Soft-Decision Sequential Sensing for Optimization of Interweave Cognitive Radio Networks

Luis M. Lopez-Ramos, Antonio G. Marques, and Javier Ramos
King Juan Carlos University, Dep. of Signal Theory and Communications, Madrid, Spain

***Abstract*—Under probability-of-interference constraints, proper spectrum sensing is crucial in Cognitive Radios (CRs). However, the capability of a CR to sense the spectrum is limited, especially when multiple users try to access multiple channels. As a consequence, control and resource allocation schemes should optimize not only transmitting resources, but also sensing resources. In this paper, the cost of such sensing resources is incorporated into the optimization, with the aim of dynamically adapting the power (energy) devoted to sense each channel. More precisely, the tradeoff among: throughput, power devoted to sense, power devoted to transmit, and probability of interference is optimized. A soft-decision Bayesian sequential sensing scheme is used to exploit the time-correlation of the primary occupancy. The joint design leverages tools of dynamic programming to solve the sequential sensing problem and relies on reinforcement learning to develop a stochastic solution.**

***Index Terms*—Cognitive radios, Dynamic Programming, sequential sensing, resource allocation.**

I. Introduction

Cognitive radios (CRs) are viewed as the next-generation solution to alleviate the perceived spectrum scarcity. When CRs are deployed, the secondary users (SUs) have to sense their radio environment to optimize their communication performance while avoiding (limiting) the interference to the primary users (PUs). As a result, effective operation of CRs requires: i) sensing the spectrum and ii) dynamically adaptation of the available resources according to the sensed information [1]. To carry out the *sensing task* two important challenges are C1) the presence of noise in the measurements that render harmless SU transmissions impossible, and C2) the inability to sense the totality of the time-frequency lattice due to the scarcity of resources (power, time or device availability). A challenge that arises to carry out the *resource allocation* (RA) *task* is C3) the need of the RA algorithms to deal with channel model imperfections. An additional challenge is C4) the coupling between sensing and transmission resources.

To deal with C1, a level of tolerance to interference has to be established, and SUs need to keep track of the state information of the primary network (SIPN), e.g., PU presence. To overcome C2, advanced sensing schemes aim at optimally selecting the subset of sensed channels. Many works consider probabilistic SIPN to manage the information obtained from sensing, although fewer exploit the statistical model of the channel imperfections (especially for the time correlation) to mitigate them; see, e.g., [7]. Regarding C3, adaptive stochastic algorithms provide robustness to non-stationarities and lack of knowledge of the channel distributions [6]. Regarding C4, RA policies have been traditionally designed separately from sensing; however, a globally optimum design requires designing those tasks jointly. Joint design is necessary when the decision of allocating a specific resource to transmission or sensing is critical for the overall efficiency. Clearly, more accurate sensing enables more efficient RA, but at the expense of higher time and/or power consumption. The sensing-throughput tradeoff is investigated in [4] for a simple CR setup.

In this paper we jointly optimize transmit resources (scheduling coefficients and transmission power) and sensing resources (power devoted to sense each band). The main difficulty is that optimization of the sensing resources is a sequential-decision problem because sensing decisions at any time instant have an impact not only on the present RA, but also on future decisions. To efficiently exploit time correlation in the SIPN, dynamic programming (DP) and reinforcement learning (RL) [9] tools have to be used. We leverage results from our previous work in [3], which addressed the joint optimization of *hard-decision* sensing and RA for a simpler setup. Specifically, the schemes in [3] were designed to minimize the cost of sensing, to maximize the throughput of the SUs and to hold the probability of interference under a pre-specified limit. The sensing variable to be optimized was binary and the model for the observations was binary too. Motivated by the results in [4], the present article goes beyond binary decision and aims at adapting the amount of power devoted to sense each channel. More specifically, at each time instant the sensing power is optimized as a function of the SIPN and state information of the secondary network (SISN). To accomplish this, we will rely on expressions that relate the sensing performance with the sensing power, the SIPN and SISN. Such expressions clearly depend on the sensor performance and the statistical model for the state information. A recursive Bayesian estimator is used to estimate the probability (*belief*) of a channel being busy. Key to handle the DP is the computation of the so-called value function. Two methods (one off-line, and one online) are proposed to accomplish this computation.

Cooperative soft-decision sensing is investigated in [5], where analog measures are combined to obtain a fine estimation of the occupancy belief. This belief is used explicitly in a joint optimization of secondary network's parameters, obtaining significant performance improvement. The main difference of the present paper with respect to [4], [5] is that

This work was partially supported by the Spanish Ministry of Science (FPU Grant AP2010-1050) and the EU-FP7 (ICT-2011-9-TUCAN3G).

978-1-4673-5577-3/13/$31.00 ©2013 IEEE

the sensing scheme is designed sequentially, so that the time-correlation in the SIPN can be optimally exploited.

The rest of the manuscript is organized as follows. The system model is detailed in Section II and the joint optimization is formulated in Section III. Optimal transmit power and access allocation for *any* sensing scheme are presented in Section IV. Section V builds on those results to design the optimal sensing. Two methods to compute the value function required to solve the DP formulated in Section V are proposed in Section VI. Numerical results in Section VII close the paper.

II. System Model

A CR with several PUs and SUs is considered. The frequency band of interest is divided into K frequency-flat orthogonal subchannels (indexed by k), so that if a SU is transmitting, no other SU can be active in the same subchannel. Access is opportunistic; hence, during a time slot (indexed by n) each of the M secondary users (indexed by m) can access any number of these channels. For simplicity, we assume that the secondary network has a network controller (NC) which performs the sensing task. At every time slot the following tasks are run sequentially: T1) the NC acquires the SISN; T2) the NC relies on the output of T1 (and previous measurements) to allocate sensing resources, then the output of the sensing (if any) is used to update the SIPN; and T3) the outputs of T1 and T2 are used to find the optimal RA for instant n.

In this section, we describe 1) the model for the SISN and SIPN; 2) the variables to be designed; and 3) the constraints that such variables need to satisfy.

1) Model for the SISN and SIPN: Starting with the SISN, the power gain of the channel between the mth SU and the NC in the kth channel at time n is denoted as $h_k^m[n]$; similarly, the power gain of the channel between the PU transmitting in the kth channel, and the secondary NC (which performs the sensing) is denoted as g_k. These variables represent the noise-normalized square magnitude of the respective fading coefficient; channels are assumed to be ergodic and independent across bands and time, and their instantaneous gain is known[2]. Moving to the SIPN, let $a_k[n]$ denote a binary variable, which is one if primary link k is active at time n, and zero otherwise. Whenever convenient, the alternative notation $\mathcal{H}_0$ (for $a_k[n]=0$) and $\mathcal{H}_1$ (for $a_k[n]=1$) will be used. Process $a_k[n]$ is modeled as a two-state, time invariant Markov chain with $P_k^{xy} := \Pr(a_k[n]=x|a_k[n-1]=y)$. Non-Markovian models can also be accommodated, at the expense of a higher computational load to solve the DP and estimate the SIPN [7].

The NC senses the band to acquire information about $a_k[n]$. Let $z_k[n]$ denote the sensor output; the probability density function (PDF) of $z_k[n]$ conditioned on the idle and busy channel hypotheses ($\mathcal{H}_0$ and $\mathcal{H}_1$) is assumed to be known. Three sources of imperfections render deterministic knowledge of $a_k[n]$ impossible: I1) noise and fading in the received primary signal; I2) limited sensing power, and I3) outdated information (for the cases when $s_k[n]=0$). Moreover, I1 and I2 affect the PDF of $z_k[n]$. While noise and fading in I1 are purely random state variables, the sensing power is a design variable. The power devoted to sense channel k at time n is denoted as $s_k[n]$. The conditional PDF of $z_k[n]$ depends on $a_k[n]$, $g_k[n]$ and $s_k[n]$, and it will be denoted as $f_z(z|\mathcal{H}_0,g,s)$ and $f_z(z|\mathcal{H}_1,g,s)$ for idle and busy channel hypotheses, respectively. The expression for the conditional PDFs depends not only on the statistical model for the state information, but also on the sensor performance. This way, if convenient, in our formulation $s_k[n]$ could be used to represent sensing resources other than power (e.g., sensing time).

Since $a_k[n]$ is a partially observable state variable, the probability of $a_k[n]=1$ will be referred to as belief. Two different belief variables are considered: the *pre-decision* belief $b_k[n] := \Pr(a_k[n]|n-1)$; and the *post-decision* belief, denoted by $B_k[n] := \Pr(a_k[n]|n)$. Intuitively, $b_k[n]$ contains the information about $a_k[n]$ before the sensing decision has been made (i.e., at the beginning of task T2), while $B_k[n]$ contains the information about $a_k[n]$ once $s_k[n]$ and $z_k[n]$ are known (i.e., at the end of task T2). Considering these two variables will simplify the DP formulation in Section VI. Using the Markov transition matrix $\mathbf{P}_k$, the pre-decision belief at time slot n is computed as $b_k[n]=\mathcal{P}_k(B_k[n-1])$, where

$$\mathcal{P}_k(B_k[n-1]) := P_k^{10}(1-B_k[n-1]) + P_k^{11}B_k[n-1]. \quad (1)$$

If sensing is performed with power $s_k[n]$, post-decision belief is updated as $B_k[n]=\mathcal{B}\left(b_k[n],g_k[n],s_k[n],z_k[n]\right)$, where

$$\mathcal{B}(b,g,s,z) := \frac{bf_z(z|\mathcal{H}_1,g,s)}{bf_z(z|\mathcal{H}_1,g,s)+(1-b)f_z(z|\mathcal{H}_0,g,s)}. \quad (2)$$

While $b_k[n]$ only depends on the previous belief and the time-correlation model, $B_k[n]$ depends on $z_k[n]$, which itself depends on $s_k[n]$. This means that $B_k[n]$ is not purely random, since, e.g., higher values of $s_k[n]$ bring $B_k[n]$ closer to $a_k[n]$.

2) Design variables: Apart from $s_k[n]$, the other design variables are the access (scheduling) coefficients $w_k^m[n]$, and the power loadings $p_k^m[n]$. Specifically, $w_k^m[n]$ is one if the mth SU is scheduled to transmit into the kth band at time n and zero otherwise. Moreover, if $w_k^m[n]=1$, $p_k^m[n]$ denotes the instantaneous nominal power transmitted over the kth band by the mth SU. The fixed sensing and transmitting power costs are denoted as ξ_k and π^m. This means that a cost $\xi_k s_k[n]$ is paid every time the NC senses channel k, while a cost $\pi^m p_k^m[n]$ is paid when $w_k^m[n]=1$. Moreover, under bit error rate or capacity constraints, instantaneous rate and power variables are coupled. This rate-power coupling will be represented by the non-decreasing function $C_k^m(h_k^m[n],p_k^m[n])$ and β^m will denote the benefit (price) associated with rate.

Scheduling decisions may generate interference. Since an interweave setup is considered, interference occurs when $a_k[n]=1$ and a SU is transmitting into the kth band; i.e., if $\sum_m w_k^m[n]=1$. To protect the PUs, let θ_k denote the price of interfering PU k, so that if at instant n a SU is transmitting, a cost $\theta_k a_k[n]$ is paid. Due to the SIPN uncertainties, the expected interference cost is $\theta_k B_k[n]$.

[2]Different models for g_k can be used. For simplicity, we assume it constant and perfectly known. With minimal changes, our formulation can accommodate a stochastic $g_k[n]$, provided that its distribution is known.

978-1-4673-5577-3/13/$31.00 ©2013 IEEE

The prices $\{\xi_k, \pi^m, \beta^m, \theta_k\}$ can be pre-specified by the NC or computed based on quality-of-service (QoS) requirements. Indeed, they can correspond to Lagrange multipliers associated with QoS constraints. For example, the value of θ_k can be set to guarantee that the long-term probability of interfering the corresponding PU is below a given value [2].

3) Constraints: Sensing and transmit powers are non-negative, so that $s_k[n] \geq 0$ and $p_k^m[n] \geq 0$. Moreover, orthogonal access requires

$$w_k^m[n] \in \{0,1\} \text{ and } \textstyle\sum_m w_k^m[n] \leq 1. \tag{3}$$

III. Problem Formulation

The last step to formulate the optimization is to identify the metric to be maximized. To this end, the prices introduced in the previous section will play a critical role. Let first define the (short-term) utility at instant n for the (m,k) pair as

$$\begin{aligned}\varphi_k^m(h_k^m[n], B_k[n], p_k^m[n]) &:= \beta^m C_k^m(h_k^m[n], p_k^m[n]) \\ &- \pi^m p_k^m[n] - \theta_k B_k[n], \end{aligned} \tag{4}$$

which can be viewed as a per-user link quality indicator representing a tradeoff among the rate, transmit power, and short-term probability of interfering. Such a utility should only be accounted for when $w_k^m[n] = 1$, and has to be penalized with the instantaneous sensing cost and then averaged across time. Mathematically, this amounts to consider

$$\begin{aligned}\bar{U} &:= \textstyle\sum_{n=0}^{\infty} \gamma^n \mathbb{E}\Big[\sum_k \big(-\xi_k s_k[n] \\ &+ \textstyle\sum_m w_k^m[n] \varphi_k^m(h_k^m[n], B_k[n], p_k^m[n])\big)\Big]. \end{aligned} \tag{5}$$

where $\gamma \in (0,1)$ represents a discount factor that places more emphasis on early instants and can also be used to handle non-stationarities. The optimal joint design is then

$$\mathsf{P}^* := \max_{\{w_k^m[n], p_k^m[n], s_k[n]\}_{\forall n}} \bar{U} \tag{6a}$$

$$\text{s. to : (3), } p_k^m[n] \geq 0, \; s_k[n] \geq 0, \tag{6b}$$

which is a DP because the allocation at time instant n has to be designed taking into account the utility (cost) not only at n but also at $n' > n$. The challenge arises because $a_k[n]$ is only partially observable, so that setting the value of $s_k[n]$ changes the value of variables at instant n ($B_k[n]$) and at instants $n' > n$ too ($b_k[n']$). Differently, optimization of variables $w_k^m[n]$ and $p_k^m[n]$ can be separated across time. Since the only source for the sequential optimization is the partially observable $a_k[n]$, whose time dynamics are Markovian, our DP is in fact a partially observable Markov decision process (POMDP). This fact, together with several structural properties of (6) will be leveraged to decrease the computational complexity required to solve the formulated DP.

To solve (6), we start by finding the optimal RA scheme for *any* fixed sensing policy; mathematically, we derive the optimal $w_k^{m*}[n]$ and $p_k^{m*}[n]$ as a function of $s_k[n]$. This is useful because: i) DP tools are not required to carry out such an optimization, and ii) it corresponds to an actual task that has to be carried out by the CR system (cf. T3). Note that there is no loss of optimality because in Section V the expression for $w_k^{m*}[n]$ and $p_k^{m*}[n]$ is used as input for the optimal sensing.

IV. Optimal power allocation and access decision

To formulate the problem that gives rise to the optimal RA, two important facts have to be taken into account: a) the expression for the optimal RA needs to hold for any sensing policy; since $B_k[n]$ depends on $s_k[n]$, the former implies that the optimal RA needs to be a function of $B_k[n]$; and b) $s_k[n]$ is assumed to be given, so that terms and constraints depending on $s_k[n]$ can be dropped; hence, constraint $s_k[n] \geq 0$ and the sensing cost term in (5) are dropped. Under these considerations, the optimal RA is obtained as the solution of

$$\begin{aligned}\mathsf{P}_{RA}^* := \max_{\{w_k^m[n], p_k^m[n]\}_{\forall n}} &\textstyle\sum_{n=0}^{\infty} \gamma^n \mathbb{E}\big[\sum_{k,m} w_k^m[n] \\ &\times \varphi_k^m(h_k^m[n], B_k[n], p_k^m[n])\big] \end{aligned} \tag{7a}$$

$$\text{s. to : (3), } p_k^m[n] \geq 0. \tag{7b}$$

The approach[3] to find the optimal RA is to decompose the objective across time and channels. The optimal nominal power is $p_k^{m*}[n] = \arg\max_p \varphi_k^m(h_k^m[n], B_k[n], p)$. To find the optimal scheduling, we define first $L_k^{m*}[n] := \max_p \varphi_k^m(h_k^m[n], B_k[n], p)$, then it holds that $w_k^{m*}[n] = \mathbb{1}_{\{(L_k^{m*}[n] = \max_u L_k^{u*}[n]) \wedge (L_k^{m*}[n] > 0)\}}$. The previous dictates that the channel should be assigned to the SU with highest utility, except if $L_k^{m*}[n] \leq 0 \; \forall m$. In that case, channel k should not be used by any SUs.

Once the optimal RA has been presented, we introduce some notation that will be useful for the optimization of the sensing. Since the problem can be decomposed across time and channels, we will use $\mathcal{R}_k(\mathbf{h}_k[n], B_k[n])$ to denote the *optimum* value of the utility at time n provided that the optimal RA is implemented; i.e.,

$$\begin{aligned}\mathcal{R}_k(\mathbf{h}_k[n], B_k[n]) &:= \textstyle\sum_m w_k^{m*}[n] \varphi_k^m(h_k^m[n], B_k[n], p_k^{m*}[n]) \\ &= \Big[\max_{m,p} \big\{\varphi_k^m(h_k^m[n], B_k[n], p)\big\}\Big]_+, \end{aligned} \tag{8}$$

where $[\cdot]_+ := \max\{0, \cdot\}$ and $\mathbf{h}_k[n]$ is a vector containing $h_k^m[n] \, \forall m$. From the point of view of the SIPN, $\mathcal{R}_k(\mathbf{h}_k[n] B_k[n])$ can be viewed as an instantaneous *expected* reward indicator, where the expectation is carried over the uncertainty on $a_k[n]$. Clearly, (8) encapsulates (via $B_k[n]$) the way in which the sensing affects the optimal RA and P_{RA}^*.

V. Optimal Sensing

Since the expressions for the optimal RA hold for any sensing scheme $s_k[n]$, the aim here is to obtain $s_k^*[n]$. Key to accomplish this are two facts. The first one is that, instead of solving (6), it suffices to solve for all n

$$\mathsf{P}_{\text{DP}}^* := \max_{\{s_k[n] \geq 0\}} \sum_{n=0}^{\infty} \gamma^n \sum_k \mathbb{E}\big[\mathcal{R}_k(\mathbf{h}_k[n], B_k[n]) - \xi_k s_k[n]\big], \tag{9}$$

[3]Due to space limitations, we keep the derivations in this section at minimum, so that more details on how to solve the DP can be provided. Interested readers can check [2] for a very similar setup.

which has smaller dimensionality and can be optimized separately per channel. The second fact is that that the instantaneous reward $\mathcal{R}_k[n]$ depends on the sensing only through the belief. As a result, the value of $s_k[n]$ impacts the term in (9) corresponding to instant n via $B_k[n]$ and $\xi_k s_k[n]$, but the terms in (9) corresponding to instants $n' > n$ only via $b_k[n']$.

To be rigorous and solve this sequential optimization, we leverage techniques described in [9, Ch. 2]. Let us first identify the generic elements of a POMDP in (9). The state space is the Cartesian product of the domains of $\mathbf{h}_k[n]$ and $b_k[n]$ (replacing the partially observable state $a_k[n]$). The transition functions that describe the dynamics of the system over time are the functions (1) and (2). The action space is the domain of $s_k[n]$; there is no need to include $p_k^m[n]$ and $w_k^m[n]$ because i) their values do not impact the future states, and ii) their optimal expression, as a function of $s_k[n]$, was already found.

The POMDP in (9) can be split into k unconstrained POMDPs, each having $\mathcal{R}_k(B_k[n], \mathbf{h}[n]) - \xi_k s_k[n]$ as a reward function. Since the latter depends on $B_k[n]$, which in turn depends on $s_k[n]$, to design $s_k[n]$ we will need to compute the *a priori* expectation of the reward function conditioned on $s_k[n]$. Let us define for brevity

$$\begin{aligned}\bar{\mathcal{R}}_k(b_k[n], \mathbf{h}, s) &:= \mathbb{E}\left[\mathcal{R}_k(\mathbf{h}, B_k[n]) \big| b_k[n], \mathbf{h}, g_k, s\right] \\ &\quad - \xi_k s_k[n] = \int_0^1 f_B(B\big|b_k, g_k, s_k) \\ &\times \Big[\max_{m,p}\left\{\varphi_k^m(h_k^m[n], p)\right\} - \theta_k B\Big]_+ dB - \xi_k s_k[n].\end{aligned} \tag{10}$$

Suppose for now that the sensing is designed as $s_k^*[n] = \arg\max_{s\geq 0} \bar{\mathcal{R}}_k(b_k[n], \mathbf{h}_k[n], s)$. Despite being computationally simple, this approach (typically referred to as *myopic policy*) ignores the impact that current sensing decisions have in future time instants. To account for future time instants we leverage the concept of value function [9]. To be specific, for any sensing policy Π, there exists a value function (more specifically, a Q-function[4]) $Q_k^{\Pi}(b_k, \mathbf{h}_k, s_k)$ representing the discounted, expected reward resulting from following policy Π sequentially from the current state. The policy optimizing (9) is then denoted as Π^*, while the associated Q-function is denoted as $Q_k^{\Pi *}(b_k, \mathbf{h}_k, s_k)$ or simply $Q_k^*(b_k, \mathbf{h}_k, s_k)$. The existence of this stationary policy is guaranteed due to the discounted formulation in (9). Once $Q_k^*(b_k, \mathbf{h}_k, s_k)$ is available, Π^* is

$$s_k^*[n] = \arg\max_{s\geq 0}\left\{Q_k^*(b_k[n], \mathbf{h}_k[n], s)\right\}, \tag{11}$$

for every k, n. This maximization is computationally affordable because it is a line search over s. In other words, the joint design has been reduced to computing $Q_k^*(b_k, \mathbf{h}_k, s_k)$. This is accomplished in the ensuing section.

VI. Computing the optimal Q-function

Since the joint design is separable across channels and the method for computing Q_k^* is the same for every k, subindex k will be dropped for clarity. Moreover, since the optimal policy is stationary, time index n will be dropped too. Every variable refers to instant n except *prime* variables, which refer to $n+1$. The first step to compute $Q^*(b, \mathbf{h}, s)$ is to write the Bellman equation for the optimal policy [9, Eq. (2.18)]

$$Q^*(b, \mathbf{h}, s) = \mathbb{E}_{b', \mathbf{h}'}\Big[\bar{\mathcal{R}}(b, \mathbf{h}, s) + \gamma \max_{s'\geq 0} Q^*(b', \mathbf{h}', s')\big|b, \mathbf{h}, s\Big], \tag{12}$$

which can be simplified as (recall that b' and $\mathbf{h}'$ stand for the value of b and $\mathbf{h}$ in the immediate future time instant)

$$Q^*(b, \mathbf{h}, s) = \bar{\mathcal{R}}(b, \mathbf{h}, s) + \mathbb{E}_{b', \mathbf{h}'}\Big[\gamma \max_{s'\geq 0} Q^*(b', \mathbf{h}', s')\big|b, s\Big]. \tag{13}$$

Two methods to compute the optimal Q-function will be developed. The first one is an off-line method, based on the value iteration algorithm, that relies on (13) to iteratively compute Q^*. More specifically, we use the model-based Q-iteration algorithm [9, Sec. 2.3.1]. With ℓ denoting an iteration index, this amounts to computing for every $(b, \mathbf{h}, s)$

$$Q_{\ell+1}(b, \mathbf{h}, s) = \bar{\mathcal{R}}(b, \mathbf{h}, s) + \mathbb{E}_{b', \mathbf{h}'}\Big[\gamma \max_{s'\geq 0} Q_\ell(b', \mathbf{h}', s')\big|b, s\Big]. \tag{14}$$

To reduce the cost of computing $Q_\ell(b, \mathbf{h}, s)$, we will define a function with a smaller dimensionality, such that $Q_\ell(b, \mathbf{h}, s)$ can be obtained from it [3]. Specifically, let define $\mathcal{Q}_\ell(b, s) := \mathbb{E}_{b', \mathbf{h}'}[\max_{s'\geq 0} Q_\ell(b', \mathbf{h}', s')|b, s]$ and rewrite (13) as

$$Q_\ell(b, \mathbf{h}, s) = \bar{\mathcal{R}}(b, \mathbf{h}, s) + \gamma \mathcal{Q}_\ell(b, s). \tag{15}$$

Substituting (15) into both sides of (14) and simplifying yields

$$\mathcal{Q}_{\ell+1}(b, s) = \mathbb{E}_{b', \mathbf{h}'}\Big[\max_{s'}\left\{\bar{\mathcal{R}}(b', \mathbf{h}', s') + \gamma \mathcal{Q}_\ell(b', s')\right\}\big|b, s\Big] \tag{16}$$

where $\mathcal{Q}_\ell$ now depends only on two *scalar* variables. Provided that the distribution of B conditioned on (b, s) is known, and that $b_k[n+1]$ depends deterministically on $B_k[n]$ [cf. (2)], the iterate in (16) can be rewritten as

$$\begin{aligned}\mathcal{Q}_{\ell+1}(b, s) &= \int_{B=0}^1 \int_{\forall \mathbf{h}'} f_B(B\big|b, g, s) f_{\mathbf{h}}(\mathbf{h}') \\ &\times \max_{s'}\left\{\bar{\mathcal{R}}(\mathcal{P}(B), \mathbf{h}', s') + \gamma \mathcal{Q}_\ell(\mathcal{P}(B), s')\right\} d\mathbf{h}' dB.\end{aligned} \tag{17}$$

The Q-iteration algorithm starts from an arbitrary Q-function Q_0 and at each iteration ℓ updates the Q-function indirectly by using (17) to update $\mathcal{Q}_\ell$. For simplicity we choose $\mathcal{Q}_0 = 0$ which corresponds to $Q_0(b, \mathbf{h}, s) = \bar{\mathcal{R}}(b, \mathbf{h}, s)$. Convergence of $\mathcal{Q}_\ell$ to $\mathcal{Q}^*$when $\ell \to \infty$ can be shown based on the fact that the mapping defined in (14) is a contraction with factor $\gamma_Q < 1$ in the infinity norm [9, Ch. 2]. Summarizing, an iterative off-line method has been proposed to compute the Q-function. Computational cost has been lowered by splitting Q into two terms [cf. (15)]. The first one can be computed directly, while the second one still has to be computed iteratively, but has much smaller dimensionality. To implement this method, the following information was assumed to be known: (i) the Markov matrix for the primary occupancy $\mathbf{P}_k$; (ii) the distribution of z conditioned on the SIPN;

[4]To express the cost-to-go or value function, we prefer the Q-function form (value function corresponding to a state-action pair) over the V-function form (value function corresponding to a state [9, Sec. 2.2.1]) because it facilitates the mathematical analysis and the design of an online algorithm.

The second method to compute the Q-function will further reduce the computational complexity, and bypass the need to know (iii). Stochastic approximation is leveraged to design an online algorithm. With $\hat{\mathcal{Q}}_n(b,s)$ representing the online approximation of $\mathcal{Q}^*(b,s)$, the stochastic update is

$$\hat{\mathcal{Q}}_n(b,s) = (1-\alpha)\hat{\mathcal{Q}}_{n-1}(b,s) + \alpha \int_0^1 f_B(B|b,g,s) \times \max_{s'\geq 0} \left\{ \bar{\mathcal{R}}(\mathcal{P}(B), \mathbf{h}[n], s') + \gamma \hat{\mathcal{Q}}_{n-1}(\mathcal{P}(B), s') \right\} dB, \quad (18)$$

where $\alpha \in (0,1]$ is a learning rate. The proposed rule is a variant of the Q-learning algorithm. More specifically, it is a model-free value iteration [9, Sec. 2.3.2]. The proposed update is model-free for $\mathbf{h}_k$, but it is still model-based for b_k. The main advantages of this mixed algorithm are that it does not need to know the distribution of $\mathbf{h}_k[n]$, makes the system robust to channel gain non-stationarities, and avoids the need for exploration that affects some Q-learning algorithms.

The Q-iteration (off-line) and Q-learning (online) algorithms just presented, will be evaluated using numerical simulations in the next section.

VII. Numerical results

In this section the sensor is particularized as the power detector used in [4]. The test statistic is $z_k[n] = \sum_{t=1}^{s_k[n]} |y_k(t)|^2$ where $s_k[n]$ is the number of received samples; hence, $z_k[n]$ follows a χ^2 distribution with $2s_k[n]$ degrees of freedom. Under $\mathcal{H}_1$ the variable is scaled by a factor of $(g_k + 1)$. The conditional distribution of $z_k[n]$ can be written as $f_z(z|\mathcal{H}_0, g_k, s_k) = f_{\chi^2,2s_k}(z)$, and $f_z(z|\mathcal{H}_1, g_k, s_k) = (1/(1+g_k)) f_{\chi^2,2s_k}(z/(1+g_k))$.

The operating conditions of the system under test are as follows: $M = 4$ users; $K = 4$ channels; the Markov matrices that drive primary users' behavior are $\mathbf{P}_1 = [0.96, 0.06; 0.04, 0.94]$, $\mathbf{P}_2 = [0.97, 0.05; 0.03, 0.95]$, $\mathbf{P}_3 = [0.97, 0.05; 0.03, 0.95]$, and $\mathbf{P}_4 = [0.8, 0.15; 0.2, 0.85]$. The secondary user-secondary NC gains are Rayleigh distributed with $\mathbb{E}[h_1^m, h_2^m, h_3^m, h_4^m] = [7, 10, 10, 7] \ \forall m \in [1,4]$, and the primary user-secondary NC gains are $[g_1, g_2, g_3, g_4] = [1, 2, 2, 0.5]$. Transmitting power, sensing power and interference costs are set to: $[\pi^1, \pi^2, \pi^3, \pi^4] = [0.1, 0.4, 1.6, 5]$; $[\xi_1, \xi_2, \xi_3, \xi_4] = [0.25, 1, 0.1, 1]$, and $[\theta_1, \theta_2, \theta_3, \theta_4] = [6, 8, 8, 10]$. These values can be pre-specified constants or correspond to Lagrange multipliers associated with QoS constraints, see e.g. [3]. Here the values have been chosen in a way such that they give rise to informative results.

Two experiments are performed. The first one evaluates the utility gain of the optimal off-line scheme (17) with respect to the myopic policy. Since the optimization is separable across channels, to gain more intuition, the results for each individual channel are also provided in Table I.

The second experiment analyzes the performance of the online stochastic policy (18) for a non-stationary scenario. In particular, we consider a single-channel system and set the simulation time to 50000 slots. During the first 25000 slots, the channel conditions are those of $k = 1$ for the previous experiment. At $n = 25000$, the channel conditions switch to those of $k = 2$. Fig. 1 depicts the achieved utility, averaged by a rectangular window of length 9000.

TABLE I
Average achieved utility

Policy	k=1	k=2	k=3	k=4	Overall
Myopic	0.6718	0.4900	1.5234	1.7398	4.4250
Optimal	0.7972	0.9239	1.6237	1.7411	5.0859
Utility gain	15.2 %	88.5 %	6.7 %	0.02 %	14.94 %

The stochastic policy performs better than the myopic one and close to the optimal off-line policy corresponding to each of the two cases considered. Utility gaps and speed of converge depend heavily on the simulated scenario and, especially, on the correlation of the primary user behavior (stronger correlation brings slower convergence).

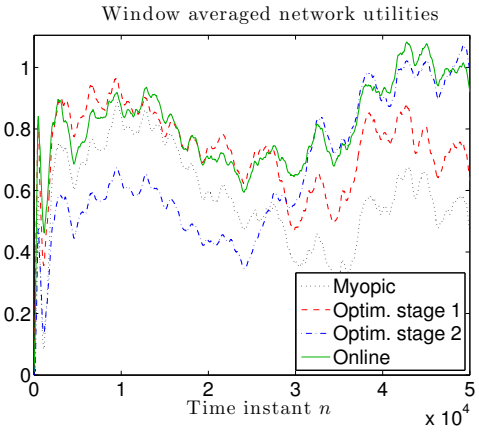

Fig. 1. Performance comparison of the optimal and stochastic iterates.

References

[1] S. Haykin, "Cognitive radio: brain-empowered wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 2, pp. 201–220, Feb. 2005.

[2] A. G. Marques, L. M. Lopez-Ramos, G. B. Giannakis, and J. Ramos, "Resource allocation for interweave and underlay CRs under probability-of-interference constraints," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 10, pp. 1922–1933, Nov. 2012.

[3] L. M. Lopez-Ramos, A. G. Marques, and J. Ramos, "Jointly optimal sensing and resource allocation for multiuser overlay cognitive radios," *arXiv preprint arXiv:1211.0954* (2012).

[4] Y.-C. Liang, Y. Zeng, E.C.Y. Peh, and A.T. Hoang, "Sensing-throughput tradeoff for cognitive radio networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 4, pp. 1326–1337, Apr. 2008

[5] H. Mu and J. K. Tugnait, "Joint soft-decision cooperative spectrum sensing and power control in multiband cognitive radios," *IEEE Trans. Signal Process.* , vol. 60, no. 10, pp. 5334–5346, Oct. 2012.

[6] A. G. Marques, G. B. Giannakis, L. M. Lopez-Ramos, and J. Ramos, "Stochastic resource allocation for cognitive radio networks based on imperfect state information," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Process.*, Prague, Czech Rep., May. 22–27, 2011.

[7] Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: A POMDP framework," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 3, pp. 589–600, Apr. 2007.

[8] A. Goldsmith, *Wireless Communications*, Cambridge Univ. Press, 2005.

[9] L. Busoniu, R. Babuska, B. De Schutter, and D. Ernst, *Reinforcement Learning and Dynamic Programming Using Function Approximators*, CRC Press, 2010.

[10] S.-J. Kim and G. Giannakis, "Sequential and cooperative sensing for multi-channel cognitive radios," *IEEE Trans. Signal Process.*, vol. 58, no. 8, pp. 4239–4253, Aug. 2010.