ii) the NE is prone to have equilibrium selection issues; and (iii) the social welfare performance of an NE may be relatively inferior. Here, we introduce a more general concept, named correlated equilibrium (CE), to characterized the adversarial game.

Definition 1 (Correlated Equilibrium): A CE solution of the adversarial game $\mathcal{G}$ with a specific network state $\{n, m\}$ is a joint strategy profile $(\mathbf{p}_n^*)^{\mathrm{T}}\mathbf{q}_m^* = [p_{n,a_{\mathrm{U}}}^* q_{m,a_{\mathrm{A}}}^*]_{|\mathcal{A}_{\mathrm{U}}|\times|\mathcal{A}_{\mathrm{A}}|}$ over the joint action space $\mathcal{A}_{\mathrm{U}} \times \mathcal{A}_{\mathrm{A}}$, such that for the two agents U and A, it holds that

$$\begin{cases}\sum_{[a_{\mathrm{U}},a_{\mathrm{A}}]\in\mathcal{A}_{\mathrm{U}}\times\mathcal{A}_{\mathrm{A}}} p_{n,a_{\mathrm{U}}}^* q_{m,a_{\mathrm{A}}}^* [w_{\mathrm{U}}(a_{\mathrm{U}}, a_{\mathrm{A}}) - w_{\mathrm{U}}(a'_{\mathrm{U}}, a_{\mathrm{A}})] \geqslant 0\\ \sum_{[a_{\mathrm{U}},a_{\mathrm{A}}]\in\mathcal{A}_{\mathrm{U}}\times\mathcal{A}_{\mathrm{A}}} p_{n,a_{\mathrm{U}}}^* q_{m,a_{\mathrm{A}}}^* [w_{\mathrm{A}}(a_{\mathrm{U}}, a_{\mathrm{A}}) - w_{\mathrm{A}}(a_{\mathrm{U}}, a'_{\mathrm{A}})] \geqslant 0.\end{cases} \tag{11}$$

The anti-hybrid attack trajectory optimization problem with incomplete information can be transformed to an adversarial game with partial observation, in which the two intelligent agents aim to obtain the CE solutions to maximize their utilities, i.e.,

$$\begin{cases}\mathbf{p}_n^* = \arg\max \sum\limits_{a_{\mathrm{U}}\in\mathcal{A}_{\mathrm{U}}} p_{n,a_{\mathrm{U}}} w_{\mathrm{U}}(a_{\mathrm{U}}, a_{\mathrm{A}}), n \in \{1, 2\}\\ \mathbf{q}_m^* = \arg\max \sum\limits_{a_{\mathrm{A}}\in\mathcal{A}_{\mathrm{A}}} q_{m,a_{\mathrm{A}}} w_{\mathrm{A}}(a_{\mathrm{U}}, a_{\mathrm{A}}), m \in \{1, 2, 3\}.\end{cases} \tag{12}$$

Remark: It has been shown that every finite game has a nonempty set of CEs [13]. For our established adversarial game $\mathcal{G}$, with the finite number of agents and their corresponding action spaces, it is indeed a finite game. Therefore the existence of CEs of the constructed adversarial game $\mathcal{G}$ can be guaranteed. Moreover, (11) shows that the CE can be regarded as a more general mixed-strategy equilibrium solution of the game. In this case, the CE solution is usually not unique.

B. Counterfactual Regret Minimization Scheme

Considering the dynamics and diversity of the network state as characterized by (12), the agents U and A are expected to apply different strategies according to their different observations, i.e., $\mathbf{p}_n^*$ for $n \in \{1, 2\}$ and $\mathbf{q}_m^*$ for $m \in \{1, 2, 3\}$. To this end, a CRM scheme is proposed, which can cope with the incomplete information of agents and achieve CE solutions of the adversarial game, thus maximizing the long-term utilities of the two agents with only their partial observations. The key idea of the CRM is to learn the environment from the perspective of *partial observation* rather than the *global state*, and to track regrets for past plays and then to take future actions proportional to the positive regrets [14]. In this case, the challenges of the dynamical state and incomplete information can be addressed. The CRM algirthm is designed as follows. Let $\mathrm{X} \in \{\mathrm{U}, \mathrm{A}\}$ denote an arbitrary agent, $o \in \{n, m\}$ represent the observation of agent X, and $\mathbf{s}_o = [s_{o,a_{\mathrm{X}}}]_{1\times|\mathcal{A}_{\mathrm{X}}|} \in \{\mathbf{p}_n, \mathbf{q}_m\}$ denote the strategy profile of agent X with observation o, where a_{X} and $\mathcal{A}_{\mathrm{X}}$ are the selected action and the action space of agent X with $a_{\mathrm{X}} \in \mathcal{A}_{\mathrm{X}}$. To enable strategy learning just relying on partial information, the CRM leverages the *action-value* with a specific observation and the *observation-value* of all possible observations to calculate the counterfactual advantage of actions with different observations. Specifically, in iteration i, with a specific observation o, agent X takes the action a_{X} according to the current strategy profile $\mathbf{s}_o(i)$. Then the action-value $\varphi_{\mathrm{X}}[o, a_{\mathrm{X}}|\mathbf{s}_o(i)]$ of action a_{X} with observation o, and the observation-value $\psi_{\mathrm{X}}[o|\mathbf{s}_o(i)]$ of observation o are respectively given by:

$$\begin{cases} \varphi_{\mathrm{X}}[o, a_{\mathrm{X}}|\mathbf{s}_o(i)] \triangleq w_{\mathrm{X}}(i) \\ \psi_{\mathrm{X}}[o|\mathbf{s}_o(i)] = \sum_{a_{\mathrm{X}}\in\mathcal{A}_{\mathrm{X}}} s_{o,a_{\mathrm{X}}}(i)\varphi_{\mathrm{X}}[o, a_{\mathrm{X}}|\mathbf{s}_o(i)]. \end{cases} \quad (13)$$

Then for agent X with observation o, the instantaneous regret $e_{\mathrm{X}}(o, a_{\mathrm{X}}, i)$ of not taking action a_{X} is expressed as:

$$e_{\mathrm{X}}(o, a_{\mathrm{X}}, i) = \varphi_{\mathrm{X}}[o, a_{\mathrm{X}}|\mathbf{s}_o(i)] - \psi_{\mathrm{X}}[o|\mathbf{s}_o(i)]. \quad (14)$$

To fully exploit the past experience, the CRM scheme applies the cumulative counterfactual regret for strategy updating. Over a period of I iterations, the cumulative regret $E_{\mathrm{X}}(o, a_{\mathrm{X}}, I)$ of agent X with observation o is given by:

$$E_{\mathrm{X}}(o, a_{\mathrm{X}}, I) = \sum\nolimits_{i=1}^{I} e_{\mathrm{X}}(o, a_{\mathrm{X}}, i). \quad (15)$$

In the CRM scheme, agent X with observation o updates its strategy profile $\mathbf{s}_o$ for all actions as follows:

$$s_{o,a_{\mathrm{X}}}(I+1) = \begin{cases} \frac{[E_{\mathrm{X}}(o,a_{\mathrm{X}},I)]^+}{\Phi_{\mathrm{X}}(o,I)}, & \text{if} \Phi_{\mathrm{X}}(o,I) > 0, \\ \frac{1}{|\mathcal{A}_{\mathrm{X}}|}, & \text{otherwise}, \end{cases} \quad (16)$$

where $\Phi_{\mathrm{X}}(o, I) \triangleq \sum_{a_{\mathrm{X}}\in\mathcal{A}_{\mathrm{X}}}[E_{\mathrm{X}}(o, a_{\mathrm{X}}, I)]^+$ is the total nonnegative cumulative counterfactual regret from the first iteration to the Ith iteration of all actions with observation o.

Next, we analyze the convergence of the proposed CRM algorithm from the perspective of the average action selection probability, which is defined as follows.

Definition 2 (Average Action Selection Probability): For agent X, consider a period of plays from $i = 1$ to $i = I$ and denote the set of iteration index that a specific observation o occurs as $\mathcal{I}_o$, i.e., $\mathcal{I}_o \triangleq \{i = 1, 2, \ldots, I|o(i) = o\}$. Then, for action a_{X} with observation o, the average action selection probability from the first play to the Ith play is given by:

$$\bar{s}_{o,a_{\mathrm{X}}}(I) = \frac{\sum_{i\in\mathcal{I}_o} s_{o,a_{\mathrm{X}}}(i)}{|\mathcal{I}_o|}. \quad (17)$$

Theorem 1: Applying the CRM scheme to the formulated adversarial game, as the number of iterations $I \to \infty$, the average action selection probability $\bar{s}_{o,a_{\mathrm{X}}}(I)$ of action a_{X} with observation o converges almost surely to the set of CEs.

Proof: The proof process is based on a lemma from [15], which gives the regret bound.

Lemma 1: With the proposed CRM algorithm, for an observation o of agent X, denote the nonnegative average counterfactual regret from $i = 1$ to $i = I$ as $\mu_{\mathrm{X}}(o, a_{\mathrm{X}}, I)$, i.e., $\mu_{\mathrm{X}}(o, a_{\mathrm{X}}, I) \triangleq [E_{\mathrm{X}}(o, a_{\mathrm{X}}, I)]^+/I$. Then we have

$$\sum\nolimits_{a_{\mathrm{X}}\in\mathcal{A}_{\mathrm{X}}} [\mu_{\mathrm{X}}(o, a_{\mathrm{X}}, I)]^2 \leqslant \frac{1}{I^2}\sum_{i=1}^{I} |\mathcal{A}_{\mathrm{X}}|[\xi_{\mathrm{X}}(o, i)]^2, \quad (18)$$

where $\xi_{\mathrm{X}}(o, i) \triangleq \max_{a_{\mathrm{X}}\in\mathcal{A}_{\mathrm{X}}} e_{\mathrm{X}}(o, a_{\mathrm{X}}, i)$ is the maximum term of the instantaneous counterfactual regrets $e_{\mathrm{X}}(o, a_{\mathrm{X}}, i)$ of all actions in the ith iteration [15].

According to **Lemma 1**, we have

$$\left\{\max_{a_{\mathrm{X}}\in\mathcal{A}_{\mathrm{X}}} [\mu_{\mathrm{X}}(o, a_{\mathrm{X}}, I)]\right\}^2 = \max_{a_{\mathrm{X}}\in\mathcal{A}_{\mathrm{X}}} [\mu_{\mathrm{X}}(o, a_{\mathrm{X}}, I)]^2$$

$$\leqslant \sum\nolimits_{a_{\mathrm{X}}\in\mathcal{A}_{\mathrm{X}}} [\mu_{\mathrm{X}}(o, a_{\mathrm{X}}, I)]^2 \leqslant \frac{|\mathcal{A}_{\mathrm{X}}|}{I^2}\sum_{i=1}^{I}[\xi_{\mathrm{X}}(o, i)]^2$$

$$\leqslant \frac{|\mathcal{A}_{\mathrm{X}}|}{I^2}[\Xi_{\mathrm{X}}(o)]^2 = \frac{|\mathcal{A}_{\mathrm{X}}|}{I}[\Xi_{\mathrm{X}}(o)]^2, \quad (19)$$

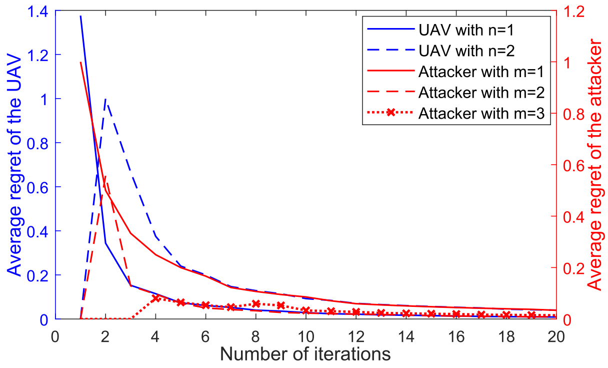


Fig. 2. Evolutions of the average regrets of the two agents U and A with the number of iterations increasing.

where $\Xi_{\mathrm{X}}(o) \triangleq \max_{\{i=1,2,\ldots,I\}} \xi_{\mathrm{X}}(o, i)$ is the maximum value of all $\xi_{\mathrm{X}}(o, i)$ from the first iteration to the Ith iteration.

According to (19), we conclude that as $I \to \infty$, for any a_{X} of agent X with observation o, the nonnegative average counterfactual regret $\mu_{\mathrm{X}}(o, a_{\mathrm{X}}, I)$ will approach zero, i.e.,

$$\lim_{I\to\infty} \mu_{\mathrm{X}}(o, a_{\mathrm{X}}, I) \leqslant \lim_{I\to\infty} \max_{a'_{\mathrm{X}}\in\mathcal{A}_{\mathrm{X}}} [\mu_{\mathrm{X}}(o, a'_{\mathrm{X}}, I)]$$

$$\leqslant \lim_{I\to\infty} \sqrt{\frac{|\mathcal{A}_{\mathrm{X}}|\,[\Xi_{\mathrm{X}}(o)]^2}{I}} = \lim_{I\to\infty} \frac{\sqrt{|\mathcal{A}_{\mathrm{X}}|}\Xi_{\mathrm{X}}(o)}{\sqrt{I}} = 0. \quad (20)$$

Therefore, we can state that the CRM is a regret minimization algorithm with guaranteed convergence. Moreover, according to [16], the conclusion can be drawn that if all agents obey the recommendation of some regret minimizers, the average probability $\bar{s}_{o,a_{\mathrm{X}}}(I)$ of all actions approaches the set of CEs as $I \to \infty$. Thus the proof is completed.□

IV. Simulation Results

In this section, we evaluate the performance of our proposed CRM scheme for anti-hybrid attack trajectory optimization via simulations. The default simulation parameters are set as: $\alpha = 3$, $\beta_{\mathrm{LoS}} = 1$ dB, $\beta_{\mathrm{NLoS}} = 20$ dB, $\kappa = 2$, $\rho_0 = -30$ dBm, $P_{\mathrm{U}} = 0.1$W, $P_{\mathrm{A}} = 0.05$W, $\sigma^2 = 0.1$mW, $V_{\mathrm{U}} = 2$m/s, $V_{\mathrm{A}} = 1$m/s, $c_{\mathrm{U}} = 2$mW and $c_{\mathrm{A}} = 1$mW.

First, to demonstrate the convergence performance of our proposed CRM scheme, we present the evolution of the average regrets of the two agents under different observations in Fig. 2. Clearly, with more iterations, the average regrets of the UAV and the attacker with all their observations both approach zero, which verifies **Theorem 1**. That is to say, with the proposed CRM algorithm, each agent can learn the appropriate strategies just relying on partial observation of the network state, by which the agents can take actions with no regrets. Thus the optimized strategies of the two agents are the CE solutions.

Furthermore, within a period of T iterations, the optimized trajectories of the UAV and the attacker given by the CRM scheme are presented in Fig. 3. It is observed that the attacker selectively approaches the UAV transmitter and the GBS receiver according to whether its current attack mode is eavesdropping or jamming, and its two-dimensional location starts at (−13 m, 20 m) and ends at (6 m, 7 m). The UAV would keep a safe distance from the attacker while approaching the GBS, and its three-dimensional location starts at (32 m, 13 m, 36 m) and ends at (2 m, −7 m, 10 m).

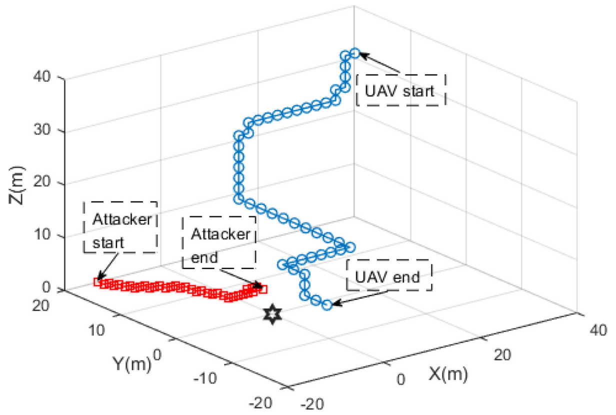
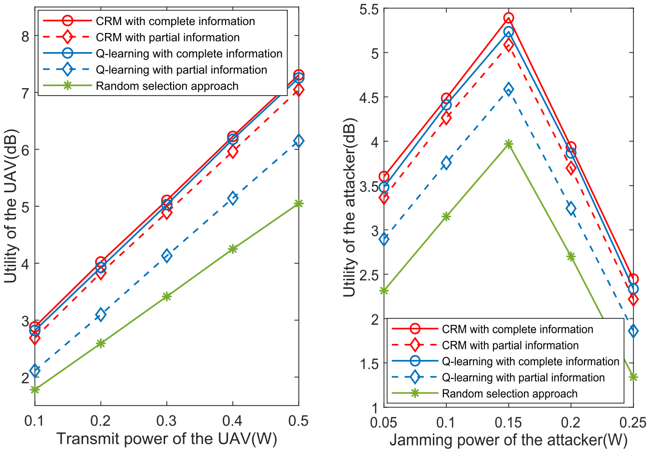


Fig. 3. The optimal trajectories of the UAV and the attacker given by the CRM scheme.



(a) (b)

Fig. 4. Utilities of the two agents under different learning schemes. (a) The UAV's utility. (b) The attacker's utility.

Moreover, to evaluate the optimization performance in terms of long-term utility, we use the Q-learning method as a benchmark, and consider two scenarios that each agent has complete or partial information of the network state. The performance of the random selection approach is regarded as a lower bound. The performance trends of the UAV and the attacker are provided in Fig. 4(a) and Fig. 4(b), respectively, where their own transmit or jamming power is increased and the other's power remains unchanged. It can be seen that for the UAV, its utility can be enhanced by straightly increasing the transmit power. Yet for the attacker, since higher jamming power would raise the risk of being detected, its utility decreases if the jamming power is larger than a certain value. More important, Fig. 4 shows that the performance of the CRM with complete and partial information are almost the same, while that of the Q-learning in the two scenarios are significantly different. The performance of Q-learning degrades severely under partial information. These results validate that our proposed CRM scheme is capable of handling incomplete information and hence achieving superior performance in both scenarios with partial information over the benchmark scheme.

V. Conclusion

In this paper, we investigated the anti-hybrid attack trajectory optimization problem for secure communications in UAV networks. To overcome the challenges of dynamic network state and incomplete information, an adversarial game with partial observation was established, and a CRM scheme that focused on minimizing the regret of actions under all possible observations was proposed to achieve the CE of the game. Numerical results validated the advantages of our proposed CRM scheme over the benchmark especially in practical scenarios with partial information. Transmit (jamming) power optimization and information type (attack mode) selection of the UAV (the attacker) are interesting directions and worthy of further study.

References

[1] Y. Kawamoto, H. Nishiyama, N. Kato, F. Ono, and R. Miura, "Toward future unmanned aerial vehicle networks: Architecture, resource allocation and field experiments," *IEEE Wireless Commun.*, vol. 26, no. 1, pp. 94–99, Feb. 2019.

[2] Y. Zhou, F. Zhou, H. Zhou, D. W. K. Ng, and R. Q. Hu, "Robust trajectory and transmit power optimization for secure UAV-enabled cognitive radio networks," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4022–4034, Jul. 2020.

[3] B. Ahuja, D. Mishra, and R. Bose, "Optimal green hybrid attacks in secure IoT," *IEEE Wireless Commun. Lett.*, vol. 9, no. 4, pp. 457–460, Apr. 2020.

[4] 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.

[5] X. Lu, L. Xiao, C. Dai, and H. Dai, "UAV-aided cellular communications with deep reinforcement learning against jamming," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 48–53, Aug. 2020.

[6] G. Zhang, Q. Wu, M. Cui, and R. Zhang, "Securing UAV communications via joint trajectory and power control," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 1376–1389, Feb. 2019.

[7] S. Li, B. Duo, M. Di Renzo, M. Tao, and X. Yuan, "Robust secure UAV communications with the aid of reconfigurable intelligent surfaces," *IEEE Trans. Wireless Commun.*, vol. 20, no. 10, pp. 6402–6417, Oct. 2021.

[8] 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.

[9] Y. Sun, M. Peng, Y. Zhou, Y. Huang, and S. Mao, "Application of machine learning in wireless networks: Key techniques and open issues," *IEEE Commun. Surv. Tut.*, vol. 21, no. 4, pp. 3072–3108, Oct.–Dec. 2019.

[10] H. Chen, X. Tao, N. Li, Y. Hou, J. Xu, and Z. Han, "Secrecy performance analysis for hybrid wiretapping systems using random matrix theory," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 1101–1114, Feb. 2019.

[11] J. Liu, H. Nishiyama, N. Kato, and J. Guo, "On the outage probability of device-to-device-communication-enabled multichannel cellular networks: An RSS-threshold-based perspective," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 1, pp. 163–175, Jan. 2016.

[12] N. Gao, Z. Qin, X. Jing, Q. Ni, and S. Jin, "Anti-intelligent UAV jamming strategy via deep Q-networks," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 569–581, Jan. 2020.

[13] S. Hart and A. Mas-Colell, "A simple adaptive procedure leading to correlated equilibrium," *Econometrica*, vol. 68, no. 5, pp. 1127–1150, Sep. 2000.

[14] C. Fan, B. Li, C. Zhao, and Y.-C. Liang, "Regret matching learning based spectrum reuse in small cell networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 1060–1064, 2020.

[15] M. Lanctot, "Monte Carlo sampling and regret minimization for equilibrium computation and decision-making in large extensive form games," Ph.D. dissertation, Univ. Alberta, Edmonton, Canada, 2013.

[16] A. Celli, A. Marchesi, G. Farina, and N. Gatti, "No-regret learning dynamics for extensive-form correlated equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, Canada, Dec. 2020, pp. 1–11.