Adv. Appl. Prob. **38,** 643–672 (2006)
Printed in Northern Ireland

SOME INDEXABLE FAMILIES OF RESTLESS BANDIT PROBLEMS

K. D. GLAZEBROOK,* *Lancaster University*
D. RUIZ-HERNANDEZ,** *Universitat Pompeu Fabra*
C. KIRKBRIDE,*** *Lancaster University*

Abstract

In 1988 Whittle introduced an important but intractable class of *restless bandit problems* which generalise the *multiarmed bandit problems* of Gittins by allowing state evolution for passive projects. Whittle's account deployed a Lagrangian relaxation of the optimisation problem to develop an index heuristic. Despite a developing body of evidence (both theoretical and empirical) which underscores the strong performance of Whittle's index policy, a continuing challenge to implementation is the need to establish that the competing projects all pass an indexability test. In this paper we employ Gittins' index theory to establish the indexability of (*inter alia*) general families of restless bandits which arise in problems of machine maintenance and stochastic scheduling problems with switching penalties. We also give formulae for the resulting Whittle indices. Numerical investigations testify to the outstandingly strong performance of the index heuristics concerned.

Keywords: Bandit problem; dynamic programming; Gittins index; machine maintenance; restless bandit; stochastic scheduling; switching cost

2000 Mathematics Subject Classification: Primary 90C40
Secondary 49L20; 90C39; 49M20

1. Introduction

The classical index result of Gittins (1979), (1989, pp. 1–77) concerns the optimal allocation over time of a single key resource among a collection of projects (or *bandits*) which are in competition for it. Application of the resource to a project at any stage earns a reward and effects a transition in the project's state. If a project is not in receipt of processing effort, it earns nothing and its state is unchanged. The optimisation goal is the identification of a policy for sequentially selecting projects on the basis of current state information to maximise the expected total discounted reward earned over an infinite horizon. Such problems are called *multiarmed bandit problems*. Gittins (1979), (1989, pp. 1–77) constructed a collection of *calibrating indices* (now usually called *Gittins indices*), one for each project, in the form of a real-valued function on the project's state space. He showed that the *index policy* which always directs the key resource to the project or projects with the highest current index is optimal.

The requirement in Gittins' models that passive projects should remain frozen inhibits applicability. Motivated by this, Whittle (1988) introduced a class of *restless bandit problems* which allow for state evolution among passive projects. This class of problems is now understood

Received 10 August 2004; revision received 30 March 2006.
* Postal address: Department of Mathematics and Statistics, Lancaster University, Lancaster, LA1 4YF, UK.
Email address: k.glazebrook@lancaster.ac.uk
** Postal address: Department of Economics and Business, Universitat Pompeu Fabra, E-08005 Barcelona, Spain.
*** Postal address: Department of Management Science, Lancaster University, Lancaster, LA1 4YX, UK.

to be intractable, and Papadimitriou and Tsitsiklis (1999) showed restless bandit problems to be PSPACE-hard. Whittle's original analysis elucidated (under stated conditions) an index-based solution to a Lagrangian relaxation of the restless bandit problem of interest in which the Lagrange multiplier has an economic interpretation as a *subsidy for passivity* or a *charge for service*. The indices identified by this route generalise those of Gittins, and Whittle proposed their use in the construction of heuristics for restless bandit problems. The importance of Whittle's contribution was underscored by Weber and Weiss (1990) who established a form of asymptotic optimality for the index heuristic and, more recently, in a range of empirical studies which have demonstrated its outstandingly strong performance in a range of application domains; see, for example, Ansell *et al.* (2003) and Glazebrook *et al.* (2005). Furthermore, Glazebrook *et al.* (2002) have discussed the development of bounds on the degree of reward suboptimality of Whittle's index policy.

A major challenge to the deployment of Whittle's powerful ideas is that his index function is only defined for those projects which pass a test of *indexability*. Although this requirement (given in Definition 1, below) seems plausible and natural, it can be very difficult to establish and, indeed, need not hold. Even when indexability is established there are few cases where the Whittle index (as we shall call it) is available in closed form. For an instance of the latter see the queueing control problem discussed by Ansell *et al.* (2003). Niño-Mora (2001), (2002) has identified sufficient conditions for project indexability by the use of polyhedral methods. The primary contribution of the current paper lies in the demonstration that general classes of restless bandit problems which arise in some major application domains are indeed indexable. Having established indexability, we proceed to identify the corresponding Whittle indices. Our tools for analysis are (generally) those of stochastic dynamic programming and (more particularly) those of Gittins' index theory. In every case, the Whittle index identified is given as a function of a corresponding Gittins index and/or in closed form. Given the ease with which Gittins indices may be computed, this is more than enough for implementation.

The paper is structured as follows. In Section 2 we present a class of restless bandit problems with a discounted reward criterion and present Whittle's notion of indexability along with a definition of his indices. In Section 3 we establish the indexability of a class of restless bandits designed to model machine maintenance problems in which maintenance interventions (the active action) have to be scheduled to mitigate escalating costs as machines deteriorate (when passive). Whittle (1996, p. 221, p. 280) and Glazebrook *et al.* (2005) have previously given index-based analyses of particular models, but we show here that indexability is guaranteed in general. Identification of the Whittle indices concerned is followed by a numerical investigation which demonstrates the very strong performance of Whittle's index heuristic in 8000 problems.

Section 4 concerns the development of policies for multiarmed bandits in which cost penalties are incurred whenever processing effort is switched between projects. This is a famously intractable problem to which many important contributions have been made; see, for example, Glazebrook (1980), Agrawal *et al.* (1988), Van Oyen and Teneketzis (1994), and Reiman and Wein (1998). Banks and Sundaram (1994) established a negative result concerning the (non)existence of optimal policies of index form for classes of such problems, even very simple ones. To the authors' knowledge, it was Niño-Mora (personal communication (2001)) who first suggested that multiarmed bandits with switching costs can be viewed as restless bandits. In Section 4 we establish indexability for general classes of such restless bandit problems, including some which provide for switching penalties dependent upon the states of *both* projects that are party to a switch. In the very simple case of constant set-up costs, the Whittle index policy developed coincides with a heuristic studied by Asawa and Teneketzis (1996), who

established some partial optimality results for their model. Numerical investigation of 11 000 problems testifies to the very strong performance of Whittle's index heuristic.

The paper concludes in Section 5 with an indexability analysis of a class of models in which the reward-generating capacity of a project may deplete with use but recovers (at a cost) when passive.

2. Discounted restless bandit problems and indexability

In this section we shall describe discounted restless bandit problems and a notion of indexability due to Whittle (1988). For restless bandit problems passing the indexability test, Whittle's index heuristic will then be described.

A discounted (reward-based) restless bandit problem is a *discounted Markov decision problem* $\{(\Omega_i, P_i^a, P_i^b, R_i^a, R_i^b, \beta),\ 1 \le i \le N, M\}$ specified as follows.

(i) Decision epochs occur at times $t \in \mathbb{N}$.

(ii) The (countable) state space is the Cartesian product $\times_{i=1}^{N} \Omega_i$, with Ω_i the state space for *bandit i* (sometimes *machine i* or *job i*), $1 \le i \le N$. The *state* of the process at time t is $\boldsymbol{X}(t) = \{X_1(t), X_2(t), \ldots, X_N(t)\}$, with $X_i(t) \in \Omega_i$ the state of bandit i at time t.

(iii) We suppose that $N > M$. At each decision epoch t, one of $\binom{N}{M}$ admissible actions is taken. A typical action is an N-vector $\boldsymbol{d} = \{d_1, d_2, \ldots, d_N\}$, where $d_i \in \{a, b\}$, $1 \le i \le N$, and

$$|\{i : d_i = a\}| = M. \tag{1}$$

Under action a bandit i is *active*, while under action b it is *passive*. We denote by A the admissible action space for the process. Equation (1) indicates that an admissible action for the process activates exactly M bandits while leaving the remaining N–M passive.

(iv) Suppose that action $\boldsymbol{d}(t)$ is taken at time $t \in \mathbb{N}$. If $d_i(t) = a$ then bandit i evolves at t according to Markov law P_i^a. We write

$$\begin{aligned} \mathrm{P}\{X_i(t+1) = y \mid X_i(t) = x,\ d_i(t) = a\} &\equiv \mathrm{P}\{X_i(t+1) = y \mid X_i(t) = x, a\} \\ &= P_i^a(x, y), \qquad x, y \in \Omega_i. \end{aligned} \tag{2}$$

If $d_i(t) = b$ then bandit i evolves at t according to Markov law P_i^b, and we write

$$\begin{aligned} \mathrm{P}\{X_i(t+1) = y \mid X_i(t) = x,\ d_i(t) = b\} &\equiv \mathrm{P}\{X_i(t+1) = y \mid X_i(t) = x, b\} \\ &= P_i^b(x, y), \qquad x, y \in \Omega_i. \end{aligned} \tag{3}$$

The N bandits evolve independently.

(v) For all i, $R_i^a : \Omega_i^2 \to \mathbb{R}^+$ and $R_i^b : \Omega_i^2 \to \mathbb{R}^+$ are bounded reward functions. If, as in (2), a transition from x to y occurs in bandit i under action a at time t, then a discounted reward $\beta^t R_i^a(x, y)$ is earned. If, as in (3), a transition from x to y occurs in bandit i under action b at time t, then a discounted reward $\beta^t R_i^b(x, y)$ is earned. Rewards are additive across bandits and over time. We shall frequently use the abbreviated notation

$$R_i^a(x) := \sum_{y \in \Omega_i} R_i^a(x, y) P_i^a(x, y), \qquad R_i^b(x) := \sum_{y \in \Omega_i} R_i^b(x, y) P_i^b(x, y),$$

$$x \in \Omega_i,\ 1 \le i \le N,$$

https://doi.org/10.1239/aap/1158684996 Published online by Cambridge University Press

to denote expected rewards earned from a single transition under actions a and b, respectively. Furthermore, $\beta \in (0, 1)$ is a discount rate.

(vi) A *policy* π is a rule for taking actions at each decision epoch. Such a rule can in principle be a function of the entire history of the process (actions taken and states occupied) to date. The goal of our analysis is the determination of a policy to maximise total expected reward over an infinite horizon. The theory of stochastic dynamic programming (see, for example, Puterman (1994)) asserts the existence of an optimal (reward-maximising) policy which is *stationary* (i.e. makes decisions in the light of the current state only) and satisfies the optimality equations of dynamic programming.

We write $V(\boldsymbol{x})$ for the *value function* of the process evaluated at $\boldsymbol{x} \in \times_{i=1}^{N} \Omega_i$, namely the maximal expected reward earned over an infinite horizon from initial state $\boldsymbol{x}$. The optimality equations may be expressed as

$$V(\boldsymbol{x}) = \max_{\boldsymbol{d} \in A} \Bigg\{ \sum_{i=1}^{N} \{1(d_i = a) R_i^a(x_i) + 1(d_i = b) R_i^b(x_i)\} + \beta \sum_{\boldsymbol{y}} \prod_{\{i\,:\,d_i = a\}} P_i^a(x_i, y_i) \prod_{\{j\,:\,d_j = b\}} P_j^b(x_j, y_j) V(\boldsymbol{y}) \Bigg\}, \qquad \boldsymbol{x} \in \bigtimes_{i=1}^{N} \Omega_i. \quad (4)$$

In (4), $1(\cdot)$ is an indicator function and the second summation on the right-hand side is over $\boldsymbol{y} \in \times_{i=1}^{N} \Omega_i$. Equation (4) notwithstanding, a pure dynamic programming approach is unlikely to yield insight and will be computationally intractable for problems of reasonable size. Hence, the primary quest is for good *heuristic policies*.

The subclass of models for which $M = 1$ and $P_i^b(x, y) = 0$, $x \neq y$, $1 \leq i \leq N$ (i.e. with only one bandit activated at each decision epoch and no state evolution under the passive action) are known as *multiarmed bandit problems*. Gittins (1979), (1989, pp. 1–77) famously demonstrated the optimality of *index policies* for multiarmed bandit problems, i.e. that there exist calibrating index functions $G_i : \Omega_i \to \mathbb{R}$ (one for each bandit) such that at time t the bandit for optimal activation is the one whose associated index $G_i(X_i(t))$ is maximal. In the event of ties, all corresponding choices are optimal. Building on this classical result, Whittle (1988) proposed a class of index heuristics for those restless bandit problems which pass an *indexability test*. These heuristics emerge naturally from a Lagrangian relaxation of the original optimisation problem. The interested reader is referred to Whittle (1988) for full details. The indices which result from Whittle's analysis generalise those of Gittins. We now outline the key notions in Whittle's approach.

Indexability and indices are properties of individual bandits. Hence, in the Markov decision problem (i)–(vi) above, we isolate an individual bandit, $(\Omega_i, P_i^a, P_i^b, R_i^a, R_i^b, \beta)$, and drop the bandit identifier i. This bandit generates a collection of Markov decision problems parametrised by a *passive subsidy*, $W \in \mathbb{R}$. We shall refer to the *subsidy-W problem* for bandit $(\Omega, P^a, P^b, R^a, R^b, \beta)$ to mean the reward-discounted Markov decision problem specified as follows.

1. Decision epochs occur at times $t \in \mathbb{N}$.
2. The (countable) state space is Ω. We use $X(t)$ for the state of the process at time t.
3. At each decision epoch t, either action a (active) or b (passive) is applied to the process.

https://doi.org/10.1239/aap/1158684996 Published online by Cambridge University Press

4. If a is chosen at t then evolution is according to P^a, with

$$\mathrm{P}\{X(t+1) = y \mid X(t) = x, a\} = P^a(x, y), \qquad x, y \in \Omega.$$

   If b is chosen at t then evolution is according to P^b, with

$$\mathrm{P}\{X(t+1) = y \mid X(t) = x, b\} = P^b(x, y), \qquad x, y \in \Omega.$$

5. If a transition from x to y occurs under action a at time t, then a discounted reward $\beta^t R^a(x, y)$ is earned. Should a transition from x to y occur under action b at time t, a discounted reward $\beta^t\{R^b(x, y) + W\}$, where W is the passive subsidy, is earned.

6. The goal of optimisation is to choose a policy to maximise the total expected reward (including passive subsidies) earned over an infinite horizon. We again assert the existence of optimal policies for the subsidy-W problem which are stationary and whose value functions satisfy the optimality equations of dynamic programming. We shall restrict to stationary policies throughout.

We denote by $V(x, W)$ the value function for the subsidy-W problem evaluated at $x \in \Omega$. The dynamic programming optimality equations may be expressed as

$$V(x, W) = \max\left\{ R^a(x) + \beta \sum_{y\in\Omega} P^a(x, y)V(y, W), \right.$$
$$\left. R^b(x) + W + \beta \sum_{y\in\Omega} P^b(x, y)V(y, W) \right\}, \qquad x \in \Omega. \tag{5}$$

The active action is optimal in x when the first term in the function max on the right-hand side of (5) is the maximum and the passive action is optimal when the second term is the maximum. We denote by $\Pi(W)$ the subset of Ω for which the passive action is optimal:

$$\Pi(W) = \left\{ x \in \Omega\colon R^b(x) + W + \beta \sum_{y\in\Omega} P^b(x, y)V(y, W) \geq R^a(x) + \beta \sum_{y\in\Omega} P^a(x, y)V(y, W) \right\}.$$

The following definition describes the indexability test for both bandit $(\Omega, P^a, P^b, R^a, R^b, \beta)$ and the restless bandit problem of which it is a part.

Definition 1. Bandit $(\Omega, P^a, P^b, R^a, R^b, \beta)$ is *indexable* if $\Pi(W)$ is increasing in W, i.e. if

$$W_1 \geq W_2 \quad \Longrightarrow \quad \Pi(W_1) \supseteq \Pi(W_2).$$

A restless bandit problem is *indexable* if each of its constituent bandits is indexable.

Hence, a restless bandit is indexable if, as the level of passive subsidy increases, so does the collection of states for which the passive action is optimal. However plausible and natural this requirement may appear, it is typically very challenging to establish and sometimes fails to hold.

Definition 2. If bandit $(\Omega, P^a, P^b, R^a, R^b, \beta)$ is indexable then its *Whittle index*, $W\colon \Omega \to \mathbb{R}$, is given by $W(x) = \inf\{W\colon x \in \Pi(W)\}$, $x \in \Omega$.

https://doi.org/10.1239/aap/1158684996 Published online by Cambridge University Press

Note that the assumed boundedness of rewards in (v) guarantees that the Whittle index must also be bounded. The value $W(x)$ represents a *fair subsidy* in state x in the sense that it renders both actions (active and passive) optimal in the subsidy-W problem.

Remark 1. A. We could equivalently define a charge-W problem in which part 5 of the subsidy-W problem above is replaced by the following.

5′. If a transition from x to y occurs under action a at time t, then a discounted reward $\beta^t\{R^a(x, y) - W\}$, where W is a charge for activity, is earned. Should a transition from x to y occur under b at time t, a discounted reward $\beta^t R^b(x, y)$ is earned.

Clearly the subsidy-W and charge-W problems are equivalent in the sense of having identical optimal policies. The value functions differ by $W(1-\beta)^{-1}$ in all states.

B. We can, of course, have cost-based restless bandit problems $\{(\Omega_i, P_i^a, P_i^b, C_i^a, C_i^b, \beta),\ 1 \le i \le N, M\}$. These are specified as in (i)–(vi) above, barring the fact that the bounded rewards R_i^a and R_i^b in (v) are replaced by the bounded costs C_i^a and C_i^b. The goal of the analysis is now the determination of policies to minimise the total expected cost incurred over an infinite horizon. The development of a subsidy-W problem follows steps 1–6 above, except now a passive subsidy W always reduces the instantaneous cost incurred under the passive action by that amount. The corresponding optimality equation replacing (5) is

$$V(x, W) = \min\Bigg\{C^a(x) + \beta \sum_{y\in\Omega} P^a(x, y)V(y, W),$$
$$C^b(x) - W + \beta \sum_{y\in\Omega} P^b(x, y)V(y, W)\Bigg\}, \qquad x \in \Omega.$$

Definitions 1 and 2 remain unchanged.

As in remark A above, we can equivalently define a charge-W problem, in the obvious way. Throughout the paper we shall use the convention that cost-based restless bandit problems will be analysed via the corresponding charge-W problems – thus yielding a wholly cost-based decision structure. Similarly, reward-based restless bandit problems (or those which are primarily so) will be analysed via the corresponding subsidy-W problems.

C. Restore the bandit identifiers and consider an indexable restless bandit problem with Whittle index $W_i : \Omega_i \to \mathbb{R}$ for bandit i, $1 \le i \le N$. The *Whittle index heuristic* operates as follows: at each time $t \in \mathbb{N}$, apply the active action to the M bandits with largest index $W_i(X_i(t))$ and apply the passive action to the remaining $N - M$ bandits.

3. Model 1: machine maintenance

In light of the development in Section 2, it will be enough to conduct our discussions of indexability and indices by reference to individual bandits as described in 1–6 above. Our goal here is to develop an indexability analysis of a general class of structured bandits designed to model machine maintenance problems. It will be convenient to discuss their indexability in terms of the charge-W problem (see Remark 1.B) associated with the bandit.

We use (Ω, P, C, D, β) to denote the cost-based bandit of interest. Here a bandit will represent a single machine evolving under the influence of maintenance interventions. The general structure specified in 1–6 above is specialised to this case as follows.

https://doi.org/10.1239/aap/1158684996 Published online by Cambridge University Press

I. Designated state $0' \in \Omega$ represents the (pristine) state of the machine following a maintenance intervention. In general, $X(t)$ represents the state of the machine at time $t \in \mathbb{N}$.

II. If action a (active, or intervention) is taken at time t, then we have

$$\mathrm{P}\{X(t+1) = y \mid X(t) = x, a\} = P(0', y), \qquad x, y \in \Omega, \tag{6}$$

where $P(\cdot, \cdot)$ is a Markov law. Equation (6) should be understood as follows: under action a the machine is instantaneously returned to state $0'$, after which it performs a single transition under P. If action b (passive, or nonintervention) is taken at time t, then we have

$$\mathrm{P}\{X(t+1) = y \mid X(t) = x, b\} = P(x, y), \qquad x, y \in \Omega. \tag{7}$$

Since action b is nonintervention, the Markov law P can be understood to model machine deterioration under the passive action.

III. The costs incurred under transitions (6) and (7) are respectively $D(x) + C(0', y)$ and $C(x, y)$. Here $D(x)$ represents the intervention cost of returning a state-x machine to state $0'$. The cost, $C(x, y)$, incurred when a transition from x to y occurs under P will incorporate ongoing maintenance costs and may include (typically very large) costs from an unexpected catastrophic breakdown of the machine. All costs are bounded. The corresponding costs for the charge-W problem are respectively $D(x) + C(0', y) + W$ and $C(x, y)$, where W is a charge for service.

As above, we denote by $V(x, W)$ the value function for the charge-W problem evaluated at $x \in \Omega$. From I–III, the optimality equations may be expressed as

$$V(x, W) = \min\Bigg\{ D(x) + C(0') + W + \beta \sum_{y \in \Omega} P(0', y) V(y, W), \\ C(x) + \beta \sum_{y \in \Omega} P(x, y) V(y, W) \Bigg\}, \qquad x \in \Omega,$$

where here and hereafter we use the abbreviated notation

$$C(x) := \sum_{y \in \Omega} C(x, y) P(x, y).$$

To begin our analysis of the charge-W problem, consider a set-up in which $X(0) = 0'$ and passive action b is taken at time 0 with an optimal (cost-minimising) policy pursued thereafter. Since we restrict to stationary policies for intervention, it follows that

$$\tau^* := \min\{t : t \geq 1 \text{ and it is optimal to choose action } a\}$$

is a stationary stopping time. Write $\Theta(0', W)$ for the corresponding expected discounted cost incurred over an infinite horizon for the charge-W problem. From I–III above we infer that

$$W + \Theta(0', W) = W + C(0', \tau^*) + \mathrm{E}[\beta^{\tau^*}[D(X(\tau^*)) + W + \Theta(0', W)] \mid 0'] \tag{8}$$

$$= W + \inf_{\tau} \{C(0', \tau) + \mathrm{E}[\beta^{\tau}[D(X(\tau)) + W + \Theta(0', W)] \mid 0']\}. \tag{9}$$

https://doi.org/10.1239/aap/1158684996 Published online by Cambridge University Press

In (8) and (9), and throughout, we use ' $\mid x$' as a notational shorthand for ' $\mid X(0) = x$' and

$$C(x, \tau) = \mathrm{E}\left[\sum_{t=0}^{\tau-1} \beta^t C(X(t)) \;\middle|\; x\right]$$

for the expected cost incurred under the passive action from initial state x up to some positive-valued stopping time τ. Note that the terms on the right-hand sides of (8) and (9) record both the expected costs incurred during the initial period of machine evolution under the passive action (up to τ^* or τ) and the costs to go from the intervention at τ^* (or τ) onwards. Note also that the infimum in (9) is over all stationary, positive, integer-valued stopping times on the machine state process evolving from $0'$ under the passive action. A stopping time is said to be *stationary* if it is the time of first entry of the process into some specified subset of Ω. Let us write

$$\Delta(0', W) := W + \Theta(0', W).$$

The following result records straightforward consequences of the foregoing discussion.

Theorem 1. *The quantity* $\Delta(0', W)$ *is given by*

$$\Delta(0', W) = (W + C(0', \tau^*) + \mathrm{E}[\beta^{\tau^*} D(X(\tau^*)) \mid 0'])\{1 - \mathrm{E}[\beta^{\tau^*} \mid 0']\}^{-1} \tag{10}$$

$$= \inf_{\tau} (W + C(0', \tau) + \mathrm{E}[\beta^{\tau} D(X(\tau)) \mid 0'])\{1 - \mathrm{E}[\beta^{\tau} \mid 0']\}^{-1}, \tag{11}$$

where the infimum is taken over all stationary, positive-valued stopping times on the machine state process evolving from $0'$ *under the passive action. Moreover,* $\Delta(0', W)$ *is continuous and strictly increasing in* W.

Proof. Equations (10) and (11) are immediate consequences of the discussion of (9). Suppose that $W_1 > W_2$ and that $\tau(W_1)$ achieves the infimum for $\Delta(0', W_1)$. Standard index theory guarantees the existence of such a stopping time. We then have

$$\begin{aligned}\Delta(0', W_1) &> (W_2 + C(0', \tau(W_1)) + \mathrm{E}[\beta^{\tau(W_1)} D(X(\tau(W_1))) \mid 0'])\{1 - \mathrm{E}[\beta^{\tau(W_1)}] \mid 0'\}^{-1} \\ &\geq \Delta(0', W_2)\end{aligned}$$

and conclude that $\Delta(0', W)$ is strictly increasing in W. Continuity of $\Delta(0', \cdot)$ is straightforward to demonstrate. This concludes the proof.

Before proceeding to the main indexability result, we develop a form of Gittins index appropriate for the analysis. In order to develop $G(x)$, the so-called *Gittins index for passivity* in state $x \in \Omega$, consider a set-up in which $X(0) = x$ and passive action b is taken at times $0, 1, 2, \ldots, \tau - 1$, where τ is a stopping time on the machine state process $\{X(t),\ t \geq 0\}$ and satisfies $\tau \geq 1$ almost surely.

Definition 3. The Gittins index for passivity, $G\colon \Omega \to \mathbb{R}$, is given by

$$G(x) := \inf_{\tau} (C(x, \tau) - D(x) + \mathrm{E}[\beta^{\tau} D(X(\tau)) \mid x])\{1 - \mathrm{E}[\beta^{\tau} \mid x]\}^{-1}, \qquad x \in \Omega, \tag{12}$$

where the infimum is taken over all stationary, positive-valued stopping times on the machine state process evolving from x under the passive action.

https://doi.org/10.1239/aap/1158684996 Published online by Cambridge University Press

Remark 2. It will assist the reader if we first characterise the bandit for which $G(x)$ in (12) is the Gittins index. It is one in which $X(0) = x$ with the machine state process $\{X(t),\ t \geq 0\}$ evolving under the passive action, as in (7). Furthermore, a transition from state y to state z incurs a cost of $C(y, z) - D(y) + \beta D(z)$. With these choices, the expected cost incurred by the machine during $[0, \tau)$ is given by

$$C(x, \tau) - D(x) + \mathrm{E}[\beta^{\tau} D(X(\tau)) \mid x].$$

We now appeal to Gittins' index theory to characterise the set of stopping times achieving the infimum in (12). They are developed as follows. Fix $W \in \mathbb{R}$ and denote by $\Gamma(W)$ the subset of Ω given by

$$\Gamma(W) = \{y \in \Omega\colon G(y) > W\},$$

and by $\Sigma(W)$ the subset of Ω given by

$$\Sigma(W) = \{y \in \Omega\colon G(y) = W\}.$$

Note that either or both of $\Gamma(W)$ and $\Sigma(W)$ may be empty. Now suppose that $X(0) = x$ and $\Sigma \subseteq \Sigma(W)$. We write τ^{Σ} for the stationary, positive-valued stopping time defined on the process $\{X(t),\ t \geq 0\}$ evolving under the passive action, given by

$$\tau^{\Sigma} = \min\{t\colon t > 0 \text{ and } X(t) \in \Gamma(W) \cup \Sigma\}.$$

Furthermore, we write $T(x, W)$ for the collection of stopping times given by

$$T(x, W) = \bigcup_{\Sigma \subseteq \Sigma(W)} \{\tau^{\Sigma}\}. \tag{13}$$

On the basis of Gittins' index theory we can assert that all stopping times in $T(x, G(x))$ achieve the infimum in (12). These are the only stationary stopping times which do so.

Theorem 2. (Indexability and indices.) (a) *Bandit* (Ω, P, C, D, β) *is indexable.*

(b) *The Whittle index in state* x, $W(x)$, *is the unique* W*-solution to the equation* $\Delta(0', W) = G(x)$.

(c) *The orderings of members of* Ω *determined by the Whittle index and the Gittins index for passivity coincide.*

Proof. Let $X(0) = x \in \Omega$. Action b is optimal for the charge-W problem at time $t = 0$ in state x if and only if there exists some stationary, positive stopping time τ on the machine state process evolving under the passive action such that any policy which

- chooses action b at times $t = 0, 1, 2, \ldots, \tau - 1$,
- chooses action a at time τ, and
- chooses optimally at all times $t \geq \tau + 1$

has total expected costs no greater than the best policy among those which choose action a at time $t = 0$. However, by I–III above and the discussion of (9), the minimum achievable cost when action a is taken at time $t = 0$ is

$$D(x) + W + \Theta(0', W) = D(x) + \Delta(0', W).$$

https://doi.org/10.1239/aap/1158684996 Published online by Cambridge University Press

Furthermore, the total expected cost incurred under any policy as described in the three points itemised above is given by

$$C(x,\tau)+\mathrm{E}[\beta^{\tau}[D(X(\tau))+W+\Theta(0',W)]\mid x]$$
$$=C(x,\tau)+\mathrm{E}[\beta^{\tau}D(X(\tau))\mid x]+\mathrm{E}[\beta^{\tau}\mid x]\Delta(0',W).$$

Hence, action b is optimal in state x for the charge-W problem if and only if there exists a stationary, positive-valued stopping time τ on the state process evolving from x under the passive action such that

$$C(x,\tau)+\mathrm{E}[\beta^{\tau}D(X(\tau))\mid x]+\mathrm{E}[\beta^{\tau}\mid x]\Delta(0',W)\le D(x)+\Delta(0',W)$$
$$\iff\quad (C(x,\tau)-D(x)+\mathrm{E}[\beta^{\tau}D(X(\tau))\mid x])\{1-\mathrm{E}[\beta^{\tau}\mid x]\}^{-1}\le\Delta(0',W). \quad (14)$$

Clearly, from Definition 3 and the fact that the infimum in (12) is always achieved, the requirement in (14) is met precisely when

$$G(x)\le\Delta(0',W). \quad (15)$$

Now let $\Pi(W)$ denote the set of states in which it is optimal to take the passive action for the charge-W problem. From the analysis leading to (15), it follows that

$$\Pi(W)=\{x\colon G(x)\le\Delta(0',W)\}. \quad (16)$$

By Theorem 1, $\Delta(0',W)$ is strictly increasing in W. From (16) it then follows that $\Pi(W)$ is increasing, and indexability follows from Definition 1. It also holds, from the continuity of $\Delta(0',\cdot)$, that the Whittle index for state x, namely $W(x)=\inf\{W\colon x\in\Pi(W)\}$, satisfies the equation

$$\Delta(0',W(x))=G(x). \quad (17)$$

Owing to the strictly increasing nature of $\Delta(0',W)$, (17) specifies $W(x)$ uniquely. This establishes parts (a) and (b) of the theorem. Part (c) follows simply from the resultant fact that $W(x)$ is strictly increasing in $G(x)$. This concludes the proof.

Remark 3. Note that it is an immediate consequence of Theorems 1 and 2 and (12) that $W(0')=-D(0')$. Hence, the pristine state has a negative Whittle index. Subsequent analysis will focus on developing indices for nonpristine states.

To proceed further, recall the ideas and notation established in (13) and the discussion thereof.

Lemma 1. (a) *Any stopping time in* $T(0',\Delta(0',W))$ *achieves the infimum in (11).*

(b) *Any stopping time in* $T(x,G(x))$ *achieves the infimum in (12), for* $x\in\Omega$.

In both cases, these are the only stationary stopping times which achieve the infima concerned.

Proof. Part (b) is summarised in the comments around (13) and is a feature of the Gittins index structure. To prove part (a), suppose that $X(0)=0'$ and that the machine state evolves under the passive action. Extend the bandit's state space to $\Omega\cup\{0^*\}$, where '0^*' is used specifically to designate state $0'$ at time 0, with '$0'$' reserved for the pristine state at other times. Furthermore, we impose the following costs: a transition from 0^* to state x incurs a cost of $W+C(0',x)+\beta D(x)$ and a transition from state $y\neq 0^*$ to state z incurs a cost

https://doi.org/10.1239/aap/1158684996 Published online by Cambridge University Press

of $C(y, z) - D(y) + \beta D(z)$. With these choices, the expected cost incurred by the machine during $[0, \tau)$ is given by

$$W + C(0', \tau) + \mathrm{E}[\beta^{\tau} D(X(\tau)) \mid 0'],$$

where τ is a positive-valued stopping time. Moreover, the expected cost incurred by the machine during $[0, \tau)$, from some initial state $x \neq 0^*$, is given by

$$C(x, \tau) - D(x) + \mathrm{E}[\beta^{\tau} D(X(\tau)) \mid x].$$

If we regard the so-constructed object as a Gittins-type bandit, $\Delta(0', W)$ is by definition the Gittins index for the initial state 0^* (see (11)). Furthermore, $G(x)$, in (12), is the Gittins index for any state $x \neq 0^*$. Lemma 1(a) is now seen to be an application to this bandit of the comments around (13). This concludes the proof.

We now explore index structure in the context of two model types, both of which rest on assumptions which are plausible in practice.

3.1. Monotone models

In addition to that specified in I–III above, *monotone models* have the following structure.

IV. The state space Ω is the set of natural numbers, $\mathbb{N}$, with 0 the designated pristine state.

V. Evolution under the passive action is *right-skip free*. This means that

$$P(x, y) = 0, \qquad y > x + 1, \ x \in \mathbb{N}.$$

VI. The Gittins index for passivity, $G : \mathbb{N} \to \mathbb{R}$, is (strictly) increasing.

Hence, under such models an increase in state corresponds to deterioration of the machine, resulting in higher cost rates (as measured by the Gittins index).

Now suppose that $X(0) = x$ and that the machine state evolves under the passive action. We use $\tau_{x,y}$ for the time of the first entry after 0 into state y. Note that the assumption (in item V) that passive evolution be right-skip free implies that $\tau_{x,y} < \tau_{x,y+1}$ almost surely, for all $x < y$.

Theorem 3. (Whittle indices for monotone models.) *For monotone models, the Whittle index is given, for $x \in \mathbb{Z}^+$, by*

$$\begin{aligned} W(x) &= G(x)\{1 - \mathrm{E}[\beta^{\tau_{0,x}} \mid 0]\} - C(0, \tau_{0,x}) - \mathrm{E}[\beta^{\tau_{0,x}} \mid 0] D(x) \qquad (18) \\ &= [C(x, \tau_{x,x+1})\{1 - \mathrm{E}[\beta^{\tau_{0,x}} \mid 0]\} - C(0, \tau_{0,x})\{1 - \mathrm{E}[\beta^{\tau_{x,x+1}} \mid x]\} \\ &\quad - \{1 - \mathrm{E}[\beta^{\tau_{0,x+1}} \mid 0]\} D(x) + \{\mathrm{E}[\beta^{\tau_{x,x+1}} \mid x] - \mathrm{E}[\beta^{\tau_{0,x+1}} \mid 0]\} D(x+1)] \\ &\quad \times \{1 - \mathrm{E}[\beta^{\tau_{x,x+1}} \mid x]\}^{-1} \qquad (19) \end{aligned}$$

and is increasing in x.

Proof. By Theorem 2, $W(x)$ is the unique W-solution to $\Delta(0, W) = G(x)$. Furthermore, by Lemma 1, $\Delta(0, W)$ is achieved by every member of $T(0, \Delta(0, W)) \equiv T(0, G(x))$. However, the right-skip-free nature of passive evolution and the monotonicity of Gittins indices means that $\tau_{0,x} \in T(0, G(x))$. We conclude from (11) that $W(x)$ is the W-solution to

$$\Delta(0, W) = \{W + C(0, \tau_{0,x}) + \mathrm{E}[\beta^{\tau_{0,x}} \mid 0] D(x)\}\{1 - \mathrm{E}[\beta^{\tau_{0,x}} \mid 0]\}^{-1} = G(x). \qquad (20)$$

Solving (20) for W yields (18).

https://doi.org/10.1239/aap/1158684996 Published online by Cambridge University Press

To obtain (19) we observe that V and VI and Lemma 1 imply that $G(x)$ is achieved by $\tau_{x,x+1} \in T(x, G(x))$ and, hence, from (12), that

$$G(x) = \{C(x, \tau_{x,x+1}) - D(x) + \mathrm{E}[\beta^{\tau_{x,x+1}} \mid x]D(x+1)\}\{1 - \mathrm{E}[\beta^{\tau_{x,x+1}} \mid x]\}^{-1}. \tag{21}$$

Equation (19) now follows upon substitution of (21) into (18) and utilisation of the identity

$$\mathrm{E}[\beta^{\tau_{0,x+1}} \mid 0] = \mathrm{E}[\beta^{\tau_{0,x}} \mid 0]\,\mathrm{E}[\beta^{\tau_{x,x+1}} \mid x].$$

Finally, that $W(x)$ is increasing in x follows from the facts that $W(x)$ is strictly increasing in $G(x)$ and $G(x)$ is increasing in x for monotone models. This concludes the proof.

3.2. Breakdown/deterioration models

In addition to that specified in I–III above, *breakdown/deterioration models* have the following structure.

VII. The state space Ω is the set of natural numbers, $\mathbb{N}$, with 0 the designated pristine state.

VIII. Evolution under the passive action is such that

$$P(x, 0) + P(x, x) + P(x, x+1) = 1, \qquad x \in \mathbb{Z}^+.$$

Hence, under the passive action, a machine currently in state x may remain there (with probability $P(x, x)$), have a catastrophic breakdown followed by immediate maintenance/replacement (with probability $P(x, 0)$), or deteriorate by a single unit (with probability $P(x, x+1)$). It will simplify the discussion if we further suppose that $P(x, 0) + P(x, x+1)$ is strictly positive for all $x \in \mathbb{N}$. For the pristine state, 0, we use $P(0)$ for the probability of a catastrophic breakdown to distinguish it from $P(0, 0)$, the probability of a nondeparture from the pristine state. We suppose that $P(0) + P(0, 0) + P(0, 1) = 1$.

For a fixed $x \in \mathbb{Z}^+$, we define $\underline{x} \le x$ and $\overline{x} > x$ as follows:

$$\underline{x} = \min\{y\colon G(y) \ge G(x)\},$$
$$\overline{x} = \min\{y\colon y \ge x+1,\ G(y) \ge G(x)\}.$$

We take $\overline{x} = \infty$ if $G(y) < G(x),\ y \ge x+1$. Note that for monotone models we have $\underline{x} = x$ and $\overline{x} = x+1$. Consider the bandit evolving from state x at time 0 under the passive action. We now introduce the positive-valued stopping time

$$\begin{aligned}\tau(x; \underline{x}, \overline{x}) &= \min\{\tau_{x,\underline{x}}, \tau_{x,\overline{x}}\} \\ &= \min\{t \ge 1\colon X(t) = \underline{x} \text{ or } X(t) = \overline{x}\}.\end{aligned} \tag{22}$$

Theorem 4. (Whittle indices for breakdown/deterioration models.) *For breakdown/deterioration models, the Whittle index is given, for $x \in \mathbb{Z}^+$, by*

$$W(x) = G(x)\{1 - \mathrm{E}[\beta^{\tau_{0,\underline{x}}} \mid 0]\} - C(0, \tau_{0,\underline{x}}) - \mathrm{E}[\beta^{\tau_{0,\underline{x}}} \mid 0]D(\underline{x}), \tag{23}$$

where

$$\begin{aligned}G(x) = \{&C(x, \tau(x; \underline{x}, \overline{x})) - D(x) + \mathrm{E}[\beta^{\tau(x;\underline{x},\overline{x})} D(X(\tau(x; \underline{x}, \overline{x}))) \mid x]\} \\ &\times \{1 - \mathrm{E}[\beta^{\tau(x;\underline{x},\overline{x})} \mid x]\}^{-1}.\end{aligned} \tag{24}$$

https://doi.org/10.1239/aap/1158684996 Published online by Cambridge University Press

For the special case $D(x) = D,\ x \in \mathbb{N}$,

$$W(x) = C(x, \tau(x; \underline{x}, \overline{x}))\{1 - \mathrm{E}[\beta^{\tau_{0,\underline{x}}} \mid 0]\}\{1 - \mathrm{E}[\beta^{\tau(x;\underline{x},\overline{x})} \mid x]\}^{-1} - C(0, \tau_{0,\underline{x}}) - D, \qquad x \in \mathbb{Z}^+. \tag{25}$$

Proof. If $X(0) = x$ then, under VIII above, evolution under the passive action will be to states above x until a transition to state 0 occurs. From state 0, evolution will be to states above 0 until further transitions to 0 occur. It then follows that $\tau(x; \underline{x}, \overline{x}) \in T(x, G(x))$ and, hence, from Lemma 1(b), that it achieves the infimum in (12). This gives (24). Furthermore, if W satisfies $\Delta(0, W) = G(x)$, then $\tau_{0,\underline{x}} \in T\{0, \Delta(0, W)\}$ and, so, achieves the infimum in (11). It now follows from Theorem 2 that $W(x)$ is the W-solution to

$$\Delta(0, W) = \{W + C(0, \tau_{0,\underline{x}}) + \mathrm{E}[\beta^{\tau_{0,\underline{x}}} \mid 0]D(\underline{x})\}\{1 - \mathrm{E}[\beta^{\tau_{0,\underline{x}}} \mid 0]\}^{-1} = G(x). \tag{26}$$

Solving (26) for W yields (23). Equation (25) follows from (23) upon substitution into the latter of (24) and utilisation of the condition $D(x) = D,\ x \in \mathbb{N}$. This concludes the proof.

3.3. Examples

We now proceed to illustrate and extend the above material by deriving explicit formulae for the Whittle indices in important special cases. We consider examples of the breakdown/deterioration models (as in VII and VIII above) for which transitions to 0 under the passive action correspond to catastrophic unexpected breakdowns of the machine (followed by its immediate replacement/renewal). Such transitions may incur great costs. Hence, costs incurred by the bandit undergoing transitions from state x to state 0, state x, or state $x + 1$ under the passive action will be taken to be of the respective forms $K + C(x)$, $C(x)$, and $C(x)$, where K is the cost of a catastrophic breakdown. The cost of a transition from state x under the active action will be $D(x) + C(0)$ for the bandit and $D(x) + C(0) + W$ for the charge-W problem. It will simplify matters if we suppose that $P(0) = 0$, i.e. that there are no catastrophic breakdowns in the pristine state. We shall explore instances of such models which are also monotone, satisfying VI above.

Calculations can be presented more economically in notational terms if we write

$$\begin{aligned} \beta P(x, 0)\{1 - \beta P(x, x)\}^{-1} &= \delta(x), \qquad x \in \mathbb{Z}^+, \\ \beta P(x, x + 1)\{1 - \beta P(x, x)\}^{-1} &= \gamma(x), \qquad x \in \mathbb{N}, \\ \{1 - \beta P(x, x)\}^{-1} &= \varepsilon(x), \qquad x \in \mathbb{N}. \end{aligned}$$

All the following formulae hold for $x \in \mathbb{Z}^+$. Simple conditioning arguments yield the conclusion that the expected cost $C(0, \tau_{0,x})$ satisfies the equation

$$C(0, \tau_{0,x}) = \sum_{y=0}^{x-1} C(y)\varepsilon(y) \prod_{z=0}^{y-1} \gamma(z) + \{K\beta^{-1} + C(0, \tau_{0,x})\} \sum_{y=1}^{x-1} \delta(y) \prod_{z=0}^{y-1} \gamma(z)$$

and, hence, that

$$C(0, \tau_{0,x}) = \sum_{y=0}^{x-1} \{C(y)\varepsilon(y) + K\beta^{-1}\delta(y)1(y \neq 0)\} \prod_{z=0}^{y-1} \gamma(z) \left\{1 - \sum_{y=1}^{x-1} \delta(y) \prod_{z=0}^{y-1} \gamma(z)\right\}^{-1}, \tag{27}$$

https://doi.org/10.1239/aap/1158684996 Published online by Cambridge University Press

where straightforward algebra yields

$$1-\sum_{y=1}^{x-1}\delta(y)\prod_{z=0}^{y-1}\gamma(z)=(1-\beta)\sum_{y=0}^{x-1}\varepsilon(y)\prod_{z=0}^{y-1}\gamma(z)+\prod_{y=0}^{x-1}\gamma(y). \tag{28}$$

Similarly, we have

$$\mathrm{E}[\beta^{\tau_{0,x}}\mid 0]=\prod_{y=0}^{x-1}\gamma(y)\left\{1-\sum_{y=1}^{x-1}\delta(y)\prod_{z=0}^{y-1}\gamma(z)\right\}^{-1}. \tag{29}$$

As noted just above (22), in monotone cases we have $\underline{x}=x$ and $\overline{x}=x+1$. It then follows that

$$C(x,\tau(x;\underline{x},\overline{x}))=C(x;\tau(x;x,x+1))=C(x)+P(x,0)\{K+\beta C(0,\tau_{0,x})\}$$

and

$$\mathrm{E}[\beta^{\tau(x;\underline{x},\overline{x})}\mid x]=\mathrm{E}[\beta^{\tau(x;x,x+1)}\mid x]=\beta\{1-P(x,0)\}+\beta P(x,0)\,\mathrm{E}[\beta^{\tau_{0,x}}\mid 0]. \tag{30}$$

We note further that, using (29),

$$\begin{aligned}&\mathrm{E}[\beta^{\tau(x;x,x+1)}D(X(\tau(x;x,x+1)))\mid x]\\&\quad=\beta P(x,x)D(x)+\beta P(x,x+1)D(x+1)\\&\qquad+\beta P(x,0)D(x)\prod_{y=0}^{x-1}\gamma(y)\left\{1-\sum_{y=1}^{x-1}\delta(y)\prod_{z=0}^{y-1}\gamma(z)\right\}^{-1}.\end{aligned} \tag{31}$$

Now, combining (27), (28), (30), and (31), we infer that

$$\begin{aligned}H(x)&:=\{C(x,\tau(x;x,x+1))-D(x)+\mathrm{E}[\beta^{\tau(x;x,x+1)}D(X(\tau(x;x,x+1)))\mid x]\}\\&\quad\times\{1-\mathrm{E}[\beta^{\tau(x;x,x+1)}\mid x]\}^{-1}\\&=\Bigg(KP(x,0)+\sum_{y=0}^{x-1}\Bigg[\Bigg([\{C(x)-D(x)\}(1-\beta)+\beta P(x,0)C(y)\\&\qquad+\{\beta P(x,x)D(x)+\beta P(x,x+1)D(x+1)\}(1-\beta)]\\&\qquad\times\varepsilon(y)\prod_{z=0}^{y-1}\gamma(z)\Bigg)\\&\qquad+\{C(x)-D(x)+\beta(1-P(x,x+1))D(x)\\&\qquad+\beta P(x,x+1)D(x+1)\}\prod_{y=0}^{x-1}\gamma(y)\Bigg]\Bigg)\\&\quad\times\left\{(1-\beta)\left[(1-\beta+\beta P(x,0))\sum_{y=0}^{x-1}\varepsilon(y)\prod_{z=0}^{y-1}\gamma(z)+\prod_{y=0}^{x-1}\gamma(y)\right]\right\}^{-1},\qquad x\in\mathbb{Z}^+.\end{aligned} \tag{32}$$

https://doi.org/10.1239/aap/1158684996 Published online by Cambridge University Press

We extend the domain of H to $\mathbb{N}$ by defining

$$\begin{aligned} H(0) &= \{C(0,\tau_{0,1}) - D(0) + \mathrm{E}[\beta^{\tau_{0,1}} D(1) \mid 0]\}\{1 - \mathrm{E}[\beta^{\tau_{0,1}} \mid 0]\}^{-1} \\ &= \{C(0) - D(0) + \beta P(0,0)D(0) + \beta P(0,1)D(1)\}(1-\beta)^{-1}. \end{aligned}$$

The following result utilises a self-consistency result for Gittins indices due to Nash (1979).

Lemma 2. *If $H : \mathbb{N} \to \mathbb{R}$ is increasing then $H(x) = G(x),\ x \in \mathbb{N}$.*

We now consider two special cases to which Lemma 2 can be applied.

3.3.1. *Special case 1:* $P(x,0) = 0,\ x \in \mathbb{Z}^+$. Here we have no catastrophic breakdowns, and between interventions the machine is subject only to gradual deterioration and (typically) increasing maintenance costs. Note that, from (32),

$$\begin{aligned} P(x,0) = 0 \quad \Longrightarrow \quad H(x) = \{&C(x) - D(x) + \beta P(x,x)D(x) + \beta P(x,x+1)D(x+1)\} \\ &\times \{1-\beta\}^{-1}, \qquad x \in \mathbb{N}. \end{aligned} \tag{33}$$

Corollary 1. *If $P(x,0) = 0$ and*

$$H(x) = \{C(x) - D(x) + \beta P(x,x)D(x) + \beta P(x,x+1)D(x+1)\}\{1-\beta\}^{-1}$$

is increasing, then $H(x) = G(x),\ x \in \mathbb{N}$. Furthermore, if $D(x) = D,\ x \in \mathbb{N}$, then the Whittle index is given by

$$W(x) = \sum_{y=0}^{x-1} \{C(x) - C(y)\}\varepsilon(y) \prod_{z=0}^{y-1} \gamma(z) - D, \qquad x \in \mathbb{Z}^+. \tag{34}$$

Proof. The first statement of the corollary is an immediate consequence of Lemma 2 and (33). For case $D(x) = D,\ x \in \mathbb{N}$, it follows from (25), current model assumptions, and the above calculations that

$$W(x) + D = C(x)\{1 - \mathrm{E}[\beta^{\tau_{0,x}} \mid 0]\}\{1-\beta\}^{-1} - C(0,\tau_{0,x}). \tag{35}$$

The expression in (34) then follows from (27), (28), (29), and (35).

Remark 4. Note that if $D(x) = D,\ x \in \mathbb{N}$, then $H(x)$ in (33) will be increasing if the maintenance cost $C(x)$ is increasing in x. It will then follow from Theorem 2(c) that $W(x)$ is also increasing in x.

3.3.2. *Special case 2:* $C(x) = 0$ *and* $D(x) = D,\ x \in \mathbb{N}$. In contrast to special case 1, the focus of special case 2 is predominantly on the minimisation of costs due to catastrophic breakdowns. Upon substitution of $C(x) = 0$ and $D(x) = D,\ x \in \mathbb{N}$, into (32), we obtain

$$\begin{aligned} &H(0) = -D, \\ &H(x) = KP(x,0)\left\{(1-\beta)\left[(1-\beta+\beta P(x,0))\sum_{y=0}^{x-1}\varepsilon(y)\prod_{z=0}^{y-1}\gamma(z) + \prod_{y=0}^{x-1}\gamma(y)\right]\right\}^{-1} - D, \\ &\qquad\qquad x \in \mathbb{Z}^+. \end{aligned} \tag{36}$$

https://doi.org/10.1239/aap/1158684996 Published online by Cambridge University Press

Corollary 2. *If $C(x) = 0$ and $D(x) = D$, $x \in \mathbb{N}$, and $P(x,0)$ is increasing in x, then $H : \mathbb{N} \to \mathbb{R}$ in (36) is increasing and $H(x) = G(x)$, $x \in \mathbb{N}$. The Whittle index is then given by*

$$W(x) = K\left[P(x,0)\varepsilon(0) + \sum_{y=1}^{x-1}\{P(x,0) - P(y,0)\}\varepsilon(y)\prod_{z=0}^{y-1}\gamma(z)\right]$$
$$\times\left[\{1-\beta+\beta P(x,0)\}\sum_{y=0}^{x-1}\varepsilon(y)\prod_{z=0}^{y-1}\gamma(z) + \prod_{y=0}^{x-1}\gamma(y)\right]^{-1} - D, \qquad x \in \mathbb{Z}^+.$$

Proof. We first note from (36) that $H(0) \le H(x)$, $x \in \mathbb{Z}^+$. If $C(x) = 0$ and $D(x) = D$, $x \in \mathbb{Z}^+$, then (36) further implies that

$$\begin{aligned}
&K(1-\beta)^{-1}\{H(x+1)+D\}^{-1}\\
&\quad= \{P(x+1,0)\}^{-1}\left[\{1-\beta+\beta P(x+1,0)\}\sum_{y=0}^{x}\varepsilon(y)\prod_{z=0}^{y-1}\gamma(z) + \prod_{y=0}^{x}\gamma(y)\right]\\
&\quad= \{P(x+1,0)\}^{-1}\{1-\beta+\beta P(x+1,0)\}\sum_{y=0}^{x-1}\varepsilon(y)\prod_{z=0}^{y-1}\gamma(z)\\
&\qquad+ \{P(x+1,0)\}^{-1}\{1-\beta+\beta P(x+1,0)\}\varepsilon(x)\prod_{y=0}^{x-1}\gamma(y)\\
&\qquad+ \{P(x+1,0)\}^{-1}\prod_{y=0}^{x}\gamma(y)\\
&\quad= \{P(x+1,0)\}^{-1}\{1-\beta+\beta P(x+1,0)\}\sum_{y=0}^{x-1}\varepsilon(y)\prod_{z=0}^{y-1}\gamma(z)\\
&\qquad+ \{P(x+1,0)\}^{-1}\prod_{y=0}^{x-1}\gamma(y)[\{1-\beta+\beta P(x+1,0)\}\varepsilon(x)+\gamma(x)]\\
&\quad\le \{P(x,0)\}^{-1}\{1-\beta+\beta P(x,0)\}\sum_{y=0}^{x-1}\varepsilon(y)\prod_{z=0}^{y-1}\gamma(z)\\
&\qquad+ \{P(x,0)\}^{-1}\prod_{y=0}^{x-1}\gamma(y)[\{1-\beta+\beta P(x,0)\}\varepsilon(x)+\gamma(x)]\\
&\quad= \{P(x,0)\}^{-1}\left[\{1-\beta+\beta P(x,0)\}\sum_{y=0}^{x-1}\varepsilon(y)\prod_{z=0}^{y-1}\gamma(z) + \prod_{y=0}^{x-1}\gamma(y)\right]\\
&\quad= K(1-\beta)^{-1}\{H(x)+D\}^{-1} \qquad (37)
\end{aligned}$$

whenever $P(x,0) \le P(x+1,0)$. It follows from (37) that if $P(x,0)$ is increasing then so is $H(x)$. The inference that $H(x) = G(x)$ uses Lemma 2.

https://doi.org/10.1239/aap/1158684996 Published online by Cambridge University Press

TABLE 1: Level of performance of the Whittle index policy (percentage suboptimality) for a class of machine maintenance problems with varying intervention costs.

D	MIN	LQ	MEDIAN	UQ	MAX
25	0.0000	0.0162	0.0515	0.1139	0.7039
50	0.0000	0.0360	0.1015	0.2265	1.7127
75	0.0000	0.0469	0.1700	0.3869	1.7455
100	0.0000	0.0300	0.1359	0.3690	1.6559
125	0.0000	0.0220	0.1164	0.3273	1.5220
150	0.0000	0.0244	0.1261	0.3035	1.2946
175	0.0000	0.0199	0.1073	0.2649	1.1338
200	0.0000	0.0158	0.0882	0.2376	0.9409

From (23) we then have

$$W(x) + D = \{H(x) + D\}\{1 - \mathrm{E}[\beta^{\tau_{0,x}} \mid 0]\} - C(0, \tau_{0,x}). \tag{38}$$

Substitution into (38) from (27), (29), and (36), together with straightforward algebra, yields the expression for the Whittle index given in the statement of the result.

3.4. Numerical study

Here we describe some numerical results which assess the quality of performance of an index policy based on the set-up considered in Corollary 1. To be precise, we explore a scenario in which a single repairman is maintaining four machines. This is modelled as a restless bandit problem with $N = 1$ and $M = 5$, i.e. a single server choosing among five bandits. Four of the five bandits model machine state evolution, as above. The fifth is an idling option modelled as a single-state bandit with no costs incurred under either action (active or passive). The Whittle index for the idling option is trivially 0. Hence, the Whittle index heuristic chooses idleness when the four bandits modelling machine evolution all have negative index values. Otherwise, the repairman works on whichever machine has the largest (positive) index.

In Table 1 we present results summarising the performance of the Whittle index heuristic as (fixed) intervention cost D increases from 25 to 200. Each row of the table summarises the results of 1000 problems studied for a fixed D. These 1000 problems were generated at random as follows: each of the four machines has ten states, labelled $0, 1, \ldots, 9$. The cost rate for machine i in state x takes the form

$$C_i(x) = A_i + B_i x, \qquad 1 \le i \le 4,\ 0 \le x \le 9,$$

where the A_i and B_i were obtained by sampling independently from a U(25, 50) distribution. The probabilities $P_i(x, x)$, $1 \le i \le 4, 0 \le x \le 8$, were obtained by sampling independently from a U(0.1, 0.9) distribution. We then set

$$P_i(x, x+1) = 1 - P_i(x, x), \quad 1 \le i \le 4,\ 0 \le x \le 8, \qquad P_i(9, 9) = 1, \quad 1 \le i \le 4.$$

The discount rate, β, was set to be 0.95 throughout.

For each problem generated, the optimal expected cost, $V^{\mathrm{opt.}}$, was computed for the initial state in which all machines are assumed to be in pristine state 0, along with the corresponding expected cost, $V^{\mathrm{ind.}}$, for the Whittle index heuristic. All computations of expected cost were

https://doi.org/10.1239/aap/1158684996 Published online by Cambridge University Press

performed using dynamic programming value iteration. The percentage cost suboptimality, $100\{V^{\text{ind.}} - V^{\text{opt.}}\}\{V^{\text{opt.}}\}^{-1}$, was then calculated. For each value of D, the 1000 percentage suboptimalites were then summarised in terms of the order statistics MIN (minimum), LQ (lower quartile), MEDIAN (median), UQ (upper quartile), and MAX (maximum), and it is these which appear in Table 1. Note the uniformly excellent level of performance of the index heuristic. In none of the 8000 problems studied was the index policy more than 1.75% suboptimal.

4. Model 2: stochastic scheduling with switching costs

We now seek an indexability analysis of a class of structured bandits aimed at the incorporation of switching penalties into a general stochastic scheduling model in the form of a multiarmed bandit. Hence, we shall initially consider a bandit in which a state-dependent switching cost is paid whenever the passive action (bandit not processed) is followed by the active action (bandit processed). After this initial analysis we shall describe a range of further model developments which preserve indexability. These include (a) switching times in addition to switching costs, (b) tear-down costs whenever the active action is followed by the passive action, and (c) losses from the system under the passive action.

We use $\{\Omega, P, R, S, \beta\}$ to denote the reward-based bandit of interest. The general structure specified in 1–6 in Section 2 is specialised to this case as follows:

I′. The state space of the bandit takes the form $\{a, b\} \times \Omega$, where Ω is countable. If $X(t) = (a, x),\ x \in \Omega$, when $t \geq 1$, then the job modelled by the bandit is in state x and action a (active) was taken at time $t - 1$. If $X(t) = (b, x),\ x \in \Omega$, when $t \geq 1$, then the job modelled by the bandit is in state x and action b (passive) was taken at time $t - 1$. It is usual to assume that $X(0) = (b, x)$ for some x (no processing before time 0), but this is only required at one point in the discussion (see Section 4.2). We write $\bar{X}(t)$ for the job state, omitting the action information, so

$$\bar{X}(t) = x \quad \Longrightarrow \quad X(t) \in \{(a, x), (b, x)\}, \qquad x \in \Omega.$$

II′. If action a is taken at time t, then we have

$$\mathrm{P}\{X(t+1) = (a, y) \mid X(t) = (a, x), a\} = P(x, y), \qquad x, y \in \Omega, \tag{39}$$
$$\mathrm{P}\{X(t+1) = (a, y) \mid X(t) = (b, x), a\} = P(x, y), \qquad x, y \in \Omega, \tag{40}$$

where P is the Markov law determining the evolution of the process $\{\bar{X}(t),\ t \geq 0\}$ under the active action. If action b is taken at time t, then we have

$$\begin{aligned}&\mathrm{P}\{X(t+1) = (b, x) \mid X(t) = (a, x), b\}\\ &\quad = \mathrm{P}\{X(t+1) = (b, x) \mid X(t) = (b, x), b\} = 1, \qquad x \in \Omega.\end{aligned} \tag{41}$$

III′. The rewards earned by the bandit under the transitions in (39) and (40) are respectively $R(x, y)$ and $-S(x) + R(x, y)$. Here $S(x)$ is the (strictly) positive cost incurred when switching processing to the bandit when in state x, and $R(x, y)$ is the reward earned under active processing from a transition from x to y. The transitions in (41) earn no rewards for the bandit. The corresponding rewards earned by the transitions in (39), (40), and (41) for the subsidy-W problem are respectively $R(x, y)$, $-S(x) + R(x, y)$, and W, where W is the subsidy for passivity.

https://doi.org/10.1239/aap/1158684996 Published online by Cambridge University Press

As above, we denote by $V(\hat{x}, W)$ the value function of the subsidy-W problem evaluated at $\hat{x} \in \{a, b\} \times \Omega$. From I′–III′, the optimality equations may be expressed as

$$V((a,x),W) = \max\left\{R(x) + \beta \sum_{y\in\Omega} P(x,y)V((a,y),W),\ W + \beta V((b,x),W)\right\}, \quad x \in \Omega, \tag{42}$$

$$V((b,x),W) = \max\left\{-S(x) + R(x) + \beta \sum_{y\in\Omega} P(x,y)V((a,y),W),\right.$$
$$\left. W + \beta V((b,x),W)\right\}, \quad x \in \Omega. \tag{43}$$

On the right-hand sides of both (42) and (43), the first argument of $\max\{\cdot,\cdot\}$ corresponds to the choice of the active action in the current state and the second argument to the choice of the passive action. Because of the appearance of $V((b,x),W)$ on both sides of (43), it is trivial to see that it is equivalent to the simpler equation

$$V((b,x),W) = \max\left\{-S(x)+R(x)+\beta \sum_{y\in\Omega} P(x,y)V((a,y),W),\ W(1-\beta)^{-1}\right\}, \quad x \in \Omega. \tag{44}$$

In order to develop optimal policies for the subsidy-W problem, we develop the *Gittins index for activity* in an analogous way to that of the Gittins index for passivity (see Definition 3). Hence, suppose that $\bar{X}(0) = x \in \Omega$ and that active action a is taken at times $0, 1, 2, \ldots, \nu - 1$, where ν is a stationary, positive-valued stopping time on the process $\{\bar{X}(t),\ t \geq 0\}$. We write

$$R(x,\nu) = \mathrm{E}\left[\sum_{t=0}^{\nu-1} \beta^t R(\bar{X}(t)) \,\middle|\, x\right]$$

for the expected reward earned during this processing.

Definition 4. The Gittins index for activity, $G\colon \{a, b\} \times \Omega \to \mathbb{R}$, is given by

$$G(a,x) = \sup_{\nu}[R(x,\nu)\{1 - \mathrm{E}[\beta^\nu \mid x]\}^{-1}], \qquad x \in \Omega, \tag{45}$$

$$G(b,x) = \sup_{\nu}[\{-S(x) + R(x,\nu)\}\{1 - \mathrm{E}[\beta^\nu \mid x]\}^{-1}], \qquad x \in \Omega, \tag{46}$$

where the suprema are taken over all stationary, positive-valued stopping times on the process $\{\bar{X}(t),\ t \geq 0\}$ evolving under the active action.

The suprema in (45) and (46) are guaranteed to be achieved. To see this we modify the material around (13) as follows. Fix a $W \in \mathbb{R}$ and denote by $\Gamma'(W)$ the subset of Ω given by

$$\Gamma'(W) = \{y \in \Omega\colon G(a,y) < W\}$$

and by $\Sigma'(W)$ the subset of Ω given by

$$\Sigma'(W) = \{y \in \Omega\colon G(a,y) = W\}.$$

https://doi.org/10.1239/aap/1158684996 Published online by Cambridge University Press

Now suppose that $\bar{X}(0) = x$ and $\Sigma \subseteq \Sigma'(W)$, and use ν^Σ to denote the stationary, positive-valued stopping time, defined on the process $\{\bar{X}(t),\ t \geq 0\}$ evolving under the active action, given by

$$\nu^\Sigma = \min\{t : t > 0 \text{ and } \bar{X}(t) \in \Gamma'(W) \cup \Sigma\}.$$

Now write $T'(x, W)$ for the collection of stopping times given by

$$T'(x, W) = \bigcup_{\Sigma \subseteq \Sigma'(W)} \{\nu^\Sigma\}.$$

The following result combines straightforward calculations with standard features of Gittins' index theory. We omit the proof.

Lemma 3. (a) $G(a, x) > G(b, x),\ x \in \Omega$.

(b) *Any stopping time in* $T'(x, G(a, x))$ *achieves the supremum in (45).*

(c) *Any stopping time in* $T'(x, G(b, x))$ *achieves the supremum in (46).*

The stationary stopping times in (b) and (c) are the only ones which achieve the suprema concerned.

Before proceeding to the main result of this section, we pause to recollect the work of Whittle (1980), who utilised a set of decision problems involving a notion of *retirement* to characterise the Gittins index. Suppose that $X(0) = \hat{x} \in \{a, b\} \times \Omega$. Consider a decision problem in which, at each time $t \in \mathbb{N}$, a choice has to be made between the active action a and retirement. Once retirement is chosen, it must continue to be chosen thereafter. The effect of choosing the active action (in terms of stochastic evolution and rewards earned) is precisely as in I′ and III′ above. A reward W is earned on each occasion that retirement is chosen. If we write ν for the first time at which retirement is chosen, then we may express the value function for the retirement problem as

$$\hat{V}((a, x), W) = \sup_\nu [R(x, \nu)1(\nu > 0) + \mathrm{E}[\beta^\nu \mid x]W(1 - \beta)^{-1}], \qquad x \in \Omega,$$

$$\hat{V}((b, x), W) = \sup_\nu [(-S(x) + R(x, \nu))1(\nu > 0) + \mathrm{E}[\beta^\nu \mid x]W(1 - \beta)^{-1}], \qquad x \in \Omega.$$

The following result may be established straightforwardly from Whittle's analysis.

Lemma 4. (Optimal retirement.) (a) *If* $X(0) = (a, x)$ *and* $W(1 - \beta)^{-1} > G(a, x)$, *then it is optimal to retire at time* 0.

(b) *If* $X(0) = (a, x)$ *and* $W(1 - \beta)^{-1} = G(a, x)$, *then retirement at time* 0 *is optimal along with retirement at any stopping time in* $T'(x, G(a, x))$.

(c) *If* $X(0) = (a, x)$ *and* $W(1 - \beta)^{-1} < G(a, x)$, *then retirement at any stopping time in* $T'(x, W(1 - \beta)^{-1})$ *is optimal.*

(d) *Statements (a)–(c) all hold with action a replaced by action b throughout.*

(e) *Statements (a)–(d) describe all optimal stationary policies for retirement.*

We are now in a position to establish the structure of all optimal stationary policies for the subsidy-W problem. Indexability will follow straightforwardly from Theorem 5.

https://doi.org/10.1239/aap/1158684996 Published online by Cambridge University Press

Theorem 5. (Optimal policies for the subsidy-W problem.) *The following statements hold for all $x \in \Omega$ and $W \in \mathbb{R}$, and describe all optimal stationary policies for the subsidy-W problem.*

(a) *If $W(1-\beta)^{-1} > G(a,x)$ then the passive action b is optimal in states (a,x) and (b,x).*

(b) *If $G(a,x) > W(1-\beta)^{-1} > G(b,x)$ then the active action a is optimal in state (a,x) with the passive action b optimal in (b,x).*

(c) *If $G(b,x) > W(1-\beta)^{-1}$ then the active action a is optimal in states (a,x) and (b,x).*

(d) *If $W(1-\beta)^{-1} = G(a,x)$ then actions a and b are both optimal in (a,x) with action b optimal in (b,x).*

(e) *If $W(1-\beta)^{-1} = G(b,x)$ then action a is optimal in (a,x) with actions a and b both optimal in (b,x).*

Proof. Fix a $W \in \mathbb{R}$ and suppose that state $x \in \Omega$ is such that the passive action b is optimal for the subsidy-W problem when the bandit is in state (a,x). It now follows from (42) and (43) that

$$\begin{aligned} W + \beta V((b,x),W) &\geq R(x) + \beta \sum_{y\in\Omega} P(x,y) V((a,y),W) \\ &> -S(x) + R(x) + \beta \sum_{y\in\Omega} P(x,y) V((a,y),W), \end{aligned}$$

from which it follows that passive action b must also be (strictly) optimal for the subsidy-W problem when the bandit is in state (b,x).

Now suppose that $X(0) = \hat{x} \in \{a,b\} \times \Omega$ and let $\tilde{\pi}$ be some stationary optimal policy for the subsidy-W problem. Write $\nu(\tilde{\pi}, \hat{x})$ for the first time at which $\tilde{\pi}$ chooses the passive action b, i.e.

$$\nu(\tilde{\pi}, \hat{x}) = \min\{t : t \geq 0 \text{ and } \tilde{\pi}\{X(t)\} = b\}.$$

It follows simply from the above that $\tilde{\pi}$ *must* choose passive action b at *all* decision epochs following $\nu(\tilde{\pi}, \hat{x})$. Making the identification of the passive action with retirement, it is now clear that optimal policies for the subsidy-W problem exactly coincide with optimal retirement policies when W is the retirement reward, with the passive action optimal for the former if and only if retirement is optimal for the latter. With this identification, the result follows immediately from Lemma 4.

To illustrate this correspondence, suppose, for example, that $G(a,x) > W(1-\beta)^{-1} > G(b,x)$. It follows from Lemma 4(c) that nonretirement in state (a,x) is optimal in this range for the retirement problem and, hence, that the active action a is optimal in state (a,x) for the subsidy-W problem. However, from Lemma 4(a) and Lemma 4(d) we see that retirement is optimal in state (b,x) and, hence, that the passive action b is optimal for the subsidy-W problem. This establishes Theorem 5(b). The other cases are dealt with similarly. This concludes the proof.

Theorem 6. (Indexability and indices.) *Bandit (Ω, P, R, S, β) is indexable. The Whittle index, $W : \{a,b\} \times \Omega \to \mathbb{R}$, is given by*

$$W(a,x) = (1-\beta)G(a,x), \quad W(b,x) = (1-\beta)G(b,x), \qquad x \in \Omega.$$

https://doi.org/10.1239/aap/1158684996 Published online by Cambridge University Press

Proof. Now write $\Pi(W)$ for the set of states in which it is optimal to take the passive action for the subsidy-W problem. By Theorem 5, we have

$$\Pi(W) = \{(a, x)\colon W(1-\beta)^{-1} \geq G(a, x)\} \cup \{(b, x)\colon W(1-\beta)^{-1} \geq G(b, x)\}. \tag{47}$$

Clearly $\Pi(W)$ is increasing in W. Furthermore, it follows from (47) that the Whittle index for state (a, x) is given by

$$W(a, x) = \inf\{W\colon (a, x) \in \Pi(W)\} = (1-\beta)G(a, x),$$

with a similar expression for the Whittle index of $W(b, x)$. This proves the theorem.

We now introduce a succession of model elaborations to the bandit (Ω, P, R, S, β) specified in I′–III′. All of these preserve indexability and the essential index structure.

4.1. Switching times

We suppose that the application of active action a in state (b, x) (namely, a switch from passive to active in state x) takes a random *switching time*, $\sigma(x)$, to accomplish. All such times are deemed independent both of each other and of the state evolution of the bandit. Here $\sigma(x)$ will be a positive, integer-valued random variable for all $x \in \Omega$. Equation (40) is now replaced by

$$\mathrm{P}\{X(t+1+\sigma(x)) = (a, y) \mid X(t) = (b, x), a\} = P(x, y), \qquad x, y \in \Omega,$$

and optimality equation (44) for the subsidy-W problem by

$$V((b, x), W) = \max\left\{-S(x) + \mathrm{E}[\beta^{\sigma(x)}]\left[R(x) + \beta \sum_{y\in\Omega} P(x, y)V((a, y), W)\right], \right.$$
$$\left. W(1-\beta)^{-1}\right\}, \qquad x \in \Omega.$$

Evolution equations (39) and (41) and the optimality equation (42) remain unchanged. With this modification, the Gittins index for activity $G(a, x)$ is unchanged for all $x \in \Omega$, but we now have

$$G(b, x) = \sup_{\nu}[\{-S(x) + \mathrm{E}[\beta^{\sigma(x)}]R(x, \nu)\}\{1 - \mathrm{E}[\beta^{\sigma(x)}]\,\mathrm{E}[\beta^{\nu} \mid x]\}^{-1}], \qquad x \in \Omega.$$

With these changes, Theorems 5 and 6 continue to hold with no major changes required in the proofs.

4.2. Tear-down costs

In the bandit (Ω, P, R, S, β) described in I′–III′ above, the cost $S(x)$ is incurred whenever processing is switched to the bandit when in state x in the underlying multiproject problem. It thus has the role of a *set-up cost* for the bandit. Suppose now that a *tear-down cost*, $S'(x)$, is also incurred whenever processing is switched away from the bandit in state x. Switching between bandits thus incurs a cost which comprises a tear-down cost for the bandit abandoned and a set-up cost for the bandit to which processing effort is transferred.

Suppose now that $X(0) = (b, x)$; that the active action is taken for the bandit throughout $[0, \nu)$, for some positive stopping time ν; and that the passive action is taken at ν. The resultant

https://doi.org/10.1239/aap/1158684996 Published online by Cambridge University Press

expected reward earned for this active period is now

$$\begin{aligned} &-S(x)+R(x,\nu)-\mathrm{E}[\beta^{\nu}S'(\bar{X}(\nu))\mid x]\\ &\quad=-\{S(x)+S'(x)\}+R(x,\nu)+S'(x)-\mathrm{E}[\beta^{\nu}S'(\bar{X}(\nu))\mid x]. \end{aligned} \tag{48}$$

It now follows from (48) that if, in the underlying multiproject scheduling problem, all constituent projects are deemed passive at time $t=0$, then the following modifications to bandit (Ω, P, R, S, β) are sufficient to accommodate tear-down costs:

- the switching cost $S(x)$ now becomes $S(x)+S'(x)$;
- the reward $R(x,y)$ is enhanced to $R(x,y)+S'(x)-\beta S'(y)$.

With these changes, the respective rewards earned under transitions (39)–(41) above are now $R(x,y)+S'(x)-\beta S'(y)$, $-S(x)+R(x,y)-\beta S'(y)$, and 0 for the bandit, and $R(x,y)+S'(x)-\beta S'(y)$, $-S(x)+R(x,y)-\beta S'(y)$, and W for the subsidy-W problem. Optimality equations (42) and (44) become

$$\begin{aligned} V((a,x),W)=\max\Bigg\{&R(x)+S'(x)+\beta\sum_{y\in\Omega}P(x,y)[-S'(y)+V((a,y),W)],\\ &W+\beta V((b,x),W)\Bigg\}, \qquad x\in\Omega,\\ V((b,x),W)=\max\Bigg\{&-S(x)+R(x)+\beta\sum_{y\in\Omega}P(x,y)[-S'(y)+V((a,y),W)],\\ &W(1-\beta)^{-1}\Bigg\}, \qquad x\in\Omega. \end{aligned}$$

Appropriate versions of the Gittins indices for activity are given by

$$\begin{aligned} G(a,x)&=\sup_{\nu}[(R(x,\nu)+S'(x)-\mathrm{E}[\beta^{\nu}S'(\bar{X}(\nu))\mid x])\{1-\mathrm{E}[\beta^{\nu}\mid x]\}^{-1}], \qquad x\in\Omega,\\ G(b,x)&=\sup_{\nu}[(-S(x)+R(x,\nu)-\mathrm{E}[\beta^{\nu}S'(\bar{X}(\nu))\mid x])\{1-\mathrm{E}[\beta^{\nu}\mid x]\}^{-1}], \qquad x\in\Omega. \end{aligned}$$

With these modifications, Theorems 5 and 6 continue to hold.

4.3. Losses from the system

We now consider models in which projects may be 'impatient' and can leave the system when not being processed. In order to accommodate this we enhance the state space Ω to include new state $*$ (denoting project absence) in which only the passive action is admissible. Evolution equations (41) now expand to

$$\begin{aligned} \mathrm{P}\{X(t+1)=*\mid X(t)=(a,x),b\}&=\mathrm{P}\{X(t+1)=*\mid X(t)=(b,x),b\}\\ &=\theta(x), \qquad x\in\Omega,\\ \mathrm{P}\{X(t+1)=(b,x)\mid X(t)=(a,x),b\}&=\mathrm{P}\{X(t+1)=(b,x)\mid X(t)=(b,x),b\}\\ &=1-\theta(x), \qquad x\in\Omega, \end{aligned}$$

https://doi.org/10.1239/aap/1158684996 Published online by Cambridge University Press

where $\theta(x)$ is the probability of loss of the project in state x under the passive action. With this modification, optimality equation (42) becomes

$$V((a,x),W) = \max\Bigg\{R(x) + \beta\sum_{y\in\Omega} P(x,y)V((a,x),W),$$
$$W + \beta\theta(x)V(*,W) + \beta\{1-\theta(x)\}V((b,x),W)\Bigg\}$$
$$= \max\Bigg\{R(x) + \beta\sum_{y\in\Omega} P(x,y)V((a,y),W),$$
$$W\{1-\beta+\beta\theta(x)\}(1-\beta)^{-1} + \beta\{1-\theta(x)\}V((b,x),W)\Bigg\}, \qquad x\in\Omega,$$

utilising the identity

$$V(*,W) = W(1-\beta)^{-1}.$$

Optimality equation (44) remains unchanged. The Gittins indices for activity remain as given in (45) and (46), and Theorems 5 and 6 continue to hold.

4.4. Numerical study

The second author is engaged on an extensive numerical study of the performance of the index policies developed through the above analysis. In Table 2 we present a summary of some numerical experiments aimed at assessing the quality of performance as the general level of switching costs varies. All problems studied are of restless bandit problems with $N = 4$ and $M = 1$, i.e. a single server choosing between four bandits. Each row of the table corresponds to a general level of switching cost, with this level increasing as one moves down the table. The columns of the table are subdivided to include three model types, labelled A, B, and C as follows:

- for A, each bandit has $|\Omega| = 3$;
- for B, we enhance case A to allow for a small rate of losses from the system under the passive action, as in Section 4.3;
- for C, each bandit has $|\Omega| = 4$.

Each entry in the table summarises the results of 150 problems studied. In all cases, active transition matrices were obtained by sampling entries independently from a U(0, 1) distribution and normalising across rows. Active rewards in each state were obtained by sampling independently from a U(200, 250) distribution. In case B, loss probabilities $\theta(\cdot)$ in each state were obtained by sampling from a U(0, 0.025) distribution. Once the active rewards have been chosen, the switching costs in each state were set to be a fixed percentage of the corresponding active reward. This percentage varies from 1% (top row) to 25% (bottom row). Hence, for each of three cases (A, B, and C) and twenty-five levels of switching cost (1% of active reward to 25%), 150 random problems were generated and studied. This makes a total of 11 250 problems studied. The discount rate, β, was set to be 0.9 throughout.

Entries in the table are obtained as follows: fix the case (A, B, or C) and the level of switching cost (1% to 25%). Generate 150 problems at random. For each problem compute the optimal expected reward, $V^{\text{opt.}}(\boldsymbol{x})$, and the expected reward from implementing the Whittle index heuristic, $V^{\text{ind.}}(\boldsymbol{x})$, for every possible initial state $\boldsymbol{x}$ in which all bandits are assumed passive

https://doi.org/10.1239/aap/1158684996 Published online by Cambridge University Press

TABLE 2: Level of performance of the Whittle index policy (percentage suboptimality) for a range of stochastic scheduling problems with state-dependent switching costs.

Switching cost (%)	Case *A*		Case *B*		Case *C*	
	Ave. (%)	Max. (%)	Ave. (%)	Max. (%)	Ave. (%)	Max. (%)
1	0.0040	0.0293	0.0097	0.1223	0.0062	0.0462
2	0.0106	0.0637	0.0125	0.1534	0.0153	0.0988
3	0.0150	0.0868	0.0142	0.1838	0.0223	0.1397
4	0.0163	0.0942	0.0147	0.1969	0.0246	0.1585
5	0.0159	0.0951	0.0151	0.2136	0.0255	0.1691
6	0.0162	0.0979	0.0150	0.2152	0.0258	0.1807
7	0.0164	0.0994	0.0144	0.2205	0.0240	0.1785
8	0.0145	0.0928	0.0123	0.1932	0.0199	0.1534
9	0.0099	0.0663	0.0089	0.1611	0.0183	0.1434
10	0.0058	0.0416	0.0056	0.1096	0.0114	0.1043
11	0.0039	0.0305	0.0037	0.0843	0.0089	0.0837
12	0.0036	0.0295	0.0029	0.0678	0.0064	0.0641
13	0.0032	0.0209	0.0027	0.0519	0.0042	0.0464
14	0.0023	0.0176	0.0019	0.0383	0.0027	0.0276
15	0.0009	0.0100	0.0011	0.0320	0.0014	0.0188
16	0.0009	0.0071	0.0011	0.0227	0.0001	0.0012
17	0.0011	0.0087	0.0009	0.0176	0.0003	0.0047
18	0.0003	0.0027	0.0004	0.0115	0.0002	0.0038
19	0.0000	0.0000	0.0001	0.0019	0.0000	0.0000
20	0.0000	0.0000	0.0000	0.0012	0.0000	0.0000
21	0.0000	0.0000	0.0000	0.0007	0.0000	0.0000
22	0.0000	0.0000	0.0000	0.0009	0.0000	0.0000
23	0.0000	0.0000	0.0000	0.0006	0.0000	0.0000
24	0.0000	0.0000	0.0000	0.0003	0.0000	0.0000
25	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

before time 0. The number of such states is 81 for cases *A* and *B* and 256 for case *C*. Calculate the percentage reward suboptimality, $100\{V^{\text{opt.}}(\boldsymbol{x}) - V^{\text{ind.}}(\boldsymbol{x})\}\{V^{\text{opt.}}(\boldsymbol{x})\}^{-1}$, for each state and record the maximum percentage suboptimality (Max.) and average percentage suboptimality (Ave.) over all states. Take averages of Max. and Ave. over the 150 problems generated to obtain entries in the table. Note that all computations of expected reward are performed using dynamic programming value iteration.

The reader should observe the outstandingly strong performance of the Whittle index policy throughout Table 2. Note that the percentage suboptimalities grow as one moves down the table, achieving a maximum when the switching costs are around 6% or 7% of the corresponding active rewards. From there the percentage suboptimalities decrease to 0% as the switching costs increase further. To understand this behaviour, note that when switching costs are 0%, the Whittle index policy becomes a Gittins index policy which in cases *A* and *C* is then optimal. As switching costs become large, optimal policies have the feature that each bandit is activated from passivity at most once over the decision horizon. The switching cost then essentially becomes a *set-up cost* for the bandit, to be incurred once only. When this happens, we can exploit Gittins' index theory to show that the Whittle index policy must be optimal in cases *A* and *C*. We should thus expect the percentage suboptimalities for cases *A* and *C* in Table 2 to

https://doi.org/10.1239/aap/1158684996 Published online by Cambridge University Press

decrease to 0% both as the switching cost decreases to 0% and when it becomes large. Case *B* with losses also follows this pattern.

5. Model 3: reward depletion and replenishment

Our final model class combines some of the features of models 1 and 2. Here, the switching of processing effort away from a bandit (job) has the effect that processing is lost and the bandit reverts to some initial 'unprocessed' state. A switching cost has to be paid before processing can then be resumed. An alternative interpretation/application of the same model is that the (active) bandit is a reward-generating mechanism which deteriorates with use. A period of passivity for the bandit allows for replenishment/recovery (at a cost) in the form of a return to the reinvigorated unprocessed state.

We use $\{\Omega, P, R, D, \beta\}$ to denote the bandit of interest. The general structure specified in 1–6 above is specialised to this case as follows.

I″. Countable set Ω is the state space of the bandit evolving under the active action, with $0' \in \Omega$ the designated *initial* or *return* state. We use '$*$' to denote a state in which the bandit is passive.

II″. If action a is taken at time t, then we have

$$\mathrm{P}\{X(t+1) = y \mid X(t) = x, a\} = P(x, y), \qquad x, y \in \Omega, \tag{49}$$

$$\mathrm{P}\{X(t+1) = y \mid X(t) = *, a\} = P(0', y), \qquad y \in \Omega. \tag{50}$$

If action b is taken at time t, then we have

$$\mathrm{P}\{X(t+1) = * \mid X(t) = x, b\} = \mathrm{P}\{X(t+1) = * \mid X(t) = *, b\} = 1, \qquad x \in \Omega. \tag{51}$$

Note from (50) that, when the active action is applied to a previously passive bandit, an instantaneous return to initial state $0'$ is followed by a single transition.

III″. The respective rewards earned by the bandit under transitions (49) and (50) are $R(x, y)$ and $-D + R(0', y)$. Here D is a replenishment/switching cost. No rewards are earned under transitions (51). The corresponding rewards for the subsidy-W problem are respectively $R(x, y)$, $-D + R(0', y)$, and W.

The value function $V(\cdot, W)$ for the subsidy-W problem satisfies the *optimality equations*

$$V(x, W) = \max\left\{R(x) + \beta \sum_{y \in \Omega} P(x, y) V(y, W), W + \beta V(*, W)\right\}, \qquad x \in \Omega,$$

$$V(*, W) = \max\left\{-D + R(0') + \beta \sum_{y \in \Omega} P(0', y) V(y, W), W(1-\beta)^{-1}\right\}.$$

Consider a set-up in which $X(0) = 0'$ and action a is taken at time 0 with an optimal (reward-maximising) policy pursued thereafter. Since we restrict to stationary policies, it follows that

$$\tau^* := \min\{t : t \geq 1 \text{ and it is optimal to take action } b\}$$

is a stationary stopping time. Write $\Theta(0', W)$ for the corresponding expected discounted reward earned over an infinite horizon for the subsidy-W problem. From I″–III″ above we have

$$\Theta(0', W) = R(0', \tau^*) + \mathrm{E}[\beta^{\tau^*} \mid 0'] \max\{W(1-\beta)^{-1}, W - \beta D + \beta \Theta(0', W)\}. \tag{52}$$

https://doi.org/10.1239/aap/1158684996 Published online by Cambridge University Press

The first term within the maximisation in the last term on the right-hand side of (52) will achieve the maximum whenever the passive action is optimal in passive state $*$. The second term achieves the maximum whenever the active action is optimal in state $*$.

Lemma 5. *If passive action b is optimal for state $*$ for the subsidy-$\bar{W}$ problem then it is also optimal for state $*$ for the subsidy-W problem, with $W \geq \bar{W}$.*

Proof. Action b is optimal for state $*$ for the subsidy-$\bar{W}$ problem if and only if

$$\bar{W}(1-\beta)^{-1} \geq \bar{W} - \beta D + \beta\Theta(0', \bar{W}) \tag{53}$$

or, equivalently,

$$\bar{W}(1-\beta)^{-1} \geq -D + \Theta(0', \bar{W}). \tag{54}$$

It must further hold that, for any stationary, positive-valued stopping time τ,

$$\begin{aligned}\bar{W}(1-\beta)^{-1} &\geq -D + R(0', \tau) + \mathrm{E}[\beta^{\tau} \mid 0']\max\{\bar{W}(1-\beta)^{-1}, \bar{W} - \beta D + \beta\Theta(0', \bar{W})\} \\ &= -D + R(0', \tau) + \mathrm{E}[\beta^{\tau} \mid 0']\bar{W}(1-\beta)^{-1},\end{aligned} \tag{55}$$

where the inequality in (55) follows from (53). We re-express (55) as

$$R(0', \tau) - D \leq \bar{W}(1-\beta)^{-1}\{1 - \mathrm{E}[\beta^{\tau} \mid 0']\} \tag{56}$$

for all τ.

We suppose now that there exists some $W > \bar{W}$ for which action a is strictly optimal in state $*$ for the subsidy-W problem, and obtain a contradiction. From (54) it must follow that

$$\Theta(0', W) > D + W(1-\beta)^{-1}. \tag{57}$$

From (52) and (57) we infer that

$$\Theta(0', W) = \{R(0', \tau^*) + \mathrm{E}[\beta^{\tau^*} \mid 0'](W - \beta D)\}\{1 - \beta\,\mathrm{E}[\beta^{\tau^*} \mid 0']\}^{-1} \tag{58}$$

for the optimal τ^*. However, from (56) we have

$$R(0', \tau) < W(1-\beta)^{-1}\{1 - \mathrm{E}[\beta^{\tau^*} \mid 0']\} + D,$$

which, when combined with (58), yields

$$\Theta(0', W) < D + W(1-\beta)^{-1},$$

which contradicts (57), as required. This concludes the proof.

In the statement of Theorem 7, the *Gittins index* for state x is given by

$$G(x) = \sup_{\tau}[R(x, \tau)\{1 - \mathrm{E}[\beta^{\tau} \mid x]\}^{-1}], \qquad x \in \Omega,$$

and coincides with the Gittins index for activity, defined in (45), for model 2.

Theorem 7. (Indexability and indices.) *Bandit $\{\Omega, P, R, D, \beta\}$ is indexable. The Whittle index for state x is given by*

$$W(x) = \min\{(1-\beta)G(x), \tilde{W}(x)\},$$

https://doi.org/10.1239/aap/1158684996 Published online by Cambridge University Press

where $G(x)$ is the Gittins index for x and $\tilde{W}(x)$ is the unique W-solution to the equation

$$W - \beta D + \beta\Theta(0', W) = G(x).$$

The Whittle index for state $*$ is the unique W-solution to the equation

$$W(1-\beta)^{-1} = -D + \Theta(0', W). \tag{59}$$

Proof. A modest development of the analysis of Lemma 5 yields the conclusion that there does indeed exist a unique solution to (59). Call this solution $\tilde{W}$. Hence, we have

$$\Psi(W) := \max\{W(1-\beta)^{-1}, W - \beta D + \beta\Theta(0', W)\} = \begin{cases} W - \beta D + \beta\Theta(0', W), & W \le \tilde{W}, \\ W(1-\beta)^{-1}, & W \ge \tilde{W}. \end{cases}$$

Let $X(0) = x \in \Omega$. Action b is optimal for the subsidy-W problem at time $t = 0$ in state x if and only if $\Psi(W)$ is no less than the expected reward from among those policies which choose action a at time $t = 0$. This will happen if and only if

$$\Psi(W) \ge R(x, \tau) + \mathrm{E}[\beta^{\tau} \mid x]\Psi(W) \tag{60}$$

for all stationary, positive-valued stopping times τ. Inequality (60) is achieved precisely when

$$\begin{aligned} \Psi(W) &\ge \sup_{\tau}[R(x, \tau)\{1 - \mathrm{E}[\beta^{\tau} \mid x]\}^{-1}] \\ &= G(x), \end{aligned}$$

the Gittins index for x. Furthermore, action b is optimal in state $*$ exactly when

$$\Psi(W) = W(1-\beta)^{-1}.$$

If we now use $\Pi(W)$ to denote the set of states in which the passive action is optimal for the subsidy-W problem, then from the above analysis we have

$$\Pi(W) = \begin{cases} \{x\colon G(x) \le \Psi(W)\}, & W < \tilde{W}, \\ \{x\colon G(x) \le \Psi(W)\} \cup \{*\}, & W \ge \tilde{W}. \end{cases} \tag{61}$$

From (61) and the evident fact that $\Psi(W)$ is strictly increasing in W, we conclude that $\Pi(W)$ is increasing in W and, hence, that the bandit $\{\Omega, P, R, D, \beta\}$ is indexable. That $\tilde{W}$ is the Whittle index for $*$ follows immediately from (61). We further see that the Whittle index for state $x \in \Omega$ is the unique W-solution to the equation $\Psi(W) = G(x)$. Call this solution $W(x)$. If $W(x) \le \tilde{W}$ then

$$G(x) = W(x) - \beta D + \beta\Theta\{0', W(x)\} \quad \Longrightarrow \quad W(x) = \tilde{W}(x) \le (1-\beta)G(x).$$

If $W(x) \ge \tilde{W}$ then

$$G(x) = W(x)(1-\beta)^{-1} \quad \Longrightarrow \quad W(x) = (1-\beta)G(x) \le \tilde{W}(x).$$

Hence, the Whittle index is as specified in the statement of the theorem. This concludes the proof.

https://doi.org/10.1239/aap/1158684996 Published online by Cambridge University Press

The next result modifies Theorem 7 in a way which assists computation.

Theorem 8. (Alternative index specification.) *The Whittle index for state x is given by*

$$W(x) = \min\{(1-\beta)G(x), \hat{W}(x)\},$$

where $G(x)$ is the Gittins index for x and $\hat{W}(x)$ is the unique W-solution to the equation $\Xi(0', W) = G(x)$. Here

$$\Xi(0', W) = \sup_{\tau}[(W - \beta D + \beta R(0', \tau))\{1 - \beta\,\mathrm{E}[\beta^{\tau} \mid 0']\}^{-1}],$$

with the supremum taken over the set of positive-valued stopping times. The Whittle index for state $*$ is the unique W-solution to the equation $\Xi(0', W) = W(1-\beta)^{-1}$.

Proof. It is straightforward to establish that

$$\Xi(0', W) = W - \beta D + \Theta(0', W), \qquad W \leq \tilde{W},$$

and, hence, that

$$\Psi(W) = \max\{W(1-\beta)^{-1}, \Xi(0', W)\}, \qquad W \leq \tilde{W}. \tag{62}$$

Furthermore,

$$\Xi(0', W) \leq W - \beta D + \Theta(0', W), \qquad W > \tilde{W},$$

and, hence,

$$\Psi(W) = \max\{W(1-\beta)^{-1}, \Xi(0', W)\}, \qquad W > \tilde{W}. \tag{63}$$

The result now follows from (61), (62), and (63) and a suitable modification of the argument in the proof of Theorem 7.

Remark 5. The advantage of Theorem 8 is that the Gittins index-like nature of $\Xi(0', W)$ means that specifications of $\hat{W}(x)$ of the sort given in Theorems 3 and 4 for model 1 are now possible. Since the details are akin to those for model 1, we omit them here. We then recover the Whittle index as $W(x) = \min\{(1-\beta)G(x), \hat{W}(x)\}$.

Acknowledgements

The first and third authors acknowledge the support of the Engineering and Physical Sciences Research Council through the award of grant GR/S45188/01. The second author acknowledges the support of the Servei de Recerca of Universitat Pompeu Fabra through the award of grant EBES-REI2645. All authors acknowledge with gratitude the comments of an anonymous referee and also the contributions of John Gittins and José Niño-Mora to their thinking about bandit problems.

References

Agrawal, R., Hedge, M. and Teneketzis, D. (1988). Asymptotically efficient adaptive allocation rules for the multiarmed bandit problem with switching cost. *IEEE Trans. Automatic Control* **33,** 899–906.

Ansell, P. S., Glazebrook, K. D., Niño-Mora, J. and O'Keeffe, M. (2003). Whittle's index policy for a multi-class queueing system with convex holding costs. *Math. Meth. Operat. Res.* **57,** 21–39.

Asawa, M. and Teneketzis, D. (1996). Multi-armed bandits with switching penalties. *IEEE Trans. Automatic Control* **41,** 328–348.

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

GITTINS, J. C. (1979). Bandit processes and dynamic allocation indices (with discussion). *J. R. Statist. Soc. B* **41,** 148–177.

GITTINS, J. C. (1989). *Multi-Armed Bandit Allocation Indices*. John Wiley, Chichester.

GLAZEBROOK, K. D. (1980). On stochastic scheduling with precedence relations and switching costs. *J. Appl. Prob.* **17,** 1016–1024.

GLAZEBROOK, K. D., MITCHELL, H. M. AND ANSELL, P. S. (2005). Index policies for the maintenance of a collection of machines by a set of repairmen. *Europ. J. Operat. Res.* **165,** 267–284.

GLAZEBROOK, K. D., NIÑO-MORA, J. AND ANSELL, P. S. (2002). Index policies for a class of discounted restless bandits. *Adv. Appl. Prob.* **34,** 754–774.

NASH, P. (1979). Optimal allocation of resources between research projects. Doctoral Thesis, University of Cambridge.

NIÑO-MORA, J. (2001). Restless bandits, partial conservation laws and indexability. *Adv. Appl. Prob.* **33,** 76–98.

NIÑO-MORA, J. (2002). Dynamic allocation indices for restless projects and queueing admission control: a polyhedral approach. *Math. Program.* **93,** 361–413.

PAPADIMITRIOU, C. H. AND TSITSIKLIS, J. N. (1999). The complexity of optimal queuing network control. *Math. Operat. Res.* **24,** 293–305.

PUTERMAN, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley, New York.

REIMAN, M. I. AND WEIN, L. M. (1998). Dynamic scheduling of a two-class queue with setups. *Operat. Res.* **46,** 532–547.

VAN OYEN, M. P. AND TENEKETZIS, D. (1994). Optimal stochastic scheduling of forest networks with switching penalties. *Adv. Appl. Prob.* **26,** 474–479.

WEBER, R. R. AND WEISS, G. (1990). On an index policy for restless bandits. *J. Appl. Prob.* **27,** 637–648. (Addendum: *Adv. Appl. Prob.* **23** (1991), 429–430.)

WHITTLE, P. (1980). Multi-armed bandits and the Gittins index. *J. R. Statist. Soc. B* **42,** 143–149.

WHITTLE, P. (1988). Restless bandits: activity allocation in a changing world. In *A Celebration of Applied Probability* (J. Appl. Prob. Spec. Vol. **25A**), ed. J. Gani, Applied Probability Trust, Sheffield, pp. 287–298.

WHITTLE, P. (1996). *Optimal Control: Basics and Beyond*. John Wiley, Chichester.