Online Learning-Based Optimal Primary User Emulation Attacks in Cognitive Radio Networks

Monireh Dabaghchian, Amir Alipour-Fanid and Kai Zeng
Electrical and Computer Engineering Department
George Mason University
Fairfax, Virginia 22030
Email: {mdabaghc, aalipour, kzeng2}@gmu.edu

Qingsi Wang
Qualcomm Research
San Diego, California
Email: qingsiw@qti.qualcomm.com

***Abstract*—In a cognitive radio (CR) network, a secondary user learns the spectrum environment and dynamically accesses the channel where the primary user is inactive. At the same time, a primary user emulation (PUE) attacker can send falsified primary user signals and prevent the secondary user from utilizing the available channel. Although there is a large body of work on PUE attack detection and defending strategies, the best attacking strategies that an attacker can apply have not been well studied. In this paper, for the first time, we study the optimal PUE attack strategies without any prior knowledge on the primary user activity characteristics and the secondary user access strategies. We formulate the problem as a non-stochastic online learning problem where the attacker needs to dynamically decide the attacking channel in each time slot based on its attacking experience in previous slots. The challenge in our problem is that the PUE attacker cannot observe the reward on the attacked channel because it never knows if a secondary user ever tries to access it. To solve this challenge, we propose an attack-but-observe-another (ABOA) scheme, in which the attacker attacks one channel in the spectrum sensing phase, but observes at least one other channel in the data transmission phase. We propose two non-stochastic online learning-based attacking algorithms, EXP3-DO and OPT-RO, which select the observing channel deterministically based on the attacking channel and uniform randomly, respectively. EXP3-DO employs an existing theoretical framework and is suboptimal. OPT-RO is based on the new proposed theoretical framework and is optimal. They achieve regret in the order of $O(T^{\frac{2}{3}})$ and $O(\sqrt{T})$, respectively. T is the number of slots the CR network operates. We also generalize OPT-RO to multichannel observation cases. We show consistency between simulation and analytical results under various system parameters.**

I. Introduction

Nowadays the demand to increased wireless bandwidth is growing rapidly due to the increasing growth in mobile applications, which raises the spectrum shortage problem. To address this problem, Federal Communications Commission (FCC) has authorized opening spectrum bands owned by licensed primary users (PU) to unlicensed secondary users (SU) when the primary users are inactive [1]. Cognitive radio (CR) is a key technology that enables secondary users to learn the spectrum environment and dynamically access the best available channel. Meanwhile, an attacker can send signals emulating the primary users to manipulate the spectrum environment, and can thus prevent a secondary user from utilizing the available channel. This attack is called primary user emulation (PUE) attack [2]–[7].

Existing works on PUE attacks mainly focus on PUE attack detection [8]–[10] and defending strategies [4]–[7]. However, there is a lack of study on the optimal PUE attacking strategies. Better understanding of the optimal attacking strategies will enable us to quantify the severeness or impact of a PUE attacker on the secondary user's throughput. It will also shed light on the design of defending strategies.

In practice, an attacker may not have any prior knowledge of the primary user activity characteristics or the secondary user's dynamic spectrum access strategies. Therefore, it needs to learn the environment and attack at the same time. In this paper, for the first time, we study the optimal PUE attacking strategies without any assumption on the prior knowledge of the primary user activity or secondary user accessing strategies. We formulate this problem as a non-stochastic online learning problem.

Different from all the existing works on online learning based PUE attack defending strategies [4]–[7], in our problem, an attacker cannot observe the reward on the attacking channel. Considering a time-slotted system, the PUE attack usually happens in the channel sensing period, in which a secondary user attempting to access a channel conducts spectrum sensing to decide the presence of a primary user. If a secondary user is sensing the attacking channel, it will believe the primary user is active so that it will not transmit the data in order to avoid interfering with the primary user. In this case, the PUE attack is effective since it disrupts the secondary user's communication degrading its throughput and affects the knowledge of the spectrum availability of the secondary user. In the other case, if there is no secondary user attempting to sense or access the attacking channel, the attacker makes no impact on the secondary user, and the attack is ineffective. However, the attacker cannot differentiate between the two cases when it launches a PUE attack on a channel because sensing is a passive behavior.

To address this challenge, we propose a new attacking scheme called attack-but-observe-another (ABOA), in which an attacker selects one channel to attack in the channel sensing period but selects at least one other channel to observe in the data transmission phase in the same slot. This strategy is motivated by the observation that a short channel sensing phase is usually followed by a longer data transmission phase in which an attacker is able to switch to at least one other

channel to observe the secondary user's activity.

Based on the basic idea of the ABOA scheme, we propose two non-stochastic online learning algorithms to dynamically decide the attacking and observing channels in each time slot in a slotted system. One online learning algorithm, EXP3-DO, decides the observing channel deterministically based on the attacking channel, and the other, OPT-RO, decides the observing channel randomly. Both of the algorithms dynamically choose the attacking channel in each slot according to the observed past activity of the secondary user.

The proposed EXP3-DO algorithm applies the existing theoretical framework [11] in which it can be categorized as a specific group of graphs called partially observable graphs. [11] drives a regret in the order of $O(T^{\frac{2}{3}})$ for partially observable graphs. OPT-RO is based on a new theoretical framework we propose and we prove its regret is in the order of $O(\sqrt{T})$. T is the number of slots the CR network operates. Since the regret of any non-stochastic online learning algorithm in this problem is $\Omega(\sqrt{T})$ [11], [12], OPT-RO is the optimal PUE attacking strategy without any prior knowledge of the primary user activity and secondary user access strategies. We further generalize this optimal learning algorithm to multichannel observation case and analyze its regret.

We summarize the contributions of this paper as follows:

- We propose a new PUE attack scheme, ABOA, in which a PUE attacker dynamically selects one channel to attack but chooses at least another channel to observe in each time slot.
- We formulate the PUE attack as a non-stochastic online learning problem without any assumption on the prior knowledge of either primary user activity characteristics or secondary user dynamic channel access strategies.
- We propose two online learning algorithms, EXP3-DO and OPT-RO, to dynamically decide the attacking and observing channels. EXP3-DO has a suboptimal regret of $O(T^{\frac{2}{3}})$ and we prove that OPT-RO achieves an optimal regret of $O(\sqrt{T})$. The algorithm and the proof for the optimal one are further generalized to multichannel observation cases.
- Theoretical contribution: For the first time we propose an online-learning algorithm that despite observing the actions partially in each time slot, can achieve an optimal regret order. We accomplish it by making randomization on the actions rather than deterministically since it can lead to full observation during several time-slots.
- We conduct extensive simulations to evaluate the performance of the proposed algorithms under various system parameters.

Through theoretical analysis and extensive simulations under various system parameters, we find that the number of observing channels has an important impact on the attacker's regret. Our theoretical analysis shows that the regret is proportional to $\sqrt{1/m}$, where m is the number of channels that can be observed by the attacker. This non-linear relationship implies that the regret can be tremendously reduced when the number of observing channels increases in the beginning. However, when more observing channels are added, the reduction in regret becomes marginal. Therefore, a relatively small number of observing channels is sufficient to approach the optimal regret. Furthermore, the attacker's regret is also proportional to $\sqrt{K \ln K}$, where K is the total number of channels.

The rest of the paper is organized as follows. Section II discusses the related work. Section III describes the system model and problem formulation. Section IV proposes the attack strategies and derives the upper-bound one their regret. Simulation results are presented in section V. Finally section VI concludes the paper and discusses future work.

II. Related Work

A. PUE Attacks in Cognitive Radio Networks

Existing work on PUE attacks mainly focus on PUE attack detection [8]–[10] and defending strategies [4]–[7].

There are few works discussing attacking strategies under dynamic spectrum access scenarios. In [4], [5], the attacker applies partially observable Markov decision process (POMDP) framework to find the attacking channel in each time slot. It is assumed that the attacker can observe the reward on the attacking channel. That is, the attacker knows if a secondary user is ever trying to access a channel or not. In [6], [7], it is assumed that the attacker is always aware of the best channel to attack. However, there is no methodology proposed or proved on how the best attacking channel can be decided.

The optimal PUE attack strategy without any prior knowledge of the primary user activity and secondary user access strategies is not well understood. In this paper, we fill this gap by formulating this problem as a non-stochastic online learning problem. Our problem is also unique in that the attacker cannot observe the reward on the attacking channel due to the nature of PUE attack.

B. Multi-armed Bandit Problems

There is a rich literature about online learning algorithms. The most related ones to our work are multi-armed bandit (MAB) problems [13]–[17]. The MAB problems have many applications in cognitive radio networks with learning capabilities [6], [18], [19]. In such problems, an agent plays a machine repeatedly and obtains a reward when it takes a certain action at each time. Any time when choosing an action the agent faces a dilemma of whether to take the best rewarding action known so far or to try other actions to find even better ones. Trying to learn and optimize his actions, the agent needs to trade off between exploration and exploitation. On one hand the agent needs to explore all the actions often enough to learn which is the most rewarding one and on the other hand he needs to exploit the believed best rewarding action to minimize his overall regret.

MAB problems can be categorized into stochastic ones and non-stochastic ones. In stochastic ones [13], the reward of the actions is assumed under a parametrized distribution, but the parameter of the distribution is unknown. In non-stochastic ones [14], [15], no probabilistic model is assumed for the reward and the reward can be arbitrary. Alternatively, the non-stochastic MAB problems can be interpreted as focused on some unknown sample path of reward.

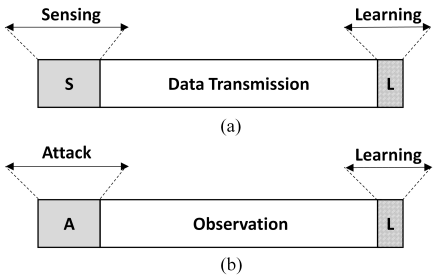


Figure 1: Time slot structure of a) the SU and b) the attacker.

In our case, the PUE attacker does not have any prior knowledge of the primary user activity or secondary user channel accessing strategies, and consequently the non-stochastic MAB problem serves as a better framework. For most existing non-stochastic MAB frameworks [14], [15], the agent needs to observe the reward on the taken action. Therefore, these frameworks cannot be directly applied to our problem where a PUE attacker cannot observe the reward on the attacking channel.

Most recently, Alon et al. generalize the cases of MAB problems [11]. They show in their work that if an agent takes an action without observing its reward, but observes the reward of all the other actions, it can achieve an optimal regret in the order of $O(\sqrt{T})$. However, the agent can only achieve a suboptimal regret of $O(T^{\frac{2}{3}})$ if it cannot observe the rewards on all the other actions simultaneously and if no action is left unobserved.

In this paper, we advance this theoretical study by proposing a strategy, OPT-RO, that can achieve the optimal regret in the order of $O(\sqrt{T})$ without observing the rewards on all the channels other than the attacking one simultaneously. In OPT-RO, the attacker uniform randomly selects at least one channel other than the attacking one to observe in each time slot.

C. Jamming Attacks

There are several works formulating jamming attacks and anti-jamming strategies as online learning problems [19]–[22]. In jamming attacks, an attacker can usually observe the reward on the attacking channel where an ongoing communication between legitimate users can be detected. Also it is possible for the defenders whether they are defending against a jammer or a PUE attacker to observe the reward on the accessed channel. PUE attacks are different in that the attacker attacks the channel sensing phase and prevents a secondary user from utilizing an available channel. As a result, a PUE attacker cannot observe the instantaneous reward on the attacking channel. That is, it cannot decide if an attack is effective or not.

III. System Model and Problem Formulation

We consider a cognitive radio network consisting of several primary users, multiple secondary users, and one attacker. There are K ($K > 1$) channels in the network. We assume the system is operated in a time-slotted fashion.

A. Primary User

We assume the primary users arbitrarily get on and off on all the channels. In each time slot, each primary user is either active (on) or inactive (off). We assume the on-off sequence of PUs on the channels is unknown to the attacker a priori. In other words, the PU activity can follow any distribution or can even be arbitrary.

B. Secondary User

The secondary users may apply any dynamic spectrum access policy [23]–[25]. In each time slot, each SU conducts spectrum sensing, data transmission, and learning in three consecutive phases as shown in Fig. 1(a).

At the beginning of each time slot, each secondary user senses a channel it attempts to access. If it finds the channel idle (i.e., the primary user is inactive), then accesses this channel; Otherwise, it remains silent till the end of the current slot in order to avoid interference to the primary user. At the end of the slot, it applies a learning algorithm to decide which channel it will attempt to access in the next time slot based on its past channel access experience.

We assume the secondary users cannot differentiate the attacking signal from the genuine primary user signal. That is, in a time slot, when the attacker launches a PUE attack on the channel a secondary user attempts to access, the secondary user will not transmit any data on that channel.

C. Attacker

We assume a smart attacker with learning capability. In one time slot, the attacker conducts the following actions in three consecutive phases: Attacking, observation, and learning as shown in Fig. 1(b).

In the attacking phase, the attacker launches the PUE attack by sending signals emulating the primary user's signals [5] to attack the SUs. We do not consider attack on the PUs. Note that, in this phase, the attacker has no idea if its attack is effective or not. That is, it does not know if a secondary user is ever trying to access the attacking channel or not.

In the observation phase following the attacking phase, the attacker switches to at least one other channel to observe the communication on that channel. It may detect a primary user signal, a secondary user signal, or nothing on the observing channel.

In the learning phase at the end of the time slot, the attacker decides which channel it attempts to attack in the next slot based on its past observations.

D. Problem Formulation

Given the above system model, the challenge for the attacker is to determine in each time slot *which channel to launch attack and which channel to observe.*

Since the attacker needs to learn and attack at the same time and it has no prior knowledge of the primary user activity or secondary user access strategies, we formulate this problem as a non-stochastic online learning problem.

We consider T as the total number of time slots the network operates. We define $x_t(j)$ as the attacker's reward on channel

Table I: Main Notation

T	total number of time slots
K	total number of channels
k	index of the channel
R	total reward achieved by the attacker
I_t	index of the channel to be attacked
J_t	index of the channel to be observed
γ	exploration rate used in algorithms 1 and 2
η	learning rate used in algorithms 1 and 2
$\omega_t(i)$	weight assigned to channel i at time slot t

j at time slot t ($1 \leq j \leq K$, $1 \leq t \leq T$). Without loss of generality, we normalize $x_t(j) \in [0, 1]$.

More specifically:

$$x_t(j) = \begin{cases} 1, & \textit{SU is on channel } j \textit{ at time } t \\ 0, & o.w. \end{cases} \tag{1}$$

Suppose the attacker applies a learning policy φ to select the attacking and observing channels. The aggregated expected reward of attacks by time slot T is equal to:

$$G^{\varphi}(T) = E^{\varphi}\left[\sum_{t=1}^{T} x_t(I_t)\right] \tag{2}$$

The attacker's goal is to maximize the expected value of the aggregated attacking reward, thus to minimize the throughput of the secondary user.

$$\text{maximize} \quad G^{\varphi}(T) \tag{3}$$

For a learning algorithm, regret is commonly used to measure its performance. The regret of the attacker can be defined as follows.

$$Regret = G_{max} - G^{\varphi}(T) \tag{4}$$

where

$$G_{max} = \max_{j} \sum_{t=1}^{T} x_t^{\varphi}(j) \tag{5}$$

The regret measures the gap between a learning algorithm and the maximum accumulated reward the attacker can obtain when it stays on one optimal channel. This regret is usually called the weak regret [15]. Then the problem can be transformed to minimize the regret.

Table 1 summarizes the notation used in this paper.

IV. Online Learning-Based Attacking Strategy

In this section we propose two non-stochastic online learning algorithms for the attacker to decide which channel to attack and observe in each time slot. These algorithms do not require the attacker to observe the reward on the attacked channel, but assuming the attacker can observe the reward on at least one other channel.

In the following algorithms, we assume the attacker can observe the reward on only one other channel. We will generalize it to the case of multiple channel observation in Section IV-C.

Both of the algorithms are considered as no-regret algorithms, in which the incremental regret between two consecutive time slots diminishes to zero as time goes to infinity. The first one, EXP3-DO, achieves a regret in the order of $O(T^{\frac{2}{3}})$ which is a suboptimal learning algorithm and the latter, OPTRO, is optimal with regret in the order of $O(\sqrt{T})$.

A. Attacking Strategy 1: EXP3-DO

In this algorithm, the attacking channel selection is based on the accumulated reward distribution on all the channels. While the observing channel is deterministically dependent on the attacking channel. The attacker always observes the channel next to the attacking one. It rounds to channel 1 when it attacks channel K. So we call this algorithm EXP3-DO (EXP3 with deterministic observation). This is in comparison to EXP3 [15] in which the rewards are observed on the same chosen action. Fig. 2(a) shows the observation strategy employed by the attacker when $K = 6$. The Algorithm 1 shows the online learning-based attack strategy employed by the attacker.

Algorithm 1: EXP3-DO, EXP3 with Deterministic Observation

Parameters: $\gamma \in [0, e-2]$, $\eta \in (0, \gamma/K]$

Initialization: $\omega_1(i) = 1, \quad i = 1, ..., K$

For each $t = 1, 2, ...$

1. Set $p_t(i) = (1-\gamma)\frac{\omega_t(i)}{\sum_{j=1}^{K}\omega_t(j)} + \frac{\gamma}{K}, \quad i = 1, ..., K$

2. Attack channel $I_t \sim p_t$ and accumulate the unobservable reward $x_t(I_t)$.

3. In the observation phase, choose channel $J_t := 1 + (I(t) \ mod \ K)$, and observe its reward $x_t(J_t)$ based on equation (1).

4. In the learning phase, for $j = 1, 2, ..., K$

$$\hat{x}_t(j) = \begin{cases} \frac{x_t(j)}{p_t(I_t)}, & j = J_t \\ 0, & o.w. \end{cases}$$

$$\omega_{t+1}(j) = \omega_t(j)\exp(\eta\hat{x}_t(j))$$

The design of the EXP3-DO is motivated by [11] in which the idea of bandit graphs is presented. A bandit graph is a generalization of EXP3 algorithm in which the actions and the following observations from the chosen action are presented by the nodes and the edges of a graph, respectively.

Based on the theoretical analysis in [11], any bandit graph that leaves no node un-observed, achieves a suboptimal regret order. The only way for a bandit graph to have an optimal regret order is if all the nodes or all except the chosen one

are observable at each action selection simultaneously or if always the chosen node is observable, similar to EXP3.

EXP3-DO as an attacking strategy is a baseline of applying current technologies for the PUE attacker's attack strategy. In other words, since the PUE attacker cannot scan through all the channels in each time slot due to the limited duration of a time slot, achieving an optimal regret order for the attacker is not possible by applying the current theoretical frameworks. As a result, in this algorithm the observation strategy is designed such that at least no action is left un-observed. So a suboptimal regret upper-bound can be guaranteed. The observation strategy stated here in not the only observation structure and any permutation of channels that creates a partially observable graph leads to the same regret bound.

We note that in Step 4 of Algorithm 1, only the weight of the observed channel is updated. The estimated reward $\hat{x}_t(j)$ is an unbiased estimate of the actual reward $x_t(j)$, i.e., conditional on all previously chosen channels before t, we have $E[\hat{x}_j(t)|I_1,\dots,I_{t-1}] = \frac{x_t(j)}{p_t(j')}p_t(j') = x_t(j)$ where $j' = (j-2)\ mod\ K+1$, i.e., the neighboring channel chosen for attack. To simplify our presentation, we will denote EXP3-DO by A1 in the following.

Theorem 1. *For any $K \geq 2$ and for any $\eta \leq \frac{\gamma}{K}$, the upper bound on the expected regret of Algorithm 1*

$$G_{max} - E[G_{A1}] \leq (\gamma + (e-2)\frac{\eta K}{\gamma})G_{max} + \frac{\ln K}{\eta}.$$

holds for any assignment of rewards for any $T > 0$.

By choosing appropriate values for γ and η, the above upper bound on the regret can be minimized.

Corollary 1. *For any $T > 0$ and the following values $\eta = \frac{\gamma^2}{K(e-2)}$, and $\gamma = \sqrt[3]{(e-2)K\ln K/g}$ where g is an upper-bound on the G_{max}. Then*

$$G_{max} - E[G_{A1}] \leq 3\sqrt[3]{(e-2)}\sqrt[3]{T^2 K \ln K}$$

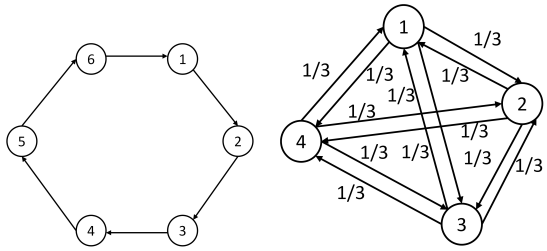
holds for any arbitrary assignment of rewards.

Proof. We sketch the proof as follows. Since $\gamma \leq (e-2)$, in order for the regret bound to be non-trivial, we need $g \geq \frac{K\ln K}{(e-2)^2}$. Then by getting the derivative, we find the optimal values for η and γ. Also T is an upper bound on the g since all the rewards are in $[0,1]$ and the network runs for T time slots, which gives us the result. □

Proof of Theorem 1. Our proposed observing strategy can be categorized as a partially observable graph in [11]. Because by choosing an action, the reward on only one other channel is observed not on all the actions. Also no node is left un-observed. In Theorem two of [11] the upper-bound of the regret of weakly observable graphs is proved. □

B. Attacking Strategy 2: OPT-RO

In this section, we propose an attacking strategy for the PUE attacker that despite observing channels partially, achieves an optimal regret order. The proposed optimal attacking strategy, called OPT-RO, observes all the channels uniformally at random. Fig. 2(b) shows the channel observation policy



(a) Deterministic, $K = 6$. (b) Random, $K = 4$

Figure 2: Channel observation strategy

for $K = 4$. For example, if channel 2 is attacked, one of the three other channels will be chosen uniformally at random to be observed. The values on the edges indicate the observation probabilities. Based on our analysis the upper bound on the attacker's regret is in the order of $O(\sqrt{T})$. The proposed algorithm is presented in Algorithm 2.

Algorithm 2 : OPT-RO, Optimal Online-Learning with Uniformly Randomized Observation

Parameters: $\gamma \in (0, 1/2]$,
$\eta \in \left(0, \frac{\gamma}{2(K-1)}\right]$

Initialization: $\omega_1(i) = 1, \quad i = 1, ..., K$

For each $t = 1, 2, ...$

1. Set $p_t(i) = (1-\gamma)\frac{\omega_t(i)}{\sum_{j=1}^{K}\omega_t(j)} + \frac{\gamma}{K}, \qquad i = 1, ..., K$

2. Attack channel $I_t \sim p_t$ and accumulate the unobservable reward $x_t(I_t)$.

3. Choose a channel J_t other than the attacked one uniformly at random and observe its reward $x_t(J_t)$ based on equation (1)

4. For $j = 1, ..., K$

$$\hat{x}_t(j) = \begin{cases} \frac{x_t(j)}{(1/(K-1))(1-p_t(j))}, & j = J_t \\ 0, & o.w. \end{cases}$$

$$\omega_{t+1}(j) = \omega_t(j)\exp(\eta\hat{x}_t(j)).$$

In Step 4 of Algorithm 2, in order to create $\hat{x}_t(j)$, an unbiased estimate of the actual reward $x_t(j)$, we divide the observed reward, $x_t(J_t)$, by $(1/(K-1))(1-p_t(J_t))$ which is the probability of choosing channel J_t to be observed. In other words, channel J_t will be chosen to be observed if it has not been chosen for attacking $(1-p_t(J_t))$ and second if it gets chosen uniformally at random from the rest of the channels, $(1/(K-1))$. In the following analysis, we denote OPT-RO by A2 for simplicity.

Theorem 2. *For any $K \geq 2$ and for any $\eta \leq \frac{\gamma}{2(K-1)}$, for the given randomized observation structure for the attacker the upper bound on the expected regret of Algorithm 2,*

$$G_{max} - E[G_{A2}] \leq 2(e-2)(K-1)\eta G_{max} + \frac{\ln K}{\eta}$$

holds for any assignment of rewards for any $T > 0$.

Similarly, we can minimize the regret bound by choosing an appropriate value for η.

Corollary 2. *For any $T > 0$, we consider g as an upper-bound on the G_{max} and consider the following value for η*

$$\eta = \sqrt{\frac{\ln K}{2(e-2)(K-1)g}}$$

Then

$$G_{max} - E[G_{A2}] \leq 2\sqrt{2(e-2)}\sqrt{T(K-1)\ln K}$$

holds for any arbitrary assignment of rewards.

Proof. The proof can be done by taking the parallel steps as in the proof of Corollary 1. We omit it due to limited space. □

Proof of Theorem 2. We define $W_t = \omega_t(1) + \cdots + \omega_t(K) = \sum_{i=1}^{K} \omega_t(i)$, then at each time t we have,

$$\begin{aligned}
\frac{W_{t+1}}{W_t} &= \sum_{i=1}^{K} \frac{p_t(i) - \gamma/K}{1-\gamma} \exp(\eta \hat{x}_t(i)) \\
&\leq \sum_{i=1}^{K} \frac{p_t(i) - \gamma/K}{1-\gamma} \\
&\qquad [1 + \eta \hat{x}_t(i) + (e-2)(\eta \hat{x}_t(i))^2] \\
&\leq 1 + \frac{\eta}{1-\gamma} \sum_{i=1}^{K} p_t(i)\hat{x}_t(i) \\
&\qquad + \frac{(e-2)\eta^2}{1-\gamma} \sum_{i=1}^{K} p_t(i)\hat{x}_t^2(i) \\
&\leq \exp\left(\frac{\eta}{1-\gamma} \sum_{i=1}^{K} p_t(i)\hat{x}_t(i) \right. \\
&\qquad \left. + \frac{(e-2)\eta^2}{1-\gamma} \sum_{i=1}^{K} p_t(i)\hat{x}_t^2(i) \right) \qquad (6)
\end{aligned}$$

The equality follows from the definition of W_{t+1}, $\omega_{t+1}(i)$, and $p_t(i)$ respectively in Algorithm 2. Also the last inequality follows from the fact that $e^x \geq 1 + x$. Finally, the first inequality holds since $e^x \leq 1 + x + (e-2)x^2$ for $x \leq 1$. When $\eta \leq \frac{\gamma}{2(K-1)}$, the result, $\eta \hat{x}_t(i) \leq 1$, follows from the observation that either $\eta \hat{x}_t(i) = 0$ or $\eta \hat{x}_t(i) = \eta \frac{x_t(i)}{\frac{1}{K-1}(1-p_t(i))} \leq \eta(K-1)\frac{2}{\gamma} \leq 1$, since $x_t(i) \leq 1$ and $p_t(i) = (1-\gamma)\frac{\omega_t(i)}{\sum_{j=1}^{K} \omega_t(j)} + \frac{\gamma}{K} \leq 1 - \gamma + \frac{\gamma}{2} \leq 1 - \frac{\gamma}{2}$.

Now we take the logarithm of both sides of (6) and sum over t from 1 to T. We derive the following inequality on the left hand side of the equation which holds for any action j,

$$\begin{aligned}
\sum_{t=1}^{T} \ln \frac{W_{t+1}}{W_t} &= \ln \frac{W_{T+1}}{W_1} \\
&\geq \ln \omega_{T+1}(j) - \ln K \\
&= \eta \sum_{t=1}^{T} \hat{x}_t(j) - \ln K. \qquad (7)
\end{aligned}$$

As a result the inequality (6) will be equal to:

$$\begin{aligned}
\sum_{t=1}^{T} \hat{x}_t(j) - \sum_{t=1}^{T}\sum_{i=1}^{K} p_t(i)\hat{x}_t(i) &\leq \gamma \sum_{t=1}^{T} \hat{x}_t(j) + \\
(e-2)\eta \sum_{t=1}^{T}\sum_{i=1}^{K} p_t(i)\hat{x}_t^2(i) &+ \frac{\ln K}{\eta} \qquad (8)
\end{aligned}$$

Let $\acute{x}_t(i) = \hat{x}_t(i) - f_t$ where $f_t = \sum_{i=1}^{K} p_t(i)\hat{x}_t(i)$. We make the pivotal observation that (8) also holds for $\acute{x}_t(i)$ since $\eta \acute{x}_t(i) \leq 1$, which is the only key to obtain (8).

We also note that

$$\begin{aligned}
\sum_{i=1}^{K} p_t(i)\acute{x}_t^2(i) &= \sum_{i=1}^{K} p_t(i)(\hat{x}_t(i) - f_t)^2 \\
&= \sum_{i=1}^{K} p_t(i)\hat{x}_t^2(i) - f_t^2 \\
&\leq \sum_{i=1}^{K} p_t(i)\hat{x}_t^2(i) - \sum_{i=1}^{K} p_t^2(i)\hat{x}_t^2(i) \\
&= \sum_{i=1}^{K} p_t(i)(1 - p_t(i))\hat{x}_t^2(i) \qquad (9)
\end{aligned}$$

Substituting $\acute{x}_t(i)$ in equation (8) and combining with (9), we have

$$\begin{aligned}
&\sum_{t=1}^{T} (\hat{x}_t(j) - f_t) - \sum_{t=1}^{T}\sum_{i=1}^{K} p_t(i)(\hat{x}_t(i) - f_t) \\
&= \sum_{t=1}^{T} \hat{x}_t(j) - \sum_{t=1}^{T}\sum_{i=1}^{K} p_t(i)\hat{x}_t(i) \\
&\leq \gamma \sum_{t=1}^{T} \left(\hat{x}_t(j) - \sum_{i=1}^{K} p_t(i)\hat{x}_t(i) \right) \\
&+ (e-2)\eta \sum_{t=1}^{T}\sum_{i=1}^{K} p_t(i)(1 - p_t(i))\hat{x}_t^2(i) + \frac{\ln K}{\eta} \qquad (10)
\end{aligned}$$

Observe that $\hat{x}_t(j)$ is similarly designed as an unbiased estimate of $x_t(j)$. Then for the expectation with respect to the sequence of channels attacked by the horizon T, we have the following relations:

$$E[\hat{x}_t(j)] = x_t(j), E\left[\sum_{i=1}^{K} p_t(i)\hat{x}_t(i)\right] = E[x_t(I_t)]$$

and

$$\begin{aligned}
&E\left[\sum_{i=1}^{K} p_t(i)(1 - p_t(i))\hat{x}_t^2(i)\right] \\
&= E\left[\sum_{i=1}^{K} p_t(i)(K-1)x_t^2(i)\right] \leq (K-1).
\end{aligned}$$

We now take the expectation with respect to the sequence of channels attacked by the horizon T in both sides of the last inequality of (10). On the left hand side, we have

$$E\left[\sum_{t=1}^{T}\hat{x}_t(j)\right]-E\left[\sum_{t=1}^{T}\sum_{i=1}^{K}p_t(i)\hat{x}_t(i)\right]=G_{max}-E[G_{A2}]. \tag{11}$$

And for the right hand side we have

$$E[\text{R.H.S}]\leq\gamma(G_{max}-E[G_{A_2}])+(e-2)(K-1)\eta T+\frac{\ln K}{\eta}$$

Combining the last two equations we get,

$$(1-\gamma)(G_{max}-E[G_{A2}])\leq(e-2)(K-1)\eta T+\frac{\ln K}{\eta},$$

since G_{max} can be substituted by T, the above relation gives us the proof assuming $\gamma\leq 1/2$. □

The important observation is that, based on [11] such an algorithm, Algorithm 2, should give a suboptimal regret since it can be categorized as a partially observable graph. However, our analysis gives a tighter bound and shows not only it is tighter but also it is optimal.

C. Extension to Multiple Channel Observations

We generalize Algorithm 2 to the case of observing multiple channels. The attacker can observe more than only one channel in each time slot. The number of channels that can be observed by the attacker other than the attacked channel depends on the length of the time slot. At least one channel and at most $K-1$ channels will be observed by the attacker.

Corollary 3. *If the attacker can observe more than one channel at each time slot, then its regret will be equal to the following:*

$$G_{max}-E\left[G_{A2}\right]\leq 2\sqrt{2\left(e-2\right)}\sqrt{T\tfrac{K-1}{m}\ln K}$$

where m is the number of channels that can be observed by the attacker. This regret upper-bound shows a faster convergence rate by making more observations as is also shown in our simulations.

Proof. In order to do the proof, we substitute $1/\left(K-1\right)$ by $m/\left(K-1\right)$ in Step 4 of Algorithm 2 since in each time slot, m channels are being chosen uniformally at random. Then the regret is derived by following the analysis. □

V. Performance Evaluation

In this section, we present the simulation results to evaluate the validity of the proposed attack strategies and the theorems provided in the previous sections. All the simulations are conducted in MATLAB and the results achieved are averaged over 10,000 independent random runs.

We compare the performance of the two proposed learning algorithms, EXP3-DO and OPT-RO, and show that their regrets scale as $O(T^{\frac{2}{3}})$ and $O(\sqrt{T})$, respectively. We then examine the impact of different system parameters on its performance. The parameters include the number of time slots, total number of channels in the network, number of possible observations by the attacker in each time slot and the distribution on the PU activities. Also a secondary user's accumulated traffic is evaluated with and without the presence of a PUE attacker.

K primary users are considered, each acting on one channel. The primary users' on-off activity follows a Markovian chain or i.i.d. distribution in the network. Also the PU activities on different channels are independent. K idle probabilities are generated using MATLAB's rand function, one for each channel if PUs follow an i.i.d. Bernoulli distributions. If the channels follow Markovian chain, for each channel we generate three probabilities, $p01$, $p10$, and $p1$ as the transition probabilities from state 0 (off) to 1 (on), from 1 to 0, and the initial idle probability.

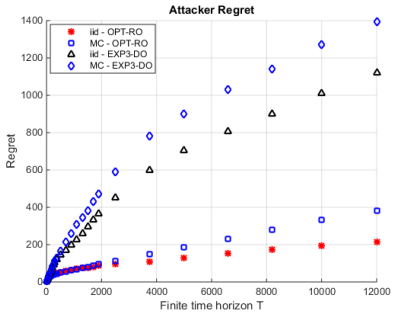
Since the goal here is to evaluate the PUE attacker's performance, for simplicity we consider one SU in the network. Throughout the simulations, we assume the SU employs an online learning algorithm called Hedge [14]. The assumption on the Hedge algorithm is that the secondary user is able to observe the rewards on all the channels in each time slot. Hedge provides the minimum regret among all the non-stochastic optimal learning based algorithms. As a result the performance of our proposed learning algorithms can be evaluated in the worst case scenario for the attacker.

The PUE attacker employs either of the proposed attacking strategies, EXP3-DO or OPT-RO as explained in each section. Throughout the simulations, we assume the attacker observes $m=1$ channel in each time slot, unless otherwise stated.

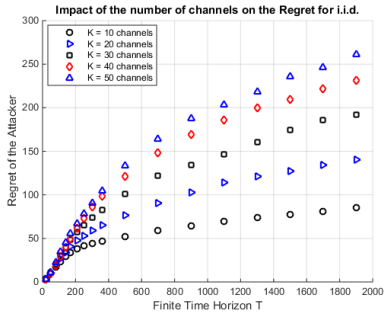
A. Comparison of the performance of EXP3-DO and OPT-RO

We compare the performance of each of the two proposed learning algorithms with the theoretical analysis from section IV. We consider a network of $K=10$ channels. Fig. 3(a) shows the simulation results. We can observe the following from the simulations.

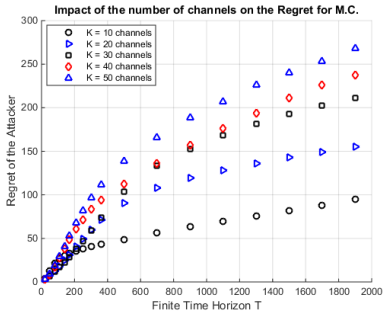
- For EXP3-DO, from Corollary (1), for the given system settings the theoretical upper-bound on the regret is equal to 4006.3. If we compare it to the actual regret occurred in the simulations (below 1400), we observe that it is less than the theoretical upper-bound for any type of PU activity which complies with our analysis. We also note that when we derived the theoretical upper-bound we did not make any assumption on the PU activity. The regret is only dependent on the K and T.
- The same discussion holds when the PUE attacker employs the OPT-RO. In this case, the theoretical regret upper-bound is equal to 1195.4 from Corollary (2). By comparing it to the actual regret happened in the simulations (below 400), we observe that it is less that the theoretical upper-bound regardless of the PU activity type.
- Both the learning based algorithms, EXP3-DO and OPT-RO have logarithmic shape; EXP3-DO has a higher slope which complies with our analysis.
- By comparing the results of the two algorithms, EXP3-DO has a much higher regret than the OPT-RO as is expected from the theoretical analysis.



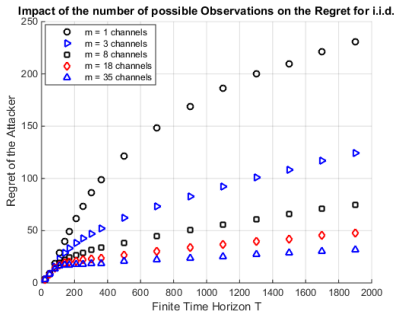
(a) Optimal versus Sub-Optimal learning



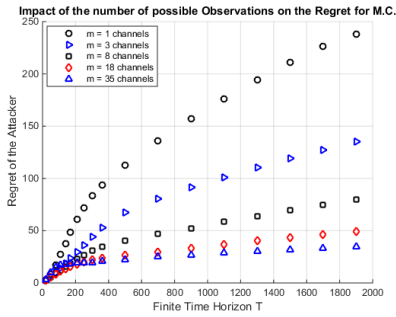
(b) Number of channels variable for i.i.d. case



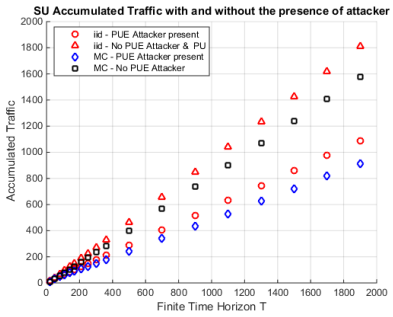
(c) Number of channels variable for M.C. case



(d) Different number of observations for i.i.d. case



(e) Different number of observations for M.C. case



(f) Accumulated Traffic of an SU

Figure 3: Simulation results under different PU activity assumptions

Since OPT-RO has a better theoretical regret bound and is also empirically better than EXP3-RO, in the following evaluations, we only examine the performance of the OPT-RO.

B. Impact of the number of channels

In this section we examine the impact of the number of channels in the network on the attacker's performance. We consider K variable from 10 to 50. Fig. 3 (b) and (c) show the attacker's regret when PUs follow i.i.d. distribution and Markovian chain respectively. In order to better observe the results we have plotted the figures for T from 1 to 2000. We can observe the following from the figures.

- In both figures regardless of the PUs activity type, the occurred regrets are below the regret upper-bound achieved from the theoretical analysis.
- In both figures, as the number of channels increases the regret increases as well as is expected based on the theoretical analysis from Corollary (2).
- As the number of channels increases, the regret does not increase linearly with it. Instead, the increment in the regret becomes marginal which complies with the theoretical analysis. Based on Corollary (2) the regret is proportional to $(\sqrt{K \ln K})$.

C. Impact of the number of observations in each time slot

We consider $K = 40$ channels in the network. The number of observing channels, m, varies from 1 to 35. Fig. 3 (d) and (e) show the performance of the OPT-RO for $m = 1, 3, 8, 18, 35$ when the PUs follow i.i.d. distribution and Markovian chain on the channels, respectively. We can observe the following from the simulation results.

- As the number of observing channels increases, the attacker achieves a lower regret. This observation complies with the Corollary (3) provided in Section IV-C.
- In the beginning, even adding a couple of more observing channels (from $m = 1$ to $m = 3$), the regret decreases significantly. The decrement in the regret becomes marginal as the number of observing channels becomes sufficiently large (e.g., from $m = 18$ to $m = 35$). This observation implies that, in order to achieve a good attacking performance, the attacker does not need to observe many channels in each time slot as is the case with Hedge [14]. In the simulation, when the number of observing channels ($m = 10$) is $\frac{1}{4}$ of the number of all channels ($K = 40$), the regret is approaching to the optimal.

D. Accumulated Traffic of SU with and without attacker

We set $K = 10$ and measure the accumulated traffic achieved by the SU with and without the presence of the attacker. Fig. 3 (f) shows that the accumulated traffic of the SU is largely decreased when there is a PUE attacker in the network for both types of PU activities.

VI. Conclusions and Future Work

In this paper, we studied the optimal PUE attacking strategies without any prior knowledge of the primary user activity characteristics and secondary user channel access policies. We formulated the PUE attack as a non-stochastic online learning

problem. We identified the uniqueness of PUE attack that a PUE attacker cannot observe the reward on the attacking channel, but is able to observe at least one other channel. We then proposed two online learning algorithms, EXP3-DO and OPT-RO, to dynamically choose attacking and observing channels for a PUE attacker in order to minimize its regret. EXP3-DO is based on the existing theoretical frameworks and it is suboptimal with regret in the order of $O(T^{\frac{2}{3}})$. OPT-RO introduces a new framework and we proved it is optimal having regret in the order of $O(\sqrt{T})$. Through theoretical analysis, we also found that the attacker's regret is proportional to $\sqrt{\frac{1}{m}}$ and $\sqrt{K \ln K}$. That is, the regret decreases when the attacker can observe more channels in each time slot and the regret increases when there are more channels in the network. Furthermore, when the number of observing channels is small, the regret decreases tremendously if we add a few more observing channels. However, the decreased regret will become marginal when more observing channels are added. This finding implies that an attacker may only need a small number of observing channels to achieve a good attacking performance.

The proposed optimal learning algorithm, OPT-RO, also advances the study of online learning algorithms. It deals with the situation where a learning agent cannot observe the reward on the action that is taken but can partially observe the reward of other actions. Before our work, the regret is proved to be in the order of $O(T^{\frac{2}{3}})$. Our algorithm achieves a tighter regret bound of $O(\sqrt{T})$ by randomizing the observing actions (channels) which is also optimal.

As of future work, we believe that our work can serve as a stepping stone to study many other problems. How to deal with multiple attackers will be interesting, especially distribution cases. One other interesting direction is to study the equilibrium between the PUE attacker(s) and secondary user(s) when both of them employ learning based algorithms. Integrating non-perfect spectrum sensing and the ability of PUE attack detection into our model will also be interesting especially that the PUE attacker may interfere with the PU during spectrum sensing period which can make it detectable by the PU.

ACKNOWLEDGMENT

This work was partially supported by the NSF under grant No. CNS-1502584 and CNS-1464487.

REFERENCES

[1] I.F. Akyildiz, Won-Yeol Lee, Mehmet C. Vuran, and S. Mohanty. A survey on spectrum management in cognitive radio networks. *Communications Magazine, IEEE*, 46(4):40–48, April 2008.

[2] Zhaoyu Gao, Haojin Zhu, Shuai Li, Suguo Du, and Xu Li. Security and privacy of collaborative spectrum sensing in cognitive radio networks. *Wireless Communications, IEEE*, 19(6):106–112, 2012.

[3] Qiben Yan, Ming Li, T. Jiang, Wenjing Lou, and Y.T. Hou. Vulnerability and protection for distributed consensus-based spectrum sensing in cognitive radio networks. In *INFOCOM, 2012 Proceedings IEEE*, pages 900–908, March 2012.

[4] Husheng Li and Zhu Han. Dogfight in spectrum: Jamming and anti-jamming in multichannel cognitive radio systems. In *Global Telecommunications Conference, 2009. GLOBECOM 2009. IEEE*, pages 1–6, Nov 2009.

[5] Husheng Li and Zhu Han. Dogfight in spectrum: Combating primary user emulation attacks in cognitive radio systems, part i: Known channel statistics. *Wireless Communications, IEEE Transactions on*, 9(11):3566–3577, November 2010.

[6] Husheng Li and Zhu Han. Dogfight in spectrum: Combating primary user emulation attacks in cognitive radio systems; part ii: Unknown channel statistics. *Wireless Communications, IEEE Transactions on*, 10(1):274–283, January 2011.

[7] Husheng Li and Zhu Han. Blind dogfight in spectrum: Combating primary user emulation attacks in cognitive radio systems with unknown channel statistics. In *Communications (ICC), 2010 IEEE International Conference on*, pages 1–6, May 2010.

[8] Kaigui Bian and Jung-Min "Jerry" Park. Security vulnerabilities in ieee 802.22. In *Proceedings of the 4th Annual International Conference on Wireless Internet*, WICON '08, pages 9:1–9:9, ICST, Brussels, Belgium, Belgium, 2008. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[9] Ruiliang Chen, Jung-Min Park, and Kaigui Bian. Robust distributed spectrum sensing in cognitive radio networks. In *INFOCOM 2008. The 27th Conference on Computer Communications. IEEE*, pages –, April 2008.

[10] Ruiliang Chen, Jung-Min Park, and J.H. Reed. Defense against primary user emulation attacks in cognitive radio networks. *Selected Areas in Communications, IEEE Journal on*, 26(1):25–37, Jan 2008.

[11] Noga Alon, Nicolò Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. *CoRR*, abs/1502.07617, 2015.

[12] Gábor Bartók, Dean P. Foster, Dávid Pál, Alexander Rakhlin, and Csaba Szepesvári. Partial monitoring—classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39(4):967–997, 2014.

[13] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2-3):235–256, May 2002.

[14] P. Auer, N. Cesa-Bianchi, Y. Freund, and Robert E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Foundations of Computer Science, 1995. Proceedings., 36th Annual Symposium on*, pages 322–331, Oct 1995.

[15] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, January 2003.

[16] K. Wang, Q. Liu, and L. Chen. Optimality of greedy policy for a class of standard reward function of restless multi-armed bandit problem. *Signal Processing, IET*, 6(6):584–593, August 2012.

[17] Yi Gai and B. Krishnamachari. Distributed stochastic online learning policies for opportunistic spectrum access. *Signal Processing, IEEE Transactions on*, 62(23):6184–6193, Dec 2014.

[18] Monireh Dabaghchian, Songsong Liu, Amir Alipour-Fanid, Kai Zeng, Xiaohua Li, and Yu Chen. Intelligence measure of cognitive radios with learning capabilities. In *Global Communications Conference, 2016 GLOBECOM 2016. IEEE*, Dec 2016.

[19] Q. Wang and M. Liu. Learning in hide-and-seek. *Networking, IEEE/ACM Transactions on*, PP(99):1–14, 2015.

[20] Hai Su, Qian Wang, Kui Ren, and Kai Xing. Jamming-resilient dynamic spectrum access for cognitive radio networks. In *Communications (ICC), 2011 IEEE International Conference on*, pages 1–5, June 2011.

[21] Q. Wang, K. Ren, and P. Ning. Anti-jamming communication in cognitive radio networks with unknown channel statistics. In *2011 19th IEEE International Conference on Network Protocols*, pages 393–402, Oct 2011.

[22] Q. Wang, P. Xu, K. Ren, and X. y. Li. Delay-bounded adaptive ufh-based anti-jamming wireless communication. In *INFOCOM, 2011 Proceedings IEEE*, pages 1413–1421, April 2011.

[23] K. Ezirim, S. Sengupta, and E. Troia. (multiple) channel acquisition and contention handling mechanisms for dynamic spectrum access in a distributed system of cognitive radio networks. In *Computing, Networking and Communications (ICNC), 2013 International Conference on*, pages 252–256, Jan 2013.

[24] Pochiang Lin and Tsungnan Lin. Optimal dynamic spectrum access in multi-channel multi-user cognitive radio networks. In *21st Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, pages 1637–1642, Sept 2010.

[25] Y. Liao, T. Wang, K. Bian, L. Song, and Z. Han. Decentralized dynamic spectrum access in full-duplex cognitive radio networks. In *2015 IEEE International Conference on Communications (ICC)*, pages 7552–7557, June 2015.