Slow Fading Channel Selection: A Restless Multi-Armed Bandit Formulation

Konstantin Avrachenkov
INRIA, Maestro Team
BP95, 06902 Sophia Antipolis, France
Email: k.avrachenkov@sophia.inria.fr

Laura Cottatellucci, Lorenzo Maggi
Eurecom
Mobile Communications Department
BP193, F-06560 Sophia Antipolis, France
Email: {laura.cottatellucci,lorenzo.maggi}@eurecom.fr

***Abstract*—We deal with a multi-access wireless network in which transmitters dynamically select a frequency band to communicate on. The slow fading channel attenuations follow an autoregressive model. In the single user case, we formulate this selection problem as a restless multi-armed bandit problem and we propose two strategies to dynamically select a band at each time slot. Our objective is to maximize the SNR in the long run. Each of these strategies is close to the optimal strategy in different regimes. In the general case with several users, we formulate the problem as a stochastic game with uncountable state space, where the objective is the SINR. Then we propose two strategies to approximate the best response policy for one user when the other users' strategy is fixed.**

I. Introduction

Next generation of wireless networks is expected to be characterized by a high decentralization/distribution of control functions among nodes to support self-organizing and self-healing capabilities. Network devices shall be able to monitor and sense the surroundings, learn from their monitoring and smartly and dynamically allocate resources. This perspective scenario is attracting a considerable amount of research efforts to develop learning techniques able to optimize the trade-off between exploration and exploitation of environment and resources. A relevant class of learning algorithms is the Multi-Armed Bandit (MAB) one. In the classic MAB problem there exist several "arms" that offer a reward when pulled (in analogy with gambling on bandits in casinos). Each arm is associated with a Markov process, and the reward of an arm is a function of its state. Gittins provided in [1] a dynamic allocation procedure, then dubbed Gittins index, which is optimal if the arms that are not pulled do not evolve over time. The more general case when the arms that are not pulled keep evolving in time is known as Restless MAB. It was proven by Papadimitriou and Tsitsiklis in [2] that restless MAB are PSPACE-hard in general. In [3], Whittle proposed to adopt a heuristic Lagrangian relaxation to extend the Gittins index to the restless case, which is asymptotically optimal under certain limiting regime [4].

In this work, we consider a wireless network where transmitters can select a frequency band from a shared pool to communicate on. The evolution of the slow fading channel attenuation associated to each frequency band and each transmitter is a random process that can be well approximated by an autoregressive process [5]. We assume that all such random processes are independent of each other. The goal of each transmitter is to maximize its average Signal to Interference and Noise Ratio (SINR) in the long run.

To get insight into this problem, first we focus on a single transmitter system to investigate the exploration-exploitation trade-off for the randomness introduced in the system by the autoregressive channel attenuations. Then, we consider the multi-transmitter case where the problem is further complicated by the randomness introduced by the autonomous band selections of multiple transmitters. For the single terminal case, the problem of dynamic frequency allocation for SNR maximization can be modeled as a restless Multi Armed Bandit (MAB) since the transmitter only knows the instantaneous attenuations on the bands utilized in the past and they evolve also when not utilized. To the best of the authors' knowledge, there are no available results on the MAB problem for autoregressive processes. We propose two heuristic frequency allocation strategies, one called "myopic" and the other "randomized". When the AR processes possess similar autocorrelation functions, we suggest to use the myopic strategy. Instead, when there is one AR process having a much higher autocorrelation, we suggest to use the randomized strategy. In the scenario with multiple transmitters the problem is formulated as a stochastic game with uncountable state space. We focus on a two-user system and we assume that user 1 is oblivious of the presence of user 2 and follows a plain single-user myopic approach. Then we propose two strategies for user 2 to approximate its best response against user 1's strategy. Again, one strategy is myopic and the other is randomized, with respect to the SINR objective function.

A lexical remark. We say that we "sample" a frequency band when we utilize it for the communication in a certain time slot.

II. Model

In Section III we consider one transmitters and one receiver, while in Section IV we deal with a model with two transmitters. Time is divided into slots and, at the beginning of a time slot, each transmitter (or user) selects a frequency band, out of a pool of M different ones, to transmit. At the receiver, a single-user decoder per transmitter is deployed. In the two-transmitter case, when a both users access the same frequency band i at time slot t, they interfere with each other, and the SINR (Signal to Interference plus Noise Ratio) for user $j = 1, 2$ at time t is

$$\mathrm{SINR}_{i,j}[t] = \frac{P_j |h_{i,j}[t]|^2}{N_0 + \sum_{q \neq j} P_q |h_{i,q}[t]|^2}$$

where P_j is the transmit power of user j, $h_{i,j}[t] \in \mathbb{C}$ is the i-th channel coefficient of user j at time t and N_0 is the variance of the additive white Gaussian noise at the receiver. When only one user is present, the SINR definition boils down to the classic SNR. For simplicity of notation, henceforth we will denote the channel attenuation coefficient $|h_{i,j}[t]|^2$ as $g_{i,j}[t]$.
Let us describe now our channel model. In [5] it is shown that, under slow fading conditions, the SNR (Signal to Noise Ratio) of indoor wireless channels can be well approximated by an autoregressive (AR) model. This means that, under such conditions, we can model the channel attenuations as

$$g_{i,j}[t] = \sum_{k=1}^{p_{i,j}} \phi_{i,j}^{(k)}\, g_{i,j}[t-k] + c_{i,j} + \epsilon_{i,j}[t]$$

where $\phi_{i,j} \in \mathbb{R}$, $\{\epsilon_{i,j}[t]\}_t$ is an *i.i.d.* Gaussian process with zero mean and variance $\sigma_{i,j}^2$, $c_{i,j} > 0$, and $p_{i,j}$ is the order of the model. Moreover, all the channels considered are independent of each other, i.e. $\epsilon_{i_1,j_1}[t]$ is independent of $\epsilon_{i_2,j_2}[t]$ when either $i_1 \neq i_2$ or $j_1 \neq j_2$.
We assume the AR process to be wide sense stationary (WSS), i.e. the roots of the polynomial $z^p - \sum_{k=1}^{p} \phi_{i,j}^{(k)} z^{p-k}$ must lie inside the unit circle.

III. Single user: MDP formulation

In this section we consider the single user case. In order to simplify the notation, we drop the user index. In our study we consider an AR(1) channel attenuation model, i.e.

$$g_i[t] = \phi_i\, g_i[t-1] + c_i + \epsilon_i[t]$$

For $|\phi_i| < 1$, the process is WSS, and the (unconditioned) expected value of channel attenuation $g_i[t]$ at any time instant t can be expressed as

$$m_i = \mathbb{E}(g_i[t]) = \frac{c_i}{1-\phi_i} \qquad \forall\, t.$$

Therefore we can say that $\mathbb{E}(\mathrm{SNR}_i[t]) = Pm_i/N_0$, for all t. Straightforward computations show that the autocovariance function of the channel attenuation can be written as

$$\mathbb{E}\big((g_i[t]-m_i)(g_i[t-n]-m_i)\big) = \phi_i^{|n|}\, \frac{\sigma_i^2}{1-\phi_i^2}. \quad (1)$$

We now illustrate the two fundamental assumptions of this paper. First, the coefficients ϕ_i and σ_i are known by the transmitter, which might have estimated them during a training phase. Second, the transmitter, at time t, *only knows the instantaneous attenuations of the frequency bands utilized* up to time $t-1$. Indeed, we assume that the receiver estimates $g_{i,j}$ and broadcast this information on the channel along with an identifier for the transmitter and the frequency band. The *goal* of the user is to dynamically switch among the channels at each time slot in order to maximize the expected average SNR over an infinite horizon. Equivalently, it wants to maximize the expected average over time of channel attenuations, denoted by $r(\pi)$:

$$\max_\pi \Big\{ r(\pi) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}_\pi(g_{\pi(t)}[t]) \Big\} \quad (2)$$

where π is a dynamic sampling strategies over the channels $1,\dots,M$. The reader should notice that a channel sampling strategy π at time t may depend on the whole history of the observed channels and of the sampling decisions up to time t. This class also includes static strategies, that choose one channel once for all. Intuitively, when there exists a channel i with much lower *unconditioned* expected attenuation, i.e. $m_i \gg m_k$ for all $k \neq i$, a static selection of the channel i is the nearly optimal strategy, since with high probability $g_i[t] > g_k[t]$ for all $k \neq i$ for almost all t.
In this section, we want to study how to dynamically select the band on which to transmit when, *a priori*, all of them are nearly equivalent, i.e. there exists $m \approx m_k$, for all k. At each time slot, there is always one channel better than the others, hence we wish to track dynamically the evolution over time of the best channel.
Intuitively, the sampling choice at each instant has to be a trade-off between exploration and exploitation. To give a hint, the most natural policy, that we will call *myopic*, at each time step t aims at maximizing the expected value of $\mathrm{SNR}[t]$, given all the previous channel observations. On the other hand, the statical information about channels that are not used becomes more and more obsolete, therefore in some cases it might be better to explore different channels with a *randomized* strategy.
We can formulate the optimization problem (2) as a restless Multi Armed Bandit problem (MAB for short), in which a user at each time instant t selects an arm (here, frequency band) which gives a reward (here, the SNR) and all the arms, including the ones that have not been selected, evolve according to a certain stochastic process (here, an autoregressive process). More specifically, we can describe the decision problem at hand as a Markov Decision Process (MDP) with an uncountable set of states $\mathcal{S}$ or, equivalently, as a Partially Observable MDP. Let us describe it in detail. At time t, we call $n_i(t)$ the number of steps ago in which channel i has been last used. The attenuation of channel i at time t conditioned on its last observation is a Gaussian r.v., and we denote its mean and variance as $\mu_i(t)$ and $\nu_i(t)$, respectively:

$$\begin{aligned}\mu_i(t) &= \mathbb{E}\big(g_i[t] \,\big|\, g_i[t-n_i(t)]\big) \\ &= \phi_i^{n_i(t)} g_i[t-n_i(t)] + c_i \frac{1-\phi_i^{n_i(t)}}{1-\phi_i} \quad (3)\end{aligned}$$

$$\nu_i(t) = \mathrm{Var}\left(g_i[t] \,\middle|\, g_i[t-n_i(t)]\right) = \sigma_i^2\, \frac{1-\phi_i^{2n_i(t)}}{1-\phi_i^2} \quad (4)$$

where $g_i[t-n_i(t)]$ is the attenuation of channel i during its last utilization. At time step t, thanks to the Markov property of the AR(1) process, the whole statistical information about channel i is hence contained in $(\mu_i(t), \nu_i(t))$. We observe that $\mu_i \in \mathbb{R}$, while ν_i is bounded between $[\sigma_i^2; \sigma_i^2/(1-\phi_i^2)]$. The decision on which channel to utilize at time t hinges on the set $S[t]$:

$$S[t] = \{\mu_1(t), \nu_1(t),\ \mu_2(t), \nu_2(t), \dots\ ,\ \mu_M(t), \nu_M(t)\}. \quad (5)$$

By utilizing the MDP jargon [6], we call by $S[t]$ the state of the decision problem at time t. The state space $\mathcal{S}$ is the uncountable collection of all the possible states. In each state $S \in \mathcal{S}$, a set of actions $\mathcal{A} = \{1, 2, \dots, M\}$ is available to

the transmitter, which represents the collection of channels that can be selected at time slot t. If channel i is selected, then we map the "reward" for the user in state $S[t]$ to the expected channel attenuation at time t conditioned on the last observation of channel i itself, i.e. $\mu_i(t)$. The state of the system at time $t+1$ evolves stochastically, according to the following Markovian rule. If channel i is selected at time t, then at time $t+1$,

$$\mu_i(t+1) = \phi_i Y + c_i, \quad \text{where } Y \sim \mathcal{N}\big(\mu_i(t), \nu_i(t)\big)$$
$$\nu_i(t+1) = \sigma_i^2.$$

Instead if channel i is not selected at time t,

$$\mu_i(t+1) = \phi_i \mu_i(t) + c_i$$
$$\nu_i(t+1) = \phi_i^2 \nu_i(t) + \sigma_i^2.$$

A. Heuristic algorithms

The theory of MDP allows us to claim that there exists an optimal stationary strategy π^O for the problem (2). Unfortunately, the computation of π^O turns out to be a difficult task. Indeed the solution to a Markov Decision Problem with uncountable state can only be approximated by means of discretization algorithms [6], and even in this case the curse of dimensionality entails that the size of the discretized state space increases exponentially with the number of arms. A different approach would be to compute the Whittle index [3] of each channel, but this approach is not guaranteed to be optimal. Hence, it becomes crucial to devise a simple policy whose performance is reasonably close to the optimal $r(\pi^O)$.

In the following we propose the most natural stationary strategy one can think of, i.e. the myopic policy π^M that aims at maximizing the instantaneous expected SNR in each state. Such a policy does not take into account that the statistics of the channel that have not been selected for a long period might become too stale. First, we need to initialize the algorithm, and we choose to sample the coefficient of each channel once.

Algorithm 1. Myopic policy π^M.
For $0 \le t \le M-1$ *select channel t, i.e.* $\pi^M(S[t]) = t+1$.
For $t \ge M$,

$$\pi^M(S[t]) = \underset{i\in\{1,\dots,M\}}{\operatorname{argmax}} \ \mu_i(t).$$

We intend to compare the performance of the myopic policy with a more sophisticated one, that we call randomized strategy and is inspired by the Thompson sampling strategy for Bayesian Multi Armed Bandit problems [7]. We suggests to draw, in each state $S[t]$, one realization of the random variable $\xi_i = g_i[t] \big| g_i[t-n_i(t)]$, for each channel $i = 1, \dots, M$. Then, the arm corresponding to the highest realization of ξ is chosen. This procedure does not always follow the myopic rule, but with a certain probability explore the arms that, though possessing a lower μ, might be optimal since their last observation is too stale.

Algorithm 2. Randomized policy π^R.
For $0 \le t \le M-1$ *select channel t, i.e.* $\pi^R(S[t]) = t+1$.
For $t \ge M$, *draw a realization of the Gaussian variable* $\xi_i \sim \mathcal{N}(\mu_i(t), \nu_i(i))$ *for all* $i = 1, \dots, M$. *Select*

$$\pi^R(S[t]) = \underset{i=1,\dots,M}{\operatorname{argmax}} \ \xi_i.$$

B. Simulations

In this section we show the results of some simulations, giving a hint about the performance of the myopic and the randomized policies, described respectively in Algorithm 1 and 2. Given a stationary policy π, we want to assess its average reward $r(\pi)$. We compare the myopic and randomized policies with *i)* the optimal policy π^O, approximated by means of a state discretization technique [6], with *ii)* the upper bound for the performance of any strategy, computed by selecting the channel with the highest coefficient g at each time step:

$$\pi^U(t) = \underset{i=1,\dots,M}{\operatorname{argmax}} \ g_i[t], \quad \forall\, t \ge 0 \tag{6}$$

and with *iii)* the static policy π^S, that selects off-line the arm with the highest expected value, and no longer switches to other channels, i.e.

$$\pi^S[t] = \underset{i=1,\dots,M}{\operatorname{argmax}} \, m_i, \quad \forall\, t \ge 0.$$

Of course, the strategy π^U is not applicable, since it is not causal. In theory, its performance is achievable only when the channels are deterministic hence perfectly predictable, i.e. $\sigma_i = 0$ for all $i = 1, \dots, M$. We now show the performance of the five policies under scrutiny, the myopic π^M, the randomized π^R, the static π^S, the optimal π^O, and the upper bound policy π^U, under different channel conditions.
First, we consider 3 arms, where arms 2,3 are statistically equivalent, and $\phi_2 = \phi_3 = 0.3$, $\sigma_2^2 = \sigma_3^2 = 1$, and $m_2 = m_3 = 8$. Arm 1 has the same coefficients $\phi_1 = 0.3$, $\sigma_1^2 = 1$ as arms 2,3. In Figure 1 we show the performance of the five policies when m_1 varies within $[7; 9]$. We see that, under these conditions, the myopic policy outperforms the randomized one since the latter wastes too much time in exploring arms that are not optimal. As intuition confirms, the static policy π^S performs as well as the myopic π^M when arm 1 has the highest expected value $m_1 > m_2 = m = 3$. Instead, for $m_1 < m_2 = m_3$, dynamically switching between the arms 2,3 is beneficial with respect to statically selecting one of the two.
As we see in Figure 1, when all the arms are characterized by the same unconditioned expectation, i.e. $m_i = 8$, for $i = 1, 2, 3$, the static policy π^S is outperformed by both the myopic and the randomized strategies. It is indeed better to switch among the channels to attempt to track the best instantaneous channel at *each* time instant, based on the previous observations. Remarkably, the performance of the myopic policy π^M is close to the optimal π^O.

Hence, we evaluate our algorithms in a different scenario, in which the value m's are the same for all the channels, but there exists one channel (say, 1) whose autocovariance function (1) decays considerably more slowly than the others. It is clear from Figure 2 that there are lapses in which channel 1 is by far the best, and some others in which its channel coefficient g_1 plummets below the others. From Figure 2 we observe that the myopic strategy often fails to track channel 1 when it is the best. The reason is quite intuitive: during the lapse in which channel 1 is the worst one, the myopic strategy does not choose it, then its last observation become obsolete, and consequently the prediction $\mu_1(t)$ tends to $m_1 = 10$. Thus, it is highly

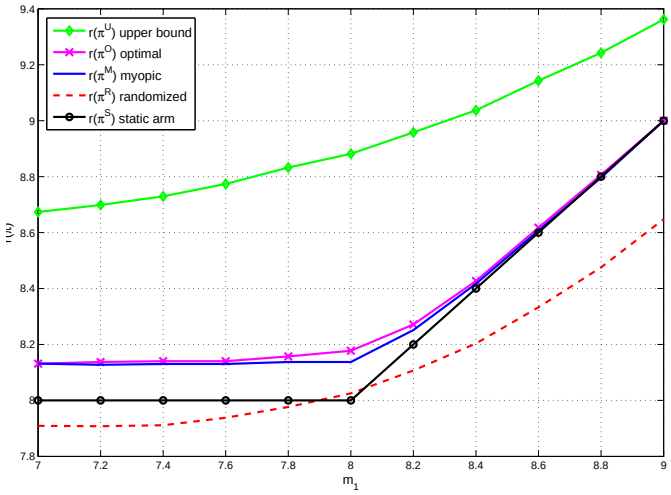

Figure 1. Performance of myopic and randomized algorithm with 3 arms (channels). Arms 2 and 3 are statistically equivalents, with $\phi_2 = \phi_3 = 0.3$, $\sigma_2^2 = \sigma_3^2 = 1$, and $m_2 = m_3 = 8$. Arm 1 has the same $\phi_1 = 0.3$, $\sigma_1^2 = 1$ as arms 2,3, while the performance of the proposed algorithms are assessed when m_1 varies within $[7; 9]$.

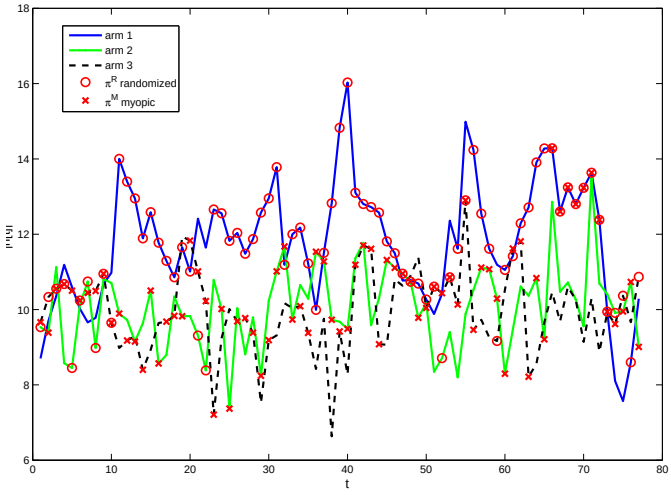

Figure 2. Channel (or arm) selection when $c_i = 10$ for all i, $\phi_1 = 0.9, \phi_2 = \phi_3 = 0.3$, $\sigma_1^2 = 1.5, \sigma_2^2 = \sigma_3^2 = 0.5$. The randomized strategy succeeds in tracking the first channel with higher autocorrelation, when it is the best one.

probable (and this probability increases with M) that one of the other, suboptimal, channels, having a fresher observation, offers a higher prediction. It easily follows that, for its inherent features, the randomized policy is more suitable to such kind of situations, as results in Figure 3 confirm. We considered 3 arms (frequency bands). Arms 2 and 3 are statistically equivalents, with $\phi_2 = \phi_3 = 0.3$, $\sigma_2^2 = \sigma_3^2 = 1$, $c_2 = c_3 = 10$. Arm 1 has the same coefficients $c_1 = 10$, $\sigma_1^2 = 1$ as arms 2,3, while the performance of the proposed algorithms are assessed when the coefficient ϕ_1 varies within $[0.3; 0.98]$. As we intuitively explained before, when the coefficient ϕ_1 is sufficiently high, i.e. $\phi > 0.85$, the randomized strategy outperforms the myopic one. Notably, the myopic policy is quasi-optimal for $\phi_1 < 0.6$, while the the randomized one is nearly optimal for $\phi_1 > 0.9$.

IV. Multi user: Stochastic game formulation

In this section we discuss the more general scenario described in Section II, in which two transmitters dynamically select one among M channels at each time slot. If some users choose the same channel in one time slot, they interfere with each other. Therefore, the objective function for each user is its SINR, and no longer its SNR. Since in the single user case the decision process can be described as an MDP, then the scenario with two users can be formalized as a stochastic game, also called competitive MDP [8], with uncountable state space. In general, a stochastic game is an MDP in which the instantaneous rewards for each player and the transition probabilities among the states are controlled by the joint actions of the players in each state.

In our case, the set of channels $h_{1,j}, \ldots, h_{M,j}$ for player j evolve independently from the ones available to any other player $k \neq j$, and the action space for each player is still $\mathcal{A} = \{1, \ldots, M\}$, i.e. the channel indices to be selected at each slot. Therefore, we are allowed to formulate the game as a stochastic game in which each user j controls its own Markov chain on the state space $\mathcal{S}_j$. As in the single user case $\mathcal{S}_j$ is the set of all the possible states (5). Formally, the state space of the stochastic game at hand is the Cartesian product $\mathcal{S}^* = \mathcal{S}_1 \times \mathcal{S}_2$.

Let us denote by π_j a sampling strategy for user j and by π_{-j} the one for the other users. Possibly, π_j, π_{-j} are randomized policies. We define the instantaneous reward for user j in state $S^*[t] \in \mathcal{S}^*$ as the expected reward

$$\mathbb{E}\big(\mathrm{SINR}_j[t] \big| S^*[t], \pi_i, \pi_{-i}\big).$$

Thus, the interaction on the players occurs only on the instantaneous rewards gained in each state, through the SINR expression. Thus we can say that our model is a reward-coupled stochastic game. This model is very similar with the one dealt with in [9], except that here the state space is uncountable and there are no constraints on the rewards.

A. Heuristic Best Response

We now propose a heuristic best response policy for user 2. Suppose that user 1 is oblivious of the presence of user 2 and performs a myopic policy π_1^M to maximize the expected average of channel attenuations over time, as in the single user case. On the other hand, user 2 knows the parameters of the channels, the current state, and the strategy of user 1. Thus, user 2 still faces an MDP with uncountable states, which is equivalent to the stochastic game described before, when user 1 fixes its own stationary strategy. Let us give an insight on a possible strategy for user 2. Assume that, for user 2, channel i_1 presents at time t the highest coefficient $g_{i_1,2}[t]$, but the expected SINR guaranteed by channel i_2 with suboptimal attenuation is higher, since the interference is much weaker. Then, it is in general not clear what user 2 should do. A myopic solution would suggest to switch to the free channel i_2, but on the other hand, in such a way the information about channel i_1 becomes stale, and moreover channel i_1 itself might become free in a near future. Then, in analogy with the single player case, we propose two strategies, one myopic and one randomized, to approximate the best response for user 2

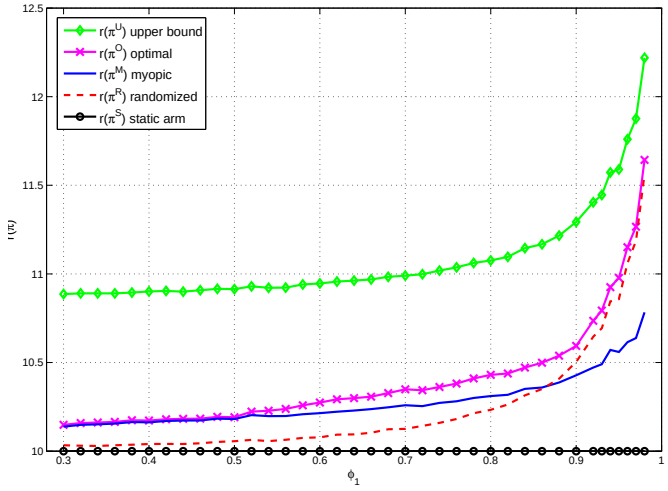

Figure 3. Performance of myopic and randomized algorithm with 3 arms (frequency bands). Arms 2 and 3 are statistically equivalents, with $\phi_2 = \phi_3 = 0.3$, $\sigma_2^2 = \sigma_3^2 = 1$, $c_2 = c_3 = 10$. Arm 1 has the same coefficients $c_1 = 10$, $\sigma_1^2 = 1$ as arms 2,3. ϕ_1 varies within $[0.3; 0.98]$.

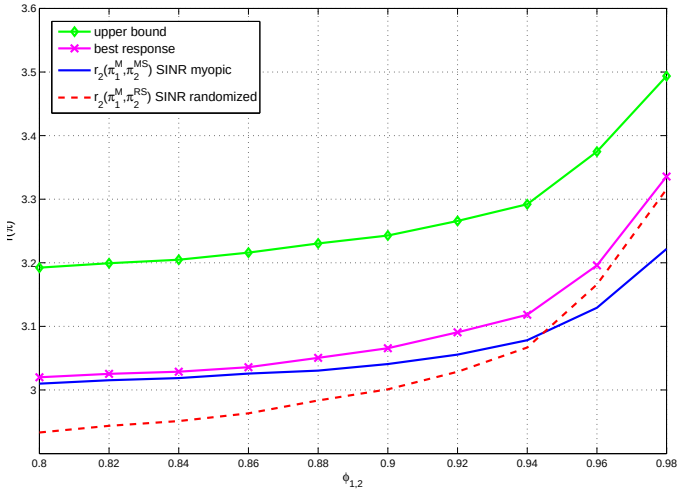

Figure 4. Best response strategy of user 2 against a myopic policy for user 1. For user 1, $\sigma_{1,1}^2 = \sigma_{2,1}^2 = 0.1$ and $\phi_{1,1} = \phi_{2,1} = 0.3$, $m_{1,1} = 2$, $m_{2,1} = 0.5$. For user 2, $\sigma_{1,2}^2 = 0.8, \sigma_{2,2}^2 = 0.4$, $m_{1,2} = 8, m_{2,2} = 3$, $\phi_{2,2} = 0.3$. $\phi_{1,2}$ varies within $[0.8; 0.98]$. $r_2(\pi_1^M, \pi_2)$ is the expected long run average SINR for user 2 when user 1 adopts strategy π_1^M.

against a myopic policy π_1^M that user 1 implements regardless of user 2's behavior. We suppose the algorithms are initialized by sampling each channel once.

Algorithm 3. SINR myopic policy π_2^{MS} for user 2, against myopic policy π_1^M for user 1.

$$\pi_2^{MS}(S^*[t], \pi_1^M) = \underset{i \in \{1,\dots,M\}}{\operatorname{argmax}} \mathbb{E}(\mathrm{SINR}_{i,2}[t] \big| S^*[t], \pi_1^M, i).$$

Algorithm 4. Randomized policy π_2^{RS} for user 2, against myopic policy π_1^M for user 1.
Draw a realization of the random variable $\xi_i = \mathrm{SINR}_{i,2}[t] \big| (S^*[t], \pi_1^M, i)$, *for all* $i = 1, \dots, M$. *Select*

$$\pi_2^{RS}(S^*[t], \pi_1^M) = \underset{i=1,\dots,M}{\operatorname{argmax}} \xi_i.$$

About the performance of policies π^{MS}, π^{RS}, we can do similar considerations to the one made for the myopic and randomized algorithms in the single user case. Let us explain the results illustrated in Figure 4. We considered 2 users and 2 channels. The noise variance is $N_0 = 1$ and $P_1 = P_2 = 1$. The channels for user 1 are almost deterministic, i.e. $\sigma_{1,1}^2 = \sigma_{2,1}^2 = 0.1$ and $\phi_{1,1} = \phi_{2,1} = 0.3$, $m_{1,1} = 2$, $m_{2,1} = 0.5$. Thus user 1, that is unaware of the presence of user 2 and adopts a myopic policy π_1^M, selects channel 1 almost always. For user 2, $\sigma_{1,2}^2 = 0.8, \sigma_{2,2}^2 = 0.4$, $m_{1,2} = 8, m_{2,2} = 3$, $\phi_{2,2} = 0.3$. Hence, a static strategy for user 2 would suggest not to collide and to select channel 2. Anyway, sometimes it is beneficial for user 2 to select channel 1 when this is good enough. Indeed, for values of $\phi_{1,2}$ approaching 1, the autocorrelation of channel 1 for user 2 increases, and the randomized policy π^{RS} succeeds in tracking channel 1 in the time slots in which its coefficient g is large enough to overwhelm the interference caused by user 1.

V. Conclusions

We proposed two strategies to dynamically select one out of a pool of M slow fading channels, modeled as autoregressive processes of order 1. The decision process is modeled as a restless bandit, or equivalently as a Markov Decision Process. The myopic channel selection strategy is nearly optimal when the channels are similarly correlated. Instead we suggest to adopt a randomized strategy when one channel shows higher autocorrelation. When two users are present, they interfere with each other, and we model the competitive learning process as a stochastic game. We finally propose two ways to approximate a best response selection strategy for the transmitters.

Acknowledgments: This research was supported by "Agence Nationale de la Recherche", with reference ANR-09-VERS-001, and Orange France Telecom Grant on Content-Centric Networking. We would like to thank Alexey Piunovskiy for very helpful discussion.

References

[1] J. C. Gittins, R. Weber, and K. D. Glazebrook, *Multi-armed bandit allocation indices*. Wiley Online Library, 1989, vol. 25.
[2] C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of optimal queueing network control," *Mathematics of Operations Research*, vol. 24, 1999.
[3] P. Whittle, "Restless bandits: Activity allocation in a changing world," *Journal of applied probability*, pp. 287–298, 1988.
[4] R. Weber and G. Weiss, "On an index policy for restless bandits," *Journal of Applied Probability*, pp. 637–648, 1990.
[5] R. Aguero, M. Garcia, and L. Mufioz, "BEAR: A bursty error auto-regressive model for indoor wireless environments," in *Personal, Indoor and Mobile Radio Communications, 2007. PIMRC 2007. IEEE 18th International Symposium on*. IEEE, 2007, pp. 1–5.
[6] N. Bäuerle and U. Rieder, *Markov Decision Processes with applications to finance*. Springer Verlag, 2011.
[7] W. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3/4, pp. 285–294, 1933.
[8] J. Filar and K. Vrieze, *Competitive Markov decision processes*. Springer Verlag, 1997.
[9] E. Altman, K. Avrachenkov, N. Bonneau, M. Debbah, R. El-Azouzi, D. Sadoc Menasche, "Constrained cost-coupled stochastic games with independent state processes," *Operations Research Letters*, vol. 36, no. 2, pp. 160–164, 2008.