

Optimally Sensing a Single Channel Without Prior Information: The Tiling Algorithm and Regret Bounds

Sarah Filippi, Olivier Cappé, Aurélien Garivier

▶ **To cite this version:**

Sarah Filippi, Olivier Cappé, Aurélien Garivier. Optimally Sensing a Single Channel Without Prior Information: The Tiling Algorithm and Regret Bounds. IEEE Journal of Selected Topics in Signal Processing, 2010, 5 (1), pp.68 - 76. 10.1109/JSTSP.2010.2058091 . hal-00408867v2

HAL Id: hal-00408867
https://hal.science/hal-00408867v2

Submitted on 23 Mar 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Optimally Sensing a Single Channel Without Prior Information: The Tiling Algorithm and Regret Bounds

Sarah Filippi*, Olivier Cappé and Aurélien Garivier
LTCI, TELECOM ParisTech and CNRS, 46 rue Barrault, 75013 Paris, France*
(filippi, cappe, garivier)@telecom-paristech.fr

Abstract

We consider the task of optimally sensing a two-state Markovian channel with an observation cost and without any prior information regarding the channel's transition probabilities. This task is of interest in the field of cognitive radio as a model for opportunistic access to a communication network by a secondary user. The optimal sensing problem may be cast into the framework of model-based reinforcement learning in a specific class of Partially Observable Markov Decision Processes (POMDPs). We propose the Tiling Algorithm, an original method aimed at reaching an optimal tradeoff between the exploration (or estimation) and exploitation requirements. It is shown that this algorithm achieves finite horizon regret bounds that are as good as those recently obtained for multi-armed bandits and finite-state Markov Decision Processes (MDPs).

Index Terms: Cognitive Radio, Opportunistic Channel Access, POMDPs, Regret Bounds, Reinforcement learning, Restless Bandit.

1 Introduction

In recent years, opportunistic spectrum access for cognitive radio has been the focus of significant research efforts [1, 10, 16]. These researches propose to improve spectral efficiency by making smarter use of the large

*This publication is partially supported by Orange Labs under contract n°289365.

portion of the frequency bands that remains unused. In Licensed Band Cognitive Radio, the goal is to share the bands licensed to primary users with non primary users called secondary users or cognitive users. These secondary users must carefully identify available spectrum resources and communicate avoiding to disturb the primary network. Opportunistic spectrum access thus has the potential for significantly increasing the spectral efficiency of wireless networks.

The opportunistic communication model previously considered by [13, 24] consists of N independent channels with time-varying states in which a single secondary user searches for idle channels temporarily unused by primary users. These N channels are licensed to a primary network whose users communicate according to a synchronous slot structure. The state $X_t(i)$ of the i-th channel is modelled by a Markov chain: at each time slot, the channel is either idle or occupied and the availability of the channel evolves in a Markovian way (see Fig.1).

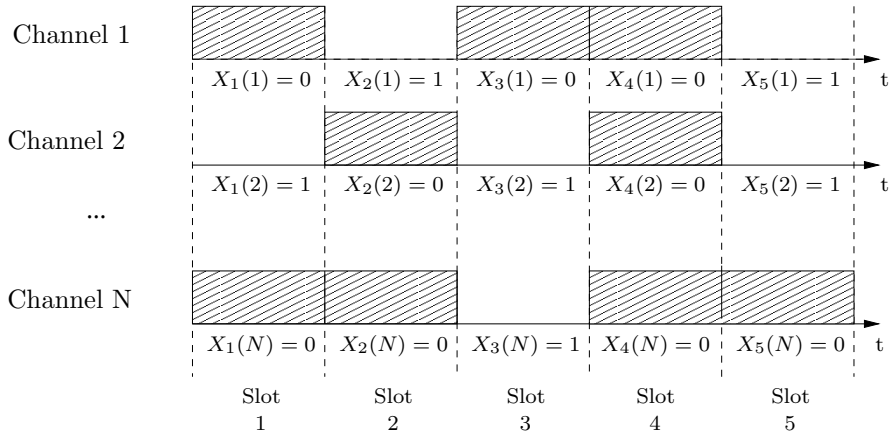


Figure 1: Representation of the primary network

Consider now a secondary user seeking opportunities of transmitting in the free slots of these N channels without disturbing the primary network. Due to hardware limitations and the energy cost of spectrum monitoring, the secondary user can not sense all the channels simultaneously [11, 13, 23], his main task is then to choose which channel to sense at each time aiming to maximize its expected long-term transmission efficiency. Under this model, channel access may be interpreted as a planning task in a particular class of Partially Observable Markov Decision Process (POMDP) also called restless bandits [13, 24]. The restless

bandit model differs from the simpler multi-armed bandit model [3] by the fact that the state of each arm (here, each channel) evolves in a Markovian way - even the arms that are not played (here, the channels that are not accessed).

Papadimitriou and Tsitsiklis [17] have established that the planning task in the restless bandit model is PSPACE-hard, and hence that optimal planning is not practically achievable for a large number N of channels. Nevertheless, recent publications have focused on near-optimal so-called *index strategies* [9, 12, 13], which have a reduced implementation cost. An index strategy consists in splitting the optimization task into N channel-specific sub-problems, following the idea originally proposed by Whittle [22]. Interestingly, to determine the Whittle index pertaining to each channel, one has to solve the planning problem in a single channel model with an additional cost (or penalty) term associated with the action of actually observing the channel, whether it is found idle or not [12, 13]. Besides, another suboptimal strategy consists in relaxing the constraint that the number of observable channels at each time slot is fixed and in introducing a cost that is proportional to the number of observed channels. Under this latter formulation, the planning task is fully decoupled and the optimal channel access policy is found by determining the optimal policies for the N single channel models with observation cost already mentioned above. In the following, we focus on the single channel model with observation cost, which we refer to as the *channel sensing model*. As we will see in Section 2, in the channel sensing model, the optimal sensing policies are computable, though already non-trivial.

It is usually assumed that the statistical information about the primary users' traffic is fully available to the secondary user [7, 13, 23, 24]. In practice however, the statistical characteristics of the traffic are not fixed a priori and must be somehow estimated by the secondary user. The goal of the present work is to determine the optimal policy in the channel sensing model *without any prior information on the channel parameters*, a task which is usually referred to as reinforcement learning [20]. Related works include [15] which proposed a heuristic rule based on the asymptotic behavior of the parameter estimate. Lai et al [11] also considered the learning task in the opportunistic channel access model but in the simpler case where each channel is memoryless. In addition, an asymptotically efficient algorithm has been proposed by [2] for both the memoryless and the Markovian model. Some similar settings with multiple cognitive users have

also raised some interest: in particular, an algorithm for a decentralized model in the case where the channels are memoryless has been recently introduced by [14].

We focus on the scenario where the secondary user first carries out an *exploration phase* aimed at estimating the channel parameter and then follows a fixed sensing policy, based on the estimated parameters. This second phase is called the *exploitation phase*. The key issue is to reach the proper balance between exploration and exploitation so as to interrupt the exploration phase as soon as enough statistical evidence is available to determine the optimal sensing policy. To evaluate the proposed algorithm, we will consider the so-called finite-horizon regret criterion, which, for any time horizon n, compares the expectations of the accumulated reward to that gained by the "oracle" agent who knows the channel parameters beforehand and thus always applies the optimal sensing policy (the reward scheme appropriate in the channel sensing model will be fully described in Section 2 below).

In the field of reinforcement learning, several approaches have been proposed recently to explicitely balance exploration and exploitation. [4, 19, 21]. This is in particular the aim of *model-based reinforcement learning*. Auer *et al.* and Tewari and Bartlett [4, 21] provide finite horizon regret bounds that apply for finite-state MDPs (Markov Decision Processes). The bounds given in [4] are of the form $C|\mathsf{X}|^2 \log(n)$, where n is the time horizon, $|\mathsf{X}|$ is the size of the state-space and C is a constant that does depend on the (unknown) underlying MDP model. [4] also provides a uniform bound of the form $C|\mathsf{X}|\sqrt{n \log n}$, where C this time refers to a universal constant. Despite its apparent simplicity, the channel sensing model corresponds to a POMDP that can only be represented as an MDP by rewriting it as a function of an internal state that takes an infinite number of distinct values [5] (see also Section 2 below). Hence, none of the approaches proposed so far for model-based reinforcement learning in MDPs appears to be usable for the channel sensing model. However, we show that it is possible to profit from the specificities of the channel sensing model, namely that (1) the state is partially observable by means of the sensing action, (2) the model is parametered by a low-dimensional parameter vector, and (3) the state transitions do not depend on the agent's actions. The resulting *Tiling Algorithm*, described in Section 3 below, achieves $C \log(n)$ regret (where C depends on the actual parameter values) and $C(\log n)^{1/3} n^{2/3}$ uniform regret for the channel sensing model. To the best of our knowledge, this is the first algorithm that obtains such strong performance guarantees for the channel

sensing model.

The rest of the article is organized as follows. The channel sensing model is formally described in Section 2. In Section 3, the Tiling Algorithm is presented and its performance in terms of finite-horizon regret are analyzed. Section 4 provides a detailed account of the use of the approach for the channel sensing access as well as some numerical experiments.

2 Channel Sensing Model

Let X_t denote the state of the channel which is equal to 0 when the channel is occupied and 1 when it is idle. Let α (resp.β) be the transition probability from state 0 (resp. 1) to state 1 (see Fig. 2). Additionally, denote by (ν_0, ν_1) the stationary probability of the Markov chain $(X_t)_t$.

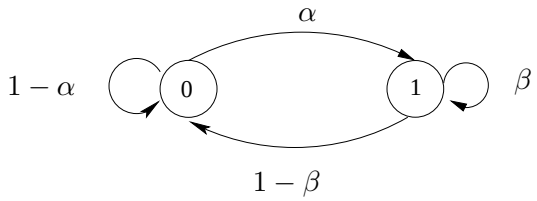


Figure 2: Transition probabilities in the i-th channel.

At each time slot, the secondary user can choose to sense the channel ($A_t = 1$) or to not observe it ($A_t = 0$). The observation Y_t is equal to the state X_t if the channel has been observed, and is void otherwise. For discussions of alternative models, including in particular sensing errors, see [6]. The reward gained at each time slot is defined as follows

$$r(X_t, A_t) = \begin{cases} 1 & \text{if } A_t = 1,\ X_t = Y_t = 1 \\ 0 & \text{if } A_t = 1,\ X_t = Y_t = 0 \\ \lambda & \text{otherwise} \end{cases},$$

which depends on X_t only through Y_t. The reward $0 \leq \lambda \leq 1$ associated to the action of not observing (called "subsidy" by Whittle [22]) may also be interpreted as a fixed cost for sensing the channel. Indeed,

the model would be equivalent upon redefining $r(X_t, A_t)$ as the difference between a reward for utilizing the channel (1 if the channel is free and 0 othwerwise) minus a fixed sensing cost equal to λ. The channel sensing model is a particular POMDP (see [5]) in which the state transition probabilities do not depend on the actions and where an action enables the secondary user to observe a part of the channel state. Note also that the model is fully determined by the two unknown transition parameters α and β and by the cost λ. It is possible to reformulate the channel sensing model as an MDP by introducing the so-called *belief state* p_t [5] that summarizes all past decisions and observations: $p_t = \mathbb{P}[X_t = 1 \,|\, A_{0:t-1}, Y_{0:t-1}]$. The belief state satisfies the following recursion

$$p_{t+1} = \begin{cases} \alpha & \text{if } A_t = 1, Y_t = 0 \\ \beta & \text{if } A_t = 1, Y_t = 1 \\ p_t\beta + (1-p_t)\alpha & \text{otherwise} \end{cases} . \tag{1}$$

Equivalently, the belief state p_t is completely determined by the pair of variables (K_t, U_t), where K_t is the lag to the latest channel observation and U_t is the latest observed status of the channel, i.e., the last time the channel was observed was at time $t - K_t$ and it was then in state $U_t \in \{0, 1\}$. We refer to the pair (K_t, U_t) as the *internal state* of the system. Note that the internal state takes its values in $\mathbb{N}^* \times \{0, 1\}$ which is a countably infinite set. Further denote by $p^{k,u}_{\alpha,\beta}$ the conditional probability that the channel is free given that $(K_t, U_t) = (k, u)$: for $k > 1$, $p^{k,u}_{\alpha,\beta} = \mathbb{P}[X_t = 1 | A_{t-k+1:t-1} = 0, A_{t-k} = 1, Y_{t-k} = u]$ and $p^{1,u}_{\alpha,\beta} = \mathbb{P}[X_t = 1 \,|\, A_{t-1} = 1, Y_{t-1} = u]$. Eq. (1) implies that, for $k > 1$, $p^{k,u}_{\alpha,\beta} = p^{k-1,u}_{\alpha,\beta}\beta + (1 - p^{k-1,u}_{\alpha,\beta})\alpha$ and $p^{1,u}_{\alpha,\beta} = \beta^u\alpha^{1-u}$. It is then easily shown by induction that these probabilities may be written as follows:

$$p^{k,0}_{\alpha,\beta} = \frac{\alpha(1 - (\beta - \alpha)^k)}{1 - \beta + \alpha} , \tag{2}$$

$$p^{k,1}_{\alpha,\beta} = \frac{(\beta - \alpha)^k(1 - \beta) + \alpha}{1 - \beta + \alpha} . \tag{3}$$

The belief state (and thus the internal state) is a *sufficient statistic* in the sense that there exists an optimal policy depending only on it [7]. Let thus $\pi : \mathbb{N}^* \times \{0, 1\} \to \mathsf{A}$ denote a policy which assigns an action according to the current internal state (K_t, U_t), and let Π be the set of the policies. A policy in Π

is characterized by the pair (m_0, m_1) which defines how long the secondary user decides to wait (i.e. not observe the channel) before observing the channel again, depending on the outcome of the last observation. Denote by $\pi_{(m_0,m_1)}$ the policy which consists in waiting $m_0 - 1$ (resp. $m_1 - 1$) time slots before observing the channel again if, last time the channel was sensed, it was occupied (resp. idle). Let π_∞ be the policy which consists in never observing the channel. The average reward received following such a policy can be exactly computed (see Appendix C for details) depending on the transition probabilities (α, β):

$$V_{\alpha,\beta}^{\pi_{(m_0,m_1)}} = \frac{p_{\alpha,\beta}^{m_0,0} + \lambda[(m_1-1)p_{\alpha,\beta}^{m_0,0} + (m_0-1)p_{\alpha,\beta}^{m_1,1}]}{m_0(1-p_{\alpha,\beta}^{m_1,1}) + m_1 p_{\alpha,\beta}^{m_0,0}} \text{ , for } m_0,\, m_1 \in \mathbb{N}^*, \tag{4}$$

$$V_{\alpha,\beta}^{\pi_\infty} = \lambda \,. \tag{5}$$

For each value of the parameter (α, β), one can identify the optimal policy $\pi_{\alpha,\beta}^*$ such that

$$V_{\alpha,\beta}^{\pi_{\alpha,\beta}^*} = \max_{\pi \in \Pi} V_{\alpha,\beta}^{\pi} \,.$$

Similarly, in both [12] and an extended version of [13], the optimal policies for this model has been studied as a function of (α, β). It is then possible to determine *policy zones* that are regions of the parameter space $[0,1] \times [0,1]$ that correspond to a single optimal policy. Fig. 3 displays the policy zones for $\lambda = 0.3$. Let $Z_{(m_0,m_1)}$ (resp. Z_∞) denote the region of the parameter space such that $\pi_{(m_0,m_1)}$ (resp. π_∞) is the optimal policy. Note that for $\alpha > \lambda$ and $\beta > \lambda$, the optimal policy $\pi_{(1,1)}$ consists in always observing the channel since the expected received reward if the channel is observed (equal to α or β) is larger than the reward λ received if it is not observed (see Fig. 3). In contrast, when $\alpha < \lambda$ and $\beta < \lambda$, it is optimal never to observe the channel. For $\alpha < \lambda < \beta$, there are an infinity of policy zones. Each of them consists in observing the channel if it has been observed to be free and wait $m_0 - 1$ times before observing it otherwise, with different values of m_0 between 2 and infinity.

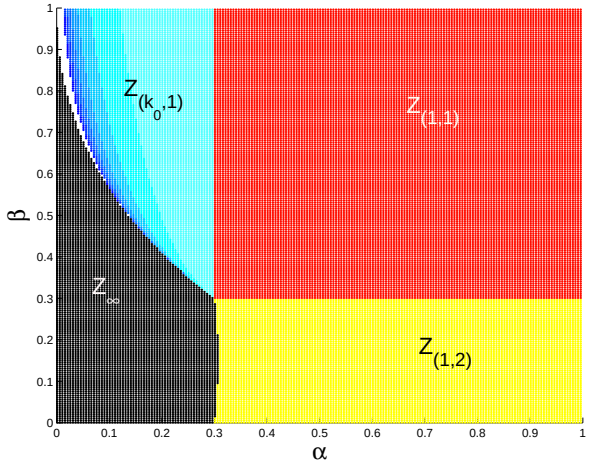


Figure 3: The optimal policy regions in the one channel model with $\lambda = 0.3$.

3 The Tiling Algorithm

As can be seen in Fig. 3, the exact configuration of the policy zones in the channel sensing model is quite complex and, in addition, it depends on the value selected for λ. As a matter of fact, we propose here an algorithm that is relevant in a more abstract framework, which does not rely on the exact shapes of the policy zones in the channel model. This abstract framework also highlights the features of the problem that are of some importance for optimal sensing; these are summarized as Assumptions 1 and 3 in Section 3.3 below. It is also conjectured that the Tiling Algorithm could be useful for other models, although we do not consider this issue in the current paper.

3.1 The Abstract Model

Consider a POMDP defined by $(\mathsf{X}, \mathsf{A}, \mathsf{Y}, Q_\theta, f, r)$, where X is the discrete state space, Y is the observation space, A is the finite set of actions, $Q_\theta : \mathsf{X} \times \mathsf{A} \times \mathsf{X} \to [0, 1]$ is the transition probability, $f : \mathsf{X} \times \mathsf{A} \to \mathsf{Y}$ is the observation function, $r : \mathsf{X} \times \mathsf{A} \to \mathbb{R}$ is the bounded reward function and $\theta \in \Theta$ denotes an unknown parameter. Given the current hidden state $x \in \mathsf{X}$ of the system, and a control action $a \in \mathsf{A}$, the probability of the next state $x' \in \mathsf{X}$ is given by $Q_\theta(x, a; x')$. At each time step t, one chooses an action $A_t = \pi(A_{0:t-1}, Y_{0:t-1})$

according to a *policy* π, and hence observes $Y_t = f(X_t, A_t)$ and receives the reward $r(X_t, A_t)$. Without loss of generality, we assume that for all $x \in \mathsf{X}$, for all $a \in \mathsf{A}$, $r(x, a) \leq 1$.

Since we are interested in rewards accumulated over finite but large horizons, we will consider the average (or long-term) reward criterion defined by

$$V_\theta^\pi = \lim_{n\to\infty} \frac{1}{n} \mathbb{E}_\theta^\pi \left(\sum_{t=1}^{n} r(X_t, A_t) \right) , \tag{6}$$

where π denotes a fixed policy. The notation V_θ^π is meant to highlight the fact that the average reward depends on both the policy π and the actual parameter value θ. For a given parameter value, the optimal long-term reward is defined as $V_\theta^* = \sup_\pi V_\theta^\pi$ and π_θ^* denotes the associated optimal policy. We assume that the dependence of V_θ^π and π_θ^* with respect to θ is fully known. In addition, there exists a particular default policy π_0 under which the parameter θ can be consistently estimated. In the channel sensing problem, this policy π_0 consists in continually observing the channel so as to estimate the transition probabilities by direct counting.

Given the above, one can partition the parameter space Θ into non-intersecting subsets, $\Theta = \bigcup_i Z_i$, such that each policy zone Z_i corresponds to a single optimal policy, which we denote by π_i^*. In other words, for any $\theta \in Z_i$, $V_\theta^* = V_\theta^{\pi_i^*}$. In each policy zone Z_i, the corresponding optimal policy π_i^* is assumed to be known as well as the long-term reward function $V_\theta^{\pi_i^*}$ for any $\theta \in \Theta$.

3.2 The Tiling Algorithm (TA)

We denote by $\hat{\theta}_t$ the parameter estimate obtained after t steps of the exploration policy and by Δ_t the associated confidence region, whose construction will be made more precise below. The principle of the Tiling Algorithm is to use the policy zones $(Z_i)_i$ to determine the length of the exploration phase: basically, the exploration phase will last until the estimated confidence region Δ_t fully enters one of the policy zones. It turns out however that this naive principle does not allow for a sufficient control of the expected duration of the exploration phase, and, hence, of the algorithm's regret. In order to deal with parameter values located close to the borders of policy zones, one needs to introduce additional *frontier zones* $(F_j(n))_j$ that will shrink at a suitable rate with the time horizon n. In Fig. 4, we represent the tiling of the parameter space for an

hypothetical example with three distinct optimal policy zones. In this case, there are four frontier zones: one between each pair of policy zones ($F_1(n)$, $F_2(n)$ and $F_3(n)$) and another ($F_4(n)$) for the intersection of all the policy zones.

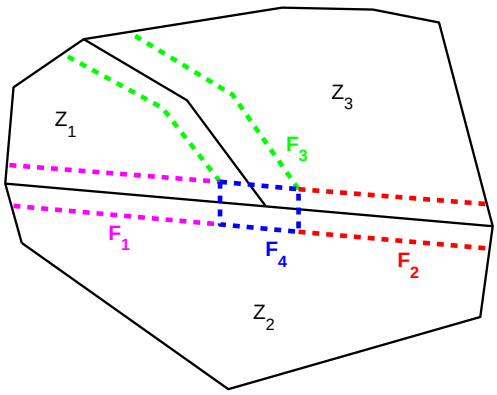


Figure 4: Tiling of the parameter space for an example with three distinct optimal policy zones.

Let

$$T_n = \inf\{t \geq 1 : \exists i,\ \Delta_t \subset Z_i \text{ or } \exists j,\ \Delta_t \subset F_j(n)\} \tag{7}$$

denote the random instant where the exploration terminates. Note that the frontier zones $(F_j(n))_j$ depend on n. Indeed, the larger n the smaller the frontier zones can be in order to balance the length of the exploration phase and the loss due to the possible choice of a suboptimal policy. The Tiling Algorithm consists in using the default exploratory policy π_0 until the occurrence of the stopping time T_n, according to (7). From T_n onward, the algorithm then selects a policy to use during the remaining time as follows: if at the end of the exploration phase, the confidence region is fully included in a policy zone Z_i, then the selected policy is π_i^*; otherwise, the confidence region is included in a frontier zone $F_j(n)$ and the selected policy is any optimal policy π_k^* *compatible* with the frontier zone $F_j(n)$. An optimal policy π_k^* is said to be *compatible* with the frontier zone $F_j(n)$ if the intersection between the policy zone Z_k and the frontier zone is non empty. In the example of Fig. 4, for instance, π_1^* and π_2^* are compatible with the frontier zone $F_1(n)$, while all the optimal policies $(\pi_i^*)_{i=1,2,3}$ are compatible with the central frontier zone $F_4(n)$. If the exploration terminates in a frontier zone, then one basically does not have enough statistical evidence to favor a particular optimal

policy and the Tiling Algorithm simply selects one of the optimal policies compatible with the frontier zone. Hence, the purpose of frontier zones is to guarantee that the exploration phase will stop even for parameter values for which discriminating between several neighboring optimal policies is challenging. Of course, in practice, there may be other considerations that suggest to select one compatible policy rather than another but the general regret bound below simply assumes that any compatible policy is selected at the termination of the exploration phase.

3.3 Performance Analysis

To evaluate the performance of this algorithm, we will consider the regret, for the prescribed time horizon n, defined as the difference between the expected cumulated reward obtained under the optimal policy and the one obtained following the algorithm,

$$R_n(\theta^*) = \mathbb{E}_{\theta^*}^{\pi^*_{\theta^*}} \left[\sum_{t=1}^{n} r(X_t, A_t) \right] - \mathbb{E}_{\theta^*}^{\mathrm{TA}} \left[\sum_{t=1}^{n} r(X_t, A_t) \right] , \tag{8}$$

where θ^* is the unknown parameter value. For any subset B of Θ, denote by $\delta(B) = \sup\{\|\theta - \theta'\|_\infty,\ \theta, \theta' \in B\}$ the diameter of B. To obtain bounds for $R_n(\theta^*)$ that do not depend on θ^*, we will need the following assumptions.

Assumption 1. *The confidence region Δ_t is constructed so that there exists constants $c_1, c_1', n_{\min} \in \mathbb{R}_+$ such that, for all $\theta \in \Theta$, for all $n \geq n_{\min}$, for all $t \leq n$, $\mathbb{P}_\theta \left(\theta \in \Delta_t,\ \delta(\Delta_t) \leq c_1 \frac{\sqrt{\log n}}{\sqrt{t}} \right) \geq 1 - c_1' \exp\{-\frac{1}{3} \log n\}$.*

Assumption 2. *Given a size $\epsilon(n)$, one may construct the frontier zones $(F_j(n))_j$ such that there exists constants $c_2, c_2' \in \mathbb{R}_+$ for which*

- *$\delta(\Delta_t) \leq c_2 \epsilon(n)$ implies that there exists either i such that $\Delta_t \subset Z_i$ or j such that $\Delta_t \subset F_j(n)$,*
- *if $\theta \in F_j(n)$, there exists $\theta' \in Z_i$ such that $\|\theta - \theta'\|_\infty \leq c_2' \epsilon(n)$, for all policy zones Z_i compatible with $F_j(n)$ (i.e., such that $Z_i \bigcap F_j(n) \neq \emptyset$).*

Assumption 3. *For all i, there exists $d_i \in \mathbb{R}_+$ such that for all $\theta, \theta' \in \Theta$, $|V_\theta^{\pi_i^*} - V_{\theta'}^{\pi_i^*}| \leq d_i \|\theta - \theta'\|_\infty$.*

Assumption 1 pertains to the construction of the confidence region and is usually met by standard applications of the Hoeffding inequality. The constant 1/3 is meant to match the worst-case rate given in Theorem 1 below. Assumption 2 formalizes the idea that the frontier zones should allow any confidence region of diameter less than $\epsilon(n)$ to be fully included either in an original policy zone or in a frontier zone, while at the same time ensuring that, locally, the size of the frontier is of order $\epsilon(n)$. The applicability of the Tiling Algorithm crucially depends on the construction of these frontiers. Finally, Assumption 3 is a standard regularity condition (Lipschitz continuity) which is usually met in most applications. The performance of the tiling approach is given by the following theorem, which is proved in Appendix A.

Theorem 1. *Under Assumptions 1, 2 and 3, and for all $n \geq n_{\min}$, the duration of the exploration phase is bounded, in expectation, by*

$$\mathbb{E}_{\theta^*}(T_n) \leq c \frac{\log n}{\epsilon^2(n)} \,, \tag{9}$$

and the regret by

$$R_n(\theta^*) \leq \mathbb{E}_{\theta^*}(T_n) + c' n \epsilon(n) + c'' n^{2/3} \,, \tag{10}$$

where $c = (c_1/c_2)^2$, $c' = c_2' \max_{i,k}(d_i + d_k)$ and $c'' = c_1'$. The minimal worst-case regret is obtained when selecting $\epsilon(n)$ of the order of $(\log n/n)^{1/3}$, which yields the bound $R_n(\theta^*) \leq C(\log n)^{1/3}\, n^{2/3}$ for some constant C.

The duration bound in (9) follows from the observation that exploration is guaranteed to terminate only when the confidence region defined by Assumption 1 reaches a size which is of the order of the diameter of the frontier, that is, $\epsilon(n)$. The second term in the right-hand side of (10) corresponds to the maximal regret if the exploration terminates in a frontier zone. The rate $(\log n)^{1/3} n^{2/3}$ is obtained when balancing these two terms ($\mathbb{E}_{\theta^*}(T_n)$ and $c' n \epsilon(n)$). A closer examination of the proof in Appendix A shows that if one can ensure that the exploration indeed terminates in one of the policy regions Z_i, then the regret may be bounded by an expression similar to (10) but without the $c' n \epsilon(n)$ term. In this case, by modifying slightly Assumption 1, one can obtain logarithmic regret bounds.

Assumption 4. *The confidence region Δ_t is constructed so that there exist constants $c_1, c_1', n_{\min} \in \mathbb{R}_+$, $x > 1$ such that, for all $\theta \in \Theta$, for all $n \geq n_{\min}$, for all $t \leq n$, $\mathbb{P}_\theta \left(\theta \in \Delta_t,\ \delta(\Delta_t) \leq c_1 \frac{\sqrt{x}}{\sqrt{t}} \right) \geq 1 - c_1' \exp\{-2x\}$.*

Doing so, however, requires to introduce additional constraints to guarantee that exploration terminates into a policy region rather than in a frontier. These constraints typically take the form of an assumed sufficient margin between the actual parameter value θ^* and the borders of the associated policy zone. This is formalized in Theorem 2 which is proved in Appendix B.

Theorem 2. *Consider θ^* in a policy zone Z such that there exists κ for which $\min_{\theta \notin Z} \|\theta^* - \theta\|_\infty > \kappa$. Under assumptions 2, 3 and 4, the regret is bounded by $R_n(\theta^*) \leq C(\kappa)\log(n) + C'(\kappa)$ for all $n \geq n_{\min}$ and for some constants $C(\kappa)$ and $C'(\kappa)$ which decrease as κ increases.*

4 Application to Channel Sensing

4.1 Applying the Tiling Algorithm

In the following, we apply the Tiling Algorithm to the channel sensing model presented in Section 2. We previously showed that, for a fixed value of λ, one can tile the parameter space $[0,1] \times [0,1]$ into distinct policy zones. In Fig. 3, this tiling is represented for $\lambda = 0.3$. The frontier zones may be constructed in many ways; as explained in Section 3.3, the only requirement is that Assumption 2 be satisfied. Fig. 5 displays a

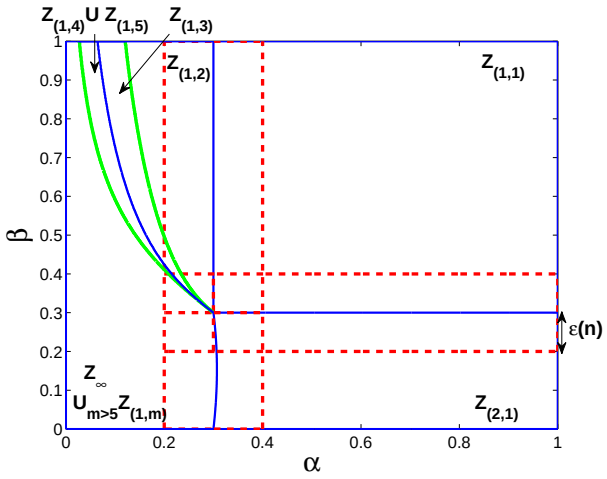


Figure 5: The proposed frontier zones for $\lambda = 0.3$.

possible contruction of the frontier zones. We introduce:

- a rectangular frontier zone between $Z_{(1,1)}$ and $Z_{(1,2)}$;
- a rectangular frontier zone between $Z_{(1,1)}$ and $Z_{(2,1)}$;
- a rectangular frontier zone between $Z_{(1,2)}$ and Z_{∞};
- a square zone centered at the joining point (λ, λ) of those three frontier zones.

The width $\epsilon(n)$ of the rectangular frontier zones depends on the time horizon n. As mentioned above, there is an accumulation of policy zones in the upper left corner (see Fig. 3); to address this issue, we aggregate the zones $Z_{(2,1)}$, and $Z_{(3,1)}$ on one hand and the non-observation zone Z_{∞} with the zones $Z_{(m_0,1)}$ for $m_0 \geq 4$ on the other hand. Then, we introduce as a frontier zone between them the union $Z_{(3,1)} \cup Z_{(4,1)} \cup Z_{(5,1)}$. The equation of the two curves delimiting this frontier zone are $V_{\alpha,\beta}^{\pi_{(2,1)}} = V_{\alpha,\beta}^{\pi_{(3,1)}}$ and $V_{\alpha,\beta}^{\pi_{(5,1)}} = V_{\alpha,\beta}^{\pi_{(6,1)}}$ (see (4)). More zones could of course be constructed but for practical purposes the proposed tiling is already satisfactory as the the value function has very limited variations among the zones that are agregated.

Note that this tiling construction only needs to be done once, prior to parameter estimation. The Tiling Algorithm then consists in estimating the parameter $\theta = (\alpha, \beta)$ until the estimated confidence region fully enters, either, one of the policy zones, or, one of the frontier zones. The exploration policy, denoted by π_0 in Section 3, consists in always sensing the channel. In that way, the parameter is easily estimated by direct counting: at time t, the estimated parameter is given by

$$\hat{\alpha}_t = \frac{N_t^{0,1}}{N_t^0} \text{ and } \hat{\beta}_t = \frac{N_t^{1,1}}{N_t^1}, \tag{11}$$

where N_t^0 (resp. N_t^1) is the number of visits to 0 (resp. 1) until time t and $N_t^{0,1}$ (resp. $N_t^{1,1}$) is the number of visits to 0 (resp. 1) followed by a visit to 1 until time t. Once the exploration phase ended, the secondary user follows the optimal policy pertaining to the estimated parameter.

In order to verify that this model satisfies the conditions of Theorem 1, we need to make an irreducibility assumption on the Markov chain.

Assumption 5. *There exists η such that $(\alpha, \beta) \in \Theta = [\eta, 1-\eta]^2$.*

We then define the confidence region as the rectangle

$$\Delta_t = \left[\hat{\alpha}_t \pm \sqrt{\frac{\log n}{6N_t^0}}\right] \times \left[\hat{\beta}_t \pm \sqrt{\frac{\log n}{6N_t^1}}\right] . \tag{12}$$

Assumption 5 bounds the expected time during which the channel's state stay fixed. It is related to the "diameter" assumption introduced in Definition 1 of [4] and is necessary to obtain confidence intervals of the form given in (12).

To prove that the regret of the Tiling Algorithm in the channel sensing model is bounded, we need to check that the three assumptions of Theorem 1 are satisfied. It is shown in Appendix D that Assumption 1 holds. We show that the parameter space partitioning scheme discussed at the beginning of Section 4.1 does satisfy Assumption 2. The first part of this assumption requires that any confidence region of diameter less than $c_2\epsilon(n)$ is fully included either in a policy zone or in a frontier zone. This is trivially satisfied for all the proposed rectangular frontier zones taking $c_2 = 1$. The difficulty concerning the frontier zone $Z_{(3,1)} \cup Z_{(4,1)} \cup Z_{(5,1)}$ is that the width of the frontier decreases when α and β approach λ. However, the central zone addresses this problem; in fact, the two curves defined by $V_{\alpha,\beta}^{\pi_{(3,1)}} = V_{\alpha,\beta}^{\pi_{(4,1)}}$ and $V_{\alpha,\beta}^{\pi_{(5,1)}} = V_{\alpha,\beta}^{\pi_{(6,1)}}$ both intersect the vertical line $\alpha = \lambda - \epsilon(n)$ at the left border of the rectangular frontier zones; it is sufficient to choose c_2 such that $c_2\epsilon(n)$ is equal to the distance between those two intersection points. Moreover, the second item of Assumption 2, requiring that the distance between any θ in the frontier zone and all the compatible policy zones is upper-bounded, is obviously satified. Finally, for all optimal policy, the average reward, defined in (4) and (5), is a Lipschitz continuous function of (α, β) for $\alpha, \beta \in [\eta, 1-\eta]$, and, hence, the third condition is also satisfied.

4.2 Experimental Results

As suggested by Theorems 1–2, the length of the exploration phase following the Tiling Algorithm depends on the value of the true parameter (α^*, β^*). In addition, for a fixed value of (α^*, β^*), the length of the exploration varies from one run to another, depending on the size of the confidence region. To illustrate these effects, we take $\lambda = 0.3$ and we consider two different values of the parameters: $(\alpha^*, \beta^*) = (0.8, 0.05)$, which is included in the policy zone $Z_{(1,2)}$ and far from any frontier zone, and, $(\alpha^*, \beta^*) = (0.8, 0.2)$ which lies in the frontier

zone between $Z_{(1,1)}$ and $Z_{(1,2)}$ and is close to the border of the frontier zone. The corresponding empirical distributions of the length of the exploration phase are represented in Fig. 6. Remark that the shape of these two distributions are quite different and that the empirical mean of the length of the exploration phase is lower for a parameter which is far from any frontier zone than for a parameter which is close to the border of a frontier zone.

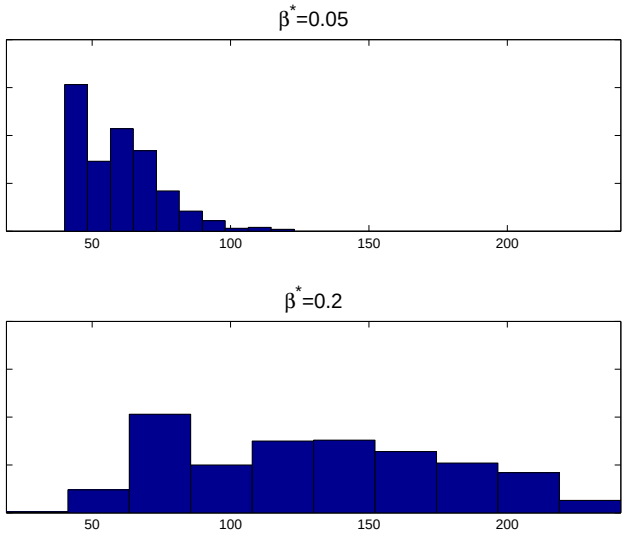


Figure 6: Distribution of the length of the exploration phase following the Tiling Algorithm for $(\alpha^*, \beta^*) = (0.8, 0.05)$ and for $(\alpha^*, \beta^*) = (0.8, 0.2)$.

In Fig. 7, we compare the cumulated regrets R_n^{TA} of the Tiling Algorithm to the regrets $R_n^{DL}(l_{expl})$ of an algorithm with a deterministic length of exploration phase l_{expl}. Both algorithms are run with $(\alpha^*, \beta^*) = (0.8, 0.05)$. We use two values of l_{expl}: one lower ($l_{expl} = 20$) and the other larger ($l_{expl} = 300$) than the average length of the exploration phase following the tiling algorithm which ranges between 40 and 150 for this value of the parameter (see Fig. 6). The algorithms are run four times independently and every cumulated regret are represented in Fig. 7. Note that, (α^*, β^*) being in the interior of a policy zone (i.e. not in a frontier zone), the regret of the Tiling Algorithm is null during the exploitation phase since the optimal policy for the true parameter is used. Similarly, when the deterministic length l_{expl} of the exploration phase is sufficiently large, the estimation of the parameter is quite precise, therefore the regret during the

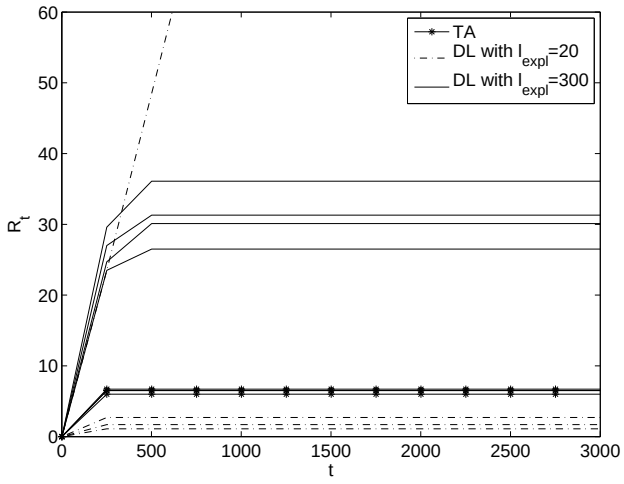


Figure 7: Comparison of the cumulated regret of the Tiling Algorithm (shaped markers) and an algorithm with a deterministic length of exploration phase equal to 20 (dashed line) or equal to 300 (solid line) for $(\alpha^*, \beta^*) = (0.8, 0.05)$

exploitation phase is null. On the other hand, a too large value of l_{expl} increases the regret during the exploration phase: we observe in Fig. 7 that the regret $R_n^{DL}(l_{expl})$ with $l_{expl} = 300$ is larger than R_n^{TA}. When the deterministic length of the exploration phase is smaller than the average length of the exploration phase following the tiling algorithm, either the parameter is estimated precisely enough and then $R_n^{DL}(l_{expl})$ is smaller than R_n^{TA}, or, the estimated value is too far away from the actual value and the policy followed during the exploitation phase is not the optimal one. In the latter case, the regret is not null during the exploitation phase and $R_n^{DL}(l_{expl})$ is noticeably large. This can be observed in Fig. 7: in three of the four runs, the cumulated regret $R_n^D L(l_{expl})$ with $l_{expl} = 20$ (dashed line) are small, whereas in the remaining run it sharply and constantly increases.

5 Conclusions

The Tiling Algorithm is a model-based reinforcement learning algorithm applicable to channel sensing. This algorithm is meant to adequately balance exploration and exploitation by adaptively monitoring the duration of the exploration phase so as to guarantee a $(\log n)^{1/3}\, n^{2/3}$ worst-case regret bounds for a pre-specified

finite horizon n. Furthermore, it has been shown in Theorem 2 that in large regions of the parameter space, the regret can be guaranteed to be logarithmic. In numerical experiments, it has been observed that the Tiling Algorithm is indeed able to adapt the length of the exploration phase, depending on the sequence of observations.

Although, we have focussed in this paper on the single channel case, the Tiling Algorithm can also be used to address at least some cases of the original N channel model depicted in Fig. 1. In fact, as mentioned in Sec. 3, the Tiling Algorithm applies in any situation where the planning problem can be solved explicitly. Unfortunately, the general multi-channel model does not fall into this category [14]; but, in the particular case where the channels are stochastically identical (i.e., share common transition parameters), an explicit near-optimal planning strategy based on the so-called Whittle index has been pointed out by [23]. This policy only depends on whether the system is positively correlated ($\alpha \leq \beta$) or negatively correlated ($\beta \leq \alpha$). It is therefore possible to apply the Tiling Algorithm considering the two resulting policy zones separated by the line $\alpha = \beta$ and defining the frontier zone as $F(n) = \{(\alpha, \beta), |\alpha - \beta| \leq \epsilon(n)\}$. As in Sec. 4.1, the three assumptions of Theorem 1 are easily checked. The Tiling Algorithm thus provides a formal decision rule which ensures a small regret: the secondary user senses a channel to estimate the parameters folowing (11) until the confidence region is fully included either in one of the two policy zones or in the frontier zone, then he applies the policy related to the estimated parameter.

A Appendix: Proof of Theorem 1

The confidence zone is such that, at the end of the exploration phase, $\mathbb{P}_{\theta^*}\left(\theta^* \in \Delta_t \ , \ \delta(\Delta_t) \leq c_1\sqrt{\log n}/\sqrt{t}\right) \geq 1 - c_1' \exp\{-\frac{1}{3}\log n\}$. At the end of the exploration phase, if the true parameter θ^* is in the confidence region, there are two possibilities: either the confidence zone Δ_t is included in a policy zone Z_i or it is included in a frontier zone $F_j(n)$. If the confidence zone is in a policy region, the regret is equal to the sum of the duration of the exploration phase and of the loss corresponding to the case where the confidence region is violated: $R_n(\theta^*) = \mathbb{E}_{\theta^*}(T_n) + c_1' n \exp\{-\frac{1}{3}\log n\}$. If the confidence zone is in a frontier region $F_j(n)$, an additional term of the regret is the loss due to the fact that the policy selected at the end of the exploration phase is not necessarily the optimal one for the true parameter θ^*. Let π_i^* denote the optimal

policy for θ^* and π_k^* the selected policy. Note that Z_i and Z_k are compatible with $F_j(n)$. The loss is $V_{\theta^*}^{\pi_i^*} - V_{\theta^*}^{\pi_k^*} = (V_{\theta^*}^{\pi_i^*} - V_{\theta}^{\pi_i^*}) + (V_{\theta}^{\pi_k^*} - V_{\theta^*}^{\pi_k^*}) + (V_{\theta}^{\pi_i^*} - V_{\theta}^{\pi_k^*})$, where $\theta \in Z_k \bigcap F_j(n)$. The last term is negative since π_k^* is the optimal policy for θ. The two other terms can be bounded using Assumption 3. Then, $|V_{\theta^*}^{\pi_i^*} - V_{\theta^*}^{\pi_k^*}| \leq (d_i + d_k)\|\theta^* - \theta\|_\infty$. According to Assumption 2, one can choose θ such that $\|\theta^* - \theta\|_\infty < c_2'\epsilon(n)$ for which $R_n(\theta^*) \leq \mathbb{E}_{\theta^*}(T_n) + nc'\epsilon(n) + c_1' n \exp\{-\frac{1}{3}\log n\}$, where $c' = c_2' \max_{i,k}(d_i + d_k)$.

The maximal regret is obtained when the confidence region belongs to a frontier zone. According to Assumptions 1 and 2, if t satisfies $c_1(\log n/t)^{1/2} < c_2\epsilon(n)$ then $t \geq T_n$, with large probability. Therefore, $\mathbb{E}_{\theta^*}(T_n) \leq (c_1^2 \log n)/(c_2\epsilon(n))^2$. The regret is then bounded by

$$\max_{\theta^*} R_n(\theta^*) \leq \frac{c_1^2 \log n}{c_2^2 \epsilon^2(n)} + nc'\epsilon(n) + c_1' n \exp\{-\frac{1}{3}\log n\} ,$$

which is minimized for $\epsilon(n) = \left(\frac{2c_1^2 \log n}{c_2^2 c' n}\right)^{1/3}$.

B Appendix: Proof of Theorem 2

The condition $\min_{\theta \notin Z} |\theta^* - \theta| > \kappa$ means that the distance between θ^* and any border of the policy zone Z is larger than κ. Hence, as soon as $\delta(\Delta_t) \leq \kappa$, the confidence region Δ_t is included in the policy zone Z. The regret of the Tiling Algorithm is then equal to $R_n(\theta^*) = \mathbb{E}_{\theta^*}(T_n) + c_1' n \exp\{-2x\}$. According to Assumption 4, if t satisfies $c_1(x/t)^{1/2} < \kappa$ then $t \geq T_n$ with large probability. Therefore, $\mathbb{E}_{\theta^*}(T_n) \leq c_1 x/\kappa^2$ and the regret is bounded by $R_n(\theta^*) = \frac{c_1 x}{\kappa^2} + c_1' n \exp\{-2x\}$, which is minimized for $x = \frac{\log(2c_1' n \kappa^2/c_1^2)}{2}$. For this value of x, we have $R_n(\theta^*) = \frac{c_1^2}{2\kappa^2}(\log(n) + \log(2c_1'\kappa^2/c_1^2) + 1)$.

C Appendix: Average Reward for the Channel Sensing Model

Let P^π and μ^π denote, respectively, the transition probability matrix and the stationary probability of the internal state $\{(K_t, U_t)\}_t$ Markov chain, when following a policy π. The average reward of a policy π, defined

in (6), can be written as a function of the stationary probability μ^π (see [18]):

$$V^\pi_{\alpha,\beta} = \sum_{k\in\mathbb{N}^*} \sum_{u\in\{0,1\}} \mu^\pi(k,u) \left[p^{k,u}_{\alpha,\beta} \mathbb{1}_{\{\pi(k,u)=1\}} + \lambda \mathbb{1}_{\{\pi(k,u)=0\}} \right] . \tag{13}$$

Therefore, to compute the average reward of the policies introduced in Section 2, it is sufficient to determine the stationary probability of the internal state markov chain under those policies.

Policy $\pi_{(m_0,m_1)}$ with m_0 , $m_1 \in \mathbb{N}^*$. Under the policy $\pi_{(m_0,m_1)}$, the internal state (K_t, U_t) can only take one of the following value: $(k,0)$ with $1 \le k \le m_0$ or $(k',1)$ with $1 \le k' \le m_1$. It is then easy to compute the transition probabilities $P^{\pi_{(m_0,m_1)}}$ between those states:

- for all $1 \le k \le m_0 - 1$, $P^{\pi_{(m_0,m_1)}}((k,0),(k+1,0)) = 1$,
- for all $1 \le k \le m_1 - 1$, $P^{\pi_{(m_0,m_1)}}((k,1),(k+1,1)) = 1$,
- $P^{\pi_{(m_0,m_1)}}((m_0,0),(1,1)) = p^{m_0,0}_{\alpha,\beta} = 1 - P^{\pi_{(m_0,m_1)}}((m_0,0),(1,0))$,
- $P^{\pi_{(m_0,m_1)}}((m_1,1),(1,1)) = p^{m_1,1}_{\alpha,\beta} = 1 - P^{\pi_{(m_0,m_1)}}((m_1,1),(1,0))$.

Solving the equation $\mu^{\pi_{(m_0,m_1)}} P^{\pi_{(m_0,m_1)}} = \mu^{\pi_{(m_0,m_1)}}$, we determine the stationary probability:

$$\mu^{\pi_{(m_0,m_1)}}(k,0) = \frac{1 - p^{m_1,1}_{\alpha,\beta}}{m_1 p^{m_0,0}_{\alpha,\beta} + m_0(1 - p^{m_1,1}_{\alpha,\beta})} \quad \text{for all } 1 \le k \le m_0 ,$$
$$\mu^{\pi_{(m_0,m_1)}}(k,1) = \frac{p^{m_0,0}_{\alpha,\beta}}{m_1 p^{m_0,0}_{\alpha,\beta} + m_0(1 - p^{m_1,1}_{\alpha,\beta})} \quad \text{for all } 1 \le k \le m_1 .$$

Finally, using (13), we obtain the average reward:

$$V^{\pi_{(m_0,m_1)}}_{\alpha,\beta} = \frac{p^{m_0,0}_{\alpha,\beta} + \lambda[(m_1-1)p^{m_0,0}_{\alpha,\beta} + (m_0-1)p^{m_1,1}_{\alpha,\beta}]}{m_0(1 - p^{m_1,1}_{\alpha,\beta}) + m_1 p^{m_0,0}_{\alpha,\beta}} , \text{ for } m_0, m_1 \in \mathbb{N}^*.$$

Policy π_∞. If m_0 or m_1 is equal to infinity, the secondary user never observes the channel after a finite number of time steps. Then, the average reward is $V^{\pi_\infty} = \lambda$.

D Appendix: Confidence Interval for Markov Chains

In this appendix, we prove that the confidence region Δ_t defined in equation (12) satisfies Assumption 1. First, remark that the event $\{\delta(\Delta_t) \leq c_1\frac{\sqrt{\log n}}{\sqrt{t}}\} = \{N_t^0 \geq c\frac{\eta t}{2},\ N_t^1 \geq c\frac{\eta t}{2}\}$ for $c_1 = 2/\sqrt{3c\eta}$. Hence, using the Hoeffding inequality, we have $\mathbb{P}_{(\alpha,\beta)}\left((\alpha,\beta) \notin \Delta_t,\ \delta(\Delta_t) \leq c_1\frac{\sqrt{\log n}}{\sqrt{t}}\right) \leq 4\exp\{-\frac{1}{3}\log n\}$. Moreover, we need to bound the probability $\mathbb{P}\left(\delta(\Delta_t) > c_1\frac{\sqrt{\log n}}{\sqrt{t}}\right)$. We apply Theorem 2 of [8] to bound $\mathbb{P}\left(N_t^1 < c\frac{\eta t}{2}\right)$. To do so, remark that $\inf_{\alpha,\beta}\nu_1 = \eta$ and that the minoration constant $1 - |\beta - \alpha|$ is lower-bounded by 2η. We then have

$$\mathbb{P}\left(N_t^1 < c\frac{\eta t}{2}\right) \leq \mathbb{P}\left(N_t^1 - \nu_1 t < -(1-c/2)\nu_1 t\right) \leq \exp\{-\frac{4\eta^2(t^2\eta(1-c/2)-1/\eta)^2}{2t}\} \leq \exp\{-\frac{1}{3}\log(n)\}\ ,$$

where the last inequality holds for $t \geq t_n \stackrel{\text{def}}{=} (8/3\log(n)\eta^{-4}(2-c)^{-2})^{1/3}$. Similarly, we can show that, for $t \geq t_n$, $\mathbb{P}(N_t^0 < c\frac{\eta t}{2}) \leq \exp\{-\frac{1}{3}\log(n)\}$. Hence, for all $t \geq t_n$, $\mathbb{P}\left(\delta(\Delta_t) > c_1\frac{\sqrt{\log n}}{\sqrt{t}}\right) \leq 2\exp\{-\frac{1}{3}\log(n)\}$. In addition, for all $t < t_n$, $c_1\sqrt{\frac{\log n}{t}} > c_1\sqrt{\frac{\log n}{t_n}} \geq 1$, for $n \geq \exp\{3 \times 2^{-3/2}c^{3/2}(2-c)^{-1}\eta^{-1/2}\} \stackrel{\text{def}}{=} n_{\min}$. Then, for $t < t_n$ and $n \geq n_{\min}$, the event $\{\delta(\Delta_t) \leq c_1\frac{\sqrt{\log n}}{\sqrt{t}}\}$ is always verified. To conclude, we have

$$\begin{aligned}
&\mathbb{P}_{(\alpha,\beta)}\left((\alpha,\beta) \in \Delta_t,\ \delta(\Delta_t) \leq c_1\frac{\sqrt{\log n}}{\sqrt{t}}\right) \\
&\quad \geq 1 - \mathbb{P}_{(\alpha,\beta)}\left(\delta(\Delta_t) > c_1\frac{\sqrt{\log n}}{\sqrt{t}}\right) - \mathbb{P}_{(\alpha,\beta)}\left((\alpha,\beta) \notin \Delta_t,\ \delta(\Delta_t) \leq c_1\frac{\sqrt{\log n}}{\sqrt{t}}\right) \geq 1 - 6\exp\{-\frac{1}{3}\log(n)\}\ .
\end{aligned}$$

References

[1] I. F. Akyildiz, L. Won-Yeol, M. C. Vuran, and S. Mohanty. A survey on spectrum management in cognitive radio networks. *IEEE Communications Magazine*, 46(4):40–48, 2008.

[2] V. Anantharam, P. Varaiya, and J. Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays. II: Markovian rewards. *IEEE Trans. Autom. Control*, 32:977–982, 1987.

[3] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002.

[4] P. Auer, T. Jaksch, and R. Ortner. Near-optimal regret bounds for reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 21, 2009.

[5] A. Cassandra, L. Kaelbling, and M. Littman. Acting optimally in partially observable stochastic domains. *Proceedings of the Twelfth National Conference on Artificial Intelligence*, 2:1023–1028, 1994.

[6] Y. Chen, Q. Zhao, and A. Swami. Joint design and separation principle for opportunistic spectrum access in the presence of sensing errors. *IEEE Transactions on Information Theory*, 54(5):2053–2071, 2008.

[7] S. Filippi, O. Cappe, F. Clerot, and E. Moulines. A near optimal policy for channel allocation in cognitive radio. In *Lecture Notes in Computer Science*, Recent Advances in Reinforcement Learning, pages 69–81. Springer, 2008.

[8] P. Glynn and D. Ormoneit. Hoeffding's inequality for uniformly ergodic Markov chains. *Statistics and Probability Letters*, 56(2):143–146, 2002.

[9] S. Guha and K. Munagala. Approximation algorithms for partial-information based stochastic control with Markovian rewards. *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on*, pages 483–493, 2007.

[10] S. Haykin. Cognitive radio: Brain-empowered wireless communications. *IEEE J. Selected Areas Commun.*, 23(2):201–220, 2005.

[11] L. Lai, H. El Gamal, H. Jiang, and H. Vicent Poor. Optimal medium access protocols for cognitive radio networks. In *6th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks and Workshops*, 2008.

[12] J. Le Ny, M. Dahleh, and E. Feron. Multi-UAV dynamic routing with partial observations using restless bandit allocation indices. In *American Control Conference, 2008*, pages 4220–4225, 2008.

[13] K. Liu and Q. Zhao. A restless bandit formulation of opportunistic access: Indexablity and index policy. *5th IEEE Annual Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks Workshops, 2008. SECON Workshops' 08*, pages 1–5, 2008.

[14] K. Liu and Q. Zhao. Decentralized multi-armed bandit with multiple distributed players. In *Information Theory and Applications*, 2010.

[15] X. Long, X. Gan, Y. Xu, J. Liu, and M. Tao. An estimation algorithm of channel state transition probabilities for cognitive radio systems. In *Cognitive Radio Oriented Wireless Networks and Communications*, 2008.

[16] J. Mitola. *Cognitive Radio - An Integrated Agent Architecture for Software Defined Radio*. PhD thesis, Royal Institute of Technology, Kista, Sweden, May 8 2000.

[17] C. Papadimitriou and J. Tsitsiklis. The complexity of optimal queueing network control. *Structure in Complexity Theory Conference, 1994., Proceedings of the Ninth Annual*, pages 318–322, 1994.

[18] M. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc. New York, NY, USA, 1994.

[19] A. Strehl and M. Littman. An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

[20] R. Sutton. *Reinforcement Learning*. Springer, 1992.

[21] A. Tewari and P. Bartlett. Optimistic linear programming gives logarithmic regret for irreducible MDPs. *Advances in Neural Information Processing Systems*, 20:1505–1512, 2008.

[22] P. Whittle. Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability*, 25:287–298, 1988.

[23] Q. Zhao, B. Krishnamachari, and K. Liu. On myopic sensing for multi-channel opportunistic access: Structure, optimality, and performance. *IEEE Transactions on Wireless Communications*, 7(12):5431–5440, 2008.

[24] Q. Zhao, L. Tong, A. Swami, and Y. Chen. Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: A POMDP framework. *IEEE Journal on Selected Areas in Communications*, 25(3):589–600, 2007.