Optimal Scheduling of Cooperative Spectrum Sensing in Cognitive Radio Networks

Tengyi Zhang, Yuan Wu, Ke Lang, and Danny H. K. Tsang, *Fellow, IEEE*

***Abstract*—In Cognitive Radio (CR) networks, secondary users can be coordinated to perform spectrum sensing so as to detect primary user activities more accurately. However, more sensing cooperations for a channel may decrease the transmission time of the secondary users, or lose opportunities for exploiting other channels. In this paper, we study this tradeoff by using the theory of Partially Observable Markov Decision Process (POMDP). This formulation leads to an optimal sensing scheduling policy that determines which secondary users sense which channels with what miss detection probability and false alarm probability. A myopic policy with lower complexity yet comparable performance is also proposed. We further analytically study the properties and the solution structure for the myopic and the optimal policies under a simplified system model. Theoretical results reveal that under certain conditions, some simple but robust structures of the value function exist, which lead to an easy way to obtain the solution of POMDP. Moreover, the cooperative sensing scheduling problem embedded in our POMDP, which is generally a hard combinatorial problem, can be analyzed in an efficient way. Numerical and simulation results are provided to illustrate that our design can utilize the spectrum more efficiently for cognitive radio users.**

***Index Terms*—cognitive radio, cooperative sensing scheduling, partially observable Markov decision process**

I. Introduction

Cognitive Radio (CR) with its intelligence in interaction with the surrounding environment and flexibility in adapting its transmission parameters (e.g. frequency agility and power control etc.) has been considered as an important technique to solve the spectrum scarcity problem [1].

Compared to the conventional systemes, the functionalities of spectrum sensing and management are novel to CR, and they determine how good Secondary Users (SUs)[1] can identify the spectrum-hole and under what level Primary Users (PUs) are influenced.

In practice, the infrastructure-based CR system (e.g. 802.22 specification[3]) relies on Base Station (BS) to manage the cell and all the associated SUs. Through the coordination of BS, SUs can perform the so-called cooperative spectrum sensing to improve the spectrum sensing accuracy significantly [4].

Excessive cooperative spectrum sensing, however, reduces the transmission time of SUs and thus impairs the system efficiency. The inefficiency may become severe if the number of SUs is limited and only sequential (narrowband) sensing is allowed[2]. For example, SUs with limited sensing duration may cooperatively sense some of the channels to obtain higher sensing accuracy, but the downside is that they lose opportunities for exploiting the other channels. Therefore, a tradeoff exists between achieving better sensing accuracy on one channel and exploring more transmission opportunities on the other channels.

Meanwhile, in practice the idle spectrum available for SUs to access is time-varying, and the information about the the dynamics of idle spectrum can only be partially observed by SUs (due to both the imperfect spectrum sensing and sensing scheduling policy which will be described in detail in sections II and IV).

Based on these considerations, in this paper we study the dynamic scheduling for cooperative sensing under time-varying spectrum environment. Specifically, we formulate our dynamic sensing scheduling problem with the Partially Observable Markov Decision Process (POMDP) and derive an optimal sensing scheduling policy (i.e. determining which SUs to sense which set of channels with what sensing accuracy) by solving the formulated problem. We also analytically study the properties and the solution structure for the myopic and the optimal policies under a simplified system model. The theoretical results show some interesting properties of the problem and lead to a simple structural policy.

In [6], the authors studied the optimal distributed MAC protocols for opportunistic spectrum access in the POMDP framework. The proposed protocols guaranteed the synchronization between the secondary transmitters and receivers without requiring a central controller. [12] studied the structure of myopic policy in their POMDP problem and compared the performance of the myopic policy with the optimal policy. [13] considered a similar problem but took the imperfect sensing performance into consideration. CR networks with energy constraint were studied in [14]. Threshold structures of the optimal sensing and access policies were found, which reduced the complexity in searching for the optimal policies. Dynamic spectrum management was also studied in [16] [17]. The theory of cooperative sensing provides a method for CR networks to improve sensing accuracy and better exploit spectrum opportunities. [5] gave a survey on various cooperative spectrum sensing schemes. Several robust cooperative spectrum sensing techniques were established and their

[1] In CR terminology, Primary User (PU) refers to the user which has an exclusive ownership of some frequency band authorized by regulatory. Meanwhile, Secondary User (SU) refers to the user which, although has no pre-authorized frequency, can opportunistically access the unused/under-utilized frequency of PU without causing severe interference.

[2] According to [9], the wideband spectrum sensing refers to that the sensing device can sense multiple spectrum bands over a wide frequency range at a time. Meanwhile, the sequential spectrum sensing refers to that the sensing device can only sense one spectrum band at a time, and thus different spectrum bands have to be sensed sequentially.

performance were analyzed. [19] studied the impact of the cooperative sensing overhead on the system throughput with the consideration of the number of reporting packets. [20] studied a similar problem, which aimed to find the optimal sensing time and the optimal parameter for the result fusion in order to maximize SUs' throughput. [21] extended the analysis to the case of multiple channels, and the soft decision rule is applied in the energy sensing. However, [19]-[21] did not consider a time-varying dynamic spectrum environment, and they did not provide an analytical insight for the cooperative sensing scheduling problem, i.e. how to assign SUs to sense the channels.

The rest of the paper is organized as follows. We present the network model and propose our protocol in section II. We then formulate the problem of the tradeoff between cooperative sensing time and transmission time as a POMDP in section III. We derive the optimal policy and myopic policy for our problem in section IV. We study the properties and solution structures of the value functions in section V. Section VI presents numerical and simulation results. Finally, we conclude this paper in section VII.

II. System Model

A. *Network Model*

In this work, we consider a centralized CR network with a base station, which manages the cooperative sensing scheduling as well as data transmission. All SUs in a cell need to be synchronized. In the following part, we further assume there exists a set of SUs $\mathcal{M} = \{1, 2, ..., M\}$, and a set of orthogonal frequency channels $\mathcal{N} = \{1, 2, ..., N\}$ with a BS in a cell.

Each SU is equipped with a single radio interface. In this work, we assume all SUs use energy detection mechanism for spectrum sensing and each SU can only carry out the sequential spectrum sensing instead of the wideband spectrum sensing due to some PHY layer limitations.

B. *Opportunistic Channel Availability Model*

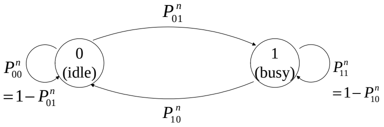

Fig. 1. DTMC model for PU channel n

In this paper, we assume primary system operates in a time slotted manner with fixed slot length L. In the PU network, each channel's occupancy (from slot to slot) follows a two-state Discrete Time Markov Chain (DTMC) as shown in Figure 1. Let $s_n(t)$ denote the availability state of channel n $(n \in \mathcal{N})$ in time slot t. $s_n(t) = 0$ denotes channel n is idle in slot t, while $s_n(t) = 1$ denotes channel n is busy in slot t. Furthermore, let the $1 \times N$ vector $\mathbf{s}(t) = (s_1(t), ..., s_N(t))$ denote the channel availability state vector for all the PU channels in slot t, which has the state space $\Omega^{\mathbf{s}} = \{(\omega_1, \omega_2, ..., \omega_N) | \omega_n = \{0, 1\}, \forall n \in \mathcal{N}\}$. By assuming independence across different channels, the dynamics of $\mathbf{s}(t)$ follow a DTMC with transition probability from state vector ω to state vector ω' given as:

$$P_{\omega\omega'} = \Pr(\mathbf{s}(t+1) = \omega' | \mathbf{s}(t) = \omega) = \prod\nolimits_{n=1}^{N} P^n_{\omega_n \omega'_n}, \quad (1)$$

$\forall \omega, \omega' \in \Omega^s$, where ω_n, ω'_n denote the nth element of state vector ω and ω', respectively. $P^n_{\omega_n \omega'_n}$ represents channel n's state transition probability. We consider the DTMC model as time homogeneous, i.e. $P^n_{01}, P^n_{10}, \forall n \in \mathcal{N}$, are time independent. We assume the PU channels' statistical behavior $P^n_{01}, P^n_{10}, \forall n \in \mathcal{N}$, can be obtained from a long term measurement by some channel parameter estimator [10], and this information is provided to CR BS. Note that for each channel n, the stationary probabilities of being idle and busy $\pi^n_0, \pi^n_1, \forall n \in \mathcal{N}$ can be calculated as $\pi^n_0 = \frac{P^n_{10}}{P^n_{01} + P^n_{10}}$, and $\pi^n_1 = \frac{P^n_{01}}{P^n_{01} + P^n_{10}}$.

C. *Spectrum Sensing Technique and Cooperative Detection*

Several well-known spectrum sensing techniques have been proposed including matched filter detection, energy detection, cyclostationary feature detection and wavelet detection [1] [5]. In this paper, we adopt the energy detection method [11]. The spectrum sensor detects the presence of PU signals by performing the binary hypothesis test: Hypothesis 0 (H_0) corresponds to no signal transmitted, while hypothesis 1 (H_1) corresponds to signal transmitted. Then, in a non-fading environment, the detection probability P_D and false alarm probability P_{FA} are given as $P_D = \Pr(Y > \lambda | H_1) = Q_u(\sqrt{2\gamma}, \sqrt{\lambda})$ and $P_{FA} = \Pr(Y > \lambda | H_0) = \Gamma(u, \frac{\lambda}{2}) / \Gamma(u)$, where Y is the test or decision statistic, λ is the decision threshold, u is the time bandwidth product, γ is the SNR, $Q_u(\cdot, \cdot)$ is the generalized Marcum Q-function, $\Gamma(\cdot)$ and $\Gamma(\cdot, \cdot)$ are the complete and incomplete gamma functions. Then, the miss detection probability is $P_{MD} = 1 - P_D$.

We adopt a simple cooperative sensing scheme called "OR" rule [5], which works as follows: every SU sends its sensing result (0 or 1) of a channel to the BS, and as long as one SU senses the channel as busy, the BS will take this channel as busy. Only if all SUs sense the channel as idle, BS will take the channel as idle. Then, the miss detection and false alarm probability of channel n are $P_{MD}(n) = \prod\limits_{m \in \mathcal{M}(n)} P_{MD}(m, n)$ and $P_{FA}(n) = 1 - \prod\limits_{m \in \mathcal{M}(n)} (1 - P_{FA}(m, n))$, where $P_{MD}(m, n)$ and $P_{FA}(m, n)$ are SU m's miss detection probability and false alarm probability of channel n, and $\mathcal{M}(n)$ is the set of SUs sensing this channel.

D. *Proposed protocol*

Figure 2 shows an example to illustrate the operation process of our proposed protocol. At the beginning of each slot, each channel will have a state transition according to the DTMC model described in section II B. The BS decides which SU senses which set of channels with what probabilities of miss detection and false alarm based on the optimal policy obtained. After receiving the decisions from the BS, the SUs will sequentially sense the assigned channels, and the channel sensing sequence can be arbitrarily determined. Since the

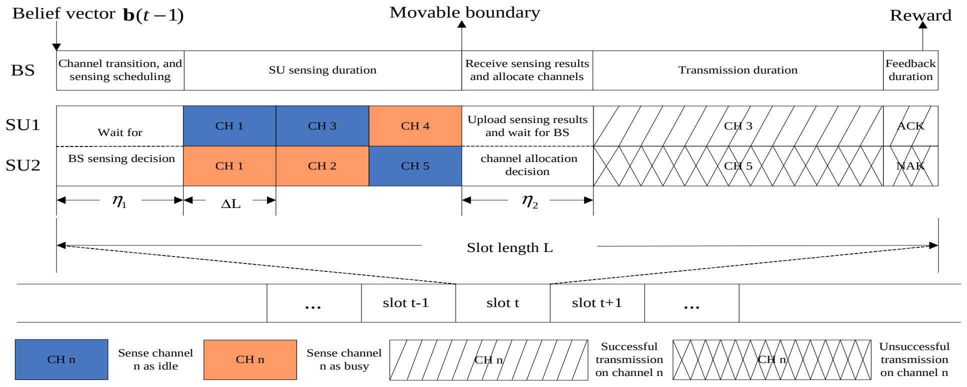
Belief vector $\mathbf{b}(t-1)$
Movable boundary
Reward
BS
Channel transition, and sensing scheduling
SU sensing duration
Receive sensing results and allocate channels
Transmission duration
Feedback duration
SU1
SU2
Wait for BS sensing decision
CH 1
CH 3
CH 4
Upload sensing results and wait for BS channel allocation decision
CH 3
ACK
CH 1
CH 2
CH 5
CH 5
NAK
η_1
ΔL
η_2
Slot length L
...
slot t-1
slot t
slot t+1
...
CH n
Sense channel n as idle
CH n
Sense channel n as busy
CH n
Successful transmission on channel n
CH n
Unsuccessful transmission on channel n

Fig. 2. An example of the operation process of our proposed protocol

sensing duration for each channel is a fixed value ΔL, and we have limited number of SUs, if the BS decides some channels are sensed by more SUs in order to increase these channels' sensing accuracy, then each SU may need to sense more channels accordingly, thus causing less time for transmission (Notice that in Figure 2, the slot length is L. The time duration for BS scheduling cooperative sensing is η_1. The time duration for SUs uploading their sensing results and BS allocating channels to SUs is η_2. All these three values are constant).

In our proposed protocol, data transmission works as follows: if the BS decides a channel as idle, then the BS will allocate this channel to one of SUs. Our protocol requires sensing synchronization for all SUs, i.e., each SU senses the same number of channels.

At the end of a slot, the SU using the channel will send an ACK or NAK to the BS. In this paper, we only consider the case of downlink transmission, but our proposed protocol can also be applied to the uplink transmission.

III. Problem Formulation

At the beginning of each time slot, based on previous actions and observations, the BS could have a belief state over every channel, which is the probability of a PU channel being in that state in the previous time slot. This is different from traditional Markov decision process since the BS may not know the exact state of a channel. For instance, if BS determines the channel as busy, it can not be sure if it is busy due to the probability of false alarm. Besides, if some channels are not sensed by any SU, the exact state of these channels will not be known either. Our problem thus fits into the POMDP framework.

A. Action

In our formulation, at the beginning of time slot t there are two actions, a^I and a^{II}. a^I determines which SU senses which channels. a^{II} determines how to tune the sensor operating point of each SU (i.e. miss detection probability and false alarm probability) when sensing a channel.

$$a^I(t) = \begin{bmatrix} a^1_{11}(t) & ... & a^1_{1N}(t) \\ ... & ... & ... \\ a^1_{M1}(t) & ... & a^1_{MN}(t) \end{bmatrix}$$

where $a^1_{mn}(t) \in \{0,1\}$, $a^1_{mn}(t) = 1$ denotes SU m senses channel n in time slot t, and $a^1_{mn}(t) = 0$ means the opposite. We define the set of SUs that are scheduled to sense channel n in slot t as $\mathcal{M}(n,t) = \{m|a^1_{mn}(t) = 1\}$. Similarly, we have $a^{II}(t)$, where each element $a^2_{mn}(t) = P_{MD}(m,n,t)$ denotes the specified miss detection probability for SU m on channel n in time slot t. Notice that by setting the value of miss detection probability $P_{MD}(m,n,t)$, we actually determine the sensor operating point for SU m on channel n in slot t, because both the sensing decision threshold λ and the false alarm probability $P_{FA}(m,n,t)$ for SU m on channel n in slot t can be calculated from $P_{MD}(m,n,t)$. Specifically, in this work we choose the value of miss detection probability $P_{MD}(m,n,t)$ from a set of discrete values, which is practical for most spectrum sensing modules' operation because the sensing operation point cannot be tuned continuously. We define $\mathbf{a}(t)$= $[a^I(t), a^{II}(t)]$.

B. Observation

Let $\theta_n(t)$ denote the observation result of channel n in time slot t. Then $\theta_n(t)$ could have the following 4 possible observations,

- BS determines the channel as idle and receives ACK after transmission; denote this as observation 0.
- BS determines the channel as idle and receives NAK after transmission due to miss detection; denote this as observation 1.
- BS determines the channel as busy, does not use the channel, and thus the BS receives no ACK or NAK; denote this as observation 2.
- BS decides not to sense the channel and thus observes nothing; denote this as observation 3.

Further, let the $1 \times N$ vector $\theta(t) = (\theta_1(t), ...\theta_N(t))$ denote the channel observation vector for all the PU channels at the end of slot t, which has the observation space $\mathbf{Z}^\theta = \{(z_1, z_2, ..., z_N)|z_n = \{0,1,2,3\}, \forall n \in \mathcal{N}\}$.

The individual channel observation probability $\Pr(\theta_n(t)|\mathbf{a}(t), s_n(t))$ is defined as the probability of the observation given the action we take and the current state of channel n. Let $P_{MD}(n,t)$ denote the miss detection probability and $P_{FA}(n,t)$ denote the false alarm probability of channel n in time slot t, respectively. Because of the "OR" rule in cooperative sensing, we have

$$\begin{aligned} P_{MD}(n,t) &= \prod_{m\in\mathcal{M}(n,t)} P_{MD}(m,n,t), \\ P_{FA}(n,t) &= 1 - \prod_{m\in\mathcal{M}(n,t)} (1 - P_{FA}(m,n,t)) \end{aligned} \tag{2}$$

Then the observation probability of the system is given by

$$\begin{aligned} &\Pr(\theta(t) = \mathbf{z}|\mathbf{a}(t), \mathbf{s}(t) = \omega) \\ &= \prod_{n=1}^{N} \Pr(\theta_n(t) = z_n|\mathbf{a}(t), s_n(t) = \omega_n) \end{aligned} \tag{3}$$

where $\mathbf{z} \in \mathbf{Z}^{\theta}$ is the observation vector, and z_n denotes the nth element of observation vector $\mathbf{z}$.

For the sake of simplicity, we assume every SU being scheduled to sense channel n should tune to the same sensor operating point (i.e. $P_{MD}(m,n,t) = \sqrt[|\mathcal{M}(n,t)|]{P_{MD}(n,t)}, \forall m \in \mathcal{M}(n,t)$).

C. Belief vector

Because of the partial spectrum sensing decisions and the presence of sensing errors, a BS may not observe the true system state. However, the BS can infer the system state based on all its past decisions and observations, and summarize this information into a belief vector [6], $\mathbf{b}(t) \triangleq \{b_{\omega}(t)\}_{\omega\in\Omega^{\mathbf{s}}}$[3] where $b_{\omega}(t) \triangleq \Pr(\mathbf{s}(t) = \omega|\mathbf{b}(0), \{\mathbf{a}(\tau), \theta(\tau)\}_{\tau=1}^{t}) \in [0,1]$ is the conditional probability (given all past decisions and observations) that the system state is ω in the current time slot t. $b_{\omega}(t)$ can only be computed at the end of the current time slot t when $\theta(t)$ is known (as shown in Figure 2). The BS will make actions at slot $t+1$ based on its belief vector of the system state $\mathbf{b}(t)$. Note that $\mathbf{b}(0)$ and $\mathbf{s}(0)$ denote the initial belief value and the initial system state at the beginning of each decision circle, respectively.

We define the updated belief vector as follows:

$$\mathbf{b}(t) \triangleq \mathcal{T}(\mathbf{b}(t-1), \mathbf{a}(t), \theta(t)) \triangleq \{b_{\omega'}(t)\}_{\omega'\in\Omega^{\mathbf{s}}} \tag{4}$$

where $\mathcal{T}(\mathbf{b}(t-1), \mathbf{a}(t), \theta(t))$ represents the updated knowledge of the network state after incorporating the action and observation obtained in slot t. Then, from Bayes rule, we have

$$\begin{aligned} &b_{\omega'}(t) = \Pr(\mathbf{s}(t) = \omega'|\mathbf{b}(t-1), \mathbf{a}(t), \theta(t)) \\ &= \frac{\sum\limits_{\omega\in\Omega^{\mathbf{s}}} b_{\omega}(t-1) P_{\omega\omega'} \Pr(\theta(t)|\mathbf{a}(t), \mathbf{s}(t) = \omega')}{\sum\limits_{\omega\in\Omega^{\mathbf{s}}} \sum\limits_{\omega''\in\Omega^{\mathbf{s}}} b_{\omega}(t-1) P_{\omega\omega''} \Pr(\theta(t)|\mathbf{a}(t), \mathbf{s}(t) = \omega'')} \end{aligned} \tag{5}$$

From these equations, we know that we have regained the Markov property for the belief state in that the next belief state depends only on the previous belief state, the current action and the current observation received.

D. Reward function

There will be a reward when the channel is sensed and finally the BS receives an ACK, the immediate reward for channel n ($n \in \mathcal{N}$) is

$$\begin{aligned} &R_n(\mathbf{a}(t), \theta_n(t)) \\ &= \begin{cases} \frac{L-k-\eta}{L}, & \text{if } \sum_{m=1}^{M} a^1_{mn}(t) > 0, \theta_n(t) = 0 \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{6}$$

where $k = \Delta L \cdot \sum_{n=1}^{N} a^1_{mn}(t),\ \forall m \in \mathcal{M}$, is the sensing duration (note that to keep synchronization, each SU should sense same number of channels, then $\sum_{n=1}^{N} a^1_{1n}(t) = \sum_{n=1}^{N} a^1_{2n}(t) = ... = \sum_{n=1}^{N} a^1_{Mn}(t)$), ΔL is the sensing duration for one channel, and $\eta = \eta_1 + \eta_2$ is a constant time used for BS decisions and getting sensing results from SUs. The reward function represents the ratio between the actual transmission time and the total slot length. Assume the number of bits delivered is proportional to the transmission time, then the physical meaning of the reward function is the normalized throughput achieved when the transmission is successful. The immediate reward for all the channels in time slot t is

$$R(\mathbf{a}(t), \theta(t)) = \sum_{n=1}^{N} R_n(\mathbf{a}(t), \theta_n(t)) \tag{7}$$

Finally, the expected reward for BS to make a decision at the beginning of slot t is

$$\begin{aligned} &\widetilde{R}(\mathbf{a}(t), \mathbf{s}(t-1) = \omega) = \\ &\sum_{\omega'\in\Omega^{\mathbf{s}}} \sum_{\mathbf{z}\in\mathbf{Z}^{\theta}} P_{\omega\omega'} \Pr(\theta(t) = \mathbf{z}|\mathbf{a}(t), \mathbf{s}(t) = \omega') R(\mathbf{a}(t), \theta(t) = \mathbf{z}) \end{aligned}$$

E. Complete problem formulation

We aim to develop the optimal policy that can maximize the expected total throughput of the SUs over a finite time horizon T, and at the same time it must satisfy the synchronization constraint and primary user interference constraint. The problem is formulated as follows:

$$\max\ E\{\sum_{t=1}^{T} R(\mathbf{a}(t), \theta(t))|\mathbf{b}(0) = \mathbf{b}\} \tag{8}$$

$$\text{subject to: } \sum_{n=1}^{N} a^1_{1n}(t) = \sum_{n=1}^{N} a^1_{2n}(t) = ... = \sum_{n=1}^{N} a^1_{Mn}(t) \tag{9}$$

$$L - \Delta L \cdot \sum_{n=1}^{N} a^1_{mn}(t) - \eta > 0,\ \forall m \in \mathcal{M} \tag{10}$$

$$P_c(n,t) \le \zeta,\ \forall n \in \mathcal{N} \tag{11}$$

In the above formulation, $\mathbf{b}$ is the initial belief vector which could be set according to the channels' statistical behavior. Constraint (9) is a synchronization constraint, which guarantees that each SU senses the same number of channels. Constraint (10) guarantees the transmission time is positive. Constraint (11) is the interference constraint, which aims to satisfy primary users' interference tolerance. Here, $P_c(n,t)$ is the collision probability of channel n in slot t, if we want to guarantee this value below the prescribed primary channel collision probability ζ, it is equivalent to require the miss detection probability of channel n in slot t below this threshold

$$P_{MD}(n,t) \le \zeta,\ \forall n \in \mathcal{N}. \tag{12}$$

[3] Here we abuse the notation a little since we just want to list all the elements in the set $\Omega^{\mathbf{s}}$ and assign them to the vector $\mathbf{b}(t)$, and the element order is not important.

Then the formulation with objective (8) constrained by (9), (10), and (11) changes into the formulation with objective (8) constrained by (9), (10), and (12).

IV. Optimal policy and Myopic policy

A. *Optimal policy*

In order to solve the objective function (8), we could solve the following value function $V_t(\mathbf{b}(t-1))$ which denotes the maximum expected remaining reward that can be obtained from the beginning of slot t when the current belief vector is $\mathbf{b}(t-1)$. We use backward induction method to calculate the value function that consists of two parts. One part is the expected immediate reward $\widetilde{R}(\mathbf{a}(t), \mathbf{s}(t-1) = \omega)$ in the current time slot, and the other part is the expected future reward $V_{t+1}(\mathcal{T}(\mathbf{b}(t-1), \mathbf{a}(t), \theta(t) = \mathbf{z}))$. An action taken in a slot will influence these two parts of the total reward, and the optimal policy finds a balance between obtaining the immediate reward and obtaining the information for future use.
(i) When $t = 1, 2, ...T-1$,

$$\begin{aligned}
&V_t(\mathbf{b}(t-1)) \\
&= \max_{\mathbf{a}(t)} \{ \sum_{\omega \in \Omega^{\mathbf{s}}} b_\omega(t-1)[\widetilde{R}(\mathbf{a}(t), \mathbf{s}(t-1) = \omega) + \\
&\sum_{\omega' \in \Omega^{\mathbf{s}}} P_{\omega\omega'} \sum_{\mathbf{z} \in \mathbf{Z}^\theta} \Pr(\theta(t) = \mathbf{z} | \mathbf{a}(t), \mathbf{s}(t) = \omega') \\
&\times V_{t+1}(\mathcal{T}(\mathbf{b}(t-1), \mathbf{a}(t), \theta(t) = \mathbf{z}))]\}
\end{aligned} \tag{13}$$

subject to: (9), (10), (12)

(ii) When $t = T$,

$$V_t(\mathbf{b}(t-1)) = \max_{\mathbf{a}(t)} \sum_{\omega \in \Omega^{\mathbf{s}}} b_\omega(t-1) \widetilde{R}(\mathbf{a}(t), \mathbf{s}(t-1) = \omega) \\ \text{subject to: (9), (10), (12)} \tag{14}$$

where $\widetilde{R}(\mathbf{a}(t), \mathbf{s}(t-1) = \omega)$ is given by (8), $P_{\omega\omega'}$ is given by (1), $\Pr(\theta(t) = \mathbf{z} | \mathbf{a}(t), \mathbf{s}(t) = \omega')$ is given by (3), and $\mathcal{T}(\mathbf{b}(t-1), \mathbf{a}(t), \theta(t) = \mathbf{z})$ is given by (4) and (5).

The optimal policy could be obtained from the value function. We use the incremental pruning algorithm to solve the value function. Detailed algorithm and its complexity analysis could be referred to [8].

B. *Myopic policy*

Although the optimal scheduling policy for cooperative spectrum sensing can be derived from the value function, the required computation complexity grows tremendously with the numbers of SUs and channels, even using the incremental pruning algorithm. Moreover, the optimal scheduling policy requires maintaining a table that specifies the optimal actions in every time slot. Therefore, the table could become very large as the time horizon increases. One solution for this computational complexity problem is the divide and conquer approach. For example, we separate all SUs into two smaller SU groups and also separate all channels into two smaller channel groups. Then, we can carry out two POMDP algorithms each consisting of one group of SUs and one group of channels.

To further address the computational complexity problem, in this work we also consider a myopic scheduling policy for cooperative spectrum sensing, which can be expressed as follows:

$$\mathbf{a}^*(t) = \arg\max_{\mathbf{a}(t)} \sum_{\omega \in \Omega^{\mathbf{s}}} b_\omega(t-1) \widetilde{R}(\mathbf{a}(t), \mathbf{s}(t-1) = \omega)$$

subject to: (9), (10), (12)

Essentially, in our myopic policy, BS aims to maximize its immediate expected reward in each time slot t. (Notice that BS also uses (4) and (5) to update its belief state $b_\omega(t)$). Our following simulation results show that the myopic scheduling policy can achieve a comparable performance as that of the optimal policy based on POMDP formulation.

V. Policy Structures of A Simplified System Model

In the previous section we show the methods for obtaining the optimal policy and the myopic policy. However, obtaining the optimal policy requires a recursive computation of the value function and the BS has to keep a table containing all the action profile, which could be quite huge when the system dimensionality increases. Furthermore, although the myopic sensing reduces the computation complexity at the expense of loosing optimality, the BS has to keep track of the belief vector and update it at each time slot. In this section, we aim to seek an efficient way to compute the optimal and myopic decisions by exploiting the inherent properties of the problem and the structure of the value function.

We focus on a simplified system model in this section. We consider a network consists of two SUs (i.e. $M = 2$) and two channels (i.e. $N = 2$) with a BS. The other system settings are the same as mentioned in section II with the additional assumptions as stated in 1) below.

1) Limited Action: Here we limit the choice of a^I to obtain some insights of the problem studied. We assume the sensing duration $k = \Delta L$ for all t, i.e. at each time slot t the BS will assign each SU to sense one channel. Under the assumptions above, the actions can be redefined as $\widetilde{a}^I(t)$:

$$\widetilde{a}^I(t) = \begin{bmatrix} a^1_{11}(t) & a^1_{12}(t) \\ a^1_{21}(t) & a^1_{22}(t) \end{bmatrix} \tag{15}$$

where $a^1_{mn}(t) \in \{0, 1\}$, $\forall m, n$, and $\sum_{n=1}^{N} a^1_{mn}(t) = 1$, $\forall m$. For presentation convenience, we define set $\mathbb{A} \triangleq \{0, 1, 2, 3\}$, where $\widetilde{a}^I(t) \in \mathbb{A}$ and

$$\begin{aligned}
\widetilde{a}^I(t) = 0 &\triangleq [a^1_{11}(t) = 1 \;\; a^1_{12}(t) = 0; a^1_{21}(t) = 1 \;\; a^1_{22}(t) = 0] \\
\widetilde{a}^I(t) = 1 &\triangleq [a^1_{11}(t) = 0 \;\; a^1_{12}(t) = 1; a^1_{21}(t) = 0 \;\; a^1_{22}(t) = 1] \\
\widetilde{a}^I(t) = 2 &\triangleq [a^1_{11}(t) = 1 \;\; a^1_{12}(t) = 0; a^1_{21}(t) = 0 \;\; a^1_{22}(t) = 1] \\
\widetilde{a}^I(t) = 3 &\triangleq [a^1_{11}(t) = 0 \;\; a^1_{12}(t) = 1; a^1_{21}(t) = 1 \;\; a^1_{22}(t) = 0].
\end{aligned}$$

When $\widetilde{a}^I(t) = 0$, the BS assigns both SUs to cooperatively sense channel 1; similarly both SUs sense channel 2 when $\widetilde{a}^I(t) = 1$. Since the two SUs are homogeneous, to the decision maker $\widetilde{a}^I(t) = 2$ and $\widetilde{a}^I(t) = 3$ have the same meaning that both channel 1 and channel 2 are sensed individually by a SU. Therefore we can remove one of them and express the actions as $\widetilde{a}^I(t) \in \mathbb{A} \triangleq \{0, 1, 2\}$.

As mentioned in section III, we assume every SU being scheduled to sense channel n will tune to the same sensor operating point. As a result the operating point selection action a^{II} can be removed from the problem formulation. We denote the miss detection probability under the chosen sensor operating point as P_{MD}, and the false alarm probability for channel n when sensed by one SU as P^1_{FA} and sensed by two SUs cooperatively as P^2_{FA}. Then the action profile at time slot t is reduced from a vector $\mathbf{a}(t)$ to a scalar $\widetilde{a}(t)$ and $\widetilde{a}(t) = \widetilde{a}^I(t)$.

2) Belief Vector Separation: The dimension of the belief vector $\mathbf{b}(t)$ grows exponentially with the number of channels, which will introduce difficulties to calculation and analysis. Since the channels are independent and the occupancy of which evolve according to their own transition probabilities, we can alternatively adopt the marginal distribution which also serves as the sufficient statistic of the system state [6]. Denote the marginal distribution as $\widetilde{\mathbf{b}}(t) \triangleq \{b^1_0(t), ..., b^N_0(t)\}$, where $b^n_0(t)$ denotes the conditional probability that channel n is idle in time slot t given all past decisions and observations:

$$b^n_0(t) \triangleq \Pr(s_n(t) = 0|\widetilde{\mathbf{b}}(0), \{\widetilde{a}(\tau), \theta(\tau)\}^t_{\tau=1}) \in [0, 1]. \quad (16)$$

The belief update of each channel can be separated as well. Similar to (4), the updated belief vector is:

$$\widetilde{\mathbf{b}}(t) \triangleq \widetilde{\mathcal{T}}(\widetilde{\mathbf{b}}(t-1), \widetilde{a}(t), \theta(t)). \quad (17)$$

3) Fixed Reward Function: Since in the simplified system model we adopt fixed sensing duration for all time slot t, the variable k disappears from the reward function and the reward function can then be simplified as

$$\begin{aligned} &R_n(\widetilde{a}(t), \theta_n(t)) \\ &= \begin{cases} \frac{L-\Delta L-\eta}{L}, & \text{if } \sum_{m=1}^M a^1_{mn}(t) > 0, \theta_n(t) = 0 \\ 0, & \text{otherwise} \end{cases} \end{aligned} \quad (18)$$

We define $R = \frac{L-\Delta L-\eta}{L}$ for easy presentation in the rest of the paper.

4) Unconstrained POMDP: It has been proved in [6] that by choosing the sensor operating point to be $P_{MD} = \zeta$, the optimal access policy is to transmit if the sensing outcome is idle and not to transmit otherwise. In this case the constraint (12) can be relaxed. To this end, all the constraints (9), (10) and (12) can be relaxed from the original problem formulation and value function. Then the problem becomes an unconstrained POMDP.

A. The Structure of the Myopic Policy

In this section we aim to exploit the structure of the myopic policy. The myopic policy loses optimality but can serve as a suboptimal solution for the problem with lower computation complexity. We first show that some of the actions will not become an option in the myopic policy if certain condition is satisfied. Then we will further show that the myopic policy admits a simple structure similar to the one proposed in [13] and [12]. Here the occupancies of the two channels evolve according to two independent and identical Markov processes. As a result the state transition probabilities of both channels are the same, i.e. $P^n_{ij} = P_{ij}\ \forall n$.

Proposition 1: Consider two i.i.d channels. At any time slot t, the myopic action determined by the BS will not include $\widetilde{a}(t) = 2$, i.e. the myopic action can only be either $\widetilde{a}(t) = 0$ or $\widetilde{a}(t) = 1$, if the following condition (C0) holds:

$$\text{(C0):} \qquad 2P^1_{FA} - 1 > P^2_{FA} \quad (19)$$

Proof: See Appendix A.

Proposition 1 provides a condition for the BS to choose either $\widetilde{a}(t) = 0$ or $\widetilde{a}(t) = 1$ as the myopic action at any time slot t. The intuitive physical meaning of (C0) is that if the false alarm probability obtained by cooperative sensing P^2_{FA} is small enough, i.e. lower than some threshold, then performing cooperative sensing by two SUs on the same channel will be more beneficial than other actions. In this case, at any time slot t only one channel will be sensed cooperatively by the two SUs. The BS will not ask the SUs to sense both channels at the same time since the expected immediate reward obtained is smaller than the one obtained under cooperative sensing. The dimensionality of the action space is further reduced by this proposition.

Corollary 1: Assume condition (C0) holds. At any time slot t, the myopic action for the BS is to:

(i) assign both SUs to sense channel n^* where $n^* = \arg\max_n b^n_0(t-1)$, when $P_{00} > P_{10}$; (ii) assign both SUs to sense channel n^* where $n^* = \arg\min_n b^n_0(t-1)$, when $P_{00} < P_{10}$; (iii) assign both SUs to sense either one of the channels, when $P_{10} = P_{00}$.

Proof: See Appendix A.

This corollary provides a simple way to decide the myopic policy for time slot t, if the BS keeps track of the belief vector of both channels.

Proposition 2: Consider two i.i.d channels. Assume condition (C0) holds and the false alarm probability P^2_{FA} given by the sensor operating point satisfies

$$P^2_{FA} < \frac{P_{10}P_{01}}{P_{00}P_{11}}, \quad \text{if} \quad P_{00} > P_{10} \quad (20)$$

and

$$P^2_{FA} < \frac{P_{00}P_{11}}{P_{10}P_{01}}, \quad \text{if} \quad P_{00} < P_{10}, \quad (21)$$

Then at any slot t, based on the action the BS took and the observation obtained in the previous slot $t-1$, the myopic action $\widetilde{a}^*(t)$ for the BS at the current slot t is given by:

(i) When $P_{00} > P_{10}$, which means the occupancy of the channel is positively correlated across time,

$$\widetilde{a}^*(t) = \begin{cases} \widetilde{a}(t-1) & \text{if} \quad \theta_{\widetilde{a}(t-1)+1} = 0 \\ 1 - \widetilde{a}(t-1) & \text{otherwise} \end{cases} \quad (22)$$

(ii) When $P_{00} < P_{10}$, which means the occupancy of the channel is negatively correlated across time,

$$\widetilde{a}^*(t) = \begin{cases} 1 - \widetilde{a}(t-1) & \text{if} \quad \theta_{\widetilde{a}(t-1)+1} = 0 \\ \widetilde{a}(t-1) & \text{otherwise} \end{cases} \quad (23)$$

which is independent of the belief vector and maximizes the expected immediate reward at time slot t.

Proof: See Appendix B.

Here we arrive at a similar result as mentioned in [13] [12] and in addition prove the case of $P_{00} > P_{10}$. Proposition 2 reveals that for i.i.d channels when the false alarm probability satisfies (20) under the case of $P_{00} > P_{10}$, the myopic policy at time slot t is to stay in the same channel $\widetilde{a}(t-1)$ as in time slot $t-1$ by receiving an ACK (i.e. $\theta_{\widetilde{a}(t-1)+1} = 0$), and switch to the other channel by receiving $\theta_{\widetilde{a}(t-1)+1} = 1$ and $\theta_{\widetilde{a}(t-1)+1} = 2$. This proposition can further reduce the computation complexity of the myopic policy since it does not require the BS to keep track of and update the belief vector. A general assumption on the initial information of the system is that only the stationary distribution of the underlying Markov chain is available [6] [12]. Since the two Markov chains of the two channels have the same parameters, the BS should randomly choose a channel to sense in the first time slot.

B. The Structure of the Optimal Policy

The analysis of the myopic policy has provided us some insights about the property and the structure of the value function. We are interested to see whether such properties and structure results can be extended to the optimal policy. Obtaining the optimal policy requires the usage of backward induction to solve the value function, which is much more complicated than obtaining the myopic policy. We show next some properties of the optimal policy which can help better understand the problem studied and reduce the complexity. The assumptions and the system parameters mentioned at the beginning of section V are also applied here.

Proposition 3: Consider two i.i.d channels and the system states of which evolve independently. If $P_{00} > P_{10}$, then the value function given in (13)(14) is monotonically increasing with the belief vector $\widetilde{\mathbf{b}}(t-1) = \{b_0^1(t-1), b_0^2(t-1)\}$, i.e. $V_t(\widetilde{\mathbf{b}}(t-1)) \geq V_t(\widetilde{\mathbf{c}}(t-1))$ for $\widetilde{\mathbf{b}}(t-1) \geq \widetilde{\mathbf{c}}(t-1)$, where $\widetilde{\mathbf{c}}(t-1) \triangleq \{c_0^1(t-1), c_0^2(t-1)\}$ denotes another belief vector.

Proof: See Appendix C.

The condition $P_{00} > P_{10}$ provides a sufficient condition for the value function to be monotonically increasing with the belief vector. Similar to [14], the rationale behind Proposition 3 is that a larger current belief vector implies a higher probability the channel will be idle in future slots and hence higher rewards can be received due to the occupancy of the channel is positively correlated across time. Next we will show another monotonic property of the value function.

Proposition 4: Consider two i.i.d channels and the system states of which evolve independently. If $P_{00} > P_{10}$ and channel n is sensed at time slot t, then the value function given in (13)(14) is monotonically decreasing with the false alarm probability $P_{FA}(n,t)$, i.e. $V_t(\widetilde{\mathbf{b}}(t-1), P_{FA}(n,t) = P_a) \geq V_t(\widetilde{\mathbf{b}}(t-1), P_{FA}(n,t) = P_b)$ for $P_a \leq P_b$.

Proof: See Appendix D.

Proposition 4 reveals that if we want to obtain higher immediate and future rewards, we should improve our sensing performance to reduce the false alarm probability. Intuitively, smaller false alarm probability means the BS may miss fewer spectrum chances and hence receive higher rewards. In our system model, smaller false alarm probability can be obtained by assigning the SUs to sense the same channel cooperatively. In other words, the BS has a temptation to require the SUs to perform cooperative sensing instead of sense different channel individually. We show next the sufficient condition for the BS to always assign the SUs to cooperatively sense the same channel.

Theorem 1: Assume $P_{00} > P_{10}$ and condition (C0) holds. Then for the system analyzed in this section, the optimal policy $\widetilde{a}^*(t)$ for the BS at any time slot t satisfies:

$$\widetilde{a}^*(t) = 0 \quad \text{or} \quad \widetilde{a}^*(t) = 1, \quad \forall t \tag{24}$$

In other words, action $\widetilde{a}(t) = 2$ will never be chosen in the optimal policy.

Proof: See Appendix E.

Theorem 1 extends Proposition 1 in the myopic case. It reveals that when the channel occupancy is positively correlated across the time and condition (C0) holds, the BS will always require the SUs to sense the same channel cooperatively. In this case, the advantage of cooperative sensing is obvious.

It is natural to ask whether the simple solution structure of the myopic policy shown in Proposition 2 can be applied to the optimal policy. It has been shown to be true [12] and we have the following conclusion.

Theorem 2: The myopic policy is optimal for the problem we study in this section, i.e. for two i.i.d channels, if the following conditions hold

(i) $P_{00} > P_{10}$; (ii) condition (C0); (iii) $P_{FA}^2 < \frac{P_{10}P_{01}}{P_{00}P_{11}}$, then the optimal policy is given by

$$\widetilde{a}^*(t) = \begin{cases} \widetilde{a}(t-1) & \text{if } \theta_{\widetilde{a}(t-1)+1} = 0 \\ 1 - \widetilde{a}(t-1) & \text{otherwise} \end{cases} \tag{25}$$

Proof: The proof is similar to the proof of Proposition 2 and [12], and hence it is omitted here for brevity. Readers are referred to [12] for details.

Theorem 2 shows the simple structure of the myopic policy is also optimal in this case. The computation complexity of the optimal policy can hence be much reduced since the simple structure does not require the recursive computation of the value function and updating the belief vector.

Based on the structure of the problem and numerical results, we conjecture that all the propositions, corollary and theorems for the myopic and optimal policies will hold for $M > 2$ and $N > 2$, under certain conditions with similar form as (C0). We will show the condition for the situation that $M = 3$ and $N = 3$ in the next subsection.

C. Analysis for the System with Multiple SUs and Multiple Channels

In this subsection, we study the solution structure for a more complicated system, which has multiple SUs and multiple channels. Specifically, we assume $M = 3$ and $N = 3$ in this system. Similar to the situation that $M = 2$ and $N = 2$, we first investigate the condition that the BS assigns all the SUs to sense one channel cooperatively.

Proposition 5: Consider three i.i.d channels. At any time slot t, the myopic action determined by the BS is to assign all the SUs to sense only one of the channels, if condition (C0) and the following condition (C1) hold:

$$\text{(C1):} \qquad P_{FA}^1 + P_{FA}^2 - 1 > P_{FA}^3, \tag{26}$$

where P_{FA}^3 denotes the false alarm probability achieved by three SUs sensing cooperatively.

Proof: See Appendix F.

Similar to Proposition 1, Proposition 5 shows that if the false alarm probability P_{FA}^3 is small enough, i.e. lower than some threshold, then performing cooperative sensing by all the three SUs on the same channel will be more beneficial than other actions. When (C1) holds, all the corollaries, propositions we established in section V subsection A can be easily extended to the current system with $M = 3$ and $N = 3$, following the same argument of proof. Specifically, for the case that $P_{00} > P_{10}$, the propositions and theorems established in the previous section also hold, which can be easily proved by incorporating our argument of proof and the results in [18]. In other words, the optimality of the myopic policy also holds for the system with $M = 3$ and $N = 3$.

Another interesting property related to the cooperative sensing scheduling problem is given in the following corollary:

Corollary 2: Let P_{FA}^m denote the false alarm probability achieved by m SUs sensing cooperatively. Assume $m \in [1, +\infty)$ is a continuous variable and P_{FA}^m is convex and nonincreasing on m, then (C1) will not hold and the BS will never assign all the SUs to sense one channel cooperatively if (C0) does not hold.

Proof: See Appendix F.

Note that the assumption that P_{FA}^m is convex and nonincreasing on m holds for most of the cases one may encounter [5] and hence is reasonable. The physical meaning of Corollary 2 is obvious from the proof: as long as assigning all the m SUs to sense one channel cooperatively is not that beneficial in terms of the value function, then the BS has no reason to assign more SUs, i.e. larger than m, to sense one channel cooperatively. On the contrary, the BS should spread out the SUs to sense more channels in order to gain the largest value function. Generally, when m is large, the problem of finding the best action for this sensing scheduling problem will become a difficult combinatorial problem. However the proof of Corollary 2 reveals that by making use of the property of P_{FA}^m, we can find the best action with low complexity even when m is large and P_{FA}^m is not always convex on the domain of m. The results of $M = 3$ and $N = 3$ provide us insight for extending the results in section V. Extensions to the general M and N will be left for our future work.

VI. Numerical and Simulation Results

We set the CR network with a set of SUs $\mathcal{M} = \{1, 2\}$, and a set of orthogonal frequency channels $\mathcal{N} = \{1, 2\}$. Both of the channels have the same unit bandwidth. We set $\Delta L = 0.2L$ as the sensing duration for each channel. We also set $\eta = 0.1L$ as the duration for BS's sensing decision, receiving sensing results from SUs, and channel allocation decision. We assume Rayleigh fading channels with the same average receiver SNR=10dB [5] [11] and we use the same time bandwidth product $u = 5$. In Figure 3, the time horizon is 15 slots, and in the other simulations the time horizon is 10 slots. In these simulations we set every channel as homogeneous for different SUs. In this case there may exist several optimal policies, the BS will just pick one of them randomly.

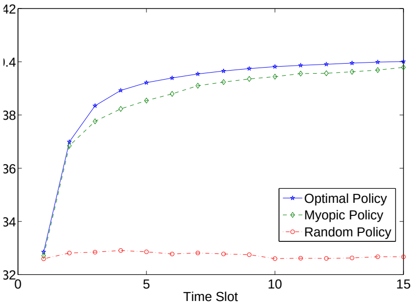

Fig. 3. SUs throughput performance comparison with $P_{00}^1 = P_{00}^2 = 0.8, P_{10}^1 = P_{10}^2 = 0.2$, and the same $\zeta = 0.1$

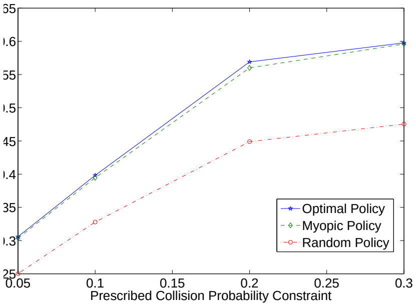

Fig. 4. SUs throughput performance comparison with $P_{00}^1 = P_{00}^2 = 0.8, P_{10}^1 = P_{10}^2 = 0.2$

To compare with the optimal and myopic policies, we consider a simple random policy which randomly picks an action while it satisfies all the constraints (9), (10), (12).

Figure 3 shows the throughput comparison of the theoretical results of optimal policy, the simulation results of myopic policy and random policy. We set the scenario that both channels have the same statistical behavior (i.e. $P_{00}^n = 0.8, P_{10}^n = 0.2,\ n = 1, 2$), and the same prescribed collision probability $\zeta = 0.1$. This figure shows the advantages of the optimal policy and myopic policy over the random one with time horizon increasing. It also shows the optimal policy and myopic policy have very similar throughput performance in this scenario.

In Figure 4, we also set the scenario that both channels have the same statistical behavior (i.e. $P_{00}^n = 0.8, P_{10}^n = 0.2,\ n = 1, 2$), and the same prescribed collision probability ζ from 0.05 to 0.3. This figure shows that with the increase of the prescribed collision probability ζ, SUs' throughput becomes larger because of PUs' more collision tolerance. Nevertheless,

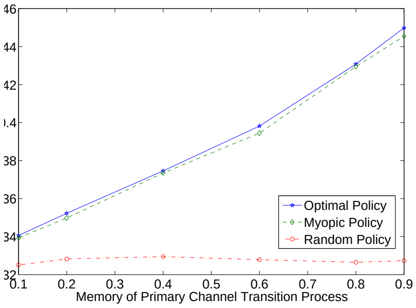

Fig. 5. SUs throughput performance comparison with $\pi_0^1 = \pi_0^2 = 0.5$, and the same $\zeta = 0.1$

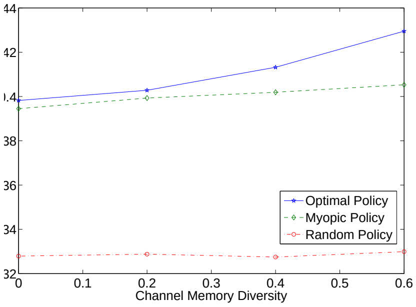

Fig. 6. SUs throughput performance comparison with $\pi_0^1 = \pi_0^2 = 0.5$, and the same $\zeta = 0.1$, and $\mu_1 + \mu_2 = 1.2$

when the prescribed collision probability reaches some level, SUs' throughput will stop increasing, this is because it has already arrived at the maximum point of the primary channels' unutilized opportunity.

In Figure 5, we study the SUs' throughput performance under different memories of PU channel transition process. According to [7], the memory of channel n's transition process is defined as $\mu_n = 1 - P_{01}^n - P_{10}^n,\ n \in \mathcal{N}$, which is the probability of remaining in the same channel state. In this paper, we set $\mu_n > 0,\ n \in \mathcal{N}$, which means all the channels have positive transition process memories. The larger the memory, the higher tendency a channel will remain in the same state. We also consider the case of both channels having the same statistical behavior, the same stationary idle probability (i.e. $\pi_0^n = 0.5,\ n = 1, 2$), and the same prescribed collision probability $\zeta = 0.1$. Figure 5 shows that when the channels' transition process memories grow larger, the throughput performance of optimal policy and myopic policy grow much better than the random policy. This indicates that if all the channels have positive transition process memories, then the larger the memories, the better throughput performance we can get by using our optimal and myopic policies.

In Figure 6, we study the SUs' throughput performance when the two channels' statistical behaviors become different. Here, we set the prescribed collision probability ζ as 0.1 for each channel, and we set the sum of the two channels' transition process memories as a constant $(i.e.\ \mu_1 + \mu_2 = 1.2)$. Besides, their stationary idle probabilities are the same $(i.e.\ \pi_0^1 = \pi_0^2 = 0.5)$. This figure shows that although their stationary idle probabilities are the same and the sum of the two channels' transition process memories does not change, using the optimal policy can obtain a better throughput performance than the myopic policy when the diversity of the two channels' transition process memories (i.e. $|\mu_1 - \mu_2|$) grows larger. This is because that when the two channels' statistical behaviors are similar, the myopic policy will be similar to the optimal policy. However, when the two channels' statistical behaviors become more different, the myopic policy will get different decisions.

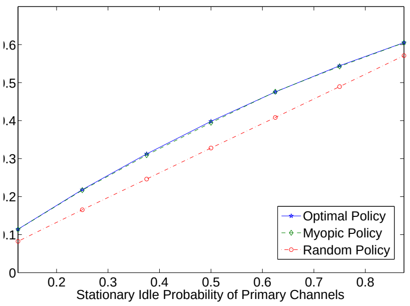

Fig. 7. SUs throughput performance comparison with the same $\zeta = 0.1$, and $\mu_1 = \mu_2 = 0.6$

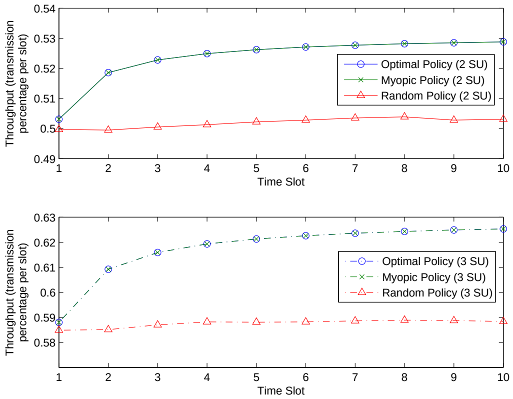

Fig. 8. Performance comparison of the optimal policy and the myopic policy (in the upper figure, the network has two SUs; in the lower figure, the network has three SUs).

Figure 7 shows the SUs' throughput performance under different stationary idle probability of PU channels. Here we set the two PU channels with the same statistical behavior and the same prescribed collision probability $\zeta = 0.1$, and then we change their stationary idle probability while maintaining their channel transition process memories $(i.e.\ \mu_1 = \mu_2 = 0.6)$. It is shown from the figure that when PU channels' stationary idle probability increases, SUs' throughput increases accordingly. This is because SUs will get more opportunities as the PU channels' idle probability increases.

In Figure 8, we compare the performance of the optimal policy and the myopic policy under the simplified system model in section V. We first consider the network with 2 SUs and 2 channels, as studied in the theoretical analysis. From the upper figure one can notice that under two i.i.d channels and the same prescribed collision probability $\zeta = 0.1$, the throughput obtained by the myopic policy agrees with the one obtained by the optimal policy. Our theoretical analysis is thus supported by the numerical results. In the lower figure, we consider a more general situation. We still utilize the same simplified model used in the upper figure, but we add one more SU to the network, i.e. 3 SUs present in the network. We also assume each SU will spend ΔL in each slot as the sensing duration, and the BS may assign each of the SUs to sense one of the two channels in each slot. As proved in section V subsection C, using the same parameters setting for the simulation described in the beginning of this section, we find this new network has a similar behavior as the previous network with 2 SUs, i.e. in each slot the BS will simply assign all the 3 SUs to sense either of the two channels cooperatively and no other actions will be chosen. As a result the simple solution structure obtained for the network with 2 SUs are actually proved to be applicable to the network with 3 SUs, and the lower figure shows the performance of the myopic policy matches the performance of the optimal policy for the network with 3 SUs. Figure 8 implies that although the analysis in section V is based on a simplified model, it can be further extended to the more general case.

VII. Conclusion

In this paper, we study the cooperative sensing scheduling problem in cognitive radio networks. We first formulate this problem as a POMDP which aims to maximize the total CR system throughput with the guarantee of primary users' prescribed collision probability. Then, we derive the optimal policy and a myopic policy that determines which SUs sense which channels with what miss detection probability and false alarm probability. We further analytically study the solution structure and properties of the value function. We have shown that under the simplified system model, several interesting properties hold for the value function and some simple but robust methods exist for finding the optimal action. A generally hard combinatorial problem is analytically studied, and the solution of the POMDP can be obtained by the simple method with low-complexity. Although the system model for analysis is simple, it provides us with many useful insights. Finally, numerical and simulation results are provided to illustrate the throughput performance of our optimal and myopic scheduling policies for cooperative spectrum sensing.

The direction for the future work is to extend the analysis of the simplified model to the general case, where more than three SUs and channels are present in the network and more actions can be chosen from for the BS.

Appendix A
Proof of Proposition 1 and Corollary 1

Proof of Proposition 1: We prove by induction. The myopic policy only maximizes the expected immediate reward while ignoring the future reward. Note that only when $s_n(t) = 0$ and $\theta_n(t) = 0$ would the BS receive nonzero reward. Let $V_t^{(\widetilde{a}(t))}(\widetilde{\mathbf{b}}(t-1))$ denote the maximum expected immediate reward received at time slot t given the belief vector $\widetilde{\mathbf{b}}(t-1)$ and action $\widetilde{a}(t)$, then from (14) we have equation (27).

For the three different actions of $\widetilde{a}(t)$, we have the following equations after some manipulations:

$$V_t^{\widetilde{a}(t)=0}(\widetilde{\mathbf{b}}(t-1)) = \sum\nolimits_{i=0}^{1} b_i^1(t-1)P_{i0}(1-P_{FA}^2)R \quad (29)$$

$$V_t^{\widetilde{a}(t)=1}(\widetilde{\mathbf{b}}(t-1)) = \sum\nolimits_{i=0}^{1} b_i^2(t-1)P_{i0}(1-P_{FA}^2)R \quad (30)$$

$$V_t^{\widetilde{a}(t)=2}(\widetilde{\mathbf{b}}(t-1)) = \sum_{n=1}^{2}\sum_{i=0}^{1} b_i^n(t-1)P_{i0}(1-P_{FA}^1)R. \quad (31)$$

Assume $\max\{V_t^{\widetilde{a}(t)=0}(\widetilde{\mathbf{b}}(t-1)), V_t^{\widetilde{a}(t)=1}(\widetilde{\mathbf{b}}(t-1))\} = V_t^{\widetilde{a}(t)=0}(\widetilde{\mathbf{b}}(t-1))$. It implies $\sum_{i=0}^{1} b_i^1(t-1)P_{i0} > \sum_{i=0}^{1} b_i^2(t-1)P_{i0}$. By comparing the difference between $V_t^{\widetilde{a}(t)=0}(\widetilde{\mathbf{b}}(t-1))$ and $V_t^{\widetilde{a}(t)=2}(\widetilde{\mathbf{b}}(t-1))$, we have equation (28). From (28), we can arrive at $V_t^{\widetilde{a}(t)=0}(\widetilde{\mathbf{b}}(t-1)) - V_t^{\widetilde{a}(t)=2}(\widetilde{\mathbf{b}}(t-1)) \geq 0$ when condition (C0) mentioned in Proposition 1 holds. For the case that $\max\{V_t^{\widetilde{a}(t)=0}(\widetilde{\mathbf{b}}(t-1)), V_t^{\widetilde{a}(t)=1}(\widetilde{\mathbf{b}}(t-1))\} = V_t^{\widetilde{a}(t)=1}(\widetilde{\mathbf{b}}(t-1))$, the same conclusion can be obtained in a similar way. This completes the proof.

Proof of Corollary 1: Let $\hat{b}_0^n(t)$ denote the idle probability of channel n at slot t after the state transition period, i.e.

$$\hat{b}_0^n(t) = \sum_{i=0}^{1} b_i^n(t-1)P_{i0} = P_{10} + b_0^n(t-1)(P_{00} - P_{10}). \quad (32)$$

Applying (4)(5), we can derive the upper bound and lower bound of $\hat{b}_0^n(t)$ as follows:

$$\begin{aligned} &P_{10} \leq \hat{b}_0^n(t) \leq P_{00}, \text{ if } \ P_{10} < P_{00} \\ &P_{00} \leq \hat{b}_0^n(t) \leq P_{10}, \text{ if } \ P_{00} < P_{10}. \end{aligned} \quad (33)$$

From Proposition 1 we know that only action $\widetilde{a}(t) = 0$ and action $\widetilde{a}(t) = 1$ are possible for the BS to choose. For $P_{00} > P_{10}$, $\hat{b}_0^n(t)$ increases with $b_0^1(t-1)$. From (29) and (30) one can notice that a larger $b_0^1(t-1)$ will result in a higher immediate reward. For $P_{10} > P_{00}$, $\hat{b}_0^n(t)$ decreases with $b_0^1(t-1)$, which means a smaller $b_0^1(t-1)$ will result in a higher immediate reward. For $P_{10} = P_{00}$, $\hat{b}_0^n(t) = P_{10} = P_{00}$ and the immediate reward is independent of the belief vector and hence the same for $\widetilde{a}(t) = 0$ and $\widetilde{a}(t) = 1$.

Appendix B
Proof of Proposition 2

We first consider the case that $P_{00} > P_{10}$. Without loss of generality we assume $\widetilde{a}(t-1) = 0$. The key of the proof here is to show that different observation received in slot $t-1$ will result in different ordering of the immediate reward obtained by the actions in slot t. Consider first $\theta_1(t) = 0$, we know that channel 1 is idle at time $t-1$ and $b_0^1(t) = 1$ in this case. The immediate reward of staying in channel 1 is $P_{00}R$, and the immediate reward of switching to channel 2 is $(P_{10} + b_0^2(t-1)(P_{00} - P_{10}))R$ with the inequality

$$(P_{10} + b_0^2(t-1)(P_{00} - P_{10}))R \leq P_{00}R \quad (34)$$

where $b_0^2(t) \in [0,1]$ and the inequality follows from $P_{00} > P_{10}$. As a result the BS should stay in channel 1 to get the maximum immediate reward.

$$V_t^{\widetilde{a}(t)}(\widetilde{\mathbf{b}}(t-1)) = \sum_{n=1}^{2}\sum_{i=0}^{1} b_i^n(t-1)P_{i0}\Pr(\theta_n(t)=0|\widetilde{a}(t),s_n(t)=0)R_n(\widetilde{a}(t),\theta_n(t)=0) \tag{27}$$

$$\begin{aligned} V_t^{\widetilde{a}(t)=0}(\widetilde{\mathbf{b}}(t-1)) - V_t^{\widetilde{a}(t)=2}(\widetilde{\mathbf{b}}(t-1)) &= \sum_{i=0}^{1} b_i^1(t-1)P_{i0}(P_{FA}^1 - P_{FA}^2) - \sum_{i=0}^{1} b_i^2(t-1)P_{i0}(1-P_{FA}^1) \\ \geq \sum_{i=0}^{1} b_i^1(t-1)P_{i0}(P_{FA}^1 - P_{FA}^2) - \sum_{i=0}^{1} b_i^1(t-1)P_{i0}(1-P_{FA}^1) &= \sum_{i=0}^{1} b_i^1(t-1)P_{i0}(2P_{FA}^1 - P_{FA}^2 - 1) \end{aligned} \tag{28}$$

Consider then $\theta_1(t)=1$ and $\theta_1(t)=2$. Careful inspection of these two observations reveals that these two observations actually refer to the same situation that "no ACK is received after sensing the channel 1". Hence we make some manipulation of the observation probability and obtain

$$\begin{aligned} &\Pr(\widetilde{\theta}_n(t)|\widetilde{a}(t),s_n(t)) \\ &= \begin{cases} 1, \ \ \text{if} \ \ \sum_{m=1}^{2} a_{mn}^1(t)=0, \ \widetilde{\theta}_n(t)=2 \\ 1-P_{FA}(n,t), \\ \quad \text{if} \ \ \sum_{m=1}^{2} a_{mn}^1(t)=1, \ s_n(t)=0, \ \widetilde{\theta}_n(t)=0 \\ P_{FA}(n,t), \\ \quad \text{if} \ \ \sum_{m=1}^{2} a_{mn}^1(t)=1, \ s_n(t)=0, \ \widetilde{\theta}_n(t)=1 \\ 1, \\ \quad \text{if} \ \ \sum_{m=1}^{2} a_{mn}^1(t)=1, \ s_n(t)=1, \ \widetilde{\theta}_n(t)=1 \end{cases} \end{aligned} \tag{35}$$

where the new observation and its probability can provide the same information as the original one. $\widetilde{\theta}_n(t)$ is the modified observations, where $\widetilde{\theta}_n(t)=0$ denotes ACK is received, $\widetilde{\theta}_n(t)=1$ denotes no ACK is received after sensing channel n, and $\widetilde{\theta}_n(t)=2$ denotes the BS decides not to sense the channel and observes nothing. The modification has not changed the system model but will allow us to express the update of the belief vector in a simpler form for the rest of the paper.

At the beginning of time slot $t-1$, the belief vector is $\widetilde{\mathbf{b}}(t-2)=\{b_0^1(t-2), b_0^2(t-2)\}$. From (32) and (35), the belief vector at time slot t can be expressed as

$$b_0^1(t-1) = \frac{\hat{b}_0^1(t-1)P_{FA}^2}{\hat{b}_0^1(t-1)P_{FA}^2 + (1-\hat{b}_0^1(t-1))} \tag{36}$$

$$b_0^2(t-1) = \hat{b}_0^2(t-1). \tag{37}$$

Then considering $P_{00} > P_{10}$ and the condition (20), we can obtain the inequality that $b_0^2(t-1) \geq b_0^1(t-1)$ by applying (33). Again from $P_{00} > P_{10}$ we arrive at $\hat{b}_0^2(t)R \geq \hat{b}_0^1(t)R$, which means the maximum immediate reward at time slot t is obtained by switching to channel 2 when $\theta_1(t)=1$ and $\theta_1(t)=2$ is observed. The structure for the case that $P_{00} < P_{10}$ can be similarly obtained.

Appendix C
Proof of Proposition 3

We prove by induction. Suppose the proposition holds for all slots $t+1 \leq T$, i.e. $V_{t+1}(\widetilde{\mathbf{b}}(t)) \geq V_{t+1}(\widetilde{\mathbf{c}}(t))$ where $\widetilde{\mathbf{b}}(t) \geq \widetilde{\mathbf{c}}(t)$. We show next the proposition holds for slot t. As seen from (17), $\widetilde{\mathbf{b}}(t-1) \geq \widetilde{\mathbf{c}}(t-1)$ implies $\widetilde{\mathcal{T}}(\widetilde{\mathbf{b}}(t-1),\widetilde{a}(t),\widetilde{\theta}(t)) \geq \widetilde{\mathcal{T}}(\widetilde{\mathbf{c}}(t-1),\widetilde{a}(t),\widetilde{\theta}(t))$, $\forall \widetilde{a}(t),\widetilde{\theta}(t)$ due to $P_{00} > P_{10}$. Applying this argument, we then have

$$\begin{aligned} &V_t(\widetilde{\mathbf{b}}(t-1)) \\ &\geq \max_{\widetilde{a}(t)}\{\sum_{n=1}^{2}\sum_{i=0}^{1} b_i^n(t-1)P_{i0}\Pr(\theta_n(t)=0|\widetilde{a}(t),s_n(t)=0) \\ &\times R_n(\widetilde{a}(t),\theta_n(t)=0) \\ &+ \sum_{n=1}^{2}\sum_{i=0}^{1} b_i^n(t-1)\sum_{j=0}^{1} P_{ij}\sum_{k=0}^{2}\Pr(\widetilde{\theta}(t)=k|\widetilde{a}(t),s_n(t)=j) \\ &\times V_{t+1}(\widetilde{\mathcal{T}}(\widetilde{\mathbf{c}}(t-1),\widetilde{a}(t),\widetilde{\theta}(t)))\} \\ &\geq \max_{\widetilde{a}(t)}\{\sum_{n=1}^{2}\sum_{i=0}^{1} c_i^n(t-1)P_{i0}\Pr(\theta_n(t)=0|\widetilde{a}(t),s_n(t)=0) \\ &\times R_n(\widetilde{a}(t),\theta_n(t)=0) \\ &+ \sum_{n=1}^{2}\sum_{i=0}^{1} b_i^n(t-1)\sum_{j=0}^{1} P_{ij}\sum_{k=0}^{2}\Pr(\widetilde{\theta}(t)=k|\widetilde{a}(t),s_n(t)=j) \\ &\times V_{t+1}(\widetilde{\mathcal{T}}(\widetilde{\mathbf{c}}(t-1),\widetilde{a}(t),\widetilde{\theta}(t)))\} = V_t(\widetilde{\mathbf{c}}(t-1)) \end{aligned}$$

where $V_{t+1}(\widetilde{\mathcal{T}}(\widetilde{\mathbf{b}}(t-1),\widetilde{a}(t),\widetilde{\theta}(t))) = V_{t+1}(\widetilde{\mathbf{b}}(t))$, $V_{t+1}(\widetilde{\mathcal{T}}(\widetilde{\mathbf{c}}(t-1),\widetilde{a}(t),\widetilde{\theta}(t))) = V_{t+1}(\widetilde{\mathbf{c}}(t))$. The second inequality follows from (29)(30)(31) and the condition $P_{00} > P_{10}$. Proposition 3 thus follows.

Appendix D
Proof of Proposition 4

We first introduce the following Lemmas.

Lemma 1: The value function given in (13)(14) is convex in the belief vector. Specifically, for the set of value functions $\{V_t(\widetilde{\mathbf{b}}_1(t-1)),...,V_t(\widetilde{\mathbf{b}}_I(t-1))\}$ with their corresponding belief vectors $\{\widetilde{\mathbf{b}}_1(t-1),...,\widetilde{\mathbf{b}}_I(t-1)\}$, the following inequality

$$\sum_{i=1}^{I}\tau_i V_t(\widetilde{\mathbf{b}}_i(t-1)) \geq V_t(\sum_{i=1}^{I}\tau_i\widetilde{\mathbf{b}}_i(t-1)), \tag{40}$$

$\forall\ \tau_i \in [0,1]$, $\sum_{i=1}^{I}\tau_i = 1$ is satisfied, where the set of belief vectors labeled as $\{\widetilde{\mathbf{b}}_1(t-1),...,\widetilde{\mathbf{b}}_I(t-1)\}$ is defined for the convenience of showing the convexity.

Proof: Smallwood and Sondik have demonstrated in [15] that the value function $V_t(\widetilde{\mathbf{b}}_i(t-1))$ is piece-wise linear and convex (PWLC) with respect to the belief vector $\widetilde{\mathbf{b}}_i(t-1)$. Since the value function given in (13)(14) is a standard value function, then Lemma 1 follows.

$$\begin{aligned}&\tau_1 V_{t+1}(\widetilde{\mathcal{T}}(\widetilde{\mathbf{b}}(t-1),\widetilde{\theta}(t)=0))+\tau_2 V_{t+1}(\widetilde{\mathcal{T}}(\widetilde{\mathbf{b}}(t-1),\widetilde{\theta}(t)=1,P_{FA}(t)=P_{FA}^2))\\&\geq V_{t+1}(\widetilde{\mathcal{T}}(\widetilde{\mathbf{b}}(t-1),\widetilde{\theta}(t)=1,P_{FA}(t)=P_{FA}^1))\end{aligned} \tag{38}$$

$$\begin{aligned}&V_{t+1}(\widetilde{\mathcal{T}}(\widetilde{\mathbf{b}}(t-1),\widetilde{\theta}(t),P_{FA}(t)=P_{FA}^2))-V_{t+1}(\widetilde{\mathcal{T}}(\widetilde{\mathbf{b}}(t-1),\widetilde{\theta}(t),P_{FA}(t)=P_{FA}^1))\\&=V_{t+1}(\widetilde{\mathcal{T}}(\widetilde{\mathbf{b}}(t-1),\widetilde{\theta}(t)=2,P_{FA}(t)=P_{FA}^2))-V_{t+1}(\widetilde{\mathcal{T}}(\widetilde{\mathbf{b}}(t-1),\widetilde{\theta}(t)=2,P_{FA}(t)=P_{FA}^1))\geq 0\end{aligned} \tag{39}$$

Lemma 2: Consider the system with only one channel, i.e. $N=1$ (channel 1 only). The action becomes whether to sense channel 1 at slot t. Since the observation can fully distinguish the two different actions, we omit the action term in the belief update equation. In any time slot t, the future rewards $V_{t+1}(\widetilde{\mathcal{T}}(\widetilde{\mathbf{b}}(t-1),\widetilde{\theta}(t),P_{FA}(t)=P_{FA}^1))$ and $V_{t+1}(\widetilde{\mathcal{T}}(\widetilde{\mathbf{b}}(t-1),\widetilde{\theta}(t),P_{FA}(t)=P_{FA}^2))$ satisfy the inequality (38), where $\tau_2=1-\tau_1$ and τ_1 is given by

$$\tau_1=\frac{\sum_{i=0}^{1}b_i^1(t-1)P_{i0}(P_{FA}^1-P_{FA}^2)}{\sum_{i=0}^{1}b_i^1(t-1)(P_{i0}P_{FA}^1+P_{i1})}. \tag{41}$$

Proof: Applying (35), we can obtain the following equality after some algebraic manipulations

$$\begin{aligned}&\tau_1\widetilde{\mathcal{T}}(\widetilde{\mathbf{b}}(t-1),\widetilde{\theta}(t)=0)\\&+\tau_2\widetilde{\mathcal{T}}(\widetilde{\mathbf{b}}(t-1),\widetilde{\theta}(t)=1,P_{FA}(t)=P_{FA}^2)\\&=\widetilde{\mathcal{T}}(\widetilde{\mathbf{b}}(t-1),\widetilde{\theta}(t)=1,P_{FA}(t)=P_{FA}^1)\end{aligned} \tag{42}$$

where τ_1 is given by (41). Then Lemma 2 follows from the conclusion of Lemma 1.

The proof of Proposition 4 is based on the two Lemmas above. Consider the case that $\widetilde{\theta}(t)=2$. Note that from (17), $P_{FA}^2<P_{FA}^1$ implies that $\widetilde{\mathcal{T}}(\widetilde{\mathbf{b}}(t-1),\widetilde{\theta}(t)=2,P_{FA}(t)=P_{FA}^2)\geq\widetilde{\mathcal{T}}(\widetilde{\mathbf{b}}(t-1),\widetilde{\theta}(t)=2,P_{FA}(t)=P_{FA}^1)$. By applying Proposition 3, we have $V_{t+1}(\widetilde{\mathcal{T}}(\widetilde{\mathbf{b}}(t-1),\widetilde{\theta}(t)=2,P_{FA}(t)=P_{FA}^2))\geq V_{t+1}(\widetilde{\mathcal{T}}(\widetilde{\mathbf{b}}(t-1),\widetilde{\theta}(t)=2,P_{FA}(t)=P_{FA}^1))$.

We now compare the value function of different false alarm probability at time slot t. Consider first the action in slot t is not to sense the channel, we have equation (39) where the inequality follows the result in the previous paragraph. Consider then the action is to sense the channel, we have

$$\begin{aligned}&V_{t+1}(\widetilde{\mathcal{T}}(\widetilde{\mathbf{b}}(t-1),\widetilde{\theta}(t),P_{FA}(t)=P_{FA}^2))\\&-V_{t+1}(\widetilde{\mathcal{T}}(\widetilde{\mathbf{b}}(t-1),\widetilde{\theta}(t),P_{FA}(t)=P_{FA}^1))\\&=\{\sum_{i=0}^{1}P_{i0}(1-P_{FA}^2)R-\sum_{i=0}^{1}P_{i0}(1-P_{FA}^1)R\}\\&\quad+\{\sum_{i=0}^{1}b_i^1(t-1)\sum_{j=0}^{1}P_{ij}\sum_{k=0}^{1}\Pr(\widetilde{\theta}(t)=k|s_n(t)=j,\\&P_{FA}(t)=P_{FA}^2)V_{t+1}(\widetilde{\mathcal{T}}(\widetilde{\mathbf{b}}(t-1),\widetilde{\theta}(t)=k,P_{FA}(t)=P_{FA}^2))\\&\quad-\sum_{i=0}^{1}b_i^1(t-1)\sum_{j=0}^{1}P_{ij}\sum_{k=0}^{1}\Pr(\widetilde{\theta}(t)=k|s_n(t)=j,\\&P_{FA}(t)=P_{FA}^1)V_{t+1}(\widetilde{\mathcal{T}}(\widetilde{\mathbf{b}}(t-1),\widetilde{\theta}(t)=k,P_{FA}(t)=P_{FA}^1))\}\\&\geq 0.\end{aligned}$$

For the immediate reward part, it is straightforward that $\sum_{i=0}^{1}P_{i0}(1-P_{FA}^2)R-\sum_{i=0}^{1}P_{i0}(1-P_{FA}^1)R\geq 0$. Then applying Lemma 2 to the future reward part, we arrive at the inequality above. To this end we prove the monotonic property for one channel. By using the similar method used for proving Proposition 3, we can easily extend the result for one channel to two channels in order to complete the proof of Proposition 4. For brevity we omit the procedure here.

Appendix E
Proof of Theorem 1

We first establish the following lemma.

Lemma 3: Consider the actions $\widetilde{a}(t)=0$ and $\widetilde{a}(t)=1$. To obtain $\max\{V_t^{\widetilde{a}(t)=0}(\widetilde{\mathbf{b}}(t-1)),V_t^{\widetilde{a}(t)=1}(\widetilde{\mathbf{b}}(t-1))\}$, $\widetilde{a}(t)$, the channel to be chosen, should be given by $\arg\max_n b_0^n(t-1)$.

Proof: First note that Proposition 3 can be applied to the case of single channel. Without loss of generality we assume $b_0^1(t-1)>b_0^2(t-1)$. Since only one channel will be sensed at any slot t here, the two value functions which have the same action set: to sense the corresponding channel or not, can be separated. Proposition 3 shows the value function with higher channel idle probability has larger value. As a result $V_t^{\widetilde{a}(t)=0}(\widetilde{\mathbf{b}}(t-1))\geq V_t^{\widetilde{a}(t)=1}(\widetilde{\mathbf{b}}(t-1))$ and Lemma 3 follows.

Lemma 3 shows that the monotonic property not only holds between different belief vectors $\widetilde{\mathbf{b}}(t-1)$ and $\widetilde{\mathbf{c}}(t-1)$, but also holds between different components of the belief vector $\widetilde{\mathbf{b}}(t-1)$. We are now ready to prove Theorem 1.

At the last time slot $t=T$, the optimal action is actually the myopic action. As mentioned in Proposition 1, only action 0 and action 1 may be chosen from if condition (C0) holds. Suppose all the optimal actions $\widetilde{a}^*(t+1)$ for $t+1<T$ only contain action 0 and 1. Without loss of generality we assume $b_0^1(t-1)>b_0^2(t-1)$, hence $V_t^{\widetilde{a}(t)=0}(\widetilde{\mathbf{b}}(t-1))\geq V_t^{\widetilde{a}(t)=1}(\widetilde{\mathbf{b}}(t-1))$. For $\widetilde{a}(t)=2$, we know from Lemma 3 that although both channels are sensed, the value of $V_t^{\widetilde{a}(t)=2}(\widetilde{\mathbf{b}}(t-1))$ is determined by the reward obtained in channel 1, since $b_0^1(t-1)\geq b_0^2(t-1)$. Furthermore, the false alarm probability on channel 1 of action $\widetilde{a}(t)=0$ is smaller than that of action $\widetilde{a}(t)=2$, by applying Proposition 4 we have $V_t^{\widetilde{a}(t)=0}(\widetilde{\mathbf{b}}(t-1))>V_t^{\widetilde{a}(t)=2}(\widetilde{\mathbf{b}}(t-1))$, which means $\widetilde{a}^*(t)=0$. Similarly, we have $\widetilde{a}^*(t)=1$ when $b_0^1(t-1)<b_0^2(t-1)$. The BS can choose either 0 or 1 when the idle probabilities of the two channels are equal. To this end we show at slot t the optimal action is also either 0 or 1, hence completes the proof.

Appendix F
Proof of Proposition 5 and Corollary 2

Proof of Proposition 5: We first consider the situation that $P_{00}>P_{10}$. We order the three channels according to $b_0^n(t)$

in a descending way. It can be seen from Corollary 1 that if we have two group of users G_1 and G_2, where $|G_1| > |G_2|$ and $|G|$ denotes the size of group $|G|$, then we should assign G_1 to sense the first channel (the one having the largest belief value) and G_2 to sense the second channel. As a result, denote the three possible sensing scheduling actions of the BS as $\hat{a}(t)$. Specifically, $\hat{a}(t) = 0$ represents all the three SUs sense the first channel, $\hat{a}(t) = 1$ represents two SUs sense the first channel and the remaining one senses the second channel, and $\hat{a}(t) = 2$ represents each SU senses one channel. Note that the action that two SUs sense the second channel and the remaining one senses the first channel will never be selected due to the reason mentioned above.

We follow the similar argument of the proof of Proposition 1. First, it is easy to derive $V_t^{\hat{a}(t)=1}(\widetilde{\mathbf{b}}(t-1)) \geq V_t^{\hat{a}(t)=2}(\widetilde{\mathbf{b}}(t-1))$ since condition (C0) holds. Then similar to (28), we have $V_t^{\hat{a}(t)=0}(\widetilde{\mathbf{b}}(t-1)) - V_t^{\hat{a}(t)=1}(\widetilde{\mathbf{b}}(t-1)) \geq 0$ when condition (C1) mentioned in Proposition 5 holds. Note that if $N = 2$, this result also holds. This completes the proof.

Proof of Corollary 2: The proof is by contradiction. Since we assume condition (C0) does not hold, we have $V_t^{\hat{a}(t)=1}(\widetilde{\mathbf{b}}(t-1)) \leq V_t^{\hat{a}(t)=2}(\widetilde{\mathbf{b}}(t-1))$ and hence $2P_{FA}^1 - 1 < P_{FA}^2$. In order for the action $\hat{a}(t) = 0$ to be the best action, we should have $V_t^{\hat{a}(t)=0}(\widetilde{\mathbf{b}}(t-1)) - V_t^{\hat{a}(t)=2}(\widetilde{\mathbf{b}}(t-1)) \geq 0$, which implies $3P_{FA}^1 - 2 > P_{FA}^3$. If this is true, then by comparing with $2P_{FA}^1 - 1 < P_{FA}^2$, we arrive at $2P_{FA}^2 > P_{FA}^1 + P_{FA}^3$ after some manipulations. However from the convexity of P_{FA}^m, one should have $2P_{FA}^2 \leq P_{FA}^1 + P_{FA}^3$ and contradiction is shown, which means the BS will never assign all the SUs to sense one channel cooperatively. The conclusion that (C1) also will not hold can be proved in the same way.

References

[1] I.F. Akyildiz, W.Y. Lee, M.C. Vuran, S. Mohanty, "Next Generation/Dynamic Spectrum Access/Cognitive Radio Wireless Networks: A Survey," *Comput. Netw.: Int. J. Comput. Telecommun. Netw.* vol. 50, no. 13, pp. 2127-2159, 2006.

[2] S. Haykin, "Cognitive Radio: Brain-Empowered Wireless Communications," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 2, pp. 201-220, 2005.

[3] C. Cordeiro, K. Challapali, D. Birru, S. Shankar, "IEEE 802.22: The First Worldwide Wireless Standard based on Cognitive Radios," *J. Commun.*, vol. 1, no. 1, pp. 60-67, 2006.

[4] S.M. Mishra, A. Sahai, R.W. Brodersen, "Cooperative Sensing among Cognitive Radios," in *Proc. IEEE Int. Conf. Commun.*, vol. 4, pp. 1658-1663, 2006.

[5] K.B. Letaief, W. Zhang, "Cooperative Spectrum Sensing," *Book chapter in Cognitive Wireless Communication Networks. Springer*, pp.115-138, 2007.

[6] Q. Zhao, L. Tong, A. Swami, 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, 2007.

[7] A.T. Hoang, Y.C. Liang, D.T.C. Wong, Y.H. Zeng, R. Zhang, "Opportunistic Spectrum Access for Energy-Constrained Cognitive Radios," *IEEE Trans. Wireless Commun.*, vol. 8, no. 3, 2009.

[8] A. Cassandra, M.L. Littman, N.L. Zhang, "Incremental Pruning: A Simple, Fast, Exact Method for Partially Observable Markov Decision Processes", in *Proc. 13th Conf. Uncertainty in Artificial Intelligence*, pp. 54-61, 1997.

[9] W.Y. Lee, I.F. Akyildiz, "Optimal Spectrum Sensing Framework for Cognitive Radio Networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 10, 2008.

[10] H. Kim, K.G. Shin, "Efficient Discovery of Spectrum Opportunities with MAC-Layer Sensing in Cognitive Radio Networks," *IEEE Trans. Mobile Computing*, vol. 7, no. 5, 2008.

[11] F.F. Digham, M.S. Alouini, M.K. Simon, "On the Energy Detection of Unknown Signals over Fading Channels," in *Proc. IEEE Int. Conf. Commun.*, vol. 5, pp. 3575-3579, 2003.

[12] Q. Zhao, B. Krishnamachari and K. Liu, "Low-Complexity Approaches to Spectrum Opportunity Tracking," in *Proc. of the 2nd International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CrownCom)*, pp. 27-35, 2007.

[13] Q. Zhao, B. Krishnamachari, "Structure and Optimality of Myopic Sensing for Opportunistic Spectrum Access," in *Proc. IEEE Int. Conf. Commun.*, pp. 6476-6481, 2007.

[14] Y. Chen, Q. Zhao and A. Swami, "Distributed Spectrum Sensing and Access in Cognitive Radio Networks With Energy Constraint," *IEEE Trans. Sig. Process.*, vol. 57, no. 2, pp. 783-797, 2009.

[15] R. Smallwood and E. Sondik, "The optimal control of partially observable Markov processes over a finite horizon," *Operations Research*, pp. 1071-1088, 1971.

[16] Y. Xing, R. Chandramouli, and C.M. Cordeiro, "Price Dynamics in Competitive Agile Spectrum Access Markets," *IEEE J. Select. Areas Commun.*, vol. 25, no. 3, pp. 613-621, Apr. 2007.

[17] Y. Wu, D. H.K. Tsang, "Dynamic Rate Allocation, Routing and Spectrum Sharing for Multi-hop Cognitive Radio Network", in *Proc. of IEEE ICC'09 Workshop on Cognitive Radio Networks and Systems*, pp. 1-6, Jun. 2009.

[18] S.H. Ahmad, M. Liu, T. Javidi, Q. Zhao and B. Krishnamachari, "Optimality of Myopic Sensing in Multi-Channel Opportunistic Access," *IEEE Trans. Information Theory*, vol. 55, No. 9, pp. 4040-4050, Sep. 2009.

[19] Y.J. Choi, Y. Xin and S. Rangarajan, "Overhead-Throughput Tradeoff in Cooperative Cognitive Radio Networks," in *Proc. of IEEE WCNC*, pp. 1-6, Apr. 2009.

[20] E.C.Y. Peh, Y.C. Liang, Y.L. Guan and Y. Zeng, "Optimization of Cooperative Sensing in Cognitive Radio Networks: A Sensing-Throughput Tradeoff View," *IEEE Trans. Vehicular Technology*, vol. 58, no. 9, pp. 5294-5299, Nov. 2009.

[21] R. Fan and H. Jiang, "Optimal Multi-Channel Cooperative Sensing in Cognitive Radio Networks," *IEEE Trans. Wireless Commun.*, vol. 9, no. 3, pp. 1128-1138, Mar. 2010.