Applied Probability Trust (4 February 2008)

INDEX POLICIES FOR DISCOUNTED BANDIT PROBLEMS WITH AVAILABILITY CONSTRAINTS

SAVAS DAYANIK,* *Princeton University*

WARREN POWELL,** *Princeton University*

KAZUTOSHI YAMAZAKI,*** *Princeton University*

Abstract

Multi-armed bandit problem is studied when the arms are not always available. The arms are first assumed to be intermittently available with some state/action-dependent probabilities. It is proven that no index policy can attain the maximum expected total discounted reward in every instance of that problem. The Whittle index policy is derived, and its properties are studied. Then it is assumed that arms may break down, but repair is an option at some cost, and the new Whittle index policy is derived. Both problems are indexable. The proposed index policies cannot be dominated by any other index policy over all multi-armed bandit problems considered here. Whittle indices are evaluated for Bernoulli arms with unknown success probabilities.

Keywords: Optimal resource allocation, multi-armed bandit problems, Gittins index, Whittle index, restart-in problem

2000 Mathematics Subject Classification: Primary 93E20

Secondary 90B36

1. Introduction

The classical multi-armed bandit problem considers the trade-off between exploration and exploitation. It deals with the situations in which one needs to decide between the alternative actions of either maximizing immediate reward or acquiring information that may help increase one's total reward in the future. In its typical formulation, decision maker has to choose at each stage one of N-arms of a slot machine to play in order to maximize expected total discounted reward over an infinite time horizon. The reward obtained from an arm depends on the state of a stochastic process that changes only when the arm is played.

* Postal address: Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544

* Email address: sdayanik@princeton.edu

Gittins and Jones [6] showed that each arm is associated with an index that is a function of the state of the arm, and that the expected total discounted reward over an infinite-horizon is maximized if an arm with the largest index is played every time. We call an arm *active* if it is played and *passive* otherwise. The proof (see Whittle [25] and Tsitsiklis [22]) relies on the condition that only active arms change their states. Due to this limitation, the range of problems where some index policies are guaranteed to be optimal is small. Nevertheless, the Gittins index policy is important because, by splitting the optimization problem into N independent smaller subproblems, it reduces the problem dimension considerably. Moreover, at each stage only one arm changes its state, and so at most one index has to be re-evaluated. For those reasons, many authors have generalized the classical multi-armed bandit problem and studied the performance of index policies designed for them.

In this paper, we study bandit problems where passive *and* active arms may become unavailable temporarily or permanently. Therefore, these are not classical multi-armed bandit problems, and the Gittins index policy is not optimal in general.

For example, in a flexible manufacturing system, a foreman may be unable to schedule some of the jobs on a day when certain raw materials are absent because of an unexpected transportation delay. In a construction project, bad weather may not allow the contractors to perform certain outside jobs that can otherwise be done simultaneously with other jobs. For a related but different set of scheduling problems involving servers subject to periods of unavailability, see Glazebrook [7, 8].

Bandit problems are also common in economics. Faced with a trade-off between exploration and exploitation, rational decision makers are assumed to act optimally using Gittins index policy. This framework was used to explain, for example, insufficient learning (Rothschild [21], Banks and Sundaram [1], Brezzi and Lai [4]), matching and job search (Jovanovic [13] and Miller [17]) and mechanism design (Bergemann and Valimaki [3]). However, decision makers do not act in the same way when the alternatives are likely to become unavailable in the future. Intuitively, the more pessimistic a decision maker is about the future availability of the alternatives, the more attention he pays to the immediate payoffs. Therefore, it is unlikely to expect that decision makers will use the Gittins index policy in these situations.

In a variation of the above-mentioned problem, we assume that arms may break down, but the decision maker has the option to fix them. For example, if an energy company loses its access to oil due to an unexpected international conflict, is it better to reestablish the access or to turn to an alternative energy source, e.g., natural gas or coal? A classical bandit problem with switching costs (see, for example, Jun [14]) is

a special case; arms break down immediately if they are not engaged, and if a broken arm is engaged, then the switching cost is incurred to pay for the repair. According to Bank and Sundaram [2], it is difficult to imagine an economic problem where the agent can switch between alternatives without incurring a cost. They also showed that, in the presence of switching costs, there does not exist an index policy that is optimal for every multi-armed bandit problem.

We generalize the classical multi-armed bandit problem as follows. There are N arms, and each arm is available with some state/action-dependent probability. At each stage, the decision maker chooses M arms to play simultaneously and collects rewards from each played arm. The reward from a particular arm n depends on a stochastic process $X_n = (X_n(t))_{t\geq 0}$, whose state changes only when the arm is played. The process X_n may represent, for example, the state of the knowledge about the reward obtainable from arm n.

At every stage, only a subset of the arms is available. We denote by $Y_n(t)$ the availability of arm n at time t; it is one if the arm is available at time t, and zero otherwise. Unlike X_n, the stochastic process Y_n changes even when the arm is not played. The objective is to find an optimal policy that chooses M arms so as to maximize the expected total discounted reward collected over the infinite time horizon. We study the following two problems:

Problem 1. Each arm is intermittently available. Its availability at time t is unobservable before time t. The conditional probability that an arm is available at time $t+1$ given

(i) the state $X(t)$ and its availability $Y(t)$ of the arm, and

(ii) whether or not the arm is played at time t

is known at time t. An arm cannot be played when it is unavailable.

This problem will not be well-defined unless there are at least M available arms to play at each stage. We can, however, let the decision maker pull fewer than M arms at a time by introducing sufficient number of arms that are always available and always give zero reward.

Problem 2. The arms are subject to failure, and the decision maker has the option to repair a broken arm. Irrespective of whether an arm is played at time t, it may break down and may not be available at time $t+1$ with some probability that depends on

(i) the state $X(t)$ of the arm at time t, and

(ii) whether or not the arm is played at time t.

If an arm is broken, the decision maker then has the option to repair it at some cost (or negative reward) that depends on $X(t)$. Repairing an arm is equivalent to playing the arm when it is broken. If a broken arm is repaired at time t, then it will become available at time $t+1$ with some conditional probability that depends only on the state $X(t)$ of the arm at time t. On the other hand, if it is not repaired, then the arm remains broken at time $t+1$.

We show that there does not exist a single index policy which is optimal for every instance of either problems. We propose a competitive index policy based on the Whittle index for restless bandit problems and show that there is not a single index policy that is better than the Whittle index policy for every instance of either problems. We evaluate the performance of the Whittle index policy for each type of problem both analytically and numerically.

The restless bandit problem was introduced by Whittle [26], and it is a generalization of the classical bandit problem in three directions: (i) the states of passive arms may change, (ii) rewards may be collected from passive arms, and (iii) $M \geq 1$ arms can be played simultaneously. Therefore, Problems 1 and 2 fall in the class of restless bandit problems, which are computationally intractable; Papadimitriou and Tsitsiklis [19] proved that they are PSPACE-hard. As in a typical restless bandit problem, we assume that rewards may be collected from passive arms and that more than one arm may be pulled simultaneously.

Whittle [26] introduced the so-called Whittle index to maximize the long-term average reward and characterized the index as a Lagrange multiplier for a relaxed conservation constraint, which ensures that on average M arms are played at each stage. See Niño-Mora [18] for the discounted case. The Whittle index policy makes sense if the problem is *indexable*. Weber and Weiss [23, 24] proved that under indexability, the Whittle index policy is asymptotically optimal as M and N tend to infinity while M/N is constant. The verification of indexability is difficult in general. Whittle [26] gave an example of an unindexable problem. However, indexability can be verified, and Whittle index policy can be developed analytically, for example, for the dual-speed restless bandit problem (see Glazebrook and Mitchell [10]) and a special problem with improving active arms and deteriorating passive arms (see Glazebrook et al. in [11]).

Glazebrook et al. [9] considered a problem in which passive arms are subject to permanent failure. They modeled it as a restless bandit, showed its indexability, and developed the corresponding Whittle index policy. Problems 1 and 2 are generalizations of their problem in that a broken arm is allowed to get back to the system, and that

both passive *and* active arms may break down. We prove that Problems 1 and 2 are indexable and derive the Whittle indices for them. Glazebrook et al.'s [9] and Gittins' indices turn out to be their special cases.

We also evaluate Whittle index policies numerically. Like the Gittins index, the Whittle indices for Problems 1 and 2 are also the solutions to suitable optimal stopping problems. We generalize Katehakis and Veinott's [15] restart-in formulation of the Gittins index to Problem 1's Whittle index. Problem 2's Whittle index turns out to be similar to the Gittins index, and we use the original restart-in problem to calculate the index for Problem 2.

In Section 2, we start by modeling Problems 1 and 2 as restless bandits. In Section 3, we review Whittle index and indexability. In Sections 4 and 5, we verify the indexability of Problems 1 and 2 and develop corresponding Whittle indices. We prove that no index policy can attain the maximum expected total discounted reward over the infinite time horizon in general in the class of Problems 1 and 2. A generalization of the restart-in problem to calculate Problem 1's Whittle index is discussed in Section 6. In Section 7 we introduce a numerical example with a Bernoulli reward process whose success probability is unknown and evaluate the index policies for Problems 1 and 2. Section 8 concludes with remarks.

2. Model

Using the same notation, X and Y defined in the previous section, the state of arm n at time t can be denoted by $S_n(t) = (X_n(t), Y_n(t))$. Then $(S_1(t), \dots, S_N(t))$ is the state at time $t \geq 0$ of the system with N arms. Suppose that X_n takes values in a countable state space $\mathcal{X}_n$ for every $n = 1, \dots, N$, and let $\mathcal{S}_n = \mathcal{X}_n \times \{0, 1\}$. Each process $(S_n(t))_{t\geq 0}$ is a controlled time-homogeneous Markov chain with $(S_n(t))_{t\geq 0}$-adapted control process

$$a_n(t) = \begin{cases} 1, & \text{if arm } n \text{ is played at time } t, \\ 0, & \text{otherwise.} \end{cases}$$

For every $1 \leq n \leq N$, the process $(X_n(t))_{t\geq 0}$ evolves according to some transition probability matrix $P^{(n)} = (p^{(n)}_{xx'})_{x,x'\in\mathcal{X}_n}$, if arm n is available and is played, and does not change otherwise; that is, for every $x, x' \in \mathcal{X}_n$,

$$\mathbb{P}\left\{X_n(t+1) = x' | X_n(t) = x, Y_n(t) = y, a_n(t) = a\right\} = \begin{cases} p^{(n)}_{xx'}, & \text{if } y = a = 1, \\ \delta_{xx'}, & \text{if } y = 0 \text{ or } a = 0, \end{cases} \tag{1}$$

where $\delta_{xx'}$ equals one if $x = x'$ and zero otherwise. Hence, even if arm n is active, the process X_n does not change if the arm is unavailable. In Problem 2, activating an

unavailable arm is equivalent to repairing it. In that case, the process X_n does not change; namely, repairing an arm changes only its availability. In Problem 1, activating an unavailable arm is not allowed.

The conditional probability that arm n is available at time $t+1$, given $X_n(t)$, $Y_n(t)$, and $a_n(t)$, is denoted by

$$\theta_n^a(x,y) := \mathbb{P}\left\{Y_n(t+1)=1 | X_n(t)=x, Y_n(t)=y, a_n(t)=a\right\} \tag{2}$$

for every $(x,y)\in\mathcal{S}_n$, $a\in\{0,1\}$, $t\geq 0$, and $1\leq n\leq N$. The random variable $Y_n(t+1)$ is conditionally independent of $X_n(t+1)$ and has conditionally a Bernoulli distribution with success probability $\theta_n^{a_n(t)}(X_n(t),Y_n(t))$ given $X_n(t)$, $Y_n(t)$, and $a_n(t)$. Let

$$R_n^a(x,y) := \text{the expected reward collected from arm } n \\ \text{given that } X_n(t)=x, Y_n(t)=y, a_n(t)=a \quad \text{for every } (x,y)\in\mathcal{S}_n,\ a\in\{0,1\},$$

and as in the classical bandit problem, we assume that $R_n^a(x,y)$ is bounded uniformly in $(x,y)\in\mathcal{S}_n$. Let $0<\gamma<1$ be a given discount rate. Then the expected discounted immediate reward at time t equals $\mathbb{E}\left[\gamma^t\sum_{n=1}^N R_n^{a_n(t)}(X_n(t),Y_n(t))\right]$.

The process $(S_1(t),\ldots,S_N(t))_{t\geq 0}$ is time-homogeneous and Markov; hence, we consider stationary policies $\pi:\mathcal{S}_1\times\ldots\times\mathcal{S}_N\mapsto\mathcal{A}:=\left\{a\in\{0,1\}^N : a_1+\ldots+a_N=M\right\}$. Denote for every fixed $((x_1,y_1),\ldots,(x_N,y_N))\in\mathcal{S}_1\times\ldots\times\mathcal{S}_N$, the value under a stationary policy π by $J^\pi(((x_1,y_1),\ldots,(x_N,y_N)))$, and it equals

$$\mathbb{E}^\pi\left[\sum_{t=0}^\infty\gamma^t\sum_{n=1}^N R_n^{a_n(t)}(X_n(t),Y_n(t))\,\middle|\,(X_n(0),Y_n(0))=(x_n,y_n), n=1,2,\ldots,N\right],$$

where $(a_1(t),\ldots,a_N(t))=\pi((X_1(t),Y_1(t)),\ldots,(X_N(t),Y_N(t)))$ for every $t\geq 0$. A policy $\pi^*\in\Pi$ is optimal if it maximizes $J^\pi(((x_1,y_1),\ldots,(x_N,y_N)))$ over $\pi\in\Pi$ for every initial state $((x_1,y_1),\ldots,(x_N,y_N))$ in $\mathcal{S}_1\times\ldots\times\mathcal{S}_N$.

3. The Whittle index and indexability

Let us fix an arm and drop the arm-specific indices from the notation in Section 2; instead, write $S(t)=(X(t),Y(t))$, $t\geq 0$ for its state process on $\mathcal{S}=\mathcal{X}\times\{0,1\}$, controlled by the $\{0,1\}$-valued $(S(t))_{t\geq 0}$-adapted process $a(t)$ according to transition probabilities

$$p^a_{(x,y),(x',y')} := \mathbb{P}\{(X(t+1),Y(t+1))=(x',y') \mid (X(t),Y(t))=(x,y), a(t)=a\}, \tag{3}$$

for every $(x,y),(x',y')\in\mathcal{S}$ and $a\in\{0,1\}$, which are determined by the transition probabilities $(p_{xx'})_{x,x'\in\mathcal{X}}$ of the Markov chain $(X(t))_{t\geq 0}$ on the space $\mathcal{X}$ and the

availability probabilities $\theta^a(x,y)$ $(a,y \in \{0,1\}, x \in \mathcal{X})$, as described in (1) and (2), respectively. Finally, let $R^a(x,y)$ $(a,y \in \{0,1\}, x \in \mathcal{X})$ denote the expected reward collected from the arm, and

$$\{0,1\} \supseteq \mathcal{A}(x,y) := \text{the set of actions available in state } (x,y) \in \mathcal{S}. \tag{4}$$

Recall that $\mathcal{A}(x,1) = \{0,1\}$ and $\mathcal{A}(x,0) = \{0\}$ for every $x \in \mathcal{X}$ in Problem 1, and $\mathcal{A}(x,y) = \{0,1\}$ for every $x \in \mathcal{X}$ and $y \in \{0,1\}$ in Problem 2.

Consider the following auxiliary problem. At each time, the decision maker can either activate the arm or leave it resting. Suppose that the current state of the arm is $(x,y) \in \mathcal{S}$. If $1 \in \mathcal{A}(x,y)$ and the arm is activated, then reward $R^1(x,y)$ is obtained. If it is rested, then a passive reward $R^0(x,y)$ *and a subsidy in the amount of* $W \in \mathbb{R}$ are obtained. The objective is to maximize the expected total discounted reward. Whittle [25] called this problem as the W*-subsidy problem*, which is a variation of the retirement problem; see, for example, Ross [20, Chapter VII]. The so-called Whittle index in state (x,y) corresponds by definition to the smallest subsidy amount W for which it is optimal to rest the arm.

After the Whittle index is calculated for every arm in their current states, the Whittle index policy is to activate M arms with the largest indices. However, this policy makes sense only if any arm rested under a subsidy W remains rested under every subsidy W' greater than W. Namely, the set of states at which it is optimal to rest the arm increases as the value of subsidy W increases. This property is called the *indexability*. These concepts were introduced originally by Whittle [26] in the average-reward case, and in the discounted case they were described by other authors; see, e.g., Niño-Mora [18].

For every fixed $W \in \mathbb{R}$, the value function $V((x,y),W)$, $(x,y) \in \mathcal{S}$ of the W-subsidy problem satisfies the dynamic programming equation

$$V((x,y),W) = \max_{a \in \mathcal{A}(x,y)} (L^a V)((x,y),W), \tag{5}$$

where

$$\begin{aligned} (L^1 V)((x,y),W) &= R^1(x,y) + \gamma \sum_{(x',y') \in \mathcal{S}} p^1_{(x,y),(x',y')} V((x',y'),W), \\ (L^0 V)((x,y),W) &= W + R^0(x,y) + \gamma \sum_{(x',y') \in \mathcal{S}} p^0_{(x,y),(x',y')} V((x',y'),W) \end{aligned} \tag{6}$$

are the maximum expected total discounted rewards if the initial action is to activate or to rest the arm, respectively, whenever those actions are allowed in state $(x,y) \in \mathcal{S}$.

Let $\Pi(W)$ be the subset of $\mathcal{S}$ in which it is optimal to rest the arm when the subsidy is W; namely,

$$\Pi(W) := \{(x,y) \in \mathcal{S} : \mathcal{A}(x,y) = \{0\}\} \bigcup \left\{(x,y) \in \mathcal{S} : \mathcal{A}(x,y) = \{0,1\} \text{ and } L^1((x,y),W) \le L^0((x,y),W)\right\}, \quad W \in \mathbb{R}. \tag{7}$$

If the arm is *indexable* and resting it is optimal for the subsidy amount W, then doing the same is also optimal whenever the subsidy amount is greater than W.

Definition 1. (*Indexability.*) An arm is indexable if $\Pi(W)$ is increasing in W; namely,

$$W_2 < W_1 \quad \Longrightarrow \quad \Pi(W_2) \subseteq \Pi(W_1). \tag{8}$$

Definition 2. (*Whittle index.*) The Whittle index of an indexable arm is defined as

$$W(x,y) := \inf\{W \in \mathbb{R} : (x,y) \in \Pi(W)\} \qquad \text{in every state } (x,y) \in \mathcal{S}. \tag{9}$$

Under indexability and whenever the infimum in (9) is attained, the Whittle index $W(x,y)$ is the smallest subsidy amount W for which both active and passive actions are optimal in state $(x,y) \in \mathcal{S}$.

Definition 3. (*Whittle index policy.*) Suppose that the arms of a restless bandit problem are indexable. The Whittle index policy plays M arms with the largest Whittle-indices.

The W-subsidy problem is one particular instance of Problems 1 and 2. If an index policy is optimal for every instance of Problems 1 or 2, then it must also be optimal for every W-subsidy problem. This observation will imply the non-existence of an index policy which is optimal for every instance of Problems 1 or 2; see Proposition 4 and 7.

4. The Whittle index for Problem 1

This section presents an index policy for Problem 1. We obtain the Whittle index by studying the W-subsidy problem and prove that no single index policy can be optimal for every instance of Problem 1.

Because in Problem 1 an unavailable arm cannot be activated, we have

$$\mathcal{A}(x,0) = \{0\} \quad \text{and} \quad \mathcal{A}(x,1) = \{0,1\} \qquad \text{for every} \quad x \in \mathcal{X}, \tag{10}$$

and (7) guarantees that no amount of passive subsidy is enough to change this constraint:

$$(x,0) \in \Pi(W) \qquad \text{for every} \quad x \in \mathcal{X},\ W \in \mathbb{R}. \tag{11}$$

As in Section 3 let us fix an arm and drop the indices. For every subsidy amount $W \in \mathbb{R}$, the value function $V((x,y),W)$, $(x,y) \in \mathcal{S}$ of the W-subsidy problem satisfies the dynamic programming equation in (5), where (6) becomes

$$\begin{aligned}(L^1V)((x,1),W) &= R^1(x,1) \\ &\quad + \gamma \sum_{x' \in \mathcal{X}} p_{xx'} \left[(1-\theta^1(x,1))V((x',0),W) + \theta^1(x,1)V((x',1),W)\right], \\ (L^0V)((x,y),W) &= W + R^0(x,y) \\ &\quad + \gamma \left[\left(1-\theta^0(x,y)\right) V\left((x,0),W\right) + \theta^0(x,y)V\left((x,1),W\right)\right]\end{aligned}$$

for every $x \in \mathcal{X}$ and $y \in \{0,1\}$. Let $\mathbb{P}^{1,0}$ be the probability law induced by the policy that the arm is active whenever it is available, and it is passive otherwise. Similarly, let $\mathbb{P}^{0,0}$ be the probability law induced when the arm is always rested. That is, for every $(x,y) \in \mathcal{S}$,

$$\begin{aligned}&\mathbb{P}^{1,0}\{X(t+1)=x', Y(t+1)=y' | X(t)=x, Y(t)=y\} \\ &\qquad = \left\{ \begin{array}{ll} p_{xx'}\left[\theta^1(x,1)\right]^{y'}\left[1-\theta^1(x,1)\right]^{1-y'}, & y=1 \\ \delta_{xx'}\left[\theta^0(x,0)\right]^{y'}\left[1-\theta^0(x,0)\right]^{1-y'}, & y=0 \end{array} \right\}, \quad \text{and} \\ &\mathbb{P}^{0,0}\{X(t+1)=x', Y(t+1)=y' | X(t)=x, Y(t)=y\} \\ &\qquad = \delta_{xx'}\left[\theta^0(x,y)\right]^{y'}\left[1-\theta^0(x,y)\right]^{1-y'}.\end{aligned}$$

Let $\mathbb{E}^{1,0}_{x,y}[\cdot]$ and $\mathbb{E}^{0,0}_{x,y}[\cdot]$ be the expectations under $\mathbb{P}^{1,0}$ and $\mathbb{P}^{0,0}$, respectively, given that $X(0)=x$ and $Y(0)=y$. Denote by $\rho(x,y)$ the expected total discounted reward from a passive arm whose current state is $(x,y) \in \mathcal{S}$; namely,

$$\rho(x,y) := \mathbb{E}^{0,0}_{x,y}\left[\sum_{t=0}^{\infty} \gamma^t R^0(X(t),Y(t))\right] = \mathbb{E}^{0,0}_{x,y}\left[\sum_{t=0}^{\infty} \gamma^t R^0(x,Y(t))\right]. \tag{12}$$

Let $(\mathcal{F}_t)_{t\geq 0}$ be the filtration generated by $(X(t),Y(t))_{t\geq 0}$, and $\mathscr{S}$ be the set of all $\mathbb{P}^{1,0}$-almost-surely (a.s.) positive stopping times of $(\mathcal{F}_t)_{t\geq 0}$, and define

$$\overline{\mathscr{S}} := \left\{ \tau \in \mathscr{S} : Y(\tau) = 1,\ \mathbb{P}^{1,0}\text{-a.s. on } \{\tau < \infty\} \right\}$$

as the set of positive stopping times at which the arm is available.

The following proposition states the existence of an index policy which is optimal for the W-subsidy problem, the indexability and the Whittle index for Problem 1. Most of the proofs are given in the appendix.

Proposition 1. *In the W-subsidy problem, it is optimal to rest an arm in state $(x,1)$ (namely, $(x,1)\in\Pi(W)$) for some $x\in\mathcal{X}$ if and only if*

$$W \geq (1-\gamma)\sup_{\tau\in\overline{\mathscr{T}}}\frac{\mathbb{E}^{1,0}_{x,1}\left[\sum_{t=0}^{\tau-1}\gamma^t R^{Y(t)}(X(t),Y(t))+\gamma^\tau\rho(X(\tau),1)\right]-\rho(x,1)}{1-\mathbb{E}^{1,0}_{x,1}\left[(1-\gamma)\sum_{t=1}^{\tau-1}\gamma^t 1_{\{Y(t)=0\}}+\gamma^\tau\right]} =: W(x,1). \quad (13)$$

Moreover, the arm is indexable, and its Whittle index $W(x,y)$ in state $(x,y)\in\mathcal{S}$ equals $-\infty$ if $y=0$ and is given by the righthand side of (13) if $y=1$.

The index $W(x,y)$, $(x,y)\in\mathcal{S}$ generalizes Glazebrook et al.'s [9] index for a problem where only passive arms may become unavailable and unavailable arms never become available.

It is easy to check that the Whittle index $W(x,1)$, $x\in\mathcal{X}$ coincides with the Gittins index if $\theta^0(x,y)=\theta^1(x,y)=1$ and $R^0(x,y)=0$ for every $(x,y)\in\mathcal{S}$ and with the immediate reward $R^1(x,1)$ if $\theta^0(x,y)=\theta^1(x,y)=R^0(x,y)=0$ for every $(x,y)\in\mathcal{S}$. The next proposition shows that, if $R^0(x,y)=0$ for every $(x,y)\in\mathcal{S}$, then $W(x,1)$ converges to the same limits as $\theta^0(x,y)=\theta^1(x,y)\equiv\theta$, $(x,y)\in\mathcal{S}$ tend simultaneously to the corresponding extreme values.

Proposition 2. *Suppose that $\theta^0(x,y)=\theta^1(x,y)=\theta$ and $R^0(x,y)=0$ for every $(x,y)\in\mathcal{S}$. Then the Whittle index $W(x,1)$ in (13) converges to the Gittins index as $\theta\nearrow 1$, and to the one-time reward $R^1(x,1)$ as $\theta\searrow 0$, uniformly in $x\in\mathcal{X}$.*

The next two propositions show that there cannot be a single index policy that is optimal in every instance of Problem 1. It turns out that, if there exists such an index policy, then its index must be a strict monotone transformation of the Whittle index $W(\cdot,\cdot)$ defined in Proposition 1. Then the non-existence of an index policy which is optimal in every instance of Problem 1 follows from an example in which the Whittle index policy is not optimal.

Proposition 3. *The index function of every index policy that performs better than other index policies for every instance of Problem 1 must be a strict monotone transformation of the Whittle index $W(\cdot,\cdot)$ of Proposition 1.*

Proposition 4. *There does not exist a single index policy that performs as good as the Whittle index policy in every instance of Problem 1 and strictly better in at least one of the instances. Therefore, there is not an index policy optimal in every instance of Problem 1.*

Proof of Proposition 4. According to Proposition 3, the index function of every index policy that performs, in every instance of Problem 1, at least as good as the Whittle index policy is a strictly monotone transformation of the Whittle index; therefore, the performance of those index policies cannot be strictly better than that of the Whittle index policy in any given instance of Problem 1. This proves the first part of Proposition 4. By the same token, if there is an index policy which is optimal in every instance of Problem 1, then so must the Whittle index policy be. We will now give a counter-example to the latter.

Consider a case with two arms. Arm 1 is always available, and arm 2 is available with probability $\epsilon \in (0,1)$. Passive arms do not give rewards and $M = 1$ arm is played in each period. The reward from arm 1 changes deterministically under the active action as in $1 \to 100 \to 10 \to 10 \to \ldots \to 10 \to \ldots$. Let the corresponding states of arm 1 be $x_{11}, x_{12}, x_{13}, \ldots$. The state x_2 of arm 2 never changes and gives a constant reward of 40 when it is available and activated. Arms 1 and 2 are initially available in states x_{11} and x_2, respectively. Let $\epsilon = 0.01$ and $\gamma = 0.7$. After obvious choices of stopping times τ in (13), the Whittle index $W_1(\cdot, 1)$ for arm 1 satisfies the bounds

$$\begin{aligned} W_1(x_{11},1) \geq \frac{1+\gamma 100}{1+\gamma} = 41.76, \quad W_1(x_{12},1) \geq \frac{100}{1} = 100, \\ W_1(x_{1n},1) = \ 10, \ n \geq 3, \end{aligned} \tag{14}$$

and $W_2(x_2, 1) = 40$ and $W_2(x_2, 0) = -\infty$ for arm 2.

According to the Whittle index policy, arm 1 must be pulled when $X_1 = x_{11}$ and x_{12}, and arm 2—if it is available and arm 1 otherwise—when $X_1 = x_{13}, x_{14}, \ldots$. That is, the optimal policy is to pull arm 1 twice first and later pull arm 2 always if arm 2 is available and arm 1 otherwise. Therefore, the value function $U(\cdot, \cdot)$ of the Whittle index policy becomes

$$\begin{aligned} U((x_{11},1),(x_2,1)) &= 1 + \gamma 100 + \gamma^2[10(1-\epsilon) + 40\epsilon] + \gamma^3[10(1-\epsilon) + 40\epsilon] + \ldots \\ &\approx 87.8233, \end{aligned}$$

and $U((x_{11}, 1), (x_2, 0)) = U((x_{11}, 1), (x_2, 1))$. However, pulling arm 2 initially in the state $((x_{11}, 1), (x_2, 1))$ and then executing Whittle index policy gives a better value:

$$\begin{aligned} 40 + \gamma[\epsilon U((x_{11},1),(x_2,1)) + (1-\epsilon)U((x_{11},1),(x_2,0))] \\ \approx 101.4763 > U((x_{11},1),(x_2,1)). \end{aligned}$$

Therefore, the Whittle index policy is not always optimal, and there is not an index policy which is optimal in every instance of Problem 1.

5. The Whittle index for Problem 2

Unlike as in the previous section, here we assume that the active action is always available—even when the arm is unavailable. An unavailable arm is regarded as a broken arm that needs a repair before it can start giving rewards. Activating a broken arm is equivalent to repairing it. Therefore, $-R_n^1(x,0)$ denotes the repair cost when arm n is broken in state $x \in \mathcal{X}$. If a broken arm is not repaired, then it will remain unavailable and stay broken until the next stage with probability one. We assume that passive arms do not give rewards, and that the reward obtained from activating an available arm is positive as in the following condition.

Condition 1. *For every $(x,y) \in \mathcal{S}$, and $n = 1,\dots,N$, suppose that $R_n^1(x,1) \geq 0$, $R_n^1(x,0) < 0$, $R_n^0(x,y) = 0$, and $\theta_n^0(x,0) = 0$.*

Let us fix an arm, drop all of the indices identifying the arm, and consider the W-subsidy problem. In every state $(x,y) \in \mathcal{S}$ of the arm both the active and passive actions are available; i.e., $\mathcal{A}(x,y) = \{0,1\}$ for every $(x,y) \in \mathcal{S}$, and the value function $V((x,y),W)$, $(x,y) \in \mathcal{S}$, $W \in \mathbb{R}$ of the W-subsidy problem satisfies (5), where (6) becomes under Condition 1 that

$$\begin{aligned}
(L^1V)((x,1),W) &= R^1(x,1) \\
&\quad + \gamma \sum_{x' \in \mathcal{X}} p_{xx'} \left[(1-\theta^1(x,1))V((x',0),W) + \theta^1(x,1)V((x',1),W) \right], \\
(L^1V)((x,0),W) &= R^1(x,0) + \gamma \left[(1-\theta^1(x,0))V((x,0),W) + \theta^1(x,0)V((x,1),W) \right], \\
(L^0V)((x,1),W) &= W + \gamma \left[(1-\theta^0(x,1))V((x,0),W) + \theta^0(x,1)V((x,1),W) \right], \\
(L^0V)((x,0),W) &= W + \gamma V((x,0),W).
\end{aligned}$$

Let $\mathbb{P}^{1,1}$ be the probability law induced by the policy that activates the arm forever and $\mathbb{E}^{1,1}$ denote the expectation under $\mathbb{P}^{1,1}$. Let $\psi(x)$ be the expected total discounted reward if the arm is active forever starting in state $(x,1)$ at time zero; namely,

$$\psi(x) := \mathbb{E}_{x,1}^{1,1} \left[\sum_{t=0}^{\infty} \gamma^t R^1(X(t),Y(t)) \right], \qquad x \in \mathcal{X}.$$

Condition 2. *Suppose $\psi(x) \geq R^1(x,0)/(1-\gamma) \equiv \sum_{t=0}^{\infty} \gamma^t R^1(x,0)$ for every $x \in \mathcal{X}$.*

Condition 2 is satisfied if (i) the arm never breaks down under the active action; i.e., $\theta^1(x,1) = 1$ for every $x \in \mathcal{X}$, or (ii) $R^1(X(t),0)$ is constant or non-decreasing almost surely under the active action.

Proposition 5. *Under Condition 2, activating the arm in state $(x, y) \in \mathcal{S}$ is optimal in the W-subsidy problem for Problem 2 if and only if*

$$W \geq (1-\gamma) \sup_{\tau \in \mathscr{S}} \frac{\mathbb{E}^{1,1}_{x,y}\left[\sum_{t=0}^{\tau-1} \gamma^t R^1(X(t), Y(t))\right]}{1 - \mathbb{E}^{1,1}_{x,y}\left[\gamma^\tau\right]} =: W(x, y). \tag{15}$$

The arm is indexable with Whittle index $W(x, y)$ defined by the righthand side of (15).

If active arms do not break down and broken arms cannot be repaired, then Problem 2 reduces to a problem studied by Glazebrook et al. [9] where passive arms do not give rewards, and the indices coincide.

Problem 2 reduces to the bandit problem with switching costs if $\theta^1(x, 1) = 1$ and $\theta^0(x, 1) = \theta^0(x, 0) = 0$, and $-R^1(x, 0)$ is the cost of switching to the arm, currently idling in state x. Because $\theta^1(x, 1) = 1$, Condition 2 is satisfied, and this version of Problem 2 is indexable. Glazebrook et al. [12] formulated the same problem, slightly differently from us, as a restless bandit problem where one does not wait for the broken arm to be fixed before the reward stream is again available: if one plays a broken arm, then he obtains its immediate reward minus the switching cost, and the arm is guaranteed to be available in the next period. However, the forms of their and our Whittle indices are the same. Their numerical studies suggest that the Whittle index policy is near-optimal for their version of Problem 2.

As in Problem 1, an index policy which is optimal in every instance of Problem 2 does not exist. We show that if there exists one, then its index function must be a strictly monotone transformation of the Whittle index, and we give an example where the Whittle index policy is not optimal. The proof of Proposition 6 is very similar to that of Proposition 3.

Proposition 6. *If an index policy is optimal in every instance of Problem 2, then its index function is a strictly monotone transformation of Whittle index $W(\cdot, \cdot)$ in (15).*

Proposition 7. *An index policy optimal in every instance of Problem 2 does not exist.*

Proof of Proposition 7. Suppose that γ, arms 1 and 2 are the same as in the proof for Proposition 4, except that arm 2 does not break down if it is active but it breaks down as soon as it is passive; if it is repaired, then it becomes available the next period with probability one (namely, $\theta^1_2(x_2, 1) = \theta^1_2(x_2, 0) = 1$, $\theta^0_2(x_2, 0) = \theta^0_2(x_2, 0) = 0$), and the repair cost equals $R^1_2(x_2, 0) = 100$. The Whittle index of arm 1 still satisfies the

inequalities in (14) while the Whittle index of arm 2 satisfies $W_2(x_2,1) \leq 40$, and

$$W_2(x_2,0) = \sup_{\tau\in\mathscr{S}} \frac{\mathbb{E}^{1,1}_{x_2,0}\left[\sum_{t=0}^{\tau-1}\gamma^t R(X(t),Y(t))\right]}{\mathbb{E}^{1,1}_{x_2,0}\left[\sum_{t=0}^{\tau-1}\gamma^t\right]} \leq \sup_{\tau\in\mathscr{S}} \mathbb{E}^{1,1}_{x_2,0}\left[\sum_{t=0}^{\tau-1}\gamma^t R(X(t),Y(t))\right]$$
$$\leq R_2^1(x_2,0) + \gamma 40 + \gamma^2 40 + \ldots = -100 + \frac{40\gamma}{1-\gamma} \leq 10.$$

Hence, Whittle index policy pulls arm 1 forever because it starts with pulling arm 1 at time 0, and arm 2 breaks down immediately as a result. Thus, its value function $U(\cdot)$ satisfies $U((x_{11},1),(x_2,1)) = 1 + 100\gamma + 10\gamma^2 + 10\gamma^3 + \ldots \approx 87.3$. However, pulling arm 2 forever gives $40 + 40\gamma + 40\gamma^2 \ldots \approx 133 > 87.3$, and Whittle index policy in (15) is not optimal, and by Proposition 6 no index policy is optimal in every instance of Problem 2.

6. The restart-in problem

We have developed the Whittle indices for Problems 1 and 2 in the previous sections. Here we discuss how to compute the indices in (13) and (15). For this purpose, we develop the *restart-in problem* representation of the indices. The restart-in problem representation of the Gittins index for the classical multi-armed bandit problem was introduced by Katehakis and Veinott [16]. The index in (15) is similar to the Gittins index; therefore, we first formulate it as a restart-in problem. We then propose a generalization of restart-in problem representation for the Whittle index in (13).

6.1. The restart-in problem for the Gittins index

We first review Katehakis and Veinott's [16] formulation of the Gittin's index as a restart-in problem. Consider a classical multi-armed bandit problem where the state of a fixed arm evolves according to a Markov chain $S = (S(t))_{t\geq 0}$ on some countable state space $\mathcal{S}$ with transition probability matrix $P = (p_{ss'})_{s,s'\in\mathcal{S}}$ under the active action, and let $R(s)$ be the one-time reward obtained if the arm is activated in state $s \in \mathcal{S}$.

Katehakis and Veinott [16] showed that the Gittins index of the arm in state $\widetilde{s} \in \mathcal{S}$ equals $(1-\gamma)\nu_{\widetilde{s}}^{(\widetilde{s})}$ with

$$\nu_{\widetilde{s}}^{(\widetilde{s})} = \sup_{\tau>0} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1}\gamma^t R(S(t))\middle|S(0)=\widetilde{s}\right]}{(1-\gamma)\mathbb{E}\left[\sum_{t=0}^{\tau-1}\gamma^t\right]} = \sup_{\tau>0} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1}\gamma^t R(S(t))\middle|S(0)=\widetilde{s}\right]}{\mathbb{E}\left[1-\gamma^\tau\right]}, \quad (16)$$

where the suprema are taken over strictly positive stopping times of the state process S, and $(\nu_s^{(\widetilde{s})})_{s\in\mathcal{S}}$ is the value function of the so-called *restart-in-state-$\widetilde{s}$ problem*. In a

restart-in-state-$\widetilde{s}$ problem, the state process S evolves according to transition probability matrix P and a reward is collected in each state until every time we decide to restart the process S in state $\widetilde{s}$ and then continue to collect the rewards afterwards. The objective is to choose the restart times so as to maximize the expected total discounted reward over an infinite time horizon, and the value function $(\nu_s^{(\widetilde{s})})_{s\in\mathcal{S}}$ of this Markov decision process is easily shown to satisfy the optimality equations

$$\nu_s^{(\widetilde{s})} = \max\left\{R(s) + \gamma \sum_{s'\in\mathcal{S}} p_{ss'}\nu_{s'}^{(\widetilde{s})},\ R(\widetilde{s}) + \gamma \sum_{s'\in\mathcal{S}} p_{\widetilde{s}s'}\nu_{s'}^{(\widetilde{s})}\right\}, \quad s\in\mathcal{S}. \tag{17}$$

The Gittins index $(1-\gamma)\nu_{\widetilde{s}}^{(\widetilde{s})}$ for every fixed state $\widetilde{s}\in\mathcal{S}$ is obtained after solving $|\mathcal{S}|$ equations in (17) simultaneously for $(\nu_s^{(\widetilde{s})})_{s\in\mathcal{S}}$, for example, by applying the value-iteration algorithm to (17). We will now characterize the Whittle indices in (13) and (15) of a potentially unavailable arm in terms of the value function of a restart-in problem.

6.2. The representation of the Whittle index of Problem 2 in terms of restart-in problems

Because the Whittle index $W(x,y)$ in (15) and Gittins index in (16) are similar, we can use the restart-in problem in (17) associated with the Gittins index. Let (X,Y) be the state process of a fixed arm on the state space $\mathcal{S}$ as described in Section 5. Then, for every fixed state $(\widetilde{x},\widetilde{y})\in\mathcal{S}$, the Whittle index $W(\widetilde{x},\widetilde{y})$ in (15) equals $(1-\gamma)\nu_{\widetilde{x},\widetilde{y}}^{(\widetilde{x},\widetilde{y})}$, where $(\nu_{x,y}^{(\widetilde{x},\widetilde{y})})_{(x,y)\in\mathcal{S}}$ is the value function of the restart-in-state-$(\widetilde{x},\widetilde{y})$ problem for the process (X,Y) under probability measure $\mathbb{P}^{1,1}$ (namely, the arm is always activated—both when it is available and when it is unavailable) and satisfies the optimality equations

$$\nu_{x,y}^{(\widetilde{x},\widetilde{y})} = \max\left\{R^1(x,y) + \gamma \sum_{x'\in\mathcal{X}} p_{xx'}\left[(1-\theta^1(x,y))\nu_{x',0}^{(\widetilde{x},\widetilde{y})} + \theta^1(x,y)\nu_{x',1}^{(\widetilde{x},\widetilde{y})}\right],\right.$$
$$\left. R^1(\widetilde{x},\widetilde{y}) + \gamma \sum_{x'\in\mathcal{X}} p_{\widetilde{x}x'}\left[(1-\theta^1(\widetilde{x},\widetilde{y}))\nu_{x',0}^{(\widetilde{x},\widetilde{y})} + \theta^1(\widetilde{x},\widetilde{y})\nu_{x',1}^{(\widetilde{x},\widetilde{y})}\right]\right\}, \qquad \forall (x,y)\in\mathcal{S}. \tag{18}$$

6.3. The representation of the Whittle index of Problem 1 in terms of restart-in problems

Let now (X,Y) be the state process of a fixed arm on the state space $\mathcal{S}$ as described in Section 4. For every fixed $\widetilde{x}\in\mathcal{X}$, recall from Proposition 1 that the Whittle index $W(\widetilde{x},0)$ in state $(\widetilde{x},0)\in\mathcal{S}$ equals $-\infty$, and $W(\widetilde{x},1)$ in state $(\widetilde{x},1)\in\mathcal{S}$ is given by the

expression in (13). In Proposition 8 below, we show that $W(\widetilde{x},1) = (1-\gamma)\nu^{(\widetilde{x},1)}_{(\widetilde{x},1)}$ if $(\nu^{(\widetilde{x},1)}_{x,y})_{(x,y)\in\mathcal{S}}$ is the solution of the equations

$$\nu^{(\widetilde{x},1)}_{x,1} = \max\left\{(L\nu^{(\widetilde{x},1)}_{\bullet})_{x,1}, (L\nu^{(\widetilde{x},1)}_{\bullet})_{\widetilde{x},1}\right\}, \tag{19}$$

$$\nu^{(\widetilde{x},1)}_{x,0} = (1-\gamma)\left[\nu^{(\widetilde{x},1)}_{\widetilde{x},1} - \rho(x,1)\right] + R^0(x,0) + \gamma\left[(1-\theta^0(x,0))\nu^{(\widetilde{x},1)}_{x,0} + \theta^0(x,0)\nu^{(\widetilde{x},1)}_{x,1}\right] \tag{20}$$

where $\rho(\cdot,\cdot)$ and $(Lw_{\bullet})_{x,1}$ for any $w : \mathcal{X}\times\{1\} \mapsto \mathbb{R}$ are defined by (12) and by

$$R^1(x,1) - \rho(x,1) + \gamma\sum_{x'\in\mathcal{X}} p_{xx'}\rho(x',1) + \gamma\sum_{x'\in\mathcal{X}} p_{xx'}\left[(1-\theta^1(x,1))w^{(\widetilde{x},1)}_{x',0} + \theta^1(x,1)w^{(\widetilde{x},1)}_{x',1}\right],$$

respectively, for every $x \in \mathcal{X}$. If we substitute $\nu^{(\widetilde{x},1)}_{x',0}$, $x' \in \mathcal{X}$ from (20) into (19), then

$$\nu^{(\widetilde{x},1)}_{x,1} = \max\left\{R(x) + \gamma\sum_{x'\in\mathcal{X}} p^{(\widetilde{x})}_{xx'}\nu^{(\widetilde{x},1)}_{x',1},\ R(\widetilde{x}) + \gamma\sum_{x'\in\mathcal{X}} p^{(\widetilde{x})}_{\widetilde{x}x'}\nu^{(\widetilde{x},1)}_{x',1}\right\}, \quad \forall x\in\mathcal{X}, \tag{21}$$

where

$$p^{(\widetilde{x})}_{xx'} := \begin{cases} p_{xx'}q_{xx'}, & \text{if} \quad x' \in \mathcal{X}\setminus\{\widetilde{x}\}, \\ p_{x\widetilde{x}}q_{x\widetilde{x}} + \displaystyle\sum_{x''\in\mathcal{X}} p_{xx''}\widetilde{q}_{xx''}, & \text{if} \quad x' = \widetilde{x}, \end{cases} \tag{22}$$

$$\left\{\begin{array}{l} R(x) := R^1(x,1) - \rho(x,1) + \gamma\displaystyle\sum_{x'\in\mathcal{X}} p_{xx'}\left[\frac{\theta^1(x,1)-\gamma(1-\theta^0(x,0))}{1-\gamma(1-\theta^0(x,0))}\rho(x',1)\right. \\ \qquad\qquad\qquad\qquad \left. + \dfrac{1-\theta^1(x,1)}{1-\gamma(1-\theta^0(x,0))}R^0(x',0)\right] \\ q_{xx'} := \theta^1(x,1) + (1-\theta^1(x,1))\dfrac{\gamma\theta^0(x',1)}{1-\gamma(1-\theta^0(x',0))} \\ \widetilde{q}_{xx'} := (1-\theta^1(x,1))\dfrac{1-\gamma}{1-\gamma(1-\theta^0(x',0))} \qquad \text{for every } x,x'\in\mathcal{X} \end{array}\right\}. \tag{23}$$

Let $(\widetilde{X}(t))_{t\geq 0}$ be a new Markov chain on the state space $\mathcal{X}$ with one-step transition probabilities $(p^{(\widetilde{x})}_{xx'})_{x,x'\in\mathcal{X}}$ in (22); note that $q_{xx'} + \widetilde{q}_{xx'} = 1$, and $\sum_{x'\in\mathcal{X}} p^{(\widetilde{x})}_{xx'} = 1$. Then (21) shows that $(\nu^{(\widetilde{x},1)}_{x,1})_{x\in\mathcal{X}}$ is the value function of the restart-in-state-$\widetilde{x}$ problem for the Markov chain $\widetilde{X}$ with running-reward function $(R(x))_{x\in\mathcal{X}}$. Indeed, because this is a discounted Markov decision process problem with a finite number of actions ("continue" or "restart") and a bounded running-reward function on a countable state space, the function $(\nu^{(\widetilde{x},1)}_{x,1})_{x\in\mathcal{X}}$ is the unique solution of the equations in (21) and is the uniform

limit of a sequence of functions obtained successively by applying the value-iteration algorithm to any initial bounded function defined on the state space; see Katehakis and Veinott [16], Ross [20, Chapter II].

Proposition 8. *For every fixed $\widetilde{x} \in \mathcal{X}$, let $(\nu_{x,y}^{(\widetilde{x},1)})_{x\in\mathcal{X},y\in\{0,1\}}$ be the unique solution of (19) and (20), equivalently, (21). Then for Problem 1 the Whittle index $W(\widetilde{x},1)$ of the arm (X,Y) in state $(\widetilde{x},1) \in \mathcal{S}$ equals $(1-\gamma)\nu_{\widetilde{x},1}^{(\widetilde{x},1)}$; namely,*

$$\nu_{\widetilde{x},1}^{(\widetilde{x},1)} = \frac{W(\widetilde{x},1)}{1-\gamma} \equiv \sup_{\tau\in\overline{\mathscr{T}}} \frac{\mathbb{E}_{x,1}^{1,0}\left[\sum_{t=0}^{\tau-1}\gamma^t R^{Y(t)}(X(t),Y(t)) + \gamma^\tau \rho(X(\tau),1)\right] - \rho(x,1)}{1 - \mathbb{E}_{x,1}^{1,0}\left[(1-\gamma)\sum_{t=1}^{\tau-1}\gamma^t 1_{\{Y(t)=0\}} + \gamma^\tau\right]}.$$

Remark 1. Suppose that $R^0(x,y) = 0$, $\theta^1(x,1) = 1$ for every $x \in \mathcal{X}$, $y \in \{0,1\}$ as in the classical multi-armed bandit problem. Then (23) and (22) become $R(x) = R^1(x,1)$, $q_{xx'} = 1 - \widetilde{q}_{xx'} = 1$, and $p_{xx'}^{(\widetilde{x})} = p_{xx'}$ for every $x, x', \widetilde{x} \in \mathcal{X}$, and (19) and (21) reduce to

$$\nu_{x,1}^{(\widetilde{x},1)} = \max\left\{R^1(x,1) + \gamma\sum_{x'\in\mathcal{X}} p_{xx'}\nu_{x',1}^{(\widetilde{x},1)},\ R^1(\widetilde{x},1) + \gamma\sum_{x'\in\mathcal{X}} p_{\widetilde{x}x'}\nu_{x',1}^{(\widetilde{x},1)}\right\},\ \forall x \in \mathcal{X}, \tag{24}$$

which is the restart-in-state-$\widetilde{x}$ problem uniquely solved by $(\nu_{x,1}^{(\widetilde{x},1)})_{x\in\mathcal{X}}$ in the Gittins index $(1-\gamma)\nu_{\widetilde{x},1}^{(\widetilde{x},1)}$ of the arm in state $(\widetilde{x},1)$, as shown by Katehakis and Veinott [16]. Thus, the problem in (21) is the natural generalization of that in (24) from an arm that is always available to an arm that is intermittently available as in the description of Problem 1.

7. Numerical example

We evaluate the performance of the Whittle index policies defined by (13) and (15) for Problems 1 and 2, respectively, through an example in which the reward of each active arm is a Bernoulli random variable with some unknown success probability. The success probability of arm n is a random variable λ_n, having a beta posterior distribution with parameters a and b, which depend on the prior distribution at time 0 and the number of successes in the previous trials with the same arm; namely,

$$\mathbb{P}\{\lambda_n \in dr \mid a,b\} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} r^{a-1}(1-r)^{b-1}dr, \qquad r \in (0,1).$$

More precisely, if the prior distribution of λ_n is beta with parameters (a,b), then after c successes and d failures in the plays so far with this arm, the posterior probability distribution of λ_n is also beta with parameters $(a+c, b+d)$. Thus, the parameters

$X_n(t)$ of the posterior beta distribution of λ_n after t plays is a Markov chain with one-step transition probabilities

$$p_{(a,b),(a+1,b)} = \frac{a}{a+b}, \qquad p_{(a,b),(a,b+1)} = \frac{b}{a+b}, \qquad a, b > 0.$$

Let $Y_n(t)$ be the indicator of whether the arm is available at time t, and $(X_n, Y_n) = (X_n(t), Y_n(t))_{t\geq 0}$ be a Markov process as in the model of Section 2. The conditional expected reward from active arm n at time t given $X_n(t) = (a, b)$ and $Y_n(t) = 1$ is $a/(a + b)$. Finally, let $-R^1_n((a, b), 0) = C_n > 0$ be the constant repair cost for arm n in Problem 2.

We calculate Whittle indices by solving the Bellman equations for the restart-in problem defined in Section 6. In a classical bandit problem with two Bernoulli arms, Katehakis and Derman [15] calculated Gittins index for each arm after truncating the state space of X to

$$Z_L = \big\{(a, b) \in \mathbb{N}^2 : a + b \leq L\big\} \qquad \text{for some fixed integer } L > 0. \tag{25}$$

Similarly, we only consider states $s = (x, y)$ where $x = (a, b) \in Z_L$ and $y \in \{0, 1\}$. Katehakis and Derman [15] proved that as L increases, their approximation converges to the value of the Gittins index. It is easy to prove that the same result holds in our settings.

We now evaluate the Whittle index policies for Problems 1 and 2 and compare them to other index policies based on Monte Carlo simulations with $1,000,000$ samples. See Dayanik et al. [5] for more numerical results and comparisons to the optimal policies.

7.1. Numerical results for Problem 1

We compare the value function of the Whittle index policy with that of the Gittins index policy defined below.

Definition 4. Gittins index policy chooses M available arms with the largest Gittins indices, and ties are broken randomly. Gittins and Whittle indices coincide when $\theta = 1$.

We compared Whittle and Gittins index policies in three cases shown in Table 1, where M arms have $\theta = 1$ and every time M of N arms are played. The parameters of the initial beta distribution of the reward from each arm is $(a, b) = (1, 1)$. Whittle index policy outperforms Gittins index policy in most examples. Gittins index policy does not use the likelihood of each arm's future availability, but Whittle index policy does. Gittins index policy should still give tight lower bounds because it is optimal when each arm is always available.

N	M	Whittle (95% CI)	Gittins (95% CI)	M	Whittle (95% CI)	Gittins (95% CI)
Case 1: $N/2$ **arms are available with probability** $\theta = 1.0$ **and** 0.5						
12	1	6.9357 (6.9245, 6.9464)	6.5600 (6.5502, 6.5698)	2	13.7052 (13.6885, 13.7219)	13.3439 (13.3288, 13.3590)
36	1	7.0297 (7.0193, 7.0401)	6.6346 (6.6246, 6.6446)	6	41.5967 (41.5687, 41.6254)	40.3890 (40.3643, 40.4137)
Case 2: $N/3$ **arms are available with probability** $\theta = 1.0$**,** 0.7**, and** 0.3						
12	1	6.9309 (6.9199, 6.9419)	6.5544 (6.5446, 6.5642)	2	13.5260 (13.5088, 13.5432)	13.1447 (13.1298, 13.1596)
36	1	7.0266 (7.0162, 7.0370)	6.4499 (6.4401, 6.4597)	6	41.1573 (41.1283, 41.1863)	39.7806 (39.7561, 39.8051)
Case 3: $N/6$ **arms are available with probability** $\theta = 1.0$**,** 0.9**,** 0.7**,** 0.5**,** 0.3**, and** 0.1						
12	1	6.8716 (6.8604, 6.8828)	6.5676 (6.5574, 6.5778)	2	13.3253 (13.3079, 13.3427)	12.9986 (12.9829, 13.0143)
36	1	7.0178 (7.0071, 7.0283)	6.6399 (6.6299, 6.6499)	6	40.7183 (40.6842, 40.7434)	39.5345 (39.5079, 39.5609)

TABLE 1: Expected total discounted rewards with Whittle and Gittins index policies for Problem 1.

N	M	Whittle (95% CI)	Policy 1 (95% CI)	Policy 2 (95% CI)
Case 1: N arms w/ $\theta = (1, 0, 0.5)$**,** $C = 0.5$				
2	1	2.6785 (2.6746, 2.6824)	1.1976 (1.1954, 1.1998)	2.6789 (2.6750, 2.6828)
6	1	3.4055 (3.4024, 3.4086)	1.7559 (1.7535, 1.7583)	3.4119 (3.4088, 3.4150)
10	1	3.6048 (3.6017, 3.6079)	2.0158 (2.0134, 2.0182)	3.6103 (3.6072, 3.6134)
Case 2: N arms w/ $\theta = (0.5, 0.5)$**,** $C = 1.0$				
2	1	1.1374 (1.1354, 1.1394)	1.1968 (1.1946, 1.1990)	-0.8995 $(-0.9038, -0.8952)$
6	1	1.6653 (1.6631, 1.6675)	1.7555 (1.7531, 1.7579)	0.1800 (0.1761, 0.1839)
10	1	1.9111 (1.9087, 1.9135)	2.0134 (2.011, 2.0158)	0.5867 (0.5828, 0.5906)
Case 3: N/2 arms w/ $\theta = (1.0, 0.5)$**,** $C = 2.0$**; N/2 arms w/** $\theta = (0.5, 0.5)$**,** $C = 1.0$				
2	1	1.1636 (1.1616, 1.1656)	1.1974 (1.1952, 1.1996)	-1.2804 $(-1.2849, -1.2759)$
6	1	1.7071 (1.7047, 1.7095)	1.7563 (1.7539, 1.7587)	-0.1007 $(-0.1048, -0.0966)$
10	1	1.9605 (1.9581, 1.9629)	2.0142 (2.0118, 2.0166)	0.3258 (0.3219, 0.3297)

TABLE 2: Expected total discounted rewards under Whittle index policy and Policies 1 and 2 for Problem 2.

7.2. Numerical results for Problem 2

Problem 2 will be more realistic if the controller has the option to retire, because fixing some of the broken arms may not be worthwhile. We introduce M dummy arms (of type 0), which are initially broken and have zero repair costs, and they always break down immediately after a repair (i.e., $\theta^1(x,0)=0$). Their Whittle indices are always zero. Choosing one of those arms is equivalent to retiring (i.e., collecting zero reward from) one of the original arms. The retirement option can be added to Problem 2 in this way. We compare the Whittle index policy with Policies 1 and 2 defined below.

(i) Policy 1 chooses up to M available arms from those with the largest Gittins indices. Every broken arm is retired permanently.

(ii) Policy 2 chooses M arms with the largest Gittins indices regardless of availability.

The Gittins index is calculated regardless of the value of the breakdown probability or repair cost. Policy 1 is pessimistic while Policy 2 is optimistic about repairing arms.

Table 2 compares the performances of the Whittle index policy and Policies 1 and 2 when $M=1$, the parameters of the beta prior distribution of each arm is $(a,b)=(1,1)$, each arm is initially available (i.e., $Y(0)=1$), and the probability that an arm is available does not depend on the state of X or whether or not it is active; namely, $\theta_n^1(x,1)=\theta_n^0(x,1)=\theta_n(1)$, $\theta_n^1(x,0)=\theta_n(0)$, $\theta_n^0(x,0)=0$ for every $x\in\mathcal{X}$ and $n=1,\ldots,N$ for some $\theta_n=(\theta_n(0),\theta_n(1))$.

We expect Policies 1 and 2 to work well because they are optimal if arms never break down. However, they behave oppositely when all the arms are unavailable; Policy 1 does well when Policy 2 does not, and vice versa. As observed from Table 2, the Whittle index policy handles the trade-off between repairing and retiring arms effectively.

8. Conclusion

We have studied an important extension of classical multi-armed bandit problem, in which arms may become intermittently unavailable, or they may break down, and repair is an option at some cost. Bandit problems with switching costs can be handled with this extension.

We showed that multi-armed bandit problems with availability constraints considered here do not admit an index policy which is optimal in every instance of the problem. However, the Whittle index policies we derived for each problem cannot be outperformed uniformly by any other index policy and are optimal for the classical bandit and the W-subsidy problems. Moreover, their index policies converge to the

Gittins index as the probability of availability approaches to one and to the immediate reward as it approaches to zero.

The Whittle indices can be computed by the value-iteration algorithm applied to a suitable restart-in problem reformulation. Finally, the numerical results suggest that the Whittle index policies perform well in general.

Acknowledgements

We acknowledge support from the U.S. Department of Homeland Security through the Center for Dynamic Data Analysis for Homeland Security administered through ONR grant number N00014-07-1-0150 to Rutgers University. We thank Dirk Bergemann, Faruk Gul, Ricardo Reis, Yosuke Yasuda, and an anonymous referee for helpful suggestions and remarks.

Appendix A. Proofs

A.1. Proof of Proposition 1

Consider the W-subsidy problem with fixed $W \in \mathbb{R}$. We show that if the process (X, Y) enters a state $(x, 1) \in \mathcal{S}$ where the passive action is optimal, then the passive action remains optimal at every future stage.

Suppose that $(X(0), Y(0)) = (x, 1)$ for some $(x, 1) \in \Pi(W)$, and the passive action is initially optimal. Then the state of X does not change, and the passive action remains optimal as long as the arm is available. However, if the arm becomes unavailable one day (i.e., (X, Y) enters $(x, 0)$), then the passive action is still optimal by (11). When the arm becomes available again, the next state will become $(x, 1)$, and the passive action remains optimal. Consequently, once the process (X, Y) enters the state $(x, 1) \in \Pi(W)$, the arm must be rested forever.

Thus, the W-subsidy problem reduces to an optimal stopping problem, where the optimal time to switch, while the arm is available, to the passive action has to be found. For every $(\mathcal{F}_t)_{t\geq 0}$-stopping time τ such that $Y(\tau) = 1$ $\mathbb{P}^{1,0}$-a.s., we consider the following strategy: activate the arm whenever it is available (and leave it rested otherwise) until time $\tau - 1$, and never activate it again at and after time τ. The corresponding expected total discounted reward is

$$\mathbb{E}_{x,1}^{1,0}\left[\sum_{t=0}^{\tau-1}\gamma^t\left(R^{Y(t)}(X(t),Y(t)) + W1_{\{Y(t)=0\}}\right) + \sum_{t=\tau}^{\infty}\gamma^t W + \gamma^\tau \rho(X(\tau),Y(\tau))\right].$$

On the other hand, immediate stopping in state $(x, 1)$ gives $\sum_{t=0}^{\infty}\gamma^t W + \rho(x, 1) = W/(1 - \gamma) + \rho(x, 1)$. Therefore, $(x, 1) \in \Pi(W)$ if and only if for every $\tau \in \overline{\mathscr{S}}$ the

quantity $W/(1-\gamma)+\rho(x,1)$ is greater or equal to the displayed expectation, which can be rewritten as

$$\mathbb{E}^{1,0}_{x,1}\left[\sum_{t=0}^{\tau-1}\gamma^t R^{Y(t)}(X(t),Y(t))+\gamma^\tau\rho(X(\tau),Y(\tau))\right]+W\mathbb{E}^{1,0}_{x,1}\left[\sum_{t=1}^{\tau-1}\gamma^t 1_{\{Y(t)=0\}}+\frac{\gamma^\tau}{1-\gamma}\right],$$

and some algebra gives

$$W\geq(1-\gamma)\frac{\mathbb{E}^{1,0}_{x,1}\left[\sum_{t=0}^{\tau-1}\gamma^t R^{Y(t)}(X(t),Y(t))+\gamma^\tau\rho(X(\tau),Y(\tau))\right]-\rho(x,1)}{1-\mathbb{E}^{1,0}_{x,1}\left[(1-\gamma)\sum_{t=1}^{\tau-1}\gamma^t 1_{\{Y(t)=0\}}+\gamma^\tau\right]},\quad\forall\tau\in\overline{\mathscr{S}}.$$

Thus, $(x,1)\in\Pi(W)$ if and only if (13) holds.

If the arm is unavailable, then the Whittle index follows from its definition and (11). Suppose now that the arm is available. By the first part, $(x,1)\in\Pi(W)$ if and only if (13) is satisfied. Then $\{(x,1)\in\mathcal{S}:(x,1)\in\Pi(W_1)\}\supseteq\{(x,1)\in\mathcal{S}:(x,1)\in\Pi(W_2)\}$ if $W_1>W_2$, and (11) implies that $\Pi(W_1)\supseteq\Pi(W_2)$ whenever $W_1>W_2$. Therefore, the arm is indexable, and $W(x,1)\equiv\inf\{W:(x,1)\in\Pi(W)\}$ in (13) gives the Whittle index.

A.2. Proof of Proposition 2

In order to emphasize the dependence of W and $\mathbb{P}^{1,0}$ on $\theta\in[0,1]$, we replace them with W_θ and $\mathbb{P}^\theta$, respectively. Then $W_\theta(x,1)=(1-\gamma)\sup_{\tau\in\overline{\mathscr{S}}}\Gamma(\theta,\tau,x)$, where

$$\Gamma(\theta,\tau,x):=\frac{\mathbb{E}^\theta_{x,1}\left[\sum_{t=0}^{\tau-1}\gamma^t R^1(X(t),1)1_{\{Y(t)=1\}}\right]}{[1-\gamma(1-\theta)]\mathbb{E}^\theta_{x,1}\left[1-\gamma^\tau\right]},\quad\forall\theta\in[0,1],\ \tau\in\overline{\mathscr{S}},\ x\in\mathcal{X}.$$

The Gittins index corresponds to $M(x)=W_1(x,1)\equiv(1-\gamma)\sup_{\tau\in\overline{\mathscr{S}}}\Gamma(1,\tau,x)$. Let $\bar{R}$ be a finite constant such that $|R(x,1)|<\bar{R}$ for every $x\in\mathcal{X}$.

We first prove the convergence to the immediate reward as $\theta\searrow 0$. Because immediate stopping gives $R^1(x,1)$, we have $W_\theta(x,1)\geq R(x,1)$. Note $W_\theta(x,1)$ is less than or equal to

$$\frac{(1-\gamma)R^1(x,1)+(1-\gamma)\bar{R}\,\mathbb{E}^\theta_{x,1}\left[\sum_{t=1}^{\infty}\gamma^t 1_{\{Y(t)=1\}}\right]}{[1-\gamma(1-\theta)]\mathbb{E}^\theta_{x,1}(1-\gamma^\tau)}=\frac{(1-\gamma)R^1(x,1)+\gamma\bar{R}\theta}{[1-\gamma(1-\theta)]\mathbb{E}^\theta_{x,1}(1-\gamma^\tau)}.$$

Suppose first that $R^1(x,1)\geq 0$. Let $K=\min\{t\geq 1;\ Y(t)=1\}$ be the first time in the future that the arm is available. Then $\mathbb{E}^\theta_{x,1}\gamma^K=\sum_{t=1}^{\infty}\gamma^t(1-\theta)^{t-1}\theta=\gamma\theta/[1-\gamma(1-\theta)]$. Because $\tau\geq K$ a.s. for every $\tau\in\overline{\mathscr{S}}$, we obtain

$$0\leq W_\theta(x,1)-R^1(x,1)\leq\frac{(1-\gamma)R^1(x,1)+\gamma\bar{R}\theta}{[1-\gamma(1-\theta)]\mathbb{E}^\theta_{x,1}(1-\gamma^K)}-R^1(x,1)=\frac{\gamma\bar{R}\theta}{1-\gamma}.$$

Suppose now that $R^1(x,1) < 0$. For every $0 \leq \theta < (1-\gamma)/\gamma$, we have $(1-\gamma)R^1(x,1) + \gamma\bar{R}\theta \leq -(1-\gamma)\bar{R} + \gamma\bar{R}\theta < 0$, and

$$0 \leq W_\theta(x,1) - R^1(x,1) \leq \frac{(1-\gamma)R^1(x,1) + \gamma\bar{R}\theta}{[1-\gamma(1-\theta)]\sup_{\tau\in\overline{\mathscr{T}}}\mathbb{E}^\theta_{x,1}(1-\gamma^\tau)} - R^1(x,1) = \frac{\gamma\bar{R}\theta}{1-\gamma(1-\theta)}.$$

Thus, for $0 \leq \theta < (1-\gamma)/\gamma$ we have $\sup_{x\in\mathcal{X}} |W_\theta(x,1) - R^1(x,1)| \leq \frac{\gamma\bar{R}\theta}{1-\gamma(1-\theta)} \xrightarrow{\theta\downarrow 0} 0$.

To show the convergence of $W_\theta(x,1)$ to the Gittins index as $\theta \nearrow 1$ uniformly in $x \in \mathcal{X}$, we will need the following lemma.

Lemma 1. *For some function $B : (0,1) \mapsto \mathbb{R}$ we have $\sup_{\tau\in\overline{\mathscr{T}},\, x\in\mathcal{X}} |\Gamma(1,\tau,x) - \Gamma(\theta,\tau,x)| \leq B(\theta)$ for every $\theta \in (0,1)$, and that $B(\theta) \to 0$ as $\theta \to 1$.*

Proof. For $\theta \in (0,1)$, $\tau \in \overline{\mathscr{T}}$, $x \in \mathcal{X}$, we have $|\Gamma(1,\tau,x) - \Gamma(\theta,\tau,x)|$ less than or equal to

$$\frac{\left|\mathbb{E}^1_{x,1}\left[\sum_{t=0}^{\tau-1}\gamma^t R^1(X(t),1)\right] - \mathbb{E}^\theta_{x,1}\left[\sum_{t=0}^{\tau-1}\gamma^t R^1(X(t),1)1_{\{Y(t)=1\}}\right]\right|}{1-\gamma} + \frac{\bar{R}}{1-\gamma}\left|\frac{1}{1-\mathbb{E}^1_{x,1}[\gamma^\tau]} - \frac{1}{[1-\gamma(1-\theta)]\mathbb{E}^\theta_{x,1}[1-\gamma^\tau]}\right|.$$

It is now sufficient to prove the existence of functions $B_1(\theta) \xrightarrow{\theta\uparrow 1} 0$ and $B_2(\theta) \xrightarrow{\theta\uparrow 1} 0$ such that

$$\left|\mathbb{E}^1_{x,1}\left[\sum_{t=0}^{\tau-1}\gamma^t R^1(X(t),1)\right] - \mathbb{E}^\theta_{x,1}\left[\sum_{t=0}^{\tau-1}\gamma^t R^1(X(t),1)1_{\{Y(t)=1\}}\right]\right| \leq B_1(\theta), \tag{26}$$

$$\left|\frac{1}{1-\mathbb{E}^1_{x,1}[\gamma^\tau]} - \frac{1}{[1-\gamma(1-\theta)]\mathbb{E}^\theta_{x,1}[1-\gamma^\tau]}\right| \leq B_2(\theta). \tag{27}$$

Let L be the first time the arm is unavailable. For every $l \geq 1$, the joint conditional $\mathbb{P}^\theta$-distribution of $\{(X(t),Y(t)); 0 \leq t \leq l-1\}$ given $L = l$ is the same as the joint unconditional $\mathbb{P}^1$-distribution of $\{(X(t),Y(t)); 0 \leq t \leq l-1\}$, and we have $\mathbb{P}^\theta\{L = l\} = \theta^{l-1}(1-\theta)$, $l \geq 1$ and $\mathbb{E}^\theta_{x,1}[\gamma^L] = \gamma(1-\theta)/(1-\gamma\theta) < (1-\theta)\gamma/(1-\gamma)$. The inequality in (26) holds with $B_1(\theta) := (1-\theta)2\gamma\bar{R}/(1-\gamma)^2$ because the lefthand side is less than or equal to

$$\sum_{l=1}^{\infty}\mathbb{P}^\theta_{x,1}\{L=l\}\left|\mathbb{E}^1_{x,1}\left[\sum_{t=0}^{\tau-1}\gamma^t R^1(X(t),1)\right] - \mathbb{E}^\theta_{x,1}\left[\sum_{t=0}^{\tau-1}\gamma^t R^1(X(t),1)1_{\{Y(t)=1\}}\middle| L=l\right]\right|,$$

where for every $l \geq 1$, the absolute difference is less than or equal to

$$\Big|\mathbb{E}^1_{x,1}\Big[\sum_{t=0}^{l-1}\gamma^t R^1(X(t),1)1_{\{\tau>l-1\}}\Big]-\mathbb{E}^\theta_{x,1}\Big[\sum_{t=0}^{l-1}\gamma^t R^1(X(t),1)1_{\{Y(t)=1\}}1_{\{\tau>l-1\}}\Big|L=l\Big]\Big|$$
$$+\Big|\mathbb{E}^1_{x,1}\Big[\sum_{t=l}^{\tau-1}\gamma^t R^1(X(t),1)1_{\{\tau>l-1\}}\Big]-\mathbb{E}^\theta_{x,1}\Big[\sum_{t=l}^{\tau-1}\gamma^t R^1(X(t),1)1_{\{Y(t)=1\}}1_{\{\tau>l-1\}}\Big|L=l\Big]\Big|$$
$$=\Big|\mathbb{E}^1_{x,1}\Big[\sum_{t=l}^{\tau-1}\gamma^t R^1(X(t),1)1_{\{\tau>l-1\}}\Big]-\mathbb{E}^\theta_{x,1}\Big[\sum_{t=l}^{\tau-1}\gamma^t R^1(X(t),1)1_{\{Y(t)=1\}}1_{\{\tau>l-1\}}\Big|L=l\Big]\Big|,$$

which is less than $2\bar{R}\sum_{t=l}^\infty\gamma^t$; therefore, the left-hand side of (26) is less than or equal to $\sum_{l=1}^\infty \mathbb{P}^\theta_{x,1}\{L=l\}\sum_{t=l}^\infty\gamma^t 2\bar{R}=\frac{2\bar{R}}{1-\gamma}\mathbb{E}^\theta_{x,1}\left[\gamma^L\right]=(1-\theta)\frac{2\gamma\bar{R}}{(1-\gamma)^2}\equiv B_1(\theta)$. The inequality in (27) holds with $B_2(\theta):=\gamma(1-\gamma)^2(1-\theta)/[1-\gamma(1-\theta)]\xrightarrow{\theta\uparrow 1}0$ because

$$\left|\frac{1}{1-\mathbb{E}^1_{x,1}\left[\gamma^\tau\right]}-\frac{1}{[1-\gamma(1-\theta)]\mathbb{E}^\theta_{x,1}\left[1-\gamma^\tau\right]}\right|$$
$$=\frac{\left|[1-\gamma(1-\theta)]\mathbb{E}^\theta_{x,1}\left[1-\gamma^\tau\right]-\left(1-\mathbb{E}^1_{x,1}\left[\gamma^\tau\right]\right)\right|}{\left(1-\mathbb{E}^1_{x,1}\left[\gamma^\tau\right]\right)[1-\gamma(1-\theta)]\mathbb{E}^\theta_{x,1}\left[1-\gamma^\tau\right]}$$
$$\leq\frac{\left|\mathbb{E}^1_{x,1}\left[\gamma^\tau\right]-\mathbb{E}^\theta_{x,1}\left[\gamma^\tau\right]\right|+\gamma(1-\theta)\mathbb{E}^\theta_{x,1}\left[1-\gamma^\tau\right]}{\left(1-\mathbb{E}^1_{x,1}\left[\gamma^\tau\right]\right)[1-\gamma(1-\theta)]\mathbb{E}^\theta_{x,1}\left[1-\gamma^\tau\right]}\leq\frac{\left|\mathbb{E}^1_{x,1}\left[\gamma^\tau\right]-\mathbb{E}^\theta_{x,1}\left[\gamma^\tau\right]\right|+\gamma(1-\theta)}{[1-\gamma(1-\theta)](1-\gamma)^2},$$

and $\left|\mathbb{E}^1_{x,1}\left[\gamma^\tau\right]-\mathbb{E}^\theta_{x,1}\left[\gamma^\tau\right]\right|\leq\sum_{l=1}^\infty\mathbb{P}^\theta_{x,1}\{L=l\}\left|\mathbb{E}^1_{x,1}\left[\gamma^\tau\right]-\mathbb{E}^\theta_{x,1}\left[\gamma^\tau|L=l\right]\right|$ equals

$$\sum_{l=1}^\infty\mathbb{P}^\theta_{x,1}\{L=l\}\left|\mathbb{E}^1_{x,1}\left[\gamma^\tau 1_{\{\tau>l-1\}}\right]-\mathbb{E}^\theta_{x,1}\left[\gamma^\tau 1_{\{\tau>l-1\}}|L=l\right]\right|$$
$$\leq\sum_{l=1}^\infty\mathbb{P}^\theta_{x,1}\{L=l\}\sum_{t=l}^\infty\gamma^t=\frac{1}{1-\gamma}\mathbb{E}^\theta_{x,1}\left[\gamma^L\right]<(1-\theta)\frac{\gamma}{(1-\gamma)^2}.$$

Finally, Lemma 1 follows with $B(\theta):=B_1(\theta)/(1-\gamma)+B_2(\theta)\bar{R}/(1-\gamma)$.

Now thanks to Lemma 1, we have $\sup_{x\in\mathcal{X}}|W_\theta(x,1)-M(x)|$ equals

$$\sup_{x\in\mathcal{X}}\left|\sup_{\tau\in\overline{\mathscr{T}}}\Gamma(\theta,\tau,x)-\sup_{\tau\in\overline{\mathscr{T}}}\Gamma(1,\tau,x)\right|\leq\sup_{\tau\in\overline{\mathscr{T}},\,x\in\mathcal{X}}|\Gamma(\theta,\tau,x)-\Gamma(1,\tau,x)|$$
$$\leq B(\theta)\xrightarrow{\theta\downarrow 0}0.$$

A.3. Proof of Proposition 3

Suppose that there are two arms. Arm 1 follows a stochastic process $(X(t),Y(t))_{t\geq 0}$ as in Section 2, and arm 2 is always available and gives some constant reward a. Let

$(x_1,1)$ and $(x_2,1)$ be the current states of arms 1 and 2, respectively. Then $W(x_2,1) =$

$$(1-\gamma)\sup_{\tau\in\overline{\mathscr{T}}}\frac{\mathbb{E}^{1,0}_{x_2,1}\left[\sum_{t=0}^{\tau-1}\gamma^t R^1(X(t),1)1_{\{Y(t)=1\}}\right]}{1-\mathbb{E}^{1,0}_{x_2,1}\left[\sum_{t=1}^{\tau-1}(1-\gamma)\gamma^t 1_{\{Y(t)=0\}}+\gamma^\tau\right]}=\sup_{\tau\in\mathscr{T}}\frac{\mathbb{E}^{1,0}_{x_2,1}\left[\sum_{t=0}^{\tau-1}\gamma^t a\right]}{\mathbb{E}^{1,0}_{x_2,1}\left[\sum_{t=0}^{\tau-1}\gamma^t\right]}=a.$$

By Proposition 1 resting arm 1 is optimal if and only if $W(x_1,1)\le a\equiv W(x_2,1)$. If there is an index policy which is optimal for every instance of Problem 1, then it must also be optimal for the above problem; therefore, an optimal index policy's index function must be a strictly monotone transformation of the Whittle index $W(\cdot,\cdot)$.

A.4. Proof of Proposition 5

We prove the indexability and obtain Whittle index under Condition 1. We consider cases $W<R^1(x,0)$ and $W\ge R^1(x,0)$ separately after the following lemmas.

Lemma 2. *For every $x\in\mathcal{X}$ and $W\in\mathbb{R}$, if $(x,0)\in\Pi(W)$, then $V((x,0),W)=W/(1-\gamma)$, which is obtained by taking the passive action all the time.*

Proof. The states of X and Y do not change under passive action. If $(x,0)\in\Pi(W)$, then the passive action remains optimal forever. Consequently, the expected total discounted reward starting in $(x,0)$ becomes $V((x,0),W)=\sum_{t=0}^{\infty}\gamma^t W=W/(1-\gamma)$.

Lemma 3. *For every $x\in\mathcal{X}$ and $W\in\mathbb{R}$, if $(x,1)\in\Pi(W)$, then the stochastic process (X,Y) starting in $(x,1)$ visits only $(x,1)$ and/or $(x,0)$ under the optimal policy, and*

$$V((x,1),W)=\max\left\{\mathbb{E}^{0,1}_{x,1}\left[\sum_{t=0}^{\infty}\gamma^t\left\{W1_{\{Y(t)=1\}}+R^1(x,0)1_{\{Y(t)=0\}}\right\}\right],\frac{W}{1-\gamma}\right\},\quad(28)$$

where $\mathbb{P}^{0,1}$ is the probability law induced by the policy that activates the arm as long as it is unavailable and leaves it rested otherwise.

Proof. The state of process X does not change under the passive action or if the arm is unavailable. Therefore, at the next time the arm becomes available, the process X is in state x. Because we assume that $(x,1)\in\Pi(W)$, the passive action must be taken whenever the arm is available. Therefore, the process (X,Y) can visit only the states $(x,0)$ and $(x,1)$. The optimal policy must be one of the following two: either that the arm always remains passive or that the arm remains passive in state $(x,1)$, but is activated in state $(x,0)$. Their values are $W/(1-\gamma)$ by Lemma 2 and the expectation in (28), respectively.

Lemma 4. *If $W<R^1(x,0)$, and Condition 2 holds, then $(x,y)\notin\Pi(W)$ for every $y\in\{0,1\}$; therefore, $\Pi(W)=\emptyset$.*

Proof. Suppose that $(x,0) \in \Pi(W)$ for some $x \in \mathcal{X}$. However, a lower bound on $V((x,0),W)$ can be obtained by considering the policy under which the arm is active at $(x,0)$ and passive otherwise. Because $W < R^1(x,0)$, this policy gives $V((x,0),W) > W/(1-\gamma)$, which contradicts with $(x,0) \in \Pi(W)$ by Lemma 2 and implies

$$(x,0) \notin \Pi(W), \quad x \in \mathcal{X}. \tag{29}$$

Suppose now $(x,1) \in \Pi(W)$ for some $x \in \mathcal{X}$. Then, by Lemma 3 and (29), we obtain $V((x,1),W) = \mathbb{E}^{0,1}_{x,1}\left[\sum_{t=0}^{\infty}\gamma^t\left\{W1_{\{Y(t)=1\}} + R^1(x,0)1_{\{Y(t)=0\}}\right\}\right] < R^1(x,0)/(1-\gamma)$. However, this contradicts with the lower bound obtained by applying the policy under which the arm is always active; namely, $V((x,1),W) = \mathbb{E}^{1,1}_{x,1}\left[\sum_{t=0}^{\infty}\gamma^t R^1(X(t),Y(t))\right] \geq R^1(x,0)/(1-\gamma)$, where the last inequality holds under Condition 2. Therefore, $(x,1) \notin \Pi(W), x \in \mathcal{X}$.

Lemma 5. *Suppose that $W \geq R^1(x,0)$. Then $(x,y) \in \Pi(W)$ if and only if $W \geq (1-\gamma)\,\mathbb{E}^{1,1}_{x,y}\left[\sum_{t=0}^{\tau-1}\gamma^t R^1(X(t),Y(t))\right]/(1-\mathbb{E}^{1,1}_{x,y}[\gamma^\tau])$ for every $\tau \in \mathscr{S}$.*

Proof. We show that in the W-subsidy problem, once the passive action is optimal, it remains optimal thereafter. This follows for $y = 0$ from Lemma 2.

Suppose $(x,1) \in \Pi(W)$ for some $x \in \mathcal{X}$. By Lemma 3, starting at $(x,1)$ under the optimal policy, only $(x,0)$ and $(x,1)$ will be visited by (X,Y), and (28) holds. Since $W \geq R^1(x,0)$, in (28) we have $\mathbb{E}^{0,1}_{x,1}\left[\sum_{t=0}^{\infty}\gamma^t\left\{W1_{\{Y(t)=1\}} + R^1(x,0)1_{\{Y(t)=0\}}\right\}\right] \leq W/(1-\gamma)$; thus, $V((x,1),W) = W/(1-\gamma)$, and passive action is always optimal.

As in the proof of Proposition 1, the W-subsidy problem now reduces to an optimal stopping problem. The optimal strategy must choose the active action until some stopping time τ and the passive action at and after time τ. Differently from Problem 1, it may now be optimal to stop when the arm is unavailable. If we switch from the active action to the passive action at some stopping time τ, then the expected total discounted reward will be $\mathbb{E}^{1,1}_{x,y}\left[\sum_{t=0}^{\tau-1}R^1(X(t),Y(t)) + \sum_{t=\tau}^{\infty}\gamma^t W\right]$. As in the proof of Proposition 1, $(x,y) \in \Pi(W)$ if and only if immediate stopping achieves greater value than the previous expectation for every positive $\tau \in \mathscr{S}$. Because immediate-stopping yields $W/(1-\gamma)$, we have $(x,y) \in \Pi(W)$ if and only if $W/(1-\gamma) \geq \mathbb{E}^{1,1}_{x,y}\left[\sum_{t=0}^{\tau-1}R^1(X(t),Y(t)) + \sum_{t=\tau}^{\infty}\gamma^t W\right]$ for every $\tau \in \mathscr{S}$.

The inequality in (15) follows from Lemmas 4 and 5 and implies the monotonicity of $W \mapsto \Pi(W)$, the indexability of the arm, and the form of the Whittle index $W(x,y)$.

A.5. Proof of Proposition 8

Fix any state $\widetilde{x} \in \mathcal{X}$. The Whittle index in state $(\widetilde{x},0) \in \mathcal{S}$ equals $W(\widetilde{x},0) = -\infty$, and we want to calculate $W(\widetilde{x},1)$ given by (13) for state $(\widetilde{x},1) \in \mathcal{S}$. From the proof

of Proposition 1 the value function $V((x,1),W)$ of the W-subsidy problem equals $\sup_{\tau\in\overline{\mathscr{F}}}\mathbb{E}_{x,1}^{1,0}\left[\sum_{t=0}^{\tau-1}\gamma^t\left(R^{Y(t)}(X(t),Y(t))+W1_{\{Y(t)=0\}}\right)+\gamma^\tau\left(\frac{W}{1-\gamma}+\rho(X(\tau),1)\right)\right]$ for every $x\in\mathcal{X}$ and $W\in\mathbb{R}$, which satisfies the optimality equation

$$V((x,1),W)=\max\left\{R^1(x,1)+\gamma\sum_{x'\in\mathcal{X}}p_{xx'}\left[(1-\theta^1(x,1))V((x',0),W)\right.\right. \tag{30}$$
$$\left.\left.+\theta^1(x,1)V((x',1),W)\right],\frac{W}{1-\gamma}+\rho(x,1)\right\},$$
$$V((x,0),W)=W+R^0(x,0)+\gamma\left[(1-\theta^0(x,0))V((x,0),W)+\theta^0(x,0)V((x,1),W)\right]. \tag{31}$$

By Definition 2, Whittle index $W(\widetilde{x},1)$ of an arm in state $(\widetilde{x},1)$ is the smallest $W\in\mathbb{R}$ in (30) for which one is indifferent in $(x,1)=(\widetilde{x},1)$ between stopping and continuation:

$$V((\widetilde{x},1),W(\widetilde{x},1))=\frac{W(\widetilde{x},1)}{1-\gamma}+\rho(\widetilde{x},1)=R^1(\widetilde{x},1)$$
$$+\gamma\sum_{x'\in\mathcal{X}}p_{\widetilde{x}x'}\left[(1-\theta^1(\widetilde{x},1))V((x',0),W(\widetilde{x},1))+\theta^1(\widetilde{x},1)V((x',1),W(\widetilde{x},1))\right]. \tag{32}$$

In (30) and (31), let us subtract $\rho(x,1)$ from both sides, and in the righthand side of both equations add and subtract $\rho(\cdot,1)$ to and from the functions $V((\cdot,0),W)$ and $V((\cdot,1),W)$. Rearranging the terms gives $V((x,1),W)-\rho(x,1)=$

$$\max\left\{R^1(x,1)+\gamma\sum_{x'\in\mathcal{X}}p_{xx'}\rho(x',1)+\gamma\sum_{x'\in\mathcal{X}}p_{xx'}\cdot\right.$$
$$\left.\left[(1-\theta^1(x,1))(V((x',0),W)-\rho(x',1))+\theta^1(x,1)(V((x',1)-\rho(x',1),W)\right],\frac{W}{1-\gamma}\right\},$$
$$V((x,0),W)-\rho(x,1)=W+R^0(x,0)-(1-\gamma)\rho(x,1)$$
$$+\gamma\left[(1-\theta^0(x,0))(V((x,0),W)-\rho(x,1))+\theta^0(x,0)(V((x,1),W)-\rho(x,1))\right].$$

In the last displayed equations, set $W=W(\widetilde{x},1)$, substitute

$$\frac{W(\widetilde{x},1)}{1-\gamma}=V(\widetilde{x},1),W(\widetilde{x},1))-\rho(\widetilde{x},1)=R^1(\widetilde{x},1)$$
$$+\gamma\sum_{x'\in\mathcal{X}}p_{\widetilde{x}x'}\rho(x',1)+\gamma\sum_{x'\in\mathcal{X}}p_{\widetilde{x}x'}\cdot\left[(1-\theta^1(\widetilde{x},1))(V((x',0),W(\widetilde{x},1))-\rho(x',1))\right.$$
$$\left.+\theta^1(\widetilde{x},1)(V((x',1)-\rho(x',1),W(\widetilde{x},1))\right] \tag{33}$$

from (32), and to get (19) and (20) rewrite the resulting equations in terms of $\nu_{x,y}^{(\widetilde{x},1)}:=V((x,y),W(\widetilde{x},1))-\rho(x,1)$, $(x,y)\in\mathcal{S}$. As shown before Proposition 8, (19) and (20) have unique solution, and $(1-\gamma)\nu_{\widetilde{x},1}^{(\widetilde{x},1)}=(1-\alpha)\left[V((\widetilde{x},1),W(\widetilde{x},1))-\rho(\widetilde{x},1)\right]=(1-\gamma)[W(\widetilde{x},1)/(1-\gamma)]=W(\widetilde{x},1)$ by (33).

References

[1] Banks, J. S. and Sundaram, R. K. (1992). Denumerable-armed bandits. *Econometrica* **60,** 1071–1096.

[2] Banks, J. S. and Sundaram, R. K. (1994). Switching costs and the Gittins index. *Econometrica* **62,** 687–694.

[3] Bergemann, D. and Valimaki, J. (2006). Efficient dynamic auctions. *Cowles Found. Disc. Paper* **1584,**.

[4] Brezzi, M. and Lai, T. L. (2000). Incomplete learning endogenous data in dynamic allocation. *Econometrica* **68,** 1511–1516.

[5] Dayanik, S., Powell, W. and Yamazaki, K. (2007). Index policies for discounted bandit problems with availability constraints. *Technical report.* Princeton University.

[6] Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *J. Roy. Statist. Soc. Ser. B* **41,** 148–177. With discussion.

[7] Glazebrook, K. D. (1984). Scheduling stochastic jobs on a single machine subject to breakdowns. *Naval Res. Logist. Quart.* **31,** 251–264.

[8] Glazebrook, K. D. (1987). Evaluating the effects of machine breakdowns in stochastic scheduling problems. *Naval Res. Logist.* **34,** 319–335.

[9] Glazebrook, K. D., Ansell, P. S., Dunn, R. T. and Lumley, R. R. (2004). On the optimal allocation of service to impatient tasks. *J. Appl. Probab.* **41,** 51–72.

[10] Glazebrook, K. D. and Mitchell, H. M. (2002). An index policy for a stochastic scheduling model with improving/deteriorating jobs. *Naval Res. Logist.* **49,** 706–721.

[11] Glazebrook, K. D., Niño-Mora, J. and Ansell, P. S. (2002). Index policies for a class of discounted restless bandits. *Adv. in Appl. Probab.* **34,** 754–774.

[12] Glazebrook, K. D., Ruiz-Hernandez, D. and Kirkbride, C. (2006). Some indexable families of restless bandit problems. *Adv. in Appl. Probab.* **38,** 643–672.

[13] Jovanovic, B. (1979). Job matching and the theory of turnover. *The Journal of Political Economy* **87, Part 1.,** 972–990.

[14] Jun, T. (2004). A survey on the bandit problem with switching costs. *De Economist* **1524,** 513–541.

[15] Katehakis, M. N. and Derman, C. (1986). Computing optimal sequential allocation rules in clinical trials. In *Adaptive statistical procedures and related topics (Upton, N.Y., 1985).* vol. 8 of *IMS Lecture Notes Monogr. Ser.* Inst. Math. Statist., Hayward, CA pp. 29–39.

[16] Katehakis, M. N. and Veinott, Jr., A. F. (1987). The multi-armed bandit problem: decomposition and computation. *Math. Oper. Res.* **12,** 262–268.

[17] Miller, R. A. (1984). Job matching and occupational choice. *The Journal of Political Economy* **926,** 1086–1120.

[18] Niño-Mora, J. (2001). Restless bandits, partial conservation laws and indexability. *Adv. in Appl. Probab.* **33,** 76–98.

[19] Papadimitriou, C. H. and Tsitsiklis, J. N. (1999). The complexity of optimal queuing network control. *Math. Oper. Res.* **24,** 293–305.

[20] Ross, S. (1983). *Introduction to Stochastic Dynamic Programming.* Academic Press, INC.

[21] Rothschild, M. (1974). A two-armed bandit theory of market pricing. *J. Econom. Theory* **9,** 185–202.

[22] Tsitsiklis, J. N. (1994). A short proof of the Gittins index theorem. *Ann. Appl. Probab.* **4,** 194–199.

[23] Weber, R. R. and Weiss, G. (1990). On an index policy for restless bandits. *J. Appl. Probab.* **27,** 637–648.

[24] Weber, R. R. and Weiss, G. (1991). Addendum to: "On an index policy for restless bandits". *Adv. in Appl. Probab.* **23,** 429–430.

[25] Whittle, P. (1980). Multi-armed bandits and the Gittins index. *J. Roy. Statist. Soc. Ser. B* **42,** 143–149.

[26] Whittle, P. (1988). Restless bandits: activity allocation in a changing world. *J. Appl. Probab.* 287–298. A celebration of applied probability.