On the Optimality of Myopic Sensing in Multi-State Channels

Yi Ouyang, *Student Member, IEEE*, and Demosthenis Teneketzis, *Fellow, IEEE*

***Abstract*— We consider the channel sensing problem arising in opportunistic scheduling over fading channels, cognitive radio networks, and resource constrained jamming. The same problem arises in many other areas of science and technology as it is an instance of restless bandit problems. The communication system consists of N channels. Each channel is modeled as a multi-state Markov chain. At each time instant a user selects one channel to sense and uses it to transmit information. A reward depending on the state of the selected channel is obtained for each transmission. The objective is to design a channel sensing policy that maximizes the expected total reward collected over a finite or infinite horizon. This problem can be viewed as an instance of restless bandit problems, for which the form of optimal policies is unknown in general. We discover sets of conditions sufficient to guarantee the optimality of a myopic sensing policy; we show that under one particular set of conditions the myopic policy coincides with the Gittins index rule.**

***Index Terms*— Myopic sensing, Markov chain, POMDP, restless bandits, stochastic order.**

I. Introduction and Literature Survey

A. Motivation

Consider a communication system consisting of N independent channels. Each channel is modeled as a K-state (K finite) Markov chain (M.C.) with known matrix of transition probabilities. At each time period a user selects one channel to sense and uses it to transmit information. A reward depending on the state of the selected channel is obtained for each transmission. The objective is to design a channel sensing policy that maximizes the expected total reward (respectively, the expected total discounted reward) collected over a finite (respectively, infinite) time horizon.

The above channel sensing problem arises in cognitive radio networks, opportunistic scheduling over fading channels, as well as on resource-constrained jamming ([2]). In cognitive radio networks a secondary user may transmit over a channel only when the channel is not occupied by the primary user. Thus, at any time instant t, state 1 of the M.C. describing the channel can indicate that the channel is occupied at t by the primary user, and states 2 through K indicate the quality of the channel that is available to the secondary user at t. In opportunistic transmission over fading channels, states 1 through K of the M.C. describe, at any time instant, the quality of the fading channel. In resource-constrained jamming a jammer can only jam one channel at a time, and any given jamming/channel sensing policy results in an expected reward for the jammer due to successful jamming. The physical channels in all of the above problems have memory. Introducing a finite state (K-state) Markovian model for each channel allows us to capture the effect of the channel's memory on its current quality by allowing K to take large values.[1]

This channel sensing problem is also an instance of restless bandit problems (see [3], [4]). Restless bandit problems arise in many areas, including wired and wireless communication systems, manufacturing systems, economic systems, statistics, biomedical engineering, business, computer science, information systems etc. (see [3], [4]).

The problem described above can be formulated as a Partially Observed Markov Decision Process (POMDP) (see [5]) and can be solved, for any selection of the channels' transition probabilities and any selection of the reward process, by numerical methods. Such an approach has two drawbacks: (i) it does not provide any insight into the nature of optimal sensing strategies; (ii) it has very high computational complexity (PSPACE-complete, see [6]). For this reason we focus on identifying instances of the general problem where it is possible to explicitly characterize optimal sensing strategies. In this paper we discover sets of conditions under which the optimal sensing strategy is the myopic policy, that is, the policy that selects at every time instant the best (in the sense of stochastic order [7]) channel.

B. Related Work

The channel sensing problem has been studied in [5] using a POMDP framework. For channels described by two-state Markov chains (henceforth called two-state channels), the myopic policy was studied in [8], where its optimality was established when the number of channels is two. For more than two channels, the optimality of the myopic policy was proved in [9] under certain conditions on channel parameters. This result for two-state channels was extended in [10] under a relaxed "positively correlated" condition. In [11], under the same "positively correlated" channel condition, the myopic

Manuscript received May 29, 2013; revised October 8, 2013; accepted October 15, 2013. Date of publication November 5, 2013; date of current version December 20, 2013. This work was supported in part by the National Science Foundation under Grant CCF-1111061 and in part by NASA under Grant NNX12A0546. This paper was presented at the 50th Annual Allerton Conference on Communication, Control, and Computing [1].

The authors are with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: ouyangyi@umich.edu; teneket@umich.edu).

Communicated by R. Sundaresan, Associate Editor for Communications.

Digital Object Identifier 10.1109/TIT.2013.2288636

[1] We can create a Markovian model of a finite-memory system by appropriate state expansion.

policy was proved to be optimal for two-state channels when the user can select multiple channels at each time instance.

For general restless bandit problems, there is a rich literature; however, contrary to classical multi-armed bandit problems (see [4] and [12]), the structure (if any) of optimal strategies for general restless bandit problems is not currently known. To gain insight into the nature of restless bandit problems, research has focused on identifying instances where an optimal strategy or qualitative properties of optimal strategies can be explicitly determined. In [3] it has been shown that the Gittins index rule (see [4] and [12] for the definition of the Gittins index rule) is not optimal for general restless bandit problems. Moreover, this class of problems is PSPACE-hard in general [6]. In [3] Whittle introduced an index policy (referred to as Whittle's index) and an "indexability condition"; the asymptotic optimality of Whittle's index was addressed in [13]. Issues related to Whittle's indexability condition were discussed in [3], [4], [13]–[16]. For the two-state channel sensing problem, Whittle's index was computed in closed-form in [15], [16], where performance simulation of that index was provided. For some special classes of restless bandit problems, the optimality of index-type policies was established under certain conditions (see [17], [18]). Approximation algorithms for the computation of optimal policies for a class of restless bandit problems similar to the one studied in this paper were investigated in [19].

C. Contribution of the Paper

This paper contributes to the modeling and analysis of channel sensing problems, and to the state of the art of the theory of restless bandit problems. Specifically:

(i) Our model is more general than the two-state channels model considered so far in the literature for the same channel sensing problem (see Section I-B). Several communication channels, such as fading channels have memory. In order to have a Markovian description of a channel that captures its memory characteristics we need more than two (possibly a large number of) states in the Markov chain that describes channel's evolution. Our model considers Markovian channels with arbitrary (but finite) number of states; thus, it can capture memory characteristics of a large class of communication channels.

(ii) We discover sets of conditions under which the policy that chooses at every time instant the best (in the sense of stochastic order [7]) channel maximizes the total expected reward collected over a finite time horizon. We also show that under one particular set of conditions the above-described policy coincides with the Gittins index rule. Since our model is more general than previously studied models, our results are a contribution to the state of the art in cognitive radio networks, opportunistic scheduling and resource constrained jamming.

(iii) The results of this paper are a contribution to the state of the art of the theory of restless bandit problems. We show in Section II-C that the optimization problem formulated in this paper is an instance of restless bandit problems. Restless bandit problems are an important class of problems that arise in many areas of science and technology and very little is known about the structure of their optimal strategies in general. Our results reveal several instances of restless bandit problems where: (a) the myopic policy is optimal; and (b) the myopic policy is optimal and coincides with the Gittins index rule. Thus, the results of this paper can be useful in many areas of science and technology.

(iv) Our methodology (in particular the development of ordering-based policies in Section III-F) can be useful in stochastic scheduling problems where the optimality of "list policies" is investigated (see [20] for an example of list policies).

D. Organization

The rest of this paper is organized as follows. In Section II, we present the model and the formulation of the optimization problem associated with the channel sensing problem. In Section III, we consider the finite horizon problem and identify sets of conditions sufficient to guarantee the optimality of the myopic policy; we briefly discuss the extension of our results to the infinite horizon. In Section IV we show that under one particular set of conditions the myopic policy coincides with the Gittins index rule. We conclude in Section V. The proofs of several intermediate results needed to establish the optimality of the myopic policy appear in Appendices A-D.

II. Model and the Optimization Problem

A. The Model

Consider a communication system consisting of N identical channels. Each channel is modeled as a K-state (K finite) Markov chain (M.C.) with (the same) matrix of transition probabilities P,

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1K} \\ p_{21} & p_{22} & \cdots & p_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ p_{K1} & p_{K2} & \cdots & p_{KK} \end{bmatrix} = \begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_K \end{bmatrix}, \qquad (1)$$

where $P_1, P_2, \ldots, P_K$ are row vectors. As pointed out in Section I-C, channels that have memory can still be modeled by Markov chain by expanding the number of states in the M.C. to account for the channel's memory. The K-state M.C. model here captures the memory characteristics of a larger class of communication channels. We assume that the channel's quality increases as the number of its state increases. We want to use this communication system to transmit information. For that matter, at each time $t = 0, 1, ..., T$, we can select one channel, observe its state, and use it to transmit information.

Let X_t^n denote the state of channel n at time t, and let U_t denote the decision made at time t; $U_t \in \{1, 2, ..., N\}$, where $U_t = n$ means that channel n is chosen for data transmission at time t.

Initially, before any channel selection is made, we assume that we have probabilistic information about the state of each of the N channels. Specifically, we assume that at $t = 0$ the decision-maker (the entity that decides which channel to sense at each time instant) knows the probability mass function (PMF) on the state space of each of the N channels; that is, the decision-maker knows $\pi_0 := (\pi_0^1, \pi_0^2, \ldots, \pi_0^N)$, where

$$\pi_0^n := (\pi_0^n(1), \pi_0^n(2), \ldots, \pi_0^n(K)), \quad n = 1, 2, \ldots, N, \quad (2)$$

$$\pi_0^n(i) := P(X_0^n = i), \quad i = 1, 2, \ldots, K. \quad (3)$$

Then, in general,

$$U_0 = g_0(\pi_0), \quad (4)$$

$$U_t = g_t(Y^{t-1}, U^{t-1}, \pi_0), \quad t = 1, 2, \ldots, \quad (5)$$

where

$$Y^{t-1} := (Y_0, Y_1, \ldots, Y_{t-1}), \quad U^{t-1} := (U_0, U_1, \ldots, U_{t-1}), \quad (6)$$

and $Y_t = X_t^{U_t}$ denotes the observation at time t; Y_t gives the state of the channel that is chosen at time t (that is, if $U_t = 2$, Y_t gives the state of channel 2 at time t).

Let $R(t)$ denote the reward obtained by the transmission at time t. We assume that $R(t)$ depends on the state of the channel chosen at time t. That is

$$R(t) = R_i, \quad i = 1, 2, \ldots, K, \quad (7)$$

if the state of the channel chosen at t is i.

B. The Optimization Problem

Under the above assumptions, the objective is to solve the following finite horizon (T) optimization problem:

Problem (P1)

$$\max_{g \in \mathcal{G}_s} E^g \left[\sum_{t=0}^{T} \beta^t R(t) \right], \quad (8)$$

where β is the discount factor ($0 < \beta \leq 1$) and $\mathcal{G}_s$ is the set of separated policies $g := (g_0, g_1, \ldots)$ (see [21], Chapter 6), that are such that

$$U_t = g_t(\pi_t) \text{ for all } t, \quad (9)$$

$$\pi_t := (\pi_t^1, \pi_t^2, \ldots, \pi_t^N), \quad (10)$$

$$\pi_t^n := (\pi_t^n(1), \pi_t^n(2), \ldots, \pi_t^n(K)), \quad n = 1, 2, \ldots, N, \quad (11)$$

$$\pi_t^n(i) := P(X_t^n = i | Y^{t-1}, U^{t-1}), \quad i = 1, 2, \ldots, K, \quad (12)$$

and π_t evolves as follows. If $U_t = n, Y^n = i$, then

$$\pi_{t+1}^n = P_i, \quad (13)$$

$$\pi_{t+1}^j = \pi_t^j P, \quad \text{for all } j \neq n. \quad (14)$$

C. Characteristics of the Optimization Problem

The optimization problem (P1) formulated above is a POMDP; it can be solved by numerical methods, but such an approach has the drawbacks pointed out in Section I-A.

Problem (P1) can also be viewed as an instance of restless bandit problems as follows. We can view the N channels as N arms with their PMFs as the states of the arms. The decision maker knows perfectly the states of the N arms at every time instant. One arm is operated (selected) at each time t, and an expected reward depending on the state of the selected arm is received. If arm n (channel n) is not selected at t, its PMF π_t^n evolves according to

$$\pi_{t+1}^n = \pi_t^n P; \quad (15)$$

if arm n (channel n) is selected at t, its PMF evolves according to

$$\pi_{t+1}^n = P_{Y_t}, \quad P(Y_t = x) = \pi_t^n(x). \quad (16)$$

Since the selected bandit process evolves in a way that differs from the evolution of the non-selected bandit processes, this problem is a restless bandit problem.

In general, restless bandit problems are difficult to solve because forward induction (the solution methodology for the classical multi-armed bandit problem) does not result in an optimal policy [4]. Consequently, optimal policies may not be of the index type, and the form of optimal policies for general restless bandit problems (hence, the channel sensing problem) is still unknown.

To gain insight into the nature of the channel sensing problem (as well as general restless bandit problems), it is important to discover special instances of the problem where it is possible to explicitly determine optimal strategies or the structure of optimal strategies. For this season, in this paper we focus on the "myopic policy" and we discover sets of conditions under which it is optimal. We define the myopic policy as follows. Let Π denote the set of PMFs on the state space $S = \{1, 2, \ldots, K\}$. We define the concept of stochastic dominance/order (see [7]). Stochastic dominance $\geq_{st}$ between two row vectors $x, y \in \Pi$ is defined as follows:
$x \geq_{st} y$ if

$$\textstyle\sum_{j=i}^K x(j) \geq \sum_{j=i}^K y(j), \quad \text{for } i = 2, 3, \ldots, K. \quad (17)$$

Definition 1: The myopic policy $g^m := (g_0^m, g_1^m, \ldots, g_T^m)$ is the policy that selects at each time instant the best (in the sense of stochastic order) channel; that is,

$$g_t^m(\pi_t) = i \quad \text{if } \pi_t^i \geq_{st} \pi_t^j \quad \forall j \neq i. \quad (18)$$

III. Analysis of the Finite Horizon Problem

We will prove the optimality of the myopic policy g^m for Problem (P1) under certain specific assumptions on the structure of the Markov chains describing the channels, on the instantaneous rewards $R = [R_1, R_2, R_3, \ldots, R_K]^T$ and on the initial PMFs $\pi_0^1, \pi_0^2, \ldots, \pi_0^N$. We proceed as follows. In Section III-A we discuss why the problem under consideration is not a trivial extension of the instance where each channel has only two states (studied in [10]). This discussion helps to justify the key assumptions/conditions we make in Section III-B. These assumptions/conditions reduce to those of [10] when $K = 2$. The main result of the paper is stated in Section III-C; its proof appears in Section III-D to III-G. The key features of the solution approach and the role of the conditions in the approach are discussed in Section III-H, where the extension to the infinite horizon problem is also presented briefly.

A. Difficulties in Establishing the Optimality of the Myopic Policy

The situation where each channel has two states, i.e. $K = 2$, has been previously investigated in [10] where the optimality of the myopic policy is established under some conditions. In the two-state channels situation, the PMF in equation (11) (called the information state of the POMDP, see [21]) can be described by a number, the conditional probability of the "best state". As a result of this feature, the information states of all channels can be totally ordered at any time regardless of channels' evolution. Such an ordering is needed for the derivation of the results in [10]. In our problem the information state defined by equation (11) is a $(K-1)$-dimensional vector; $(K - 1)$-dimensional vectors can not, in general, be ordered at every time instant. This difference between the information state of two-state channels and the one in our paper results in a lot of complications in extending the results of [10] to multi-state channels.

In general, an extension of the results on the optimality of the myopic policy for two-state channels to multi-state channels would require: (i) An ordering of the channels' information states (PMFs defined by eq. (1)) at every time instant. Such an ordering can only be ensured under certain conditions (Conditions (A1)-(A3) appearing in Section III-B) on the evolution of the channels. (ii) If the myopic policy is to be optimal, the instantaneous expected gain incurred by choosing the best channel (say channel n) versus any other channel (say channel m) must overcompensate expected future losses in performance resulting in when channel m is chosen instead of channel n. We have K channel states and this leads to $K - 1$ inequalities in Condition (A4) (appearing in Section III-B) on the separation of instantaneous rewards. Condition (A4) describes how much the instantaneous rewards obtained in states i and $i - 1, i = 2, 3, ..., K$, should be separated so as to ensure the optimality of the myopic policy.

The above discussion provides the rationale for Conditions (A1)-(A4) appearing below.

B. Key Assumptions/Conditions

We make the following assumptions/conditions

(A1)

$$P_K \geq_{st} P_{K-1} \geq_{st} \cdots \geq_{st} P_1. \tag{19}$$

Note that the quality of a channel state increases as its number increases. Assumption (A1) ensures that the higher the quality of the channel's current state the higher is the likelihood that the next channel state will be of high quality. This requirement is the same as the "positively correlated" condition when $K = 2$ in [10].

(A2) Let $\Pi P := \{\pi P : \pi \in \Pi\}$. At time 0,

$$\pi_0^1, \pi_0^2, \ldots, \pi_0^N \in \Pi P, \tag{20}$$

$$\text{and } \pi_0^1 \leq_{st} \pi_0^2 \leq_{st} \cdots \leq_{st} \pi_0^N. \tag{21}$$

Assumption (A2) states that initially the channels can be ordered in terms of their quality, expressed by the PMF on S. Moreover, the initial PMFs of the channels are in ΠP. The requirement expressed by (20) is always satisfied since the channels evolve before we begin sensing them. Requirement (20) also ensures that the initial PMFs on the channel states are in the same space as all subsequent PMFs.

(A3) There exists some L, $2 \leq L \leq K$ such that

$$P_1 P \geq_{st} P_{L-1}, \tag{22}$$

$$P_K P \leq_{st} P_L. \tag{23}$$

Assumption (A3) along with (A2) ensure that, any PMF π reachable from a non-selected channel has quality between P_{L-1} and P_L, that is $P_L \geq_{st} \pi \geq_{st} P_{L-1}$ (see also Property 2, Section III-D).

As pointed out in Section III-A, (A1)-(A3) ensure that the channels' information states are ordered at any time t (see Property 3, Section III-D).

(A4)

$$\begin{aligned} R_i - R_{i-1} &\geq \beta(P_i - P_{i-1})M \\ &\geq \beta(P_i - P_{i-1})U \geq 0, \quad \text{for } i \neq L, \end{aligned} \tag{24}$$

$$R_L - R_{L-1} \geq \beta(h - P_{L-1}R) \geq 0, \tag{25}$$

where M and U are vectors given by

$$M := U + \beta \sum_{i \geq L} p_{Ki} P U, \tag{26}$$

$$U_i := R_i \quad \text{for } i = 1, 2, \ldots, L-1, \tag{27}$$

$$U_i := R_i + \beta(P_i - P_{L-1})U, \quad \text{for } i = L, L+1, \ldots, K, \tag{28}$$

and h is given by

$$h = \frac{P_K R - \beta \sum_{i<L} p_{Ki} P_i R}{1 - \beta \sum_{i<L} p_{Ki}}. \tag{29}$$

Assumption (A4) states that the instantaneous rewards obtained at different states of the channel are sufficiently separated (see (24) and (25)). The reason for such a separation was discussed in Section III-A.

We note that (A1)-(A4) describe sets of sets of assumptions/conditions; for every value of $L, L = 2, 3, \ldots, K$, we have a distinct set of conditions. In [1] we show, via an example, that Conditions (A1)-(A4) can be simultaneously satisfied.

We now compare the above conditions with those made in [1] and [10]. When $L = K$, the above conditions are exactly the same as those in [1]. In [1] we did not address situations where $L \neq K$ that is, situation where the quality of the information state resulting from a non-selected channel is between P_L and P_{L-1} for $L \neq K$. Consequently, the result of this paper subsumes the results obtained in [1].

When $K = 2$ the above Conditions (A1)-(A4) reduce to those of [10] as follows. When $K = 2, L = K$. Then, our Conditions (A1)-(A4) reduce to

$$p_{2,2} \geq p_{1,2}, \tag{30}$$

$$\pi_0^n = (1 - p^n, p^n), \quad p_{1,2} \leq p^n \leq p_{2,2} \text{ for } n = 1, 2, \ldots, N, \tag{31}$$

$$p^1 \leq p^2 \leq \cdots \leq p^N. \tag{32}$$

Condition (30) is precisely the "positively correlated" condition in [10]. Condition (31) is satisfied, if the channels evolve before we begin sensing them (before time $t = 0$). Condition (32) is always satisfied by renumbering of the channels. (For more details see [22]).

C. The Main Result

The main result we establish in this paper is given by Theorem 1 below.

Theorem 1: Under assumptions (A1)-(A4), the myopic policy g^m, that is, the policy that selects at every time instant the best (in the sense of stochastic order) channel is optimal for Problem (P1).

We proceed to establish the optimality of the myopic policy g^m as follows. In sections III-D-III-F we develop preliminary results needed for the proof of Theorem 1. Specifically: In section III-D we present three properties of the evolution of the PMFs on the channel states. In section III-E we present a property of the instantaneous expected reward. In section III-F we define a class of ordering-based channel sensing policies $\mathcal{G}^O$ which includes the myopic policy g^m; using the results of sections III-D and III-E we discover four properties of the expected reward resulting from any policy in $\mathcal{G}^O$. In section III-G we use the results of section III-F to establish the optimality of the myopic policy for Problem (P1). The properties' proofs appear in Appendices A-D.

D. Properties of the Channels' Evolution

Under assumptions/conditions (A1)-(A4) stated in section III-B, the following properties hold.

Property 1: Let $x, y \in \Pi$. Under Assumption (A1),

$$x \geq_{st} y \Longrightarrow xP \geq_{st} yP. \tag{33}$$

An implication of Property 1 is the following. If at any time t the information states of two channels (expressed by the PMFs on their state space) are stochastically ordered and none of these channels is sensed at t, then the same stochastic order between the information states at time $t+1$ is maintained.

Property 2: Let $\pi = xP^2 \in \Pi P^2$, $\Pi P^2 := \{\pi = xP^2, x \in \Pi\}$. Under (A1)-(A3),

$$P_L \geq_{st} xP^2 \geq_{st} P_{L-1}. \tag{34}$$

Property 2 says the following. By Condition (A2) a channel's information state (the PMF on its state space) is always in ΠP. If the channel is not sensed at time t, then at time $t+1$ its information state is in ΠP^2, moreover it is stochastically always between P_{L-1} and P_L. If the channel is sensed at time t and its observed state is larger than or equal to L (respectively smaller than L), then at time $t+1$ this channel is in the stochastically largest (respectively stochastically smallest) information state among all channels.

Property 3: Under (A1)-(A3), we have either $\pi_t^n \leq_{st} \pi_t^m$ or $\pi_t^m \leq_{st} \pi_t^n$ for all $n, m \in \{1, 2, \ldots, N\}$ for all t.

Property 3 states that under (A1)-(A3) the information states of all channels can be ordered stochastically at all times.

The proofs of Properties 1-3 appear in Appendix A.

E. A Property of the Instantaneous Expected Reward

A direct consequence of Condition (A4) is the following property of the instantaneous expected reward:

Property 4: Let $x, y \in \Pi$. Let υ be a column vector in increasing order, i.e. $\upsilon_i \geq \upsilon_{i-1}$ for $i = 2, 3, \ldots, K$. If $x \geq_{st} y$, we have

(i) $(x - y)\upsilon \geq 0$.

(ii) $(x - y)M \geq (x - y)U \geq (x - y)R \geq 0$, where M, U, R are defined by equations (24)-(28).

(iii) $(x - y)M \geq \beta(x - y)PM$.

(iv) If $x(i) = y(i)$ for all $i \geq L$ or $x(i) = y(i)$ for all $i < L$, then

$$(x - y)R \geq \beta(x - y)PM. \tag{35}$$

Part (i) of Property 4 says the following. Consider a reward vector such that the reward increases as the quality of the channel state increases. Then the expected reward increases as the information state of the channel increases stochastically.

Part (ii) is a restatement of part (i) when the reward vector υ takes the values $M - U, U - R, R$.

Part (iii) can be interpreted as follows. Consider the reward vector M defined by (26). Consider two channels, channel i and channel j, that have information states x and y respectively, such that $x \geq_{st} y$. Consider the following scenarios: (SC1) Sense channel i first, then sense channel j; (SC2) Sense channel j first, then sense channel i. Afterwards, continue with the same channel selection sequence under both scenarios. Then part (iii) of Property 4 asserts that scenario (SC1) is better than scenario (SC2) in terms of the expected accumulated rewards; that is, it is better to sense the best (in the sense of stochastic order) channel first.

Part (iv) has an interpretation similar to that of part (iii) when x, y satisfy the condition of part (iv).

The proof of Property 4 appears in Appendix B.

F. Properties of the Reward Associated With Ordering-Bathe Optimality of the Myopicsed Channel Sensing Polices

In this section we introduce ordering-based policies and study their properties. The reason for considering this class of policies is because under Conditions (A1)-(A4) we obtain the following: (i) The performance of any sensing policy can be upper-bounded by an appropriately chosen ordering-based policy (see Section III-G); thus, for the solution of the original optimization problem (Problem (P1)) we can restrict attention to ordering-based policies. (ii) The myopic policy is an optimal ordering-based policy. Combining (i) and (ii) we establish the optimality of the myopic policy for Problem (P1).

We note that Properties 1-4, developed so far, are essential for the discovery of the properties of ordering-based policies that lead eventually to the solution of Problem (P1) (see discussion in Section III-H).

Let $\mathcal{O}$ be the set of all orderings/permutations of the N channels $\{1, 2, \ldots, N\}$. Consider the ordering-based selection function $\hat{g} : \mathcal{O} \mapsto \{1, 2, \ldots, N\}$ and the ordering update mapping $\hat{m} : \mathcal{O} \times \{1, 2, \ldots, K\} \mapsto \mathcal{O}$ defined as follows.

For every $O := (O(1), O(2), \ldots, O(N)) \in \mathcal{O}$,

$$\hat{g}(O) = O(N), \tag{36}$$

$$\hat{m}(O, y) = \begin{cases} O & \text{if } y \geq L, \\ SO & \text{if } y < L, \end{cases} \tag{37}$$

where S is the cyclic shift operator on $\mathcal{O}$ such that

$$SO =: (O(N), O(1), O(2), \ldots, O(N-1)). \tag{38}$$

Given a channel ordering $O_t \in \mathcal{O}$ at time t, we define an ordering-based channel sensing policy $g_{t:T}^{O_t} := (g_t^{O_t}, g_{t+1}^{O_t}, \ldots, g_T^{O_t})$ as follows.

$$U_t = g_t^{O_t}(O_t) = \hat{g}(O_t) = O(N), \tag{39}$$

$$O_s = \hat{m}(O_{s-1}, Y_{s-1}), \quad \text{for } s = t+1, t+2, \ldots, T, \tag{40}$$

$$\begin{aligned} U_s &= g_s^{O_t}(Y_{t:s-1}, U_{t:s-1}) \\ &= g_s^{O_t}(O_s) = \hat{g}(O_s), \quad \text{for } s = t+1, t+2, \ldots, T. \end{aligned} \tag{41}$$

At time $s, t \leq s \leq T$, $g_s^{O_t}$ chooses the last channel in O_s; the ordering O_s is shifted to the right by the update mapping $\hat{m}$ whenever the observed state is less than L, and remains the same otherwise. As a result of the above specification of $g_{t:T}^{O_t}$, if at time t channel n is on the right of channel m in the ordering O_t, channel n will be sensed by policy $g_{t:T}^{O_t}$ before channel m.

Note that, the policy $g_{t:T}^{O_t}$ is not a separated policy in general. However, if the ordering $O_0 = (O_0(1), O_0(2), \ldots, O_0(N))$ at time 0 is such that $\pi_0^{O_0(1)} \leq_{st} \pi_0^{O_0(2)} \leq_{st} \cdots \leq_{st} \pi_0^{O_0(N)}$, then $g_{0:T}^{O_0}$ is the myopic policy g^m, therefore; $g_{0:T}^{O_0} = g^m \in \mathcal{G}_s$, as the following property shows.

Property 5: At time $t = 0$ consider the ordering O_0 such that $\pi_0^{O_0(1)} \leq_{st} \pi_0^{O_0(2)} \leq_{st} \cdots \leq_{st} \pi_0^{O_0(N)}$. Then, the ordering based policy $g_{0:T}^{O_0}$ is just the myopic policy g^m.

The validity of Property 5 crucially depends on Properties 1 and 2, which say that stochastic order is maintained under the evolution of unobserved channels (Property 1), and the observed channel is either the stochastically best or the stochastically worst among all channels (Property 2). Without Properties 1 and 2 the myopic policy is not an ordering-based policy. The proof of Property 5 appears in Appendix C.

Define $V_t(O_t, \pi_t^1, \pi_t^2, \ldots, \pi_t^N)$ to be the expected reward collected from time t up to and including T due to the ordering-based policy $g_{t:T}^{O_t}$. That is,

$$\begin{aligned} &V_t(O_t, \pi_t^1, \pi_t^2, \ldots, \pi_t^N) \\ &\quad := E^{g_{t:T}^{O_t}}\left[\sum_{l=t}^{T} \beta^{l-t} R(l) | \pi_t^1, \pi_t^2, \ldots, \pi_t^N\right]. \end{aligned} \tag{42}$$

Then, $V_t(O_t, \pi_t^1, \pi_t^2, \ldots, \pi_t^N)$ can be written recursively as follows.

$$V_T(O_t, \pi_T^1, \pi_T^2, \ldots, \pi_T^N) = \pi_T^{O_t(N)} R, \tag{43}$$

$$\begin{aligned} &V_t(O_t, \pi_t^1, \pi_t^2, \ldots, \pi_t^N) \\ &= \pi_t^{O_t(N)} R + \beta \sum_{i<L} \pi_t^{O_t(N)}(i) V_{t+1}(SO_t, \pi_{t+1}^1, \ldots, \pi_{t+1}^N) \\ &\quad + \beta \sum_{i \geq L} \pi_t^{O_t(N)}(i) V_{t+1}(O_t, \pi_{t+1}^1, \ldots, \pi_{t+1}^N), \end{aligned} \tag{44}$$

$$\text{where } \pi_{t+1}^n = \begin{cases} P_i & \text{for } n = O_t(N), \\ \pi_t^n P & \text{otherwise.} \end{cases} \tag{45}$$

The function $V_t(O_t, \pi_t^1, \pi_t^2, \ldots, \pi_t^N)$ defined above possesses Properties 6-9 below. We will explain the role of these properties in Section III-H after we prove the main result on the optimality of the myopic policy in Section III-G.

Property 6: Let $\hat{\pi}_t^1, \pi_t^1, \pi_t^2, \ldots, \pi_t^N \in \Pi P$ and $O_t \in \mathcal{O}$. Define

$$\begin{aligned} &L_t(O_t, \hat{\pi}_t^1, \pi_t^1, \pi_t^2, \ldots, \pi_t^N) \\ &:= V_t(O_t, \hat{\pi}_t^1, \pi_t^2, \ldots, \pi_t^N) - V_t(O_t, \pi_t^1, \pi_t^2, \ldots, \pi_t^N). \end{aligned} \tag{46}$$

If $\hat{\pi}_t^1 \geq_{st} \pi_t^1$, and $O_t(n) = 1$, then for all $m < n$

$$\begin{aligned} 0 &\leq L_t(O_t, \hat{\pi}_t^1, \pi_t^1, \pi_t^2, \ldots, \pi_t^N) \\ &\quad - L_t(S^{-m} O_t, \hat{\pi}_t^1, \pi_t^1, \pi_t^2, \ldots, \pi_t^N) \\ &\leq (\hat{\pi}_t^1 - \pi_t^1) U, \end{aligned} \tag{47}$$

where $S^{-m} O_t$ is the counter-clockwise cyclic shift of O_t by m positions, that is,

$$\begin{aligned} &S^{-m} O_t \\ &= (O_t(m+1), O_t(m+2), \ldots, O_t(N), O_t(1), \ldots, O_t(m)). \end{aligned} \tag{48}$$

Property 7: For $O_t \in \mathcal{O}$, define the operator W_{nm} as follows.

$$W_{nm} O_t(i) := \begin{cases} O_t(n) & \text{for } i = m, \\ O_t(m) & \text{for } i = n, \\ O_t(i) & \text{otherwise.} \end{cases} \tag{49}$$

If $\hat{\pi}_t^1 \geq_{st} \pi_t^1$, and $O_t(n) = 1$, then for $m < n$

$$\begin{aligned} 0 &\leq L_t(O_t, \hat{\pi}_t^1, \pi_t^1, \pi_t^2, \ldots, \pi_t^N) \\ &\quad - L_t(W_{nm} O_t, \hat{\pi}_t^1, \pi_t^1, \pi_t^2, \ldots, \pi_t^N) \\ &\leq (\hat{\pi}_t^1 - \pi_t^1) M. \end{aligned} \tag{50}$$

The meaning of Properties 6 and 7 is the following. Restrict attention to ordering-based policies. Take any channel, say channel 1. Replace it with a better quality (in the sense of stochastic order) channel. Such a replacement will result in an improvement in performance. This improvement is different for different channel orderings. The earlier channel 1 is used (that is, the closer to the right-most position in the ordering channel 1 is) the higher is the improvement. Properties 6 and 7 also provide bounds on the difference between maximum and minimum improvement. These bounds are useful in proving Properties 6 and 7 by induction.

Property 8: If $\pi_t^{O_t(n)} \geq_{st} \pi_t^{O_t(m)}$, then for $m < n$ then

$$V_t(O_t, \pi_t^1, \pi_t^2, \ldots, \pi_t^N) \geq V_t(W_{nm} O_t, \pi_t^1, \pi_t^2, \ldots, \pi_t^N). \tag{51}$$

Property 8 states that if the position of two channels in any arbitrary but fixed channel ordering are interchanged so that the better (in the stochastic order sense) channel comes closer to the right-most position (i.e. it is used earlier) in the new ordering, the performance due to the ordering-based policy improves.

Property 9: For $O_t \in \mathcal{O}$, define the operator A_{nm} as follows.

$$A_{nm} O_t(i) := \begin{cases} O_t(n) & \text{for } i = m, \\ O_t(i-1) & \text{for } i = m+1, m+2, \ldots, n, \\ O_t(i) & \text{otherwise.} \end{cases} \tag{52}$$

If $\pi_t^1 \leq_{st} \pi_t^1 P$, and $O_t(n) = 1$, then

$$\begin{aligned} &V_t(A_{nm} O_t, \pi_t^1, \pi_t^2, \ldots, \pi_t^N) - V_t(O_t, \pi_t^1, \pi_t^2, \ldots, \pi_t^N) \\ &\leq h - \pi_t^1 P^{N-n} R. \end{aligned} \tag{53}$$

Property 9 states the following. Suppose that a channel, say channel 1, is such that as long as it is not sensed its quality is continuously improving (i.e. its PMF is continuously increasing stochastically). Then, no matter how late this channel is sensed (that is, no matter how much we move the channel to the left from its initial position in the original channel ordering) the change in performance due to an ordering-based policy can not exceed a certain bound, given by the right hand side of (53).

Properties 6-9 are proved simultaneously in Appendix D. The idea of their proof may be useful in stochastic scheduling problems where the optimality of "list polices" ([20]) is investigated. In the analysis of "list polices", it is important to compare the performance due to different orders of task processing/scheduling. To do this we consider an initial ordering of the tasks to be processed. We perturb the ordering and study the resulting change in performance. Several types of perturbation need to be examined. Typical types of such perturbations are described in the statements of Properties 6-9. The proof of Properties 6-9 indicates that such perturbations can not be analyzed in isolation but have to be considered simultaneously.

G. Proof of the Main Result (Theorem 1)

Proof: We proceed by induction.
At T, the expected reward is the instantaneous expected reward. Since by part (ii) of Property 4 a better channel (in the sense of stochastic order) gives larger instantaneous expected reward, the myopic policy g^m is optimal at T. This establishes the basis of induction.

The induction hypothesis is that the myopic policy g^m is optimal at $t+1, t+1, \ldots, T$.

Without loss of generality, we assume $\pi_t^1 \leq_{st} \pi_t^2 \leq_{st} \cdots \leq_{st} \pi_t^N$. Consider any policy g. If g picks channel n at time t, then the expected reward collected from t on due to the policy g is given by

$$\begin{aligned} &E^g\left[\sum_{l=t}^{T} \beta^{l-t} R(l) | \pi_t^1, \ldots, \pi_t^N\right] \\ &= \pi^n R + \sum_{i=1}^{K} \pi_t^n(i) \\ &\quad \times E^g\left[\sum_{l=t+1}^{T} \beta^{l-t} R(l) | \pi_{t+1}^n = P_i, \pi_{t+1}^m = \pi_t^m P \text{ for } m \neq n\right] \\ &\leq \pi^n R + \sum_{i=1}^{K} \pi_t^n(i) \\ &\quad \times E^{g^m}\left[\sum_{l=t+1}^{T} \beta^{l-t} R(l) | \pi_{t+1}^n = P_i, \pi_{t+1}^m = \pi_t^m P \text{ for } m \neq n\right] \\ &= \pi_t^n R + \beta \sum_{i<L} \pi_t^n(i) V_{t+1}(S O_t, \pi_{t+1}^1, \ldots, \pi_{t+1}^N) \\ &\quad + \beta \sum_{i \geq L} \pi_t^n(i) V_{t+1}(O_t, \pi_{t+1}^1, \ldots, \pi_{t+1}^N) \\ &= V_t(O_t, \pi_t^1, \ldots, \pi_t^N). \end{aligned} \tag{54}$$

The inequality in (54) follows from the induction hypothesis and the ordering $O_t := (1, 2, \ldots, n-1, n+1, \ldots, N, n)$.

Since $\pi_t^n \leq_{st} \pi_t^m$ for all $m = n+1, n+2, \ldots, N$, repeatedly applying Property 8 we get

$$\begin{aligned} &V_t(O_t, \pi_t^1, \ldots, \pi_t^N) \\ &\leq V_t((1, 2, \ldots, n-1, n, n+1, \ldots, N), \pi_t^1, \ldots, \pi_t^N) \\ &= E^{g^m}\left[\sum_{l=t}^{T} R(l) | \pi_t^1, \pi_t^2, \ldots, \pi_t^N\right]. \end{aligned} \tag{55}$$

Combining (54), (55) we obtain

$$\begin{aligned} &E^g\left[\sum_{l=t}^{T} \beta^{l-t} R(l) | \pi_t^1, \pi_t^2, \ldots, \pi_t^N\right] \\ &\qquad \leq E^{g^m}\left[\sum_{l=t}^{T} \beta^{l-t} R(l) | \pi_t^1, \pi_t^2, \ldots, \pi_t^N\right], \end{aligned} \tag{56}$$

which completes the proof. ■

H. Remarks

1) The key steps in establishing the optimality of the myopic policy, under the assumptions made in the problem formulation, are the following:

(K1) The assertion that the performance of any separated policy can be upper-bounded by the performance of an ordering-based policy. Consequently, for the solution of the original optimization problem, one can restrict attention to ordering-based policies.

(K2) The assertion that the performance of an ordering-based policy improves when a better (in the sense of stochastic order) channel is used earlier. This assertion implies the optimality of the myopic policy.

The assertion of (K1) is established in Theorem 1 (its induction step). The assertion of (K2) is established by Property 8, provided that the myopic policy is an ordering-based policy, and that stochastic order is maintained among all channels at every time. The fact that the myopic policy is an ordering-based policy is ensured by Property 5. The existence of a stochastic ordering among all channels at any time t is ensured by Property 3. To establish these properties we need Properties 1-9.

We now elaborate on the interdependence of Properties 1-9. Property 3, which asserts that channels can be ordered stochastically, is a consequence of Properties 1 and 2 for the unobserved channels and the observed channel, respectively. Properties 1 and 2 also ensure that the myopic policy g^m belongs to the class of ordering-based policies (Property 5). Property 8 is a special case of Property 7 when $\hat{\pi}_t^1 = \pi_t^{O_t(m)} \geq_{st} \pi_t^1 = \pi_t^{O_t(n)}$. Property 7 is coupled with Properties 6 and 9, that is, Properties 6, 7 and 9 need to be proven simultaneously. The proof of Properties 6, 7 and 9 requires Property 4.

The upper bounds that appear in Properties 6, 7 and 9 are essential in establishing the optimality of the myopic policy. These bounds along with Condition (A4) ensure that the instantaneous advantage in expected reward obtained by the use of the myopic policy g^m over any other policy g, overcompensates any future possible expected reward losses of g^m as compared to g.

2) The result of Theorem 1 is also valid for the infinite horizon expected discounted cost problem (see [22] for details).

IV. Myopic Policy vs. Gittins Index Rule

In this section we investigate conditions under which the myopic policy coincides with the Gittins index rule.

Select a channel, say channel $n, n = 1, 2, \ldots, N$. For PMF $\pi \in \Pi$, the Gittins index ([4][12]) of channel n is defined is defined by

$$\nu^n(\pi) := \max_{\tau} \frac{E^{g^\tau}\left[\sum_{t=0}^{\tau-1} \beta^t \pi_t^n R | \pi_0^n = \pi\right]}{E^{g^\tau}\left[\sum_{t=0}^{\tau-1} \beta^t | \pi_0^n = \pi\right]}, \tag{57}$$

where τ is any stopping time with respect to $\{\pi_t^n, t = 0, 1, \ldots\}$ and g^τ chooses channel n from $t = 0$ up to $t = \tau - 1$. The Gittins index rule ([4][12]) chooses the channel with the highest Gittins index at every time instant t.

In condition (A3) (Section III-B) L is fixed; it can be any number form 2 to K. In this section we show that when $L = K$, under conditions (A1)-(A4), after time 0 the myopic policy coincides with the Gittins index rule. We establish this result via Theorems 2 and 2.

Theorem 2:

(i) For $\pi \in \Pi P$, $P_{K-1} \leq_{st} \pi \leq_{st} P_K$, the Gittins index $\nu(\pi)$ is given by

$$\nu(\pi) = \frac{\pi R + \beta\pi(K)\frac{P_K R}{1-\beta p_{KK}}}{1 + \beta\pi(K)\frac{1}{1-\beta p_{KK}}}. \tag{58}$$

(ii) If $\pi_x, \pi_y \in \Pi P$ and $P_{K-1} \leq_{st} \pi_y \leq_{st} \pi_x \leq_{st} P_K$, then $\nu(\pi_x) \geq \nu(\pi_y)$.

(iii) If $\pi \in \Pi P$ and $P_{K-1} \leq_{st} \pi \leq_{st} P_K$, then $\nu(\pi) \geq \nu(P_i)$ for $i < K$.

Proof: (i) From Property 2 and part (ii) of Property 4 we know that

$$\pi R \leq P_K R \quad \text{for all } \pi \in \Pi P. \tag{59}$$

Using (59) in the definition of Gittins index (57) we get

$$\nu(\pi) \leq P_K R \quad \text{for all } \pi \in \Pi P. \tag{60}$$

Letting $\tau = 1$ in (57), we get an lower bound on the Gittins index of P_K

$$\nu(P_K) \geq E[R(\pi_0)|\pi_0 = P_K] = P_K R. \tag{61}$$

Combining (60) and (61), $\nu(P_K) = P_K R$ and the PMF P_K has the largest Gittins index among all PMFs.

From Theorem 4.1 in [23] we know that the second largest Gittens index among PMFs
$\{\pi, P_1, P_2, .., P_{K-1}, P_K\}$ is given by

$$\max_{x=\{\pi, P_1, P_2, .., P_{K-1}\}} \nu_K(x), \tag{62}$$

where

$$\nu_K(x) : = \frac{A_K(x)}{B_K(x)}, \tag{63}$$

$$A_K(x) : = xR + \beta x(K) A_K(P_K),\ A_K(P_K) = \frac{P_K R}{1 - \beta P_{KK}}, \tag{64}$$

$$B_K(x) : = 1 + \beta x(K) B_K(P_K),\ B_K(P_K) = \frac{1}{1 - \beta P_{KK}}. \tag{65}$$

We now show that for $P_{K-1} \leq_{st} \pi \leq_{st} P_K$

$$\nu_K(\pi) = \max_{x=\{\pi, P_1, P_2, .., P_{K-1}\}} \nu_K(x). \tag{66}$$

For that matter we need to show that $\nu(\pi_x) \geq \nu(\pi_y)$ whenever $\pi_x \geq_{st} \pi_y, \pi_x, \pi_y \in \Pi P$. From (63),

$$\begin{aligned} \nu_K(\pi_x) &= \frac{\pi_x R + \beta\pi_x(K) A_K(P_K)}{1 + \beta\pi_x(K) B_K(P_K)} \\ &= P_K R + \frac{\pi_x R - P_K R}{1 + \beta\pi_x(K) B_K(P_K)} \\ &\geq P_K R + \frac{\pi_y R - P_K R}{1 + \beta\pi_x(K) B_K(P_K)} \\ &\geq P_K R + \frac{\pi_y R - P_K R}{1 + \beta\pi_y(K) B_K(P_K)} \\ &= \nu_K(\pi_y). \end{aligned} \tag{67}$$

The first inequality in (67) follows from part (ii) of Property 4 and $\pi_x \geq_{st} \pi_y$. The second inequality in (67) holds because $\pi_y R - P_K R \leq 0$ as $\pi_y \leq_{st} P_K$.

Since $\pi \geq_{st} P_i$ for $i = 1, 2, \ldots, K-1$, (67) ensures that $\nu_K(\pi) \geq \nu_K(P_i)$ for $i = 1, 2, \ldots, K-1$. Thus, π is the PMF with the second largest Gittins index among $\{\pi, P_1, P_2, .., P_{K-1}, P_K\}$.

The Gittins index for $\pi \in \Pi P, P_{K-1} \leq_{st} \pi \leq_{st} P_K$ is given by

$$\nu(\pi) = \nu_K(\pi) = \frac{\pi R + \beta\pi(K)\frac{P_K R}{1-\beta p_{KK}}}{1 + \beta\pi(K)\frac{1}{1-\beta p_{KK}}}. \tag{68}$$

This completes the proof of (i).

(ii) If $\pi_x, \pi_y \in \Pi P$ and $P_{K-1} \leq_{st} \pi_y \leq_{st} \pi_x \leq_{st} P_K$, by (67) and (68), we get

$$\nu(\pi_y) = \nu_K(\pi_y) \leq \nu_K(\pi_x) = \nu(\pi_x). \tag{69}$$

(iii) From part (i) we know that for $\pi \in \Pi P$ and $P_{K-1} \leq_{st} \pi \leq_{st} P_K$, π gives the second largest Gittins index among $\{\pi, P_1, P_2, .., P_{K-1}, P_K\}$. Consequently, $\nu(\pi) \geq \nu(P_i)$ for $i < K$. ■

Theorem 2: Under Conditions (A1)-(A4) and $L = K$, after time $t = 0$ the Gittins index rule is an optimal channel sensing policy for Problems (P1).

Proof: Consider any time $t > 0$. If the channel observed at time $t-1$ is in state K then the PMF of that channel at t is P_K. The myopic policy senses this channel at t. The Gittins index rule senses the same channel at t as P_K is the PMF with the largest Gittins index by Theorem 2, part (ii).

If the channel observed at time $t-1$ is in state $i, i < K$, then the PMF of that channel at t is P_i and the PMFs of all other channels are stochastically ordered and are stochastically larger than P_{K-1} and stochastically smaller than P_K by Property 2. The myopic policy will choose the channel with the stochastically largest PMF (among all channels that are not observed at $t-1$). By Theorem 2 (ii), the Gittins index of the same channel is the largest among the Gittins indices of all channels that are not observed at $t-1$. By Theorem 2 (iii), the Gittins index of the channel observed at time $t-1$ is $\nu(P_i) \le \nu(\pi)$ for all $P_{K-1} \le_{st} \pi \le_{st} P_K$. Therefore, the Gittins index chooses the same channel as the myopic policy. From the optimality of the myopic policy, under Conditions (A1)-(A4) (Theorem 2) and the condition $L = K$, after time $t = 0$ the Gittins index rule is an optimal channel sensing strategy for problem (P1) and (P2). ■

Note that, if two channels, say channel 1 and 2 are such that $\pi_0^1, \pi_0^2 \in \{P_1, P_2, \ldots, P_{K-1}\}$ then $\pi_0^1, \pi_0^2 \in \Pi P$ and thus, (A2) is satisfied. Nevertheless π_0^1, π_0^2 do not necessarily satisfy the condition $P_{k-1} \le_{st} \pi_0^i \le_{st} P_K$ of Theorem 2. Thus, at $t = 0$, the assertion of Theorem 2 may not be true for channels 1 and 2, thus the Gittins index rule may not be optimal at time 0.

V. Conclusion

The channel sensing problem we investigated in this paper arises in communications and many other fields of science and technology, as it is an instance of restless bandit problems. We identified conditions sufficient to guarantee the optimality of the myopic policy, the policy that selects at each time instant the channel with the stochastically largest PMF on its states. We also identified conditions under which the Gittins index rule coincides with the myopic policy (and is optimal).

Our results on the optimality of the myopic policy extend previously existing results on the same problem when each channel has two states. As pointed out in Section III-A, such an extension is non-trivial and requires significant insight into the nature of the problem (so as to identify the appropriate assumptions/conditions), and much more careful and complicated analysis (so as to discover qualitative properties of optimal sensing policies, such as the optimality of the myopic policy).

Our results on the optimality of the Gittins index rule rely on: (i) the fact that the information state of any channel after $t > 0$ lies stochastically between P_{K-1} and P_K, i.e. $P_{K-1} \le_{st} \pi \le_{st} P_K$; and (ii) the fact that $\nu(\hat{\pi}) \ge \nu(\pi)$ whenever $\hat{\pi} \ge_{st} \pi$ and both $\hat{\pi}$ and π are stochastically ordered between P_{K-1} and P_K. We have not been able to prove whether or not the Gittins index rule coincides with the myopic policy when conditions (A1)-(A4) are valid and $L \ne K$ in (A3).

Appendix A

Proof of Property 1:

$$xP - yP = \sum_{i=2}^{K} \left[\left(\sum_{j=i}^{K} (x(j) - y(j)) \right) (P_i - P_{i-1}) \right]. \quad (70)$$

Note that $\sum_{j=i}^{K}(x(j) - y(j)) \ge 0$ since $x \ge_{st} y$. Then, by assumption (A1) $P_i \ge_{st} P_{i-1}$ we get

$$\left(\sum_{j=i}^{K} (x(j) - y(j)) \right) (P_i - P_{i-1}) \ge_{st} \mathbf{0}. \quad (71)$$

■

Proof of Property 2: From Property 1, (A1) and (A3) we obtain

$$P_iP \le_{st} P_KP \le_{st} P_L, \quad (72)$$

$$P_iP \ge_{st} P_1P \ge_{st} P_{L-1}. \quad (73)$$

Therefore, (72) and (73) give

$$P_{L-1} \le_{st} \sum_{i=1}^{K} x(i)P_iP = xP^2 \le_{st} P_L. \quad (74)$$

■

Proof of Property 3: We prove this Property by induction. The Property is true at $t = 0$ by (A2).

Assume the Property is true at t. If n, m are not selected at t, $\pi_{t+1}^n = \pi_t^n P$, $\pi_{t+1}^m = \pi_t^m P$.

By the induction hypothesis we have $\pi_t^n \le_{st} \pi_t^m$ or $\pi_t^m \le_{st} \pi_t^n$. Then from Property 1 we obtain $\pi_{t+1}^n \le_{st} \pi_{t+1}^m$ or $\pi_{t+1}^m \le_{st} \pi_{t+1}^n$.

Suppose, without loss of generality, that channel n is selected at t. Since channel m is not selected at t, $\pi_{t+1}^m = \pi_t^m P \in \Pi P^2$. Then from Property 2 we have either $\pi_{t+1}^n \le_{st} \pi_{t+1}^m$ or $\pi_{t+1}^m \le_{st} \pi_{t+1}^n$. ■

Appendix B

Proof of Property 4:

(i) Since $x \ge_{st} y$ and $\upsilon_i \ge \upsilon_{i-1}$, $i = 2, 3, \ldots, K-1$, by summation by parts we have

$$(x-y)\upsilon = \sum_{i=2}^{K} \left[\left(\sum_{j=i}^{K} (x(j) - y(j)) \right) (\upsilon_i - \upsilon_{i-1}) \right] \ge 0. \quad (75)$$

$$\text{For } i < L, U_i - U_{i-1} = R_i - R_{i-1}. \quad (76)$$

$$\begin{aligned} \text{For } i \ge L, U_i - U_{i-1} &= R_i - R_{i-1} + \beta(P_i - P_{i-1})U \\ &\ge R_i - R_{i-1}. \end{aligned} \quad (77)$$

Then, for all i, from the definition of M we obtain

$$M_i - M_{i-1} \ge U_i - U_{i-1} \ge R_i - R_{i-1} \ge 0. \quad (78)$$

Since $x \ge_{st} y$, from (78) and the result of part (i) we have

$$(x-y)M \ge (x-y)U \ge (x-y)R \ge 0. \quad (79)$$

(ii) Because of Assumption (A4) and the result of part (ii) we have:

$$\begin{aligned}\text{For } i < L, U_i - U_{i-1} &= R_i - R_{i-1}\\ &\geq \beta(P_i - P_{i-1})M\\ &\geq \beta(P_i - P_{i-1})U. &(80)\\ \text{For } i \geq L, U_i - U_{i-1} &= R_i - R_{i-1} + \beta(P_i - P_{i-1})U\\ &\geq \beta(P_i - P_{i-1})U. &(81)\end{aligned}$$

Then, (80) and (81) imply that $U - \beta PU$ is in increasing order, consequently,

$$(x-y)U \geq \beta(x-y)PU. \tag{82}$$

Since $M = U + \beta \sum_{i \geq L} p_{Ki} PU$,

$$\begin{aligned}(x-y)M &\geq \beta(x-y)PU + \beta \sum_{i \geq L} p_{Ki} \beta(xP - yP)PU\\ &= \beta(x-y)PM. &(83)\end{aligned}$$

(iii) If $x(i) = y(i)$ for all $i \geq L$, then $x(i) - y(i) = 0$ for $i \geq L$.

Define $\upsilon := (\upsilon_1, \upsilon_2, \ldots, \upsilon_K)$ such that

$$\upsilon_i = R_i - \beta P_i M, \quad \text{for } i = 1, 2, \ldots, L-1, \tag{84}$$

$$\upsilon_i = \upsilon_{L-1}, \quad \text{for } i \geq L. \tag{85}$$

From assumption (24) in (A4) we know that $\upsilon_i - \upsilon_{i-1} = R_i - R_{i-1} - \beta(P_i - P_{i-1})M \geq 0$ for $i \leq L-1$ and $\upsilon_i - \upsilon_{i-1} = 0$ for $i \geq L$. Then from the result of part (i) we obtain

$$(x-y)(R - \beta PM) = (x-y)\upsilon \geq 0. \tag{86}$$

The case where $x(i) = y(i)$ for all $i < L$ can be proved in the same way. ■

Appendix C

Proof of Property 5: We want to show that under $g_{0:T}^{O_0}$, at any time t the ordering O_t has the property that $\pi_t^{O_t(1)} \leq_{st} \pi_t^{O_t(2)} \leq_{st} \cdots \leq_{st} \pi_t^{O_t(N)}$.

At $t = 0$, by the statement of Property 5, the initial ordering O_0 is such that $\pi_0^{O_0(1)} \leq_{st} \pi_0^{O_0(2)} \leq_{st} \cdots \leq_{st} \pi_0^{O_0(N)}$.

Suppose at time t, the ordering O_t is such that $\pi_t^{O_t(1)} \leq_{st} \pi_t^{O_t(2)} \leq_{st} \cdots \leq_{st} \pi_t^{O_t(N)}$.

If the observation is $Y_t \geq L$, the new ordering is $O_{t+1} = \hat{m}(O_t, Y_t) = O_t$ and the PMFs of the channels evolves to

$$\pi_{t+1}^n = \pi_t^n P, \quad \text{for } n \neq O_t(N), \tag{87}$$

$$\pi_{t+1}^{O_t(N)} = P_{Y_t} \geq_{st} P_L. \tag{88}$$

From Properties 1 and 2 we know that

$$\begin{aligned}&\pi_t^{O_t(1)} P \leq_{st} \pi_t^{O_t(2)} P \leq_{st} \cdots \leq_{st} \pi_t^{O_t(N-1)} P\\ &\leq_{st} P_L \leq_{st} P_{Y_t}. &(89)\end{aligned}$$

On the other hand, if the observation is $Y_t < L$, the new ordering is $O_{t+1} = \hat{m}(O_t, Y_t) = SO_t$ and the PMFs of the channels become

$$\pi_{t+1}^n = \pi_t^n P, \quad \text{for } n \neq O_t(N), \tag{90}$$

$$\pi_{t+1}^{O_t(N)} = P_{Y_t} \leq_{st} P_{L-1}. \tag{91}$$

Again, from Properties 1 and 2 we get

$$\begin{aligned}&P_{Y_t} \leq_{st} P_{L-1}\\ &\leq_{st} \pi_t^{O_t(1)} P \leq_{st} \pi_t^{O_t(2)} P \leq_{st} \cdots \leq_{st} \pi_t^{O_t(N-1)} P. &(92)\end{aligned}$$

Thus, the ordering-based policy $g_{0:T}^{O_0}$ selects at any time t the channel $O_t(N)$ from the ordering O_t with $\pi_t^{O_t(1)} \leq_{st} \pi_t^{O_t(2)} \leq_{st} \cdots \leq_{st} \pi_t^{O_t(N)}$. This ordering-based policy is exactly the same as the myopic policy g^m. ■

Appendix D

For a more detailed version of this Appendix we refer the reader to [22].

We first establish a lemma that is needed for the proof of Properties 6-9.

Lemma 1: The functions $V_t(O_t, \pi_t^1, \pi_t^2, \ldots, \pi_t^N)$, $t = 1, 2, \ldots, T$ (defined by eq. (42)), are linear in every component $\pi_t^n, n = 1, 2, \ldots, N$.

That is, for all $n = 1, 2, \ldots, N$

$$\begin{aligned}&V_t(O_t, \pi_t^1, \pi_t^2, \ldots, \pi_t^N)\\ &= \sum_{i=1}^{K} \pi^n(i) V_t(O_t, \pi_t^1, \ldots, \pi_t^{n-1}, e_i, \pi_t^{n+1}, \ldots, \pi_t^N), &(93)\end{aligned}$$

where e_i is the vector with 1 in the ith position and 0 otherwise, i.e. $\underset{\uparrow\, i\text{th position}}{e_i = [0, \ldots, 0,\ 1, 0, \ldots, 0]}$.

Furthermore, $L_t(O_t, \hat{\pi}_t^1, \pi_t^1, \pi_t^2, \ldots, \pi_t^N)$ satisfies for $n = 2, 3, \ldots, N$

$$\begin{aligned}&L_t(O_t, \hat{\pi}_t^1, \pi_t^1, \pi_t^2, \ldots, \pi_t^N)\\ &= \sum_{i=1}^{K} \pi^n(i) L_t(O_t, \hat{\pi}_t^1, \pi_t^1, \ldots, \pi_t^{n-1}, e_i, \pi_t^{n+1}, \ldots, \pi_t^N), &(94)\end{aligned}$$

$$\begin{aligned}&L_t(O_t, \hat{\pi}_t^1, \pi_t^1, \pi_t^2, \ldots, \pi_t^N)\\ &= \sum_{i=1}^{K} (\hat{\pi}_t^1(i) - \pi_t^1(i)) V_t(O_t, e_i, \pi_t^2, \ldots, \pi_t^N). &(95)\end{aligned}$$

Proof: From the definition of V_t (eq. (42)) we have

$$\begin{aligned}&V_t(O_t, \pi_t^1, \pi_t^2, \ldots, \pi_t^N)\\ &= \sum_{i=1}^{K} \pi_t^n(i) E^{g_{t:T}^{O_t}} \left[\sum_{s=t}^{T} \beta^{s-t} R(s) | \pi_t^1, \pi_t^2, \ldots, \pi_t^N, X_t^n = i \right]\\ &= \sum_{i=1}^{K} \pi_t^n(i)\\ &\quad \times E^{g_{t:T}^{O_t}} \left[\sum_{s=t}^{T} \beta^{s-t} R(s) | \pi_t^1, \ldots, \pi_t^{n-1}, \pi_t^{n+1}, \ldots, \pi_t^N, \pi_t^n = e_i \right]\\ &= \sum_{i=1}^{K} \pi_t^n(i) V_t(O_t, \pi_t^1, \ldots, \pi_t^{n-1}, e_i, \pi_t^{n+1}, \ldots, \pi_t^N). &(96)\end{aligned}$$

The third equality in (96) follows from the specification of the ordering-based policy $g_{t:T}^{O_t}$ and the fact that conditional on $\{X_t^n = i, \pi_t^n\}$ the evolution of channel n is the same as that conditional on $\{\pi_t^n = e_i\}$.

Furthermore, L_t is the difference of two V_t's, so the linearity of V_t leads directly to equations (94) and (95). ■

We proceed now with the proof of Properties 6-9. In the following proof, we use the notation

$$\pi_t^{k_1:k_2} := (\pi_t^{k_1}, \pi_t^{k_1+1}, \dots, \pi_t^{k_2}), \tag{97}$$

$$\pi_t^{k_1:k_2} P := (\pi_t^{k_1} P, \pi_t^{k_1+1} P, \dots, \pi_t^{k_2} P). \tag{98}$$

Proof of Properties 6-9: First note that Property 8 is a special case of Property 7. This can be seen as follows.

Without loss of generality, let $O_t(n) = 1, O_t(m) = 2$, and $\pi_t^1 \geq_{st} \pi_t^2$. Note that

$$V_t(O_t, \pi_t^2, \pi_t^2, \dots, \pi_t^N) = V_t(W_{nm} O_t, \pi_t^2, \pi_t^2, \dots, \pi_t^N). \tag{99}$$

Applying Property 7 at time t, we have

$$\begin{aligned}
&V_t(O_t, \pi_t^1, \pi_t^2, \dots, \pi_t^N) - V_t(W_{nm} O_t, \pi_t^1, \pi_t^2, \dots, \pi_t^N)\\
&= V_t(O_t, \pi_t^1, \pi_t^2, \dots, \pi_t^N) - V_t(O_t, \pi_t^2, \pi_t^2, \dots, \pi_t^N)\\
&\quad + V_t(W_{nm} O_t, \pi_t^2, \pi_t^2, \dots, \pi_t^N)\\
&\quad - V_t(W_{nm} O_t, \pi_t^1, \pi_t^2, \dots, \pi_t^N)\\
&= L_t(O_t, \pi_t^1, \pi_t^2, \pi_t^2, \dots, \pi_t^N)\\
&\quad - L_t(W_{nm} O_t, \pi_t^1, \pi_t^2, \pi_t^2, \dots, \pi_t^N) \geq 0.
\end{aligned} \tag{100}$$

Therefore, Property 8 is true at time t once Property 7 is true at time t.

We prove all three Properties 6, 7 and 9 simultaneously by induction.

For both the basis of induction and the induction we consider two cases.

(i) When channel 1 is not the right-most channel in O_t (i.e. $n \neq N$ and $O_t(N) \neq 1$).

(ii) When channel 1 is the right-most channel in O_t (i.e. $n = N$ and $O_t(N) = 1$).

Basis of induction

It can be verified that Properties 6, 7 and 9 are true at time $t = T$. For details see [22].

Induction hypothesis

Assume that the assertions of Properties 6, 7 and 9 are true for time $t+1, t+2, \dots, T$.

Induction step

We prove here Properties 6, 7 and 9 for t.

We first develop five expressions (105), (107), (108), (109) and (112) for L_t and L_{t+1}, defined by eq. (46), that will be useful in the sequel.

For any PMF $\pi \in \Pi$ we define

$$\underline{\pi} := (\pi(1), \pi(2), \dots, \pi(L-2), \sum_{i=L-1}^{K} \pi(i), 0, \dots, 0), \tag{101}$$

$$\bar{\pi} := (0, \dots, 0, \sum_{i=1}^{L} \pi(i), \pi(L+1), \dots, \pi(K)). \tag{102}$$

Then, $\underline{\pi}, \bar{\pi} \in \Pi$, and

$$\pi = \underline{\pi} + \bar{\pi} - e_L + \sum_{i=L}^{K} \pi(i)(e_L - e_{L-1}). \tag{103}$$

Furthermore, if $\hat{\pi} \geq_{st} \pi$, it follows that

$$\underline{\hat{\pi}} \geq_{st} \underline{\pi}, \quad \bar{\hat{\pi}} \geq_{st} \bar{\pi}. \tag{104}$$

Consider any arbitrary ordering $O \in \mathcal{O}$. When $O(N) \neq 1$, assume $O(N) = 2$ without any loss of generality. Then,

$$\begin{aligned}
&L_t(O, \hat{\pi}_t^1, \pi_t^1, \pi_t^{2:N})\\
&= (\pi_t^2 R - \pi_t^2 R) + \beta \sum_{i<L} \pi_t^2(i)(V_{t+1}(SO, \hat{\pi}_t^1 P, P_i, \pi_t^{3:N} P)\\
&\quad - V_{t+1}(SO, \pi_t^1 P, P_i, \pi_t^{3:N} P))\\
&\quad + \beta \sum_{i \geq L} \pi_t^2(i)(V_{t+1}(O, \hat{\pi}_t^1 P, P_i, \pi_t^{3:N} P)\\
&\quad - V_{t+1}(O, \pi_t^1 P, P_i, \pi_t^{3:N} P))\\
&= \beta \sum_{i<L} \pi_t^2(i) L_{t+1}(SO, \hat{\pi}_t^1 P, \pi_t^1 P, P_i, \pi_t^{3:N} P)\\
&\quad + \beta \sum_{i \geq L} \pi_t^2(i) L_{t+1}(O, \hat{\pi}_t^1 P, \pi_t^1 P, P_i, \pi_t^{3:N} P).
\end{aligned} \tag{105}$$

Furthermore, by the induction hypothesis for Property 6, we get, for all $i = 1, 2, \dots, K$,

$$\begin{aligned}
&L_{t+1}(SO, \hat{\pi}_t^1 P, \pi_t^1 P, P_i, \pi_t^{3:N} P)\\
&\qquad \geq L_{t+1}(O, \hat{\pi}_t^1 P, \pi_t^1 P, P_i, \pi_t^{3:N} P).
\end{aligned} \tag{106}$$

Therefore,

$$\begin{aligned}
&\beta L_{t+1}(SO, \hat{\pi}_t^1 P, \pi_t^{1:N} P)\\
&= \beta \sum_{i=1}^{L} \pi_t^2(i) L_{t+1}(SO, \hat{\pi}_t^1 P, \pi_t^1 P, P_i, \pi_t^{3:N} P)\\
&\geq L_t(O, \hat{\pi}_t^1, \pi_t^{1:N}).
\end{aligned} \tag{107}$$

$$\begin{aligned}
&L_t(O, \hat{\pi}_t^1, \pi_t^{1:N})\\
&\geq \beta \sum_{i=1}^{L} \pi_t^2(i) L_{t+1}(O, \hat{\pi}_t^1 P, \pi_t^1 P, P_i, \pi_t^{3:N} P)\\
&= \beta L_{t+1}(O, \hat{\pi}_t^1 P, \pi_t^{1:N} P).
\end{aligned} \tag{108}$$

When $O(N) = 1$,

$$\begin{aligned}
&L_t(O_t, \hat{\pi}_t^1, \pi_t^{1:N})\\
&:= V_t(O_t, \hat{\pi}_t^1, \pi_t^{2:N}) - V_t(O_t, \pi_t^1, \pi_t^{2:N})\\
&= (\hat{\pi}_t^1 R - \pi_t^1 R) + \beta \sum_{i<L} (\hat{\pi}_t^1(i) - \pi_t^1(i)) V_{t+1}(SO_t, P_i, \pi_t^{2:N} P)\\
&\quad + \beta \sum_{i \geq L} (\hat{\pi}_t^1(i) - \pi_t^1(i)) V_{t+1}(O_t, P_i, \pi_t^{2:N} P)\\
&= (\hat{\pi}_t^1 - \pi_t^1) R + \beta L_{t+1}(SO_t, \underline{\hat{\pi}}_t^1 P, \underline{\pi}_t^1 P, \pi_t^{2:N} P)\\
&\quad + \beta L_{t+1}(O_t, \bar{\hat{\pi}}_t^1 P, \bar{\pi}_t^1 P, \pi_t^{2:N} P)\\
&\quad + \beta \left[V_{t+1}(O_t, P_L, \pi_t^{2:N} P) - V_{t+1}(SO_t, P_{L-1}, \pi_t^{2:N} P)\right]\\
&\quad \times \left[\sum_{i=L}^{K} (\hat{\pi}_t^1(i) - \pi_t^1(i))\right].
\end{aligned} \tag{109}$$

The last equality in (109) follows from the linearity of L_t (Lemma 1) and the definition of $\underline{\pi}, \bar{\pi}$ given by (101)-(102).

Furthermore, using (109) we get

$$\begin{aligned}
&L_t(O,\hat{\pi}_t^1,\pi_t^{1:N})-\beta L_{t+1}(SO,\hat{\pi}_t^1P,\pi_t^{1:N}P)\\
&=(\hat{\pi}_t^1-\pi_t^1)R+\beta L_{t+1}(SO,\underline{\hat{\pi}}_t^1P,\underline{\pi}_t^1P,\pi_t^{2:N}P)\\
&\quad+\beta L_{t+1}(O,\bar{\hat{\pi}}_t^1P,\bar{\pi}_t^1P,\pi_t^{2:N}P)\\
&\quad+\beta\left[V_{t+1}(O,P_L,\pi_t^{2:N}P)-V_{t+1}(SO,P_{L-1},\pi_t^{2:N}P)\right]\\
&\quad\times\left[\sum_{i=L}^{K}(\hat{\pi}_t^1(i)-\pi_t^1(i))\right]\\
&\quad-\beta L_{t+1}(SO,\hat{\pi}_t^1P,\pi_t^1P,\pi_t^{2:N}P)\\
&=(\hat{\pi}_t^1-\pi_t^1)R+\beta L_{t+1}(O,\bar{\hat{\pi}}_t^1P,\bar{\pi}_t^1P,\pi_t^{2:N}P)\\
&\quad-\beta L_{t+1}(SO,\bar{\hat{\pi}}_t^1P,\bar{\pi}_t^1P,\pi_t^{2:N}P)\\
&\quad+\beta\left[V_{t+1}(O,P_L,\pi_t^{2:N}P)-V_{t+1}(SO,P_L,\pi_t^{2:N}P)\right]\\
&\quad\times\left[\sum_{i=L}^{K}(\hat{\pi}_t^1(i)-\pi_t^1(i))\right]\\
&\le(\hat{\pi}_t^1-\pi_t^1)R+\beta(\bar{\hat{\pi}}_t^1-\bar{\pi}_t^1)PU\\
&\quad+\beta\left[V_{t+1}(O,P_L,\pi_t^{2:N}P)-V_{t+1}(SO,P_L,\pi_t^{2:N}P)\right]\\
&\quad\times\left[\sum_{i=L}^{K}(\hat{\pi}_t^1(i)-\pi_t^1(i))\right].
\end{aligned}\tag{110}$$

The second equality in (110) follows from (103) and the linearity of L_t (Lemma 1). The inequality in (110) follows from the induction hypothesis for the upper bound of Property 6 at $t+1$ and the fact that $\bar{\hat{\pi}}_t^1P\ge_{st}\bar{\pi}_t^1P$.

For the last term in (110), note that

$$\begin{aligned}
&V_{t+1}(O,P_L,\pi_t^{2:N}P)-V_{t+1}(SO,P_L,\pi_t^{2:N}P)\\
&=L_{t+1}(O,P_L,P_{L-1},\pi_t^{2:N}P)-L_{t+1}(SO,P_L,P_{L-1},\pi_t^{2:N}P)\\
&\quad+V_{t+1}(O,P_{L-1},\pi_t^{2:N}P)\\
&\quad-V_{t+1}(W_{12}\cdots W_{(N-1)(N-2)}W_{N(N-1)}O,P_{L-1},\pi_t^{2:N}P)\\
&\le L_{t+1}(O,P_L,P_{L-1},\pi_t^{2:N}P)-L_{t+1}(SO,P_L,P_{L-1},\pi_t^{2:N}P)\\
&\le(P_L-P_{L-1})U.
\end{aligned}\tag{111}$$

The equality in (111) follows from the definition of L_{t+1} and the fact that $SO=W_{12}\cdots W_{(N-1)(N-2)}W_{N(N-1)}O$. The inequalities in (111) follow by the induction hypothesis for Property 8 and Property 6 at $t+1$.

Therefore, using (111) and (110) we get

$$\begin{aligned}
&L_t(O,\hat{\pi}_t^1,\pi_t^1,\pi_t^{2:N})-\beta L_{t+1}(SO,\hat{\pi}_t^1P,\pi_t^1P,\pi_t^{2:N}P)\\
&\le(\hat{\pi}_t^1-\pi_t^1)R+\beta(\bar{\hat{\pi}}_t^1-\bar{\pi}_t^1)PU\\
&\quad+\beta(P_L-P_{L-1})U\sum_{i=L}^{K}(\hat{\pi}_t^1(i)-\pi_t^1(i))\\
&=(\hat{\pi}_t^1-\pi_t^1)U.
\end{aligned}\tag{112}$$

The last equality in (112) follows from the definition of the vector U.

Induction step for Property 6:

We first consider the lower bound of Property 6.

(i) When $O_t(N)\neq 1$ (i.e. $n\neq N$), we also have $S^{-m}O_t(N)=O_t(m)\neq 1$. Then,

$$\begin{aligned}
L_t(O_t,\hat{\pi}_t^1,\pi_t^{1:N})&\ge\beta L_{t+1}(O_t,\hat{\pi}_t^1P,\pi_t^{1:N}P)\\
&\ge\beta L_{t+1}(S^{1-m}O_t,\hat{\pi}_t^1P,\pi_t^{1:N}P)\\
&\ge L_t(S^{-m}O_t,\hat{\pi}_t^1,\pi_t^{1:N}).
\end{aligned}\tag{113}$$

The first inequality in (113) follows from (107) and the fact that $O_t(N)\neq 1$. The second inequality in (113) follows from the induction hypothesis for Property 6 at $t+1$. The last inequality in (113) follows from (108) and the fact that $S^{-m}O_t(N)\neq 1$.

(ii) When $O_t(N)=1$ (i.e. $n=N$).

Since $S^{-m}O_t(N)=O_t(m)\neq 1$, from (108) we get

$$\begin{aligned}
&L_t(S^{-m}O_t,\hat{\pi}_t^1,\pi_t^{1:N})\\
&\le\beta L_{t+1}(S^{1-m}O_t,\hat{\pi}_t^1P,\pi_t^{1:N}P)\\
&=\beta L_{t+1}(S^{1-m}O_t,\underline{\hat{\pi}}_t^1P,\underline{\pi}_t^1P,\pi_t^{2:N}P)\\
&\quad+\beta L_{t+1}(S^{1-m}O_t,\bar{\hat{\pi}}_t^1P,\bar{\pi}_t^1P,\pi_t^{2:N}P)\\
&\quad+\beta\sum_{i=L}^{K}(\hat{\pi}_t^1(i)-\pi_t^1(i))L_{t+1}(S^{1-m}O_t,P_L,P_{L-1},\pi_t^{2:N}P).
\end{aligned}\tag{114}$$

Since $O_t(N)=1$, applying (109) we obtain

$$\begin{aligned}
&L_t(O_t,\hat{\pi}_t^1,\pi_t^{1:N})-L_t(S^{-m}O_t,\hat{\pi}_t^1,\pi_t^{1:N})\\
&=(\hat{\pi}_t^1-\pi_t^1)R+\beta L_{t+1}(SO_t,\underline{\hat{\pi}}_t^1P,\underline{\pi}_t^1P,\pi_t^{2:N}P)\\
&\quad+\beta L_{t+1}(O_t,\bar{\hat{\pi}}_t^1P,\bar{\pi}_t^1P,\pi_t^{2:N}P)\\
&\quad+\beta\left[V_{t+1}(O_t,P_L,\pi_t^{2:N}P)-V_{t+1}(SO_t,P_{L-1},\pi_t^{2:N}P)\right]\\
&\qquad\times\sum_{i=L}^{K}(\hat{\pi}_t^1(i)-\pi_t^1(i))-L_t(S^{-m}O_t,\hat{\pi}_t^1,\pi_t^1,\pi_t^{2:N})\\
&\ge(\hat{\pi}_t^1-\pi_t^1)R+\beta L_{t+1}(SO_t,\underline{\hat{\pi}}_t^1P,\underline{\pi}_t^1P,\pi_t^{2:N}P)\\
&\quad-\beta L_{t+1}(S^{1-m}O_t,\underline{\hat{\pi}}_t^1P,\underline{\pi}_t^1P,\pi_t^{2:N}P)\\
&\quad+\beta L_{t+1}(O_t,\bar{\hat{\pi}}_t^1P,\bar{\pi}_t^1P,\pi_t^{2:N}P)\\
&\quad-\beta L_{t+1}(S^{1-m}O_t,\bar{\hat{\pi}}_t^1P,\bar{\pi}_t^1P,\pi_t^{2:N}P)\\
&\quad+\beta\Big[V_{t+1}(O_t,P_L,\pi_t^{2:N}P)-V_{t+1}(SO_t,P_{L-1},\pi_t^{2:N}P)\\
&\qquad-L_{t+1}(S^{1-m}O_t,P_L,P_{L-1},\pi_t^{2:N}P)\Big]\sum_{i=L}^{K}(\hat{\pi}_t^1(i)-\pi_t^1(i))\\
&\ge(\hat{\pi}_t^1-\pi_t^1)R+\beta L_{t+1}(SO_t,\underline{\hat{\pi}}_t^1P,\underline{\pi}_t^1P,\pi_t^{2:N}P)\\
&\quad-\beta L_{t+1}(S^{1-m}O_t,\underline{\hat{\pi}}_t^1P,\underline{\pi}_t^1P,\pi_t^{2:N}P)\\
&\quad+\beta\Big[V_{t+1}(O_t,P_L,\pi_t^{2:N}P)-V_{t+1}(SO_t,P_{L-1},\pi_t^{2:N}P)\\
&\qquad-L_{t+1}(S^{1-m}O_t,P_L,P_{L-1},\pi_t^{2:N}P)\Big]\sum_{i=L}^{K}(\hat{\pi}_t^1(i)-\pi_t^1(i)).
\end{aligned}\tag{115}$$

The equality in (115) follows from (109) and the fact that $O_t(N)=1$. The first inequality in (115) follows from (114). The second inequality in (115) follows from the induction hypothesis for Property 6 at $t+1$.

Letting $\underline{O}_{t+1}:=S^{1-m}O_t$ and $\underline{n}:=N+1-m,\underline{m}:=N-m$, we have $\underline{m}<\underline{n}$ and

$$\underline{O}_{t+1}(\underline{n})=S^{1-m}O_t(\underline{n})=1,\quad SO_t=S^{-(\underline{m})}\underline{O}_{t+1}.\tag{116}$$

Consequently, the induction hypothesis for the upper bound of Property 6 at $t+1$ gives

$$\begin{aligned} &L_{t+1}(S^{1-m}O_t, \underline{\hat{\pi}}_t^1 P, \underline{\pi}_t^1 P, \pi_t^{2:N} P) \\ &-L_{t+1}(SO_t, \underline{\hat{\pi}}_t^1 P, \underline{\pi}_t^1 P, \pi_t^{2:N} P) \\ =\ &L_{t+1}(\underline{O}_{t+1}, \underline{\hat{\pi}}_t^1 P, \underline{\pi}_t^1 P, \pi_t^{2:N} P) \\ &-L_{t+1}(S^{-\underline{(m)}}\underline{O}_{t+1}, \underline{\hat{\pi}}_t^1 P, \underline{\pi}_t^1 P, \pi_t^{2:N} P) \\ \leq\ &(\underline{\hat{\pi}}_t^1 P - \underline{\pi}_t^1 P)U. \end{aligned} \tag{117}$$

Letting $\underline{m}' := 1$, we have $\underline{m}' < n = N$ and $A_{\underline{m}'n}O_t = SO_t$. Therefore,

$$\begin{aligned} &V_{t+1}(SO_t, P_{L-1}, \pi_t^{2:N} P) - V_{t+1}(O_t, P_L, \pi_t^{2:N} P) \\ &+L_{t+1}(S^{1-m}O_t, P_L, P_{L-1}, \pi_t^{2:N} P) \\ \leq\ &V_{t+1}(SO_t, P_{L-1}, \pi_t^{2:N} P) - V_{t+1}(O_t, P_L, \pi_t^{2:N} P) \\ &+L_{t+1}(O_t, P_L, P_{L-1}, \pi_t^{2:N} P) \\ =\ &V_{t+1}(A_{\underline{m}'n}O_t, P_{L-1}, \pi_t^{2:N} P) - V_{t+1}(O_t, P_{L-1}, \pi_t^{2:N} P) \\ \leq\ &h - P_{L-1}R. \end{aligned} \tag{118}$$

The first inequality in (118) follows from the induction hypothesis for the lower bound of Property 6 at $t+1$. The last inequality in (118) follows from the induction hypothesis for Property 9 at $t+1$.

Using (117) and (118) in (115) we obtain

$$\begin{aligned} &L_t(O_t, \hat{\pi}_t^1, \pi_t^{1:N}) - L_t(S^{-m}O_t, \hat{\pi}_t^1, \pi_t^{1:N}) \\ \geq\ &(\hat{\pi}_t^1 - \pi_t^1)R - \beta(\underline{\hat{\pi}}_t^1 P - \underline{\pi}_t^1 P)U \\ &-\beta\sum_{i=L}^{K}(\hat{\pi}_t^1(i) - \pi_t^1(i))(h - P_{L-1}R) \\ =\ &(\underline{\hat{\pi}}_t^1 - \underline{\pi}_t^1)(R - \beta U) + (\bar{\hat{\pi}}_t^1 - \bar{\pi}_t^1)R \\ &+\sum_{i=L}^{K}(\hat{\pi}_t^1(i) - \pi_t^1(i))(R_L - R_{L-1} - \beta(h - P_{L-1}R)) \\ \geq\ &0. \end{aligned} \tag{119}$$

The last inequality in (119) is true because: the terms $(\underline{\hat{\pi}}_t^1 - \underline{\pi}_t^1)(R - \beta U)$ and $(\bar{\hat{\pi}}_t^1 - \bar{\pi}_t^1)R$ are positive by parts (iv) and (ii) of Property 4; the term $(R_L - R_{L-1} - \beta(h - P_{L-1}R))$ is positive by Condition (A4).

The proof of the lower bound of Property 6 is now complete.

Now consider the upper bound of Property 6.

Let $O_t' := S^{N-n}O_t$;, then $O_t'(N) = 1$ and $SO_t'(1) = 1$. Consequently,

$$\begin{aligned} &L_t(O_t, \hat{\pi}_t^1, \pi_t^{1:N}) - L_t(S^{-m}O_t, \hat{\pi}_t^1, \pi_t^{1:N}) \\ \leq\ &L_t(O_t', \hat{\pi}_t^1, \pi_t^{1:N}) - L_t(SO_t', \hat{\pi}_t^1, \pi_t^{1:N}) \\ \leq\ &L_t(O_t', \hat{\pi}_t^1, \pi_t^{1:N}) - \beta L_{t+1}(SO_t', \hat{\pi}_t^1 P, \pi_t^{1:N} P) \\ \leq\ &(\hat{\pi}_t^1 - \pi_t^1)U. \end{aligned} \tag{120}$$

The first inequality in (120) is true because of the lower bound of Property 6 at t. The second inequality in (120) follows from (108) and the fact that $SO_t'(N) \neq 1$. The third inequality in (120) follows from (112) and the fact that $O_t'(N) = 1$.

This completes the proof of Property 7 at time t.

Induction step for Property 7:

(i) When $O_t(N) \neq 1$ (i.e. $n \neq N$), assume $O_t(N) = 2$ without loss of generality. Then because of (105),

$$\begin{aligned} &L_t(O_t, \hat{\pi}_t^1, \pi_t^{1:N}) - L_t(W_{nm}O_t, \hat{\pi}_t^1, \pi_t^{1:N}) \\ =\ &\beta\sum_{i<L}\pi^2(i)\Big[L_{t+1}(SO_t, \hat{\pi}_t^1 P, \pi_t^1 P, P_i, \pi_t^{3:N} P) \\ &-L_{t+1}(W_{(n+1)(m+1)}(SO_t), \hat{\pi}_t^1 P, \pi_t^1 P, P_i, \pi_t^{3:N} P)\Big] \\ &+\beta\sum_{i\geq L}\pi^2(i)\Big[L_{t+1}(O_t, \hat{\pi}_t^1 P, \pi_t^1 P, P_i, \pi_t^{3:N} P) \\ &-L_{t+1}(W_{nm}O_t, \hat{\pi}_t^1 P, \pi_t^1 P, P_i, \pi_t^{3:N} P)\Big]. \end{aligned} \tag{121}$$

By the induction hypothesis for Property 7, each term in (121) is positive and smaller than $(\hat{\pi}_t^1 P - \pi_t^1 P)M$. Thus,

$$\begin{aligned} 0 &\leq L_t(O_t, \hat{\pi}_t^1, \pi_t^{1:N}) - L_t(W_{nm}O_t, \hat{\pi}_t^1, \pi_t^{1:N}) \\ &\leq \beta(\hat{\pi}_t^1 P - \pi_t^1 P)M \leq (\hat{\pi}_t^1 - \pi_t^1)M. \end{aligned} \tag{122}$$

The last inequality in (122) holds by part (iii) of Property 4.

(ii) $O_t(N) = 1$ (i.e. $n = N$).

We first consider the lower-bound. Using (103) and the linearity of L_t (Lemma 1) we get

$$\begin{aligned} &L_t(O_t, \hat{\pi}_t^1, \pi_t^{1:N}) - L_t(W_{Nm}O_t, \hat{\pi}_t^1, \pi_t^{1:N}) \\ =\ &L_t(O_t, \underline{\hat{\pi}}_t^1, \underline{\pi}_t^1, \pi_t^{2:N}) - L_t(W_{Nm}O_t, \underline{\hat{\pi}}_t^1, \underline{\pi}_t^1, \pi_t^{2:N}) \\ &+L_t(O_t, \bar{\hat{\pi}}_t^1, \hat{\pi}_t^1, \pi_t^{2:N}) - L_t(W_{Nm}O_t, \bar{\hat{\pi}}_t^1, \hat{\pi}_t^1, \pi_t^{2:N}) \\ &+\Big[L_t(O_t, e_L, e_{L-1}, \pi_t^{2:N}) - L_t(W_{Nm}O_t, e_L, e_{L-1}, \pi_t^{2:N})\Big] \\ &\times\left[\sum_{i=L}^{K}(\hat{\pi}_t^1(i) - \pi_t^1(i))\right]. \end{aligned} \tag{123}$$

We consider each of the terms

(a) $L_t(O_t, \underline{\hat{\pi}}_t^1, \underline{\pi}_t^1, \pi_t^{2:N}) - L_t(W_{Nm}O_t, \underline{\hat{\pi}}_t^1, \underline{\pi}_t^1, \pi_t^{2:N})$.
(b) $L_t(O_t, \bar{\hat{\pi}}_t^1, \hat{\pi}_t^1, \pi_t^{2:N}) - L_t(W_{Nm}O_t, \bar{\hat{\pi}}_t^1, \hat{\pi}_t^1, \pi_t^{2:N})$.
(c) $[L_t(O_t, e_L, e_{L-1}, \pi_t^{2:N}) - L_t(W_{Nm}O_t, e_L, e_{L-1}, \pi_t^{2:N})]$ $[\sum_{i=L}^{K}(\hat{\pi}_t^1(i) - \pi_t^1(i))]$.

that appear in the right hand side of (123) separately.

(a) Consider the first term.

Let $O_t' = S(W_{Nm}O_t) = W_{1m+1}(SO_t)$, then $O_t'(m+1) = 1$ and $W_{m+1,1}O_t' = SO_t$. Therefore,

$$\begin{aligned} &L_t(O_t, \underline{\hat{\pi}}_t^1, \underline{\pi}_t^1, \pi_t^{2:N}) - L_t(W_{Nm}O_t, \underline{\hat{\pi}}_t^1, \underline{\pi}_t^1, \pi_t^{2:N}) \\ =\ &(\underline{\hat{\pi}}_t^1 - \underline{\pi}_t^1)R + \beta L_{t+1}(SO_t, \underline{\hat{\pi}}_t^1 P, \underline{\pi}_t^1 P, \pi_t^{2:N} P) \\ &-L_t(W_{Nm}O_t, \underline{\hat{\pi}}_t^1, \underline{\pi}_t^1, \pi_t^{2:N}) \\ \geq\ &(\underline{\hat{\pi}}_t^1 - \underline{\pi}_t^1)R + \beta L_{t+1}(SO_t, \underline{\hat{\pi}}_t^1 P, \underline{\pi}_t^1 P, \pi_t^{2:N} P) \\ &-\beta L_{t+1}(S(W_{Nm}O_t), \underline{\hat{\pi}}_t^1 P, \underline{\pi}_t^1 P, \pi_t^{2:N} P) \\ \geq\ &(\underline{\hat{\pi}}_t^1 - \underline{\pi}_t^1)R - \beta(\underline{\hat{\pi}}_t^1 P - \underline{\pi}_t^1 P)M \geq 0. \end{aligned} \tag{124}$$

The first equality in (124) follows from (109) and that fact that $\underline{\hat{\pi}}_t^1(i) = \underline{\pi}_t^1(i) = 0$ for $i \geq L$. The first inequality in (124) follows from (107). The second inequality in (124) follows from the induction hypothesis for the upper bound of Property 7 at $t+1$. The last inequality in (124) holds by part (iv) of Property 4.

(b) Consider the second term. From the induction hypothesis for Property 6 at $t+1$ and the fact that $SO_t = S^{-(N-1)}O_t$

and $O_t(N) = 1$, we obtain.

$$\begin{aligned}
&L_t(O_t, \bar{\hat{\pi}}_t^1, \bar{\pi}_t^1, \pi_t^{2:N}) - L_t(W_{Nm}O_t, \bar{\hat{\pi}}_t^1, \bar{\pi}_t^1, \pi_t^{2:N}) \\
&= (\bar{\hat{\pi}}_t^1 - \bar{\pi}_t^1)R + \beta L_{t+1}(O_t, \bar{\hat{\pi}}_t^1 P, \bar{\pi}_t^1 P, \pi_t^{2:N} P) \\
&\quad -L_t(W_{Nm}O_t, \bar{\hat{\pi}}_t^1, \bar{\pi}_t^1, \pi_t^{2:N}) \\
&\geq (\bar{\hat{\pi}}_t^1 - \bar{\pi}_t^1)R + \beta L_{t+1}(SO_t, \bar{\hat{\pi}}_t^1 P, \bar{\pi}_t^1 P, \pi_t^{2:N} P) \\
&\quad -L_t(W_{Nm}O_t, \bar{\hat{\pi}}_t^1, \bar{\pi}_t^1, \pi_t^{2:N}). \qquad (125)
\end{aligned}$$

Then, similar to the first term (a), the second term is positive.

(c) Consider the third term.

Assume $O_t(m) = 2$ without any loss of generality. Then $W_{Nm}O_t(N) = 2$. Therefore,

$$\begin{aligned}
&L_t(O_t, e_L, e_{L-1}, \pi_t^{2:N}) - L_t(W_{Nm}O_t, e_L, e_{L-1}, \pi_t^{2:N}) \\
&= R_L - R_{L-1} + \beta \sum_{i<L} \pi^2(i) \\
&\quad \times \Big[V_{t+1}(O_t, P_L, P_i, \pi_t^{3:N} P) - V_{t+1}(SO_t, P_{L-1}, P_i, \pi_t^{3:N} P) \\
&\quad -L_{t+1}(SW_{Nm}O_t, P_L, P_{L-1}, P_i, \pi_t^{3:N} P)\Big] \\
&\quad +\beta \sum_{i\geq L} \pi^2(i) \\
&\quad \times \Big[V_{t+1}(O_t, P_L, P_i, \pi_t^{3:N} P) - V_{t+1}(SO_t, P_{L-1}, P_i, \pi^{3:N}{}_t P) \\
&\quad -L_{t+1}(W_{Nm}O_t, P_L, P_{L-1}, P_i, \pi_t^{3:N} P)\Big]. \qquad (126)
\end{aligned}$$

Let $O_t' := S(W_{Nm}O_t) = W_{1m+1}(SO_t)$; then $O_t'(m+1) = 1$ and $W_{m+1,1}O_t' = SO_t$.

For each term in the first sum in (126), we have $P_{L-1} \geq_{st} P_i$ ($i < L$ in the first sum in (126)). Therefore, by the induction hypothesis for Property 8 at $t+1$ we get

$$\begin{aligned}
&V_{t+1}(O_t, P_L, P_i, \pi_t^{3:N} P) - V_{t+1}(SO_t, P_{L-1}, P_i, \pi_t^{3:N} P) \\
&\quad -L_{t+1}(SW_{Nm}O_t, P_L, P_{L-1}, P_i, \pi_t^{3:N} P) \\
&\geq V_{t+1}(O_t, P_L, P_i, \pi_t^{3:N} P) - V_{t+1}(O_t', P_L, P_i, \pi_t^{3:N} P). \quad (127)
\end{aligned}$$

Furthermore, since $P_L \geq_{st} \pi_t^{O_t(l)} P$ for all $l = 1, 2, \ldots, N$ by Property 2, repeatedly applying Property 8 at $t+1$ we obtain

$$\begin{aligned}
&V_{t+1}(O_t, P_L, P_i, \pi_t^{3:N} P) \\
&\geq V_{t+1}(W_{(m+2)(m+1)} \cdots W_{N(N-1)} O_t, P_L, P_i, \pi_t^{3:N} P). \quad (128)
\end{aligned}$$

Note that $W_{(m+2)(m+1)} \cdots W_{N(N-1)} O_t = A_{N(m+1)} O_t$ and $A_{m1}(A_{N(m+1)} O_t) = S(W_{Nm}O_t) = O_t'$. Consequently, the induction hypothesis for Property 9 at $t+1$ gives

$$\begin{aligned}
&V_{t+1}(W_{(m+2)(m+1)} \cdots W_{N(N-1)} O_t, P_L, P_i, \pi_t^{3:N} P) \\
&\quad -V_{t+1}(O_t', P_L, P_i, \pi_t^{3:N} P) \\
&= V_{t+1}(A_{N(m+)1} O_t, P_L, P_i, \pi_t^{3:N} P) \\
&\quad -V_{t+1}(A_{m1}(A_{N(m+1)} O_t), P_L, P_i, \pi_t^{3:N} P) \\
&\geq -(h - P_i P^{N-m} R). \qquad (129)
\end{aligned}$$

For each term in the second sum in (126), we have

$$\begin{aligned}
&V_{t+1}(O_t, P_L, P_i, \pi_t^{3:N} P) - V_{t+1}(SO_t, P_{L-1}, P_i, \pi_t^{3:N} P) \\
&\quad -L_{t+1}(W_{Nm}O_t, P_L, P_{L-1}, P_i, \pi_t^{3:N} P) \\
&\geq V_{t+1}(O_t, P_L, P_i, \pi_t^{3:N} P) - V_{t+1}(SO_t, P_{L-1}, P_i, \pi_t^{3:N} P) \\
&\quad -L_{t+1}(O_t, P_L, P_{L-1}, P_i, \pi_t^{3:N} P) \\
&= V_{t+1}(O_t, P_{L-1}, P_i, \pi_t^{3:N} P) - V_{t+1}(SO_t, P_{L-1}, P_i, \pi_t^{3:N} P) \\
&\geq -(h - P_{L-1} R). \qquad (130)
\end{aligned}$$

The inequalities in (130) follows from the induction hypothesis at $t+1$ for the lower bound of Property 7 and Property 9 respectively.

Using the lower bounds provided by (129) and (130) for terms in (126), we obtain

$$\begin{aligned}
&L_t(O_t, e_L, e_{L-1}, \pi_t^{2:N}) - L_t(W_{Nm}O_t, e_L, e_{L-1}, \pi_t^{2:N}) \\
&\geq R_L - R_{L-1} - \beta \sum_{i<L} \pi_t^2(i)(h - P_i P^{N-m} R) \\
&\quad -\beta \sum_{i\geq L} \pi_t^2(i)(h - P_{L-1} R) \\
&\geq R_L - R_{L-1} - \beta(h - P_{L-1} R) \geq 0. \qquad (131)
\end{aligned}$$

The second and the last inequalities in (131) follows from part (ii) of Property 4 and condition (A**??**) respectively.

Since the three terms (a), (b) and (c) in (123) are positive, the proof for the lower bound of Property 7 is complete when $O_t(N) = 1$ (case (ii)).

We now proceed to establish the upper bound of Property 7 when $O_t(N) = 1$ (case (ii)).

Assume $O_t(m) = 2$ without any loss of generality; then $W_{Nm}O_t(N) = 2$. Therefore,

$$\begin{aligned}
&L_t(O_t, \hat{\pi}_t^1, \pi_t^{1:N}) - L_t(W_{Nm}O_t, \hat{\pi}_t^1, \pi_t^{1:N}) \\
&= L_t(O_t, \hat{\pi}_t^1, \pi_t^{1:N}) - \beta L_{t+1}(SO_t, \hat{\pi}_t^1 P, \pi_t^{1:N} P) \\
&\quad +\beta L_{t+1}(SO_t, \hat{\pi}_t^1 P, \pi_t^{1:N} P) - L_t(W_{Nm}O_t, \hat{\pi}_t^1, \pi_t^{1:N}) \\
&\leq (\hat{\pi}_t^1 - \pi_t^1)U + \beta L_{t+1}(SO_t, \hat{\pi}_t^1 P, \pi_t^{1:N} P) \\
&\quad -L_t(W_{Nm}O_t, \hat{\pi}_t^1, \pi_t^{1:N}) \\
&\leq (\hat{\pi}_t^1 - \pi_t^1)U + \beta L_{t+1}(S(W_{Nm}O_t), \hat{\pi}_t^1 P, \pi_t^{1:N} P) \\
&\quad -L_t(W_{Nm}O_t, \hat{\pi}_t^1, \pi_t^{1:N}) \\
&= (\hat{\pi}_t^1 - \pi_t^1)U \\
&\quad +\beta \sum_{i\geq L} \pi_t^2(i) \Big[L_{t+1}(S(W_{Nm}O_t), \hat{\pi}_t^1 P, \pi_t^1 P, P_i, \pi_t^{3:N} P) \\
&\qquad\qquad -L_{t+1}(W_{Nm}O_t, \hat{\pi}_t^1 P, \pi_t^1 P, P_i, \pi_t^{3:N} P)\Big] \\
&\leq (\hat{\pi}_t^1 - \pi_t^1)U + \beta \sum_{i\geq L} \pi_t^2(i)(\hat{\pi}_t^1 P - \pi_t^1 P)U \\
&\leq (\hat{\pi}_t^1 - \pi_t^1)U + \beta \sum_{i\geq L} p_{Ki}(\hat{\pi}_t^1 P - \pi_t^1 P)U \\
&= (\hat{\pi}_t^1 - \pi_t^1)M. \qquad (132)
\end{aligned}$$

The first inequality in (132) follows from (112). The second inequality in (132) follows from the induction hypothesis for the lower bound of Property 7 at $t+1$. The second equality in (132) follows from (105). The third inequality in (132) follows from the induction hypothesis for the upper bound of Property 6 and the fact that $\hat{\pi}_t^1 P \geq_{st} \pi_t^1 P$ (since $\hat{\pi}_t^1 \geq_{st} \pi_t^1$ and Property 1). The last inequality in (132) is true because $\pi_t^2 \leq_{st} P_K$. The last equality in (132) follows from the definition of M.

The proof of the upper bound of Property 7 at t is now complete. The proof of the induction step for Property 7 at t is also complete.

Induction step for Property 9:

(i) When $O_t(N) \neq 1$ (i.e. $n \neq N$), assume $O_t(N) = N$ without loss of generality. Then,

$$\begin{aligned}
&V_t(A_{nm}O_t, \pi_t^{1:N}) - V_t(O_t, \pi_t^{1:N})\\
&= \sum_{i<L} \pi_t^N(i) \Big[V_{t+1}(S(A_{nm}O_t), \pi_t^{1:N-1}P, P_i)\\
&\qquad\qquad - V_{t+1}(SO_t, \pi_t^{1:N-1}P, P_i) \Big]\\
&\quad + \sum_{i\geq L} \pi_t^N(i) \Big[V_{t+1}(A_{nm}O_t, \pi_t^{1:N-1}P, P_i)\\
&\qquad\qquad - V_{t+1}(O_t, \pi_t^{1:N-1}P, P_i) \Big]\\
&\leq \sum_{i<L} \pi_t^N(i) \Big(h - \pi_t^1 P(P^{N-n-1}R) \Big)\\
&\quad + \sum_{i\geq L} \pi_t^N(i) \Big(h - \pi_t^1 P(P^{N-n}R) \Big)\\
&\leq h - \pi_t^1 P^{N-n} R.
\end{aligned} \tag{133}$$

The inequalities in (133) follows from the induction hypothesis for Property 9 and part (ii) of Property 4 respectively.

(ii) When $O_t(N) = 1$ (i.e. $n = N$), assume $O_t(N-1) = N$ without loss of generality. Then $A_{Nm}O_t(N) = O_t(N-1) = N$.

By the recursive equation and the linearity of the function V_{t+1} (eq. (44) and Lemma 1) we obtain

$$\begin{aligned}
&V_t(A_{Nm}O_t, \pi_t^{1:N}) - V_t(O_t, \pi_t^{1:N})\\
&= (\pi_t^N - \pi_t^1)R\\
&\quad + \beta \sum_{i<L} \pi_t^N(i) \Big[V_{t+1}(S(A_{Nm}O_t), \pi_t^{1:N-1}P, P_i)\\
&\qquad\qquad - V_{t+1}(A_{Nm}O_t, \pi_t^{1:N-1}P, P_i) \Big]\\
&\quad + \beta \sum_{i<L} \pi_t^1(i) \Big[V_{t+1}(A_{Nm}O_t, P_i, \pi_t^{2:N}P)\\
&\qquad\qquad - V_{t+1}(SO_t, P_i, \pi_t^{2:N}P) \Big]\\
&\quad + \beta \sum_{i\geq L} \pi_t^1(i) \Big[V_{t+1}(A_{Nm}O_t, P_i, \pi_t^{2:N}P)\\
&\qquad\qquad - V_{t+1}(O_t, P_i, \pi_t^{2:N}P) \Big].
\end{aligned} \tag{134}$$

Furthermore, each term in the second and the third sum in (134) is negative from repeatedly using Property 8 at $t+1$. Therefore,

$$\begin{aligned}
&V_t(A_{Nm}O_t, \pi_t^{1:N}) - V_t(O_t, \pi_t^{1:N}) \leq (\pi_t^N - \pi_t^1)R\\
&\quad + \beta \sum_{i<L} \pi_t^N(i) \Big[V_{t+1}(S(A_{Nm}O_t), \pi_t^{1:N-1}P, P_i)\\
&\qquad\qquad - V_{t+1}(A_{Nm}O_t, \pi_t^{1:N-1}P, P_i) \Big]\\
&\leq (\pi_t^N - \pi_t^1)R + \beta \sum_{i<L} \pi_t^N(i)(h - P_i R)\\
&= \pi_t^N v - \pi_t^1 R.
\end{aligned} \tag{135}$$

The second inequality in (135) follows from the induction hypothesis for Property 9 and v is the vector such that

$$v_i = \begin{cases} R_i + \beta(h - P_i R), & \text{for } i < L,\\ R_i, & \text{for } i \geq L. \end{cases} \tag{136}$$

It can be verified that v_i increases with i. Then, from part (i) of Property 4 and the fact that $\pi_t^N \leq_{st} P_K$ we obtain

$$\begin{aligned}
&V_t(A_{Nm}O_t, \pi_t^{1:N}) - V_t(O_t, \pi_t^{1:N})\\
&\quad \leq \pi_t^N v - \pi_t^1 R \leq P_K v - \pi_t^1 R = h - \pi_t^1 R.
\end{aligned} \tag{137}$$

The last equality in (137) follows from the definition of h.

This completes the proof of the induction step for Property 9 at t, and the proof of the entire induction step. ■

Acknowledgment

The authors would like to thank the associate editor and the anonymous reviewers for their thoughtful comments and suggestions on the original draft of this paper.

References

[1] Y. Ouyang and D. Teneketzis, "On the optimality of a myopic policy in multi-state channel probing," in *Proc. 50th Annu. Allerton Conf. Commun., Control, Comput.*, Monticello, IL, USA, 2012, pp. 342–349.

[2] Q. Zhao and B. Sadler, "A survey of dynamic spectrum access," *IEEE Signal Process. Mag.*, vol. 24, no. 3, pp. 79–89, May 2007.

[3] P. Whittle, "Restless bandits: Activity allocation in a changing world," *J. Appl. Probab.*, vol. 25A, pp. 287–298, 1988.

[4] J. Gittins, K. Glazebrook, and R. Weber, *Multi-Armed Bandit Allocation Indices*. Oxford, U.K.: Blackwell, 2011.

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

[6] V. D. Blondel and J. N. Tsitsiklis, "A survey of computational complexity results in systems and control," *Automatica*, vol. 36, no. 9, pp. 1249–1274, 2000.

[7] A. Marshall, I. Olkin, and B. Arnold, *Inequalities: Theory of Majorization and its Applications*. New York, NY, USA: Springer-Verlag, 2010.

[8] 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.

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

[10] S. Ahmad, M. Liu, T. Javidi, Q. Zhao, and B. Krishnamachari, "Optimality of myopic sensing in multichannel opportunistic access," *IEEE Trans. Inf. Theory*, vol. 55, no. 9, pp. 4040–4050, Sep. 2009.

[11] S. Ahmad and M. Liu, "Multi-channel opportunistic access: A case of restless bandits with multiple plays," in *Proc. 47th Annu. Allerton Conf. Commun., Control, Comput.*, Monticello, IL, USA, 2009, pp. 1361–1368.

[12] J. Gittins, "Bandit processes and dynamic allocation indices," *J. R. Stat. Soc.*, vol. 41, no. 2, pp. 148–177, 1979.

[13] R. Weber and G. Weiss, "On an index policy for restless bandits," *J. Appl. Probab.*, vol. 27, no. 3, pp. 637–648, 1990.

[14] J. Niño-Mora, "Dynamic priority allocation via restless bandit marginal productivity indices," *TOP*, vol. 15, no. 2, pp. 161–198, 2007.

[15] J. L. Ny, M. Dahleh, and E. Feron, "Multi-UAV dynamic routing with partial observations using restless bandit allocation indices," in *Proc. ACC*, Seattle, WA, USA, 2008, pp. 4220–4225.

[16] K. Liu and Q. Zhao, "Indexability of restless bandit problems and optimality of Whittle index for dynamic multichannel access," *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5547–5567, Nov. 2010.

[17] C. Lott and D. Teneketzis, "On the optimality of an index rule in multichannel allocation for single-hop mobile networks with multiple service classes," *Probab. Eng. Inf. Sci*, vol. 14, no. 3, pp. 259–297, 2000.

[18] N. Ehsan and M. Liu, "Server allocation with delayed state observation: Sufficient conditions for the optimality of an index policy," *IEEE Trans. Wireless Commun.*, vol. 8, no. 4, pp. 1693–1705, Apr. 2009.

[19] S. Guha, K. Munagala, and P. Shi, "Approximation algorithms for restless bandit problems," *J. ACM*, vol. 58, no. 1, pp. 3:1–3:50, 2010.

[20] M. P. Van Oyen and D. Teneketzis, "Optimal stochastic scheduling of forest networks with switching penalties," *Adv. Appl. Probab.*, vol. 26, no. 2, pp. 474–497, 1994.

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

[22] Y. Ouyang and D. Teneketzis, "On the optimality of myopic sensing in multi-state channels," 2013, available at arXiv:1305.6993.

[23] P. Varaiya, J. Walrand, and C. Buyukkoc, "Extensions of the multiarmed bandit problem: The discounted case," *IEEE Trans. Autom. Control*, vol. 30, no. 5, pp. 426–439, May 1985.

Yi Ouyang (S'13) received the B.S. degree in Electrical Engineering from the National Taiwan University, Taipei, Taiwan in 2009. He is currently a Ph.D. student in Electrical Engineering and Computer Science at the University of Michigan, Ann Arbor, MI, USA. His research interests include stochastic scheduling and decentralized stochastic control.

Demosthenis Teneketzis (M'87–SM'97–F'00) received the diploma in electrical engineering from the University of Patras, Patras, Greece, and the M.S., E.E., and Ph.D. degrees, all in electrical engineering, from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 1974, 1976, 1977, and 1979, respectively.

He is currently a Professor of Electrical Engineering and Computer Science at the University of Michigan, Ann Arbor, MI, USA. In winter and spring 1992, he was a Visiting Professor at the Swiss Federal Institute of Technology (ETH), Zurich, Switzerland. Prior to joining the University of Michigan, he worked for Systems Control, Inc., Palo Alto, CA,USA, and Alphatech, Inc., Burlington, MA, USA. His research interests are in stochastic control, decentralized systems, queueing and communication networks, stochastic scheduling and resource allocation problems, mathematical economics, and discrete-event systems.