IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 12, 2018, accepted January 21, 2019, date of publication February 1, 2019, date of current version March 20, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2896709

Selfish Bandit-Based Cognitive Anti-Jamming Strategy for Aeronautic Swarm Network in Presence of Multiple Jammer

HAITAO LI, JIAWEI LUO, AND CHANGJUN LIU
Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China

Corresponding author: Haitao Li (lihaitao@bjut.edu.cn)

This work was supported in part by the Program of the Science and Technology on Avionics Integration Laboratory, and in part by the Aeronautical Science Foundation of China under Grant 2018ZC15.

ABSTRACT In order to enhance the anti-jamming capability of aeronautic swarm tactical network in the complicated electromagnetic environment, we address the problem of bandit-based cognitive anti-jamming strategy for enabling reliable information transmission. We first present an adversarial multiuser multi-armed bandit model for the aeronautic swarm network employing airborne cognitive radios with the same-frequency simultaneous transmit and receive feature. Then, we utilize the improved energy detection method to perform jamming sensing and derive the closed expression of false alarm probability, false detection probability, and the optimal decision threshold in the case of single and multi-jammer. Finally, using the jamming sensing output to calculate reward and with the objective of maximizing the throughput of each airborne radio, a decentralized selfish doubling trick kl-UCB^{++} anti-jamming strategy is developed to allocate an optimal configuration of transmitting power and spectrum channel to each radio. This anytime bandit strategy is simultaneously minimaxed optimal and asymptotically optimal. The simulation results validate that the aggregate average throughput, cumulative regret obtained with the proposed anti-jamming strategy outperform the well-known UCB, kl-UCB^{++} bandit algorithm.

INDEX TERMS Cognitive anti-jamming, aeronautic swarm, adversarial multi-armed bandit, improved energy detection, doubling trick kl-UCB^{++}.

I. INTRODUCTION

To enable the emerging Net-Centric Warfare (NCW) needs, the next generation of airborne tactical networks (ATN) must evolve with multi-unmanned aerial vehicle (UAV) systems to provide swarm combat capability. Aeronautic swarm network (ASN) consists of multi-UAVs is a new kind of airborne tactical networks inspired by biological swarm behaviors, the intensive use of UAVs combat system will be standard practice in the next decade. In military operations, aeronautic swarm networks may vary from slow dynamic to dynamic and have intermittent links and fluid topology, which would bring new challenges to the design of mission-centric ATN. While it is believed that ad hoc mesh network would be most suitable for aeronautic swarm system and offers the promise of improved capacity and maintaining reliable communications for multi-UAVs [1].

Aeronautic swarm network is used for exchange of constantly growing amount of battlefield situation information and it also causes a lot of interferences so coexistence among swarm nodes becomes a demand. Moreover, with the aerial battlefield electromagnetic environment getting increasingly complex and intentional jamming, swarm nodes are non-permanent, wireless channels may be impaired, and communication links connectivity between peer nodes is intermittent. This necessitates resiliency anti-jamming technologies to be closely integrated into ASN to provide robust connectivity and gain competitive advantage of future electromagnetic spectrum warfare (EMSW).

The traditional anti-jamming solutions do not work as well in complicated electromagnetic scenarios, due to current airborne tactical radios are statically configured to operate within a pre-allocated spectrum channel prior to deployment

The associate editor coordinating the review of this manuscript and approving it for publication was Xin Su.

in temporal, frequency and geographical domains. The paradigm of static spectrum allocation results in a situation where some frequency bands are utilized effectively where as some portions of spectrum remain under-utilized. Aeronautic tactical radios need share spectrum with other in- and out-of network radios to improve frequency spectrum utilization. Latterly, the conception of cognitive radio (CR) based anti-jamming communication technology was bring forward to improve spectral efficiency of tactical network in a congested electromagnetic environment [2]–[3]. Cognitive anti-jamming (CAJ) radio can sense the jamming signal and opportunistically avoids the jammer spectrum for secure data transmission in the presence of intentional and accidental interferences, and emerge as an intelligent aeronautical wireless communication system through dynamic spectrum access (DSA) feature of CR, that has some autonomy to make decisions about the spectrum usage.

Cognitive anti-jamming technology has attracted widespread attention and considerable researches. To address the interactive competition between the legitimate users and the jammers, game theory and Markov decision process (MDP) has been firstly used for cognitive anti-jamming network. A stochastic zero-sum game framework is proposed in [2] and Minimax-Q learning algorithm is utilized to explore an optimal channel accessing strategy in dynamic anti-jamming game. For the same stochastic game model, Singh *et al*. presents the use of state-action-reward-state-action learning and QV learning that are the on-policy and non-greedy variant of Q-learning algorithm which outperform the Minimax-Q algorithm [3]. Further, assumed system model allows multiple tactical radios to simultaneously operate over the same spectrum band, and each radio attempts to evade the transmissions of other radios as well as avoiding jamming signal, a multi-agent reinforcement learning (MARL) algorithm based on Q-learning is proposed to find optimal anti-jamming and interference avoidance policies in [4]. Moreover, a new decision policy for the sub-band spectrum state to reduce the computational complexity of learning is developed in the multi-agent environment.

All these above-mentioned cognitive anti-jamming works is mainly based on Game theory model and utilizes Q-learning to solve. The stateful Q-learning approach requires explicit modeling of network states and actions from an underlying MDP. Unfortunately, for the aeronautic swarm network, it is difficult to deal with this model directly because of the more state of the environment. Wang *et al.* [5] have modeled the DSA problem with partially observable Markov decision process (POMDP) framework which considers channel quality to decide about the channel to sense, however, it has comparatively higher complexity. On the contrary, the cognitive anti-jamming problem is modeled under the multi-armed bandit (MAB) framework which turns to be very easy and less complex to implement. Therefore, we investigate the stateless MAB model that address the exploration-exploitation dilemma for allocating power and channel selection on sequential reward sampling.

The classical MAB models a sequential interaction scheme between a learner and an environment. The learner sequentially selects one out of K actions (arms) and obtains some rewards determined by the chosen action and also influenced by the environment. Under various assumptions made on the environment and the structure of the arms, several MAB settings have been considered such as stochastic bandits, adversarial bandits, restless bandits and contextual bandits. In these bandits setting, the most important basic case is the stochastic bandit problem where, for each particular action, the rewards are i.i.d. of random variables from a fixed distribution. However, the assumption on i.i.d. processes does not always apply to the real battlefield environment. On the other hand, the adversarial (or non-stochastic) bandit problem do not make any assumptions on the payoffs, where the rewards are chosen arbitrarily by the environment. Since aerial combat applications where the swarm network nodes would be highly mobile and would establish the network topology in an ad hoc manner to communicate and cooperate. Acquiring accurate context information may be extremely challenging and even unfeasible due to the frequent change of network topology. Therefore, we are more interested in the case where no context can be inferred, and the ASN anti-jamming communication problem would be modeled as an adversarial multi-armed bandit.

Some of the related bandit-based anti-jamming studies have been reported recently. From a multi-domain perspective, the anti-jamming defense scheme which includes both power domain and spectrum domain is proposed [6]. To be more specific, a Stackelberg power game is formulated to fight against the jamming attacks in the power domain, and a UCB1 bandit algorithm-based channel selection scheme with a channel switching cost is designed to achieve anti-jamming in the spectrum domain. In [7], an adversarial multi-player MAB game is employed to model the problem of joint channel and power allocation in underwater acoustic communication networks, and presents a game-based distributed hierarchical exponential learning algorithm that effectively improves user learning ability and decreases learning time. Based on multi-player bandit model, Sawant *et al*. [8] study distributed algorithms that are robust against malicious jamming attack and give constant regret with high confidence.

The bandit algorithms used in the above anti-jamming approaches rely on the knowledge of this horizon T to sequentially select arms (*one time*) and also can't be simultaneously asymptotically optimal and minimax optimal. kl-UCB^{++} algorithm, a slightly modified version of kl-UCB^{+}, is the first algorithm proved to be asymptotically- and minimax-optimal at the same time [9]. An online learning algorithm is anytime if it does not need to know in advance the horizon T. It is necessary to design bandit-based anti-jamming strategy with any time feature due to each swarm node is difficulty to decide the accurate time horizon in dynamic combat scenario. Note that a well-known technique to obtain an anytime algorithm from any non-anytime algorithm is the ''doubling trick'' (DT). In this paper, we merge

doubling trick and kl-UCB^{++} to design a novel multi-domain cognitive anti-jamming strategy for aeronautic swarm network.

On the other hand, most of contemporary research work has been done in the context of bandit-based anti-jamming assume that perfect spectrum channel sensing in physical layer (PHY), and the key to anti-jamming operation is the radio's ability to sense its surrounding electromagnetic environment, this functionality is known as jamming sensing. However, imperfect sensing has some limitations concerning the anti-jamming capability. There have been some attempts in [10], to consider the energy detector (ED) output as a reward for general reinforcement learning algorithms, but they lack from significant theoretical guarantee and a relation with achievable throughput. In contrast, we jointly investigate DT kl-UCB^{++} bandit algorithm and jamming sensing, with the objective of maximizing the throughput of each airborne radio, design a optimal configuration of transmit power and spectrum channel for enabling ASN anti-jamming communication.

The remainder of this paper is organized as follows. Section II, we describe the aeronautic swarm network model consists of in-band full-duplex (IBFD) enabled CRs, which has a good advantage of increased throughput and real-time sensing ability. In Section III, the detection/false alarm probability of jamming sensing based on improved energy detection (IED) is analyzed theoretically. Further, according to the accurate reward calculation from jamming sensing output, the distributed anti-jamming scheme using DT kl-UCB^{++} algorithm is proposed in Section IV. In Section V, the performance evaluation of the presented bandit anti-jamming is analyzed with simulations. Finally, the conclusions are drawn in Section VI.

II. SYSTEM MODEL

We consider the aeronautic swarm network is illustrated in Figure 1. UAV nodes of ASN are hovering over a geographical area and are equipped with tactical cognitive radio. In the battlefield the ally and enemy tactical radios face each other in a competition to dominate an open spectrum resource to achieve higher throughput. Using accurate local spectrum situation sensing information, airborne radio applies a strategy to perform transmit or silence action. Similarly, we assume the smart jammers have cognitive features such as spectrum sensing, learning and reconfigure ability, subsequently causing more damages than the conventional jammers. During the operation, the radio nodes periodically exchange control information to select the best radios as local controllers, i.e. cluster heads (CH) in the swarm ad hoc network. If operational conditions of a specific CH degrade, its role can be taken over by another radio node of the network. Then, the radio nodes are also selected to act as gateways (GW) between clusters if required.

FIGURE 1. ASN model.

In the following, we present mathematical formulation for the ASN with N tactical radios and M jammer. Let K designate the number of non-overlapping channels in the frequency band for open access, which is partitioned in time and frequency and located at the center frequency f_k(MHz) with bandwidth B(Hz) for $k = 1, \ldots, K$. A transmission slot at channel k and time t with time duration T_d (msec) is represented by a tuple $<f_k, B, t, T_d>$. We define the action set A such that $a^t \in A$ at time t. And the power domain and frequency domain-based anti-jamming scheme is considered in this paper, hence, an ith element in a^t designates a configuration of power and frequency channel that the ith radio tries to transmit at t. The tactical radio actions result in an outcome $\Omega : A \rightarrow R$. Subsequently, the outcome can be mapped to a reward r. For convenience, Table 1 lists the notations used in this study.

TABLE 1. Summarizes the used notation.

N	Number of tactical radios
M	Number of jammers
K	Number of frequency channels
A	Action set
a	Action element
r	Reward
R_T	Accumulated regret
π_i	The strategy of anti-jamming radio i
Ω	The energy test statistics
λ	Decision threshold
Λ_{opt}	optimal sensing threshold
P_{md}	Miss detection probability of jamming sensing
P_f	False alarm probability of jamming sensing
P_e	Total error probability of jamming sensing
B_i	Channel bandwidth of radio i
C_i	Instantaneous throughput of radio i
C_i^*	Maximum theoretical throughput of radio i
p_i	Transmit power of radio i,
p_J	Jammer power
μ_a	Expectation of action a
N_a	Number of performing action a
T_i	Doubling sequence

For the presented aeronautic swarm network configuration, a suitable mathematical formulation needs to be created. Since there are multiple radios in the tactical network, our problem is classified as multi-player MAB. In the bandit model, the radios are the players (agent) in the swarm network, and they play (i.e., transmit) the channels, an arm (action) corresponds to be a frequency channel and transmit

power level that the anti-jamming radio may choose under competition. In the case of decentralized decision making, each radio computes its own action. For radio i, we can write as, $x_i^t, \left\{x^t, a^j, \Omega^j\right\}_{j=1}^{t-1} \xrightarrow{\pi_i^t} a_i^t$, where x_i^t is the sensing information only available to radio i at time t, and π_i^t is the strategy of radio i's own. For the decentralized ad hoc swarm network, it is an adversarial bandit problem, in which the actions of a given radio affect the reward distributions of the others' actions.

For the aeronautic swarm network in which multiple tactical radios and jammers have to coexist, it generally operates in highly congested and contested electromagnetic environments, which may result the spectrum resources is scarce. Therefore, the same-frequency simultaneous transmit and receive (SF-STAR) technologies is employed for cognitive radio in this paper. It is worth noting that SF-STAR CAJ (SCAJ) radio is transmitting and receiving information signals at the same time and at the same center frequency, and promise to double the network throughput of a wireless link, compared with traditional half-duplex operation. The military IBFD radios will have the progressive capability for SF-STAR by which they can conduct electronic warfare at the same time when they are also using the same frequency band for communication. It is quite obvious that, by utilizing the STAR capability, SCAJ radio could gain a major technical advantage over an opponent that does not possess similar technology [11]–[12], and the use of artificial noise generated by FD receiver technology has been presented to enhance physical layer security [13].

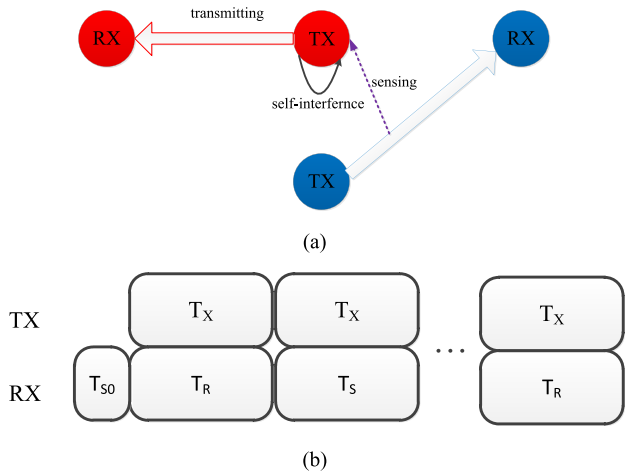

FIGURE 2. Operation mode of SCAJ radio.

We design IBFD transmitter-based transmit-sense-receive (T-S-R) mode for SCAJ radio as depicted in Fig. 2(a)-(b). Firstly, to check channel availability, the radio initially senses in a half-duplex fashion for a duration T_{S0}. Based on the sensing outcome, the transmit side (TX) will decide the current operational center frequency of transmit signal for duration T_X. Simultaneously, the receiver side (RX) continue to sense jamming or receive signal. This sensing/receiving process may be divided into K short sensing/receiving periods T_S/T_R, which can be dynamically allocated to account for the tradeoff between sensing efficiency and timeliness in detecting jamming activity. The current operating frequency of airborne radio can be continuously monitored during sensing time T_S to improve the ability to situation awareness. The T-S-R mode is effectively and practical way to do long sensing for detecting whether the channel has been interfered by jammer.

III. JAMMING SENSING

The design of cognitive anti-jamming strategy started from the premise that the frequency bands usage information can be available. Such information gives an advantage during the operation mission because not only helps to ensure information transmission but could be used for electromagnetic warfare too. Spectrum situation awareness of jamming signal is a part of cognitive anti-jamming communication system and would be utilized to learn and adapt to the environment. Likewise, sensing accuracy indicates the detecting probability when the jammer is present. There are several methods of channel sensing including energy detection, matched filtering based detection and cyclostationarity-based detection are the popular methods of sensing and estimation used in the CR implementation. However, these sensing approaches can't achieve the trade off between performance and complexity. To perform well in jamming sensing, we make use of an improved energy detector [14], i.e. a p-norm energy detector, where the conventional energy detector is modified by replacing the squaring operation of the received signal amplitude with an arbitrary positive power p may yield a performance gain.

For the swarm network with N tactical radios and M jammer, where M jammers operate in the same frequency band and is sensed by each tactical radio. Hence, at SCAJ radio the $n-$th sample of the baseband equivalent received signal can be expressed as

$$\mathrm{y}\,(n) = \begin{cases} h_{SI}\,u\,(n) + w\,(n), & H_0 \\ \sum_{i=1}^{M} h_i(n)s_i(n) + h_{SI}\,u\,(n) + w\,(n), & H_1 \end{cases} \quad (1)$$

where $s_i(n)$ denotes the n-th sample of signal transmitted by the i-th jammer, $u\,(n)$ is the n-th sample of self-interfere signal. We assume that s_i and u are zero-mean circular symmetric complex white Gaussian processes with variances σ_s^2 and σ_u^2. h_i is the zero mean complex-valued channel coefficient with variance σ_h^2, h_{SI} is the self-interference channel coefficient from radio transmitter to receiver with variance $\sigma_{h_{SI}}^2$, while w represent Gaussian noise signal with zero mean and variance σ_w^2. H_0 and H_1 correspond to the decision about the presence and absence, respectively, of the jamming signal in current frequency channel.

Based on the signal model described above, and defining $\eta(\mathrm{n}) = \frac{|\mathrm{y(n)}|^p}{\sigma_w^p}$, where p is an arbitrary positive real number and is a tunable parameter that gives the decision statistics some flexibility. Then the improved energy detector

calculates the energy test statistics as

$$\Omega = \sum_{n=0}^{N_s} \eta(\mathrm{n}) \tag{2}$$

where N_s is the number of samples used for jamming sensing. The energy test statistic Ω is compared against a threshold λ to yield the sensing decision, i.e., the IED decides that the channel is busy if $\Omega > \lambda$ or idle, otherwise. When $p = 2$, Ω reduces to the statistic $\sum_{n=0}^{N_s} \frac{|\mathrm{y(n)}|^2}{\sigma_w^2}$ corresponding to the conventional energy detection method.

Since $|\mathrm{y}(\mathrm{n})|^2/\sigma_w^2$ is exponentially distributed and the probability distribution function (PDF) $f_{|\mathrm{y(n)}|^2/\sigma_w^2}(.)$ is an exponentially distributed random variable with parameters $\theta = \frac{1}{1+\gamma_{inr}}$ and $\theta = \frac{1}{1+\gamma_{inr}+\sum_{i=1}^{M}\gamma_{snr}(i)}$ under hypotheses H_0 and H_1, respectively, where $\gamma_{inr} = \sigma_{hSI}^2\sigma_u^2/\sigma_w^2$ is the self-interference to noise ratio, $\gamma_{snr}(i) = \sigma_h^2(i)\sigma_s^2(i)/\sigma_w^2$ is the signal to noise ratio of the ith-jammer-radio link. We can make an equivalent transformation on the cumulative distribution function (CDF) of η by $\Pr(\eta \le \mathrm{x}) = \Pr\left(\frac{|\mathrm{y(n)}|^p}{\sigma_w^p} \le \mathrm{x}\right) = \Pr\left(\frac{|\mathrm{y(n)}|^2}{\sigma_w^2} \le \mathrm{x}^{\frac{2}{p}}\right) = \int_0^{x^{\frac{2}{p}}} \theta \exp(\theta t)\, dt = 1 - \exp(-\theta x^{\frac{2}{p}})$, where $\Pr(.)$ denotes the probability.

The probability distribution function of η can be obtained by differentiating the preceding equation, resulting in $f_\eta(\mathrm{x}) = \frac{2}{p}\theta x^{\frac{2}{p}-1}\exp(-\theta x^{\frac{2}{p}})$. Therefore, we can obtain the conditional PDF $f_{\eta|H_0}(\mathrm{x})$ and $f_{\eta|H_1}(\mathrm{x})$ under hypotheses H_0 and H_1 as

$$f_{\eta|H_0}(\mathrm{x}) = \frac{2}{p}\left(\frac{1}{1+\gamma_{inr}}\right)x^{\frac{2}{p}-1} exp\left[-\left(\frac{1}{1+\gamma_{inr}}\right)x^{\frac{2}{p}}\right] \tag{3}$$

$$f_{\eta|H_1}(\mathrm{x}) = \frac{2}{p}\left(\frac{1}{1+\gamma_{inr}+\sum_{i=1}^{M}\gamma_{snr}(i)}\right)x^{\frac{2}{p}-1} \times exp\left[-\left(\frac{1}{1+\gamma_{inr}+\sum_{i=1}^{M}\gamma_{snr}(i)}\right)x^{\frac{2}{p}}\right] \tag{4}$$

We know that $f_{\eta|H_0}(\mathrm{x})$ and $f_{\eta|H_1}(\mathrm{x})$ are Weibull distributed [15]. According to the central limit theorem, if the number of received samples are large, the decision variable Ω will be normal distributed with mean and variance as

$$H_0: \begin{cases} \mu_0 = N_s(1+\gamma_{inr})^{\frac{p}{2}}\Gamma\left(1+\frac{p}{2}\right) \\ \sigma_0^2 = N_s(1+\gamma_{inr})^{\frac{p}{2}}\left[\Gamma\left(1+\frac{p}{2}\right) - \Gamma^2\left(1+\frac{p}{2}\right)\right] \end{cases} \tag{5}$$

and

$$H_1: \begin{cases} \mu_1 = N_s\left[1+\gamma_{inr}+\sum_{i=1}^{M}\gamma_{snr}(i)\right]^{\frac{p}{2}}\Gamma\left(1+\frac{p}{2}\right) \\ \sigma_1^2 = N_s\left[1+\gamma_{inr}+\sum_{i=1}^{M}\gamma_{snr}(i)\right]^{\frac{p}{2}} \\ \quad \times\left[\Gamma\left(1+\frac{p}{2}\right) - \Gamma^2\left(1+\frac{p}{2}\right)\right] \end{cases} \tag{6}$$

After some algebraic manipulations, the probability of miss detection in each SCAJ radio can be obtained as

$$P_{md} = Pr\{\Omega \ge \lambda\,|H_1\} = 1 - Q\left(\frac{\mu_1}{\sigma_1}\right) - Q\left(\frac{\lambda-\mu_1}{\sigma_1}\right) \tag{7}$$

where $Q(\cdot)$ is the Q-function. Similarly, the probability of false alarm in each radio can be obtained as

$$P_f = Pr\{\Omega \ge \lambda\,|H_0\} = Q\left(\frac{\lambda-\mu_0}{\sigma_0}\right) \tag{8}$$

Hence the total error probability of SCAJ radio jamming sensing can be calculated as

$$\begin{aligned} P_e &= P_f + P_{md} \\ &= 1 + Q\left(\frac{\lambda-\mu_0}{\sigma_0}\right) - Q\left(\frac{\mu_1}{\sigma_1}\right) - Q\left(\frac{\lambda-\mu_1}{\sigma_1}\right) \end{aligned} \tag{9}$$

By differentiating the preceding equation (10), we can get

$$\begin{aligned} \frac{dP_e(\lambda)}{d(\lambda)} = &\frac{1}{\sigma_0}\frac{1}{\sqrt{2\pi}\sigma_0} exp\left[-\left(\frac{\lambda-\mu_0}{\sigma_0}\right)^2\right] \\ &- \frac{1}{\sigma_1}\frac{1}{\sqrt{2\pi}\sigma_1} exp\left[-\left(\frac{\lambda-\mu_1}{\sigma_1}\right)^2\right] \end{aligned} \tag{10}$$

Let $\frac{dP_e(\lambda)}{d(\lambda)} = 0$, after some transformations, we can obtain the optimal jamming sensing threshold

$$\lambda_{opt} = \frac{\frac{\mu_0}{\sigma_0^2} - \frac{\mu_1}{\sigma_1^2}}{\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}} - \sqrt{\frac{\left(\frac{\mu_0}{\sigma_0^2} - \frac{\mu_1}{\sigma_1^2}\right)^2 - \left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\right)\left(\frac{\mu_0}{\sigma_0^2} - \frac{\mu_1}{\sigma_1^2} + 2ln\frac{\sigma_0}{\sigma_1}\right)}{\left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\right)^2}} \tag{11}$$

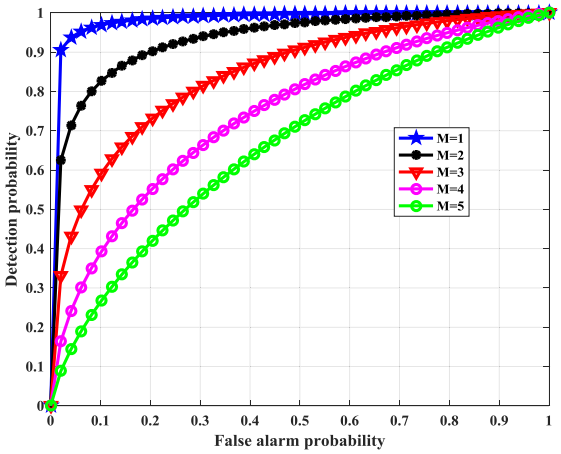

FIGURE 3. ROCs for IED with $p = 3$.

In Fig. 3, the receiver operating characteristic curves (ROCs) for improved energy detection are illustrated for different number of jammers, simulation parameters are $p = 3$, $N_s = 10$, $INR = -2$dB. We observe that as the number of jammers increases, the detection probability reduces due to the interference form the jammers increase for a fixed false alarm probability.

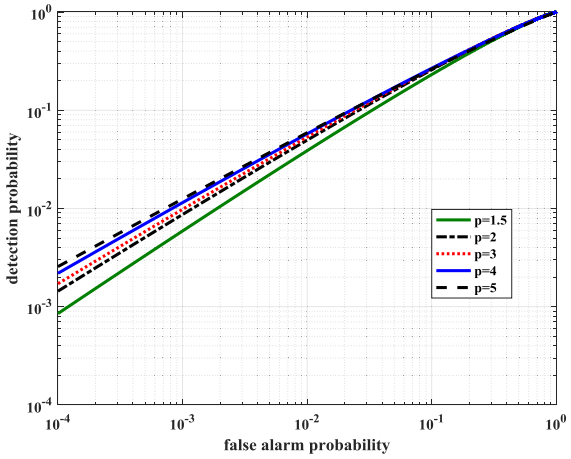

FIGURE 4. ROCs for multiple jammer with M = 5.

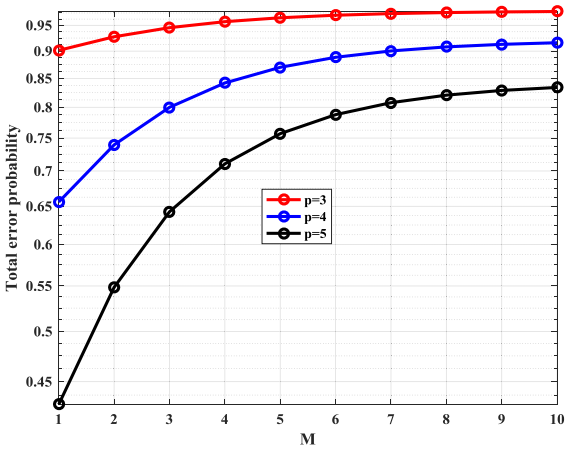

FIGURE 6. Total error probability w.r.t. M.

Next, we consider that the SCAJ radio operates in 5-jammers environment, where the simulation parameter $M = 5$. We assume the same p for the jammers, it is observed that the detection probability increases as the p increase for a fixed false alarm due to the interferences from the jammers are suppressed.

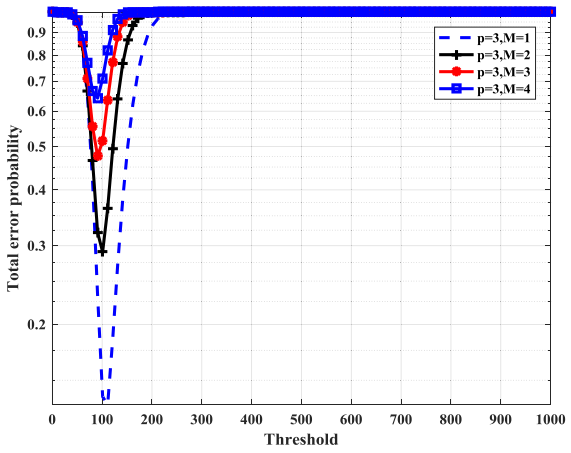

FIGURE 5. The total error probability w.r.t. threshold.

Fig. 5 plots the jamming sensing total error probability of SCAJ radio versus threshold by setting $\gamma_{inr} = -10$dB. As show in the figure, for a fixed $p = 3$, M increase, the minimum value of the total error probability increase.

Fig. 6 plots total error probability of jamming sensing under threshold $\lambda = \lambda_{opt}$. It is observed that as M increase, the total error probability increase, this due to the increase of interferences from jammer increase; for a fixed M, as the increase of p, the total error probability decrease.

IV. COGNITIVE ANTIJAMMING STRATEGY

In section II, we have presented the multiuser bandit model for anti-jamming communication in swarm network. Normally, this problem can be considered as an approximation of contextual MAB, and the conventional contextual bandit considers the existence of a context that influences the action-selection process. As a consequence, the available strategies vary with the context and the probability distribution of a given reward. However, due to the dynamic change of aeronautical swarm network topology, such information is difficulty to be obtained in practice. Therefore, we are more focus on the case where no context can be inferred, and the anti-jamming communication problem is modeled as an adversarial bandit in which no stochastic assumption is taken and several tactical radios compete against each other. Especially, recent research shows that bandit algorithms tailored for a stochastic model is still useful in non-stochastic adversarial bandit problem [16]. This is a very encouraging and beneficial result, and we will explore the cognitive anti-jamming strategies based on stochastic bandit learning algorithm. In the following, we present a selfish doubling trick KL-UCB^{++} algorithm to cope with this kind of bandit problem.

A. REWARD DEFINITION

In swarm network, the radio shapes an anti-jamming strategy according to the obtained rewards. And a reward function allows a radio conducting its action towards a given performance metric. When choosing an action in anti-jamming scheme based on bandit learning, the SCAJ radio has access to the history of rewards and actions. The radio's objective is to choose a strategy that maximizes the expected reward over a finite time horizon T. Therefore, accurate reward is important to design anti-jamming strategy, and we carry out reward calculation using the above-mentioned jamming sensing output. However, defining a reward function may be a very complex task. If a precise definition of reward perfectly matches with the desired goal, the reward would improve the learning procedure and reduce the probability of falling into a local minimum.

Let $a_i \in A$ be an action that a SCAJ radio may choose. Each action is a configuration of frequency channel and transmit power, and grants a reward that depends on the others' action.

We define C_i be the instantaneous throughput experienced by radio i at time t, and C_i^* is the maximum achievable throughput of SCAJ radio. The maximum theoretical throughput can be calculated as

$$C_i^* = B_i log\left(1 + SNR_i\right) \tag{12}$$

where B_i is the access channel bandwidth of radio i on channel k, and SNR_i is the receive signal-to-noise ratio (SNR) of radio i. In the ASN, each radio opportunistically access to the idle frequency channel f_i with the transmit power p_i under the local sensing result, thus the opportunistic instantaneous transmission throughput of radio i is given by

$$C_i = \left(1 - P_f\right) r_i^{(1)} + P_{md} r_i^{(2)} \tag{13}$$

where $r_i^{(1)} = B_i log\left(1 + \frac{|h_{ii}|^2 p_i}{\sigma_w^2 + \sum_{j \neq i}^{N} |h_{ji}|^2 p_j}\right)$, h_{ji} is the channel gain for the link from radio i to radio j, p_i is the transmit power of radio i, and $r_i^{(2)} = B_i log\left(1 + \frac{|h_{ii}|^2 p_i}{\sigma_w^2 + \sum_{j \neq i}^{N} |h_{ji}|^2 p_j + p_J}\right)$, p_J is the jamming power, the P_{md} and P_f are defined as (7) and (8), respectively. After achieving the instantaneous throughput C_i and the theoretical throughput C_i^*, the reward can be defined as

$$\begin{aligned} r_i &= \frac{C_i}{C_i^*} \\ &= \frac{\left(1 - P_f\right) r_i^{(1)} + P_{md} r_i^{(2)}}{B_i log\left(1 + SNR_i\right)} \end{aligned} \tag{14}$$

This reward definition characterize a selfish behavior which purely reflect the decentralized and adversarial problem. Through selfish learning, each tactical radio tries to learn the best configuration for their own gain, regardless of the performance experienced by other radios in swarm network. Under these circumstances, each radio ignores the existence of other learners. In particular, the accumulated regret $R_{i,T}$ that a given radio i experiences until time T can be characterized as follows

$$R_{i,T} = \sum_{t=1}^{T} \left(r_{i,t}^* - r_{i,t}\right) \tag{15}$$

where $r_{i,t}^*$ is the optimal reward granted by the best possible action in iteration t, and $r_{i,t}$ is the reward granted by the action chosen by radio i at that iteration.

Since the agent in multiuser MAB model of aeronautical swarm network can't get a priori information about the state transition probabilities, KL-UCB^{++}-based model-free reinforcement learning algorithms would be suitable to solve this game through trial-and-error interactions. Accordingly, we introduce the KL-UCB^{++} algorithm to present a decentralized bandit anti-jamming strategy.

B. KL-UCB^{++} ALGORITHM

The KL-UCB^{++} algorithm is a slight modification of algorithm KL-UCB^{+}. We first present some definition of a bandit problem with K actions indexed by $a \in \{a_1, \ldots, a_K\}$. Each action is assumed to be a probability distribution of some canonical one-dimensional family v_θ indexed by $\theta \in \Theta$. The expectation of action a is denoted by $\mu_a \in [\mu^-, \mu^+] \subset I$ and the best mean is $\mu^* = max_{a=1,\ldots,K} \mu_a$. At each round 1, an agent performs an action a_t and receives an independent reward r_t of the distribution $v_{\theta_{A_t}}$. Let $N_a(T) = \sum_{t=1}^{T} \mathbb{1}_{\{A_t = a\}}$ be the number of performing action a up to and including time T. The goal of KL-UCB^{++} algorithm is to minimize the expected accumulated regret

$$\begin{aligned} R_T &= T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} r_t\right] = \mathbb{E}\left[\sum_{t=1}^{T} \left(\mu^* - \mu_{A_t}\right) r_t\right] \\ &= \left[\sum_{t=1}^{T} \left(\mu^* - \mu_a\right)\right] \mathbb{E}\left[N_a(T)\right] \end{aligned} \tag{16}$$

Let $\mu_{a,n}$ be the empirical mean of the first n rewards from action a, and the empirical mean of action a after t rounds is

$$\bar{\mu}_a(t) = \bar{\mu}_{a,N_a(T)} = \frac{1}{N_a(T)} \sum_{s=1} Y_s \mathbb{1}_{\{A_s = a\}} \tag{17}$$

Algorithm 1 The KL-UCB++ Algorithm

Parameters: The horizon T and an exploration function g: $\mathbb{N} \mapsto \mathbb{R}^+$

Initialization: Pull each arm of $\{1, .., K\}$ once.

1: For $t = K$ to $T - 1$, do

2: Compute for each arm a the quantity

$$I_a(t) = sup\left\{\mu \in I : \mathrm{kl}\left(\hat{\mu}_a(t), \mu\right) \leq \frac{g\left(N_a(T)\right)}{N_a(T)}\right\}$$

3: Play $A_{t+1} \in argmax_{a \in \{1,\ldots,K\}} I_a(t)$

4: end

The KL-UCB^{++} algorithm is described as Algorithm 1, where $\mathrm{kl}\left(\bar{\mu}_a(t), \mu\right)$ is the Kullback-Leibler divergence on the set of action expectations. And the KL-UCB^{++} algorithm uses the exploration function g given by

$$\mathrm{g}(n) = log_+\left(\frac{T}{Kn}\left(log_+^2\left(\frac{T}{Kn}\right) + 1\right)\right) \tag{18}$$

where $log_+(x) := max\left(log(x), 0\right)$.

The following results state that the kl-UCB^{++} algorithm is simultaneously minimax- and asymptotically-optimal.

Lemma 1 (Minimax Optimality [9]): For any family $\mathcal{F}$ and for any bandit model $v \in \mathcal{F}$, the expected regret of the KL-UCB^{++} algorithm is upper-bounded as

$$R_T \leq 76\sqrt{VKT} + \left(\mu^+ - \mu^-\right) K \tag{19}$$

Lemma 2 (Asymptotic Optimality [9]): For any bandit model $v \in \mathcal{F}$, for any suboptimal arm a and any δ such that $\sqrt{22VK/T} \leq \delta \leq \left(\mu^* - \mu_a\right)/3$,

$$\mathbb{E}\left[N_a(T)\right] \leq \frac{log(T)}{kl\left(\mu_a + \delta, \mu^* - \delta\right)} + O\left(\frac{loglog(T)}{\delta^2}\right) \tag{20}$$

which implies the asymptotic optimality.

The kl-UCB^{++} algorithm is simultaneously minimax optimal and asymptotically optimal, but it is not *anytime* due to the total number of decisions making is unknown for anti-jamming communication in ASN. Hence in such cases,

it is crucial to devise *anytime* kl-UCB^{++} algorithms which do no rely on the knowledge of this horizon T to sequentially select actions. A general way to implement an *anytime* algorithm is the use of the doubling trick (DT), first proposed by [17], that utilize geometric sequence $T_i = 2^i$ to consist in repeatedly running the base algorithm with increasing horizons, in which the horizon is actually *doubling*.

The doubling trick is a well known idea in online learning, and the key to guarantee regret is to choose correctly the doubling sequence. Empirically, the term doubling trick usually refers to the geometric sequence $T_i = \lfloor T_0 b^i \rfloor$, is a general procedure to convert a non-anytime algorithm into an anytime algorithm. A geometric doubling sequence allows to conserve a minimax bound of the form $T^{\varepsilon}\,(logT)^{\rho}$ for any $0 < \varepsilon < 1$ and $\rho \geq 0$. Specific, unlike the previous geometric sequences, the exponential sequence $T_i := \left\lfloor \tau a^{b^i} \right\rfloor$ can indeed be used to conserve minimax regret bounds $(logT)^{\rho}$. It has been proved that the regret bounds of exponential doubling tricks is better than that geometric doubling trick [18]. Next, we utilize the kl-UCB^{++} algorithm based on exponential doubling trick to design the anti-jamming strategy.

C. DT KL-UCB^{++} ANTI-JAMMING STRATEGY

For our K-armed adversarial bandit model with N users, where the arms (actions) are refer to as the configuration of spectrum channels and transmit power, the players are the SCAJ radios. The idea of multi-domain cognitive anti-jamming strategy is that each radio utilizes bandit learning algorithm to successfully learn a frequency-power selection policy to avoid the smart jammer. For the classical MAB framework, an agent interacts with the environment in order to maximize the reward according to its actions. However, the presence of other radios in our adversarial bandit model adds an extra complexity. In the dynamic swarm network environment, spectrum channel quality may not be the same for each radio, and channel-power selection should be done by each SCAJ radio independently. Hence, a decentralized selfish anti-jamming communication technology should be implemented, where different radios aim to find the best configuration by their own.

In our decentralized anti-jamming framework, some important implications must be considered with regards to practical application of bandit learning to ASN. Since each radio attempts to learn by its own in highly dynamic environments, the action selection procedure is held in a disorganized way and the competition unleashed by the adversarial radios exits among the SCAJ radios. Although the radio can sense a channel to detect the presence of jammers before deciding to transmit data on this channel. However, distinct radios may transmit on the same frequency band leading to intensive collision, which may reduce throughput and cannot always guarantee a sublinear regret. Therefor, the decision making strategies which guarantee collision-free transmissions in the ASN are desired.

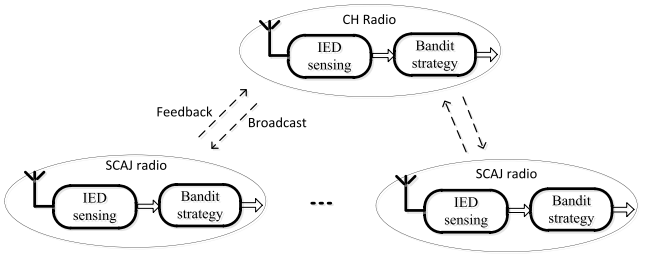

FIGURE 7. Schematic of decentralized anti-jamming.

To reduce collision and speed up convergence, we propose a decentralized anti-jamming strategy with finite coordination, which is shown schematically in Figure 7. The basic principle can be described as follows. Firstly, the available frequency channel is achieved by IED jamming sensing and shapes a channel list. It is assumed that the list is stored in cluster heads of ASN. If a radio have accessed one channel, it will feedback the channel occupation information to cluster heads. Then, this channel index would be canceled from current list to avoid collision, and cluster heads broadcast the updated channel list information to other radios to access. Finally, the opportunity of channel collision can be reduced by this partial coordination. The detailed anti-jamming strategy is illustrate in Algorithm 2.

Algorithm 2 DT KL-UCB++ anti-jamming strategy

Input: KL-UCB^{++} algorithm $\mathcal{A}^{(0)}$
exponential sequence $(T_i)_{i\in\mathbb{N}}$
channel list $\{f_1, \cdots, f_K\}$ using IED sensing
action set $\{a_1, \cdots, a_K\}$
Initialization: Let $i = 0$, and $\mathcal{A}^{(0)} = A_{T_0}$
1: for $t = 1, ..., T$ do
2: if $t > T_i$ then
3: $i = i + 1$.
4: Initialize KL-UCB^{++} algorithm $\mathcal{A}^i = A_{Ti-Ti-1}$.
5: Update $\mathcal{A}^i$ with the history of actions and rewards from all the steps from $t = 1$ to $t = T_i$.
6: end
7: Perform $\mathcal{A}^i(t - T_i)$ using Algorithm 1.
8: Compute an index I_a for each action.
9: Choose the action with highest index.
10: Computing the instantaneous throughput using (7), (8) and (13).
11: Computing the theoretical throughput using (12).
12: Computing the reward using (14).
13: Update $t = t + 1$.
14: end

Clearly, we find that the doubling trick kl-UCB^{++} strategy depends on a non-decreasing diverging *doubling sequence* $(T_i)_{i\in\mathbb{N}}$ and reinitializes its underlying algorithm $\mathcal{A}$ at each time T_i. Hence, the total regret is upper bounded by the regret on each sequence $\{T_i, \cdots, T_{i+1} - 1\}$ and is illustrated in Lemma 3.

Lemma 3 (Regret Upper Bounds [18]): For any bandit model and algorithm $\mathcal{A}$ and horizon T, doubling trick algorithm has the generic upper bound,

$$R_T(\mathcal{DT}(\mathcal{A}(T_i))_{i\in\mathbb{N}}) \leq \sum_{i-0}^{L_T} R_{\boldsymbol{T_i}-\boldsymbol{T_{i-1}}}(A_{\boldsymbol{T_i}-\boldsymbol{T_{i-1}}}) \quad (21)$$

where $L_T(T_i)_{i\in\mathbb{N}} := \min\{i \in \mathbb{N} : T_i > T\}$.

Further, it can be observed that Algorithm 2 is a heuristic doubling trick algorithm, in which a fresh algorithm $\mathcal{A}^{(i)}$ is created by the history from all the steps from $t = 1$ to $t = T_i$, then fed with successive observations. However, it is much harder to present theoretical result on this heuristic doubling trick algorithm. We only conjecture that a regret upper bound similar to that from Lemma 3, but it is still an open problem to be solved.

V. SIMULATION RESULTS

In this Section, we evaluate the performance of selfish bandit-based cognitive anti-jamming strategy for aeronautic swarm network. In our numeric simulations, the available bandwidth is $B = 10$MHz, it is divided into $K = 2$ frequency channels. And we set the number of radio nodes $N = 4$ and the number of jammers to $N_J = 5$. The ASN radios compete for access to two orthogonal channels at three possible power levels. Hence, denoting that the action sets {channel number f_k, transmit power $p_t(dB)$: $a_1 = \{1, 5\}$, $a_2 = \{1, 10\}$, $a_3 = \{1, 15\}$, $a_4 = \{2, 5\}$, $a_5 = \{2, 10\}$, $a_6 = \{2, 15\}$, respectively. Let $P_{\mathrm{d}} = 0.9$ and the mean rewards $\bar{\mu} = [0.1, 0.5, 0.6, 0.9, 0.7,0.8]$ for distinct actions,. The doubling sequences we consider are a exponential sequences $T_i = 200 \times 2^i$.

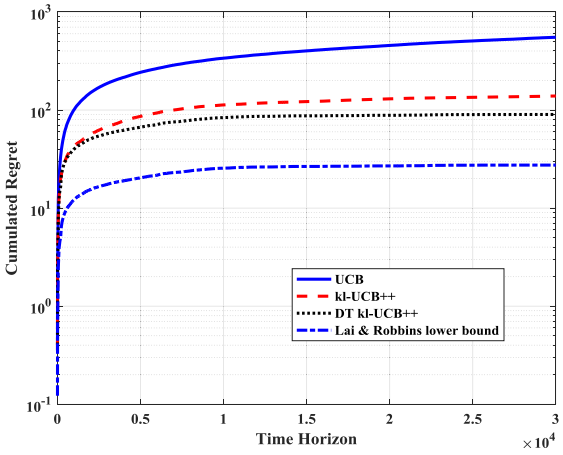

FIGURE 8. Cumulated regret w.r.t time horizon.

In bandit learning-based strategy, a quantity termed as expected cumulative regret is often used to characterize the learning performance, which represents the cumulative difference between the reward of the chosen actions and the maximum expected reward. Accordingly, the objective of anti-jamming strategy is equivalent to minimizing the expected cumulative regret. We compare DT-kl-UCB^{++}-based anti-jamming strategy with the UCB and kl-UCB^{++} strategies. Figure 8 presents the growth of cumulated regret with time of all these anti-jamming strategies. As expected, it can be observed that the cumulative regret performance of DT-kl-UCB^{++} anti-jamming strategy clearly outperforms the UCB, kl-UCB^{++} strategies and Lai & Robbins lower bound.

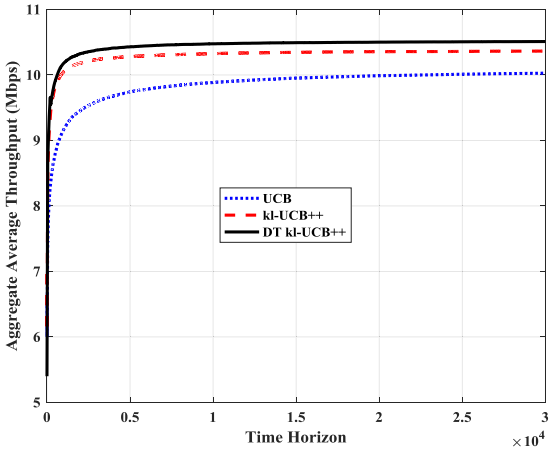

FIGURE 9. Aggregate average throughput w.r.t. time horizon.

Figure 9 compares the aggregate average throughput achieved by UCB and kl-UCB^{++} and DT-kl-UCB^{++} strategies. In the figure, we find that the UCB and kl-UCB^{++} strategies perform slightly worse than the anytime DT-kl-UCB^{++} strategy.

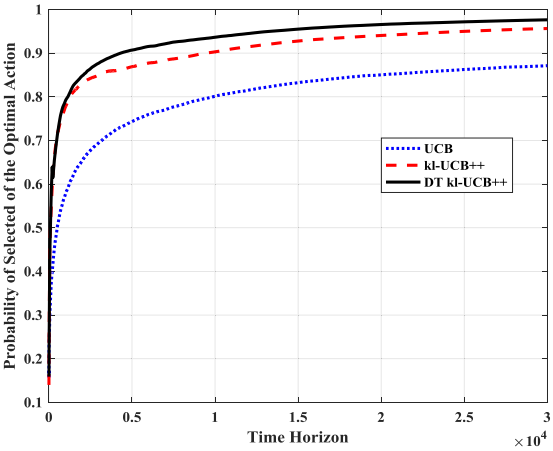

FIGURE 10. Probability of selection of the optimal action.

The probability of selection of the optimal action is shown in Figure 10 for different strategies. Similarly, it can be observed that the proposed DT kl-UCB^{++} strategy enjoys more opportunity to select the optimal action than the other non-anytime strategies.

VI. CONCLUSION

This paper has dealt with the potential and feasibility of applying decentralized selfish bandit anti-jamming strategy to aeronautic swarm network. We analyze the main characteristics of ASN in electromagnetic spectrum warfare scenario and establish an adversarial multiuser multi-armed

bandit model. Then, we propose a doubling trick kl-UCB^{++} bandit-based multidomain anti-jamming strategy to cope with this model. We highlight practical issues such as accurate reward generation from jamming sensing results and anytime kl-UCB^{++} algorithm design when applying bandit learning methods into ASN. Our studies show that the proposed multidomain anti-jamming strategy is able to achieve larger average throughput and low cumulative regret than state-of-the-art bandit learning strategies. Even though each radio performs anti-jamming by selfish and has no knowledge of the number of players, which is appealing to engineering implementation in dynamic ASN scenarios. In addition, the presented DT-kl-UCB^{++} bandit strategy is only an heuristic and lacks systematic theoretical proof, which is left for our future work.

REFERENCES

[1] L. Gupta, R. Jain, and G. Vaszkun, "Survey of important issues in UAV communication networks," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1123–1152, 2nd Quart., 2016.

[2] B. Wang, Y. Wu, K. J. R. Liu, and T. C. Clancy, "An anti-jamming stochastic game for cognitive radio networks," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 4, pp. 877–889, Apr. 2011.

[3] S. Singh and A. Trivedi, "Anti-jamming in cognitive radio networks using reinforcement learning algorithms," in *Proc. 9th Int. Conf. Wireless Opt. Commun. Netw. (WOCN)*, Indore, India, Sep. 2012, pp. 1–5.

[4] M. A. Aref, S. K. Jayaweera, and S. Machuzak, "Multi-agent reinforcement learning based cognitive anti-jamming," in *Proc.IEEE Wireless Commun. Netw. Conf. (WCNC)*, San Francisco, CA, USA, Mar. 2017, pp. 1–6.

[5] Y. Wang, Y. Xu, L. Shen, C. Xu, and Y. Cheng, "Two-dimensional pomdp-based opportunistic spectrum access in time-varying environment with fading channels," *J. Commun. Netw.*, vol. 16, no. 2, pp. 217–226, Apr. 2014.

[6] L. Jia, Y. Xu, Y. Sun, S. Feng, L. Yu, and A. Anpalagan, "A multi-domain anti-jamming defense scheme in heterogeneous wireless networks," *IEEE Access*, vol. 6, pp. 40177–40188, 2018.

[7] S. Han, X. Li, L. Yan, J. Xu, Z. Liu, and X. Guan, "Joint resource allocation in underwater acoustic communication networks: A game-based hierarchical adversarial multiplayer multiarmed bandit algorithm," *Inf. Sci.*, vols. 454–455, pp. 382–400, Jul. 2018.

[8] S. Sawant, R. Kumar, M. K. Hanawal, and S. J. Darak. (2018). "Learning to coordinate in a decentralized cognitive radio network in presence of jammer." [Online]. Available: https://arxiv.org/abs/1803.06810

[9] P. Ménard and A. Garivier, "A Minimax and asymptotically optimal algorithm for stochastic bandits," in *Proc. 28th Int. Conf. Algorithmic Learn. Theory*. Kyoto, Japan, 2017, pp. 223–237.

[10] K. Wanuga, N. Gulati, H. Saarnisaari, and K. R. Dandekar, "Online learning for spectrum sensing and reconfigurable antenna control," in *Proc. 9th Int. Conf. Cogn. Radio Oriented Wireless Netw. Commun. (CROWNCOM)*, Oulu, Finland, Jun. 2014, pp. 508–513.

[11] T. Riihonen, D. Korpi, O. Rantula, H. Rantanen, T. Saarelainen, and M. Valkama, "Inband full-duplex radio transceivers: A paradigm shift in tactical communications and electronic warfare?" *IEEE Commun. Mag.*, vol. 55, no. 10, pp. 30–36, Oct. 2017.

[12] T. Riihonen, D. Korpi, M. Turunen, and M. Valkama, "Full-duplex radio technology for simultaneously detecting and preventing improvised explosive device activation," in *Proc. Int. Conf. Military Commun. Inf. Syst. (ICMCIS)*, Warsaw, Poland, May 2018, pp. 1–4.

[13] J. Hu, K. Shahzad, S. Yan, X. Zhou, F. Shu, and J. Li, "Covert communications with a full-duplex receiver over wireless fading channels," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kansas City, MO, USA, May 2018, pp. 1–6.

[14] Y. Chen, "Improved energy detector for random signals in Gaussian noise," *IEEE Trans. Wireless Commun.*, vol. 9, no. 2, pp. 558–563, Feb. 2010.

[15] A. Singh, M. R. Bhatnagar, and R. K. Mallik, "Performance of an improved energy detector in multihop cognitive radio networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 2, pp. 732–743, Feb. 2016.

[16] R. Bonnefoi, L. Besson, C. Moy, E. Kaufmann, and J. Palicot, "Multi-armed bandit learning in iot networks: Learning helps even in non-stationary settings," in *Proc. Int. Conf. Cogn. Radio Oriented Wireless Netw.*, Lisbon, Portugal, 2017, pp. 173–185.

[17] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "Gambling in a rigged casino: The adversarial multi-armed bandit problem," in *Proc. IEEE 36th Annu. Found. Comput. Sci.*, Milwaukee, WI, USA, Oct. 1995, pp. 322–331.

[18] L. Besson and E. Kaufmann. (2018). "What doubling tricks can and can't do for multi-armed bandits." [Online]. Available: https://arxiv.org/abs/1803.06971

[19] F. Wilhelmi, S. Barrachina-Muñoz, C. Cano, B. Bellalta, A. Jonsson, and G. Neu. (2018). "Potential and pitfalls of multi-armed bandits for decentralized spatial reuse in WLANs." [Online]. Available: https://arxiv.org/abs/1805.11083

[20] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The non-stochastic multiarmed bandit problem," *SIAM J. Comput.*, vol. 32, no. 1, pp. 48–77, 2002.

[21] M. Li, D. Yang, J. Li, M. Li, and J. Tang, "SpecWatch: A framework for adversarial spectrum monitoring with unknown statistics," *Comput. Netw.*, vol. 143, pp. 176–190, Oct. 2018.

[22] T. Lattimore and C. Szepesvári, *Bandit Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2018.[Online]. Available: http://downloads.tor-lattimore.com/book.pdf

HAITAO LI received the Ph.D. degree from the University of Electronic Science Technology of China, in 2001. From 2002 to 2004, he was a Postdoctoral Researcher with Tsinghua University. He is currently an Associate Professor with the Faculty of Information Technology, Beijing University of Technology. His research interests include wireless communication, signal processing, the Internet of Things, and machine learning application.

JIAWEI LUO received the B.S. degree in Internet of Things engineering from Henan Polytechnic University, Henan, China, in 2016. He is currently pursuing the M.S. degree in communications and information systems with the Faculty of Information Technology, Beijing University of Technology. His research interest includes reinforcement learning for anti-jamming communication techniques.

CHANGJUN LIU is currently pursuing the M.S. degree in communications and information systems with the Faculty of Information Technology, Beijing University of Technology. His research interest includes deep learning for anti-jamming communication techniques.