Multiuser Scheduling in Centralized Cognitive Radio Networks: A Multi-Armed Bandit Approach

Amir Alipour-Fanid, *Member, IEEE*, Monireh Dabaghchian, *Member, IEEE*, Raman Arora, and Kai Zeng, *Member, IEEE*

***Abstract*—In wireless communication networks, the network provider serves certain licensed primary users who pay for a dedicated use of the frequency channels. However, not all the channels are occupied by the primary users at all times. For efficient spectrum utilization, in centralized cognitive radio networks (CRNs), a cognitive base station (CBS) dynamically identifies the spectrum holes and allocates the frequency channels to the on-demand unlicensed secondary users known as cognitive radios (CRs). Although existing literature has developed various dynamic spectrum access mechanisms for CBS, there is still a dearth of studies due to the wide range of assumptions made in the solutions. Most of the existing works study the CBS scheduling problem scheme by adopting optimization-based methods and rely on the prior knowledge of the network parameters such as primary users' activity. Moreover, the impact of channel switching costs on the network throughput has not been well studied. In this paper, we aim to maximize the CRNs total throughput, and we formulate the CBS scheduling problem as a non-stochastic (i.e., adversarial) combinatorial multi-armed bandit problem with semi-bandit feedback and arm switching costs. We propose two novel online learning algorithms for CBS scheduling with and without channel switching costs, where their regret performances are proved sublinear order-optimal in time as $T^{1/2}$ and $T^{2/3}$, respectively, offering throughput-optimal scheduling for CRNs. Experiments on the synthetic and real-world spectrum measurement data complement and validate our theoretical findings.**

***Index Terms*—Cognitive radio network, multi-armed bandit, channel switching costs, network throughput, multichannel wireless communication.**

Manuscript received June 21, 2021; revised December 3, 2021; accepted January 7, 2022. Date of publication February 4, 2022; date of current version June 9, 2022. This work was supported in part by the Commonwealth Cyber Initiative (CCI) and its Northern Virginia (NOVA) node, US Army Research Office (ARO) under Grant No. W911NF-21-1-0187, NSF Networking Technology and Systems (NeTS) program under Grant No. 2131507, NSF Research Initiation Award under Grant No. 2100804, and Microsoft Research Award. The associate editor coordinating the review of this article and approving it for publication was T. Chen. *(Amir Alipour-Fanid and Monireh Dabaghchian contributed equally to this work.) (Corresponding author: Amir Alipour-Fanid.)*

Amir Alipour-Fanid is with the Architectures and Security Team, General Motors Research and Development, Warren, MI 48092 USA (e-mail: amir.alipour-fanid@gm.com).

Monireh Dabaghchian is with the Department of Computer Science, Morgan State University, Baltimore, MD 21251 USA (e-mail: monireh.dabaghchian@morgan.edu).

Raman Arora is with the Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: arora@cs.jhu.edu).

Kai Zeng is with the Department of Electrical and Computer Engineering, George Mason University, Fairfax, VA 22030 USA (e-mail: kzeng2@gmu.edu).

Digital Object Identifier 10.1109/TCCN.2022.3149113

I. Introduction

A FLOURISH in the number of Internet-of-Things (IoT) and mobile applications in recent years has led to explosive growth in the demand for spectrum resources. To address the imminent spectrum shortage problem, Federal Communications Commission (FCC) has authorized opening spectrum bands (e.g., 3550-3700 MHz and TV white space) owned by licensed primary users (PUs) to unlicensed secondary users (SUs), when the PUs are not active on frequency bands [1], [2], [3]. This authorization has led to the emergence of cognitive radio networks (CRNs) as a promising paradigm to shift the spectrum utilization efficiency and provide ubiquitous connections for many growing numbers of applications such as smart city, smart home, intelligent vehicles, smart grid, smart farming, healthcare systems, etc. [2], [4], [5].

Spectrum sharing networks are usually categorized into centralized and decentralized CRNs [6], [7]. In decentralized CRNs, secondary user cognitive radios (CRs) access the spectrum in an opportunistic fashion by running their own internal dynamic spectrum access (DSA) policy. In centralized CRNs, which is the focus of this paper, a cognitive base station (CBS) assigns the frequency channels to the CRs by dynamically searching for the unused portions of the licensed spectrum (a.k.a. spectrum hole or white space). This type of spectrum sharing approach in centralized CRNs is called *CBS scheduling* [8].

In CRNs, the knowledge about PUs activity (i.e., ON/OFF or busy/idle on the channels) is a fundamental building block for designing and implementing efficient spectrum sharing mechanisms. In fixed spectrum assignment, full knowledge of PUs spectrum occupancy is queried by the CBS from an external white space database [9]. However, this method incurs higher communication overhead between the CBS and database and assumes that the information stored in the database is always reliable and has not been breached.

A recurring theme in much of prior work is to model the PUs activity using a parametric family of probability distributions [10] and to employ statistical methods for parameter estimation [11], [12]. In particular, the PUs activity is typically modeled using the classical two-state Markov chain wherein the maximum likelihood estimation techniques are applied to estimate the PUs state transition probabilities [13], [14]. More recently, some works have explored supervised learning approaches by leveraging prior data to predict PUs activity with a reasonable level of accuracy [15]. However, in

practical scenarios, PUs channel occupancy model may not be known a priori and is likely to evolve in an arbitrary manner. Further, different PUs may not have the same ON/OFF behavior, and they may change their activity patterns depending on the demands and frequent changes in the regulations. It is, therefore, unrealistic to model the PUs activity as stochastic.

Another key practical feature of CBS scheduling is that CRs' switching from a certain frequency channel to a different channel incurs a cost in terms of lost throughput as the radio takes time to actuate and settle [16], [17]. Mainly, channel switching latency in CRs depends on hardware limitations and imperfections of the front-end frequency synthesizers, and the type of CRs designed to support a specific application. Studies show that channel switching latency can range from 0.224 ms in typical CRs [18, Sec. 14.6.12] to 160 ms in industrial CRs [16], [19]. This latency range is significant compared to the time slot length in communication systems which is usually between 1 ms to 200 ms, depending on the protocol specification. Therefore, CRs channel switching introduces a non-negligible delay resulting in lost throughput and CRNs performance degradation. As a result, it is of paramount importance to address the channel switching latency in the CBS scheduling problem and investigate its impact on the network's total achievable throughput.

Most of the existing work has adopted optimization-based methods and considered the channel switching costs as a constraint [16], [17], [19], [20]. However, the solutions of these approaches are either subject to some constraints or are heuristic (see, e.g., [16]), and the theoretical performance guarantees of the global solutions are not readily available. Such methods also require the knowledge of the system model parameters and rely on the prior knowledge of the PUs activity.

We study the problem in a non-stochastic setting wherein we assume no knowledge of the PUs channel utilization behavior. Moreover, we take the impact of channel switching latency into account and integrate it with the CBS scheduling framework. We model the CBS scheduling problem as a combinatorial multi-armed bandit problem and propose two online learning algorithms with and without channel switching costs. To design the algorithms, we adapt the well-known non-stochastic EXP3 algorithm – the seminal work by Auer *et al.* [21] – where our algorithms achieve order-optimal performance and enjoy simpler implementation and analysis.

Based on the proposed scheduling policy, at each time slot the CBS picks a subset of s channels out of K available channels, which we refer to as a *slate* [22], [23], and then assigns each channel to a CR for data transmission. If in any time slot the CBS picks a different subset of channels than the previous time slot, it incurs a switching cost in terms of lost throughput for each of the CRs. To handle the channel switching costs, we let the CBS to switch the slate according to a Bernoulli stochastic process with a time-decaying parameter. We show that this method avoids excessive throughput loss by optimal channel switching. Finally, at the end of each time slot, each CR reports back to the CBS the attained throughput (realized as an observed reward on the assigned channel) with which the CBS updates its learning parameters.

Various types of feedback are possible in combinatorial bandit problems [24], [25]. Our setting fits into the *semi-bandit* feedback as the CBS observes feedback for each assigned channel and their combinations. Hence, we name the proposed online learning framework as *s-set semi-bandit with switching costs*. The goal for the CBS is to minimize the empirical *regret*, defined to be the difference of maximal cumulative throughput of the *best slate*, in hindsight, and that achieved by the CBS. Intuitively, this quantity measures how much the CBS regrets not following a different competing strategy (e.g., a constant selection of a particular slate). We say that the CBS is learning (equivalently, has zero regret) if its regret is a sublinear function of the total number of time slots (equivalently, the average regret goes to zero asymptotically).

We note that as with any successful online learning algorithm, a multi-armed bandit requires a careful tradeoff between exploration (i.e., to acquire more information about the expected throughput of the other channels) and exploitation (i.e., to utilize the channel that is likely to yield the highest throughput). This challenge is further compounded by the need to account for switching costs which makes exploration expensive, and the semi-bandit feedback which can potentially help with exploration. The algorithm(s) we present here strike an optimal balance and yield order-optimal regret bounds.

Our main contributions in this paper are as follows.

- We formulate the problem of cognitive base station (CBS) scheduling in centralized CRNs as a non-stochastic combinatorial multi-armed bandit problem with semi-bandit feedback and switching costs. Our setting has no assumption on the primary users' activity while accounting for the impact of CRs' channel switching costs in the network's total throughput.
- We propose two novel online learning algorithms for the CBS scheduling problem described above as follows: 1) CBS scheduling without channel switching costs which achieves an order-optimal regret of $O(\sum_{i=1}^{s}\sqrt{(K-i+1)T\ln(K-i+1)})$, where K is the number of frequency channels, s is the number of CRs served by the CBS, and T is the total number of time slots; 2) CBS scheduling with switching costs which obtains an order-optimal throughput regret of $O(s(K\ln K)^{1/3}T^{2/3})$.
- Our algorithms follow the principal of exponentially weighted average method in the popular EXP3 algorithm [21]. Hence, they are computationally efficient, simpler to implement, and easy to analyze.
- To validate our findings, we conduct experimental evaluations on both the synthetic and real-world spectrum measurement data and demonstrate the consistency between theoretical analysis and empirical evaluations.

The rest of the paper is organized as follows. In Section II, we discuss the related work. Section III presents the system model and problem formulation. The proposed CBS scheduling algorithms and the main results are presented in Section IV. We present experimental evaluations and empirical comparisons in Section V. Finally, we briefly discuss the future work in Section VI, and conclude the paper in Section VII.

II. Related Work and Background

A. Motivation: Centralized Versus Decentralized CRNs

Cognitive radio networks are mainly configured in centralized or decentralized (i.e., distributed) settings [7], [26], [27], depending on the application needs, requirements, multiple trade-off parameters such as network throughput, signaling overhead, secondary users' power constraints and their intelligence capabilities [28], etc. In the following, we point out the main characteristics of each architecture.

1) Centralized CRNs: i) This configuration is suitable for small-scale networks (e.g., a typical indoor scenario) with power constrained secondary user devices such as smart home IoT users where an access point acts as a cognitive base station (CBS) scheduler. The access point then can manage the dynamic spectrum sharing between the users in the local network and the primary users such as TV station signal. ii) Due to the centralized spectrum management, this configuration yields to a better performance in heterogeneous cognitive radio networks. iii) Since the spectrum decision making algorithm is run on the cognitive base station, the secondary users do not need to consume power for spectrum availability inference. This structure then reduces the power consumption significantly, highlighting the cost effectiveness of the centralized model. iv) The main drawback of such schemes is that it is necessary to collect all the information at the access point which may generate overhead for exchanging the context and control information of the entire network.

2) Decentralized CRNs: i) In this configuration secondary users run their own spectrum sharing algorithms and each device can select the best resource without any centralized management, so the setting is suitable for ad-hoc cognitive radio networks. ii) Since each node handles the decision-making algorithms individually the computation overhead and power consumption in the device is increased significantly. iii) The devices require additional hardware resources to be implemented at the node level, which is very costly, in practice. iv) Decision information sharing and synchronization among the users are costly as the network requires $s(s-1)/2$ communication links among the users (s denotes the number of secondary users). v) Each node may act selfishly and can endure various network failures independently.

B. Cognitive Base Station Scheduling in Centralized CRNs

Cognitive base station (CBS) scheduling is a fundamental building block of efficient spectrum utilization in centralized CRNs. Previous studies primarily have relied on optimization-based methods and typically formulated the CBS scheduling problem as a throughput maximization and energy efficient resource allocation problem [8], [16], [20], [29]. In these works, the solution ended up on solving a nonlinear integer programming problem which yielded to a set of heuristic CBS scheduling algorithms. The work by [30] assumed that each CR is allowed to transmit on multiple channels. This assumption helped to relax the constraints in the optimization problem and simplify the channel assignment mechanism. However, the theoretical performance guarantee of the scheduler is not available. Despite many advances in developing CBS scheduling mechanisms, existing works still lag behind in designing efficient frameworks to effectively meet the CRs traffic requirements while considering throughput loss due to the channel switching latency.

CBS scheduling with channel switching costs has been studied by the previous work [16], [17], [19], [20], [31]. However, again the solutions are heuristic algorithms which solve a set of optimization problems in offline manner. The methods also rely on prior knowledge of the system model parameters. Instead, we propose to use online learning approach where we provide theoretical performance guarantees for the proposed algorithms. To address the channel switching costs which result lower spectrum efficiently, we implement an efficient channel switching algorithm in the CBS which avoids excessive costs by optimally restricting the channel switching and reducing the unavoidable overhead.

The knowledge on PUs activity (ON/OFF status) is another crucial factor in designing effective and efficient CBS scheduling algorithms. Existing work either assume that PUs activity is fixed and available a priori [9] or aim to model the occupancy state of PUs with known statistical models [13], [14], [32]. In these works, the PUs activity are typically modeled using the classical two-state Markov chain (a.k.a. Gilbert-Elliott Model) wherein the maximum likelihood estimation techniques are applied to estimate the PUs state transition probabilities [10], [13], [14]. The work by [10] presents an alternative, stochastic differential equation based spectrum utilization model that captures dynamic changes in channel conditions induced by PUs activity. The recent work by [11] has also utilized generalized Pareto distributions for long and short time PUs occupancy sequence estimation. Estimating the primary user behavior by adaptive length of the sample sequence is proposed in [14]. The method dynamically estimates the required length of the sample sequence which is adaptive to the changing of the PUs behavior. However, estimation accuracy determined by the confidence level of this method still suffers from the low prediction accuracy.

Recently, machine learning approaches have been leveraged to predict the PUs activity [11], [12]. In [12] the PUs' signal feature is extracted using standard energy and cumulant calculations, while the work by [11] applies neural network model of long short term memory. The performance guarantees and computation complexity of these methods are still under investigation. In this paper, we relax the above assumptions and assume no prior knowledge nor statistical models on the PUs activity. In particular, within a family of online learning-based CBS scheduling algorithms we assume non-stochastic PUs activity.

Most of the work which develop online learning algorithms for spectrum sharing in CRNs, mainly assume a specific probability distribution for the PUs activity and model it as a stationary stochastic process [33], [34], [35]. The work by [33] models the PUs activity as an arbitrarily-distributed random variable with bounded support but unknown mean, i.i.d. over time. However, after we examined the real-world spectrum measurement data [36], [37] and extracted the PUs activity over multiple frequency channels across various frequency bands (refer to Section V), we found that the PUs activity are

not behaving well and are not stable over the time to follow a specific and known stochastic model. Hence, we designed novel combinatorial algorithms which not only tackles the non-stochastic PUs activity, but also takes the channel switching costs into the consideration and provides provable performance guarantees. Next, we give a brief background and related work information on multi-armed bandits (MAB) which is the main tool we leveraged to develop the CBS scheduling algorithms for centralized CRNs.

C. Multi-Armed Bandit

1) Multi-Armed Bandit Problem: One of the most fundamental online learning problems is that of multi-armed bandits, wherein, at each round a player (or learner) chooses an action out of K available actions and observes the reward associated with the chosen arm. The reward at each round can either be stochastic or non-stochastic (a.k.a., adversarial) [21], [38]. In this paper, we focus on non-stochastic setting as the PUs activities are assumed to follow no statistical distribution. The goal of the player in the MAB problems is to minimize empirical regret, defined to be the difference of maximal cumulative reward of any arm, in hindsight, and that collected by the player. We say that the player is learning if its regret is a sublinear function of the total number of rounds.

The problem of multi-armed bandits was introduced by [39] in the context of studying medical trials, and popularized further by [40]. The problem was first studied in a non-stochastic setting in the seminal work of [21]. The popular Exponential-weight algorithm for Exploration and Exploitation (EXP3) was proposed by [21], and was inspired by prior work on weighted majority algorithm [41] and Hedge algorithm [42]. EXP3 achieves regret of $O(\sqrt{KT\ln K})$ for K-armed bandits over T rounds. Later on, Audibert and Bubeck [43] considered a new class of randomized policies and proposed INF (Implicitly Normalized Forecaster) algorithm which improved the EXP3 by a factor of $\sqrt{\ln K}$ and achieved a minimax regret of $\Theta(\sqrt{KT})$.

2) Bandit Learning With Switching Costs: The problem of multi-armed bandits with switching costs was introduced by Dekel *et al.* [44]. They were primarily interested in establishing a lower bound to match the regret upper bound of a mini-batching algorithm proposed by Arora *et al.* [45]. In particular, [44] shows that the mini-batched variant of EXP3 studied by [45], achieves the minimax regret of $\Theta((K\ln K)^{1/3}T^{2/3})$. This method has also been applied in the application of mobility management of users in communication systems by [46]. Learning-wise adopting mini-batching algorithm of [45] to the semi-bandit feedback setting of [23], [47] provides semi-bandit with switching costs wherein its regret results in the same order as of this paper. The analysis of such setting could be involved which affects the learning parameters. Application-wise mini-batching forces the CBS to select the same set of frequency channels for a fixed period of time. In that case, if an attacker (such as primary user emulation attacker [48], [49]) finds out the selected channel set, the attacker can launch the PU signal over the channel until the end of mini-batch size and degrade the CRN throughput significantly. However, since our approach adopts randomness in deciding to switch or not switching the transmission channels, it adds robustness to the communication systems and achieves secure CRNs.

Recently, Arora *et al.* [50] investigated the bandit learning with feedback graphs and switching cost, and proposed an adaptive mini-batching strategy to achieve the minimax regret. In [50] the mini-batch size is proportional to the probability that the arm is sampled to be played. That means the mini-batch size could vary over time. Adopting this framework to the s-semi bandit setting, multiple arm selection with different mini-batch size for each arm may result in collision in subsequent arm selections (i.e., assigning a channel for more than one user in the CBS scheduling problem). In this paper, we propose an s-semi bandit algorithm with probabilistic switching policy which controls the switching of an s-set arm without arm selection collision with optimized learning parameters and order-optimal learning regret.

3) Online Learning With Semi-Bandit Feedback: Many of the real-world problems, especially those that involve sequential decision making, can be posed by MAB with pulling multiple arms simultaneously. For example, in an online ads display problem, a learner can choose multiple ads to display on a given webpage. The task for learner, therefore, includes selecting a subset of s arms out of K available arms ($s < K$). This online learning problem is called bandit slate [22], [23] or combinatorial MAB [24], [51] problem which shows up in many other applications including spectrum sharing in wireless communication networks, routing in computer networks, search engines, personalized matching, etc. The bandit slate problems were first studied by [23], which studies both the ordered (permutation problem) and unordered settings. Several other researchers build on that work to give algorithms that yield optimal algorithms in both stochastic and non-stochastic environments simultaneously [52], yield data-dependent regret bounds [53]. The work by [47] proposed a variant of EXP3 for multiple players with the focus of running time and space efficiency improvements where they obtained the regret upper bound as $O(\sqrt{sKT\ln(K/s)})$.

In other line of research [54], [55], [56], non-stationary is introduced within the stochastic bandit problem by allowing the mean rewards to change at some time-step while staying stationary between those changes. This setting is called switching bandit [54]. These algorithms hold strong assumptions for the application of CBS scheduling problem. First, they rely on abrupt changes in the arms reward distributions while staying stationary for the time intervals wherein the distribution is not changing. However, in the experimental evaluations we saw that the PUs states do not change abruptly and do not stay fixed over some time intervals. Second, they require a priori knowledge of number of distribution changes [54], [55] and the gap variable which depends on the distributions of arm outcomes [55]. Third, distribution-independent regret in the dynamic environment of [55] achieves regret order of $T^{2/3}$ which is worse than $T^{1/2}$ achieved for the non-stochastic combinatorial s-set semi-bandit setting studied in this paper.

Various types of feedback are possible in combinatorial bandit including *full feedback*, *bandit feedback*, and *semi-bandit*

feedback [24]. In full-feedback the player observes the reward on all the K arms at each round. In bandit feedback the player observes only the summation of the rewards on the selected s arms, and in the semi-bandit feedback the player not only observes the summation of the rewards, but also it observes the reward on each arm in the selected subset. Our CBS scheduling problem in this paper fits into the bandit slate with semi-bandit feedback setting. There is a rich literature on combinatorial multi-armed bandit problem with semi-bandit feedback [24], [51], [52], [57], [58].

However, the above works do not consider the arms' switching costs in the semi-bandit feedback setting and its impact on the regret bound. In addition, the algorithms are fairly involved (see e.g., online stochastic mirror descent method in [24]) and are not easily amenable to account for switching costs. On the other hand, the algorithms we present to address the CBS scheduling problem results in a unified framework to handle both semi-bandit feedback and switching costs. The proposed approach admits simpler analysis and implementation as it follows the same fundamental design procedure which has been proposed by Auer *et al.* to design the EXP3 algorithm [21]. We prove that our solution provides an order-optimal regret bound for CBS scheduling problem, offering throughput-optimal scheduling for CRNs.

III. System Model and Problem Formulation

A. System Model

We consider a centralized cognitive radio network (CRN) consisting of a cognitive base station (CBS), multiple primary users (PUs) and secondary users cognitive radios (CRs). Centralized CBS scheduling typically targets smaller cognitive networks in terms of limited geographical coverage (e.g., a typical indoor scenario such as smart home cognitive IoT network controlled by a local access point). Hence, in this paper the location impact of the PUs and CRs are not considered, so it is assumed that the throughput is the same for each CR on the same frequency channel. Our objective is to design a family of online learning algorithms for the CBS scheduler and achieve optimal network throughput. The application scenario is illustrated in Fig. 1. In the following, we point out the main assumptions and key components in this scenario, assuming that the communication system operates in time-synchronized manner with discrete-time units, called time-slots.

- *Spectrum Resource:* We consider that the spectrum resource in the target CRN is partitioned into K non-overlapping orthogonal frequency channels. The channels are licensed to the PUs, however, for efficient spectrum utilization they are shared with unlicensed CRs via the CBS scheduler.
- *CRs Spectrum Utilization Schemes:* In cognitive radio networks the CRs may seek three different approaches to share the spectrum with PUs: 1) Underlay: in this approach simultaneous CRs and PUs transmissions are allowed as long as the interference level at the PUs side remains acceptable. 2) Overlay: in this approach, PUs share the knowledge of their signal codebooks with the CRs, allowing CRs to be aware of spectrum utilization a prior. 3) Interweave: in this approach the CRs sense the frequency channels and access them for data transmission as long as the PUs remain idle. Our CBS scheduling problem in this paper targets the interweave spectrum sharing approach. For further information about these paradigms interested readers may refer to [59], [60].
- *CBS Scheduling:* At each time slot, the CBS selects a set of channels and assigns each channel to a CR. The CRs then sense the assigned channel and transmit data over the channel if PUs signal are not present (OFF status) on that specific time slot and channel. However, if a PU signal is present (ON status), then the associated CR stays on sleep mode (no data transmission) without interfering with the PUs. Then, at the end of the same time slot, all the CRs report back to the CBS the throughput that each could achieve as a result of the CBS channel assignment.
- *PUs Activity:* We assume the PUs activity information (ON/OFF or busy/idle status) is not known to the CBS scheduler a priori, nor it follows any statistical distribution. In other words, the PUs spectrum occupancy is considered to be non-stochastic and unknown. The CBS then runs its own built-in online learning algorithms (proposed in this paper) to learn the PUs activity and opportunistically schedule the transmission channels for the CRs.

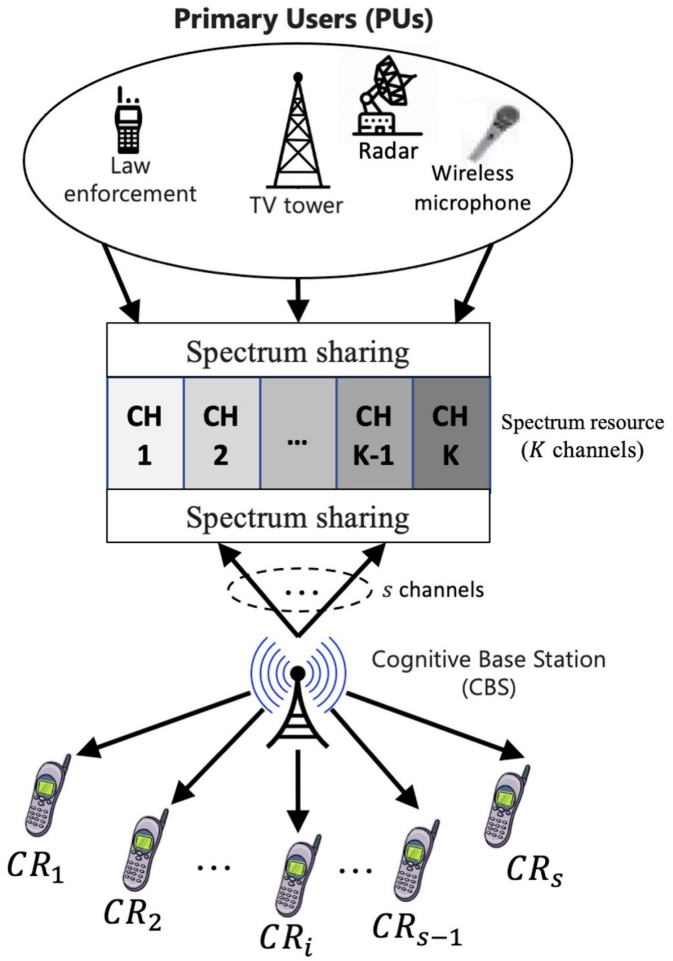


Fig. 1. Application scenario of cognitive base station (CBS) scheduling for cognitive radios (CRs).

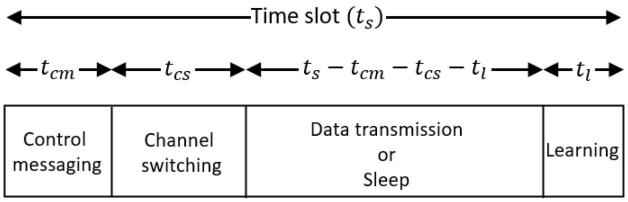


Fig. 2. Time slot structure.

- *Channel Switching Cost:* If the target channel scheduled by the CBS in the next time slot is different from the current one, the CRs have to switch the frequency channel. Reassigning the communication on the new channel incurs a non-negligible latency which results in throughput loss and ultimately the CRNs' performance degradation [16], [17].
- *Time Slot Structure:* Fig. 2 illustrates the time slot structure. A time slot t_s is partitioned into four non-overlapping portions including: control messaging time t_{cm}, channel switching time t_{cs}, learning time t_l, and data transmission or sleep time $t_s - t_{cm} - t_{cs} - t_l$. If a CR does not switch the channel, the channel switching time becomes zero ($t_{cs} = 0$), and then the CR transmits or sleeps for a duration of $t_s - t_{cm} - t_l$, depending on whether PUs signal is ON or OFF on the channel. The duration of both control messaging and learning time are assumed to be fixed in the whole communication time.
- *Throughput for CRN:* The throughput of a CR during a time slot depends on data transmission duration and capacity of the channel. Assuming additive white Gaussian noise (AWGN) channel with *W* Hz bandwidth and a given signal-to-noise ratio (SNR), the channel capacity denoted by *B* bps is obtained by Shannon's well-known formula as

$$B = W \log_2(1 + SNR). \tag{1}$$

The throughput for CRN then is defined as the number of bits transmitted by all the CRs through the available channels within a given time frame.

B. Problem Formulation

We formulate the CBS scheduling problem as an online learning problem posed by a family of non-stochastic combinatorial multi-armed bandits with semi-bandits feedback and arm switching costs. The main notation is summarized in Table I. Let $[K] = \{1, 2, \ldots, K\}$ denote the set of primary network's channels. The set of secondary user cognitive radios $C = \{CR_1, CR_2, \ldots, CR_s\}$, of size $|C| = s$, are registered in the CBS. At each time slot, $t = 1, \ldots, T$, the CBS picks a subset of channels, $S(t) \subset [K]$ which we refer to as *slate*, of size $|S(t)| = s$, and assigns each channel to a CR in set *C*. The slate is a set of channels with $S_i(t)$ denoting the channel assigned to the CR_i (i.e., the CR which is in the i^{th} position on the set *C*). After channel assignment, the CR_i attains the throughput of $Y_{CR_i}(t)$ at time slot *t* as follows:

$$Y_{CR_i}(t) = x_{S_i(t)}(t) - c(t)\mathbb{1}_{\{S_i(t) \neq S_i(t-1)\}}, \tag{2}$$

TABLE I
SUMMARY OF MAIN NOTATION

Notation	Definition
$[K]$	the set of channels $[K] := \{1, ..., K\}$.
C	the set of CRs, $C = \{CR_1, CR_2, \ldots, CR_s\}$.
$[s]$	the set of index for CRs, $[s] := \{1, ..., s\}$.
T	the given total time slots.
$S(t)$	the slate: a subset of channels $S(t) \subset [K]$.
$S_i(t)$	the channel in the slate $S_i(t) \in S(t)$ assigned to the CR_i.
ν	a given online learning policy.
$\sigma(t)$	the channel switching policy.
$p_j^{(i)}(t)$	the probability of channel j selection for the CR_i.
$\delta(t)$	the channel switching probability.
γ	the exploration rate.
η	the learning rate.
$\mathbb{1}_{\{A\}}$	Indicator of event A.
$\mathbb{E}[\cdot]$	Expectation operator.
$f(T) = O(g(T))$	$\lvert f\rvert$ is bounded above by g (up to constant factor) asymptotically.
$f(T) = \Omega(g(T))$	$\lvert f\rvert$ is bounded below by g asymptotically.
$f(T) = \Theta(g(T))$	f is bounded both above and below by g asymptotically.

where

$$x_{S_i(t)}(t) = \begin{cases} 0, \ \textit{If PUs signal is ON on the channel } S_i(t), \\ B(t_s - t_{cm} - t_l) \ o.w., \end{cases} \tag{3}$$

is the throughput without channel switching costs, and

$$c(t) = Bt_{cs}, \tag{4}$$

denotes the lost throughput due to the channel switching latency if PUs signal is OFF and CR_i is transmitting on a different channel at time *t* than the one at time $t - 1$, and $\mathbb{1}_{\{A\}}$ denotes the indicator of event *A*. For mathematical brevity, we assume the normalized value of $x_{S_i(t)}(t) \in [0, 1]$ and $c(t) \in [0, 1]$, and define vector $\mathrm{X}(t) \in [0, 1]^K$ with the j^{th} channel, $x_j(t)$, to denote the throughput for channel *j*. The CBS scheduler accumulates the throughput $\sum_{i=1}^{s} Y_{CR_i}(t)$, at time slot *t*. The objective is to maximize the CRNs total throughput over time by finding the best slate and assigning them to the CRs for data transmission.

Let ν be the online learning policy which the CBS employs for *s*-set channel selection over time. Then, the expected accumulated throughput in the CRN after *T* time slots is

$$G_\nu(T) := \mathbb{E}_\nu \left[\sum_{t=1}^{T} \sum_{i=1}^{s} x_{S_i(t)}(t) - \sum_{t=1}^{T} \sum_{i=1}^{s} c(t) \mathbb{1}_{\{S_i(t) \neq S_i(t-1)\}} \right]. \tag{5}$$

where the expectation is taken with respect to the internal randomness of the policy ν.

We evaluate the performance of the proposed policy with respect to the *best slate* in hindsight which has the highest accumulated throughput up to time *T*. Assuming a genie with full prior knowledge, the optimal static policy then is the one that CBS persistently applies to select the best slate over the

Algorithm 1 CBS Scheduling Without Channel Switching Costs (s-Set Semi-Bandits Feedback Without Switching Costs)

Parameters: $\gamma_i \in (0,1]$, $\eta_i \in (0,1], i = 1, \ldots, s$.

Initialization: $w_j^{(i)}(1) = 1$, for all $i \in [s]$, and $j \in [K]$.

1: **while** $t \leq T$ **do**
2: Set $\mathcal{M}_i = \emptyset$ for all $i \in [s]$.
3: **for** $i = 1, \ldots, s$ **do**
4: Set $p_j^{(i)}(t) = (1-\gamma_s)\frac{w_j^{(i)}(t)\mathbb{1}_{j\notin\mathcal{M}_i}}{\sum_{r=1}^{K} w_r^{(i)}(t)\mathbb{1}_{r\notin\mathcal{M}_i}} + \frac{\gamma_i}{K-i+1}\mathbb{1}_{j\notin\mathcal{M}_i}$, for all $j \in [K]$.
5: Sample $S_i(t) \sim p^{(i)}(t) = (p_1^{(i)}(t), \ldots, p_K^{(i)}(t))$, and form $\mathcal{M}_{i+1} = \mathcal{M}_i \cup \{S_i(t)\}$.
6: **end for**
7: Play the slate $S(t)$, and receive the reward $x_{S_i(t)}(t) \in [0,1]$, for all $i \in [s]$.
8: Set $\hat{x}_j(t) = \frac{x_j(t)}{p_j^{(i)}(t)\prod_{r=1}^{i-1} 1-p_j^{(r)}(t)}\mathbb{1}_{S_i(t)=j}\mathbb{1}_{j\notin\mathcal{M}_i}$, for all $i \in [s]$, and $j \in [K]$.
9: Update $\omega_j^{(i)}(t+1) = \omega_j^{(i)}(t)\exp(\eta_i \hat{x}_j(t))$, for all $i \in [s]$, and $j \in [K]$.
10: $t = t + 1$.
11: **end while**

time. Then, the maximum accumulated throughput on the best slate is defined as follows:

$$G_{max}(T) \coloneqq \max_{\substack{S\subset[K]\\|S|=s}} \sum_{t=1}^{T}\sum_{i=1}^{s} x_{S_i}(t). \tag{6}$$

We measure the performance of the learning policy ν with the notion of *regret* which is the performance difference between the proposed policy and the optimal static policy in hindsight [21]. In other words, the regret measures the gap between the accumulated throughput achieved by applying a learning policy and the maximum accumulated throughput the CBS can obtain when it keeps playing on the best slate. Our goal is to minimize the *regret of the s-set semi-bandit with switching costs after T rounds* defined as follows:

$$\min_{\nu}\ R(T) \coloneqq G_{max}(T) - G_{\nu}(T). \tag{7}$$

In the next section, we present a family of multi-armed bandit algorithms that generate the order-optimal policy for CBS scheduling with switching costs and show they achieve *sub-linear* throughput regret upper bound over time. That is, the proposed solution performs no worse than the optimal static policy on average, asymptotically.

IV. Online Multiuser Cognitive Base Station Scheduling Policy

In this section, we design efficient online learning algorithms for CBS scheduling which strikes an optimal balance between searching for spectrum holes for data transmission and channel switching costs to maximize the CRNs' throughput. We illustrate our algorithm design in three key steps. First, in Section IV-A, we focus on the special case with no channel switching costs $c(t) = 0$ (i.e., no channel switching latency $t_{cs} = 0$) for all t in (5), recovering the multi-armed semi-bandit problem studied by [57], [58]. Next, in Section IV-B1, we study the special case of the problem in (7), for a single CR, i.e., $s = 1$, thereby recovering the well-studied problem of multi-armed bandits with switching costs [44], [45]. Finally, in Section IV-B2, we combine the key algorithmic ideas from Sections IV-A and IV-B1, to design an order-optimal online CBS scheduling algorithm for minimizing the regret in (7). All proofs are deferred to the Appendix.

A. CBS Scheduling Without Channel Switching Costs

We focus on the CBS scheduling problem where we assume that the CRs' channel switching latency is negligible. In this case, considering that the CBS is serving s number of CRs, the problem of throughput maximization is modeled with s-set semi-bandit feedback and zero switching costs, i.e., $c(t) = 0$. The objective of the CBS scheduler is to choose a slate of s channels out of K and assign them to the CRs such that the regret in (7) is minimized.

Considering no prior knowledge on the PUs activity, we propose a non-stochastic online learning algorithm which is a modified version of EXP3 algorithm [21]. The proposed CBS scheduling algorithm maintains a distribution over K channels for each of the CRs in the i^{th} position on the set C for $i = 1, \ldots, s$. Then, at each time slot t, the CBS scheduler fills the channel slate using the following sequential sampling. Starting at position $i = 1$ in the set C, the scheduler first samples the channel to be assigned to the CR which has first position in the slate according to $S_1(t) \sim p^{(i)}(t)$. Then for $i = 2$, we exclude the channel $S_1(t)$ from the support of $p^{(2)}(t)$ and re-normalize $p^{(2)}(t)$ before sampling $S_2(t) \sim p^{(2)}(t)$. Proceeding in this manner, for the CR in the i^{th} position in the the set C, we restrict the support of $p^{(i)}(t)$ to $[K]\backslash\{S_1(t), \ldots, S_{i-1}(t)\}$, re-normalize, and sample $S_i(t) \sim p^{(i)}(t)$. After assigning the slate $S(t)$, the CBS receives the report from the CRs about the throughput for each channel assignment, i.e., s-set semi-bandit feedback, and updates the weight for each $p^{(i)}(t)$. The pseudocode is presented in Algorithm 1.

We now show that the estimated throughput $\hat{x}_j(t)$ by the CBS scheduler (line 8 in Algorithm 1), for any CR_i in the set C is unbiased. Let $\mathcal{M}_i = \{S_1(t), \ldots, S_{i-1}(t)\}$, and $\mathcal{M}_1 = \emptyset$. Let also

$$\mathbb{1}_{j\notin\mathcal{M}_i} = \begin{cases} 1, & \text{with probability } \prod_{r=1}^{i-1} 1 - p_j^{(r)}(t), \\ 0, & \text{with probability } 1 - \prod_{r=1}^{i-1} 1 - p_j^{(r)}(t), \end{cases} \tag{8}$$

where $\mathbb{1}_{j\notin\mathcal{M}_i}$ indicates a random variable defined for CR_i in the slate on each channel *j* which takes value 1 with probability that this channel is not selected to be assigned to neither of its preceding CRs. Then we denote $\mathbb{E}_{j\notin\mathcal{M}_i}[\hat{x}_j(t)]$ as the expectation that channel $S_i(t)$ is not chosen to be assigned to neither of the CRs in position 1 to $i-1$ at time *t* as follows:

$$\begin{aligned}\mathbb{E}_{j\notin\mathcal{M}_i}\big[\hat{x}_j(t)\big] &= \left(\prod_{r=1}^{i-1}\Big(1-p_j^{(r)}(t)\Big)\right)\\ &\quad\cdot\frac{x_j(t)}{p_j^{(i)}(t)\prod_{r=1}^{i-1}1-p_j^{(r)}(t)}\mathbb{1}_{S_i(t)=j}\\ &\quad+\left(1-\prod_{r=1}^{i-1}\Big(1-p_j^{(r)}(t)\Big)\right)\\ &\quad\cdot 0=\frac{x_j(t)}{p_j^{(i)}(t)}\mathbb{1}_{S_i(t)=j}. \end{aligned}\tag{9}$$

Then by taking the expectation w.r.t. to the randomness of channel $S_i(t)$, we show that the estimated throughput $\hat{x}_j(t)$ is unbiased:

$$\begin{aligned}\mathbb{E}_{S_i(t)\sim p^{(i)}(t)}\big[\mathbb{E}_{j\notin\mathcal{M}_i}\big[\hat{x}_j(t)\big]\big] &= \sum_{\substack{r=1\\ r\notin\mathcal{M}_i}}^{K} p_r^{(i)}(t)\frac{x_j(t)}{p_j^{(i)}(t)}\mathbb{1}_{r=j}\\ &= x_j(t),\quad j\notin\mathcal{M}_i. \end{aligned}\tag{10}$$

Our main result is as follows.

Theorem 1: For any $K>s>0$, with exploration parameter of $\gamma_i=\sqrt{\frac{(K-i+1)\ln(K-i+1)}{T}}$, and learning rate $\eta_i=\sqrt{\frac{\ln(K-i+1)}{(e-2)(K-i+1)T}}$, for $i=1,\ldots,s$, the regret of CBS scheduler without channel switching cost presented in Algorithm 1, after $T\geq K\ln K$ time slots, is bounded as

$$\mathbb{E}[R(T)]\leq 2.7\sum_{i=1}^{s}\sqrt{(K-i+1)T\ln(K-i+1)},$$

where expectation is with respect to the internal randomization of the algorithm.

Proof: See the Appendix. ∎

A few remarks are in order. First, the bound above is order-optimal due to the lower bound of $\Omega(T^{1/2})$ in [24]. Second, the bound is only worse by a factor of $\sqrt{s}$ due to [23], but it slightly improves the regret upon the known result of $O(s\sqrt{KT\ln K})$ in [58]. The key for achieving this is in restricting the exploration set to the size of $K-i+1$ for each position $i\in[s]$ using sequential sampling. Third, the proposed algorithm fares favorably in terms of the computational complexity, requiring $O(sK)$ computation per iteration. Forth, it is noted that the proposed spectrum sharing scheme for the centralized CRNs achieves the regret bound in the order of $T^{1/2}$. However, if a decentralized setting is considered then the very recent work by Bubeck *et al.* [61] on distributed online learning algorithm can be adopted where it obtains the regret bound in the order of $T^{(1-\frac{1}{2s})}$ (*s* is the number of CRs). As we can see, in this case the order of regret depends on the number of CRs, and as the number of CRs increase the regret order increases as well. However, the regret order is a constant of 1/2 for the centralized setting using s-set semi-bandits online learning framework.

B. CBS Scheduling With Channel Switching Costs

In this subsection, we build on the algorithmic idea from the previous subsection as well as a stochastic policy to handle the problem of CBS scheduling with channel switching costs. In particular, we use a probabilistic switching policy as a wrapper function around Algorithm 1 to form the *s*-set bandit feedback with switching costs framework which is presented in Algorithm 2.

Our randomized switching policy $\sigma(t)$ follows a stochastic Bernoulli process as

$$\sigma(t)=\begin{cases}\text{Switch}, & \text{with probability } \delta(t),\\ \text{No switch}, & \text{with probability } 1-\delta(t).\end{cases}\tag{11}$$

Based on this policy, at each time slot the CBS switches the entire slate with probability $\delta(t)$ (line 5-11 in Algorithm 2), in which case it runs the sequential sampling procedure of Algorithm 1, or it decides to play the same slate as the previous time slot with probability $1-\delta(t)$ (line 12-16 in Algorithm 2). The probability to switch, $\delta(t)$, depends on the number of channels *K*, and decays with time as $t^{-\alpha}$. The choice of α is crucial – a slow decaying $\delta(t)$ would allow frequent channel switching and help with exploration, but at the expense of potentially not exploiting a high throughput channel and incurring additional lost throughput due to switching. On the other hand, a fast decaying $\delta(t)$ may hurt exploration and, therefore, overall achievable CRNs throughput. We show that $\alpha=1/3$, i.e., $\delta(t)$ proportional to $t^{-1/3}$, offers a good trade-off between exploration and exploitation yielding order-optimal regret upper bound. Note, though, that the switching probability cannot exceed $1-(K\ln K/T)^{1/3}$. We find $\delta(t)=\min\{1-\epsilon,\sqrt[3]{\frac{K\ln K}{t}}\}$, where $\epsilon=\sqrt[3]{\frac{K\ln K}{T}}$. Then, the CBS scheduler maintains a distribution $p^{(i)}(t)\in\Delta^{K-1}:=\{p\in[0,1]^K:\sum_{j=1}^{K}p_j^{(i)}=1\}$, $i\in[s]$, and at every time slot that it decides to switch, samples $S^{(i)}(t)\sim p^{(i)}(t)$. The distribution $p^{(i)}$ depends on the throughput obtained and involves mixing exploration proportional to a certain parameter $\gamma_i>0$. The only other difference in Algorithm 2 is the way in which the CBS constructs an unbiased estimator of the throughput. At time slot *t*, the achieved throughput for the j^{th} channel, i.e., $x_j(t)$, is scaled by switching probability and the probability of selecting the channel *j* in proceeding position, $1,\ldots,i-1$.

Next, to better understand the regret analysis of the proposed algorithm, we first consider the setting of scheduling for a single CR ($s=1$) with channel switching costs. Then, we extend the analysis for the setting of multiple CRs ($s>1$) and give the main results of upper bound regret for Algorithm 2.

1) CBS Scheduling for a Single CR With Channel Switching Costs: In Algorithm 2 we set $s=1$ and then show that the estimated throughput $\hat{x}_j(t)$ (line 11 and 16 in Algorithm 2) is unbiased. We first take the expectation w.r.t. the randomness of switching as follows:

$$\begin{aligned}\mathbb{E}_{\sigma(t)\sim\delta(t)}\big[\hat{x}_j(t)\big] &= \frac{x_j(t)}{2\delta(t)p_j(t)}\mathbb{1}_{S(t)=j}\delta(t)\\ &\quad+\frac{x_j(t)}{2(1-\delta(t))p_j(t)}\mathbb{1}_{S(t)=j}(1-\delta(t))\\ &= \frac{x_j(t)}{p_j(t)}\mathbb{1}_{S(t)=j}. \end{aligned}\tag{12}$$

Algorithm 2 CBS Scheduling With Channel Switching Costs (s-Set Semi-Bandits Feedback With Switching Costs)

Parameters: $\gamma_i \in \left(\sqrt[3]{\frac{K\ln K}{T}}, 1\right]$, $\eta_i \in \left(0, \frac{1}{2^{i-2}(K-i+1)}\sqrt[3]{\frac{(K\ln K)^{i+1}}{T^{i+1}}}\right)$, $\epsilon = \sqrt[3]{\frac{K\ln K}{T}}$.

Initialization: $w_j^{(i)}(1) = 1$, $p_j^{(i)}(0) = \frac{1}{K}$, $S_i(0) \sim p^{(i)}(0) = (p_1^{(i)}(0), \ldots, p_K^{(i)}(0))$, for all $i \in [s]$, and $j \in [K]$.

while $t \leq T$ **do**
  Set $\mathcal{M}_i = \emptyset$ for all $i \in [s]$.
  Set $\delta(t) = \min\left\{1-\epsilon, \sqrt[3]{\frac{K\ln K}{t}}\right\}$.
  Draw $u \sim \mathcal{U}[0,1]$.
  if $\delta(t) \geq u$ **then** {\\Switch}
    for $i = 1, \ldots, s$ **do**
      Set $p_j^{(i)}(t) = (1-\gamma_s)\frac{w_j^{(i)}(t)\mathbb{1}_{j\notin\mathcal{M}_i}}{\sum_{r=1}^{K} w_r^{(i)}(t)\mathbb{1}_{r\notin\mathcal{M}_i}} + \frac{\gamma_s}{K-s+1}\mathbb{1}_{j\notin\mathcal{M}_i}$, for all $j \in [K]$.
      Sample $S_i(t) \sim p^{(i)}(t) = (p_1^{(i)}(t), ..., p_K^{(i)}(t))$, and form $\mathcal{M}_{i+1} = \mathcal{M}_i \cup \{S_i(t)\}$.
    end for
    Play the slate $S(t)$, and receive the reward $x_{S_i(t)}(t) \in [0,1]$ for all $i \in [s]$.
    Set $\hat{x}_j(t) = \frac{x_j(t)}{2\delta(t)p_j^{(i)}(t)\prod_{r=1}^{i-1}1-p_r^{(i)}(t)}\mathbb{1}_{S_i(t)=j}\mathbb{1}_{j\notin\mathcal{M}_i}$, for all $i \in [s]$, and $j \in [K]$.
  else {\\No Switch}
    Set $p_j^{(i)}(t) = p_j^{(i)}(t-1)$, for all $j \in [K]$ and $i \in [s]$.
    Set $S_i(t) = S_i(t-1)$, for all $i \in [s]$.
    Play the slate $S(t)$, and receive the reward $x_{S_i(t)}(t) \in [0,1]$, for all $i \in [s]$.
    Set $\hat{x}_j(t) = \frac{x_j(t)}{2(1-\delta(t))p_j^{(i)}(t)\prod_{r=1}^{i-1}1-p_r^{(i)}(t)}\mathbb{1}_{S_i(t)=j}\mathbb{1}_{j\notin\mathcal{M}_i}$, for all $i \in [s]$ and $j \in [K]$.
  end if
  Update $\omega_j^{(i)}(t+1) = \omega_j^{(i)}(t)\exp(\eta_i\hat{x}_j(t))$, for all $i \in [s]$ and $j \in [K]$.
  $t = t+1$.
end while

Then by taking the expectation w.r.t. to the randomness of channel $S(t)$, we have

$$\mathbb{E}_{S(t)\sim p(t)}\left[\mathbb{E}_{\sigma(t)\sim\delta(t)}\left[\hat{x}_j(t)\right]\right] = \sum_{r=1}^{K} p_r(t)\frac{x_j(t)}{p_j(t)}\mathbb{1}_{r=j} = x_j(t), \tag{13}$$

which shows that estimated throughput $\hat{x}_j(t)$ is unbiased.

Theorem 2: For any $K \geq 2$, $s = 1$, $\eta \leq \frac{2}{K}\sqrt[3]{\frac{(K\ln K)^2}{T^2}}$, and $\gamma \in (\sqrt[3]{\frac{K\ln K}{T}}, 1]$, the regret of the CBS scheduler for a single CR with channel switching costs of $c(t)$ is bounded as

$$\mathbb{E}[R(T)] \leq \gamma T + \frac{(e-2)K\eta T^{\frac{4}{3}}}{16} \times \left(\frac{7}{(K\ln K)^{\frac{1}{3}}} + \frac{K\ln K}{\left(T^{\frac{1}{3}} - (K\ln K)^{\frac{1}{3}}\right)^4}\right) + \frac{3}{2}(K\ln K)^{\frac{1}{3}}T^{\frac{2}{3}} + \frac{\ln K}{\eta},$$

where expectation is with respect to the internal randomization of the algorithm, and $T \geq 8K\ln K$.

Proof: See the Appendix. ∎

For an optimal choice of the learning rate, η, we obtain the following bound.

Corollary 1: For $\gamma = \sqrt[3]{\frac{K\ln K}{T}}$ and $\eta = \frac{4}{T^{2/3}}\sqrt{\frac{\ln K}{(e-2)K}}\left(\frac{7}{(K\ln K)^{\frac{1}{3}}} + \frac{K\ln K}{(T^{\frac{1}{3}}-(K\ln K)^{\frac{1}{3}})^4}\right)^{-1/2}$ we have

$$\mathbb{E}[R(T)] \leq 3.62(K\ln K)^{1/3}T^{2/3},$$

for $T \geq 8K\ln K$.

Proof: See the Appendix. ∎

A few remarks are in order. First, note that [44] shows a lower bound for multi-armed bandits with switching costs in the order of $\Omega(T^{2/3})$, so the bound in Corollary 1 is order-optimal. Second, the algorithm of [45] achieves the same regret bound but it involves mini-batching over epochs of size $T^{1/3}$. Our algorithm is a rather simple, and much easier to understand and implement as it admits standard analysis techniques. Also, the mechanism for ensuring that the CBS avoids switching too often is fundamentally different as it relies on a learning policy whose exploration rate diminishes over time, whereas [45] resorts to playing a constant arm in epochs of fixed size.

2) CBS Scheduling for Multiple CRs With Channel Switching Costs: We now consider the setting in which the CBS is serving multiple CRs which results in slate size of $s > 1$. We show the following regret bound for Algorithm 2.

Theorem 3: For any $K > s > 0$, with exploration parameter of $\gamma_i = (\frac{K\ln K}{T})^{1/3}$, channel switching cost of $c(t)$, and learning rates $\eta_i = \frac{4}{T^{2/3}}\sqrt{\frac{\ln(K-i+1)}{(e-2)(K-i+1)}}\left(\frac{7}{(K\ln K)^{\frac{1}{3}}} +\right.$

$\frac{K \ln K}{(T^{\frac{1}{3}}-(K \ln K)^{\frac{1}{3}})^4})^{-1/2}$, for $i = 1, \ldots, s$, the regret of the CBS scheduling algorithm presented in Algorithm 2 after $T \geq 8K \ln K$ time slots is bounded as

$$\mathbb{E}[R(T)] \leq 3.62s(K \ln K)^{1/3} T^{2/3},$$

where the expectation is with respect to the internal randomization of the algorithm.

Proof: See the Appendix. ∎

Note that the bound above is order-optimal due to the lower bound of $\Omega(T^{2/3})$ in the case of multi-armed bandits with switching costs [44]. To the best of our knowledge, the above regret bound is the first which we derive for the *s*-semi bandit with switching costs setting to address the CBS scheduling problem in centralized CRNs.

V. Performance Evaluations

In this section, we validate our theoretical findings by numerically evaluating the performance of the proposed CBS scheduling algorithms and measuring the regret in various CRNs with different settings. Our evaluation is conducted over a spectrum in which the PUs activity are created by computer simulations (synthetic data), as well as a spectrum in which the PUs activity is collected from a set of real-world spectrum data measurements.

A. CBS Scheduling on Synthetic Spectrum Data

Since we have considered no statistical assumptions on the PUs activity, we first create and validate a non-stochastic PUs activity on the frequency channels. Then, we run the proposed algorithms on the constructed non-stochastic environment, and compare the analytical and simulation results, as well as the baseline solutions, where available.

1) Non-Stochastic PU Activity Environment Setup: We simulate a stochastically constrained non-stochastic environment by adopting the approach of [52] to create the PUs. This method has been demonstrably effective in testing non-stochastic online learning algorithms via extensive experiments [62]. We adopt the framework nearly as is except that we generate normalized throughput (i.e., reward) in [0, 1] instead of $[-1, +1]$. This difference also changes the mean of throughput distribution on the channels. We describe the non-stochastic environment setup in detail as follows. Given the total number of time slots *T*, we split it into *n* consecutive (odd and even) phases as follows:

$$\underbrace{1, \ldots, t_1}_{T_1}, \underbrace{t_1+1, \ldots, t_2}_{T_2}, \ldots, \underbrace{t_n+1, \ldots, T}_{T_n}, \tag{14}$$

where $T_r = \lfloor 1.6^r \rfloor$, for $r = 1, \ldots, n$, is increasing exponentially with *r*. Let $\mu_j(t)$ denote the average throughput for data transmission on channel *j* at time slot *t* for the odd and even phases as follows:

$$\text{In odd phases:} \Rightarrow \mu_i(t) = \begin{cases} 1, & if \quad j \leq s, \\ 1-\Delta, & o.w., \end{cases} \tag{15}$$

$$\text{In even phases:} \Rightarrow \mu_i(t) = \begin{cases} \Delta, & if \ j \leq s, \\ 0, & o.w., \end{cases} \tag{16}$$

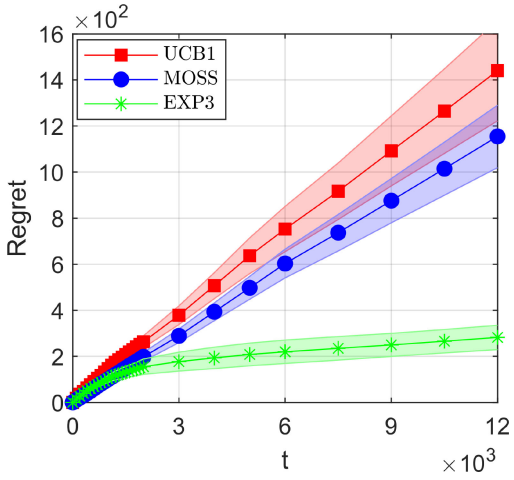


Fig. 3. Non-stochastic PUs activity verification.

where $\Delta = 1/K$ represents the mean throughput gap. The above setting means that for the channels $j \leq s$, we switch between $\mu_j(t) = 1$ and $\mu_j(t) = \Delta$ over consecutive phases while keeping the means fixed during a phase. Similarly, for every other channels, we switch between $\mu_j(t) = 1 - \Delta$ and $\mu_j(t) = 0$. Then, at time slot *t*, we generate a random vector $\mathrm{X}(t) \in [0, 1]^K$ to indicate the PUs activity on the frequency channels as follows: With probability $\mu_j(t)$ the PUs signal is OFF (i.e., setting $x_j(t)$ equal to 1), and with probability $1 - \mu_j(t)$ the PUs signal is ON (i.e., setting $x_j(t)$ equal to 0).

For our first set of experiments, we set $K = 10$ and $s = 1$. We set the CBS to run two well-known stochastic algorithms for channel scheduling, UCB1 [38] and MOSS [43], and the popular non-stochastic algorithm EXP3 [21] on the simulated spectrum with non-stochastic PUs activity. We set the time horizon to $T = 12 \times 10^3$ time slots, and average over 1,000 random trials.

Fig. 3 illustrates the empirical regret of the CBS running the three algorithms, with the shaded areas representing the two standard deviation of the empirical expected regret. Based on the plots, we can see that the algorithms designed for a stochastic settings, i.e., UCB1 and MOSS, exhibit a nearly-linear regret, failing in the non-stochastic environment, whereas EXP3 achieves a sublinear regret. This confirms the non-stochastic PUs activity on the simulated synthetic spectrum data.

2) CBS Scheduling on the Spectrum With Non-Stochastic PUs Activity: We now evaluate the proposed CBS scheduling algorithms empirically on the synthetic spectrum data created by the method described in the previous subsection. First we seek to evaluate the performance of the proposed Algorithm 1, for CBS scheduling without channel switching costs ($c(t) = 0$). We consider three different settings for the CRN with various number of frequency channels *K* and CRs *s*. The CBS then runs the Algorithm 1 for channel scheduling and computes the total CRN throughput. The plot in Fig. 4 shows the empirical regret of the proposed Algorithm 1 along with the theoretical upper bound from Theorem 1. The simulation results are consistent with the analytical results in obtaining a sublinear regret upper bound, that is, the CBS

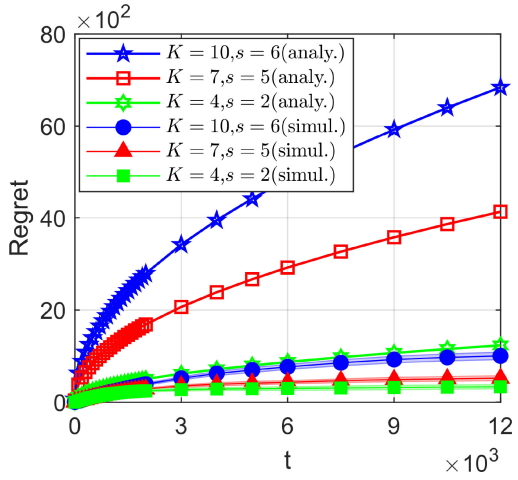


Fig. 4. Algorithm 1 (analytical and simulation comparison).

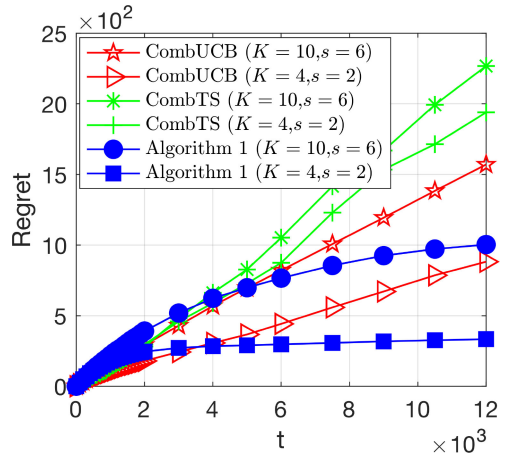


Fig. 5. Comparison of the proposed Algorithm 1 with the baseline solutions.

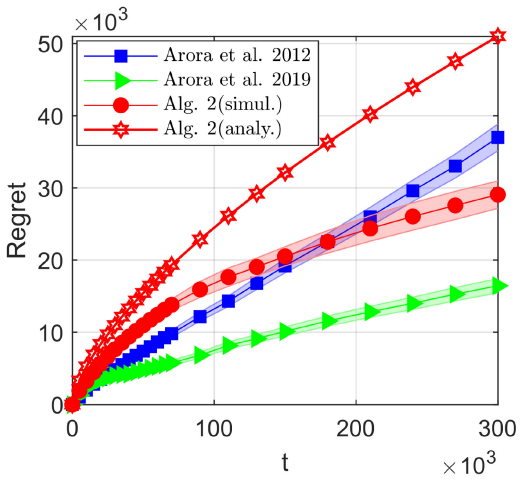


Fig. 6. Comparison of the proposed Algorithm 2 with the baseline solutions for $K = 10$ and $s = 1$.

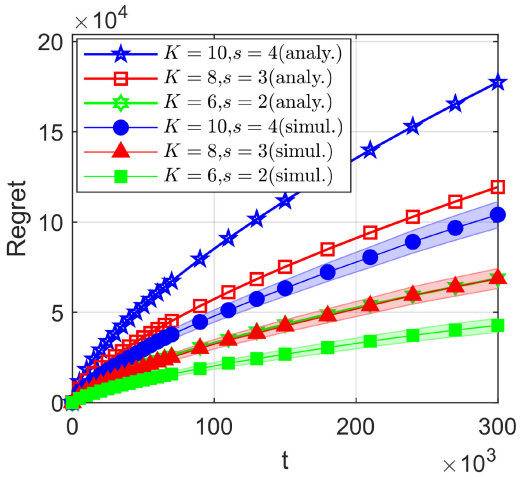


Fig. 7. Algorithm 2 with different K and s.

asymptotically converges to select the best slate to serve the CRs and maximize the CRNs total throughput.

We also empirically compare the performance of the proposed combinatorial semi-bandit algorithm (Algorithm 1) against two baseline algorithms: combinatorial UCB (CombUCB) [63], [64] and combinatorial Thompson Sampling (CombTS) [65]. We first set the setting of the non-stochastic PUs activity environment created by the synthetic simulations in the previous subsection, and then run the algorithms in the CBS to schedule the channels for the CRs. Fig. 5 illustrates the simulation results of the regret upper bound for the CBS scheduling when the above three algorithms run over various numbers of frequency channels K and CRs s. We observe that the proposed non-stochastic algorithm achieves sublinear regret while the both CombUCB and CombTS fail to converge to the best s-set channels to serve the CRs and admit a relatively linear regret upper bound.

We then evaluate the performance of the proposed Algorithm 2, for handling channel switching costs on a single CR, i.e., $s = 1$, by comparing its empirical regret against that of prior work which includes the fixed-size mini-batch algorithm of Arora *et al.* 2012 [45], and the adaptive mini-batching algorithm of Arora *et al.* 2019 [50]. We set $K = 10$, and channel switching cost $c(t) = 1$. Fig. 6 shows that our proposed switching policy compares favorably with other algorithms. Finally, we evaluate the performance of Algorithm 2 for CBS scheduling with channel switching costs. Fig. 7 plots the observed empirical regret and compares it with theoretical upper bound from Theorem 3. The results consistency and sublinear regret confirm the analysis. Also, as expected, the regret increases with the number of channels and CRs.

B. CBS Scheduling on Real-World Spectrum Measurement Data

In this subsection, we evaluate performance of the proposed CBS scheduling algorithms on the real-world spectrum measurement data. The data has been collected by a group of researchers in 5G-Xcast project in Turku city, Finland [36], using CRFS RFeye spectrum measurement receiver for a continuous 8-day period from January 20th to January 27th from four years, 2015 to 2018.[1]

From the five wide frequency bands provided in the spectrum dataset by [36], we pick three bands as: Band 1: 130-800

[1]Detailed information about the spectrum measurement setting, equipment, and data can be found in [37] and the following website: https://zenodo.org/record/1293283/files/Open%20Spectrum%20data.pdf?download=1.

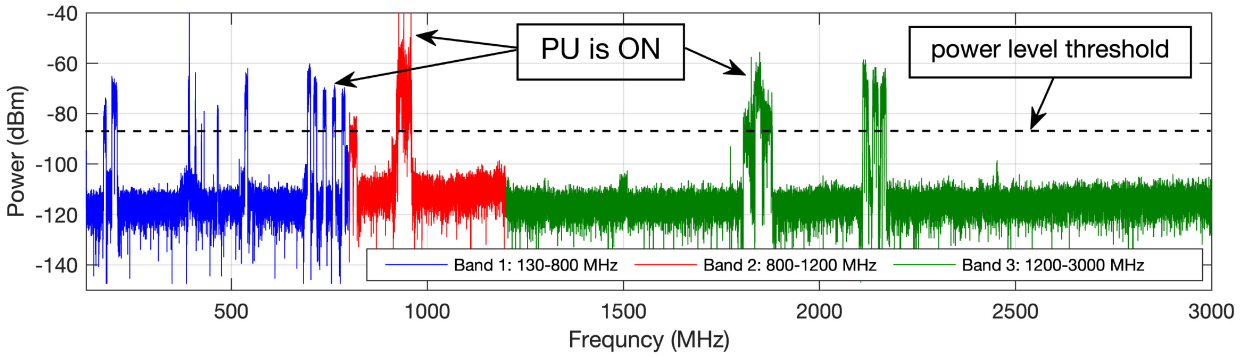


Fig. 8. A snapshot of spectrum measurement data from paging bands.

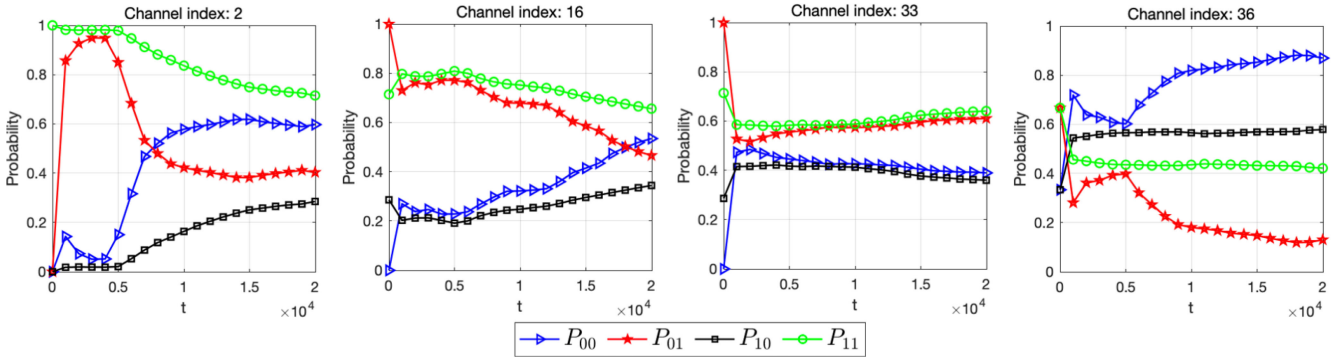


Fig. 9. Real-world PUs' state transition probability over the time in different channels.

MHz, Band 2: 800-1200 MHz, and Band 3: 1200-3000 MHz. Fig. 8 illustrates a sample of power measurement across the whole three bands. We consider that PUs signals are ON on a frequency channel if power measurement is higher than the threshold power level -90 dBm, and they are OFF if it is less than this threshold. For channel selection within the bands, we examine the PUs activity over the whole spectrum and randomly pick a set of $K = 36$, $K = 24$, and $K = 12$ frequency channels which are mostly occupied by the PUs over time. The frequency bandwidth of each channel is considered $W = 39.0625$ KHz, according to the dataset information.

To better understand the behavior of the PUs activity in practice, we examine the spectrum measurement data and extract the PUs activity over multiple frequency channels. We consider the two-state Markov Chain model for PUs activity, i.e., states "1" and "0" referring to ON and OFF state, respectively. Then, using maximum likelihood estimation we estimate the state transition probability of the PUs activities as $P_{00}, P_{01}, P_{10}, P_{11}$ where Pij denotes the transition probability from state i to state j, and $i, j \in \{0, 1\}$. We select 4 channels out of 36 channels across the frequency bands uniformly at random and show the PUs state transition probability in Fig. 9. We find that the transition probability is changing over the time, referring to the non-stationary behavior of the PUs activities and implying that the PUs are not behaving well and are not stable over the time to follow a specific and known stochastic model. This observation motivated us to search for non-stochastic learning algorithms which can tackle the practical scenarios.

We set the number of CRs to $s = 24$, $s = 15$ and $s = 8$, accordingly. Based on the measurement information, signal

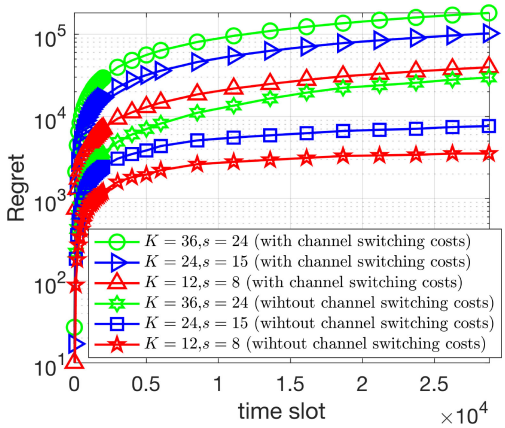


Fig. 10. Regret measurements of CBS scheduling for both with and without channel switching costs settings.

power level is measured every three seconds across the whole spectrum. This fixes the time slot length to $t_s = 3$ seconds for our experimental evaluation. Data transmission time and channel switching latency t_{cs} are considered to be 70% and 30% of the total time slot length while the control messaging duration and learning time, which are not the focus of this paper, are assumed to be negligible (i.e., $t_{cm} = 0$, $t_l = 0$). With an acceptable SNR if the channel assigned to a CR by the CBS was idle, then the CR obtains a throughput using (1); otherwise if the channel is busy, no throughput is acquired as the CR stays in sleep mode to not interfere with the PU's signal.

In Fig. 10, we show the regret upper bound over $T = 28,812$ time slots for the settings of CBS scheduling without and with

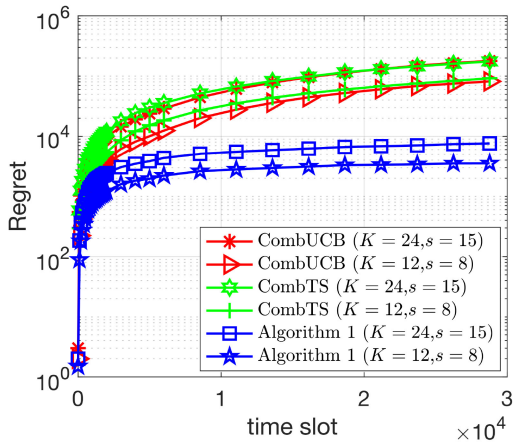


Fig. 11. Comparison of Algorithm 1 with the baseline solutions on the real-world spectrum measurement data.

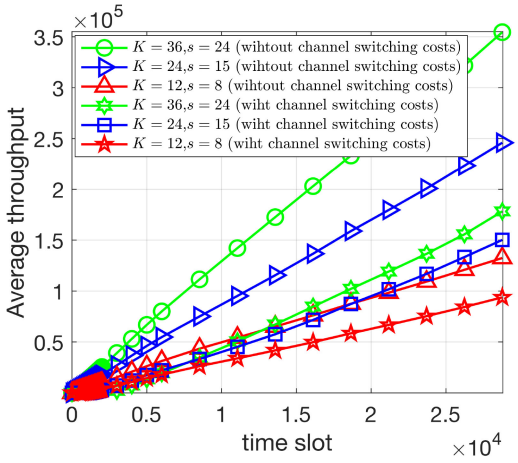


Fig. 12. Throughput measurement of CBS scheduling for both with and without channel switching costs settings.

switching costs (i.e., Algorithm 1 and Algorithm 2, respectively). For better comparison the regret is plotted in log scale. From the figure, we see that first, the regret results are sublinear. Second, scheduling with channel switching costs has higher regret compared to the case of without switching costs. This is consistent with the analytical regret order bound of their settings which are $T^{2/3}$ and $T^{1/2}$, respectively. Third, the regret increases as the number of channels and CRs increase.

We further compare the regret upper bound of the proposed CBS scheduler algorithm against the CombUCB [63], [64] and CombTS [65] algorithms over the real-world spectrum measurement data. From Fig. 11, we observe that the regret bound of the proposed algorithm significantly outperforms the state-of-the-art solutions (note that the regret is plotted in log scale, so the differences are noteworthy). The reason is that the real-world PUs activity, as shown in Fig. 9, does not follow a stationary stochastic process. However, the proposed non-stochastic algorithm captures the PUs dynamics and maximizes the network throughput by achieving sublinear regret upper bound.

Fig. 12 also shows the average accumulated throughput by the CBS over time. Again, as expected CBS accumulates more

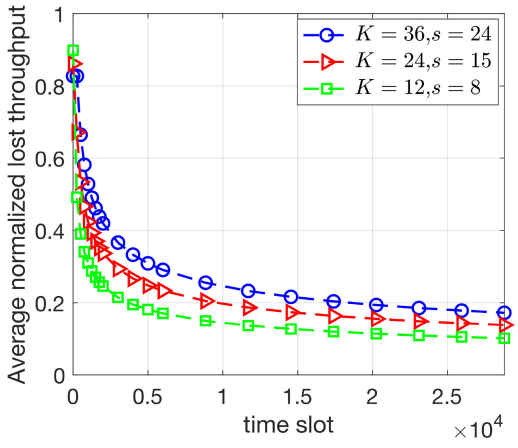


Fig. 13. Lost throughput due to channel switching latency.

throughput in the case of without switching costs. Fig. 13 also illustrates the average normalized lost throughput due to channel switching latency for each network. We observe that the lost throughput decreases over time. This is because as the CBS learns the best slate, the frequency of channel switching is reduced.

VI. Discussion on Future Work

- We investigated the throughput maximization problem in centralized CRNs using optimal CBS scheduling. We considered the switching costs in terms of lost throughput, however, future work can study joint throughput loss and energy consumption minimization as channel switching consumes comparable energy in RF front-end circuits [16], [20].
- We assumed fixed switching costs between any pair of actions (i.e., frequency channels). This is similar to the assumptions made in the literature [44], [45], [50]. However, switching latency may depend on frequency separation distance between the pairs. Koren *et al.* [66] proposed a new metric called *movement costs* wherein the switching cost is linearly proportional to the arm index differences between the pair of actions. We believe the future work on CBS scheduling problem can adopt the work by [66] to address the setting of different switching costs between each pair of frequency channels.
- We assumed the number of CRs in CBS coverage are smaller than the number of channels, i.e., $s < K$. The case of $s > K$ creates an interesting setting which involves decision on prioritizing the serving CRs per time slot, accounting the the traffic requirements of the CRs.
- In this paper, we omitted the impact of PUs and CRs location which yielded to the combinatorial problem formulation. When the users' location impact is integrated into the system model, it is formulated as an online permutation problem ([22, Sec. 5.3]) which requires a different setting and regret analysis. The future work can further investigate the integration of channel switching costs into the online permutation problem with the consideration of different frequency separation costs between the pairs of channels.

- Decentralized CRNs with online learning-based CRs is of great importance to be investigated. This setting involves channel access collision among the CRs. The recent work by Bubeck *et al.* [61] can tackle both the scenarios in which the CRs either communicate the collision information with each other or not.

VII. Conclusion

We modeled the cognitive base station (CBS) scheduling problem in centralized cognitive radio networks (CRNs) with non-stochastic multi-armed bandits and semi-bandit feedback with channel switching costs. We proposed two novel online learning algorithms with and without accounting the channel switching costs, where we proved the order-optimal regret upper bound of $T^{1/2}$ and $T^{2/3}$, respectively. By employing the proposed algorithms, the CBS achieved the maximum network throughput over time by learning the best slate of the frequency channels. Our algorithms are simple and intuitive and admit a much easier analysis than prior work. Our solution relaxed the assumptions on the prior knowledge of primary users' activity and their statistical model. We validated our theoretical findings with extensive experimental evaluation using synthetic and real-world spectrum measurement data.

Appendix

Proof of Theorem 1: We first verify the following equalities:

$$\mathbb{E}_{j\sim p^{(i)}(t)}\big[\mathbb{E}_{j\notin\mathcal{M}_i}\big[\hat{x}_j(t)\big]\big] = \sum_{\substack{j=1\\ j\notin\mathcal{M}_i}}^{K} p_j^{(i)}(t)\frac{x_j(t)}{p_j^{(i)}(t)}\mathbb{1}_{S_i(t)=j} = x_{S_i(t)}(t), \quad S_i(t)\notin\mathcal{M}_i, \tag{17}$$

and

$$\mathbb{E}_{j\notin\mathcal{M}_i}\Big[\hat{x}_j^2(t)\Big] = \frac{x_j^2(t)}{(p_j^{(i)}(t))^2\prod_{r=1}^{i-1}1-p_j^{(r)}(t)}\mathbb{1}_{S_i(t)=j}, \tag{18}$$

and for any $j\notin\mathcal{M}_i$,

$$\mathbb{E}_{S_i(t)\sim p^{(i)}(t)}\Big[\mathbb{E}_{j\notin\mathcal{M}_i}\Big[\hat{x}_j^2(t)\Big]\Big] = \frac{x_j^2(t)}{p_j^{(i)}(t)\prod_{r=1}^{i-1}1-p_j^{(r)}(t)}. \tag{19}$$

Let $W^{(i)}(t)=\sum_{j=1}^{K}\omega_j^{(i)}(t)\mathbb{1}_{j\notin\mathcal{M}_i}$, and $W^{(i)}(t+1)=\sum_{j=1}^{K}\omega_j^{(i)}(t+1)\mathbb{1}_{j\notin\mathcal{M}_i}$. Using these definitions, we have

$$\begin{aligned}\frac{W^{(i)}(t+1)}{W^{(i)}(t)} &= \sum_{j=1}^{K}\frac{\omega_j^{(i)}(t)\mathbb{1}_{j\notin\mathcal{M}_i}}{\sum_{r=1}^{K}\omega_r^{(i)}(t)\mathbb{1}_{r\notin\mathcal{M}_i}}\exp\big(\eta\hat{x}_j(t)\big)\\ &\le \sum_{j=1}^{K}\frac{\omega_j^{(i)}(t)\mathbb{1}_{j\notin\mathcal{M}_i}}{\sum_{r=1}^{K}\omega_r^{(i)}(t)\mathbb{1}_{r\notin\mathcal{M}_i}}\\ &\quad\times\Big(1+\eta_i\hat{x}_j(t)+(e-2)\eta_i^2\hat{x}_j^2(t)\Big).\end{aligned} \tag{20}$$

We then take the expectation from both sides of (20) in (21), as shown in bottom of the page where we used $\sum_{\substack{j=1\\ j\notin\mathcal{M}_i}}^{K}p_j^{(i)}=1$, $\sum_{\substack{j=1\\ j\notin\mathcal{M}_i}}^{K}1=K-i+1$, $\prod_{r=1}^{i-1}1-p_j^{(r)}(t)\le 1$, the definition of $p_j^{(i)}(t)$ in Algorithm 1, and the following facts:

$$\frac{\omega_j^{(i)}(t)\mathbb{1}_{j\notin\mathcal{M}_i}}{\sum_{r=1}^{K}\omega_r^{(i)}(t)\mathbb{1}_{r\notin\mathcal{M}_i}} = \begin{cases}\frac{p_j^{(i)}(t)-\frac{\gamma_i}{K-i+1}}{1-\gamma_i}, & \text{with probability } \prod_{r=1}^{i-1}1-p_j^{(r)}(t),\\ 0, & \text{with probability } 1-\prod_{r=1}^{i-1}1-p_j^{(r)}(t),\end{cases} \tag{22}$$

hence, for any $j\notin\mathcal{M}_i$,

$$\mathbb{E}_{j\notin\mathcal{M}_i}\left[\frac{\omega_j^{(i)}(t)\mathbb{1}_{j\notin\mathcal{M}_i}}{\sum_{r=1}^{K}\omega_r^{(i)}(t)\mathbb{1}_{r\notin\mathcal{M}_i}}\right] = \frac{p_j^{(i)}(t)-\frac{\gamma_i}{K-i+1}}{1-\gamma_i}\prod_{r=1}^{i-1}1 - p_j^{(r)}(t). \tag{23}$$

Then, by taking the logarithms from both sides of (21) and using $1+x\le e^x$, we get

$$\begin{aligned}&\ln\left[\mathbb{E}_{j\notin\mathcal{M}_i}\left[\frac{W^{(i)}(t+1)}{W^{(i)}(t)}\right]\right]\\ &\le\frac{\eta_i}{1-\gamma_i}\sum_{\substack{j=1\\ j\notin\mathcal{M}_i}}^{K}p_j^{(i)}(t)\mathbb{E}_{j\notin\mathcal{M}_i}\big[\hat{x}_j(t)\big]\\ &\quad+\frac{(e-2)\eta_i^2}{1-\gamma_i}\sum_{\substack{j=1\\ j\notin\mathcal{M}_i}}^{K}p_j^{(i)}(t)\mathbb{E}_{j\notin\mathcal{M}_i}\Big[\hat{x}_j^2(t)\Big]\prod_{r=1}^{i-1}1-p_j^{(r)}(t).\end{aligned} \tag{24}$$

$$\begin{aligned}\mathbb{E}_{j\notin\mathcal{M}_i}\left[\frac{W^{(i)}(t+1)}{W^{(i)}(t)}\right] &\le \sum_{j=1}^{K}\mathbb{E}_{j\notin\mathcal{M}_i}\left[\frac{\omega_j^{(i)}(t)\mathbb{1}_{j\notin\mathcal{M}_i}}{\sum_{r=1}^{K}\omega_r^{(i)}(t)\mathbb{1}_{r\notin\mathcal{M}_i}}\right]\Big(1+\eta_i\mathbb{E}_{j\notin\mathcal{M}_i}\big[\hat{x}_j(t)\big]+(e-2)\eta_i^2\mathbb{E}_{j\notin\mathcal{M}_i}\Big[\hat{x}_j^2(t)\Big]\Big)\\ &= \sum_{\substack{j=1\\ j\notin\mathcal{M}_i}}^{K}\frac{p_j^{(i)}(t)-\frac{\gamma_i}{K-i+1}}{1-\gamma_i}\prod_{r=1}^{i-1}1-p_j^{(r)}(t)\Big(1+\eta_i\mathbb{E}_{j\notin\mathcal{M}_i}\big[\hat{x}_j(t)\big]+(e-2)\eta_i^2\mathbb{E}_{j\notin\mathcal{M}_i}\Big[\hat{x}_j^2(t)\Big]\Big)\\ &\le 1+\frac{\eta_i}{1-\gamma_i}\sum_{\substack{j=1\\ j\notin\mathcal{M}_i}}^{K}p_j^{(i)}(t)\mathbb{E}_{j\notin\mathcal{M}_i}\big[\hat{x}_j(t)\big]+\frac{(e-2)\eta_i^2}{1-\gamma_i}\sum_{\substack{j=1\\ j\notin\mathcal{M}_i}}^{K}p_j^{(i)}(t)\mathbb{E}_{j\notin\mathcal{M}_i}\Big[\hat{x}_{i,j}^2(t)\Big]\prod_{r=1}^{i-1}1-p_j^{(r)}(t)\end{aligned} \tag{21}$$

Using Jensen's inequality and summing over t we then get

$$\mathbb{E}_{j\notin\mathcal{M}_i}\left[\ln\left[\frac{W^{(i)}(T+1)}{W_1^{(i)}(1)}\right]\right] \le \frac{\eta_i}{1-\gamma_i}\sum_{t=1}^{T}\sum_{\substack{j=1\\ j\notin\mathcal{M}_i}}^{K} p_j^{(i)}(t)\mathbb{E}_{j\notin\mathcal{M}_i}\left[\hat{x}_j(t)\right] + \frac{(e-2)\eta_i^2}{1-\gamma_i}\sum_{t=1}^{T}\sum_{\substack{j=1\\ j\notin\mathcal{M}_i}}^{K} p_j^{(i)}(t)\mathbb{E}_{j\notin\mathcal{M}_i}\left[\hat{x}_j^2(t)\right]\prod_{r=1}^{i-1}1-p_j^{(r)}(t). \tag{25}$$

For any $j\notin\mathcal{M}_i$,

$$\mathbb{E}_{j\notin\mathcal{M}_i}\left[\ln\left[\frac{W^{(i)}(T+1)}{W^{(i)}(1)}\right]\right] \ge \mathbb{E}_{j\notin\mathcal{M}_i}\left[\ln\left[\frac{\omega_j^{(i)}(T+1)}{W^{(i)}(1)}\right]\right] = \eta_i\sum_{t=1}^{T}\mathbb{E}_{j\notin\mathcal{M}_i}\left[\hat{x}_j(t)\right] - \ln(K-i+1), \tag{26}$$

where we used the initialization of $\omega_j^{(i)}(1)=1$, and the definition of $W^{(i)}(1)=\sum_{j=1}^{K}\omega_j^{(i)}(1)\mathbb{1}_{j\notin\mathcal{M}_i}=K-i+1$ as well as the definition of $\omega_j^{(i)}(t+1)$ in Algorithm 1. Combining (25) and (26), and some rearrangements we get

$$\sum_{t=1}^{T}\mathbb{E}_{j\notin\mathcal{M}_i}\left[\hat{x}_j(t)\right]-\sum_{t=1}^{T}\sum_{\substack{j=1\\ j\notin\mathcal{M}_i}}^{K} p_j^{(i)}(t)\mathbb{E}_{j\notin\mathcal{M}_i}\left[\hat{x}_j(t)\right] \le \gamma_i\sum_{t=1}^{T}\mathbb{E}_{j\notin\mathcal{M}_i}\left[\hat{x}_j(t)\right] + (e-2)\eta_i\sum_{t=1}^{T}\sum_{\substack{j=1\\ j\notin\mathcal{M}_i}}^{K} p_j^{(i)}(t)\mathbb{E}_{j\notin\mathcal{M}_i}\left[\hat{x}_j^2(t)\right]\prod_{r=1}^{i-1}1-p_j^{(r)}(t) + \frac{\ln(K-i+1)}{\eta_i}. \tag{27}$$

Now considering that $S_1(t), S_2(t), \ldots, S_{i-1}(t)$ are given, we take the expectation w.r.t. randomness of $S_i(t)$, and use the equalities in (9), (10), (17), (18), and (19), we then get

$$\sum_{t=1}^{T}x_{j_i^*}(t)-\sum_{t=1}^{T}\mathbb{E}_{S_i(t)\sim p^{(i)}(t)}\left[x_{S_i(t)}(t)\right] \le \gamma_i\sum_{t=1}^{T}x_{j_i^*}(t)+(e-2)\eta_i\sum_{t=1}^{T}\sum_{\substack{j=1\\ j\notin\mathcal{M}_i}}^{K}x_j(t)+\frac{\ln(K-i+1)}{\eta_i}. \tag{28}$$

Since $\sum_{t=1}^{T}\sum_{\substack{j=1\\ j\notin\mathcal{M}_i}}^{K}x_j(t)\le(K-i+1)T$, we obtain the expected regret of position i in the slate as follows:

$$\mathbb{E}[R_i(T)]\le\gamma_i T+(e-2)(K-i+1)\eta_i T+\frac{\ln(K-i+1)}{\eta_i}. \tag{29}$$

By getting the derivative from the above inequality w.r.t. η_i, we find the optimal value of learning rate $\eta_i=\sqrt{\frac{\ln(K-i+1)}{(e-2)(K-i+1)T}}$. We choose $\gamma_i=\sqrt{\frac{(K-i+1)\ln(K-i+1)}{T}}$ and find that $T\ge(K-i+1)\ln(K-i+1)$ satisfies both $\gamma_i\le 1$ and $\eta_i\le 1$. With these input parameters and summing the regret over i, we get the total regret upper bound in statement of the theorem.

Proof of Theorem 2: First, we verify the following equalities:

$$\mathbb{E}_{j\sim p(t)}\left[\mathbb{E}_{\sigma(t)\sim\delta(t)}\left[\hat{x}_j(t)\right]\right]=\sum_{j=1}^{K}p_j(t)\frac{x_j(t)}{p_j(t)}\mathbb{1}_{S(t)=j}=x_{S(t)}(t), \tag{30}$$

and

$$\mathbb{E}_{\sigma(t)\sim\delta(t)}\left[\hat{x}_j^2(t)\right]=\left(\frac{x_j(t)}{2\delta(t)p_j(t)}\mathbb{1}_{S(t)=j}\right)^2\delta(t)+\left(\frac{x_j(t)}{2(1-\delta(t))p_j(t)}\mathbb{1}_{S(t)=j}\right)^2(1-\delta(t))=\frac{x_j^2(t)}{4p_j^2(t)}\left(\frac{1}{\delta(t)}+\frac{1}{1-\delta(t)}\right)\mathbb{1}_{S(t)=j}, \tag{31}$$

and

$$\mathbb{E}_{S(t)\sim p(t)}\left[\mathbb{E}_{\sigma(t)\sim\delta(t)}\left[\hat{x}_j^2(t)\right]\right]=\sum_{r=1}^{K}p_r(t)\left(\frac{x_j^2(t)}{4p_j^2(t)}\left(\frac{1}{\delta(t)}+\frac{1}{1-\delta(t)}\right)\mathbb{1}_{r=j}\right)=\frac{x_j^2(t)}{4p_j(t)}\left(\frac{1}{\delta(t)}+\frac{1}{1-\delta(t)}\right). \tag{32}$$

In the following we compute the upper bound for the terms $\sum_{t=1}^{T}\frac{1}{\delta(t)}$, $\sum_{t=1}^{T}\frac{1}{1-\delta(t)}$, and $\sum_{t=1}^{T}\delta(t)$, which later we will use them in the derivation of the regret upper bound:

$$\sum_{t=1}^{T}\frac{1}{\delta(t)}=\sum_{t=1}^{\frac{K\ln K}{(1-\epsilon)^3}-1}\frac{1}{1-\epsilon}+\sum_{t=\frac{K\ln K}{(1-\epsilon)^3}}^{T}\sqrt[3]{\frac{t}{K\ln K}} \le\frac{K\ln K}{(1-\epsilon)^4}+\frac{1}{\sqrt[3]{K\ln K}}\int_{\frac{K\ln K}{(1-\epsilon)^3}}^{T+1}t^{\frac{1}{3}} \simeq\frac{K\ln K}{4(1-\epsilon)^4}+\frac{3\,T^{\frac{4}{3}}}{4(K\ln K)^{\frac{1}{3}}}. \tag{33}$$

$$\sum_{t=1}^{T}\frac{1}{1-\delta(t)}\le\sum_{t=1}^{T}\frac{1}{\epsilon}=\frac{T}{\epsilon}=\frac{T^{\frac{4}{3}}}{(\mathrm{K}\ln K)^{\frac{1}{3}}}. \tag{34}$$

$$\sum_{t=1}^{T}\delta(t)=\sum_{t=1}^{\frac{K\ln K}{(1-\epsilon)^3}}(1-\epsilon)+\sum_{t=\frac{K\ln K}{(1-\epsilon)^3}+1}^{T}\sqrt[3]{\frac{K\ln K}{t}} \le\frac{K\ln K}{(1-\epsilon)^2}+(K\ln K)^{\frac{1}{3}}\int_{t=\frac{K\ln K}{(1-\epsilon)^3}}^{T}t^{-\frac{1}{3}} =\frac{3}{2}(K\ln K)^{\frac{1}{3}}T^{\frac{2}{3}}-\frac{K\ln K}{2(1-\epsilon)^2} \le\frac{3}{2}(K\ln K)^{\frac{1}{3}}T^{\frac{2}{3}}. \tag{35}$$

In Algorithm 2 with $s = 1$, the regret at each time t is a random variable as follows:

$$r(t) = \begin{cases} x_{j^*}(t) - x_{S(t)}(t) + c(t), & \text{with probability } \delta(t), \\ x_{j^*}(t) - x_{S(t)}(t), & \text{with probability } 1-\delta(t). \end{cases} \tag{36}$$

The expected value of the regret w.r.t. the randomness of switching policy at time t is equal to

$$\begin{aligned} \mathbb{E}_{\sigma(t)\sim\delta(t)}[r(t)] &\le \Big(x_{j^*}(t) - x_{S(t)}(t) + c(t)\Big)\delta(t) \\ &\quad + \Big(x_{j^*}(t) - x_{S(t)}(t)\Big)(1-\delta(t)) \\ &\le x_{j^*}(t) - x_{S(t)}(t) + \delta(t), \end{aligned} \tag{37}$$

where we used $c(t) \le 1$. Hence, the expected value of the accumulated regret $R(T)$ w.r.t. the randomness of taken action $S(t)$ is

$$\begin{aligned} \mathbb{E}_{S(t)\sim p(t)}[R(T)] &= \mathbb{E}_{S(t)\sim p(t)}\left[\mathbb{E}_{\sigma(t)\sim\delta(t)}\left[\sum_{t=1}^{T} r(t)\right]\right] \\ &\le \sum_{t=1}^{T} x_{j^*}(t) - \sum_{t=1}^{T}\mathbb{E}_{S(t)\sim p(t)}\Big[x_{S(t)}(t)\Big] \\ &\quad + \sum_{t=1}^{T}\delta(t). \end{aligned} \tag{38}$$

In the above equation, the expected regret upper bound consists of three terms. We first derive the upper bound on the first two terms then add it with the bound of the third term which we have computed in (35). Below is the derivation of the upper bound for the first two terms in (38). For a single t,

$$\begin{aligned} \frac{W(t+1)}{W(t)} &= \sum_{j=1}^{K}\frac{\omega_j(t)}{W(t)}\exp\big(\eta\hat{x}_j(t)\big) \\ &\le \sum_{j=1}^{K}\frac{p_j(t)-\gamma/K}{1-\gamma}\Big(1+\eta\hat{x}_j(t)+(e-2)\eta^2\hat{x}_j(t)^2\Big) \\ &\le \exp\left(\frac{\eta}{1-\gamma}\sum_{j=1}^{K}p_j(t)\hat{x}_j(t)\right. \\ &\quad \left. + \frac{(e-2)\eta^2}{1-\gamma}\sum_{j=1}^{K}p_j(t)\hat{x}_j(t)^2\right), \end{aligned} \tag{39}$$

where the equality follows from the definition of $W(t+1) = \sum_{j=1}^{K}\omega_j(t+1)$, and $\omega_j(t+1)$ in Algorithm 2 with $s = 1$. Also, the last inequality follows from the fact that $e^x \ge 1+x$. Finally, the first inequality holds by the definition of $p_j(t)$ in Algorithm 2 and the fact that $e^x \le 1 + x + (e-2)x^2$ for $x \le 1$. In this case, we need $\eta\hat{x}_j(t) \le 1$. From the definition of $\hat{x}_j(t)$ in Algorithm 2 and knowing that since $p_j(t) \ge \frac{\gamma}{K}$, $\delta(t) \ge \epsilon$, $1-\delta(t) \ge \epsilon$, for $T \ge 8K\ln K$ where $\epsilon = \sqrt[3]{\frac{K\ln K}{T}}$, by choosing $\gamma \ge \sqrt[3]{\frac{K\ln K}{T}}$, we find $\eta \le \frac{2}{K}\sqrt[3]{\frac{(K\ln K)^2}{T^2}}$ which satisfies the required condition (i.e., $\eta\hat{x}_j(t) \le 1$). By taking the logarithms and summing over T on both sides of equation (39), for the left hand side (LHS) of the equation, and for any j we have

$$\begin{aligned} \sum_{t=1}^{T}\ln\frac{W(t+1)}{W(t)} &= \ln\frac{W(T+1)}{W(1)} \ge \ln\omega_j(T+1) - \ln K \\ &= \eta\sum_{t=1}^{T}\hat{x}_j(t) - \ln K. \end{aligned} \tag{40}$$

By combining (39) with (40), we get

$$\begin{aligned} &\sum_{t=1}^{T}\hat{x}_j(t) - \sum_{t=1}^{T}\sum_{j=1}^{K}p_j(t)\hat{x}_j(t) \\ &\quad \le \gamma\sum_{t=1}^{T}\hat{x}_j(t) + (e-2)\eta\sum_{t=1}^{T}\sum_{j=1}^{K}p_j(t)\hat{x}_j^2(t) + \frac{\ln K}{\eta}. \end{aligned} \tag{41}$$

We take the expectation w.r.t. the both randomness of switching policy $\sigma(t)$ and $S(t)$, substitute the j with j^* (best channel index) and use (12), (13), (30), (31), (32), then we get

$$\begin{aligned} &\sum_{t=1}^{T}x_{j^*}(t) + \sum_{t=1}^{T}\mathbb{E}_{S(t)\sim p(t)}\Big[x_{S(t)}(t)\Big] \\ &\quad \le \gamma\sum_{t=1}^{T}x_{j^*}(t) + \frac{(e-2)K\eta}{4}\left(\sum_{t=1}^{T}\frac{1}{\delta(t)} + \sum_{t=1}^{T}\frac{1}{1-\delta(t)}\right) \\ &\qquad + \frac{\ln K}{\eta}, \end{aligned} \tag{42}$$

which gives us the bound on the first two terms of regret upper bound in (38). Next, by adding (35) with (42), we get the upper bound for (38) as follows:

$$\begin{aligned} \mathbb{E}[R(T)] &\le \gamma\sum_{t=1}^{T}x_{j^*}(t) + \frac{(e-2)K\eta}{4} \\ &\quad \times \left(\sum_{t=1}^{T}\frac{1}{\delta(t)} + \sum_{t=1}^{T}\frac{1}{1-\delta(t)}\right) \\ &\quad + \sum_{t=1}^{T}\delta(t) + \frac{\ln K}{\eta}. \end{aligned} \tag{43}$$

Substituting the bounds in (33), (34), and (35) into the above equation, we obtain the expected regret upper bound in the statement of the theorem.

Proof of Corollary 1: By getting the derivative from the statement in Theorem 2 w.r.t. η, we find the optimal value of $\eta = \frac{4}{T^{2/3}}\sqrt{\frac{\ln K}{(e-2)K}}\left(\frac{7}{(K\ln K)^{\frac{1}{3}}} + \frac{K\ln K}{(T^{\frac{1}{3}} - (K\ln K)^{\frac{1}{3}})^4}\right)^{-1/2}$ which also satisfies $\eta \le \frac{2}{K}\sqrt[3]{\frac{(K\ln K)^2}{T^2}}$. By choosing $\gamma = \sqrt[3]{\frac{K\ln K}{T}}$ we get the expected regret upper bound in the corollary for any $T \ge 8K\ln K$ as follows:

$$\begin{aligned} \mathbb{E}[R(T)] &\le (K\ln K)^{\frac{1}{3}}T^{\frac{2}{3}} + \frac{\sqrt{7}}{2}\sqrt{e-2}(K\ln K)^{\frac{1}{3}}T^{\frac{2}{3}} \\ &\quad + \frac{3}{2}(K\ln K)^{\frac{1}{3}}T^{\frac{2}{3}} \\ &= 3.62(K\ln K)^{\frac{1}{3}}T^{\frac{2}{3}}. \end{aligned}$$

Proof of Theorem 3: The proof follows similar steps as in the proof of Theorem 2, Corollary 1, and Theorem 1 with a difference which arises in satisfying the condition of $\eta_i\hat{x}_j(t) \le 1$ for every CR at position i in set C. Considering the definition of $\hat{x}_j(t)$ in Algorithm 2 and verifying the the following inequalities: $p_j^{(i)}(t) \ge \frac{\gamma_i}{K-i+1}$ for $j \notin \mathcal{M}_i$, $\delta(t) \ge \epsilon$, $1-\delta(t) \ge \epsilon$, for $T \ge 8K\ln K$ where $\epsilon = \sqrt[3]{\frac{K\ln K}{T}}$, and

$\prod_{r=1}^{i-1} 1 - p_j^{(r)}(t) \geq (\frac{\gamma_i}{2})^{i-1}$ for $i > 1$, by choosing $\gamma_i \geq (\frac{K \ln K}{T})^{1/3}$ we find the upper bound of learning rate $\eta_i \leq \frac{1}{2^{i-2}(K-i+1)} \sqrt[3]{\frac{(K \ln K)^{i+1}}{T^{i+1}}}$. We then find the optimal value of the learning rate $\eta_i = \frac{4}{T^{2/3}} \sqrt{\frac{\ln(K-i+1)}{(e-2)(K-i+1)}} (\frac{7}{(K \ln K)^{\frac{1}{3}}} + \frac{K \ln K}{(T^{\frac{1}{3}} - (K \ln K)^{\frac{1}{3}})^4})^{-1/2}$ which satisfies the required condition. By choosing $\gamma_i = (\frac{K \ln K}{T})^{1/3}$, considering $K - i + 1 \leq K$, and summing the regret over all the channels in the slate, we get the total expected regret upper bound in the theorem.

REFERENCES

[1] A. Burkitt-Gray, "U.S. Frees 3.5GHz for 5G in Dynamic Spectrum Sharing Plan." Jan. 2020. [Online]. Available: https://www.capacitymedia.com/articles/3824823/us-frees-35ghz-for-5g-in-dynamic-spectrum-sharing-plan

[2] D. Anderson, K. Shruthi, D. Crawford, and R. W. Stewart, "Evolving spectrum sharing methods, standards and trials," in *Spectrum Sharing: The Next Frontier in Wireless Networks*. Hoboken, NJ, USA: Wiley, 2019, pp. 59–74.

[3] "Report and order and second further notice of proposed rule making, 15-47 GN Docket no. 12-354," Federal Commun. Commission (FCC), Washington, DC, USA, Rep. FCC-15-47, Apr. 2015. [Online]. Available: https://docs.fcc.gov/public/attachments/FCC-15-47A1.pdf

[4] H. Yu and Y. B. Zikria, "Cognitive radio networks for Internet of Things and wireless sensor networks," *Sensors*, vol. 20, no. 18, p. 5288, 2020.

[5] J. Wang, M. Ghosh, and K. Challapali, "Emerging cognitive radio applications: A survey," *IEEE Commun. Mag.*, vol. 49, no. 3, pp. 74–81, Mar. 2011.

[6] K. Kumar and K. K. Kiran, "Centralized and decentralized cognitive radio network optimization," *Int. J. Res.*, vol. 3, no. 14, pp. 321–327, 2016.

[7] M. Hasegawa, H. Hirai, K. Nagano, H. Harada, and K. Aihara, "Optimization for centralized and decentralized cognitive radio networks," *Proc. IEEE*, vol. 102, no. 4, pp. 574–584, Apr. 2014.

[8] S. Bayhan and F. Alagoz, "Scheduling in centralized cognitive radio networks for energy efficiency," *IEEE Trans. Veh. Technol.*, vol. 62, no. 2, pp. 582–595, Feb. 2013.

[9] R. Murty, R. Chandra, T. Moscibroda, and P. Bahl, "SenseLess: A database-driven white spaces network," in *Proc. IEEE Int. Symp. Dyn. Spectr. Access Netw. (DySPAN)*, Aachen, Germany, May 2011, pp. 10–21.

[10] S.-P. Sheng, M. Liu, and R. Saigal, "Data-driven channel modeling using spectrum measurement," *IEEE Trans. Mobile Comput.*, vol. 14, no. 9, pp. 1794–1805, Sep. 2015.

[11] D. K. Patel, B. Soni, and M. López-Benítez, "On the estimation of primary user activity statistics for long and short time scale models in cognitive radio," *Wireless Netw.*, vol. 25, no. 8, pp. 5099–5111, Nov. 2019.

[12] H. Kyerematen-Boateng, M. Conn, D. Josyula, and M. Mareboyana, "Prediction of radio frequency spectrum occupancy," in *Proc. IEEE 19th Int. Conf. Trust Security Privacy Comput. Commun. (TrustCom)*, Guangzhou, China, Dec. 2020, pp. 2028–2034.

[13] X. Li, D. Wang, X. Mao, and J. McNair, "On the accuracy of maximum likelihood estimation for primary user behavior in cognitive radio networks," *IEEE Commun. Lett.*, vol. 17, no. 5, pp. 888–891, May 2013.

[14] X. Li, X. Mao, D. Wang, J. McNair, and J. Chen, "Primary user behavior estimation with adaptive length of the sample sequence," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Anaheim, CA, USA, Dec. 2012, pp. 1308–1313.

[15] D. Roy, T. Mukherjee, M. Chatterjee, and E. Pasiliao, "Primary user activity prediction in DSA networks using recurrent structures," in *Proc. IEEE Int. Symp. Dyn. Spectr. Access Netw. (DySPAN)*, Newark, NJ, USA, 2019, pp. 1–10.

[16] S. Demirci and D. Gözöpek, "Switching cost-aware joint frequency assignment and scheduling for industrial cognitive radio networks," *IEEE Trans. Ind. Informat.*, vol. 16, no. 7, pp. 4365–4377, Jul. 2020.

[17] D. Gözöpek, S. Buhari, and F. Alagöz, "A spectrum switching delay-aware scheduling algorithm for centralized cognitive radio networks," *IEEE Trans. Mobile Comput.*, vol. 12, no. 7, pp. 1270–1280, Jul. 2013.

[18] *Part II: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, ANSI/IEEE Standard 802.11, 1999. [Online]. Available: https://www.wardriving.ch/hpneu/info/doku/802.11-1999.pdf

[19] M. Çamurli and D. Gözüpek, "Channel switching cost-aware resource allocation for multi-hop cognitive radio networks with a single transceiver," in *Ad Hoc Networks*. Cham, Switzerland: Springer Int., 2014, pp. 158–168. [Online]. Available: https://doi.org/10.1007

[20] S. Eryigit, S. Bayhan, and T. Tugcu, "Channel switching cost aware and energy-efficient cooperative sensing scheduling for cognitive radio networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Budapest, Hungary, 2013, pp. 2633–2638.

[21] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The non-stochastic multiarmed bandit problem," *SIAM J. Comput.*, vol. 32, no. 1, pp. 48–77, Jan. 2003.

[22] N. Cesa-Bianchi and G. Lugosi, "Combinatorial bandits," *J. Comput. Syst. Sci.*, vol. 78, no. 5, pp. 1404–1422, 2012.

[23] S. Kale, L. Reyzin, and R. E. Schapire, "Non-stochastic bandit slate problems," in *Advances in Neural Information Processing Systems 23*. Red Hook, NY, USA: Curran Assoc., Inc., 2010, pp. 1054–1062.

[24] J.-Y. Audibert, S. Bubeck, and G. Lugosi, "Regret in online combinatorial optimization," *Math. Oper. Res.*, vol. 39, no. 1, pp. 31–45, Feb. 2014. [Online]. Available: https://doi.org/10.1287/moor.2013.0598

[25] N. Alon, N. Cesa-Bianchi, C. Gentile, S. Mannor, Y. Mansour, and O. Shamir, "Nonstochastic multi-armed bandits with graph-structured feedback," *SIAM J. Comput.*, vol. 46, no. 6, pp. 1785–1826, 2017.

[26] W. Na, J. Yoon, S. Cho, D. Griffith, and N. Golmie, "Centralized cooperative directional spectrum sensing for cognitive radio networks," *IEEE Trans. Mobile Comput.*, vol. 17, no. 6, pp. 1260–1274, Jun. 2018.

[27] I. F. Akyildiz, B. F. Lo, and R. Balakrishnan, "Cooperative spectrum sensing in cognitive radio networks: A survey," *Phys. Commun.*, vol. 4, no. 1, pp. 40–62, 2011.

[28] M. Dabaghchian, S. Liu, A. Alipour-Fanid, K. Zeng, X. Li, and Y. Chen, "Intelligence measure of cognitive radios with learning capabilities," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Washington, DC, USA, 2016, pp. 1–6.

[29] T. Manna and I. S. Misra, "A prediction and scheduling framework in centralized cognitive radio network for energy efficient non-real time communication," *Int. J. Commun. Syst.*, vol. 31, no. 13, p. e3716, 2018.

[30] N. Shami and M. Rasti, "A joint multi-channel assignment and power control scheme for energy efficiency in cognitive radio networks," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Doha, Qatar, 2016, pp. 1–6.

[31] O. Sweileh, M. S. Hassan, H. S. Mir, and M. H. Ismail, "A switching-based and delay-aware scheduling algorithm for cognitive radio networks," *Int. J. Interdiscip. Telecommun. Netw.*, vol. 11, no. 3, pp. 34–48, 2019.

[32] M. M. Rashid, M. J. Hossain, E. Hossain, and V. K. Bhargava, "Opportunistic spectrum scheduling for multiuser cognitive radio: A queueing analysis," *IEEE Trans. Wireless Commun.*, vol. 8, no. 10, pp. 5259–5269, Oct. 2009.

[33] Y. Gai, B. Krishnamachari, and R. Jain, "Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation," in *Proc. IEEE Symp. New Front. Dyn. Spectr. (DySPAN)*, Singapore, 2010, pp. 1–9.

[34] M. Lelarge, A. Proutière, and M. S. Talebi, "Spectrum bandit optimization," in *Proc. IEEE Inf. Theory Workshop (ITW)*, 2013, pp. 1–5.

[35] S. Kang and C. Joo, "Low-complexity learning for dynamic spectrum access in multi-user multi-channel networks," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, Honolulu, HI, USA, 2018, pp. 1367–1375.

[36] J. Kalliovaara. "5G-Xcast Open Spectrum Data 1/6." Zenodo. Aug. 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1293283

[37] T. Taher *et al.*, "Global spectrum observatory network setup and initial findings," in *Proc. 9th Int. Conf. Cogn. Radio Orient. Wireless Netw. Commun. (CROWNCOM)*, Oulu, Finland, 2014, pp. 79–88.

[38] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, pp. 235–256, May 2002.

[39] H. Robbins, "Some aspects of the sequential design of experiments," *Bull. Amer. Math. Soc.*, vol. 58, no. 5, pp. 527–535, Sep. 1952.

[40] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, no. 1, pp. 4–22, 1985.

[41] N. Littlestone and M. Warmuth, "The weighted majority algorithm," *Inf. Comput.*, vol. 108, no. 2, pp. 212–261, 1994.

[42] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, 1997.

[43] J.-Y. Audibert and S. Bubeck, "Minimax policies for adversarial and stochastic bandits," in *Proc. 22nd Annu. Conf. Learn. Theory (COLT)*, Jan. 2009, pp. 773–818.

[44] O. Dekel, J. Ding, T. Koren, and Y. Peres, "Bandits with switching costs: $T^{2/3}$ regret," in *Proc. 46th Annu. ACM Symp. Theory Comput.*, 2014, pp. 459–467.

[45] R. Arora, O. Dekel, and A. Tewari, "Online bandit learning against an adaptive adversary: From regret to policy regret," in *Proc. 29th Int. Conf. Mach. Learn.*, Jan. 2012, pp. 1747–1754.

[46] C. Shen, C. Tekin, and M. van der Schaar, "A non-stochastic learning approach to energy efficient mobility management," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3854–3868, Dec. 2016.

[47] T. Uchiya, A. Nakamura, and M. Kudo, "Algorithms for adversarial bandit problems with multiple plays," in *Proc. 21st Int. Conf. Algorithmic Learn. Theory*, 2010, pp. 375–389.

[48] M. Dabaghchian, A. Alipour-Fanid, K. Zeng, Q. Wang, and P. Auer, "Online learning with randomized feedback graphs for optimal PUE attacks in cognitive radio networks," *IEEE/ACM Trans. Netw.*, vol. 26, no. 5, pp. 2268–2281, Oct. 2018.

[49] M. Dabaghchian, A. Alipour-Fanid, K. Zeng, and Q. Wang, "Online learning-based optimal primary user emulation attacks in cognitive radio networks," in *Proc. IEEE Conf. Commun. Netw. Security (CNS)*, Philadelphia, PA, USA, 2016, pp. 100–108.

[50] R. Arora, T. V. Marinov, and M. Mohri, "Bandits with feedback graphs and switching costs," in *Proc. Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, BC, Canada, Dec. 2019, pp. 10397–10407.

[51] R. Combes, M. S. Talebi, A. Proutiere, and M. Lelarge, "Combinatorial bandits revisited," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran, 2015, pp. 2116–2124.

[52] J. Zimmert, H. Luo, and C.-Y. Wei, "Beating stochastic and adversarial semi-bandits optimally and simultaneously," in *Proc. 36th Int. Conf. Mach. Learn.*, vol. 97, Jun. 2019, pp. 7683–7692.

[53] C.-Y. Wei and H. Luo, "More adaptive algorithms for adversarial bandits," in *Proc. 31st Conf. Learn. Theory*, vol. 75, Jul. 2018, pp. 1263–1291.

[54] R. Allesiardo, R. Féraud, and O.-A. Maillard, "The non-stationary stochastic multi-armed bandit problem," *Int. J. Data Sci. Anal.*, vol. 3, pp. 267–283, Mar. 2017.

[55] W. Chen, L. Wang, H. Zhao, and K. Zheng, "Combinatorial semi-bandit in the non-stationary environment," 2021, *arXiv:2002.03580*.

[56] A. Garivier and E. Moulines, "On upper-confidence bound policies for switching bandit problems," in *Algorithmic Learning Theory*, J. Kivinen, C. Szepesvári, E. Ukkonen, and T. Zeugmann, Eds. Heidelberg, Germany: Springer, 2011, pp. 174–188.

[57] G. Neu, "First-order regret bounds for combinatorial semi-bandits," in *Proc. 28th Annu. Conf. Learn. Theory (COLT)*, vol. 40. Paris, France, Jul. 2015, pp. 1360–1375.

[58] G. Neu and G. Bartók, "An efficient algorithm for learning with semi-bandit feedback," in *Algorithmic Learning Theory*. Heidelberg, Germany: Springer, 2013, pp. 234–248.

[59] A. Goldsmith, S. A. Jafar, I. Maric, and S. Srinivasa, "Breaking spectrum gridlock with cognitive radios: An information theoretic perspective," *Proc. IEEE*, vol. 97, no. 5, pp. 894–914, May 2009.

[60] B. Kumar, S. K. Dhurandher, and I. Woungang, "A survey of overlay and underlay paradigms in cognitive radio networks," *Int. J. Commun. Syst.*, vol. 31, p. e3443, Jan. 2018.

[61] S. Bubeck, Y. Li, Y. Peres, and M. Sellke, "Non-stochastic multi-player multi-armed bandits: Optimal rate with collision information, sublinear without," in *Proc. 33rd Conf. Learn. Theory*, vol. 125, Jul. 2020, pp. 961–987.

[62] J. Zimmert and Y. Seldin, "Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits," *J. Mach. Learn. Res.*, vol. 22, no. 28, pp. 1–49, 2021.

[63] W. Chen, Y. Wang, and Y. Yuan, "Combinatorial multi-armed bandit: General framework and applications," in *Proc. 30th Int. Conf. Mach. Learn.*, vol. 28. Atlanta, GA, USA, Jun. 2013, pp. 151–159.

[64] B. Kveton, Z. Wen, A. Ashkan, and C. Szepesvari, "Tight regret bounds for stochastic combinatorial semi-bandits," in *Proc. 18th Int. Conf. Artif. Intell. Stat.*, vol. 38. San Diego, CA, USA, May 2015, pp. 535–543.

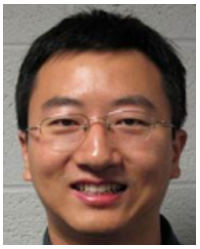
[65] S. Wang and W. Chen, "Thompson sampling for combinatorial semi-bandits," in *Proc. 35th Int. Conf. Mach. Learn. (ICML)*, Jul. 2018, pp. 5114–5122. [Online]. Available: https://www.microsoft.com/en-us/research/publication/thompson-sampling-for-combinatorial-semi-bandits/

[66] T. Koren, R. Livni, and Y. Mansour, "Multi-armed bandits with metric movement costs," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran, 2017, pp. 4122–4131.

Amir Alipour-Fanid (Member, IEEE) received the Ph.D. degree in electrical and computer engineering from George Mason University, Fairfax, VA, USA. He is currently a Senior Researcher with the Architectures and Security Team, General Motors (GM) Research and Development. Prior to joining the GM, he was an Assistant Professor with the Department of Computer Science and Electrical Engineering, West Virginian University, Morgantown, WV, USA. His research interests focus on wireless cyber–physical systems security, Internet-of-Things communication security, connected and autonomous vehicles security, 5G wireless communication, theoretical multi-armed bandits, and applied machine learning.

Monireh Dabaghchian (Member, IEEE) received the Ph.D. degree in electrical and computer engineering from George Mason University, Fairfax, VA, USA, in 2019. She is currently an Assistant Professor with the Department of Computer Science, Morgan State University (MSU), Baltimore, MD, USA, where she is also a member of the Cybersecurity Assurance and Policy Center. Her research interests lie at the intersection of cybersecurity and machine learning. Specifically, her work focuses on designing machine learning algorithms including multi-armed bandits learning and robust adversarial machine learning techniques with applications in cybersecurity, Internet-of-Things security, cyber–physical systems security, and network security. She was a recipient of the U.S. National Science Foundation Research Initiation Award in 2021 and the Outstanding Academic Achieved Award from George Mason University, in 2020.

Raman Arora received the Ph.D. degree from the University of Wisconsin–Madison. He is an Assistant Professor with the Department of Computer Science, Johns Hopkins University, where he is also affiliated with the Mathematical Institute for Data Science, the Center for Language and Speech Processing, and the Institute for Data Intensive Engineering and Science. Prior to joining Johns Hopkins, he was a Research Assistant Professor/Postdoctoral Scholar with Toyota Technological Institute, Chicago; a Visiting Researcher with Microsoft Research, Redmond; and a Research Associate with the University of Washington, Seattle. His research interests are in machine learning, online learning, robustness and privacy. He received the NSF CAREER Award in 2019.



Kai Zeng (Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the Worcester Polytechnic Institute (WPI) in 2008. He was a Postdoctoral Scholar with the Department of Computer Science, University of California at Davis (UCD) from 2008 to 2011. He was with the Department of Computer and Information Science, University of Michigan–Dearborn as an Assistant Professor from 2011 to 2014. He is currently an Associate Professor with the Department of Electrical and Computer Engineering, Cyber Security Engineering, and the Department of Computer Science, George Mason University. His current research interests are in cyber–physical system/IoT security and privacy, 5G wireless security, machine learning, and spectrum sharing. He received the Excellence in Postdoctoral Research Award from UCD in 2011 and the Sigma Xi Outstanding Ph.D. Dissertation Award from WPI in 2008. He was a recipient of the U.S. National Science Foundation Faculty Early Career Development Award in 2012. He is an Associate Editor of the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY and IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING.