Jamming-resistant Multi-radio Multi-channel Opportunistic Spectrum Access in Cognitive Radio Networks

Qian Wang, *Member, IEEE,* Kui Ren, *Senior Member, IEEE,*
Peng Ning, *Member, IEEE* and Shengshan Hu, *Student Member, IEEE*

Abstract—For achieving optimized spectrum usage, most existing opportunistic spectrum sensing and access protocols model the spectrum sensing and access problem as a partially observed Markov decision process (POMDP) by assuming that the information states and/or the primary users' (PUs) traffic statistics are known *a priori* to the secondary users (SUs). While theoretically sound, the existing solutions may not be effective in practice due to two main concerns. First, the assumptions are not practical, as before the communication starts, PUs' traffic statistics may not be readily available to the SUs. Secondly and more seriously, existing approaches are extremely vulnerable to malicious jamming attacks. By leveraging the same statistic information and stochastic dynamic decision making process that the SUs would follow, a cognitive attacker with sensing capability can sense and jam the channels to be accessed by SUs while not interfering PUs. To address the above concerns, we formulate the anti-jamming multi-channel access problem as a non-stochastic multi-armed bandit (NS-MAB) problem, where the SU sender and the SU receiver adaptively choose channels to operate respectively. By leveraging probabilistically-shared information between the sender and the receiver, the proposed spectrum sensing and access protocol enables them to hop to the same set of channels with high probability while gaining resilience to jamming attacks without affecting PUs' activities. We analytically show the convergence of the learning algorithms and derive the performance bound based on *regret*. We further discuss the problem of tracking the best adaptive strategy and characterize the performance bound based on a new *regret*. Extensive simulations are conducted to validate the theoretical analysis. The results show that the probabilistic spectrum sensing and access protocol can overcome the limitation of existing solutions and is highly resilient to various jamming attacks even with jammed ACK information.

Index Terms—Anti-jamming, cognitive radio networks, multi-radio multi-channel.

1 Introduction

Cognitive radio is an emerging advanced radio technology in wireless access, with many promising benefits including dynamic spectrum sharing, robust cross-layer adaption and collaborative networking. Opportunistic spectrum access (OSA), which is at the core of cognitive radio technologies, has recently received increasing attention due to its great potential to improve the spectrum utilization efficiency and reliability [2]–[6]. The basic idea of OSA is that individual secondary users (SUs) dynamically search and access the spectrum vacancy to maximize the spectrum utilization while introducing limited interference to the primary users (PUs). In existing literature, the optimality of the channel sensing and access problem has been extensively studied from the single-channel access setting to the multi-channel access setting and from perfect sensing to imperfect sensing using various optimization tools. Most of existing solutions, however, inevitably assumed that traffic statistics are pre-known to SUs. In practice, such assumption may not always hold and more seriously, solutions based on this assumption are vulnerable to malicious jamming attacks. First, PU's traffic statistics (*i.e.,* initial information states, transition probabilities and the order of transition probabilities) may not be readily available to the SUs prior to the start of sensing actions. Without *a priori* information on the traffic patterns, those opportunistic spectrum sensing and access protocols cannot work. Moreover, a cognitive jammer with sensing capabilities can choose channels to sense by leveraging the same statistic information and stochastic dynamic decision making process. Based on the sensing results, the attackers then jam the idle channels potentially used by SUs without affecting activities of PUs. This is due to the fact that the structure of those sensing policies is fixed and the channel selection procedure that SUs follow is publicly known. Therefore, a jammer can predict which channels the SUs are going to use in *each* timeslot and prevent the spectrum from being utilized efficiently.

- *Qian Wang and Shengshan Hu are with the School of Computer Science, Wuhan University, China. E-mail: qianwang@whu.edu.cn.*
- *Kui Ren is with the Department of Computer Science and Engineering, The State University of New York at Buffalo, NY 14260, USA. E-mail: kuiren@buffalo.edu.*
- *Peng Ning is with the Department of Computer Science, North Carolina State University Raleigh, NC 27695, USA. E-mail: pning@ncsu.edu.*

A preliminary version [1] of this paper was presented at the 19th IEEE International Conference on Network Protocols (ICNP'11), Vancouver, BC Canada, 2011.

Traditional anti-jamming schemes, including both frequency hopping spread spectrum (FHSS) and direct-sequence spread spectrum (DSSS) [7], commonly rely on some pre-shared secrets (*i.e.*, hopping sequences and spreading codes) to achieve jamming-resistant communication. However, they are not directly applicable to CRNs due to the fact that the pre-sharing of secrets are not applicable in a dynamic SU network since SUs may never meet each other before the start of communication. Recently, uncoordinated frequency hopping (UFH) and uncoordinated direct-sequence spread spectrum (UDSSS) and their variations were proposed to eliminate the reliance on the pre-shared secrets [8]–[13]. The major problem with UFH and UDSSS is that they are both very expensive. For UFH, it takes a long time for an SU sender to transmit a message to an SU receiver. This is not practical for CRNs where SUs need to finish transmission quickly to yield the channel to PUs. On the other hand, UDSSS may take less time to deliver a message, but the message decoding process at the receiver side will incur a large cost. Moreover, applying UDSSS directly to the anti-jamming problem in CRNs results in a problem. UDSSS is commonly used in a broadcast communication setting where the communication channel is publicly known and SUs are using randomly-selected spreading codes to defend against jamming. In CRNs, it will cause large interference to PUs when they are also active on the same communication channel. In [14], [15], the problem of defending jamming attacks in cognitive radio networks was investigated using game-theoretic approaches. However, they only explored the single-channel case and assumed that the SU receiver can always communicate with the secondary sender (*i.e.*, they are considered as a single player) and sensing is perfect. In [16], the spectrum sensing problem was formulated under time-varying channels as an adversarial bandit problem. Similar to [14], [15], the authors only considered the case of single sensing channel and assumed that the SU receiver and the SU sender were considered as a single player. In [17], [18], anti-jamming games were investigated in CRNs. However the SU sender and the SU receiver are still considered as a single player, *i.e.*, they are assumed to stay coordinated by initialization with the same random seed.

To address the above limitations, in this paper we propose a decentralized and robust anti-jamming multi-channel spectrum access protocol for ad hoc CRNs, which can accommodate both the environment dynamics and the strategic behaviors of the jammers. Compared to existing UFH protocols, in a CRN setting, our protocol can adaptively choose the most likely "free" channels with high probability instead of randomly sensing and accessing channels. That is, the transceivers will selectively sense channels with high probability of non-occupancy by the jammer and the PUs, based on the history information of sensing and access. However, if the SU sender detects the presence of a PU on a sensed channel, it will remain silent and does not access that channel in the current timeslot. The sensing results together with the immediately following access results will be feedbacked to the sensing actions in the future timeslots. Therefore, communication efficiency can be significantly improved without affecting PU's activities. Different from existing deterministic dynamic spectrum access protocols, we adopt a probabilistic spectrum sensing and access approach, where the sender and the receiver sense/access channels in an unpredictable way and their knowledge of channels can converge due to shared information the jammer doesn't have. As will be shown, by leveraging probabilistically-shared information between the sender and the receiver, the communication efficiency can be retained while gaining resilience to jamming attacks. Compared to the preliminary version [1], in this paper we have made substantial improvements including both the theoretical performance analysis and experiments. New experimental results and full proofs of performance bounds are provided. We also discuss the problem of tracking the best compound (adaptive) strategy and characterize the bounds on the new regret. The main contributions of this paper are:

We propose the first online adaptive multi-channel jamming-resistant spectrum access protocol for ad hoc CRNs by formulating the anti-jamming problem as a non-stochastic MAB problem. We analytically show the convergence of the learning algorithms as T goes to infinity, *i.e.*, the time-averaged performance difference between the SU sender and the SU receiver's optimal strategies is no more than $\frac{20k}{\sqrt{T}\varepsilon}\sqrt{n \ln n}$, where $k = \max\{{k_s}^2, k_r\}$, k_r and k_s are the number of channels the receiver and the sender can access simultaneously in each timeslot, respectively, ε is the probability of sensing and n is the total number of channels. The proposed algorithm can be efficiently implemented in polynomial time.

We further consider the problem of tracking the best adaptive strategy and present an extension of our construction for anti-jamming spectrum access. We analyze the performance bound on the new regret defined based on the adaptive optimal strategy. We analytically show the time-averaged performance difference between the SU sender and the SU receiver's optimal strategies is upper bounded by $O(12k\sqrt{n \ln n})$, where $k = \max\{k_s, k_r\}$. Since k_s, k_r and n are pre-set system parameters, the performance bound is constant as T goes to infinity. The extended algorithm for tracking the best compound strategy can also be implemented in polynomial time.

We present a thorough quantitative performance characterization of the proposed scheme. The performance is evaluated by analyzing a practical metric– the expected time for message delivery with *high*

probability. We also perform an extensive simulation study to validate our theoretical results. It is shown that the proposed algorithm is efficient and highly effective against various jamming attacks even with jammed ACK information.

2 Models and Assumptions

2.1 System Model and Assumptions

In a typical cognitive radio network (CRN), there exist a primary user (PU) network and a secondary user (SU) network. To facilitate dynamic spectrum access, the spectrum is divided into n channels, each of which evolves independently and has the same total bandwidth. Different from most existing works, in our model we assume the channel statistics are not necessarily the same for n channels. In the system, PUs occupy and vacate the spectrum following a discrete-time Markov process, where channel i transits from busy state ("0") to idle state ("1") with probability p_{01}^i and stays in idle state ("1") with probability p_{11}^i. In the SU network, SUs seek spectrum opportunities among n channels. Specifically, SUs reserve a sensing interval in each timeslot to detect the presence of a PU. Based on the outcomes of sensing, the SU senders decide whether to take the opportunity to access the currently idle channels or not, and vacate the spectrum whenever PUs reclaim them. At the end of a timeslot, the SU receiver sends a short acknowledgement to the SU sender on the channel where a packet transmission is successful.

It is worth noting that we investigate the problem of robust spectrum sensing and access in an ad hoc SU network without a central controller for coordinating the SUs. Therefore, each autonomous SU aims to maximize its own performance by sensing and accessing the spectrum independently [2]. Different from most existing opportunistic spectrum access (OSA) protocols [2]–[6] where traffic statistics are known a *priori*, we consider a more general and practical scenario where traffic statistics are not available to SUs before the start of communication. For ease of exposition, in the following discussion we term one pair of communicating SUs as the sender and the receiver. In a multi-radio setting, the sender and the receiver are equipped with $k_s < n$ and $k_r < n$ radios, respectively, enabling them to access multiple channels simultaneously in each timeslot. Since SUs must not interference with active PUs in each timeslot, a SU sender thus senses $k_s < n$ and accesses only $k_a \leq k_s$ channels sequentially. At the receiver side, various efficient message verification schemes can be used for packet verification to defend against pollution attacks, and fragments that have passed integrity checks are reassembled to reconstruct the original message. To relax the strict synchronization between the sender and the receiver, we can let the hopping frequency of the receiver be much slower than the hopping frequency of the sender, so packet losses caused by the lack of synchronization between sender and receiver can be neglected. Note that in our model, we do not consider node authentication and message privacy, which are orthogonal to the security problems this work addresses.

2.2 Threat Model and Assumptions

In CRNs, PUs such as TV users are *licensed* users (*i.e.*, being protected by law) and usually well physically protected. From the jammer's perspective, it is very difficult to launch effecitve attacks, and there will be heavy penalties on the attackers if being detected [17]. Therefore, we assume the jammer does not have high incentive to attack PUs and risk itself in jamming the licensed bands when PUs are active. Instead, the jammer's target is on the secondary users (SUs), who are unlicensed users and only permitted to access the spectrum when not interfering with PUs. The SUs' access to the spectrum is opportunistic in nature without clear legal protection. Besides, SU networks are usually dynamic ad hoc networks formed by randomly deployed self-organizing wireless devices, where it is difficult to implement effective security countermeasures. A stealthy attacker can choose to jam any targeted SUs and prevent the targets from using the spectrum for communication. Note that such attacks against SUs by the jammer are stealthy and do not affect PUs' communications. That is, the attacker utilizes the same sensing interval to detect (sense) the activity of the PUs and only jam the idle channels (which are potentially used by SUs) based on the sensing outcomes.

We assume the jammer has similar radio capabilities as SUs. That is, in each timeslot, the jammer is capable of sensing and jamming k_j $(k_j < n)$ channels simultaneously. Assuming the jammer knows the whole spectrum access protocol, his objective then is to prevent the spectrum from being utilized efficiently by the legitimate SUs. Specifically, we consider four types of jammers with different jamming strategies:

Static jammer. A static jammer is an oblivious attacker, who selects the same set of channels in each timeslot to sense. Based on the sensing results he emits jamming signals on the sensed idle channels. Note that the jamming action is made independent of the sensing history the jammer may have observed in the past.

Random jammer. A random jammer is also an oblivious attacker, who selects a set of channels uniformly at random from the public set of n channels in each timeslot to sense. Based on the sensing results he emits jamming signals on the sensed idle channels. Similar to the static jammer, the jamming action is made independent of the sensing history he may have observed in the past.

Myopic jammer. An *myopic* jammer is a powerful cognitive attacker running the *myopic* algorithm, which

is a well-known OSA strategy and can achieve suboptimal performance (The principle of myopic policy will be shown later in Section 3). Initially, the jammer selects k_j channels to sense in each timeslot until all the n channels have been sufficiently sensed. Then he can make an accurate estimation of the traffic statistics using the sensing results, based on which he utilizes *myopic policy* to predict PUs' channel occupancy pattern and emits jamming signals on the *most* likely idle channels. Obviously, in each timeslot the jamming strategy is selected based on the sensing history and pre-known channel occupancy statistics.

Adaptive jammer. An adaptive jammer is also a cognitive attacker running an multi-armed bandit (MAB) algorithm, which is a online learning protocol (The MAB based learning protocol will be shown in section 4). The jammer selects k_j channels to sense in each timeslot and jams the sensed idle channels based on his sensing history and past observations.

Note that, in the power *adaptive jamming* attack model we assume the jammer can adjust his sensing and jamming strategies by leveraging the outcomes of jamming. In other words, we assume that the jammer knows whether he succeeds in jamming the transmitting channels (where both the sender and the receiver reside on in a timeslot) for all the past timeslots. We emphasize that it is almost impossible to implement such a powerful jammer in practice. However, for the purpose of performance comparison we show that SUs equipped with our anti-jamming spectrum sensing and access protocol are still resilient to such adaptive jamming attacks.

3 Vulnerability Analysis of Multi-channel Opportunistic Spectrum Access Protocols

In this section, we analyze the weakness of the existing multi-channel opportunistic spectrum access protocols under jamming attacks due to their *deterministic* feature, which motivates us to develop a *probabilistic* spectrum sensing and access approach in the next section. For ease of illustration, in the following we consider a SU network with a single sender-receiver pair, but the same ideas can also be applied and extended to a multi-user setting.

3.1 Opportunistic spectrum access with known channel traffic statistics

In the context of cognitive radio for opportunistic spectrum access, a single-channel access problem within the framework of POMDP was investigated, and myopic policies under both perfect and imperfect sensing cases have been investigated in [2]–[6]. The main idea of these schemes is that the sender chooses a subset of n channels to sense based on its past observations and gains a fixed reward if a channel is sensed idle. The objective of the sender then is to maximize the rewards that it can gain over a (potentially infinite) number of timeslots. It has been shown that this optimization problem can be solved by a stochastic dynamic programming (SDP) approach [19] to obtain optimal performance. To reduce the computation complexity of SDP caused by the expensive *backward induction* procedure, many researches has been focused on *index policies* and *myopic policy* that maximizes the conditional expected reward acquired at t was first proposed and explored in [2], [3]. By concentrating only on the present and completely ignores the future, *myopic approaches* achieves suboptimal performance in general. In myopic policy, it has also been shown that a sufficient statistic or the *information state* of the system for the optimal decision making is the belief vector $\Omega(t) = [\omega_1(t), \omega_2(t), \ldots, \omega_n(t)]$, where $\omega_i(t)$ is the conditional probability that channel i is idle in timeslot t. In timeslot t, a sensing action $a(t)$ denotes the k_s channels to be sensed. Let $K_i(t) \in \{0, 1\}$ denote whether an ACK on channel i is received or not in timeslot t. Given $a(t)$ and $K_i(t)$, the belief state in timeslot $t+1$ is given by [2]

$$\omega_i(t+1) = \begin{cases} p_{11}^i, & i \in a(t), K_i(t) = 1 \\ p_{01}^i, & i \in a(t), K_i(t) = 0 \\ \omega_i(t)p_{11}^i + (1-\omega_i(t))p_{01}^i, & i \notin a(t). \end{cases}$$

Assume all channels have the same transmission rate B_i, the myopic policy under Ω is defined as $\hat{a}(t) = \arg\max_{a(t)} \sum_{i \in a(t)} \omega_i(t) B_i$. Recently, the dynamic multi-channel access problem was studied under a special class of restless multi-armed bandit problems (RMBP) in [6], based on which an *index policy* called *Whittle's index policy* has also been applied in the dynamic spectrum access. Similar to myopic policy, the proposed *Whittle's index policy* enables the SU sender to choose those channels whose current states have the largest indices to sense and access. However, a strict constraint which requires the activating of exact $m = k_s$ arms/channels at each time step may cause the optimality to be lost, but even so the *Whittle's index policy* has the near optimal performance. Another interesting observation is that the Whittle's index policy has the same structure as the myopic policy when channels are stochastically identical.

3.2 Analysis of OSA Under Malicious Jamming Attacks

Although theoretically sound, almost all the existing OSA protocols (including index based policies) only work well in non-malicious environments. Among others, one key assumption made by the existing solutions is that the traffic statistics should be known a *priori*. Take index based policies for example, it is required that the initial belief vectors $\Omega(0)$ and the order of state transition probabilities (*i.e.*, p_{01}^i is greater

or less than p_{11}^i) on all channels be pre-known to SUs. In practice, however, these statistics may not be readily available [5]. More seriously, due to the deterministic nature of the channel/frequency selection procedure, those OSA protocols are vulnerable to malicious jamming attacks. That is, an intelligent jammer, who knows the traffic statistics of all channels or learns them through sensing and estimation by observing all channels, can leverage such information to predict which channel to be used. Since the index policies always choose the first k_s channels with the largest indices for sensing and accessing, the jammer can use the same dynamic decision process to perform effective jamming attacks. In the worst case, the communication can be completely jammed as the jammer maintains the same updates information for channel "index" as SUs in each timeslot. From a theoretical perspective, most of OSA protocols are formulated as optimization problems with deterministic solutions. For example, the *index policies* are established based on the stochastic model of the channel statistics. Consider the Whittle's index policy developed under the restless multi-armed bandit problems (RMBP) [20]. Since the evolvement of information state (belief vector) is known, the players (the sender and the receiver) can compute ahead of time exactly what payoffs (rewards) will be received from each arm (channel). Based on the above analysis, it is necessary and important to develop *probabilistic* OSA protocols that are resistant to various jamming attacks and can accommodate the special characteristics of CRNs.

To enhance the robustness of OSA, the problem of defending jamming attacks in cognitive radio networks was investigated using game-theoretic approaches [14], [15]. However, they only explored the single-channel case and assumed that the SU receiver can always communicate with the secondary sender (*i.e.*, they are considered as a single player) and sensing is perfect. In [16], the spectrum sensing problem was formulated under time-varying channels as an adversarial bandit problem. Similar to [14], [15], the authors only considered the case of single sensing channel and assumed that the SU receiver and the SU sender were considered as a single player. In this paper, we consider a more practical model and make a step towards the development of robust multi-radio multi-channel OSA protocols for CRNs.

4 Jamming-Resistant Multi-Radio Multi-Channel Opportunistic Spectrum Access

4.1 Scheme Overview

Based on the above analysis, we can see that when an attacker launches malicious jamming attacks to disrupt legitimate communications in SU networks, the channel statistics (which are determined by activities of PUs when there exists no jamming) cannot correctly reflect the true state (idle or busy) of the channel. That is, the rewards (*i.e.*, indications of successful packet receptions) associated with each channel cannot be modeled by a stationary distribution or no statistical assumptions can be made about the transition of information state and the generation of rewards. This is due to the dynamic behaviors of both PUs and jammers, *i.e.*, PUs occasionally occupy and free the channels and a jammer may adjust his sensing and jamming strategy to maximize the effect of jamming. These effects will make the generation of rewards arbitrarily change on channels in each timeslot. Motivated by this observation, it is necessary to keep an *exploration* of the best possible set of channels for transmission to adapt the dynamics of jammers and PUs. Meanwhile, it is also necessary to *exploit* the previously-chosen favorable set of channels as too much exploration will potentially underutilize them. Obviously, the proposed anti-jamming problem is thus the one balancing between *exploitation* and *exploration*, rather than only optimizations.

4.2 Problem Formulation: An Multi-player Game

In this paper, we consider a jamming and anti-jamming game among a SU sender, a SU receiver and a jammer under dynamic PU behaviors. To fully utilize the vacant spectrum, the objective of the SU sender-receiver is to choose the sensing, access/receiving actions in each timeslot to maximize the total expected rewards (*i.e.*, successfully received packets) over T timeslots. On the contrary, the jammer's objective is to minimize the total expected rewards to disrupt the legitimate SU communications. Since channel states (idle or busy) are not directly observable before channel sensing, the sender chooses k_s channels to sense during the sensing interval, where the sensing action is made based on all the past decisions and observations. Due to PUs' dynamic actions on a channel, the sender only chooses k_a ($k_a \leq k_s$) idle channels to access. At the receiver side, the receiver independently chooses k_r channels to receive, where the selection is also made based on all the past decisions and observations. During the same timeslot, the jammer chooses k_j channels to sense and jam the *sensed idle* channels based on his chosen jamming strategy.

Note that, although we consider a single SU pair in our anti-jamming problem, the proposed scheme can be directly applied to a SU communication network with multiple SU sender-receiver pairs. This is because each SU, which is autonomous in an ad hoc SU network, can utilize our proposed scheme to maximize its own performance by taking interference/collistions caused by other SU pairs as jamming signals. It is easy to see that when the number of other SU pairs in the neighborhood of the receiver which use the same channels is much less than n, the impact of unintentional interference can be negligible.

We next formalize the jamming and anti-jamming game using mathematical notation. We first number the channels/frequencies from 1 to n and construct the vector space $\{0,1\}^n$. Obviously, the sender's sensing and access strategy space and the receiver's receiving strategy space are denoted as $S_s \subseteq \{0,1\}^n$ of size $\binom{n}{k_s}$ and $S_r \subseteq \{0,1\}^n$ of size $\binom{n}{k_r}$, respectively. In a SU's strategy/vector, the value of the f-th ($f \in \{1,\ldots,n\}$) entry of a vector is 1 if the f-th channel is chosen for sending and access or receiving; 0 otherwise. Accordingly, the jamming strategy space for the jammer is denoted as $S_j \subseteq \{0,1\}^n$ of size $\binom{n}{k_j}$. Different from a SU's strategy, the value 0 in the f-th entry denotes that the jammer chooses the f-th channel to sense and jam and the value is 1 otherwise. Different from the above three parties, PUs' activities on the channels are independent of other parties's actions, and a PU's action/strategy can also be denoted as a vector $s_p \in \{0,1\}^n$, where the value 1 denotes the channel is idle and the value 0 denotes the channel is occupied.

During each timeslot, the sender, the receiver and the jammer choose their own respective strategies $s_s \in S_s$, $s_r \in S_r$ and $s_j \in S_j$, respectively. In each timeslot, assume the PU's strategy or activity is s_p. From the receiver's perspective, $s_s \wedge s_p \wedge s_j$ can be considered as a joint decision made by the sender, the PU and the jammer, where $\wedge$ denotes bitwise "AND" operation. We say that in timeslot t the sender, a *reward* "$g_{f,t} = 1$" is introduced for channel f if the f-th entry of $s_s \wedge s_p \wedge s_j$ is 1; otherwise no reward is received, *i.e.*, "$g_{f,t} = 0$". On the receiver side, the reception of a reward depends on the state of the channel f the receiver has *chosen* for packet reception. In addition, we use erasure coding combined with short signatures to verify/authenticate the received packets, reassemble message and defend against pollution-based DoS attacks [9]. Note that, we do not differentiate between packet jamming and packet collisions as they both cause interference to the legitimate packets, and packet coding can be used to recover bit errors in received packets.

After the receiver chooses a strategy s_r, a reward on channel f is revealed to the receiver if and only if f is chosen as a receiving channel. There are four possible cases:
Case 1: No packet is received on f. This is because f has not been selected by the sender for transmission. In this case, *reward* 0 is obtained.
Case 2: A packet is received on f. If the received packet fails to pass the verification, *reward* 0 is obtained.
Case 3: A packet is received on f. Jammed or collided packets that cannot be recovered will be discarded, resulting in 0 reward.
Case 4: A packet is received on f. If no jamming is detected or corrupted packets due to jamming can be recovered via packet coding, a *reward* 1 is obtained.

Similarly, after the sender chooses a strategy s_s, a reward on channel f is revealed to the sender if and only if f is chosen as a sending channel. A reward 1 is obtained if an ACK is received on f, otherwise the reward is 0.

In this paper, we formally formulate the jamming-resistant spectrum sensing and access problem as a non-stochastic MAB problem (NS-MAB) [21]–[23], where each channel can be considered as an arm of an multi-arm bandit. Due to the jamming effect and dynamics behaviors of PUs, each channel f is then associated with an arbitrary and unknown sequence of *rewards*, which can be obtained on a channel if the sender and the receiver choose f for sending and receiving simultaneously.

For ease of analysis and presentation, we first define some important notation. In each timeslot $t \in \{1,\ldots,T\}$, the sender (receiver) independently selects a strategy I_t from his strategy sets. We write $f \in i$ if channel f is chosen in strategy i, *i.e.*, the value of the fth entry of i is 1. Note that a strategy is a vector of dimension n, I_t denotes a particular strategy chosen for timeslot t, and i denotes a general strategy in the strategy set. The total rewards of a strategy i during timeslot t is $g_{i,t} = \sum_{f \in i} g_{f,t}$, and the cumulative rewards up to timeslot t of each strategy i is $G_{i,t} = \sum_{s=1}^{t} g_{i,s} = \sum_{f \in i} \sum_{s=1}^{t} g_{f,s}$. The total rewards over all chosen strategies up to timeslot t is thus $\widehat{G}_t = \sum_{s=1}^{t} g_{I_s,s} = \sum_{s=1}^{t} \sum_{f \in I_s} g_{f,s}$, where I_s is chosen randomly according to certain distribution over the strategy set. To quantify the performance, we use the following metric called *regret*:

$$\max_{i \in S_x} G_{i,T} - \widehat{G}_T^x, \quad x \in \{s,r\},$$

where the superscript is used to differentiate the sender from the receiver, and the maximum is taken over all strategies available to the sender or the receiver. The *regret* is defined as the accumulated rewards (or successfully received packets) *difference* over T timeslots between the proposed strategy and the optimal *static* one. The static optimal strategy denotes the best fixed solution (*i.e.*, the best set of channels that if keeping to use them largest rewards will be generated.) for message reception in the presence of jamming. Note that the sender and the receiver will adaptively choose their own strategies in each timeslot based on the updated probability distributions over the strategy set. As for the sender (receiver), the updates of the probability distribution are determined by the outcomes of joint actions of PU, the jammer and the receiver (sender). Thus, the accumulated rewards of the sender (receiver) along the time depend on the actions of the other three parties in each timeslot.

4.3 Our Construction

In this subsection, we present our jamming-resistant spectrum sensing and access protocol. Our algorithm is a probabilistic one that can accommodate the

Algorithm 1 A Jamming-resistant Multi-radio Multi-channel Spectrum Sensing and Access Protocol.

Input: n, k_r, k_s, T, $\varepsilon \in (0,1]$, $\delta \in (0,1)$, $\beta^s, \beta^r \in (0,1]$, $\gamma^s, \gamma^r \in (0, 1/2]$, $\eta^s, \eta^r > 0$.
Initialization: Initialize all system parameters, setting the channel weight $w^x_{f,0} = 1\ \forall f \in [1,n]$, the strategy weight $w^x_{i,0} = 1\ \forall i \in [1, N^x]$, and the total strategy weight $W^x_0 = N^x = \binom{n}{k_s}$, where $x = s, t$.
For timeslot $t = 1, 2, \ldots, T$
1: Select a strategy I^x_t according to $p^x_{i,t}$ ($\forall i \in [1, N^x]$), with $p^x_{i,t}$ computed following Eq. (5).
2: Compute channel selection probability $q^x_{f,t}$ ($\forall f \in [1,n]$) as $q^x_{f,t} = \sum_{i:f\in i} p^x_{i,t}$.
3: Transmit a packet if and only if the channel is sensed to be idle.
4: Perform verification and jamming detection once a packet is received on channel f. Transmit back an acknowledgement on f if the received packet passes the check.
5: Compute rewards $g^x_{f,t}$ ($\forall f \in I^x_t$) and virtual rewards $g^{x'}_{f,t}$ with the revealed $g_{f,t}$ ($\forall f \in [1,n]$), following Eqs. (3) and (4).
6: Update channel weight $w^x_{f,t}$ and strategy weight $w^x_{i,t}$ following Eqs. (1) and (2), respectively. Update the total strategy weight as $W^x_t = \sum_{i=1}^{N^x} w^x_{i,t}$.
End

changes of channel status caused by a (potentially) malicious jammer. The dynamic property of the proposed solution lies in the trade-off between exploration action and exploitation action, which will both affect the system performance.

As shown in Algorithm 1, the algorithm comprises two subalgorithms: $\mathcal{A}^s$ at the sender side and $\mathcal{A}^r$ at the receiver side. In Algorithm 1, the system parameters β, γ and η are determined by the *regret bound*, and the derivation of them will be shown in proof of Theorem 1.

Let N^x ($x \in \{s,r\}$) denote the total number of strategies. Each strategy is assigned a strategy weight, and each channel is assigned a channel weight. During each timeslot, the channel weight is dynamically adjusted based on the virtual channel rewards revealed to the sender and the receiver:

$$w^x_{f,t} = w^x_{f,t-1} e^{\eta^x g^{x'}_{f,t}}, \quad x \in \{s,r\}. \tag{1}$$

We use exponentially weighted forecasters which follow the Exp3 ("Exponential-weight algorithm for Exploration and Exploitation") first proposed in [21]. In a multi-armed bandit setting, at time t, an expert is chosen with probability that increases with the past performance of the expert. In practice, the most popular choice of such kind of function is exponential function. It is easy to see that the increase of the virtual channel rewards leads to larger channel weights.

A strategy indicates the choices of channels for use, so we define the weight of a strategy as the product of the weights of all channels:

$$w^x_{i,t} = \Pi_{f\in i} w^x_{f,t} = w^x_{i,t-1} e^{\eta^s g^{x'}_{i,t}}, \quad x \in \{s,r\}. \tag{2}$$

where $g^{x'}_{i,t} = \sum_{f\in i} g^{x'}_{f,t}$.

Here, the reason to estimate reward for each channel first instead of estimating rewards for each strategy directly is that the reward of each channel can provide useful information about the other unchosen strategies containing the same channels. The parameter β is used to control the bias in estimating the channel reward $g^{s'}_{f,t}$ and $g^{r'}_{f,t}$, which are computed as:

$$g^{s'}_{f,t} = \begin{cases} \frac{g^s_{f,t}+\beta^s}{\varepsilon q^s_{f,t}} R_t & \text{if } f \in I^s_t, \\ \frac{\beta^s}{\varepsilon q^s_{f,t}} R_t & \text{otherwise,} \end{cases} \tag{3}$$

$$g^{r'}_{f,t} = \begin{cases} \frac{g^r_{f,t}+\beta^r}{q^r_{f,t}} & \text{if } f \in I^r_t, \\ \frac{\beta^r}{q^r_{f,t}} & \text{otherwise,} \end{cases} \tag{4}$$

where $q^x_{f,t}$ ($x \in \{s,t\}$) denotes the channel f's probability distribution, and R_t is a random variable under Bernoulli distribution satisfying $\mathbf{P}\{R_t = 1\} = \varepsilon$. The parameter β is a fixed value that will be determined before the execution of the protocol (see the proof of Theorem 1). Based on the true rewards revealed to the sender and the receiver, we define the virtual rewards to increase weight of "good" channels, *i.e.*, increase the access probabilities of "good" channels which have been less often sensed.

In Algorithm 1, at the beginning of each timeslot, the transceiver chooses a strategy based on the probability distribution $p^x_{i,t}$ ($x \in \{s,t\}$) as:

$$p^x_{i,t} = \begin{cases} (1-\gamma^x)\frac{w^x_{i,t-1}}{W^x_{t-1}} + \frac{\gamma^x}{|\mathcal{C}^x|} & i \in \mathcal{C}^x \\ (1-\gamma^x)\frac{w^x_{i,t-1}}{W^x_{t-1}} & \text{otherwise,} \end{cases} \tag{5}$$

where $x \in \{s,r\}$. The parameter γ^x is used to balance between $\frac{w^x_{i,t-1}}{W^x_{t-1}}$ and $\frac{1}{|\mathcal{C}|^x}$.

In the calculation of the strategy probability distribution, the first part is a distribution which assigns to each action a probability mass exponential in the estimated cumulative reward for that action, and the second part is the uniform distribution. If not mixed with the uniform distribution, the algorithm might have large deviations with high probability, *i.e.*, from time to time it may concentrate on the wrong strategy for too long and then incur a large regret. So the mixing is done to make sure that the algorithm tries out all strategies and gets good estimates of the gains for each channel [21]. Note γ^x is a fixed value that will be determined before the execution of the protocol (see the proof of Theorem 1). The *covering strategy set* $\mathcal{C}^x$ is defined to ensure that each channel/frequency is sampled sufficiently often. The covering set has the property that for each channel f, there is a strategy i in the covering set such that $f \in i$. Based on the definition of strategy, each strategy includes k_x ($x \in \{s,r\}$) "active" channels. Thus, we can construct one typical and simple covering set with size $|\mathcal{C}^x| = \lceil \frac{n}{k_s} \rceil$ ($x \in \{s,r\}$).

Discussions. In practice, the transceiver (*i.e.*, the sender and the receiver) may not have the same sensing outcomes due to sensing errors. So, in our design we let the sender perform sensing in each timeslot, and the receiver only selects channels to listen on. Note that the operating point of the spectrum sensor is set as the probability of the collision with PUs [2], which includes two types of sensing errors: *false alarm* probability and *miss detection* probability. Without loss of generality, we use τ to denote the *sensing error probability* in the following analysis, where $\tau = \mathbb{P}\{false\ alarm\}(1 - \mathbb{P}\{PU\ active\}) + \mathbb{P}\{miss\ detection\}\mathbb{P}\{PU\ active\}$.

To eliminate the information asymmetry between the sender and the receiver, the sender uses the acknowledge information to update the probability distribution over strategy set. Thus, the accumulated rewards for the sender and the receiver are equivalent, *i.e.*, $\widehat{G}_t^s = \widehat{G}_t^r$ (Note that, we make this assumption to obtain the upperbound performance of the proposed anti-jamming scheme. In Section 6, we evaluate the case where ACKs are randomly jammed by the attacker, showing the strong resilience of our proposed scheme). In addition, because the sender and the receiver are not perfectly synchronized, it is necessary and important to evaluate how close the sender's and the receiver's strategies are as time goes. Since the closer the transceivers' chosen strategies, the more rewards generated. This is equivalent to saying that how well the learning based algorithm proceeds to maximize the system throughput.

The spectrum sensing usually consumes more energy compared to reception, *i.e.*, it is costly to obtain the sensing results [24]. In certain application scenarios, legitimate nodes may only have a limited number of sensing times due to energy constraint. Let ε denote the proportion of timeslots when sensing is performed. For T timeslots, the number of sensing times is approximately $T\varepsilon$. In Algorithm 1, we introduce a Bernoulli random variable with $\mathbf{P}\{R_t = 1\} = \varepsilon$ at the sender side. Thus, the sender senses the channel with probability ε. There are two possible cases when the sender does not perform sensing in a timeslot. In the first case, the sender remains silent in this timeslot without transmitting any packets. Due to the random sensing and access strategy, it is hard for the adversary to predict the behaviors of the sender. However, as no packets are transmitted, the transmission delay may be increased. In the second case, the sender still accesses the most possibly free channels based on the previous probability distribution. In this case, there is a tradeoff between the collision probability with PUs and the number of sensing times.

4.4 Theoretical Analysis

In this subsection, we analyze the performance of Algorithm 1 in terms of both optimality and efficiency. For ease of analysis, we first give the following two important definitions.

Definition 1: An algorithm $\mathcal{A}$ is α-*static* (or α-*adaptive*) approximation of the *static* (or *adaptive*) optimal solution if and only if it can transmit the message successfully in time αT with high probability (w.h.p) $1-\frac{1}{l^\epsilon}$ when the *static* (or *adaptive*) optimal solution can transmit the same message successfully with the same probability $1-\frac{1}{l^\epsilon}$ in time T, where $\epsilon > 0$ is a constant and l is the number of packets in the message.

Definition 2: The *regret* of an algorithm $\mathcal{A}$ is the difference between the accumulated rewards using the *static* optimal strategy and that using $\mathcal{A}$ over T timeslots, *i.e.*, $G_T^{max} - \widehat{G}_T^{\mathcal{A}}$, where $G_T^{max} = \max_{i\in S} G_{i,T} = \max_{i\in S}\sum_{f\in i}\sum_{s=1}^{T} g_{f,s}$ and $\widehat{G}_T^{\mathcal{A}} = \sum_{s=1}^{T} g_{I_s,s} = \sum_{s=1}^{T}\sum_{f\in I_s} g_{f,s}$.

The first definition is used to characterize the approximation ratio between the proposed algorithm and the static and adaptive optimal solutions. The second definition is used to characterize the throughput performance between the proposed algorithm and the optimal solution. In the following analysis, we will write G^{max} instead of G_T^{max} whenever the value of T is clear from the context. In addition, we will write $G_T^{max}(s)$ and $G_T^{max}(r)$ to denote the rewards of the *static* optimal strategies for the sender and the receiver, respectively.

Due to the probabilistic strategy selection, the sender and the receiver are not perfectly synchronized in each timeslot. However, we show that the sender's sensing strategy and the receiver's receiving strategy will converge to their own optimal strategies. The following theorem measures how close their optimal strategies are as $T \to \infty$.

Theorem *1:* The normalized reward distance $\frac{1}{T}|G_T^{max}(s) - G_T^{max}(r)|$ converges to 0 at rate $O(1/\sqrt{T})$ as $T \to \infty$.

Proof: We first prove that at the receiver side, with probability at least $1-\delta$, the *regret* $G_T^{max}(r) - \widehat{G}_T^{\mathcal{A}^r}$ is at most $6k_r\sqrt{Tn\ln n}$, while $\beta^r = \sqrt{\frac{k_r}{nT}\ln\frac{n}{\delta}}$, $\gamma^r = 2\eta^r n$ and $\eta^r = \sqrt{\frac{\ln n}{4Tn}}$ and $T \geq \max\{\frac{k_r}{n}\ln\frac{n}{\delta}, 4n\ln n\}$.

We use a superscript for η, γ, β to differentiate between the sender and the receiver. However, for ease of exposition, we do not differentiate between the other notation since they are independent in the proofs for the sender and the receiver. Now we introduce some notation for performance analysis: $G_{i,T} = \sum_{t=1}^{T} g_{i,t}$ and $G'_{i,T} = \sum_{t=1}^{T} g'_{i,t}$ for all $1 \leq i \leq N$, where $G_{i,T}$ ($G'_{i,T}$) denotes the total gain (virtual gain, respectively) of strategy i in T timeslots, and $G_{f,T} = \sum_{t=1}^{T} g_{f,t}$ and $G'_{f,T} = \sum_{t=1}^{T} g'_{f,t}$ for all $1 \leq f \leq n$, where $G_{f,T}$ ($G'_{f,T}$) denotes the total gain (virtual gain, respectively) on channel f in T timeslots. The relationship between gain and virtual gain is derived as follows.

The proof is applicable for any channel f. $\forall u > 0$ and $c > 0$, by using the bound of Chernoff, we get $\mathbb{P}[G_{f,T} > G'_{f,T} + u] \leq e^{-cu}\mathbb{E}[e^{c(G_{f,T}-G'_{f,T})}]$. Let $u = \ln\frac{n}{\delta}/\beta$ and $c = \beta$, we obtain $e^{-cu}\mathbb{E}[e^{c(G_{f,T}-G'_{f,T})}] = \frac{\delta}{n}\mathbb{E}[e^{\beta(G_{f,T}-G'_{f,T})}]$. Thus, we can show that $e^{\beta(G_{f,T}-G'_{f,T})} \leq 1$ for all T. Let $Z_t = e^{\beta(G_{f,t}-G'_{f,t})}$. By showing $\mathbb{E}[Z_t] \leq Z_{t-1}$ for all $t \geq 2$ and $\mathbb{E}[Z_1] \leq 1$, it is easy to prove $\forall\delta \in (0,1)$, $0 \leq \beta < 1$ and $1 \leq f \leq n$,

$$\mathbb{P}[G_{f,T} > G'_{f,T} + \frac{1}{\beta}\ln\frac{n}{\delta}] \leq \frac{\delta}{n} \tag{6}$$

Next we show the regret bound by using $\ln\frac{W_T}{W_0}$. First, we directly can obtain the lower bound by the definition $\ln\frac{W_T}{W_0} = \ln\sum_{i=1}^{N} e^{\eta^r G'_{i,T}} - \ln N \geq \eta^r \max_{1\leq i\leq N} G'_{i,T} - \ln N$. Then we derive the upper bound as follows: $\eta^r g'_{i,t} = \eta^r \sum_{f\in i} g'_{f,t} \leq \eta^r \sum_{f\in i}\frac{1+\beta^r}{q_{f,t}} \leq \frac{\eta^r k_r(1+\beta^r)|\mathcal{C}|}{\gamma^r} \leq 1$, where the second inequality term holds due to $q_{f,t} \geq \frac{\gamma^r}{|\mathcal{C}|}$ $(\forall f)$ by the definition.

Using the fact that $e^x \leq 1 + x + x^2$ for all $x \leq 1$, for all $t = 1, 2, \cdots, T$ we have $\ln\frac{W_t}{W_{t-1}} = \ln\sum_{i=1}^{N}\frac{w_{i,t-1}}{W_{t-1}}e^{\eta^r g'_{i,t}} \leq \ln(\sum_{i=1}^{N}\frac{w_{i,t-1}}{W_{t-1}}(1 + \eta^r g'_{i,t} + (\eta^r)^2 g'^2_{i,t})) \leq \ln(1 + \sum_{i=1}^{N}\frac{p_{i,t}}{1-\gamma^r}(\eta^r g'_{i,t} + (\eta^r)^2 g'^2_{i,t})) \leq \frac{\eta^r}{1-\gamma^r}\sum_{i=1}^{N} p_{i,t}g'_{i,t} + \frac{(\eta^r)^2}{1-\gamma^r}\sum_{i=1}^{N} p_{i,t}g'^2_{i,t}$. We derive the inequalities using the following two common facts: $\sum_{i=1}^{N} p_{i,t} \leq 1 - \gamma^r$ and inequality $\ln(1+x) \leq x$ for all $x > -1$.

Let $\mathcal{N}$ denote the strategy set $\{1, \ldots, N\}$. On the one hand, we have $\sum_{i=1}^{N} p_{i,t}g'_{i,t} = \sum_{i=1}^{N} p_{i,t}\sum_{f\in i} g'_{f,t} = \sum_{f=1}^{n} g'_{f,t}\sum_{i\in\mathcal{N}:f\in i} p_{i,t} = \sum_{f=1}^{n} g'_{f,t}q_{f,t} = g_{I_t,t} + n\beta^r$. On the other hand, $\sum_{i=1}^{N} p_{i,t}g'^2_{i,t} = \sum_{i=1}^{N} p_{i,t}(\sum_{f\in i} g'_{f,t})^2 \leq \sum_{i=1}^{N} p_{i,t}k_r\sum_{f\in i} g'^2_{f,t} = k_r\sum_{f=1}^{n} g'^2_{f,t}\sum_{i\in\mathcal{N}:f\in i} p_{i,t} = k_r\sum_{f=1}^{n} g'^2_{f,t}q_{f,t} \leq k_r(1+\beta^r)\sum_{f=1}^{n} g'_{f,t}$, which holds the fact that $g'_{f,t} \leq \frac{1+\beta^r}{q_{f,t}}$. Note that for clearly differentiating between the *regret* bounds for the sender and the receiver, in the derivation we loose the bounds a little bit by choosing k_r instead of $\min\{k_r, k_s\varepsilon(1-\tau), n - k_j\}$. Therefore, $\ln\frac{W_t}{W_{t-1}} \leq \frac{\eta^r}{1-\gamma^r}(g_{I_t,t} + n\beta^r) + \frac{(\eta^r)^2 k_r(1+\beta^r)}{1-\gamma^r}\sum_{f=1}^{n} g'_{f,t}$.

Summing for $t = 1, \cdots, T$, we have the following inequality $\ln\frac{W_T}{W_0} \leq \frac{\eta^r}{1-\gamma^r}(\widehat{G}_T + n\beta^r T) + \frac{(\eta^r)^2 k_r(1+\beta^r)}{1-\gamma^r}\sum_{f=1}^{n} G'_{f,T} \leq \frac{\eta^r}{1-\gamma^r}(\widehat{G}_T + n\beta^r T) + \frac{(\eta^r)^2 k_r(1+\beta^r)}{1-\gamma^r}|\mathcal{C}|\max_{1\leq i\leq N} G'_{i,T}$. Here, $\widehat{G}_T$ denotes the expected total gain of the proposed algorithm in T timeslots. By using the upper and the lower bounds together, we get $\widehat{G}_T \geq (1 - \gamma^r - \eta^r k_r(1 + \beta^r)|\mathcal{C}|)\max_{1\leq i\leq N} G'_{i,T} - \frac{1-\gamma^r}{\eta^r}\ln N - n\beta^r T$.

Applying Eq. (6), we can show with probability at least $1 - \delta$, $\widehat{G}_T \geq (1 - \gamma^r - \eta^r k_r(1 + \beta^r)|\mathcal{C}|)(\max_{1\leq i\leq N} G_{i,T} - \frac{k_r}{\beta^r}\ln\frac{n}{\delta}) - \frac{1-\gamma^r}{\eta^r}\ln N - n\beta^r T$. Here, we used the fact $1 - \gamma^r - \eta^r k_r(1+\beta^r)|\mathcal{C}| > 0$ which follows the assumptions of the theorem.

By doing some transpositions and using the following fact $\max_{1\leq i\leq N} G_{i,T} \leq Tk_r$, we have $\max_{1\leq i\leq N} G_{i,T} - \widehat{G}_T \leq (\gamma^r + \eta^r(1+\beta^r)k_r|\mathcal{C}|)Tk_r + (1 - \gamma^r - \eta^r(1+\beta^r)k_r|\mathcal{C}|)\frac{k_r}{\beta^r}\ln\frac{n}{\delta} + + \frac{1-\gamma^r}{\eta^r}\ln N + n\beta^r T$ with probability at least $1 - \delta$. Let $K = \min\{k_s, n - k_j, k_r\}$. Since $\hat{G}_T = KT - \hat{L_T}$ and $\max_{1\leq i\leq N} G_{i,T} = KT - \min_{1\leq i\leq N} L_{i,T}$, we have $\hat{L_T} \leq KT(\gamma^r + \eta^r(1 + \beta^r)k_r|\mathcal{C}|) + (1 - \gamma^r - \eta^r(1+\beta^r)k_r|\mathcal{C}|)\min_{1\leq i\leq N} L_{i,T} + (1 - \gamma^r - \eta^r(1+\beta^r)k_r|\mathcal{C}|)\frac{k_r}{\beta^r}\ln\frac{n}{\delta} + \frac{1-\gamma^r}{\eta^r}\ln N + n\beta^r T$ with probability $1 - \delta$. By simplifying the inequality, we have $\hat{L_T} - \min_{1\leq i\leq N} L_{i,T} \leq k_r T\gamma^r + 2\eta^r T k_r n + \frac{k_r}{\beta^r}\ln\frac{n}{\delta} + \frac{1-\gamma^r}{\eta^r}k_r\ln n + n\beta^r T$ with probability $1 - \delta$

Setting $\beta^r = \sqrt{\frac{k_r}{nT}\ln\frac{n}{\delta}}$ and $\gamma^r = 2\eta^r k_r|\mathcal{C}|$, we can get $G_T^{max}(r) - \widehat{G}_T^{\mathcal{A}^r} = \max_{1\leq i\leq N} G_{i,T} - \widehat{G}_T \leq 4\eta^r T{k_r}^2|\mathcal{C}| + \frac{\ln N}{\eta^r} + 2\sqrt{k_r nT\ln\frac{n}{\delta}}$ which holds with probability $1 - \delta$ if $T \geq \frac{k_r}{n}\ln(\frac{n}{\delta})$. Finally, by using the facts $|\mathcal{C}| = \lceil\frac{n}{k_r}\rceil$ and $N \leq n^{k_r}$, and setting $\eta^r = \sqrt{\frac{\ln N}{4k_r^2 T|\mathcal{C}|}}$, we can prove $\max_{1\leq i\leq N} G_{i,T} - \widehat{G}_T \leq 6k_r\sqrt{Tn\ln n}$ by properly choosing δ.

Similarly, at the sender side we first show the connection between the true and the estimated cumulative rewards. The only difference is that the computation of estimated channel rewards is involved with a random variable ε. We prove that with probability at least $1 - \delta$, the *regret* $G_T^{max}(s) - \widehat{G}_T^{\mathcal{A}^s}$ is bounded by $14{k_s}^2\sqrt{\frac{Tn\ln n}{\varepsilon}}$ by property choosing δ, $\beta^s = \sqrt{\frac{k_s}{nT\varepsilon}\ln\frac{2n}{\delta}}$, $\gamma^s = \frac{2\eta^s n}{\varepsilon}$ and $\eta^s = \sqrt{\frac{\varepsilon\ln n}{4Tn}}$ and $T \geq \max\{\frac{{k_s}\ln^2\frac{2n}{\delta}}{\varepsilon n\ln n}, \frac{n\ln\frac{2n}{\delta}}{k_s}, 4n\ln n\}$. To clearly differentiate the *regret* bounds for the sender and the receiver, we loosen the bounds a little bit by choosing k_r and k_s instead of $\min\{k_r, k_s\varepsilon(1-\tau), n - k_j\}$. Hence, the sensing error probability τ does not appear in the final expression of performance bound.

Finally, as $\widehat{G}_T^{\mathcal{A}^s} = \widehat{G}_T^{\mathcal{A}^r}$, $|G_T^{max}(s) - G_T^{max}(r)| \leq 6k_r\sqrt{Tn\ln n} + 14{k_s}^2\sqrt{\frac{Tn\ln n}{\varepsilon}} \leq \frac{20k}{\sqrt{\varepsilon}}\sqrt{Tn\ln n}$, where $k = \max\{{k_s}^2, k_r\}$. Thus, $\frac{1}{T}|G_T^{max}(s) - G_T^{max}(r)| \to 0$ at rate $O(1/\sqrt{T})$ as $T \to \infty$. □

Theorem *2:* Algorithm 1 has time complexity $O(k_x nT)$ and space complexity $O(k_x n)$, where $x \in \{s, r\}$.

Proof: In the proposed algorithm, the computation of probability distributions of strategy and channel are the most time-consuming steps due to the exponential number of possible strategies. In this proof, we show that the time complexity can be reduced by using dynamic programming. We use $S(\bar{f}, \bar{k})$ to denote the strategy set of which each strategy selects $\bar{k}$ channels from $\bar{f}, \bar{f}+1, \cdots, n$. Also, we use $\bar{S}(\bar{f}, \bar{k})$ to denote the strategy set of which each strategy selects $\bar{k}$ channels from $1, 2, \cdots, \bar{f}$. We define $W_t(\bar{f}, \bar{k}) = \sum_{i\in S(\bar{f},\bar{k})}\prod_{f\in i} w_{f,t}$ and $\bar{W}_t(\bar{f}, \bar{k}) = \sum_{i\in\bar{S}(\bar{f},\bar{k})}\prod_{f\in i} w_{f,t}$. Note $W_t(\bar{f}, \bar{k}) = W_t(\bar{f}+1, \bar{k}) + w_{\bar{f},t}W_t(\bar{f}+1, \bar{k}-1)$ and $\bar{W}_t(\bar{f}, \bar{k}) = \bar{W}_t(\bar{f}-1, \bar{k}) + w_{\bar{f},t}W_t(\bar{f}-1, \bar{k}-1)$, which implies both $W_t(\bar{f}, \bar{k})$

and $\bar{W}_t(\bar{f},\bar{k})$ can be calculated in $O(k_x n)$ (letting $W_t(\bar{f},0)=1$, $W(n+1,\bar{k})=\bar{W}(0,\bar{k})=0$) by using dynamic programming approach $\forall 1 \le \bar{f} \le n$ and $1 \le \bar{k} \le k_x$.

In step 1, a strategy should be drawn from $\binom{n}{k_x}$ strategies. Instead of drawing a strategy, we choose channel for the strategy one by one. Here, we choose channels one by one in the increasing order of channel indices, *i.e.*, we determine whether the channel 1 should be selected, and the channel 2, and so on. $\forall f$, if $k \le k_x$ channels have already been selected from channels $1,\cdots,f-1$, we select a channel f with probability $\frac{w_{f,t-1}W_{t-1}(f+1,k_x-k-1)}{W_{t-1}(f,k_x-k)}$ and not select f with probability $\frac{W_{t-1}(f+1,k_x-k)}{W_{t-1}(f,k_x-k)}$. Let $w(f)=w_{f,t-1}$ if channel $f \in i$; $w(f)=0$ otherwise. Obviously, $w(f)$ is actually the weight of f in the strategy weight. In our algorithm, $w_{i,t-1}=\prod_{f=1}^{n} w(f)$. Let $c(f)=1$ if f is selected in i; $c(f)=0$ otherwise. $\sum_{f=1}^{\bar{f}} c(f)$ denotes the total number of channels selected from channels $1,2,\cdots,\bar{f}$ in i. By this implementation, the probability of choosing i is $\prod_{\bar{f}=1}^{n} \frac{w(\bar{f})W_{t-1}(\bar{f}+1,k_x-\sum_{f=1}^{\bar{f}} c(f))}{W_{t-1}(\bar{f},k_x-\sum_{f=1}^{\bar{f}-1} c(f))} = \frac{\prod_{\bar{f}=1}^{n} w(\bar{f})}{W_{t-1}(1,k_x)} = \frac{w_{i,t-1}}{W_{t-1}}$. This probability is equivalent to that in Algorithm 1, which implies the implementation is correct.

Because we do not maintain $w_{i,t}$, it is impossible to compute $q_{f,t}$ as we described in Algorithm 1. Then, $q_{f,t}$ can be calculated in $O(n)$ as $q_{f,t}=(1-\gamma)\frac{\sum_{k=0}^{k_x-1} W_{t-1}(f-1,k)w_{f,t-1}W_{t-1}(f+1,k_x-k-1)}{W_{t-1}(1,k_x)} + \gamma\frac{|\{i\in\mathcal{C};f\in i\}|}{|\mathcal{C}|}$ for each round. □

In practice, the transmitted messages, which may have much larger size than the length of timeslots, have to be split into small fragments to fit the timeslots. As shown above, the proposed jamming-resistant spectrum sensing and access protocol is probabilistic in nature, so we cannot guarantee the transmitted message is delivered in certain number of timeslots with probability one. So, to evaluate the transmission efficiency, we consider the expected time for a message delivery with *high* probability, which implies the probability goes to one when the total number of packets goes to infinity. Based on the acknowledgement information, in each timeslot the sender will pick up a packet that has not been delivered. Without loss of generality, assuming a message M is partitioned into l packets $M_1, M_2, \cdots, M_l$, each of which has size $|M_i| = |M|/l$ $(1 \le i \le l)$. Then, the transmitted message M can be reconstructed at the receiver if and only if all l packets are successfully received. The following theorems characterize the approximation factors for the static optimal and adaptive optimal solutions.

Theorem *3:* When $l \ge 36(1+c\epsilon)k_r n \ln n/(c-1)^2\epsilon^2$, our algorithm is $(1+c\epsilon)$-static approximation for any constant $c>1$.

Proof: See [1]. □

Theorem *4:* When $l \ge 36\frac{n^3 \ln nK(1+c\epsilon)}{k_s\varepsilon(1-\tau)(n-k_j)(c-1)^2\epsilon^2}$, our algorithm is $\frac{n^2}{k_r k_s \varepsilon(1-\tau)(n-k_j)}K(1+c\epsilon)$-adaptive approximation for any constant $c>1$, where $K=\min\{k_r, k_s\varepsilon(1-\tau), n-k_j\}$, ε is the probability of sensing a channel and τ is the sensing error probability.

Proof: See [1]. □

Discussions. As can be seen in the proof of Theorem 1, the parameters β, η and γ are fixed values and they are all pre-computed before the protocol execution. If we aim at ensuring that with probability at least $1-\delta$ the regret bound can be achieved, we can set a preferable value for δ. The parameter selection process is as follows. We have $\beta^x=\sqrt{\frac{k_x}{nT}\ln\frac{n}{\delta}}$, $\gamma^x=2\eta^x n$ and $\eta^x=\sqrt{\frac{\ln n}{4Tn}}$. Here, n and k_r are pre-selected system parameters. Once T is obtained, the specific values of β^x, η^x and γ^x can be determined such that the regret bound holds (or asymptotic optimality is achieved). To determine T, in our protocol design, we let the sender determine a feasible T and encode it in each packet for transmission. The receiver obtains T by successfully decoding any received packet and begins to run the algorithm. Assume p is the probability of message delivery, the sender determines T by first estimating a lower bound $\underline{k_r}$ of k_r and an upper bound $\overline{k_j}$ of k_j. It then calculates ϵ such that $1-\frac{1}{l^\epsilon}=p$ and determines the constant $c>1$ such that $l=36(1+c\epsilon)\underline{k_r}n\ln n/(c-1)^2\epsilon^2$. Finally, the expected time for message delivery is $T=(1+c\epsilon)l/(\underline{k_r}\frac{k_s\varepsilon(1-\tau)}{n}\frac{n-\overline{k_j}}{n})$. By theorem 3, with probability at least p the message M can be successfully recovered at the receiver.

5 Tracking the Adaptive Compound Strategy for Anti-jamming Spectrum Access

In the above discussions, *regret* is computed as the accumulated reward difference between the proposed anti-jamming strategy and the *static* optimal strategy. We have shown that the proposed jamming-resistant in Algorithm 1 can track the *static* optimal strategy and converge to it as time goes. According to the definition, the *static* optimal strategy is selected as the fixed "best" strategy used for all timeslots. However, for each timeslot there always exists the best strategy against the "joint" strategy of the other parties involved in the anti-jamming game. Linking these strategies from all timeslots together, the best compound strategy is formulated, and this is so-called the *adaptive* optimal strategy. So, an interesting question can be raised here: *at each timeslot, the good strategy may change, is it possible to select a sequence of strategies to approximate the adaptive/compound strategy*?

5.1 The Proposed Construction

In Algorithm 1, the size of the static optimal strategy set is $\binom{n}{k_r}$ ($\binom{n}{k_s}$). However, for all possible compound

strategies, the strategy set is extremely large, *i.e.*, with size approximately as large as $\binom{n}{k_r}^T$ ($\binom{n}{k_s}^T$). Therefore, by using the previous protocol it is computationally expensive to track the best compound (adaptive) strategy. In this section, we will consider an extension of the anti-jamming protocol using the tracking the best expert problem and develop an efficient algorithm to approximate the best compound strategy.

Different from the static optimal strategy, the best compound strategy is allowed to change its strategy m times in T timeslots, *i.e.*, a strategy from $\binom{n}{k_r}$ ($\binom{n}{k_s}$) is assigned in a timeslot. Consider the compound strategy $\mathbf{i} = (i_1, i_2, \ldots, i_m)$ corresponding to the timeslot vector $\mathbf{t} = (t_1, t_2, \ldots, t_m)$, strategy i_j is used to predict the best strategy at time instant $t_j \leq t \leq t_{j+1}$. Then the new *regret* is defined as

$$\max_{(\mathbf{i},\mathbf{t})} G_{i,T} - \widehat{G}_T^x, \quad x \in \{s, r\}, \tag{7}$$

where $\max_{(\mathbf{i},\mathbf{t})} G_{i,T}$ denotes the accumulated rewards obtained by using the adaptive compound strategy with respect to $(\mathbf{i}, \mathbf{t})$. For ease of analysis, we assume $\varepsilon = 1$, *i.e.*, the sender performs sensing in each timeslot. The new algorithm for tracking the adaptive compound strategy differs from Algorithm 1 in step 6. For ease of notation, we eliminate the superscript x in the following expressions. In step 6, the sender and the receiver both update the strategy weight as

$$\begin{aligned} v_{i,t} &= w_{i,t-1} e^{\eta g'_{i,t}}, \\ w_{i,t} &= (1-\alpha) v_{i,t} + \frac{\alpha}{N} W_t, \end{aligned}$$

where $N = \binom{n}{k_s}$ ($N = \binom{n}{k_r}$), $g'_{i,t} = \sum_{f \in i} g'_{f,t}$ and W_t is the sum of the total weights, *i.e.*,

$$W_t = \sum_{i=1}^{N} v_{i,t}.$$

5.2 A Fast Implementation of The Proposed Construction

As can be seen, the time complexity of the proposed construction for tracking the adaptive compound strategy is $O(n^{k_s} T)$. In this section, we present an alternative method of implementing the above algorithm in $O(k_s n T^2)$ time.

The basic idea of our fast implementation is to select channels one by one in each timeslot/round, instead of computing each strategy from a large strategy set. We let $S(\overline{f}, \overline{k})$ denote the strategy set in which each strategy chooses $\overline{k}$ channels from channels $\overline{f}, \overline{f}+1, \ldots, n$ and $\overline{S}(\overline{f}, \overline{k})$ denote the strategy set in which each strategy chooses $\overline{f}$ channels from channel $1, 2, \ldots, \overline{f}$. In addition, we let $G'_{t',t-1}(f)$ denote the sum of cumulative gains in the interval $[t', t-1]$ of channel f, $G'([t', t-1], i)$ denote the sum of cumulative gains in the interval $[t', t-1]$ of strategy i, $M_{t',t-1}(\overline{f}, \overline{k})$ denote the sum of exponential cumulative gains in the interval $[t', t-1]$ of all the strategies in $S(\overline{f}, \overline{k})$. Formally, we have

$$M_{t',t-1}(\overline{f}, \overline{k}) = \sum_{i \in S(\overline{f}, \overline{k})} e^{\eta \sum_{f \in i} G'_{t',t-1}(f)},$$

where $G'_{t',t-1}(f) = \sum_{j=t'}^{t-1} g'_{f,j}$.

Similarly, we define

$$\overline{M}_{t',t-1}(\overline{f}, \overline{k}) = \sum_{i \in \overline{S}(\overline{f}, \overline{k})} e^{\eta \sum_{f \in i} G'_{t',t-1}(f)}.$$

Correspondingly, we have the following properties

$$\begin{aligned} M_{t',t-1}(\overline{f}, \overline{k}) &= M_{t',t-1}(\overline{f}+1, \overline{k}) \\ &\quad + e^{\eta G'_{t',t-1}(f)} M_{t',t-1}(\overline{f}+1, \overline{k}-1) \end{aligned} \tag{8}$$

$$\begin{aligned} \overline{M}_{t',t-1}(\overline{f}, \overline{k}) &= \overline{M}_{t',t-1}(\overline{f}-1, \overline{k}) \\ &\quad + e^{\eta G'_{t',t-1}(f)} \overline{M}_{t',t-1}(\overline{f}-1, \overline{k}-1) \end{aligned} \tag{9}$$

At timeslot t, for any $t' \in [1, t-1]$, if $k < k_s$ channels have been chosen from channels $1, \ldots, f-1$, we choose channel f with probability

$$e^{\eta G_{t',t-1}(f)} \frac{M_{t',t-1}(f+1, k_s-k-1)}{M_{t',t-1}(f, k_s-k)}. \tag{10}$$

If $t' = t$, all channels are chosen with the same probability $\frac{1}{nN}$.

Note that, t' is chosen before the computation of Eq. (10) according to the following distribution,

$$p_{t'} = \begin{cases} \frac{(1-\alpha)^{t-1} Z_{1,t-1}}{N W_t}, & \text{if } t' = 1 \\ \frac{\alpha (1-\alpha)^{t-t'} W_{t'} Z_{t',t-1}}{N W_t}, & \text{if } t' = 2, \ldots, t, \end{cases} \tag{11}$$

where $Z_{t',t-1} = \sum_{i=1}^{N} e^{\eta G'([t', t-1], i)}$ and $Z_{t,t-1} = N$. Here, W_t can be computed efficiently as follows

$$W_t = \frac{\alpha}{N} \sum_{t'=2}^{t-1} (1-\alpha)^{t-t'-1} W_{t'} Z_{t',t-1} + \frac{(1-\alpha)^{t-2}}{N} Z_{1,t-1}. \tag{12}$$

Note that $G'_{t',t}(f) = G'_{t',t-1}(f) + g'_{f,t-1}$, so $Z_{t',t} = M_{t',t}(1, k_s)$.

Instead of maintaining the weight of each strategy $w_{i,t}$, we compute the probability $q_{f,t}$ for each channel as follows

$$\begin{aligned} (1-\gamma) &\frac{\sum_{k=0}^{k_s-1} \overline{M}_{t-1,t-1}(f-1, k) e^{\eta G'_{t',t-1}(f)}}{M_{t-1,t-1}(1, k_r)} \\ &\cdot M_{t-1,t-1}(f+1, k_s-k-1) + \gamma \frac{|i \in \mathcal{C}|}{\mathcal{C}}. \end{aligned} \tag{13}$$

Algorithm 2 shows a complete description of the fast implementation algorithm. It is easy to see that, when calculating Eqs. (8) and (9) for a given t', it only requires $O(nk)$ computations. So, at each timeslot t, the computational cost of calculating $M_{t',t}(f, k)$ and $\overline{M}_{t',t}(f, k)$ for all $t' = 1, \ldots, t (t \in [1, T])$ and $f \in [1, n]$ is approximately $O(Tnk_s)$. In addition, the

Algorithm 2 A Fast Implementation of Tracking the Adaptive Compound Strategy for Anti-jamming Spectrum Access.

Input: $n, k, \delta \in (0,1), T, \beta \in (0,1], \gamma \in (0, 1/2], \eta \in (0,1), m \in \{0, 1, \ldots, T-1\}, \alpha = \frac{m}{T-1}$.
Initialization: Set the initial gain $G'_{t',0}(f) = g'_{f,0} = 0$ and the total weight $W_1 = 1$. Let $M_{t',t-1}(\overline{f}, 0) = 1$ and $M_{t,t-1}(n+1, \overline{k}) = \overline{M}_{t',t-1}(0, \overline{k}) = 0$ and compute $M_0(f,k)$ and $\overline{M}_0(f,k)$ following Eqs. (8) and (9), respectively.
For timeslot $t = 1, \ldots, T$,

1: Choose t' from $[1, t]$ randomly following Eq. (11).
2: Select channel f ($\forall f \in [1, n]$) one by one following Eq. (10) until a strategy I_t with k chosen channels is obtained.
3: Compute probability $q_{f,t}$ ($\forall f \in [1, n]$) following Eq. (13).
4: Obtain the channel reward $g_{f,t-1}$ and compute the virtual reward $g'_{f,t}$ ($\forall f \in [1, n]$) as $g'_{f,t} = \begin{cases} \frac{g_{f,t}+\beta}{q_{f,t}} & \text{if } f \in I_t \\ \frac{\beta}{q_{f,t}} & \text{otherwise.} \end{cases}$
5: Update $M_{t',t}(f,k)$ and $\overline{M}_{t',t}(f,k)$ for $t' = 1, \ldots, t$ following Eqs. (8) and (9), respectively.
6: Update W_t following Eq. (12).

End

computation of W_t and $q_{f,t}$ can be done in $O(T)$ and $O(k_s)$, respectively. Therefore, for all timeslots, the total time complexity is approximately $O(T^2 n k_s)$ while the space complexity is $O(Tnk_s)$.

We next show the correctness of Algorithm 2. Let $G'(f) = G'_{t',t-1}(f)$ and $c(f) = 1$ if channel f is chosen in the strategy i; otherwise $G'(f) = c(f) = 0$. Then, the number of channels chosen among channels $1, 2, \ldots, \overline{f}$ is denoted by $\sum_{f=1}^{f=\overline{f}} c(f)$. It is obvious that the virtual reward of strategy i is $G'_{t',t-1}(i) = \sum_{f=1}^{n} G'(f)$. Therefore, the probability that a strategy i is chosen for any $t' \in [1, t-1]$ at timeslot t is

$$\prod_{\overline{f}=1}^{n} \frac{e^{\eta G'(f) M_{t',t-1}(\overline{f}+1, k_s - \sum_{f=1}^{f=\overline{f}} c(f))}}{M_{t',t-1}(\overline{f}, k_s - \sum_{f=1}^{f=\overline{f}-1} c(f))} = \frac{e^{\eta \sum_{f=1}^{n} G'(f)}}{M_{t',t-1}(1, k_s)} = \frac{e^{\eta G'_{t',t-1}(i)}}{Z_{t',t-1}}. \quad (14)$$

Besides, if $t' = t$ the probability to choose strategy i is $\frac{1}{N}$. Thus, according to the conditional probability formula and Eqs. (11) and (14), we can derive the probability to choose the strategy i as $\frac{w_{i,t-1}}{W_{t-1}}$, which is exactly the same as the original algorithm shown in Section 5.1. Therefore, this fast implementation and the original algorithm are equivalent in the sense that the prediction sequences of strategies have the same distribution.

Finally, we have the following theorem to characterize the performance bound on the new *regret* defined in Eq. (7), which measures the difference between the proposed algorithm and the adaptive optimal compound strategy. As before, we first show how the SU sender and the SU receiver both approach their own adaptive optimal strategies and then derive the performance bound between these two optimal strategies in terms of the accumulated rewards over time.

Theorem *5:* For the new algorithm tracking the best compound strategy, the normalized reward distance $\frac{1}{T}|G_T^{max}(s) - G_T^{max}(r)|$ is upper bounded by $O(12k\sqrt{n \ln n})$, where $k = \max\{k_s, k_r\}$.

Proof: Following the same proof strategy in Theorem 1 and with a slight modification of proof of tracking the best expert in [23], we can show at the receiver side, with probability $1 - \delta$, the *regret* for the compound strategy $G_T^{max}(r) - \widehat{G}_T^{\mathcal{A}^r}$ is at most $2\sqrt{Tk_r}(\sqrt{4k_r|\mathcal{C}|D} + \sqrt{n(T+1)\ln\frac{n(T+1)}{\delta}})$ when $\beta^r = \sqrt{\frac{k_r T}{Tn}\ln\frac{nT}{\delta}}$, $\gamma^r = 2\eta^r k_r|\mathcal{C}|$, $\eta^r = \sqrt{\frac{D}{4Tk_r^2|\mathcal{C}|}}$, $D = T\ln N + T - 1$, and $T \geq \max\{\frac{k_r T}{n}\ln\frac{nT}{\delta}, 4|\mathcal{C}|D\}$. Using the facts $|\mathcal{C}| = \lceil \frac{n}{k_r} \rceil$ and $N \leq n^{k_r}$, we prove that when $T \to \infty$, the *regret* for the compound strategy is at most $6k_r\sqrt{T^2 n \ln n}$ by properly choosing k_r, n and δ.

Similarly, we obtain the bound $6k_s\sqrt{T^2 n \ln n}$ at the sender side. Finally, as $\widehat{G}_T^{\mathcal{A}^s} = \widehat{G}_T^{\mathcal{A}^r}$, $|G_T^{max}(s) - G_T^{max}(r)| \leq 12k\sqrt{T^2 n \ln n}$, where $k = \max\{k_s, k_r\}$. Thus, $\frac{1}{T}|G_T^{max}(s) - G_T^{max}(r)|$ is bounded by $O(12k\sqrt{n \ln n})$. □

Discussion. In comparison to the *regret* bound obtained using the static optimal strategy, the proposed algorithm cannot guarantee that the normalized reward distance converges to 0 when tracking the best compound strategy. This is because the best strategy may always change in each timeslot, and it is hard for the decision maker to adapt his choices to the adaptive optimal strategy. However, Theorem 5 guarantees that the reward distance between the sender and the receiver is at most $O(12k\sqrt{n \ln n})$ when T goes to infinity. In practice, k and n are pre-set system parameters, so the reward difference between two transceivers can achieve constant performance.

6 Numerical and Simulation Results

To demonstrate the effectiveness and robustness of the proposed jamming-resistant spectrum sensing and access protocol, in this section we provide an extensive numerical and simulation study under various jamming models.

In our simulation, we assume both the sender and the receiver use the proposed probabilistic anti-jamming protocol, *i.e.*, MAB-based online channel selection strategy. Meanwhile, the PU dynamically access the whole spectrum with $p_{11}^i > p_{01}^i$. The jammer, however, chooses his jamming strategy from static jamming, random jamming, myopic jamming and adaptive jamming (*i.e.*, MAB-based jamming). For ease of illustration, we let a four-element tuple denote the four parties' respective strategies. For example, "mab sta dyn mab" is used to denote the simulation setting that the sender uses the MAB-based strategy, the jammer uses static jamming strategy,

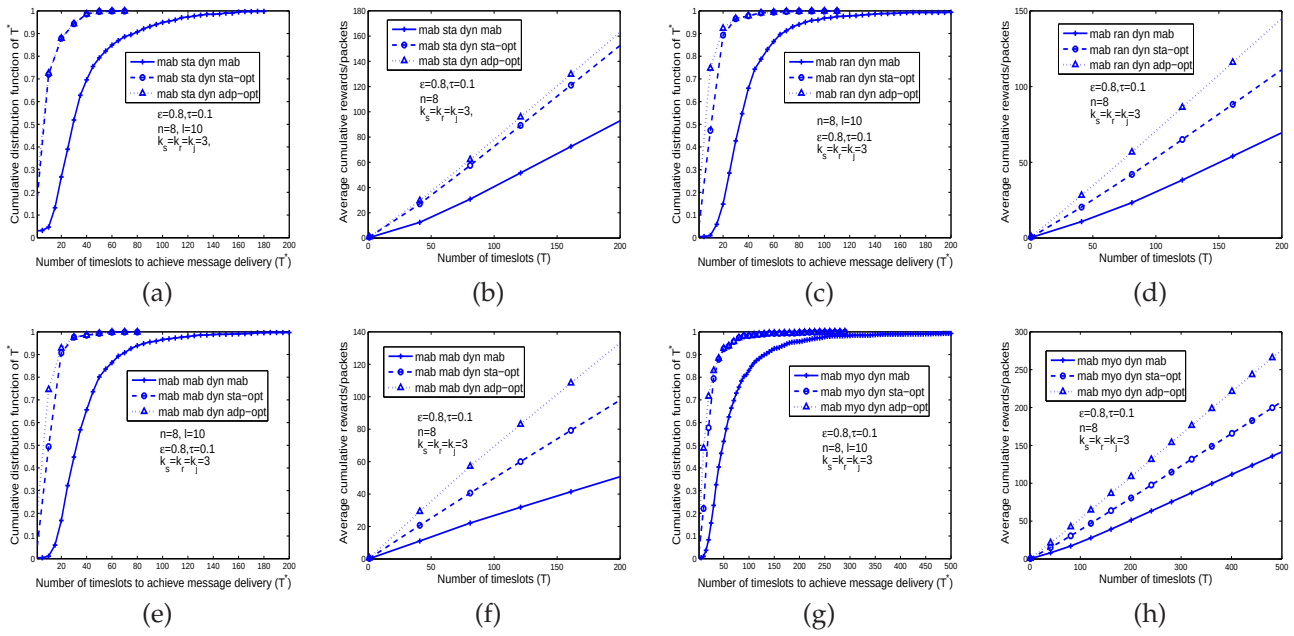


Fig. 1. Average number of received packets vs. the number of timeslots (T) and CDF of expected time to achieve message delivery under different strategy settings with $p_i^{11} > p_i^{01}$.

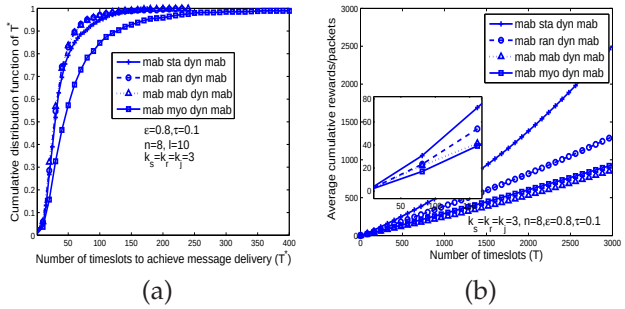


Fig. 3. The comparisons of the different jamming strategies on the system performance.

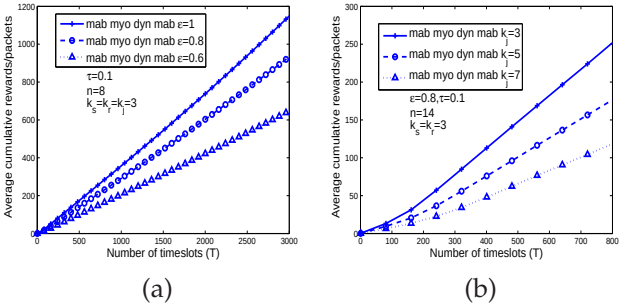


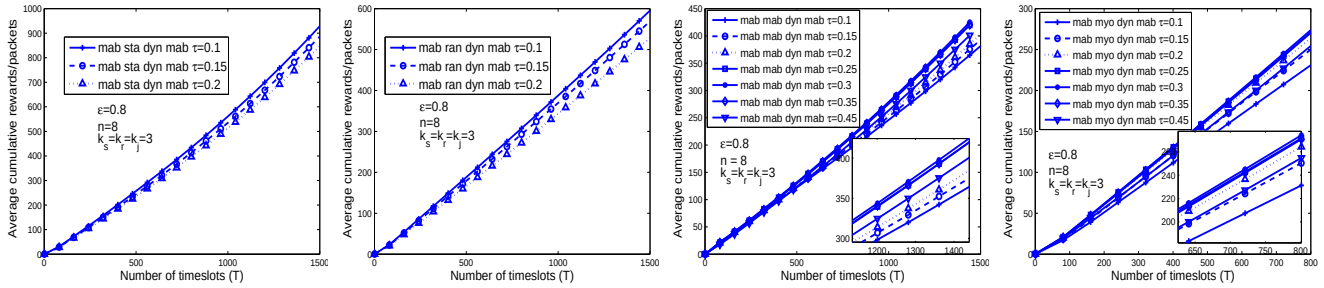
Fig. 4. The effects of sensing probability ε and jamming capability k_j on the system performance under "mab myo dyn mab".

the PU dynamically uses the spectrum according to certain traffic statistics and the receiver uses the MAB-based strategy. Without loss of generality, we assume $k_s = k_r = 3$, under which we vary the jammer's jamming capabilities and the total number of channels in the simulation.

6.1 Message Delivery Performance Evaluation

We first evaluate the performance of Algorithm 1. Fig. 1 shows (i) the average number of delivered packets as a function of T and (ii) the CDF of the expected time to achieve message delivery when $l = 10$, $k_j = 3$, $n = 8$ and $p_i^{11} > p_i^{01}$. Fig. 1 (a), (c), (e), (g) show that the performances of *static opt* and *adaptive opt* remain nearly the same especially when the jammer uses static jamming strategy. This implies the PU's dynamics incur relatively "static" channel status from the perspective of SUs. So, we cannot gain much more by using the *adaptive opt* than the *static opt*. We also compare the effect of different jamming strategies on the throughput performance in Fig. 3. In Fig. 3 (a), it is shown that when static, random or MAB-based jamming strategies are adopted and the number of packets to be transmitted is relatively small, the whole message can be delivered with high probabilities before $T = 150$. As for the myopic jamming attack, it takes $T = 250$ for the receiver to recover the whole message with high probability. However, as shown in Fig. 3 (b), if T continues to increase to 150 timeslots, the adaptive jammer incurs nearly the same performance deterioration as the myopic jammer. Among others, the key reason why the myopic jammer and the adaptive jammer are the most effective jammers is that they can make use of traffic statistics and/or acknowledgement information to dynamically adjust their jamming strategies.

Fig. 4 (a) and (b) illustrates how the sensing probability ε and the jamming capability k_j affect the performance, respectively. Not surprisingly, the increase of k_j will lead to less number of delivered packets,



(a) (b) (c) (d)

Fig. 2. The effects of sensing error probability τ on the system performance.

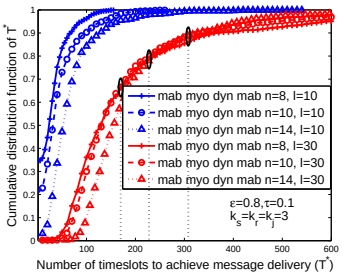


Fig. 5. The effects of n and l on the system performance with respect to the CDF of the expected time to achieve message delivery.

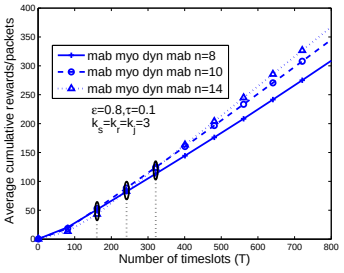


Fig. 6. The effect of n on the system performance with respect to the average cumulative rewards/packets.

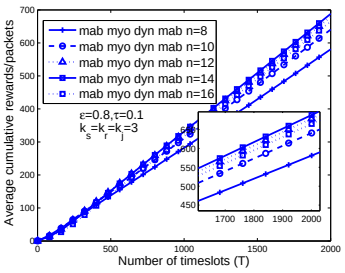


Fig. 7. The effect of n on the system performance with respect to the average cumulative rewards/packets.

and a larger sensing probability will enable the sender to update the strategy distributions using the sensing outcomes. In Fig. 2, we evaluate how the sensing error probability τ affects the system performance. It has been shown that under static jamming or random jamming attacks, the average number of cumulative delivered packets decreases when τ increases. Interestingly, if adaptive jamming and myopic jamming attacks are launched, the system performance is first improved as τ increases and then deteriorates when τ reaches a certain threshold. This is because a smaller τ can help disrupt the predictions of the two types of intelligent jammers on the available channels. If the sensing error probability τ continues to increase, sensing errors begins to dominate the performance and causes a performance deterioration.

In Fig. 5, Fig. 6 and Fig. 7, we use the setting "mab myo dyn mab" as an example to show how the parameters n and l affect the system performance. Fig. 5 shows that when l increases (*i.e.*, from 10 to 30), the expected time to received the message w.h.p. increases correspondingly. On the other hand, different values of n will also affect performance as T increases. For example, see the circle point in Fig. 5 and Fig. 6. When $T < 180$, the case of $n = 8$ gives the best performance; After $T > 180$, the case of $n = 10$ outperforms that of $n = 8$; When the time reaches $T = 240$, the case of $n = 14$ outperforms the case of $n = 8$ and it gives the best performance after $T = 320$. That means that it is better to choose a small n when the message size is short; a larger n is preferred when the message size is relatively large. However, it does not imply that the larger n will always give the best performance. As shown in Fig. 7, when n increases from 12 to 14, the performance gain is very small, and when n further increases to $n = 16$, the performance deteriorates. This is because the use of a large n also makes it difficult for the sender and the receiver to hop to the same set of channels.

We next evaluate the performance of Algorithm 2. In Fig. 8 (a), (b), (c) and (d), we show the impact of parameter m on the system performance. Let m_s, m_r denote the number of times to change the strategy in T timeslots for the sender and the receiver, respectively. As expected, the larger m_s will help to improve the system performance, which indicates

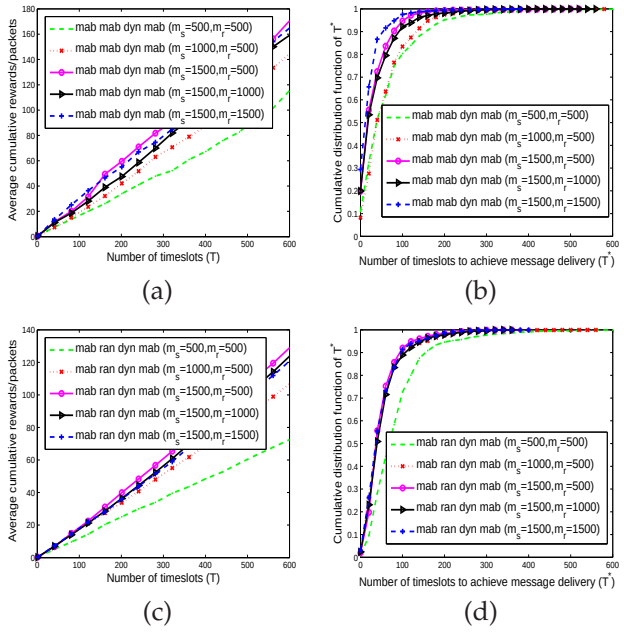


Fig. 8. Average number of received packets vs. the number of timeslots (T) and CDF of expected time to achieve message delivery under different strategy settings with $p_i^{11} > p_i^{01}$ and the change of m.

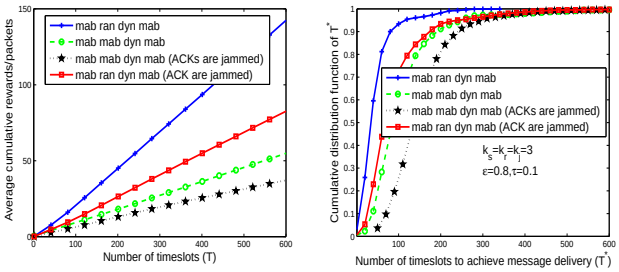


Fig. 9. The effect of jamming ACK information on the system performance.

that the transceiver requires more time to learn to choose good channels. The selection of a new strategy will contribute to the update of system parameters so that available channels are chosen with a higher probability. However, when $m_s = 1500$, the larger m_r will lead to less number of received packets and it requires more time to deliver the message with high probability. This is because it is difficult for the sender and the receiver to hop on the same channels when both parties choose new channels too frequently.

We also evaluate the system performance when the jammer randomly jams the ACK information. In practice, this is the best strategy the jammer can adopt to disrupt the strategy convergence between the sender and the receiver. In Fig. 9, we can see that when ACKs get jammed, the number of successfully received packets will decrease and it requires more time to deliver the whole message. However, it also indicates that our proposed anti-jamming spectrum sensing and access protocol can still defend against such a powerful jammer.

7 Conclusion

In this paper, we identified the vulnerability of the existing OSA protocols under malicious jamming attacks. Motivated by this observation, we studied jamming-resistant multi-radio multi-channel spectrum sensing and access in CRNs. We designed efficient and robust online OSA algorithms and analytically showed the regret bounds and approximation ratios of our methods with respect to optimal strategies. Our extensive simulations validate the theoretical analysis, showing that our methods perform extremely well and are very effective in defending against malicious jamming attacks.

References

[1] Q. Wang, K. Ren, and P. Ning, "Anti-jamming communication in cognitive radio networks with unknown channel statistics," in *Proc. of ICNP'11*, 2011, pp. 393–402.

[2] Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized cognitive mac for opportunistic spectrum access in ad hoc networks: A pomdp framework," *IEEE JSAC*, vol. 25, no. 3, pp. 589–600, 2007.

[3] S. H. A. Ahmad, M. Liu, T. Javidi, Q. Zhao, and B. Krishnamachari, "Optimality of myopic sensing in multi-channel opportunistic access," *IEEE Transactions on Information Theory*, vol. 55, no. 9, pp. 4040–4050, 2009.

[4] K. Liu, Q. Zhao, and B. Krishnamachari, "Dynamic multichannel access with imperfect channel state detection," *IEEE Transactions on Signal Processing*, vol. 58, no. 5, pp. 2795–2808, 2010.

[5] J. Unnikrishnan and V. V. Veeravalli, "Algorithms for dynamic spectrum access with learning for cognitive radio," *IEEE Transactions on Signal Processing*, vol. 58, no. 2, pp. 750–760, 2010.

[6] K. Liu and Q. Zhao, "A restless bandit formulation of multichannel opportunistic access: Indexablity and index policy," *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5547–5567, 2010.

[7] A. J. Viterbi, *CDMA: Principles of Spread Spectrum Communication*. Addison Wesley, 1995.

[8] M. Strasser, C. Pöpper, S. Capkun, and M. Cagalj, "Jamming-resistant key establishment using uncoordinated frequency hopping," in *Proc. of IEEE Security and Privacy*, May 2008.

[9] M. Strasser, C. Pöpper, and S. Capkun, "Efficient uncoordinated fhss anti-jamming communication," in *Prob. of ACM MobiHoc'09*, 2009.

[10] D. Slater, P. Tague, R. Poovendran, and B. J. Matt, "A coding-theoretic approach for efficient message verification over insecure channels," in *Proc. of ACM WISEC'09*. ACM, 2009.

[11] A. Liu, P. Ning, H. Dai, and Y. Liu, "Usd-fh: Jamming-resistant wireless communication using frequency hopping with uncoordinated seed disclosure," in *Proc. of MASS'10*, 2010.

[12] Y. Liu, P. Ning, H. Dai, and A. Liu, "Randomized differential dsss: Jamming-resistant wireless broadcast communication," in *Proc. of IEEE INFOCOM'10*, 2010.

[13] A. Liu, P. Ning, H. Dai, Y. Liu, and C. Wang, "Defending dsss-based broadcast communication against insider jammers via delayed seed-disclosure," in *Proc. of ACSAC*, 2010, pp. 367–376.

[14] H. Li and Z. Han, "Dogfight in spectrum: Combating primary user emulation attacks in cognitive radio systems, part i: Known channel statistics," *IEEE Transactions on Wireless Communications*, vol. 9, no. 11, pp. 3566–3577, 2010.

[15] ——, "Dogfight in spectrum: Combating primary user emulation attacks in cognitive radio systems - part ii: Unknown channel statistics," *IEEE Transactions on Wireless Communications*, vol. 10, no. 1, pp. 274–283, 2011.

[16] S. Gao, L. Qian, D. R. Vaman, and Z. Han, "Distributed cognitive sensing for time varying channels: Exploration and exploitation," in *Proc. of WCNC*, 2010.
[17] B. Wang, Y. Wu, K. J. R. Liu, and T. C. Clancy, "A stochastic anti-jamming game in cognitive radio networks," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 4, pp. 877–889, 2011.
[18] Y. Wu, B. Wang, K. R. Liu, and T. C. Clancy, "Anti-jamming games in multi-channel cognitive radio networks," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 1, pp. 4–15, 2012.
[19] A. O. Hero, D. A. Castan, D. Cochran, and K. Kastella, *Foundations and Applications of Sensor Management*. Springer Publishing Company, Incorporated, 2007.
[20] P. Whittle, "Restless bandits: activity allocation in a changing world," *Journal of Applied Probability*, vol. 25A, pp. 287–298, 1988.
[21] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The nonstochastic multiarmed bandit problem," *SIAM J. Comput.*, 2002.
[22] B. Awerbuch and R. D. Kleinberg, "Adaptive routing with end-to-end feedback: distributed learning and geometric approaches," in *Proc. of ACM STOC'04*, 2004, pp. 45–53.
[23] A. György, T. Linder, G. Lugosi, and G. Ottucsák, "The on-line shortest path problem under partial monitoring," *J. Mach. Learn. Res.*, 2007.
[24] V. Namboodiri, "Are cognitive radios energy efficient? a study of the wireless lan scenario," in *Proc. of IPCCC'09*, 2009, pp. 437–442.