Learning-based Cognitive Radio Access via Randomized Point-Based Approximate POMDPs

Bharath Keshavamurthy *Student Member, IEEE* and Nicolò Michelusi *Senior Member, IEEE*

***Abstract*—In this paper, a novel spectrum sensing and access strategy based on approximate Partially Observable Markov Decision Processes (POMDPs) is proposed, wherein a cognitive radio learns the time-frequency correlation model defining the occupancy behavior of incumbents, via the Baum-Welch algorithm, and concurrently devises an optimal strategy to perform spectrum sensing and access that exploits this learned correlation model. To ameliorate the complexity of the POMDP optimization, the PERSEUS algorithm, a randomized point-based value iteration method, is designed, with fragmentation and Hamming distance state filters. Evaluating the cognitive radio throughput against incumbent interference, we demonstrate that, with sensing restrictions, our framework achieves a 6% performance gain over that attained by a maximum a-posteriori (MAP) state estimator with prior model knowledge, and outperforms correlation-coefficient based clustering algorithms by an average of 60%; additionally, it surpasses a Neyman-Pearson Detector that assumes independence among channels with no sensing restrictions, by an average of 25%. Furthermore, unlike state-of-the-art algorithms, the proposed design facilitates the regulation of the trade-off between cognitive radio throughput and incumbent interference via a penalty parameter in the underlying MDP.**

***Index Terms*—Cognitive Radio, HMM, POMDP**

I. Introduction

With the advent of fifth-generation wireless communication networks, the problem of spectrum scarcity has been exacerbated [2]. For some time now, cognitive radio 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 licensed users while ensuring rigorous incumbent non-interference compliance [3].

A crucial aspect underlying the design of cognitive radio networks is the channel access protocol in the MAC layer of the stack. Additionally, physical design limitations are imposed on the cognitive radio node's spectrum sensor because of quick turnaround times and energy efficiency [4], which restrict the number of channels that can be sensed at any given time. This has led to research in algorithms that first determine which channels to sense and then, via aggregation or correlation exploitation, use the gathered information to perform optimal channel access. In this regard, the current state-of-the-art involves channel sensing and access strategies dictated by multi-armed bandits [5], reinforcement learning agents [6], and other custom heuristics [7], [8]. Yet, almost all these works, such as [5], [6], assume independence across frequencies 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 time (temporal correlation) [9]–[11]. This pattern in the occupancy behavior of the incumbents imputes a high correlation among channels, which may be learned and leveraged for more accurate predictions of spectrum holes. In this paper, we propose a parameter estimation algorithm to learn the aforementioned correlation model, and an algorithm based on approximate POMDPs to determine the optimal channel sensing and access policy that exploits this learned correlation structure.

Related Work: The works [5], [12], [13] develop spectrum sensing and access algorithms under the assumption that the occupancy behavior of incumbents is independent across both time and frequencies. In our work, we exploit both frequency and temporal correlations. In [14], a compressed spectrum sensing scheme is devised that exploits sparse temporal dynamics in the occupancy of licensed users; in [15], an efficient spectrum sensing strategy is proposed for dense multi-cell cognitive networks, that also exploits the spatial structure of interference; yet, both works neglect frequency correlation.

Spectrum sensing and access strategies in a distributed multi-agent cognitive radio setting have been considered in [6] and solved using TD-SARSA with Linear Function Approximation (LFA). However, frequency correlation is precluded, and errors in state estimation are neglected in the decision process. Unlike [6], we consider a model with correlation across frequencies, and we account for uncertainty in the occupancy state via a POMDP formulation. Although the spectrum sensing algorithms detailed in [7] consider the correlation in incumbent occupancy behavior across frequencies, the authors assume a perfect, noise-free observation model; instead, we account for the impact of noisy observations in our design. In both [7], [8], the authors account for occupancy correlation across both time and frequencies, yet their algorithms use a data-driven strategy wherein pre-loaded databases are employed offline to estimate the correlation models – which is impractical in non-stationary settings; instead, we present a *fully online framework* that learns the correlation model and simultaneously solves for the optimal channel sensing and access policy.

Non-adaptive strategies like the Viterbi algorithm in [8] employ a fixed channel sensing set throughout their period of operation and require a-priori knowledge of the transition model of the underlying MDP. In contrast, our solution learns this transition model and concurrently adapts the channel sensing set based on the estimated occupancy transitions and reward/penalty feedback. Next, highlighting our solution against the model-free RL model described in [16] which frames the problem under an unknown Markovian time-frequency correlated PU occupancy structure, our solution achieves superior

Extensions to this work have been submitted to IEEE TCCN [1].
Part of this research has been funded by NSF under grant CNS-1642982.
B. Keshavamurthy is with ECE, Purdue University, West Lafayette, IN.
N. Michelusi is with the School of ECEE, Arizona State University, AZ.
Email: bkeshava@purdue.edu, nicolo.michelusi@asu.edu

performance owing to more accurate estimation of the MDP transition model parameters and more nuanced approximations based on this correlation in PU occupancy behavior – namely, fragmentation (frequency correlation) and Hamming distance state filters (temporal correlation). It is also worth noting that none of the works in the state-of-the-art provides a mechanism to *regulate the trade-off* between the throughput attained by the cognitive radio and the interference caused to incumbent transmissions due to inaccurate/overlapping access by the cognitive radio. Our proposed solution facilitates this regulatory mechanism by enabling the tuning of a penalty/cost parameter associated with the underlying MDP model.

Novelty: In a nutshell, the contributions of this paper are as follows: we develop an approximate POMDP framework for spectrum sensing and access in a radio environment with multiple licensed users exhibiting Markovian 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 with fragmentation and Hamming distance state filters to find the optimal spectrum sensing and access policy in a computationally tractable way; finally, we benchmark our solution with state-of-the-art algorithms, and demonstrate superior performance.

Extensions: In an elaboration of our work [1], the learning-based approximate POMDP framework outlined in this paper has been extended to multi-agent settings, with evaluations in centralized deployments performed on the DARPA SC2 Colosseum, and in distributed deployments performed on a custom ad-hoc test-bed of ESP32 WiFi radios.

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 throughput by opportunistically accessing portions of the spectrum left unused by the PUs, see Fig. 1. To this end, the SU should learn how to intelligently access spectrum holes, while ensuring nominal interference with PU 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; $V_k(i) \sim \mathcal{CN}(0, \sigma_V^2)$ is circularly symmetric additive complex Gaussian noise, i.i.d across frequency and time, and independent of channel H and PU signal X. We further assume that the J PUs employ orthogonal spectrum access (e.g., OFDMA) so that $X_{j,k}(i) X_{g,k}(i) = 0, \forall j \neq g$. Thus,

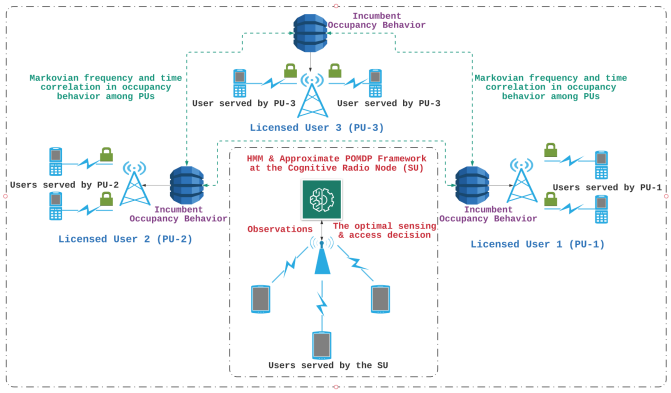

Fig. 1. The radio ecosystem under analysis: an exemplification of the system model discussed in Sec. II with J=3, K=18, and κ=3.

letting $j_{k,i}$ be the index of the PU that occupies the kth spectrum band at time i, $H_k(i) \triangleq H_{j_{k,i},k}(i)$ and $X_k(i) \triangleq X_{j_{k,i},k}(i)$ (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 $H_k(i)$ as Rayleigh fading 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 $X_k(i) = \sqrt{P_{tx}} B_k(i) S_k(i)$, 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 is modeled as the vector

$$\vec{B}(i) = [B_1(i), B_2(i), B_3(i), \cdots, B_K(i)]^T \in \{0, 1\}^K. \tag{3}$$

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

$$\mathbb{P}(\vec{B}(i+1) | \vec{B}(j), \forall j \leq i) = \mathbb{P}(\vec{B}(i+1) | \vec{B}(i)). \tag{4}$$

Additionally, when joining the spectrum pool, the PUs may occupy several 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. 2, and is described as

$$\mathbb{P}(\vec{B}(i+1) | \vec{B}(i)) \tag{5}$$
$$= \mathbb{P}(B_1(i+1) | B_1(i)) \prod_{k=2}^{K} \mathbb{P}(B_k(i+1) | B_k(i), B_{k-1}(i+1)).$$

Overall, the correlation model is expressed by two coupled Markov chains: one across time and the other

[1] If $S_k(i)$ does not have constant amplitude, we can model $H_k(i) S_k(i) \sim \mathcal{CN}(0, \sigma_H^2 \mathbb{E}[|S_k(i)|^2])$, without any modifications to the subsequent analysis.

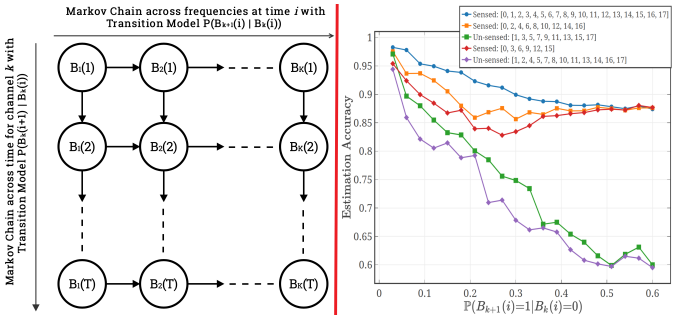

Fig. 2. The correlation model across time and frequencies underlying the occupancy behavior of PUs in the network (a); and an illustration of the variation of estimation accuracy with channel correlation, using the Viterbi algorithm as a state estimator in HMM settings for a Markov chain across channels ($\mathbb{P}(\vec{B}_{k+1}(i)|\vec{B}_k(i))$), with $\mathbb{P}(\vec{B}_k(i)=1)=0.6$

across frequencies. We parameterize this two-dimensional Markov chain structure by the vector $\vec{\theta}=[\vec{p}\ \ \vec{q}]^\top$, where $\vec{p}=[p_{uv}=\mathbb{P}(B_k(i+1)=1|B_{k-1}(i+1)=u, B_k(i)=v):u,v\in\{0,1\}]^\top$ and $\vec{q}=[q_w=\mathbb{P}(B_1(i+1)=1|B_1(i)=w):w\in\{0,1\}]^\top$. We estimate $\vec{\theta}$ using the parameter estimation algorithm described in Sec. III to obtain the transition model underlying the MDP, given by (5). Fig. 2 (b) illustrates three intuitively obvious causes impeding the dexterity of a state estimator to correctly estimate channel occupancies: a decrease in the amount of correlation between successive channels, a lack of sensed data, and size of the discontinuities between successive sensed channels.

Spectrum Sensing Model: To detect the available spectrum holes, the SU performs spectrum sensing. However, owing to physical design limitations of its spectrum sensor [4], 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

$$\vec{Y}(i)=[Y_k(i)]_{k\in\mathcal{K}_i}, \tag{6}$$

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

$$f(\vec{Y}(i)|\vec{B}(i),\mathcal{K}_i)=\prod_{k\in\mathcal{K}_i} f(Y_k(i)|B_k(i)), \tag{7}$$

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

$$Y_k(i)|B_k(i)\sim\mathcal{CN}(0,\sigma_H^2 P_{tx}B_k(i)+\sigma_V^2). \tag{8}$$

POMDP Agent Model: We model the SU as a POMDP agent, whose goal is to devise an optimal sensing and access policy to maximize its throughput while ensuring minimal interference with PU transmissions. The agent's limited sensing capabilities coupled with its noisy observations result in uncertainty on the PU spectrum occupancy state. The transition model of the underlying MDP, given by (5), is denoted by $\mathbf{A}$ and is learned by the agent by interacting with the radio environment (see Sec. III). The observation model in (7) is denoted by $\mathbf{M}$, with $f(Y_k(i)|B_k(i))$ given by (8).

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 (3); $\mathcal{A}$ 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-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 $\mathcal{K}_i=\pi(\beta_i)$, which defines 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

$$\begin{aligned}\hat{\beta}_i(\vec{B}')&=\mathbb{P}(\vec{B}(i)=\vec{B}'|\vec{Y}(i),\mathcal{K}_i,\beta_i)\\&=\frac{\mathbb{P}(\vec{Y}(i)|\vec{B}',\mathcal{K}_i)\beta_i(\vec{B}')}{\sum_{\vec{B}''\in\{0,1\}^K}\mathbb{P}(\vec{Y}(i)|\vec{B}'',\mathcal{K}_i)\beta_i(\vec{B}'')}.\end{aligned} \tag{9}$$

Given the posterior belief $\hat{\beta}_i$, we estimate the spectrum occupancy $\vec{B}(i)$ based on the MAP criterion

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

with the kth spectrum channel estimate given by $\phi_k(\hat{\beta}_i)$. If the kth channel is deemed idle, 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 PUs, 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 $\lambda>0$ is a penalty cost. After performing data transmission, the SU computes the prior belief for the next time-slot based on the learned Markov chain dynamics (see Sec. III) as

$$\beta_{i+1}(\vec{B}'')=\sum_{\vec{B}'}\mathbb{P}(\vec{B}(i+1)=\vec{B}''|\vec{B}(i)=\vec{B}')\hat{\beta}_i(\vec{B}'). \tag{10}$$

We denote the functions that map the prior belief β_i to the posterior belief $\hat{\beta}_i$ through the spectrum sensing action $\mathcal{K}_i$ and the observation signal $\vec{Y}(i)$, and the posterior belief $\hat{\beta}_i$ to the next prior belief β_{i+1} as

$$\hat{\beta}_i=\hat{\mathbb{B}}(\beta_i,\mathcal{K}_i,\vec{Y}(i)),\ \ \beta_{i+1}=\mathbb{B}(\hat{\beta}_i),\ \text{respectively.}$$

The goal is to determine an optimal spectrum sensing policy to maximize the infinite-horizon discounted reward, with discount factor $0<\gamma<1$,

$$\pi^*=\arg\max_{\pi}V^\pi(\beta)\triangleq\mathbb{E}_\pi\Big[\sum_{i=1}^{\infty}\gamma^i R(\vec{B}(i),\hat{\beta}_i)|\beta_0=\beta\Big], \tag{11}$$

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

$$V_{t+1}(\beta) = \max_{\mathcal{K}\in\mathcal{A}} \sum_{\vec{B}\in\mathcal{B}} \beta(\vec{B}) \mathbb{E}_{\vec{Y}|\vec{B},\mathcal{K}} \Big[R(\vec{B}, \hat{\mathbb{B}}(\beta, \mathcal{K}, \vec{Y})) + \gamma V_t(\mathbb{B}(\hat{\mathbb{B}}(\beta, \mathcal{K}, \vec{Y}))) \Big], \ \forall \beta. \quad (12)$$

This problem can be solved using the value iteration algorithm, i.e., by solving (12) 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 high-dimensional belief space, hence, solving equation (12) exactly is computationally intractable.

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: an approximate, randomized point-based value iteration algorithm [17].

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 is then fed back to the POMDP agent to compute the optimal spectrum sensing and access policy. Inherently, the approach constitutes solving the Maximum Likelihood Estimation (MLE) problem

$$\vec{\theta}^* = \arg\max_{\vec{\theta}} \log \Big(\sum_{\mathbf{B}} \mathbb{P}(\mathbf{B}, \mathbf{Y}|\vec{\theta}) \Big), \quad (13)$$

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 may 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 Baum-Welch algorithm, a special instance of the Expectation-Maximization (EM) algorithm used to find the unknown parameters of a HMM. Specifically, the E-step is given by

$$Q(\vec{\theta}|\vec{\theta}^{(t)}) = \mathbb{E}_{\mathbf{B}|\mathbf{Y},\vec{\theta}^{(t)}} \Big[\log \Big(\mathbb{P}(\mathbf{B}, \mathbf{Y}|\vec{\theta}^{(t)}) \Big) \Big]. \quad (14)$$

The term $Q(\vec{\theta}|\vec{\theta}^{(t)})$ can be computed by employing the Forward-Backward algorithm [18] using the current estimate $\vec{\theta}^{(t)}$. The M-step constitutes

$$\vec{\theta}^{(t+1)} = \arg\max_{\vec{\theta}} Q(\vec{\theta}|\vec{\theta}^{(t)}), \quad (15)$$

which involves re-estimation of the maximum likelihood parameter vector $\vec{\theta}$ [18] 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 (12) for POMDPs with large state and action spaces using exact value iteration is computationally infeasible [17]. 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 [17], 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 (12) can be approximated by a piece-wise linear and convex function [17], 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

$$V_t(\beta) \approx \beta \cdot \vec{\alpha}_t^{u^*}, \ u^* = \arg\max_{u\in\{1,2,\ldots,|\tilde{\mathcal{B}}|\}} \beta \cdot \vec{\alpha}_t^u, \quad (16)$$

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 [17],

$$\vec{\alpha}_{t+1}^u = \Xi_{\mathcal{K}_{t+1}^u}^u, \ \mathcal{K}_{t+1}^u = \arg\max_{\mathcal{K}\in\mathcal{A}} \beta_u \cdot \Xi_{\mathcal{K}}^u, \quad (17)$$

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

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

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

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

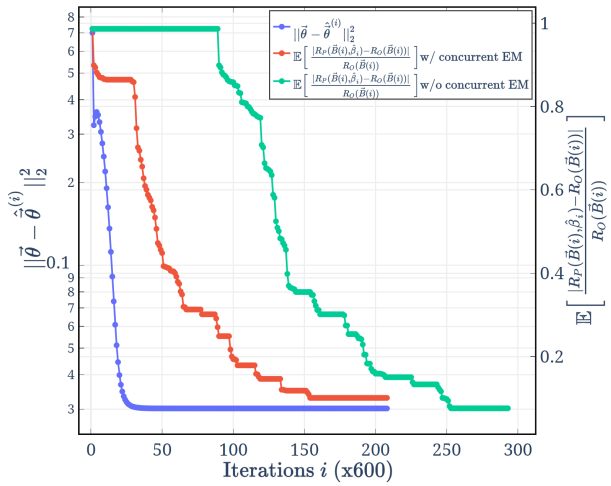

Fig. 3. Convergence of: MSE of the EM algorithm to estimate $\vec{\theta}$; and Loss of the fragmented PERSEUS algorithm with belief update simplification

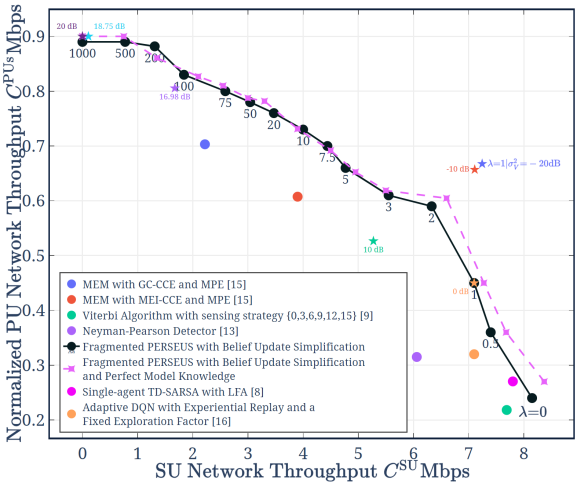

Fig. 4. Evaluation of C^{SU} versus C^{PUs} for different values of the penalty λ and noise power σ_V^2; comparisons with state-of-the-art and our solution with model foresight

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, which constitutes a single PERSEUS iteration. 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 (9) 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 update simplification heuristic*, i.e., we allow only those state transitions we deem to be the most probable – for example, only those 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 with parameters $\vec{p}=[p_{00}=0.1, p_{01}=p_{10}=0.3, p_{11}=0.7]^\top$ and $\vec{q}=[q_0=0.3, q_1=0.8]^\top$. An SU intelligently tries to access the available white spaces in order to maximize its own throughput, while limiting interference to licensed users. 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, $\text{SINR}_{\text{PU}}=17$dB when there is no interference from the SU, and $\text{SINR}_{\text{PU}}=6$dB under SU interference; at the SU receiver, $\text{SINR}_{\text{SU}}=11$dB when the channel is not occupied by a PU, and $\text{SINR}_{\text{SU}}=-6$dB under PU interference. We denote the actual SINR realizations in frequency k and time index i by $\text{SINR}_{\text{SU}}(k,i)$ and $\text{SINR}_{\text{PU}}(k,i)$, with $\text{SINR}_{\text{SU}}(k,i)=0$ if the SU remains idle. The SU can sense at most $\kappa=6$ out of $K=18$ channels at any given time, it is backlogged and accesses all the channels deemed idle. Hence, the average SU throughput over T time-slots 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 $R_{\text{SU}}=0.6$Mbps 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 transmission attempts, is given by

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

where $R_{\text{PU}}=0.9$Mbps is the transmission rate of the PUs on each frequency channel. To solve for the optimal PERSEUS policy, we employ a discount factor of $\gamma=0.9$ and a termination threshold of $\epsilon=10^{-5}$.

Evaluations: The plot depicted in Fig. 3 shows the Mean Squared Error (MSE) convergence of the parameter estimation algorithm to determine the time-frequency correlation parameters $\vec{\theta}$, averaged over $45 \cdot 10^3$ iterations. Assuming a time-slot duration of 3ms, this corresponds to an observation and estimation period of 135s. Starting with initial estimates of 0.5 for the parameters p_{uv} and q_w, $\forall u, v, w \in \{0, 1\}$, 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. 3 also illustrates the convergence of the PERSEUS algorithm, on the same time scale and in the same simulation run. The loss metric is defined as the difference

in utility obtained by our PERSEUS algorithm, denoted by $R_P(\vec{B}(i), \hat{\beta}_i)$ at time-slot i, and an *Oracle* which knows the exact occupancy behavior of PUs in the network, with utility $R_O(\vec{B}(i))$.

In Fig. 4, we evaluate the performance of our solution in terms of SU and PU network throughputs over varying values of the penalty term λ: we find that our POMDP agent decides to limit channel access when the penalty is high, leading to lower SU network throughput and PU interference; conversely, it follows a more lenient channel access strategy when the penalty is low, resulting in higher SU network throughput and PU interference. In a similar vein, the agent is apprehensive about channel access when the noise power is high; and is sure-footed in low-noise settings. Compared to a variant of PERSEUS in which the agent has perfect knowledge of the correlation model, we observe small performance degradation due to the concurrent model learning. We appraise the performance of our framework – in terms of the *achieved SU throughput vis-à-vis PU throughput degradation* – against state-of-the-art algorithms including RL-based solutions, all with sensing restrictions of 6 unless otherwise stated:

- Two variants 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 correlation threshold 0.77: we show an average improvement of 60%;
- The Viterbi algorithm [8] with a-priori model information: we demonstrate a 6% upswing;
- Neyman-Pearson Detector [12], which neglects time-frequency correlation among channels and senses all channels (*no sensing restrictions*); AND fusion rule across 300 different samplings, and a detection threshold determined to achieve a false alarm probability of 0.3: we demonstrate 25% enhancement;
- TD-SARSA with LFA [6] in single-agent deployment settings, with belief update heuristic constant 0.9, discount factor 0.9, exploration factor 0.01, and raw false alarm probability 5%: we perform 3% better; and
- Adaptive Deep Q-Network (DQN) [16] with experiential replay (memory size 10^6), 2048 input neurons, 2×4096 neurons with ReLU activation functions, MSE cost function with an Adam optimizer, exploration factor 0.1, learning rate 10^{-4}, and batch size 32: we perform 9% better.

Besides, our design facilitates the regulation of the trade-off between cognitive radio throughput and incumbent interference via the penalty term λ.

V. Conclusion

In this paper, we formulate the optimal spectrum sensing and access problem as an approximate POMDP, which leverages learning of the spectrum occupancy correlation model of the PU via the Baum-Welch algorithm. Through system simulations, we demonstrate the advantages of exploiting the correlation structure – as opposed to Neyman-Pearson Detector which assumes independence – and of adapting the spectrum sensing decision to optimize the performance – as opposed to Viterbi, which uses a fixed sensing strategy. We demonstrate the feasibility of a concurrent learning and decision-making framework, as opposed to correlation-coefficient based clustering algorithms which rely on pre-loaded datasets for determining the PU occupancy correlation model. Our framework enables a critical feature: the ability of the SU to regulate the interference caused to PUs, by adjusting a penalty parameter.

References

[1] B. Keshavamurthy and N. Michelusi, "Learning-based Spectrum Sensing in Cognitive Radio Networks via Approximate POMDPs," 2021, Under review at IEEE Transactions on Cognitive Communications and Networking.

[2] 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.

[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] 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.

[5] 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.

[6] 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.

[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] 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.

[9] 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.

[10] 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.

[11] 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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] S. Wang, H. Liu, P. H. Gomes, and B. Krishnamachari, "Deep Reinforcement Learning for Dynamic Multichannel Access in Wireless Networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 2, pp. 257–265, 2018.

[17] 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

[18] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE 77, no. 2*, pp. 257–286, Feb 1989.