Spectrum Sensing in Cognitive Radio Networks via Approximate POMDP ~~methods~~

Bharath Keshavamurthy, Nicolò Michelusi

[NM: abstract is way too long.. limit to 200 words]
***Abstract*—Cognitive radio technologies will be critical to the wireless communication infrastructure due to their potential to alleviate the spectrum crunch, ~~in the near future due to the increasingly incredible number of applications being added to the computer networking ecosystem,~~ both in the commercial and the military spheres, ~~resulting in increased pressure on the available spectrum, which is a limited physical resource~~. In this paper, we propose a novel spectrum sensing and ~~channel~~ access strategy ~~in networks with multiple licensed users~~ based on partially observable Markov decision processes (POMDPs) wherein a cognitive radio node learns the correlation model defining the occupancy behavior of the incumbents and devises an optimal strategy to perform spectrum sensing and access that exploits the learned correlation model. To alleviate the complexity of the POMDP optimization, ~~Since the computational complexity associated with solving for the optimal spectrum sensing and channel access strategy scales exponentially~~ [NM: it is actually doubly-exponential for POMDPs..] ~~with the number of spectrum bands under consideration,~~ we employ ~~propose a system employing approximate POMDP value iteration methods, namely,~~ the PERSEUS algorithm, an approximate value iteration method. ~~Furthermore, through system simulations,~~ We compare numerically the performance of the proposed algorithms with state-of-the-art ~~we compare the performance of standard~~ MAP-based state estimators and correlation-coefficient based clustering algorithms ~~in the state-of-the-art against our proposed system employing a customized PERSEUS algorithm, with respect to the secondary network throughput and the number of collisions with the incumbent transmissions.~~ and demonstrate that....[NM: fix] These simulations show that the proposed system, in terms of the episodic utilities[NM: what do you mean by episodic utility?], outperforms the existing correlation-coefficient based state-of-the-art by an average of 50% and a Neyman-Pearson Detector that assumes independence among channels by an average of 33%. Finally, the proposed system, on average, matches the episodic utilities provided by standard MAP-based state estimators which possess prior knowledge about the transition model of the underlying MDP.**

***Index Terms*—Hidden Markov Models, Cognitive radio, spectrum sensing, POMDP**

I. Introduction

With the advent of fifth-generation wireless communication networks, the problem of spectrum scarcity has been exacerbated [?]. For some time now, cognitive radio (CR) technologies have been in the spotlight as a potential solution to this problem in commercial and military applications [?]. Cognitive radio networks facilitate efficient spectrum utilization by intelligently accessing "white spaces" left unused by the sparse and infrequent transmissions of the licensed users, while satisfying interference constraints with respect to the incumbents in the network [?]. A crucial aspect underlying the design of cognitive radio networks is the channel access protocol in the MAC layer of the stack. In this regard, the current state-of-the-art involves channel access strategies dictated by multi-armed bandits [?], reinforcement learning agents [?], and other custom heuristics [?], [?], [?]. However, almost all these works, such as [?], [?], [?], [?], assume independence among channels in the discretized spectrum which is imprudent because licensed users exhibit correlation across both frequency and time in their channel occupancy behavior: the primary users frequently occupy a set of adjacent channels (frequency correlation), repeating similar motifs in behavior over an extended period of time (temporal correlation) [?], [?], [?]. This pattern in occupancy behavior of the incumbents imputes very high levels of correlation among channels which need to be leveraged for more accurate predictions of spectrum holes. In this paper, we propose a parameter estimation algorithm to learn the aforementioned correlation model, along with a state estimation algorithm to infer channel occupancy from noisy and incomplete information, and an approach to solve for the optimal channel sensing and access policy to be followed by the cognitive radio node, that exploits the learned correlation structure.

The works [?], [?] develop spectrum sensing and access algorithms under the assumption that the occupancy behavior of ~~the~~ incumbents ~~in the radio environment~~ is independent across both time and ~~across~~ frequencies. In our work, we exploit both frequency and temporal correlations. In [?], a compressed spectrum sensing scheme is devised that exploits sparse temporal dynamics in the occupancy of licensed users, and in [?], an efficient spectrum sensing strategy is proposed for dense multi-cell cognitive networks, that also exploits the spatial structure of interference; however, both works assume independence across frequencies. Spectrum sensing and access strategies in a distributed multiple CR setting have been considered in [?] and solved using SARSA with linear value function approximation. However, frequency correlation is precluded, and errors in state estimation are neglected in the decision process. Unlike [?], we consider a model with correlation across frequencies, and we account for uncertainty in the occupancy state via a partially observable Markov decision process (POMDP) formulation.

Although the spectrum sensing algorithms detailed in [?] consider the correlation in incumbent occupancy behavior across frequencies, the authors assume a perfect, noise-free observation model, unlike which we construct a more reliable and realistic AWGN model; and like other works modeling correlation across both time and frequencies, such as [?], [?], the algorithms operate based on a data-driven strategy wherein pre-loaded databases are employed offline to estimate

This research has been funded in part by NSF under grant CNS-1642982.
The authors are with the School of Electrical and Computer Engineering, Purdue University. email: {bkeshava,michelus}@purdue.edu.

the correlation models, while we, in our work, present a fully online framework that estimates the correlation models and simultaneously**[NM: "Simultaneously" or on a "two-step" process? It can't be both.]** solves for the optimal channel sensing and access policy in a two-step iterative routine. **[NM: I am not understanding your point: Those papers estimate the correlation offline. In your paper, you FIRST estimate the correlation, and only AFTER that you optimize the policy..Practically, isn't it the same thing as estimating online?]**

The rest of the paper is organized as follows: in Sec. **??**, we define the signal model, followed by the formulations, approaches, and algorithms in Sec. **??**; in Sec. **??**, we present numerical evaluations, followed by our conclusions in Sec. **??**.

II. System Model

A. Signal Model

We consider a network consisting of P licensed users termed the Primary Users (PUs) and one cognitive radio node termed the Secondary User (SU) equipped with a spectrum sensor. The objective of the SU is to opportunistically access portions of the spectrum left unused by the PUs in order to maximize its own throughput. To this end, the SU should learn how to intelligently access spectrum holes (white-spaces) intending to maximize its throughput while maintaining strict non-interference compliance with incumbent transmissions. The wideband signal received at the SU receiver at time n is denoted as $y(n)$ and is given by

$$y(n) \;=\; \sum_{p=1}^{P} \sum_{l=0}^{L_p-1} h_p(l) x_p(n-l) + v(n), \qquad (1)$$

where $y(n)$ is expressed as a convolution of the signal $x_p(n)$ of the pth PU with the channel impulse response $h_p(n)$, and $v(n)$ denotes additive white Gaussian noise (AWGN) with variances σ_v^2. Eq. (**??**) can be written in the frequency domain by taking a K-point DFT which decomposes the observed wideband signal into K discrete narrow-band components as

$$Y_k(i) \;=\; \sum_{p=1}^{P} H_{p,k}(i) X_{p,k}(i) + V_k(i), \qquad (2)$$

where $i \in \{1,2,3,\ldots,T\}$ represents the time index; $k \in \{1,2,3,\ldots,K\}$ represents the index of the components in the frequency domain; $V_k(i) \sim \mathcal{CN}(0,\sigma_V^2)$ represents ~~a~~ circularly symmetric additive complex Gaussian noise ~~sample~~, i.i.d across frequency and across time, and independent of H and X; $X_{p,k}(i)$ is the signal of the pth PU in the frequency domain, and $H_{p,k}(i)$ is its frequency domain channel. ~~The noise samples are assumed to be independent of the occupancy state of the channels.~~ We further assume that the P PUs employ an orthogonal access to the spectrum (e.g., OFDMA) so that $X_{p,k}(i)X_{q,k}(i) = 0,\ \forall p \neq q$. Thus, letting p_k be the index of the PU that contributes to the signal in the kth spectrum band (possibly, $p_k = 0$ if no PU is transmitting in the kth spectrum band), and letting $H_k(i) = H_{p_k,k}(i)$ and $X_k(i) = X_{p_k,k}(i)$, we can rewrite (**??**) as

$$Y_k(i) \;=\; H_k(i)X_k(i) + V_k(i). \qquad (3)$$

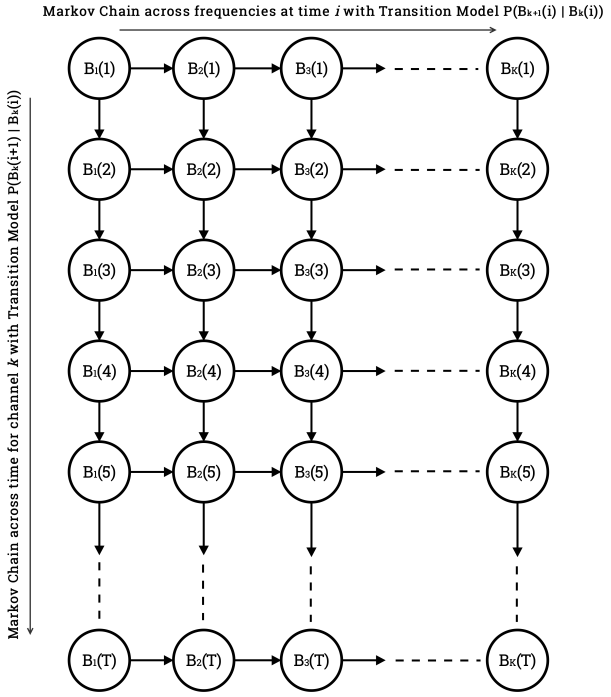

Fig. 1. The correlation model across time and across frequencies underlying the occupancy behavior of incumbents in the network **[NM: make this figure smaller, i.e., remove a few rows, such as time indices 3,4,5.. they add not information]**

Thus, $H_k(i)$ represents the kth DFT coefficient of the impulse response $h_{p_k,k}(n)$ of the channel in between the PU operating on the kth spectrum band and the SU, at time i; we model it as a zero-mean circularly symmetric complex Gaussian random variable with variance σ_H^2, $H_k \sim \mathcal{CN}(0,\sigma_H^2)$, i.i.d. across frequency bands, over time, and independent of the occupancy state of the channels.

B. PU Spectrum Occupancy Model

We now introduce the model of PU occupancy over time and across the frequency domain. We model each $X_k(i)$ as

$$X_k(i) = \sqrt{P_{tx}} B_k(i) S_k(i), \qquad (4)$$

where P_{tx} is the transmission power of the PUs, $S_k(i)$ is the transmitted symbol modelled as a constant amplitude signal, $|S_k(i)|=1$, i.i.d. over time and across frequency bands;[1] $B_k(i) \in \{0,1\}$ is the binary spectrum occupancy variable, with $B_k(i)=1$ if the kth spectrum band is occupied by a PU at time i, and $B_k(i)=0$ otherwise. Therefore, the PU occupancy behavior in the entire wideband spectrum of interest at time i, discretized into narrow-band frequency components can be modeled as the vector

$$\vec{B}(i) \;=\; [B_1(i), B_2(i), B_3(i), \cdots, B_K(i)]^T \in \{0,1\}^K. \quad (5)$$

PUs join and leave the spectrum at random times. To capture this temporal correlation in the spectrum occupancy dynamics

[1] In the case where $S_k(i)$ does not have constant amplitude, we may approximate $H_k(i)S_k(i)$ as complex Gaussian with zero mean and variance $\sigma_H^2 \mathbb{E}[|S_k(i)|^2]$, without any modification to the subsequent analysis.

of PUs, we model $\vec{B}(i)$ ~~the spectrum occupancy dynamics~~ as a Markov process: ~~given $\vec{B}(i)$, the spectrum occupancy state at time index i, $\vec{B}(i+1)$ is independent of the past, $\vec{B}(j), j < i$; $j, i \in \{1, 2, 3, \ldots, T\}$,~~ i.e.

$$\mathbb{P}(\vec{B}(i+1)|\vec{B}(j),\ \forall j \leq i) \ = \ \mathbb{P}(\vec{B}(i+1)|\vec{B}(i)). \tag{6}$$

Additionally, when joining the spectrum pool, PUs occupy a number of adjacent spectrum bands, and may vary their spectrum needs depending on traffic demands, channel conditions, etc. To capture this behavior, we model $\vec{B}(i)$ as having Markovian correlation across the bands as,

$$\mathbb{P}(\vec{B}(i+1)|\vec{B}(i)) \tag{7}$$
$$= \mathbb{P}(B_1(i+1)|B_1(i)) \prod_{k=2}^{K} \mathbb{P}(B_k(i+1)|B_k(i), B_{k-1}(i+1)).$$

That is, the spectrum occupancy at time $i+1$ in frequency band k, $B_k(i+1)$, depends on the occupancy state of the adjacent spectrum band at the same time, $B_{k-1}(i+1)$, and that of the same spectrum band k in the previous time index i, $B_k(i)$ as shown in Fig. **??**.

C. Spectrum Sensing Model

In order to detect the available spectrum holes, the SU performs spectrum sensing. However, owing to physical design limitations at the SU's spectrum sensor [**?**], not all channels in the discretized spectrum can be sensed at once, so that ~~Therefore, due to limited sensing capabilities,~~ the SU can sense only κ out of K spectrum bands at any given time, with $1 \leq \kappa \leq K$. Let $\mathcal{K}_i \subseteq \{1, 2, \ldots, K\}$ with $|\mathcal{K}_i| \leq \kappa$ be the set of indices ~~of corresponding to the~~ spectrum bands sensed by the SU at time i, which is part of our design. Then, we define the observation vector

$$\vec{Y}(i) \ = \ [Y_k(i)]_{k \in \mathcal{K}_i}, \tag{8}$$

where $Y_k(i)$ is given by (**??**). The true states $\vec{B}(i)$ encapsulate the actual occupancy behavior of the PU and the measurements at the SU are noisy observations of these true states which are modeled to be the observed states of a Hidden Markov Model (HMM). Conditional on ~~Given the spectrum occupancy vector~~ $\vec{B}(i)$ and ~~the set of sensed spectrum bands~~ $\mathcal{K}_i$, the probability density function of $\vec{Y}(i)$ is expressed as

$$f(\vec{Y}(i)|\vec{B}(i), \mathcal{K}_i) = \prod_{k \in \mathcal{K}_i} f(Y_k(i)|B_k(i)), \tag{9}$$

owing to the independence of channels, noise, and transmitted symbols across frequency bands. Moreover, from (**??**), ~~we find that~~

$$Y_k(i)|B_k(i) \sim \mathcal{CN}(0, \sigma_H^2 P_{tx} B_k(i) + \sigma_V^2). \tag{10}$$

D. POMDP Agent Model

In this section, we model the spectrum access scheme of the SU as a ~~Partially Observable Markov Decision Process~~ POMDP, whose goal ~~wherein the goal of the POMDP agent~~ is to devise an optimal sensing and access policy in order to maximize its throughput while maintaining strict non-interference compliance with incumbent transmissions. In fact, the agent's limited sensing capabilities coupled with its noisy observations result in an increased level of uncertainty at the agent's end about the occupancy state of the spectrum under consideration and the exact effect of executing an action on the radio environment. The transition model of the underlying MDP as described by (**??**), is denoted by $\mathbf{A}$ and is learned by the agent by interacting with the radio environment (see Sec. **[NM:]**). The emission model is denoted by $\mathbf{M}$ and is given by (**??**), with $f(Y_k(i)|B_k(i))$ given by (**??**).

We model the POMDP as a tuple $(\mathcal{B}, \mathcal{A}, \mathcal{Y}, \mathbf{P}, \mathbf{M})$ where $\mathcal{B} \equiv \{0,1\}^K$ represents the state space of the underlying MDP with states $\vec{B}$ given by all possible realizations of the spectrum occupancy vector as described by (**??**), $\mathcal{A}$ represents the action space of the agent, given by all ~~(K, κ)~~ possible combinations in which the κ spectrum bands are chosen to be sensed out of K at any given time; and $\mathcal{Y}$ represents the observation space of the agent based on the signal model outlined in the previous subsection. The state of the POMDP at time i is given by the *prior belief* β_i, which represents the probability distribution of the underlying MDP state $\vec{B}(i)$, given the information collected by the agent up to time i, but before collecting the new information in slot i. At the beginning of each time index i, given β_i, the agent selects κ spectrum bands out of K according to a policy $\pi(\beta_i)$, thus defining the sensing set $\mathcal{K}_i$, performs spectrum sensing on these spectrum bands, observes $\vec{Y}(i) \in \mathcal{Y}$, and updates its *posterior belief* $\hat{\beta}_i$ of the current spectrum occupancy $\vec{B}(i)$ as

$$\hat{\beta}_i(\vec{B}') = \mathbb{P}(\vec{B}(i) = \vec{B}'|\vec{Y}(i), \mathcal{K}_i, \beta_i) \tag{11}$$
$$= \frac{\mathbb{P}(\vec{Y}(i)|\vec{B}', \mathcal{K}_i)\ \beta_i(\vec{B}')}{\sum_{\vec{B}'' \in \{0,1\}^K} \mathbb{P}(\vec{Y}(i)|\vec{B}'', \mathcal{K}_i)\ \beta_i(\vec{B}'')}.$$

We denote the function that maps the prior belief β_i to the posterior belief $\hat{\beta}_i$ through the spectrum sensing action $\mathcal{K}_i$ and the observation signal $\vec{Y}(i)$ as $\hat{\beta}_i = \hat{\mathbb{B}}(\beta_i, \mathcal{K}_i, \vec{Y}(i))$. ~~where $\mathbb{P}(\vec{Y}(i)|\mathcal{K}_i, \beta_{i-1})$ denotes the normalization constant and β_{i-1} represents the belief of the agent prior to the observation $\vec{Y}(i)$, defined as a probability distribution over all possible states.~~

Given the posterior belief $\hat{\beta}_i$, we employ a threshold based detection mechanism to determine whether a given channel is occupied ($\phi_k(\hat{\beta}_i) = 1$ for channel k) or idle ($\phi_k(\hat{\beta}_i) = 0$). **[NM: can you add more details here on how ϕ is defined, or a reference to the section where it is discussed?]** If a channel k is deemed to be idle by this threshold based detection mechanism, the SU accesses it for the delivery of its network flows. ~~On the other hand, if a channel k is deemed to be occupied, the SU~~ Otherwise, it leaves it untouched. ~~Furthermore, for utility evaluation, since relying on feedback from the radio environment will introduce unforeseen variables and additional dynamics into the problem, we use the state estimator detailed in Sec. refIII to determine the reference occupancy metrics of the incumbents (the reference estimated state vector) $\hat{\vec{B}}$ based on the observations $\vec{Y}(i)$ obtained from the POMDP agent's sensing action, i.e. κ.~~ **[NM: ??? I don't understand this last statement..]** Given the PU occupancy state $\vec{B}(i)$ and posterior belief $\hat{\beta}_i$, the reward metric of the POMDP is given by the number of *truly idle* bands detected by the SU accounting for the throughput maximization aspect

of the agent's end-goal and a penalty for *missed detections* accounting for the incumbent non-interference constraint, ~~the reward to the agent is modeled as~~ i.e.

$$R(\hat{\vec{B}}(i)\vec{B}(i),\hat{\beta}_i)=\sum_{k=1}^{K}(1-\hat{B}_k B_k(i))(1-\phi_k(\hat{\beta}_i))-\lambda\hat{B}_k B_k(i)(1-\phi_k(\hat{\beta}_i)), \tag{12}$$

[NM: why $\hat{\vec{B}}(i)$? It should be the true state $\vec{B}(i)$] where $\lambda > 0$ represents the cost term penalizing the agent for missed detections, i.e. interference with the incumbent. After performing data transmission, the SU computes the prior belief for the next slot based on the dynamics of the Markov chain as

$$\beta_{i+1}(\vec{B}')=\mathbb{P}(\vec{B}(i+1)=\vec{B}'|\mathcal{K}_{i+1},\hat{\beta}_i). \tag{13}$$

We denote the function that maps the posterior belief $\hat{\beta}_i$ to the prior belief $\hat{\beta}_{i+1}$ as $\beta_{i+1}=\mathbb{B}(\hat{\beta}_i)$. **[NM: why $\mathcal{K}_{i+1}$? It should be $\mathcal{K}_i$.. and actually it only depends on $\hat{\beta}_i$ but is independent of $\mathcal{K}_i$.]** The ~~optimization objective~~ goal of the problem at hand is to determine an optimal spectrum sensing policy to maximize the infinite-horizon discounted reward, ~~can be written as~~

$$\pi^*=\arg\max_{\mathcal{K}_i\in\mathcal{A}^\pi} V^\pi(\beta)\triangleq\mathbb{E}_\pi\Big[\sum_{i=0}^{\infty}\gamma^i R(\vec{B}(i),\mathcal{K}_i)R(\vec{B}(i),\beta_i)\mid\beta_0=\beta\Big], \tag{14}$$

[NM: why are you changing the reward definition? Before it has been defined as $R(\vec{B}(i),\hat{\beta}_i)$!] where $0<\gamma<1$ is the discount factor, β_0 is the initial belief, and $\hat{\beta}_i$ is the posterior belief induced by policy π, which ~~The action policy π of the agent~~ maps the beliefs β_i to actions $\mathcal{K}_i$ at time i, ~~and is characterized by a~~ and we have defined the value function $V^\pi(\beta)$ under policy π starting from belief β.

$$V^\pi(\beta)=\mathbb{E}_\pi\Big[\sum_{i=0}^{\infty}\gamma^i R(\vec{B}(i),\pi(\beta_i))\mid\beta_0=\beta\Big], \tag{15}$$

[NM: again, why are you changing the reward? Be consistent!] ~~where $\pi(\beta_i)$ is the action taken by the agent at time i under policy π. The optimal policy π^* specifies the optimal action to take at the current time index assuming that the agent behaves optimally at future time indices as well. It is evident from equation eqref15 that we have an infinite-horizon discounted reward problem formulation and in order to solve for the optimal policy we need to solve the Bellman equation given by~~ The optimal policy π^* and the corresponding optimal reward $V^*(\beta)$ are the solutions of Bellman's optimality equation $V^*=H[V^*]$, where we have defined the operator $V_{n+1}=H[V_n]$ as

$$V_{n+1}(\beta)=\max_{\mathcal{K}\in\mathcal{A}}\Big[\sum_{\vec{B}\in\mathcal{B}}R(\vec{B},\mathcal{K})\beta(\vec{B})\mathbb{E}_{\vec{Y}|\vec{B},\mathcal{K}}\Big[R(\vec{B},\hat{\mathbb{B}}(\beta,\mathcal{K},\vec{Y}))+\gamma\sum_{\vec{Y}\in\mathcal{Y}}\mathbb{P}(\vec{Y}|\mathcal{K},\beta)V_n(\hat{\beta}\mathbb{B}(\hat{\mathbb{B}}(\beta,\mathcal{K},\vec{Y})))\Big]\Big],\ \forall\beta. \tag{16}$$

[NM: why are you changing the reward?] This problem can be solved using the value iteration algorithm, i.e., by solving (**??**) iteratively until convergence to a fixed point. However, given the high dimensionality of the spectrum sensing and access problem, i.e. the number of states of the underlying MDP scales exponentially with the number of bands in the spectrum, solving equation (**??**) using Exact Value Iteration and Policy Iteration algorithms is computationally infeasible. Additionally, solving for the optimal policy from equation (**??**) requires prior knowledge about the underlying MDP's transition model. Therefore, in this paper we present a framework to estimate the transition model of the underlying MDP online, while utilizing this learned model to solve for the optimal policy by employing Randomized Point-Based Value Iteration techniques, namely, the PERSEUS algorithm [**?**].

III. Approaches and Algorithms

A. Occupancy Behavior Transition Model Estimation

In real-world implementations of cognitive radio systems, the transition model of the occupancy behavior of the PUs is not known to the SUs in the network and needs to be learned over time. The learned model then needs to be fed back to the POMDP agent as a stride in a two-step iterative routine in order to solve for the optimal policy. The system can learn the model either before triggering or during the operation of the POMDP agent. Inherently, the approach constitutes solving a parameter estimation problem formulated as

$$\mathbf{A}^*=\arg\max_{\mathbf{A}}\ \mathbb{P}([\vec{Y}(i)]_{i=1}^{\tau}|\mathbf{A}), \tag{17}$$

which is a Maximum Likelihood Estimation (MLE) problem, where $\mathbf{A}$ is defined as $\mathbb{P}(\vec{B}(i+1)|\vec{B}(i))$ and τ refers to the learning period of the parameter estimator: this, as mentioned earlier, can be equal to the entire duration of the POMDP agent's interaction with the radio environment or can be a predefined parameter learning period before triggering the POMDP agent. In order to facilitate better readability, for the description of this parameter estimator, we denote $[\vec{Y}(i)]_{i=1}^{\tau}$ as $\mathbf{Y}$ and $[\vec{B}(i)]_{i=0}^{\tau}$ as $\mathbf{B}$. Re-framing (**??**) as an optimization of the log-likelihood, ~~using the definition of marginal probability, and focusing on the joint instead of the conditional,~~ we get,

$$\mathbf{A}=\arg\max_{\mathbf{A}}\ \log\Big(\sum_{\mathbf{B}}\mathbb{P}(\mathbf{B},\mathbf{Y}|\mathbf{A})\Big). \tag{18}$$

In order to exploit the characteristics of the stated Markov model, we multiply and divide the operand of the logarithm by ρ**[NM: what is this?]** which from the equality constraint of Jensen's Inequality turns out to be $\mathbb{P}(\mathbf{B}|\mathbf{Y},\hat{\mathbf{A}})$. The optimization problem in (**??**) is then restated as,

$$\mathbf{A}=\arg\max_{\mathbf{A}}\ \sum_{\mathbf{B}}\mathbb{P}(\mathbf{B}|\mathbf{Y},\hat{\mathbf{A}})\ \log(\mathbb{P}(\mathbf{B},\mathbf{Y},\mathbf{A})). \tag{19}$$

[NM: this is NOT the same as (??)...this is just the EM algorithm, you dont need to go through its derivations.. just say that you use the EM algorithm and provides the two steps of the algorithm (Estep and Mstep)] Applying the characteristics of the Markov model discussed in Sec. **??**, we write (**??**) as

$$\mathbf{A}=\arg\max_{\mathbf{A}}\sum_{\mathbf{B}}\mathbb{P}(\mathbf{B}|\mathbf{Y},\hat{\mathbf{A}})\sum_{L}\sum_{R}\sum_{i=1}^{\tau}\sum_{j=1}^{\tau}\mathcal{I}\log(M_R(\vec{Y}(i)))+\mathcal{J}\log(a_{LR}), \tag{20}$$

[NM: where is the dependence on A here? Everything seems to depend on $\hat{A}$] where $L, R \in \{0,1\}^K$ represent iterables**[NM: ?]** for the occupancy state vectors,

$$\mathbf{M}_R(\vec{Y}(i)) = \mathbb{P}(\vec{Y}(i)|\vec{B}(i) = R), \tag{21}$$

represents the emission model outlined in (**??**),

$$a_{LR} = \mathbb{P}(\vec{B}(i) = R|\vec{B}(i-1) = L, \hat{\mathbf{A}}), \tag{22}$$

represents the unknown transition model which is the subject of this estimation, and $\mathcal{I}$ and $\mathcal{J}$ detailed below are indicator random variables introduced to bring in specificity into the estimation procedure.

$$\mathcal{I} = \begin{cases} 1, & \text{if } \vec{Y}(i) \text{ and } \vec{B}(i) = R\text{[NM : whatdoes"ifY"mean?IfYwhat?]} \\ 0, & \text{otherwise} \end{cases} \tag{23}$$

$$\mathcal{J} = \begin{cases} 1, & \text{if } \vec{B}(i-1) = L \text{ and } \vec{B}(i) = R \\ 0, & \text{otherwise} \end{cases} \tag{24}$$

We impose a constraint on the transition probability in (**??**) as

$$\sum_R a_{LR} = 1, \tag{25}$$

and formulate the Lagrangian as

$$\begin{aligned} \mathcal{L} = \Big\{ \sum_{\mathbf{B}} \mathbb{P}(\mathbf{B}|\mathbf{Y}, \hat{\mathbf{A}}) \sum_L \sum_R \sum_{i=1}^{\tau} \sum_{j=1}^{\tau} \mathcal{I} \log(\mathbf{M}_R(\vec{Y}(i))) \\ + \mathcal{J} \log(a_{LR}) \Big\} \\ + \sum_L \lambda_L (1 - \sum_R a_{LR}). \end{aligned} \tag{26}$$

Solving for a_{LR}, we get,

$$a_{LR} = \frac{\sum_{i=1}^{\tau} \mathbb{P}(\mathbf{Y}, \mathbf{A}, \vec{B}(i) = R, \vec{B}(i-1) = L)}{\sum_{i=1}^{\tau} \mathbb{P}(\mathbf{Y}, \mathbf{A}, \vec{B}(i-1) = L)}. \tag{27}$$

[NM: this is pretty standard HMM stuff, no need to repeat the derivations.. just name a relevant reference and provide the mathematical steps of the algorithm specialized to your scenario, but without going through all derivations..] In order to further simplify (**??**) and bring it into an iterative algorithmic form, we introduce the forward and backward probabilities. We define the forward probability as

$$\begin{aligned} F(i,R) &= \mathbb{P}([\vec{Y}(t)]_{t=1}^{i}, \vec{B}(i) = R) \\ &= \sum_L \mathbb{P}(\vec{B}(i) = R, \vec{Y}(i)|\vec{B}(i-1) = L) F(i-1, L), \end{aligned} \tag{28}$$

and the backward probability as

$$\begin{aligned} D(i,L) &= \mathbb{P}([\vec{Y}(t)]_{t=i}^{\tau}|\vec{B}(i-1) = L) \\ &= \sum_R \mathbb{P}(\vec{B}(i) = R, \vec{Y}(i)|\vec{B}(i-1) = L) D(i+1, R). \end{aligned} \tag{29}$$

Using these definitions, (**??**) can be rewritten as,

$$a_{LR} = \frac{\sum_{i=1}^{\tau} F(i-1, L) \mathbf{M}_R(\vec{Y}(i)) a_{LR} D(i+1, R)}{\sum_R \sum_{i=1}^{\tau} F(i-1, L) \mathbf{M}_R(\vec{Y}(i)) a_{LR} D(i+1, R)}. \tag{30}$$

[NM: why is a_{LR} appearing on both sides of the equation?]

B. Occupancy Behavior State Estimation

During the reward evaluation phase of the POMDP agent at time i, the observations made based on the sensing action $\mathcal{K}_i$ are employed at a state estimator to determine the occupancy state of the spectrum bands. Based on this estimated state vector, we formulate the false alarm and missed detection metrics which allow us to capture the throughput maximization and PU non-interference requirements essential for the operation of our POMDP agent. We formulate the state estimation problem as**[NM: which value function? You have already defined the value function V for the POMDP, dont confuse the two things]**

$$\vec{B}(i)^* = \arg\max_{\vec{B}} \ \mathbb{P}(\vec{B}|\vec{Y}(i)), \tag{31}$$

[NM: isnt this conditional on the belief as well?] which is a Maximum-A-Posteriori (MAP) estimation problem. This optimization problem can be restated in terms of the value functions as **[NM: I dont understand why you need to go through all of this.. if you have the posterior belief $\hat{\beta}$, that is all you need to do your MAP estimate, i.e. $\vec{B}(i)^* = \arg\max_{\vec{B}} \hat{\beta}_i(\vec{B})$]**

$$W_{i,k}^r = \max_{\tilde{\mathbf{B}}_{[t-1,k-1]}} \ \mathbb{P}(\tilde{\mathbf{Y}}_{[t-1,k-1]}, \tilde{\mathbf{B}}_{[t-1,k-1]}, Y_k(i), B_k(i)), \tag{32}$$

where for the estimation of the occupancy in spectrum band k at time i,

$$\tilde{\mathbf{Y}}_{[t-1,k-1]} \equiv \{ \ [\vec{Y}_v(i)]_{v=1}^{k-1}, \ [\vec{Y}_k(t)]_{t=1}^{t-1} \ \} \tag{33}$$

denotes the set of all essential past observations which for readability purposes is denoted simply as $\tilde{\mathbf{Y}}$ and

$$\tilde{\mathbf{B}}_{[t-1,k-1]} \equiv \{ \ [\vec{B}_v(i)]_{v=1}^{k-1}, \ [\vec{B}_k(t)]_{t=1}^{t-1} \ \} \tag{34}$$

denotes set of all essential past states which is henceforth simply referred to as $\tilde{\mathbf{B}}$ for readability. Applying the characteristics of the Markov model detailed in (**??**) to (**??**), we get

$$\begin{aligned} W_{i,k}^r = \mathbf{M}_r(Y_k(i)) \max_{\tilde{\mathbf{B}}} \mathbb{P}(B_k(i) = r|B_k(i-1) = m, \\ B_{k-1}(i) = n) \mathbb{P}(\tilde{\mathbf{Y}}, \tilde{\mathbf{B}}), \end{aligned} \tag{35}$$

which can be simplified further to show that,

$$W_{i,k}^r = \mathbf{M}_r(Y_k(i)) \max_{m,n} \ a_{mnr} W_{i-1,k}^m W_{i,k-1}^n, \tag{36}$$

where

$$a_{mnr} = \mathbb{P}(B_k(i) = r|B_k(i-1) = m, B_{k-1}(i) = n), \tag{37}$$

which can be evaluated from the estimated transition model. Equation (**??**) corresponds to the forward recursion aspect of the double Markov chain Viterbi algorithm. Next, similar to the backtracking procedure in the one dimensional (single Markov chain) Viterbi algorithm, the Trellis diagram is traversed backwards to recover the most probable previous neighbours of $B_k(i)$. This is done recursively until the entire Trellis diagram has been traversed to yield the most probable state sequence, i.e. the Viterbi path. Mathematically, the backtracking step with respect to the neighbours of $B_k(i)$ is represented as

$$m^*, n^* = \arg\max_{m,n} \ a_{mnr} W_{i-1,k}^m W_{i,k-1}^n, \tag{38}$$

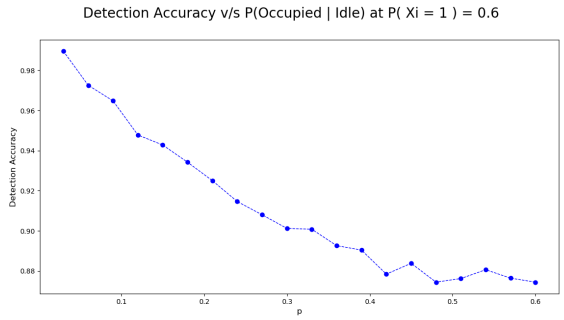

Fig. 2. The detection accuracy of the unconstrained Viterbi algorithm over varying values of $\mathbb{P}$(Occupied|Idle)

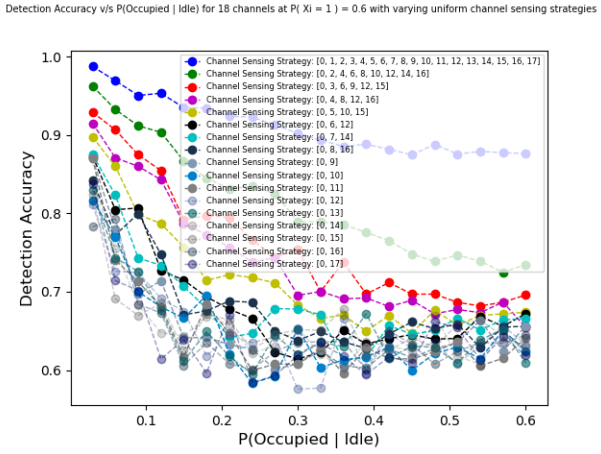

Fig. 3. The detection accuracies of the constrained Viterbi algorithm for different sensing strategies over varying values of$\mathbb{P}$(Occupied|Idle)

where m^* is the most probable state of channel k in time index $i-1$ and n^* is the most probable state of channel $k-1$ in time index i, given that channel k in time index i is in state r; $m, n, r \in \{0, 1\}$.

C. The PERSEUS Algorithm

As discussed in Sec. **??** of this article, solving the Bellman equation (**??**) for POMDPs with large state and action space using exact value iteration and policy iteration techniques [**?**] is computationally infeasible [**?**], [**?**]. Hence, we resort to approximate value iteration techniques to ensure that the system scales well to a large number of bands in the spectrum of interest. For infinite-horizon POMDPs, V^* in (**??**) can be approximated by a Piece-Wise Linear and Convex function (PWLC) [**?**]. The core idea behind the PERSEUS algorithm is that the value function in time index i can be parameterized by a set of hyperplanes $\{\vec{\alpha}_i^u\}$, $u = \{0, 1, 2, \ldots, |V_i|\}$, each of which represents a region of the belief space for which it is the maximizing element. The backup step is defined as the process of determining the optimal hyperplane out of the set of available hyperplanes in time index i as

$$\vec{\alpha}_i(\beta) = \arg\max_{\vec{\alpha}_i^u} \beta \cdot \vec{\alpha}_i^u, \tag{39}$$

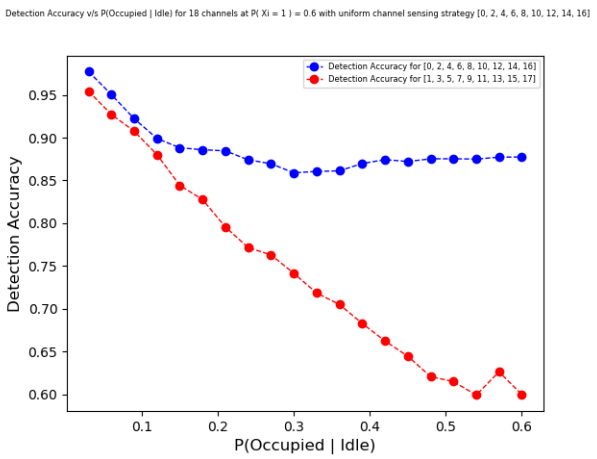

Fig. 4. The detection accuracies of the constrained Viterbi algorithm for sensed and un-sensed channels under a given channel sensing strategy

[NM: argmax from which set?] which implies that,

$$\begin{aligned} V_i(\beta) &= \max_{\vec{\alpha}_i^u} \beta \cdot \vec{\alpha}_i^u, \\ \pi_i(\beta) &= a(\vec{\alpha}_i^{u*}), \end{aligned} \tag{40}$$

[NM: max from which set?] where $a(\vec{\alpha}_i^u)$ ~~refers to~~ is the action corresponding to ~~the optimal~~ hyperplane $\vec{\alpha}_i^u$. The PERSEUS algorithm constitutes an Exploration phase**[NM: i am not sure I understand this thing about exploration... PERSEUS is NOT an online algorithm]** wherein the POMDP agent randomly interacts with the radio environment to collect a set of so-called reachable beliefs which are to be improved over numerous iterative backup stages. In each backup stage, the agent samples a belief β uniformly at random from the set of unimproved points and performs a backup on this sampled belief point according to (**??**)**[NM: how do you compute the new hyperplanes? You are not showing that]** to determine the optimal hyperplane $\vec{\alpha}$. Considering an arbitrary time index i, if $V_{i+1}(\beta) = \beta \cdot \vec{\alpha} \geq V_i(\beta)$, then the belief point β is said to be improved along with any other belief points β' in the unimproved set for which $V_{i+1}(\beta') = \beta' \cdot \vec{\alpha} \geq V_i(\beta')$. If $V_{i+1}(\beta) = \beta \cdot \vec{\alpha} < V_i(\beta)$, then a copy of the maximizing hyperplane for $V_i(\beta)$ is used for $V_{i+1}(\beta)$ and the belief point β is then removed from the set of unimproved points. The backup stage continues until the set of unimproved points is empty and the agent performs a series of backup stages until there the number of policy changes between iterations is below a specified threshold η.

The belief update procedure outlined in (**??**) is an essential aspect of the PERSEUS algorithm which can turn into a performance bottleneck for large state spaces due to the inherent iteration over all possible states. In order to circumvent this problem, we fragment the spectrum into much smaller, independent sets of correlated channels and then run the PERSEUS algorithm on these fragments by leveraging multi-processing and multi-threading tools available at our disposal in software frameworks. Furthermore, we avoid iterating over all possible states and allow only those state transitions we deem to be the most probable - for example, we allow only

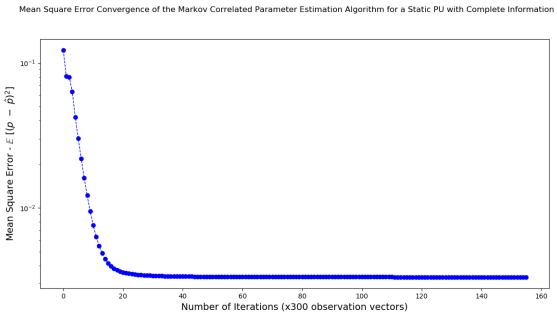

Fig. 5. The mean square error convergence plot of the parameter estimation algorithm**[NM: Show x axis up to 40]**

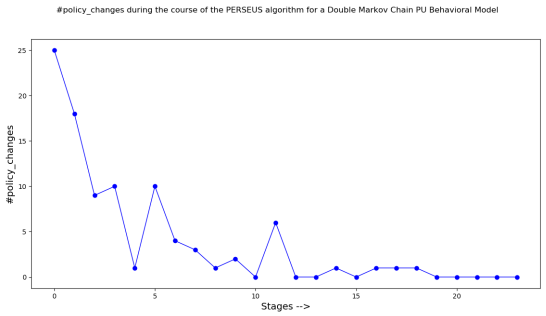

Fig. 7. The number of policy changes involved in the PERSEUS algorithm as it transitions toward convergence over numerous backup and wrapper stages

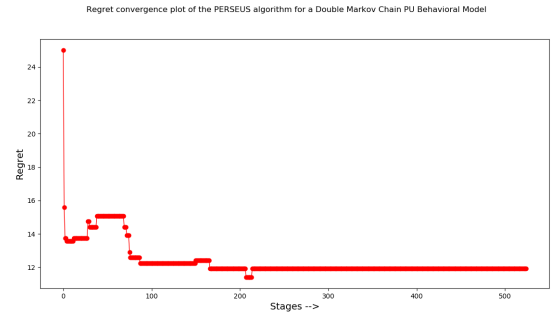

Fig. 6. The regret convergence plot of the PERSEUS algorithm over several backup and wrapper stages

those state transitions that involve a Hamming distance of up to 3 between the previous state vector and the current state vector in an 18 channel radio environment.**[NM: you need to provide more details here, including math..]**

IV. Numerical Evaluation

The given framework is simulated in Python for a system with 18 channels and a channel model constituting an SNR of 19dB when an incumbent occupies a specific channel. The same Markovian transition model, emission model, and steady-state model is employed across both the chain across time and the chain across frequencies. The plot depicted in Fig. **??** illustrates the detection accuracy of the Viterbi algorithm wherein the SU makes observations of all the channels in the radio environment and estimates the occupancy states of these channels over varying values of $\mathbb{P}(\text{Occupied}|\text{Idle})$, i.e., as the channels transition toward independence. We note that the detection accuracy of the state estimator degrades as the channels transition toward independence, which is as surmised, because the Viterbi algorithm's structure begins to crumble if the Markovian correlation ceases to exist among cells in the grid, where a cell corresponds to the state of a certain channel at a given time index.

Fig. **??** illustrates the detection accuracies of the Viterbi algorithm wherein the SU makes observations of only the channels in the given channel sensing strategy and estimates the occupancy states of both the sensed and the un-sensed channels over varying values of $\mathbb{P}(\text{Occupied}|\text{Idle})$. In Fig. **??**, the plot depicted shows the detection accuracies of the estimation of sensed and un-sensed channels for the constrained Viterbi algorithm in which the SU only senses channels whose indices correspond to the multiples of 2 and uses these channels to estimate the occupancy behavior of the incumbents across all 18 channels over varying values of $p = \mathbb{P}(\text{Occupied}|\text{Idle})$. In both these plots, we observe that, as anticipated, the detection accuracy of the state estimator deteriorates as the amount of information available for estimation reduces.

The plot depicted in Fig. **??** shows the Mean Square Error convergence of the Parameter Estimation algorithm while determining the transition model of the MDP underlying the problem under consideration. Starting with an initial estimate of 0.0, the EM algorithm detailed in Sec. **??** converges to the true transition model with an error of $\epsilon \leq 10^{-8}$ over numerous iterations, each iteration corresponding to an averaging operation of 300 observation vectors. Here, we observe the mean square error given by $\mathbb{E}[(p-\hat{p})^2]$ iteratively reduces as it goes through the E-step which finds the distribution for the parameters in the current time-step, given the previous estimate and the M-step which determines the maximum likelihood estimate of the parameters, given the distribution obtained in the E-step. It has been theoretically shown to converge, i.e., each iteration either improves the true likelihood or leaves it unchanged [**?**]. Since the EM algorithm is susceptible to premature convergence to local optima and saddle points, we mitigate this by averaging the procedure over several cycles.

Fig. **??** illustrates the Regret Convergence plot of the PERSEUS algorithm over several backup and wrapper stages wherein the regret metric corresponds to the difference in utility between the PERSEUS algorithm at a certain stage and an Oracle which has complete information with respect to the occupancy behavior of the incumbents in the radio environment. Furthermore, the algorithm involves an online estimation of the transition model of the underlying MDP and a random exploration strategy to gather the initial set of reachable beliefs. Fig. **??** depicts the number of policy changes involved in each individual stage of the PERSEUS algorithm as it moves toward convergence. The termination condition for the PERSEUS algorithm is that the number of policy changes over several consecutive backup stages should be zero, i.e., $\eta = 0$. These plots, similar to the *Reward v/s Time* and $\Delta\pi$ *v/s time* plots in [**?**], serve as a measure of convergence for our fragmented PERSEUS algorithm with simplified belief

updates, and an online transition model estimation.

V. Conclusion

In this paper, we formulate the optimal spectrum sensing and access problem in an AWGN observation model with multiple licensed users and a cognitive radio node restricted in terms of its sensing capabilities, as a partially observable stochastic game. In a radio environment wherein the occupancy behavior of the incumbents is correlated across time and frequencies, we present a consolidated framework that employs the Expectation-Maximization algorithm to estimate the transition model of this occupancy behavior and leverage a fragmented PERSEUS algorithm with belief update heuristics to simultaneously solve for the optimal spectrum sensing and access policy. Through system simulations, we show that our framework out-performs the existing correlation-coefficient based state-of-the-art; surpasses Neyman-Pearson based occupancy detection schemes that fail to exploit the correlation across time and frequencies; and matches the performance of standard MAP-estimators which possess the transition model statistics as an apriori.

References

[1] C. Pradhan, K. Sankhe, S. Kumar, and G. R. Murthy, "Revamp of enodeb for 5g networks: Detracting spectrum scarcity," in *2015 12th Annual IEEE Consumer Communications and Networking Conference (CCNC)*, Jan 2015, pp. 862–868.

[2] M. Danneberg, R. Datta, A. Festag, and G. Fettweis, "Experimental testbed for 5g cognitive radio access in 4g lte cellular systems," in *2014 IEEE 8th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, June 2014, pp. 321–324.

[3] F. Xu, L. Zhang, Z. Zhou, and Y. Ye, "Architecture for next-generation reconfigurable wireless networks using cognitive radio," in *2008 3rd International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CrownCom 2008)*, May 2008, pp. 1–5.

[4] K. Cohen, Q. Zhao, and A. Scaglione, "Restless multi-armed bandits under time-varying activation constraints for dynamic spectrum access," in *2014 48th Asilomar Conference on Signals, Systems and Computers*, Nov 2014, pp. 1575–1578.

[5] J. Lundén, S. R. Kulkarni, V. Koivunen, and H. V. Poor, "Multiagent reinforcement learning based spectrum sensing policies for cognitive radio networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 5, pp. 858–868, Oct 2013.

[6] M. Gao, X. Yan, Y. Zhang, C. Liu, Y. Zhang, and Z. Feng, "Fast spectrum sensing: A combination of channel correlation and markov model," in *2014 IEEE Military Communications Conference*, Oct 2014, pp. 405–410.

[7] C. Park, S. Kim, S. Lim, and M. Song, "Hmm based channel status predictor for cognitive radio," in *2007 Asia-Pacific Microwave Conference*, Dec 2007, pp. 1–4.

[8] G. Ding, J. Wang, Q. Wu, L. Yu, Y. Jiao, and X. Gao, "Joint spectral-temporal spectrum prediction from incomplete historical observations," in *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Dec 2014, pp. 1325–1329.

[9] J. Oksanen, V. Koivunen, J. Lundén, and A. Huttunen, "Diversity-based spectrum sensing policy for detecting primary signals over multiple frequency bands," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2010, pp. 3130–3133.

[10] S. Chaudhari, V. Koivunen, and H. V. Poor, "Autocorrelation-based decentralized sequential detection of ofdm signals in cognitive radios," *IEEE Transactions on Signal Processing*, vol. 57, no. 7, pp. 2690–2700, July 2009.

[11] S. Yin, D. Chen, Q. Zhang, M. Liu, and S. Li, "Mining spectrum usage data: A large-scale spectrum measurement study," *IEEE Transactions on Mobile Computing*, vol. 11, no. 6, pp. 1033–1046, June 2012.

[12] R. I. C. Chiang, G. B. Rowe, and K. W. Sowerby, "A quantitative analysis of spectral occupancy measurements for cognitive radio," in *2007 IEEE 65th Vehicular Technology Conference - VTC2007-Spring*, April 2007, pp. 3016–3020.

[13] M. A. McHenry, P. A. Tenhula, D. McCloskey, D. A. Roberson, and C. S. Hood, "Chicago spectrum occupancy measurements & analysis and a long-term studies proposal," in *Proceedings of the First International Workshop on Technology and Policy for Accessing Spectrum*, ser. TAPAS '06. New York, NY, USA: ACM, 2006. [Online]. Available: http://doi.acm.org/10.1145/1234388.1234389

[14] L. Ferrari, Q. Zhao, and A. Scaglione, "Utility maximizing sequential sensing over a finite horizon," *IEEE Transactions on Signal Processing*, vol. 65, no. 13, pp. 3430–3445, July 2017.

[15] N. Michelusi and U. Mitra, "Cross-layer estimation and control for cognitive radio: Exploiting sparse network dynamics," *IEEE Transactions on Cognitive Communications and Networking*, vol. 1, no. 1, pp. 128–145, March 2015.

[16] N. Michelusi, M. Nokleby, U. Mitra, and R. Calderbank, "Multi-Scale Spectrum Sensing in Dense Multi-Cell Cognitive Networks," *IEEE Transactions on Communications*, vol. 67, no. 4, pp. 2673–2688, April 2019.

[17] S. Maleki, S. P. Chepuri, and G. Leus, "Energy and throughput efficient strategies for cooperative spectrum sensing in cognitive radios," in *2011 IEEE 12th International Workshop on Signal Processing Advances in Wireless Communications*, June 2011, pp. 71–75.

[18] M. T. J. Spaan and N. A. Vlassis, "Perseus: Randomized point-based value iteration for pomdps," *CoRR*, vol. abs/1109.2145, 2011. [Online]. Available: http://arxiv.org/abs/1109.2145

[19] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artif. Intell.*, vol. 101, no. 1-2, pp. 99–134, May 1998. [Online]. Available: http://dx.doi.org/10.1016/S0004-3702(98)00023-X

[20] J. Pineau, G. Gordon, and S. Thrun, "Point-based value iteration: An anytime algorithm for pomdps," in *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, ser. IJCAI'03. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2003, pp. 1025–1030. [Online]. Available: http://dl.acm.org/citation.cfm?id=1630659.1630806

[21] R. M. Neal and G. E. Hinton, *A View of the Em Algorithm that Justifies Incremental, Sparse, and other Variants*. Dordrecht: Springer Netherlands, 1998, pp. 355–368. [Online]. Available: https://doi.org/10.1007/978-94-011-5014-9_12