Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 8825120, 15 pages
https://doi.org/10.1155/2020/8825120

Research Article

Hierarchical Q-Learning Based UAV Secure Communication against Multiple UAV Adaptive Eavesdroppers

Jue Liu,[1,2] Nan Sha,[1] Weiwei Yang,[1] Jia Tu,[3] and Lianxin Yang[1]

[1]*College of Communications Engineering, Army Engineering University of PLA, Nanjing 210007, China*
[2]*School of Information Science and Engineering, Jinshen College of Nanjing Audit University, Nanjing 210023, China*
[3]*College of International Studies, National University of Defense Technology, Nanjing 210039, China*

Correspondence should be addressed to Weiwei Yang; wwyang1981@163.com

Received 16 June 2020; Revised 19 August 2020; Accepted 13 September 2020; Published 8 October 2020

Academic Editor: Ashok Kumar Das

In this paper, we investigate secure unmanned aerial vehicle (UAV) communication in the presence of multiple UAV adaptive eavesdroppers (AEs), where each AE can conduct eavesdropping or jamming adaptively by learning others' actions for degrading the secrecy rate more seriously. The one-leader and multi-follower Stackelberg game is adopted to analyze the mutual interference among multiple AEs, and the optimal transmit powers are proven to exist under the existing conditions. Following that, a mixed-strategy Stackelberg Equilibrium based on finite and discretized power set is also derived and a hierarchical Q-learning based power allocation algorithm (HQLA) is proposed to obtain the optimal power allocation strategy of the transmitter. Numerical results show that secrecy performance can be degraded severely by multiple AEs and verify the availability of the optimal power allocation strategy. Finally, the effect of the eavesdropping cost on the AE's attack mode strategies is also revealed.

1. Introduction

With the inherent advantages in mobility, flexibility, and adaptive altitude, unmanned aerial vehicle (UAV) wireless communication has experienced an upsurge of interest in both military and civilian applications [1–6]. However, both the broadcast nature of the wireless medium and the malicious attackers make the electromagnetic environment of UAV communication hostile. Hence, the security issue of UAV communications is of paramount importance yet a significant challenge [7].

As an option, the physical layer security (PLS) technique with lower computation complexity has been proven that it can protect wireless communication networks from wiretapping and interfering by exploiting the random characteristics of wireless channels in recent years [8–11]. Naturally, due to the payload-limited characteristic of UAV, many PLS approaches [12–23] combining with the high altitude and mobility of UAV have been applied widely in UAV-involved communications.

However, most approaches in the above work that mainly focused on the single-mode scenarios are not fully suitable for the novel attackers, named as "adaptive eavesdroppers (AEs)," "active eavesdropper," or "smart eavesdropper." They use programmable radio devices to flexibly choose their attack methods, such as eavesdropping, jamming, and spoofing, according to the ongoing transmission status and the radio channel states. For example, an AE sends spoofing signals if she has a similar channel state with Alice or sends jamming signals if she is very close to Bob. Compared with the traditional single-mode attackers each performing a single-mode attack, an AE can be more harmful to the UAV transmission by reducing the secrecy capacity. Therefore, it is urgent to investigate the effective countermeasures against this type of eavesdropper.

In recent years, some literature began to investigate AE. One form of the AE is achieved by the manner of multiantenna full-duplex (FD) technology [24–27], which can assign one part of antennas to wiretap and the other antennas interfere simultaneously. Another type of AE emerges during the

channel estimation phase in the form of time-division duplex, and it leads to pilot contamination by sending the same pilot sequence as the legitimate node [28–31]. While in the data transmission phase, the AE reverts to passive eavesdropper again. Nevertheless, it should be noticed that the attack modes of these two forms of AE are predefined which means it cannot change the attack mode adaptively. The third form of AE in [32–36] can adjust its attack strategies adaptively. But there are still several problems remaining unsolved. Firstly, current work rarely considered multiple AEs case and the mutual interference between themselves. Secondly, existing studies neglected the AE adaptivity supported by the learning ability in searching for the optimal strategies of the transmitter and did not reveal the impact of the learning ability of AE on the secrecy performance of the considered system. How to search the transmitter's optimal power strategies in the face of multiple AEs with the learning ability and how to handle the mutual interference from AEs to improve the security capacity of the UAV communication system are necessary to be considered.

In our work, we mainly concentrate on a secure UAV communication scenario in the presence of multiple UAV AEs, which can eavesdrop or jam adaptively by learning others' strategies as well as dynamic environments. For the considered scenario, each AE's attack activity may affect the signal to interference plus noise ratio (SINR) of others. This implies that each AE's decision-making is not only coupled with the interactions from the transmitter but also from other AEs. Considering these hierarchical interactions between the transmitter-side and AEs-side, the Stackelberg game [37–41] is a suitable framework to capture the sequential interactions between the transmitter and AEs. Then, the Stackelberg Equilibrium (SE) points of the formulated game turn to be the feasible solutions to the transmit power allocation problem. However, the SE points solely provide theoretic solutions and it is challenging to obtain the SE solutions. In particular, the AE with the learning ability in this paper makes decisions spontaneously and independently, which results in unpredictable attack modes of the whole AE set. In this context, it is not feasible to handle this problem by centralized means because the number of each attack mode and locations of AEs are unknown, which motivates applying the idea of reinforcement learning (RL). So, we incorporate RL technology into the proposed game and a hierarchical Q-learning based power allocation algorithm is proposed to obtain the mixed-strategy equilibrium solution. The main contributions of this paper are summarized as follows:

(i) We propose a secure UAV communication model which constitutes of one transmitter-receiver pair and multiple UAV AEs. Each AE decides to eavesdrop or jam adaptively by learning the other nodes' strategies as well as the dynamic environment to maximize its damage. Also, the interference among AEs is investigated.

(ii) We formulate the UAV secure transmission problem as a one-leader and multi-follower Stackelberg game where the transmitter acts as the leader and all AEs are followers. The optimal transmit power of leader are obtained by analyzing the pure strategy SEs under the existing conditions. Besides, the mixed-strategy SE is also derived for the finite and discretized power set. Then, we apply a hierarchical RL framework in which each player chooses its attack strategy based on a probability distribution and a hierarchical Q-learning based power allocation algorithm is proposed to discover the mixed-strategy equilibrium of the formulated game. Besides, we provide rigorous theoretical proof about the convergence of the proposed algorithm.

(iii) Numerical results show the availability of the optimal power allocation strategy of the legitimate transmitter in the more hostile situation and reveal the impact of AE's learning ability on the secrecy rate. Meanwhile, we show that the proposed algorithm has a significant convergence advantage over the single-agent RL algorithm. Finally, the effect of the eavesdropping cost on the AE's attack mode strategies is also revealed.

(iv) We organize the rest of this paper as follows. In Section 2, we present the related work. Then, we present the system model in Section 3. In Section 4, we formulate the UAV secure transmission game and investigate a power allocation policy in Section 5. In Section 6, we provide the simulation results and conclude the work in Section 7.

2. Related Work

In UAV communication, there have been abundant approaches, such as 3D beamforming [12–14], trajectory optimization [5, 15–19], multi-UAV cooperation [17, 20], and resource management techniques [21–23], concerning on the single attack mode. Whereas, it is inappropriate to apply them directly to defend the novel attacker that has the multiple abilities of eavesdropping, jamming, spoofing, and so on.

As a novel attacker, the AE can eavesdrop and jam simultaneously by the FD capability [24–27]. Specifically, Tang et al. investigated the physical layer security issue in the presence of an FD AE within a hierarchical game framework in [24]. In [25], Mukherjee and Swindlehurst examined the design of an FD active eavesdropper in the 3-user MIMOME wiretap channel, where the adversary intends to optimize its transmit and receive sub-arrays and jamming signal parameters to minimize the MIMO secrecy rate of the main channel. In [26], the potential benefits of an FD receiver node in the presence of an active FD eavesdropper was studied. The optimal receive/transmit antennas allocation at the receiver against active eavesdropper in an FD pattern is provided in [27]. The second AE scenario adopts time-division duplex technology. The adaptive eavesdropper sent the same pilot sequence as the legitimate user node in the training phase leading to pilot contamination [28–31]. Zhou et al. discussed how an AE attacked the training phase in wireless communication to improve its eavesdropping performance in [28]. A

simple protocol to determine whether an AE is present or not using the channel properties of MMIMO is proposed in [29]. A novel random-training-assisted (RTA) pilot spoofing detection algorithm and a zero-forcing based secure transmission scheme is proposed to protect the confidential information from the active eavesdropper in [30]. Unfortunately, all AEs in the above scenarios cannot adjust attack mode adaptively. More recently, the AE that can determine the attack mode autonomously has been studied in [32–36]. To be specific, Li et al. studied the secure communication game under the AE from UAV with the imperfect channel estimation but ignored the mobility of UAV in [32]. Li et al. formulated the MIMO transmission in the presence of AE as a noncooperative game and obtained the power control strategy based on Q-learning in [33]. Zhu et al. proposed a noncooperative strategic game to make a complex decision between users that perform uplink transmission via relay and an active malicious node in [34]. In [35], Xiao et al. formulated a subjective smart attack game for the UAV transmission and proposed a deep Q-learning RL based UAV power allocation strategies. However, these above researches did not refer to the multiple AEs' scenario, and the mutual interference between AEs is hardly considered. Moreover, these AEs cannot learn from others' strategies and the dynamic environment. A summary of the proposed literature about AE has been given in Table 1.

Our work in this paper is different from the above researches that we focus on the AE with learning ability that can choose the attack mode independently and investigate the secure transmission problem of UAV communication in the presence of multiple AEs. Note that the approach of defending multiple AEs using the Stackelberg game in UAV communication networks was presented in our previous work [37], and the main differences and new contributions are (i) aim to the actual UAV communication, we introduce the mixed-strategies for the discretized transmit power set, and (ii) we assume that each AE has the learning ability and reveal the impact of the AE's learning ability on the secrecy rate. Besides, the similarity between the most related work in [32] and our work is that the Stackelberg game-based power allocation problem in the secure transmission of UAV communication is investigated. The main differences are (i) we consider the multiple AEs case which is more actual in UAV communication while the work in [32] ignores it, and (ii) the mutual interference among themselves is considered.

3. System Model

As shown in Figure 1, we consider the downlink of a UAV communication system consisting of a transmitter (Alice), a receiver (Bob), and M number of UAV AEs randomly distributed around transmitter-receiver pairs, where all nodes are single-antenna and UAVs are all hovering. Here, we adopt a 3D Cartesian coordinate system with the Alice, Bob, and the AE_m located at (x_a, y_a, h_a), (x_b, y_b, h_b), and (x_m, y_m, h_m). Alice communicates with Bob by using transmit power that is denoted by $P_s \in [0, P_{\max}]$, where $P_{\max}$ is the maximum transmit power. Without the loss of generality, being a programmable radio device, when Alice is transmitting a signal to Bob, some AEs act as passive eavesdroppers to overhear Alice's signals if they can derive enough information. The rest of the AEs send jamming signals if they can effectively block Alice's signal to Bob. Each AE can either eavesdrop on Alice or jam Bob, under a half-duplex constraint. Here $q_m \in \{e, j\}, m \in [1, M]$, corresponding to eavesdropping and jamming, denotes the specific attack mode of AE_m. Hence, the sets of the passive eavesdroppers and the active jammers can be denoted by Φ_E and Φ_J, respectively, where $|\Phi_J| + |\Phi_E| = M$.

Considering the low mobility of low-altitude UAVs, all the channels are assumed to be quasi-static fading, i.e., the channel gains are constant with each transmission block. Besides, the channel gains between the UAVs follow the free-space path loss model, which is determined by the distance between the UAVs, i.e.,

$$g_{i,j} = \beta_0 d_{i,j}^{-\eta} = \frac{\beta_0}{\left(\sqrt{\|\zeta_i - \zeta_j\|^2}\right)^{\eta}}, \tag{1}$$

where β_0 is the channel power gain at the reference distance of $d_0 = 1m$, $d_{i,j}$ is the distance from node i to node j, ζ_i is the coordinate of node i, and η is the path loss exponent.

Remark 1. Since each AE can eavesdrop or interfere adaptively by learning the communication environment, Alice and other AEs can monitor the AE's position when it chooses to jam. So, we assume that the number and locations of all nodes (legitimate communication pairs and all AEs) are available between each other via a priori measurement following the above analysis. In addition, as the AE considered in this paper can only eavesdrop and interfere, at each time slot, each node judged other's actions by sensing the jamming signal. If one AE does not jam, other nodes consider that it chooses to eavesdrop.

At each time slot, Alice first sends a normalized signal x_a with transmit power P_s. Then, all AEs conduct different attack modes by learning others' strategies. The legitimate link and all passive eavesdroppers suffer interference from all active jammers. The interference to legitimate link and the k^{th} passive eavesdropper ($k \in \Phi_E$) is given by $\sum_{j\in\Phi_J} P_j g_{j,b}$ and $\sum_{j\in\Phi_J} P_j g_{j,k}$, where P_j is the jamming power.

The received signal at Bob can be expressed as

$$y_b = \sqrt{P_s g_{a,b}} x_a + \sum_{j\in\Phi_J} \sqrt{P_j g_{j,b}} x_j + n_b, \tag{2}$$

where $n_b \sim CN(0, \sigma_n^2)$ is the additive white Gaussian noise (AWGN) at Bob. The received SINR at Bob can be expressed as

$$r_{ab} = \frac{P_s \omega_0 d_{a,b}^{-\eta}}{I_{J,B} + 1}, \tag{3}$$

where $\omega_0 = \beta_0/\sigma_n^2$ and $I_{J,B} = \sum_{j\in\Phi_J} P_j \omega_0 d_{j,b}^{-\eta}$ that denotes the interference from all AEs who choose to jam. We can obtain the data rate of the Alice-Bob link as

Table 1: A summary of the proposed literature about AE.

Ref.	FD/HD	Attacker's antennas	Attack mode and phase	Attacker number	Solution		Advantages	Disadvantages
					Theory	Algorithm		
24 (2016)	FD	Multiple antennas	Eavesdrop and jam during data transmission	1	Stackelberg game	Primal-dual interior-point method	Consider the physical layer security issue for multi-channel wireless communications in the presence of an FD active eavesdropper.	Not adaptive
25 (2011)	FD	Multiple antennas	Eavesdrop and jam during data transmission	1	Gradient projection + fixed-point iteration algorithm		Solve the worst-case jamming covariance for arbitrary and Gaussian input signaling.	Not adaptive
26 (2017)	FD	Multiple antennas	Eavesdrop and jam during data transmission	1	Convex optimization	Primal decomposition iterative algorithm	The channel state information is uncertain.	Not adaptive
27 (2017)	HD	Multiple antennas	Eavesdrop and jam during data transmission	1	Constructing the precoding matrix pair at Alice and Bob.		Optimal antennas allocation of the receiver to maximize the achievable secrecy degrees of freedom.	Not adaptive
28 (2012)	HD	Single antenna	Jam during pilot training and eavesdrop during data transmission	1	Convex optimization		Introduce the active eavesdropper into the pilot contamination phenomenon.	Not adaptive
30 (2017)	HD	Single antenna	Jam during pilot training and eavesdrop during data transmission	1	Random-training-assisted spoofing detection scheme		Adding a random training phase after the conventional pilot training phase.	Not adaptive
32 (2018)	HD	Single antenna	Silence, spoof, eavesdrop, and jam during data transmission	1	Stackelberg game	Single-agent Q-learning algorithm	Solve the physical layer security issue under imperfect channel estimation.	Single-agent Q-learning
33 (2017)	FD	Multiple antennas	Eavesdrop, jam, spoof, or keep silent during data transmission	1	Noncooperative game	Single-agent Q-learning algorithm	Solve the physical layer security issue under imperfect channel estimation.	Not adaptive
34 (2011)	HD	Single antenna	Eavesdrop and jam during data transmission	1	Mixed-strategy based noncooperative nonzero-sum game	Fictitious play-based iterative algorithm	Reveal the impact of the presence of a malicious node on the multi-relay to multi-user choices.	Adaptive
35 (2017)	HD	Single antenna	Eavesdrop, jam, and spoof during data transmission	1	PT-based dynamic smart attack game	Q-learning\WoLF-PHC\DQN algorithms	Apply the prospect theory.	Adaptive

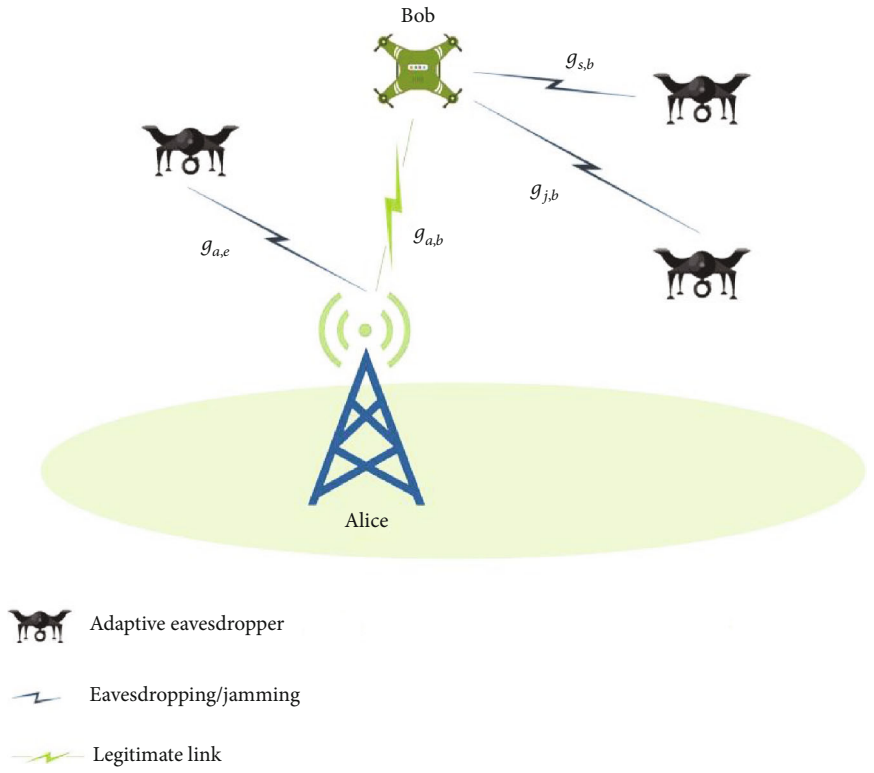


Figure 1: System Model.

$$R_{ab} = \log_2(1 + r_{ab}). \tag{4}$$

Due to Remark 1, each AE can get the other AEs' actions. So, the signal received at the k^{th} passive eavesdropper can be expressed as

$$y_e = \sqrt{P_s g_{a,k}} x_a + \sum_{j \in \Phi_J} \sqrt{P_j g_{j,k}} x_j + n_e, \tag{5}$$

where $n_e \sim CN(0, \sigma_n^2)$ is the AWGN at the k^{th} passive eavesdropper. Similarly, the received SINR at the k^{th} passive eavesdropper can be expressed a

$$r_k = \frac{P_s \omega_0 d_{a,k}^{-\eta}}{I_{J,E} + 1}, \tag{6}$$

where $I_{J,E} = \sum_{j \in \Phi_J} P_j \omega_0 d_{j,k}^{-\eta}$.

Assuming the maximal eavesdropped information is determined by the maximal SINR among all passive eavesdroppers, i.e., $r_E = \max_{k \in \Phi_E} r_k$. We obtain the maximal data rate of the Alice-AE links, which is given as

$$R_{ae} = \log_2(1 + r_E). \tag{7}$$

From (4) and (7), the secrecy rate of Alice can be written as

$$R_a = [R_{ab} - R_{ae}]^+, \tag{8}$$

where $[X]^+$ returns X if X is positive, while returns 0 otherwise.

4. Secure Transmission Game

In this section, we investigate the secure transmission problem with multiple UAV AEs. The interactions between the transmitter and multiple UAV AEs are formulated under the Stackelberg game framework. The optimal power allocations and secrecy rate of Alice and the best attack modes of all AEs are derived by analyzing the equilibrium of the game.

4.1. Secure Transmission Game Formulation. The secure transmission problem of this proposed system can be formulated as a two-stage Stackelberg game. Specifically, Alice is a leader and all AEs are followers. Alice decides its transmit power firstly and all AEs take their action adaptively based on the observation of the leader's action in the sequel. The secure transmission game is formulated as

$$\mathcal{G} = \{\mathcal{N}, \mathcal{P}, \mathcal{Q}, \mathcal{U}_a, \mathcal{U}_m\}. \tag{9}$$

Here, $\mathcal{N} = \{\text{Alice}, \text{AE}_1, \cdots, \text{AE}_m, \cdots, \text{AE}_M\}$ is modeled as the players, and $\mathcal{P} \in [0, P_{\max}]$ and $\mathcal{Q} = \{e, j\}$ are the strategy space of Alice and AE, respectively. Also, $\mathcal{U}_a$ and $\mathcal{U}_m$ are the utility of Alice and AE, respectively.

In this system, Alice wants to send a confidential message and thus naturally intends to maximize its secrecy rate. Meanwhile, the transmission cost is inevitable during the transmission. Therefore, the utility of the leader is the trade-off of the secrecy rate and transmission cost, which can be formulated as

$$\mathcal{U}_a = R_a \ \ln 2 - C_a P_s, \tag{10}$$

where C_a denotes the cost of the unit transmit power of Alice. For computational convenience, we multiply the data rate by a coefficient ln 2.

The objective of the leader is to solve the following problem to obtain the optimal power allocation:

$$P_s^* = \arg \max_{P_s \in [0, P_{\max}]} \mathcal{U}_a(P_s, q_m^*, q_{-m}^*), \tag{11}$$

where $q_1^*, \cdots, q_m^*$ denotes the optimal action of all AEs.

On the other hand, each AE attempts to minimize the secrecy rate of Alice by changing its attack mode adaptively according to Alice's transmit power. Therefore, we formulate the utility of AE_m with the trade-off of the secrecy rate and its attack cost as follows

$$\mathcal{U}_m = -R_a - \theta_{q_m}, \theta_{q_m} \in \{e, j\}, \tag{12}$$

where θ_e and θ_j denotes the cost of each AE to perform as the passive eavesdropper and active jammer, respectively. We assume that θ_e is related to the R_{ae}, i.e., $\theta_e = C_e R_{ae}$, where C_e denotes the cost of unit rate of R_{ae}. $\theta_j = C_j P_j$, where C_j denotes the cost of the unit transmit power of jammer.

To calculate the utility of a single AE accurately, at each time slot, when Alice is transmitting a signal to Bob, we divide all AEs into three parts, which are denoted as Φ_E^{-m}, Φ_J^{-m}, and AE_m, respectively, i.e., $|\Phi_E^{-m}| + |\Phi_J^{-m}| + |\mathrm{AE}_m| = M$. Φ_E^{-m} is the set of passive eavesdroppers except AE_m and Φ_J^{-m} is the set of active jammers except AE_m.

If AE_m decides to act as a passive eavesdropper, the R_a can be expressed as

$$\begin{aligned} R_a &= [R_{ab} - R_{ae}]^+ = [\log_2(1 + r_{ab}) - \log_2(1 + r_E)]^+ \\ &= \left[\log_2\left(1 + \frac{P_s \omega_0 d_{a,b}^{-\eta}}{I_{J,B}^{-m} + 1}\right) - \log_2\left(1 + \max\left(\max_{k \in \Phi_E^{-m}}(r_k), r_m\right)\right)\right]^+ \\ &= \left[\log_2\left(1 + \frac{P_s \omega_0 d_{a,b}^{-\eta}}{I_{J,B}^{-m} + 1}\right) - \log_2(1 + \max(r_E^{-m}, r_m))\right]^+, \end{aligned} \tag{13}$$

where $I_{J,B}^{-m}$ is the interference received at Bob from Φ_J^{-m}, and r_m is the SINR of AE_m.

Similarly, if AE_m selects to jam, R_a can be expressed as

$$\begin{aligned} R_a &= [R_{ab} - R_{ae}]^+ = [\log_2(1 + r_{ab}) - \log_2(1 + r_E)]^+ \\ &= \left[\log_2\left(1 + \frac{P_s \omega_0 d_{a,b}^{-\eta}}{I_{J,B}^{-m} + I_m + 1}\right) - \log_2\left(1 + \max_{k \in \Phi_E^{-m}}(r_k)\right)\right]^+, \end{aligned} \tag{14}$$

where I_m is the jamming power of AE_m, and r_E^{-m} is the maximal SINR among all passive eavesdroppers in Φ_E^{-m}.

In conclusion, $\mathcal{U}_m$ can be expressed as

$$\mathcal{U}_m = \begin{cases} \left[\log_2\left(1 + \frac{P_s \omega_0 d_{a,b}^{-\eta}}{I_{J,B}^{-m} + 1}\right) - \log_2(1 + \max(r_E^{-m}, r_m))\right]^+ - C_e R_{ae}, q_m = e \quad (a), \\ \left[\log_2\left(1 + \frac{P_s \omega_0 d_{a,b}^{-\eta}}{I_{J,B}^{-m} + I_m + 1}\right) - \log_2\left(1 + \max_{k \in \Phi_E^{-m}}(r_k)\right)\right]^+ - C_j P_j, q_m = j \quad (b). \end{cases} \tag{15}$$

Similarly, the objective of AE_m is to solve the following problem:

$$q_m^* = \arg \max_{q_m \in \{e,j\}} \mathcal{U}_m(P_s^*, q_m, q_{-m}^*), \tag{16}$$

where q_{-m}^* denotes the optimal action of all AEs except AE_m.

4.2. Analysis of Strategy Equilibrium. Now, we will analyze the proposed Stackelberg game model and solve the optimization subproblems of (11) and (16). As a follower, each AE will adjust its attack mode after sensing Alice's strategy. Therefore, the subgame of followers is analyzed firstly.

Proposition 1. *Given the strategy of Alice, the optimal attack mode strategy of AE_m is expressed as (17) if (17(a)) and (17(b)) hold.*

$$q_m^*(P_s) = \begin{cases} e, \quad \text{if} \quad C_j P_j \geq \log_2\left[\frac{\left(1 + P_s^* \omega_0 d_{a,b}^{-\eta} + I_{J,B}^{-m*}\right)\left(I_{J,B}^{-m*} + I_m + 1\right)\left(1 + r_E^{-m*}\right)}{\left(1 + P_s^* \omega_0 d_{a,b}^{-\eta} + I_{J,B}^{-m*} + I_m\right)\left(I_{J,B}^{-m*} + 1\right)\left(1 + \max(r_E^{-m*}, r_m)\right)^{1-C_e}}\right] \quad (a), \\ j, \quad \text{if} \quad C_j P_j \leq \log_2\left[\frac{\left(1 + P_s^* \omega_0 d_{a,b}^{-\eta} + I_{J,B}^{-m*}\right)\left(I_{J,B}^{-m*} + I_m + 1\right)\left(1 + r_E^{-m*}\right)}{\left(1 + P_s^* \omega_0 d_{a,b}^{-\eta} + I_{J,B}^{-m*} + I_m\right)\left(I_{J,B}^{-m*} + 1\right)\left(1 + \max(r_E^{-m*}, r_m)\right)^{1-C_e}}\right] \quad (b), \end{cases} \tag{17}$$

where P_s^* is the optimal power allocation, and $I_{J,B}^{-m*} = \sum_{j\in\Phi_J^{-m*}} P_j\omega_0 d_{j,b}^{-\eta}$ denotes the interference from Φ_J^{-m*}, in which each AE chooses to jam as an optimal strategy, $r_E^{-m*} = \max_{k\in\Phi_E^{-m*}} (r_k)$ denotes the maximal SINR among all AEs in Φ_E^{-m*} where each AE chooses to overhear as an optimal strategy.

Proof. If (17(a)) holds, from (12), we have

$$\begin{aligned}\mathcal{U}_m(P_s^*, e, q_{-m}^*) - \mathcal{U}_m(P_s^*, j, q_{-m}^*) &= -\left[\log_2\left(1+\frac{P_s^*\omega_0 d_{a,b}^{-\eta}}{I_{J,B}^{-m*}+1}\right) - \log_2(1+\max(r_E^{-m*}, r_m))\right] + \left[\log_2\left(1+\frac{P_s\omega_0 d_{a,b}^{-\eta}}{I_{J,B}^{-m*}+I_m+1}\right) - \log_2(1+r_E^{-m*})\right] \\ &\quad - C_e\log_2(1+\max(r_E^{-m*}, r_m)) + C_jP_j \\ &= -\log_2\left[\frac{\left(1+P_s^*\omega_0 d_{a,b}^{-\eta}+I_{J,B}^{-m*}\right)\left(I_{J,B}^{-m*}+I_m+1\right)\left(1+r_E^{-m*}\right)}{\left(1+P_s^*\omega_0 d_{a,b}^{-\eta}+I_{J,B}^{-m*}+I_m\right)\left(I_{J,B}^{-m*}+1\right)\left(1+\max(r_E^{-m*}, r_m)\right)^{1-C_e}}\right] + C_jP_j \geq 0.\end{aligned} \tag{18}$$

If (17(b)) holds, from (12), we have

$$\begin{aligned}&\mathcal{U}_m(P_s^*, j, q_{-m}^*) - \mathcal{U}_m(P_s^*, e, q_{-m}^*) \\ &= \left[\log_2\left(1+\frac{P_s^*\omega_0 d_{a,b}^{-\eta}}{I_{J,B}^{-m*}+1}\right) - \log_2(1+\max(r_E^{-m*}, r_m))\right] \\ &\quad - \left[\log_2\left(1+\frac{P_s\omega_0 d_{a,b}^{-\eta}}{I_{J,B}^{-m*}+I_m+1}\right) - \log_2(1+r_E^{-m*})\right] \\ &\quad + C_e\log_2(1+\max(r_E^{-m*}, r_m)) - C_jP_j \\ &= \log_2\left[\frac{\left(1+P_s^*\omega_0 d_{a,b}^{-\eta}+I_{J,B}^{-m*}\right)\left(I_{J,B}^{-m*}+I_m+1\right)\left(1+r_E^{-m*}\right)}{\left(1+P_s^*\omega_0 d_{a,b}^{-\eta}+I_{J,B}^{-m*}+I_m\right)\left(I_{J,B}^{-m*}+1\right)\left(1+\max(r_E^{-m*}, r_m)\right)^{1-C_e}}\right] \\ &\quad - C_jP_j \geq 0.\end{aligned} \tag{19}$$

Thus (17) holds.

As shown in Proposition 1, if passive eavesdropping can bring worse secrecy rate and less cost than active jamming, the AE will select to overhear and vice versa.

As the leader of the game, Alice first chooses to transmit power. The optimal power strategy of Alice can be derived by solving (11), which is revealed in Proposition 2.

Proposition 2. *The optimal power allocation is P_s^*, which satisfies the following equation:*

$$\begin{cases}\dfrac{\omega_0 d_{a,b}^{-\eta} - \omega_0\max\limits_{k\in\Phi_E^*}\left(d_{a,k}^{-\eta}/I_{J,E}^*+1\right)I_{J,B} - \omega_0\max\limits_{k\in\Phi_E^*}\left(d_{a,k}^{-\eta}/I_{J,E}^*+1\right)}{\left(1+I_{J,B}^*+P_s^*\omega_0 d_{a,b}^{-\eta}\right)\left(1+P_s^*\omega_0\max\limits_{k\in\Phi_E^*}\left(d_{a,k}^{-\eta}/I_{J,E}^*+1\right)\right)} = C_a \quad (a), \\ 0\leq P_s^*\leq P_{\max} \quad (b),\end{cases} \tag{20}$$

if (21(a)) and (21(b)) hold.

$$\begin{cases}\max\limits_{k\in\Phi_E^*}\left(\dfrac{d_{a,k}^{-\eta}}{I_{J,E}^*+1}\right) < \dfrac{d_{a,b}^{-\eta}}{I_{J,B}^*+1} \quad (a), \\ \dfrac{\omega_0 d_{a,b}^{-\eta} - \omega_0\max\limits_{k\in\Phi_E^*}\left(d_{a,k}^{-\eta}/I_{J,E}^*+1\right)I_{J,B} - \omega_0\max\limits_{k\in\Phi_E^*}\left(d_{a,k}^{-\eta}/I_{J,E}^*+1\right)}{\left(1+I_{J,B}^*+P_s^*\omega_0 d_{a,b}^{-\eta}\right)\left(1+P_s^*\omega_0\max\limits_{k\in\Phi_E^*}\left(d_{a,k}^{-\eta}/I_{J,E}^*+1\right)\right)} \leq C_a \leq \dfrac{\omega_0 d_{a,b}^{-\eta}}{1+I_{J,B}^*} - \omega_0\max\limits_{k\in\Phi_E^*}\left(\dfrac{d_{a,k}^{-\eta}}{I_{J,E}^*+1}\right) \quad (b),\end{cases} \tag{21}$$

where $I_{J,B}^*$ and $I_{J,E}^*$ denotes the interference from Φ_J^* in which each AE chooses to jam as an optimal strategy to Bob and the k^{th} passive eavesdropper, respectively.

Proof. We obtain the following differential equation describing the evolution of the utility of Alice:

$$\frac{\partial\mathcal{U}_a}{\partial P_s} = \frac{\omega_0 d_{a,b}^{-\eta} - \omega_0\max\limits_{k\in\Phi_E^*}\left(d_{a,k}^{-\eta}/I_{J,E}^*+1\right)I_{J,B} - \omega_0\max\limits_{k\in\Phi_E^*}\left(d_{a,k}^{-\eta}/I_{J,E}^*+1\right)}{\left(1+I_{J,B}^*+P_s^*\omega_0 d_{a,b}^{-\eta}\right)\left(1+P_s^*\omega_0\max\limits_{k\in\Phi_E^*}\left(d_{a,k}^{-\eta}/I_{J,E}^*+1\right)\right)} - C_a, \tag{22}$$

$$\frac{\partial^2 \mathcal{U}_a}{\partial P_s^2} = \left[\frac{\omega_0 d_{a,b}^{-\eta}}{\left(1 + I_{J,B}^* + P_s \omega_0 d_{a,b}^{-\eta}\right)} + \frac{\omega_0 \max\limits_{k \in \Phi_E^*} \left(d_{a,k}^{-\eta} / I_{J,E}^* + 1\right)}{\left(1 + P_s \omega_0 \max\limits_{k \in \Phi_E^*} \left(d_{a,k}^{-\eta} / I_{J,E}^* + 1\right)\right)} \right] \cdot \left[\frac{\omega_0 \max\limits_{k \in \Phi_E^*} \left(d_{a,k}^{-\eta} / I_{J,E}^* + 1\right)}{\left(1 + P_s \omega_0 \max\limits_{k \in \Phi_E^*} \left(d_{a,k}^{-\eta} / I_{J,E}^* + 1\right)\right)} - \frac{\omega_0 d_{a,b}^{-\eta}}{\left(1 + I_{J,B}^* + P_s \omega_0 d_{a,b}^{-\eta}\right)} \right]. \tag{23}$$

If (21(a)) holds, (23) is less than zero. Thus, we have

$$\frac{\partial^2 \mathcal{U}_a}{\partial P_s^2} < 0, \tag{24}$$

which indicates that $\partial \mathcal{U}_a / \partial P_s$ monotonically decreases with P_s. Therefore, if (21(b)) holds, we have

$$\left. \frac{\partial \mathcal{U}_a}{\partial P_s} \right|_{P_s = 0} = \frac{\omega_0 d_{a,b}^{-\eta}}{\left(1 + I_{J,B}^*\right)} - \omega_0 \max_{k \in \Phi_E^*} \left(\frac{d_{a,k}^{-\eta}}{I_{J,E}^* + 1} \right) - C_a > 0, \tag{25}$$

$$\left. \frac{\partial^2 \mathcal{U}_a}{\partial P_s^2} \right|_{P_s = P_{\max}} = \frac{\omega_0 d_{a,b}^{-\eta}}{\left(1 + I_{J,B}^* + P_{\max} \omega_0 d_{a,b}^{-\eta}\right)} - \frac{\omega_0 \max\limits_{k \in \Phi_E^*} \left(d_{a,k}^{-\eta} / I_{J,E}^* + 1\right)}{\left(1 + P_{\max} \omega_0 \max\limits_{k \in \Phi_E^*} \left(d_{a,k}^{-\eta} / I_{J,E}^* + 1\right)\right)} - C_a < 0, \tag{26}$$

indicating that there is a sole solution to $\partial \mathcal{U}_a / \partial P_s = 0$, given in (20(a)). From (22)–(24), we can find that $\mathcal{U}_a(P_s, q_m^*, q_{-m}^*)$ increases with P_s, if $P_s < P_s^*$, while it decreases otherwise. Thus, (11) also holds and (P_s, q_m^*, q_{-m}^*) is a Nash Equilibrium (NE) of the game. In this way, we have completed the proof of Proposition 2.

As shown in Proposition 2, Alice stops the transmission when (21(b)) does not hold. In other words, Alice will stop the transmission under the circumstances that radio channel degradation is serious and the security cannot be guaranteed.

Another NE $(P_{\max}, q_m^*, q_{-m}^*)$ is revealed in Proposition 3.

Proposition 3. *The secure game has the NE $(P_{\max}, q_m^*, q_{-m}^*)$ if (21(a)) and the following equation hold:*

$$\frac{\omega_0 d_{a,b}^{-\eta}}{\left(1 + I_{J,B}^* + P_{\max} \omega_0 d_{a,b}^{-\eta}\right)} - \frac{\omega_0 \max\limits_{k \in \Phi_E^*} \left(d_{a,k}^{-\eta} / I_{J,E}^* + 1\right)}{\left(1 + P_{\max} \omega_0 \max\limits_{k \in \Phi_E^*} \left(d_{a,k}^{-\eta} / I_{J,E}^* + 1\right)\right)} > C_a. \tag{27}$$

Proof. (21(a)) has been discussed above.

Therefore, if (27) holds, we have

$$\frac{\partial \mathcal{U}_a}{\partial P_s} \geq \left. \frac{\partial \mathcal{U}_a}{\partial P_s} \right|_{P_s = P_{\max}} \geq 0, \forall 0 \leq P_s \leq P_{\max}, \tag{28}$$

which indicates that $\mathcal{U}_a$ monotonically increases with P_s, $(P_{\max}, q_m^*, q_{-m}^*)$ is also an NE of the game. In this way, we have completed the proof of Proposition 3.

As shown in Proposition 3, low transmission costs in (27) will make Alice select the maximum transmit power to transmit the signals.

5. Hierarchical Reinforcement Learning Framework for Secure Transmission Game

The proposed UAV secure communication problem with multiple AEs has been formulated as a Stackelberg game, which belongs to the category of two-stage dynamic game and has a significant two-layer game structure. Alice and all AEs become intelligent agents and have the learning ability to automatically optimize their configuration. Besides, the mixed-strategy is applied by both sides of the communication to confuse each other. In this section, we apply a hierarchical RL framework to derive the mixed-strategy equilibrium and implement the UAV secure communication.

5.1. Analysis of Mixed-Strategy Equilibrium. Considering the actual wireless communication scenario, we assume that Alice has a finite and discretized power set. Specifically, a policy of Alice at time slot t is defined to be a probability vector $\pi^t = (\pi_1^t, \pi_2^t, \cdots \pi_L^t)$, where π_l^t means the probability with which Alice chooses action (power level) P_l from a finite discrete set $\mathcal{P}$, which satisfies $\sum_{l=1}^{L} \pi_l^t = 1$. Similarly, $\delta_m^t = (\delta_{m,1_m}^t, \delta_{m,2_m}^t)$ denotes the policy of AE_m at time slot t, where δ_{m,i_m}^t means the probability with which AE_m chooses action (attack mode) $\mathcal{Q}_i$ from a finite discrete set $\mathcal{Q}$, which satisfies $\sum_{i=1}^{2} \delta_{m,i_m}^t = 1$.

Based on the above analysis, we have the following definition of an SE for the hierarchical RL framework based on Eqs. (10) and (12). Alice's objective is to maximize its revenue as

$$\pi^* = \arg \max_{\pi} \mathcal{U}_a(\pi, \delta_m^*, \delta_{-m}^*), \tag{29}$$

Similarly, each AE's objective is

$$\delta_m^* = \arg \max_{\delta_m} \mathcal{U}_m(\pi^*, \delta_m, \delta_{-m}^*), \tag{30}$$

Then, we will define the SE in a hierarchical reinforcement learning framework.

Definition 1. A stationary policy profile $(\pi^*, \delta_m^*, \delta_{-m}^*)$ is the SE for hierarchical RL framework if the followings hold.

$$\begin{cases} \mathcal{U}_a(\pi^*, \delta_m^*, \delta_{-m}^*) \geq \mathcal{U}_a(\pi, \delta_m^*, \delta_{-m}^*)(a) \\ \mathcal{U}_m(\pi^*, \delta_m^*, \delta_{-m}^*) \geq \mathcal{U}_m(\pi^*, \delta_m, \delta_{-m}^*)(b) \end{cases}. \tag{31}$$

Proposition 4. *For the proposed hierarchical RL framework, there exists Alice's stationary policy and an AEs' NE policy that form an SE.*

Hierarchical Q-learning Based Power Allocation Algorithm

Initialize $t = 0$, $Q_a^t(P_l) = 0$, $Q_m^t(q_{m,i_m}^t) = 0$, $\pi_l^t(P_l) = 1/L$, $\delta_{m,i_m}^t = 1/2$, $m \in \mathcal{M} \setminus \{Alice\}$;
Loop:
$t = t + 1$;
Update Alice's policies $\pi_l^t(P_l)$ and $AE_m{}'s$ policies $\delta_{m,i_m}^t(q_{m,i_m})$ according to (35) and (33), respectively;
Alice chooses the action P_l with π_l^t;
Each AE_m sensing the Alice's transmit power, and selects q_{m,i_m} with δ_{m,i_m}^t;
Alice updates $\mathcal{U}_a(P_l, q_{m,i_m}^{t+1}, q_{-m,i_{-m}}^{t+1})$ according to (10), and each AE_m updates $\mathcal{U}_m(P_l^{t+1}, q_{m,i_m}^t, q_{-m,i_{-m}}^{t+1})$ according to (12);
Alice and all AEs update Q-values according to (34) and (32), respectively;
End Loop;

Algorithm 1.

Proof. If the Alice follows a stationary policy π, the Stackelberg game is simplified into an M-player hierarchical RL game. It has been shown in [42] that every finite strategic-form game has a mixed policy equilibrium. As a result, there always exists an NE (π) in our formulation of the discrete power allocation game given Alice's policy π. The rest of the proof follows directly from the definition of an SE and is thus omitted for brevity.

5.2. Hierarchical Q-Learning Based Power Allocation Algorithm. In the proposed UAV secure transmission game, since there is no information exchange between Alice and AEs, both sides can only maximize their expected utilities through repeated interactions with each other. When the action taken by the agent (Alice or AEs) brings positive feedback to the agent, the agent will strengthen the action, otherwise the agent will weaken the action. Agents constantly adjust their strategies based on the feedback to achieve optimal long-term returns. Thus, a hierarchical Q-learning based power allocation algorithm (HQLA) is adopted, where each agent's policy is parameterized through the Q-function that characterizes the relative expected utility of a particular action.

To be specific, for the follower's learning, let $Q_m^t(q_{m,i_m}^t)$ denote the corresponding Q-function of $\mathrm{AE}_m{}'s$ action q_{m,i_m}^t based on current police δ_{m,i_m}^t at time slot t. Then, after conducting the action q_{m,i_m}^t, the corresponding Q-value is updated as follows

$$Q_m^{t+1}\left(q_{m,i_m}\right) = Q_m^t\left(q_{m,i_m}^t\right) + \alpha\left(\mathcal{U}_m\left(P_l^{t+1}, q_{m,i_m}^t, q_{-m,i_{-m}}^{t+1}\right) - Q_m^t\left(q_{m,i_m}^t\right)\right), \tag{32}$$

where $\alpha \in [0, 1)$ is the learning rate and $\mathcal{U}_m(P_l^{t+1}, q_{m,i_m}^t, q_{-m,i_{-m}}^{t+1})$ is the utility of AE_m at time slot $t + 1$.

Each AE updates its policy based on Boltzmann distribution

$$\delta_{m,i_m}^{t+1}\left(q_{m,i_m}\right) = \frac{\exp\left[Q_m^t\left(q_{m,i_m}^t\right)/\tau\right]}{\sum_{q_{s,i_s}\in\mathcal{Q}}\exp\left[Q_m^t\left(q_{s,j_s}^t\right)/\tau\right]}, \tag{33}$$

where temperature τ controls the trade-off between exploration and exploitation, i.e., for $\tau \to 0$, AE_m greedily chooses the policy corresponding to the maximum Q-value which means pure exploitation, whereas for $\tau \to \infty$, $\mathrm{AE}_m{}'s$ policy is completely random which means pure exploration [43]. Accordingly, the Q-value of Alice is updated as follows

$$Q_a^{t+1}(P_l) = Q_a^t(P_l) + \alpha\left(\mathcal{U}_a\left(P_l, q_{m,i_m}^{t+1}, q_{-m,i_{-m}}^{t+1}\right) - Q_a^t(P_l)\right), \tag{34}$$

where $\mathcal{U}_a(P_l, q_{m,i_m}^{t+1}, q_{-m,i_{-m}}^{t+1})$ is the utility of Alice at time slot $t + 1$. Then, Alice updates its policy based on Boltzmann distribution

$$\pi_l^{t+1}(P_l) = \frac{\exp\left[Q_a^t(P_l)/\tau\right]}{\sum_{P_j\in\mathcal{P}}\exp\left[Q_a^t(P_j)/\tau\right]}. \tag{35}$$

Now, we present the detailed description of the Q-learning based hierarchical RL algorithm.

5.3. Convergence Analysis of Algorithm 1. The learning algorithm results in a stochastic process of choosing a power level, so we need to investigate the long-term behavior of the learning procedure. Along with the discussion in [43], we obtain the following differential equation describing the evolution of the Q-values:

$$\frac{dQ_a^{t+1}(P_l)}{dt} = \alpha\left(\mathcal{U}_a\left(P_l, q_{m,i_m}^{t+1}, q_{-m,i_{-m}}^{t+1}\right) - Q_a^t(P_l)\right), \tag{36}$$

$$\frac{dQ_m^{t+1}\left(q_{m,i_m}\right)}{dt} = \alpha\left(\mathcal{U}_m\left(P_l^{t+1}, q_{m,i_m}^t, q_{-m,i_{-m}}^{t+1}\right) - Q_m^t\left(q_{m,i_m}^t\right)\right). \tag{37}$$

In the following, we would like to express the dynamics in terms of strategies rather than the Q-values. Toward this end, we differentiate (35) with respect to time t and use (36). Similarly, we differentiate (33) with respect to time t and use (37).

We can obtain the equations like (38) and (39).

$$\frac{d\pi_l^{t+1}(P_l)}{dt} = \pi_l^{t+1}(P_l)\frac{\alpha}{\tau}\left\{\left[\mathcal{U}_a(P_l) - \sum_{j\in\mathcal{P}}\pi_j^{t+1}(P_j)\mathcal{U}_a(P_j)\right] - \tau\sum_{j\in\mathcal{P}}\pi_j^{t+1}(P_j)\ln\left(\frac{\pi_l^{t+1}(P_l)}{\pi_j^{t+1}(P_j)}\right)\right\}, \tag{38}$$

$$\frac{d\delta_{m,i_m}^{t+1}\left(q_{m,i_m}\right)}{dt} = \delta_{m,i_m}^{t+1}\left(q_{m,i_m}\right)\frac{\alpha}{\tau}\left\{\left[\mathcal{U}_m\left(q_{m,i_m}\right) - \sum_{i_s\in\mathcal{Q}}\delta_{s,i_s}^{t+1}\left(q_{s,i_s}\right)\mathcal{U}_m\left(q_{s,i_s}\right)\right] - \tau\sum_{i_s\in\mathcal{Q}}\delta_{s,i_s}^{t+1}\left(q_{s,i_s}\right)\ln\left(\frac{\delta_{m,i_m}^{t+1}\left(q_{m,i_m}\right)}{\delta_{s,i_s}^{t+1}\left(q_{s,i_s}\right)}\right)\right\}. \tag{39}$$

The steady-state strategy profile $z^s = (\pi^s(P_l), \delta_{m,i_m}^s(q_{m,i_m}))$ can be obtained [43].

$$\pi^s(P_l) = \frac{\exp\left[Q_a^t(P_l)/\tau\right]}{\sum_{P_j\in\mathcal{P}}\exp\left[Q_a^t(P_j)/\tau\right]}, \tag{40}$$

$$\delta_{m,i_m}^s\left(q_{m,i_m}\right) = \frac{\exp\left[Q_m^t\left(q_{m,i_m}^t\right)/\tau\right]}{\sum_{q_{s,i_s}\in\mathcal{Q}}\exp\left[Q_m^t\left(q_{s,j_s}^t\right)/\tau\right]}. \tag{41}$$

Let $Z^t = (z_1^t, \cdots, z_N^t)$ the strategy profile of all players at time slot t. In the following analysis, we resort to an ordinary differential equation (ODE) whose solution approximates the convergence of Z^t. The right-hand side of (38) and (39) can be represented by a function $f(Z^t)$ as $\alpha \to 0$. Z^t will converge weakly to $Z^* = (\pi_0^*, \delta_0^*)$, which is the solution to

$$\frac{dZ}{dt} = f(Z), Z^0 = Z_0. \tag{42}$$

Proposition 5. *The HQLA can discover a mixed-strategy SE.*

Proof. We prove this by contradiction. Suppose that the process generated by (33) and (35) converges to a non-SE. But the solutions of (42) are by definition stationary points. This implies that HQLA will only converge to stationary points. This means that stationary points that are not SEs are stable, which is contradicting Proposition 4.

6. Simulation Results

Simulations are carried out to evaluate the performance of the proposed power allocation strategies against multiple UAV AEs. This scenario has one transmitter-receiver pair and three UAV AEs denoted as Alice, Bob, AE_1, AE_2, and AE_3, respectively. We set up a scenario network where all the UAVs are distributed in a 200 m $*$ 200 m region. The system parameters are chosen for some typical scenarios including the cost of unit transmit power and jamming power, i.e., $C_a = C_j = 0.1$ and $C_e = 0.5$, the path loss exponent $\eta = 2$ and $\omega_0 = 80$.

Figure 2 shows the expected utilities of the leader under different algorithms. We can find that the expected utility achieved by the proposed HQLA is significantly lower than the single-agent Q-learning algorithm (SAQL). This is because in SAQL, only Alice applies the reinforcement learning mechanism to maximize the secrecy rate but all AEs' behaviors constituting joint actions are considered to be stated in the Q-learning algorithm which means each AE cannot choose the optimal strategy adaptively to maximize its utility. While in HQLA, each AE with reinforcement learning ability can maximize its damages to the secrecy rate of the considered system through repeated interactions with Alice's and other AEs' strategies. The comparison of the leader's expected utilities with SAQL implies that the agent's learning ability has a significant impact on its utility. So, the proposed HQLA provides an optimal power allocation strategy in a more hostile case that suffers the adaptive attacks from multiple AEs. On the other hand, the proposed HQLA is superior to the random selection algorithm (RS) because the proposed HQLA may converge to a desirable solution, whereas the RS is an instinctive approach.

Figure 3 shows the cumulative distribution function (CDF) of the convergence of HQLA and SAQL. As observed from Figure 3, we can find that the proposed algorithm converges at about 500 iterations, while the contrast algorithm converges at about 1000 iterations, which means the convergence rate of HQLA is significantly better than SAQL. This is because that all AEs in SAQL select action randomly without learning ability whereas in HQLA, taking the interactions between two sides of communication into account, all AEs make decisions according to the mixed-strategy derived by RL which can obtain the optimal strategy via trials-and-errors. This also means that the learning ability has a significant positive impact on the convergence rate.

Figure 4 presents the strategy selection probabilities evolution of the leader's transmit power. At the very beginning, Alice randomly selects transmit power according to a uniform distribution. As Algorithm 1 iterates, the strategy selection probabilities keep on updating until convergence after about 500 iterations. It is worthy to note that Algorithm 1 under this scenario converges to pure strategy NE points since the probability of selecting one power level is equal to 1, while the probabilities of the other levels of transmit power decrease to 0 if the time slots are large enough. So, the theoretical prediction in Proposition 4 is verified under the existing conditions. Specifically, the $P_{\max}$ in Figure 4(a) as the optimal transmit power is consistent with Proposition 3, and the P_s^* in Figure 4(b) is consistent with Proposition 2.

The leader's expected utility comparison under different C_e is shown in Figure 5(a). It is noted that the steady value of the leader's expected utility increase with the value of C_e growing because that C_e leads to changes in AEs' attack strategies. Specifically, as a rational agent, all AEs choose to interfere with Bob finally in Figure 5(b) because they find the utility of the jammer is higher than eavesdropper according

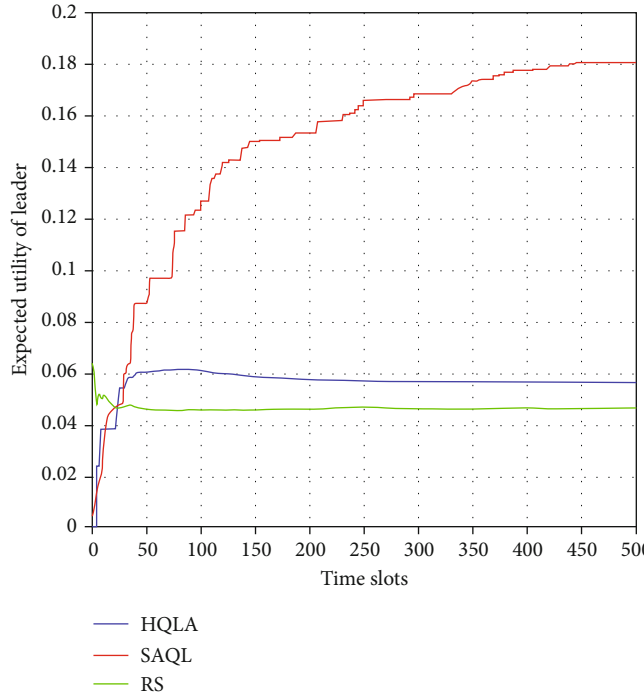


Figure 2: Expected utility of leader.

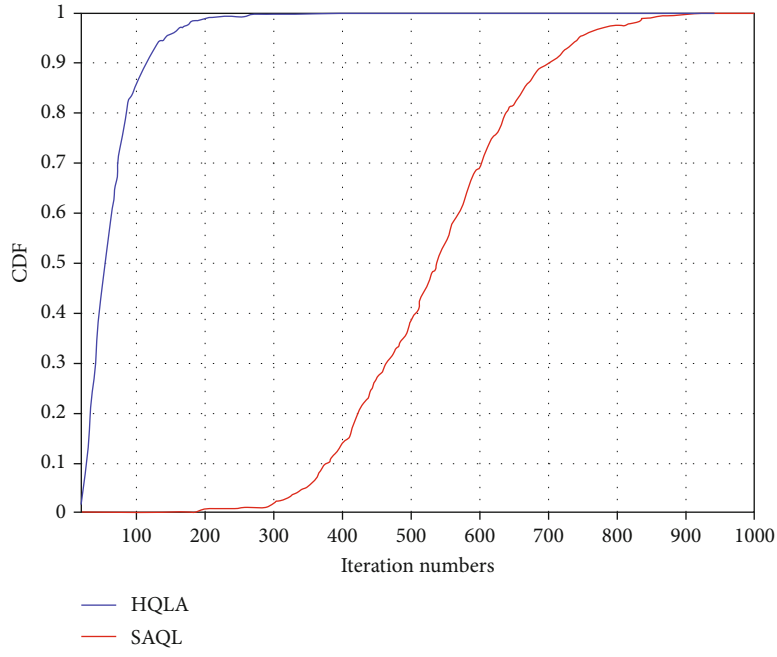
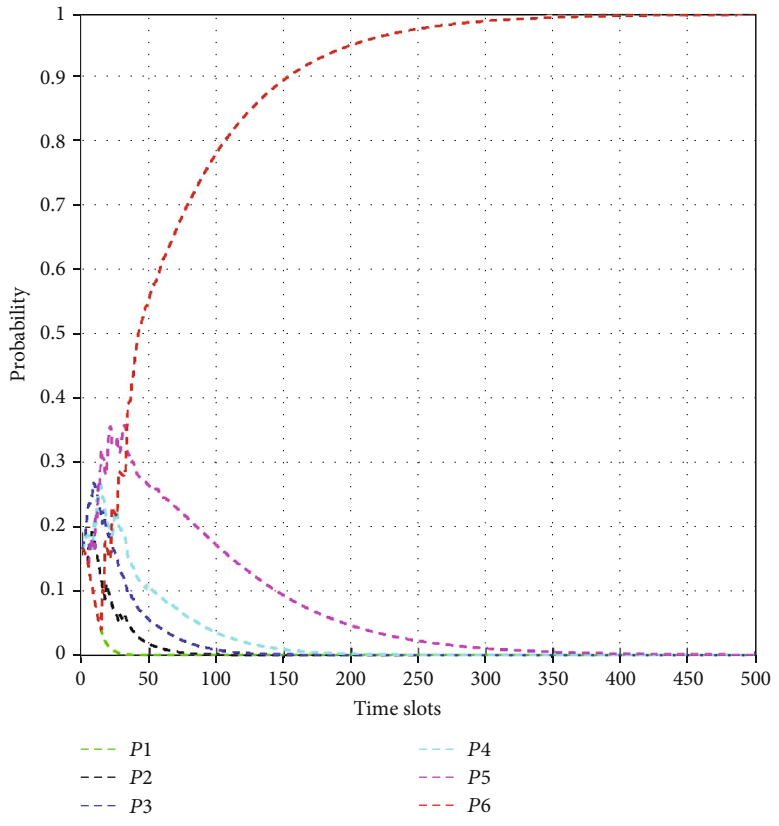
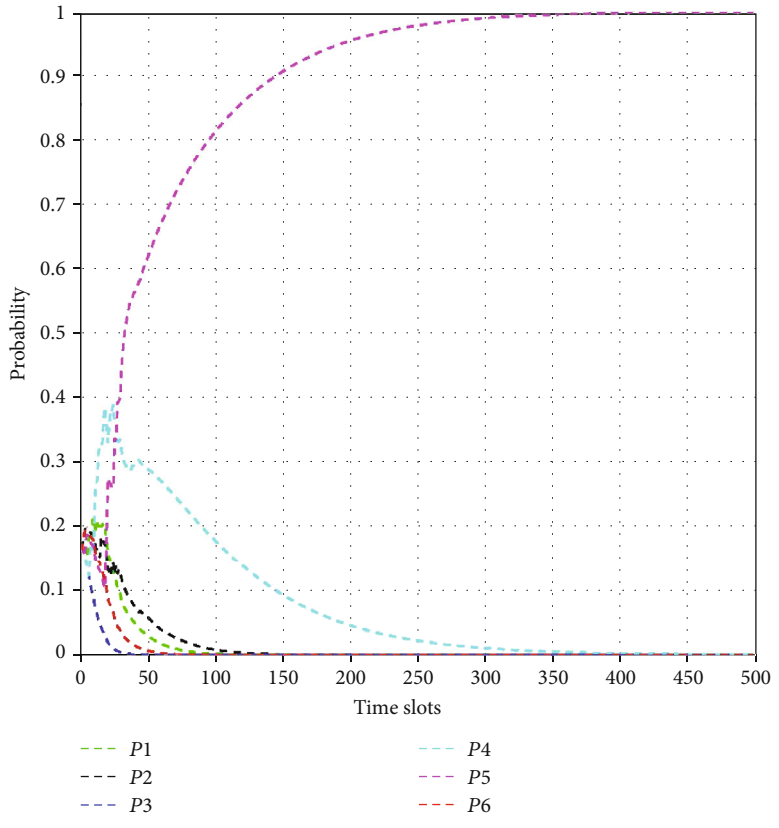


Figure 3: The CDF of convergence of HQLA and SAQL.

to the learning process when the difference of $C_e = 0.8$. Thus, the maximal data rate of the Alice-AE link is zero which means the leader will obtain the maximal secrecy rate and expected utility. Similarly, in Figure 5(d), all AEs find the utility of eavesdropper is higher than jammer when $C_e = 0.2$ and every AE choose to eavesdrop on Alice. As a result, the maximal data rate of the Alice-AE link between all AEs is achieved and the leader suffers the lowest utility. When $C_e = 0.5$ (in Figure 5(c)), according to the utilities of themselves, AE_1 and AE_3 always choose to interfere with Bob and AE_2 prefers eavesdropping which makes the expected utility of leader is between $C_e = 0.8$ and $C_e = 0.2$. In addition, it is worthy to note that the attack strategies of all AEs have a pure strategy equilibrium since the probability of selecting one attack mode is equal to 1 while the probability of another attack mode decreases to 0.

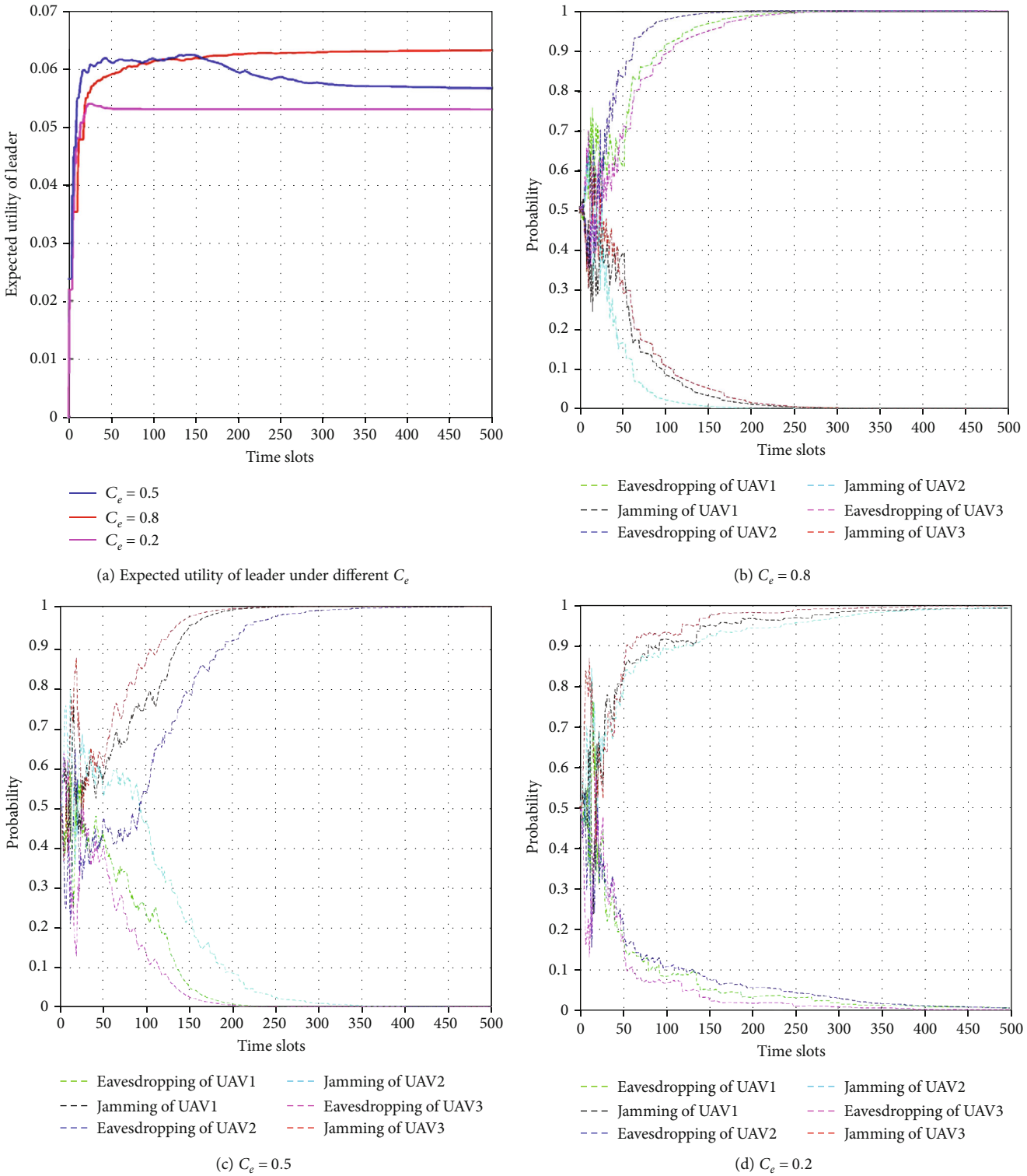


(a)



(b)

Figure 4: Power allocation probabilities of leader.



(a) Expected utility of leader under different C_e

(b) $C_e = 0.8$

(c) $C_e = 0.5$

(d) $C_e = 0.2$

FIGURE 5: Expected utility of leader and attack mode probabilities of all AEs under different C_e.

7. Conclusions and Future Work

In this paper, we have investigated the transmit power optimization problem of secure UAV communication in the presence of multiple UAV AEs. A secure transmission game is formulated to prove the existence of the NE by analyzing the interactions between the legitimate user and AEs. Within a hierarchical game framework, we obtain the optimal transmit power solutions for the legitimate transmissions. Numerical results verified the theoretical analysis and shown that the secrecy performance could be degraded severely by AEs' learning ability. Moreover, the outperformance of the

HQLA's convergence and the impact of the eavesdropping cost on the decision of AE's attack mode is also demonstrated. To take advantage of the UAV's mobility that can bring the potential performance enhancement, in future work, we will devote our efforts to joining the UAV's trajectory and resource allocation optimization against multiple AEs.

Data Availability

The data (figures) used to support the findings of this study are included within the article. Further details can be provided upon request.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (no. 61771487 and no. 61471393) and the National Key R&D Program of China under Grant 2018YFB1801103.

References

[1] V. Mayor, R. Estepa, A. Estepa, and G. Madinabeitia, "Deploying a reliable UAV-aided communication service in disaster areas," *Wireless Communications and Mobile Computing*, vol. 2019, Article ID 7521513, 20 pages, 2019.

[2] X. Fan, C. Huang, B. Fu, S. Wen, and X. Chen, "UAV-assisted data dissemination in delay-constrained VANETs," *Mobile Information Systems*, vol. 2018, Article ID 8548301, 12 pages, 2018.

[3] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Unmanned aerial vehicle with underlaid device-to-device communications: performance and trade-offs," *IEEE Transactions on Wireless Communications*, vol. 15, no. 6, pp. 3949–3963, 2016.

[4] J. Lyu, Y. Zeng, R. Zhang, and T. J. Lim, "Placement optimization of uav-mounted mobile base stations," *IEEE Communications Letters*, vol. 21, no. 3, pp. 604–607, 2017.

[5] Q. Wu, Y. Zeng, and R. Zhang, "Joint trajectory and communication design for multi-uav enabled wireless networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 3, pp. 2109–2121, 2018.

[6] N. H. Motlagh, M. Bagaa, and T. Taleb, "Uav-based iot platform: a crowd surveillance use case," *IEEE Communications Magazine*, vol. 55, no. 2, pp. 128–134, 2017.

[7] Yingbin Liang, H. V. Poor, and S. Shamai, "Secure communication over fading channels," *IEEE Transactions on Information Theory*, vol. 54, no. 6, pp. 2470–2492, 2008.

[8] W. Yang, L. Tao, X. Sun, R. Ma, Y. Cai, and T. Zhang, "Secure on-off transmission in mmwave systems with randomly distributed eavesdroppers," *IEEE Access*, vol. 7, pp. 32681–32692, 2019.

[9] Z. Xiang, W. Yang, G. Pan, Y. Cai, and Y. Song, "Physical layer security in cognitive radio inspired Noma network," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 3, pp. 700–714, 2019.

[10] X. Sun, W. Yang, Y. Cai, L. Tao, Y. Liu, and Y. Huang, "Secure transmissions in wireless information and power transfer millimeter-wave ultra-dense networks," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 7, pp. 1817–1829, 2019.

[11] R. Ma, W. Yang, X. Sun, L. Tao, and T. Zhang, "Secure communication in millimeter wave relaying networks," *IEEE Access*, vol. 7, pp. 31218–31232, 2019.

[12] J. Huang and A. L. Swindlehurst, "Robust secure transmission in miso channels based on worst-case optimization," *IEEE Transactions on Signal Processing*, vol. 60, no. 4, pp. 1696–1707, 2012.

[13] Q. Yuan, Y. Hu, C. Wang, and Y. Li, "Joint 3d beamforming and trajectory design for uav-enabled mobile relaying system," *IEEE Access*, vol. 7, pp. 26488–26496, 2019.

[14] L. Zhu, J. Zhang, Z. Xiao, X. Cao, D. O. Wu, and X.-G. Xia, "3-D beamforming for flexible coverage in millimeter-wave uav communications," *IEEE Wireless Communications Letters*, vol. 8, no. 3, pp. 837–840, 2019.

[15] G. Zhang, Q. Wu, M. Cui, and R. Zhang, "Securing uav communications via trajectory optimization," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, pp. 1–6, Singapore, Singapore, Jan. 2017.

[16] G. Zhang, Q. Wu, M. Cui, and R. Zhang, "Securing uav communications via joint trajectory and power control," *IEEE Transactions on Wireless Communications*, vol. 18, no. 2, pp. 1376–1389, 2019.

[17] C. Zhong, J. Yao, and J. Xu, "Secure uav communication with cooperative jamming and trajectory control," *IEEE Communications Letters*, vol. 23, no. 2, pp. 286–289, 2019.

[18] Q. Wang, Z. Chen, H. Li, and S. Li, "Joint power and trajectory design for physical-layer secrecy in the uav-aided mobile relaying system," *IEEE Access*, vol. 6, pp. 62849–62855, 2018.

[19] X. Sun, C. Shen, D. W. K. Ng, and Z. Zhong, "Robust trajectory and resource allocation design for secure uav-aided communications," in *2019 IEEE International Conference on Communications Workshops (ICC Workshops)*, pp. 1–6, Shanghai, China, China, 2019.

[20] A. Li, Q. Wu, and R. Zhang, "UAV-enabled cooperative jamming for improving secrecy of ground wiretap channel," *IEEE Wireless Communications Letters*, vol. 8, no. 1, pp. 181–184, 2019.

[21] X. Sun, C. Shen, T.-H. Chang, and Z. Zhong, "Joint resource allocation and trajectory design for uav-aided wireless physical layer security," in *2018 IEEE Globecom Workshops (GC Wkshps)*, pp. 1–6, Abu Dhabi, United Arab Emirates, United Arab Emirates, 2018.

[22] X. Tang, P. Ren, Y. Wang, Q. Du, and L. Sun, "Securing wireless transmission against reactive jamming: a stackelberg game framework," in *2015 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, San Diego, CA, USA, 2015.

[23] J. Zheng, Y. Cai, and A. Anpalagan, "Astochastic game theoretic approach for interference mitigation in small cell networks," *IEEE Communications Letters*, vol. 19, no. 2, pp. 251–254, 2015.

[24] X. Tang, P. Ren, Y. Wang, and Z. Han, "Combating full-duplex active eavesdropper: a hierarchical game perspective," *IEEE Transactions on Communications*, vol. 65, no. 3, pp. 1379–1395, 2017.

[25] A. Mukherjee and A. L. Swindlehurst, "A full-duplex active eavesdropper in mimo wiretap channels: construction and

countermeasures," in *2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, pp. 265–269, Pacific Grove, CA, USA, 2011.

[26] M. R. Abedi, N. Mokari, H. Saeedi, and H. Yanikomeroglu, "Robust resource allocation to enhance physical layer security in systems with full-duplex receivers: active adversary," *IEEE Transactions on Wireless Communications*, vol. 16, no. 2, pp. 885–899, 2017.

[27] L. Li, A. P. Petropulu, and Z. Chen, "MIMO secret communications against an active eavesdropper," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 10, pp. 2387–2401, 2017.

[28] X. Zhou, B. Maham, and A. Hjorungnes, "Pilot contamination for active eavesdropping," *IEEE Transactions on Wireless Communications*, vol. 11, no. 3, pp. 903–907, 2012.

[29] A. Al-nahari, "Physical layer security using massive multiple-input and multiple-output: passive and active eavesdroppers," *IET Communications*, vol. 10, no. 1, pp. 50–56, 2016.

[30] X. Tian, M. Li, and Q. Liu, "Random-training-assisted pilot spoofing detection and security enhancement," *IEEE Access*, vol. 5, pp. 27384–27399, 2017.

[31] Y. Wu, R. Schober, D. W. K. Ng, C. Xiao, and G. Caire, "Secure massive mimo transmission with an active eavesdropper," *IEEE Transactions on Information Theory*, vol. 62, no. 7, pp. 3880–3900, 2016.

[32] 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.

[33] Y. Li, L. Xiao, H. Dai, and H. V. Poor, "Game theoretic study of protecting MIMO transmissions against smart attacks," in *2017 IEEE International Conference on Communications (ICC)*, pp. 1–6, Paris, France, May 2017.

[34] Q. Zhu, W. Saad, Z. Han, H. V. Poor, and T. Basar, "Eavesdropping and jamming in next-generation wireless networks: a game-theoretic approach," in *2011 - MILCOM 2011 Military Communications Conference*, pp. 119–124, Baltimore, MD, USA, 2011.

[35] L. Xiao, C. Xie, M. Min, and W. Zhuang, "User-centric view of unmanned aerial vehicle transmission against smart attacks," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 4, pp. 3420–3430, 2018.

[36] L. Yang, J. Chen, H. Jiang, S. A. Vorobyov, and H. Zhang, "Optimal relay selection for secure cooperative communications with an adaptive eavesdropper," *IEEE Transactions on Wireless Communications*, vol. 16, no. 1, pp. 26–42, 2017.

[37] J. Liu, W. Yang, S. Xu, J. Liu, and Q. Zhang, "Q-learning based UAV secure communication in presence of multiple UAV active eavesdroppers," in *2019 11th International Conference on Wireless Communications and Signal Processing (WCSP)*, pp. 1–6, Xi'an, China, China, 2019.

[38] D. Yang, G. Xue, J. Zhang, A. Richa, and X. Fang, "Coping with a smart jammer in wireless networks: a Stackelberg game approach," *IEEE Transactions on Wireless Communications*, vol. 12, no. 8, pp. 4038–4047, 2013.

[39] L. Jia, Y. Xu, Y. Sun, S. Feng, and A. Anpalagan, "Stackelberg game approaches for anti-jamming defence in wireless networks," *IEEE Wireless Communications*, vol. 25, no. 6, pp. 120–128, 2018.

[40] H. Fang, L. Xu, Y. Zou, X. Wang, and K.-K. R. Choo, "Three-stage Stackelberg game for defending against full-duplex active eavesdropping attacks in cooperative communication," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 11, pp. 10788–10799, 2018.

[41] C. Cheng, Z. Zhu, B. Xin, and C. Chen, "A multi-agent reinforcement learning algorithm based on Stackelberg game," in *2017 6th Data Driven Control and Learning Systems (DDCLS)*, pp. 727–732, Chongqing, China, 2017.

[42] D. Fudenberg and T. Jean, *Tirole: Game theory*, vol. 726, MIT Press, 1991.

[43] A. Kianercy and A. Galstyan, "Dynamics of Boltzmann Q learning in two-player two-action games," *Physical Review E*, vol. 85, no. 4, article 041145, 2012.