Optimality of Myopic Sensing in Multi-Channel Opportunistic Access

Tara Javidi[†], Bhaskar Krishnamachari[§], Qing Zhao[‡], Mingyan Liu[♯]
tara@ece.ucsd.edu, bkrishna@usc.edu, qzhao@ece.ucdavis.edu, mingyan@eecs.umich.edu
[†]Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92093
[§]Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089
[‡]Department of Electrical and Computer Engineering, University of California, Davis, CA 95616
[♯]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109

***Abstract*— We consider opportunistic communications over multiple channels where the state ("good" or "bad") of each channel evolves as independent and identically distributed Markov processes. A user, with limited sensing and access capability, chooses one channel to sense and subsequently access (based on the sensed channel state) in each time slot. A reward is obtained when the user senses and accesses a "good" channel. The objective is to design the optimal channel selection policy that maximizes the expected reward accrued over time. This problem can be generally formulated as a Partially Observable Markov Decision Process (POMDP) or a restless multi-armed bandit process, to which optimal solutions are often intractable. We show in this paper that the myopic policy, with a simple and robust structure, achieves optimality under certain conditions. This result finds applications in opportunistic communications in fading environment, cognitive radio networks for spectrum overlay, and resource-constrained jamming and anti-jamming.**

Opportunistic access, cognitive radio, POMDP, restless multi-armed bandit process, myopic policy.

I. Introduction

We consider a fundamental communication context in which a sender has the ability to access many channels, but is limited to sensing and transmitting only on one at a given time. We explore how a smart sender should exploit past observations and the knowledge of the stochastic state evolution of these channels to maximize its transmission rate by switching opportunistically across channels.

We model this problem in the following manner. As shown in Figure 1, there are n channels, each of which evolves as an independent, identically-distributed, two-state discrete-time Markov chain. The two states for each channel — "good" (1) and "bad" (0) — indicate the desirability of transmitting at a given time slot. In each time period the sender picks one of the channels to sense based on its prior observations, and obtains some fixed reward if it is in the good state. The basic objective of the sender is to maximize the reward that it can gain over a given finite time horizon. This problem can be described as a partially observable Markov decision process (POMDP) [4] or a restless multi-armed bandit process [8]. We discuss the implications of each formulation and the relationship of our work to the relevant bodies of literature in Section V. Generalizations of this problem, such as considering non-i.i.d. channels, imperfect sensing, and more than two states, are also of interest, but we do not treat them in this work.

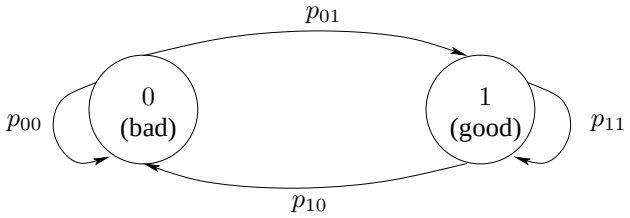

Fig. 1. The Markov channel model.

This formulation is broadly applicable to several domains. It arises naturally in opportunistic spectrum access (OSA) [1], [2], [3], where the sender is a secondary user, and the channel states describe the occupancy by primary users. In the OSA problem, the secondary sender may send on a given channel only when there is no primary user occupying it. It pertains to communication over parallel fading channels as well, if a two-state Markovian fading model is employed. Another interesting application of this formulation is in the domain of communication security, where it can be used to developing bounds on the performance of resource-constrained Jamming. A jammer that has access to only one channel at a time could also use the same stochastic dynamic decision making process to maximize the number of times that it can successfully jam communications that occur on these channels. In this application, the "good" state for the jammer is precisely when the channel is being utilized by other senders (in contrast with the OSA problem).

In prior work [2], it has been shown that when there are two channels, a simple myopic policy offers optimal performance. It has been conjectured (based on simulation results) that this result holds generally for all n. It has also been shown in [2] that for all n the myopic policy also has an elegant and robust structure that obviates the need to know the channel state transition probabilities exactly. Specifically, it suffices to know the sign of the auto-correlation of the channel process over one unit of time, or equivalently, whether $p_{01} > p_{11}$ or *vice versa*. We make progress towards solving the conjecture in this work: we show that the simple myopic policy is optimal for all n, under certain conditions on p_{01}, p_{11}. We also generalize the

result to related formulations involving discounted rewards and infinite horizons.

II. Problem Formulation

We consider the scenario where a user is trying to access the wireless spectrum to maximize its throughput or data rate. The spectrum consists of n independent and statistically identical channels. The state of a channel is given by a two-state discrete time Markov chain shown in Figure 1.

The system operates in discrete time steps indexed by t, $t = 1, 2, \cdots, T$, where T is the time horizon of interest. At time t^-, the channels (i.e., the Markov chains representing them) go through state transitions, and at time t the user makes the channel sensing and access decision. Specifically, at time t the user selects one of the n channels to sense, say channel i. If the channel is sensed to be in the "good" state (state 1), the user transmits and collects one unit of reward. Otherwise the user does not transmit (or transmits at a lower rate), collects no reward, and waits till $t+1$ to make another choice. This process repeats sequentially till the time horizon expires. Here we have assumed that sensing errors are negligible. The optimality of the myopic policy in the presence of sensing errors has been shown for $n = 2$ in [3].

The state of the above system at time t (or more precisely at t^-) is given by the vector $\mathbf{s}(t) = [s_1(t), s_2(t), \cdots, s_n(t)] \in \{0, 1\}^n$. Note that $\mathbf{s}(t)$ is not directly observable to the user. However, it can be shown (see e.g., [4], [5], [10]) that a sufficient statistics of the system for optimal decision making, or the *information state* of the system [10], [5], is given by the conditional probabilities that each channel is in state 1 given all past observations. We denote this information state or belief vector by $\bar{\omega}(t) = [\omega_1(t), \cdots, \omega_n(t)] \in [0, 1]^n$, where $\omega_i(t)$ is the conditional probability that channel i is in state 1 at time t^-.

The user's action space is given by the finite set $\{1, 2, \cdots, n\}$, and we will use $a(t) = i$ to denote that the user selects channel i to sense at time t. It follows that the information state of the system is governed by the following transition upon an action a in the state $\bar{\omega}(t)$ with an observation outcome $s_a(t)$:

$$\omega_i(t+1) = \begin{cases} p_{11} & \text{if } a = i, s_a(t) = 1 \\ p_{01} & \text{if } a = i, s_a(t) = 0 \\ \tau(\omega_i(t)) & \text{if } a \neq i \end{cases}. \quad (1)$$

In Equation (1) the first case denotes the conditional probability when the channel is observed to be in the "good" state; the second case when the channel is observed to be "bad"; and the last case when the channel is not observed, where we have used the operator $\tau : [0, 1] \to [0, 1]$ defined as

$$\tau(\omega) := \omega p_{11} + (1 - \omega) p_{01}, \quad 0 \leq \omega \leq 1. \quad (2)$$

We will further use the operator $\mathcal{T}$ to denote the above state transition. More precisely, we have

$$\bar{\omega}(t+1) = \mathcal{T}(\bar{\omega}(t)|a, s_a(t)), \quad (3)$$

with the understanding that the above notation implies the operation given in (1) applied to $\bar{\omega}(t)$ element-by-element.

The user's policy π is given by the vector $\pi = [\pi(1), \pi(2), \cdots, \pi(T)]$, where $\pi(t) = i \in \{1, 2, \cdots, n\}$ denotes the decision to select channel i at time t. Such decisions are based on the current information state $\bar{\omega}(t)$.

The objective of the user is to maximize its total (discounted or average) expected reward over a finite (or infinite) horizon. Let $J_T^\pi(\bar{\omega})$, $J_\beta^\pi(\bar{\omega})$, and $J_\infty^\pi(\bar{\omega})$ denote, respectively, these cost criteria (namely, finite horizon, infinite horizon with discount, and infinite horizon average reward) under policy π starting in state $\bar{\omega} = [\omega_1, \cdots, \omega_n]$. The associated optimization problems ((P1)-(P3)) are formally defined as follows.

$$\text{(P1): } \max_\pi J_T^\pi(\bar{\omega}) = \max_\pi E^\pi[\sum_{t=1}^{T} \beta^{t-1} R_{\pi(t)}(\bar{\omega}(t))|\bar{\omega}(1) = \bar{\omega}]$$

$$\text{(P2): } \max_\pi J_\beta^\pi(\bar{\omega}) = \max_\pi E^\pi[\sum_{t=1}^{\infty} \beta^{t-1} R_{\pi(t)}(\bar{\omega}(t))|\bar{\omega}(1) = \bar{\omega}]$$

$$\text{(P3): } \max_\pi J_\infty^\pi(\bar{\omega}) = \max_\pi E^\pi[\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} R_{\pi(t)}(\bar{\omega}(t))|\bar{\omega}(1) = \bar{\omega}]$$

where β ($0 \leq \beta \leq 1$ for (P1) and $0 \leq \beta < 1$ for (P2)) is the discount factor, the maximizations are over all admissible policies π, and $R_{\pi(t)}(\bar{\omega}(t))$ is the reward collected under state $\bar{\omega}(t)$ when channel $\pi(t)$ is selected, and is given by $R_{\pi(t)}(\bar{\omega}(t)) = s_{\pi(t)}(t)$.

In subsequent sections we will first focus on problem (P1), and then extend the results to (P2) and (P3).

III. Optimal Policy

A. Preliminaries

Consider (P1). An optimal policy obviously exists since the number of admissible policies are finite. In theory, such a policy may be found by using dynamic programming:

Fact 1: Define recursively the functions

$$\begin{aligned} V_T(\bar{\omega}) &= \max_{a=1,2,\cdots,n} E[R_a(\bar{\omega})] \\ V_t(\bar{\omega}) &= \max_{a=1,2,\cdots,n} E[R_a(\bar{\omega}) + \beta V_{t+1}(\mathcal{T}(\bar{\omega}|a, s_a(t)))] \\ &= \max_{a=1,\cdots,n} (\omega_i + \omega_i V_{t+1}(\mathcal{T}((\bar{\omega})|i, 1))) \\ &\qquad + (1 - \omega_i) V_{t+1}(\mathcal{T}((\bar{\omega})|i, 0)). \end{aligned} \quad (4)$$

Then,

i) $V_t(\bar{\omega}) = \max_\pi J_{T-t+1}^\pi(\bar{\omega})$ with probability 1. Furthermore, $V_1(\bar{\omega}) = \max_\pi J_T^\pi(\bar{\omega})$.

ii) A Markov policy $\pi^* = \{\pi_1^*, \pi_2^*, \ldots, \pi_T^*\}$ is optimal if and only if for $t = 1, \ldots, T$, $a = \pi_t^*(\bar{\omega})$ achieves the maximum in (4) .

Note that $V_t(\bar{\omega})$ is the value function, or the maximum expected remaining reward that can be accrued starting from time t when the current information state is $\bar{\omega}$. It has two parts: (i) the immediate reward $R_a(\bar{\omega})$ obtained in slot t when the user senses channel a; and (ii) the maximum expected remaining reward starting from time $t + 1$, given by $V_{t+1}(\mathcal{T}(\bar{\omega}|a, s_a(t)))$ where the new state represents the

updated knowledge of the system state after incorporating the action a and the observation $s_a(t)$.

Similarly, the dynamic programming principle holds for (P2) and (P3) as given below. The existence of optimal stationary Markov policies in (P2) and (P3) is a consequence of i) finiteness of action space, ii) countability of the (unobservable) state space, and iii) the bounded condition on the immediate reward (see [5]).

Fact 2: Consider (P2). There exists a unique function $V_\beta(\cdot)$ satisfying the following equation:

$$\begin{aligned} V_\beta(\bar{\omega}) &= \max_{a=1,\cdots,n} E[R_a(\bar{\omega}) + \beta V_\beta(\mathcal{T}(\bar{\omega}|a, s_a(t)))] \\ &= \max_{a=1,\cdots,n} (\omega_i + \beta\omega_i V_\beta(\mathcal{T}((\bar{\omega})|i,1))) \\ &\quad +\beta(1-\omega_i)V_\beta(\mathcal{T}((\bar{\omega})|i,0)) \quad (5) \end{aligned}$$

Furthermore,

i) $V_\beta(\bar{\omega}) = \max_\pi J_\beta^\pi(\bar{\omega})$ with probability 1.
ii) A Markov stationary policy π^* is optimal if and only if $a = \pi^*(\bar{\omega})$ achieves the maximum in (5).

Fact 3: Consider (P3). There exist function $h_\infty(\cdot)$ and constant scalar J satisfying the following equation:

$$\begin{aligned} J + h_\infty(\bar{\omega}) &= \max_{a=1,2,\cdots,n} E[R_a(\bar{\omega}) + h_\infty(\mathcal{T}(\bar{\omega}|a, s_a(t)))] \\ &= \max_{a=1,\cdots,n} (\omega_i + \omega_i h_\infty(\mathcal{T}((\bar{\omega})|i,1))) \\ &\quad +(1-\omega_i)h_\infty(\mathcal{T}((\bar{\omega})|i,0)) \quad (6) \end{aligned}$$

Furthermore,

i) $J = \max_\pi J_\infty^\pi(\bar{\omega})$ with probability 1.
ii) A Markov stationary policy π^* is optimal if and only if $a = \pi^*(\bar{\omega})$ achieves the minimum in (6).

Unfortunately, due to the impact of the current action on the future reward, the uncountable space of the information state $\omega(t)$, and the non-stationary nature of the optimal policy (owing to the finiteness of the horizon), obtaining the optimal solution using the above equations directly is in general computationally prohibitive.

For the remainder of this paper, we will focus on obtaining structural properties of the optimal policy. Specifically, we will consider a myopic policy that aims at maximizing the immediate reward at each time step, and show its optimality under certain conditions.

B. The Myopic Policy

A myopic policy ignores the impact of the current action on the future reward, focusing solely on maximizing the expected immediate reward. Myopic policies are thus stationary. For (P1), the myopic policy under state $\bar{\omega} = [\omega_1, \omega_2, \cdots, \omega_n]$ is simply given by

$$a^*(\bar{\omega}) = \arg \max_{a=1,\cdots,n} E[R_a(\bar{\omega})] = \arg \max_{a=1,\cdots,n} \omega_a. \quad (7)$$

In general, obtaining the myopic action in each time slot requires the recursive update of the information state as given in (1), which requires the knowledge of the transition probabilities $\{p_{ij}\}$. Interestingly, it has been shown in [2], [3] that for this problem at hand, the myopic policy has a simple structure that does not need the update of the information state 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 in the next slot if the channel in the current slot is sensed to be "good". Otherwise, the user switches to the channel visited the longest time ago. When $p_{11} < p_{01}$, the myopic action is to stay after observing a "bad" channel 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 that has not been visited for the longest time, which can be any of the channels that have never been visited if such channels exist. Note that this simple structure of the myopic policy reveal that other than the order of p_{11} and p_{01}, the knowledge of the transition probabilities are unnecessary.

IV. The Optimality of Myopic Policy

In this section, we show that the myopic policy, with a simple and robust structure, is optimal under certain conditions. For convenience, we adopt the following notation.

$$\begin{aligned} V_t(\bar{\omega}; a = i) &:= E[R_i(\bar{\omega}) + \beta V_{t+1}(\mathcal{T}(\bar{\omega}|i, s_i(t)))] \\ &= \omega_i + \beta\omega_i V_{t+1}(\mathcal{T}(\bar{\omega}|i,1)) \\ &\quad +\beta(1-\omega_i)V_{t+1}(\mathcal{T}(\bar{\omega}|i,0)). \end{aligned}$$

Note that $V_t(\bar{\omega}) = \max_a V_t(\bar{\omega}; a)$.

We first prove the optimality of the myopic policy for (P1) under the following assumption/condition, and then extend the result to (P2) and (P3).

Assumption 1: The transition probabilities p_{01} and p_{11} are such that

$$\begin{aligned} p_{11} - p_{01} &\geq 0 \\ 1 + p_{01}(x - x^2) - 2x &\geq 0 \end{aligned}$$

where $x = \beta(p_{11} - p_{01})$.
The first inequality in (8) ensures that $\tau(\omega - \omega')$ is increasing in $\omega - \omega'$. In particular,

$$\tau(\omega - \omega') = (\omega - \omega')(p_{11} - p_{01}) \geq (\omega - \omega') \quad (8)$$

We also note that when $\beta \leq 0.5$, the second inequality in (8) always holds.

A. Finite Horizon

Our main results are summarized in the following theorem.

Theorem 1: Consider Problem (P1). Under Assumption 1, the myopic policy is optimal, i.e. for $\forall t, 0 \leq t < T$, and $\forall \bar{\omega} = [\omega_1, \cdots, \omega_N] \in [0,1]^N$,

$$V_t(\bar{\omega}; u(t) = 1) - V_t(\bar{\omega}; u(t) = i) \geq 0, \quad (9)$$

if $\omega_1 \geq \omega_i$ for $i = 1, \cdots, n$.

Proof: The proof is inductive and follows the following steps: assuming the optimality of myopic policy at times $t, t+1, \ldots, T$, we prove a set of inequalities, given by Lemmas 1-2. Using these inequalities, we then prove the optimality of myopic policy at time $t-1$ in Lemma 4. Note that

the optimality of myopic policy at time T is straightforward. The proofs of Lemmas 1-4 are given in the Appendix.

Lemma 1 (Monotonicity of Value Function:): Assume that the value function is monotone at times $t+1, t+2, \ldots, T$. Consider $\bar{\omega}$ and $\bar{\omega}'$ such that $\omega_1 \geq \omega'$ and $\omega_j = \omega_j^{'}, \forall j \neq 1$. We have

$$V_t(\bar{\omega}) = V_t(\bar{\omega}') \geq 0. \tag{10}$$

Lemma 2: Suppose the myopic policy is optimal at time t. Consider $\bar{\omega}$ and $\bar{\omega}'$ such that $\omega_1 \geq \omega'$ and $\omega_j = \omega_j^{'}, \forall j \neq 1$. We have

$$V_t(\bar{\omega}) - V_t(\bar{\omega}') \leq \frac{\omega_1 - \omega_1'}{1 - \beta(p_{11} - p_{01})}. \tag{11}$$

We also identify a tighter upper bound when $\omega_2 = p_{11}$:

Lemma 3: Suppose the myopic policy is optimal at time t. Consider $\bar{\omega}$ and $\bar{\omega}'$ such that $\omega_1 \geq \omega'$ and $\omega_j = \omega_j^{'}, \forall j \neq 1$, and $\omega_2 = \omega_2^{'} = p_{11}$. We have

$$V_t(\bar{\omega}) - V_t(\bar{\omega}') \leq \beta \frac{(p_{11} - p_{01})(\omega_1 - \omega_1')}{1 - \beta(p_{11} - p_{01})}. \tag{12}$$

The next lemma provides a comparison between two actions followed by optimal policies and establishes the advantage of the myopic action.

Lemma 4: Consider $\bar{\omega}$ where $\omega_1 \geq \omega_i,\ i = 1, \cdots, n$. If the myopic policy is optimal at times $t, t+1, \ldots, T$, then

$$V_{t-1}(\bar{\omega}; a(t) = 1) - V_{t-1}(\bar{\omega}; a(t) = 2) \geq 0. \tag{13}$$

■

B. Infinite Horizon

Now we consider extensions of the above result to (P2) and (P3), i.e., to show that the myopic policy is also optimal for (P2) and (P3) under the same condition. Intuitively, this holds due to the fact that the stationary optimal policy of the finite horizon problem is independent of the horizon as well as the discount factor. The theorems below concretely establish this.

We point out that the proofs of Theorems 2 and 3 do not rely on Assumption 1, but rather the optimality of the myopic policy for (P1). Indeed if the optimality of the myopic policy for (P1) can be established under weaker conditions, the proofs of Theorems 2 and 3 can be easily modified/extended to established its optimality under the same weaker condition for (P2) and (P3), respectively.

Theorem 2: Consider (P2) for $0 \leq \beta < 1$. Under Assumption 1 the myopic policy is optimal. Furthermore, its optimal policy value is the limiting optimal policy value of (P1) as the time horizon goes to infinity, i.e., we have $\max_\pi J_\beta^\pi(\bar{\omega}) = \lim_{T\to\infty} \max_\pi J_T^\pi(\bar{\omega})$.

Proof: We first use the Bounded Convergence Theorem to establish the fact that under any deterministic stationary Markov policy π, we have $J_\beta^\pi(\bar{\omega}) = \lim_{T\to\infty} J_T^\pi(\bar{\omega})$. We prove this by noting

$$\begin{aligned} J_\beta^\pi(\bar{\omega}) &= E^\pi[\lim_{T\to\infty} \sum_{t=1}^{T} \beta^{t-1} R_{\pi(t)}(\bar{\omega}(t)) | \bar{\omega}(1) = \bar{\omega}] \\ &= \lim_{T\to\infty} E^\pi[\sum_{t=1}^{T} \beta^{t-1} R_{\pi(t)}(\bar{\omega}(t)) | \bar{\omega}(1) = \bar{\omega}] \\ &= \lim_{T\to\infty} J_T^\pi(\bar{\omega}) \end{aligned} \tag{14}$$

where the second equality is due to the bounded convergence theorem. This proves the second part of the theorem by noting that we can interchange maximization and limit since the action space is finite.

Denote the myopic policy by π^*. We now establish the optimality of π^* for (P2). From Theorem 1, we know:

$$\begin{aligned} J_T^{\pi^*}(\bar{\omega}) = \max_{a=i} (&\omega_i + \beta\omega_i J_{T-1}^{\pi^*}(\mathcal{T}(\bar{\omega}|i,1)) \\ &+ \beta(1-\omega_i) J_{T-1}^{\pi^*}(\mathcal{T}(\bar{\omega}|i,0)). \end{aligned}$$

Taking limit of both sides, we have

$$\begin{aligned} J_\beta^{\pi^*}(\bar{\omega}) = \max_{a=i} (&\omega_i + \beta\omega_i J_\beta^{\pi^*}(\mathcal{T}(\bar{\omega}|i,1)) \\ &+ \beta(1-\omega_i) J_\beta^{\pi^*}(\mathcal{T}(\bar{\omega}|i,0)). \end{aligned} \tag{15}$$

Note that (15) is nothing but the dynamic programming equation for the infinite horizon discounted reward problem given in (5). From the uniqueness of the dynamic programming solution, then, we have

$$J_\beta^{\pi^*}(\bar{\omega}) = V_\beta(\bar{\omega}) = \max_\pi J_\beta^\pi(\bar{\omega})$$

hence, the optimality of the myopic policy.

■

Theorem 3: Consider (P3) with the expected average reward and under the ergodicity assumption $|p_{11} - p_{00}| < 1$. Myopic policy is optimal for problem (P3).

Proof: We consider the infinite horizon discounted cost for $\beta < 1$:

$$\begin{aligned} J_\beta^{\pi^*}(\bar{\omega}) = \max_{a=i} \Big\{ &\omega_i + \beta\omega_i J_\beta^{\pi^*}(\mathcal{T}((\bar{\omega})|i,1)) \\ &+ \beta(1-\omega_i) J_\beta^{\pi^*}(\mathcal{T}((\bar{\omega})|i,0)) \Big\}. \end{aligned} \tag{16}$$

This can be written as

$$\begin{aligned} &(1-\beta) J_\beta^{\pi^*}(\bar{\omega}) \\ &= \max_{a=i} (\omega_i + \beta\omega_i \left[J_\beta^{\pi^*}(\mathcal{T}((\bar{\omega})|i,1)) - J_\beta^{\pi^*}(\bar{\omega}) \right] \\ &\quad + \beta(1-\omega_i) \left[J_\beta^{\pi^*}(\mathcal{T}((\bar{\omega})|i,0)) - J_\beta^{\pi^*}(\bar{\omega}) \right]. \end{aligned}$$

Notice that the boundedness of the reward function and compactness of information state implies that the sequence of $(1-\beta) J_\beta^{\pi^*}(\bar{\omega})$ is bounded. In other words, there exist a converging sequence $\beta_k \to 1$ such that

$$\lim_{k\to\infty} (1-\beta_k) J_{\beta_k}^{\pi^*}(\bar{\omega}^*) := J^*, \tag{17}$$

where $\omega_i^* := \frac{p_{01}}{1-p_{11}+p_{01}}$ is the stationary belief (the limiting belief when channel i is not sensed for a long time). Also define

$$h^{\pi^*}(\bar{\omega}) := \lim_{k\to\infty}\left[J_{\beta_k}^{\pi^*}(\bar{\omega}) - J_{\beta_k}^{\pi^*}(\bar{\omega}^*)\right]. \quad (18)$$

Note that applying Lemma 2 (in the limit of $T \to \infty$) together with the assumption that $-1 < p_{11} - p_{00} < 1$ implies that there exists $K := \frac{n|p_{11}-p_{01}|}{1-|p_{11}-p_{01}|}$ such that

$$\left|J_\beta^{\pi^*}\left(\mathcal{T}\left((\bar{\omega})|i,0\right)\right) - J_\beta^{\pi^*}(\bar{\omega})\right| \leq K. \quad (19)$$

This implies that

$$\begin{aligned} J^* &= \lim_{k\to\infty}(1-\beta_k)J_{\beta_k}^{\pi^*}(\bar{\omega}) \\ &= \lim_{k\to\infty}(1-\beta_k)J_{\beta_k}^{\pi^*}(\bar{\omega}^*) + (1-\beta_k)\left[J_{\beta_k}^{\pi^*}(\bar{\omega}) - J_{\beta_k}^{\pi^*}(\bar{\omega}^*)\right]. \end{aligned}$$

In other words,

$$\begin{aligned} J^* &= \lim_{k\to\infty}(1-\beta_k)J_{\beta_k}^{\pi^*}(\bar{\omega}) \\ &= \lim_{k\to\infty}\max_{a=i}(\omega_i + \beta_k\omega_i\left[J_\beta^{\pi^*}\left(\mathcal{T}\left(\bar{\omega}|i,1\right)\right) - J_{\beta_k}^{\pi^*}(\bar{\omega})\right] \\ &\quad +\beta_k(1-\omega_i)\left[J_{\beta_k}^{\pi^*}\left(\mathcal{T}\left(\bar{\omega}|i,0\right)\right) - J_{\beta_k}^{\pi^*}(\bar{\omega})\right] \\ &= \max_{a=i}\left[\omega_i + \omega_i h^{\pi^*}\left(\mathcal{T}\left(\bar{\omega}|i,1\right)\right) + (1-\omega_i)h^{\pi^*}\left(\mathcal{T}\left(\bar{\omega}|i,0\right)\right)\right] \\ &\quad -h^{\pi^*}(\bar{\omega}). \end{aligned} \quad (20)$$

It thus follows that

$$\begin{aligned} J^* + h^{\pi^*}(\bar{\omega}) = \max_{a=i}\Big[&\omega_i + \omega_i h^{\pi^*}\left(\mathcal{T}\left(\bar{\omega}|i,1\right)\right) + \\ &(1-\omega_i)h^{\pi^*}\left(\mathcal{T}\left(\bar{\omega}|i,0\right)\right)\Big] \end{aligned} \quad (21)$$

Note that (21) is nothing but the DP equation as given by (6). This implies that J^* is the maximum average reward, i.e. $J^* = \max_\pi J_\infty^\pi(\bar{\omega}(t))$.

Replacing (16) with

$$\begin{aligned} J_\beta^{\pi^*}(\bar{\omega}) = \omega_{\pi(\bar{\omega})} &+ \beta\omega_{\pi(\bar{\omega})}J_\beta^{\pi^*}\left(\mathcal{T}\left((\bar{\omega})|\pi(\bar{\omega}),1\right)\right) \\ &+\beta(1-\omega_{\pi(\bar{\omega})})J_\beta^{\pi^*}\left(\mathcal{T}\left((\bar{\omega})|\pi(\bar{\omega}),0\right)\right), \end{aligned}$$

we repeat (17)-(20) to arrive at the following:

$$\begin{aligned} J^* + h^{\pi^*}(\bar{\omega}) = \omega_{\pi(\bar{\omega})} + \omega_{\pi(\bar{\omega})}h^{\pi^*}&\left(\mathcal{T}\left(\bar{\omega}|\pi(\bar{\omega}),1\right)\right) + \\ &(1-\omega_{\pi(\bar{\omega})})h^{\pi^*}\left(\mathcal{T}\left(\bar{\omega}|\pi(\bar{\omega}),0\right)\right). \end{aligned} \quad (22)$$

From part (ii) of Fact 3, we now have the optimality of myopic policy. ■

V. Discussion and Future Work

The general problem of opportunistic sensing and access arises in many multi-channel communication contexts. For cases where the stochastic evolution of channels can be modelled as i.i.d. two-state Markov chains, we have shown that a simple and robust myopic policy is optimal for several related problem formulations, under some assumptions on the channel state transition probabilities. The main open problem pertaining to our work is whether these assumptions can be further relaxed to prove the general conjecture that the myopic policy is optimal for all 2-state transition matrices.

The studied problem lies at the intersection of two well studied problems in stochastic control, namely, POMDP [4] and the restless bandit problems [8]. Viewing the problem as a POMDP was key in allowing us to establish many important structural properties of the solution. For instance, the finiteness of the underlying (unobservable) state space was key in establishing the existence of an optimal stationary Markov policy (P2) and (P3): a very useful fact whose validity can be difficult to establish in a general restless bandit context. Similarly, identifying belief as an information state is a consequence of the POMDP formulation.

At the same time, our problem can be viewed as a special case of restless bandit problem. Differing from the classical multi-armed bandit problem where only one project can be activated at a time and the passive projects do not change state, a restless multi-armed bandit process allows passive projects to change states. While the classical bandit problems can be solved optimally using the Gittin's Index [6], restless bandit problems are known to be PSPACE-hard in general [7]. Whittle proposed a Gittin's-like indexing heuristic for the restless bandits problem [8] which is known to be asymptotically optimal under certain limiting regime [9]. Beyond this asymptotic result, relatively little is known about the structure of the optimal policies for the general restless bandit problems (see [11] for near-optimal heuristics). In fact, even the indexability of a bandit can be rather complicated to establish [12]. The optimality of the myopic policy shown in this paper not only suggests the indexability of certain special cases of restless bandit processes, where the index is expected to be closely related to the expected immediate reward, but also identifies (non-asymptotic) conditions where an index policy is, in fact, optimal.

In the future, we plan to explore further the connections between the myopic policy and Whittle's indexing heuristic, and investigate whether the optimality of myopic policy can be established for restless bandit problems under relaxed conditions.

References

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

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

[3] 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.

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

[5] E. Fernandez-Gaucherand, and A. Arapostathis and S. I. Marcus, "On the average cost optimality equation and the structure of optimal policies for partially observable Markov decision processes", in *Annals of Operations Research*, Volume 29, December, 1991.

[6] J.C. Gittins, "Bandit Processes and Dynamic Allocation Indices," *Journal of the Royal Statistical Society*, Series B, 41, pp. 148-177, 1979.

[7] C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of optimal queueing network control." in*Mathematics of Operations Research*, Volume. 24, 1999
[8] P. Whittle, "Restless bandits: Activity allocation in a changing world", in *Journal of Applied Probability*, Volume 25, 1988.
[9] R. R. Weber and G. Weiss, "On an index policy for restless bandits," *Journal of Applied Probability*, 27:637–648, 1990.
[10] R. Kumar and P. Varaiya, *Stochastic Control*, Prentice-Hall, 1986.
[11] D. Betsimas and J. E. Niño-Mora, "Restless bandits, linear programming relaxations, and a primal-dual heuristic," in *Operations Research*, Volume. 48, No. 1, January-February 2000.
[12] J. E. Niño-Mora, "Restless Bandits, Partial Conservation Laws and Indexability" *Adv. Appl. Probab.* 33, 76–98, 2001.

APPENDIX

Before proceeding, we note that due to the fact that all channels are identical, the information state vector is unordered. That is, $V_t(\bar{\omega}) = V_t(\bar{\omega}_p)$ where $\bar{\omega}_p$ is any permutation of $\bar{\omega}$, for any t. Using this result in the proofs below we frequently reorder the state vector for convenience. We also frequently write $V_t(\bar{\omega})$ as $V_t(\omega_1, \cdots, \omega_n)$.

Proof of Lemma 1

For convenience, define

$$\tau^{-j}(\bar{\omega}) := (\tau(\omega_1), \ldots, \tau(\omega_{j-1}), \tau(\omega_{j+1}), \ldots, \tau(\omega_n)).$$

Using induction, the basis is obviously true. Assume the monotonicity holds for $t+1, t+2, \cdots, T$. Let j^* be the optimal action for the state $\bar{\omega}'$ at time t. Since j^* is in general not necessarily optimal for the state $\bar{\omega}$, we have the following.

$$\begin{aligned}
&V_t(\bar{\omega}) - V_t(\bar{\omega}') \\
&\geq V_t(\bar{\omega}; a(t) = j^*) - V_t(\bar{\omega}'; a(t) = j^*) \\
&= \omega_{j^*} + \beta\omega_{j^*} V_{t+1}\left(p_{11}, \tau^{-j^*}(\bar{\omega})\right) \\
&\quad + \beta(1-\omega_{j^*}) V_{t+1}\left(p_{01}, \tau^{-j^*}(\bar{\omega})\right) \\
&\quad - \omega'_{j^*} - \beta\omega'_{j^*} V_{t+1}\left(p_{11}, \tau^{-j^*}(\bar{\omega}')\right) \\
&\quad - \beta(1-\omega'_{j^*}) V_{t+1}\left(p_{01}, \tau^{-j^*}(\bar{\omega}')\right) \\
&= \omega_{j^*} - \omega'_{j^*} \\
&\quad + \beta\omega'_{j^*}\left[V_{t+1}\left(p_{11}, \tau^{-j^*}(\bar{\omega})\right) - V_{t+1}\left(p_{11}, \tau^{-j^*}(\bar{\omega}')\right)\right] \\
&\quad + \beta(1-\omega'_{j^*})\left[V_{t+1}\left(p_{01}, \tau^{-j^*}(\bar{\omega})\right) - \right. \\
&\qquad \left. V_{t+1}\left(p_{01}, \tau^{-j^*}(\bar{\omega}')\right)\right] + \beta(\omega_{j^*} - \omega'_{j^*}) \\
&\qquad \left[V_{t+1}\left(p_{11}, \tau^{-j^*}(\bar{\omega})\right) - V_{t+1}\left(p_{01}, \tau^{-j^*}(\bar{\omega})\right)\right] \\
&\geq 0, \qquad (23)
\end{aligned}$$

where the last inequality holds due to monotonicity of operator τ and the induction hypothesis.

Proof of Lemma 2

For $t = T$ is true.

Induction hypothesis: The assertion of lemma is true for $t+1, t+2, \ldots, T$, we need to establish the upper bound for t.

Case 1: $\omega_1 \geq \omega_j, j = 2, \ldots, n$.

$$\begin{aligned}
&V_t(\omega_1, \omega_2, \ldots, \omega_n) - V_t(\omega'_1, \omega_2, \ldots, \omega_n) \\
&\leq \omega_1 + \beta\omega_1 V_{t+1}\left(p_{11}, \tau(\omega_2), \ldots, \tau(\omega_n)\right) \\
&\quad + \beta(1-\omega_1) V_{t+1}\left(p_{01}, \tau(\omega_2), \ldots, \tau(\omega_n)\right) \\
&\quad - \beta\omega'_1 V_{t+1}\left(p_{11}, \tau(\omega_2), \ldots, \tau(\omega_n)\right) \\
&\quad - \beta(1-\omega'_1) V_{t+1}\left(p_{01}, \tau(\omega_2), \ldots, \tau(\omega_n)\right) \\
&= \omega_1 - \omega'_1 + \beta(\omega_1 - \omega'_1)\left[V_{t+1}\left(p_{11}, \tau(\omega_2), \ldots, \tau(\omega_n)\right) - \right. \\
&\quad \left. V_{t+1}\left(p_{01}, \tau(\omega_2), \ldots, \tau(\omega_n)\right)\right] \\
&\leq (\omega_1 - \omega'_1)\left[1 + \beta\frac{p_{11} - p_{01}}{1 - \beta p_{11} + \beta p_{01}}\right] = \frac{\omega_1 - \omega'_1}{1 - \beta p_{11} + \beta p_{01}}.
\end{aligned}$$

Case 2: $\max_k \omega_k = \omega_j > \omega_1 \geq \omega'_1$.

$$\begin{aligned}
&V_t(\omega_1, \omega_2, \ldots, \omega_n) - V_t(\omega'_1, \omega_2, \ldots, \omega_n) \\
&= \omega_j + \beta\omega_j V_{t+1}\left(\tau(\omega_1), \ldots, p_{11}, \tau(\omega_{j+1}), \ldots, \tau(\omega_n)\right) \\
&\quad + \beta(1-\omega_j) V_{t+1}\left(\tau(\omega_1), \ldots, p_{01}, \tau(\omega_{j+1}), \ldots, \tau(\omega_n)\right) \\
&\quad - \omega_j - \beta\omega_j V_{t+1}\left(\tau(\omega'_1), \ldots, p_{11}, \tau(\omega_{j+1}), \ldots, \tau(\omega_n)\right) \\
&\quad - \beta(1-\omega_j) V_{t+1}\left(\tau(\omega'_1), \ldots, p_{01}, \tau(\omega_{j+1}), \ldots, \tau(\omega_n)\right) \\
&= \beta\omega_j\left[V_{t+1}\left(\tau(\omega_1), \ldots, p_{11}, \tau(\omega_{j+1}), \ldots, \tau(\omega_n)\right)\right. \\
&\qquad \left. - V_{t+1}\left(\tau(\omega'_1), \ldots, p_{11}, \tau(\omega_{j+1}), \ldots, \tau(\omega_n)\right)\right] \\
&\quad + \beta(1-\omega_j)\left[V_{t+1}\left[\tau(\omega_1), \ldots, p_{01}, \tau(\omega_{j+1}), \ldots, \tau(\omega_n)\right)\right. \\
&\qquad \left. - V_{t+1}\left(\tau(\omega'_1), \ldots, p_{01}, \tau(\omega_{j+1}), \ldots, \tau(\omega_n)\right)\right] \\
&\leq \frac{\beta\tau(\omega_1 - \omega'_1)}{1 - \beta p_{11} + \beta p_{01}} = \beta(\omega_1 - \omega'_1)\frac{p_{11} - p_{01}}{1 - \beta p_{11} + \beta p_{01}} \\
&\leq (\omega_1 - \omega'_1)\frac{p_{11} - p_{01}}{1 - \beta p_{11} + \beta p_{01}}. \qquad (24)
\end{aligned}$$

Proof of Lemma 3

This lemma is a corollary of Lemma 2, as (12) is nothing but a special case of (24) (we use here the tighter bound that includes β as the multiplicative factor).

Proof of Lemma 4

The myopic policy is optimal at time t. We appeal to Lemmas 1, 2, and 3 to establish inequalities at time t:

$$
\begin{aligned}
&V_{t-1}(\omega_1,\omega_2,\ldots,\omega_n;a(t)=1)-V_{t-1}(\omega_1,\omega_2,\ldots,\omega_n;a(t)=2)\\
&= \omega_1-\omega_2+\beta\omega_1 V_t\left(p_{11},\tau(\omega_2),\ldots,\tau(\omega_n)\right)\\
&\quad +\beta(1-\omega_1)V_t\left(p_{01},\tau(\omega_2),\ldots,\tau(\omega_n)\right)\\
&\quad -\beta\omega_2 V_t\left(\tau(\omega_1),p_{11},\ldots,\tau(\omega_n)\right)\\
&\quad -\beta(1-\omega_2)V_t\left(\tau(\omega_1),p_{01},\ldots,\tau(\omega_n)\right)\\
&= \omega_1-\omega_2\\
&\quad +\beta(\omega_1-\omega_2)\left[V_t\left(p_{11},\tau(\omega_2),\ldots,\tau(\omega_n)\right)-\right.\\
&\quad \left.V_t\left(p_{01},\tau(\omega_2),\ldots,\tau(\omega_n)\right)\right]\\
&\quad +\beta\omega_2\left[V_t\left(p_{11},\tau(\omega_2),\ldots,\tau(\omega_n)\right)-\right.\\
&\quad \left.V_t\left(\tau(\omega_1),p_{11},\ldots,\tau(\omega_n)\right)\right]\\
&\quad +\beta(1-\omega_2)\left[V_t\left(p_{01},\tau(\omega_2),\ldots,\tau(\omega_n)\right)-\right.\\
&\quad \left.V_t\left(\tau(\omega_1),p_{01},\ldots,\tau(\omega_n)\right)\right]\\
&\geq \omega_1-\omega_2\\
&\quad -\beta^2\omega_2(\omega_1-\omega_2)\frac{(p_{11}-p_{01})^2}{1-\beta p_{11}+\beta p_{01}}\\
&\quad -\beta(1-\omega_2)(\omega_1-\omega_2)\frac{p_{11}-p_{01}}{1-\beta p_{11}+\beta p_{01}}\\
&= (\omega_1-\omega_2)\left[1-\beta^2\omega_2\frac{(p_{11}-p_{01})^2}{1-\beta p_{11}+\beta p_{01}}\right.\\
&\quad \left.-\beta(1-\omega_2)\frac{p_{11}-p_{01}}{1-\beta p_{11}+\beta p_{01}}\right]\\
&= (\omega_1-\omega_2)\left[1+\omega_2\left(\frac{\beta(p_{11}-p_{01})-\beta^2(p_{11}-p_{01})^2}{1-\beta p_{11}+\beta p_{01}}\right)\right.\\
&\quad \left.-\beta\frac{p_{11}-p_{01}}{1-\beta p_{11}+\beta p_{01}}\right]\\
&\geq (\omega_1-\omega_2)\left[1+p_{01}\left(\frac{\beta(p_{11}-p_{01})-\beta^2(p_{11}-p_{01})^2}{1-\beta p_{11}+\beta p_{01}}\right)\right.\\
&\quad \left.-\beta\frac{p_{11}-p_{01}}{1-\beta p_{11}+\beta p_{01}}\right]
\end{aligned}
$$

Let $x=\beta(p_{11}-p_{01})$. It can be shown that the term in the brackets is non-negative if and only if $1+p_{01}(x-x^2)-2x\geq 0$, as per Assumption 1.