Spectrum Sensing in Cognitive Radio Networks via Approximate POMDP

Bharath Keshavamurthy, Nicolò Michelusi

***Abstract*—In this letter, a novel spectrum sensing and access strategy based on POMDPs is proposed. A cognitive radio learns the correlation model defining the occupancy behavior of incumbents, based on which it devises an optimal sensing and access policy. The optimization complexity is ameliorated via point-based value iteration methods. Numerical evaluations demonstrate that the proposed framework achieves higher throughput for cognitive radios, with lower interference to incumbents, compared to state-of-the-art clustering algorithms and a Neyman-Pearson detector that does not exploit correlation among channels. Furthermore, it achieves performance comparable to HMM MAP estimators with a priori knowledge of the correlation model.**

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

I. Introduction

The advent of fifth-generation wireless communication networks has exacerbated the problem of spectrum scarcity [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 incumbent non-interference compliance [2].

A crucial aspect underlying the design of cognitive radio networks is the ability to perform spectrum sensing. 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 to determine the best channels to sense before making channel access decisions. The state-of-the-art are based on multi-armed bandits [4], reinforcement learning [5], and custom heuristics [6], [7]. However, most of these works, such as [4], [5], [8]–[11], assume independence across frequency bands, which is imprudent because licensed users may exhibit correlations across both frequency and time in their channel occupancy behavior: they may occupy a set of adjacent frequency bands, and for an extended period of time [12]. This correlation structure may be leveraged for more accurate predictions of spectrum holes. In this letter, we propose a parameter estimation algorithm to learn the time-frequency correlation structure, based on which we solve for the optimal sensing and access policy via partially observable Markov decision processes (POMDPs).

Related Work: Distributed spectrum sensing has been considered in [5] 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 [7], the frequency correlation is exploited, but the observation model is noise-free. Compared to [5], [7], we account for the uncertainty in the occupancy state and for noisy observations via a POMDP formulation. Standard MAP-based state estimators for Hidden Markov Models (HMMs) such as the Viterbi algorithm can be employed to estimate spectrum occupancy [6]; however, these estimators rely on knowledge of the transition model, which may be unknown in practice. To this end, in [6], [7], the time-frequency correlation structure is estimated offline based on pre-loaded databases. However, this solution may be impractical in non-stationary settings. To fill this gap, in this letter, we develop a fully online framework to learn the parametric time-frequency correlation model, and embed it directly into the sensing and access optimization of the decision engine. It is also worth noting that [6] does not impose sensing restrictions on the cognitive radio, as we do. Finally, unlike [4]–[7], our POMDP framework allows to regulate the trade-off between the throughput achieved by the cognitive radio and the interference to the incumbents.

In a nutshell, the contributions of this paper are as follows: we develop a POMDP framework for spectrum sensing and access in a radio environment with multiple licensed users exhibiting correlations in their occupancy behavior across both time and frequency, assuming a linear, Gaussian observation model with sensing limitations; we develop an online parameter estimation algorithm to learn the incumbents' occupancy correlation model; concurrently, we propose a randomized point-based value iteration algorithm to find the optimal spectrum sensing and access policy; finally, we compare numerically the proposed framework with state-of-the-art algorithms, and demonstrate superior performance. The rest of the paper is organized as follows: in Sec. II, we define the system 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

Signal Model: We consider a network of J incumbents (Primary Users, PUs) and one cognitive radio (Secondary User, SU) equipped with a spectrum sensor. The goal of the SU is to maximize its own throughput by opportunistically accessing portions of the spectrum left unused by the PUs. To this end, the SU should learn how to intelligently access spectrum holes, while ensuring nominal interference with incumbent transmissions. We express the discretized wideband frequency domain signal received at the SU at time index i as

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

where $k \in \{1, 2, \ldots, K\}$ is the frequency domain channel index; $X_{j,k}(i)$ is the signal of the jth PU in the frequency domain, and $H_{j,k}(i)$ is the frequency domain channel between the jth PU and the SU; $V_k(i) \sim \mathcal{CN}(0, \sigma_V^2)$ represents circularly symmetric additive complex Gaussian noise, i.i.d across frequency and across time, and independent of channel H and PU signal X. We further assume that the J PUs employ an orthogonal spectrum access (e.g., OFDMA) so that $X_{j,k}(i) X_{g,k}(i) = 0, \forall j \neq g$. Thus, letting $j_{k,i}$ be the index of the

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.

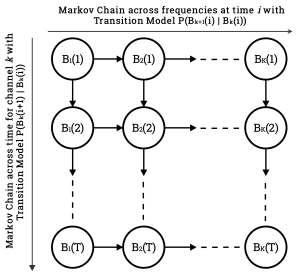

Fig. 1. The correlation model across time and frequencies underlying the occupancy behavior of incumbents in the network

PU that occupies the kth spectrum band at time i, and letting $H_k(i)=H_{j_{k,i},k}(i)$ and $X_k(i)=X_{j_{k,i},k}(i)$ (with $X_k(i)=0$, if no PU is transmitting in the kth spectrum band at time i), we can rewrite (1) as

$$Y_k(i) = H_k(i)X_k(i) + V_k(i). \tag{2}$$

We model the channel as Rayleigh fading, so that $H_k(i)$ is 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 and time.

PU Spectrum Occupancy Model: We model $X_k(i)$ as

$$X_k(i) = \sqrt{P_{tx}} B_k(i) S_k(i), \tag{3}$$

where P_{tx} is the transmission power of the PUs, $S_k(i)$ is the transmitted symbol modeled 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 state in the entire wideband spectrum of interest at time i can be modeled as the vector

$$\vec{B}(i) = [B_1(i), B_2(i), B_3(i), \cdots, B_K(i)]^T \in \{0,1\}^K. \tag{4}$$

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

$$\mathbb{P}(\vec{B}(i+1)|\vec{B}(j), \forall j \leq i) = \mathbb{P}(\vec{B}(i+1)|\vec{B}(i)). \tag{5}$$

Additionally, when joining the spectrum pool, the PUs occupy a number of adjacent spectrum bands, and may vary their spectrum needs over time depending on time-varying traffic demands, channel conditions, etc. To capture this behavior, we model $\vec{B}(i)$ as a Markov chain across spectrum bands, i.e., 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, $B_k(i)$. This structure is depicted in Fig. 1, and is described as

$$\begin{aligned} &\mathbb{P}(\vec{B}(i+1)|\vec{B}(i)) \\ &= \mathbb{P}(B_1(i+1)|B_1(i)) \prod_{k=2}^{K} \mathbb{P}(B_k(i+1)|B_k(i), B_{k-1}(i+1)). \end{aligned} \tag{6}$$

Overall, the correlation model is expressed by two Markov chains: one across time and the other across frequencies. We parameterize this double Markov chain structure by the vector $\vec{\theta}=[\vec{p} \quad \vec{q}]^\intercal$, where $\vec{p}=[p_{uv}=\mathbb{P}(B_k(i+1)=1|B_{k-1}(i+1)=u, B_k(i)=v): u,v \in \{0,1\}]^\intercal$ and $\vec{q}=[q_w=\mathbb{P}(B_1(i+1)=1|B_1(i)=w): w \in \{0,1\}]^\intercal$. We estimate this parameter vector $\vec{\theta}$, using the parameter estimation algorithm described in Sec. III in order to obtain the transition model underlying the MDP, given by (6).

Spectrum Sensing Model: In order to detect the available spectrum holes, the SU performs spectrum sensing. However, owing to physical design limitations of its spectrum sensor [3], it can sense at most 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 spectrum bands sensed by the SU at time i, part of our design. Then, we define the observation vector

$$\vec{Y}(i) = [Y_k(i)]_{k \in \mathcal{K}_i}, \tag{7}$$

with $Y_k(i)$ given by (2). The true states $\vec{B}(i)$ encapsulate the actual occupancy state of the PUs, and the measurements $\vec{Y}(i)$ 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

$$f(\vec{Y}(i)|\vec{B}(i), \mathcal{K}_i) = \prod_{k \in \mathcal{K}_i} f(Y_k(i)|B_k(i)), \tag{8}$$

owing to the independence of channels (given the occupancy states), noise, and transmitted symbols across frequency bands. Moreover, from (2),

$$Y_k(i)|B_k(i) \sim \mathcal{CN}(0, \sigma_H^2 P_{tx} B_k(i) + \sigma_V^2). \tag{9}$$

POMDP Agent Model: We model the SU as a POMDP agent, whose goal is to devise an optimal sensing and access policy in order to maximize its throughput, while ensuring minimal interference with incumbent transmissions. The agent's limited sensing capabilities coupled with its noisy observations result in uncertainty on the PU spectrum occupancy state and on the exact effect of executing a spectrum access action on the radio environment. The transition model of the underlying MDP, as described by (6), is denoted by $\mathbf{A}$ and is learned by the agent by interacting with the radio environment (see Sec. III). The observation model is denoted by $\mathbf{M}$ and is given by (8), with $f(Y_k(i)|B_k(i))$ given by (9).

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, given by all possible realizations of the spectrum occupancy vector $\vec{B}$, as described by (4); $\mathcal{A}$ represents the action space of the agent, given by all possible combinations in which 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 aforementioned signal model. 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 time-step 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

$$\hat{\beta}_i(\vec{B}') = \mathbb{P}(\vec{B}(i) = \vec{B}' | \vec{Y}(i), \mathcal{K}_i, \beta_i)$$

[1] If $S_k(i)$ 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 modifications to the subsequent analysis.

$$= \frac{\mathbb{P}(\vec{Y}(i)|\vec{B}', \mathcal{K}_i)\beta_i(\vec{B}')}{\sum_{\vec{B}''\in\{0,1\}^K} \mathbb{P}(\vec{Y}(i)|\vec{B}'', \mathcal{K}_i)\beta_i(\vec{B}'')}. \quad (10)$$

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 $\hat{\beta}_i = \hat{\mathbb{B}}(\beta_i, \mathcal{K}_i, \vec{Y}(i))$.

Given the posterior belief $\hat{\beta}_i$, we estimate $\vec{B}(i)$ as

$$\vec{\phi}(\hat{\beta}_i) \triangleq \arg\max_{\vec{B}\in\mathcal{B}} \hat{\beta}_i(\vec{B}),$$

with the kth spectrum channel estimate given by $\phi_k(\hat{\beta}_i)$. If the kth 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 to deliver 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 interference to the incumbents,

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

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

$$\beta_{i+1}(\vec{B}'') = \sum_{\vec{B}'} \mathbb{P}(\vec{B}(i+1) = \vec{B}''|\vec{B}(i) = \vec{B}')\hat{\beta}_i(\vec{B}'). \quad (11)$$

We denote the function that maps the posterior belief $\hat{\beta}_i$ to the prior belief β_{i+1} as $\beta_{i+1} = \mathbb{B}(\hat{\beta}_i)$. The goal is to determine an optimal spectrum sensing policy to maximize the infinite-horizon discounted reward, with discount factor $0<\gamma<1$,

$$\pi^* = \arg\max_{\pi} V^{\pi}(\beta) \triangleq \mathbb{E}_{\pi}\Big[\sum_{i=1}^{\infty} \gamma^i R(\vec{B}(i), \hat{\beta}_i)|\beta_0 = \beta\Big], \quad (12)$$

where β_0 is the initial belief, $\hat{\beta}_i$ 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 π starting from belief β. The optimal policy π^* and the corresponding optimal reward $V^*(\beta)$ are the solutions of Bellman's optimality equation $V^* = \mathcal{H}[V^*]$, where the operator $V_{t+1} = \mathcal{H}[V_t]$ is defined as

$$V_{t+1}(\beta) = \max_{\mathcal{K}\in\mathcal{A}} \sum_{\vec{B}\in\mathcal{B}} \beta(\vec{B}) \mathbb{E}_{\vec{Y}|\vec{B},\mathcal{K}}\Big[R(\vec{B}, \hat{\mathbb{B}}(\beta, \mathcal{K}, \vec{Y})) + \gamma V_t(\mathbb{B}(\hat{\mathbb{B}}(\beta, \mathcal{K}, \vec{Y})))\Big], \ \forall \beta. \quad (13)$$

This problem can be solved using the value iteration algorithm, i.e., by solving (13) iteratively until convergence to a fixed point. However, two challenges arise in our formulation:

- The parameter vector $\vec{\theta}$ is unknown, hence the belief updates $\hat{\mathbb{B}}$ and $\mathbb{B}$ cannot be computed;
- The number of states of the underlying MDP scales exponentially with the number of spectrum bands, resulting in a high-dimensional belief space, hence, solving equation (13) exactly is computationally infeasible.

To overcome these challenges, in the next section, we present a framework to estimate the transition model of the underlying MDP online, while concurrently utilizing this learned model to solve for the optimal policy via PERSEUS, a low-complexity, randomized point-based value iteration algorithm [13].

III. Approaches and Algorithms

Occupancy Behavior Model Estimation: In real-world implementations of cognitive radios, the correlation model defining the occupancy behavior of the PUs is unknown to the SUs, and therefore 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

$$\vec{\theta}^* = \arg\max_{\vec{\theta}} \log\Big(\sum_{\mathbf{B}} \mathbb{P}(\mathbf{B}, \mathbf{Y}|\vec{\theta})\Big), \quad (14)$$

where $\vec{\theta} = [\vec{p}\ \vec{q}]^T$, $\mathbf{Y} = [\vec{Y}(i)]_{i=1}^{\tau}$, $\mathbf{B} = [\vec{B}(i)]_{i=1}^{\tau}$, and τ 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. This problem can be solved using the Expectation-Maximization (EM) algorithm [14], with the E-step given by

$$Q(\vec{\theta}|\vec{\theta}^{(t)}) = \mathbb{E}_{\mathbf{B}|\mathbf{Y},\vec{\theta}^{(t)}}\Big[\log\Big(\mathbb{P}(\mathbf{B}, \mathbf{Y}|\vec{\theta}^{(t)})\Big)\Big]. \quad (15)$$

The term $Q(\vec{\theta}|\vec{\theta}^{(t)})$ can be computed by employing the Forward-Backward algorithm [14] using the current estimate $\vec{\theta}^{(t)}$. The M-step constitutes

$$\vec{\theta}^{(t+1)} = \arg\max_{\vec{\theta}} Q(\vec{\theta}|\vec{\theta}^{(t)}), \quad (16)$$

which involves re-estimation of the maximum likelihood parameter vector $\vec{\theta}$ 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 both the EM and the POMDP algorithms converge. As discussed in Sec. II of this article, solving the Bellman equation (13) for POMDPs with large state and action spaces using exact value iteration is computationally infeasible [13]. Hence, we resort to approximate value iteration methods to ensure that the system scales well to a large number of bands in the spectrum of interest. One such method, the PERSEUS algorithm [13] is a randomized point-based approximate value iteration technique 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 PERSEUS 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 finite-horizon POMDPs, V^* in (13) can be approximated by a piece-wise linear and convex function [13], PERSEUS operates on the core idea that the value function at iteration t can be parameterized by a set of hyperplanes $\{\vec{\alpha}_t^u\}$, $u\in\{1, 2, \ldots, |\tilde{\mathcal{B}}|\}$, each representing a region of the belief space for which it is the maximizing

element, and each associated with an optimal spectrum sensing action $\mathcal{K}_t^u$. That is, when operating with belief β, the value function is approximated as

$$V_t(\beta) \approx \beta \cdot \vec{\alpha}_t^{u^*},\ u^* = \underset{u\in\{1,2,\ldots,|\tilde{\mathcal{B}}|\}}{\arg\max}\ \beta \cdot \vec{\alpha}_t^u, \tag{17}$$

and the optimal spectrum sensing action is $\mathcal{K}_t^{u^*}$, where $\beta \cdot \vec{\alpha} = \sum_{\vec{B}} \beta(\vec{B})\vec{\alpha}(\vec{B})$ denotes inner product. The set of hyperplanes $\{\vec{\alpha}_t^u\}$ associated with the set $\tilde{\mathcal{B}}$ are improved over multiple iterations of PERSEUS: given $\{\vec{\alpha}_t^u\}$ and the optimal spectrum sensing actions $\{\mathcal{K}_t^u\}$ at iteration t, a PERSEUS iteration generates a new set of hyperplanes $\{\vec{\alpha}_{t+1}^u\}$ and associated spectrum sensing actions $\{\mathcal{K}_{t+1}^u\}$, as we now describe. Let $\tilde{\mathcal{U}}$ be the set of unimproved belief points (initially, $\tilde{\mathcal{U}}=\tilde{\mathcal{B}}$). Then, a belief β_u is picked randomly from $\tilde{\mathcal{U}}$. Next, a *backup* operation is performed on β_u to determine a new associated hyperplane and spectrum sensing action as in [13],

$$\vec{\alpha}_{t+1}^u = \Xi_{\mathcal{K}_{t+1}^u}^u,\ \mathcal{K}_{t+1}^u = \underset{\mathcal{K}\in\mathcal{A}}{\arg\max}\, \beta_u \cdot \Xi_{\mathcal{K}}^u, \tag{18}$$

where $\Xi_{\mathcal{K}}^u$ is the hyperplane corresponding to a one-step look-ahead under action $\mathcal{K}$ and belief β_u, given by

$$\Xi_{\mathcal{K}}^u(\vec{B}) = \mathbb{E}_{\vec{Y}|\vec{B},\mathcal{K}}\Big[R(\vec{B},\hat{\mathbb{B}}(\beta_u,\mathcal{K},\vec{Y})) + \gamma \sum_{\vec{B}'} \mathbb{P}(\vec{B}(i+1)=\vec{B}'|\vec{B}(i)=\vec{B})\Xi_{\mathcal{K},\vec{Y}}^u(\vec{B}')\Big]$$

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

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

under the new belief $\mathbb{B}(\hat{\mathbb{B}}(\beta_u,\mathcal{K},\vec{Y}))$ reached from β_u, under action $\mathcal{K}$ and observation $\vec{Y}$, in one step. At this point, $\beta_u\cdot\vec{\alpha}_{t+1}^u$ is the approximate value function associated with the belief β_u. If $\beta_u\cdot\vec{\alpha}_{t+1}^u \geq V_t(\beta_u)$, the newly defined hyperplane generates an improved value function; otherwise ($\beta_u\cdot\vec{\alpha}_{t+1}^u < V_t(\beta_u)$), the value function is worsened and the previous hyperplane is kept, hence $\vec{\alpha}_{t+1}^u := \vec{\alpha}_t^u$ and $\mathcal{K}_{t+1}^u := \mathcal{K}_t^u$. Finally, the belief β_u is removed from $\tilde{\mathcal{U}}$, along with all belief points that are improved by the newly added hyperplane:

$$\tilde{\mathcal{U}} \leftarrow \tilde{\mathcal{U}} \setminus \{\beta_u\} \setminus \{\beta' \in \tilde{\mathcal{U}} : \beta'\cdot\vec{\alpha}_{t+1}^u \geq V_t(\beta')\}.$$

This operation continues until the set $\tilde{\mathcal{U}}$ is empty. The PERSEUS iterations continue until a termination condition is met, i.e., $|V_{t+1}(\beta)-V_t(\beta)|<\epsilon,\ \forall\beta\in\tilde{\mathcal{B}},\ \epsilon>0$. The belief update procedure outlined in (10) is an essential aspect of POMDPs, 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 employ a *fragmentation heuristic*, i.e., we fragment the spectrum into 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 by employing a *belief simplification heuristic*, i.e., allow only those state transitions we deem to be the most probable – for example, only those state transitions that involve a Hamming distance of up to 3 between the previous and current state vectors in a radio environment with 18 frequency channels.

IV. Numerical Evaluations

Simulation Setup: We simulate a radio environment with $J=3$ PUs occupying a set of $K=18$ channels, each of bandwidth BW=160kHz, according to a Markovian time-frequency correlation structure defined by the parameter vector $\vec{\theta}=[\vec{p}\ \vec{q}]^\intercal$, where $\vec{p}=[p_{00}=0.1\ p_{01}=0.3\ p_{10}=0.3\ p_{11}=0.7]^\intercal$ and $\vec{q}=[q_0=0.3\ q_1=0.8]^\intercal$, and a SU intelligently trying to access the available white spaces. We model the expected SINRs at the SU and PU receivers, averaged out with respect to fading and conditional on the SU and PU access decisions, as follows: at the PU receivers, SINR_{PU}=17dB when there is no interference from the SU, and SINR_{PU}=6dB under SU interference; at the SU receiver, SINR_{SU}=11dB when the channel is not occupied by an incumbent, and SINR_{SU}=−6dB under PU interference. We denote by $\text{SINR}_{\text{SU}}(k,i)$ and $\text{SINR}_{\text{PU}}(k,i)$ the actual SINR realizations in frequency k and time index i. We assume that the SU is always backlogged and that it accesses all the channels deemed idle by the optimal POMDP policy. With this notation, the average SU throughput over a time interval of duration T is given by

$$C^{\text{SU}} = \frac{1}{T}\sum_{i=1}^{T}\sum_{k=1}^{K} R_{\text{SU}}\cdot\mathcal{I}\left(\text{SINR}_{\text{SU}}(k,i) \geq 2^{R_{\text{SU}}/\text{BW}}-1\right),$$

where R_{SU}=0.6Mbps is the transmission rate of the SU on each frequency channel. Similarly, the throughput attained by the PUs over the same time interval, normalized over time and the number of transmissions, is given by

$$C^{\text{PUs}} = \frac{\sum_{i=1}^{T}\sum_{k=1}^{K} R_{\text{PU}}\cdot B_k(i)\mathcal{I}\left(\text{SINR}_{\text{PU}}(k,i)\geq 2^{R/\text{BW}}-1\right)}{\sum_{i=1}^{T}\sum_{k=1}^{K} B_k(i)},$$

where R_{PU}=0.9Mbps is the transmission rate of the PUs on each frequency channel. The SU can sense at most $\kappa=6$ out of $K=18$ channels in any given time-step i. Finally, while solving for the optimal policy in the PERSEUS algorithm, we employ a discount factor of $\gamma=0.9$ and a termination threshold of $\epsilon=10^{-5}$.

Evaluations: The plot depicted in Fig. 2 shows the mean squared error (MSE) convergence of the parameter estimation algorithm to determine the parametric time-frequency correlation structure $\vec{\theta}$, averaged over $60,000$ iterations. Assuming a time-slot duration of 3ms, this corresponds to an observation and estimation period of 180s. Starting with initial estimates of 10^{-8} for all $\theta\in\vec{\theta}$, the EM algorithm iteratively reduces the MSE, as it goes through the E and M steps, and converges to the true transition model with an error of 10^{-8}. Fig. 2 also illustrates the loss convergence plot of the PERSEUS algorithm, on the same time scale and in the same simulation run, wherein the loss metric corresponds to the difference in utility obtained by our PERSEUS algorithm at a certain time-step i in our simulation, denoted by $R_P(\vec{B}(i),\hat{\beta}_i)$, and an *Oracle* which knows the exact occupancy behavior of incumbents in the network, with utility $R_O(\vec{B}(i))$. This trace in Fig. 2, similar to the *Reward v/s Time* plot in [13], serves as a measure of convergence for our fragmented PERSEUS algorithm with belief simplification and online model estimation.

In Fig. 3, we evaluate the performance of the proposed framework in terms of SU and PU network throughputs over varying values of the penalty term λ. As surmised,

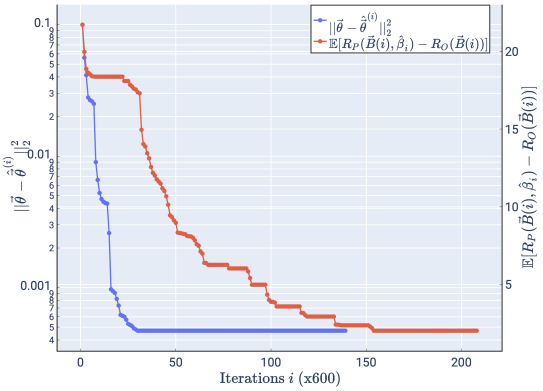

Fig. 2. Mean squared error convergence of the parameter estimation algorithm to determine the correlation model $\vec{\theta}$, and Loss convergence of the fragmented PERSEUS algorithm with belief simplification.

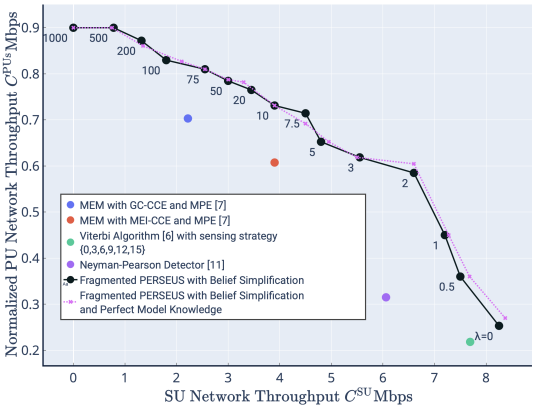

Fig. 3. SU Network Throughput versus PU Interference evaluation of the proposed framework over varying values of the penalty λ; comparison with state-of-the-art.

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. In general, we observe the trend of rising throughput and increasing interference as the penalty for missed detections λ is lowered. We also compare the performance of our proposed framework with state-of-the-art algorithms, namely: two separate versions of Minimum Entropy Merging (MEM) with Channel Correlation Estimation (CCE) and Markov Process Estimation (MPE) – one with Greedy Clustering (GC) and the other with Minimum Entropy Increment (MEI) Clustering [7], both with a correlation threshold of ρ_{th}=0.77 and a sensing restriction of 6; the Viterbi algorithm [6] with a priori model information and a sensing restriction of 6; a Neyman-Pearson Detector [11] assuming independence among channels across both frequency and time, no sensing restrictions, an AND fusion rule across 300 different samplings, and whose threshold is determined with respect to a false alarm probability of 0.3; and a variant of our PERSEUS framework in which the agent has perfect knowledge of the correlation model. We note that our proposed algorithm outperforms the state-of-the-art, by achieving a higher SU throughput for a certain PU throughput degradation: for the same level of PU throughput degradation, our framework achieves the following performance with respect to SU throughput: a 37.5% increase over the MEM with MEI-CCE and MPE algorithm, a 25% increase over the Neyman-Pearson Detector, and a performance comparable to that of the Viterbi algorithm. Additionally, our framework allows the SU to regulate the trade-off between its throughput and the interference caused to PUs, by adjusting the parameter λ, which is critical in practical scenarios, where different interference constraints may be imposed on incumbents.

V. Conclusion

In this letter, we formulate the optimal spectrum sensing and access problem in cognitive radios networks via POMDPs. In a radio environment wherein the occupancy behavior of the incumbents is correlated across time and frequencies, we present a 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 concurrently solve for the optimal spectrum sensing and access policy. In addition to its superior performance compared to state-of-the-art algorithms, our framework facilitates regulation of the trade-off between SU throughput and PU interference.

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] 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] 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.

[7] 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.

[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. Mosleh, A. A. Tadaion, and M. Derakhtian, "Performance analysis of the neyman-pearson fusion center for spectrum sensing in a cognitive radio network," in *IEEE EUROCON 2009*, May 2009, pp. 1420–1425.

[12] 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.

[13] 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

[14] 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.