Utility Maximization in Cognitive Radio Networks using POMDP Approximate Value Iteration methods

Bharath Keshavamurthy, Nicolò Michelusi

[NM: title is a bit too long..]

***Abstract*—**

***Index Terms*—Hidden Markov Model, POMDP, and the PERSEUS Algorithm**

I. Introduction

II. System Model

[NM: add figure with system model]

A. The Observation Model

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

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

where $y(n)$ is expressed as a convolution of the signal $x_p(n)$ of the pth PU with the channel impulse response $h_p(n)$, and $v(n)$ denotes additive white Gaussian noise (AWGN) with variances σ_v^2. Eq. (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

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

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

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

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

B. PU Spectrum occupancy model

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

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

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] Therefore, the PU occupancy behavior in the entire wideband spectrum of interest discretized into narrow-band frequency components, at time index i can be modeled as the vector

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

PUs join and leave the spectrum at random times. To capture this temporal correlation in the spectrum occupancy dynamics of PUs, we model the spectrum occupancy dynamics as a Markov process: given $\vec{B}(i)$, the spectrum occupancy state at time index i, $\vec{B}(i+1)$ is independent of the past, $\vec{B}(j), \ j < i$; $j, i \in \{1, 2, 3, \dots, T\}$, i.e.

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

Additionally, when joining the spectrum pool, PUs occupy a number of adjacent spectrum bands, and may vary their

This research has been funded by —–.

The authors are with the School of Electrical and Computer Engineering, Purdue University. email: {bkeshava,michelus}@purdue.edu.

[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, without any modification to the subsequent analysis. $B_k(i) \in \{0, 1\}$ 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.

spectrum needs depending on traffic demands, channel conditions, etc. To capture this behavior, we model $\vec{B}(i)$ as having Markovian correlation across sub-bands as,

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

That is, $B_{k+1}(i+1)$ depends on the occupancy state of the adjacent spectrum band $B_k(i+1)$ at the same time, and that of the same spectrum band $k+1$ in the previous time index i. The true states encapsulate the actual occupancy behavior of the PU and the measurements at the SU are noisy observations of these true states which are modeled to be the observed states of a Hidden Markov Model (HMM). Owing to physical design limitations at the SU's spectrum sensor **[NM: citation?]**, not all sub-bands in the discretized spectrum can be sensed. Let κ with $1 \leq \kappa \leq K$ be the number of spectrum bands that can be sensed by the SU at any time, capturing the spectrum sensing constraint. Let $\mathcal{K}_i \subseteq \{1, 2, \dots, K\}$ with $|\mathcal{K}_i| \leq \kappa$ be the set of indices corresponding to the spectrum bands sensed by the SU, which is part of our design. Then, we model the emission process of the HMM as

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

Given the spectrum occupancy vector $\vec{B}(i)$ and the set of sensed spectrum bands $\mathcal{K}_i$, the probability density function of $\vec{Y}(i)$ is expressed as

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

owing to the channels, noise and transmitted symbols across frequency bands. Moreover,

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

Now, we model the spectrum access scheme of the SU as a Partially Observable Markov Decision Process (POMDP) wherein the goal of the POMDP agent is to devise an optimal sensing and access policy in order to maximize its throughput while maintaining strict non-interference compliance with incumbent transmissions.

C. The POMDP Agent Model

The agent's limited observational 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 in the previous subsection, is denoted by A and is learnt by the agent by interacting with the radio environment. The emission model B is given by (9), with $f(Y_k(i)|B_k(i))$ given by (10). We model the POMDP as a tuple $(\mathcal{X}, \mathcal{A}, \mathcal{Y}, A, B)$ where $\mathcal{X} \equiv \{0,1\}^K$ represents the state space of the underlying MDP with states $\vec{x}$ given by all possible realizations of the spectrum occupancy vector as given by (3), $\mathcal{A}$ represents the action space of the agent, given by all $\binom{K}{\kappa}$ possible combinations in which the κ spectrum bands are chosen to be sensed out of K at any given time index; and $\mathcal{Y}$ represents the observation space of the agent based on the Observation Model outlined in the previous subsection. The run-time or interaction time of the agent is quantized into discrete time indices. At the beginning of each time index i, the agent selects κ spectrum bands out of K, thus defining the sensing set $\mathcal{K}_i$, performs spectrum sensing on these spectrum bands, observes $\vec{y} \in \mathcal{Y}$, and updates its belief $\forall \vec{x}'$ as

$$b_a^{\vec{y}}(\vec{x}') = \mathbb{P}(\vec{x}'|\vec{y}, a, \vec{b}) = \frac{\mathbb{P}(\vec{y}|\vec{x}', a)}{\mathbb{P}(\vec{y}|a, \vec{b})} \sum_{\vec{x} \in \mathcal{X}} \mathbb{P}(\vec{x}'|\vec{x}, a) b(\vec{x}), \quad (11)$$

where, $\mathbb{P}(\vec{y}|a, \vec{b})$ is the normalization constant, $\vec{b} \in \mathcal{B}$ represents the belief vector of the agent, i.e. a probability distribution over all states, in the previous time-step, $b(\vec{x}) \in \vec{b}$ is termed the belief and it represents the degree of certainty assigned to world state $\vec{x} \in \mathcal{X}$ by the belief vector $\vec{b}$. Considering sub-band k, the set of available actions to the agent at time index i is given by

$$a_k(i) = \begin{cases} 1, & \text{sense and access sub-band } k, \\ 0, & \text{do nothing with respect to sub-band } k. \end{cases} \quad (12)$$

Based on the number of truly idle sub-bands found accounting for the throughput maximization aspect of our end-goal and a penalty for missed detections accounting for the incumbent non-interference constraint, the reward to the agent is modelled as

$$R(\vec{x}(i), a(i)) = (1 - P_{FA}(i)) + \lambda P_{MD}(i), \quad (13)$$

where, $P_{FA}(i)$ represents the False Alarm Probability across all channels at time index i, $P_{MD}(i)$ represents the Missed Detection Probability across all channels at time index i, and $\lambda < 0$ represent the cost term penalizing the agent for missed detections, i.e. interference with the incumbent. The action policy of the agent $\pi : \mathcal{B} \rightarrow \mathcal{A}$ maps the belief vectors $\vec{b} \in \mathcal{B}$ to actions $a \in \mathcal{A}$ and is characterized by a Value Function

$$V^{\pi}(\vec{b}) = \mathbb{E}_{\pi}\Big[\sum_{i=0}^{\infty} \gamma^i R(\vec{b}_i,\ \pi(\vec{b}_i))|\vec{b}_0 = \vec{b}\Big], \quad (14)$$

where, $0 < \gamma < 1$ is the discount factor, $\pi(\vec{b}_i)$ is the action taken by the agent at time index i under policy π, and $\vec{b}_0$ is the initial belief vector. The optimal policy π^* specifies the optimal action to take at the current time index assuming that the agent behaves optimally at future time indices as well. It is evident from equation (14) that we have an infinite-horizon discounted reward problem formulation and in order to solve for the optimal policy we need to solve the modified Bellman equation given as

$$V^*(\vec{b}) = \max_{a \in \mathcal{A}} \Big[\sum_{\vec{x} \in \mathcal{X}} R(\vec{x}, a) b(\vec{x}) + \gamma \sum_{\vec{y} \in \mathcal{Y}} \mathbb{P}(\vec{y}|a, \vec{b}) V^*(\vec{b}_a^{\vec{y}})\Big], \quad (15)$$

$\forall \vec{b} \in \mathcal{B}$. 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 sub-bands, 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 and then utilize this learned model to solve for the optimal policy by employing Randomized Point-Based Value Iteration techniques, namely, the PERSEUS algorithm. **[NM: All the citations should go into the file ref.bib with the specific bibtex format (you can typically get the entry from IEEE Xplore)]**