A Restless Bandit Formulation of Opportunistic Access: Indexablity and Index Policy

Keqin Liu,   Qing Zhao
University of California, Davis, CA 95616
kqliu@ucdavis.edu, qzhao@ece.ucdavis.edu

Abstract—We focus on an opportunistic communication system consisting of multiple independent channels with time-varying states. With limited sensing, a user can only sense and access a subset of channels and accrue rewards determined by the state of the sensed channels. We formulate the problem of optimal sequential channel probing as a restless multi-armed bandit process, for which a powerful index policy–Whittle's index policy–can be implemented based on the indexability of the system. Exploiting the underlying structure of the multi-channel opportunistic access problem, we establish the indexability and obtain the Whittle's index in closed-form, which leads to a direct implementation of Whittle's index policy with little complexity. Furthermore, we show that Whittle's index policy is equivalent to the myopic policy when channels are statistically identical.

Index Terms—Opportunistic access, optimal channel probing, restless multi-armed bandit, Whittle's index policy, indexability.

I. INTRODUCTION

We consider an opportunistic communication system with N parallel channels. These N channels are modeled as independent but not necessarily identical Gilbert-Elliot channels [1] as illustrated in Fig. 1. The state of a channel — "good" (1) or "bad" (0) — indicates the desirability of accessing this channel and determines the resulting reward. With limited sensing and access capability, a user chooses M out of these N channels to sense and access in each slot, aiming to maximize its expected long-term reward. Such an opportunistic communication system arises in the applications of cognitive radios for spectrum overlay (also referred to as opportunistic spectrum access), where secondary users search in the spectrum for idle channels temporarily unused by primary users [2]. Other applications include transmission over fading channels and resource-constrained jamming and anti-jamming.
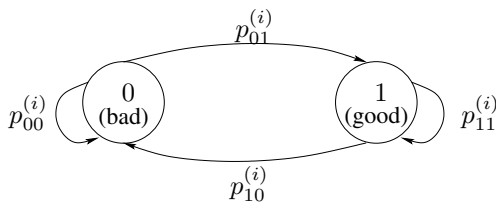

Fig. 1.   The Gilber-Elliot channel model.

A. Restless Multi-armed Bandit Problem

The optimal design of a sensing policy for channel probing can be formulated as a partially observable Markov decision process (POMDP), where channels can be generally correlated [3]. For independent channels, the problem can also be viewed as a restless multi-armed bandit process as shown in section II. Unfortunately, the optimal solution to a restless multi-armed bandit problem is often intractable: the problem has been shown to be PSPACE-hard in general [4].

Whittle in 1988 proposed an index policy (referred to as Whittle's index policy) for the restless multi-armed bandit problem and introduced the concept of indexability [5]. The remarkable nature of Whittle's index policy lies in that it decomposes the N-dimensional problem into N 1-dimensional problems. Furthermore, Whittle's index policy is optimal if the constraint is on the *average* number of arms to activate at a given time. Under the strict constraint where a *constant* number of arms are activated at a given time, Whittle's index policy is shown to be asymptotically optimal under certain conditions [6]. In a large range of empirical studies, the near-optimal performance of the Whittle's index policy has been demonstrated in non-asymptotic regimes, see for example [7], [8].

Unfortunately, Whittle's index policy does not always exist. It is necessary that a restless bandit is *indexable* in order to use Whittle's index policy, and the indexability of a general restless bandit can be complicated to establish [9].

B. Contribution

Indexability We formulate the design of the optimal sensing policy as a restless multi-armed bandit problem where the state of each arm is defined as the belief value (or the information state) of each channel, *i.e.*, the conditional probability that this channel is in the good state given the entire observation and decision history. The state of each arm is thus *uncountable*, which further complicates the establishment of indexability. By exploiting the rich structure of the problem, we prove that the restless bandit formulation of multi-channel opportunistic access is indexable.

Whittle's index in closed-form Even when the indexability of a restless bandit problem can be established, Whittle's index is generally difficult to obtain even numerically. In this paper, we show that for the problem at hand, Whittle's index can be obtained in closed-form. Whittle's index policy can then

be implemented with simple evaluations of the closed-form expressions.

Equivalence between Whittle's index policy and myopic policy for identical channels When channels are statistically identical, we show that the myopic policy coincides with the Whittle's index policy. Interestingly, the myopic policy has a simple structure and is shown to be optimal under certain conditions [10]–[12]. As a consequence, this equivalence provides examples under which Whittle's index policy is optimal.

II. PROBLEM FORMULATION

Consider N independent Gilbert-Elliot channels with bandwidth B_i ($i = 1, \cdots, N$). The state of channel i–"good"(1) or "bad"(0)– evolves as a Markov chain from slot to slot as shown in Fig.1. The transition matrix of the Markov chain is denoted by P_i ($P_i = \begin{bmatrix} p_{00}^{(i)}, & p_{01}^{(i)} \\ p_{10}^{(i)}, & p_{11}^{(i)} \end{bmatrix}$).

At the beginning of slot t, the user selects M out of N channels to sense. If the state $S_i(t)$ of sensed channel i is 1, the user transmits and collects B_i units of reward from this channel. Otherwise, the use collects no reward and waits for the next slot to make another selection of M channels. Our objective is to maximize the long-term total reward by choosing a policy which sequentially selects M channels to sense in each slot based on the decision and observation history.

A. Restless Multi-armed Bandit Formulation

Due to limited sensing, the channel state $\mathcal{S} = [S_1(t), ..., S_N(t)] \in \{0,1\}^N$ is not fully observable. If we treat $\mathcal{S}$ as the system state, then we have a POMDP formulation of this problem [13]. In a restless multi-armed bandit formulation, the system state has to be fully observable. Thus we cannot treat $\mathcal{S}$ as the system state of the restless bandit. However, it has been shown that the conditional probabilities that each channel is in state 1 given all past decisions and observations is a sufficient statistic for optimal decision making [3]. The vector $\Omega(t) \triangleq [\omega_1(t), ..., \omega_N(t)]$ which comprises of the conditional probability $\omega_i(t)$ that $S_i(t) = 1$ is called the belief vector. Moreover, given the sensing action and the observation in slot t, the belief vector $\Omega(t+1)$ for slot $t+1$ can be obtained as follows:

$$\omega_i(t+1) = \begin{cases} p_{11}^{(i)}, & i \in I(t), S_i(t) = 1 \\ p_{01}^{(i)}, & i \in I(t), S_i(t) = 0 \\ \omega_i(t)p_{11}^{(i)} + (1 - \omega_i(t))p_{01}^{(i)}, & i \notin I(t) \end{cases}$$
(1)

where $I(t)$ is the set of the M channels sensed in slot t.

We formulate the problem as a restless multi-armed bandit problem. Each channel is considered as an arm. The state of arm i in slot t is $\omega_i(t)$. The user chooses an action in each slot: activate (sense) M out of N channels and make others passive (not sense). The expected reward obtained from the activated arm i in slot t is $\omega_i(t)B_i$. And the system state $\Omega(t)$ transits as a Markov chain as given in (1). Our objective is to design the optimal policy which maps the belief vector to

action in each slot in order to maximize the expected total discounted reward.

Formally, we have the following restless multi-armed bandit $(\Omega, \{P_i : 1 \le i \le N\}, R, \beta)$ as defined below.

(i) The user takes action $I(t)$ in slot t ($t = 0, 1, 2, ...$), where $I(t)$ denotes the set of the M arms which the user activates in slot t.

(iii) The user collects an expected reward

$$R(t) = \Sigma_{i \in I(t)} \omega_i(t) B_i$$

in slot t.

(iv) The state of each arm transits as a Markov chain according to (1).

(v) A policy $\pi : \Omega(t) \to I(t)$ specifies the action to take in each slot given the current belief state $\Omega(t)$.

The objective is to maximize the expected total discounted reward over an infinite horizon:

$$\max_{\pi}\{\mathbb{E}_{\pi}\{\Sigma_{t=0}^{\infty}\beta^t R(t)\}\},$$
(2)

where $0 \le \beta < 1$ is the discount factor.

B. Value Function

Let $V_t(\Omega)$ be the value function, which denotes the maximum expected total remaining reward obtained starting from slot t given the current belief vector Ω. Given that the user takes action I and observes $O = \{S_i : i \in I\}$, the reward that can be accumulated starting from slot t consists of two parts: the immediate reward $R_I(t) = \Sigma_{i \in I(t)} \omega_i(t) B_i$ and the maximum expected future reward $V_{t+1}(\Omega(t+1)|I(t), O(t))$. Averaging over all possible observations O and maximizing over all action I, we have the following optimality equation:

$$V_t(\Omega) = \max_{I(t)}\{R_I(t) + \beta \mathbb{E}_{O(t)}(V_{t+1}(\Omega(t+1)|I(t), O(t)))\}.$$
(3)

It has been shown that $V_t(\Omega)$ does not depend on t under the criterion of discounted reward over an infinite horizon [14]. Thus we can drop the subscript t of V_t in (3):

$$V(\Omega) = \max_{I(t)}\{R_I(t) + \beta \mathbb{E}_{O(t)}(V(\Omega(t+1)|I(t), O(t)))\}.$$
(4)

C. Myopic Policy

A myopic policy ignores the impact of the current action on the future reward, focusing solely on maximizing the expected immediate reward R_I. The myopic action $\hat{I}$ under belief state $\Omega = [\omega_1, \cdots, \omega_N]$ is simply given by

$$\hat{I}(\Omega) = \arg\max_{I} \Sigma_{i \in I} \omega_i B_i.$$
(5)

Interestingly for identical channels under single-channel sensing[1], it has been shown in [10] that the myopic policy has a simple structure that does not need the update of the belief vector or the precise knowledge of the transition probabilities.

Specifically, when $p_{11} \geq p_{01}$, the myopic action is to stay in the same channel if the channel in the current slot is in state 1. Otherwise, the user switches to the channel visited the longest time ago. The channel selection is thus in a round robin fashion as illustrated in Fig. 2: sense N channels in turn with a random switching time (when the current channel transits to state 0).
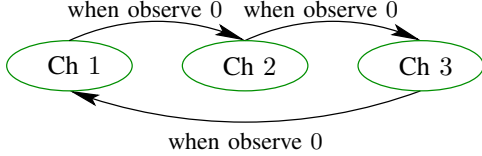

Fig. 2. The structure of the myopic policy for $p_{11} \geq p_{01}$ ($N = 3$).

When $p_{11} < p_{01}$, the myopic action is to stay in the same channel when the channel is in state 0 and switch otherwise. When a channel switch is needed, the user chooses, among those channels to which the last visit occurred an even number of slots ago, the one most recently visited. If there are no such channels, the user chooses the channel visited the longest time ago.

Surprisingly for identical channels with single-channel sensing, the myopic policy with such a simple and robust structure achieves the optimal performance for $N = 2$ [10]. In a recent work [12], the optimality of the myopic policy has been extended to $N = 3$, and $N > 3$ under the condition of $p_{11} > p_{01}$.

However for independent but not identical channels, the myopic policy suffers from a performance loss as shown in section IV.

D. Whittle's Index Policy

Whittle introduced a heuristic policy referred to as Whittle's index policy. The basic idea is to introduce an index that measures how attractive it is to activate a particular arm at its current state, and then activate those M arms with the largest index at each time. Whittle's index for an arm at state ω is defined as the extra amount of reward ν that we should provide to the passive action (referred to as the subsidy for passivity) in order to make the active and passive actions equally attractive at the current state ω.

The significance of Whittle's index policy and its strong performance have been discussed in section I. The main challenge is that a restless multi-armed bandit may not have a well-defined Whittle's index. We present the definition of indexability as follows.

Consider arm i of the bandit $(\Omega, \{P_i : 1 \leq i \leq N\}, R, \beta)$. In each slot, the user either activates the arm or not. Assume that a constant subsidy m for passivity is obtained if the

---

[1]It has been shown in [15] that a similar structure of the myopic policy holds with multi-channel sensing.

user makes the arm passive. The objective is to maximize the expected total discounted reward. Let $U_i(m)$ denote the passive set consisting of states (belief values) in which the optimal action is to make the arm passive.

Definition 1: Bandit $(\Omega, \{P_i : 1 \leq i \leq N\}, R, \beta)$ is *indexable* if $U_i(m)$ is monotonically increasing from $\emptyset$ to the whole set $[0,1]$ as m goes from $-\infty$ to $+\infty$, for all $i \in \{1, 2, .., N\}$.

Definition 2: If a bandit $(\Omega, \{P_i : 1 \leq i \leq N\}, R, \beta)$ is indexable, the *Whittle's index* $W(\omega_i(t))$ of arm i in slot t is the subsidy m such that it is optimal to make the arm either active or passive in slot t given the current belief $\omega_i(t)$.

Definition 3: In each slot, the user activates M arms with the largest Whittle's index. This policy is called *Whittle's index policy*.

III. INDEXABILITY AND WHITTLE'S INDEX POLICY

In this section, we present our main theorem which shows the above restless multi-armed bandit problem is indexable. Furthermore, we give the Whittle's index in closed-form.

To prove the indexability, we only need to consider a single arm. To simplify the notation, we assume that the arm has bandwidth B, subsidy m for passivity, and transition matrix P ($P = \begin{bmatrix} p_{00}, & p_{01} \\ p_{10}, & p_{11} \end{bmatrix}$). Theorems 1-3 given below summarize the results on the indexability and Whittle's index policy for multi-channel opportunistic access.

Theorem 1: The optimal policy for a single-armed bandit (ω, P, B, m, β) is a threshold policy: there exists a $\omega^*(m) \in \mathbb{R}$ such that it is optimal to activate the arm if the current belief $\omega > \omega^*(m)$; otherwise it is optimal to make the arm passive.

Theorem 2: The bandit $(\Omega, \{P_i, 1 \leq i \leq N\}, R, \beta)$ is indexable.

Theorem 3: Whittle's index $W(\omega) \in \mathbb{R}$ as a function of the belief state $\omega \in [0, 1]$ for arm i of bandit $(\Omega, \{P_i, 1 \leq i \leq N\}, R, \beta)$ is given as follows.

Let $\mathcal{T}_i^k(\pi)$ denote the $k-$step transition of a belief state π of arm i under passive actions:

$$\mathcal{T}_i^k(\pi) = \frac{p_{01}^{(i)} - (p_{11}^{(i)} - p_{01}^{(i)})^k (p_{01}^{(i)} - (1 + p_{01}^{(i)} - p_{11}^{(i)})\pi)}{1 + p_{01}^{(i)} - p_{11}^{(i)}}.$$

For simplicity, let $\mathcal{T}_i(\pi)$ denote $\mathcal{T}_i^1(\pi)$.

Case 1: $p_{11}^{(i)} \geq p_{01}^{(i)}$

$$W(\omega) = \begin{cases} \omega B_i, & \text{if } \omega \leq p_{01}^{(i)} \text{ or } \omega \geq p_{11}^{(i)}; \\[2mm] \frac{\omega}{1 - \beta p_{11}^{(i)} + \beta \omega} B_i, & \text{if } \omega_o \leq \omega < p_{11}^{(i)}; \\[2mm] \frac{\omega - \beta \mathcal{T}_i(\omega) + C(1-\beta)(\beta(1 - \beta p_{11}^{(i)}) - \beta(\omega - \beta \mathcal{T}_i(\omega)))}{1 - \beta p_{11}^{(i)} - A(1-\beta)(\beta(1 - \beta p_{11}^{(i)}) - \beta(\omega - \beta \mathcal{T}_i(\omega)))} B_i, \\ \quad \text{if } p_{01}^{(i)} < \omega < \omega_o; \end{cases}$$

where $A = \dfrac{(1 - \beta p_{11}^{(i)})(1 - \beta^{L+1})}{(1 - \beta p_{11}^{(i)})(1-\beta)(1 - \beta^{L+2}) + (1-\beta)^2 \beta^{L+2} \mathcal{T}_i^{L+1}(p_{01}^{(i)})}$;

$$C = \frac{(1-\beta)\beta^{L+1}\mathcal{T}_i^{L+1}p_{01}^{(i)}}{(1-\beta p_{11}^{(i)})(1-\beta)(1-\beta^{L+2})+(1-\beta)^2\beta^{L+2}\mathcal{T}_i^{L+1}(p_{01}^{(i)})};$$

$$L = \left\lfloor \log_{\frac{p_{01}^{(i)}-\omega(1-p_{11}^{(i)}+p_{01}^{(i)})}{p_{01}^{(i)}(p_{11}^{(i)}-p_{01}^{(i)})}} \frac{p_{01}^{(i)}-\omega(1-p_{11}^{(i)}+p_{01}^{(i)})}{p_{11}^{(i)}-p_{01}^{(i)}} \right\rfloor;$$

$$\omega_o = \frac{p_{01}^{(i)}}{p_{01}^{(i)}+p_{10}^{(i)}}.$$

Case 2: $p_{11}^{(i)} < p_{01}^{(i)}$

$$W(\omega) = \begin{cases} \omega B_i, & \text{if } \omega \le p_{11}^{(i)} \text{ or } \omega \ge p_{01}^{(i)}; \\[2mm] \frac{\beta p_{01}^{(i)}+\omega(1-\beta)}{1+\beta(p_{01}^{(i)}-\omega)}B_i, & \text{if } \mathcal{T}_i(p_{11}^{(i)}) \le \omega < p_{01}^{(i)}; \\[2mm] \frac{(1-\beta)(1+\beta E)(\beta p_{01}^{(i)}+\omega(1-\beta))}{1-\beta(1-p_{01}^{(i)})-D(1-\beta)(\beta^2 p_{01}^{(i)}+\beta\omega-\beta^2\omega)}B_i, \\ \quad \text{if } \omega_o \le \omega < \mathcal{T}_i(p_{11}^{(i)}); \\[2mm] \frac{(1-\beta)(\beta p_{01}^{(i)}+\omega-\beta\mathcal{T}_i(\omega))-E(1-\beta)\beta(\beta\mathcal{T}(\omega)-\beta p_{01}^{(i)}-\omega)}{1-\beta(1-p_{01}^{(i)})+D(1-\beta)\beta(\beta\mathcal{T}_i(\omega)-\beta p_{01}^{(i)}-\omega)}B_i, \\ \quad \text{if } p_{11}^{(i)} < \omega < \omega_o; \end{cases}$$

where $D = \frac{1-\beta(1-p_{01}^{(i)})}{1-\beta(1-p_{01}^{(i)})-\beta^2\mathcal{T}_i(p_{11}^{(i)})(1-\beta)-\beta^3 p_{01}^{(i)}};$

$$E = \frac{\beta\mathcal{T}_i(p_{11}^{(i)})(1-\beta)+\beta^2 p_{01}^{(i)}}{1-\beta(1-p_{01}^{(i)})-\beta^2\mathcal{T}_i(p_{11}^{(i)})(1-\beta)-\beta^3 p_{01}^{(i)}}.$$

Corollary 1: If channels are identical, then the myopic policy is equivalent to Whittle's index policy.

IV. SIMULATION EXAMPLES

From Theorem 3, the mapping from belief to Whittle's index varies when the belief ω is in different regions. To illustrate this mapping, we give an example as shown in Figure 3, where we assume that the arm has bandwidth 1.
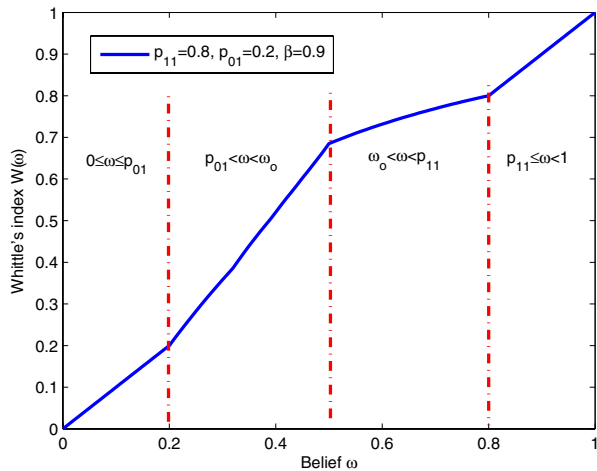

Fig. 3. Whittle's index in different regions.

Figure 4 compares Whittle's index for different transition probabilities. Assume that $B_i = 1$. This figure clearly shows that when channels are non-identical, the channels with the largest belief values may not have the largest Whittle's index; the myopic policy is thus different from Whittle's index policy.
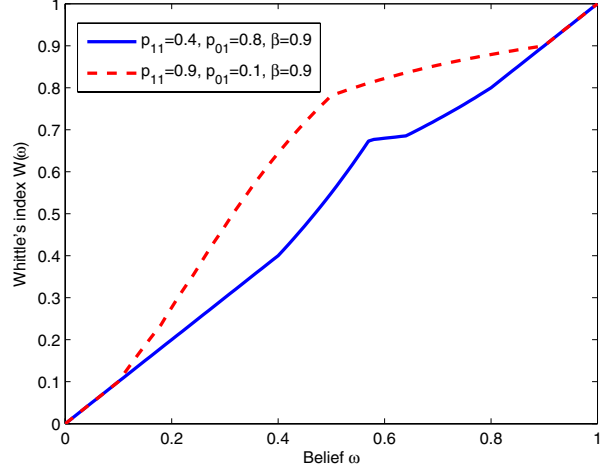

Fig. 4. Whittle's index for different network parameters.

Shown in Figure 5 is a simulation example where we evaluate the performance of Whittle's index policy for independent but not identical channels. We observe that the Whittle's index policy yields a near-optimal performance.
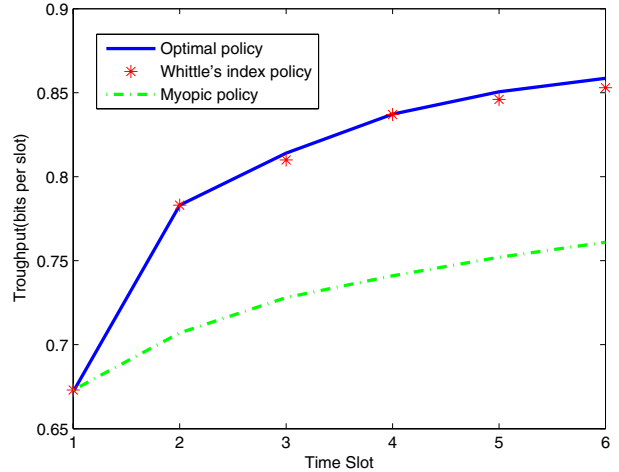

Fig. 5. Performance of the Whittle's index policy ($M = 1, N = 7, \beta = 0.999$).

V. CONCLUSION

In this paper, we have formulated the multi-channel opportunistic access problem as a restless multi-armed bandit problem. We proved the indexability of the restless bandit and obtained Whittle's index in closed-form. When channels are identical, Whittle's index policy coincides with the myopic

policy. When channels are non-identical, the closed-form expression of Whittle's index provides a simple low-complexity index policy with near-optimal performance.

REFERENCES

[1] E.N. Gilbert, "Capacity of burst-noise channels," Bell Syst. Tech. J., vol. 39, pp. 1253-1265, Sept. 1960.

[2] Q. Zhao and B. Sadler, "A Survey of Dynamic Spectrum Access," *IEEE Signal Processing magazine*, vol. 24, pp. 79–89, May 2007.

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

[4] C. H. Papadimitriou and J. N. Tsitsiklis, "The Complexity of Optimal Queueing Network Control," in *Mathematics of Operations Research*, Vol. 24, No. 2, May 1999, pp. 293-305.

[5] P. Whittle, "Restless bandits: Activity allocation in a changing world", in *Journal of Applied Probability*, Volume 25, 1988.

[6] Richard R. Weber and Gideon Weiss, "On an Index Policy for Restless Bandits," in *Journal of Applied Probability*, Vol.27, No.3, pp. 637-648, Sep 1990.

[7] Ansell, P. S., Glazebrook, K. D., Niño-Mora, J. and O'Keeffe, M. "Whittle's index policy for a multi-class queueing system with convex holding costs," in *Math. Meth. Operat. Res.* 57, 21–39, 2003.

[8] Glazebrook KD, Mitchell HM, Ansell PS,"Index policies for the maintenance of a collection of machines by a set of repairmen." *Eur J Oper Res* 165:267284, 2005.

[9] J.E. Nio-Mora, "Restless bandits, partial conservation laws and indexability," in *Advances in Applied Probability*, 33:7698, 2001.

[10] Q. Zhao and B. Krishnamachari, "Structure and optimality of myopic sensing for opportunistic spectrum access," in *Proc. of IEEE Workshop on Towards Cognition in Wireless Networks(CogNet)*, pp. 6476–6481, June, 2007.

[11] 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)*, August 2007.

[12] T. Javidi, B. Krishnamachari, Q. Zhao, and M. Liu, "Optimality of Myopic Sensing in Multi-Channel Opportunistic Access," to appear in *Proc. of IEEE International Conference on Communications(ICC)*, May, 2008.

[13] Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized Cognitive MAC for Opportunistic Spectrum Access in Ad Hoc Networks: A POMDP Framework," in *IEEE Journal on Selected Areas in Communications (JSAC): Special Issue on Adaptive, Spectrum Agile and Cognitive Wireles Networks* , April 2007.

[14] Edward J. Sondik, "The Optimal Control at Partially Observable Markov Processes Over the Infinite Horizon: Discounted Costs," in *Operations Research*, Vol.26, No.2 (Mar. - Apr,. 1978), 282 - 304.

[15] K. Liu and Q. Zhao, "Channel Probing for Opportunistic Access with Multi-channel Sensing," to appear in *Proc. of IEEE Asilomar Conference on Signals, Systems, and Computers*, October, 2008.