THE MYOPIC SOLUTION OF THE MULTI-ARMED BANDIT COMPRESSIVE SPECTRUM SENSING PROBLEM

Saeed Bagheri and Anna Scaglione

Electrical and Computer Engineering Department
University of California, Davis
Davis, CA, USA

ABSTRACT

In this paper we formulate a Multi-Armed Bandit Compressive Spectrum Sensing (MAB-CSS) problem, in which a Cognitive Receiver (CR) decides dynamically how to best sense N sub-channels states, that switch from being occupied to being available as independent and statistically identical Markov chains. We assume that the CR is endowed with K CSS samplers each sensing an arbitrary mixture of the N signals in the sub-channels, and upon deciding what channels are available, it collects an equal reward from each channel unoccupied that is sensed. The MAB-CSS problem accounts for the ability of the CR of sweeping a large spectrum and being able to reconstruct the exact support of the N channels occupancy pattern, as long as the latter is sufficiently sparse. This is a generalization of the typical model in which the CR can sense K out of the N sub-channels. In choosing the compressive sensing strategy, the CR needs to consider how to gather the most informative statistics on the spectrum while not exceeding the limits beyond which the occupancy is no longer identifiable. In this work, we study a simplified and noiseless discrete sensing model and establish the structure of the optimum MAB-CSS myopic policy.

***Index Terms*—** Opportunistic access, multi-channel sensing, cognitive radio, compressive sensing, myopic policy.

1. INTRODUCTION

The well established framework of the multi-armed bandit (MAB) problem models the situation of a cognitive radio agent that simultaneously attempts to acquire new knowledge and to optimize its decisions based on what it has previously learnt [1–10]. Both Bayesian [2–8] and non-Bayesian [9,10] formulations of MAB have been considered for channel sensing and selection in cognitive radio systems. When the occupancy of each channel is modeled as a Markov chain with known transition probabilities, the dynamic channel selection problem becomes a restless multi-armed bandit (RMAB) within the Bayesian framework [4,5,11]. Most of the literature on online cognitive radio algorithms assumes that the receiver can be tuned to filter and sample different portions of the spectrum at different times; in these models there is implicitly a one to one correspondence between the total width of the spectrum explored and the number of samples per second available (i.e. the Nyquist limit).

In contrast, advances in Compressive Spectrum Sensing (CSS) [12–15] and finite rate of innovation (FRI) sampling [13, 16–19] are being applied to expand the spectrum sensing range further than the Nyquist limit, modeling the received signals as having a sparse structure, due to the spectrum holes the secondary users (SU) are wishing to detect. While FRI sampling and CS are well established sensing options for cognitive radio applications [20–22], active learning has been considered somewhat in antithesis with the FRI or CS sensing approach [23]. The drawback of static FRI or CSS front-ends is that sparsity is desired but, unfortunately, not guaranteed.

Our objective is to combine the perspective of online learning and MAB in particular, with new receiver architecture that not only can leverage on sparsity to learn, but also can use what it has learnt in the past to modify the way the spectrum is queried. We formulate as a MAB problem the optimal selection of a compressive sensing *arm* tuning the K CSS branches. To gain insight on the MAB-CSS optimal policy, we focus on a simplified noiseless discrete sensing model where the receiver has N sub-bands to sense overall, and as a sensing action, it can select only K linear combinations of a subset of them at each time slot. This means that, in principle, the cognitive receiver under the MAB-CSS architecture can choose to sense strategically a spectrum of variable size. We leave the study of noisy observations and of a detailed MAB-FRI sampling structure as future work.

2. PROBLEM FORMULATION

We are motivated by solving the standard problem of a CR user trying to opportunistically access a wideband spectrum assigned to a primary user (PU). The spectrum is divided into N non-overlapping narrow-band sub-channels which are assumed to be independent and statistically identical. In a particular geographical region and within a particular time interval, some of the N sub-bands are idle and available for opportunistic access. We further assume that the set of occupied sub-bands is potentially sparse. The CR objective is to access the empty sub-channels using a slotted transmission structure. Each sub-channel evolves as an i.i.d., two-state discrete time Markov chain. The state $s_i[t]$ of channel i in time slot t - "idle" (empty or state 0) and "busy" (occupied or state 1) - indicates the desirability of transmission over that channel at that time slot. The full system state in slot t is denoted as $\mathbf{s}[t]$ by collecting all N channel states, $\mathbf{s}[t] = [s_1[t], \ldots, s_N[t]] \in \{0,1\}^N$. The state transition probabilities are given by p_{ij}, $i,j \in \{0,1\}$. We assume that the probability of staying in state 0 or in state 1 is greater than that of switching, meaning that $p_{11} > p_{10}$ and $p_{00} > p_{01}$.

Given the existing restriction on the number of observations (due to K A/D converters), we choose to sense strategically a set of L_t sub-channels in slot t where $K \leq L_t \leq N$. At the beginning of each slot, the CR selects a subset $\mathcal{A}_t \subseteq \mathcal{N} \triangleq \{1, \ldots, N\}$ of the N channels to sense and each MAB *arm* selects to activate a linear combination of the frequency bands in $\mathcal{A}_t$, resulting in the following noiseless observation model

$$\boldsymbol{\theta}_{\mathcal{A}_t}[t] = \mathbf{B}_{\mathcal{A}_t}[t]\boldsymbol{\alpha}_{\mathcal{A}_t}[t]\,, \tag{1}$$

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

where $\boldsymbol{\theta}_{\mathcal{A}_t}[t]$ is a $K \times 1$ vector of observation gathered at the output of the CSS sampler and the $K \times L_t$ matrix $\mathbf{B}_{\mathcal{A}_t}[t]$ is the sensing matrix. In this work, we simply consider a random sensing matrix where any K columns of $\mathbf{B}_{\mathcal{A}_t}$ are linearly independent with probability 1. The $L_t \times 1$ vector $\boldsymbol{\alpha}_{\mathcal{A}_t}[t]$ represents the potentially sparse vector containing the samples (in frequency domain) of the selected sub-bands with nonzero elements for indices in the set $\{i \in \mathcal{A}_t : s_i[t] = 1\}$. We define $\mathbf{s}_{\mathcal{A}_t}$ as the support of the vector $\boldsymbol{\alpha}_{\mathcal{A}_t}[t]$, where $\mathbf{s}_{\mathcal{A}_t}$ includes the entries in $\mathbf{s}[t]$ corresponding to the indices in the set $\mathcal{A}_t$. The cognitive receiver recovers the vector $\boldsymbol{\alpha}_{\mathcal{A}_t}[t]$ and its support $\mathbf{s}_{\mathcal{A}_t}$ (system state) based on the observation vector $\boldsymbol{\theta}_{\mathcal{A}_t}[t]$ by exploiting its sparsity.

For each channel detected to be idle, the user transmits and collects one unit of reward. If none is sensed empty, the user does not transmit on the channels, collects no reward, and waits until the next slot to make another choice. This process repeats sequentially until the time horizon expires. The objective is to maximize the average reward (throughput) over a horizon of T slots by choosing strategically a sensing policy that governs channel selection in each slot.

Because $K \leq N$, the full system state in slot t is not observable and this problem falls into the general model of POMDP (Partially Observable Markov Decision Process) [24]. It has been shown that a sufficient statistic for optimal decision is the conditional probability that each channel is in state 0 (idle) given all past decisions and observations. Referred to as the belief vector, this sufficient statistic is denoted by $\boldsymbol{\Omega}[t] \triangleq [\omega_1[t], \ldots, \omega_N[t]]$, where $\omega_i[t]$ is the conditional probability that $s_i[t] = 0$. Given the sensing action $\mathcal{A}_t$ and the observation $\boldsymbol{\theta}_{\mathcal{A}_t}[t]$ in slot t, the belief vector for slot $t+1$, $\boldsymbol{\Omega}[t+1]$, can be obtained.

In multi-channel opportunistic access, the objective is to find a sensing policy π which specifies a sequence of functions $\pi \triangleq [\pi_1, \ldots, \pi_T]$, where π_t is the decision rule at time t that maps a belief vector $\boldsymbol{\Omega}[t]$ to a sensing action $\mathcal{A}_t \subseteq \mathcal{N}$. This is equivalent to the stochastic optimization problem of maximizing the total expected reward over a finite horizon, i.e.

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\pi} \left[\sum_{t=1}^{T} R_{\pi_t}\big(\boldsymbol{\Omega}[t]\big) \big| \boldsymbol{\Omega}[1] \right], \tag{2}$$

where $R_{\pi_t}(\boldsymbol{\Omega}[t])$ is the reward obtained under state $\boldsymbol{\Omega}[t]$ when channels in the set $\mathcal{A}_t = \pi_t(\boldsymbol{\Omega}[t])$ are selected. For a given sensing policy π, the belief vectors $\{\boldsymbol{\Omega}[t]\}_{t=1}^{T}$ form a Markov process with an uncountable state space. The expectation in (2) is with respect to this Markov process which determines the reward process. The vector $\boldsymbol{\Omega}[1]$ is the initial belief vector and if no information about the initial system state is available, each entry of $\boldsymbol{\Omega}[1]$ can be set to the stationary distribution ω_o of the underlying Markov chain:

$$\omega_o = \frac{p_{10}}{p_{01} + p_{10}}. \tag{3}$$

2.1. Identifiability of $\boldsymbol{\alpha}_{\mathcal{A}_t}[t]$

Integrating CSS into MAB learning significantly enlarges both the action space and the observation space of the learning engine. Specifically, the subset $\mathcal{A}_t$ of channels to be sensed at each time t is no longer limited to have a cardinality $L_t = K$. In fact, the action space consists of all subsets of the entire spectrum of interest with $K \leq L_t \leq N$. Thus, the CR dilemma is not only what bands to explore, but also to what extent the cardinality L_t should be chosen in time slot t so as to gather the most informative statistics on the spectrum, while not exceeding the limits beyond which the occupancy is no longer identifiable.

The identifiability of the channel occupancy is a direct consequence of the constraints for the exact sparse support recovery. Sparse vectors recovery methods exhibit a phase transition, that depends on the number of active components in the vector $\mathcal{A}_t[t]$ compared to the number of observations K. From results on compressive sensing [25], it is well-known that if there is no noise, as long as any K columns of the sub-matrix $\mathbf{B}_{\mathcal{A}_t}$ are linearly independent, any K observations $\boldsymbol{\theta}_{\mathcal{A}_t}[t]$ can recover *uniquely* an arbitrary $K/2$-sparse vector $\boldsymbol{\alpha}_{\mathcal{A}_t}[t]$ via an exhaustive search. Clearly, as long as the number of active sub-bands K_α in the chosen set $\mathcal{A}_t$ satisfies $K_\alpha < K/2$, the support is recovered exactly.

This ambiguity affects our study of the MAB-CSS in two essential ways: 1) In order to express the expected immediate reward $R_{\pi_t}(\boldsymbol{\Omega}[t])$, we need to explicitly know $\mathbf{s}_{\mathcal{A}_t}$, the support of the sparse vector $\boldsymbol{\alpha}_{\mathcal{A}_t}[t]$, given the observation vector $\boldsymbol{\theta}_{\mathcal{A}_t}[t]$. However, due to the ambiguity in sparse vector recovery, we cannot express all the elements for all possible cases. In fact, the capability to uniquely express all the elements in $\boldsymbol{\alpha}_{\mathcal{A}_t}[t]$, depends on the number of active sub-bands K_α in the chosen set $\mathcal{A}_t$; 2) Updating the belief vector as $\boldsymbol{\Omega}[t+1] \triangleq \mathcal{T}(\boldsymbol{\Omega}[t]|\mathcal{A}_t, \boldsymbol{\theta}_{\mathcal{A}_t}[t])$ is not as direct. Given the sensing action $\mathcal{A}_t$ and the observations $\boldsymbol{\theta}_{\mathcal{A}_t}[t]$ in slot t, the belief vector for slot $t+1$ should be obtained. In order to express $\omega_i[t+1]$ for $i \in \mathcal{A}_t$, we require to have an explicit mapping from $\boldsymbol{\theta}_{\mathcal{A}_t}[t]$ to $\mathbf{s}_{\mathcal{A}_t}$, which is not straightforward due to the existing ambiguity in sparse recovery.

3. OPTIMAL POLICY AND MYOPIC POLICY

3.1. Value Function and Optimal Policy

Let $V_t(\boldsymbol{\Omega}[t])$ be the value function, which represents the maximum expected total reward that can be obtained starting from slot t given the current belief vector $\boldsymbol{\Omega}[t]$. The reward that can be accumulated starting from slot t consists of expected immediate reward $\mathbb{E}[R_{\mathcal{A}_t}[t]]$ and the maximum expected future reward $V_{t+1}(\mathcal{T}(\boldsymbol{\Omega}[t]|\mathcal{A}_t, \boldsymbol{\theta}_{\mathcal{A}_t}[t]))$, given that the user takes action $\mathcal{A}_t$ and observes $\boldsymbol{\theta}_{\mathcal{A}_t}[t]$ in slot t.

Averaging over all possible observations $\theta_{\mathcal{A}_t}[t]$ and maximizing over all actions $\mathcal{A}_t$, we arrive at the following optimality equations

$$\begin{aligned} V_T(\boldsymbol{\Omega}[T]) &= \max_{\mathcal{A}} \quad \mathbb{E}\big[R_{\mathcal{A}}[T]\big] \\ V_t(\boldsymbol{\Omega}[t]) &= \max_{\mathcal{A}} \quad \Big[\mathbb{E}\big[R_{\mathcal{A}}[t]\big] + \sum_{\boldsymbol{\theta}} p(\boldsymbol{\theta}_{\mathcal{A}}[t] = \boldsymbol{\theta}|\mathcal{A}) \\ &\qquad \cdot V_{t+1}\big(\mathcal{T}(\boldsymbol{\Omega}[t]|\mathcal{A}, \boldsymbol{\theta}_{\mathcal{A}}[t] = \boldsymbol{\theta})\big)\Big], \end{aligned} \tag{4}$$

where the summation is over all possibilities for the observation vector $\boldsymbol{\theta}$. In theory, the optimal policy π^* and its performance $V_1(\boldsymbol{\Omega}[1])$ can be obtained by solving the above dynamic program. Brute force, this approach is computationally prohibitive due to the impact of the current action on the future reward and the uncountable space of the belief vector $\boldsymbol{\Omega}[t] \in [0,1]^N$, which is a vector of probabilities. It is thus common to consider suboptimal policies that are easier to compute and implement. One of the simplest such heuristics is a greedy policy where at each time step we take an action that maximizes the immediate one-step reward.

3.2. Myopic Policy

The myopic policy $\hat{\pi}$ ignores the effect of the current action on the future reward, and entirely focuses on maximizing the expected immediate reward $\mathbb{E}\big[R_{\mathcal{A}_t}[t]\big]$. Myopic policies are thus stationary and seek to maximize the reward as if there were only one step left in the

horizon. In the following, we first express the expected immediate reward for the MAB-CSS problem in hand and in Section 3.3, we derive the myopic policy.

For $L_t = K$, assuming that $\mathbf{B}_{\mathcal{A}_t}$ is full rank with probability 1, we can uniquely recover $\boldsymbol{\alpha}_{\mathcal{A}_t}[t]$ and its support vector $\mathbf{s}_{\mathcal{A}_t}$. As a result, the problem becomes equivalent to the MAB problem with K-arm selection [8]. In this case, the expected immediate reward can be expressed as $\mathbb{E}\big[R_{\mathcal{A}_t}[t]\big] = \sum_{i\in\mathcal{A}_t}\omega_i[t]$. For $L_t > K$, as we discussed in Section 2.1, we are faced with the identifiability problem in recovering $\boldsymbol{\alpha}_{\mathcal{A}_t}[t]$. In this paper, we simplify this dilemma by assuming that as long as the number of active sub-bands in the set $\mathcal{A}_t$ is smaller than or equal to $\Gamma \triangleq \lceil K/2\rceil - 1$ ($K_\alpha \leq \Gamma$), our sparse recovery algorithm is able to uniquely recover the support vector $\mathbf{s}_{\mathcal{A}_t}$ from the observation $\boldsymbol{\theta}_{\mathcal{A}_t}[t]$. If the sparse recovery algorithm determines that $K_\alpha > \Gamma$ (*failure event*), the CR conservatively does not transmit on the sub-bands sensed in $\mathcal{A}_t$ and collects no reward. In addition, in this case, the belief vector is updated as if $\mathcal{A}_t$ was an empty set. Under this assumption:

$$\mathbb{E}\big[R_{\mathcal{A}_t}[t]\big] = \begin{cases} \sum\limits_{i\in\mathcal{A}_t}\omega_i[t], & |\mathcal{A}_t| = K \\ \sum\limits_{k=0}^{\Gamma}(|\mathcal{A}_t|-k)P_{k|\mathcal{A}_t}, & |\mathcal{A}_t| > K \end{cases} \tag{5}$$

where $P_{k|\mathcal{A}_t} \triangleq \Pr(K_\alpha = k|\mathcal{A}_t, \boldsymbol{\Omega}[t])$ depends on the elements of the belief vector $\boldsymbol{\Omega}[t]$ with indices in $\mathcal{A}_t$. Thus, deriving (5) requires deriving $\boldsymbol{\Omega}[t]$. For $L_t = K$, $\omega_i[t+1]$ can be expressed as

$$\omega_i[t+1] = \begin{cases} p_{10}, & i\in\mathcal{A}_t,\ \mathbf{s}_{\mathcal{A}_t}(i) = 1 \\ p_{00}, & i\in\mathcal{A}_t,\ \mathbf{s}_{\mathcal{A}_t}(i) = 0 \\ \tau(\omega_i[t]), & i\notin\mathcal{A}_t \end{cases} \tag{6}$$

where $\tau(\omega) \triangleq \omega p_{00} + (1-\omega)p_{10}$. For $L_t > K$, we can express the belief update $\boldsymbol{\Omega}[t+1] = \mathcal{T}(\boldsymbol{\Omega}[t]|\mathcal{A}, \boldsymbol{\theta}_{\mathcal{A}}[t] = \boldsymbol{\theta})$ in terms of the *recovered support vector* $\mathbf{s}_{\mathcal{A}_t}$ as follows

$$\omega_i[t+1] = \begin{cases} p_{10}, & i\in\mathcal{A}_t,\ K_\alpha \leq \Gamma,\ \mathbf{s}_{\mathcal{A}_t}(i) = 1 \\ p_{00}, & i\in\mathcal{A}_t,\ K_\alpha \leq \Gamma,\ \mathbf{s}_{\mathcal{A}_t}(i) = 0 \\ \tau(\omega_i[t]), & i\in\mathcal{A}_t,\ K_\alpha > \Gamma \\ \tau(\omega_i[t]). & i\notin\mathcal{A}_t \end{cases} \tag{7}$$

3.3. Structure of the Myopic Policy

The myopic action under belief vector $\boldsymbol{\Omega}[t]$ is given by

$$\mathcal{A}_t^* = \arg\max_{\mathcal{A}}\ \mathbb{E}\big[R_{\mathcal{A}}[t]\big]\,. \tag{8}$$

Finding the myopic policy brute force is also a computationally intensive task, since it requires the search over all possible sets with cardinality $K \leq L_t \leq N$. In this Section, we derive the structure of the myopic policy for the general values of K and N by solving the optimization problem in (8).

Assume that at time t, we have the ordered belief vector as $\omega_{n_1}[t] \geq \omega_{n_2}[t] \geq \ldots \geq \omega_{n_N}[t]$. Then, for $L_t = K$, we have

$$\widetilde{\mathcal{R}}_t \triangleq \max_{\mathcal{A}}\ \mathbb{E}\big[R_{\mathcal{A}}[t]\big] = \max_{\substack{\mathcal{A}\\|\mathcal{A}|=K}}\ \sum_{i\in\mathcal{A}}\omega_i[t] = \sum_{i=1}^{K}\omega_{n_i}[t]\,, \tag{9}$$

which corresponds to the set $\widetilde{\mathcal{A}}_t = \{n_1, n_2, \ldots, n_K\}$. To find the myopic policy, we need to solve the following optimization problem for $L_t > K$ and compare its corresponding expected immediate reward with $\widetilde{\mathcal{R}}_t$ in (9):

$$\widehat{\mathcal{A}}_t = \arg\max_{\mathcal{A}}\ \mathbb{E}\big[R_{\mathcal{A}}[t]\big] = \arg\max_{\mathcal{A}}\ \sum_{k=0}^{\Gamma}(L_t-k)P_{k|\mathcal{A}}\,. \tag{10}$$

In order to derive the structure of the greedy policy, we first express the optimal set to be observed for a fixed cardinality $L_t = |\mathcal{A}_t|$. Then, the procedure is completed by providing conditions to determine the optimal value for L_t.

Lemma 1 *For a fixed or given cardinality $K \leq L_t \leq N$, $\mathbb{E}\big[R_{\mathcal{A}}[t]\big]$ is maximized for $\mathcal{A} = \{n_1, n_2, \ldots, n_{L_t}\}$ which senses the L_t sub-bands with the largest belief values from the vector $\boldsymbol{\Omega}[t]$.*

Proof The Lemma is certainly true for $L_t = K$. Let us define $\mathcal{A}' \triangleq \mathcal{A}\cup\{i\}$ for a fixed set $\mathcal{A}$ with cardinality $M \geq K$ where $i \in \mathcal{N}\setminus\mathcal{A}$. Using the law of total probability for conditional probabilities, $P_{k|\mathcal{A}'}$ can be expressed as

$$P_{k|\mathcal{A}'} = \begin{cases} \omega_i[t]P_{k|\mathcal{A}}\,, & k = 0 \\ (1-\omega_i[t])P_{k-1|\mathcal{A}} + \omega_i[t]P_{k|\mathcal{A}}\,. & 1 \leq k \leq \Gamma \end{cases} \tag{11}$$

Using (11) in (5), after some mathematical simplifications, leads to the following expression

$$\mathbb{E}\big[R_{\mathcal{A}'}\big] = G_{\mathcal{A}} + \omega_i[t](T_{\mathcal{A}} + (M-\Gamma)P_{\Gamma|\mathcal{A}}) - (M-\Gamma)P_{\Gamma|\mathcal{A}}\,, \tag{12}$$

where $G_{\mathcal{A}} \triangleq \sum_{k=0}^{\Gamma}(L_t-k)P_{k|\mathcal{A}}$ and $T_{\mathcal{A}} \triangleq \sum_{k=0}^{\Gamma}P_{k|\mathcal{A}}$. Since, $T_{\mathcal{A}} + (M-\Gamma)P_{\Gamma|\mathcal{A}} > 0$ and the terms $G_{\mathcal{A}}$, $T_{\mathcal{A}} + (M-\Gamma)P_{\Gamma|\mathcal{A}}$ and $(M-\Gamma)P_{\Gamma|\mathcal{A}}$ do not depend on $\omega_i[t]$, we can conclude that when $\mathcal{A}$ is fixed, $\mathbb{E}\big[R_{\mathcal{A}'}\big]$ is maximized when i is switched with

$$i^* \triangleq \arg\max_{k\in\mathcal{N}\setminus\mathcal{A}}\ \omega_k[t]\,. \tag{13}$$

This implies that $\mathbb{E}\big[R_{\mathcal{A}'}\big]$ can be further increased by sequentially switching the elements in $\mathcal{A}'$ with the elements in $\mathcal{N}\setminus\mathcal{A}'$ with higher belief values. This establishes that $\mathbb{E}\big[R_{\mathcal{A}'}\big]$ is maximized with respect to all the entries in $\mathcal{A}'$, when no switching is possible and $\mathcal{A}' = \{n_1, \ldots, n_{M+1}\}$ which completes the proof. ■

A direct consequence of Lemma 1 is that the optimization in (10) over $\mathcal{A}$ reduces to the maximization over the cardinality of the set $\mathcal{A}$ for $K+1 \leq L_t \leq N$:

$$\widehat{L}_t = \arg\max_{M}\ \mathbb{E}\big[R_{\mathcal{B}_M}[t]\big] = \arg\max_{M}\ \sum_{k=0}^{\Gamma}(M-k)P_{k|\mathcal{B}_M}\,, \tag{14}$$

where $\mathcal{B}_M \triangleq \{n_1, n_2, \ldots, n_M\}$ and $\widehat{\mathcal{A}}_t = \{n_1, \ldots, n_{\widehat{L}_t}\}$. As a result, at time slot t, the search space is reduced significantly to $N-K+1$ sets. The CR still needs to compute $\mathbb{E}\big[R_{\mathcal{B}_M}[t]\big]$ for $K+1 \leq M \leq N$ and compare their maximum $\mathbb{E}\big[R_{\widehat{\mathcal{A}}_t}[t]\big]$ with $\widetilde{\mathcal{R}}_t$ which is the expected immediate reward for $L_t = K$. This procedure requires the explicit computation of the probabilities $P_{k|\mathcal{B}_M}$ for $0 \leq k \leq \Gamma$, $K \leq M \leq N$ which may be computationally expensive since the number of active sub-bands is a Poisson binomial random variable. In the following Remark, a simple procedure is proposed to generate the probabilities $P_{k|\mathcal{B}_M}$ recursively.

Remark 1 *At time slot t, based on the belief vector $\boldsymbol{\Omega}[t]$, the CR only evaluates the probabilities $P_{k|\mathcal{B}_K}$, $0 \leq k \leq \Gamma$ using the recursive formula [26] for computing the probabilities in Poisson binomial distribution. Then, the probabilities $P_{k|\mathcal{B}_M}$, $K+1 \leq M \leq N$, are calculated sequentially according to the expressions in* (11) *replacing i with n_M, $\mathcal{A}$ with $\mathcal{B}_{M-1}$ and $\mathcal{A}'$ with $\mathcal{B}_M$.*

To further reduce the complexity of the process of finding the optimal L_t, we propose a suboptimal variant of the introduced myopic policy. The following Lemma establishes the idea that motivates the simplified and sequential procedure.

Lemma 2 $\mathbb{E}\big[R_{\mathcal{B}_{M+1}}\big] \geq \mathbb{E}\big[R_{\mathcal{B}_M}\big]$ *under a threshold policy* $\omega_{n_{M+1}}[t] \geq \eta_M$*, where the threshold is defined as*

$$\eta_M \triangleq \begin{cases} \dfrac{\widetilde{\mathcal{R}}_t - G_{\mathcal{B}_K} + (K-\Gamma)P_{\Gamma|\mathcal{B}_K}}{T_{\mathcal{B}_K} + (K-\Gamma)P_{\Gamma|\mathcal{B}_K}}, & M = K \\ \dfrac{(M-\Gamma)P_{\Gamma|\mathcal{B}_M}}{T_{\mathcal{B}_M} + (M-\Gamma)P_{\Gamma|\mathcal{B}_M}}. & M \geq K+1 \end{cases} \tag{15}$$

Proof The proof directly follows by replacing i with n_{M+1}, $\mathcal{A}'$ with $\mathcal{B}_{M+1}$ and $\mathcal{A}$ with $\mathcal{B}_M$ in (12). ■

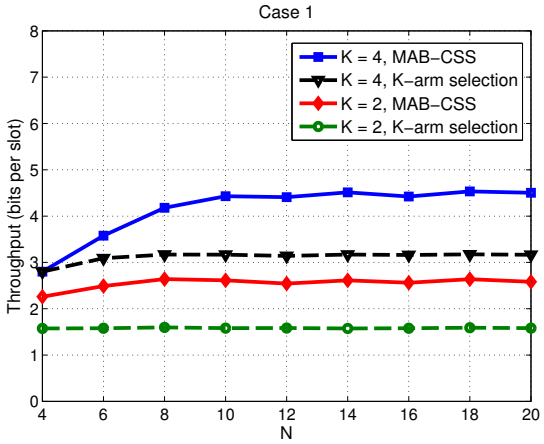
To find the optimal L_t, the CR first orders the belief vector and using the ordered beliefs computes $\widetilde{\mathcal{R}}_t$ and $P_{k|\mathcal{B}_K}$, $0 \leq k \leq \Gamma$. Afterwards, inspired by Lemma 2, the CR sequentially compares $\omega_{n_{M+1}}[t]$, $K \leq M \leq N-1$ with the threshold η_M defind in (15) to decide whether to increase L_t or not. The first time that this condition is not satisfied, the CR stops and selects the current value of M as the optimal value of L_t[1]. To compute the threshold η_M, the CR only needs to evaluate $T_{\mathcal{B}_M}$ and $P_{\Gamma|\mathcal{B}_M}$. However, using (11), we can easily discover the sequential update formulas as $T_{\mathcal{B}_{M+1}} = T_{\mathcal{B}_M} - (1-\omega_{n_{M+1}}[t])P_{\Gamma|\mathcal{B}_M}$ and $P_{\Gamma|\mathcal{B}_{M+1}} = (1-\omega_{n_{M+1}}[t])P_{\Gamma-1|\mathcal{B}_M} + \omega_{n_{M+1}}[t]P_{\Gamma|\mathcal{B}_M}$ which reduces the computational burden of evaluating η_M.

4. NUMERICAL EXPERIMENTS

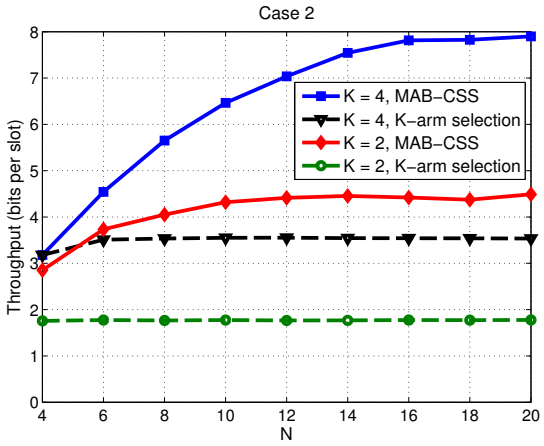
In this Section, we evaluate numerically the performance of the greedy approach for MAB-CSS architecture and specifically compare it with the myopic policy for the K-arm selection problem [8] where the CR selects exactly K out of the N sub-channels to sense at each time slot t. In [8], the authors have shown that for K-arm selection problem, the myopic policy is optimal when $p_{00} \geq p_{10}$.

In the numerical experiments, we assume $T = 30$ and the number of arms is equal to $K = 2$ and 4. We consider N independent sub-channels with the same transition probabilities and bandwidth $B = 1$. In the simulations, the value of N is set to vary from 4 to 20 and we compute the normalized expected total reward achieved over 500 simulation trials. For better comparison and visualization reasons, the results are normalized by T to reflect the throughput per slot. We consider two scenarios for the transition probabilities to capture the sparsity in spectrum occupancy and study the effect of the sparse channel occupancy on the performance of MAB-CSS. In Case 1, we set the transition probabilities as $p_{10} = 0.42$ and $p_{00} = 0.82$, which in the steady state corresponds to spectrum occupancy rate of 30%. In Case 2, we investigate a sparser scenario with transition probabilities $p_{10} = 0.4$ and $p_{00} = 0.9$, which in the steady state corresponds to channel occupancy rate of 20%.

In Fig. 1(a), the performance of MAB-CSS and K-arm selection are presented for $K = 2, 4$ in Case 1. The greedy approach in MAB-CSS outperforms the myopic policy in K-arm selection for all values of N and for both $K = 2$ and 4. Fig. 1(b) shows the performance comparison for Case 2. In this case, with sparser channel occupancy, the performance improvement is more significant. We also observe that myopic MAB-CSS with $K = 2$ outperforms K-arm selection with 4 arms when $N \geq 6$. The experiments showcase the capability of MAB-CSS architecture to improve the expected total throughput when the channel occupancy is sparse. Evidently, the myopic policy in MAB-CSS problem requires more processing and is more computationally extensive. However, our experiments demonstrate that in sparse enough settings (e.g. Case 2), it can double the expected throughput which is a considerable enhancement.



(a) Case 1: $p_{10} = 0.42$ and $p_{00} = 0.82$



(b) Case 2: $p_{10} = 0.4$ and $p_{00} = 0.9$

Fig. 1. Performance comparison of the myopic MAB-CSS with K-arm selection.

5. CONCLUSION

In this paper, we combined the perspective of MAB with FRI sampling structure. We specifically formulated the selection of a compressive sensing *arm* with K branches as a MAB problem. We assumed that when the number of active sub-bands in the selected subset to sense is limited by $K/2$, the states of the sensed sub-channels are perfectly identifiable. For the complexity reduced and noiseless MAB-CSS problem we considered in this work, the myopic policy was established and investigated numerically. The numerical experiments demonstrate that in finite horizon setting and when the channel occupancy is sparse, exploiting sparsity in MAB-CSS problem improves the expected total reward.

[1]We have experimentally observed that this procedure leads to the optimal selection of L_t. Since, we do not have a proof, we refer to this approach as a "suboptimal" version of the myopic policy.

6. REFERENCES

[1] Q. Zhao and B.M. Sadler, "A survey of dynamic spectrum access," *Signal Processing Magazine, IEEE*, 24(3):79–89, May 2007.

[2] Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized Cognitive MAC for Opportunistic Spectrum Access in Ad Hoc Networks: A POMDP Framework," *Selected Areas in Communications, IEEE Journal on*, 25(3):589–600, 2007.

[3] C. Tekin and M. Liu, "Online Learning in Opportunistic Spectrum Access: A Restless Bandit Approach," in Proc. of the 30th IEEE International Conference on Computer Communications (INFOCOM 2011), China, April, 2011.

[4] Q. Zhao, B. Krishnamachari, and K. Liu, "On myopic sensing for multi-channel opportunistic access: structure, optimality, and performance," *IEEE Transactions on Wireless Communications*, 2008.

[5] S. Ahmad, M. Liu, T. Javidi, Q. Zhao, and B. Krishnamachari, "Optimality of Myopic Sensing in Multichannel Opportunistic Access," *IEEE Transactions on Information Theory*, vol. 55, pp. 4040–4050, 2009.

[6] K. Liu and Q. Zhao, "Distributed Learning in Multi-Armed Bandit With Multiple Players," *IEEE Transactions on Signal Processing*, vol. 58, no. 11, pp. 5667–5681, November, 2010.

[7] J. Unnikrishnan and V. V. Veeravalli, "Algorithms for Dynamic Spectrum Access with Learning for Cognitive Radio," *IEEE Transactions on Signal Processing*, 58 (2):750–760, Feb, 2010.

[8] S. Ahmad, and M. Liu, "Multi-channel Opportunistic Access: A Case of Restless Bandits with Multiple Plays," Forty-Seventh Annual Allerton Conference, Allerton House, UIUC, Illinois, USA, September 30 - October 2, 2009.

[9] H. Liu, K. Liu, and Q. Zhao, "Learning in A Changing World: Non-Bayesian Restless Multi-Armed Bandit," IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2011), May, 2011.

[10] W. Dai, Y. Gai, B. Krishnamachari and Q. Zhao, "The Non-Bayesian Restless Multi-Armed Bandit: a Case of Near-Logarithmic Regret," IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2011), May, 2011.

[11] K. Liu and Q. Zhao, "Indexability of Restless Bandit Problems and Optimality of Whittle Index for Dynamic Multi-channel Access," *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5547–5567, November, 2010.

[12] V. Havary-Nassab, S. Hassan, and S. Valaee, "Compressive detection for wide-band spectrum sensing," IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2010), pages 3094–3097, 2010.

[13] M. Mishali and Y.C. Eldar, "Wideband spectrum sensing at sub-nyquist rates," [applications corner], *Signal Processing Magazine, IEEE*, 28(4):102–135, 2011.

[14] Z. Tian and G.B. Giannakis, "Compressed sensing for wideband cognitive radios," IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2007), volume 4, pages 1357–1360, 2007.

[15] Z. Tian, Y. Tafesse, and B.M. Sadler, "Cyclic feature detection with sub-nyquist sampling for wideband spectrum sensing," *Selected Topics in Signal Processing, IEEE Journal of*, 6(1):58–69, 2012.

[16] I. Maravic and M. Vetterli, "Sampling and reconstruction of signals with finite rate of innovation in the presence of noise," *Signal Processing, IEEE Transactions on*, 53(8):2788–2805, 2005.

[17] X. Li, A. Rueetschi, A. Scaglione, and Y. C. Eldar, "Optimal sampling structure for asynchronous multi-access channels," IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2012), 2012.

[18] E. Matusiak and Y.C. Eldar, "Sub-nyquist sampling of short pulses," IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2011), pages 3944–3947, 2011.

[19] K. Gedalyahu, R. Tur, and Y.C. Eldar, "Multichannel sampling of pulse streams at the rate of innovation," *Signal Processing, IEEE Transactions on*, 59(4):14911504, 2011.

[20] M. Mishali and Y.C. Eldar, "Blind multiband signal reconstruction: Compressed sensing for analog signals," *Signal Processing, IEEE Transactions on*, 57(3):9931009, 2009.

[21] P. Shukla and P.L. Dragotti, "Sampling schemes for multidimensional signals with finite rate of innovation," *Signal Processing, IEEE Transactions on*, 55(7):3670–3686, 2007.

[22] J.A. Tropp, J.N. Laska, M.F. Duarte, J.K. Romberg, and R.G. Baraniuk, "Beyond nyquist: Efficient sampling of sparse bandlimited signals," *Information Theory, IEEE Transactions on*, 56(1):520–544, 2010.

[23] R. Castro, J. Haupt, and R. Nowak, "Compressed sensing vs. active learning," IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2006), volume 3, may 2006.

[24] Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized cognitive MAC for opportunistic spectrum access in Ad Hoc networks: A POMDP framework," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 3, pp.589–600, Apr. 2007.

[25] M.A. Davenport, M.F. Duarte, YC Eldar, and G. Kutyniok, "Introduction to compressed sensing," in Compressed Sensing: Theory and Applications, Cambridge University Press, 2012.

[26] B.K. Shah, "On the distribution of the sum of independent integer valued random variables," American Statistician 27 (3): 123–124, 1994.