# Indexability of Restless Bandit Problems and Optimality of Whittle Index for Dynamic Multichannel Access

Keqin Liu and Qing Zhao

***Abstract*—In this paper, we consider a class of restless multiarmed bandit processes (RMABs) that arises in dynamic multichannel access, user/server scheduling, and optimal activation in multiagent systems. For this class of RMABs, we establish the indexability and obtain Whittle index in closed form for both discounted and average reward criteria. These results lead to a direct implementation of Whittle index policy with remarkably low complexity. When arms are stochastically identical, we show that Whittle index policy is optimal under certain conditions. Furthermore, it has a semiuniversal structure that obviates the need to know the Markov transition probabilities. The optimality and the semiuniversal structure result from the equivalence between Whittle index policy and the myopic policy established in this work. For nonidentical arms, we develop efficient algorithms for computing a performance upper bound given by Lagrangian relaxation. The tightness of the upper bound and the near-optimal performance of Whittle index policy are illustrated with simulation examples.**

***Index Terms*—Dynamic channel selection, indexability, myopic policy, opportunistic access, restless multiarmed bandit (RMAB), Whittle index.**

## I. Introduction

### A. Restless Multiarmed Bandit Problem

RESTLESS MULTIARMED BANDIT PROCESSES (RMABs) are generalizations of the classical multiarmed bandit processes (MAB), which have been studied since 1930s [1]. In a MAB, a player, with full knowledge of the current state of each arm, chooses *one* out of $N$ arms to activate at each time and receives a reward determined by the state of the activated arm. Only the activated arm changes its state according to a Markovian rule while the states of passive arms are frozen. The objective is to maximize the long-run reward over an infinite horizon by choosing which arm to activate at each time.

The classical MAB problem remained open for almost 40 years until Gittins showed in [2] and [3] that the optimal policy has an index structure.[1] Specifically, a priority index (now known as Gittins index) can be assigned to each state of each arm, and the optimal action at each time is to activate the arm whose current state has the largest index. The significance of this result is that arms are decoupled when computing the index, thus reducing an $N$-dimensional problem to $N$ independent 1-D problems. As a consequence, the complexity of finding the optimal policy for a MAB is reduced from exponential with $N$ to linear with $N$.

Whittle generalized MAB to RMAB by allowing multiple $(K \geq 1)$ arms to be activated simultaneously and allowing passive arms to also change states and offer rewards [5]. Either of these two generalizations would render Gittins index policy suboptimal in general, and finding the optimal solution to a general RMAB has been shown to be PSPACE-hard by Papadimitriou and Tsitsiklis [6].

By considering the Lagrangian relaxation of the problem, Whittle proposed a heuristic index policy for RMABs [5]. Whittle index policy is the optimal solution to RMABs under a relaxed constraint: the number of activated arms can vary over time provided that its average over the infinite horizon equals to $K$. This average constraint leads to decoupling among arms, subsequently, the optimality of an index policy. Under the strict constraint that exactly $K$ arms are to be activated at each time, Whittle index policy has been shown to be asymptotically (as $N$ approaches infinity) optimal under certain conditions [7]. These conditions have been shown to always hold for two-state and three-state RMABs [7], [8]. In the finite regime, extensive empirical studies have demonstrated the near-optimal performance of Whittle index policy (see, for example, [9]–[11]).

Unfortunately, not every RMAB has a well-defined Whittle index; those that admit Whittle index policy are called *indexable* [5]. The indexability of an RMAB is often difficult to establish, the optimality of Whittle index policy in the finite regime is generally unknown, and computing Whittle index can be complex, often relying on numerical approximations that do not apply to RMABs with an infinite state space considered in this paper.

In this paper, we show that for a significant class of RMABs most relevant to dynamic multichannel access applications, the indexability can be established, Whittle index can be obtained in closed form, and, under certain conditions, Whittle index policy

Manuscript received November 13, 2008; revised January 26, 2010. Date of current version October 20, 2010. This work was supported by the Army Research Office under Grant W911NF-08-1-0467 and by the National Science Foundation under Grants ECS-0622200 and CCF-0830685. The material in this paper was presented in part at the 5th IEEE Conference on Sensor, Mesh, and Ad Hoc Communications and Networks (SECON), San Francisco, CA, June 2008 and the IEEE Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, October 2008.

The authors are with the Department of Electrical and Computer Engineering, University of California at Davis, Davis, CA 95616 USA (e-mail: kqliu@ ucdavis.edu; qzhao@ucdavis.edu).

Communicated by R. A. Berry, Associate Editor for Communication Networks.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2010.2068950

[1]According to Whittle [4], Gittins had the result in the late 1960s, but, deferred formal publication until 1974.

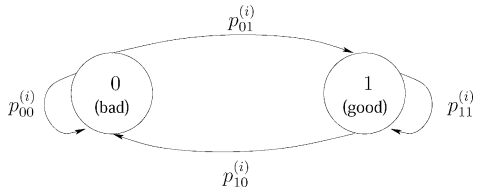


Fig. 1. The Gilber–Elliot channel model.

achieves the optimal performance with a simple semiuniversal structure that is robust against model mismatch and variations. This class of RMABs is described next.

### B. Dynamic Multichannel Access

Consider the problem of probing $N$ independent Markov chains. Each chain has two states—good $(1)$ and bad $(0)$—with different transition probabilities $\{p_{01}^{(i)}, p_{11}^{(i)}\}$ across chains (see Fig. 1). At each time, a player chooses $K$ $(1 \leq K < N)$ chains to probe and receives a reward for each probed chain that is in the good state. The objective is to design an optimal policy that governs the selection of $K$ chains at each time to maximize the long-run reward.

The above general problem arises in a wide range of communication systems, including cognitive radio networks, downlink scheduling in cellular systems, opportunistic transmission over fading channels, and resource-constrained jamming and antijamming. In the communications context, the $N$-independent Markov chains correspond to $N$ communication channels under the Gilbert–Elliot channel model [12], which has been commonly used to abstract physical channels with memory (see, for example, [13] and [14]). The state of a channel models the communication quality of this channel and determines the reward of accessing this channel. For example, in cognitive radio networks where secondary users search in the spectrum for idle channels temporarily unused by primary users [15], the state of a channel models the occupancy of the channel by primary users. For downlink scheduling in cellular systems, the user is a base station, and each channel is associated with a downlink mobile receiver. Downlink receiver scheduling is thus equivalent to channel selection.

The application of this problem also goes beyond communication systems. For example, it has applications in target tracking as considered in [16], where $K$ unmanned aerial vehicles are tracking the states of $N$ $(N > K)$ targets in each slot.

### C. Main Results

Fundamental questions concerning Whittle index policy since the day of its invention have been its existence, its performance, and the complexity in computing the index. What are the necessary and/or sufficient conditions on the state transition and the reward structure that make an RMAB indexable? When can Whittle index be obtained in closed form? For which special classes of RMABs is Whittle index policy optimal? When numerical evaluation has to be resorted to in studying its performance, are there easily computable performance benchmarks?

In this paper, we attempt to address these questions for the class of RMABs described above. As will be shown, this class of RMABs has an *uncountable* state space (when considering all possible initial conditions), making the problem highly nontrivial. The underlying two-state Markov chain that governs the state transition of each arm, however, brings rich structures into the problem, leading to positive and surprising answers to the above questions. The wide range of applications of this class of RMABs makes the results obtained in this paper generally applicable.

Under both discounted and average reward criteria, we establish the indexability of this class of RMABs. The basic technique of our proof is to bound the total amount of time that an arm is made passive under the optimal policy. The general approach of using the total passive time in proving indexability was considered by Whittle [5] when showing that a classic MAB is always indexable. Applying this approach to a nontrivial RMAB is, however, much more involved, and our proof appears to be the first that extends this approach to RMABs. We hope that this work contributes to the set of possible techniques for establishing indexability of RMABs.

Based on the indexability, we show that Whittle index can be obtained in closed form for both discounted and average reward criteria. This result reduces the complexity of implementing Whittle index policy to simple evaluations of these closed-form expressions. This result is particularly significant considering the uncountable state space which would render numerical approaches impractical. The monotonicity and piecewise concavity [for positively correlated arms[2] (i.e., $p_{11} \geq p_{01}$)] or piecewise convexity [for negatively correlated arms (i.e., $p_{11} < p_{01}$)] of Whittle index are also established. The monotonicity of Whittle index leads to an interesting equivalence with the myopic policy—the simplest nontrivial index policy—when arms are stochastically identical (i.e., all arms have the same Markovian dynamics and reward structure). This equivalence allows us to work on the myopic index, which has a much simpler form, when establishing the structure, the optimality, and the performance (in terms of system parameters) of Whittle index policy for stochastically identical arms. A sufficient condition for the equivalence between Whittle index policy and the myopic policy for a general RMAB is also established.

When arms are stochastically identical, we show that Whittle index policy is optimal under certain conditions. This result provides an example for the optimality of Whittle index policy in the finite regime. The approximation factor of Whittle index policy (the ratio of the performance of Whittle index policy to that of the optimal policy) is analyzed when the optimality conditions do not hold. The performance of Whittle index policy in terms of the system parameters is also analyzed. Furthermore, we show that Whittlle index policy has a semiuniversal structure that obviates the need to know the Markov transition probabilities. The only required knowledge about the Markovian model is the order of $p_{11}$ and $p_{01}$. This semiuniversal structure reveals the robustness of Whittle index policy against model mismatch and variations. The optimality and the structure of Whittle index policy for stochastically identical arms are obtained based on its

[2]It is easy to show that $p_{11} > p_{01}$ corresponds to the case where the channel states in two consecutive slots are positively correlated, i.e., for any distribution of $S(t)$, we have $\mathbb{E}[(S(t) - \mathbb{E}[S(t)])(S(t+1) - \mathbb{E}[S(t+1)])] > 0$, where $S(t)$ is the state of the Gilbert–Elliot channel in slot $t$. Similarly, $p_{11} < p_{01}$ corresponds to the case where $S(t)$ and $S(t+1)$ are negatively correlated, and $p_{11} = p_{01}$ the case where $S(t)$ and $S(t+1)$ are independent.

equivalence to the myopic policy and prior findings in [17]–[19] on the myopic policy for this class of RMAB.

When arms are nonidentical, numerical evaluations are resorted to when assessing the performance of Whittle index policy. To this end, we develop an efficient algorithm for computing an upper bound of the optimal performance given by Lagrangian relaxation. We show that this algorithm runs in at most $O(N(\log N)^2)$ time to compute the performance upper bound within $\epsilon$-accuracy for any $\epsilon > 0$. When every arm is negatively correlated, this algorithm produces the exact performance upper bound in $O(N^2 \log N)$ time. Simulation examples demonstrate both the tightness of the upper bound and the near-optimal performance of Whittle index policy.

### D. Related Work

Dynamic multichannel access in the context of cognitive radio systems has been studied in [20] and [21] where the problem is formulated as a partially observable Markov decision process (POMDP) to take into account potential correlations among channels. For stochastically identical and independent channels and under the assumption of single-channel sensing ($K = 1$), the structure, optimality, and performance of the myopic policy were studied in [17], where the semiuniversal structure and the performance characterization of the myopic policy were established for all $N$ and the optimality of the myopic policy proved for $N = 2$ (both positive and negative correlation cases). The optimality of the myopic policy was extended to $N > 2$ positively correlated channels with $K = 1$ in [18], and then to arbitrary $K$ in [19]. Extensions to cases with probing errors were addressed in [22]. The equivalence between the myopic policy and Whittle index policy established in this paper for stochastically identical channels shows that the results obtained in [17]–[19] for the myopic policy are directly applicable to Whittle index policy. Furthermore, we also address extensions to negatively correlated arms. Specifically, we show that Whittle index policy is optimal for $K = N - 1$. For a general $K$, we establish the approximation factor of Whittle index policy.

Other examples of applying the general RMAB framework to communication systems can be found in [23]–[25]. In [23], the problem of multichannel allocation in single-hop mobile networks with multiple service classes was formulated as an RMAB, and sufficient conditions for the optimality of a myopic-type index policy were established. In [24], multicast scheduling in wireless broadcast systems with strict deadlines was formulated as an RMAB with a finite state space. The indexability was established and Whittle index was obtained in closed form. In [25], a bandwidth allocation problem arisen in queuing systems was formulated as an RMAB with countable sate space; the indexability, closed-form Whittle index, and sufficient conditions for the optimality of Whittle index policy were obtained. The RMAB framework has also been applied to economic systems for handling inventory regulation [26].

In the general context of RMAB, there is a rich literature on indexability. See [10] for examples of specific indexable restless bandit processes and [27] and [28] for a numerical approach of testing indexability and calculating Whittle index. We point out that the numerical approach established in [27] and [28] only applies to RMABs with a finite state space and under specific values of the system parameters (such as the transition probabilities and the reward of each arm). Consequently, if any parameter (in particular, the transition probabilities) takes infinite possible values, the procedure cannot enumerate all possible system settings. In this paper, we show that for the class of RMAB considered here, indexability holds regardless of the system parameters and the closed-form Whittle index is obtained in terms of general system parameters.

Constant-factor approximation algorithms for RMABs have also been explored in the literature. For the same class of RMABs as considered in this paper, Guha and Munagala [29] developed a constant-factor $(1/68)$ approximation via linear programming (LP) relaxation under the condition of $p_{11} > \frac{1}{2} > p_{01}$ for each arm. In [30], Guha *et al.* developed a factor-2 approximation policy via LP relaxation for the so-called monotone restless bandit processes.

In [16], Le Ny *et al.* have considered the same class of RMABs motivated by the applications of target tracking. They have independently established the indexability and obtained the closed-form expressions for Whittle index under the discounted reward criterion. A conference version of our result was published at the same time as [16]. Our approach to establishing indexability and obtaining Whittle index is, however, different from that used in [16], and the two approaches complement each other. Indeed, the fact that two completely different applications lead to the same class of RMABs lends support for a detailed investigation of this particular type of RMABs. We also include several results that were not considered in [16]. In particular, we consider both discounted and average reward criteria, develop algorithms for and analyze the complexity of computing the optimal performance under the Lagrangian relaxation, and establish the semiuniversal structure, the optimality, and the performance of Whittle index policy for stochastically identical arms.

### E. Organization

The rest of the paper is organized as follows. In Section II, the RMAB formulation is presented. In Section III, we introduce the basic concepts of indexability and Whittle index. In Section IV, we address the total discounted reward criterion, where we establish the indexability, obtain Whittle index in closed form, and develop efficient algorithms for computing an upper bound on the performance of the optimal policy. Simulation examples are provided to illustrate the tightness of the upper bound and the near-optimal performance of Whittle index policy. In Section V, we consider the average reward criterion and obtain results parallel to those obtained under the discounted reward criterion. In Section VI, we consider the special case when channels are stochastically identical. We show that Whittle index policy is optimal under certain conditions and has a simple and robust structure. The approximation factor of Whittle index policy is also analyzed. Section VII concludes this paper.

## II. Problem Statement and Restless Bandit Formulation

### A. Dynamic Multichannel Access

We motivate this class of RMABs by considering the application of dynamic multichannel access. Consider $N$ independent Gilbert–Elliot channels, each with transmission rate $B_i$ $(i = 1, \ldots, N)$. Without loss of generality, we normalize the maximum data rate: $\max_{i\in\{1,2,\ldots,N\}}\{B_i\} = 1$. The state of channel $i$—good $(1)$ or bad $(0)$—evolves from slot to slot as a Markov chain with transition matrix $\mathbf{P}_i = \{p_{j,k}^{(i)}\}_{j,k\in\{0,1\}}$ as shown in Fig. 1.

At the beginning of slot $t$, the user selects $K$ out of $N$ channels to sense. If the state $S_i(t)$ of the sensed channel $i$ is 1, the user transmits and collects $B_i$ units of reward in this channel. Otherwise, the user collects no reward in this channel. Let $U(t)$ denote the set of $K$ channels chosen in slot $t$. The reward obtained in slot $t$ is thus given by

$$R_{U(t)}(t) = \Sigma_{i\in U(t)} S_i(t) B_i.$$

Our objective is to maximize the expected long-run reward by designing a sensing policy that sequentially selects $K$ channels to sense in each slot.

### B. Restless Multiarmed Bandit Formulation

The channel states $[S_1(t), \ldots, S_N(t)] \in \{0,1\}^N$ are not directly observable before the sensing action is made. The user can, however, infer the channel states from its decision and observation history. It has been shown that a sufficient statistic for optimal decision making is given by the conditional probability that each channel is in state 1 given all past decisions and observations [31]. Referred to as the belief vector or information state, this sufficient statistic is denoted by $\Omega(t) \triangleq [\omega_1(t), \ldots, \omega_N(t)]$, where $\omega_i(t)$ is the conditional probability that $S_i(t) = 1$. Given the sensing action $U(t)$ and the observation in slot $t$, the belief state in slot $t+1$ can be obtained recursively as follows:

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

where

$$\mathcal{T}(\omega_i(t)) \triangleq \omega_i(t) p_{11}^{(i)} + (1 - \omega_i(t)) p_{01}^{(i)}$$

denotes the operator for the one-step belief update for unobserved channels.

If no information on the initial system state is available, the $i$th entry of the initial belief vector $\Omega(1)$ can be set to the stationary distribution $\omega_o^{(i)}$ of the underlying Markov chain

$$\omega_o^{(i)} = \frac{p_{01}^{(i)}}{p_{01}^{(i)} + p_{10}^{(i)}}. \quad (2)$$

It is now easy to see that we have an RMAB, where each channel is considered as an arm and the state of arm $i$ in slot $t$ is the belief state $\omega_i(t)$. The user chooses an action $U(t)$ consisting of $K$ arms to activate (sense) in each slot, while other arms are made passive (unobserved). The states of both active and passive arms change as given in (1). A policy $\pi : \Omega(t) \to U(t)$ is a function that maps from the belief vector $\Omega(t)$ to the action $U(t)$ in slot $t$.

There are two commonly used performance measures. One is the expected total *discounted* reward over the infinite horizon

$$\mathbb{E}_\pi \left[ \sum_{t=1}^{\infty} \beta^{t-1} R_{\pi(\Omega(t))}(t) \,|\, \Omega(1) \right] \quad (3)$$

where $0 \le \beta < 1$ is the discount factor and $R_{\pi(\Omega(t))}(t)$ is the reward obtained in slot $t$ under action $U(t) = \pi(\Omega(t))$ determined by the policy $\pi$. This performance measure applies when rewards in the future are less valuable, for example, in delay-sensitive communication systems. It also applies when the horizon length is a geometrically distributed random variable with parameter $\beta$. For example, a communication session may end at a random time, and the user aims to maximize the number of packets delivered before the session ends.

The other performance measure is the expected *average* reward over the infinite horizon [32]

$$\mathbb{E}_\pi \left[ \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} R_{\pi(\Omega(t))}(t) \,\middle|\, \Omega(1) \right]. \quad (4)$$

This is the common measure of throughput in the context of communications.

For notation convenience, let $(\Omega(1), \{\mathbf{P}_i\}_{i=1}^N, \{B_i\}_{i=1}^N, \beta)$ denote the RMAB with the discounted reward criterion, and $(\Omega(1), \{\mathbf{P}_i\}_{i=1}^N, \{B_i\}_{i=1}^N, 1)$ the RMAB with the average reward criterion.

## III. Indexability and Index Policies

In this section, we introduce the basic concepts of indexability and Whittle index policy.

### A. Index Policy

An *index policy* assigns an index to each state of each arm to measure how rewarding it is to activate an arm at a particular state. In each slot, the policy activates those $K$ arms whose current states have the largest indices.

For a strongly decomposable index policy, the index of an arm only depends on the characteristics (transition probabilities, reward structure, etc.) of this arm. Arms are thus decoupled when computing the index, reducing an $N$-dimensional problem to $N$ independent 1-D problems.

A myopic policy is a simple example of strongly decomposable index policies. This policy ignores the impact of the current action on the future reward, focusing solely on maximizing the expected immediate reward. The index is thus the expected immediate reward of activating an arm at a particular state. For the problem at hand, the myopic index of each state $\omega_i(t)$ of arm $i$ is simply $\omega_i(t)B_i$. The myopic action $\hat{U}(t)$ under the belief state $\Omega(t) = [\omega_1(t), \ldots, \omega_N(t)]$ is given by

$$\hat{U}(t) = \arg\max_{U(t)} \Sigma_{i\in U(t)} \omega_i(t) B_i. \quad (5)$$

### B. Indexability and Whittle Index Policy

To introduce indexability and Whittle index, it suffices to consider a single arm due to the strong decomposability

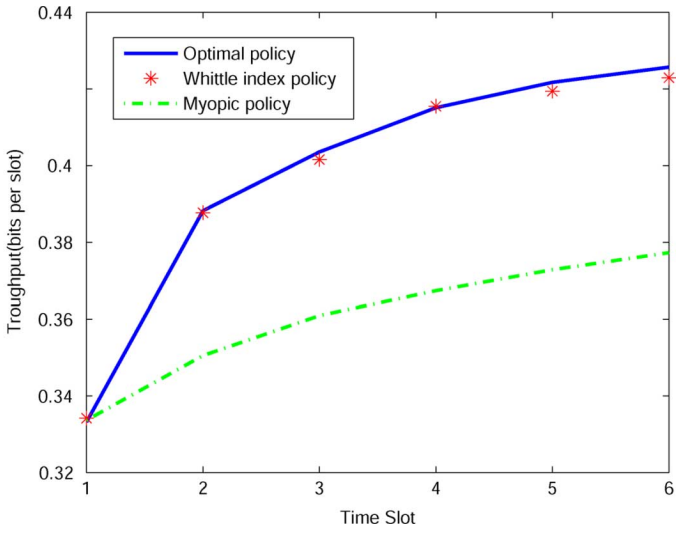


Fig. 2. The performance by Whittle index policy ($K = 1, N = 7, \{p_{01}^{(i)}\}_{i=1}^{7} = \{0.8, 0.6, 0.4, 0.9, 0.8, 0.6, 0.7\}, \{p_{11}^{(i)}\}_{i=1}^{7} = \{0.6, 0.4, 0.2, 0.2, 0.4, 0.1, 0.3\}$, and $B_i = \{0.4998, 0.6668, 1.0000, 0.6296, 0.5830, 0.8334, 0.6668\}$).

of Whittle index. Consider a single-armed bandit process (a single channel) with transition probabilities $\{p_{j,k}\}_{j,k\in 0,1}$ and bandwidth $B$ (here we drop the channel index for notation simplicity). In each slot, the user chooses one of two possible actions—$u \in \{0(\text{passive}), 1(\text{active})\}$—to make the arm passive or active. An expected reward of $\omega B$ is obtained when the arm is activated at belief state $\omega$, and the belief state transits according to (1). The objective is to decide whether to active the arm in each slot to maximize the total discounted or average reward. The optimal policy is essentially given by an optimal partition of the state space $[0, 1]$ into a passive set $\{\omega : u^*(\omega) = 0\}$ and an active set $\{\omega : u^*(\omega) = 1\}$, where $u^*(\omega)$ denotes the optimal action under belief state $\omega$.

Whittle index measures how attractive it is to activate an arm based on the concept of *subsidy for passivity* [5]. Specifically, we construct a single-armed bandit process that is identical to the above specified bandit process except that a constant subsidy $m$ is obtained whenever the arm is made passive. Obviously, this subsidy $m$ will change the optimal partition of the passive and active sets, and states that remain in the active set under a larger subsidy $m$ are more attractive to the user. The minimum subsidy $m$ that is needed to move a state from the active set to the passive set under the optimal partition thus measures how attractive this state is.

We now present the formal definition of indexability and Whittle index. We consider the discounted reward criterion. Their definitions under the average reward criterion can be similarly obtained.

Denoted by $V_{\beta,m}(\omega)$, the value function represents the maximum expected total discounted reward that can be accrued from a single-armed bandit process with subsidy $m$ when the initial belief state is $\omega$. Considering the two possible actions in the first slot, we have

$$V_{\beta,m}(\omega) = \max\{V_{\beta,m}(\omega; u = 0), V_{\beta,m}(\omega; u = 1)\} \quad (6)$$

where $V_{\beta,m}(\omega; u)$ denotes the expected total discounted reward obtained by taking action $u$ in the first slot followed by the optimal policy in future slots. Consider $V_{\beta,m}(\omega; u = 0)$. It is given by the sum of the subsidy $m$ obtained in the first slot under the passive action and the total discounted future reward $\beta V_{\beta,m}(\mathcal{T}(\omega))$ which is determined by the updated belief state $\mathcal{T}(\omega)$ [see (1)]. $V_{\beta,m}(\omega; u = 1)$ can be similarly obtained, and we arrive at the following dynamic programming:

$$V_{\beta,m}(\omega; u = 0) = m + \beta V_{\beta,m}(\mathcal{T}(\omega)), \quad (7)$$

$$V_{\beta,m}(\omega; u = 1) = \omega + \beta(\omega V_{\beta,m}(p_{11}) + (1 - \omega)V_{\beta,m}(p_{01})). \quad (8)$$

The optimal action $u_m^*(\omega)$ for belief state $\omega$ under subsidy $m$ is given by

$$u_m^*(\omega) = \begin{cases} 1, & \text{if } V_{\beta,m}(\omega; u = 1) > V_{\beta,m}(\omega; u = 0) \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

The passive set $\mathcal{P}(m)$ under subsidy $m$ is given by

$$\mathcal{P}(m) = \{\omega : u_m^*(\omega) = 0\} \quad (10)$$

$$= \{\omega : V_{\beta,m}(\omega; u = 0) \geq V_{\beta,m}(\omega; u = 1)\}. \quad (11)$$

*Definition 1:* An arm is *indexable* if the passive set $\mathcal{P}(m)$ of the corresponding single-armed bandit process with subsidy $m$ monotonically increases from $\emptyset$ to the whole state space $[0, 1]$ as $m$ increases from $-\infty$ to $+\infty$. An RMAB is indexable if every arm is indexable.

Under the indexability condition, Whittle index is defined as follows.

*Definition 2:* If an arm is indexable, its *Whittle index* $W(\omega)$ of the state $\omega$ is the infimum subsidy $m$ such that it is optimal to make the arm passive at $\omega$. Equivalently, Whittle index $W(\omega)$ is the infimum subsidy $m$ that makes the passive and active actions equally rewarding

$$W(\omega) = \inf_m \{m : u_m^*(\omega) = 0\} \quad (12)$$

$$= \inf_m \{m : V_{\beta,m}(\omega; u = 0) = V_{\beta,m}(\omega; u = 1)\}. \quad (13)$$

In Fig. 2, we compare the performance (throughput) of the myopic policy, Whittle index policy, and the optimal policy for the RMAB formulated in Section II. We observe that Whittle index policy achieves a near-optimal performance while the myopic policy suffers from a significant performance loss.

## IV. Whittle Index Under Discounted Reward Criterion

In this section, we focus on the discounted reward criterion. We establish the indexability, obtain Whittle index in closed form, and develop efficient algorithms for computing an upper bound of the optimal performance to provide a benchmark for evaluating the performance of Whittle index policy.

### A. Properties of Belief State Transition

To establish indexability and obtain Whittle index, it suffices to consider the single-armed bandit process with subsidy $m$. Again, we drop the channel index from all notations and set $B = 1$.

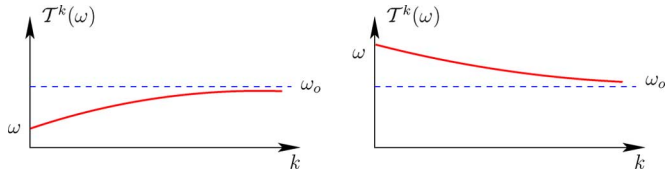


Fig. 3. The $k$-step belief update of an unobserved arm ($p_{11} \geq p_{01}$).

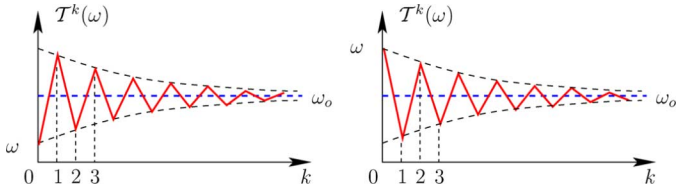


Fig. 4. The $k$-step belief update of an unobserved arm ($p_{11} < p_{01}$).

The following lemma establishes properties of belief state transition that reveal the basic structure of the RMAB considered in this paper. We resort often to these properties when deriving the main results.

*Lemma 1:* Let $\mathcal{T}^k(\omega(t)) \triangleq \Pr[S(t+k) = 1 \,|\, \omega(t)] (k = 0, 1, 2, \ldots)$ denote the $k$-step belief update of $\omega(t)$ when the arm is unobserved for $k$ consecutive slots. We have

$$\mathcal{T}^k(\omega) = \frac{p_{01} - (p_{11} - p_{01})^k (p_{01} - (1 + p_{01} - p_{11})\omega)}{1 + p_{01} - p_{11}}, \quad (14)$$

$$\min\{p_{01}, p_{11}\} \leq \mathcal{T}^k(\omega) \leq \max\{p_{01}, p_{11}\}, \quad \forall \omega \in [0,1], \quad \forall k \geq 1. \quad (15)$$

Furthermore, the convergence of $\mathcal{T}^k(\omega)$ to the stationary distribution $\omega_o = \frac{p_{01}}{p_{01}+p_{10}}$ has the following property.

Case 1) *Positively correlated channel* ($p_{11} \geq p_{01}$). For any $\omega \in [0,1]$, $\mathcal{T}^k(\omega)$ monotonically converges to $\omega_o$ as $k \to \infty$ (see Fig. 3).

Case 2) *Negatively correlated channel* ($p_{11} < p_{01}$). For any $\omega \in [0,1]$, $\mathcal{T}^{2k}(\omega)$ and $\mathcal{T}^{2k+1}(\omega)$ converge, from opposite directions, to $\omega_o$ as $k \to \infty$ (see Fig. 4).

*Proof:* $\mathcal{T}^k(\omega) = \omega \mathcal{T}^k(1) + (1-\omega)\mathcal{T}^k(0)$, where $\mathcal{T}^k(1) = \Pr[S(t+k) = 1 \,|\, S(t) = 1]$ is the $k$-step transition probability from 1 to 1, and $\mathcal{T}^k(0) = \Pr[S(t+k) = 1 \,|\, S(t) = 0]$ is the $k$-step transition probability from 0 to 1. From the eigendecomposition of the transition matrix $\mathbf{P}$ (see [33]), we have $\mathcal{T}^k(1) = \frac{p_{01} + (1-p_{11})(p_{11}-p_{01})^k}{1+p_{01}-p_{11}}$ and $\mathcal{T}^k(0) = \frac{p_{01}(1-(p_{11}-p_{01})^k)}{1+p_{01}-p_{11}}$, which leads to (14). Other properties follow directly from (14). □

Next, we define an important quantity $L(\omega, \omega')$. Referred to as the *first crossing time*, $L(\omega, \omega')$ is the minimum amount of time required for a passive arm to transit across $\omega'$ starting from $\omega$

$$L(\omega, \omega') \triangleq \min\{k : \mathcal{T}^k(\omega) > \omega'\}.$$

For a positively correlated arm, we have, from Lemma 1, (16), shown at the bottom of the page. For a negatively correlated arm, we have

$$L(\omega, \omega') = \begin{cases} 0, & \text{if } \omega > \omega' \\ 1, & \text{if } \omega \leq \omega' \text{ and } \mathcal{T}(\omega) > \omega' \\ \infty, & \text{if } \omega \leq \omega' \text{ and } \mathcal{T}(\omega) \leq \omega'. \end{cases} \quad (17)$$

In the next section, we will show that the optimal policy for the single-armed bandit process with subsidy is a threshold policy: the arm will be activated if its belief state crosses a certain threshold $\omega^*$. In other words, starting from an arbitrary belief state $\omega$, the first active action on the arm is taken after $L(\omega, \omega^*)$ slots. After the active action is taken, belief state of this arm is either $p_{01}$ or $p_{11}$ in the next slot. Consequently, the value function of the arm starting from an arbitrary belief state only depends on the first crossing time $L(\omega, \omega^*)$ and the value functions of $p_{01}$ and $p_{11}$. These results lead to sufficient equations to solve for the value function and the total passive time in closed-form (see Sections IV-C and IV-D), which are the key quantities in establishing the indexability and the closed-form Whittle index (see Section IV-E).

### B. The Optimal Policy for the Single-Armed Bandit Process With Subsidy

In this section, we show that the optimal policy for the single-armed bandit process with subsidy $m$ is a threshold policy. This threshold structure is obtained by examining the value functions $V_{\beta,m}(\omega; u = 0)$ and $V_{\beta,m}(\omega; u = 1)$ given in (7) and (8). From (8), we observe that $V_{\beta,m}(\omega; u = 1)$ is a linear function of $\omega$. Following the general result on the convexity of the value function of a partially observable Markov decision process (POMDP) [34], we conclude that $V_{\beta,m}(\omega; u = 0)$ given in (7) is convex in $\omega$. These properties of $V_{\beta,m}(\omega; u = 1)$ and $V_{\beta,m}(\omega; u = 0)$ lead to the following lemma.

*Lemma 2:* The optimal policy for the single-armed bandit process with subsidy $m$ is a threshold policy, i.e., there exists an $\omega_\beta^*(m) \in \mathbb{R}$ such that

$$u_m^*(\omega) = \begin{cases} 1, & \text{if } \omega > \omega_\beta^*(m) \\ 0, & \text{if } \omega \leq \omega_\beta^*(m) \end{cases}$$

and $V_{\beta,m}(\omega_\beta^*(m); u = 0) = V_{\beta,m}(\omega_\beta^*(m); u = 1)$.

$$L(\omega, \omega') = \begin{cases} 0, & \text{if } \omega > \omega' \\ \left\lfloor \log_{p_{11}-p_{01}} \frac{p_{01} - \omega'(1-p_{11}+p_{01})}{p_{01} - \omega(1-p_{11}+p_{01})} \right\rfloor + 1, & \text{if } \omega \leq \omega' < \omega_o \\ \infty, & \text{if } \omega \leq \omega' \text{ and } \omega' \geq \omega_o. \end{cases} \quad (16)$$

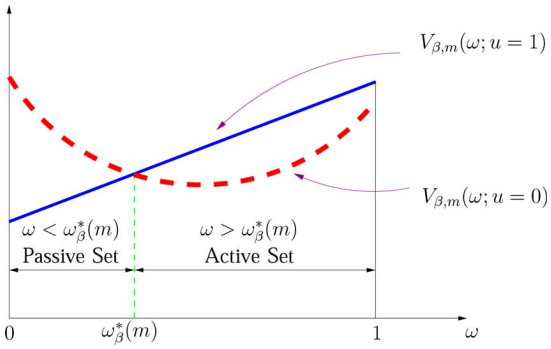


Fig. 5. The optimality of a threshold policy ($0 \le m < 1$).

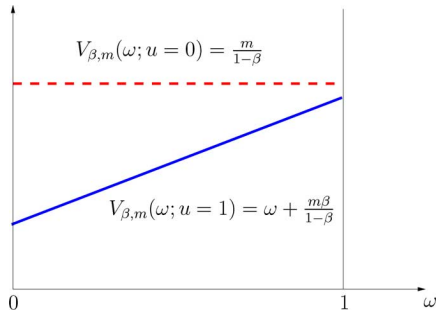


Fig. 6. The optimality of a threshold policy ($m \ge 1$).

*Proof:* Consider first $0 \le m < 1$. We have the following inequality regarding the end points of $V_{\beta,m}(0;u=1)$ and $V_{\beta,m}(0;u=0)$ (see Fig. 5):

$$\begin{aligned} V_{\beta,m}(0;u=1) &= \beta V_{\beta,m}(p_{01}) \le m+\beta V_{\beta,m}(p_{01}) \\ &= V_{\beta,m}(0;u=0) \qquad (18) \\ V_{\beta,m}(1;u=1) &= 1+\beta V_{\beta,m}(p_{11}) > m+\beta V_{\beta,m}(p_{11}) \\ &= V_{\beta,m}(1;u=0). \qquad (19) \end{aligned}$$

Since $V_{\beta,m}(\omega;u=1)$ is linear in $\omega$ and $V_{\beta,m}(\omega;u=0)$ is convex in $\omega, V_{\beta,m}(\omega;u=1)$ and $V_{\beta,m}(\omega;u=0)$ must have one unique intersection at some point $\omega^*_\beta(m)$ as shown in Fig. 5.

When $m \ge 1$, it is optimal to make the arm passive all the time since the expected immediate reward $\omega$ by activating the arm is uniformly upper bounded by 1 (see Fig. 6). We can thus choose $\omega^*_\beta(m)=c$ for any $c>1$.

When $m<0$, we have (see Fig. 7)

$$\begin{aligned} V_{\beta,m}(0;u=1) &= \beta V_{\beta,m}(p_{01}) > m+\beta V_{\beta,m}(p_{01}) \\ &= V_{\beta,m}(0;u=0) \qquad (20) \\ V_{\beta,m}(1;u=1) &= 1+\beta V_{\beta,m}(p_{11}) > m+\beta V_{\beta,m}(p_{11}) \\ &= V_{\beta,m}(0;u=0). \qquad (21) \end{aligned}$$

Based on the convexity of $V_{\beta,m}(\omega;u=0)$ in $\omega$, we have $V_{\beta,m}(\omega;u=1) > V_{\beta,m}(\omega;u=0)$ for any $\omega \in [0,1]$. It is thus optimal to always activate the arm, and we can choose $\omega^*_\beta(m)=b$ for any $b<0$. Lemma 2 thus follows. The expressions of $V_{\beta,m}(0;u=1)$ and $V_{\beta,m}(0;u=0)$ given in Figs. 6 and 7 are obtained from the closed-form expression of the value function, which will be shown in Section IV-C. □

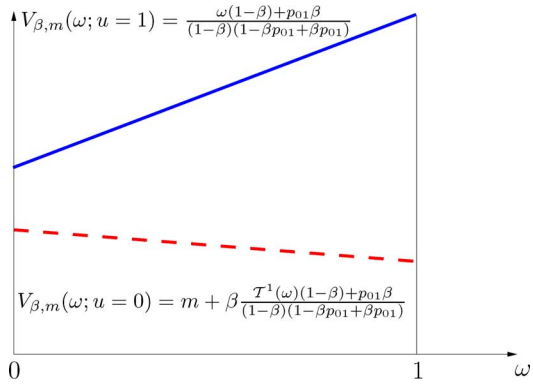


Fig. 7. The optimality of a threshold policy ($m < 0$).

### C. Closed-Form Expression of the Value Function

In this section, we obtain closed-form expressions for the value function $V_{\beta,m}(\omega)$. This result is fundamental to calculating Whittle index in closed-form and analyzing the performance of Whittle index policy.

Based on the threshold structure of the optimal policy, the value function $V_{\beta,m}(\omega)$ can be expressed in terms of $V_{\beta,m}(\mathcal{T}^{t_0-1}(\omega);u=1)$ for some $t_0 \in \mathcal{Z}^+ \cup \{\infty\}$, where $t_0 = L(\omega,\omega^*_\beta(m))+1$ is the index of the slot when the belief $\omega$ transits across the threshold $\omega^*_\beta(m)$ for the first time [recall that $L(\omega,\omega^*_\beta(m))$ is the first crossing time given in (16) and (17)]. Specifically, in the first $L(\omega,\omega^*_\beta(m))$ slots, the subsidy $m$ is obtained in each slot. In slot $t_0 = L(\omega,\omega^*_\beta(m))+1$, the belief state transits across the threshold $\omega^*_\beta(m)$ and the arm is activated. The total reward thereafter is $V_{\beta,m}(\mathcal{T}^{L(\omega,\omega^*_\beta(m))}(\omega);u=1)$. We thus have, considering the discount factor

$$\begin{aligned} V_{\beta,m}(\omega) = {} & \frac{1-\beta^{L(\omega,\omega^*_\beta(m))}}{1-\beta}m \\ & + \beta^{L(\omega,\omega^*_\beta(m))}V_{\beta,m}\left(\mathcal{T}^{L(\omega,\omega^*_\beta(m))}(\omega);u=1\right). \qquad (22) \end{aligned}$$

Since $V_{\beta,m}(\mathcal{T}^{t_0-1}(\omega);u=1)$ is a function of $V_{\beta,m}(p_{01})$ and $V_{\beta,m}(p_{11})$ as shown in (7), we only need to solve for $V_{\beta,m}(p_{01})$ and $V_{\beta,m}(p_{11})$. Note that $p_{01}$ and $p_{11}$ are simply two specific values of $\omega$; both $V_{\beta,m}(p_{01})$ and $V_{\beta,m}(p_{11})$ can be written as functions of themselves through (22). We can thus solve for $V_{\beta,m}(p_{01})$ and $V_{\beta,m}(p_{11})$ as given in Lemma 3.

*Lemma 3:* Let $\omega^*_\beta(m)$ denote the threshold of the optimal policy for the single-armed bandit process with subsidy $m$. The value functions $V_{\beta,m}(p_{01})$ and $V_{\beta,m}(p_{11})$ can be obtained in closed-form as given below.

Case 1) *Positively correlated channel* ($p_{11} \ge p_{01}$) [see (23)–(24), shown at the bottom of the next page]. Note that $V_{\beta,m}(p_{01})$ is given explicitly in (23) while $V_{\beta,m}(p_{11})$ is given in terms of $V_{\beta,m}(p_{01})$ for the ease of presentation.

Case 2) *Negatively correlated channel* ($p_{11} < p_{01}$) [see (25)–(26), shown at the bottom of the next page]. Note that $V_{\beta,m}(p_{11})$ is given explicitly in (25) while $V_{\beta,m}(p_{01})$ is given in terms of $V_{\beta,m}(p_{11})$ for the ease of presentation.

*Proof:* The key to the closed-form expressions for $V_{\beta,m}(p_{01})$ and $V_{\beta,m}(p_{11})$ is to find the first slot in which the

optimal policy activates the arm [i.e., the belief state transits across the threshold $\omega^*_\beta(m)$]. This can be done by applying the transition properties of the belief state given in Lemma 1. See Appendix I for the complete proof. □

### D. The Total Discounted Time of Being Passive

In this section, we study the total discounted time that the single-armed bandit process with subsidy $m$ is made passive. This quantity plays the central role in our proof of indexability and in the algorithms of computing an upper bound of the optimal performance as shown in Sections IV-E and IV-F, respectively.

Let $D_{\beta,m}(\omega)$ denote the (expected) total discounted time that the single-armed bandit process with subsidy $m$ is made passive under the optimal policy when the initial belief state is $\omega$. It has been shown by Whittle that $D_{\beta,m}(\omega)$ is the derivative of the value function $V_{\beta,m}(\omega)$ with respect to $m$ [5]

$$D_{\beta,m}(\omega)=\frac{d(V_{\beta,m}(\omega))}{dm}.$$

This result is intuitive: when the subsidy for passivity $m$ increases, the rate at which the total discounted reward $V_{\beta,m}(\omega)$ increases is determined by how often the arm is made passive.

Based on the threshold structure of the optimal policy, we can obtain the following dynamic programming equation for $D_{\beta,m}(\omega)$ similar to that for $V_{\beta,m}(\omega)$ given in (22):

$$\begin{aligned}D_{\beta,m}(\omega)=&\frac{1-\beta^{L(\omega,\omega^*_\beta(m))}}{1-\beta}\\&+\beta^{L(\omega,\omega^*_\beta(m))+1}\Big(\mathcal{T}^{L(\omega,\omega^*_\beta(m))}(\omega)D_{\beta,m}(p_{11})\\&+(1-\mathcal{T}^{L(\omega,\omega^*_\beta(m))}(\omega))D_{\beta,m}(p_{01})\Big).\end{aligned}\tag{27}$$

Specifically, the first term in (27) is the total discounted time of the first $L(\omega,\omega^*_\beta(m))$ slots when the arm is made passive. In slot $L(\omega,\omega^*_\beta(m))+1$, the arm is activated. With probability $\mathcal{T}^{L(\omega,\omega^*_\beta(m))}(\omega)$, the channel is in the good state in this slot, and the total future discounted passive time is $D_{\beta,m}(p_{11})$. With probability $1-\mathcal{T}^{L(\omega,\omega^*_\beta(m))}(\omega)$, the channel is in the bad state in this slot, and the total future discounted passive time is $D_{\beta,m}(p_{01})$.

By considering $\omega=p_{01}$ and $\omega=p_{11}$, both $D_{\beta,m}(p_{01})$ and $D_{\beta,m}(p_{11})$ can be written as functions of themselves through (27). We can thus solve for $D_{\beta,m}(p_{01})$ and $D_{\beta,m}(p_{11})$ as given in Lemma 4.

*Lemma 4:* Let $\omega^*_\beta(m)$ denote the threshold of the optimal policy for the single-armed bandit process with subsidy $m$. The total discounted passive times $D_{\beta,m}(p_{01})$ and $D_{\beta,m}(p_{11})$ are given as follows.

$$V_{\beta,m}(p_{01})=\begin{cases}\dfrac{p_{01}}{(1-\beta)(1-\beta p_{11}+\beta p_{01})}, & \text{if } \omega^*_\beta(m)<p_{01}\\ \dfrac{(1-\beta p_{11})(1-\beta^{L(p_{01},\omega^*_\beta(m))})m+(1-\beta)\beta^{L(p_{01},\omega^*_\beta(m))}\mathcal{T}^{L(p_{01},\omega^*_\beta(m))}(p_{01})}{(1-\beta p_{11})(1-\beta)(1-\beta^{L(p_{01},\omega^*_\beta(m))+1})+(1-\beta)^2\beta^{L(p_{01},\omega^*_\beta(m))+1}\mathcal{T}^{L(p_{01},\omega^*_\beta(m))}(p_{01})}, & \text{if } p_{01}\le\omega^*_\beta(m)<\omega_o\\ \dfrac{m}{1-\beta}, & \text{if } \omega^*_\beta(m)\ge\omega_o\end{cases}\tag{23}$$

$$V_{\beta,m}(p_{11})=\begin{cases}\dfrac{p_{11}+\beta(1-p_{11})V_{\beta,m}(p_{01})}{1-\beta p_{11}}, & \text{if } \omega^*_\beta(m)<p_{11}\\ \dfrac{m}{1-\beta}, & \text{if } \omega^*_\beta(m)\ge p_{11}.\end{cases}\tag{24}$$

$$V_{\beta,m}(p_{11})=\begin{cases}\dfrac{p_{11}(1-\beta)+\beta p_{01}}{(1-\beta)(1-\beta p_{11}+\beta p_{01})}, & \text{if } \omega^*_\beta(m)<p_{11}\\ \dfrac{m(1-\beta(1-p_{01}))+\beta\mathcal{T}(p_{11})(1-\beta)+\beta^2p_{01}}{1-\beta(1-p_{01})-\beta^2\mathcal{T}(p_{11})(1-\beta)-\beta^3p_{01}}, & \text{if } p_{11}\le\omega^*_\beta(m)<\mathcal{T}(p_{11})\\ \dfrac{m}{1-\beta}, & \text{if } \omega^*_\beta(m)\ge\mathcal{T}(p_{11})\end{cases}\tag{25}$$

$$V_{\beta,m}(p_{01})=\begin{cases}\dfrac{p_{01}+\beta p_{01}V_{\beta,m}(p_{11})}{1-\beta(1-p_{01})}, & \text{if } \omega^*_\beta(m)<p_{01}\\ \dfrac{m}{1-\beta}, & \text{if } \omega^*_\beta(m)\ge p_{01}.\end{cases}\tag{26}$$

Case 1) *Positively correlated channel* ($p_{11} \geq p_{01}$) [see (28)–(29), shown at the bottom of the page].

Case 2) *Negatively correlated channel* ($p_{11} < p_{01}$) [see (30)–(31), shown at the bottom of the page].

*Proof:* The process of solving for $D_{\beta,m}(p_{01})$ and $D_{\beta,m}(p_{11})$ is similar to that of solving for $V_{\beta,m}(p_{01})$ and $V_{\beta,m}(p_{11})$. Details are omitted. $D_{\beta,m}(p_{01})$ and $D_{\beta,m}(p_{11})$ can also be obtained by taking the derivatives of $V_{\beta,m}(p_{01})$ and $V_{\beta,m}(p_{11})$ with respect to $m$. □

We point out that $V_{\beta,m}(\omega)$ is not differentiable in $m$ at every point (i.e., the left derivative may not equal to the right derivative). Suppose that $V_{\beta,m}(\omega)$ is not differentiable at $m_0$. Then, it can be shown that the left derivative at $m_0$ corresponds to the case when the threshold $\omega^*_\beta(m_0)$ is included in the active set while the right derivative corresponds to the case when $\omega^*_\beta(m_0)$ is included in the passive set. In this paper, we include the threshold in the passive set [see (11)], i.e., we choose the passive action when both actions are optimal. As a consequence, we consider the right derivative of $V_{\beta,m}(\omega)$ when it is not differentiable.

The following lemma shows the piecewise constant (a stair function) and monotonically increasing properties of $D_{\beta,m}(\omega)$ as a function of $m$ (see an illustration in Figs. 10 and 11). These properties allow us to develop an efficient algorithm for computing a performance upper bound as shown in Section IV-F.

*Lemma 5:* The total discounted passive time $D_{\beta,m}(\omega)$ as a function of $m$ is monotonically increasing and piecewise constant (with countable pieces for $p_{11} \geq p_{01}$ and finite pieces for $p_{11} < p_{01}$). Equivalently, the value function $V_{\beta,m}(\omega)$ is piecewise linear and convex in $m$.

*Proof:* The piecewise constant property follows directly from (27) and Lemma 4. The monotonicity of $D_{\beta,m}(\omega)$ applies to a general restless bandit and has been stated without proof by Whittle [5]. We provide a proof below for completeness.

We show that $V_{\beta,m}(\omega)$ is convex in $m$, i.e., for any $0 \leq \alpha \leq 1, m_1, m_2 \in \mathcal{R}$

$$\alpha V_{\beta,m_1}(\omega) + (1-\alpha)V_{\beta,m_2}(\omega) \geq V_{\beta,\alpha m_1+(1-\alpha)m_2}(\omega). \quad (32)$$

Consider the optimal policy $\pi$ under subsidy $\alpha m_1+(1-\alpha)m_2$. If we apply $\pi$ to the system with subsidy $m_1$, the total discounted reward will be

$$V_{\beta,\alpha m_1+(1-\alpha)m_2}(\omega)+D_{\beta,\alpha m_1+(1-\alpha)m_2}(\omega)((1-\alpha)(m_1-m_2)).$$

Since $\pi$ may not be the optimal policy under subsidy $m_1$, we have

$$V_{\beta,m_1}(\omega) \geq V_{\beta,\alpha m_1+(1-\alpha)m_2}(\omega) + D_{\beta,\alpha m_1+(1-\alpha)m_2}(\omega)((1-\alpha)(m_1-m_2)). \quad (33)$$

Similarly

$$V_{\beta,m_2}(\omega) \geq V_{\beta,\alpha m_1+(1-\alpha)m_2}(\omega) + D_{\beta,\alpha m_1+(1-\alpha)m_2}(\omega)(\alpha(m_2-m_1)). \quad (34)$$

Equation (32) thus follows from (33) and (34). □

$$D_{\beta,m}(p_{01}) = \begin{cases} 0, & \text{if } \omega^*_\beta(m) < p_{01} \\ \dfrac{(1-\beta p_{11})\left(1-\beta^{L(p_{01},\omega^*_\beta(m))}\right)}{(1-\beta p_{11})(1-\beta)\left(1-\beta^{L(p_{01},\omega^*_\beta(m))+1}\right)+(1-\beta)^2\beta^{L(p_{01},\omega^*_\beta(m))+1}\mathcal{T}^{L(p_{01},\omega^*_\beta(m))}(p_{01})}, & \text{if } p_{01} \leq \omega^*_\beta(m) < \omega_o \\ \dfrac{1}{1-\beta}, & \text{if } \omega^*_\beta(m) \geq \omega_o \end{cases} \quad (28)$$

$$D_{\beta,m}(p_{11}) = \begin{cases} \dfrac{\beta(1-p_{11})D_{\beta,m}(p_{01})}{1-\beta p_{11}}, & \text{if } \omega^*_\beta(m) < p_{11} \\ \dfrac{1}{1-\beta}, & \text{if } \omega^*_\beta(m) \geq p_{11}. \end{cases} \quad (29)$$

$$D_{\beta,m}(p_{11}) = \begin{cases} 0, & \text{if } \omega^*_\beta(m) < p_{11} \\ \dfrac{1-\beta(1-p_{01})}{1-\beta(1-p_{01})-\beta^2\mathcal{T}(p_{11})(1-\beta)-\beta^3 p_{01}}, & \text{if } p_{11} \leq \omega^*_\beta(m) < \mathcal{T}(p_{11}) \\ \dfrac{1}{1-\beta}, & \text{if } \omega^*_\beta(m) \geq \mathcal{T}(p_{11}) \end{cases} \quad (30)$$

$$D_{\beta,m}(p_{01}) = \begin{cases} \dfrac{\beta p_{01} D_{\beta,m}(p_{11})}{1-\beta(1-p_{01})}, & \text{if } \omega^*_\beta(m) < p_{01} \\ \dfrac{1}{1-\beta}, & \text{if } \omega^*_\beta(m) \geq p_{01}. \end{cases} \quad (31)$$

### E. Indexability and Whittle Index Policy

With the threshold structure of the optimal policy and the closed-form expressions of the value function and discounted passive time, we are ready to establish the indexability and solve for Whittle index.

*Theorem 1:* The RMAB $(\Omega(1), \{\mathbf{P}_i\}_{i=1}^N, \{B_i\}_{i=1}^N, \beta)$ is indexable.

*Proof:* The proof is based on Lemmas 2 and 4. Details are given in Appendix II. □

*Theorem 2:* Whittle index $W_\beta(\omega) \in \mathbb{R}$ for arm $i$ of the RMAB $(\Omega(1), \{\mathbf{P}_i\}_{i=1}^N, \{B_i\}_{i=1}^N, \beta)$ is given as follows.

Case 1) *Positively correlated channel* ($p_{11}^{(i)} \geq p_{01}^{(i)}$). See (35), shown at the bottom of the page.

Case 2) *Negatively correlated channel* ($p_{11}^{(i)} < p_{01}^{(i)}$). See (36), shown at the bottom of the page.

*Proof:* By the definition of Whittle index, for a given belief state $\omega$, its Whittle index is the subsidy $m$ that is the solution to the following equation of $m$:

$$\underbrace{\omega + \beta(\omega V_{\beta,m}(p_{11}) + (1-\omega)V_{\beta,m}(p_{01}))}_{V_{\beta,m}(\omega;u=1)} = \underbrace{m + \beta V_{\beta,m}(\mathcal{T}^1(\omega))}_{V_{\beta,m}(\omega;u=0)}. \quad (37)$$

From the closed-form expressions for $V_{\beta,m}(p_{11})$, $V_{\beta,m}(p_{01})$, and $V_{\beta,m}(\mathcal{T}^1(\omega))$ given in Lemma 3, we can solve (37) and obtain Whittle index. □

The following properties of Whittle index $W_\beta(\omega)$ follow from Theorems 1 and 2.

$$W_\beta(\omega) = \begin{cases} \omega B_i, & \text{if } \omega \leq p_{01}^{(i)} \text{ or } \omega \geq p_{11}^{(i)} \\ \dfrac{\omega}{1-\beta p_{11}^{(i)} + \beta\omega} B_i, & \text{if } \omega_o^{(i)} \leq \omega < p_{11}^{(i)} \\ \dfrac{\omega - \beta\mathcal{T}^1(\omega) + C_2(1-\beta)\left(\beta\left(1-\beta p_{11}^{(i)}\right) - \beta(\omega - \beta\mathcal{T}^1(\omega))\right)}{1-\beta p_{11}^{(i)} - C_1\left(\beta\left(1-\beta p_{11}^{(i)}\right) - \beta(\omega-\beta\mathcal{T}^1(\omega))\right)} B_i, & \text{if } p_{01}^{(i)} < \omega < \omega_o^{(i)} \end{cases} \quad (35)$$

where

$$C_1 = \frac{\left(1-\beta p_{11}^{(i)}\right)\left(1-\beta^{L\left(p_{01}^{(i)},\omega\right)}\right)}{\left(1-\beta p_{11}^{(i)}\right)\left(1-\beta^{L\left(p_{01}^{(i)},\omega\right)+1}\right) + (1-\beta)\beta^{L\left(p_{01}^{(i)},\omega\right)+1}\mathcal{T}^{L\left(p_{01}^{(i)},\omega\right)}\left(p_{01}^{(i)}\right)}$$

$$C_2 = \frac{\beta^{L\left(p_{01}^{(i)},\omega\right)}\mathcal{T}^{L\left(p_{01}^{(i)},\omega\right)}\left(p_{01}^{(i)}\right)}{\left(1-\beta p_{11}^{(i)}\right)\left(1-\beta^{L\left(p_{01}^{(i)},\omega\right)+1}\right) + (1-\beta)\beta^{L\left(p_{01}^{(i)},\omega\right)+1}\mathcal{T}^{L\left(p_{01}^{(i)},\omega\right)}\left(p_{01}^{(i)}\right)}.$$

$$W_\beta(\omega) = \begin{cases} \omega B_i, & \text{if } \omega \leq p_{11}^{(i)} \text{ or } \omega \geq p_{01}^{(i)} \\ \dfrac{\beta p_{01}^{(i)} + \omega(1-\beta)}{1+\beta\left(p_{01}^{(i)} - \omega\right)} B_i, & \text{if } \mathcal{T}^1\left(p_{11}^{(i)}\right) \leq \omega < p_{01}^{(i)} \\ \dfrac{(1-\beta+\beta C_4)\left(\beta p_{01}^{(i)} + \omega(1-\beta)\right)}{1-\beta\left(1-p_{01}^{(i)}\right) - C_3\left(\beta^2 p_{01}^{(i)} + \beta\omega - \beta^2\omega\right)} B_i, & \text{if } \omega_o^{(i)} \leq \omega < \mathcal{T}^1\left(p_{11}^{(i)}\right) \\ \dfrac{(1-\beta)\left(\beta p_{01}^{(i)} + \omega - \beta\mathcal{T}^1(\omega)\right) - C_4\beta\left(\beta\mathcal{T}^1(\omega) - \beta p_{01}^{(i)} - \omega\right)}{1-\beta\left(1-p_{01}^{(i)}\right) + C_3\beta\left(\beta\mathcal{T}^1(\omega) - \beta p_{01}^{(i)} - \omega\right)} B_i, & \text{if } p_{11}^{(i)} < \omega < \omega_o^{(i)} \end{cases} \quad (36)$$

where

$$C_3 = \frac{1-\beta(1-p_{01}^{(i)})}{1+(1+\beta)\beta p_{01}^{(i)} - \beta^2\mathcal{T}^1(p_{11}^{(i)})} \quad \text{and} \quad C_4 = \frac{\beta\mathcal{T}^1(p_{11}^{(i)})(1-\beta) + \beta^2 p_{01}^{(i)}}{1+(1+\beta)\beta p_{01}^{(i)} - \beta^2\mathcal{T}^1(p_{11}^{(i)})}.$$

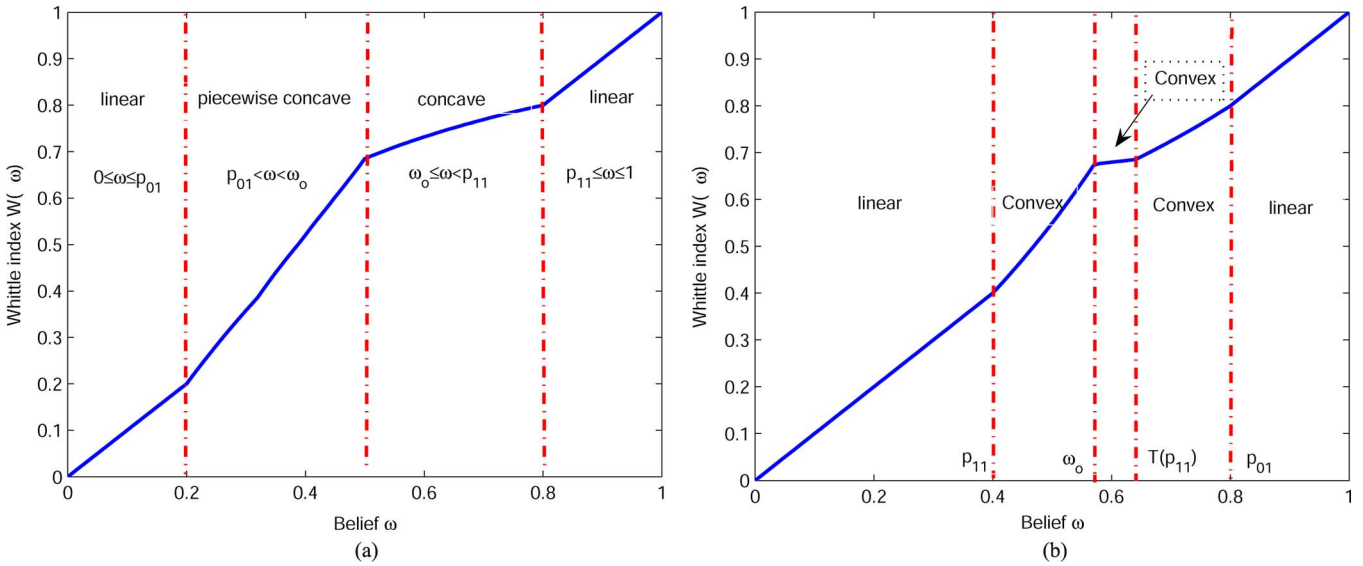


Fig. 8. Whittle index: (a) $p_{11} = 0.8, p_{01} = 0.2, \beta = 0.9$; (b) $p_{11} = 0.4, p_{01} = 0.8, \beta = 0.9$.

*Corollary 1 (Properties of Whittle Index):*

- $W_\beta(\omega)$ is a monotonically increasing function of $\omega$. As a consequence, Whittle index policy is equivalent to the myopic policy for stochastically identical arms.
- For a positively correlated channel $(p_{11} \geq p_{01})$, $W_\beta(\omega)$ is piecewise concave with countable pieces. More specifically, $W_\beta(\omega)$ is linear in $[0, p_{01}]$ and $[p_{11}, 1]$, concave in $[\omega_o, p_{11})$, and piecewise concave with countable pieces in $(p_{01}, \omega_0)$ [see Fig. 8(a)].
- For a negatively correlated channel $(p_{11} < p_{01})$, $W_\beta(\omega)$ is piecewise convex with finite pieces. More specifically, $W_\beta(\omega)$ is linear in $[0, p_{11}]$ and $[p_{01}, 1]$, concave in $(p_{11}, \omega_o)$, $[\omega_o, \mathcal{T}(p_{11}))$, and $[\mathcal{T}(p_{11}), p_{01})$ [see Fig. 8(b)].

The equivalence between Whittle index policy and the myopic policy is particularly important. It allows us to establish the structure and optimality of Whittle index policy by examining the myopic policy which has a very simple index form. The following theorem shows that this equivalence can be extended to a general RMAB under certain conditions.

*Theorem 3:* For a general RMAB, Whittle's index policy is equivalent to the myopic policy under the following conditions.

1) Arms are stochastically identical.
2) The optimal policy for the single-armed bandit with subsidy has the following threshold structure on the reward space: under any subsidy $m$, the expected immediate reward obtained from any state $\omega$ in the active set is no less than that obtained from any state $\omega'$ in the passive set, i.e., $\forall m$

$$\mathbb{E}[R(\omega)] \geq \mathbb{E}[R(\omega')], \quad \text{for all } \omega \in (\mathcal{P}(m))^C, \qquad \omega' \in \mathcal{P}(m)$$

where $(\mathcal{P}(m))^C$ denote the complement of passive set $\mathcal{P}(m)$.
3) The RMAB is indexable.

*Proof:* Based on the second and third conditions, Whittle index of a state is monotonically increasing with the reward of this state, leading to the equivalence between Whittle index policy and the myopic policy for stochastically identical arms.□

Note that for the class of RMAB considered here, the region of $[p_{01}, \omega_o)$ for a positively correlated arm is the most complex. The infinite but countable concave pieces of Whittle index in this region correspond to each possible value of the first crossing time $L(p_{01}, \omega) \in \{1, 2, \cdots\}$. This region presents most of the difficulties in analyzing the performance of Whittle index policy as shown in Section IV-F.

### F. Performance of Whittle Index Policy

*1) The Optimality of Whittle Index Policy Under a Relaxed Constraint:* Whittle index policy is the optimal solution to a Lagrangian relaxation of RMABs [5]. Specifically, the number of activated arms can vary over time provided that its discounted average over the infinite horizon equals to $K$. Let $K(t)$ denote the number of arms activated in slot $t$. The relaxed constraint is given by

$$\mathbb{E}_\pi[(1-\beta)\sum_{t=1}^{\infty}\beta^{t-1}K(t)] = K. \tag{38}$$

Let $\bar{V}_\beta(\Omega(1))$ denote the maximum expected total discounted reward that can be obtained under this relaxed constraint when the initial belief vector is $\Omega(1)$. Based on the Lagrangian multiplier theorem, we have [5]

$$\bar{V}_\beta(\Omega(1)) = \inf_m \left\{ \sum_{i=1}^{N} V_{\beta,m}^{(i)}(\omega_i(1)) - m\frac{(N-K)}{1-\beta} \right\} \tag{39}$$

where $V_{\beta,m}^{(i)}(\omega)$ is the value function of the single-armed bandit process with subsidy $m$ that corresponds to the $i$th channel.

The above equation reveals the role of the subsidy $m$ as the Lagrangian multiplier and the optimality of Whittle index policy for RMABs under the relaxed constraint given in (38). Specifically, under the relaxed constraint, Whittle index policy is implemented by activating, in each slot, those arms whose current states have a Whittle index greater than a constant $m^*$. This constant $m^*$ is the Lagrangian multiplier that makes the relaxed constraint given in (38) satisfied, or equivalently, the Lagrangian multiplier that achieves the infimum in (39). It is not difficult to see that Whittle index policy implemented by comparing to a

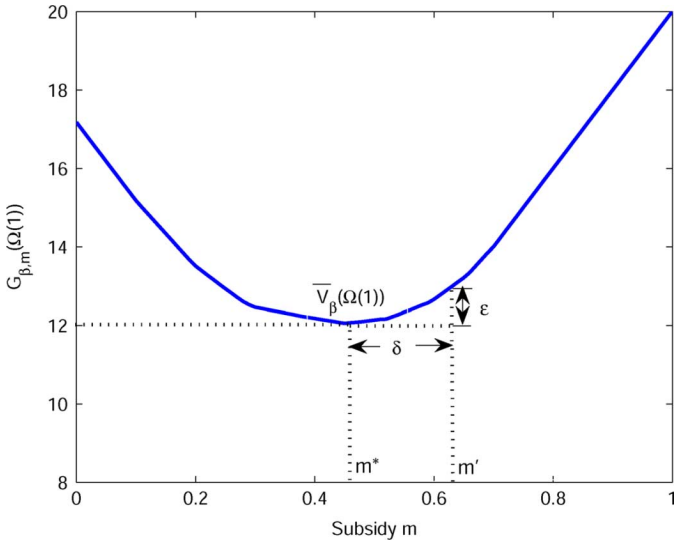


Fig. 9. The optimal performance under relaxed constraint ($N = 8, M = 4$, $\{p_{01}^{(i)}\}_{i=1}^{8} = [0.2, 0.5, 0.8, 0.1, 0.6, 0.2, 0.3, 0.8]$, $\{p_{11}^{(i)}\}_{i=1}^{8} = [0.4, 0.1, 0.3, 0.6, 0.2, 0.8, 0.7, 0.6]$, $B_i = 1$for all $i = 1, \ldots, 8, \beta = 0.8$).

constant $m^*$ is the optimal policy [i.e., achieves $\bar{V}_\beta(\Omega(1))$] for RMABs under the relaxed constraint.

*2) An Upper Bound of the Optimal Performance:* Under the strict constraint of $K(t) = K$ for all $t$, Whittle index policy is implemented by activating those $K$ arms with the largest indices in each slot. Its optimality may be lost.

Let $V_\beta(\Omega(1))$ denote the maximum expected total discounted reward of the RMAB under the strict constraint that $K(t) = K$ for all $t$. It is obvious that

$$V_\beta(\Omega(1)) \le \bar{V}_\beta(\Omega(1)).$$

$\bar{V}_\beta(\Omega(1))$ thus provides a performance benchmark for all RMAB policies, including Whittle index policy. Unfortunately, $\bar{V}_\beta(\Omega(1)$ as given in (39) is, in general, difficult to obtain due to the complexity of calculating the value functions of all arms and searching for the infimum over an uncountable space. For the problem at hand, however, we have obtained $V_{\beta,m}^{(i)}(\omega_i(1))$ in closed form as given in Lemma 3. Furthermore, the piecewise constant structure of the discounted passive time $D_{\beta,m}^{(i)}(\omega_i(1))$ given in Lemma 5 leads to efficient algorithms for searching for the infimum of the value functions over $m$ as shown below.

Let

$$G_{\beta,m}(\Omega(1)) = \sum_{i=1}^{N} V_{\beta,m}^{(i)}(\omega_i(1)) - m\frac{(N-K)}{1-\beta}.$$

We then have $\bar{V}_\beta(\Omega(1)) = \inf_m G_{\beta,m}(\Omega(1), m)$. From Lemma 5, it is easy to see that $G_{\beta,m}(\Omega(1))$ is convex in $m$ as illustrated in Fig. 9. The infimum of $G_{\beta,m}(\Omega(1))$ is achieved at $m^*$ at which the derivative of $G_{\beta,m}(\Omega(1))$ with respect to $m$ becomes nonnegative for the first time [note that $G_{\beta,m}(\Omega(1))$ is not differentiable at every $m$, and we consider the right derivative when it is not differentiable]. Equivalently

$$m^* = \sup\left\{m : \frac{d(G_{\beta,m}(\Omega(1)))}{dm} = \Sigma_{i=1}^{N} D_{\beta,m}^{(i)}(\omega_i(1)) - \frac{(N-K)}{1-\beta} \le 0\right\}.$$

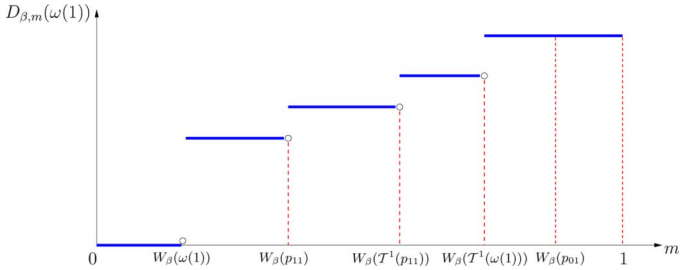


Fig. 10. The passive time for different regions ($p_{11} < p_{01}$).

From Lemma 5, $D_{\beta,m}^{(i)}(\omega_i(1))$ is piecewise constant for each channel (see Figs. 10 and 11). We can thus partition the range of $m$ into disjoint regions such that $\frac{d(G_{\beta,m}(\Omega(1)))}{dm}$ is constant in each region. To obtain $m^*$, we only need to check each region successively until $\frac{d(G_{\beta,m}(\Omega(1)))}{dm}$ becomes nonnegative for the first time [due to the monotonically increasing property of $D_{\beta,m}^{(i)}(\omega_i(1))$ in $m$]. The difficulty is that for a positively correlated channel, there are infinite constant regions of $D_{\beta,m}^{(i)}(\omega_i(1))$ (see Fig. 11). However, we can find an arbitrarily small interval $(W_\beta(\bar{\omega}), W_\beta(\omega)]$—referred to as the gray area—outside which there are only finite number of constant regions of $D_{\beta,m}^{(i)}(\omega_i(1))$. By setting the gray area for each positively correlated channel small enough, we can find an $m'$ that is arbitrarily close to $m^*$ so that $G_{\beta,m'}(\Omega(1))) - G_{\beta,m^*}(\Omega(1)) \le \epsilon$ for any $\epsilon > 0$. Specifically, we set the length of the gray area for each positively correlated channel to $\frac{\delta}{N}$ (i.e., $W_\beta(\omega_o) - W_\beta(\bar{\omega}) \le \frac{\delta}{N}$) where $\delta = \frac{\epsilon(1-\beta)}{K}$. The total length of the gray area over all channels is thus at most $\delta$, i.e., $m' - m^* \le \delta$. Based on the convexity of $G_{\beta,m}(\Omega(1))$, the maximum derivative of $G_{\beta,m}(\Omega(1))$ for $m^* \le m \le 1$ is achieved at $m = 1$, which is equal to $\frac{K}{1-\beta}$. Thus, we have

$$\begin{aligned} G_{\beta,m'}(\Omega(1))) - G_{\beta,m^*}(\Omega(1)) &\le \frac{K}{1-\beta}(m' - m^*) \\ &\le \frac{\delta K}{1-\beta} = \epsilon. \end{aligned}$$

We point out that if $m^*$ does not fall into the gray area, the algorithm will obtain $m^*$ and $\bar{V}_\beta(\Omega(1))$ without error. In the special case when every channel is negatively correlated, the algorithm will always output the exact value of $m^*$ and $\bar{V}_\beta(\Omega(1))$. The detailed algorithm is given in Fig. 12. The complexity of this algorithm is given in Theorem 4.

*Theorem 4:* For any $\epsilon > 0$, the algorithm given in Fig. 12 runs in at most $O(N^2 \log N)$ time to output a value $G$ that is within $\epsilon$ of $\bar{V}_\beta(\Omega(1))$ for any $\epsilon > 0$.

*Proof:* See Appendix III. □

To find the infimum of $G_\beta(\Omega(1), m)$, we can also carry out a binary search on subsidy $m$. It can be shown that this algorithm runs in $O(N(\log N)^2)$ time. However, it cannot output the exact value of $m^*$ and $\bar{V}_\beta(\Omega(1))$.

Fig. 13 shows an example of the performance of Whittle index policy. It demonstrates the near-optimal performance of Whittle index policy and the tightness of the performance upper bound.

## V. Whittle Index Under Average Reward Criterion

In this section, we investigate Whittle index policy under the average reward criterion and establish results parallel to those obtained under the discounted reward criterion in Section IV.

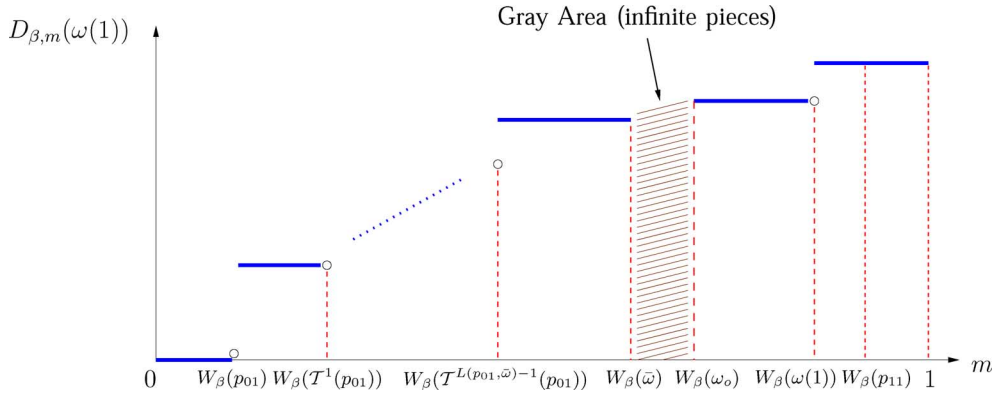


Fig. 11. The passive time for different regions ($p_{11} \geq p_{01}$).

**Computing the Performance Upper Bound within $\epsilon$-Accuracy**

Input an $\epsilon > 0$. Set $\delta = \frac{\epsilon(1-\beta)}{K}$ and $j = 0$.

1) For each negatively correlated channel $i$, calculate $W_\beta(p_{11}^{(i)})$, $W_\beta(p_{01}^{(i)})$, and $W_\beta(\mathcal{T}(p_{11}^{(i)}))$. If $\omega_i(1) < \omega_o^{(i)}$, calculate $W_\beta(\omega_i(1))$ and $W_\beta(\mathcal{T}^1(\omega_i(1)))$; otherwise only calculate $W_\beta(\omega_i(1))$.
2) For each positively correlated channel $i$, calculate $W_\beta(p_{01}^{(i)})$, $W_\beta(p_{11}^{(i)})$, and $W_\beta(\omega_o^{(i)})$. Search for an $\bar{\omega}^{(i)} \in [\omega_o^{(i)} - \frac{\delta}{N}, \omega_o^{(i)})$ such that $W_\beta(\omega_o^{(i)}) \geq W_\beta(\bar{\omega}^{(i)}) - \frac{\delta}{N}$. Let $l_i$ be the smallest integer such that $\mathcal{T}^{l_i}(p_{01}^{(i)}) > \bar{\omega}^{(i)}$. Calculate $W_\beta(\mathcal{T}^k(p_{01}^{(i)}))$ for all $1 \leq k \leq l_i$. If $\omega_i(1) < \omega_o^{(i)}$, then let $d_i$ be the smallest integer such that $\mathcal{T}^{d_i}(\omega_i(1)) > \bar{\omega}^{(i)}$ and calculate $W_\beta(\mathcal{T}^k(\omega_i(1)))$ for all $1 \leq k \leq d_i$; otherwise only calculate $W_\beta(\omega_i(1))$. Set the gray area $V = \cup_i[\min\{W_\beta(\mathcal{T}^{l_i}(p_{01}^{(i)})), W_\beta(\mathcal{T}^{d_i}(\omega_i(1)))\}, W_\beta(\omega_o^{(i)}))$.
3) Order all Whittle indices calculated in Step 1 and 2 by the ascending order. Let $[a_1, ...a_h]$ denote the ordered Whittle indices. Set $a_0 = 0$ and $a_{h+1} = 1$.
4) If $[a_j, a_{j+1}) \nsubseteq V$, calculate $D = \Sigma_{k=1}^N D_{\beta,m}^{(k)}(\omega_k(1)) - \frac{(N-K)}{1-\beta}$ for $m \in [a_j, a_{j+1})$ according to (27) (note that every $D_{\beta,m}^{(k)}(\omega_k(1))$ is constant for $m \in [a_j, a_{j+1})$), if $D$ is nonnegative, go to Step 5; otherwise set $j = j + 1$ and repeat Step 4.
5) Calculate $G = G_{\beta,m}(\Omega(1))$ when $m \in [a_j, a_{j+1})$ according to (22). Output $m' = a_j$ and $G$.

Fig. 12. Algorithm for computing the upper bound of the optimal performance.

### A. The Value Function and the Optimal Policy

First, we present a general result by Dutta [35] on the relationship between the value function and the optimal policy under the total discounted reward criterion and those under the average reward criterion. This result allows us to study Whittle index policy under the average reward criterion by examining its limiting behavior under the discounted reward criterion as the discount factor $\beta \to 1$.

*Dutta's Theorem* [35]. Let $\mathcal{F}$ be the belief space of a POMDP and $V_\beta(\Omega)$ the value function with discount factor $\beta$ for belief $\Omega \in \mathcal{F}$. The POMDP satisfies the value boundedness condition if there exist a belief $\Omega'$, a real-valued function $c_1(\Omega) : \mathcal{F} \to \mathcal{R}$, and a constant $c_2 < \infty$ such that

$$c_1(\Omega) \leq V_\beta(\Omega) - V_\beta(\Omega') \leq c_2$$

for any $\Omega \in \mathcal{F}$ and $\beta \in [0, 1)$. Under the value-boundedness condition, if a series of optimal policies $\pi_{\beta_k}$ for a POMDP with discount factor $\beta_k$ pointwise converges to a limit $\pi^*$ as $\beta_k \to 1$, then $\pi^*$ is the optimal policy for the POMDP under the average reward criterion. Furthermore, let $J(\Omega)$ denote the maximum expected average reward over the infinite horizon starting from the initial belief $\Omega$. We have

$$J(\Omega) = \lim_{\beta_k \to 1}(1 - \beta_k)V_{\beta_k}(\Omega)$$

and $J(\Omega) = J$ is independent of the initial belief $\Omega$.

Next, we will show that the single-armed bandit process with subsidy $m$ under the discounted reward criterion (see Section III-B) satisfies the valueboundedness condition.

*Lemma 6:* The single-armed bandit process with subsidy under the discounted reward criterion satisfies the value-boundedness condition. More specifically, we have[3]

$$|V_{\beta,m}(\omega) - V_{\beta,m}(\omega')| \leq c + 1, \qquad \text{for all } \omega, \omega' \in [0, 1] \quad (40)$$

where $c = \max\{\frac{2}{1-p_{11}}, \frac{2}{p_{01}}\}$.

*Proof:* See Appendix IV. □

Under the value boundedness condition, the optimal policy for the single-armed bandit process with subsidy under the average reward criterion can be obtained from the limit of any pointwise convergent series of the optimal policies under the discounted reward criterion. Lemma 7 shows that the optimal

[3]Here we do not consider the trivial case that the arm has absorbing states.

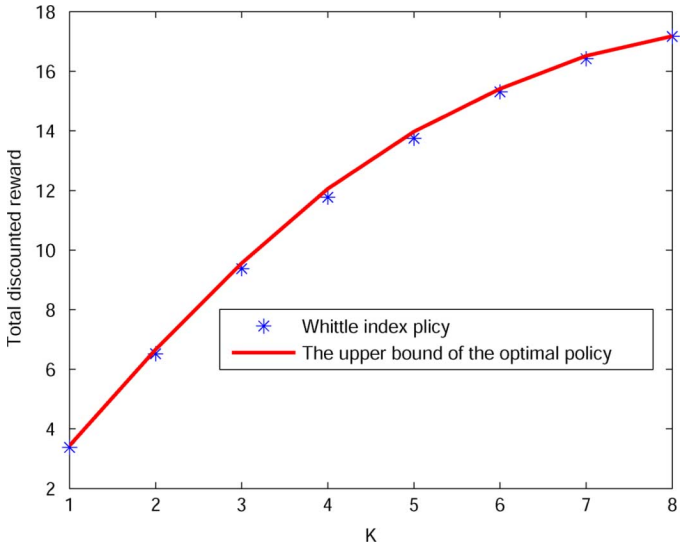


Fig. 13. The performance of Whittle index policy ($N = 8, \{p_{01}^{(i)}\}_{i=1}^{8} = \{0.2, 0.5, 0.8, 0.1, 0.6, 0.2, 0.3, 0.8\}$, $\{p_{11}^{(i)}\}_{i=1}^{8} = \{0.4, 0.1, 0.3, 0.6, 0.2, 0.8, 0.7, 0.6\}$, $B_i = 1$ for $i = 1, \ldots, 8$, and $\beta = 0.8$).

policy for the single-armed bandit process with subsidy under the average reward criterion is also a threshold policy.

*Lemma 7:* Let $\omega_\beta^*(m)$ denote the threshold of the optimal policy for the single-armed bandit process with subsidy $m$ under the discounted reward criterion. Then, $\lim_{\beta\to 1}\omega_\beta^*(m)$ exists for any $m$. Furthermore, the optimal policy for the single-armed bandit process with subsidy $m$ under the average reward criterion is also a threshold policy with threshold $\omega^*(m) = \lim_{\beta\to 1}\omega_\beta^*(m)$.

*Proof:* See Appendix V. □

### B. Indexability and Whittle Index Policy

Based on Lemma 7, the RMAB $(\Omega, \{\mathbf{P}_i\}_{i=1}^N, \{B_i\}_{i=1}^N, 1)$ is indexable if the threshold $\omega^*(m)$ of the optimal policy is monotonically increasing with subsidy $m$. Next, we show that the monotonicity holds and the RMAB $(\Omega, \{\mathbf{P}_i\}_{i=1}^N, \{B_i\}_{i=1}^N, 1)$ is indexable. Moreover, we obtain Whittle index in closed form as shown below.

*Theorem 5:* The RMAB $(\Omega(1), \{\mathbf{P}_i\}_{i=1}^N, \{B_i\}_{i=1}^N, 1)$ is indexable with Whittle index $W(\omega)$ given below.

Case 1) *Positively correlated channel* ($p_{11}^{(i)} \geq p_{01}^{(i)}$). See (41), shown at the bottom of the page.

Case 2) *Negatively correlated channel* ($p_{11}^{(i)} < p_{01}^{(i)}$). See (42), shown at the bottom of the page.

*Proof:* See Appendix VI. □

The monotonicity and piecewise concave/convex properties of Whittle index under the discounted reward criterion given in Corollary 1 are preserved under the average reward criterion. The only difference is that Whittle index under the discounted reward criterion is always strictly increasing with the belief state while Whittle index $W(\omega)$ under the average reward criterion is a constant function of $\omega$ when $\omega_o \leq \omega < \mathcal{T}^1(p_{11})$ for a negatively correlated channel [see (42)].

### C. The Performance of Whittle Index Policy

Similarly to the case under the discounted reward criterion, Whittle index policy is optimal under the average reward criterion when the constraint on the number of activated arms $K(t)(t \geq 1)$ is relaxed to the following:

$$\mathbb{E}_\pi\left[\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}K(t)\right] = K.$$

Let $\bar{J}(\Omega(1))$ denote the maximum expected average reward that can be obtained under this relaxed constraint when the initial belief vector is $\Omega(1)$. Based on the Lagrangian multiplier theorem, we have [5]

$$\bar{J} = \inf_m\left\{\sum_{i=1}^{N}J_m^{(i)} - m(N-K)\right\} \tag{43}$$

$$W(\omega) = \begin{cases} \omega B_i, & \text{if } \omega \leq p_{01}^{(i)} \text{ or } \omega \geq p_{11}^{(i)} \\ \dfrac{(\omega - \mathcal{T}^1(\omega))\left(L\left(p_{01}^{(i)},\omega\right)+1\right) + \mathcal{T}^{L(p_{01}^{(i)},\omega)}\left(p_{01}^{(i)}\right)}{1 - p_{11}^{(i)} + (\omega - \mathcal{T}^1(\omega))L\left(p_{01}^{(i)},\omega\right) + \mathcal{T}^{L(p_{01}^{(i)},\omega)}\left(p_{01}^{(i)}\right)}B_i, & \text{if } p_{01}^{(i)} < \omega < \omega_o^{(i)} \\ \dfrac{\omega}{1 - p_{11}^{(i)} + \omega}B_i, & \text{if } \omega_o^{(i)} \leq \omega < p_{11}^{(i)}. \end{cases} \tag{41}$$

$$W(\omega) = \begin{cases} \omega B_i, & \text{if } \omega \leq p_{11}^{(i)} \text{ or } \omega \geq p_{01}^{(i)} \\ \dfrac{\omega + p_{01}^{(i)} - \mathcal{T}^1(\omega)}{1 + p_{01}^{(i)} - \mathcal{T}^1\left(p_{11}^{(i)}\right) + \mathcal{T}^1(\omega) - \omega}B_i, & \text{if } p_{11}^{(i)} < \omega < \omega_o^{(i)} \\ \dfrac{p_{01}^{(i)}}{1 + p_{01}^{(i)} - \mathcal{T}^1\left(p_{11}^{(i)}\right)}B_i, & \text{if } \omega_o^{(i)} \leq \omega < \mathcal{T}^1\left(p_{11}^{(i)}\right) \\ \dfrac{p_{01}^{(i)}}{1 + p_{01}^{(i)} - \omega}B_i, & \text{if } \mathcal{T}^1\left(p_{11}^{(i)}\right) \leq \omega < p_{01}^{(i)}. \end{cases} \tag{42}$$

where $J_m^{(i)}$ is the value function of the single-armed bandit process with subsidy $m$ that corresponds to the $i$th channel.

Let $J(\Omega(1))$ denote the maximum expected average reward of the RMAB under the strict constraint that $K(t) = K$ for all $t$. Obviously

$$J(\Omega(1)) \le \bar{J}.$$

$\bar{J}$ thus provides a performance benchmark for Whittle index policy under the strict constraint. To evaluate $\bar{J}$, we consider the single-armed bandit with subsidy $m$ under the average reward criterion. The value function $J_m$ and the average passive time $D_m = \frac{d(J_m)}{dm}$ can be obtained in closed form as shown in Lemma 8.

*Lemma 8:* The value function $J_m$ and $D_m$ can be obtained in closed form as given below, where $\omega^*(m)$ is the threshold of the optimal policy. Furthermore, $D_m$ is piecewise constant and increasing with $m$. See (44)–(45), shown at the bottom of the page.

*Proof:* Under the value-boundedness condition as shown in Section V-A, we have, according to Dutta's theorem

$$J_m = \lim_{\beta_k \to 1} (1-\beta_k) V_{\beta_k}(\omega, m)$$

which leads to (44) directly. The closed-form expression for $D_m$ can be obtained from the derivative of $J_m$ with respect to $m$. The proof that $D_m$ is increasing with $m$ is similar to that given in Lemma 5. □

Based on the closed-form $D_m$ given in Lemma 8, we can obtain the subsidy $m^*$ that achieves the infimum in (43). Specifically, the subsidy $m^*$ that achieves the infimum in (43) is the supremum value of $m \in [0,1]$ satisfying $\sum_{i=1}^{N} D^{m,i} \le N-K$. After obtaining $m^*$, it is easy to calculate the infimum according to the closed-form $J_m$ given in Lemma 8. With minor changes, the algorithm in Section IV-F can be applied to evaluate the upper bound $\bar{J}$. We notice that the initial belief will not be considered in the algorithm, which leads to a shorter running time.

Simulation results similar to Fig. 9 have been observed, demonstrating the near-optimal performance of Whittle index policy under the average reward criterion.

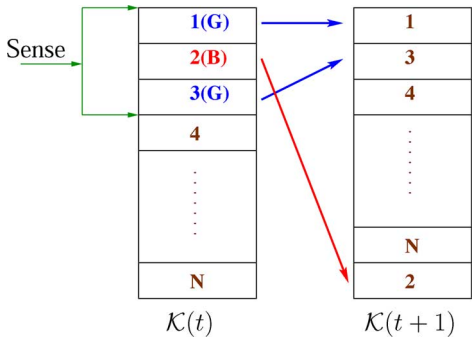


Fig. 14. The structure of Whittle index policy ($p_{11} \ge p_{01}$).

## VI. Whittle Index Policy for Stochastically Identical Channels

Based on the equivalence between Whittle index policy and the myopic policy for stochastically identical arms, we can analyze Whittle index policy by focusing on the myopic policy which has a much simpler index form. In this section, we establish the semiuniversal structure and study the optimality of Whittle index policy for stochastically identical arms.

### A. The Structure of Whittle Index Policy

The implementation of Whittle index policy can be described with a queue structure. Specifically, all $N$ channels are ordered in a queue, and in each slot, those $K$ channels at the head of the queue are sensed. Based on the observations, channels are reordered at the end of each slot according to the following simple rules.

When $p_{11} \ge p_{01}$, the channels observed in state 1 will stay at the head of the queue while the channels observed in state 0 will be moved to the end of the queue (see Fig. 14).

When $p_{11} < p_{01}$, the channels observed in state 0 will stay at the head of the queue while the channels observed in state 1 will be moved to the end of the queue. The order of the unobserved channels is reversed (see Fig. 15).

The initial channel ordering $\mathcal{K}(1)$ is determined by the initial belief vector as given below

$$\begin{aligned} &\omega_{n_1}(1) \ge \omega_{n_2}(1) \ge \cdots \ge \omega_{n_N}(1) \\ &\quad \Longrightarrow \mathcal{K}(1) = (n_1, n_2, \ldots, n_N). \end{aligned} \tag{46}$$

$$J_m = \begin{cases} \omega_o, & \text{if } \omega^*(m) < \min\{p_{01}, p_{11}\} \\ \dfrac{(1-p_{11})L(p_{01},\omega^*(m))m + \mathcal{T}^{L(p_{01},\omega^*(m))}(p_{01})}{(1-p_{11})(L(p_{01},\omega^*(m))+1) + \mathcal{T}^{L(p_{01},\omega^*(m))}(p_{01})}, & \text{if } p_{01} \le \omega^*(m) < \omega_o \\ \dfrac{p_{01}m + p_{01}}{1+2p_{01}-\mathcal{T}^1(p_{11})}, & \text{if } p_{11} \le \omega^*(m) < \mathcal{T}^1(p_{11}) \\ m, & \text{other cases} \end{cases} \tag{44}$$

and

$$D_m = \begin{cases} 0, & \text{if } \omega^*(m) < \min\{p_{01}, p_{11}\} \\ \dfrac{(1-p_{11})L(p_{01},\omega^*(m))}{(1-p_{11})(L(p_{01},\omega^*(m))+1) + \mathcal{T}^{L(p_{01},\omega^*(m))}(p_{01})}, & \text{if } p_{01} \le \omega^*(m) < \omega_o \\ \dfrac{p_{01}}{1+2p_{01}-\mathcal{T}^1(p_{11})}, & \text{if } p_{11} \le \omega^*(m) < \mathcal{T}^1(p_{11}) \\ 1, & \text{other cases.} \end{cases} \tag{45}$$

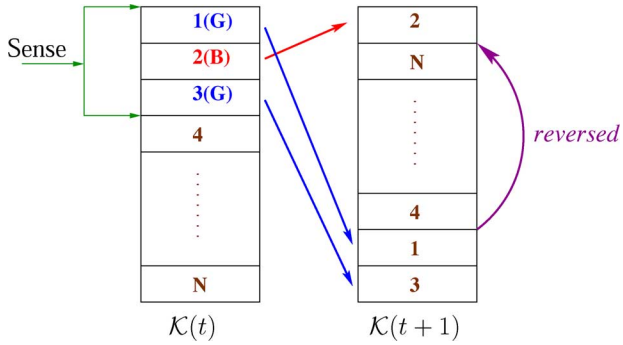


Fig. 15. The structure of Whittle index policy ($p_{11} < p_{01}$).

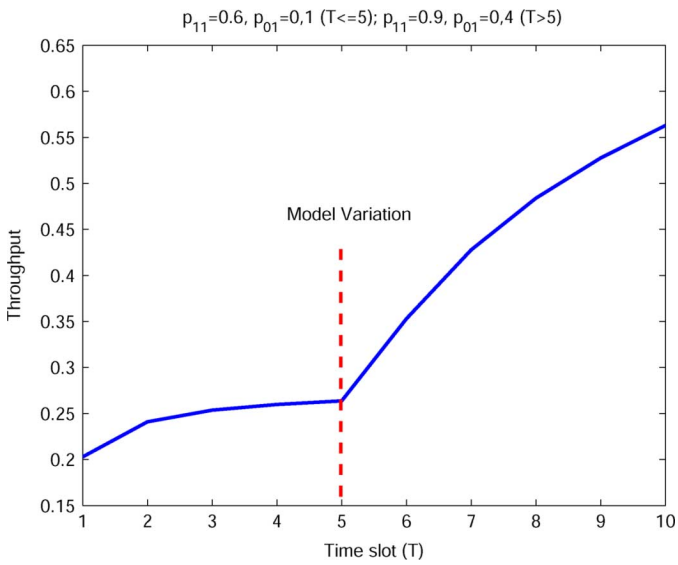


Fig. 16. Tracking the change in channel transition probabilities occurred at $t = 6$.

The proof is similar to that in [17] for the case of $K = 1$ and omitted here.

The advantage of this structure of Whittle index policy is twofold. First, it demonstrates the simplicity of Whittle index policy: channel selection is reduced to maintaining a simple queue structure that requires no computation and little memory. Second, it shows that Whittle index policy has a semiuniversal structure; it can be implemented without knowing the channel transition probabilities except the order of $p_{11}$ and $p_{01}$. As a result, Whittle index policy is robust against model mismatch and automatically tracks variations in the channel model provided that the order of $p_{11}$ and $p_{01}$ remains unchanged. As show in Fig. 16, the transition probabilities change abruptly in the fifth slot, which corresponds to an increase in the occurrence of good channel state in the system. From this figure, we can observe, from the change in the throughput increasing rate, that Whittle index policy effectively tracks the model variations.

### B. Optimality and Approximation Factor of Whittle Index Policy

The optimality of the myopic policy was first shown for $N = 2$ (both positive and negative correlation cases) in [17]. It was then extended to any number of positively correlated channels with single channel sensing ($K = 1$) in [18], and then to arbitrary $K$ in [19]. Based on the equivalence between Whittle index policy and the myopic policy, we conclude that Whittle index policy is optimal for any $N$ and $K$ when $p_{11} \geq p_{01}$.

In this section, we extend the optimality of Whittle index policy to negatively correlated channels with $K = N - 1$. For a general $K$, we establish the approximation factor of Whittle index policy. Furthermore, we characterize the performance of Whittle index policy in terms of the system parameters for both positively and negatively correlated channels. Specifically, we obtain a lower bound and an upper bound on the average reward $J_w$ achieved by Whittle index policy, as given in Theorem 6.

*Theorem 6 (Lower and Upper Bounds of the Performance of Whittle Index Policy):* Recall that $J$ denote the average reward achieved by the optimal policy. We have

$$\frac{K\mathcal{T}^{\lfloor \frac{N}{K} \rfloor - 1}(p_{01})}{1 - p_{11} + \mathcal{T}^{\lfloor \frac{N}{K} \rfloor - 1}(p_{01})} \leq J_w = J \leq \min\left\{ \frac{K\omega_o}{1 - p_{11} + \omega_o}, \omega_o N \right\}, \quad \text{if } p_{11} \geq p_{01} \tag{47}$$

$$\frac{Kp_{01}}{1 - \mathcal{T}^{2\lfloor \frac{N}{K} \rfloor - 2}(p_{11}) + p_{01}} \leq J_w \leq J \leq \min\left\{ \frac{Kp_{01}}{1 - \mathcal{T}^{1}(p_{11}) + p_{01}}, \omega_o N \right\}, \quad \text{if } p_{11} < p_{01}. \tag{48}$$

*Proof:* The upper bound of $J$ is obtained from the upper bound of the optimal performance for generally nonidentical channels as given in (43). The lower bound of $J_w$ is obtained from the structure of Whittle index policy. See Appendix VII for the complete proof. □

*Corollary 2:* Let $\eta = \frac{J_w}{J}$ be the approximation factor defined as the ratio of the performance by Whittle index policy to the optimal performance. We have, under the condition of $p_{11} < p_{01}$

$$\begin{cases} \eta = 1, & \text{for } K = N - 1 \\ \eta \geq \max\left\{ \frac{1}{2}, \frac{K}{N} \right\}, & \text{o.w.} \end{cases}$$

*Proof:* See Appendix VIII. □

## VII. Conclusion

In this paper, we considered a class of RMABs arisen in dynamic multichannel access, user/server scheduling, and optimal activation in multiagent systems. For this class of RMAB, we established the indexability and obtained Whittle index in closed form for both discounted and average reward criteria. The basic approach is on analyzing the optimal passive time for a single arm with subsidy, which extends Whittle's original proof of the indexability for the classical MAB [5]. For stochastically identical arms, we further showed that Whittle index policy is equivalent to the myopic policy that has a simple and robust semiuniversal structure. This equivalence leads to an analytical characterization of the optimality and the performance of Whittle index policy. For nonidentical arms, we developed efficient algorithms for computing a performance upper bound given by

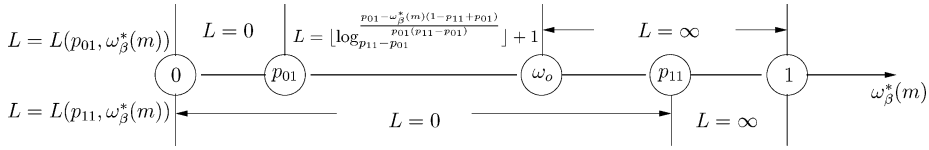


Fig. 17. The first threshold crossing time for different regions of $\omega^*_\beta(m)$ when $p_{11} \geq p_{01}$ (the top partition is for $L(p_{01}, \omega^*_\beta(m))$, the bottom for $L(p_{11}, \omega^*_\beta(m))$).

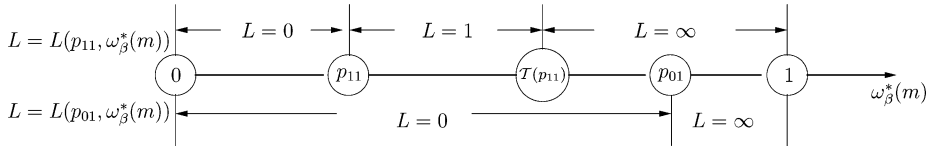


Fig. 18. The first threshold crossing time for different regions of $\omega^*_\beta(m)$ when $p_{11} < p_{01}$ (the top partition is for $L(p_{11}, \omega^*_\beta(m))$, the bottom for $L(p_{01}, \omega^*_\beta(m))$).

Lagrangian relaxation. The tightness of the upper bound and the near-optimal performance of Whittle index policy were illustrated with simulation examples. Recently, this work was extended to non-Markovian arm models in [36], where the indexability, Whittle index, and the asymptotic optimality of Whittle index policy were established under certain conditions.

## Appendix I
## Proof of Lemma 3

From (22), we have

$$V_{\beta,m}(p_{01}) = \frac{1-\beta^{L(p_{01},\omega^*_\beta(m))}}{1-\beta} m + \beta^{L(p_{01},\omega^*_\beta(m))} V_{\beta,m}\left(\mathcal{T}^{L(p_{01},\omega^*_\beta(m))}(p_{01}); u=1\right) \tag{49}$$

$$V_{\beta,m}(p_{11}) = \frac{1-\beta^{L(p_{11},\omega^*_\beta(m))}}{1-\beta} m + \beta^{L(p_{11},\omega^*_\beta(m))} V_{\beta,m}\left(\mathcal{T}^{L(p_{11},\omega^*_\beta(m))}(p_{01}); u=1\right). \tag{50}$$

As shown in (7), $V_{\beta,m}(\mathcal{T}^{L(\omega,\omega^*_\beta(m))}(\omega); u=1)$ is a function of $V_{\beta,m}(p_{01})$ and $V_{\beta,m}(p_{11})$ for any $\omega \in [0,1]$. We thus have (49) and (50) for two unknowns $V_{\beta,m}(p_{01})$ and $V_{\beta,m}(p_{11})$ provided that we can obtain the two first crossing times $L(p_{01}, \omega^*_\beta(m))$ and $L(p_{11}, \omega^*_\beta(m))$.

From (16) and (17), we can obtain these first crossing times by considering different regions that the threshold $\omega^*_\beta(m)$ may lie in (see Figs. 17 and 18). We can thus solve for $V_{\beta,m}(p_{01})$ and $V_{\beta,m}(p_{11})$ from (49) and (50) by considering each region within which both first crossing times $L(p_{01}, \omega^*_\beta(m))$ and $L(p_{11}, \omega^*_\beta(m))$ are constant.

## Appendix II
## Proof of Theorem 1

It suffices to prove that an arm with an arbitrary transition matrix $\mathbf{P}$ is indexable. Based on the threshold structure of the optimal policy for the single-armed bandit with subsidy $m$ given in Lemma 2, indexability is reduced to the monotonicity of the threshold $\omega^*_\beta(m)$, i.e., $\omega^*_\beta(m)$ is monotonically increasing with the subsidy $m$ for $m \in [0,1)$. To prove the monotonicity of $\omega^*_\beta(m)$, we first give Lemma 9.

*Lemma 9:* Suppose that for any $m \in [0,1)$ we have

$$\left.\frac{dV_{\beta,m}(\omega; u=1)}{dm}\right|_{\omega=\omega^*_\beta(m)} < \left.\frac{dV_{\beta,m}(\omega; u=0)}{dm}\right|_{\omega=\omega^*_\beta(m)}. \tag{51}$$

Then, $\omega^*_\beta(m)$ is monotonically increasing with $m$.

We prove Lemma 9 by contradiction. Assume that there exists an $m_0 \in [0,1)$ such that $\omega^*_\beta(m)$ is decreasing at $m_0$. Then, there exists an $\epsilon > 0$ such that for any $\Delta m \in [0, \epsilon]$, we have

$$V_{\beta,m_0+\Delta m}(\omega^*_\beta(m_0); u=1) \geq V_{\beta,m_0+\Delta m}(\omega^*_\beta(m_0); u=0). \tag{52}$$

Since $\omega^*_\beta(m_0)$ is the threshold of the optimal policy under subsidy $m_0$, we have

$$V_{\beta,m_0}(\omega^*_\beta(m_0); u=1) = V_{\beta,m_0}(\omega^*_\beta(m_0); u=0). \tag{53}$$

From (52) and (53), we have

$$\left.\frac{dV_{\beta,m}(\omega; u=1)}{dm}\right|_{\omega=\omega^*_\beta(m_0)} \geq \left.\frac{dV_{\beta,m}(\omega; u=0)}{dm}\right|_{\omega=\omega^*_\beta(m_0)}$$

which contradicts with (51). Lemma 9 thus holds.

According to Lemma 9, it is sufficient to prove (51). Recall that $D_{\beta,m}(\omega) = \frac{d(V_{\beta,m}(\omega))}{dm}$. From (7) and (8), we can write (51) as

$$\beta(\omega^*_\beta(m) D_{\beta,m}(p_{11}) + (1-\omega^*_\beta(m)) D_{\beta,m}(p_{01})) < 1 + \beta D_{\beta,m}(\mathcal{T}^1(\omega^*_\beta(m))). \tag{54}$$

To prove (54), we consider the following three regions of $\omega^*_\beta(m)$.

- **Region 1**: $0 \leq \omega^*_\beta(m) < \min\{p_{01}, p_{11}\}$. Based on the lower bound of the updated belief given in Lemma 1, the arm will be activated in every slot when the initial belief $\omega > \omega^*_\beta(m)$. Thus, $D_{\beta,m}(p_{11}) = D_{\beta,m}(p_{01}) = D_{\beta,m}(\mathcal{T}^1(\omega^*_\beta(m))) = 0$; (54) holds trivially.

- **Region 2**: $\omega_o \le \omega_\beta^*(m) \le 1$. In this region, the arm is made passive in every slot when the initial belief state is $\mathcal{T}^1(\omega_\beta^*(m))$. This is because $\mathcal{T}^k(\omega_\beta^*(m)) \le \omega_\beta^*(m)$ for any $k \ge 1$ (see Lemma 1 and Figs. 3 and 4). Therefore, $D_{\beta,m}(\mathcal{T}^1(\omega_\beta^*(m))) = \frac{1}{1-\beta}$. Since both $D_{\beta,m}(p_{11})$ and $D_{\beta,m}(p_{01})$ are upper bounded by $\frac{1}{1-\beta}$, it is easy to see that (54) holds.
- **Region 3**: $\min\{p_{01}, p_{11}\} \le \omega_\beta^*(m) < \omega_o$. In this region, $\mathcal{T}^1(\omega_\beta^*(m)) > \omega_\beta^*(m)$ (see Figs. 3 and 4). Thus, $\mathcal{T}^1(\omega_\beta^*(m))$ is in the active set, which gives us

$$D_{\beta,m}(\mathcal{T}^1(\omega_\beta^*(m))) = \beta(\mathcal{T}^1(\omega_\beta^*(m))D_{\beta,m}(p_{11}) + (1-\mathcal{T}^1(\omega_\beta^*(m)))D_{\beta,m}(p_{01})). \quad (55)$$

To prove (54), we consider the positively correlated and negatively correlated cases separately.

Case 1) *Negatively correlated channel* ($p_{11} < p_{01}$). Since $p_{01} > \omega_o > \omega_\beta^*(m), p_{01}$ is in the active set. We thus have

$$D_{\beta,m}(p_{01}) = \beta(p_{01}D_{\beta,m}(p_{11}) + (1-p_{01})D_{\beta,m}(p_{01})). \quad (56)$$

Substituting (55) and (56) into (54), we reduce (54) to the following:

$$\frac{\beta}{1-\beta(1-p_{01})}D_{\beta,m}(p_{11})(1-\beta) \times (\beta p_{01} + \omega_\beta^*(m) - \beta\mathcal{T}^1(\omega_\beta^*(m))) < 1. \quad (57)$$

Notice that the left-hand side of (57) is increasing with $\omega_\beta^*(m)$ and $D_{\beta,m}(p_{11})$. It thus suffices to show the inequality by replacing $\omega_\beta^*(m)$ with its upper bound $\omega_o$ and $D_{\beta,m}(p_{11})$ with its upper bound $\frac{1}{1-\beta}$. After some simplifications, it is sufficient to prove

$$f(\beta)\triangleq p_{01}(p_{01}-p_{11})\beta^2 + \beta(p_{01}+1-p_{11}-p_{01}^2+p_{01}p_{11}) - 1 - p_{01} + p_{11} < 0. \quad (58)$$

It is easy to see that $f(\beta)$ is convex in $\beta$, $f(0) = -1-p_{01}+p_{11} < 0$, and $f(1) = 0$. We thus conclude that $f(\beta) < 0$ for any $0 \le \beta < 1$.

Case 2) *Positively correlated channel* ($p_{11} > p_{01}$). Since $p_{11} \ge \omega_o > \omega_\beta^*(m), p_{11}$ is in the active set. We thus have

$$D_{\beta,m}(p_{11}) = \beta(p_{11}D_{\beta,m}(p_{11}) + (1-p_{11})D_{\beta,m}(p_{01})). \quad (59)$$

Substituting (55) and (59) into (54), we reduce (54) to the following:

$$\beta D_{\beta,m}(p_{01})(1-\beta) \times \left(1 - \frac{\omega_\beta^*(m) - \beta\mathcal{T}^1(\omega_\beta^*(m))}{1-\beta p_{11}}\right) < 1. \quad (60)$$

Substituting the closed form of $D_{\beta,m}(p_{01})$ given in (28) into (60), we end up with an inequality in terms of $L(p_{01}, \omega_\beta^*(m))$ and $\omega_\beta^*(m)$. Notice that the left-hand side of (60) is decreasing with $\omega_\beta^*(m)$. It thus suffices to show the inequality by replacing $\omega_\beta^*(m)$ with its lower bound $\mathcal{T}^{L(p_{01},\omega_\beta^*(m))-1}(p_{01})$ [by the definition of $L(p_{01}, \omega_\beta^*(m))$]. Let $x = p_{11} - p_{01}$. After some simplifications, it is sufficient to show that for any $0 \le \beta < 1, 0 \le p_{01} \le 1, 0 \le x \le 1 - p_{01}, L \in \{0, 1, 2, \ldots\}$

$$f(\beta)\triangleq \beta^{L+2}p_{01}x^{L+1}(1-x) + \beta^2(p_{01}x^{L+2} + x - x^2 - p_{01}x) + \beta(x^2 + p_{01}x - p_{01}x^{L+1} - 1) + 1 - x > 0. \quad (61)$$

Since $f(0) = 1 - x > 0$ and $f(1) = 0$, it is sufficient to prove that $f(\beta)$ is strictly decreasing with $\beta$ for $0 \le \beta \le 1$, which follows by showing $\frac{d(f(\beta))}{d(\beta)} < 0$ for $0 \le \beta < 1$

$$\frac{d(f(\beta))}{d(\beta)} = (L+2)\beta^{L+1}p_{01}x^{L+1}(1-x) + 2\beta(p_{01}x^{L+2} + x - x^2 - p_{01}x) + (x^2 + p_{01}x - p_{01}x^{L+1} - 1). \quad (62)$$

To show $\frac{d(f(\beta))}{d(\beta)} < 0$ for $0 \le \beta < 1$, we will establish the following two facts:

i) $\frac{d(f(\beta))}{d(\beta)}|_{\beta=1} \le 0$;
ii) $\frac{d(f(\beta))}{d(\beta)}$ is strictly increasing with $\beta$.

To prove (i), we set $\beta = 1$ in (62). After some simplifications, we need to prove

$$h(p_{01})\triangleq -p_{01}Lx^{L+2} + p_{01}(L+1)x^{L+1} - x^2 - p_{01}x + 2x - 1 \le 0. \quad (63)$$

Since $h(0) = -(x-1)^2 \le 0$, it is sufficient to prove that $h(p_{01})$ is monotonically decreasing with $p_{01}$, i.e., we need to prove

$$\frac{d(h(p_{01}))}{d(p_{01})} = -Lx^{L+2} + (L+1)x^{L+1} - x \le 0. \quad (64)$$

Since $Lx^{L+1} \le \sum_{k=1}^{L} x^k = \frac{x-x^{L+1}}{1-x}$, it is easy to see that (64) holds. We thus proved (i).

To prove (ii), it suffices to show that the coefficient of $\beta$ in (62) is nonnegative, i.e., we need to prove

$$x^{L+2} + x - x^2 - p_{01}x \ge 0. \quad (65)$$

Since $0 \le x \le 1 - p_{01}$, we have $p_{01}x(x^{L+1} - 1) \ge -p_{01}x \ge (x-1)x$. It is easy to see that (65) holds. We thus proved (ii).

From (i) and (ii), it is easy to see that $\frac{d(f(\beta))}{d(\beta)} < 0$ for any $0 \le \beta < 1$. We thus proved the indexability.

## APPENDIX III
## PROOF OF THEOREM 4

We notice that Step 1 runs in $O(N)$ time. In Step 2, the number of regions that needs to be calculated for each channel is at most $O(\log \frac{\delta}{N}) = O(\log N)$. It runs in constant time to find $l_i$ and

$d_i$ for channel $i$. So Step 2 runs in at most $O(N \log N)$ time. In Step 3, the ordering of all those probabilities needs at most $O(N \log N)(\log(O(N \log N))) = O(N(\log N)^2)$ time. Step 4 runs in $O(N)$ time for each region that does not belong to $V$. So Step 4 runs in at most $O(N^2 \log N)$ time. Finally, Step 5 runs in $O(N)$ time. Overall, the algorithm runs in at most $O(N^2 \log N)$ time.

## Appendix IV
## Proof of Lemma 6

From the closed-form $V_{\beta,m}(p_{01})$ (see Lemma 3), we have, for any $\beta$ $(0 \leq \beta < 1)$

$$|V_{\beta,m}(p_{01}) - V_{\beta,m}(p_{11})| \leq c. \tag{66}$$

From Figs. 6, 7, and 5, we have, for any $\omega \in [0,1]$

$$\begin{aligned}\min\{V_{\beta,m}(0;u=1), V_{\beta,m}(1;u=1)\} &\leq V_{\beta,m}(\omega) \\ &\leq \max\{V_{\beta,m}(0;u=0), V_{\beta,m}(1;u=1)\}.\end{aligned} \tag{67}$$

Consequently, we have, for any $\omega, \omega' \in [0,1]$

$$\begin{aligned}&|V_{\beta,m}(\omega) - V_{\beta,m}(\omega')| \\ &\quad\leq \max(|V_{\beta,m}(0;u=1) - V_{\beta,m}(1;u=1)|, \\ &\qquad |V_{\beta,m}(0;u=0) - V_{\beta,m}(0;u=1)|, \\ &\qquad |V_{\beta,m}(0;u=0) - V_{\beta,m}(1;u=1)|) \\ &\quad= \max(|\beta(V_{\beta,m}(p_{01}) - V_{\beta,m}(p_{11})) - 1|, \\ &\qquad |\beta(V_{\beta,m}(p_{01}) - V_{\beta,m}(p_{11}))|, 1).\end{aligned}$$

Since $|V_{\beta,m}(p_{01}) - V_{\beta,m}(p_{11})| \leq c$ for any $\beta(0 \leq \beta < 1)$, then $V_{\beta,m}(\omega) - V_{\beta,m}(\omega')| \leq c+1$ for any $\beta(0 \leq \beta < 1)$ and $\omega, \omega' \in [0,1]$. Thus, the value-boundedness condition is satisfied.

## Appendix V
## Proof of Lemma 7

The convergence of $\omega^*_\beta(m)$ is trivial for $m < 0$ and $m \geq 1$.

For $0 \leq m < 1$, let $W(\omega) = \lim_{\beta \to 1} W_\beta(\omega)$. This limit exists and is given in Theorem 5 (it is tedious and lengthy to get the limit and we skip the detailed calculation). Define $\omega^*(m)$ as the inverse function of $W(\omega)$. We notice that $W(\omega)$ is a constant function (thus not invertible) when $\omega_o \leq \omega \leq \mathcal{T}^1(p_{11})$ [see (42)]. In this case, we set $\omega^*(m) = \mathcal{T}^1(p_{11})$. Formally, we have

$$\omega^*(m) = \begin{cases} c(c<0), & \text{if } m < 0 \\ \max\{\omega : W(\omega) = m\}, & \text{if } 0 \leq m < 1 \\ b(b>1), & \text{if } m \geq 1. \end{cases} \tag{68}$$

Next, we prove that $\lim_{\beta \to 1} \omega^*_\beta(m) = \omega^*(m)$ as $\beta \to 1$ by contradiction. Since $W(\omega) = \lim_{\beta \to 1} W_\beta(\omega)$ and $W_\beta(\omega)$ is increasing with $\omega, W(\omega)$ is also increasing with $\omega$. Assume first that $W_\beta(\omega)$ is strictly increasing at point $\omega^*_\beta(m)$. We prove $\lim_{\beta \to 1} \omega^*_\beta(m) = \omega^*(m)$ by contradiction as follows.

Assume $\omega^*_\beta(m) \nrightarrow \omega^*(m)$, i.e., there exist an $\epsilon > 0$, a $\beta'(0 \leq \beta' < 1)$, and a series $\{\beta_k\}(\beta_k \to 1)$ such that $|\omega^*_{\beta_k}(m) - \omega^*(m)| > \epsilon$ for any $\beta_k > \beta'$. If $\omega^*(m) - \epsilon > \omega^*_{\beta_k}(m)$ for any $\beta_k > \beta'$, then $W_{\beta_k}(\omega^*(m) - \epsilon) \geq W_{\beta_k}(\omega^*_{\beta_k}(m))$ for any $\beta_k > \beta'$ by the monotonicity of $W_{\beta_k}(\omega)$. Since $W(\omega)$ is strictly increasing at point $\omega^*(m)$, there exists a $\delta > 0$ such that $W(\omega^*(m)) - W(\omega^*(m) - \epsilon) > \delta$. Then, we have, for any $\beta_k > \beta'$

$$\begin{aligned}W_{\beta_k}(\omega^*(m) - \epsilon) &\geq W_{\beta_k}(\omega^*_{\beta_k}(m)) = m \\ &= W(\omega^*(m)) > W(\omega^*(m) - \epsilon) + \delta\end{aligned}$$

which contradicts with the fact that $W_{\beta_k}(\omega^*_{\beta_k}(m) - \epsilon) \to W(\omega^*(m) - \epsilon)$ as $\beta_k \to 1$. The proof for the case when $\omega^*(m) + \epsilon < \omega^*_{\beta_k}(m)$ for any $\beta_k > \beta'$ is similar to the above.

Consider next that $W(\omega)$ is not strictly increasing at point $\omega^*(m)$. This case only occurs when $p_{11} < p_{01}$ and $\omega^*(m) = \mathcal{T}^1(p_{11})$. We notice that $W_\beta(\mathcal{T}^1(p_{11}))$ increasingly converges to $W(\mathcal{T}^1(p_{11}))$ as $\beta \to 1$. Thus, $\omega^*_\beta(m) \geq \mathcal{T}^1(p_{11}) = \omega^*(m)$ by the monotonicity of $W_\beta(\omega)$. Assume $\omega^*_\beta(m) \nrightarrow \omega^*(m)$, i.e., there exist an $\epsilon > 0$, a $\beta'(0 \leq \beta' < 1)$, and a series $\{\beta_k\}(\beta_k \to 1)$ such that $\omega^*_{\beta_k}(m) - \omega^*(m) > \epsilon$ for any $\beta_k > \beta'$. We have $W_{\beta_k}(\omega^*(m) + \epsilon) < W_{\beta_k}(\omega^*_{\beta_k}(m))$ for any $\beta_k > \beta'$ by the monotonicity of $W_{\beta_k}(\omega)$. Since $W(\omega)$ is strictly increasing in $[\omega^*(m), \omega^*(m)+\epsilon]$, there exists a $\delta' > 0$ such that $W(\omega^*(m)+\epsilon) - W(\omega^*(m)) > \delta'$. Then, we have, for any $\beta_k > \beta'$

$$\begin{aligned}W_{\beta_k}(\omega^*(m) + \epsilon) &\leq W_{\beta_k}(\omega^*_{\beta_k}(m)) = m \\ &= W(\omega^*(m)) < W(\omega^*(m) + \epsilon) - \delta'\end{aligned}$$

which contradicts with the fact that $W_{\beta_k}(\omega^*_{\beta_k}(m) + \epsilon) \to W(\omega^*(m) + \epsilon)$ as $\beta_k \to 1$.

Next, we show that the optimal policy $\pi^*_{\beta_k}$ for the single-armed bandit process with subsidy under the discounted reward criterion pointwise converges to a threshold policy $\pi^*$ as $\beta_k \to 1$. To see this, we construct $\pi^*$ as follows: 1) if $m < 0$, then the arm is made active all the time; 2) if $m \geq 1$, the arm is made passive all the time; and 3) if $0 \leq m < 1$, then $\omega$ is made passive when current state $\omega \leq \omega^*(m)$, otherwise it is activated. Since $\omega^*_\beta(m)$ converges to $\omega^*(m)$ as $\beta \to 1$, it is easy to see that $\pi^*_{\beta_k}$ pointwise converges to $\pi^*$ for any $\beta_k \to 1$. Because the single-armed bandit process with subsidy under the discounted reward criterion satisfies the value boundedness condition (see Lemma 6), the threshold policy $\pi^*$ is optimal for the single-armed bandit process with subsidy under the average reward criterion based on Dutta's theorem.

## Appendix VI
## Proof of Theorem 5

Since $\omega^*(m) = \lim_{\beta \to 1} \omega^*_\beta(m)$ and $\omega^*_\beta(m)$ is monotonically increasing with $m$ (see Theorem 1), it is easy to see that $\omega^*(m)$ is also monotonically increasing with $m$. Therefore, the bandit is indexable.

Next, we prove that $W(\omega) \triangleq \lim_{\beta \to 1} W_\beta(\omega)$ is indeed Whittle index under the average reward criterion. For a belief state $\omega$ of an arm, its Whittle index is the infimum subsidy $m$ such that $\omega$ is in the passive set under the optimal policy for the arm, i.e., the infimum subsidy $m$ such that $\omega \leq \omega^*(m)$ (according to Lemma 7). From (68) and the monotonicity of $W(\omega)$ with $\omega$, we have that $W(\omega)$ is the infimum subsidy $m$ such that $\omega \leq \omega^*(m)$.

## APPENDIX VII
## PROOF OF THEOREM 6

The proof for the lower bound of $J_w$ is an extension of that with single-channel sensing ($K = 1$) given in [17]. It is, however, much more complex to analyze the performance of Whittle index policy when $K \geq 1$. The lower bound obtained here is looser than that in [17] when applied to the case of $K = 1$.

Define a transmission period on a channel as the number of consecutive slots in which the channel has been continuously sensed before being moved to the end of the queue. Based on the structure of Whittle index policy, it is easy to show that

$$J_w = \begin{cases} K\left(1 - \dfrac{1}{\mathbb{E}[\tau]}\right), & \text{if } p_{11} \geq p_{01} \\ K\dfrac{1}{\mathbb{E}[\tau]}, & \text{if } p_{11} < p_{01} \end{cases} \tag{69}$$

where $\mathbb{E}[\tau]$ is the average length of the transmission period over the infinite time horizon.

To bound the throughput $J_w$, it is equivalent to bound the average length of the transmission period $\mathbb{E}[\tau]$ as shown in (69). We consider the following two cases.

Case 1) $p_{11} \geq p_{01}$. Let $\omega$ denote the belief value of the chosen channel in the first slot of a transmission period. The length $\tau(\omega)$ of this transmission period has the following distribution:

$$\Pr[\tau(\omega) = l] = \begin{cases} 1 - \omega, & l = 1 \\ \omega p_{11}^{l-2} p_{10}, & l > 1. \end{cases} \tag{70}$$

It is easy to see that if $\omega' \geq \omega$, then $\tau(\omega')$ stochastically dominates $\tau(\omega)$.

From the structure of Whittle index policy, $\omega = \mathcal{T}^k(p_{01})$, where $k$ is the number of consecutive slots in which the channel has been unobserved since the last visit to this channel. When the user leaves one channel, this channel has the lowest priority. It will take at least $\lfloor \frac{N-K}{K} \rfloor$ slots before the user returns to the same channel, i.e., $k \geq \lfloor \frac{N}{K} \rfloor - 1$. Based on the monotonically increasing property of the $k$-step transition probability $\mathcal{T}^k(p_{01})$ (see Fig. 3), we have $\omega = \mathcal{T}^k(p_{01}) \geq \mathcal{T}^{\lfloor \frac{N}{K} \rfloor - 1}(p_{01})$. Thus, $\tau(\mathcal{T}^{\lfloor \frac{N}{K} \rfloor - 1}(p_{01}))$ is stochastically dominated by $\tau(\omega)$, and the expectation of the former leads to the lower bound of $J_w$ given in (47).

Case 2) $p_{11} < p_{01}$. In this case, $\tau(\omega)$ has the following distribution:

$$\Pr[\tau(\omega) = l] = \begin{cases} \omega, & l = 1 \\ (1 - \omega) p_{00}^{l-2} p_{01}, & l > 1. \end{cases} \tag{71}$$

Opposite to Case 1, $\tau(\omega')$ stochastically dominates $\tau(\omega)$ if $\omega' \leq \omega$.

From the structure of Whittle index policy, $\omega = \mathcal{T}^k(p_{11})$, where $k$ is the number of consecutive slots in which the channel has been unobserved since the last visit to this channel. If $k$ is odd, then $\mathcal{T}^k(p_{11}) \geq \mathcal{T}^{2\lfloor \frac{N}{K} \rfloor - 2}(p_{11})$ since $2\lfloor \frac{N}{K} \rfloor - 2$ is an even number (see Fig. 4). If $k$ is even, then $k$ is at least $2\lfloor \frac{N-K}{K} \rfloor$. we have $\omega = \mathcal{T}^k(p_{11}) \geq \mathcal{T}^{2\lfloor \frac{N}{K} \rfloor - 2}(p_{11})$. Thus, $\tau(\omega)$ is stochastically dominated by $\tau(\mathcal{T}^{2\lfloor \frac{N}{K} \rfloor - 2}(p_{11}))$, and the expectation of the latter leads to the lower bound of $J_w$ as given in (48).

Next, we show the upper bound of $J$. From (43), we have $J \leq \inf_m \{NJ_m - m(N-K)\}$ since channels are stochastically identical.

When $p_{11} \geq p_{01}$, we have

$$\begin{aligned} J &\leq \min_{m \in \{\frac{\omega_o}{1 - p_{11} + \omega_o}, 0\}} \{NJ_m - m(N-K)\} \\ &= \min\left\{\frac{K\omega_o}{1 - p_{11} + \omega_o}, N\omega_o\right\}. \end{aligned} \tag{72}$$

When $p_{11} > p_{01}$, we have

$$\begin{aligned} J &\leq \min_{m \in \{\frac{p_{01}}{1 - \mathcal{T}^1(p_{11}) + p_{01}}, 0\}} \{NJ_m - m(N-K)\} \\ &= \min\left\{\frac{Kp_{01}}{1 - \mathcal{T}^1(p_{11}) + p_{01}}, N\omega_o\right\}. \end{aligned} \tag{73}$$

## APPENDIX VIII
## PROOF OF COROLLARY 2

We first prove that Whittle index policy is optimal when $K = N-1$. We construct a genie-aided system where the user knows the states $S_i(t)$ of all channels at the end of each slot $t$. In this system, Whittle index policy is clearly optimal, and the optimal performance is the upper bound of the original one. For the original system where the user only knows the states of the sensed $N-1$ channels, we notice that the channel ordering by Whittle index policy in each slot is the same as that in the genie-aided system. Whittle index policy thus achieves the same performance as in the genie-aided system. It is thus optimal.

Next, we show that Whittle index policy achieves at least $\max\{\frac{1}{2}, \frac{K}{N}\}$ of the optimal performance for negatively correlated channels ($p_{11} < p_{01}$). According to Theorem 6, we arrive at the following inequality (notice that $J_w \geq K\omega_o$):

$$\eta \geq \max\left\{\frac{1 - \mathcal{T}^1(p_{11}) + p_{01}}{1 - p_{11} + p_{01}}, \frac{K}{N}\right\}, \qquad \text{if} \quad p_{11} < p_{01}. \tag{74}$$

Note that

$$\begin{aligned} \frac{1 - \mathcal{T}^1(p_{11}) + p_{01}}{1 - p_{11} + p_{01}} &= 1 + \frac{(p_{11} - p_{01})(1 - p_{11})}{1 - (p_{11} - p_{01})} \\ &\geq 1 - \frac{(1 - p_{11})^2}{2 - p_{11}} \geq 0.5. \end{aligned} \tag{75}$$

Combining (74) and (75), we have $\eta \geq \max\{\frac{1}{2}, \frac{K}{N}\}$.

## REFERENCES


[1] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3/4, pp. 275–294, Dec. 1933.

[2] J. C. Gittins and D. M. Jones, "A dynamic allocation index for the sequential design of experiments," *Progr. Stat.*, pp. 241–266, 1974.

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

[4] P. Whittle, "Comments on: Dynamic priority allocation via restless bandit marginal productivity indices," *TOP*, vol. 15, no. 2, pp. 217–219, Dec. 2007.

[5] P. Whittle, "Restless bandits: Activity allocation in a changing world," *J. Appl. Probab.*, vol. 25, pp. 287–298, 1988.
[6] C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of optimal queueing network control," *Math. Oper. Res.*, vol. 24, no. 2, pp. 293–305, May 1999.
[7] R. R. Weber and G. Weiss, "On an index policy for restless bandits," *J. Appl. Probab.*, vol. 27, no. 3, pp. 637–648, Sept. 1990.
[8] R. R. Weber and G. Weiss, "Addendum to 'on an index policy for restless bandits'," *Adv. Appl. Probab.*, vol. 23, no. 2, pp. 429–430, June 1991.
[9] P. S. Ansell, K. D. Glazebrook, J. E. Niño-Mora, and M. O'Keeffe, "Whittle's index policy for a multi-class queueing system with convex holding costs," *Math. Meth. Oper. Res.*, vol. 57, pp. 21–39, 2003.
[10] K. D. Glazebrook, D. Ruiz-Hernandez, and C. Kirkbride, "Some indexable families of restless bandit problems," *Adv. Appl. Probab.*, vol. 38, pp. 643–672, 2006.
[11] K. D. Glazebrook and H. M. Mitchell, "An index policy for a stochastic scheduling model with improving/deteriorating jobs," *Naval Res. Logistics*, vol. 49, pp. 706–721, Mar. 2002.
[12] E. N. Gilbert, "Capacity of burst-noise channels," *Bell Syst. Tech. J.*, vol. 39, pp. 1253–1265, Sept. 1960.
[13] M. Zorzi, R. Rao, and L. Milstein, "Error statistics in data transmission over fading channels," *IEEE Trans. Commun.*, vol. 46, no. 11, pp. 1468–1477, Nov. 1998.
[14] L. A. Johnston and V. Krishnamurthy, "Opportunistic file transfer over a fading channel: A POMDP search theory formulation with optimal threshold policies," *IEEE Trans. Wireless Commun.*, vol. 5, no. 2, pp. 394–405, Feb. 2006.
[15] Q. Zhao and B. Sadler, "A survey of dynamic spectrum access," *IEEE Signal Process. Mag.*, vol. 24, no. 3, pp. 79–89, May 2007.
[16] J. Le Ny, M. Dahleh, and E. Feron, "Multi-UAV dynamic routing with partial observations using restless bandit allocation indices," in *Proc. Amer. Control Conf.*, Seattle, WA, Jun. 2008, pp. 4220–4225.
[17] 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.
[18] S. H. Ahmad, M. Liu, T. Javadi, Q. Zhao, and B. Krishnamachari, "Optimality of myopic sensing in multi-channel opportunistic access," *IEEE Trans. Inf. Theory*, vol. 55, no. 9, pp. 4040–4050, Sep. 2009.
[19] S. Ahmad and M. Liu, "Multi-channel opportunistic access: A case of restless bandits with multiple plays," in *Proc. Allerton Conf. Commun. Control Comput.*, Allerton, IL, Oct. 2009, pp. 1361–1368.
[20] 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.
[21] Y. Chen, Q. Zhao, and A. Swami, "Joint design and separation principle for opportunistic spectrum access in the presence of sensing errors," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 2053–2071, May 2008.
[22] K. Liu, Q. Zhao, and B. Krishnamachari, "Dynamic multichannel access with imperfect channel state detection," *IEEE Trans. Signal Process.*, vol. 58, no. 5, pp. 2795–2808, May 2010.
[23] 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, pp. 259–297, 2000.
[24] V. Raghunathan, V. Borkar, M. Cao, and P. R. Kumar, "Index policies for real-time multicast scheduling for wireless broadcast systems," in *Proc. IEEE Conf. Comput. Commun.*, 2008, pp. 1570–1578.
[25] N. Ehsan and M. Liu, "On the optimality of an index policy for bandwidth allocation with delayed state observation and differentiated services," in *Proc. IEEE Conf. Comput. Commun.*, 2004, vol. 3, pp. 1974–1983.
[26] M. L. Veatch and M. Wein, "Scheduling a make-to-stock queue: Index policies and hedging points," *Oper. Res.*, vol. 44, pp. 634–647, 1996, 634647.
[27] J. Niño-Mora, "Restless bandits, partial conservation laws and indexability," *Adv. Appl. Probab.*, vol. 33, pp. 76–98, 2001.
[28] J. Niño-Mora, "Dynamic priority allocation via restless bandit marginal productivity indices," *TOP*, vol. 15, pp. 161–198, 2007.
[29] S. Guha and K. Munagala, "Approximation algorithms for partial-information based stochastic control with Markovian rewards," in *Proc. 48th IEEE Symp. Found. Comput. Sci.*, 2007, pp. 483–493.
[30] S. Guha and K. Munagala, "Approximation algorithms for restless bandit problems," [Online]. Available: http://arxiv.org/abs/0711.3861
[31] R. Smallwood and E. Sondik, "The optimal control of partially ovservable Markov processes over a finite horizon," *Oper. Res.*, pp. 1071–1088, 1973.
[32] A. Arapostathis, V. S. Borkar, E. Fernández-gaucherand, M. K. Ghosh, and S. I. Marcus, "Discrete-time controlled Markov processes with average cost criterion: A survey," *SIAM J. Control Optim.*, vol. 31, no. 2, pp. 282–344, 1993.
[33] R. G. Gallager, *Discrete Stochastic Processes*. Norwell, MA: Kluwer, 1995.
[34] E. J. Sondik, "The optimal control at partially observable Markov processes over the infinite horizon: Discounted costs," *Oper. Res.*, vol. 26, no. 2, pp. 282–304, 1978.
[35] P. K. Dutta, "What do discounted optima converge to? A theory of discount rate asymptotics in economic models," *J. Econ. Theory*, vol. 55, pp. 64–94, 1991.
[36] K. Liu, R. R. Weber, and Q. Zhao, "On a class of restless bandit problems," 2010, to be submitted.

**Keqin Liu** received the B.S. degree in automation from Southeast University, Nanjing, China, in 2005 and the M.S. and Ph.D. degrees in electrical engineering from the University of California at Davis, Davis, in 2008 and 2010, respectively.

Currently, he is a Postdoctoral Researcher at the Department of Electrical Engineering, University of California at Davis. He is a Research Assistant at the Department of Electrical Engineering, University of California at Davis. His research interests are stochastic optimization in dynamic systems, distributed control in multiagent systems, and signal processing in wireless networks.

**Qing Zhao** received the Ph.D. degree in electrical engineering from Cornell University, Ithaca, NY, in 2001.

In August 2004, she joined the Department of Electrical and Computer Engineering, University of California at Davis, Davis, where she is currently an Associate Professor. Her research interests are in the general area of dynamic systems and communication networks.

Dr. 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 at Davis College of Engineering.