Sequential Learning for Multi-channel Wireless Network Monitoring with Channel Switching Costs

Thanh Le, Csaba Szepesvári, *Senior Member, IEEE*, and Rong Zheng, *Senior Member, IEEE*

***Abstract*—We consider the problem of optimally assigning p sniffers to K channels to monitor the transmission activities in a multi-channel wireless network with switching costs. The activity of users is initially unknown to the sniffers and is to be learned along with channel assignment decisions to maximize the benefits of this assignment, resulting in the fundamental trade-off between exploration and exploitation. Switching costs are incurred when sniffers change their channel assignments. As a result, frequent changes are undesirable. We formulate the sniffer-channel assignment with switching costs as a linear partial monitoring problem, a super-class of multi-armed bandits. As the number of arms (sniffer-channel assignments) is exponential, novel techniques are called for, to allow efficient learning. We use the linear bandit model to capture the dependency amongst the arms and develop a policy that takes advantage of this dependency. We prove that the proposed Upper Confident Bound-based (UCB) policy enjoys a logarithmic regret bound in time t that depends sub-linearly on the number of arms, while its total switching cost grows in the order of $O(\log\log(t))$.**

***Index Terms*—Local area networks, network monitoring, sequential learning.**

I. Introduction

Deployment and management of wireless devices and networks are often hampered by the poor visibility of PHY and MAC characteristics, and complex interactions at various layers of the protocol stack both inside a managed network and across multiple administrative domains. The "can you hear me" Verizon wireless TV commercial is a vivid demonstration of the shortage of real-time knowledge that cellular providers have regarding the condition of operational networks. Accurate and timely estimates of network conditions and performance characteristics can yield better performance in a number of applications, including the following:

- *Network resource management:* Wireless service providers and network administrators need to determine the coverage of their own networks and make critical decisions such as dimensioning and allocation of network resources.
- *Wireless advisory:* Individual devices can better adapt their operational parameters (e.g., channels, sub-carriers, hopping sequences, transmission power levels, etc.) for co-existence and better performance.
- *Trouble-shooting and diagnosis:* Availability of cross-layer information of the operational network can help network administrators to determine the root causes of service outage or performance degradation, as well as to identify malicious behavior and intrusion.

Passive monitoring is a technique where a dedicated set of hardware devices, called *sniffers*, are used to monitor activities in wireless networks. These devices capture transmissions of wireless devices or activities of interference sources in their vicinity, and store packet level or PHY layer information in trace files, which can be analyzed distributively or at a central location. Passive monitoring is advantageous in that it is not limited to a single wireless transmission technology, and it can be used to capture PHY layer characteristics and phenomena only observable on the wireless side such as electronic-magnetic interference levels, collision events and packet transmission times.

Since most, if not all infrastructure networks utilize multiple contiguous or non-contiguous channels or bands [1], an important issue is to determine which set of frequency bands each sniffer should operate on to maximize the total amount of information gathered. This is called the *sniffer-channel assignment* problem or *channel assignment* problem for short. It is a challenging problem for a number of reasons. First, monitoring resources are limited, and thus it is infeasible to monitor all channels at all locations at all times. Second, intelligent channel assignment requires the knowledge of usage patterns, i.e., the likelihood of the occurrence of interesting events. These are not known *a priori*. An interesting trade-off arises between assigning sniffers to channels known to be the busiest based on current knowledge, versus exploring channels that are under-sampled. Third, in practical systems, channel switching is not instantaneous. For example, [1] reports on the 802.11b/g AR5212 chipset that the channel switching operation requires a full hardware reset that incurs a delay of approximately 1.2 ms. An additional delay of about 3.2 ms is introduced in the channel switching operation due to other system operations, such as flushing any pending transmission buffers in the hardware queues and waiting for any pending transmit or receive direct memory access (DMA) operations to finish. Our own measurements on USRP2, a software

Copyright (c) 2014 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

Thanh Le is an Assistant Manager at Samsung Electronics, Vietnam. The work was conducted when he was a MS student with the Department of Electrical and Computer Engineering and Department of Computer Science, University of Houston, Houston, Texas USA.

Csaba Szepesvári is with the Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada (E-mail: szepesva@ualberta.ca).

Rong Zheng is with the Department of Computing and Software, McMaster University, Hamilton, ON, Canada. The work was partly conducted when the author was with the Department of Computer Science, University of Houston, Houston, Texas, USA.

[1] A channel can be a single frequency band, a code in a CDMA system, or a hopping sequence in a frequency-hopping system.

defined radio platform indicate a latency on the order of hundreds of milliseconds when shifting the central frequency of the spectrum analyzer implemented in GNURadio. During channel switching, packet receptions and transmissions are not possible. As a result, frequent switching is undesirable.

Sniffer-channel assignment with no prior knowledge of user activity is closely related to the multi-armed bandit problem (MAB) [2]. In a MAB, a gambler must decide which arm of N non-identical slot machines to play in a sequence of trials so as to maximize his payoff. In the sniffer-channel assignment problem, each of the p sniffers must be assigned to one of the K non-identical channels to monitor so as to maximize the total information gathered. The number of choices (arms) available in a round is thus $N = K^p$. In this work we assume that the pay-off is proportional to the number of distinct users detected. For simplicity, we assume that a user's activity in a given channel can be described with a sequence of IID Bernoulli random variables. However, as opposed to the standard MAB problem, the observation upon a single assignment is not only the reward associated with the assignment, but also the activity patterns observed at each monitored channel. Note that the observed pattern may have correlated components, e.g. when two sniffers observe the transmission of the same set of users.

A policy for sniffer-channel assignment determines at any point in time the assignment to be chosen based on past information. Further, to discourage the policies from frequently changing the channel assignment, we add a constant to the regret of the policy every time the assignment is changed. The efficiency of different policies is measured in terms of their associated *regret*, which is defined as the difference between the expected pay-off gained by a "genie" (an unattainable ideal) who always uses the optimal stationary sniffer-channel assignment (thus not suffering any switching cost), and that obtained by the given policy. The regret achieved by a policy can be evaluated in terms of its growth over time and how it scales with respect to the various problem parameters. A naive approach to the channel assignment problem would be to treat each sniffer-channel combination as an arm (action), and learn the statistics of each arm individually. With p sniffers and K channels in the network, the statistics of a total of K^p arm-payoffs needs to be learned. Direct application of known approaches to MAB (e.g., UCB[3], ϵ-greedy[4]) results in a problem-dependent regret-bound linear in the number of arms K^p.

In this paper, we formulate the optimal channel assignment with switching costs as a multi-agent multi-arm partial information problem with linearly parameterized pay-off. Our proposed policy is centralized and slotted in nature, namely, a fusion center collects the information from each sniffer in each time-slot and makes decisions regarding the channel assignments for the next slot. Utilizing the dependency among the arms, we reduce the dimension of the unknown parameter to $K \cdot 2^p$. We devise an order-optimal policy whose total regret grows logarithmically with respect to time with an associated constant that grows sub-linearly in the number of arms. The switching cost incurred only grows as $O(\log\log(t))$, where t denotes time. The key improvement compared to the naive approach comes from the concept of *spanner arms*, i.e, a small collection of arms which provide information about all parameters. The policies and regret bounds are derived for general correlation structures among the sniffers, and remain valid for special cases where the sniffer's observations are identical or independent.

The rest of the paper is organized as follows. In Section II, related work on wireless monitoring and sequential learning is summarized. We present the problem formulation in Section III. Details and analysis of the proposed order-optimal policy are provided in Section IV. Simulation results are presented in Section V, followed by conclusion and a list of suggestions for future work in Section VI.

II. Related Work

Wireless monitoring is an active area of research that has received much attention from several perspectives. There has been much work done on wireless monitoring from a *system-level* viewpoint, in an attempt to design complete systems, and address the interactions among the components of such systems [5], [6], [7], [8], [9]. The authors of these works have argued both qualitatively and quantitatively the need for monitoring on the wireless side.

To determine the optimal allocation of monitoring resources to maximize captured information, Shin and Bagchi consider the selection of monitoring nodes and their associated channels for monitoring wireless mesh networks [10]. The optimal monitoring is formulated as a maximum coverage problem with group budget constraints, which was previously studied by Chekuri and Kumar in [11]. In [12], we introduced a quality of monitoring (QoM) metric defined by the expected number of active users monitored, and investigated the problem of maximizing QoM by judiciously assigning sniffers to channels based on knowledge of user activities in a multi-channel wireless network. Two capture models are considered. The first one, called the *user-centric model* assumes frame-level capturing capability of sniffers such that the activities of different users can be distinguished. The second one, called the *sniffer-centric model* utilizes binary channel information only(active or not) at a sniffer.

The above works assume that certain statistics regarding the users' activity are given [10], [11] or can be inferred [12]. When such statistics are not known a priori, sequential learning is needed. Sequential decision making in presence of uncertainty, faces the fundamental trade-off between *exploration* and *exploitation*. On one hand, it is desirable to put sniffers to the channels where most activities have been observed and thus more information is likely to be gathered (exploitation). On the other hand, exploring the channels that are under-sampled helps to reduce uncertainty and thus avoid being misled by imprecise information. Such trade-offs are vividly illustrated by the famous multi-armed bandit problem (MAB). A large volume of work has been been devoted to designing good strategies for variations of the MAB problem and to the understanding of the theoretical limits of such procedures, among which, just to name a few, Lai and Robbin [3] established logarithmic upper and lower bounds for bandit problems where the pay-off distributions of arms belong to some known parametric family; Agrawal [13] considered a class of sample-mean based policies for the same setting; Auer *et al* analyzed upper confidence bound (UCB) based and ϵ-greedy policies for non-parametric stochastic bandit problems [4]. Recently, bandit problems with linear parameterized payoff are considered in [14], [15]. Regret minimization under partial monitoring is investigated in [16], where the player in a repeated game, instead of observing the action chosen by the opponent in each game round, receives a feedback generated by the combined choice of the two players. MAB with switching costs was first considered in [17]. An excellent survey on MAB with switching costs can be found in [18].

Recognizing the connection between the MAB and spectrum access in cognitive radio networks, Lai *et al.* applied the UCB1 algorithm [4] to single user-channel selection in [19], and later extended it to consider Markovian payoffs and for the case of multiple users in [20]. Liu and Zhao [21] formulated the problem of secondary user channel selection as a decentralized multi-armed bandit problem, and presented a policy that achieves asymptotically logarithmic regret in time. Anandkumar [22] proposed two policies for distributed learning and access with order-optimal cognitive system throughput under self play. In addition to learning the channel availability, the second users also learn the other users' strategies and the number of total users in the system through channel feedback. Existing work applying MAB in the cognitive radio context assumes identical channel view with the exception of Gai *et al* [23]. However, the model considered in this work, in fact makes the implicit assumption that all secondary users are co-located ("if there are multiple users on the channel, then we assume that, due to interference, at most one of the conflicting users gets reward"). Since co-located secondary users likely observe identical primary user activities, a contradiction arises to the claim of "allowing the reward process on the same channel to be different" [23]. In our earlier work [24], we proposed two order-optimal policies for the channel assignment problem without switching costs. Both policies have logarithmic regrets in time that are sub-linear to the number of arms. In this work, we extend the previous algorithms to consider switching costs. To demonstrate the importance of this extension, in Section V we will show that the direct application of the policy of [24] without consideration of switching costs leads to a significantly higher regret.

In contrast to existing work, we consider a model where sniffers are in general configuration and may observe different sets of users in the same channel. This encompasses models when either sniffers are co-located or when they are sufficiently far apart. The algorithms and analytical bounds devised are directly applicable to these specialized cases. Admittedly, due to its generality, the model suffers from a higher computation and storage complexity. Unfortunately, this is unavoidable as a result of the NP-hardness of the nominal resource allocation problem when all statistics are known, as discussed in Section III.

III. Problem Formulation

Consider p sniffers monitoring user activities in K channels. A user u operates in one of K channels, $c(u) \in \mathcal{K} = \{1, \ldots, K\}$. Let p_u denote the transmission probability of user u. We represent the relationship between users and sniffers using an undirected bi-partite graph $G = (S, U, E)$, where $S = \{1, \ldots, p\}$ is the set of sniffer nodes and U is the set of users. An edge $e = (s, u)$ exists between sniffer $s \in S$ and user $u \in U$ if u is within the reception range of sniffer s. If transmissions from a user cannot be captured by any sniffer, the user is excluded from G. For every vertex $v \in S \cup U$, we let $N(v)$ denote vertex v's neighbors in G. For users, their neighbors are sniffers, and vice versa. We assume that one sniffer can observe one user at a time. This is consistent with many existing multiple access mechanisms including FDMA and TDMA.

At any point in time, a sniffer can only observe transmissions over a single channel. We will consider *channel assignments* of sniffers to channels, $\mathbf{k} = (k_1, \ldots, k_p)$, where $1 \le k_i \le K$. Let $\mathbb{K} = \{\mathbf{k} \mid \mathbf{k} : S \to \{1, .., K\}^p\}$ be the set of all possible assignments. The set of users a sniffer s can observe is given by $N(s) \bigcap \{\, u : c(u) = k_s \,\}$.

A. Optimal channel assignment in the nominal form

We first consider the formulation of the optimal sniffer-channel assignment where the graph G and the user-activity probabilities $(p_u; u \in U)$ are both known. Since optimal channel assignment with uncertainty is inherently harder than that without uncertainty, determining the complexity of the later provides a baseline understanding of the computational aspect of the former problem.

The objective of optimal channel assignment is to maximize the expected number of active users monitored. Let MAX-EFFORT-COVER (MEC) denote the problem of determining the optimal channel assignment of sniffers to maximize the total weight of users monitored, under the constraint that each sniffer can monitor one of a set of k channels. Note that in MEC, the weights can in fact be any non-negative values and are not limited to $[0, 1]$. The MEC problem can be cast as the following integer program (IP):

$$\begin{array}{lll} \max & \sum_{u \in U} p_u y_u & \\ \text{s.t.} & \sum_{k=1}^{K} z_{s,k} \le 1 & \forall s \in S \\ & y_u \le \sum_{s \in N(u)} z_{s,c(u)} & \forall u \in U \\ & y_u, z_{s,k} \in \{0,1\} & \forall u, s, k. \end{array} \tag{1}$$

Each sniffer is associated with a set of binary decision variables: $z_{s,k} = 1$ if the sniffer is assigned to channel k; 0, otherwise. Further, y_u is a binary variable (but not a decision variable) indicating whether or not user u is monitored, and p_u is the weight associated with user u. The following result has been proven in [12]:

Theorem 1 (Theorem 1[12]): The MEC problem is NP-hard with respect to the number of sniffers, even for $K = 2$.

In other words, the computational complexity for a genie to make the optimal choice with the knowledge of all users' activity grows grows faster than any polynomial with respect to the number of sniffers, unless $P = NP$. However, when the graphs G have some specific structure, there may exist efficient algorithms. For example, when G is restricted to be a complete bipartite graph, it can be shown that MEC reduces to maximum matching in a transformed bipartite graph, which can be solved in polynomial time.

When the graph G and the user activity probabilities are given, the optimal sniffer channel assignment is stationary in time. However, when the user activity probabilities are unknown, a learning strategy must try different assignments, and as a result, intelligent schemes need to be designed to take the switching costs into account.

B. Linear bandit for optimal channel assignment with uncertainty

Now, we turn to the optimal channel assignment when there is uncertainty in both G and p_u's. We first define the structure of the instantaneous feedback and payoff of each sniffer.

Let $U_{ik}(t)$ be a nonnegative, integer-valued random variable that denotes the index of the user whose activity sniffer i observes in channel k at time t, or which takes the value of zero if there is no user activity in channel k. For simplicity, we assume that $U(t) = (U_{ik}(t); 1 \leq i \leq p, 1 \leq k \leq K)$ is a sequence of IID random variables. The instantaneous feedback (observations) received under the joint action $\mathbf{k}(t) = (k_1, \ldots, k_p)$ is $Y^{\circ}_{(k_1,\ldots,k_p)}(t) = (U_{1,k_1}(t), U_{2,k_2}(t), \ldots, U_{p,k_p}(t))$. Let $\mathbb{I}_{\{x\}}$ be the indicator function, where $\mathbb{I}_{\{x\}} = 1$ if x is true; and $\mathbb{I}_{\{x\}} = 0$, otherwise. Note that the indicator $\mathbb{I}_{\{U_{i_1,k_{i_1}}(t)=U_{i_2,k_{i_2}}(t)=\ldots=U_{i_s,k_{i_s}}(t)>0\}}$ is a function of $Y^{\circ}_{(k_1,\ldots,k_p)}(t)$ and hence can be taken as part of the observation $Y_{(k_1,\ldots,k_p)}(t)$, defined as the collection

$$\Big[\mathbb{I}_{\{U_{i_1,k_{i_1}}(t)=U_{i_2,k_{i_2}}(t)=\ldots=U_{i_s,k_{i_s}}(t)>0\}}; \\ 1 \leq s \leq p, 1 \leq i_1 < \ldots < i_s \leq p \,\Big]. \qquad (2)$$

We view $Y_{(k_1,\ldots,k_p)}$ as a vector of 2^p binary variables indicating whether the respective collection of sniffers observe the same user. Clearly, there exists a bijection between $Y^{\circ}_{(k_1,\ldots,k_p)}$ and $Y_{(k_1,\ldots,k_p)}$ (by possibly renaming the users) under the condition that each sniffer can only observe one user at a time.

Note that spatial multiplexing is allowed such that multiple users can be active at the same time in one channel (as long as they are sufficiently far apart geographically). However, we assume that only one user can be observed by one sniffer at a time. This is consistent with many existing multiple access mechanisms including FDMA, TDMA. As in Section III-A, the payoff upon selecting the joint action is the number of distinct users observed. That is, the joint payoff for selecting channels $\mathbf{k} = (k_1, k_2, \ldots, k_p)$ is

$$\begin{aligned} X_{\mathbf{k}}(t) \quad &= \quad |\{U_{1,k_1}(t), \ldots, U_{p,k_p}(t)\}| \\ &\qquad -\mathbb{I}_{\{U_{1,k_1}(t)=0,\ldots,U_{p,k_p}(t)=0\}} \\ &= \quad \textstyle\sum_{i=1}^{p} \mathbb{I}_{\{U_{1,k_i}(t)>0\}} \\ &- \quad \textstyle\sum_{i,j=1}^{p} \mathbb{I}_{\{U_{i,k_i}(t)=U_{j,k_j}(t)>0\}} \mathbb{I}_{\{k_i=k_j,i\neq j\}} \\ &\ldots \\ &- \quad (-1)^p \mathbb{I}_{\{U_{1,k_1}(t)=U_{2,k_2}(t)=\ldots=U_{p,k_p}(t)>0\}} \\ &\times \quad \mathbb{I}_{\{k_1=k_2=\ldots=k_p\}}. \end{aligned} \qquad (3)$$

The expected payoff for channels $\mathbf{k} = (k_1, k_2, \ldots, k_p)$ is given by,

$$\begin{aligned} &\mathbb{E}\left[X_{\mathbf{k}}(t)\right] \\ = \quad &\textstyle\sum_{i=1}^{p} \mathbb{P}\left(U_{1,k_i}(t) > 1\right) \\ - \quad &\textstyle\sum_{i,j=1}^{p} \mathbb{P}\big(U_{i,k_i}(t) = U_{j,k_j}(t) > 0\big)\, \mathbb{I}_{\{k_i=k_j,i\neq j\}} \\ &\ldots \\ - \quad &(-1)^p \mathbb{P}\big(U_{1,k_1}(t) = \ldots = U_{p,k_p}(t) > 0\big) \\ &\times \mathbb{I}_{\{k_1=k_2=\ldots=k_p\}} \end{aligned} \qquad (4)$$

Define a vector θ, whose components are initially unknown to the learning algorithm, with the following entries:

$$\begin{aligned} \mathbb{P}\left(U_{i,k} > 0\right), \qquad & 1 \leq i \leq p, 1 \leq k \leq K, \\ \mathbb{P}\left(U_{i_1,k} = U_{i_2,k} > 0\right), \qquad & 1 \leq i_1 < i_2 \leq p, 1 \leq k \leq K, \\ & \vdots \\ \mathbb{P}\left(U_{1,k} = U_{2,k} = \ldots = U_{p,k} > 0\right), \qquad & 1 \leq k \leq K. \end{aligned} \qquad (5)$$

We introduce the "arm features", $\phi_{\mathbf{k}} \in \mathbb{R}^M$ shown in (6), where $M = K(2^p - 1)$. Note that the jth arm feature $\phi_{\mathbf{k},j}$ is uniquely determined by the arm $\mathbf{k} = (k_1, k_2, \ldots, k_p)$. Let $\mathcal{M}_{\mathbf{k}} = \{\, i : 1 \leq i \leq M, \phi_{\mathbf{k},i} \neq 0 \,\}$ be the set of nonzero components of feature vector $\phi_{\mathbf{k}}$ and let $M_{\mathbf{k}} = |\mathcal{M}_{\mathbf{k}}|$.

To this end, we can rewrite the expected payoff in MEC as a linear function of the arm feature $\phi_{\mathbf{k}}$,

$$\mathbb{E}\left[X_{\mathbf{k}}(t)\right] = \theta^T \phi_{\mathbf{k}}, \qquad (7)$$

where $(\cdot)^T$ denotes transposition.

Knowing θ suffices to play optimally: An arm with maximal payoff is given by $\mathbf{k}^* = \operatorname{argmax}_{\mathbf{k}\in\mathbb{K}} \theta^\top \phi_{\mathbf{k}}$ (here, and in what follows, for the sake of simplicity, we assume that there is a unique optimal arm). Note that this optimization problem is just a trivial reformulation of the MEC problem in Section III. A reasonable way to estimate the parameter vector θ is to keep a running average for the components of θ. If at time t the agent chose $\mathbf{k}(t) \in \mathbb{K}$ then the current estimate, $\hat{\theta}(t-1)$, can be updated by

$$\begin{aligned} &\hat{\theta}_i(t) = \hat{\theta}_i(t-1) + \frac{1}{N_i(t)} \Big(Y_i(t) - \hat{\theta}_i(t-1)\Big)\, \mathbb{I}_{\{i\in\mathcal{M}_{\mathbf{k}(t)}\}}, \\ &N_i(t) = N_i(t-1) + \mathbb{I}_{\{i\in\mathcal{M}_{\mathbf{k}(t)}\}}. \end{aligned} \qquad (8)$$

Here $N_i(0) = 0$, $\hat{\theta}_i(0) = 0$. Thus, $N_i(t)$ counts the number of times that component i has been observed up to time t. $Y_i(t)$ is defined in (2), the binary observation for component i at time t.

Example 1 (Co-located sniffers): When the sniffers are "co-located" or are deployed at close proximity, their observations are identical. Therefore, $U(t)$ will be such that if $k_i = k_j$ then

$$\phi_{\mathbf{k},i} = \begin{cases} \mathbb{I}_{\{k_1=i\}}\ , & \text{if } 1 \le i \le K; \\ \cdots & \\ \mathbb{I}_{\{k_2=i-l\cdot K\}}\ , & \text{if } l\cdot K+1 \le i \le (l+1)\cdot K; \\ \cdots & \\ -\mathbb{I}_{\{k_1=k_2=i-p\cdot K\}}\ , & \text{if } p\cdot K+1 \le i \le (p+1)\cdot K; \\ \cdots & \\ -(-1)^p \mathbb{I}_{\{k_1=k_2=\ldots=k_p=i-K(2^p-2)\}}\ , & \text{if } K(2^p-2)+1 \le i \le K(2^p-1) \end{cases} \tag{6}$$

$U_{i,k_i}(t) = U_{j,k_j}(t)$.[2] Then, the expected payoff is maximized by putting different sniffers to different channels, i.e., $k_i \neq k_j$, $1 \le i < j \le p$. It can be proved that it is strictly better to put different sniffers to different channels. In this case it suffices to estimate $P(U_{ik} > 0)$, i.e., a total of $K \cdot p$ parameters. The problem then becomes essentially the multi-armed bandit problem with multiple plays considered in a number of previous works [26], [21], [22].

Example 2 (Independent sniffers): The opposite case is when $U_{i,k_i}(t) \neq U_{j,k_j}(t)$ whenever $i \neq j$ and when one of $U_{i,k_i}(t)$ and $U_{j,k_j}(t)$ is nonzero. This happens when all sniffers are guaranteed to observe distinct users (e.g., they are far away from one another). Then, $\mathbb{I}_{\{U_{i_1,k_{i_1}}=U_{i_2,k_{i_2}}=\ldots=U_{i_s,k_{i_s}}>0\}} = 0$, $2 \le s \le p$, $1 \le i_1 <, \ldots, < i_s \le p$. Therefore, the number of parameters is reduced to $K \cdot p$ and each sniffer can decide independently which channel to monitor. Thus the, problem reduces to p independent K-arm bandit problems.

In practice, sniffers are deployed distributedly. Their observations are typically correlated but non-identical. This motives us to consider the optimal channel assignment in general configurations. The learning efficiency of a policy is evaluated in terms of its regret, which, following [17] is decomposed into two terms, the *sampling regret* R_n^π due to the play of suboptimal arms; and the *switching regret* SW_n^π, capturing the cost of switching assignments. The sampling regret is given by,

$$R_n^\pi = \mathbb{E}\left[\sum_{t=1}^{n}\left\{\max_{\mathbf{k}\in\mathcal{A}} \phi_{\mathbf{k}}^T\theta - \phi_{\mathbf{k}_t}^T\theta\right\}\right], \tag{9}$$

where $\mathbf{k}_t$ is the assignment selected at time t. To define the switching regret, let

$$S_n(\mathbf{k}) = \sum_{t=1}^{n} \mathbb{I}_{\{\mathbf{k}_t=\mathbf{k},\mathbf{k}_{t+1}\neq\mathbf{k}\}}$$

count the number of switches from the joint action $\mathbf{k}$ to some other action during the first n rounds. The switching regret is

$$SW_n^\pi = C_{sw}\sum_{\mathbf{k}\in\mathcal{A}} \mathbb{E}\left[S_n(\mathbf{k})\right], \tag{10}$$

where C_{sw} is the switching cost. Note we assume that the switching cost is constant across all joint actions. This is reasonable in a synchronous system where all sniffers coordinate the onsets of monitoring.

An optimal monitoring policy π determines a sequence of actions in $\mathbb{K}$ over time such that the expected *total regret* is minimized:

$$Q_n^\pi = R_n^\pi + SW_n^\pi. \tag{11}$$

C. Spanners

Since some arms reveal information about other arms, it might be possible to identify a restricted set $\mathcal{E} \subset \mathbb{K}$, which is much smaller than $\mathbb{K}$, so that playing only arms in $\mathcal{E}$ gives sufficient information to identify the optimal arm. A sufficient condition for this is that $\cup_{\mathbf{k}\in\mathcal{E}}\mathcal{M}_{\mathbf{k}} = \{1, \ldots, M\}$. This condition ensures that by choosing an appropriate arm in $\mathcal{E}$ any component of $X(t)$ can be observed, which is clearly sufficient to identify θ. Since exploration is generally costly, the set $\mathcal{E}$ is ideally chosen to be small. In the monitoring problem $\mathcal{E}$ can be chosen to be $\mathcal{E} = \{\,(k, \ldots, k) : 1 \le k \le K\,\}$, i.e. all the sniffers assigned to the same channel to cover $(2^p - 1)$ parameters, whose cardinality is $K \ll K^p = |\mathbb{K}|$. The set $\mathcal{E}$ is called a *spanning set* or a *spanner* and its elements are called *spanner arms*.

IV. An Upper Confidence Bound (UCB)-based Policy

When the switching cost is not negligible, changing the joint action too often shall be discouraged. Most policies that consider switching costs utilize "block" sampling, namely, an action once selected persists for a period of time, called an *epoch*. Intuitively, the block length should be short initially when the uncertainty in the parameters is high (and thus more exploration), and increases as more knowledge is gained (and thus more exploitation). Define a function $\tau(r) = \lceil(1+\alpha)^r\rceil$, where $\alpha \ge 0$.

The algorithm first plays each arm in $\mathcal{E}$ once. From there on, the decision time instances for arm selection are denoted by t_j, $j = 1, \ldots, J_n$, where $t_1 = |\mathcal{E}|+1$ and J_n is the number of decision time instances up to time n. The quantities $t_j, j = 1, \ldots, J_n$ divide the time into epochs of length $l_j = t_{j+1} - t_j$, $j = 1, \ldots, J_n - 1$ to be defined next. At time t_j, the algorithm chooses

$$\mathbf{k}(t_j) = \operatorname*{argmax}_{\mathbf{k}\in\mathcal{E}} V_{\mathbf{k}}(t_j - 1), \tag{12}$$

where

$$\begin{aligned} V_{\mathbf{k}}(t_j - 1) &= \hat{\mu}_{\mathbf{k}}(t_j - 1) + \sum_{i\in\mathcal{M}_{\mathbf{k}}} \sqrt{\frac{\rho\log t_j}{N_i(t_j - 1)}}, & (13) \\ \hat{\mu}_{\mathbf{k}}(t_j - 1) &= \hat{\theta}(t_j - 1)^\top \phi_{\mathbf{k}}. & (14) \end{aligned}$$

[2] Clock synchronization among sniffers can be achieved online or offline using methods such as in [25].

Then, arm $\mathbf{k}(t_j)$ is played l_j times. From (13), we see that the choice of arm at time t_j is determined by two factors, namely, the estimated payoff in (14) (an approximation of the expected payoff in (7)) and a confidence bound determined by the number of times a component has been observed. Maximizing the first term gives exploitation, while the second term reflects the need for exploration. The choice in (12) trades off exploitation and exploration in order to reduce the sampling regret in (9).

Let $I(t_j) = \text{argmin}_{m \in \mathcal{M}_{\mathbf{k}(t)}} N_m(t_j - 1)$ (ties can be broken, say, in favor of the smallest index). Each component i is associated with an epoch counter $r_i(t)$ initialized to zero. At time t_j, the epoch counter of component $I(t_j)$ is updated as $r_{I(t_j)}(t_j) = r_{I(t_j)}(t_j - 1) + 1$, and remains the same for the rest of the epoch. The epoch length is given by $l_j = \tau(r_{I(t_j)}(t_j)) - \tau(r_{I(t_j)}(t_j) - 1)$. In other words, the epoch length is associated with the epoch counter of the least visited component. Note that the parameters are updated using (8) in each time slot after every observation, while the decisions are changed at the end of epochs only.

Theorem 2: Choose any ρ that satisfies $\rho > 1/1.99$. Then, there exists a constant $C > 0$ (which may depend on ρ) such that for all $n \geq 1$, the expected regret of our algorithm satisfies

$$Q_n^{\text{UCB}} \leq 4M\Delta_{\max}\left(\max_{\mathbf{k}:\Delta_{\mathbf{k}}>0} \frac{M_{\mathbf{k}}}{\Delta_{\mathbf{k}}}\right)^2 \rho \log n + M \log_{1+\alpha} \log n + C,$$

where $\Delta_{\max} = \max_{\mathbf{k}} \Delta_{\mathbf{k}}$.

Proof. Details of the proof can be found in Appendix B. The proof of the sampling regret is an adaption of the proof of Theorem 3 in [24]. The key difference lies in handling the changes in epoch length. The $\log \log$ form of the switching regret comes from the facts that the epoch lengths grow exponentially, and at the end of each epoch, suboptimal arms are chosen less and less likely over time. That the epoch lengths are chosen based on the counter of the least visited component ensures that the algorithm does not spend too much time on a suboptimal assignment. ■

Note that in the proof no attempt was made to optimize the constants. In the algorithm, the growth of the epoch length is relatively slow. The epoch counter is only updated for the component least visited in the chosen arm. Alternatively, one may track more closely the exact number of times a component have been visited.

V. Numerical Results

In this section, we illustrate the performance of the proposed UCB algorithm of Section IV and the UCB algorithm of [24] using numerical simulations.

In the simulations, wireless users are placed randomly in a 2-D plane. The area is partitioned into hexagon cells with circumcircle of radius 86 meters. Each cell is associated with a base station operating in a channel (and so are the users in the cell). The channel to base station assignment ensures that *no neighboring cells use the same channel*. Sniffers are deployed in a grid formation separated by a distance of 100 meters, with a coverage radius of 120 meters. A snapshot of this synthetic arrangement is shown in Figure 1. Wireless users are placed uniformly at random over the area. The transmission probability of users is selected uniformly from [0, 0.06], resulting in an average busy probability of 0.2685 in each cell. We vary the number of cells from 4 to 12, and the number of sniffers from 3 to 6. The switching cost per time slot is set to be 0.4. Changes in the switching cost per time slot will only change the switching regret by constant factors.

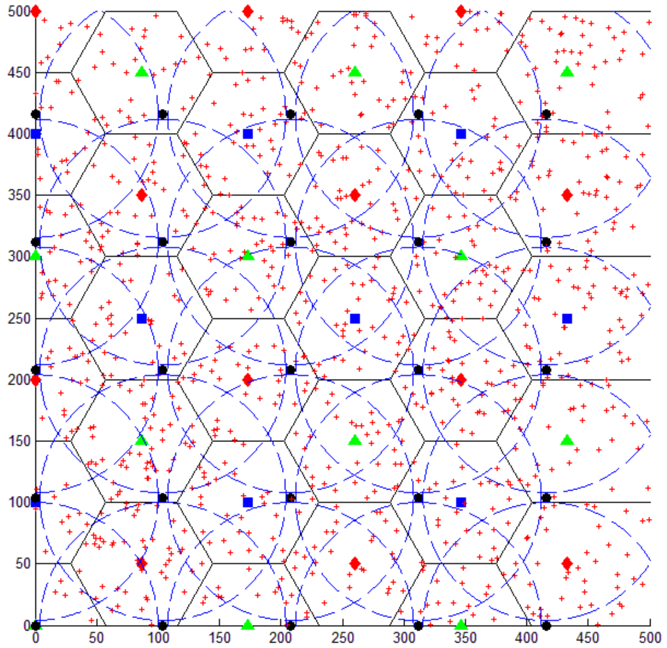
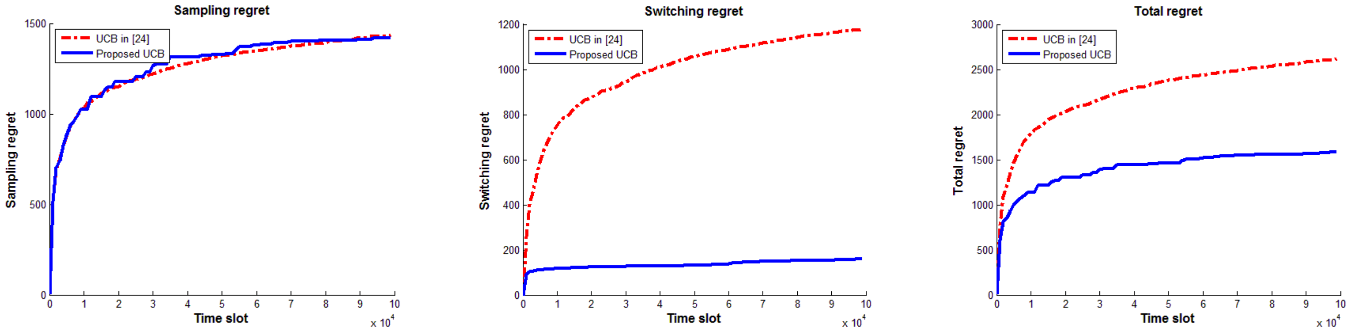


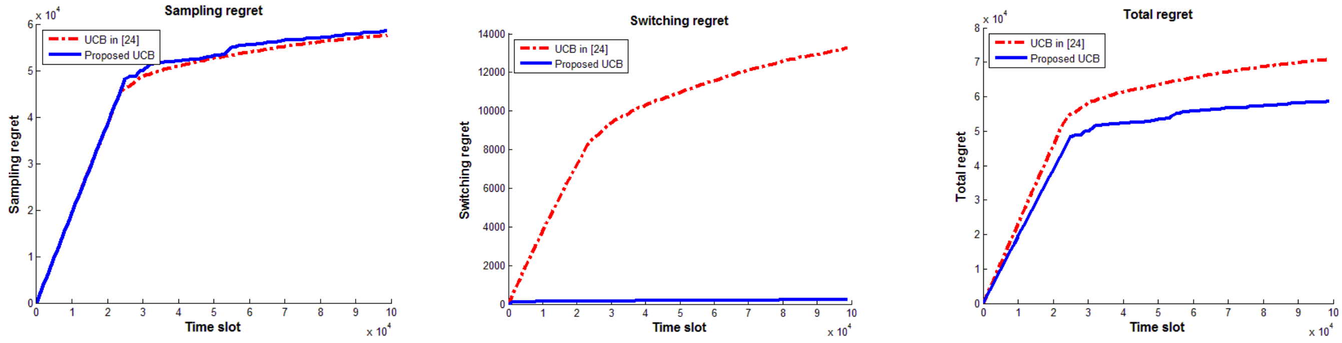
Fig. 1: Hexagonal layout with users(red '+'), sniffers (black solid dots), access points (APs) (at the center of each hexagonal cell), and channels of each cell (in different colors)

Figure 2 and 3 show the sampling and switching regrets of the proposed method and the method of [24] over time. It can be seen that in the two scenarios, the proposed method achieves similar sampling regrets as that of [24] (Figure 2(a), 3(a)) even though the new method explores less often due to the use of epochs. When comparing the switching regrets (Figure 2(b), 3(b)) and total regrets (Figure 2(c), 3(c)), the proposed method clearly outperforms the previous method. The differences in switching regrets are more pronounced when the number of sniffers increases from 3 to 6. Recalling that the number of assignments grows exponentially with the number of sniffers, we can explain this by noting that more switching is likely to occur when the number of possible assignments is larger, especially in the initial phase of learning. Furthermore, we observe the growth of the switching regrets in the proposed algorithm is much slower – roughly at the rate of $O(\log \log n)$, while that of the previous method grows roughly as $O(\log n)$. Finally, it should be noted that the computation complexity of the both algorithms grow exponentially with respect to the number of sniffers (which may be unavoidable due to the NP-hardness of the MEC problem), but the computation time is independent of the number of users. Low complexity algorithms can be devised, though at the expense of giving up



(a) Sampling regret (b) Switching regret (c) Total regret

Fig. 2: Comparison of two algorithms over the configuration of 4 APs using 3 channels, 3 sniffers and 129 users



(a) Sampling regret (b) Switching regret (c) Total regret

Fig. 3: Comparison of two algorithms over the configuration of 12 APs using 3 channels, 6 sniffers and 328 users

TABLE I: Total computation time (mins) on a desktop PC with Intel core i7-2600 CPU 3.4GHz and 8GB memory

Configuration	Proposed method	Method in [24]
4 APs, 3 channels, 3 sniffers	0.64	14.68
12APs, 3 channels, 6 sniffers	21.9	1868.5

sublinear regrets [27].

In all scenarios, the two algorithms have comparable sampling regrets. This is expected because they use the same formula to find the best arm to play whenever the system needs to make a new decision. However, using epochs, the proposed method incurs much smaller switching costs than the algorithm in [24], thus incurring lower total regret. Interestingly, we also find the overall computation time (Table I) is much less since the computation time is proportional to the number of times that (12) has to be evaluated.

VI. Conclusion and Future Work

In this paper, we considered the problem of optimally assigning p sniffers to K channels to monitor the transmission activities in a multi-channel wireless network. A new Upper Confident Bound policy is proposed that learns the user activities while making channel assignment decisions in a sequential manner. The key technique to combat switching costs is to reduce the frequency of switching and stay with one action sufficiently long. For this we proposed a specific method, which is shown to achieve logarithmic sampling regret in the number of time slots with a term sub-linear in cardinality of the action space, and a switching regrets that grow the order of $O(\log\log(n))$ where n denotes the number of time slots.

The generalization of the results to the following cases is possible: *(i)* $X_i(t)$ is sub-Gaussian with known tail behavior (e.g., $X_i(t)$ are bounded with known bounds), *(ii)* $\phi_{\mathbf{k}} \in \{0,1\}^M$. Other possible future work includes extension to non-stationarity environments where the statistics of user activities change over time, e.g., as a the result of mobility (this could be done, for instance, along the line of work of [28]), and the consideration of adversarial settings [16], [29].

Appendix A
Tail Probability Bounds

The following lemma generalizes Hoeffding's inequality to sums with a random number of terms. The lemma in the form presented here can be found as Theorem 18 of [28] (a similar statement, generalizing Bernstein's inequality can be extracted from [30]).

Lemma 3: Let $(\mathcal{F}_t; t \geq 0)$ be a filtration. Let $(X_t; t \geq 1)$ be an i.i.d. sequence taking values in some interval of length B. Let $\varepsilon_t \in \{0,1\}$ be a binary sequence. Assume that X_t is

$\mathcal{F}_t$-measurable and ε_t is $\mathcal{F}_{t-1}$-measurable ($t \geq 1$). Let $N_n = \sum_{t=1}^n \varepsilon_t$, $\overline{X}_n = \sum_{t=1}^n \varepsilon_t X_t / N_n$. Then, for any $n \geq 1$, $\eta > 0$,

$$\mathbb{P}\left(\overline{X}_n > \mathbb{E}\left[X_1\right] + z\sqrt{\frac{1}{N_n}}, N_n \geq 1\right) \leq \frac{\log n}{\log(1+\eta)} \exp\left(-\frac{2z^2}{B^2}\left(1 - \frac{\eta^2}{16}\right)\right).$$

In particular, when $\eta = 0.3$,

$$\mathbb{P}\left(\overline{X}_n > \mathbb{E}\left[X_1\right] + \frac{z}{\sqrt{N_n}}, N_n \geq 1\right) \leq \lceil 4 \ln n \rceil \exp\left(-\frac{1.99z^2}{B^2}\right)$$

Now, we consider a multi-dimensional generalization of this result:

Lemma 4: Let $(\mathcal{F}_t; t \geq 0)$ be a filtration. Let $(X_t; t \geq 1)$ be an i.i.d. sequence taking values in $\mathbb{R}^M$ such that X_{ti}, the i^{th} component of X_t, takes values in some interval of length B. Define $\mu = \sum_{i=1}^M \mathbb{E}\left[X_{1i}\right]$. Let $\varepsilon_t \in \{0,1\}^M$ be an M-dimensional binary sequence. Assume that X_t is $\mathcal{F}_t$-measurable and ε_t is $\mathcal{F}_{t-1}$-measurable ($t \geq 1$). Let $N_{ni} = \sum_{t=1}^n \varepsilon_{ti}$, $\overline{X}_{ni} = N_{ni}^{-1} \sum_{t=1}^n \varepsilon_{ti} X_{ti}$ and $\overline{X}_n = \sum_{i=1}^M \overline{X}_{ni}$. Then, for any $n \geq 1$,

$$\mathbb{P}\left(\overline{X}_n > \mu + z \sum_{i=1}^M \sqrt{\frac{1}{N_{ni}}},\, N_{n1}, \ldots, N_{nM} \geq 1\right) \leq M \lceil 4 \ln n \rceil \, \exp\left(-\frac{1.99z^2}{B^2}\right).$$

Proof. Let p denote the probability to be bounded and let $\mu_i = \mathbb{E}\left[X_{1i}\right]$. Then,

$$p \leq \sum_{i=1}^M \mathbb{P}\left(\overline{X}_{ni} > \mu_i + z\sqrt{\frac{1}{N_{ni}}},\, N_{ni} \geq 1\right).$$

The result then follows by applying Lemma 3 to each of the M terms on the right-hand side. ■

The next result can be extracted from [4] (with a slight improvement). The setting is similar to that of Lemma 3 with the deviation from the mean as a deterministic number.

Lemma 5: Let $(\mathcal{F}_t; t \geq 0)$ be a filtration. Let $(X_t; t \geq 1)$ be an i.i.d. sequence taking values in some interval of length 1. Let $\varepsilon_t \in \{0,1\}$ be a binary sequence. Assume that X_t is $\mathcal{F}_t$-measurable and ε_t is $\mathcal{F}_{t-1}$-measurable ($t \geq 1$). Let $N_n = \sum_{t=1}^n \varepsilon_t$, $\overline{X}_n = \sum_{t=1}^n \varepsilon_t X_t / N_n$. Then, for any $n \geq 1$, $x > 0$, $z > 0$,

$$\mathbb{P}\left(\overline{X}_n > \mathbb{E}\left[X_1\right] + \frac{z}{2}\right) \leq \mathbb{P}\left(N_n < x\right) + \frac{2}{z^2} \exp\left(-\frac{\lceil x \rceil z^2}{2}\right).$$

Proof. We have

$$\mathbb{P}\left(\overline{X}_n > \mathbb{E}\left[X_1\right] + \frac{z}{2}\right) \leq \mathbb{P}\left(N_n < x\right) + \mathbb{P}\left(N_n \geq x, \overline{X}_n > \mathbb{E}\left[X_1\right] + \frac{z}{2}\right).$$

Now,

$$\mathbb{P}\left(N_n \geq x, \overline{X}_n > \mathbb{E}\left[X_1\right] + \frac{z}{2}\right) = \sum_{s=\lceil x \rceil}^n \mathbb{P}\left(N_n = s, \overline{X}_n > \mathbb{E}\left[X_1\right] + \frac{z}{2}\right).$$

Let $S_n = \sum_{t=1}^n \varepsilon_t X_t$. Define $\tau(s)$ as the first time when s values of X are observed: $\tau(s) = \min\{\, t \geq 1 : N_t = s \,\}$. Further, let $S^{(1)} = S_{\tau(1)}$, $S^{(2)} = S_{\tau(2)}$, ... Note that $S^{(k)}$ has exactly k terms and $S^{(k)}$ is an $\mathcal{F}^{(k)}$-adapted martingale, where $\mathcal{F}^{(k)} = \mathcal{F}_{\tau(k)-1}$ (the so-called the "optional skipping process"). Now, $\overline{X}_n = S_n / N_n = S^{(N_n)} / N_n$. Hence,

$$\begin{aligned}
&\mathbb{P}\left(N_n = s, \overline{X}_n > \mathbb{E}\left[X_1\right] + \frac{z}{2}\right) \\
&= \mathbb{P}\left(N_n = s, S^{(N_n)}/N_n > \mathbb{E}\left[X_1\right] + \frac{z}{2}\right) \\
&= \mathbb{P}\left(N_n = s, S^{(s)}/s > \mathbb{E}\left[X_1\right] + \frac{z}{2}\right) \\
&\leq \mathbb{P}\left(S^{(s)}/s > \mathbb{E}\left[X_1\right] + \frac{z}{2}\right).
\end{aligned}$$

By the Hoeffding-Azuma inequality, $\mathbb{P}\left(S^{(s)}/s > \mathbb{E}\left[X_1\right] + \frac{z}{2}\right) \leq \exp(-s\, z^2/2)$. Using $\sum_{s=u}^{\infty} e^{-\kappa u} \leq \kappa^{-1} e^{-\kappa u}$, which holds for any integer u and $\kappa > 0$, we obtain the desired result. ■

Appendix B
Proof of Theorem 2

As the total regret consists of the sampling regret and the switching regret, we derive a bound of each part separately.

We start by introducing the necessary notation. We denote by $T_{\mathbf{k}}(n)$ the number of times arm $\mathbf{k}$ is chosen up to time n (including time n): $T_{\mathbf{k}}(n) = \sum_{t=1}^n \mathbb{I}_{\{\mathbf{k}(t)=\mathbf{k}\}}$. We let $\mu^* = \max_{\mathbf{k}} \mu_{\mathbf{k}}$, $\Delta_k = \mu^* - \mu_{\mathbf{k}}$. Then, it is easy see that $\mathbb{E}\left[R_n^{\text{UCB1}}\right] = \sum_{\mathbf{k}} \Delta_{\mathbf{k}} \mathbb{E}\left[T_{\mathbf{k}}(n)\right] \leq (\max_{\mathbf{k}} \Delta_{\mathbf{k}})\, \mathbb{E}\left[\sum_{\mathbf{k}:\Delta_{\mathbf{k}}>0} T_{\mathbf{k}}(n)\right]$. Our goal is to develop a bound on $\mathbb{E}\left[\sum_{\mathbf{k}:\Delta_{\mathbf{k}}>0} T_{\mathbf{k}}(n)\right]$ which scales linearly with M rather that with $|\mathbb{K}|$.

Let $Z_i(t_j) = \mathbb{I}_{\{\mathbf{k}(t_j) \neq \mathbf{k}^*, I(t_j)=i\}}$, and $\tilde{T}_i(t) = \tilde{T}_i(t-1) + Z_i(t_j), t_j \leq t < t_{j+1}$.[3] Note that $\sum_{\mathbf{k} \neq \mathbf{k}^*} T_{\mathbf{k}}(n) = \sum_i \tilde{T}_i(n)$, since exactly one of the counters is incremented on both sides when a suboptimal arm is chosen. Thus, it suffices to bound $\tilde{T}_i(n)$.

Therefore pick any index $1 \leq i \leq M$ and let u be an integer to be chosen later. We have $Z_i(t_j) = Z_i(t_j)\mathbb{I}_{\{\tilde{T}_i(t_j-1)>\tau(u)\}} + Z_i(t_j)\mathbb{I}_{\{\tilde{T}_i(t_j-1)\leq\tau(u)\}}$. Since $\sum_{j=1}^{J_n} Z_i(t_j)\mathbb{I}_{\{\tilde{T}_i(t_j-1)\leq\tau(u)\}} l_j \leq \tau(u)$, it suffices to deal with the first term, which we bound as follows:

$$\begin{aligned}
&Z_i(t_j)\mathbb{I}_{\{\tilde{T}_i(t_j-1)>\tau(u)\}} \\
&\leq \mathbb{I}_{\{V_{\mathbf{k}(t_j)}(t_j-1)>\mu^*, \tilde{T}_i(t_j-1)>\tau(u), I(t_j)=i\}} + \mathbb{I}_{\{V_{\mathbf{k}^*}(t_j-1)\leq\mu^*\}}.
\end{aligned}$$

Now, let the $\hat{\mathbf{k}}_{ij}$ be the arm played in the jth epoch out of the epochs j's where $I(t_{j'}) = i$, and $\hat{t}_{ij}$ is the time when

[3] We are using the assumption that there is a unique optimal arm $\mathbf{k}^*$. Note that this is assumed just for the sake of simplicity and the proof, at the price of a more complicated presentation, works without it.

such an epoch starts. Clearly, $\hat{\mathbf{k}}_{ij} = \mathbf{k}(\hat{t}_{ij})$. Denote $\delta(j) = \tau(j) - \tau(j-1)$. Thus,

$$\begin{aligned} \tilde{T}_i(n) &= 1 + \textstyle\sum_{j=1}^{J_n - 1} Z_i(t_j) l_j \\ &\leq 1 + \tau(u) + \sum_{j=u+1}^{J_{max}} Z_i(\hat{t}_{ij}) \mathbb{I}_{\{\tilde{T}_i(\hat{t}_{ij}-1) > \tau(u)\}} \delta(j), \end{aligned} \tag{15}$$

where J_{max} is the maximum possible number of epochs where $Z_i(t_j) = 1$. Clearly, $J_{max} \leq \lceil \frac{\log n}{\log(1+\alpha)} \rceil$. Therefore,

$$\begin{aligned} &\mathbb{E}\left[\tilde{T}_i(n)\right] \leq \tau(u) + 1 \\ &\quad + \sum_{j=u+1}^{J_{max}} \mathbb{P}\left(V_{\hat{\mathbf{k}}_{ij}}(\hat{t}_{ij}-1) > \mu^*, \tilde{T}_i(\hat{t}_{ij}-1) > \tau(u)\right) \delta(j) \\ &\quad + \sum_{j=u+1}^{J_{max}} \mathbb{P}\left(V_{\mathbf{k}^*}(\hat{t}_{ij}-1) \leq \mu^*\right) \delta(j). \end{aligned}$$

We will now show that both sums can be bounded logarithmically with respect to n, provided that u is sufficiently large.

The summand of the first sum is bounded as follows:

$$\begin{aligned} p_{1j} &\overset{\text{def}}{=} \mathbb{P}\left(V_{\hat{\mathbf{k}}_{ij}}(\hat{t}_{ij}-1) > \mu^*, \tilde{T}_i(\hat{t}_{ij}-1) > \tau(u), I(\hat{t}_{ij}) = i\right) \\ &\leq \mathbb{P}\Big\{\hat{\mu}_{\hat{\mathbf{k}}_{ij}}(\hat{t}_{ij}-1) > \mu_{\hat{\mathbf{k}}_{ij}} + \Delta_{\hat{\mathbf{k}}_{ij}} - c_{\hat{\mathbf{k}}_{ij}, \hat{t}_{ij}-1}, \\ &\qquad\qquad \tilde{T}_i(t-1) > \tau(u), I(\hat{t}_{ij}) = i\Big\} \end{aligned}$$

where $c_{\hat{\mathbf{k}}_{ij}, \hat{t}_{ij}-1} = \sqrt{\rho \log \hat{t}_{ij}} \sum_{m \in \mathcal{M}_{\hat{\mathbf{k}}_{ij}}} \sqrt{\frac{1}{N_m(\hat{t}_{ij}-1)}} \overset{\text{def}}{=} \sqrt{\rho \log \hat{t}_{ij}}\ W_{\hat{\mathbf{k}}_{ij}}(\hat{t}_{ij}-1)$. Now,

$$\begin{aligned} &\Delta_{\hat{\mathbf{k}}_{ij}} - c_{\hat{\mathbf{k}}_{ij}, \hat{t}_{ij}-1} \\ &= \sum_{m \in \mathcal{M}_{\hat{\mathbf{k}}_{ij}}} \left(\frac{\Delta_{\hat{\mathbf{k}}_{ij}}}{W_{\hat{\mathbf{k}}_{ij}}(\hat{t}_{ij}-1)} - \sqrt{\rho \log \hat{t}_{ij}} \right) \sqrt{\frac{1}{N_m(\hat{t}_{ij}-1)}}. \end{aligned}$$

We claim that under the condition that $\tilde{T}_i(\hat{t}_{ij}-1) > \tau(u)$ the largest value $W_{\hat{\mathbf{k}}_{ij}}(\hat{t}_{ij}-1)$ can take is bounded from above by $M_{\hat{\mathbf{k}}_{ij}}/\sqrt{\tau(u)}$. To see this note that $\tilde{T}_m(t-1) \leq N_m(t-1)$ holds for any m and t, because $N_m(\cdot)$ is always incremented when $\tilde{T}_m(\cdot)$ is incremented. Further, since $I(t) = \text{argmin}_{m \in \mathcal{M}_{\mathbf{k}(t)}} N_m(t-1)$, $N_{I(t)}(t-1) \leq N_m(t-1)$ holds for any $m \in \mathcal{M}_{\mathbf{k}(t)}$. Thus, for arbitrary $m \in \mathcal{M}_{\hat{\mathbf{k}}_{ij}}$, $\tau(u) < \tilde{T}_i(\hat{t}_{ij}-1) \leq N_i(\hat{t}_{ij}-1) \leq N_m(\hat{t}_{ij}-1)$. The claim then follows from the definition of $W_{\hat{\mathbf{k}}_{ij}}(\hat{t}_{ij}-1)$.

Hence,

$$\begin{aligned} &\Delta_{\hat{\mathbf{k}}_{ij}} - c_{\hat{\mathbf{k}}_{ij}, \hat{t}_{ij}-1} \\ &\geq \sum_{m \in \mathcal{M}_{\hat{\mathbf{k}}_{ij}}} \left(\frac{\Delta_{\hat{\mathbf{k}}_{ij}} \sqrt{\tau(u)}}{M_{\hat{\mathbf{k}}_{ij}}} - \sqrt{\rho \log \hat{t}_{ij}} \right) \sqrt{\frac{1}{N_m(\hat{t}_{-}1)}}. \end{aligned}$$

Further, $\frac{\Delta_{\hat{\mathbf{k}}_{ij}} \sqrt{\tau(u)}}{M_{\hat{\mathbf{k}}_{ij}}} - \sqrt{\rho \log \hat{t}_{ij}} \geq \sqrt{\rho \log n}$ holds for $1 \leq \hat{t}_{ij} \leq n$ if

$$\tau(u) \geq \left(2 \max_{\mathbf{k}:\Delta_{\mathbf{k}} > 0} \frac{M_{\mathbf{k}}}{\Delta_{\mathbf{k}}}\right)^2 \rho \log n.$$

Then, $\Delta_{\hat{\mathbf{k}}_{ij}} - c_{\hat{\mathbf{k}}_{ij}, t-1} \geq \sqrt{\rho \log n}\ W_{\hat{\mathbf{k}}_{ij}}(t-1)$ and thus

$$\begin{aligned} p_{1j} &\leq \mathbb{P}\left(\hat{\mu}_{\hat{\mathbf{k}}_{ij}}(t_{ij}-1) > \mu_{\hat{k}k_{ij}} + \sqrt{\rho \log n}\ W_{\hat{\mathbf{k}}_{ij}}(\hat{t}_{ij}-1)\right) \\ &\leq \sum_{\mathbf{k}} M_{\mathbf{k}} \lceil 4 \log n \rceil \exp\left(-1.99 \rho \log n\right), \end{aligned}$$

where the last inequality follows from the union bound and Lemma 4, which is presented in Appendix A.

The summand of the second sum can be bounded as follows:

$$\begin{aligned} p_{2j} &\overset{\text{def}}{=} \mathbb{P}\left(V_{\mathbf{k}^*}(\hat{t}_{ij}-1) \leq \mu^*\right) \\ &= \mathbb{P}\left(\hat{\mu}_{\mathbf{k}^*}(\hat{t}_{ij}-1) + c_{\mathbf{k}^*, \hat{t}_{ij}} \leq \mu^*\right) \\ &\leq \sum_{t=\tau(j)}^{n} \mathbb{P}\left(\hat{\mu}_{\mathbf{k}^*}(t-1) + c_{\mathbf{k}^*, t} \leq \mu^*\right) \\ &= \sum_{t=\tau(j)}^{n} \mathbb{P}\left(\hat{\mu}_{\mathbf{k}^*}(t-1) \leq \mu^* - \sqrt{\rho \log t} W_{k^*}(t)\right) \end{aligned}$$

The inequality is due to the fact that $t_{ij} \geq \tau(j)$ and the union bound. Using Lemma 4 again, we get that

$$\begin{aligned} p_{2j} &\leq \sum_{t=\tau(j)}^{n} M_{\mathbf{k}^*} \lceil 4 \log n \rceil t^{-1.99\rho} \\ &\leq M_{\mathbf{k}^*} \lceil 4 \log n \rceil \int_{\tau(j)}^{\infty} t^{-1.99\rho} \\ &= M_{\mathbf{k}^*} \lceil 4 \log n \rceil \tau(j)^{-1.99\rho + 1} \\ &\leq M_{\mathbf{k}^*} \lceil 4 \log n \rceil (1+\alpha)^{(-1.99\rho+1)j} \end{aligned}$$

Putting together the inequalities, for n sufficiently large, we have

$$\begin{aligned} &\mathbb{E}\left[\tilde{T}_i(n)\right] \leq \tau(u) + \sum_{j=u+1}^{J_{max}} (p_{1j} + p_{2j}) \delta(j) \\ &\leq \left(2 \max_{\mathbf{k}:\Delta_{\mathbf{k}} > 0} \frac{M_{\mathbf{k}}}{\Delta_{\mathbf{k}}}\right)^2 \rho \log n \\ &+ \sum_{\mathbf{k}} M_{\mathbf{k}} \lceil 4 \log n \rceil\, n^{-1.99\rho} \sum_{j=u+1}^{J_{max}} \delta(j) \\ &+ M_{\mathbf{k}^*} \lceil 4 \log n \rceil \sum_{j=u+1}^{J_{max}} (1+\alpha)^{(-1.99\rho+1)j} \delta(j) \\ &\leq \left(2 \max_{\mathbf{k}:\Delta_{\mathbf{k}} > 0} \frac{M_{\mathbf{k}}}{\Delta_{\mathbf{k}}}\right)^2 \rho \log n \\ &+ \sum_{\mathbf{k}} M_{\mathbf{k}} \lceil 4 \log n \rceil\, n^{-1.99\rho} \lceil (1+\alpha)^{J_{max}} - (1+\alpha)^{u} \rceil \\ &+ M_{\mathbf{k}^*} \lceil 4 \log n \rceil \sum_{j=u+1}^{J_{max}} (1+\alpha)^{(-1.99\rho+2)j} \\ &\leq \left(2 \max_{\mathbf{k}:\Delta_{\mathbf{k}} > 0} \frac{M_{\mathbf{k}}}{\Delta_{\mathbf{k}}}\right)^2 \rho \log n + \sum_{\mathbf{k}} M_{\mathbf{k}} \lceil 4 \log n \rceil\, n^{-1.99\rho+1} \\ &+ C' M_{\mathbf{k}^*} \lceil 4 \log n \rceil n^{(-1.99\rho+2)}, \end{aligned}$$

where C' is a proper defined constant dependent on α. Clearly, if $\rho \geq 2/1.99$, the last two terms are $o(1)$.

To this end, we have proved that the sampling regret grows logarithmically with n. Next, we analyze the asymptotic property of the switching regret. Clearly, the number of switching is bounded by the number of times a suboptimal arm is played, which is clearly logarithmic in time. However, a tighter bound can be obtained by taking into account the epoch length. Let $\Psi_i(0) = 0$ and

$$\Psi_i(t) = \left\{ \begin{array}{ll} \Psi_i(t-1) + Z_i(t), & t = t_1, t_2, \ldots \\ \Psi_i(t-1), & else \end{array} \right.$$

Recall that $Z_i(t_j) = \mathbb{I}_{\{k(t_j) \neq k^*, I(t_j) = i\}}$. Namely, $\Psi_i(t)$ is the number of epochs the ith component incurred till time t, where i is the least visited component in the chosen arm. Clearly, the switching cost is bounded by $C_{sw} \sum_i \Psi_i$. Thus,

$$\begin{array}{rcl} \Psi_i(n) & = & 1 + \sum_{j=1}^{J_n} Z_i(t_j) \\ & \leq & 1 + u + \displaystyle\sum_{j=u+1}^{J_{max}} Z_i(\hat{t}_{ij}) \mathbb{I}_{\{\tilde{T}_i(\hat{t}_{ij}-1) > \tau(u)\}}. \end{array} \qquad (16)$$

Therefore,

$$\begin{aligned} &\mathbb{E}\left[\Psi_i(n)\right] \leq 1 + u \\ &\quad + \sum_{j=u+1}^{J_{max}} \mathbb{P}\left(V_{\hat{\mathbf{k}}_{ij}}(\hat{t}_{ij} - 1) > \mu^*, \tilde{T}_i(\hat{t}_{ij} - 1) > \tau(u), I(t) = i\right) \\ &\quad + \sum_{j=u+1}^{J_{max}} \mathbb{P}\left(V_{\mathbf{k}^*}(\hat{t}_{ij} - 1) \leq \mu^*\right). \end{aligned}$$

Following the same argument as the proof of sampling regret and picking

$$u = \log_{1+\alpha}\left(2 \max_{\mathbf{k}:\Delta_{\mathbf{k}}>0} \frac{M_{\mathbf{k}}}{\Delta_{\mathbf{k}}}\right)^2 \rho \log n,$$

we can prove that,

$$\mathbb{E}\left[\Psi_i(n)\right] \leq \log_{1+\alpha} \log n + C''.$$

In summary, the sampling regret grows logarithmic with time, while the switching regret grows in $\log\log$ fashion. Combining the sampling and the switching regret, we complete the proof of the theorem.

Acknowledgment

This work was in part supported by US National Science Foundation under Grant number CNS-1117560, and Natural Sciences and Engineering Research Council of Canada (NSERC) under project ID 2012R00435.

References

[1] A. Sharma and E. M. Belding, "FreeMAC: Framework for multi-channel mac development on 802.11 hardware", in *Proceedings of the ACM Workshop on Programmable Routers for Extensible Services of Tomorrow*, Seattle WA, Aug. 2008, pp. 69–74.

[2] H. Robbins, "Some aspects of the sequential design of experiments", *Bulletin of the American Mathematical Society*, vol. 58, no. 1952, pp. 527–535, 1952.

[3] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules", *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4–22, Dec. 1985.

[4] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem", *Journal of Machine Learning*, vol. 47, no. 2–3, pp. 235–256, Jun. 2002.

[5] A. Balachandran, G. M. Voelker, P. Bahl, and P. V. Rangan, "Characterizing user behavior and network performance in a public wireless LAN", *SIGMETRICS - Performance Evaluation Review*, vol. 30, no. 1, pp. 195–205, Jun. 2002.

[6] T. Henderson, D. Kotz, and I. Abyzov, "The changing usage of a mature campus-wide wireless network", in *Proceedings of the 10th Annual International Conference on Mobile Computing and Networking*, Philadelphia PA, Sep. 2004, pp. 187–201.

[7] J. Yeo, M. Youssef, and A. Agrawala, "A framework for wireless LAN monitoring and its applications", in *Proceedings of the 3rd ACM Workshop on Wireless Security*, Philadelphia PA, Oct. 2004, pp. 70–79.

[8] M. Rodrig, C. Reis, R. Mahajan, D. Wetherall, and J. Zahorjan, "Measurement-based characterization of 802.11 in a hotspot setting", in *Proceedings of the 2005 ACM SIGCOMM Workshop on Experimental Approaches to Wireless Network Design and Analysis*, Philadelphia PA, Aug. 2005, pp. 5–10.

[9] Y. C. Cheng, J. Bellardo, P. Benkö, A. C. Snoeren, G. M. Voelker, and S. Savage, "Jigsaw: Solving the puzzle of enterprise 802.11 analysis", in *Proceedings of the 2006 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, Pisa Italy, Sep. 2006, pp. 39–50.

[10] D. H. Shin and S. Bagchi, "Optimal monitoring in multi-channel multi-radio wireless mesh networks", in *Proceedings of the 10th ACM International Symposium on Mobile Ad hoc Networking and Computing*, New Orleans LA, May 2009, pp. 229–238.

[11] C. Chekuri and A. Kumar, "Maximum coverage problem with group budget constraints and applications", in *Approximation, Randomization, and Combinatorial Optimization*, Cambridge MA, Aug. 2004, vol. 3122, pp. 72–83.

[12] A. Chhetri, H. Nguyen, G. Scalosub, and R. Zheng, "On quality of monitoring for multi-channel wireless infrastructure networks", in *Proceedings of the 11th ACM International Symposium on Mobile Ad hoc Networking and Computing*, Chicago IL, Sep. 2010, pp. 111–120.

[13] R. Agrawal, "Sample mean based index policies with o(log n) regret for the multi-armed bandit problem", *Advances in Applied Probability*, vol. 27, no. 11, pp. 1054–1078, Dec. 1995.

[14] P. Rusmevichientong and J. N. Tsitsiklis, "Linearly parameterized bandits", *Mathematics of Operations Research*, vol. 35, no. 2, pp. 395–411, May 2010.

[15] V. Dani, T. P. Hayes, and S. M. Kakade, "Stochastic linear optimization under bandit feedback", in *Proceedings of the 21st Annual Conference on Learning Theory*, Helsinki Finland, Jul. 2008, pp. 355–366.

[16] N. Cesa-Bianchi, G. Lugosi, and G. Stoltz, "Regret minimization under partial monitoring", *Mathematics of Operations Research*, vol. 31, no. 3, pp. 562–580, Aug. 2006.

[17] R. Agrawal, M.V. Hedge, and D. Teneketzis, "Asymptotically efficient adaptive allocation rules for the multiarmed bandit problem with switching cost", *IEEE Transactions on Automatic Control*, vol. 33, no. 10, pp. 899–906, Oct. 1988.

[18] T. Jun, "A survey on the bandit problem with switching costs", *De Economist*, vol. 152, no. 4, pp. 513–541, Dec. 2004.

[19] L. Lai, H. E. Gamal, H. Jiang, and H. V. Poor, "Cognitive medium access: exploration, exploitation and competition", *IEEE Transactions on Mobile Computing*, vol. 10, no. 2, pp. 239–253, Feb. 2007.

[20] L. Lai, H. Jiang, and H. V. Poor, "Medium access in cognitive radio networks: A competitive multi-armed bandit framework", in *Proceedings of the 42nd Asilomar Conference on Signals, Systems and Computers*, Pacific Grove CA, Oct. 2008, pp. 98–102.

[21] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players", *IEEE Transactions on Signal Processing*, vol. 58, no. 11, pp. 5667–5681, Nov. 2010.

[22] A. Anandkumar, N. Michael, and A. K. Tang, "Opportunistic spectrum access with multiple users: Learning under competition", in *Proceedings of IEEE International Conference on Computer Communications*, San Deigo CA, Mar. 2010, pp. 803–811.

[23] Y. Gai, B. Krishnamachari, and R. Jain, "Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation", in *IEEE Symposium on New Frontiers in Dynamic Spectrum*, Singapore Singapore, Apr. 2010, pp. 1–9.

[24] P. Arora, C. Szepesvári, and R. Zheng, "Sequential learning for optimal monitoring of multi-channel wireless networks", in *Proceedings of IEEE International Conference on Computer Communications*, Shanghai China, Apr. 2011, pp. 1152–1160.

[25] J. Elson, L. Girod, and D. Estrin, "Fine-grained network time synchronization using reference broadcasts", *SIGOPS Operating Systems Review*, vol. 36, no. SI, pp. 147–163, Dec. 2002.

[26] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: I.i.d. rewards", *IEEE Transactions on Automatic Control*, vol. 32, no. 11, pp. 968–976, Nov. 1987.

[27] R. Zheng, T. Le, and Z. Han, "Approximate online learning for passive monitoring of multi-channel wireless networks", in *Proceedings of IEEE International Conference on Computer Communications*, Turin Italy, Apr. 2013.

[28] A. Garivier and E. Moulines, "On upper-confidence bound policies for non-stationary bandit problems", Tech. Rep., LTCI, Dec 2008.

[29] N. Cesa-Bianchi and G. Lugosi, "Combinatorial bandits", in *Proceedings of the 22nd Annual Conference on Learning Theory*, Quebec Canada, Jun. 2009, pp. 237–246.

[30] J-Y Audibert, R. Munos, and C. Szepesvári, "Tuning bandit algorithms in stochastic environments", in *Proceedings of the 18th international conference on Algorithmic Learning Theory*, Sendai Japan, Oct. 2007, pp. 150–165.

Thanh Le received the B.E. degree in electronics and telecommunications from Hanoi University of Technology in 2008 and the M.S. degree in electrical engineering from the University of Houston in 2013. Since October 2013, he has been a research member at the Viettel R&D Center. His current research interests include wireless communication systems, adaptive algorithms in machine learning, and optimization.

Csaba Szepesvári (SM'09) received his PhD in 1999 from "József Attila" University, Szeged, Hungary. After his PhD, he has spent 8 years in the software industry, after which he joined the Computer and Automation Research Institute of the Hungarian Academy of Sciences as a Senior Researcher where he founded the Machine Learning Research Group. Since 2006 he is with the Department of Computing Science of the University of Alberta and a Principal Investigator of the Alberta Innovates Center for Machine Learning. He has published about 150 journal and conference papers and two books. He holds various editorial positions at leading control and machine learning journals.

Rong Zheng (S03-M04-SM10) received her Ph.D. degree from Dept. of Computer Science, Uni- versity of Illinois at Urbana-Champaign and earned her M.E. and B.E. in Electrical Engineering from Tsinghua University, P.R. China. She is on the faculty of the Department of Computing and Software, McMaster University. She was with University of Houston between 2004 and 2012. Rong Zhengs research interests include network monitoring and diagnosis, cyber physical systems, and sequential learning and decision theory. She received the National Science Foundation CAREER Award in 2006. She serves on the technical program committees of leading networking conferences including INFOCOM, ICDCS, ICNP, etc. She served as a guest editor for EURASIP Journal on Advances in Signal Processing, Special issue on wireless location estimation and tracking, Elsevlers Computer Communications Special Issue on Cyber Physical Systems; and Program co-chair of WASA12, CPSCom12, MobileHealth'14.