Optimality of Myopic Sensing in Multichannel Opportunistic Access

Sahand Haji Ali Ahmad, Mingyan Liu, *Member, IEEE*, Tara Javidi, *Member, IEEE*, Qing Zhao, *Senior Member, IEEE*, and Bhaskar Krishnamachari

Abstract—This paper considers opportunistic communication over multiple channels where the state ("good" or "bad") of each channel evolves as independent and identically distributed (i.i.d.) Markov processes. A user, with limited channel sensing capability, chooses one channel to sense and decides whether to use the channel (based on the sensing result) in each time slot. A reward is obtained whenever the user senses and accesses a "good" channel. The objective is to design a channel selection policy that maximizes the expected total (discounted or average) reward accrued over a finite or infinite horizon. This problem can be cast as a partially observed Markov decision process (POMDP) or a restless multiarmed bandit process, to which optimal solutions are often intractable. This paper shows that a myopic policy that maximizes the immediate one-step reward is optimal when the state transitions are positively correlated over time. When the state transitions are negatively correlated, we show that the same policy is optimal when the number of channels is limited to two or three, while presenting a counterexample for the case of four channels. This result finds applications in opportunistic transmission scheduling in a fading environment, cognitive radio networks for spectrum overlay, and resource-constrained jamming and antijamming.

Index Terms—Cognitive radio, Gittins index, myopic policy, opportunistic access, partially observed Markov decision process (POMDP), restless bandit, Whittle's index.

Fig. 1. The Markov channel model.

I. Introduction

WE consider a communication system in which a sender has access to multiple 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 properties of these channels to maximize its transmission rate by switching opportunistically across channels.

We model this problem in the following manner. As shown in Fig. 1, there are n channels, each of which evolves as an independent and identically distributed (i.i.d.), two-state discrete-time Markov chain. The two states for each channel—"good" (or state 1) and "bad" (or state 0)—indicate the desirability of transmitting over that channel at a given time slot. The state transition probabilities are given by p_{ij}, i, $j = 0, 1$. In each time slot, 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 finite or infinite horizon. This problem can be described as a partially observed Markov decision process (POMDP) [1] since the states of the underlying Markov chains are not fully observed. It can also be cast as a special case of the restless bandit problems [2]; more discussion on this is given in Section VII.

This formulation is broadly applicable to several domains. It arises naturally in opportunistic spectrum access (OSA) [3], [4], where the sender is a secondary user, and the channel states describe the occupancy by primary users. In the OSA problem, the secondary user 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 develop 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 this paper, we examine the optimality of a simple myopic policy for the opportunistic access problem outlined above. Specifically, we show that the myopic policy is optimal for arbitrary n when $p_{11} \geq p_{01}$. We also show that when $p_{11} < p_{01}$, it is optimal for $n = 3$ while presenting a finite horizon counterexample showing that it is in general not optimal for $n \geq 4$. We further generalize these results to related formulations involving discounted and average rewards over an infinite horizon.

These results extend and complement those reported in prior work [5]. Specifically, it has been shown in [5] that for all n, the myopic policy has an elegant and robust structure that obviates the need to know the channel state transition probabilities and reduces channel selection to a simple round-robin procedure. Based on this structure, the optimality of the myopic policy for $n = 2$ was established and the performance of the myopic policy, in particular, the scaling property with respect to n, analyzed in [5]. It was conjectured in [5] that the myopic policy is optimal for any n. This conjecture was partially addressed in a preliminary conference version [6], where the optimality was established under certain restrictive conditions on the channel parameters and the discount factor. In the present paper, we significantly relax these conditions and formerly prove this conjecture under the condition $p_{11} \geq p_{01}$. We also provide a counterexample for $p_{11} < p_{01}$.

We would like to emphasize that compared to earlier work [5], [6], the approach used in this paper relies on a coupling argument, which is the key to extending the optimality result to the arbitrary n case. Earlier techniques were largely based on exploiting the convex analytic properties of the value function, and were shown to have difficulty in overcoming the $n = 2$ barrier without further conditions on the discount factor or transition probabilities. This observation is somewhat reminiscent of the results reported in [7], where a coupling argument was also used to solve an n-queue problem while earlier versions [8] using value function properties were limited to a two-queue case. We invite the interested reader to refer to [9], an important manuscript on monotonicity in Markov decision processes (MDPs) which explores the power as well as the limitation of working with analytic properties of value functions and dynamic programming operators as we had done in our earlier work. In particular, [9, Sec. 9.5] explores the difficulty of using such techniques for multidimensional problems where the number of queues is more than $n = 2$; [9, Ch. 12] contrasts this proof technique with the stochastic coupling arguments, which our present work uses.

The remainder of this paper is organized as follows. We formulate the problem in Section II and illustrate the myopic policy in Section III. In Section IV, we prove that the myopic policy is optimal in the case of $p_{11} \geq p_{01}$, and show in Section V that it is in general not optimal when this condition does not hold. Section VI extends the results from finite horizon to infinite horizon. We discuss our work within the context of restless bandit problems as well as some related work in this area in Section VII. Section VIII concludes the paper.

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 Fig. 1.

The system operates in discrete time steps indexed by t, $t = 1, 2, \ldots, T$, 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 until $t + 1$ to make another choice. This process repeats sequentially until the time horizon expires.

This abstraction is primarily motivated by the following multichannel access scenario, where a secondary user seeks spectrum opportunity in between a primary user's activities. Specifically, time is divided into frames and at the beginning of each frame there is a designated time slot for the primary user to reserve that frame and for secondary users to perform channel sensing. If a primary user intends to use a frame it will simply remain active in a channel (or multiple channels) during that sensing time slot (i.e., reservation is by default for a primary user in use of the channel), in which case a secondary user will find the channel(s) busy and not attempt to use it for the duration of that frame. If the primary user is inactive during this sensing time slot, then the remainder of the frame is open to secondary users. Such a structure provides the necessary protection for the primary user as channel sensing (in particular active channel sensing that involves communication between a pair of users) conducted at arbitrary times can cause undesirable interference.

Within such a structure, a secondary user has a limited amount of time and capability to perform channel sensing, and may only be able to sense one or a subset of the channels before the sensing time slot ends. And if all these channels are unavailable then it will have to wait till the next sensing time slot. In this paper, we will limit our attention to the special case where the secondary user only has the resources to sense one channel within this slot. Conceptually, our formulation is easily extended to the case where the secondary user can sense multiple channels at a time within this structure, although the corresponding results differ, see, e.g., [10].

Note that we have assumed in this formulation we do not explicitly model the cost of channel sensing; it is implicit in the fact that the user is limited in how many channels it can sense at a time. Alternative formulations have been studied where sensing costs are explicitly taken into consideration in a user's sensing and access decision, see, e.g., [11] and [12].

In this formulation, we have assumed that sensing errors are negligible. Techniques used in this paper may be applicable in proving the optimality of the myopic policy under imperfect sensing. The reason behind this is that our proof exploits the simple structure of the myopic policy, which continues to hold when sensing is subject to error as shown in [13].

Note that the system is not fully observable to the user, i.e., the user does not know the exact state of the system when making the sensing decision. Specifically, a channel goes through state transition at time t^- (or any time between $(t-1, t)$), thus, when the user makes the channel sensing decision at time t, it does not have the true state of the system at time t, which we denote by $\bar{s}(t) = [s_1(t), s_2(t), \ldots, s_n(t)] \in \{0, 1\}^n$. Furthermore, even after its action (at time t^+) it only gets to observe the true state of one channel, which goes through another transition at or before time $(t+1)^-$. The user's action space at time t is given by the finite set $\{1, 2, \ldots, n\}$, and we will use $a(t) = i$ to denote that the user selects channel i to sense at time t. For clarity, we will denote the outcome/observation of channel sensing at time t following the action $a(t)$ by $h_{a(t)}(t)$, which is essentially the true state $s_{a(t)}(t)$ of channel $a(t)$ at time t since we assume channel sensing to be error free.

It can be shown (see, e.g., [1], [14], [15]) that a sufficient statistic of such a system for optimal decision making, or the *information state* of the system [14], [15], is given by the conditional probabilities of the state each channel is in given all past actions and observations. Since each channel can be in one of two states, we denote this information state or belief vector by $\bar{\omega}(t) = [\omega_1(t), \ldots, \omega_n(t)] \in [0, 1]^n$, where $\omega_i(t)$ is the conditional probability that channel i is in state 1 at time t given all past states, actions, and observations.[1] Throughout the paper $\omega_i(t)$ will be referred to as the information state of channel i at time t, or simply the channel probability of i at time t.

Due to the Markovian nature of the channel model, the future information state is only a function of the current information state and the current action; i.e., it is independent of past history given the current information state and action. It follows that the information state of the system evolves as follows. Given that the state at time t is $\bar{\omega}(t)$ and action $a(t) = i$ is taken, $\omega_i(t+1)$ can take on two values: 1) p_{11} if the observation is that channel i is in a "good" state ($h_i(t) = 1$); this occurs with probability $P\{h_i(t) = 1 | \bar{\omega}(t)\} = \omega_i(t)$; 2) p_{01} if the observation is that channel i is in a "bad" state ($h_i(t) = 0$); this occurs with probability $P\{h_i(t) = 0 | \bar{\omega}(t)\} = 1 - \omega_i$. For any other channel $j \neq i$, the corresponding $\omega_j(t+1)$ can only take on one value (i.e., with probability 1): $\omega_j(t+1) = \tau(\omega_j(t))$ where the operator $\tau : [0, 1] \to [0, 1]$ is defined as

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

These transition probabilities are summarized in the following equation for $t = 1, 2, \ldots, T - 1$:

$$\{\omega_i(t+1) | \bar{\omega}(t), a(t)\}$$
$$= \begin{cases} p_{11} & \text{with prob. } \omega_i(t) \text{ if } a(t) = i \\ p_{01} & \text{with prob. } 1 - \omega_i(t) \text{ if } a(t) = i \\ \tau(\omega_i(t)) & \text{with prob. } 1 \text{ if } a(t) \neq i, \end{cases}$$
$$i = 1, 2, \ldots, n. \quad (2)$$

Also note that $\bar{\omega}(1) \in [0, 1]^n$ denotes the initial condition (information state in the form of conditional probabilities)

of the system, which may be interpreted as the user's initial belief about how likely each channel is in the good state before sensing starts at time $t = 1$. For the purpose of the optimization problems formulated below, this initial condition is considered given, which can be any probability vector.[2]

It is important to note that although in general a POMDP problem has an uncountable state space (information states are probability distributions), in our problem the state space is countable for any given initial condition $\bar{\omega}(1)$. This is because as shown earlier, the information state of any channel with an initial probability of ω can only take on the values $\{\omega, \tau^k(\omega), p_{01}, \tau^k(\omega), p_{11}, \tau^{k-1}(\omega)\}$, where $k = 1, 2, \ldots$, and $\tau^k(\omega) := \tau(\tau^{k-1}(\omega))$, which is a countable set.

For compactness of presentation, we will further use the operator $\mathcal{T}$ to denote the above probability distribution of the information state (the entire vector)

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

by noting that the operation given in (2) is applied to $\bar{\omega}(t)$ element-by-element. We will also use the following to denote the information state given observation outcome

$$\mathcal{T}(\bar{\omega}(t), a(t) | h_{a(t)}(t) = 1)$$
$$= (\tau(\omega_1(t)), \ldots, \tau(\omega_{a(t)-1}(t)), p_{11}, $$
$$\tau(\omega_{a(t)+1}(t)), \ldots, \tau(\omega_n(t))) \quad (4)$$
$$\mathcal{T}(\bar{\omega}(t), a(t) | h_{a(t)}(t) = 0)$$
$$= (\tau(\omega_1(t)), \ldots, \tau(\omega_{a(t)-1}(t)), p_{01}, $$
$$\tau(\omega_{a(t)+1}(t)), \ldots, \tau(\omega_n(t))). \quad (5)$$

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 $\bar{\omega} = [\omega_1, \ldots, \omega_n]$. The associated optimization problems ((P1)–(P3)) are formally defined as follows.

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

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

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

where β ($0 \leq \beta \leq 1$ for (P1) and $0 \leq \beta < 1$ for (P2)) is the discount factor, and $R_{\pi_t}(\bar{\omega}(t))$ is the reward collected under state $\bar{\omega}(t)$ when channel $a(t) = \pi_t(\bar{\omega}(t))$ is selected and $h_{a(t)}(t)$ is observed. This reward is given by $R_{\pi_t}(\bar{\omega}(t)) = 1$ with probability $\omega_{a(t)}(t)$ (when $h_{a(t)}(t) = 1$), and 0 otherwise.

---

[1]Note that this is a standard way of turning a POMDP problem into a classic MDP problem by means of information state, the main implication being that the state space is now uncountable.

[2]That is, the optimal solutions are functions of the initial condition. A reasonable choice, if the user has no special information other than the transition probabilities of these channels, is to simply use the steady-state probabilities of channels being in state "1" as an initial condition (i.e., setting $\omega_i(1) = \frac{p_{10}}{p_{01} + p_{10}}$).

The maximization in (P1) is over the class of deterministic Markov policies.[3] An admissible policy π, given by the vector $\pi = [\pi_1, \pi_2, \ldots, \pi_T]$, is thus such that π_t specifies a mapping from the current information state $\bar{\omega}(t)$ to a channel selection action $a(t) = \pi_t(\bar{\omega}(t)) \in \{1, 2, \ldots, n\}$. This is done without loss of optimality due to the Markovian nature of the underlying system, and due to known results on POMDPs. Note that the class of Markov policies in terms of information state are also known as seperated policies (see [15]). Due to finiteness of (unobservable) state spaces and action space in problem (P1), it is known that an optimal policy (over all random and deterministic, history-dependent and history-independent policies) may be found within the class of separated (i.e., deterministic Markov) policies (see, e.g., [15, Theorem 7.1, Ch. 6]), thus justifying the maximization and the admissible policy space.

In Section VI, we establish the existence of a stationary separated policy π^*, under which the supremum of the expected discounted reward as well as the supremum of expected average cost are achieved, hence justifying our use of maximization in (P2) and (P3). Furthermore, it is shown that under this policy the limit in (P3) exists and is greater than the limsup of the average performance of any other policy (in general history-dependent and randomized). This is a strong notion of optimality; the interpretation is that the most "pessimistic" average performance under policy π^* $\left(\liminf \frac{1}{T} J_T^{\pi^*}(\cdot) = \lim \frac{1}{T} J_T^{\pi^*}(\cdot)\right)$ is greater than the most "optimistic" performance under any other policy π $\left(\limsup \frac{1}{T} J_T^{\pi}(\cdot)\right)$. In much of the literature on MDP, this is referred to as the *strong optimality* for an expected average cost (reward) problem; for a discussion on this, see [16, p. 344].

III. OPTIMAL POLICY AND THE MYOPIC POLICY

A. Dynamic Programming Representations

Problems (P1)–(P3) defined in the previous section may be solved using their respective dynamic programming (DP) representations. Specifically, for problem (P1), we have the following recursive equations:

$$V_T(\bar{\omega}) = \max_{a=1,\ldots,n} E[R_a(\bar{\omega})]$$
$$V_t(\bar{\omega}) = \max_{a=1,\ldots,n} E[R_a(\bar{\omega}) + \beta V_{t+1}(\mathcal{T}(\bar{\omega}, a))]$$
$$= \max_{a=1,\ldots,n} (\omega_a + \beta \omega_a V_{t+1}(\mathcal{T}(\bar{\omega}, a|1))$$
$$+ \beta(1 - \omega_a)V_{t+1}(\mathcal{T}(\bar{\omega}, a|0))) \quad (6)$$

for $t = 1, 2, \ldots, T - 1$, where $V_t(\bar{\omega})$ is known as the value function, or the maximum expected future reward that can be accrued starting from time t when the information state is $\bar{\omega}$. In particular, we have $V_1(\bar{\omega}) = \max_\pi J_T^\pi(\bar{\omega})$, and an optimal deterministic Markov policy exists such that $a = \pi_t^*(\bar{\omega})$ achieves the maximum in (6) (see, e.g., [16, Ch. 4]). Note that since $\mathcal{T}$ is a conditional probability distribution (given in (3)), $V_{t+1}(\mathcal{T}(\bar{\omega}, a))$ is taken to be the expectation over this distribution when its argument is $\mathcal{T}$, with a slight abuse of notation, as expressed in (6).

Similar dynamic programming representations hold for (P2) and (P3) as given below. For problem (P2) there exists a

unique function $V_\beta(\cdot)$ satisfying the following fixed point equation:

$$V_\beta(\bar{\omega}) = \max_{a=1,\ldots,n} E[R_a(\bar{\omega}) + \beta V_\beta(\mathcal{T}(\bar{\omega}, a))]$$
$$= \max_{a=1,\ldots,n} (\omega_a + \beta \omega_a V_\beta(\mathcal{T}(\bar{\omega}, a|1))$$
$$+ \beta(1 - \omega_a)V_\beta(\mathcal{T}(\bar{\omega}, a|0))). \quad (7)$$

We have that $V_\beta(\bar{\omega}) = \max_\pi J_\beta^\pi(\bar{\omega})$, and that a stationary separated policy π^* is optimal if and only if $a = \pi^*(\bar{\omega})$ achieves the maximum in (7) [17, Theorem 7.1].

For problem (P3), we will show that there exist a bounded function $h_\infty(\cdot)$ and a constant scalar J satisfying the following equation:

$$J + h_\infty(\bar{\omega}) = \max_{a=1,\ldots,n} E[R_a(\bar{\omega}) + h_\infty(\mathcal{T}(\bar{\omega}, a))]$$
$$= \max_{a=1,\ldots,n} (\omega_a + \omega_a h_\infty(\mathcal{T}(\bar{\omega}, a|1))$$
$$+ (1 - \omega_a)h_\infty(\mathcal{T}(\bar{\omega}, a|0))). \quad (8)$$

The boundedness of h_∞ and the immediate reward implies that $J = \max_\pi J_\infty^\pi(\bar{\omega})$, and that a stationary separated policy π^* is optimal in the context of (P3) if $a = \pi^*(\bar{\omega})$ achieves the maximum in (8) [17, Theorems 6.1–6.3].

Solving (P1)–(P3) using the above recursive equations is in general computationally heavy. Therefore, instead of directly using the DP equations, the focus of this paper is on examining the optimality property of a simple, greedy policy. We define this algorithm next and show its simplicity in structure and implementation.

B. The Myopic Policy

A myopic or greedy 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 $\bar{\omega} = [\omega_1, \omega_2, \ldots, \omega_n]$ is given by

$$a^*(\bar{\omega}) = \arg \max_{a=1,\ldots,n} E[R_a(\bar{\omega})] = \arg \max_{a=1,\ldots,n} \omega_a. \quad (9)$$

In general, obtaining the myopic action in each time slot requires the successive update of the information state as given in (2), which explicitly relies on the knowledge of the transition probabilities $\{p_{ij}\}$ as well as the initial condition $\bar{\omega}(1)$. However, due to monotonicity, the myopic policy requires only the knowledge of the initial condition and the order of p_{11} and p_{01}, but not the precise values of these transition probabilities. To make the present paper self-contained, below we briefly describe how this policy works; more details may be found in [5].

Specifically, when $p_{11} \geq p_{01}$, the conditional probability updating function $\tau(\omega)$ is a monotonically increasing function, i.e., $\tau(\omega_1) \geq \tau(\omega_2)$ for $\omega_1 \geq \omega_2$. Therefore, the ordering of information states among channels is preserved when they are not observed. If a channel has been observed to be in state "1" (respectively, "0"), its probability at the next step becomes $p_{11} \geq \tau(\omega)$ (respectively, $p_{01} \leq \tau(\omega)$) for any $\omega \in [0, 1]$. In other words, a channel observed to be in state "1" (respectively, "0") will have the highest (respectively, lowest) possible information state among all channels.

---

[3]A Markov policy is a policy that derives its action from the current (information) state, rather than the entire history of states, see, e.g., [15].

These observations lead to the following implementation of the myopic policy. We take the initial information state $\bar{\omega}(1)$, order the channels according to their probabilities $\omega_i(1)$, and probe the highest one (top of the ordered list) with ties broken randomly. In subsequent steps, we stay in the same channel if the channel was sensed to be in state "1" (good) in the previous slot; otherwise, this channel is moved to the bottom of the ordered list, and we probe the channel currently at the top of the list. This in effect creates a round-robin style of probing, where the channels are cycled through in a fixed order. This circular structure is exploited in Section IV to prove the optimality of the myopic policy in the case of $p_{11} \geq p_{01}$.

When $p_{11} < p_{01}$, we have an analogous but opposite situation. The conditional probability updating function $\tau(\omega)$ is now a monotonically decreasing function, i.e., $\tau(\omega_1) \leq \tau(\omega_2)$ for $\omega_1 \geq \omega_2$. Therefore, the ordering of information states among channels is reversed at each time step when they are not observed. If a channel has been observed to be in state "1" (respectively, "0"), its probability at the next step becomes $p_{11} \leq \tau(\omega)$ (respectively, $p_{01} \geq \tau(\omega)$) for any $\omega \in [0, 1]$. In other words, a channel observed to be in state "1" (respectively, "0") will have the lowest (respectively, highest) possible information state among all channels.

As in the previous case, these similar observations lead to the following implementation. We take the initial information state $\bar{\omega}(1)$, order the channels according to their probabilities $\omega_i(1)$, and probe the highest one (top of the ordered list) with ties broken randomly. In each subsequent step, if the channel sensed in the previous step was in state "0" (bad), we keep this channel at the top of the list but completely reverse the order of the remaining list, and we probe this channel. If the channel sensed in the previous step was in state "1" (good), then we completely reverse the order of the entire list (including dropping this channel to the bottom of the list), and probe the channel currently at the top of the list. This alternating circular structure is exploited in Section V to examine the optimality of the myopic policy in the case of $p_{11} < p_{01}$.

IV. OPTIMALITY OF THE MYOPIC POLICY IN THE CASE OF $p_{11} \geq p_{01}$

In this section, we show that the myopic policy, with a simple and robust structure, is optimal when $p_{11} \geq p_{01}$. We will first show this for the finite horizon discounted cost case, and then extend the result to the infinite horizon case under both discounted and average cost criteria in Section VI.

The main assumption is formally stated as follows.

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

$$p_{11} - p_{01} \geq 0. \quad (10)$$

The main theorem of this section is as follows.

Theorem 1: Consider Problem (P1). Define $V_t(\bar{\omega}; a) := E[R_a(\bar{\omega}) + \beta V_{t+1}(\mathcal{T}(\bar{\omega}, a))]$, i.e., the value of the value function given in (6) when action a is taken at time t followed by an optimal policy. Under Assumption 1, the myopic policy is optimal, i.e., for $\forall t, 1 \leq t < T$, and $\forall \bar{\omega} = [\omega_1, \ldots, \omega_n] \in [0, 1]^n$,

$$V_t(\bar{\omega}; a = j) - V_t(\bar{\omega}; a = i) \geq 0 \quad (11)$$

if $\omega_j \geq \omega_i$, for $i = 1, \ldots, n$.

The proof of this theorem is based on backward induction on t: given the optimality of the myopic policy at times $t + 1, t + 2, \ldots, T$, we want to show that it is also optimal at time t. This relies on a number of lemmas introduced below. The first lemma introduces a notation that allows us to express the expected future reward under the myopic policy.

Lemma 1: There exist T n-variable functions, denoted by $W_t()$, $t = 1, 2, \ldots, T$, each of which is a polynomial of order 1^4 and can be represented recursively in the following form:

$$W_t(\bar{\omega}) = \omega_n + \omega_n \beta W_{t+1}(\tau(\omega_1), \ldots, \tau(\omega_{n-1}), p_{11})$$
$$+ (1 - \omega_n)\beta W_{t+1}(p_{01}, \tau(\omega_1), \ldots, \tau(\omega_{n-1})) \quad (12)$$

where $\bar{\omega} = [\omega_1, \omega_2, \ldots, \omega_n]$ and $W_T(\bar{\omega}) = \omega_n$.

Proof: The proof is easily obtained using backward induction on t given the above recursive equation and noting that $W_T()$ is one such polynomial and the mapping $\tau()$ is a linear operation. $\square$

Corollary 1: When $\bar{\omega}$ represents the ordered list of information states $[\omega_1, \omega_2, \ldots, \omega_n]$ with $\omega_1 \leq \omega_2 \leq \cdots \leq \omega_n$, then $W_t(\bar{\omega})$ is the expected total reward obtained by the myopic policy from time t on.

Proof: This result follows directly from the description of the policy given in Section III-B.

When $\bar{\omega}$ is the ordered list of information states, the recursive expression in (12) gives the expected reward of the following policy: probe the nth channel; for the next step, if the current sensing outcome is "1," then continue to probe the nth channel; if the current sensing outcome is "0," then drop this channel to the bottom of the list (it becomes the first channel) while moving the ith channel to the $(i+1)$th position for all $i = 1, \ldots, n-1$; repeat this process. (This is essentially the same description as given in Section III for the case of $p_{11} \geq p_{01}$.) To see that this is the myopic policy, note that under the above policy, at any time the list of channel probabilities are increasingly ordered. This is because for any $0 \leq \omega \leq 1$, we have $p_{01} \leq \tau(\omega) \leq p_{11}$ when $p_{11} \geq p_{01}$. Furthermore, under the assumption $p_{11} \geq p_{01}$, $\tau(\omega)$ is a monotonically increasing function. Therefore, under this policy, when starting out with increasingly ordered information states, this order is maintained in each subsequent time step. As expressed in (12), at each step it is always the nth channel that is probed. Since the nth channel has the largest probability of being available, this is the myopic policy. $\square$

Proposition 1: The fact that W_t is a polynomial of order 1 and affine in each of its elements implies that

$$W_t(\omega_1, \ldots, \omega_{n-2}, y, x) - W_t(\omega_1, \ldots, \omega_{n-2}, x, y)$$
$$= (x - y)[W_t(\omega_1, \ldots, \omega_{n-2}, 0, 1) - W_t(\omega_1, \ldots, \omega_{n-2}, 1, 0)]. \quad (13)$$

Similar results hold when we change the positions of x and y.

To see this, consider $W_t(\omega_1, \ldots, \omega_{n-2}, x, y)$ and $W_t(\omega_1, \ldots, \omega_{n-2}, y, x)$ as functions of x and y, each having an x term, a y term, an xy term, and a constant term. Since

^{4}Each function W_t is affine in each variable, when all other variables are held constant.

we are just swapping the positions of x and y in these two functions, the constant term remains the same, and so does the xy term. Thus, the only difference is the x term and the y term, as given in the preceding equation. This linearity result will be used later in our proofs.

The next lemma establishes a necessary and sufficient condition for the optimality of the myopic policy.

Lemma 2: Consider Problem (P1) and Assumption 1. Given the optimality of the myopic policy at times $t+1, t+2, \ldots, T$, the optimality at time t is equivalent to

$$W_t(\omega_1, \ldots, \omega_{i-1}, \omega_{i+1}, \ldots, \omega_n, \omega_i) \leq W_t(\omega_1, \ldots, \omega_n),$$
$$\text{for all } \omega_1 \leq \ldots \leq \omega_i \leq \cdots \leq \omega_n.$$

Proof: Since the myopic policy is optimal from $t+1$ on, it is sufficient to show that probing ω_n followed by myopic probing is better than probing any other channel followed by myopic probing. The former is precisely given by the right-hand side (RHS) of the above equation; the latter by the left-hand side (LHS), thus completing the proof. □

Having established that $W_t(\bar{\omega})$ is the total expected reward of the myopic policy for an increasingly ordered vector $\bar{\omega} = [\omega_1, \ldots, \omega_n]$, we next proceed to show that we do not decrease this total expected reward $W_t(\bar{\omega})$ by switching the order of two neighboring elements ω_i and ω_{i+1} if $\omega_i \geq \omega_{i+1}$. This is done in two separate cases, when $i+1 < n$ (given in Lemma 4) and when $i+1 = n$ (given in Lemma 5), respectively. The first case is quite straightforward, while proving the second cased turned out to be significantly more difficult. Our proof of the second case (Lemma 5) relies on a separate lemma (Lemma 3) that establishes a bound between the greedy operation on two identical vectors but with a different starting position. The proof of Lemma 3 is based on a coupling argument and is quite instructive. Below we present and prove Lemmas 3–5.

Lemma 3: For $0 < \omega_1 \leq \omega_2 \leq \cdots \leq \omega_n < 1$, we have the following inequality for all $t = 1, 2, \ldots, T$:

$$1 + W_t(\omega_2, \ldots, \omega_n, \omega_1) \geq W_t(\omega_1, \ldots, \omega_n). \quad (14)$$

Proof: We prove this lemma using a coupling argument along any sample path. The LHS of the above inequality represents the expected reward of a policy (referred to as L below) that probes in the sequence of channels 1 followed by $n, n-1, \ldots$, and then 1 again, and so on, plus an extra reward of 1; the RHS represents the expected reward of a policy (referred to as R below) that probes in the sequence of channels n followed by $n-1, \cdots$, and 1 and then n again, and so on. It helps to imagine lining up the n channels along a circle in the sequence of $n, n-1, \ldots, 1$, clock-wise, and thus L's starting position is 1, R's starting position is n, exactly one spot ahead of L clockwise. Each will cycle around the circle till time T.

Now for any realization of the channel conditions (or any sample path of the system), consider the sequence of "0's" and "1's" that these two policies see, and consider the position they

are on the circle. The reward a policy gets along a given sample path is $R_l = \sum_{j=t}^{T} \beta_j I_{h^l}(j) = 1$ for policy L, where $I_{h^l(j=1)} = 1$ if L sees a "1" at time j, and 0 otherwise; the reward for R is $R_r = \sum_{j=t}^{T} \beta^j I_{h^r}(j) = 1$ with $I_{h^r}(j) = 1$ similarly defined. There are two cases.

Case (1): The two eventually catch up with each other at some time $K \leq T$, i.e., at some point they start probing exactly the same channel. From this point on, the two policies behave exactly the same way along the same sample path, and the reward they obtain from this point on is exactly the same. Therefore, in this case we only need to compare the rewards (L has an extra 1) leading up to this point.

Case (2): The two never manage to meet within the horizon T. In this case, we need to compare the rewards for the entire horizon (from t to T).

We will consider Case (1) first. There are only two possibilities for the two policies to meet: (Case 1a) either L has seen exactly one more "0" than R in its sequence, or (Case 1b) R has seen exactly $n-1$ more "0's" than L. This is because the moment we see a "0" we will move to the next channel on the circle. L is only one position behind R, so one more "0" will put it at exactly the same position as R. The same with R moving $n-1$ more positions ahead to catch up with L.

Case (1a): L sees exactly one more "0" than R in its sequence. The extra "0" necessarily occurs at exactly time K, $t \leq K \leq T$, meaning that at K, L sees a "0" and R sees a "1." From t to K, if we write the sequence of rewards (zeros and ones) under L and R, we observe the following: between t and K both L and R have equal number of zeros, while for $\forall t' = t, t+1, \ldots, K-1$, the number of zeros up to time t' is less (or no more) for L than for R. In other words, L and R see the same number of "0's," but L's is always lagging behind (or no earlier). That is, for every "0" R sees, L has a matching "0" that occurs no earlier than R's "0." This means that if we denote by $R_l(t_1, t_2)$ the rewards accumulated between t_1 and t_2, then for the rewards in $[t, K-1]$, we have $R_l(t, t') \geq R_r(t, t')$, for $\forall t' \leq K-1$, while $R_l(K, K) = \beta^K$ and $R_r(K, K) = 0$. Finally by definition we have $R_l(K+1, T) = R_r(K+1, T)$. Therefore, overall we have $1 + R_l(t, T) \geq R_r(t, T)$, proving the above inequality.

Case (1b): R sees $n-1$ more "0's" than L does. The comparison is simpler. We only need to note that R's "0's" must again precedes (or be no later than) L's since otherwise we will return to Case (1a). Therefore, we have $R_l \geq R_r$, and thus $1 + R_l \geq R_r$ is also true.

We now consider Case (2). The argument is essentially the same. In this case, the two do not get to meet, but they are on their way, meaning that either L has exactly the same "0's" as R and their positions are no earlier (corresponding to Case (1a)), or R has more "0's" than L (but not up to $n-1$) and their positions are no later than L's (corresponding to Case (1b)). So either way we have $1 + R_l \geq R_r$.

The proof is thus complete. □

Lemma 4: For all j, $1 \leq j \leq n-3$, and all $x \geq y$, we have

$$W_t(\omega_1, \ldots, \omega_j, x, y, \ldots, \omega_n) \leq W_t(\omega_1, \ldots, \omega_j, y, x, \ldots, \omega_n). \quad (15)$$

Proof: We prove this by induction over t. The claim is obviously true for $t = T$, since both sides will be equal to ω_n, thereby establishing the induction basis. Now suppose the claim is true for all $t+1, \ldots, T-1$. We have

$$
\begin{aligned}
W_t&(\omega_1, \ldots, \omega_{j-1}, x, y, \ldots, \omega_n) \\
&= \omega_n (1 + \beta W_{t+1}(\tau(\omega_1), \ldots, \tau(x), \\
&\qquad\qquad \tau(y), \ldots, \tau(\omega_{n-1}), p_{11})) \\
&\quad + (1 - \omega_n)\beta W_{t+1}(p_{01}, \tau(\omega_1), \ldots, \tau(x), \\
&\qquad\qquad \tau(y), \ldots, \tau(\omega_{n-1})) \\
&\leq \omega_n (1 + \beta W_{t+1}(\tau(\omega_1), \ldots, \tau(y), \\
&\qquad\qquad \tau(x), \ldots, \tau(\omega_{n-1}), p_{11})) \\
&\quad + (1 - \omega_n)\beta W_{t+1}(p_{01}, \tau(\omega_1), \ldots, \tau(y), \\
&\qquad\qquad \tau(x), \ldots, \tau(\omega_{n-1})) \\
&= W_t(\omega_1, \ldots, \omega_{j-1}, y, x, \ldots, \omega_n) \quad (16)
\end{aligned}
$$

where the inequality is due to the induction hypothesis, and noting that $\tau(\ )$ is a monotone increasing mapping in the case of $p_{11} \geq p_{01}$. $\square$

Lemma 5: For all $x \geq y$, we have

$$
\begin{aligned}
W_t(\omega_1, \ldots, \omega_j, \ldots, \omega_{n-2}, x, y) \\
\leq W_t(\omega_1, \ldots, \omega_j, \ldots, \omega_{n-2}, y, x). \quad (17)
\end{aligned}
$$

Proof: This lemma is proved inductively. The claim is obviously true for $t = T$. Assume it also holds for times $t+1, \ldots, T-1$. We have by the definition of $W_t()$ and due to its linearity property

$$
\begin{aligned}
W_t&(\omega_1, \ldots, \omega_{n-2}, y, x) - W_t(\omega_1, \ldots, \omega_{n-2}, x, y) \\
&= (x - y)(W_t(\omega_1, \ldots, \omega_{n-2}, 0, 1) \\
&\qquad\quad - W_t(\omega_1, \ldots, \omega_{n-2}, 1, 0)) \\
&= (x - y)(1 + \beta W_{t+1}(\tau(\omega_1), \ldots, \tau(\omega_{n-2}), p_{01}, p_{11}) \\
&\qquad\quad - \beta W_{t+1}(p_{01}, \tau(\omega_1), \ldots, \tau(\omega_{n-2}), p_{11})).
\end{aligned}
$$

But from the induction hypothesis we know that

$$
\begin{aligned}
W_{t+1}&(\tau(\omega_1), \ldots, \tau(\omega_{n-2}), p_{01}, p_{11}) \\
&\geq W_{t+1}(\tau(\omega_1), \ldots, \tau(\omega_{n-2}), p_{11}, p_{01}). \quad (18)
\end{aligned}
$$

This means that

$$
\begin{aligned}
1 + \beta W_{t+1}&(\tau(\omega_1), \ldots, \tau(\omega_{n-2}), p_{01}, p_{11}) \\
&- \beta W_{t+1}(p_{01}, \tau(\omega_1), \ldots, \tau(\omega_{n-2}), p_{11}) \\
\geq 1 + \beta W_{t+1}&(\tau(\omega_1), \ldots, \tau(\omega_{n-2}), p_{11}, p_{01}) \\
&- \beta W_{t+1}(p_{01}, \tau(\omega_1), \ldots, \tau(\omega_{n-2}), p_{11}) \geq 0
\end{aligned}
$$

where the last inequality is due to Lemma 3 (note that in that lemma we proved $1 + A \geq B$, which obviously implies $1 + \beta A \geq \beta B$ for $0 \leq \beta \leq 1$ that is used above). This, together with the condition $x \geq y$, completes the proof. $\square$

We are now ready to prove the main theorem.

Proof of Theorem 1: The basic approach is by induction on t. The optimality of the myopic policy at time $t = T$ is obvious. So the induction basis is established. Now assume that the myopic policy is optimal for all times $t+1, t+2, \ldots, T$, and we will show that it is also optimal at time t. By Lemma 2 this is equivalent to establishing the following:

$$W_t(\omega_1, \ldots, \omega_{i-1}, \omega_{i+1}, \ldots, \omega_n, \omega_i) \leq W_t(\omega_1, \ldots, \omega_n). \quad (19)$$

But we know from Lemmas 4 and 5 that

$$
\begin{aligned}
W_t&(\omega_1, \ldots, \omega_{i-1}, \omega_{i+1}, \ldots, \omega_n, \omega_i) \\
&\leq W_t(\omega_1, \ldots, \omega_{i-1}, \omega_{i+1}, \ldots, \omega_i, \omega_n) \\
&\leq W_t(\omega_1, \ldots, \omega_{i-1}, \omega_{i+1}, \ldots, \omega_i, \omega_{n-1}, \omega_n) \leq \cdots \\
&\leq W_t(\omega_1, \ldots, \omega_n)
\end{aligned}
$$

where the first inequality is the result of Lemma 5, while the remaining inequalities are repeated applications of Lemma 4, completing the proof. $\square$

We would like to emphasize that from a technical point of view, Lemma 3 is the key to the whole proof: it leads to Lemma 5, which in turn leads to Theorem 1. While Lemma 5 was easy to conceptualize as a sufficient condition to prove the main theorem, Lemma 3 was much more elusive to construct and prove. This, indeed, marks the main difference between the proof techniques used here versus that used in our earlier work [6]: Lemma 3 relies on a coupling argument instead of the convex analytic properties of the value function.

V. THE CASE OF $p_{11} < p_{01}$

In the preceding section we showed that a myopic policy is optimal if $p_{11} \geq p_{01}$. In this section we examine what happens when $p_{11} < p_{01}$, which corresponds to the case when the Markovian channel state process exhibits a negative autocorrelation over a unit of time. This is perhaps a case of less practical interest and relevance. However, as we shall see, this case presents a greater degree of technical complexity and richness than the previous case. Specifically, we first show that when the number of channels is three ($n = 3$) or when the discount factor $\beta \leq \frac{1}{2}$, the myopic policy remains optimal even for the case of $p_{11} < p_{01}$ (the proof for two channels in this case was given earlier in [5]). We thus conclude that the myopic policy is optimal for $n \leq 3$ or $\beta \leq 1/2$ regardless of the transition probabilities. We then present a counterexample showing that the myopic policy is not optimal in general when $n \geq 4$ and $\beta > 1/2$. In particular, our counterexample is for a finite horizon with $n = 4$ and $\beta = 1$.

A. $n = 3$ or $\beta \leq \frac{1}{2}$

We start by developing some results parallel to those presented in the previous section for the case of $p_{11} \geq p_{01}$.

Lemma 6: There exist T n-variable polynomial functions of order 1, denoted by $Z_t()$, $t = 1, 2, \ldots, T$, i.e., each function is

linear in all the elements, and can be represented recursively in the following form:

$$Z_t(\bar{\omega}) := \omega_n(1 + \beta Z_{t+1}(p_{11}, \tau(\omega_{n-1}), \ldots, \tau(\omega_1)))$$
$$+ (1 - \omega_n)\beta Z_{t+1}(\tau(\omega_{n-1}), \ldots, \tau(\omega_1), p_{01}) \quad (20)$$

where $Z_T(\bar{\omega}) = \omega_n$.

Corollary 2: $Z_t(\bar{\omega})$ given in (20) represents the expected total reward of the myopic policy when $\bar{\omega}$ is ordered in increasing order of ω_i.

Similar to Corollary 1, the above result follows directly from the policy description given in Section III-B.

It follows that the function Z_t also has the same linearity property presented earlier, i.e.,

$$Z_t(\omega_1, \ldots, \omega_{n-2}, y, x) - Z_t(\omega_1, \ldots, \omega_{n-2}, x, y)$$
$$= (x - y)(Z_t(\omega_1, \ldots, \omega_{n-2}, 0, 1)$$
$$- Z_t(\omega_1, \ldots, \omega_{n-2}, 1, 0)). \quad (21)$$

Similar results hold when we change the positions of x and y.

In the next lemma and theorem, we prove that the myopic policy is still optimal when $p_{11} < p_{01}$ if $n = 3$ or $\beta \leq 1/2$. In particular, Lemma 7 below is the analogy of Lemmas 4 and 5 combined.

Lemma 7: At time t ($t = 1, 2, \ldots, T$), for all $j \leq n - 2$, we have the following inequality for $\forall\, 1 \geq x \geq y \geq 0$ if either $n = 3$ or $\beta \leq 1/2$:

$$Z_t(\omega_1, \ldots, \omega_j, y, x, \omega_{j+3}, \ldots, \omega_n)$$
$$\geq Z_t(\omega_1, \ldots, \omega_j, x, y, \omega_{j+3}, \ldots, \omega_n). \quad (22)$$

Proof: We prove this by induction on t. The claim is obviously true for $t = T$. Now suppose it is true for $t + 1, \ldots, T$. Due to the linearity property of Z_t

$$Z_t(\omega_1, \ldots, \omega_j, y, x, \omega_{j+3}, \ldots, \omega_n)$$
$$- Z_t(\omega_1, \ldots, \omega_j, x, y, \omega_{j+3}, \ldots, \omega_n)$$
$$= (x - y)\big(Z_t(\omega_1, \ldots, \omega_j, 0, 1, \omega_{j+3}, \ldots, \omega_n)$$
$$- Z_t(\omega_1, \ldots, \omega_j, 1, 0, \omega_{j+3}, \ldots, \omega_n)\big). \quad (23)$$

Thus it suffices to show that

$$Z_t(\omega_1, \ldots, \omega_j, 0, 1, \omega_{j+3}, \ldots, \omega_n)$$
$$\geq Z_t(\omega_1, \ldots, \omega_j, 1, 0, \omega_{j+3}, \ldots, \omega_n).$$

We treat the case when $j < n - 2$ and $j = n - 2$ separately. Indeed, without loss of generality, let $j = n - 3$ (the proof follows exactly for all $j \leq n - 3$ with more lengthy notations). At time t we have

$$Z_t(\omega_1, \ldots, \omega_{n-3}, 0, 1, \omega_n) - Z_t(\omega_1, \ldots, \omega_{n-3}, 1, 0, \omega_n)$$
$$= \omega\beta(Z_{t+1}(p_{11}, p_{11}, p_{01}, \tau(\omega_{n-3}), \ldots, \tau(\omega_1))$$
$$- Z_{t+1}(p_{11}, p_{01}, p_{11}, \tau(\omega_{n-3}), \ldots, \tau(\omega_1)))$$
$$+ (1 - \omega)\beta(Z_{t+1}(p_{11}, p_{01}, \tau(\omega_{n-3}), \ldots, \tau(\omega_1), p_{01})$$
$$- Z_{t+1}(p_{01}, p_{11}, \tau(\omega_{n-3}), \ldots, \tau(\omega_1), p_{01}))$$
$$\geq 0$$

where the last inequality is due to the induction hypothesis.

Now we will consider the case when $j = n - 2$.

$$Z_t(\omega_1, \ldots, \omega_{n-2}, 0, 1) - Z_t(\omega_1, \ldots, \omega_{n-2}, 1, 0)$$
$$= 1 + \beta Z_{t+1}(p_{11}, p_{01}, \tau(\omega_{n-2}), \ldots, \tau(\omega_1))$$
$$- \beta Z_{t+1}(p_{11}, \tau(\omega_{n-2}), \ldots, \tau(\omega_1), p_{01}). \quad (24)$$

Next we show that if $\beta \leq 1/2$ or $n = 3$, the RHS of (24) is nonnegative.

If $\beta \leq 1/2$, then

$$1 + \beta Z_{t+1}(p_{11}, p_{01}, \tau(\omega_{n-2}), \ldots, \tau(\omega_1))$$
$$- \beta Z_{t+1}(p_{11}, \tau(\omega_{n-2}), \ldots, \tau(\omega_1), p_{01})$$
$$\geq 1 - \frac{\beta}{1 - \beta} \geq 0.$$

If $n = 3$, then

$$1 + \beta Z_{t+1}(p_{11}, p_{01}, \tau(\omega_1)) - \beta Z_{t+1}(p_{11}, \tau(\omega_1), p_{01})$$
$$= 1 + \beta(\tau(\omega_1) - p_{01})(Z_{t+1}(p_{11}, 0, 1) - Z_{t+1}(p_{11}, 1, 0))$$
$$\geq 1 - \beta(Z_{t+1}(p_{11}, 0, 1) - Z_{t+1}(p_{11}, 1, 0))$$
$$\geq 0$$

where the first inequality is due to the fact that $-1 \leq \tau(\omega_1) - p_{01} \leq 0$ and the last inequality is given by the induction hypothesis. $\quad\square$

Theorem 2: Consider Problem (P1). Assume that $p_{11} < p_{01}$. The myopic policy is optimal for the case of $n = 3$ and the case of $\beta \leq 1/2$ with arbitrary n. More precisely, for these two cases, $\forall\, t$, $1 \leq t \leq T$, we have

$$V_t(\bar{\omega}; a = j) - V_t(\bar{\omega}; a = i) \geq 0 \quad (25)$$

if $\omega_j \geq \omega_i$ for $i = 1, \ldots, n$.

Proof: We prove by induction on t. The optimality of the myopic policy at time $t = T$ is obvious. Now assume that the myopic policy is optimal for all times $t + 1, t + 2, \ldots, T - 1$, and we want to show that it is also optimal at time t. Suppose at time t the channel probabilities are such that $\omega_n \geq \omega_i$ for $i = 1, \ldots, n - 1$. The myopic policy is optimal at time t if and only if probing ω_n followed by myopic probing is better than probing any other channel followed by myopic probing. Mathematically, this means

$$Z_t(\omega_1, \ldots, \omega_{i-1}, \omega_{i+1}, \ldots, \omega_n, \omega_i) \leq Z_t(\omega_1, \ldots, \omega_n),$$
$$\text{for all } \omega_1 \leq \omega_i \leq \omega_n.$$

But this is a direct consequence of Lemma 7, completing the proof. $\quad\square$

B. A Four-Channel Counter Example

The following example shows that the myopic policy is not, in general, optimal for $n \geq 4$ when $p_{11} < p_{01}$.

Example 1: Consider an example with the following parameters: $p_{01} = 0.9$, $p_{11} = 0.1$, $\beta = 1$, and $\bar{\omega} = [0.97, 0.97, 0.98, 0.99]$. Now compare the following two policies at time $T - 3$: play myopically (I), or play the 0.98 channel

first, followed by the myopic policy (II). Computation reveals that

$$V_{T-3}^{\mathrm{I}}(0.97, 0.97, 0.98, 0.99) = 2.401863$$
$$< V_{T-3}^{\mathrm{II}}(0.97, 0.97, 0.98, 0.99) = 2.402968$$

which shows that the myopic policy is not optimal in this case.

It remains an interesting question as to whether such counterexamples exist in the case when the initial condition is such that all channel are in the good state with the stationary probability.

VI. INFINITE HORIZON

Now we consider extensions of results in Sections IV and V to (P2) and (P3), i.e., to show that the myopic policy is also optimal for (P2) and (P3) under the same conditions. 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. Theorems 3 and 4 below concretely establish this.

We point out that the proofs of Theorems 3 and 4 do not rely on any additional assumptions other than the optimality of the myopic policy for (P1). Indeed, if the optimality of the myopic policy for (P1) can be established under weaker conditions, Theorems 3 and 4 can be readily invoked to establish its optimality under the same weaker condition for (P2) and (P3), respectively.

Theorem 3: If myopic policy is optimal for (P1), it is also optimal for (P2) for $0 \leq \beta < 1$. Furthermore, its value function is the limiting value function 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 (BCT) 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 that

$$
\begin{aligned}
J_\beta^\pi(\bar{\omega}) &= E^\pi \left[\lim_{T \to \infty} \sum_{t=1}^T \beta^{t-1} R_{\pi(t)}(\bar{\omega}(t)) | \bar{\omega}(1) = \bar{\omega} \right] \\
&= \lim_{T \to \infty} E^\pi \left[\sum_{t=1}^T \beta^{t-1} R_{\pi(t)}(\bar{\omega}(t)) | \bar{\omega}(1) = \bar{\omega} \right] \\
&= \lim_{T \to \infty} J_T^\pi(\bar{\omega})
\end{aligned}
\tag{26}
$$

where the second equality is due to BCT for

$$\sum_{t=1}^T \beta^{t-1} R_{\pi(t)}(\bar{\omega}(t)) \leq \frac{1}{1-\beta}.$$

This proves the second part of the theorem by noting that due to the finiteness of the action space, we can interchange maximization and limit. Let π^* denote the myopic policy. We now establish the optimality of π^* for (P2). From Theorem 1, we know

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

Taking limit of both sides, we have

$$
\begin{aligned}
J_\beta^{\pi^*}(\bar{\omega}) = \max_{a=i} \Big\{ &\omega_i + \beta \omega_i J_\beta^{\pi^*}\left(\mathcal{T}(\bar{\omega}, i | 1)\right) \\
&+ \beta(1-\omega_i) J_\beta^{\pi^*}\left(\mathcal{T}(\bar{\omega}, i | 0)\right) \Big\}.
\end{aligned}
\tag{27}
$$

Note that (27) is nothing but the dynamic programming equation for the infinite horizon discounted reward problem given in (7). 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. $\square$

Theorem 4: Consider (P3) with the expected average reward and under the ergodicity assumption $|p_{11} - p_{00}| < 1$. Myopic policy is optimal for problem (P3) if it is optimal for (P1).

Proof: We consider the infinite horizon discounted cost for $\beta < 1$ under the optimal policy denoted by π^*

$$
\begin{aligned}
J_\beta^{\pi^*}(\bar{\omega}) = \max_{a=i} \Big\{ &\omega_i + \beta \omega_i J_\beta^{\pi^*}\left(\mathcal{T}(\bar{\omega}, i | 1)\right) \\
&+ \beta(1-\omega_i) J_\beta^{\pi^*}\left(\mathcal{T}(\bar{\omega}, i | 0)\right) \Big\}.
\end{aligned}
\tag{28}
$$

This can be written as

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

Notice that the boundedness of the reward function and compactness of information state implies that the sequence of $\left\{ (1-\beta) J_\beta^{\pi^*}(\bar{\omega}) \right\}$ is bounded, i.e., for all $0 \leq \beta \leq 1$

$$(1-\beta) J_\beta^{\pi^*}(\bar{\omega}) \leq 1. \tag{29}$$

Also, applying Lemma 2 from [6] (which provides an upper bound on the difference in value functions between taking two different actions followed by the optimal policy) and noting that $-1 < p_{11} - p_{00} < 1$, we have that there exists some positive constant $K := \frac{1}{1-|p_{11}-p_{01}|}$ such that

$$\left| J_\beta^{\pi^*}\left(\mathcal{T}(\bar{\omega}, i | 0)\right) - J_\beta^{\pi^*}(\bar{\omega}) \right| \leq K. \tag{30}$$

By Bolzano–Weierstrass theorem, (29) and (30) guarantee the existence of a converging sequence $\beta_k \to 1$ such that

$$\lim_{k \to \infty} (1-\beta_k) J_{\beta_k}^{\pi^*}(\bar{\omega}^*) := J^* \tag{31}$$

$$\text{and} \quad \lim_{k \to \infty} \left[J_{\beta_k}^{\pi^*}(\bar{\omega}) - J_{\beta_k}^{\pi^*}(\bar{\omega}^*) \right] := h^{\pi^*}(\bar{\omega}) \tag{32}$$

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

As a result, (31) can be written as

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

In other words

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

From (32), we can write this as

$$J^* + h^{\pi^*}(\bar{\omega}) = \max_{a=i}\left\{\omega_i + \omega_i h^{\pi^*}\left(\mathcal{T}\left(\bar{\omega}, i|1\right)\right) \right.$$
$$\left. + (1 - \omega_i)h^{\pi^*}\left(\mathcal{T}\left(\bar{\omega}, i|0\right)\right)\right\}. \quad (33)$$

Note that (33) is nothing but the DP equation as given by (8). In addition, we know that the immediate reward as well as function h are both bounded by $\max(1, K)$. This implies that J^* is the maximum average reward, i.e., $J^* = \max_\pi J_\infty^\pi(\bar{\omega}(t))$ (see [17, Theorems 6.1–6.3]).

On the other hand, we know from Theorem 3 that the myopic policy is optimal for (P2) if it is for (P1), and thus we can take π^* in (28) to be the myopic policy. Rewriting (28) gives the following:

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

Repeating steps (31)–(33) we arrive at the following:

$$J + h^{\pi^*}(\bar{\omega}) = \omega_{\pi^*(\bar{\omega})} + \omega_{\pi^*(\bar{\omega})}h^{\pi^*}\left(\mathcal{T}\left(\bar{\omega}, \pi^*(\bar{\omega})|1\right)\right)$$
$$+ (1 - \omega_{\pi^*(\bar{\omega})})h^{\pi^*}\left(\mathcal{T}\left(\bar{\omega}, \pi^*(\bar{\omega})|0\right)\right) \quad (34)$$

which shows that $\left(J^*, h^{\pi^*}, \pi^*\right)$ is a canonical triplet [17, Theorems 6.2]. This, together with boundedness of h^{π^*} and immediate reward, implies that the myopic policy π^* is optimal for (P3) [17, Theorems 6.3]. □

VII. DISCUSSION AND RELATED WORK

The problem studied in this paper may be viewed as a special case of a class of MDPs known as the restless bandit problems [2]. In this class of problems, N controlled Markov chains (also called *projects* or *machines*) are activated (or played) one at a time. A machine when activated generates a state-dependent reward and transits to the next state according to a Markov rule. A machine not activated transits to the next state according to a (potentially different) Markov rule. The problem is to decide the sequence in which these machines are activated so as to maximize the expected (discounted or average) reward over an infinite horizon. To put our problem in this context, each channel corresponds to a machine, and a channel is activated when it is probed, and its information state goes through a transition depending on the observation and the underlying channel model. When a channel is not probed, its information state goes through a transition solely based on the underlying channel model.[5]

In the case that a machine stays frozen in its current state when not played, the problem reduces to the multiarmed bandit problem, solved by Gittins in his 1970 seminal work [18]. Gittins showed that there exists an *index* associated with each machine that is solely a function of that individual machine and its state, and that playing the machine currently with the highest index is optimal. This index has since been referred to as the *Gittins index* due to Whittle [19]. The remarkable nature of this result lies in the fact that it essentially decomposes the N-dimensional problem into N 1-dimensional problems, as an index

is defined for a machine independent of others. The basic model of multiarmed bandit has been used previously in the context of channel access and cognitive radio networks. For example, in [20], Bayesian learning was used to estimate the probability of a channel being available, and the Gittins indices, calculated based on such estimates (which were only updated when a channel was observed and used, thus giving rise to a multiarmed bandit formulation rather than a restless bandit formulation), were used for channel selection.

On the other hand, relatively little is known about the structure of the optimal policies for the restless bandit problem in general. It has been shown that the Gittins index policy is not in general optimal in this case [2], and that this class of problems is PSPACE-hard in general [21]. Whittle, in [2], proposed a Gittins-like index (referred to as the Whittle's index policy), shown to be optimal under a constraint on the *average* number of machines that can be played at a given time, and asymptotically optimal under certain limiting regimes [22]. There has been a large volume of literature in this area, including various approximation algorithms, see for example [23] and [24] for near-optimal heuristics, as well as conditions for certain policies to be optimal for special cases of the restless bandit problem, see, e.g., [25], [26]. The nature of the results derived in the present paper is similar to that of [25], [26] in spirit. That is, we have shown that for this special case of the restless bandit problem an index policy is optimal under certain conditions. For the indexability (as defined by Whittle [2]) of this problem, see [27].

Recently, Guha and Munagala [28], [29] studied a class of problems referred to as the feedback multiarmed bandit problems. This class is very similar to the restless bandit problem studied in the present paper, with the difference that channels may have different transition probabilities (thus this is a slight generalization to the one studied here). While we identified conditions under which a simple greedy index policy is optimal in the present paper, Guha and Munagala in [28], [29] looked for provably good approximation algorithms. In particular, they derived a $2 + \epsilon$-approximate policy using a duality-based technique.

VIII. CONCLUSION

The general problem of opportunistic sensing and access arises in many multichannel communication contexts. For cases where the stochastic evolution of channels can be modelled as i.i.d. two-state Markov chains, we showed that a simple and robust myopic policy is optimal for the finite and infinite horizon discounted reward criteria as well as the infinite horizon average reward criterion, when the state transitions are positively correlated over time. When the state transitions are negatively correlated, we showed that the same policy is optimal when the number of channels is limited to two or three, and presented a counterexample for the case of four channels.

---

[5]The standard definition of bandit problems typically assumes finite or countably infinite state spaces. While our problem can potentially have an uncountable state space, it is nevertheless countable for a given initial state. This view has been taken throughout the paper.

REFERENCES

[1] R. Smallwood and E. Sondik, "The optimal control of partially observable Markov processes over a finite horizon," *Oper. Res.*, pp. 1071–1088, 1971.

[2] P. Whittle, "Restless bandits: Activity allocation in a changing world," *A Celebration of Applied Probability, Ed. J. Gani, J. Appl. Probab.*, vol. 25A, pp. 287–298, 1988.

[3] Q. Zhao and B. Sadler, "A survey of dynamic spectrum access," *IEEE Signal Process. Mag.*, vol. 24, pp. 79–89, May 2007.

[4] Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: A POMDP framework," *IEEE J. Sel. Areas Commun. (Special Issue on Adaptive, Spectrum Agile and Cognitive Wireless Networks)*, vol. 25, no. 3, pp. 589–600, Apr. 2007.

[5] Q. Zhao, B. Krishnamachari, and K. Liu, "On myopic sensing for multi-channel opportunistic access: Structure, optimality, and performance," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 5431–5440, Dec. 2008.

[6] T. Javidi, B. Krishnamachari, Q. Zhao, and M. Liu, "Optimality of myopic sensing in multi-channel opportunistic access," in *Proc. IEEE Int. Conf. Communications (ICC)*, Beijing, China, May 2008, pp. 2107–2112.

[7] A. Ganti, E. Modiano, and J. N. Tsitsiklis, "Optimal transmission scheduling in symmetric communication models with intermittent connectivity," *IEEE Trans. Inf. Theory*, vol. 53, no. 3, pp. 998–1008, Mar. 2007.

[8] A. Ganti, "Transmission Scheduling for Multi-Beam Satellite Systems," Ph.D. dissertation, Dept. Elec. Eng. Comp. Scie., MIT, Cambridge, MA, 2003.

[9] G. Koole, "Monotonicity in Markov reward and decision chains: Theory and applications," in *Foundations and Treends in Stochastic Systems*.   Hanover, MA: Now Publishers, 2006, vol. 1, pp. 1–76.

[10] K. Liu and Q. Zhao, "Channel probing for opportunistic access with multi-channel sensing," in *Proc. Asilomar Conf. Signals, Systems, and Computers*, Pacific Grove, CA, Oct. 2008, pp. 93–97.

[11] N. Chang and M. Liu, "Optimal channel probing and transmission scheduling for opportunistic spectrum access," in *Proc. Int. Conf. Mobile Computing and Networking (MOBICOM)*, Montreal, QC, Canada, Sep. 2007, pp. 27–38.

[12] A. T. Hoang, Y.-C. Liang, D. T. C. Wong, R. Zhang, and Y. Zeng, "Opportunistic spectrum access for energy-constrained cognitive radios," *IEEE Trans. Wireless Commun.*, vol. 8, no. 3, pp. 1206–1211, Mar. 2009.

[13] Q. Zhao, B. Krishnamachari, and K. Liu, "Low-complexity approaches to spectrum opportunity tracking," in *Proc. 2nd Int. Conf. Cognitive Radio Oriented Wireless Networks and Communications (CrownCom)*, Orlando, FL, Aug. 2007, pp. 27–35.

[14] E. Fernandez-Gaucherand, A. Arapostathis, and S. I. Marcus, "On the average cost optimality equation and the structure of optimal policies for partially observable Markov decision processes," *Ann. Oper. Res.*, vol. 29, no. 1–4, pp. 439–470, Dec. 1991.

[15] P. Kumar and P. Karaiya, *Stochastic Systems: Estimation, Identification, and Adaptive Control*.   Englewood Cliffs, NJ: Prentice-Hall, 1986.

[16] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*.   New York: Wiley-Interscience, 1994, Wiley Series in Probability and Mathematical Statistics.

[17] A. Arapostathis, V. Borkar, E. Fernandez-Gaucherand, M. K. Gosh, and S. I. Marcus, "Discrete-time controlled Markov processes with average cost criterion: A survey," *SIAM J. Contr. Optimiz.*, vol. 31, no. 2, pp. 282–344, Mar. 1993.

[18] J. C. Gittins, "Bandit processes and dynamic allocation indices," *J. Roy. Statist. Soc.*, vol. B14, pp. 148–167, 1972.

[19] P. Whittle, "Multi-armed bandits and the Gittins index," *J. Roy. Statist. Soc.*, vol. 42, no. 2, pp. 143–149, 1980.

[20] A. Motamedi and A. Bahai, "MAC protocol design for spectrum-agile wireless networks: Stochastic control approach," in *Proc. IEEE Int. Symp. New Frontiers in Dynamic Spectrum Access Networks (DySPAN)*, Dublin, Ireland, Apr. 2007, pp. 448–451.

[21] C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of optimal queueing network control," *Math. Oper. Res.*, vol. 24, pp. 293–305, May 1999.

[22] R. R. Weber and G. Weiss, "On an index policy for restless bandits," *J. Appl. Probab.*, vol. 27, pp. 637–648, 1990.

[23] D. Bertsimas and J. E. Niño-Mora, "Restless bandits, linear programming relaxations, and a primal-dual heuristic," *Oper. Res.*, vol. 48, no. 1, pp. 80–90, Jan. 2000.

[24] J. E. Niño-Mora, "Restless bandits, partial conservation laws and indexability," *Adv. Appl. Probab.*, vol. 33, pp. 76–98, 2001.

[25] C. Lott and D. Teneketzis, "On the optimality of an index rule in multi-channel allocation for single-hop mobile networks with multiple service classes," *Probab. Eng. Inf. Sci.*, vol. 14, pp. 259–297, Jul. 2000.

[26] N. Ehsan and M. Liu, "Server allocation with delayed state observation: Sufficient conditions for the optimality of an index policy," *IEEE Trans. Wireless Commun.*, to be published.

[27] K. Liu and Q. Zhao, "A restless multiarmed bandit formulation of opportunistic access: Indexability and index policy," in *Proc. 5th IEEE Conf. Sensor, Mesh and Ad Hoc Communications and Networks (SECON)*, San Francisco, CA, Jun. 2008, pp. 1–5.

[28] S. Guha and K. Munagala, "Approximation algorithms for partial-information based stochastic control with markovian rewards," in *Proc. 48th IEEE Symp. Foundations of Computer Science (FOCS)*, Providence, RI, Oct. 2007, pp. 483–493.

[29] S. Guha, K. Munagala, and P. Shi, "Approximation algorithms for restless bandit problems," in *ACM-SIAM Symp. Discrete Algorithms (SODA)*, New York, Jan. 2009, pp. 28–37.

Sahand Haji Ali Ahmad was born in Tehran, Iran, in 1980. He received the B.Sc. and M.Sc. degrees in electrical engineering from Sharif University of Technology, Tehran, and the University of Illinois at Urbana-Champaign, Urbaban IL, respectively, in 2002 and 2005.

He is now pursuing the Ph.D. degree in electrical engineering at the University of Michigan, Ann Arbor. He is the author and coauthor of several technical papers in different journals and conferences. His main interests are applications of stochastic control, game theory, and information theory to different aspects of communication.

Mr. Ahmad was the winner of the Gold medal in Iran's national mathematical olympiad, Winner of the Gold medal in the 39th International Mathematical Olympiad, and also received a Vodafone fellowship for research in wireless communication during academic year 2005–2006.

Mingyan Liu (S'96–M'00) received the B.Sc. degree in electrical engineering in 1995 from the Nanjing University of Aeronotics and Astronautics, Nanjing, China, and the M.Sc. degree in systems engineering, and the Ph.D. degree in electrical engineering from the University of Maryland, College Park, in 1997 and 2000, respectively.

In September 2000, she joined the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, where she is currently an Associate Professor. Her research interests are in performance modeling, analysis, energy efficiency, and resource allocation issues in wireless, mobile *ad hoc*, and sensor networks.

Dr. Liu is the recipient of the 2002 National Science Foundation CAREER Award, and the University of Michigan Elizabeth C. Crosby Research Award in 2003. She serves on the editorial board of the IEEE/ACM TRANSACTIONS ON NETWORKING and the IEEE TRANSACTIONS ON MOBILE COMPUTING.

Tara Javidi (S'96–M'02) studied electrical engineering at Sharif University of Technology, Tehran, Iran from 1992 to 1996. She received the M.S. degrees in electrical engineering (systems), and in applied mathematics (stochastics) from the University of Michigan, Ann Arbor, in 1998 and 1999, respectively. She received the Ph.D. degree in electrical engineering and computer science from the University of Michigan, Ann Arbor, in 2002.

From 2002 to 2004, she was an Assistant Professor at the Electrical Engineering Department, University of Washington, Seattle. She joined University of California, San Diego, in 2005, where she is currently an Assistant Professor of Electrical and Computer Engineering. She was a Barbour Scholar during 1999–2000 academic year and received an NSF CAREER Award in 2004. Her research interests are in communication networks, stochastic resource allocation, stochastic control theory, and wireless communications.

Qing Zhao (S'97–M'02–SM'08) received the Ph.D. degree in electrical engineering in 2001 from Cornell University, Ithaca, NY.

In August 2004, she joined the Department of Electrical and Computer Engineering at University of California, Davis, where she is currently an Associate Professor. Prior to that, she was a Communications System Engineer with Aware, Inc., Bedford, MA and a Postdoctoral Research Associate with Cornell University. Her research interests are in the general area of signal processing, communications, and wireless networking.

Prof. Zhao received the 2000 Young Author Best Paper Award from IEEE Signal Processing Society and the 2008 Outstanding Junior Faculty Award from the University of California, Davis, College of Engineering. She is an Associate Editor of IEEE TRANSACTIONS ON SIGNAL PROCESSING, and an elected member of IEEE Signal Processing Society SP-COM Technical Committee.

Bhaskar Krishnamachari received the B.E. degree in electrical engineering at The Cooper Union, New York, in 1998, and the M.S. and Ph.D. degrees from Cornell University, Ithaca, NY, in 1999 and 2002 respectively.

He is currently an Associate Professor in the Department of Electrical Engineering at the University of Southern California, Los Angeles. His primary research interest is in the design and analysis of algorithms for wireless networks.