# Spectrum Sensing in Cognitive Radio Networks via Approximate POMDP

Bharath Keshavamurthy, Nicolò Michelusi

***Abstract*—[NM: per journal guidelines, abstract needs to be limited to 100 words.] In this paper, a novel spectrum sensing and access strategy based on Partially Observable Markov Decision Processes is proposed. A cognitive radio learns the correlation model defining the occupancy behavior of incumbents, based on which it devises an optimal spectrum sensing and access strategy. The optimization complexity is ameliorated via point-based value iteration methods. Numerical evaluations demonstrate that the proposed scheme performs closely to a MAP-based state estimator with prior knowledge of the correlation model, and outperforms a state-of-the art correlation-coefficient based clustering algorithm by an average of 80%; additionally, it surpasses a Neyman-Pearson Detector that assumes independence among channels, by an average of 25%, thus revealing the importance of leveraging the correlation structure underlying the occupancy behavior of incumbents. [NM: the most important thing to comment about is the new figure..]**

***Index Terms*—Hidden Markov Model, Cognitive Radio, Spectrum Sensing, POMDP**

## I. Introduction

With the advent of fifth-generation wireless communications, the problem of spectrum scarcity has been exacerbated [1]. Cognitive radio networks facilitate efficient spectrum utilization by intelligently accessing *white spaces* left unused by the sparse and infrequent transmissions of licensed users, while ensuring rigorous incumbent non-interference compliance [2].

A crucial aspect underlying the design of cognitive radio networks is the ability to perform spectrum sensing. ~~channel access protocol in the MAC layer of the stack. Additionally~~ However, physical design limitations are imposed on the cognitive radio's spectrum sensor in view of quick turnaround times and energy efficiency [3], which restrict the number of channels that can be sensed at any given time. This has led to research in algorithms that first determine the best channels to sense and then~~, via aggregation or correlation exploitation,~~ use the gathered information to perform channel access. The state-of-the-art are based on ~~In this regard, the current state-of-the-art involves channel sensing and access strategies dictated by~~ custom heuristics [4], [5], multi-armed bandits [6], and reinforcement learning [7]. However, most of these works, such as [6]–[10], assume independence across frequency bands~~ies in the discretized spectrum~~, which is imprudent because licensed users may exhibit correlation across both frequency and time in their channel occupancy behavior: they may occupy a set of adjacent frequency bands for an extended period of time ~~(frequency correlation), repeating similar motifs in behavior over an extended period of time (temporal correlation)~~ [11]–[13]. This correlation structure~~This pattern in occupancy behavior of the incumbents imputes very high levels of correlation among channels which~~ may be leveraged for more accurate predictions of spectrum holes. In this paper, we propose a parameter estimation algorithm to learn the frequency and time correlation structure, based on which we solve for the optimal channel sensing policy.

Distributed spectrum sensing has been considered in [7] 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. In [5], the frequency correlation is exploited, but a noise-free observation model is assumed**[NM: what do you mean by that? The work is about spectrum sensing, I found it hard to believe that they have noise free observations.. please clarify]**. Compared to [5], [7], we account for the uncertainty in the occupancy state and for noisy observations via a partially observable Markov decision process (POMDP) formulation.~~Instead, we account for the impact of noisy observations in our design.~~ Standard MAP-based state estimators for Hidden Markov Models (HMMs) such as the Viterbi algorithm can be employed to estimate spectrum occupancy**[NM: is any of the paper you cite doing this?]**; however, these estimators rely on knowledge of the transition model, which may be unknown in practice. Therefore, in our paper, we embed a parameter estimation algorithm to learn a parametric time-frequency correlation model. In [4], [5], this model is estimated offline based on pre-loaded databases.**[NM: Is that the only difference wrt [4]? What are the other differences? Since the model is similar, are you doing a performance comparison with [4], [5]?]** Instead, in our work, we present a fully online framework to estimate it and, simultaneously, to solve for the optimal channel sensing strategy.

**[NM: what are the contributions in a few line? What are the key results?]** The contributions of this paper are as follows:....

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

## II. System Model

We consider a network consisting of $J$ 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 **[NM: No need to do this.. jump straight to OFDMA and to the frequency domain..]**

$$y(n) = \sum_{j=1}^{J} \sum_{l=0}^{L_j-1} h_j(l) x_j(n-l) + v(n), \qquad (1)$$

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.

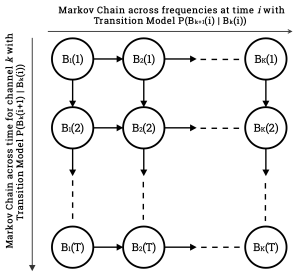


Fig. 1. The correlation model across time and frequencies underlying the occupancy behavior of incumbents in the network

where $y(n)$ is expressed as a convolution of the signal $x_j(n)$ of the $j$th PU with the channel impulse response $h_j(n)$, and $v(n)$ denotes additive white Gaussian noise (AWGN) with variances $\sigma_v^2$. Eq. (1) 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

$$Y_k(i) = \sum_{j=1}^{J} H_{j,k}(i) X_{j,k}(i) + V_k(i), \tag{2}$$

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 circularly symmetric additive complex Gaussian noise, i.i.d across frequency and across time, and independent of $H$ and $X$; $X_{j,k}(i)$ is the signal of the $j$th PU in the frequency domain, and $H_{j,k}(i)$ is its frequency domain channel. We further assume that the $J$ PUs employ an orthogonal access to the spectrum (e.g., OFDMA) so that $X_{j,k}(i)X_{g,k}(i) = 0, \forall j \neq g$. Thus, letting $j_k$ be the index of the PU that contributes to the signal in the $k$th spectrum band, and letting $H_k(i) = H_{j_k,k}(i)$ and $X_k(i) = X_{j_k,k}(i)$ (with $X_k(i) = 0$ if no PU is transmitting in the $k$th spectrum band at time $i$), we can rewrite (2) as

$$Y_k(i) = H_k(i) X_k(i) + V_k(i). \tag{3}$$

We model $H_k(i)$ as a zero-mean circularly symmetric complex Gaussian random variable with variance $\sigma_H^2$, $H_k \sim \mathcal{CN}(0, \sigma_H^2)$, i.i.d. across frequency bands, over time, and independent of the occupancy state of the channels.

**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

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

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 $k$th 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

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

[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.

PUs join and leave the spectrum at random times. To capture this temporal correlation in the spectrum occupancy dynamics of PUs, we model $\vec{B}(i)$ as a Markov process,

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

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 spectrum bands,

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

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. 1. We structure the correlation models as two Markov chains: one across time and the other across frequencies, where the chain across frequencies is parameterized by $p = \mathbb{P}(B_k(i+1) = 1 | B_{k-1}(i+1) = 0)$ and the chain across time is parameterized by $q = \mathbb{P}(B_k(i+1) = 1 | B_k(i) = 0)$. We estimate these parameters $p$ and $q$, parameterizing each of these two chains, using the parameter estimator algorithm described in Sec. III in order to obtain the transition model underlying the MDP, given by (7).

**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, the SU can sense only $\kappa$ 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 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 (3). 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 an HMM. Given $\vec{B}(i)$ and $\mathcal{K}_i$, the probability density function of $\vec{Y}(i)$ is

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

owing to the independence of channels, noise, and transmitted symbols across frequency bands. Moreover, from (3),

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

**POMDP Agent Model:** In this section, we model the spectrum access scheme of the SU as a POMDP, whose goal 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 (7), is denoted by $\mathbf{A}$ and is learned by the agent by interacting with the radio environment (see Sec. III). The emission model is denoted by $\mathbf{M}$ and is given by (9), with $f(Y_k(i)|B_k(i))$ given by (10).

We model the POMDP as a tuple $(\mathcal{B}, \mathcal{A}, \mathcal{Y}, \mathbf{A}, \mathbf{M})$ 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 (5); $\mathcal{A}$ represents the action space of the agent, given by all possible combinations in which the $\kappa$ 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 before. The state of the POMDP at time $i$ is given by the *prior belief* $\beta_i$, 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 $\beta_i$, the agent selects $\kappa$ 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

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

We denote the function that maps the prior belief $\beta_i$ to the posterior belief $\hat{\beta}_i$ through the spectrum sensing action $\mathcal{K}_i$ and the observation signal $\vec{Y}(i)$ as $\hat{\beta}_i = \hat{\mathbb{B}}(\beta_i, \mathcal{K}_i, \vec{Y}(i))$.

Given the posterior belief $\hat{\beta}_i$, we estimate the occupancy state of the discretized spectrum under consideration as

$$\vec{B}(i)^* = \arg\max_{\vec{B} \in \mathcal{B}} \hat{\beta}_i(\vec{B}). \quad (12)$$

Let $B_k(i)^* = \phi_k(\hat{\beta}_i) \in \{0,1\}$ be the estimated state of channel $k$ at time $i$. If the channel is deemed to be idle as a result of this MAP estimation procedure, i.e., $\phi_k(\hat{\beta}_i) = 0$, the SU accesses the channel for delivering its network flows. Else, it leaves it untouched. 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 objective and a penalty for *missed detections* accounting for the incumbent non-interference constraint, i.e.,

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

where $\lambda > 0$ represents a penalty factor. After performing data transmission, the SU computes the prior belief for the next slot based on the dynamics of the Markov chain as

$$\beta_{i+1}(\vec{B}') = \mathbb{P}(\vec{B}(i+1) = \vec{B}' | \hat{\beta}_i). \quad (13)$$

We denote the function that maps the posterior belief $\hat{\beta}_i$ to the prior belief $\beta_{i+1}$ as $\beta_{i+1} = \mathbb{B}(\hat{\beta}_i)$. The goal of the problem at hand is to determine an optimal spectrum sensing policy to maximize the infinite-horizon discounted reward,

$$\pi^* = \arg\max_{\pi} V^{\pi}(\beta) \triangleq \mathbb{E}_{\pi}\Big[\sum_{i=1}^{\infty} \gamma^i R(\vec{B}(i), \hat{\beta}_i) | \beta_0 = \beta\Big], \quad (14)$$

where $0 < \gamma < 1$ is the discount factor, $\beta_0$ is the initial belief, and $\hat{\beta}_i$ is the posterior belief induced by policy $\mathcal{K}_i = \pi(\beta_i)$ and the observation $\vec{Y}(i)$ via $\hat{\beta}_i = \hat{\mathbb{B}}(\beta_i, \mathcal{K}_i, \vec{Y}(i))$, and we have defined the value function $V^{\pi}(\beta)$ under policy $\pi$ starting from belief $\beta$. The optimal policy $\pi^*$ and the corresponding optimal reward $V^*(\beta)$ are the solutions of Bellman's optimality equation $V^* = H[V^*]$, where the operator $V_{n+1} = H[V_n]$ is defined as

$$V_{n+1}(\beta) = \max_{\mathcal{K} \in \mathcal{A}} \sum_{\vec{B} \in \mathcal{B}} \beta(\vec{B}) \mathbb{E}_{\vec{Y}|\vec{B},\mathcal{K}} \Big[ R(\vec{B}, \hat{\mathbb{B}}(\beta, \mathcal{K}, \vec{Y})) + \gamma V_n(\mathbb{B}(\hat{\mathbb{B}}(\beta, \mathcal{K}, \vec{Y}))) \Big], \ \forall \beta. \quad (15)$$

This problem can be solved using the value iteration algorithm, i.e., by solving (15) 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 (15) using Exact Value Iteration and Policy Iteration algorithms is computationally infeasible. Additionally, solving for the optimal policy from equation (15) 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 simultaneously utilizing this learned model to solve for the optimal policy by employing randomized point-based value iteration techniques, namely, the PERSEUS algorithm [14].

## III. Approaches and Algorithms

**Occupancy Behavior Transition Model Estimation:** In real-world implementations of cognitive radio systems, the transition model of the occupancy behavior of the PUs is unknown 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 which is solving for the optimal spectrum sensing and access policy simultaneously. Inherently, the approach constitutes solving the Maximum Likelihood Estimation (MLE) problem

$$\vec{\theta}^* = \arg\max_{\vec{\theta}} \mathbb{P}([\vec{Y}(i)]_{i=1}^{\tau} | \vec{\theta}), \quad (16)$$

where $\vec{\theta} = [p \ q]^T$ and $\tau$ refers to the learning period of the parameter estimator: this can be equal to the entire duration of the POMDP agent's interaction with the radio environment implying simultaneous model learning 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 $[\vec{B}(i)]_{i=1}^{\tau}$ as $\mathbf{B}$. Re-framing (16) as an optimization of the log-likelihood, we get,

$$\vec{\theta}^* = \arg\max_{\vec{\theta}} \log\Big(\sum_{\mathbf{B}} \mathbb{P}(\mathbf{B}, \mathbf{Y} | \vec{\theta})\Big). \quad (17)$$

This problem can be solved using the Expectation-Maximization (EM) algorithm [15], where the E-step constitutes

$$Q(\vec{\theta} | \hat{\vec{\theta}}^{(t)}) = \mathbb{E}_{\mathbf{B}|\mathbf{Y}, \hat{\vec{\theta}}^{(t)}} \Big[ \log\Big(\sum_{\mathbf{B}} \mathbb{P}(\mathbf{B}, \mathbf{Y} | \hat{\vec{\theta}}^{(t)})\Big)\Big], \quad (18)$$

which can be obtained by employing the Forward-Backward algorithm using the current estimate of $\vec{\theta}$, i.e., $\vec{\theta}^{(t)}$ [15], and the M-step constitutes

$$\hat{\vec{\theta}}^{(t+1)} = \arg\max_{\vec{\theta}} Q(\vec{\theta}|\hat{\vec{\theta}}^{(t)}), \tag{19}$$

which involves re-estimation of the maximum likelihood parameters in $\vec{\theta}$ using the statistics obtained from the Forward-Backward algorithm.

**The PERSEUS Algorithm:** We solve for the optimal spectrum sensing and access policy, formulated as a POMDP, in parallel with the parameter estimation algorithm, employing the model estimates until the EM algorithm converges; after which, we utilize this converged transition model until the POMDP value iteration algorithm converges. As discussed in Sec. II of this article, solving the Bellman equation (15) for POMDPs with large state and action spaces using exact value iteration and policy iteration techniques is computationally infeasible [14]. 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. One such technique, the PERSEUS algorithm [14] is a randomized point-based approximate value iteration method that involves an initial phase of determining a set of so-called *reachable beliefs* $\tilde{\mathcal{B}}$ by allowing the agent to randomly interact with the radio environment. The goal of the PERSEUS algorithm is to improve the value of all the belief points in this set $\tilde{\mathcal{B}}$ by updating the value of only a subset of these belief points, chosen iteratively at random. Using the notion that, for infinite-horizon POMDPs, $V^*$ in (15) can be approximated by a Piece-Wise Linear and Convex function (PWLC) [14], the PERSEUS algorithm operates on the core idea that the value function at time index $i$ can be parameterized by a set of hyperplanes $\{\vec{\alpha}_i^u\}$, $u\in\{1,2,\ldots,|\tilde{\mathcal{B}}|\}$, each of which represents a region of the belief space for which it is the maximizing element. The belief points in $\tilde{\mathcal{B}}$ are to be improved over numerous iterative *backup* stages. The optimal hyperplane $\vec{\alpha}_{i+1}^u$ at time index $i+1$ for the $u$th belief $\beta_u \in \tilde{\mathcal{B}}$ can be iteratively computed as, from [14],

$$\vec{\alpha}_{i+1}^u = \text{backup}(\beta_u) = \arg\max_{\mathcal{K}\in\mathcal{A}} \beta_u \cdot \Xi_{\mathcal{K}}^u, \tag{20}$$

where $\beta \cdot \Xi = \sum_{\vec{B}} \beta(\vec{B})\Xi(\vec{B})$ denotes inner product. $\Xi_{\mathcal{K}}^u$ is the hyperplane corresponding to a one-step look-ahead of the value iteration updates under action $\mathcal{K}$ and belief $\beta_u$, given by

$$\Xi_{\mathcal{K}}^u = \sum_{\vec{B}\in\mathcal{B}} \beta_u(\vec{B})\mathbb{E}_{\vec{Y}|\vec{B},\mathcal{K}}\Big[R(\vec{B},\hat{\mathbb{B}}(\beta_u,\mathcal{K},\vec{Y})) + \gamma\Xi_{\mathcal{K},\vec{Y}}^u(\vec{B})\Big],$$

where $\Xi_{\mathcal{K},\vec{Y}}^u$ is the hyperplane associated with the future value function, computed from the previous set of hyperplanes as

$$\Xi_{\mathcal{K},\vec{Y}}^u = \arg\max_{u'\in\{1,2,\ldots,|\tilde{\mathcal{B}}|\}} \mathbb{B}(\hat{\mathbb{B}}(\beta_u,\mathcal{K},\vec{Y}))\cdot\alpha_i^{u'}.$$

Once these hyperplanes have been computed, the new value function at a generic belief $\beta$ and the corresponding policy can be computed as

$$V_{i+1}(\beta) = \max_{u\in\{1,2,\ldots,|\tilde{\mathcal{B}}|\}} \beta\cdot\vec{\alpha}_{i+1}^u,\ \pi_{i+1}(\beta) = a(\vec{\alpha}_{i+1}^*), \tag{21}$$

where $a(\vec{\alpha}_{i+1}^*)$ is the action corresponding to the maximizing hyperplane $\vec{\alpha}_{i+1}^*$. In each backup stage, the agent samples a belief $\beta$ uniformly at random from the set of unimproved points and performs a backup on this sampled belief point according to (20), to determine the optimal hyperplane. Considering an arbitrary time index $i+1$, if $V_{i+1}(\beta)=\beta\cdot\vec{\alpha}\geq V_i(\beta)$, then the belief point $\beta$ is said to be improved along with any other belief points $\beta'$ 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 $\beta$ 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 the number of policy changes between consecutive iterations is below a specified threshold $\eta$. The belief update procedure outlined in (11) 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 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.

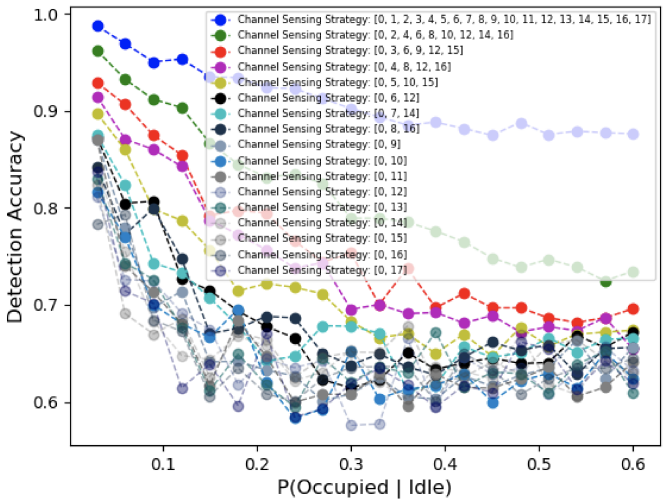


Fig. 2. The detection accuracies of the constrained Viterbi algorithm for different sensing strategies over varying values of $p$**[NM: we probably dont need this figure and we may remove it]**

## IV. Numerical Evaluations

**[NM: You need to make sure to list all parameters of your simulation. For instance, how many channels are you sensing?]** The given framework is simulated using *Python* for a system with 18 channels and a channel model constituting an SNR of 19dB when an incumbent occupies a specific channel. To begin with, we model a Viterbi algorithm described by $\vec{B}(i)=\arg\max_{\vec{B}\in\mathcal{B}} \mathbb{P}(\vec{B}(i)=\vec{B}|\vec{Y}(i))$ for the HMM PU spectrum occupancy model outlined in Sec. II, i.e. a MAP-based state estimator leveraging the correlations across both time and frequencies, functioning under constraints imposed upon the number of channels that can be observed by the

SU at any given time step $i$. We design and develop the Viterbi algorithm in order to demonstrate two conclusions that are carried forward into the PERSEUS framework: exploiting correlations in incumbent occupancy behavior across time and frequencies helps improve the utility obtained by the SU as defined by $R(\vec{B}(i), \hat{\beta}_i)$, while incomplete channel occupancy information leads to a reduction in this utility.

The first line trace (in blue) depicted in Fig. 2 illustrates the detection accuracies 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 $p$, defined in Sec. II, i.e., as the channels transition toward independence. We note that the detection accuracy of this MAP-based 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. This plot is an important illustration to prove that leveraging correlation in incumbent occupancy behavior across frequencies gives us a boost in detection accuracy, and thereby, a boost in secondary network throughput while complying with the required non-interference constraints. Additionally, Fig. 2 also illustrates the detection accuracies of the Viterbi algorithm (refer to all the other line traces) wherein the SU makes observations of only the channels in the given channel sensing strategy and estimates the occupancy states of all the channels in the discretized spectrum of interest over varying values of $p$. We observe from Fig. 2 that, as anticipated, the detection accuracy of this estimator deteriorates as the amount of information available for estimation decreases. In other words, the imposition of sensing limitations on the SU's spectrum sensor in view of quick turnaround times and improved energy efficiency, will put a dent in the detection accuracy of the cognitive radio node, and this needs to be mitigated or offset by leveraging the correlation in incumbent occupancy behavior across frequencies.

The plot depicted in Fig. 3 shows the mean square error convergence of the parameter estimation algorithm while determining the transition model of the Markov chain across frequencies, which as detailed in Sec. II, is parameterized by $p$. Starting with an initial estimate of $10^{-8}$, the EM algorithm detailed in Sec. III converges to the true transition model with an error of $\epsilon \leq 10^{-8}$ over numerous iterations, each iteration corresponding to an averaging operation constituting 300 observation vectors. We observe the mean square error given by $\mathbb{E}[(p-\hat{p})^2]$ iteratively reduces as it goes through the E-step and the M-step. It has been theoretically shown to converge, i.e., each iteration either improves the true likelihood or leaves it unchanged [16]. 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. 4 illustrates the *Regret* convergence plot of the PERSEUS algorithm over several backup stages wherein the regret metric corresponds to the difference in utility $R(\vec{B}(i), \hat{\beta}_i)$ obtained by the PERSEUS algorithm at a certain stage and an *Oracle* which has complete information about the occupancy behavior of incumbents in the network. Furthermore, the algorithm involves an online estimation of

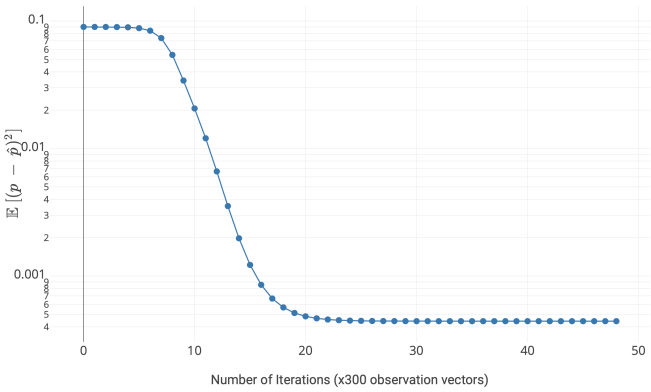


Fig. 3. Mean square error convergence of the parameter estimation algorithm while determining the correlation model across frequencies, specifically, $p$**[NM: what about $q$?] [NM: this figure and the next one can be combined into a single figure, maybe two subpolots?]**

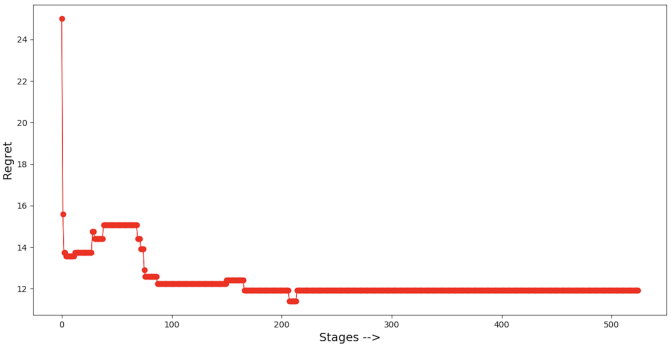


Fig. 4. Regret convergence of PERSEUS over numerous backup stages**[NM: the fontsize in your figures is too small. Can you please fix? The font size should be comparable with the caption (slightly smaller is fine)]**

the transition model of the underlying MDP and a random exploration strategy to gather the initial set of reachable beliefs $\tilde{\mathcal{B}}$. The termination condition for the PERSEUS algorithm is that the number of policy changes, denoted by $\eta$, over several consecutive backup stages should be 0. This plot, similar to the *Reward v/s Time* plot in [14], serves as a measure of convergence for our fragmented PERSEUS algorithm with simplified belief updates and an online transition model estimation.

The proposed fragmented PERSEUS algorithm with belief simplification, random exploration, and online model learning is evaluated against the Minimum Entropy Merging (MEM) algorithm with Markov Process Estimation (MPE)

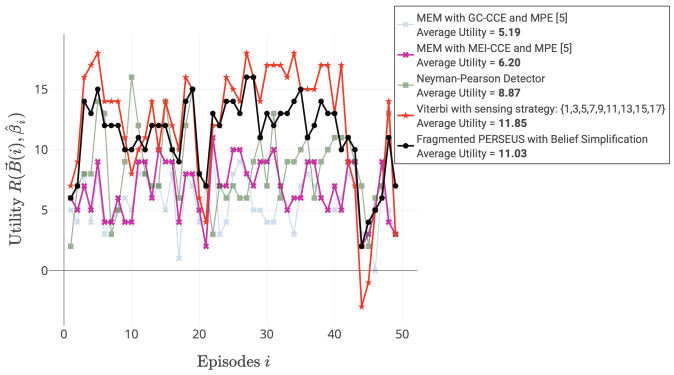


Fig. 5. The evaluation of the proposed framework against a medley of approaches in the state-of-the-art **[NM: I think we should remove this figure]**

for exploiting first-order Markovian correlation across time, and both Channel Correlation Estimation (CCE) techniques: Greedy Clustering (GC) with a correlation threshold $\rho_{th}$=0.77 and Minimum Entropy Increment (MEI) Clustering with the number of clusters metric of 6, for exploiting correlation across frequencies, from [5]; a Neyman-Pearson detector, observing all 18 channels at any given time index, and assuming independence in channel occupancy behavior across both time and frequencies; and a Viterbi algorithm that is restricted to only observe channels whose indices correspond to odd integers, that has the occupancy correlation models known apriori, and that exploits Markovian correlations across both time and frequencies. With a penalty of 1, i.e., $\lambda$=1 in $R(\vec{B}(i), \hat{\beta}_i)$, we observe from Fig. 5, that the proposed framework achieves an average utility per time step (or episode) of 11.03, which is almost 80% more than that achieved by the MEM with MEI-CCE and MPE procedure from [5], and which is almost 25% more than that achieved by the Neyman-Pearson detector scheme described earlier. Furthermore, the proposed framework, on average, matches the utility obtained by the aforementioned Viterbi algorithm.

We evaluate the performance of the proposed framework in terms of the SU network throughput and PU interference metrics over varying values of the penalty term $\lambda$ as illustrated in Fig. 6. As expected, we find that our POMDP agent decides to limit channel access when the penalty is high, leading to lower SU network throughput and lower PU interference; and on the other hand, it follows a more lenient channel access strategy when the penalty is low, resulting in higher SU network throughput and higher PU interference. Here, SU network throughput is defined as the number of truly idle channels found by our POMDP agent, and PU interference is defined as the number of channels in which the SU's transmissions collide with a licensed user. In general, we observe the trend of rising throughput and increasing interference as the penalty for missed detections $\lambda$ is lowered. Comparing this performance of our proposed framework with correlation-coefficient based state-of-the-art, namely the MEM with MEI-CCE and MPE algorithm with $\rho_{th}$=0.77 and 6 specified clusters, from [5], we find that our framework achieves higher SU network throughput and lower PU interference with $\lambda \geq 10$. Furthermore, with a penalty of 0 **[NM: what do you mean with a penalty of 0?]** the proposed framework comes very close to achieving the throughput attained by a constrained Viterbi agent while providing the same interference performance. It is worth noting that the Viterbi agent possesses prior knowledge about the transition model of the underlying MDP and senses more channels per time-step than our POMDP agent. Most importantly, the proposed framework allows to regulate the trade-off between the interference caused to PUs and the throughput of the SU, by adjusting the paraneter $\lambda$.

## 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 POMDP. 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 EM 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 conclude that our framework, in terms of the average utility obtained per time-step $i$, outperforms the existing correlation-coefficient based state-of-the-art; surpasses Neyman-Pearson detection schemes that fail to exploit the correlations across time and frequencies; and achieves the performance attained by standard MAP-estimators which possess the transition model statistics as an apriori.

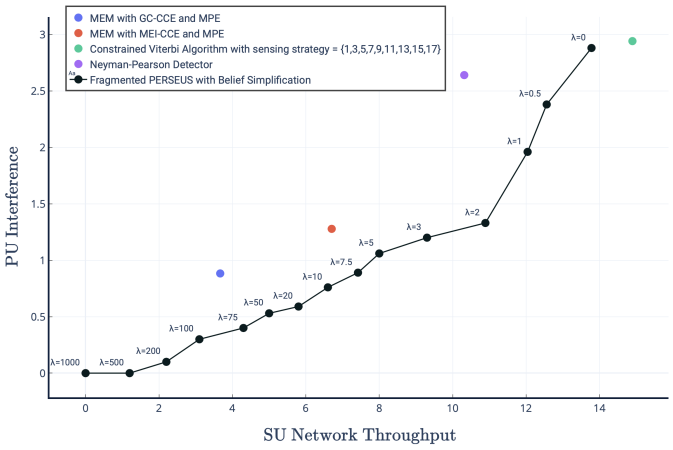


Fig. 6. SU Network Throughput versus PU Interference evaluation of the proposed framework over varying values of the penalty $\lambda$. **[NM: what do you mean by "constrained Viterbi algorithm"? In this figure, you need to have also PERSEUS with perfect knowledge of the transition model. Can you add that in?] [NM: what is the unit of the x axis? IS it Mbps, or what? You need to specify that.] [NM: Same for the y axis.. what is "interference to PU"? IT might be more helpful if you provide the interference to noise ratio at the PU..]**

## 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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[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] 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

[15] W. Turin, "Map decoding using the em algorithm," in *1999 IEEE 49th Vehicular Technology Conference (Cat. No.99CH36363)*, vol. 3, May 1999, pp. 1866–1870 vol.3.

[16] 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