Indexability and Whittle Index for Restless Bandit Problems Involving Reset Processes

Keqin Liu, Richard Weber, Qing Zhao

***Abstract*— We consider a class of restless multi-armed bandit (RMAB) problems, in which the active action resets the stochastic evolution of the system. We obtain the Whittle index in closed-form, showing that it induces a policy that is equivalent to the myopic policy, and that it is optimal for stochastically identical arms. These results find applications in opportunistic spectrum access and supervisory control systems such as anomaly detection and control.**

***Index Terms*— Reset processes, restless multi-armed bandit, Whittle index policy, myopic policy.**

I. Introduction

A. Reset Processes

Consider N independent random processes, in which the state of each random process can be either 0 or 1. At each discrete time instant, we can choose K $(K < N)$ random processes to observe, and a positive reward of r is obtained from each random process observed in state 1. Whenever a process is observed, its stochastic evolution is reset, potentially by a control action taken in response to the observed state. Specifically, after a process is observed in state i, the subsequent distribution of its state is given by a sequence $\{p_{i1}(t)\}_{t\geq 1}$, in which $p_{i1}(t)$ is the probability that the process is in state 1 after t time slots since the last observation. It shall become clear later that the marginal distribution $\{p_{i1}(t)\}_{t\geq 1}$ suffices for the optimal decision making. We also allow different stochastic dynamics and rewards across these N processes, but the parameters of the processes are not indexed for notation simplicity.

The objective is to design a policy that sequentially selects K arms to observe at each time to maximize the expected average reward over an infinite horizon.

The above general problem is motivated by two classes of applications. One is supervisory control in which $\{p_{i1}(t)\}_{t\geq 1}$ characterizes the evolution of a certain physical process and the decision maker interacts with the physical process by taking control actions that reset the stochastic evolution of the process. For example, consider a supervisory control system in which multiple chemical processes are monitored and controlled. The state 1 represents an abnormal state and 0 the normal state. The objective is to track and rectify processes that are in the abnormal state to ensure system security or product quality. Different control actions are taken according to the observed state i, which reset the state evolution of the process to $\{p_{i1}(t)\}_{t\geq 1}$.

In the other class of applications, the underlying physical processes evolve as two-state Markov chains independent of the decision maker. The decision maker decides at each time which K processes to monitor and accrue rewards from each process that is in the good state 1. The probability sequence $\{p_{i1}(t)\}_{t\geq 1}$ characterizes the knowledge of the decision maker on the state of each process, and whenever a process is observed, the knowledge about it is refreshed to $p_{i1}(1) = q_{i1}$ where $\{q_{i,j}\}_{i,j=0,1}$ are the transition probabilities of the corresponding process. It is easy to see that $p_{i1}(t)$ is the t-step transition probability of the two-state Markov chain with the following specific form.

$$\begin{cases} p_{01}(t) & = \frac{q_{01}(1-(q_{11}-q_{01})^t)}{1+q_{01}-q_{11}} \\ p_{11}(t) & = \frac{q_{01}+(1-q_{11})(q_{11}-q_{01})^t}{1+q_{01}-q_{11}} \end{cases} \tag{1}$$

One example of this second class of applications (referred to as non-interactive applications) is cognitive radio for dynamic spectrum access, in which secondary users search in the spectrum for idle channels temporarily unused by primary users [1]. For this application, the state "1" represents an idle channel and the state "0" an occupied channel. The reward obtained on an idle channel represents data delivery and is proportional to the bandwidth of the channel. The objective is to maximize the long-run throughput.

B. Main Results

The above general problem can be modeled as a restless multi-armed bandit problem (RMAB) introduced by Whittle in 1988 [2]. The actions 'observe' and 'do not observe' correspond to taking the active and passive actions on an arm, respectively, which are denoted as the controls $a = 1$ and $a = 0$. For the optimal action making, we only need to know the marginal probabilities $\{p_{i1}(t)\}_{t\geq 1}$. The arm state is thus given by $(i,\ t)$, where i is the state of the process when it was last observed and t the time lag since the last observation. The correlation of the state of the random process is not required to be known and can take an arbitrary form.

We show that the above RMAB consisting of reset processes satisfies Whittle indexability. Under certain monotone conditions on the stochastic evolution of the arm state, we obtain the Whittle index in closed-form. This result reduces the complexity of implementing Whittle index policy to simple evaluations of these closed-form expressions. When arms are stochastically identical, Whittle index policy is

The work of K. Liu and Q. Zhao was supported by the Army Research Office under Grant W911NF-08-1-0467 and by the National Science Foundation under Grant CCF-0830685.

Keqin Liu and Qing Zhao are with the Department of Electrical and Computer Engineering, University of California, Davis, CA 95616, {kqliu, qzhao}@ucdavis.edu.

Richard Weber is with the Statistical Laboratory, University of Cambridge, CB3 0WB UK, rrw1@cam.ac.uk.

shown to be equivalent to the myopic policy that has a simple and robust structure. Based on this structure, we establish the asymptotic optimality of Whittle index policy as $K/N \to 0$. In the finite regime, we focus on the non-interactive applications and the myopic policy. Specifically, we extend the optimality of the myopic policy established in [3]–[5] to a more general Markovian model that consists of time-inhomogeneous Markov chains and allows K to be time varying.

C. Related Work

The results in this paper generalize those reported in [3]–[5] and [6], [7] that consider only the non-interactive applications with time-homogeneous Markov chains. Specifically, the RMAB model for monitoring stochastically identical time-homogeneous Markov chains was considered in [3], in which the semi-universal structure of the myopic policy was established for all N and the optimality of the myopic policy was proved for $N = 2$. In [4], the optimality of the myopic policy was extended to $N > 2$ with $K = 1$ under the condition of $q_{11} \geq q_{01}$. Under the same condition, a recent work [5] extended the optimality of the myopic policy to $K > 1$. In [6], Liu and Zhao considered a general scenario of non-identical Markov chains, for which Whittle indexability was established and Whittle index was solved in closed-form under both discounted and average reward criteria [6]. Whittle index policy was shown to be optimal for stochastically identical arms [6]. This result was obtained by establishing the equivalence between Whittle index policy and the myopic policy. For the same RMAB problem considered in [6], Whittle indexability and Whittle index were also obtained under the discounted reward criterion in an independent work by Le Ny et al. in [7].

In the context of general RMAB problems, establishing indexability is still an open problem and often relies on numerical algorithms [8]. The optimality of Whittle index policy is generally intractable due to the exponential complexity of the problem [9]. Weber and Weiss showed that Whittle index policy is asymptotically optimal ($N \to \infty$ with fixed K/N) under a certain condition [10] that was shown to hold in numerous RMAB problems [10]–[12].

II. Indexability and Whittle Index Policy

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

A. Index Policy

An *index policy* assigns an index for each state of each arm to measure how rewarding it is to activate an arm at a particular state. At each time, 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-dimensional problems.

The 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 state $(i,\ t)$ of an arm is simply $p_{i1}(t)r$.

B. Definition of Indexability and Whittle Index

To introduce indexability and Whittle index, it is sufficient to focus on a single arm based on the strong decomposability of Whittle index [2]. Assume a subsidy λ is provided whenever the arm is made passive (i.e., not observed). Consider the problem of deciding whether to observe this arm at each time in order to maximize the average reward over the infinite horizon. We have the following optimality equation.

$$\begin{aligned} g + \phi(i,\ t) = \max\{&\lambda + \phi(i,\ t+1),\ p_{i1}(t)r \\ &+ p_{i1}(t)\phi(1,\ 1) + (1 - p_{i1}(t))\phi(0,\ 1)\}, \end{aligned} \tag{2}$$

where g is the maximum average reward obtained from the single arm and the function $\phi(\cdot)$ is the differential reward caused by the transient effect on the single arm. The optimal policy for this single-arm problem is essentially given by an optimal partition of the state space $\bigcup_{i=1,2}\{(i,\ t)\}_{t\geq 1}$ into a passive set

$$\begin{aligned} \mathcal{P}(\lambda) &= \{(i,\ t) : a^*_\lambda(i,\ t) = 0\} \\ &= \{(i,\ t) : \lambda + \phi(i,\ t+1) \geq p_{i1}(t)r + p_{i1}(t) \\ &\quad \times\phi(1,\ 1) + (1 - p_{i1}(t))\phi(0,\ 1)\} \end{aligned}$$

and its complement, an active set $\mathcal{A}(\lambda) = \{p_{i1}(t) : a^*_\lambda(i,\ t) = 1\}$, where $a^*_\lambda(i,\ t)$ denotes the optimal action at state $(i,\ t)$ under subsidy λ.

To define Whittle index, it is required that the RMAB is *indexable* [2]. An RMAB is indexable if for each arm, the passive set $\mathcal{P}(\lambda)$ increases monotonically from the empty set ϕ to the entire state space $\bigcup_{i=1,2}\{(i,\ t)\}_{t\geq 1}$ as the subsidy λ increases from $-\infty$ to $+\infty$.

Given indexability, Whittle index $W(i,\ t)$ of a state $(i,\ t)$ is defined as the infimum subsidy λ that makes the passive action optimal at $(i,\ t)$:

$$\begin{aligned} W(i,\ t) &= \inf\{\lambda :\ a^*_\lambda(i,\ t) = 0\} \\ &= \inf\{\lambda :\ \lambda + \phi(i,\ t+1) \geq p_{i1}(t)r \\ &\quad + p_{i1}(t)\phi(1,\ 1) + (1 - p_{i1}(t))\phi(0,\ 1)\}. \end{aligned}$$

Whittle index thus measures how attractive it is to activate an arm based on the subsidy λ. The minimum subsidy λ 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.

III. Indexability of the RMAB consisting of Reset Processes

In this section, we establish the indexability of the class of RMAB considered in this paper.

Theorem 1: The RMAB consisting of multiple reset processes is indexable.

Proof: It is helpful to rewrite (2). Given we take the active action, and observe the corresponding random process to be in state i $(i = 0, 1)$, we will wait some further time t_i until making the next observation (i.e., next taking the active action). The average-reward optimality equation is then

$$\begin{aligned}\phi(i) &= \max_{t_i \geq 1}\{-gt_i + \lambda(t_i - 1) + (1 - p_{i1}(t_i)) \\ &\quad \times\phi(0) + p_{i1}(t_i)(r + \phi(1))\}, \quad i = 1, 2. \quad (3)\end{aligned}$$

Let us set $\phi(0) = 0$ (since only $\phi(1) - \phi(0)$ is determined by the above) and rewrite the optimality equation as

$$0 = \max_{t_0 \geq 1}\{-g - (t_0 - 1)(g - \lambda) + p_{01}(t_0)(r + \phi(1))\}, \quad (4)$$

$$\phi(1) = \max_{t_1 \geq 1}\{-g - (t_1 - 1)(g - \lambda) + p_{11}(t_1)(r + \phi(1))\}. \quad (5)$$

To prove indexability, it is equivalent to show that the optimizing t_i in (3), say $t_i^*(\lambda)$, should be non-decreasing in λ. Note that it is intuitive that as λ, the subsidy for passivity increases, it should be optimal to wait longer between inspections (instances of taking $a = 1$). By the following procedure, we prove that this is true.

1. Let $g(\lambda, t_0, t_1)$ be the average reward under a policy such that whenever the active action is taken and the true state is observed to be i, the next active action is taken t_i steps later, $t_i \in \{1, 2, \cdots\}$. We can compute $g(\lambda, t_0, t_1)$ from the equations

$$0 = -g - (t_0 - 1)(g - \lambda) + p_{01}(t_0)(r + \phi(1)), \quad (6)$$

$$\phi(1) = -g - (t_1 - 1)(g - \lambda) + p_{11}(t_1)(r + \phi(1)), \quad (7)$$

   where we omit showing dependence of g and $\phi(1)$ on $\lambda,\ t_0,\ t_1$. One could explicitly solve these equations for $\phi(1)$ and g. However, this is not required. It is sufficient to observe that: $g(\lambda, t_0, t_1) - \lambda$ is a linear function of λ, nonnegative, and decreasing in λ. This is obvious, because $g(\lambda, t_0, t_1) - \lambda$ is an averaging of '$\lambda - \lambda$' (when passive), '$0 - \lambda$' and '$r - \lambda$' (when active).
2. We now combine the observation 1, with the fact that

$$g(\lambda) - \lambda = \max_{t_0, t_1}\{g(\lambda, t_0, t_1) - \lambda\} \quad (8)$$

   to conclude that $g(\lambda) - \lambda$ is nonnegative, and decreasing in λ. It is also convex in λ, but we do not need this fact.
3. $g(\lambda)$ is increasing in λ.
4. From 2 and 3, we have that $g(\lambda)/(g(\lambda) - \lambda)$ is positive, and increasing in λ.
5. From (4),

$$0 = \max_{t_0 \geq 1}\left\{-\frac{g(\lambda)}{g(\lambda) - \lambda} - (t_0 - 1) + p_{01}(t_0)\left(\frac{r + \phi(1)}{g(\lambda) - \lambda}\right)\right\}. \quad (9)$$

   Hence, using 4, we see from (9) that $\frac{r+\phi(1)}{g(\lambda)-\lambda}$ must be positive, and nondecreasing in λ.
6. Suppose the process is known to be in true state i and it is better to next take the active action at step t, rather than at any earlier step $s,\ s < t$. That is,

$$\begin{aligned}&-g - (t - 1)(g - \lambda) + p_{i1}(t)(r + \phi(1)) \\ >\ &-g - (s - 1)(g - \lambda) + p_{i1}(s)(r + \phi(1)), \\ &s = 1, \cdots, t - 1, \quad (10)\end{aligned}$$

   where for simplicity we again omit showing the dependence of $\phi(1)$ and g on λ.
   Recall that $r + \phi(1) > 0$. This fact has been established in 5. Alternatively, we can see this by supposing $\lambda < r$ and so $g(\lambda) < r$. By adding r to both sides of (5) and observing that we could take $t_1 = 1$, we thus find $(1 - p_{11}(1)(\phi(1) + r) \geq r - g(\lambda) \geq 0$.
   So (10) is equivalent to

$$p_{i1}(t) - p_{i1}(s) > (t - s)\left(\frac{g(\lambda) - \lambda}{r + \phi(1)}\right),\ s = 1, \ldots, t - 1. \quad (11)$$

   Using 5, above, we deduce that as λ increases the right hand side of (11) decreases. Thus the set of t for which (10) and (11) are true is nondecreasing in λ. Thus we conclude that $t_i^*(\lambda)$ is nondecreasing in λ. Since this implies that $\mathcal{P}(\lambda)$ is nondecreasing in λ, we proved indexability.

■

IV. Whittle Index

Based on the indexability given in Theorem 1, Whittle index policy exists for the RMAB. In this section, we solve for Whittle index in closed-form under the following monotone condition on the stochastic evolution of each arm.
The Monotone Condition: $p_{11}(t)$ is decreasing with t and $p_{01}(t)$ is increasing with t; $p_{11}(1) \geq p_{01}(t)$ for all $t \geq 1$.

Note that for the non-interactive applications with a Markovian model, the monotone condition is equivalent to the case that each arm is positively correlated (i.e., $q_{11} \geq q_{01}$). Consider, for example, sampling a continuous Markov process where the samples are always positively correlated. In this case, $p_{i1}(t)$ monotonically converges to the stationary distribution at a rate that depends on $q_{11} - q_{01}$, as shown in (1). In the general scenario, the monotone condition is more relaxed and does not require that $\{p_{i1}(t)\}_{t \geq 1}$ has any specific form. In the rest of the paper, we assume that the monotone condition is satisfied.

The following lemma is the key to solving for Whittle index. It shows that the optimal policy for a single arm with subsidy is a threshold policy.

Lemma 1: The optimal policy for a single arm with subsidy is a threshold policy: for any λ, there exists a threshold $p^*(\lambda)$ such that it is optimal to activate the arm if and only if the probability that the underlying random process is in state 1 is larger than $p^*(\lambda)$.

Proof: Recall that t_i^* $(i = 0, 1)$ maximize, respectively, the righthand side of (4) and (5). From (11), we have

$$\min_{s < t_i^*} \frac{p_{i1}(t_i^*) - p_{i1}(s)}{t_i^* - s} \geq c(\lambda) = \frac{g - \lambda}{r + \phi(1)}. \quad (12)$$

Under the monotone condition, $p_{11}(t)$ is decreasing with t. From observation 5 in the proof of Theorem 1, $c(\lambda)$ is nonnegative. To satisfy (12), we should activate on $(1,\ 1)$ immediately (i.e., $t_i^*=1$) or never activate it (i.e., $t_i^*=\infty$) depending on $c(\lambda)$. However, the latter only happens when $c(\lambda)=0$, i.e., $g=\lambda$, which can be achieved by being passive at all states. This is a trivial threshold policy.

We focus on the non-trivial case that we activate on $(1,\ 1)$ immediately. In this case, the arm state evolves in the set $\{(0,\ t)\}_{t\geq 1}\cup(1,\ 1)$. Under subsidy λ, the passive set is $\{(0,\ 1),\ldots,(0,\ t_0^*-1)\}$. By indexability, t_0^* is nondecreasing with λ. The minimum subsidy that makes $(0,\ t)$ in the passive set is thus not smaller than that makes $(0,\ s)$ for any $t>s$. Since $p_{01}(t)$ is increasing with t under the monotone condition, Whittle index for $(0,\ t)$ is increasing with $p_{01}(t)$.

From the above, the minimum subsidy that makes $(1,\ 1)$ passive also makes all states passive. So $(1,\ 1)$ has the largest Whittle index among $\{(0,\ t)\}_{t\geq 1}\cup(1,\ 1)$. Combined with the fact that Whittle index for $(0,\ t)$ is nondecreasing with $p_{01}(t)$ and $p_{11}(1)\geq p_{01}(t)$ for all $t\geq 1$ (under the monotone condition), we conclude that Whittle index is increasing with the probability that the underlying random process is in state 1. The optimal policy is thus a threshold policy on the probability space $\{p_{01}(t)\}_{t\geq 1}\cup p_{11}(1)$: for any subsidy λ, there exists an $p^*(\lambda)$ such that we activate the arm if the probability that the underlying random process is in state 1 exceeds $p^*(\lambda)$. ■

Note that states $\{(1,t)\}_{t>1}$ are transient under the optimal policy and their Whittle indices are not clearly defined. We will return to this issue later. In the following theorem, we first solve for Whittle indices of states $\{(0,t)\}_{t\geq 1}\cup(1,1)$. For simplicity, we focus on the case that the bandit is strictly indexable, i.e., there is no tie among the Whittle indexes. A sufficient and necessary condition for the strict indexability is given in the following lemma.

Lemma 2: Under C1, the restless bandit consisting of reset processes is strict indexable if and only if the following condition is satisfied:

C2: $p_{01}(t+1)-p_{01}(t)$ is strictly decreasing with t.

Proof: From equation (10) and (11), t is an optimal activation time if and only if

$$p_{i1}(t)-p_{i1}(s) \geq (t-s)\left(\frac{g(\lambda)-\lambda}{r+\phi(1)}\right),\ \forall\ s<t; \quad (13)$$

$$p_{i1}(u)-p_{i1}(t) \leq (u-t)\left(\frac{g(\lambda)-\lambda}{r+\phi(1)}\right),\ \forall\ u>t. \quad (14)$$

Note that the right-hand side of (13) (for fixed t,s) is continuously decreasing from $+\infty$ to 0 as λ increases from $-\infty$ to $+\infty$.

We first prove the necessity. Under the strict indexability, states $\{(0,t)\}_{t\geq 1}$ join the passive set one by one as λ increases. Consider an arbitrary $j\geq 1$. If both (13) and (14) hold with equality by letting $(u,t,s)=(j+2,j+1,j)$ as λ increases to a certain value, than Whittle indexes for states $(0,j)$ and $(0,j+1)$ would be the same. This contradicts the strict indexability. We thus have that $p_{01}(j+1)-p_{01}(j)$ is strictly decreasing at j.

Now we prove the sufficiency. If C2 is satisfied, there must exist a subsidy λ such that both (13) and (14) hold with strict inequality by letting $(t,s,u)=(i+1,i,i+2)$. So the Whittle index for state $(0,i)$ is smaller than this λ while the Whittle index for state $(0,i+1)$ is larger than it. This proves the strict indexability. ■

Theorem 2: Let $W(x)$ denote the Whittle index of state x. Under the strict indexability,

$$W(0,\ t) = \frac{p_{01}(t)(t+1)-p_{01}(t+1)t}{1-p_{11}(1)+tp_{01}(t)-(t-1)p_{01}(t+1)}\,r\,, \quad (15)$$

$$W(1,\ 1) = p_{11}(1)r\,. \quad (16)$$

Proof: Based on the proof of Lemma 1, we have

$$W(1,\ 1)=p_{11}(1)r\,. \quad (17)$$

To find the Whittle index for a state of the form $(0,\ t)$ we must find the value of λ for which one would be indifferent between policies with $t_0=t$ and $t_0=t+1$. Under the second of these policies we have the following optimality equations

$$\begin{aligned}
\phi(0,\ 1)&=0\\
\phi(1,\ 1)+g&=p_{11}(1)(r+\phi(1,\ 1))\\
\phi(0,\ s)+g&=\lambda+\phi(0,\ s+1)\,,\quad s=1,\ldots,t\\
\phi(0,\ t+1)+g&=p_{01}(t+1)(r+\phi(1,\ 1))
\end{aligned}$$

For indifference between passive and active actions in state $(0,\ t)$, we need

$$\lambda+\phi(0,\ t+1)=p_{01}(t)(r+\phi(1,\ 1))\,.$$

Solving the $t+4$ equations above for the variables $\phi(1,\ 1)$, $\phi(0,\ 1),\ldots,\phi(0,\ t+1)$, g and λ gives (15). We also have that

$$g=\frac{(1-p_{11}(1))(t-1)W(0,\ t)+p_{01}(t)}{(1-p_{11}(1))(t+p_{01}(t))+p_{01}(t)p_{11}(1)}. \quad (18)$$

■

Now, we consider the states $\{(1,\ t)\}_{t>1}$. When $\lambda\geq p_{11}r$, these states are in the passive set. Consider the case that $\lambda<p_{11}r$. Since one always takes the active action at state $(1,\ 1)$, these states are never reached under an optimal policy unless they are given as the initial state. If the latter happens, it is necessary to activate the arm at some time such that the system will evolve on the state space $\{(0,t)\}_{t\geq 1}\cup(1,1)$ and achieve the maximum average reward. In other words, there must exist a subsequence of states $\{(1,t_k)\}_{t_k>1}$ of which the Whittle indices are given by $p_{11}r$. For other states that do not belong to the subsequence, their Whittle indices can be set as an arbitrary value that does not exceed $p_{11}r$. Note that the subsequence can also be chosen arbitrarily. Since the states $\{(1,\ t)\}_{t>1}$ may become recurrent under Whittle index policy for multiple arms, the above results lead to a fundamental implementation issue of Whittle index policy: how do we choose Whittle indices that are transient under the optimal policy for a single arm with subsidy? How different choices would affect the performance of Whittle index policy? To the best of our knowledge, these issues

have not been discussed in the literature and are interesting for future investigations. In the next section, we show that for stochastically identical arms, the states $\{(1,t)\}_{t>1}$ are transient under Whittle index policy and the choices of their Whittle indices do not affect the system performance.

V. Structure and Optimality of Whittle Index Policy

In this section, we study the structure and performance of Whittle index policy when arms are stochastically identical. Without loss of generality, we set $r = 1$ in this section.

A. Equivalence to the Myopic Policy

Lemma 3: When arms are stochastically identical, Whittle index policy is equivalent to the myopic policy.

Proof: We first notice that under Whittle index policy, the arm states $\{(1,t)\}_{t>1}$ for which Whittle indices are not clearly defined will eventually disappear (after each arm has been observed once). It is thus sufficient to focus on the state space $\{(0,\ t)\}_{t\geq 1} \cup (1,\ 1)$. Based on the proof of Lemma 1, Whittle index of an arm is increasing with the probability that the underlying random process is in state 1. Since stochastically identical arms would have the same Whittle index if they have the same probability that the underlying random processes are in state 1, Whittle index policy is equivalent to the myopic policy. ■

B. Structure

For time-homogeneous Markov chains satisfying the monotone condition (i.e., $q_{11} \geq q_{01}$), it was shown in [3] that the myopic policy has the following simple structure: initialize a queue in which arms are ordered according to the descending order of the initial probabilities that the underlying random processes are in state 1. Each time we activate the K arms at the head of the queue, where arms of which the underlying random processes observed in state 1 will stay at the head of the queue and other observed arms will be moved to the end of the queue.

It is easy to see that the above structure of the myopic policy is preserved (after each arm has been observed once) under the more general model of reset processes considered in this paper. Whittle index policy thus has the same structure due to its equivalence to the myopic policy.

From the structure of Whittle index policy, it does not require the explicit knowledge of $p_{i1}(t)$ $(i = 0, 1)$ and is thus robust to model variations (as long as the monotone condition is satisfied). Based on this structure, we establish the optimality of Whittle index policy under certain conditions, as presented in the next two subsections.

C. Asymptotic Optimality of Whittle Index Policy

The asymptotic optimality of Whittle index policy in a limiting regime was studied in [10] by Weber and Weiss in 1990. They showed that under certain condition, Whittle index policy achieves the optimal average reward per arm as $N \to \infty$ (fixed K/N). For the problem at hand, the arm state space is infinite and the condition is difficult to check. In this section, we consider another limiting regime in which $K/N \to 0$. We show that for this case, Whittle index policy is asymptotically optimal.

The following lemma is the key to establishing the asymptotic optimality of Whittle index policy, in which the closed-form lower and upper bounds on the performance of Whittle index policy are established.

Lemma 4: Let G_w denote the average reward under Whittle index policy. Let G denote the maximum average reward under the optimal policy. We have,

$$\frac{Kp_{01}(\lfloor \frac{N}{K} \rfloor)}{1-p_{11}(1)+p_{01}(\lfloor \frac{N}{K} \rfloor)} \leq G_w \leq G \leq \frac{K\omega_o}{1-p_{11}(1)+\omega_o}, \quad (19)$$

where $\omega_o = \lim_{t\to\infty} p_{01}(t)$.

Proof: Define an *active period* on an arm as the time period when this arm is continuously activated before being moved to the end of the queue. Based on the structure of Whittle index policy, it is easy to show that

$$R_w = K\left(1 - \frac{1}{\mathbb{E}[L]}\right), \quad (20)$$

where $\mathbb{E}[L]$ is the average length of the active period over the infinite time horizon and all arms.

To bound the throughput R_w, it is equivalent to bound the average length of the transmission period $\mathbb{E}[L]$ as shown in equation (20).

Let ω denote the probability that the underlying random process of the chosen arm is in state 1 at the beginning of an active period. The length $L(\omega)$ of this active period has the following distribution.

$$\Pr[L(\omega) = l] = \begin{cases} 1-\omega, & l = 1 \\ \omega(p_{11}(1))^{l-2}(1-p_{11}(1)), & l > 1 \end{cases}. \quad (21)$$

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

From the structure of Whittle index policy, $\omega = p_{01}(s+1)$, where s is the time duration in which the arm has been unobserved since the last active action on this arm. When the arm is moved to the end of the queue, it has the lowest priority. It will take at least $\lfloor \frac{N-K}{K} \rfloor$ steps before we return to activate the same arm, i.e., $s \geq \lfloor \frac{N}{K} \rfloor - 1$. Based on the monotonically increasing property of $p_{01}(t)$ (under the monotone condition), we have $\omega = p_{01}(s+1) \geq p_{01}(\lfloor \frac{N}{K} \rfloor)$. Thus $L(p_{01}(\lfloor \frac{N}{K} \rfloor))$ is stochastically dominated by $L(\omega)$, and the expectation of the former leads to the lower bound of G_w given in (19).

Next, we show the upper bound of G. From Whittle Lagrangian relaxation [2], we have

$$G \leq \inf_{\lambda}\{Ng(\lambda) - \lambda(N-K)\}, \quad (22)$$

where $g(\lambda)$ is the average reward on a single arm with subsidy λ, as given in (18) (after replacing $W(0,\ t)$ by λ).

Consider the subsidy $\lambda_o = \lim_{t\to\infty} W(0,\ t) = \frac{\omega_o}{1-p_{11}(1)+\omega_o}$. From (18), we have

$$g(\lambda_o) = g(\lim_{t\to\infty} W(0,\ t)) = \lim_{t\to\infty} g(W(0,\ t)) = \lambda_o. \quad (23)$$

From (22) and (23), we arrive at

$$\begin{aligned} G &\leq Ng(\lambda_o) - \lambda_o(N-K) \\ &= K\lambda_o = \frac{K\omega_o}{1-p_{11}(1)+\omega_o}. \end{aligned} \tag{24}$$

■

Theorem 3: When arms are stochastically identical, Whittle index policy is asymptotically optimal in the follow sense:

(i) $G_w/G \to 1$ as $K/N \to 0$;

(ii) $G - G_w \to 0$ as $K(\omega_o - p_{01}(\lfloor N/K \rfloor)) \to 0$.

Proof: This is a direct result from the lower bound of G_w and the upper bound of G given in Lemma 4. Compared to the first optimality result, the second one is stronger but requires K does not grow too fast as $N \to \infty$. ■

D. Optimality in the Finite Regime for Markovian Processes

For a finite N, the optimality of the myopic policy was proven in [3]–[5] when the underlying random processes are time-homogeneous Markov chains satisfying the monotone condition (i.e., $q_{11} \geq q_{01}$). This result leads to the optimality of Whittle index policy based on its equivalence to the myopic policy [6]. In this subsection, we provide a simpler proof for the optimality of the myopic policy under the Markovian model and show that the optimality results can be further generalized. This generalization is twofold: i) each underlying Markov process can be time-inhomogeneous; ii) K can be time varying.

Theorem 4: Consider a general Markovian model in which each underlying Markov process is time-inhomogeneous and K is time varying, the myopic policy is optimal over both finite and infinite horizons.

Proof: In following proof, we allow the parameters K, $p_{01}(1)$ and $p_{01}(1)$ to be time varying. We adopt a notation similar to that in [4], [5], but with some small differences. Recall that K of the N arms are to be observed at each step. We consider a discounted problem over a finite horizon. Let $W_s(\omega_1, \ldots, \omega_N)$ be the discounted reward over s steps when the arms are ordered so the probabilities that the underlying random processes are in state 1 are $\omega_1, \omega_2, \ldots, \omega_N$. Suppose that we observe the K arms at the start of this list. Then

$$\begin{aligned} W_{s+1}(\omega_1, \ldots, \omega_N) = \sum_{i=1}^{K} \omega_i + \beta E\big[W_s(\underbrace{p_{11}(1), \ldots, p_{11}(1)}_{\ell_1 \text{ times}}, \\ \tau(\omega_{K+1}), \ldots, \tau(\omega_N), \underbrace{p_{01}(1), \ldots, p_{01}(1)}_{\ell_0 \text{ times}})\big], \end{aligned}$$

where $W_0(\cdot) = 0$, $\tau(x) = p_{11}(1)x + p_{01}(1)(1-x)$, and the expectation is taken over possible outcomes that can occur when the K arms that are observed are those at the left end (i.e., having probabilities $\omega_1, \ldots, \omega_K$ that the underlying random processes are in state 1), and ℓ_i of the underlying random processes are found to be in state i (and so $\ell_0 + \ell_1 = K$). Notice that if $\omega_1 \geq \cdots \geq \omega_N$, then $W_s(\omega_1, \ldots, \omega_N)$ is the value function for the myopic policy. That is, $W_s(\omega_1, \ldots, \omega_N)$ is the expected discounted reward obtained over s remaining steps under the myopic policy when $\omega_1 \geq \cdots \geq \omega_N$.

We wish to show that the myopic policy is optimal. To do this, it is sufficient to show that for $y > x$, and $\omega_1 \geq \cdots \geq \omega_{K-1} \geq x$ and $y \geq \omega_{K+2} \geq \cdots \geq \omega_N$ we have

$$\begin{aligned} & W_{s+1}(\omega_1, \ldots, \omega_{K-1}, y, x, \omega_{K+2}, \ldots, \omega_N) \\ > \; & W_{s+1}(\omega_1, \ldots, \omega_{K-1}, x, y, \omega_{K+2}, \ldots, \omega_N). \end{aligned}$$

From [4], [5],

$$\begin{aligned} & W_{s+1}(\omega_1, \ldots, \omega_{K-1}, y, x, \omega_{K+2}, \ldots, \omega_N) \\ & - W_{s+1}(\omega_1, \ldots, \omega_{K-1}, x, y, \omega_{K+2}, \ldots, \omega_N) \\ =& (y-x)[W_{s+1}(\omega_1, \ldots, \omega_{K-1}, 1, 0, \omega_{K+2}, \ldots, \omega_N) \\ & - W_{s+1}(\omega_1, \ldots, \omega_{K-1}, 0, 1, \omega_{K+2}, \ldots, \omega_N)]. \end{aligned}$$

This is because that the expression before the equality must be a function of the form $a+bx+cy$, for some a, b and c. The above is positive if the term in square brackets is positive. Thus, we can complete a step of the inductive proof (of the optimality of the myopic policy) by showing this.

However, we show something stronger. Let $\bar{\omega}_i$ denote any sequence of ω_is, possibly empty. We might partition the state vector as $(\bar{\omega}_1, y, \bar{\omega}_2, \bar{\omega}_3)$, where y is the probability that a single underlying random process is in state 1. We shall prove by induction on s that that for all valid state vectors $(\bar{\omega}_1, y, \bar{\omega}_2, \bar{\omega}_3)$

$$\text{(A)} \quad 1 + W_s(\bar{\omega}_1, y, \bar{\omega}_2, \bar{\omega}_3) - W_s(\bar{\omega}_1, \bar{\omega}_2, y, \bar{\omega}_3) \geq 0.$$

This means that the total reward loss by moving 'y' higher in the queue is at most 1.
And for all $y > x$, and valid state vectors $(\bar{\omega}_1, y, \bar{\omega}_2, x, \bar{\omega}_3)$,

$$\text{(B)} \quad W_s(\bar{\omega}_1, y, \bar{\omega}_2, x, \bar{\omega}_3) - W_s(\bar{\omega}_1, x, \bar{\omega}_2, y, \bar{\omega}_3) \geq 0.$$

These are clearly true for $s = 1$. Let us begin by proving an induction step for (B). As above, the expression in (B) is equal to

$$(y-x)\Big[W_{s+1}(\bar{\omega}_1, 1, \bar{\omega}_2, 0, \bar{\omega}_3) - W_{s+1}(\bar{\omega}_1, 0, \bar{\omega}_2, 1, \bar{\omega}_3)\Big].$$

Let us focus on the arms that have been swapped. Suppose that these occur in the ith and jth place, $i < j$. If $i, j \leq K$, then the expression in square brackets evaluates to 0. If $i, j > K$ the expression evaluates to something nonnegative, by the inductive hypothesis for (B). The interesting case is $i \leq K < j$, in which case, for some $\bar{\omega}'_1, \bar{\omega}'_2, \bar{\omega}'_3$ (which are stochastically determined by the observations from the top K arms in the queue)

$$
\begin{aligned}
&W_{s+1}(\bar{\omega}_1, 1, \bar{\omega}_2, 0, \bar{\omega}_3) - W_{s+1}(\bar{\omega}_1, 0, \bar{\omega}_2, 1, \bar{\omega}_3) \\
&= 1 + \beta E\Big[W_s(\bar{\omega}_1', p_{11}(1), \bar{\omega}_2', p_{01}(1), \bar{\omega}_3') \\
&\qquad - W_s(\bar{\omega}_1', \bar{\omega}_2', p_{11}(1), \bar{\omega}_3', p_{01}(1))\Big] \\
&\geq 1 + \beta E\Big[W_s(\bar{\omega}_1', \bar{\omega}_2', p_{11}(1), p_{01}(1), \bar{\omega}_3') \\
&\qquad - W_s(\bar{\omega}_1', \bar{\omega}_2', p_{11}(1), \bar{\omega}_3', p_{01}(1))\Big] \\
&= 1 - \beta + \beta E\Big[1 + W_s(\bar{\omega}_1', \bar{\omega}_2', p_{11}(1), p_{01}(1), \bar{\omega}_3') \\
&\qquad - W_s(\bar{\omega}_1', \bar{\omega}_2', p_{11}(1), \bar{\omega}_3', p_{01}(1))\Big],
\end{aligned}
$$

where the inequality follows from the inductive hypothesis for (B). The final line is nonnegative since the expression inside the expectation is nonnegative by the inductive hypothesis for (A).

Now consider proving an induction step for (A). Suppose that y occurs within the two expressions in the ith and jth place, $i < j$. As above, if $i, j \leq K$, then the expression of interest evaluates to 1. If $i, j > K$ the expression evaluates to something nonnegative, by the inductive hypothesis for (A). So the interesting case is $i \leq K < j$. In this case, let $\bar{\omega}_2 = (\bar{\omega}_2', x, \bar{\omega}_2'')$ so that

$$
\begin{aligned}
&1 + W_{s+1}(\bar{\omega}_1, y, \bar{\omega}_2, \bar{\omega}_3) - W_{s+1}(\bar{\omega}_1, \bar{\omega}_2, y, \bar{\omega}_3) \\
&= 1 + W_{s+1}(\bar{\omega}_1, y, \bar{\omega}_2', x, \bar{\omega}_2'', \bar{\omega}_3) \\
&\qquad - W_{s+1}(\bar{\omega}_1, \bar{\omega}_2', x, \bar{\omega}_2'', y, \bar{\omega}_3)
\end{aligned}
$$

and $\bar{\omega}_1$ and $\bar{\omega}_2'$ together account for $K - 1$ arms. As previously, the above expression above is of the form $a+bx+cy+dxy$. This is nonnegative for all $x, y \in [0, 1]$ if and only if it is nonnegative for all $(x, y) \in \{(0, 0), (0, 1), (1, 0), (1, 1)\}$. So we can prove nonnegativity by checking four cases. Two of them are the same. If $x = y = 0$ or $x = y = 1$ then

$$
\begin{aligned}
&1 + W_{s+1}(\bar{\omega}_1, y, \bar{\omega}_2', x, \bar{\omega}_2'', \bar{\omega}_3) - W_{s+1}(\bar{\omega}_1, \bar{\omega}_2', x, \bar{\omega}_2'', y, \bar{\omega}_3) \\
&= 1 + \beta E\Big[W_s(\bar{\omega}_1', \tau(x), \tau(\bar{\omega}_2''), \bar{\omega}_3') - W_s(\bar{\omega}_1', \tau(\bar{\omega}_2''), \tau(x), \bar{\omega}_3')\Big] \\
&= 1 - \beta + \beta E\Big[1 + W_s(\bar{\omega}_1', \tau(x), \tau(\bar{\omega}_2''), \bar{\omega}_3') \\
&\qquad - W_s(\bar{\omega}_1', \tau(\bar{\omega}_2''), \tau(x), \bar{\omega}_3')\Big].
\end{aligned}
$$

The term within the final expection is nonnegative, by the inductive hypothesis for (A).
If $x = 0$, $y = 1$,

$$
\begin{aligned}
&1 + W_{s+1}(\bar{\omega}_1, y, \bar{\omega}_2', x, \bar{\omega}_2'', \bar{\omega}_3) - W_{s+1}(\bar{\omega}_1, \bar{\omega}_2', x, \bar{\omega}_2'', y, \bar{\omega}_3) \\
&= 1 + \beta E\Big[1 + W_s(\bar{\omega}_1', \tau(1), \tau(0), \tau(\bar{\omega}_2''), \bar{\omega}_3') \\
&\qquad - W_s(\bar{\omega}_1', \tau(\bar{\omega}_2''), \tau(1), \bar{\omega}_3', \tau(0))\Big] \\
&\geq 1 + \beta E\Big[1 + W_s(\bar{\omega}_1', \tau(0), \tau(\bar{\omega}_2''), \tau(1), \bar{\omega}_3') \\
&\qquad - W_s(\bar{\omega}_1', \tau(0), \tau(\bar{\omega}_2''), \tau(1), \bar{\omega}_3')\Big] \\
&> 0,
\end{aligned}
$$

where the first inequality follows from the inductive hypothesis for (B).

If $x = 1$, $y = 0$,

$$
\begin{aligned}
&1 + W_{s+1}(\bar{\omega}_1, y, \bar{\omega}_2', x, \bar{\omega}_2'', \bar{\omega}_3) - W_{s+1}(\bar{\omega}_1, \bar{\omega}_2', x, \bar{\omega}_2'', y, \bar{\omega}_3) \\
&= 1 + \beta E\Big[W_s(\bar{\omega}_1', \tau(1), \tau(\bar{\omega}_2''), \bar{\omega}_3', \tau(0)) - 1 \\
&\qquad - W_s(\bar{\omega}_1', \tau(1), \tau(\bar{\omega}_2''), \tau(0), \bar{\omega}_3')\Big] \\
&= 1 - \beta + \beta E\Big[W_s(\bar{\omega}_1', \tau(1), \tau(\bar{\omega}_2''), \bar{\omega}_3', \tau(0)) \\
&\qquad - W_s(\bar{\omega}_1', \tau(1), \tau(\bar{\omega}_2''), \tau(0), \bar{\omega}_3')\Big].
\end{aligned}
$$

The term within the expectation is nonnegative by the inductive hypothesis for (B).

By contradiction, it is easy to show that the myopic policy also maximizes the expected total discounted reward and the expected average reward over the infinite horizon. ■

Remarks on Theorem 4

1. The myopic policy does something stronger than maximize the expected long-run reward. Let $r(t)$ be the total number of times we obtain the reward by time t, i.e., the number of $1s$ observed from the underlying random processes. Then by a similar proof as above, we can show that the myopic policy maximizes $\Pr(r(t) \geq m)$ for all m.
2. One can also generalize the result by allowing new arms to arrive over time. For the case of departing arms, imagine that each arm is failing into state 2, a 'dead state', e.g., the underlying Markov chain is
$$
\begin{pmatrix}
1 - \alpha - p_{01}(1) & p_{01}(1) & \alpha \\
1 - \alpha - p_{11}(1) & p_{11}(1) & \alpha \\
0 & 0 & 1
\end{pmatrix}.
$$
Thus, we can have a model with channels both arriving and leaving the population of available arms.
3. When the underlying random processes are non-Markovian, numerical examples showed that the myopic policy is not optimal over a *finite* horizon. Its optimality over the infinite horizon is, however, still open to future studies.

VI. Conclusion

In this paper, we studied a class of RMAB problems that model the monitoring and control of multiple reset processes. We showed that the RMAB satisfies Whittle indexability. When the RMAB satisfies certain monotone conditions on the stochastic evolution of the arm state, Whittle index was established in closed-form. Furthermore, when arms are stochastically identical, we showed that Whittle index policy is equivalent to the myopic policy. Based on this equivalency, the structure and optimality of Whittle index policy were established under certain conditions.

References

[1] Q. Zhao and B. Sadler, "A Survey of Dynamic Spectrum Access," *IEEE Signal Process. Mag.*, vol. 24, no. 3, pp. 79-89, May 2007.

[2] P. Whittle, "Restless Bandits: Activity Allocation in a Changing World," *J. Appl. Probab.*, vol. 25, pp. 287-298, 1988.

[3] 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, December 2008.

[4] 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, September, 2009.

[5] S. Ahmad and M. Liu, "Multi-channel Opportunistic Access: A Case of Restless Bandits with Multiple Plays," in *Proc. of Allerton Conference on Communication, Control, and Computing*, Allerton, IL, October 2009.

[6] K. Liu and Q. Zhao, "Indexability of Restless Bandit Problems and Optimality of Whittle Index for Dynamic Multichannel Access," *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5547-5567, November, 2010.

[7] J. Le Ny, M. Dahleh, E. Feron, "Multi-UAV Dynamic Routing with Partial Observations using Restless Bandit Allocation Indices," in *Proceedings of the 2008 American Control Conference*, Seattle, WA, June, 2008.

[8] J. E. Niño-Mora, "Restless Bandits, Partial Conservation Laws and Indexability," *Adv. Appl. Prob.*, vol. 33, pp. 76-98, 2001.

[9] 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.

[10] R. R. Weber and G. Weiss, "On an Index Policy for Restless Bandits," *J. Appl. Probab.*, vol.27, no.3, pp. 637-648, September 1990.

[11] R. R. Weber and G. Weiss, "Addendum to 'On an Index Policy for Restless Bandits," *Adv. Appl. Prob.*, vol. 23, no. 2, pp. 429-430, Jun., 1991.

[12] R. R. Weber, "On a Heuristic for Restless Bandits in Discounted-Reward Problems," work in progress, 2010.