Spectrum Bandit Optimization

Marc Lelarge
INRIA – Ecole Normale Superieure
Email: marc.lelarge@ens.fr

Alexandre Proutiere
KTH Royal Institute of Technology
Email: alexandre.proutiere@ee.kth.se

M. Sadegh Talebi
KTH Royal Institute of Technology
Email: mstms@kth.se

***Abstract*—We consider the problem of allocating radio channels to links in a wireless network. Links interact through interference, modelled as a conflict graph (i.e., two interfering links cannot be simultaneously active on the same channel). We aim at identifying the channel allocation maximizing the total network throughput over a finite time horizon. Should we know the average radio conditions on each channel and on each link, an optimal allocation would be obtained by solving an Integer Linear Program (ILP). When radio conditions are unknown a priori, we look for a sequential channel allocation policy that converges to the optimal allocation while minimizing on the way the throughput loss or *regret* due to the need for exploring sub-optimal allocations. We formulate this problem as a generic linear bandit problem, and analyze it in a stochastic setting where radio conditions are driven by a i.i.d. stochastic process, and in an adversarial setting where radio conditions can evolve arbitrarily. We provide, in both settings, algorithms whose regret upper bounds outperform those of existing algorithms.**

I. Introduction

Dynamic Spectrum Access techniques play an increasingly important role in wireless communication networks where large parts of the radio spectrum can be used and shared. In this paper we consider networks where transmitters can share a potentially large number of frequency bands or channels for transmission. In such networks, transmitters should be able to select a channel (i) that is not selected by neighbouring transmitters to avoid interference, and (ii) that offers good radio conditions. A spectrum allocation is defined by the channels assigned to the various transmitters or links, and our fundamental objective is to devise an optimal allocation, i.e., maximizing the network-wide throughput. If the radio conditions on each link and on each channel were known, the problem would reduce to a combinatorial optimization problem, and more precisely to an Integer Linear Program. For example, if all links interfere each other (no two links can be active on the same channel), a case referred to as *full interference*, the optimal spectrum allocation problem is an instance of a Maximum Weighted Matching in a bipartite graph (vertices on one side correspond to links and vertices on the other side to channels; the weight of an edge, i.e., a (link, channel) pair, represents the radio conditions for the corresponding link and channel). In practice, the radio conditions on the various channels are not known a priori, and they evolve over time in an unpredictable manner. Hence, we need to dynamically learn and track the optimal spectrum allocation. This task is complicated by the fact that we can gather information about the radio conditions for a particular (link, channel) pair only by actually including this pair in the selected spectrum allocation. We face a classical exploration vs. exploitation trade-off problem: we need to exploit the allocation that provided the best performance so far whilst constantly exploring other allocations. We model our sequential spectrum allocation problem as a linear multi-armed bandit problem. The challenge in this problem resides in the very high dimension of the decision action space: the size of the set of possible allocations exponentially grows with the number of links and channels.

We study generic linear bandit problems in two different settings, and apply our results to sequential spectrum allocation problems. In the stochastic setting, we assume that the radio conditions for each (link,channel) pair evolve over time according to a stationary (actually i.i.d.) process whose average is unknown. This first model is instrumental to represent scenarios where the average radio conditions evolve relatively slowly, in the sense that the spectrum allocation can be updated many times before this average exhibits significant changes. In the adversarial setting, the radio conditions evolve arbitrarily, as if they were generated by an *adversary*. This model is relevant when the channel allocation cannot be updated at the same pace as radio conditions change. In both settings, as usual for bandit optimization problems, we measure the performance of a given sequential decision policy through the notion of *regret*, defined as the difference of the performance obtained over some finite time horizon under the best static allocation and under the given sequential allocation selection policy. We make the following contributions:

For adversarial linear bandit problems: We propose ColorBand, a new sequential decision policy, and derive an upper bound on its regret. For example in the full interference case, when the number of channels c and the number of links n are identical, this bound scales as $\sqrt{n^3 \log(n) T}$ where T denotes the time horizon – this improves over the upper bounds for the best previously known algorithms [1], which scale as $\sqrt{n^5 \log(n) T}$.

For stochastic linear bandit problems: (a) We derive an asymptotic lower bound for the regret of any sequential decision policy, and show that this bound typically scales as $(n \times c) \log(T)$. (b) We propose a simple sequential decision policy, and provide upper bounds on its regret. In full interference scenario, when $n = c$, this bound scales as $n^3 \log(T)$, which significantly improves over bounds of existing algorithms [2] (the latter scales as $n^5 \log(T)$).

Proofs are omitted due to space constraints and can be found in [3].

Related work. Spectrum allocation has attracted considerable attention recently, mainly due to the increasing popularity of cognitive radio systems. In such systems, transmitters have to explore spectrum to find frequency bands free from primary users. This problem can also be formulated as a bandit problem, see e.g. [4], [5], but is simpler than our problem (in cognitive radio systems, there are basically c unknown variables, each representing the probability that a channel is free). Spectrum sharing problems similar to ours have been very recently investigated in [2], [6]. Both aforementioned papers restrict their analysis to the case of full interference, and even in this scenario, we obtain better regret bounds. As far as we know, adversarial bandit problems have not been considered to model spectrum allocation issues.

There is a vast literature on bandit problems, both in the stochastic and adversarial settings; see [7] for a quick survey. Surprisingly, there are very little work on linear bandit with discrete action space in the stochastic setting, and existing results are derived for very simple problems only; see e.g. [8] and references therein. In contrast, the problem has received more attention in the adversarial setting [1], [9]–[12]. The algorithm we devise yields a regret upper bound that beats all known bounds of algorithms previously proposed in the literature.

II. Preliminaries

A. Network and interference model

Consider a network consisting of n links indexed by $i \in [n] = \{1, \ldots, n\}$. Each link can use one of the c available radio channels indexed by $j \in [c]$. Interference is represented as a conflict graph $G = (V, E)$ where vertices are links, and edges $(i, i') \in E$ if links i and i' interfere, i.e., these links cannot be simultaneously active on the same channel. A spectrum allocation is represented as a configuration $M \in \{0,1\}^{n \times c}$, where $M_{ij} = 1$ if and only if link-i transmitter uses channel j. Configuration M is feasible if (i) for all i, the corresponding transmitter uses at most one channel, i.e., $\sum_{j \in [c]} M_{ij} \in \{0, 1\}$; (ii) two interfering links cannot be active on the same channel, i.e., for all $i, i' \in [n]$, $(i, i') \in E$ implies for all $j \in [c]$, $M_{ij} M_{i'j} = 0$.[1] Let $\mathcal{M}$ be the set of feasible configurations. For $M \in \mathcal{M}$, if link i is active, we denote by $M(i)$ the channel allocated to this link, i.e., $M_{ij} = 1$ iff $j = M(i)$. We also write $(i, j) \in M$ for $i \in [n]$ and $j \in [c]$, if $j = M(i)$. In the following we denote by $\mathcal{K} = \{\mathcal{K}_\ell, l \in [k]\}$ the set of maximal cliques of the interference graph G. We also introduce $K_{\ell i} \in \{0, 1\}$ such that $K_{\ell i} = 1$ if and only if link i belongs to the maximal clique $\mathcal{K}_\ell$.

Of particular interest is the *full interference* case, where the conflict graph G is complete. In such a case, a feasible configuration M is a matching in the complete bipartite graph $([n], [c])$, where on one side we have the set $[n]$ of links, and on the other side, the set $[c]$ of radio channels.

[1] This model assumes that the interference graph is the same over the various channels. Our analysis and results can be extended to the case where one has different interference graphs depending on the channel.

B. Fading

To model the way radio conditions evolve over time on the various channels, we consider a time slotted system, where the duration of a slot corresponds to the transmission of a fixed number m of packets. The channel allocation, i.e., the chosen feasible configuration, may change at the beginning of each slot. We denote by $r_{ij}(t)$ the proportion of packets successfully transmitted during slot t when link-i transmitter selects channel j for transmission in this slot and in the absence of interference. Depending on the ability of transmitters to rapidly switch channels, we introduce two settings.

In the *adversarial setting*, $r_{ij}(t) \in \{0, 1/m, \ldots, 1\}$ (our analysis holds in fact for any $r_{ij}(t) \in [0, 1]$) can be arbitrary as if it was generated by an *adversary*, and unknown in advance. This setting is useful to model scenarios where the duration of a slot is comparable to or smaller than the channel coherence time. In such scenarios, we assume that the channel allocation cannot change at the same pace as the radio conditions on the various links, which is of interest in practice, when the radios cannot rapidly change channels.

In the *stochastic setting*, $r_{ij}(t)$ on link i and channel j are independent over i and j, and are i.i.d. across slots t. The average proportion of successful packet transmissions per slot is denoted by $\mathbb{E}[r_{ij}(t)] = \theta_{ij}$, and is supposed to be unknown initially. In slot t, each packet is successfully transmitted with probability θ_{ij}, so that $r_{ij}(t)$ is a random variable whose distribution is that of Y_{ij}/m where Y_{ij} has a binomial distribution $\mathrm{Bin}(m, \theta_{ij})$. When $m = 1$, $r_{ij}(t)$ is a Bernoulli random variable of mean θ_{ij}. The stochastic setting models scenarios where the radio channel conditions are stationary, i.e., for any pair (i, j), θ_{ij} does not evolve over time.

In the following, we denote by $r_M(t)$ the total number (renormalized by a factor m) of packet successfully transmitted during slot t under feasible configuration $M \in \mathcal{M}$, i.e., $r_M(t) = \sum_{i \in [n]} \sum_{j \in [c]} M_{ij} r_{ij}(t) = M \bullet r(t)$.

C. Channel allocations and objectives

We analyze the performance of adaptive spectrum allocation policies that may select different feasible configurations at the beginning of each slot, depending on the observed received throughput under the various configurations used in the past. More precisely, at the beginning of each slot t, under policy π, a feasible configuration $M^\pi(t) \in \mathcal{M}$ is selected. This selection is made based on some feedback on the previously selected configurations and their observed throughput. More precisely, at the end of slot t, the number of packets successfully transmitted on the various links are observed, i.e., the feedback $f(t)$ is $(r_{ij}(t), i, j : M^\pi_{ij}(t) = 1)$.

At the beginning of slot t, the selected configuration $M(t)$ may depend on past decisions and the received feedback, i.e., on $M^\pi(1), f(1), \ldots, M^\pi(t-1), f(t-1)$. The chosen configuration can also be randomized (at the beginning of a slot, we sample a configuration from a given distribution that

depends on past observations). We denote by Π the set of feasible policies. The objective is to identify a policy maximizing over a finite time horizon T the expected number of packets successfully transmitted or simply what we call the *reward*. The expectation is here taken with respect to the possible randomness in the stochastic rewards (in the stochastic setting only) and in the probabilistic successively selected channel allocations. Equivalently, we aim at designing a sequential channel allocation policy that minimizes the *regret*. The regret of policy $\pi \in \Pi$ is defined by comparing the performance achieved under π to that of the best static policy:

$$R^{\pi}(T) = \max_{M \in \mathcal{M}} \mathbb{E}[\sum_{t=1}^{T} r_M(t)] - \mathbb{E}[\sum_{t=1}^{T} r_{M^{\pi}(t)}(t)], \quad (1)$$

where $M^{\pi}(t)$ denotes the configuration selected under π in slot t. The notion of regret quantifies the performance loss due to the need for learning radio channel conditions, and the above problem can be seen as a linear bandit problem.

D. Optimal Static Allocation

When evaluating the regret of a sequential spectrum allocation policy, the performance of the latter is compared to that of the best static allocation: $M^{\star} \in \arg\max_{M \in \mathcal{M}} \mathbb{E}[\sum_{t=1}^{T} r_M(t)]$, where in the above formula, the expectation is taken with respect to the possible randomness in the throughput $r_M(t)$ (in the stochastic setting only). To simplify the presentation, we assume that the optimal static allocation $M^{\star}$ is unique (the analysis can be readily extended to the case where several configurations are optimal, but at the expense of the use of more involved notations). To identify $M^{\star}$, we have to solve an Integer Linear Program (ILP). Indeed, $M^{\star}$ solves:

$$\begin{aligned} \max \quad & \sum_{i \in [n], j \in [c]} \gamma_{ij} M_{ij} \quad (2) \\ \text{s.t.} \quad & \sum_{j \in [c]} M_{ij} \leq 1, \quad \forall i \in [n], \\ & \sum_{i \in [n]} K_{\ell i} M_{ij} \leq 1, \quad \forall \ell \in [k], j \in [c] \\ & M_{ij} \in \{0,1\}, \quad \forall i \in [n], j \in [c], \end{aligned}$$

where for any pair (i,j), $\gamma_{ij} = \theta_{ij}$ in the stochastic setting, and $\gamma_{ij} = \sum_{t=1}^{T} r_{ij}(t)$ in the adversarial setting. It is easy to check that the ILP problem (2) is NP-complete for general interference graphs, even in case of a single available channel ($c = 1$); see Theorem 64.1 in [13]. In contrast, when the interference graph is complete, i.e., in the full interference case, the ILP problem can be interpreted as a Maximum Weighted Matching in a bipartite graph, and it can be solved in polynomial time [13].

III. Adversarial Bandit Problem

In this section, we study the problem in the adversarial setting. Applying results from [14], we can find algorithms whose regret scales as $O(\sqrt{|\mathcal{M}|T})$ where $|\mathcal{M}|$ is the number of feasible configurations. However, $|\mathcal{M}|$ grows exponentially with the number of links and channels (e.g. in the full interference case, the number of possible allocations is $\frac{n!}{(n-c)!}$ if $n \geq c$). One can achieve a much lower regret exploiting the problem structure [1], [12]. Here we propose ColorBand algorithm, and derive a regret upper bound that outperforms those of existing algorithms: for example, in the full interference case and when $n = c$, the bound scales as $\sqrt{n^3 \log(n) T}$, whereas the upper regret bound of the best algorithms so far scaled as $\sqrt{n^5 \log(n) T}$ [1].

We start with some observations about the ILP problem (2):

$$\begin{aligned} \max_{M \in \mathcal{M}} M \bullet r &= \max_{p(M) \geq 0, \sum_{M \in \mathcal{M}} p(M) = 1} \sum_{M \in \mathcal{M}} p(M) M \bullet r \\ &= \max_{\mu \in Co(\mathcal{M})} \mu \bullet r, \end{aligned}$$

where $Co(\mathcal{M})$ is the convex hull of the feasible allocation matrices $\mathcal{M}$.

We identify matrices in $\mathbb{R}^{n \times c}$ with vectors in $\mathbb{R}^{nc}$. Without loss of generality, we can always assume that c is sufficiently large (possibly adding artificial channels with zero reward) such that for all $i \in [n]$, $\sum_{j \in [c]} M_{ij} = n$ for all $M \in \mathcal{M}$, i.e. all links are allocated to a (possibly artificial) channel. Indeed, this can be done as soon as $c \geq \gamma(G)$ where $\gamma(G)$ is the chromatic number of the interference graph G. In other words, the bounds derived below are valid with c replaced by the maximum between the number of channels and the chromatic number of the interference graph. With this simplifying assumption, we can embed $\mathcal{M}$ in the simplex of distributions in $\mathbb{R}^{nc}$ by scaling all the entries by n. Let $\mathcal{P}$ be this scaled version of $Co(\mathcal{M})$.

We also define the matrix in $\mathbb{R}^{n \times c}$ with coefficients $\mu^0_{ij} = \frac{1}{n|\mathcal{M}|} \sum_{M \in \mathcal{M}} M_{ij}$. Clearly $\mu^0 \in \mathcal{P}$. We define $\mu_{\min} = \min n\mu^0_{ij} \geq \frac{1}{|\mathcal{M}|}$. Our algorithm is inspired from algorithms devised in [11] where full information is revealed and that use the projection onto convex sets using the KL divergence (see Chapter 3, I-projections in [15]). The KL divergence between distributions q and p in $\mathcal{P}$ (or more generally in the simplex of distribution in $\mathbb{R}^{nc}$) is: $KL(q\|p) = \sum_e q(e) \log \frac{q(e)}{p(e)}$, where e ranges over the couples $(i,j) \in [n] \times [c]$ and with the usual convention where $p \log \frac{p}{q}$ is defined to be 0 if $p = 0$ and $+\infty$ if $p > q = 0$. By definition, the projection of a distribution q onto a closed convex set Ξ of distributions is $p^{\star} \in \Xi$ such that $KL(p^{\star}\|q) = \min_{p \in \Xi} KL(p\|q)$. The pseudo-code of ColorBand is presented below (at the top of next page).

Theorem 1: We have for any T:

$$R^{ColorBand}(T) \leq 4n \sqrt{\mu_{\min}^{-1} T \log \mu_{\min}^{-1}}.$$

Note that in the full interference case, we have $\mu_{\min}^{-1} = \min(c,n)$. We also stress that the above result is valid for any $r \in [0,1]$.

IV. Stochastic bandit problem

This section is devoted to the analysis of our bandit problem in the stochastic setting. We first derive an asymptotic lower bound on the regret achieved by any feasible sequential

Algorithm 1: ColorBand Algorithm

Initialization: Start with distribution $q_0 = \mu^0$, $\gamma = \sqrt{\frac{\mu_{\min}^{-1} \log \mu_{\min}^{-1}}{T}}$

for *all* $t \geq 1$ **do**

- Let $p_{t-1} = (1-\gamma)q_{t-1} + \gamma\mu^0$ ($p_{t-1} \in \mathcal{P}$ so that $np_{t-1} \in Co(\mathcal{M})$).
- Select a random allocation $M(t)$ with distribution np_{t-1}.
- Get a reward $r_t = \sum_{i,j} r_{ij}(t) M_{ij}(t)$ and observe the reward vector: $r_{ij}(t)$ for all ij such that $M_{ij}(t) = 1$.
- Construct the reward matrix: $\tilde{r}_{ij}(t) = \frac{r_{ij}(t)}{np_{t-1}(ij)}$ for all i, j with $M_{ij}(t) = 1$ and all other entries are 0.
- Update $\tilde{q}_t(ij) \propto q_{t-1}(ij) \exp\left(\eta \tilde{r}_{ij}(t)\right)$.
- Set q_t to be the projection of $\tilde{q}_t$ onto the set $\mathcal{P}$ using the KL divergence.

end

spectrum allocation policy. This provides a fundamental performance limit that no policy can beat. We then present an algorithm whose regret upper bound outperforms those of existing UCB-like algorithms [2].

A. Asymptotic regret lower bound

In their seminal paper [16], Lai and Robbins consider the classical multi-armed bandit problem, where a decision maker has to sequentially select an action from a finite set of K actions whose respective rewards are independent and i.i.d. across time. For example, when the rewards are distributed according to Bernoulli distributions of respective means $\theta_1, \ldots, \theta_K$, they show that the regret of any online action selection policy π satisfies the following lower bound:

$$\liminf_{T\to\infty} \frac{R^\pi(T)}{\log(T)} \geq \sum_{i=1}^{K} \frac{\theta_1 - \theta_i}{KL(\theta_1, \theta_i)},$$

where without loss of generality $\theta_1 > \theta_i$ for all $i \neq 1$, and $KL(u, v)$ is the KL divergence number between two Bernoulli distributions of respective means u and v, $KL(u, v) = u\log(u/v) + (1-u)\log(1-u)/(1-v)$. The simplicity of this lower bound is due to the stochastic independence of the rewards obtained selecting different actions. In our linear bandit problem, the rewards obtained selecting different configurations are inherently correlated (as in these configurations, a link may be allocated with the same channel). Correlations significantly complicate the derivation and the expression of the lower bound on regret. To derive such a lower bound, we use the method developed in [17] for controlled Markov chains.

We use the following notation: $\Theta = [0,1]^{n\times c}$; $\theta = (\theta_{ij}, i \in [n], j \in [c])$; $\mu^M(\lambda) = M \bullet \lambda$, for any $M \in \mathcal{M}$ and $\lambda \in \Theta$. Recall that $\mu^\star = \max_{M\in\mathcal{M}} M \bullet \theta$, and the optimal configuration is $M^\star$, i.e., $\mu^\star = M^\star \bullet \theta$.

We introduce $B(\theta)$ as the set of *bad* parameters, i.e., the set of $\lambda \in \Theta$ such that configuration $M^\star$ provides the same reward as under parameter θ, and yet $M^\star$ is not the optimal static configuration:

$$B(\theta) = \{\lambda \in \Theta : \forall i, \lambda_{iM^\star(i)} = \theta_{iM^\star(i)},\ \mu^\star < \max_{M\in\mathcal{M}} \mu^M(\lambda)\}.$$

Then $B(\theta) = \bigcup_{M\neq M^\star} B_M(\theta)$, where

$$B_M(\theta) = \{\lambda \in \Theta : \forall i, \lambda_{iM^\star(i)} = \theta_{iM^\star(i)},\ \mu^\star < \mu^M(\lambda)\}.$$

The reward distribution for link i under configuration M and parameter θ is denoted by $p_i(\cdot; M, \theta)$. This distribution is over the set $\mathcal{S} = \{0, 1/m, \ldots, 1\}$ if m packets per slot are sent. Of course when $\sum_{j\in[c]} M_{ij} = 0$, we have $p_i(0; M, \theta) = 1$. When $\sum_{j\in[c]} M_{ij} = 1 = M_{iM(i)}$, we have, for $y_i \in \mathcal{S}$,

$$p_i(y_i; M, \theta) = \binom{m}{my_i} \theta_{iM(i)}^{my_i} (1 - \theta_{iM(i)})^{m-my_i}.$$

For example, if $m = 1$, for $y_i \in \{0, 1\}$,

$$p_i(y_i; M, \theta) = \theta_{iM(i)}^{y_i} (1 - \theta_{iM(i)})^{1-y_i},$$

We define the KL divergence number $KL^M(\theta, \lambda)$ under static configuration M as:

$$KL^M(\theta, \lambda) = \sum_{i\in[n]} \sum_{y_i\in\mathcal{S}} \log \frac{p_i(y_i; M, \theta)}{p_i(y_i; M, \lambda)} p_i(y_i; M, \theta).$$

($KL^M(\theta, \lambda) = \sum_{i\in[n], j\in[c]} M_{ij} KL(\theta_{ij}, \lambda_{ij})$ if $m = 1$). As we shall see later in this section, we can identify sequential spectrum allocations whose regret scales as $\log(T)$ when T grows large. Hence, we restrict our attention to the so-called *uniformly good* policies: $\pi \in \Pi$ is uniformly good if for all $\theta \in \Theta$, if the configuration M is sub-optimal ($M \neq M^\star$), then the number of times $T_M(t)$ it is selected up to time t satisfies: $\mathbb{E}[T_M(t)] = o(t^\gamma)$ for all $\gamma > 0$.

Theorem 2: For all $\theta \in \Theta$, for all uniformly good policy $\pi \in \Pi$,

$$\liminf_{T\to\infty} \frac{R^\pi(T)}{\log(T)} \geq C(\theta), \tag{3}$$

where $C(\theta)$ is the optimal value of the following optimization problem:

$$\inf_{x_M \geq 0, M\in\mathcal{M}} \sum_{M\in\mathcal{M}} x_M(\mu^\star - \mu^M(\theta)) \tag{4}$$

$$\text{s.t.} \inf_{\lambda\in B_M(\theta)} \sum_{Q\neq M^\star} x_Q KL^Q(\theta, \lambda) \geq 1, \forall M \neq M^\star. \tag{5}$$

The above lower bound is unfortunately not explicit. In the case of full interference, however, the bound can be simplified under the mild technical assumption that $\Theta = [0, a]^{n\times c}$ for $a < 1$. In particular, we may characterize how it scales with the numbers of links and channels.

Theorem 3: In the case of full interference and $m = 1$, we have:

$$C(\theta) = \Theta(n \times c), \quad \text{as } n, c \to \infty.$$

The above theorem states that there exist positive constants $k_1 > 0, k_2 > 0$ (that depend on θ) such that $C(\theta)/(nc) \in [k_1, k_2]$ for n, c large enough. This result is intuitive and means that the regret has to scale with the number of unknown parameters in the system.

B. A simple ϵ-greedy algorithm and its regret

Next we present a simple ϵ-greedy algorithm and show that its regret upper bound outperforms those of existing algorithms. Consider a set $\mathcal{A} \subset \mathcal{M}$ of configurations that covers all possible (link, channel) pairs. The construction of such a set is easy, and for example, in the case of full interference, we can simply use a set of $\max(n, c)$ configurations or matchings. Let A be the cardinality of $\mathcal{A}$. The algorithm consists in selecting the configuration that has provided with the maximum reward so far with probability $1 - \epsilon_t$, and a configuration selected uniformly at random among the covering set $\mathcal{A}$ of configurations. By reducing the exploration rate ϵ_t over time, a logarithmic regret can be achieved. More precisely, we will choose $\epsilon_t = \min(1, d/t)$ for some constant $d > 0$. Define $\hat{r}(t) = (\hat{r}_{ij,T_{ij}(t)}, i \in [n], j \in [c])$, where $T_{ij}(t)$ denotes the number of slots up to slot t when channel j is allocated to link i, and $\hat{r}_{ij,t} = \frac{1}{t}\sum_{s=1}^{t} X_{ij}(s)$ and $X_{ij}(s)$ is the proportion of packets successfully received during the s-th slot where j is allocated to i.

We state the following result for the regret of ϵ-greedy for the case of $n = c$. This result, however, can be easily extended for general n and c.

Theorem 4: When $d > 10An^2/\Delta_{\min}^2$, we have:

$$R^{\epsilon-greedy}(T) \leq 10A\frac{\Delta_{\max}}{\Delta_{\min}^2}n^2\log(T) + O(1) \quad \text{as } T \to \infty,$$

where $\Delta_{\max} = \max_{M\in\mathcal{M}}(\mu^\star - \mu^M(\theta))$ and $\Delta_{\min} = \min_{M\neq M^\star}(\mu^\star - \mu^M(\theta))$.

Recall that when $n = c$, we can select $\mathcal{A}$ with $A = n$. As a result, for the full interference case with this choice of $\mathcal{A}$, the regret scales as $\frac{\Delta_{\max}}{\Delta_{\min}^2}n^3\log(T)$ when T grows large. Compared to the regret bound of UCB-like algorithms (Theorem 3 of [3]), a factor n^2 has been removed. The upper bound proposed in the above theorem is the best bound derived so far, even for the full interference case.

Algorithm 2: ϵ-greedy Algorithm

Initialization: For $t = 1, \ldots, A$, select configurations in $\mathcal{A}$, observe the detailed rewards, and update $\hat{r}_{ij,t}$.
for *all* $t > A$ **do**
  Let $\epsilon_t = \min(1, d/t)$.
  Select configuration $M(t) \in \arg\max_{M\in\mathcal{M}} M \bullet \hat{r}(t)$ with probability $1 - \epsilon_t$, and a configuration uniformly selected at random in $\mathcal{A}$ with probability ϵ_t.
  Observe the detailed rewards and update $\hat{r}(t+1)$.
end

V. Conclusion

In this paper, we investigate the problem of sequential spectrum allocation in wireless networks where a potentially large number of channels are available, and whose average radio conditions are initially unknown. The design of such allocations has been mapped into a generic linear multi-armed bandit problem, for which we have devised efficient online algorithms. Lower bounds for the performance of these algorithms have been derived, and they are shown to outperform performance bounds of existing algorithms, both in the adversarial setting where no assumptions are made regarding the evolution of channel qualities, and in the stochastic setting where the radio conditions on the various channels and links are modelled as stationary processes.

Acknowledgment

M. Lelarge acknowledges the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-11-JS02-005-01 (GAP project). The work of A. Proutiere is partially supported by ERC Fluid Spectrum Access project.

References

[1] N. Cesa-Bianchi and G. Lugosi, "Combinatorial bandits," *Journal of Computer and System Sciences*, 2012.
[2] D. Kalathil, N. Nayyar, and R. Jain, "Decentralized learning for multi-player multi-armed bandits," *submitted to IEEE Trans. on Information Theory*, 2012.
[3] M. Lelarge, A. Proutiere, and S. Talebi, "Spectrum bandit optimization," *arXiv:1302.6974 [cs.LG]*, 2013. [Online]. Available: http://arxiv.org/abs/1302.6974
[4] L. Lai, H. El Gamal, H. Jiang, and H. Poor, "Cognitive medium access: Exploration, exploitation, and competition," *Mobile Computing, IEEE Transactions on*, vol. 10, no. 2, pp. 239–253, 2011.
[5] A. Anandkumar, N. Michael, A. Tang, and A. Swami, "Distributed algorithms for learning and cognitive medium access with logarithmic regret," *Selected Areas in Communications, IEEE Journal on*, vol. 29, no. 4, pp. 731–745, 2011.
[6] B. Radunovic, A. Proutiere, D. Gunawardena, and P. Key, "Dynamic channel, rate selection and scheduling for white spaces," in *Proceedings of the Seventh COnference on emerging Networking EXperiments and Technologies*, ser. CoNEXT '11. ACM, 2011.
[7] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *to appear in Foundations and Trends in Machine Learning, available online*, 2012.
[8] P. Rusmevichientong and J. N. Tsitsiklis, "Linearly parameterized bandits," *Math. Oper. Res.*, vol. 35, no. 2, May 2010.
[9] B. Awerbuch and R. Kleinberg, "Online linear optimization and adaptive routing," *J. Comput. Syst. Sci.*, vol. 74, no. 1, pp. 97–114, 2008.
[10] A. Gyorgy, T. Linder, and G. Ottucsak, "The shortest path problem under partial monitoring," in *Learning Theory*, ser. Lecture Notes in Computer Science, G. Lugosi and H. U. Simon, Eds. Springer Berlin Heidelberg, 2006, vol. 4005, pp. 468–482.
[11] D. P. Helmbold and M. K. Warmuth, "Learning permutations with exponential weights," *J. Mach. Learn. Res.*, vol. 10, pp. 1705–1736, Dec. 2009.
[12] S. Kale, L. Reyzin, and R. Schapire, "Non-stochastic bandit slate problems," *Advances in Neural Information Processing Systems*, pp. 1054–1062, 2010.
[13] A. Schrijver, *Combinatorial Optimization : Polyhedra and Efficiency (Algorithms and Combinatorics)*. Springer, Jul. 2004. [Online]. Available: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/3540204563
[14] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The nonstochastic multiarmed bandit problem," *SIAM J. Comput.*, vol. 32, no. 1, pp. 48–77 (electronic), 2002/03. [Online]. Available: http://dx.doi.org/10.1137/S0097539701398375
[15] I. Csiszár and P. Shields, *Information theory and statistics: A tutorial*. Now Publishers Inc, 2004.
[16] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4–2, 1985.
[17] T. L. Graves and T. L. Lai, "Asymptotically efficient adaptive choice of control laws in controlled markov chains," *SIAM J. Control and Optimization*, vol. 35, no. 3, pp. 715–743, 1997.