Proceedings of the
47th IEEE Conference on Decision and Control
Cancun, Mexico, Dec. 9-11, 2008

Dynamic Spectrum Access Policies for Cognitive Radio

Jayakrishnan Unnikrishnan and Venugopal V. Veeravalli

Abstract— **We study the problem of dynamic spectrum sensing and access in cognitive radio systems as a partially observed Markov decision process (POMDP). A group of cognitive users cooperatively tries to exploit vacancies in some primary (licensed) channels whose occupancies have a Markovian evolution. We first consider the scenario where the cognitive users are aware of the distribution of the signals they receive from the primary users and we obtain a greedy channel selection and access policy that maximizes the instantaneous reward, while satisfying a constraint on the probability of interfering with licensed transmissions. We also derive an analytical universal upper bound on the performance of the optimal policy.**

We then consider the more practical scenario where the distribution of the signal from the primary is characterized by an unknown random parameter. We develop an algorithm that can learn this random parameter, still guaranteeing the constraint on the interference probability. We also demonstrate the performance gains of all our schemes through simulations.

I. Introduction

Cognitive radios are smart radios that exploit vacancies in licensed spectrum by identifying times when a specific licensed band is not used at a particular place and using this band for unlicensed transmissions without causing interference to the licensed user (referred to as the 'primary'). The cognitive radio (also called the 'secondary user') needs to decide what is the best strategy for selecting the licensed channels for sensing and access. The sensing and access policies should jointly ensure that the probability of interfering with the primary's transmission meets a given constraint.

In this paper, we consider the design of such a joint sensing and access policy, assuming a Markovian structure for the primary spectrum usage on the channels being monitored. In most of the existing schemes in the literature in this field, the authors either assume error-free observations of the channel states [1], [2], [3] or assume that the channel states are learned based on the ACK signals transmitted from the secondary's receivers [4]. We adopt a different strategy in which we use the analog observations made on the channels to track the probability of occupancy of the different channels and obtain a suboptimal solution to the resultant POMDP problem.

In the second part of the paper, we propose and study a more practical problem that arises when the secondary users are not aware of the exact distributions of the signals that they see from the primary transmitters. We assume that these signals have distributions parameterized by an unknown random parameter in a known set. We develop a scheme that learns these parameters online, still satisfying a constraint on the probability of interfering with the primary signals. The learning algorithm converges almost surely to the correct value of the parameter. Through simulations, we show that this scheme gives improved performance over the naive scheme that assumes a worst-case value for the unknown parameter.

II. Problem Statement

We consider a slotted system where a group of secondary users, located close to each other, monitor a set of L primary channels. The state of each primary channel switches between 'occupied' and 'unoccupied' according to the evolution of a Markov chain. The secondary users can cooperatively sense any one out of the L channels in each slot, and can access any one out of the L channels in the same slot. In each slot, the secondary users must satisfy a strict constraint on the probability of interfering with potential primary transmissions that may be going on in any channel. When the secondary users access a channel that is free during a given time slot, they receive a reward proportional to the bandwidth of the channel that they access. The objective of the secondary users is to select the channels for sensing and access in each slot in such a way that their total expected reward accrued over all slots is maximized subject to the constraint on interfering with potential primary transmissions every time they access a channel. Since the secondary users do not have explicit knowledge of the states of the channels, the resultant problem is a constrained partially observable Markov decision process (POMDP) problem.

We assume that all the L channels have equal bandwidth B, and are statistically identical and independent in terms of primary usage. The occupancy of each channel follows a stationary Markov chain. The state of channel a in any time slot k is represented by variable $S_a(k)$ and could be either 1 or 0, where states 0 and 1 correspond to the channel being available or unavailable for secondary access, respectively. The statistics of this Markov process are assumed to be known by the secondary users.

The secondary system includes a decision center that has access to all the analog observations made by the cooperating secondary users. The decisions about which channels to sense and access in each slot are made at this decision center. When channel a is sensed in slot k, we use $\underline{X}_a(k)$ to denote the vector of observations made by the different cooperating users on channel a in slot k. The statistics of these observations are assumed to be time-invariant and

This work was supported by a Vodafone Foundation Graduate Fellowship and by NSF Grant CCF 07-29031, through the University of Illinois

The authors are with the Department of Electrical and Computer Engineering and Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, IL 61820, USA {junnikr2, vvv}@illinois.edu

978-1-4244-3124-3/08/$25.00 ©2008 IEEE

conditionally independent conditioned on the states of the channel. The observations on channel a in slot k have distinct joint probability density functions f_0 and f_1 when $S_a(k) = 0$ and $S_a(k) = 1$ respectively. The collection of all observations up to slot k is denoted by $\underline{X}^k$, and the collection of observations on channel a up to slot k is denoted by $\underline{X}^k_a$. The channel sensed in slot k is denoted by u_k and the set of time slots up to slot k when channel a was sensed is denoted by K^k_a. The decision to access channel a in slot k is denoted by a binary variable $\delta_a(k)$, which takes value 1 when channel a is accessed in slot k, and 0 otherwise.

Whenever the secondary users access a free channel in some time slot k, they get a reward B equal to the common bandwidth of each of the L channels. The secondary users should satisfy the following constraint on the probability of interfering with the primary transmissions in each slot:

$$\mathsf{P}(\{\delta_a(k) = 1\}|\{S_a(k) = 1\}) < \zeta$$

In order to simplify the structure of the access policy, we also assume that in each slot the decision to access a channel is made using only the observations made in that slot. Hence it follows that in each slot k, the secondary users can access only the channel they sense in slot k, say channel a. Furthermore, the access decision must be based on a binary hypothesis test [5] between the two possible states of channel a, performed on the observation $\underline{X}_a(k)$. The optimal test [5] is to compare the joint log-likelihood ratio,

$$\mathcal{L}(\underline{X}_a(k)) = \log\left(\frac{f_1(\underline{X}_a(k))}{f_0(\underline{X}_a(k))}\right)$$

to some threshold Δ that is chosen to satisfy,

$$\mathsf{P}\left(\{\mathcal{L}(\underline{X}_a(k)) < \Delta\}\,|\{S_a(k) = 1\}\right) = \zeta \qquad (1)$$

and the optimal access decision would be to access the sensed channel whenever the threshold exceeds the joint log-likelihood ratio. Hence,

$$\delta_a(k) = \mathcal{I}_{\left\{\mathcal{L}(\underline{X}_a(k))<\Delta\right\}}\mathcal{I}_{\{u_k=a\}} \qquad (2)$$

and the reward obtained in slot k can be expressed as

$$\hat{r}_k \;=\; B\mathcal{I}_{\left\{S_{u_k}(k)=0\right\}}\mathcal{I}_{\left\{\mathcal{L}(\underline{X}_{u_k}(k))<\Delta\right\}} \qquad (3)$$

where $\mathcal{I}_E$ represents the indicator function of event E. The main advantage of the structure of the access policy given in (2) is that we can obtain a simple sufficient statistic for the resultant POMDP without having to keep track of all the past observations, as discussed later. It also has the added advantage that the secondary users can set the thresholds Δ to meet the constraint on the probability of interfering with the primary transmissions without relying on their knowledge of the Markov statistics. This follows from the fact that the access decisions are made using only the observations from the current slot and the threshold is selected to satisfy the interference constraint using only the observations from the current slot. Therefore under this scheme, the interference constraint is satisfied even if the secondary users do not have accurate knowledge of the Markov statistics.

Our objective is to generate a policy that makes optimal use of primary spectrum subject to the interference constraint. Since we do not know the exact number of slots over which we need to optimize the expected accrued reward, we introduce a discount factor $\alpha \in (0, 1)$ and aim to solve the infinite horizon dynamic program with discounted rewards. That is, we seek the sequence of channels $\{u_0, u_1, \ldots\}$, such that the $\sum_{k=0}^{\infty} \alpha^k \mathsf{E}[\hat{r}_k]$ is maximized, where the expectation is performed over the random observations and channel state realizations. It can be shown [6] that,

$$\begin{aligned} \mathsf{E}[\hat{r}_k] &= \mathsf{E}\left[B(1-\hat{\epsilon})\mathcal{I}_{\{S_{u_k}(k)=0\}}\right] && (4)\\ \text{where } \hat{\epsilon} &= \mathsf{P}(\{\mathcal{L}(\underline{X}_a(k)) > \Delta\}|\{S_a(k) = 0\}) && (5)\end{aligned}$$

Under the assumption of identical channels and time-invariant observation-statistics, $\hat{\epsilon}$ given by (5) is a constant independent of k. From the structure of the expected reward in (4) it follows that we can redefine our problem such that the reward in slot k is now given by:

$$r_k \;=\; B(1-\hat{\epsilon})\mathcal{I}_{\{S_{u_k}(k)=0\}} \qquad (6)$$

and the optimization problem is equivalent to maximizing $\sum_{k=0}^{\infty} \alpha^k \mathsf{E}[r_k]$. Thus the problem of spectrum sensing and access boils down to choosing the optimal channel to sense in each slot. Whenever the secondary users sense some channel and make observations with log-likelihood ratio lower than the threshold, they access that channel. Thus we have converted the constrained POMDP problem into an unconstrained POMDP problem.

III. Dynamic programming

The state of the system in slot k denoted by

$$\underline{S}(k) = (S_1(k), S_2(k), \ldots, S_L(k))^\top$$

is the vector of states of the L channels that have independent and identical Markovian evolutions. The channel to be sensed in slot k is decided in slot $k-1$ and is given by

$$u_k = \mu_k(I_{k-1})$$

where μ_k is a deterministic function and $I_k := (\underline{X}^k, u^k)$ represents the net information available at slot k. The observations made in slot k can be expressed as:

$$\underline{X}_{u_k}(k) = \underline{h}(S_{u_k}(k), u_k, v_k)$$

where $\underline{h}$ is a deterministic function and v_k is a random variable whose distribution conditioned on $\underline{S}(k)$ and u_k is known. The reward obtained in slot k is a function of the state in slot k and u_k as given by (6). We seek the sequence of channels $\{u_0, u_1, \ldots\}$, such that $\sum_{k=0}^{\infty} \alpha^k \mathsf{E}[r_k]$ is maximized. Under this formulation it can be easily verified that this problem is essentially a standard dynamic programming problem with imperfect observations. It is known [7] that for such a POMDP problem, a sufficient statistic at the end of any time slot k, is the probability distribution of the system state $\underline{S}(k)$, conditioned on all the past observations and decisions, given by $\mathsf{P}(\{\underline{S}(k) = \underline{s}\}|I_k)$. Since the Markovian evolution of the different channels in our problem are independent of each other, this conditional probability distribution is equivalently represented by the set of *beliefs* about the occupancy states of each channel, i.e., the probability of occupancy of each channel in slot k, conditioned on all the past observations on channel a and times when channel a was sensed. We use $p_a(k)$ to represent the belief about channel a at end of slot k, i.e., $p_a(k)$ is the probability that the state $S_a(k)$ of channel a in slot k is 1 conditioned on all observations and decisions up to time slot k given by:

$$p_a(k) = \mathsf{P}(\{S_a(k) = 1\}|\underline{X}_a^k, K_a^k) = \mathsf{P}(\{S_a(k) = 1\}|I_k)$$

We use $\underline{p}(k)$ to denote the $L \times 1$ vector representing the beliefs about the L channels conditioned on I_k. The initial values of the belief parameters for all channels are set using the stationary distribution of the Markov chain. Now, using P to represent the transition probability matrix for the state transitions of each channel, we define:

$$q_a(k) = P(1,1)p_a(k-1) + P(0,1)(1 - p_a(k-1)) \quad (7)$$

This $q_a(k)$ is the occupancy probability of channel a in slot k, conditioned on I_{k-1}. Using Bayes' rule, the belief values are updated as follows after the observation in time slot k:

$$p_a(k) = \frac{q_a(k) f_1(\underline{X}_a(k))}{q_a(k) f_1(\underline{X}_a(k)) + (1 - q_a(k)) f_0(\underline{X}_a(k))} \quad (8)$$

when channel a was selected in slot k (i.e., $u_k = a$), and $p_a(k) = q_a(k)$ otherwise. Thus from (8) we see that updates for the sufficient statistic can be performed using only the joint log-likelihood ratio of the observations, $\mathcal{L}(\underline{X}_a(k))$, instead of the entire vector of observations. Furthermore, from (2) we also see that the access decisions also depend only on the log-likelihood ratios. Hence we can conclude that this problem with vector observations is equivalent to one with scalar observations where the scalars represent the joint likelihood ratio of the observations of all the cooperating secondary users. Therefore, in the rest of this paper, we use a scalar observation model with the observation made on channel a in slot k represented by $Y_a(k)$.

Hence the new access decisions are based on comparing the log-likelihood ratio of $Y_a(k)$ represented by $\mathcal{L}'(Y_a(k))$ to a threshold Δ' that is chosen to satisfy:

$$\mathsf{P}(\{\mathcal{L}'(Y_a(k)) < \Delta'\}|\{S_a(k) = 1\}) = \zeta \quad (9)$$

and the access decisions are given by:

$$\delta_a(k) = \mathcal{I}_{\{\mathcal{L}'(Y_a(k)) < \Delta'\}} \mathcal{I}_{\{u_k = a\}} \quad (10)$$

Similarly the belief updates are performed as in (8) with the evaluations of density functions of $\underline{X}_a(k)$ replaced with the evaluations of the density functions f_0' and f_1' of $Y_a(k)$:

$$p_a(k) = \frac{q_a(k) f_1'(Y_a(k))}{q_a(k) f_1'(Y_a(k)) + (1 - q_a(k)) f_0'(Y_a(k))} \quad (11)$$

when channel a is accessed in slot k (i.e., $u_k = a$), and $p_a(k) = q_a(k)$ otherwise. We use $G(\underline{p}(k-1), u_k, Y_{u_k}(k))$ to denote the function that returns the value of $\underline{p}(k)$ given that channel u_k was sensed in slot k. This function can be calculated using the relations (7) and (11). There is some randomness in function $G(.)$ arising from the random observation $Y_{u_k}(k)$. The reward obtained in slot k can be expressed as:

$$r_k \quad = \quad B(1-\epsilon)\mathcal{I}_{\{S_{u_k}(k) = 0\}} \quad (12)$$

where ϵ is given by

$$\epsilon = \mathsf{P}(\{\mathcal{L}'(Y_a(k)) > \Delta'\}|\{S_a(k) = 0\}) \quad (13)$$

The maximum value of $\sum_{k=0}^{\infty} \alpha^k \mathsf{E}[r_k]$, over all possible channel selection policies, is a function of $\underline{p}$, the initial value of the belief vector, i.e., the prior probability of channel occupancies in slot -1. We denote this function, called the optimal reward-to-go function, by $J(\underline{p})$. From the structure of the dynamic program, it can be seen that the observation noises v_k are i.i.d., the Markov chain that controls the state transitions is stationary, and the reward obtained in each slot is non-negative and bounded above by B. This observation suggests that the optimal solution to this dynamic program can be obtained by solving the following Bellman equation [7] for the optimal reward-to-go function:

$$J(\underline{p}) = \max_{u \in \mathcal{A}} [B(1-\epsilon)(1 - q_u) + \alpha \mathsf{E}(J(G(\underline{p}, u, Y_u)))] \quad (14)$$

where $\mathcal{A} = \{1, 2, \ldots, L\}$ is the set of channels, $\underline{p}$ represents the initial value of the belief vector and $\underline{q}$ is calculated from $\underline{p}$ as in (7) by:

$$q_a = P(1,1)p_a + P(0,1)(1 - p_a), \qquad a \in \mathcal{A} \quad (15)$$

Since it is not easy to find the optimal solution to this Bellman equation, we adopt a suboptimal strategy to obtain a channel selection policy that performs well.

In the rest of the paper we make the following assumptions on the probability transition matrix P, which gives the state transition probabilities for the Markov chain representing the state of each channel:

$$\begin{aligned} &\text{Assumption 1} \;:\; 0 < P(j,j) < 1, \qquad j \in \{0,1\} \quad (16) \\ &\text{Assumption 2} \;:\; P(0,0) > P(1,0) \quad (17) \end{aligned}$$

where $P(i,j)$ represents the probability that a channel that is in state i in slot k switches to state j in slot $k+1$. The first assumption ensures that the resultant Markov chain is irreducible and positive recurrent, while the second assumption ensures that it is more likely for a channel that is free in the current slot to remain free in the next slot than for a channel that is occupied in the current slot to switch states and become free in the next slot.

A. Greedy policy

A straightforward solution to the channel selection problem is to employ the greedy policy, i.e., the policy of maximizing the expected instantaneous reward. The expected instantaneous reward obtained by accessing some channel a in a given slot k is given by $B(1-\epsilon)(1-q_a(k))$ where ϵ is given by (13). Hence the greedy policy is to choose the channel a that maximizes $1-q_a(k)$.

$$u_k^{\text{gr}} = \operatorname*{argmax}_{u\in\mathcal{A}} \{1-q_u(k)\} \tag{18}$$

In other words, in each slot $k+1$, the greedy policy chooses the channel that is most likely to be free, conditioned on I_k.

The greedy policy for this problem is in fact equivalent to the Q_{MDP} policy, which is a standard sub-optimal solution to the POMDP problem (see, e.g., [8]). In [6] we also show that the results of [2] and [3] can be used to argue that the greedy policy is optimal in high SNR.

B. An upper bound

An upper bound on the optimal reward for a POMDP can be obtained by making the Q_{MDP} assumption [8] wherein we assume that in all future slots, the state of all channels become known exactly after making the observation in that slot. The optimal reward under this assumption is a function J^Q of the initial belief vector $\underline{p}(-1)$, i.e., the vector of prior probabilities of occupancy of the channels in slot -1. A typical choice of this initial value is given by the stationary distribution of Markov chains. Under this initialization, an upper bound for the optimal reward of the POMDP is given by $J^U = J^Q(p^*\underline{1})$ where p^* represents the stationary distribution of the transition probability matrix P and $\underline{1}$ represents an $L\times 1$ vector of all 1's.

The first step involved in evaluating J^U is to determine $\tilde{J}$, the optimal reward function under the assumption that all the channel states become known exactly after making the observation in each slot including the current slot. We have to evaluate $\tilde{J}(\underline{x})$ for all binary strings $\underline{x}$ of length L that represent the 2^L possible values of the vector representing the states of all channels in slot -1. The Q_{MDP} assumption implies that the functions J^Q and $\tilde{J}$ satisfy the equation:

$$\begin{aligned} J^Q(\underline{z}) \;=\; & \max_{u\in\mathcal{A}} \Bigg\{ \Bigg[\alpha \sum_{\underline{x}\in\{0,1\}^L} \mathsf{P}(\{\underline{S}(0)=\underline{x}\})\tilde{J}(\underline{x}) \\ & +B(1-\epsilon)(1-q_u)\Bigg]\Bigg\} \text{ s.t. } \underline{p}(-1)=\underline{z} \end{aligned}$$

Hence the upper bound $J^U = J^Q(p^*\underline{1})$ can be easily evaluated using the transition probability matrix P once the function $\tilde{J}$ is determined.

Now we describe how one can solve for the function $\tilde{J}$ under the assumption that the states of all the channels become known at the time of observation. It is easy to see that the optimal access decision in each slot k is to access some channel that is free in that slot, if any. Moreover, the optimal channel to be sensed in slot k is chosen so as to maximize the expected instantaneous reward, which is achieved by sensing the channel that is most likely to be free in the current slot. Now under the added assumption stated in (17) earlier, if some channel was free in the previous time slot, that channel would be the one that is most likely to be free in the current time slot. Hence the optimal policy would be to select some channel that was free in the previous time slot, if there is any. If not, the optimal policy would be to select any of the L channels, since all of them are equally likely to be free in the current slot. Hence the function $\tilde{J}$ can be derived in a straightforward manner as illustrated below. The argument of $\tilde{J}$ is the state of the system in the slot preceding the initial slot, i.e., $\underline{S}(-1)$.

$$\begin{aligned} \tilde{J}(\underline{x}) &= \max_{u\in\mathcal{A}} \mathsf{E}\big[\big[B(1-\epsilon)\mathcal{I}_{\{S_u(0)=0\}} + \\ &\qquad\qquad \alpha\tilde{J}(\underline{S}(0))\big]\big|\{\underline{S}(-1)=\underline{x}\}\big] \\ &= \begin{cases} B(1-\epsilon)P(0,0)+\alpha V(\underline{x}) \text{ if } \underline{x}\neq\underline{1} \\ B(1-\epsilon)P(1,0)+\alpha V(\underline{x}) \text{ if } \underline{x}=\underline{1} \end{cases} \end{aligned}$$

where $V(\underline{x}) = \mathsf{E}[\tilde{J}(\underline{S}(0))|\{\underline{S}(-1)=\underline{x}\}]$ and $\underline{1}$ is an $L\times 1$ string of all 1's. This means that we can write

$$\begin{aligned} \tilde{J}(\underline{x}) &= B(1-\epsilon)\Bigg[P(0,0)\sum_{k=0}^{\infty}\alpha^k - \\ &\qquad (P(0,0)-P(1,0))w(\underline{x})\Bigg] \end{aligned} \tag{19}$$

where

$$w(\underline{x}) := \mathsf{E}\left[\sum_{M\geq -1:\underline{S}(M)=\underline{1}} \alpha^{M+1} \middle| \{\underline{S}(-1)=\underline{x}\}\right]$$

is a scalar function of the vector state $\underline{x}$. Here the expectation is over the random slots when the system is in state $\underline{1}$. Now by stationarity we have:

$$w(\underline{x}) = \mathsf{E}\left[\sum_{M\geq 0:\underline{S}(M)=\underline{1}} \alpha^{M} \middle| \{\underline{S}(0)=\underline{x}\}\right] \tag{20}$$

We use $\mathbb{P}$ to denote the matrix of size $2^L\times 2^L$ representing the transition probability matrix of the joint Markov process that describes the transitions of the vector of channel states $\underline{S}(k)$. The $(i,j)^{\text{th}}$ element of $\mathbb{P}$ represents the probability that the state of the system switches to $\underline{y}$ in slot $k+1$ given that the state of the system is $\underline{x}$ in slot k, where $\underline{x}$ is the L-bit binary representation of $i-1$ and $\underline{y}$ is the L-bit binary representation of $j-1$. Using a slight abuse of notation we represent the $(i,j)^{\text{th}}$ element of $\mathbb{P}$ as $\mathbb{P}(\underline{x},\underline{y})$ itself. Now equation (20) can be solved to obtain:

$$w(\underline{x}) = \sum_{\underline{y}} \alpha\mathbb{P}(\underline{x},\underline{y})w(\underline{y}) + \mathcal{I}_{\{\underline{x}=\underline{1}\}} \tag{21}$$

This fixed point equation which can be solved to obtain:

$$\underline{w} = (\mathbb{I}-\alpha\mathbb{P})^{-1} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}_{2^L\times 1} \tag{22}$$

where $\underline{w}$ is a $2^L \times 1$ vector whose elements are the values of the function $w(\underline{x})$ evaluated at the 2^L different possible values of the vector state $\underline{x}$ of the system in time slot -1. Again, the i^{th} element of vector $\underline{w}$ is $w(\underline{x})$ where $\underline{x}$ is the L-bit binary representation of $i-1$. Thus $\tilde{J}$ can now be evaluated by using relation (19) and the expected reward for this problem under the Q_{MDP} assumption can be calculated by evaluating $J^U = J^Q(p^*\underline{1})$ via equation (19). This optimal value yields an analytical upper bound on the optimal reward of the original problem (14).

IV. The case of unknown distributions

In practice, the secondary users are typically unaware of the primary's signalling scheme and channel conditions [9] and have to rely on some form of non-coherent detection such as energy detection while sensing the primary signals. Furthermore, they are also unaware of their locations relative to the primary and hence the shadowing and path loss from the primary to the secondary. Hence the secondary users are often unaware of the exact distributions of the observations under the primary-present hypothesis, although they may know the distribution of the observations under the primary-absent hypothesis. We model this scenario by using a parametric description of the pdf's of the received signal under the primary-present hypothesis as follows:

$$\begin{aligned} S_a(k) = 0 \quad &: \quad Y_a(k) \sim f_{\theta_0} \\ S_a(k) = 1 \quad &: \quad Y_a(k) \sim f_{\theta_a} \\ \text{where } \theta_a \in \Theta, \forall a \quad &\in \quad \mathcal{A} \end{aligned} \tag{23}$$

where the parameters $\{\theta_a\}$ are unknown for all channels a, and $\theta_0 \in \mathbb{R}$ and $\Theta \subset \mathbb{R}$ are known. We use $\mathcal{L}_\theta(.)$ to denote the log-likelihood function under f_θ defined by:

$$\mathcal{L}_\theta(x) := \log\left(\frac{f_\theta(x)}{f_{\theta_0}(x)}\right), \qquad x \in \mathbb{R}, \theta \in \Theta \tag{24}$$

In this section, we study two different approaches for dealing with such a scenario, restricting ourselves to greedy policies for channel selection. For ease of illustration, in this section we consider a secondary system comprised of a single user, although the same ideas can be applied also for a system with multiple cooperating users.

A. Worst-case design for non-random θ_a

When the parameters $\{\theta_a\}$ are non-random and unknown, we will have to meet the constraint on the interference probability for all possible realizations of θ_a. We also need to find some means to perform the belief updates in (11). We show in [6] that for parametric families that satisfy a particular ordering condition, it is always possible to find some $\theta^* \in \Theta$ such that designing the policy assuming $\theta_a = \theta^*$ is optimal for this problem.

B. Modeling θ_a as random

Our simulations in [6] show that the worst-case design leads to a severe decline in performance relative to the scenario where the distribution parameters in (23) are known accurately. In practice it may be possible to learn the value of these parameters online. In order to learn the parameters $\{\theta_a\}$ we model the parameters $\{\theta_a\}$ as random variables, which are i.i.d. across the channels and independent of the channel states as well as the observation noise. We also assume that the cardinality of set Θ is finite[1] and let $|\Theta| = N$. Let $\{\mu_i\}_1^N$ denote the elements of set Θ. The prior distribution of the parameters $\{\theta_a\}$ is known to the secondary users. The beliefs of the different channels no longer forms a sufficient statistic for this problem. Instead, we keep track of an $L \times N \times 2$ array $Q(k)$ containing the following a posteriori probabilities which we refer to as *joint beliefs*:

$$Q_{a,i,j}(k) = \mathsf{P}(\{(\theta_a, S_a(k)) = (\mu_i, j)\}|I_k) \tag{25}$$

These joint beliefs are initialized using the product distribution of the stationary distribution of the Markov chain and the prior distribution on the parameters $\{\theta_a\}$. Now define:

$$H_{a,i,j}(k) = \sum_{\ell \in \{0,1\}} P(\ell, j) Q_{a,i,\ell}(k-1)$$

Again, the values of the array $H(k)$ represent the a posteriori probability distributions of the parameters $\{\theta_a\}$ and the channel states in slot k conditioned on I_{k-1}. The new update equations are:

$$Q_{a,i,j}(k) = \begin{cases} \kappa H_{a,i,0}(k) f_{\theta_0}(Y_a(k)) & \text{if } j = 0 \\ \kappa H_{a,i,1}(k) f_{\mu_i}(Y_a(k)) & \text{if } j = 1 \end{cases}$$

when channel a was accessed in slot k, and $Q_{a,i,j}(k) = H_{a,i,j}(k)$ otherwise. Here κ is just a normalizing factor.

In [6] we show that, for each channel a, the a posteriori probability mass function of parameter θ_a conditioned on the information up to slot k, converges to a delta-function at the true value of parameter θ_a as $k \to \infty$, provided we sense channel a frequently enough. This observation suggests that we could use this knowledge learned about parameters $\{\theta_a\}$ in order to be more liberal in our access policy than in Section IV-A. With this in mind, we propose the following algorithm for making access decisions in each slot.

Assume channel a was sensed in slot k. We first arrange the elements of set Θ in increasing order of the a posteriori probabilities of parameter θ_a conditioned on I_{k-1}. We partition set Θ into an 'upper' partition, $\Theta_a(k)$, and a 'lower' partition, $\Theta_a(k)^c$, such that all elements in $\Theta_a(k)$ have higher a posteriori probability values than all elements not in $\Theta_a(k)$. The partitioning is done such that the number of elements in $\Theta_a(k)^c$ is maximized subject to the constraint that the a posteriori probabilities of the elements in $\Theta_a(k)^c$ add up to a value lower than ζ. The elements of $\Theta_a(k)^c$ can be ignored while designing the access policy since the sum of their a posteriori probabilities is below the interference constraint. We then design the access policy such that we meet the interference constraint conditioned on parameter θ_a taking any value in $\Theta_a(k)$. The mathematical description of the algorithm is given in [6]. The access decision on channel a in slot k is given by:

$$\tilde{\delta}_a(k) = \mathcal{I}_{\{u_k = a\}} \prod_{\theta \in \Theta_a(k)} \mathcal{I}_{\{\mathcal{L}_\theta(Y_a(k)) < \tau_\theta\}} \tag{26}$$

[1] We do discuss the scenario when Θ is a compact set in [6].

where τ_θ satisfies:

$$\mathsf{P}(\{\mathcal{L}_\theta(Y_a(k)) < \tau_\theta\}|\{S_a(k) = 1, \theta_a = \theta\}) = \zeta$$

The access policy given above guarantees that

$$\mathsf{P}(\{\tilde{\delta}_a(k) = 1\}|\{S_a(k) = 1\}, I_{k-1}) < \zeta \qquad (27)$$

whence the same holds without conditioning on I_{k-1}. Hence, the interference constraint is met on an average, averaged over the posteriori distributions of θ_a. We show in [6] that the a posteriori probability mass function of parameter θ_a converges to a delta function at the true value of parameter θ_a almost surely. Hence the constraint is asymptotically met even conditioned on θ_a taking the correct value. This follows from the fact that, if μ_{i*} is the actual realization of the random variable θ_a, and $b_a^{i*}(k)$ converges to 1 almost surely, then, for sufficiently large k, (26) becomes: $\tilde{\delta}_a(k) = \mathcal{I}_{\{u_k=a\}}\mathcal{I}_{\{\mathcal{L}_{\mu_{i*}}(Y_a(k))<\tau_{\mu_{i*}}\}}$ with probability one and hence the claim is satisfied.

It is important to note that the access policy given in (26) need not be optimal for this problem. Unlike in Section II, here we allow the access decision in slot k to depend on the observations in all slots up to k via the joint beliefs. Hence, it is no longer obvious that the optimal test should be a threshold test on the likelihood ratio of the observations in the current slot even if parameter θ_a is known. However, this structure for the access policy is justified since it is simpler to implement in practice than some other policy that requires us to keep track of all the past observations. This scheme also shows substantial performance improvement over the worst-case approach in simulations, further justifying this structure for the access policy.

Under this scheme the new greedy policy for channel selection is to sense the channel which promises the highest expected instantaneous reward which is now given by:

$$\widetilde{u_k^{\mathsf{gr}}} = \underset{a\in\mathcal{A}}{\operatorname{argmax}} \left\{ \sum_{i=1}^{N} H_{a,i,0}(k)(1-\epsilon_a(k)) \right\} \qquad (28)$$

where

$$\epsilon_a(k) = \mathsf{P}\left(\bigcup_{\theta\in\Theta_a(k)} \{\mathcal{L}_\theta(Y_a(k)) > \tau_\theta\} \middle| \{S_a(k) = 0\} \right)$$

V. Results and discussion

The performance of the schemes proposed in this paper are shown in Table I for a scalar Gaussian observation model with a single secondary user. Detailed description of the simulation setup is given in [6]. The observations have unit variance under both states of the channel. When the channel is free, the mean is 0 and when it is occupied the mean is given by the θ_a parameters in (23). The transition probability matrix was chosen to be:

$$P = \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix}$$

with first and second rows corresponding to transitions from states 0 and 1 respectively. The set Θ is chosen with three elements such that the SNR values in dB lie in $\{1, 6, 10\}$. False alarm constraint $\zeta = 0.01$, discount factor $\alpha = 0.999$, and $L = 2$ with observations on both channels having equal means in the simulations.

TABLE I

SNR	UB	G_1	G_2	Worst case	Learn θ_a
1	93.87	85.5	76.7	85.5	85.8
6	304.28	291.6	249.1	89.5	243.4
10	656.07	647.3	535.5	91.2	593.8

Values under G_1 in Table I correspond to the case where the mean is known while those under UB give the analytical upper bound. Clearly, when parameters θ_a are known our greedy policy achieves performance close to the upper bound and hence is nearly optimal. For the scenario where θ_a are unknown, worst-case design can lead to a big drop in performance over G_1 when SNR is high. Much better performance is obtained by using our scheme that learns the θ_a. The caveat is that the learning procedure requires knowledge of a reliable model for the state transition process if we need to give probabilistic guarantees of the form (27) and to ensure convergence of the beliefs about the θ_a parameters. From Table I, we also see that for the problem with known mean, our greedy policy (G_1) that uses analog observations for learning channel occupancies gives significant performance gains over a scheme that uses only ACK signals proposed in [4] whose values are shown under G_2.

References

[1] 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 (JSAC): Special Issue on Adaptive, Spectrum Agile and Cognitive Wireless Networks*, vol. 25, no. 3, pp. 589–600, April 2007.

[2] Q. Zhao, B. Krishnamachari, and K. Liu, "On myopic sensing for multi-channel opportunistic access," *IEEE Transactions on Wireless Communications*, submitted for publication.

[3] T. Javidi, B. Krishnamachari, Q. Zhao, and K. Liu, "Optimality of myopic sensing in multi-channel opportunistic access," in *IEEE International Conference on Communications (ICC)*, Beijing, May 2008.

[4] 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*, vol. 54, no. 5, pp. 2053–2071, May 2008.

[5] H. V. Poor, *An Introduction to Signal Detection and Estimation*, 2nd ed. New York: Springer-Verlag, 1994.

[6] J. Unnikrishnan and V. Veeravalli, "Algorithms for dynamic spectrum access with learning for cognitive radio," *IEEE Transactions on Signal Processing*, submitted for publication.

[7] D. Bertsekas, *Dynamic Programming and Optimal Control*, 3rd ed. Athena Scientific, 2005, vol. 1.

[8] D. Aberdeen, "A (revised) survey of approximate methods for solving pomdps," National ICT Australia, Tech. Rep., December 2003. Available online at `http://users.rsise.anu.edu.au/∼daa/papers.html`.

[9] A. Sahai, N. Hoven, and R. Tandra, "Some fundamental limits on cognitive radio," in *Forty-Second Allerton Conference on Communication, Control, and Computing*, October 2004.