MATHEMATICS OF OPERATIONS RESEARCH
Vol. 34, No. 1, February 2009, pp. 26–44
ISSN 0364-765X | EISSN 1526-5471 | 09 | 3401 | 0026

DOI 10.1287/moor.1080.0342

A Generalized Gittins Index for a Class of Multiarmed Bandits with General Resource Requirements

K. D. Glazebrook, R. Minty
Department of Mathematics and Statistics, Lancaster University, Lancaster, LA1 4YF, United Kingdom
{k.glazebrook@lancaster.ac.uk, r.minty@lancaster.ac.uk}

We generalise classical multiarmed bandits to allow for the distribution of a (fixed amount of a) divisible resource among the constituent bandits at each decision point. Bandit activation consumes amounts of the available resource, which may vary by bandit and state. Any collection of bandits may be activated at any decision epoch, provided they do not consume more resource than is available. We propose suitable bandit indices that reduce to those proposed by Gittins [Gittins, J. C. 1979. Bandit processes and dynamic allocation indices (with discussion). *J. R. Statist. Soc.* **B41** 148–177] for the classical models. The index that emerges is an elegant generalization of the Gittins index, which measures in a natural way the reward earnable from a bandit per unit of resource consumed. The paper discusses both how such indices may be computed and how they may be used to construct heuristics for resource distribution. We also describe how to develop bounds on the closeness to optimality of index heuristics and demonstrate a form of asymptotic optimality for a greedy index heuristic in a class of simple models. A numerical study testifies to the strong performance of a weighted index heuristic.

Key words: asymptotic optimality; bandit problems; dynamic programming; Gittins index; resource allocation
MSC2000 subject classification: Primary: 90C40; secondary: 68M20, 90B36
OR/MS subject classification: Primary: dynamic programming/optimal control, Markov; secondary: production/scheduling, sequencing/stochastic
History: Received January 31, 2007; revised January 15, 2008 and May 3, 2008. Published online in *Articles in Advance* January 27, 2009.

1. Introduction. The paper concerns developments of the classical work of Gittins [8, 9], who elucidated index-based solutions to a family of Markov decision processes (MDPs) called *multiarmed bandit problems* (MABs). This work concerned the optimal allocation of a single indivisible resource among a collection of stochastic projects (or *bandits* as they are sometimes called). Gittins' original contribution—namely, to demonstrate that optimal project choices are those of largest index—has given rise to a substantial literature describing a range of extensions to, and reformulations of, his result. Gittins [9] gives an extensive bibliography of early contributions. A more recent survey is due to Mahajan and Teneketzis [18].

An important limitation on the modeling power of Gittins' MABs is the critical assumption that bandits are frozen while not in receipt of resource. In response to this, Whittle [30] introduced a class of *restless bandit problems* (RBPs) in which projects may change state whether active or passive, although according to different dynamics. In contrast to Gittins' MABs, RBPs are almost certainly intractable and have been shown to be PSPACE-hard by Papadimitriou and Tsitsiklis [24]. Whittle used a Lagrangian relaxation to develop an index heuristic (which reduces to Gittins' index policy in the MAB case) for those RBPs that pass an *indexability test*. Weber and Weiss [29] demonstrated a form of asymptotic optimality for Whittle's index policy, but otherwise the model's perceived difficulty inhibited further substantial progress for some time. More recently, Niño-Mora [20, 21] and Glazebrook et al. [16] have explored indexability issues from a polyhedral perspective. Further, a range of empirical studies have demonstrated the power and practicability of Whittle's approach in a range of application contexts. See, for example, Ansell et al. [2], Opp et al. [23], Glazebrook et al. [14, 15], and Glazebrook and Kirkbride [12, 13].

A significant limitation on the applicability of the above classical models is the very simple view they take of the resource to be allocated. In Gittins' MABs, a single indivisible resource is to be allocated *en bloc* to a single bandit at each decision epoch. Whittle's RBP formulation contemplates *parallel server* versions of this. Although such views of the resource may sometimes be appropriate, many applications concern a *divisible* resource (for example, money or manpower) in situations where its overconcentration would usually be far from optimal. Indeed, this was an issue for Gittins in his work on the planning of new product pharmaceutical research that provided the motivation for his pioneering contribution. See, for example, Gittins [9, Chapter 5]. Stimulated by such considerations, we develop in §2 a family of MABs with switching penalties that extend those of Glazebrook et al. [15]. In these models, a fixed amount of some divisible resource is available for distribution among a collection of projects at each decision epoch. Activation of a bandit consumes a quantity of resource which is in general both bandit and state dependent. Any collection of bandits may be activated at any decision epoch, provided they do not consume more resource than is available. The presence of switching penalties induces a simple form of restlessness in the models. Plainly, the development of good policies must now take serious account not only of the rewards that the bandits are capable of securing, but also of the amount of resource they will consume in the process. We note that the problem of developing optimal policies for MABs with switching penalties is a famously intractable one to which several important contributions have been made. See, for example, Glazebrook [10], Agrawal et al. [1], Van Oyen and Teneketzis [28], Banks and Sundaram [3], and Reiman and Wein [26]. The version of this problem with general resource requirements discussed in this paper is, to the authors' knowledge, new.

In §2, notions of bandit indexability are developed by an approach that suitably extends that of Whittle [30]. Bandit indices are defined and proposals are advanced for how the indices may be used to construct policies for our class of MABs. Section 3 contains a formal proof of indexability for our model class under mild conditions. The index that results is an elegant generalization of the Gittins index, which measures in a natural way the reward earnable from a bandit per unit of resource consumed. Succeeding sections discuss how these indices are computed (§4) and how to analyse the closeness to optimality of index-based heuristics (§5). A form of asymptotic optimality is established for a greedy index heuristic in a class of simple models. The paper concludes with a numerical study (§6).

2. Index heuristics for a family of multiarmed bandits with varying resource requirements. We shall consider a *multiarmed bandit* (MAB) model with switching costs and general resource requirements, as outlined in the introduction. More specifically, we focus on a family of *reward-discounted* Markov decision processes that model a situation in which a decision maker chooses which of N *bandits* to make active at each decision epoch $t \in \mathbb{N}$, within resource constraints. The details are as follows:

(i) At each time $t \in \mathbb{N}$ an *action* $\boldsymbol{\alpha}(t) = (\alpha_1(t), \alpha_2(t), \ldots, \alpha_N(t))$ is applied to the process with $\alpha_n(t) \in \{a, b\}$ the action applied to bandit n. The action a is *active* and calls for the positive commitment of resource. The alternative action b is *passive*.

(ii) The *state space* of the process is $\times_{n=1}^{N}[\{a, b\} \times \Omega_n]$ with Ω_n the (finite or countable) state space of bandit n. The *state* of the process is observed at each time $t \in \mathbb{N}$. At time t, we write $\mathbf{X}(t) = (\mathbf{X}_1(t), \mathbf{X}_2(t), \ldots, \mathbf{X}_N(t))$ for the process state with $\mathbf{X}_n(t) \in \{a, b\} \times \Omega_n$ the *extended state of bandit* n. The equation $\mathbf{X}_n(t) = (\bullet, x)$, $\bullet \in \{a, b\}$, $x \in \Omega_n$, means that the *state* of bandit n at time t is x and that action $\bullet$ was applied to the bandit at time $t-1$, $t \geq 1$. When needed, we use $X_n(t)$ for the (nonextended) bandit state, i.e., $\mathbf{X}_n(t) = (\bullet, x) \implies X_n(t) = x$.

(iii) State evolution is Markov, with the N bandits evolving independently under any choice of actions. That is, we have, for all choices of the quantities concerned:

$$\begin{aligned} P\big\{\mathbf{X}(t+1) = \mathbf{y} \mid \mathbf{X}(t) = \mathbf{x}, \mathbf{X}(t-1) = \mathbf{x}_{t-1}, \ldots, \mathbf{X}(0) = \mathbf{x}_0, \\ \boldsymbol{\alpha}(t) = \boldsymbol{\alpha}, \boldsymbol{\alpha}(t-1) = \boldsymbol{\alpha}_{t-1}, \ldots, \boldsymbol{\alpha}(0) = \boldsymbol{\alpha}_0\big\} \\ = P\big\{\mathbf{X}(t+1) = \mathbf{y} \mid \mathbf{X}(t) = \mathbf{x}, \boldsymbol{\alpha}(t) = \boldsymbol{\alpha}\big\} \\ = \prod_{n=1}^{N} P\big\{\mathbf{X}_n(t+1) = \mathbf{y}_n \mid \mathbf{X}_n(t) = \mathbf{x}_n, \alpha_n\big\}. \end{aligned} \tag{1}$$

The assumption of independence in (1) is commonplace in Gittins index theory (see, for example, Gittins [9]), but contrasts with many research contributions that deal with dependent bandits. For some examples of the latter, the reader is referred to Berry and Fristedt [4] and to references contained within.

We shall impose further structure upon (1) by introducing an assumption that changes to (nonextended) bandit states do not occur under the passive action. We thus have, for all choices of the quantities concerned,

$$P\big\{\mathbf{X}_n(t+1) = (b, x) \mid \mathbf{X}_n(t) = (\bullet, x), b\big\} = 1. \tag{2}$$

We shall also suppose that for bandit n, state transitions under the active action are determined by the stationary Markov law P_n. We write:

$$P\big\{\mathbf{X}_n(t+1) = (a, y) \mid \mathbf{X}_n(t) = (\bullet, x), a\big\} = P_n(x, y). \tag{3}$$

(iv) The *reward functions* r_n^a, r_n^b, S_{nr}^a, S_{nr}^b all map the state space Ω_n into the nonnegative reals $\mathbb{R}^+$ and are bounded. The expected reward earned by bandit n for the transition in (3) is $r_n^a(x)$ when $\bullet = a$ and is $r_n^a(x) - S_{nr}^a(x)$ when $\bullet = b$. Hence, $S_{nr}^a(x)$ is a cost incurred whenever bandit n is switched on (goes from passive to active) in state x. The expected reward earned by the transition in (2) is $r_n^b(x) - S_{nr}^b(x)$ when $\bullet = a$ and is $r_n^b(x)$

when $\bullet = b$. Hence, $S_{nr}^b(x)$ is a cost incurred when bandit n is switched off in state x. Although it is natural to assume that expected rewards earned under the active action a uniformly exceed those earned under the passive action b, namely

$$r_n^a(x) \geq r_n^b(x), \quad x \in \Omega_n, \ \ 1 \leq n \leq N,$$

we do not impose this as a general requirement.

Should action $\boldsymbol{\alpha}$ be applied to the system when in state $\mathbf{x}$, the expected reward earned, denoted $r(\boldsymbol{\alpha}, \mathbf{x})$, is the sum of the rewards earned by the individual bandits. We write

$$\begin{aligned} r(\boldsymbol{\alpha}, \mathbf{x}) = &\sum_{n:\,\alpha_n = a} \big[r_n^a(x_n) I(\mathbf{x}_n \in \{a\} \times \Omega_n) + \{r_n^a(x_n) - S_{nr}^a(x_n)\} I(\mathbf{x}_n \in \{b\} \times \Omega_n)\big] \\ &+ \sum_{n:\,\alpha_n = b} \big[\{r_n^b(x_n) - S_{nr}^b(x_n)\} I(\mathbf{x}_n \in \{a\} \times \Omega_n) + r_n^b(x_n) I(\mathbf{x}_n \in \{b\} \times \Omega_n)\big], \end{aligned}$$

where I is an indicator. All rewards and costs are discounted according to rate $\beta \in (0, 1)$.

(v) The *consumption functions* c_n^a, c_n^b, S_{nc}^a, S_{nc}^b all map the state space Ω_n into the nonnegative reals $\mathbb{R}^+$ and are bounded. The resource consumed by bandit n when action a is taken in enhanced state $(\bullet, n)$ is $c_n^a(x)$ when $\bullet = a$ and is $c_n^a(x) + S_{nc}^a(x)$ when $\bullet = b$. When action b is applied to bandit n in enhanced state $(\bullet, x)$, the resource consumed is $c_n^b(x) + S_{nc}^b(x)$ when $\bullet = a$ and $c_n^b(x)$ when $\bullet = b$. Hence, additional resource may be consumed when bandit n is switched on or off. We shall assume that the amount of resource consumed under the active action a uniformly exceeds that under the passive action b, that is,

$$c_n^a(x) \geq c_n^b(x) + S_{nc}^b(x), \quad x \in \Omega_n, \ \ 1 \leq n \leq N. \tag{4}$$

Should action $\boldsymbol{\alpha}$ be applied to the system in state $\mathbf{x}$, the total resource consumed, denoted $c(\boldsymbol{\alpha}, \mathbf{x})$, is the sum of the amounts of resource consumed by the individual bandits. We write:

$$\begin{aligned} c(\boldsymbol{\alpha}, \mathbf{x}) = &\sum_{n:\,\alpha_n = a} \big[c_n^a(x_n) I(\mathbf{x}_n \in \{a\} \times \Omega_n) + \{c_n^a(x_n) + S_{nc}^a(x_n)\} I(\mathbf{x}_n \in \{b\} \times \Omega_n)\big] \\ &+ \sum_{n:\,\alpha_n = b} \big[\{c_n^b(x_n) + S_{nc}^b(x_n)\} I(\mathbf{x}_n \in \{a\} \times \Omega_n) + c_n^b(x_n) I(\mathbf{x}_n \in \{b\} \times \Omega_n)\big]. \end{aligned}$$

The set of *admissible actions* in system state $\mathbf{x}$ is given by $A(\mathbf{x}) = \{\boldsymbol{\alpha};\, c(\boldsymbol{\alpha}, \mathbf{x}) \leq C\}$, where C is the total resource available at each decision epoch. We shall suppose that $A(\mathbf{x}) \neq \varnothing$, $\mathbf{x} \in \times_{n=1}^N [\{a, b\} \times \Omega_n]$. We also require that $\exists \mathbf{x} \in \times_{n=1}^N [\{a, b\} \times \Omega_n]$ for which $A(\mathbf{x}) \neq \{a, b\}^N$, namely, that resource availability strictly constrains project activation.

(vi) A *policy*, respectively, an admissible policy, is a rule u for choosing an action, respectively, an admissible action, at each decision epoch. Such a rule can in principle be a function of the entire history of the process to date. We shall seek an admissible policy to maximise the total expected reward earned over an infinite horizon. Standard theory (see, for example, Puterman [25]) asserts the existence of an optimal policy that is *stationary* (makes decisions in light of the current process state only) and which satisfies the optimality equations of dynamic programming (DP). That said, a pure DP approach to the above problem is unlikely to yield insight and will be computationally intractable for problems of realistic size. Hence the primary quest is for strongly performing *heuristic policies*.

Following the classic work of Gittins [8, 9] and Whittle [30], we shall seek heuristics in the form of *index policies*. Hence, we shall seek *calibrating functions* ν_n: $\{a, b\} \times \Omega_n \to \mathbb{R}$, $1 \leq n \leq N$, one for each bandit, that will guide the construction of good actions in each system state. To develop such indices, we shall seek a decomposition of (a relaxation of) our optimization problem into N individual problems, one for each bandit. It is these individual problems that, when suitably structured, will yield the calibrating functions required.

We proceed as follows: We state the optimization problem of interest, expressed in (i)–(v) above, as

$$R^{\text{opt}}(\mathbf{x}) = \sup_{u \in \mathcal{U}} \left\{ \sum_{n=1}^N R_n^u(\mathbf{x}) \right\}. \tag{5}$$

In (5), $\mathcal{U}$ is the class of stationary admissible policies for the problem, and $R_n^u(\mathbf{x})$ is the total expected reward yielded by bandit n under policy u from initial state $\mathbf{x}$, with successive rewards discounted at rate β. We relax (5) as follows: We extend to the class $\mathcal{U}'$ of stationary policies that are allowed a free choice of actions from $\{a, b\}^N$ in every state, and then for any $\nu \in \mathbb{R}^+$ we write

$$R^{\text{opt}}(\mathbf{x}, \nu) = \sup_{u \in \mathcal{U}'} \left\{ \sum_{n=1}^{N} [R_n^u(\mathbf{x}) - \nu C_n^u(\mathbf{x})] \right\}. \tag{6}$$

In (6), $C_n^u(\mathbf{x})$ is the total expected resource consumed by bandit n under policy u from initial state $\mathbf{x}$, with successive consumptions discounted at rate β. Explicitly, we have

$$C_n^u(\mathbf{x}) = E_u\left[\sum_{t=0}^{\infty} \beta^t c_n^u(\mathbf{X}(t)) \,\middle|\, \mathbf{X}(0) = \mathbf{x} \right], \quad \mathbf{x} \in \mathsf{X}_{n=1}^{N}[\{a, b\} \times \Omega_n], \tag{7}$$

where

$$\begin{aligned} c_n^u(\mathbf{x}) = {} & \big[c_n^a(x_n) I(\mathbf{x}_n \in \{a\} \times \Omega) + \{c_n^a(x_n) + S_{nc}^a(x_n)\} I(\mathbf{x}_n \in \{b\} \times \Omega_n)\big] I[u(\mathbf{x}) \in A_n] \\ & + \big[\{c_n^b(x_n) + S_{nc}^b(x_n)\} I(\mathbf{x}_n \in \{a\} \times \Omega) + c_n^b(x_n) I(\mathbf{x}_n \in \{b\} \times \Omega_n)\big] I[u(\mathbf{x}) \in B_n], \end{aligned} \tag{8}$$

and where, in (7), E_u denotes an expectation taken over realizations of the system evolving under stationary policy u; whereas in (8) A_n (respectively, B_n), is the subset of $\{a, b\}^N$ consisting of actions whose nth component is a (respectively, b).

In (6) the multiplier ν has an economic interpretation as a *charge* (sometimes called the *prevailing charge*) imposed per unit of resource consumed; then, for each bandit n, the difference $R_n^u(\mathbf{x}) - \nu C_n^u(\mathbf{x})$ is the total expected reward earned by n less charges imposed for resource consumption under policy u from initial state $\mathbf{x}$. In (6) we seek to maximise the aggregate total over all bandits. To motivate relaxation (6), observe that $u \in \mathcal{U}'$ is optimal for (6) if and only if it is optimal for

$$\bar{R}^{\text{opt}}(\mathbf{x}, \nu) := \sup_{u \in \mathcal{U}'} \left\{ \sum_{n=1}^{N} [R_n^u(\mathbf{x}) - \nu C_n^u(\mathbf{x})] \right\} + \nu C(1-\beta)^{-1}. \tag{9}$$

However, for any prevailing charge $\nu \in \mathbb{R}^+$ and any admissible policy $u \in \mathcal{U}$ we have

$$\sum_{n=1}^{N} C_n^u(\mathbf{x}) \leq C(1-\beta)^{-1} \tag{10}$$

and so

$$\bar{R}^{\text{opt}}(\mathbf{x}, \nu) \geq R^{\text{opt}}(\mathbf{x}). \tag{11}$$

Now, to optimize (6) and (9) it is sufficient to find an optimal policy for each individual bandit. Both of these problems decompose into N parallel problems. To express this decomposition we write:

$$R^{\text{opt}}(\mathbf{x}, \nu) = \sum_{n=1}^{N} R_n^{\text{opt}}(\mathbf{x}_n, \nu) = \bar{R}^{\text{opt}}(\mathbf{x}, \nu) - \nu C(1-\beta)^{-1}. \tag{12}$$

That is, for bandit n and prevailing charge ν, $R_n^{\text{opt}}(\mathbf{x}_n, \nu)$ is the value of the problem that optimizes $R_n(\mathbf{x}_n) - \nu C_n(\mathbf{x}_n)$ over all policies for bandit n with initial state $\mathbf{x}_n$ that choose from the action space $\{a, b\}$ in a stationary way. We call this problem $p_n(\nu)$. Plainly, from (12), an optimal policy for (6) and (9) simply runs optimal policies for the $p_n(\nu)$, $1 \leq n \leq N$, in parallel.

We now follow Whittle [30] in requiring structure in optimal policies for each $p_n(\nu)$ that will permit development of an appropriate *calibrating function* ν_n: $\{a, b\} \times \Omega_n \to \mathbb{R}^+$ for bandit n. In order to do this, we write $b_n(u_n)$ for the *passive set* corresponding to policy u_n, which is stationary for $p_n(\nu)$, namely

$$b_n(u_n) := \{\mathbf{x}_n \in \{a, b\} \times \Omega_n;\ u_n(\mathbf{x}_n) = b\}.$$

Definition 2.1 describes both the structure required and the calibrating function that results from it.

Definition 2.1. Bandit n is *indexable* if there exists a family of policies $\{u_n(\nu), \nu \in \mathbb{R}^+\}$ such that

(I) $u_n(\nu)$ is optimal for $p_n(\nu)$; and

(II) $b_n\{u_n(\nu)\}$ is nondecreasing in ν.

Under (I) and (II), the corresponding index for bandit n, ν_n: $\{a, b\} \times \Omega_n \to \mathbb{R}^+$ is given by

$$\nu_n(\mathbf{x}_n) = \inf[\nu \in \mathbb{R}^+; \mathbf{x}_n \in b_n\{u_n(\nu)\}]. \tag{13}$$

From (13), $\nu_n(\mathbf{x}_n)$ has an interpretation as a *fair charge* (per unit of resource) for the additional resource consumed in going from passive (b) to active (a) when bandit n is in state $\mathbf{x}_n$. For any prevailing charge ν, any indexable bandit n, and any state $\mathbf{x}_n$, if the prevailing charge ν exceeds or equals the fair charge $\nu_n(\mathbf{x}_n)$, then in the problem $p_n(\nu)$ the passive action is optimal; conversely, if $\nu \leq \nu_n(\mathbf{x}_n)$, the active action is optimal. In both cases we have strict optimality whenever the inequalities are strict. The index of Definition 2.1 generalises that of Gittins [8] to our more complex resource consumption setup. If we take all the c_n^a to be identically one and all the c_n^b, r_n^b, S_{nc}^a, S_{nc}^b, S_{nr}^a, S_{nr}^b to be identically zero, then all bandits are known to be indexable and (13) is the *Gittins index* of bandit n. We shall show in the following sections that for our family of MABs with varying resource requirements described in (i)–(vi) above, bandits are indeed indexable under mild conditions and that indices are easily computed.

Remark. We refer to our index as a *generalized Gittins index* in the sense of the preceding paragraph—namely, that it reduces to the Gittins index in the special cases of our model that correspond to Gittins' MABs. Other indices that are generalized Gittins *in this sense* include those of Nash [19] for *generalized bandits* and of Whittle [30] for *restless bandits*. A more recent example is due to Denardo et al. [7].

In light of this indexability, we propose a *greedy index heuristic* (*GI*) for the optimization problem (5) that in each system state $\mathbf{x}$ constructs an action by adding bandits to the *active set* in decreasing order of the indices $\nu_n(\mathbf{x}_n)$ until the point is reached where any further such addition is *either* of a zero index bandit and/or violates the resource constraint. This heuristic corresponds to the index policies of Gittins [8] and Whittle [30] in the simpler settings considered by them. In our context, *GI* runs the risk of poor utilization of the available resource, for example, in cases where the increments in resource needed to activate individual bandits are large and irregular. In analyses of some simple models in §5, we shall see that this emerges as the central concern regarding the performance of *GI*. One suggested approach in such cases is to use the *weighted index heuristic* (*WI*). In each system state $\mathbf{x}$, *WI* chooses an action to maximise the sum

$$\begin{aligned}\sum_n \big[&\{c_n^a(x_n) - c_n^b(x_n) - S_{nc}^b(x_n)\} I(\mathbf{x}_n \in \{a\} \times \Omega_n)\\ &+ \{c_n^a(x_n) + S_{nc}^a(x_n) - c_n^b(x_n)\} I(\mathbf{x}_n \in \{b\} \times \Omega_n)\big]\nu_n(\mathbf{x}_n)\end{aligned} \tag{14}$$

within the resource constraint. The sum in (14) is taken over all active bandits. It weights the index for bandit n in state $\mathbf{x}_n$ by the difference between the resource required for active and for passive treatment. As shown in the next section, the index is a certain ratio of reward to consumption. Because our goal is reward maximization, it seems heuristically plausible to weight each index by consumption.

For further discussion of the notion of indexability in the wider context of *restless bandit problems*, the reader is referred to Whittle [30]. He gives an example of a restless bandit that fails to be indexable.

3. Indexability and indices for bandits with varying resource requirements. From Definition 2.1, the investigation of indexability in relation to the class of multiarmed bandit problems of the previous section centres on the decision problems $p_n(\nu)$, $1 \leq n \leq N$. Because indexability is a characteristic of *individual bandits*, it is these that become the focus of our discussion in this section and the next. We are thus able to drop the bandit identifier n from the notation and will use B to denote an individual bandit within our MAB model. With B is associated the reward-discounted Markov decision problem $p(\nu)$, which is structured as follows:

(i)′ At each time $t \in \mathbb{N}$, either action a (active) or action b (passive) is applied to B.

(ii)′ The *state space* of B is $\{a, b\} \times \Omega$. We write $\mathbf{X}(t)$ for the *extended state* of B at time t. If $\mathbf{X}(t) = (\bullet, x)$, $\bullet \in \{a, b\}$, $x \in \Omega$ for some $t \in \mathbb{Z}^+$, then the *state* of the bandit is x and action $\bullet$ was applied to B at time $t-1$. When needed, we use $X(t)$ for the (nonextended) bandit state, i.e., $\mathbf{X}(t) = (\bullet, x) \implies X(t) = x$. We make no assumption about the initial state of B; we consider both $\mathbf{X}(0) \in \{a\} \times \Omega$ and $\mathbf{X}(0) \in \{b\} \times \Omega$.

(iii)′ Should action a be applied to B at time t, where $\mathbf{X}(t) = (\bullet, x)$, then the new state will be $\mathbf{X}(t+1) = (a, y)$, with y determined according to the Markov law P. We write

$$P\{\mathbf{X}(t+1) = (a, y) \mid \mathbf{X}(t) = (\bullet, x), a\} = P_{xy}, \quad x, y \in \Omega. \tag{15}$$

Should action b be applied to the bandit at time t, where $\mathbf{X}(t) = (\bullet, x)$, then

$$P\{\mathbf{X}(t+1) = (b, x) \mid \mathbf{X}(t) = (\bullet, x), b\} = 1, \quad x \in \Omega. \tag{16}$$

(iv)′ Should action a be applied to B when in state $(\bullet, x)$, then an expected reward is earned that is equal to $r^a(x) - \nu c^a(x)$ when $\bullet = a$ and equal to $r^a(x) - S_r^a(x) - \nu\{c^a(x) + S_c^a(x)\}$ when $\bullet = b$. Should action b be applied to B when in state $(\bullet, x)$, then an expected reward is earned that is equal to $r^b(x) - S_r^b(x) - \nu\{c^b(x) + S_c^b(x)\}$ when $\bullet = a$ and equal to $r^b(x) - \nu c^b(x)$ when $\bullet = b$. Note that we continue to interpret the functions r^a and r^b as returns earned by B; S_r^a and S_r^b as penalties paid, respectively, when switching processing toward and away from B; c^a and c^b as quantities of resource consumed by B; and S_c^a and S_c^b as additional resource consumed, respectively, when switching processing toward and away from B.

(v)′ A *policy* is a rule for choosing an action from the set $\{a, b\}$ at each decision epoch. An *optimal policy* for $p(\nu)$ maximises the total expected discounted reward earned over an infinite horizon. Puterman [25] asserts the existence of an optimal policy that is *stationary* and that satisfies the DP optimality equations. To express these, we write $V\colon \{a, b\} \times \Omega \times \mathbb{R}^+ \to \mathbb{R}$ for the *value function* for $p(\bullet)$, with $V((\bullet, x), \nu)$ the maximal return from the bandit B when $\mathbf{X}(0) = (\bullet, x)$ and the resource charge is $\nu \geq 0$. From (i)′–(iv)′ above, the optimality equations for the problem $p(\nu)$ are

$$V((a,x),\nu) = \max\left\{ r^a(x) - \nu c^a(x) + \beta \sum_{y\in\Omega} P_{xy} V((a,y),\nu); \right.$$
$$\left. r^b(x) - S_r^b(x) - \nu\{c^b(x) + S_c^b(x)\} + \beta V((b,x),\nu) \right\} \tag{17}$$

and

$$V((b,x),\nu) = \max\left\{ r^a(x) - S_r^a(x) - \nu\{c^a(x) + S_c^a(x)\} + \beta \sum_{y\in\Omega} P_{xy} V((a,y),\nu); \right.$$
$$\left. r^b(x) - \nu c^b(x) + \beta V((b,x),\nu) \right\}, \tag{18}$$

where both equations hold $\forall\, \nu \in \mathbb{R}^+$ and $x \in \Omega$. In Equation (17) the two terms within the maximization on the right-hand side are the expected returns when action a (respectively, b), is applied at time 0 to the bandit when in extended state (a, x) and when actions are taken optimally thereafter. Equation (18) is constructed similarly for the extended state (b, x).

We require the following technical assumption.

Condition 1 (Resource Consumption). The functions c^a, c^b, S_c^b and the Markov law P are such that for all $x \in \Omega$,

$$c^a(x) + \beta \sum_{y\in\Omega} P_{xy}\left\{S_c^b(y) + c^b(y)(1-\beta)^{-1}\right\} > S_c^b(x) + c^b(x)(1-\beta)^{-1}. \tag{19}$$

Condition 1 implies that if we start from extended state (a, x) and apply the passive action to B throughout, the expected total discounted amount of resource consumed will be increased if we change the action taken at time 0 from passive to active—i.e., if we choose action a at time 0 and action b thereafter. The condition is trivially satisfied when c^a is positive valued and $c^b \equiv 0$, $S_c^b \equiv 0$, namely, that B consumes no resource when passive. Inequality (19) can be rendered nonstrict at the cost of some additional complexity.

The next result describes an important property of optimal policies for $p(\nu)$. It arises because a switch of processing away from an active bandit followed immediately by a transfer of processing back to it can only introduce unnecessarily incurred switching costs and unnecessarily consumed additional resource. Such action sequences can therefore be eliminated from consideration.

Lemma 3.1. *If action b is optimal for (a, x), then it is also optimal for (b, x), strictly so if any of the switching penalties $S_r^a(x), S_r^b(x), S_c^a(x)$, or $S_c^b(x)$ are strictly positive.*

Proof. If action b is optimal for (a, x), then by standard results, the maximization on the right-hand side (r.h.s.) of (17) is achieved by the second term, namely,

$$r^b(x) - S_r^b(x) - \nu\left\{c^b(x) + S_c^b(x)\right\} + \beta V((b,x),\nu) \geq r^a(x) - \nu c^a(x) + \beta \sum_{y\in\Omega} P_{xy} V((a,y),\nu).$$

It must then follow that

$$\begin{aligned} r^b(x) - \nu c^b(x) + \beta V((b,x),\nu) &\geq r^a(x) - \nu c^a(x) + S_r^b(x) + \nu S_c^b(x) + \beta \sum_{y\in\Omega} P_{xy} V((a,y),\nu) \\ &\geq r^a(x) - \nu c^a(x) - S_r^a(x) - \nu S_c^a(x) + \beta \sum_{y\in\Omega} P_{xy} V((a,y),\nu), \end{aligned}$$

where the second inequality holds because all switching penalties are nonnegative. The inequality is strict if any penalty is strictly positive. It follows that the maximization on the r.h.s. of (18) is achieved by the second term, uniquely so if the condition on switching penalties in the statement of the lemma holds. The result now follows. □

As before, we write $b(u)$ for the *passive set* of stationary policy u. It follows from Lemma 3.1 that in the search for solutions to $p(\nu)$, we may restrict attention to stationary policies in the class Φ, where

$$\Phi := \big\{u;\, (x;\, x\in\Omega \text{ and } (a,x)\in b(u)) \subseteq (x;\, x\in\Omega \text{ and } (b,x)\in b(u))\big\}.$$

Suppose now that $\mathbf{X}(0) = (\bullet, x),\ x\in\Omega$, and consider evolution of the bandit B under some $u\in\Phi$. There are two possibilities. Either $\mathbf{X}(0)\in b(u)$, in which case u's stationarity, its membership of Φ, and the nature of its evolution under b guarantees that $\mathbf{X}(t)\in b(u),\ t\geq 0$. Alternatively, $\mathbf{X}(0)\notin b(u)$, in which case u will choose action a from time 0 until time τ, where

$$\tau = \inf\big\{t;\, t\geq 1 \text{ and } X(t)\in (y\in\Omega \text{ and } (a,y)\in b(u))\big\}, \tag{20}$$

in which case u's stationarity, its membership of Φ, and the nature of its evolution under b guarantees that $\mathbf{X}(t)\in b(u),\ t\geq\tau$. Hence, the determination of optimal actions for $p(\nu)$ may be reduced to a collection of *stopping problems*, one for each (initial) extended state, where *stopping* connotes first application of the passive action b.

In order to develop the necessary ideas, consider the bandit B evolving from some initial enhanced state under continuous application of the active action a. We now introduce the class of *stationary positive-valued stopping times* T. For $\Theta\subseteq\Omega$, define class member $\tau(\Theta)$ by

$$\tau(\Theta) := \inf\{t;\, t\geq 1 \text{ and } X(t)\in\Theta\}. \tag{21}$$

Formally, $T = \{\tau(\Theta);\, \Theta\subseteq\Omega\}$. Plainly, the time τ in (20) is in T. To summarise, it follows from Lemma 3.1 and the above discussion that for any given initial extended state for B, an optimal policy for $p(\nu)$ must *either* always choose the passive action b *or* choose action a at time 0 and thereafter until some $\tau\in T$, at and after which action b is taken. As a useful shorthand, we shall refer to the latter as *policy* τ.

In order to progress with an economy of notation, when $\tau\in T$ we shall use the notations $r^a(x,\tau)$, $c^a(x,\tau)$ for, respectively, the total discounted reward earned and resource consumed during $[0,\tau)$ under policy τ from initial enhanced state $\mathbf{X}(0) = (a,x)$. We write

$$r^a(x,\tau) := \mathbf{E}\left[\sum_{t=0}^{\tau-1} \beta^t r^a(X(t)) \,\middle|\, x\right], \tag{22}$$

and similarly for $c^a(x,\tau)$, where in (22) and throughout we shall use $|x$ as a notational shorthand for the conditioning on the initial state more fully expressed by $|X(0)=x$. We can now introduce an appropriate *index* for bandit B in the form of a real-valued function on the state space $\{a,b\}\times\Omega$. Optimal policies for $p(\nu)$ will be describable in terms of this index.

Definition 3.1. The *return/consumption index* for B is a function ν: $\{a,b\}\times\Omega\to\mathbb{R}$ given by

$$\nu(a,x) := \sup_{\tau}\{\Delta r(a,x,\tau)/\Delta c(a,x,\tau)\}, \qquad x\in\Omega,$$

where

$$\Delta r(a,x,\tau) := r^a(x,\tau) - \big\{r^b(x) - \mathbf{E}(\beta^\tau r^b(X(\tau)) \mid x)\big\}(1-\beta)^{-1} + S_r^b(x) - \mathbf{E}(\beta^\tau S_r^b(X(\tau)) \mid x)$$

and

$$\Delta c(a,x,\tau) := c^a(x,\tau) - \big\{c^b(x) - \mathbf{E}(\beta^\tau c^b(X(\tau)) \mid x)\big\}(1-\beta)^{-1} - \{S_c^b(x) - \mathbf{E}(\beta^\tau S_c^b(X(\tau)) \mid x)\};$$

and

$$\nu(b, x) := \sup_{\tau}\{\Delta r(b, x, \tau)/\Delta c(b, x, \tau)\}, \quad x \in \Omega,$$

where

$$\Delta r(b, x, \tau) := \Delta r(a, x, \tau) - S_r^a(x) - S_r^b(x)$$

and

$$\Delta c(b, x, \tau) := \Delta c(a, x, \tau) + S_c^a(x) + S_c^b(x).$$

Both of the above suprema are over $\tau \in T$. We shall write ν^+: $\{a, b\} \times \Omega \to \mathbb{R}^+$ for the *positive part of function* ν, namely,

$$\nu^+(\bullet, x) = \max\{\nu(\bullet, x), 0\}, \quad \bullet \in \{a, b\}, \quad x \in \Omega.$$

Remark. The *return/consumption indices* of Definition 3.1 are guaranteed finite and have a natural interpretation. The quantity $\Delta r(a, x, \tau)$ (respectively, $\Delta r(b, x, \tau)$) is the increase in expected return achieved when the active action a is taken during $[0, \tau)$ instead of the passive action b from enhanced state (a, x) (respectively, (b, x)). Similarly, $\Delta c(a, x, \tau)$ (respectively, $\Delta c(b, x, \tau)$) is the increase in resource consumed. Hence, the ratio may be understood as a measure of additional return achieved from such processing per unit of additional resource consumed. The *return/consumption index* may be understood as a generalized form of Gittins index that is able to take varying patterns of resource consumption into account.

Theorem 3.1 (Indexability and Indices). *The bandit B is indexable with index given by the positive part of the return/consumption index above. Further, $\nu^+(a, x) \geq \nu^+(b, x)$, $\forall\, x \in \Omega$.*

Proof. Consider extended state (a, x), $x \in \Omega$ and stopping time $\tau \in T$. The expected reward earned by bandit B from initial state (a, x) under policy τ is written:

$$\begin{aligned} r^a(x, \tau) - \nu c^a(x, \tau) - \mathbf{E}\big(\beta^\tau S_r^b(X(\tau)) \mid x\big) - \nu \mathbf{E}\big(\beta^\tau S_c^b(X(\tau)) \mid x\big) \\ + \mathbf{E}\big(\beta^\tau \big\{r^b(X(\tau)) - \nu c^b(X(\tau))\big\} \mid x\big)(1-\beta)^{-1}. \end{aligned} \tag{23}$$

Reading the expression (23) from left to right, we first account for the rewards earned net of charges for resource consumption under action a during $[0, \tau)$, then the switching costs incurred at τ when the choice of action changes from a to b, and finally, the net rewards earned under action b during $[\tau, \infty)$. From the above discussion following Lemma 3.1, the active action will be strictly optimal for (a, x) when $\exists \tau \in T$ for which the quantity in (23) exceeds the expected reward obtained when action b is taken throughout, namely,

$$\{r^b(x) - \nu c^b(x)\}(1-\beta)^{-1} - S_r^b(x) - \nu S_c^b(x).$$

Straightforward algebra yields the conclusion that this is equivalent to the requirement that $\exists \tau \in T$ for which

$$\Delta r(a, x, \tau) > \nu \Delta c(a, x, \tau).$$

Condition 1 above guarantees that $\Delta c(\bullet, x, \tau) > 0$ for all extended states $(\bullet, x) \in \{a, b\} \times \Omega$ and stationary positive-valued stopping times $\tau \in T$. It now follows that action a is optimal in (a, x) whenever there exists $\tau \in T$ for which

$$\Delta r(a, x, \tau)/\Delta c(a, x, \tau) > \nu.$$

This will be the case if $\nu^+(a, x) = \nu(a, x) > \nu$. By continuity of (optimized) returns as ν varies, action a must also be optimal in state (a, x) when $\nu^+(a, x) = \nu$. However, if $\nu(a, x) < \nu$, then

$$\Delta r(a, x, \tau)/\Delta c(a, x, \tau) < \nu \quad \forall\, \tau \in T,$$

from which it follows that action b is strictly optimal in state (a, x). We conclude that action a is optimal in state (a, x) if and only if $\nu^+(a, x) \geq \nu$. Further, action b is optimal in state (a, x) if and only if $\nu^+(a, x) \leq \nu$.

A similar argument yields the conclusion that action a is optimal in (b, x) if and only if $\nu^+(b, x) \geq \nu$, with action b optimal if and only if $\nu^+(b, x) \leq \nu$.

We infer the existence of a family of optimal policies $\{u(\nu), \nu \in \mathbb{R}^+\}$ for $p(\nu)$ whose associated passive sets

$$b\{u(\nu)\} = \big\{(\bullet, x); \bullet \in (a, b), x \in \Omega, \nu^+(\bullet, x) \leq \nu\big\}$$

are nondecreasing in ν. This establishes the indexability of B from Definition 2.1. That the index is indeed the positive part of the *return/consumption index* of Definition 3.1 is now trivial, as is the fact that $\nu^+(a, x) \geq \nu^+(b, x)\ \forall\, x \in \Omega$. The latter is an immediate consequence of the definitions of the quantities concerned. □

We shall now proceed to show that in the above definition of the *return/consumption index*, the supremum is always achieved by stopping time(s) that have a simple characterization. To facilitate the discussion, we introduce subsets of state space Ω as follows:

$$\Lambda(\nu) := \{y; y \in \Omega, \nu(a, y) < \nu\}, \quad \nu \in \mathbb{R}^+,$$

and

$$\Xi(\nu) := \{y; y \in \Omega, \nu(a, y) = \nu\}, \quad \nu \in \mathbb{R}^+.$$

Further, we use $T(\nu)$ for the collection of stationary positive-valued stopping times given by

$$T(\nu) = \big\{\tau(\Theta); \Theta = \Lambda(\nu) \cup \Xi \text{ for some } \Xi \subseteq \Xi(\nu)\big\}, \quad \nu \in \mathbb{R}^+.$$

Proposition 3.1. *If* $\nu(\bullet, x) \geq 0$, *then any stopping time in* $T\{\nu(\bullet, x)\}$ *achieves the supremum in the definition of* $\nu(\bullet, x)$, $\bullet \in \{a, b\}$, $x \in \Omega$.

Proof. Consider the problem $p\{\nu(a, x)\}$ in which ν, the prevailing charge per unit of resource consumed, is fixed at the value $\nu(a, x)$, assumed to be nonnegative. By the argument in the proof of the preceding theorem, in problem $p\{\nu(a, x)\}$ action a, respectively b, is optimal in enhanced state $(\bullet, y)$ if and only if $\nu(\bullet, y) \geq \nu(a, x)$, respectively, $\nu(\bullet, y) \leq \nu(a, x)$. In particular, both actions a and b are optimal at time 0 when $\mathbf{X}(0) = (a, x)$. Should action a be taken at 0, then it is optimal to continue taking it through $[0, t)$, provided that $\nu(a, x) \leq \nu\{a, X(s)\}$, $0 \leq s \leq t-1$. Should that be so, action b will then be optimal at t if $\nu(a, x) \geq \nu\{a, X(t)\}$. Hence, assuming that we make decisions in a stationary fashion, we deduce that only the following sequences of actions are optimal for $p\{\nu(a, x)\}$ when $\mathbf{X}(0) = (a, x)$: Firstly, take action b at all decision epochs; and secondly, take action a throughout $[0, \tau)$ and action b from τ onwards, where $\tau \in T\{\nu(a, x)\}$. Equating the expected returns from the policies outlined in the two alternatives yields the conclusion that

$$\nu(a, x) = \Delta r(a, x, \tau)/\Delta c(a, x, \tau), \quad \tau \in T\{\nu(a, x)\}, \quad x \in \Omega,$$

which establishes the result when $\bullet = a$. The case $\bullet = b$ is dealt with similarly. □

Remark. The proof of Proposition 3.1 also reveals that the stopping times in $T\{\nu(\bullet, x)\}$ are essentially the *only* stationary ones achieving the suprema concerned. Should stopping time $\tau(\Theta)$ in (21) be such that stopping set Θ contains some y for which $\nu(a, y) > \nu(\bullet, x)$ and, moreover, that when the bandit is subject to the continuous application of action a from initial state $(\bullet, x)$, then

$$P\{X(\tau(\Theta)) = (a, y)\} > 0,$$

it will follow from the argument in the above proof that

$$\nu(\bullet, x) > \Delta r\{\bullet, x, \tau(\Theta)\}/\Delta c\{\bullet, x, \tau(\Theta)\}.$$

4. Computation of the return/consumption indices.

We have noted above the Gittins indexlike nature of the *return/consumption indices* developed in the preceding section for our multiarmed bandit model with general resource requirements. We shall now proceed to give an algorithm for their computation in finite state-space cases, which is a suitably modified version of the *adaptive greedy algorithm* for Gittins indices first given by Robinson [27] and subsequently further developed by Bertsimas and Niño-Mora [5]. All of these algorithms compute the indices required for each bandit in order from largest to smallest.

Consider now the bandit B. We produce from it a *derived* Gittins-type bandit DB (i.e., one that does not change state and earns no rewards under the passive action) as follows:

(i)″ DB's state space is the enhanced state space of B, namely, $\Omega_D := \{a, b\} \times \Omega$. This is now assumed to be finite. We use $\mathbf{X}(t)$ for the state of DB at time $t \in \mathbb{N}$. At each time $t \in \mathbb{N}$ either action a (active) or action b (passive) is applied to DB.

(ii)″ The stochastic dynamics of DB under action a are exactly as for B and are given in (15) above. However, DB does not change state under action b.

(iii)″ The *reward function* r_D: $\Omega_D \to \mathbb{R}$ yields expected rewards earned by DB under action a. Specifically, when action a is taken in state (a, x), the expected reward is

$$r_D(a,x) := r^a(x) - \left\{r^b(x) - \beta\sum_{y\in\Omega} P_{xy} r^b(y)\right\}(1-\beta)^{-1} + \left\{S_r^b(x) - \beta\sum_{y\in\Omega} P_{xy} S_r^b(y)\right\}, \quad x \in \Omega,$$

and when action a is taken in state (b, x), the expected reward is

$$r_D(b,x) := r_D(a,x) - S_r^a(x) - S_r^b(x), \quad x\in\Omega.$$

The function r_D may be understood as follows. It represents the difference in expected reward between giving bandit B active treatment for exactly one step, followed by passive thereafter, and giving passive throughout from the outset. Derived bandit DB earns no rewards under action b.

Suppose now that DB is in some state $(\bullet, x)$ at time 0 and is subject to action a up to stopping time $\tau \in T$. It follows from the characterization of r_D in the preceding paragraph that the expected reward earned by DB during $[0, \tau)$ can be equated to the difference in expected reward earned by bandit B under policy τ and the policy which takes the passive action throughout. However, this difference is exactly the quantity $\Delta r(\bullet, x, \tau)$ of Definition 3.1. Formally, we write, in an obvious notation,

$$\mathbf{E}\left[\sum_{t=0}^{\tau-1}\beta^t r_D(\mathbf{X}(t))\,\middle|\,(\bullet,x)\right] = \Delta r(\bullet,x,\tau), \quad \bullet\in\{a,b\},\ x\in\Omega,\ \tau\in T. \tag{24}$$

(iv)″ From the consumption functions c^a, c^b, S_c^a, S_c^b associated with B we derive a function C_D: $\Omega_D \to \mathbb{R}^+$ for DB. We first write $C^a(x)$ for the total expected discounted resource consumed by B when action a is taken at all decision epochs from time 0, at which point it is in enhanced state (a, x). That is,

$$C^a(x) := \mathbf{E}\left[\sum_{t=0}^{\infty}\beta^t c^a(X(t))\,\middle|\,(a,x)\right], \quad x\in\Omega,$$

the expectation being taken over realizations of B under permanent application of the active action. We now define

$$C_D(a,x) := C^a(x) - c^b(x)(1-\beta)^{-1} - S_c^b(x), \quad x\in\Omega,$$

and

$$C_D(b,x) := C_D(a,x) + S_c^a(x) + S_c^b(x), \quad x\in\Omega.$$

The function C_D may be understood as follows. It represents the difference in total expected resource consumed between giving bandit B active treatment throughout and giving B passive treatment throughout. Condition 1 guarantees that C_D is positive valued.

Suppose now that DB is in some state $(\bullet, x)$ at time 0 and is subject to action a up to stopping time $\tau\in T$. It follows from the characterization of C_D in the preceding paragraph that the difference

$$C_D(\bullet,x) - \mathbf{E}[\beta^\tau C_D(\mathbf{X}(\tau)) \mid (\bullet,x)]$$

can be equated to the difference in total expected discounted resource consumed by bandit B under policy τ and the policy that takes the passive action throughout. However, this difference is exactly the quantity $\Delta c(\bullet, x, \tau)$ of Definition 3.1. Hence, we have

$$C_D(\bullet,x) - \mathbf{E}[\beta^\tau C_D(\mathbf{X}(\tau)) \mid (\bullet,x)] = \Delta c(\bullet,x,\tau), \quad \bullet\in\{a,b\},\ x\in\Omega,\ \tau\in T. \tag{25}$$

It now follows from (24), (25), and Definition 3.1 that the return/consumption indices for B may expressed in terms of quantities associated with the derived bandit DB as follows:

$$\nu(\bullet,x) = \sup_\tau\left\{\mathbf{E}\left[\sum_{t=0}^{\tau-1}\beta^\tau r_D(\mathbf{X}(t)) \mid (\bullet,x)\right]\Big/\left(C_D(\bullet,x) - \mathbf{E}[\beta^\tau C_D(\mathbf{X}(\tau)) \mid (\bullet,x)]\right)\right\}, \quad \bullet\in\{a,b\},\ x\in\Omega, \tag{26}$$

the supremum being taken over all $\tau \in T$. The expression in (26) is an index that is of the modified Gittins type. The form of modification is precisely that required by a class of *generalized bandit problems* first studied by Nash [19]. An account of this class of processes has been given by Crosbie and Glazebrook [6] using polyhedral methods. From that analysis we are able to infer that the following *adaptive greedy algorithm* yields the *return/consumption indices* for bandit B when the latter has finite state space.

For ease of notation, we now use $\mathbf{x}$, $\mathbf{y}$ for generic states of DB. We require the matrix of constants $\mathbf{A}(B) := \{A_{\mathbf{x}}(\Theta)\}_{\mathbf{x}\in\Omega_D,\Theta\subseteq\Omega_D}$ defined as follows: Suppose that $\mathbf{X}(0)=\mathbf{x}$ and that DB evolves from zero under the active action. We use $\tau_{\mathbf{x}}(\Theta)$ for the first time at or after time 1 for which the state of DB lies in subset Θ. We write

$$\tau_{\mathbf{x}}(\Theta) = \inf\{t;\, t \geq 1 \text{ and } \mathbf{X}(t) \in \Theta\}.$$

We then define

$$A_{\mathbf{x}}(\Theta) := I(\mathbf{x}\in\Theta)\big(C_D(\mathbf{x}) - \mathbf{E}[\beta^{\tau_{\mathbf{x}}(\Theta)} C_D(\mathbf{X}(\tau_{\mathbf{x}}(\Theta))) \mid \mathbf{x}]\big), \qquad \mathbf{x}\in\Omega_D, \quad \Theta\subseteq\Omega_D, \tag{27}$$

where $C_D\colon \Omega_D \to \mathbb{R}^+$ is as in (iv)″ above. The following *adaptive greedy algorithm* yields the *return/consumption* indices $\{\nu(\mathbf{x}), \mathbf{x}\in\Omega_D\}$. Inputs to the algorithm are the matrix $\mathbf{A}(B)$ and the rewards $\{r_D(\mathbf{x}), \mathbf{x}\in\Omega_D\}$ defined in (iii)″ above. The algorithm operates as follows:

Step 1.

$$\text{Set } \Theta_{|\Omega_D|} = \Omega_D \qquad \text{and} \qquad \gamma(\Theta_{|\Omega_D|}) = \max\left\{\frac{r_D(\mathbf{x})}{A_{\mathbf{x}}(\Omega_D)};\, \mathbf{x}\in\Omega_D\right\}. \tag{28}$$

Take any maximising state from (28) and call it state $|\Omega_D|$. State $|\Omega_D|$ has bandit B's largest *return/consumption index*. Set $\nu(|\Omega_D|) = \gamma(\Theta_{|\Omega_D|})$ and $\Theta_{|\Omega_D|-1} = \Theta_{|\Omega_D|}\backslash\{|\Omega_D|\}$.

Step k. For $k = 2, 3, \ldots, |\Omega_D|$, set $\Theta_{|\Omega_D|-k+1} = \Theta_{|\Omega_D|-k+2}\backslash\{|\Omega_D|-k+2\}$ and

$$\gamma(\Theta_{|\Omega_D|-k+1}) = \max\left\{\frac{r_D(\mathbf{x}) - \sum_{j=1}^{k-1} A_{\mathbf{x}}(|\Omega_D|-j+1)\gamma(\Theta_{|\Omega_D|-j+1})}{A_{\mathbf{x}}(\Theta_{|\Omega_D|-k+1})};\, \mathbf{x}\in\Theta_{|\Omega_D|-k+1}\right\}. \tag{29}$$

Take any maximising state from (29) and call it state $|\Omega_D|-k+1$. State $|\Omega_D|-k+1$ has bandit B's kth largest *return/consumption index*. Set $\nu(|\Omega_D|-k+1) = \nu(|\Omega_D|-k+2) + \gamma(\Theta_{|\Omega_D|-k+1})$.

Remark. The collection $\{\nu(k), 1\leq k\leq|\Omega_D|\}$ consists of the *return/consumption indices* of all (extended) states of the bandit B, with these states now numbered in decreasing order of their index values, state $|\Omega_D|$ having the highest *return/consumption index*. The idea behind the algorithm is as follows: Consider first the highest index state $|\Omega_D|$. From Definition 3.1 and again in (26), the index $\nu(|\Omega_D|)$ is characterized as a supremum over $\tau\in T$ of a quantity that measures additional reward earned per unit of additional resource consumed up to τ. It is easy to show that, in the case of $\nu(|\Omega_D|)$, this supremum is achieved by the stopping time $\tau(\Omega_D)$ such that $\tau\equiv 1$ almost surely. This is reflected in Step 1 of the above algorithm. The second-highest index may then be computed by constructing a new bandit, adapted from B by removing state $|\Omega_D|$ from its state space and modifying its reward/stochastic structure, such that the highest index state of this new bandit is the second-highest state for B, which is then easy to compute. Step 2 of the above algorithm implements this. The algorithm continues in this fashion. There are certainly other algorithms that will compute the *return/consumption indices*. Most existing algorithms for the computation of Gittins' indices may be modified to yield the quantities described in Definition 3.1. These include the approach of Katehakis and Veinott [17] based on restart problems and the fast pivoting algorithm of Niño-Mora [22]. We emphasise the *adaptive greedy algorithm* because its structure will be exploited in the theory of §5.

States of B are now identified as integers in the range $1\leq k\leq|\Omega_D|$, whereas subsets of Θ are members of $2^{\{1,2,\ldots,|\Omega_D|\}}$. In the above algorithm, each subset $\Theta_j = \{j, j-1, \ldots, 1\}$ contains j states of smallest return/consumption index. It is an immediate consequence of the structure of the above algorithm that the rewards $\{r_D(i), 1\leq i\leq|\Omega_D|\}$ may be reexpressed as

$$r_D(i) = \nu(|\Omega_D|)A_i(|\Omega_D|) - \sum_{j=i}^{|\Omega_D|-1}\{\nu(j+1)-\nu(j)\}A_i(\Theta_j), \qquad 1\leq i\leq|\Omega_D|, \tag{30}$$

and this representation will be used in the development of the suboptimality bounds discussed in the next section.

5. On the closeness to optimality of index policies. In the preceding two sections, the focus has been on individual bandits as we have demonstrated indexability and developed the *return/consumption indices*. We now return to the full multiarmed bandit (MAB) model of §2 and will need to reinstate the bandit identifying suffix n within the notation. This model features N bandits, with $\Omega_{nD} := \{a, b\} \times \Omega_n$ the state space of bandit n, $1 \le n \le N$. We use $\mathbf{\Omega}_D := \bigcup_{n=1}^N \Omega_{nD}$ for their union. In what follows we shall also use $\Omega_n^a := \{a\} \times \Omega_n$ for the subset of Ω_{nD} consisting of bandit n's currently active states. Please note that all results in this section up to and including Theorem 5.1 relate to the MAB model of §2 in full generality. Thereafter, a collection of simplifying assumptions are made in the interest of securing the form of asymptotic optimality of the *GI* heuristic described in Theorem 5.3.

All members of $\mathbf{\Omega}_D$ have a *return/consumption index*, computed by applying the adaptive greedy algorithm of the preceding section to each bandit in turn. We create a numerical representation of the members of $\mathbf{\Omega}_D$ by numbering them in decreasing order of their index values such that

$$\nu(|\mathbf{\Omega}_D|) \ge \nu(|\mathbf{\Omega}_D| - 1) \ge \cdots \ge \nu(2) \ge \nu(1). \tag{31}$$

We use i for a generic state within this numbering. With this state representation, subsets $\mathbf{\Theta}$ of $\mathbf{\Omega}_D$ are members of $2^{\{1,2,\ldots,|\mathbf{\Omega}_D|\}}$. We shall use $\mathbf{\Theta}_j = \{j, j-1, \ldots, 1\}$ for a subset of $\mathbf{\Omega}_D$ containing j states with smallest *return/consumption indices*. We define the matrix $\mathbf{A} := \{\mathbf{A}_i(\mathbf{\Theta})\}_{i\in\mathbf{\Omega}_D, \mathbf{\Theta}\subseteq\mathbf{\Omega}_D}$ as follows: If state i is a member of Ω_{nD} (i.e., is a bandit n state), then $\mathbf{A}_i(\mathbf{\Theta}) := A_i(\mathbf{\Theta} \cap \Omega_{nD})$, with the latter quantity being given by an appropriate form of (27). It is now a straightforward matter to check that the expression for the rewards $r_D(i)$ given (30) in terms of quantities associated with individual bandits yield the equations

$$r_D(i) = \nu(|\mathbf{\Omega}_D|)\mathbf{A}_i(\mathbf{\Omega}_D) - \sum_{j=i}^{|\mathbf{\Omega}_D|-1} \{\nu(j+1) - \nu(j)\}\mathbf{A}_i(\mathbf{\Theta}_j), \quad 1 \le i \le |\mathbf{\Omega}_D|. \tag{32}$$

We write $R^u(\mathbf{j})$ for the total expected discounted reward (net of switching penalties) earned by the MAB when stationary admissible policy $u \in \mathcal{U}$ is applied from initial state $\mathbf{j} \equiv \{j_1, j_2, \ldots, j_N\} \in \times_{n=1}^N \Omega_{nD}$. We shall develop an expression for $R^u(\mathbf{j})$ using the representation of rewards given in (32) and the *performance measures* $\{\pi^u(i \mid \mathbf{j}), i \in \mathbf{\Omega}_D\}$ defined by

$$\pi^u(i \mid \mathbf{j}) := \mathbf{E}_u\left\{\sum_{t=0}^\infty \beta^t I(i, t) \,\middle|\, \mathbf{j}\right\}. \tag{33}$$

In (33) the expectation is taken over all realizations of the MAB under policy u from initial state $\mathbf{j}$, and $I(i, t)$ is an indicator that takes the value 1 if a bandit with state i is active at epoch $t \in \mathbb{N}$ and is zero otherwise. Proposition 5.1 describes how to deploy these performance measures to compute $R^u(\mathbf{j})$.

Proposition 5.1 (Computation of Total Expected Rewards). *For any $u \in \mathcal{U}$ and $\mathbf{j} \in \times_{n=1}^N \Omega_{nD}$,*

$$R^u(\mathbf{j}) = \sum_{i\in\mathbf{\Omega}_D} r_D(i)\pi^u(i \mid \mathbf{j}) + \sum_{n=1}^N \{-S_{nr}^b(j_n) I(j_n \in \Omega_n^a) + r_n^b(j_n)(1-\beta)^{-1}\}, \tag{34}$$

where $I(\bullet)$ is an indicator.

Proof. Adopting the usage in (5) above, we write

$$R^u(\mathbf{j}) = \sum_{n=1}^N R_n^u(\mathbf{j}), \tag{35}$$

where in (35), $R_n^u(\mathbf{j})$ is the total expected discounted reward (net of switching penalties) yielded by bandit n under policy u from initial state $\mathbf{j}$. To demonstrate (34), it is enough to show that

$$R_n^u(\mathbf{j}) = \sum_{i\in\Omega_{nD}} r_D(i)\pi^u(i \mid \mathbf{j}) - S_{nr}^b(j_n) I(j_n \in \Omega_n^a) + r_n^b(j_n)(1-\beta)^{-1}, \quad 1 \le n \le N. \tag{36}$$

With policy u and initial state $\mathbf{j}$ fixed, we define the sequence of random times $\{\psi_m^n, m \ge 1\}$ as the collection of (successive) epochs at which processing is *switched into* bandit n, with $\{\phi_m^n, m \ge 1\}$ as the collection of (successive) epochs at which processing is *switched away* from bandit n. In what follows we shall use the notational shorthand $u_n(t)$ for the action applied by policy u to bandit n at epoch t. Hence, we have

$$\psi_1^n = \inf\{t \geq 0;\, u_n(t) = a\}.$$

If $\psi_1^n = \infty$, then $\phi_m^n = \infty,\ \ m \geq 1$, and $\psi_m^n = \infty,\ \ m \geq 2$. Otherwise,

$$\phi_1^n = \inf\{t > \psi_1^n;\, u_n(t) = b\}.$$

We continue inductively. If $m \geq 2$ and $\phi_{m-1}^n < \infty$, we have

$$\psi_m^n = \inf\{t > \phi_{m-1}^n;\, u_n(t) = a\},$$

whereas if $\psi_m^n < \infty$, we have

$$\phi_m^n = \inf\{t > \psi_m^n;\, u_n(t) = b\}.$$

We now utilise the form of reward function r_{nD} given in (iii)$''$ above and the nature of the performance measures in (33) to infer that when $j_n \in \Omega_{nD} \backslash \Omega_n^a$ (bandit n is deemed in a passive state at time 0),

$$\begin{aligned}
&\sum_{i \in \Omega_{nD}} r_D(i)\pi^u(i \mid \mathbf{j}) - S_{nr}^b(j_n) I(j_n \in \Omega_n^a) + r_n^b(j_n)(1-\beta)^{-1} \\
&\quad = \sum_{i \in \Omega_{nD}} r_D(i)\pi^u(i \mid \mathbf{j}) + r_n^b(j_n)(1-\beta)^{-1} \\
&\quad = \mathbf{E}\left[\sum_{m=1}^{\infty}\left\{-\beta^{\psi_m^n} S_{nr}^a(X_n(\psi_m^n)) + \left[\sum_{t=\psi_m^n}^{\phi_m^n - 1} \beta^t r_n^a(X_n(t))\right] - \beta^{\phi_m^n} S_{nr}^b(X_n(\phi_m^n))\right\} \middle|\, \mathbf{j}\right] \\
&\qquad + \mathbf{E}\left[\sum_{m=1}^{\infty}\{-\beta^{\psi_m^n} r_n^b(X_n(\psi_m^n)) + \beta^{\phi_m^n} r_n^b(X_n(\phi_m^n))\} \middle|\, \mathbf{j}\right](1-\beta)^{-1} + r_n^b(j_n)(1-\beta)^{-1}. \qquad (37)
\end{aligned}$$

Recall that $X_n(t)$ denotes the (nonextended) state of bandit n at time t. If we now use the fact that the (nonextended) state of bandit n is frozen under application of the passive action, we have that, with probability one,

$$X_n(t) = X_n(\phi_m^n), \qquad \phi_m^n \leq t \leq \psi_{m+1}^n - 1, \quad m \geq 0, \qquad (38)$$

where we take $\phi_0^n = 0$. Utilising (38) within (37), we see that the r.h.s. of the latter expression may be rewritten

$$\begin{aligned}
&\mathbf{E}\left[\sum_{m=1}^{\infty}\left\{-\beta^{\psi_m^n} S_{nr}^a(X_n(\psi_m^n)) + \left[\sum_{t=\psi_m^n}^{\phi_m^n - 1} \beta^t r_n^a(X_n(t))\right] - \beta^{\phi_m^n} S_{nr}^b(X_n(\phi_m^n))\right\} \middle|\, \mathbf{j}\right] \\
&\qquad + \mathbf{E}\left[\left\{\sum_{t=0}^{\psi_1^n - 1} \beta^t r_n^b(X_n(t)) + \sum_{m=1}^{\infty} \sum_{t=\phi_m^n}^{\psi_{m+1}^n - 1} \beta^t r_n^b(X_n(t))\right\} \middle|\, \mathbf{j}\right]. \qquad (39)
\end{aligned}$$

However, the expression (39) accounts for all rewards earned by, and switching penalties exacted from, bandit n when policy u is applied to it from initial state $\mathbf{j}$. Hence, the quantity in (39) is $R_n^u(\mathbf{j})$, as required. The case $j_n \in \Omega_n^a$ is dealt with similarly. We therefore have (36), and the result is proved. □

We observe that the second term on the r.h.s. of (34) is policy independent and are able to conclude that, for any $u, v \in \mathcal{U}$, and any initial state $\mathbf{j}$,

$$R^u(\mathbf{j}) - R^v(\mathbf{j}) = \sum_{i \in \mathbf{\Omega}_D} r_D(i)\{\pi^u(i \mid \mathbf{j}) - \pi^v(i \mid \mathbf{j})\}. \qquad (40)$$

To deploy (32) within (40), we now write, for any policy $u \in \mathcal{U}$, subset $\mathbf{\Theta} \subseteq \mathbf{\Omega}_D$, and system state $\mathbf{j} \in \times_{n=1}^N \Omega_{nD}$,

$$\mathbf{A}^u(\mathbf{\Theta} \mid \mathbf{j}) := \sum_{i \in \mathbf{\Theta}} \mathbf{A}_i(\mathbf{\Theta})\pi^u(i \mid \mathbf{j}). \qquad (41)$$

In what follows we use u^* for an optimal stationary admissible policy for our MAB and $R^{\text{opt}}(\mathbf{j})$ for the maximal expected return from initial state $\mathbf{j}$, as in (5). Theorem 5.1 is a straightforward consequence of (32), (34), and (41).

Theorem 5.1 (Policy Evaluation). *For any* $u \in \mathcal{U}$ *and* $\mathbf{j} \in \times_{n=1}^{N} \Omega_{nD}$,

$$R^{\text{opt}}(\mathbf{j}) - R^{u}(\mathbf{j}) = \nu(|\mathbf{\Omega}_D|)\{\mathbf{A}^{u^*}(\mathbf{\Omega}_D \mid \mathbf{j}) - \mathbf{A}^{u}(\mathbf{\Omega}_D \mid \mathbf{j})\} + \sum_{i=1}^{|\mathbf{\Omega}_D|-1} \{\nu(i+1) - \nu(i)\}\{\mathbf{A}^{u}(\mathbf{\Theta}_i \mid \mathbf{j}) - \mathbf{A}^{u^*}(\mathbf{\Theta}_i \mid \mathbf{j})\}. \tag{42}$$

As an illustration of the application of Theorem 5.1, we shall now use it to analyse the closeness to optimality of the greedy index heuristic *GI* introduced in §2 in simple finite state cases in which the following hold:

(A1) There are no switching penalties. Hence, the functions S^a_{nr}, S^b_{nr}, S^a_{nc}, and S^b_{nc} are all identically zero for all bandits n;

(A2) No rewards are earned or resource consumed under the passive action. Hence, the functions r^b_n and c^b_n are identically zero for all bandits n;

(A3) The consumption functions c^a_n are constant for each bandit n. We use c^a_n, $1 \le n \le N$, for the constant values.

(A4) Each Markov law P_n determining transitions in bandit n under the active action is irreducible, and hence, positive recurrent.

(A5) The total resource available, C, and the resource requirements of the individual bandits, c^a_n, are all multiples of resource quantum $\delta > 0$.

In light of (A1) and (A2), there is no need to incorporate into the state of each bandit information regarding the previous action taken. Hence, we can work with the (nonextended) state spaces Ω_n, $1 \le n \le N$, and their union $\mathbf{\Omega} := \bigcup_{n=1}^{N} \Omega_n$. Under (A1)–(A3) above, Definition 3.1 is simplified as follows: The return/consumption index for B is now a function $\nu\colon \Omega \to \mathbb{R}$ given by

$$\nu(x) = \sup_{\tau}\{r^a(x,\tau)/c^a(x,\tau)\}, \quad x \in \Omega, \tag{43}$$

the supremum being taken over $\tau \in T$. Note that $r^a(x,\tau)$ is given in (22), with $c^a(x,\tau)$ defined similarly. All indices are nonnegative and hence equal to their positive parts. Note also that under (A1) and (A2), Condition 1 is trivially satisfied. As above, we create a numerical representation of the members of $\mathbf{\Omega}$ by numbering them in decreasing order of their index values such that

$$\nu(|\mathbf{\Omega}|) \ge \nu(|\mathbf{\Omega}|-1) \ge \cdots \ge \nu(2) \ge \nu(1), \tag{44}$$

and use i, j for generic states within this numbering. With these adjustments, Equation (42) continues to hold, but with $\mathbf{\Omega}$ replacing $\mathbf{\Omega}_D$. Also note that under Assumptions (A1)–(A3) above the matrix $\mathbf{A}(B) := \{A_x(\Theta)\}_{x\in\Omega,\Theta\subseteq\Omega}$ defined in §4 has components whose form is greatly simplified. Equation (27) now becomes

$$A_x(\Theta) = c^a(1-\beta)^{-1} I(x \in \Theta)\big[1 - \mathbf{E}\{\beta^{\tau_x(\Theta)} \mid x\}\big], \quad x \in \Omega, \quad \Theta \subseteq \Omega. \tag{45}$$

Finally, in light of A5 above, we write

$$C = M\delta, \qquad c^a_n = m_n\delta, \quad 1 \le n \le N.$$

It will facilitate the analysis if we now introduce M additional single-state bandits labeled $m\delta$, $1 \le m \le M$. We shall use $m\delta$ to denote both one such bandit and its single state. It will assist in what follows if we use the notational shorthand $M^{\#} \equiv \{\delta, 2\delta, \dots, M\delta\}$. In the MAB problem, activation of $m\delta \in M^{\#}$ at some time $t \in \mathbb{N}$ indicates that the amount of resource *NOT* consumed by the N conventional bandits at time t is exactly $m\delta$. Hence, at most one member of $M^{\#}$ is activated at any epoch. None of the members of $M^{\#}$ earns rewards under activation, and hence all have *return/consumption index* equal to zero. We write $\mathbf{\Omega}^M := \mathbf{\Omega} \cup M^{\#}$. By definition, the resource consumed by the activated members of $\mathbf{\Omega}^M$ at each decision epoch is always exactly C under any policy.

The matrix $\mathbf{A} := \{\mathbf{A}_i(\boldsymbol{\theta})\}_{i\in\mathbf{\Omega}^M,\mathbf{\Theta}\subseteq\mathbf{\Omega}^M}$ appropriate when $\mathbf{\Omega}$ is extended to $\mathbf{\Omega}^M$ in this way has components given by

$$\mathbf{A}_i(\mathbf{\Theta}) = A_i(\mathbf{\Theta}\cap\Omega_n) = c^a_n(1-\beta)^{-1} I(i \in \mathbf{\Theta}\cap\Omega_n)\big[1 - \mathbf{E}\{\beta^{\tau_i(\mathbf{\Theta}\cap\Omega_n)} \mid i\}\big], \quad i \in \Omega_n, \quad 1 \le n \le N, \tag{46}$$

and

$$\mathbf{A}_{m\delta}(\mathbf{\Theta}) = m\delta I(m\delta \in \mathbf{\Theta}), \quad m\delta \in M^{\#}. \tag{47}$$

Note that (46) uses the appropriate form from (45) above.

We now write $\mathbf{\Theta}_j^M = \{j, j-1, \ldots, 1\} \cup M^\#$ for the union of the states from $\mathbf{\Omega}$ with j lowest *return/consumption indices* together with $M^\#$. Because all members of $M^\#$ have index equal to zero, $\mathbf{\Theta}_j^M$ may alternatively be thought of as consisting of $j+M$ members of $\mathbf{\Omega}^M$ of lowest index. We now extend the definition of the performance measures $\pi^u(i \mid \mathbf{j})$ of (33) to include $i \in M^\#$ and thereby extend (41) to

$$\mathbf{A}^u(\mathbf{\Theta} \mid \mathbf{j}) = \sum_{i \in \mathbf{\Theta}} \mathbf{A}_i(\mathbf{\Theta}) \pi^u(i \mid \mathbf{j}) \tag{48}$$

for subsets $\mathbf{\Theta} \subseteq \Omega^M$. Applying Theorem 5.1 to this setup, we have, for any $u \in \mathcal{U}$ and $\mathbf{j} \in \times_{n=1}^N \Omega_n$,

$$\begin{aligned} R^{\text{opt}}(\mathbf{j}) - R^u(\mathbf{j}) &= \nu(|\mathbf{\Omega}|)\big\{\mathbf{A}^{u^*}(\mathbf{\Omega}^M \mid \mathbf{j}) - \mathbf{A}^u(\mathbf{\Omega}^M \mid \mathbf{j})\big\} \\ &\quad + \sum_{i=0}^{|\mathbf{\Omega}|-1} \{\nu(i+1) - \nu(i)\}\big\{\mathbf{A}^u(\mathbf{\Theta}_i^M \mid \mathbf{j}) - \mathbf{A}^{u^*}(\mathbf{\Theta}_i^M \mid \mathbf{j})\big\}, \end{aligned} \tag{49}$$

where $\nu(0) = 0$, $\mathbf{\Theta}_0^M = M^\#$ in (49). However, it follows easily from (46)–(48) above that, for any $u \in \mathcal{U}$ and $\mathbf{j} \in \times_{n=1}^N \Omega_n$,

$$\mathbf{A}^u(\mathbf{\Omega}^M \mid \mathbf{j}) = C(1-\beta)^{-1},$$

and hence (49) simplifies to

$$R^{\text{opt}}(\mathbf{j}) - R^u(\mathbf{j}) = \sum_{i=0}^{|\mathbf{\Omega}|-1} \{\nu(i+1) - \nu(i)\}\big\{\mathbf{A}^u(\mathbf{\Theta}_i^M \mid \mathbf{j}) - \mathbf{A}^{u^*}(\mathbf{\Theta}_i^M \mid \mathbf{j})\big\}. \tag{50}$$

Before we apply (50) to the evaluation of the greedy index heuristic GI determined by the state ordering in (44) we require the following additional notation: We shall write $\Gamma(t)$ for the amount of resource *unused* at time t (i.e., by the N standard bandits). From the construction above, this coincides with the member of $M^\#$ chosen at t, if any, and is zero otherwise. We write

$$\Gamma(u \mid \mathbf{j}) := \mathbf{E}_u\left\{\sum_{t=0}^{\infty} \beta^t \Gamma(t) \,\middle|\, \mathbf{j}\right\}$$

for the expected total discounted amount of resource unused when policy $u \in \mathcal{U}$ is implemented from initial state $\mathbf{j}$. We further use N_i for the subset of the N standard bandits that have no intersection with $\mathbf{\Theta}_i$, namely,

$$N_i := \{n;\, 1 \le n \le N \text{ and } \Omega_n \cap \mathbf{\Theta}_i = \varnothing\}.$$

Further, we write

$$C(N_i) := \sum_{n \in N_i} c_n^a$$

for the total resource required by the bandits in N_i. Note that $N_{|\mathbf{\Omega}|} = \varnothing$, $N_0 = \{1, 2, \ldots, N\}$, and hence $C(N_{|\mathbf{\Omega}|}) = 0$, $C(N_0) > C$ with $C(N_i)$ decreasing in i. For Theorem 5.2, we define

$$I^* := \min\{i;\, C(N_i) < C\}$$

and use $O(1)$ to denote a quantity that remains bounded in the limit $\beta \to 1$.

Theorem 5.2 (Closeness to Optimality of GI). *Under the conditions* (A1)–(A5) *above, for any initial state* $\mathbf{j} \in \times_{n=1}^N \Omega_n$,

$$R^{\text{opt}}(\mathbf{j}) - R^{GI}(\mathbf{j}) \le \nu(I^*)\{\Gamma(GI \mid \mathbf{j}) - \Gamma(u^* \mid \mathbf{j})\} + O(1).$$

Proof. First note that it is plain from the definitions of the quantities concerned that

$$\mathbf{A}^u(\mathbf{\Theta}_i^M \mid \mathbf{j}) = \mathbf{A}^u(\mathbf{\Theta}_i \mid \mathbf{j}) + \mathbf{A}^u(M^\# \mid \mathbf{j}) = \mathbf{A}^u(\mathbf{\Theta}_j \mid \mathbf{j}) + \Gamma(u \mid \mathbf{j}) \tag{51}$$

for all choices of i, u, and $\mathbf{j}$. Further, it is clear that if $C(N_i) \ge C$, then the greedy index policy GI will never activate any members of $\mathbf{\Theta}_i$. We then conclude from (51) that $\forall\, \mathbf{j}$ and $i < I^*$,

$$\mathbf{A}^{GI}(\mathbf{\Theta}_i \mid \mathbf{j}) = 0$$

and hence that

$$\mathbf{A}^{GI}(\mathbf{\Theta}_i^M \mid \mathbf{j}) - \mathbf{A}^{u^*}(\mathbf{\Theta}_i^M \mid \mathbf{j}) \leq \Gamma(GI \mid \mathbf{j}) - \Gamma(u^* \mid \mathbf{j}). \tag{52}$$

Now suppose that $C(N_i) < C$ or, equivalently, that $i \geq I^*$. Write $T_i(GI, \mathbf{j})$ for the first time at which a member of $\mathbf{\Theta}_i$ is scheduled for processing when the greedy index policy is implemented from state $\mathbf{j}$ at time 0. The random time $T_i(GI, \mathbf{j})$ may be infinite with positive probability. It must be true, invoking the structure of *GI*, that at epoch $T_i(GI, \mathbf{j})$ all bandits in N_i are processed and that all bandits not being processed must have their current states in $\mathbf{\Theta}_i$. It is clear that under *GI* all bandits in N_i will continue to be processed at each $t \geq T_i(GI, \mathbf{j})$ because they can never be displaced by any of the bandits unprocessed at $T_i(GI, \mathbf{j})$. By definition of the quantities concerned, the processing of bandits in N_i contributes nothing to the measure $\mathbf{A}^{GI}(\mathbf{\Theta}_i^M \mid \mathbf{j})$ when $i \geq I^*$. It now follows that at any decision epoch $t \in [T_i(GI, \mathbf{j}), \infty)$, the residual resource $C - C(N_i)$ remaining once the bandits in N_i have been processed will *either* not be used (equivalently, will be allocated to members of $M^{\#}$) *and/or* will be used to process bandits in $\{1, 2, \ldots, N\} \backslash N_i$ that have previously entered states in $\mathbf{\Theta}_i$. This, together with the fact that, from assumption (A4) above, prior to time $T_i(GI, \mathbf{j})$, the bandits in $\{1, 2, \ldots, N\} \backslash N_i$ can only have been processed a finite number of times almost surely, yields the conclusion that when $C(N_i) < C$,

$$\mathbf{A}^{GI}(\mathbf{\Theta}_i^M \mid \mathbf{j}) = \{C - C(N_i)\}(1-\beta)^{-1} + O(1). \tag{53}$$

An argument very close to that used by Glazebrook and Garbe [11] in their analysis of the performance of parallel processor versions of Gittins index policies for conventional multiarmed bandits and which utilized a single machine relaxation of the optimization problem

$$\min_{u \in \mathcal{U}} \mathbf{A}^u(\mathbf{\Theta}_i^M \mid \mathbf{j})$$

yields the conclusion that when $C(N_i) < C$,

$$\mathbf{A}^u(\mathbf{\Theta}_i^M \mid \mathbf{j}) \geq \{C - C(N_i)\}(1-\beta)^{-1} + O(1), \quad u \in \mathcal{U}. \tag{54}$$

In fact, the inequality (54) is secured easily from the analysis of Glazebrook and Garbe [11] by regarding the resource available at each decision epoch ($C = M\delta$) as being made available by M parallel processors, each of which can supply δ units of resource. From (53) and (54), we conclude that when $C(N_i) < C$,

$$\mathbf{A}^{GI}(\mathbf{\Theta}_i^M \mid \mathbf{j}) - \mathbf{A}^{u^*}(\mathbf{\Theta}_i^M \mid \mathbf{j}) \leq O(1). \tag{55}$$

Using (52) and (55) within (50), we infer that

$$\begin{aligned} R^{\text{opt}}(\mathbf{j}) - R^{GI}(\mathbf{j}) &= \sum_{i<I^*} \{\nu(i+1) - \nu(i)\}\big\{\mathbf{A}^{GI}(\mathbf{\Theta}_i^M \mid \mathbf{j}) - \mathbf{A}^{u^*}(\mathbf{\Theta}_i^M \mid \mathbf{j})\big\} \\ &\quad \sum_{i \geq I^*} \{\nu(i+1) - \nu(i)\}\big\{\mathbf{A}^{GI}(\mathbf{\Theta}_i^M \mid \mathbf{j}) - \mathbf{A}^{u^*}(\mathbf{\Theta}_I^M \mid \mathbf{j})\big\} \\ &\leq \Bigg[\sum_{i<I^*} \{\nu(i+1) - \nu(i)\}\Bigg]\big\{\Gamma(GI \mid \mathbf{j}) - \Gamma(u^* \mid \mathbf{j})\big\} + O(1) \\ &= \nu(I^*)\big\{\Gamma(GI \mid \mathbf{j}) - \Gamma(u^* \mid \mathbf{j})\big\} + O(1), \end{aligned}$$

as required. □

The fact that the suboptimality bound for *GI* given in Theorem 5.1 must be nonnegative yields the following result.

Corollary 5.1 (Resource Consumption of *GI*). *Under the conditions* (*A*1)–(*A*5) *above, for any initial state* $\mathbf{j} \in \mathsf{X}_{n=1}^N \Omega_n$,

$$\Gamma(u^* \mid \mathbf{j}) - \Gamma(GI \mid \mathbf{j}) \leq O(1).$$

Remark. We conclude that for this simple family of multiarmed bandits, the degree of reward suboptimality of the greedy index heuristic *GI* is, to within an $O(1)$ quantity, bounded above by a multiple of the difference between the amount of available resource left unused by *GI* and the equivalent quantity for any optimal policy. Should this difference be zero (or, indeed, negative), then the greedy index policy must be within $O(1)$ of optimality.

It in turn must then follow that the total amount of resource left unused by an optimal policy (as measured by $\Gamma(u^* \mid \mathbf{j})$) must, for all initial states and to within an $O(1)$ quantity, be bounded above by the total amount of resource left unused by *GI*. This is Corollary 5.1. One way of seeing why this is so is to note that, by its construction, *GI* makes the best use (as measured by the *reward/consumption indices* and to within $O(1)$) of the resource it consumes. Hence, the only way that a policy can outperform *GI* is to use more resource. For this class of models, then, it is precisely any inability of *GI* to use the available resource as effectively as other good policies that might cause it to perform poorly. In examples where this is not a serious concern, *GI* will perform well.

We formalize some of the above ideas by developing a form of *asymptotic optimality* for *GI* not unlike that proposed for restless bandits by Whittle [30]. We develop a sequence of MABs, each structured as in (A1)–(A5) above and sharing a common discount rate β. The sequence is indexed by N, the number of competing bandits in the Nth problem, where $N \geq 2$. Objects related to problem N are so identified by adding N as an additional subscript. Hence, we have reward functions $\{r^a_{nN},\ 1 \leq n \leq N,\ N \geq 2\}$, consumption values $\{c^a_{nN},\ 1 \leq n \leq N,\ N \geq 2\}$, and total resources $\{C_N, N \geq 2\}$. We require the following additional conditions:

(A6) The reward functions $\{r^a_{nN}, 1 \leq n \leq N, N \geq 2\}$ are uniformly bounded above and away from zero;
(A7) The collection of consumptions $\{c^a_{nN}, 1 \leq n \leq N, N \geq 2\}$ are bounded above and away from zero;
(A8) The total resources $\{C_N, N \geq 2\}$ form a nondecreasing divergent sequence with

$$C_N < \sum_{n=1}^{N} c^a_{nN}, \quad N \geq 2.$$

We expand the notation $R^u(\mathbf{j})$ to $R^u_N(\mathbf{j}, \beta)$ include the index N and to express its dependence on the discount rate β.

Theorem 5.3 (Asymptotic Optimality of *GI*). *Under the conditions* (*A*1)–(*A*8) *above, for any initial state* $\mathbf{j} \in \times_{n=1}^{N} \Omega_n$,

$$\lim_{N\to\infty} \lim_{\beta\to 1} \{R^{\text{opt}}_N(\mathbf{j}, \beta) - R^{GI}_N(\mathbf{j}, \beta)\}\{R^{\text{opt}}_N(\mathbf{j}, \beta)\}^{-1} = 0.$$

Proof. In what follows we use $\bar{r}^a$, $\underline{r}^a$ to denote uniform upper and lower bounds on the reward functions, with $\bar{c}^a$, $\underline{c}^a$ the equivalent for consumptions. From (A6) and (A7) we may assume that $\underline{r}^a$ and $\underline{c}^a$ are both (strictly) positive. It is clear from (43) that $\bar{r}^a/\underline{c}^a$ is a uniform upper bound on all *reward/consumption indices*. It is also clear that $\bar{c}^a$ bounds above the amount of resource unused by *GI* at any epoch and in any problem. Hence, we deduce that, uniformly in $\mathbf{j}$,

$$\Gamma_N(GI \mid \mathbf{j}) - \Gamma_N(u^* \mid \mathbf{j}) \leq \bar{c}^a(1-\beta)^{-1}. \tag{56}$$

It is straightforward that, uniformly in $\mathbf{j}$,

$$R^{\text{opt}}_N(\mathbf{j}, \beta) \geq \left[\frac{C_N}{\bar{c}^a}\right] \underline{r}^a (1-\beta)^{-1}, \tag{57}$$

where in (57), $[y]$ denotes the integer part of y.

Now, from Theorem 5.2, (56) and (57), and the above discussion, we infer that, uniformly in $\mathbf{j}$,

$$\lim_{\beta\to 1} \{R^{\text{opt}}_N(\mathbf{j}, \beta) - R^{GI}_N(\mathbf{j}, \beta)\}\{R^{\text{opt}}_N(\mathbf{j}, \beta)\}^{-1} \leq \bar{r}^a \bar{c}^a / \underline{r}^a \underline{c}^a \left[\frac{C_N}{\bar{c}^a}\right] \to 0, \quad N \to \infty,$$

as required. □

6. Numerical study. We explore the power of the above ideas by means of a numerical investigation of the performance of two index-based heuristics in the context of a collection of four-armed bandits. As was the case with the simple class of models discussed at the conclusion of §5, we shall suppose that all the r^b_n, c^b_n, S^a_{nr}, S^b_{nr}, S^a_{nc}, and S^b_{nc} are identically zero. Hence, no rewards are earned or resource consumed by passive bandits, nor are any switching penalties incurred. All bandits have eight states, and hence each problem considered has a state space with $8^4 = 4{,}096$ elements. The study is based on 100 randomly generated problems with all individual elements chosen independently as follows: In each case, bandits 1 and 2 are to be thought of as low consumption, low reward, with bandits 3 and 4 thought of as high consumption, high reward. We shall assume that for all bandits resource consumption levels (under the active action) do not differ across states. For bandits 1 and 2 the (constant) resource consumption levels c^a_1, c^a_2 are sampled from a uniform $U(0.8, 1)$ distribution,

Table 1. The comparative performance of four heuristics for a collection of multiarmed bandit problems.

	Median percentage suboptimalities, $\beta = 0.9$				Max percentage suboptimalities, $\beta = 0.9$				Median percentage suboptimalities, $\beta = 0.95$				Max percentage suboptimalities, $\beta = 0.95$			
	GM	*M*	*GI*	*WI*	*GM*	*M*	*GI*	*WI*	*GM*	*M*	*GI*	*WI*	*GM*	*M*	*GI*	*WI*
min	0.168	0.612	0.000	0.000	1.302	2.409	1.049	0.000	0.068	0.403	0.000	0.000	0.489	1.294	0.323	0.000
lq	1.519	1.794	0.316	0.001	5.329	5.661	2.116	0.224	1.329	1.566	0.185	0.000	2.884	4.085	1.121	0.081
med	3.020	2.818	0.945	0.018	7.942	7.974	3.488	0.385	2.908	2.981	0.699	0.008	5.498	6.233	2.163	0.189
uq	5.916	4.371	2.436	0.052	11.610	11.370	6.103	0.585	6.782	5.354	2.434	0.048	9.730	10.090	4.472	0.359
max	14.140	11.530	11.100	0.196	23.690	21.880	15.710	1.465	16.900	14.020	13.100	0.223	22.320	21.370	15.840	0.830

whereas those for bandits 3 and 4 are sampled from a $U(1.8, 2)$ distribution. For bandits 1 and 2 the expected rewards earned when active in state 1, namely $r_1^a(1)$, $r_2^a(1)$ are sampled from a $U(8, 12)$ distribution, with the equivalent quantities for bandits 3 and 4, $r_3^a(1)$, $r_4^a(1)$ sampled from a $U(16, 24)$ distribution. The expected rewards earned by bandits 1 and 2 when active in any state other than 1, namely $r_1^a(x)$, $r_2^a(x)$, $2 \leq x \leq 8$, are sampled from a $U(2, 3)$ distribution, whereas a $U(4, 6)$ distribution is used for this purpose for bandits 3 and 4. In all cases, entries for the Markov transition matrix under the active action are obtained by sampling from a $U(0, 1)$ distribution and normalising across rows.

For each of the 100 problems so generated, the expected total reward under the optimal policy and four heuristics was estimated via the use of DP value iteration for each of the 4,096 initial states and for two choices of the discount rate β, namely 0.9 and 0.95. Hence, the number of total rewards computed in the study was $100 \times 4{,}096 \times 5 \times 2 = 4.096 \times 10^6$. The heuristics concerned were:

- greedy myopic (*GM*). In each state $\mathbf{x} = (x_1, x_2, x_3, x_4)$ choose the bandits for activation in decreasing order of the one-step returns $r_n^a(x_n)$ until no further resource may be consumed;
- myopic (*M*). In each state choose the bandits for activation to maximise the corresponding sum of one-step returns within the resource constraint;
- greedy index (*GI*). In each state $\mathbf{x}$ choose the bandits for activation in decreasing order of the *return/consumption indices* $\nu_n(x_n)$ until no further resource may be consumed;
- weighted index (*WI*). In each state $\mathbf{x}$ choose the bandits for activation to maximise the sum given in (14) within the resource constraint. Because here the only consumption of resource is by active bandits, the sum to be maximized simplifies to $\sum_n c_n^a(x_n)\nu_n(x_n)$.

The greedy index heuristic *GI* was subject to analysis for the simple class of models in §5 and shown to enjoy a form of asymptotic optimality. The heuristic *WI* was described in §2 and is justified informally by noting that, because indices are maximal return/consumption ratios, weighting them by one-step consumptions and aggregating yields decisions in favour of a collection of bandits capable of securing returns at maximal rate.

For each heuristic H and each choice of problem, initial state $\mathbf{x}$, and discount rate β, the *percentage suboptimality*

$$100\{R^{\text{opt}}(\mathbf{x}, \beta) - R^H(\mathbf{x}, \beta)\}\{R^{\text{opt}}(\mathbf{x}, \beta)\}^{-1}$$

was computed. For each choice of problem, heuristic H and discount rate β, the *median* and *maximum* percentage suboptimalities among the 4,096 computed (one for each initial state) were recorded. For each heuristic H and discount rate β, the 100 medians and 100 maxima so obtained were summarized by the order statistics minimum (min), lower quartile (lq), median (med), upper quartile (uq), and maximum (max). The results are presented in Table 1 below.

Although both of the index heuristics clearly outperform their myopic counterparts, much the most striking feature of the above results is the consistent excellence of the weighted index (*WI*) heuristic. In over 4×10^6 problems, on no occasion was it more than 1.465% suboptimal, whereas in well over half the cases its expected total reward was within 0.02% of optimal. Although the performance of *GI* is impressive overall, there plainly are problems for which its failure to use the available resource effectively yield significant suboptimalities, as flagged up in the discussion of Theorem 5.2 and Corollary 5.1.

Acknowledgments. Both authors acknowledge the support of the Engineering and Physical Sciences Research Council. The first author's support comes via the award of Grants GR/S45188/01 and EP/E049265/1, and the second author's support comes via a research studentship.

References

[1] Agrawal, R., M. Hedge, D. Teneketzis. 1988. Asymptotically efficient adaptive allocation rules for the multiarmed bandit problem with switching cost. *IEEE Trans. Aut. Ctrl.* **33** 899–906.

[2] Ansell, P. S., K. D. Glazebrook, J. Niño-Mora, M. O'Keeffe. 2003. Whittle's index policy for a multi-class queueing system with convex holding costs. *Math. Methods Oper. Res.* **57** 21–39.

[3] Banks, J. S., D. Sundaram. 1994. Switching costs and the Gittins index. *Econometrica* **62** 687–694.

[4] Berry, D. A., B. Fristedt. 2004. *Bandit Problems: Sequential Allocation of Experiments*. Chapman and Hall, London.

[5] Bertsimas, D., J. Niño-Mora. 1996. Conservation laws, extended polymatroids and multi-armed bandit problems: A polyhedral approach to indexable systems. *Math. Oper. Res.* **21** 257–306.

[6] Crosbie, J. H., K. D. Glazebrook. 2000. Index policies and a novel performance space structure for a class of generalised branching bandit problems. *Math. Oper. Res.* **25** 281–297.

[7] Denardo, E. V., U. G. Rothblum, L. van der Heyden. 2004. Index policies for stochastic search in a forest with an application to R.& D. project management. *Math. Oper. Res.* **29** 162–181.

[8] Gittins, J. C. 1979. Bandit processes and dynamic allocation indices (with discussion). *J. Roy. Statist. Soc.* **B41** 148–177.

[9] Gittins, J. C. 1989. *Multi-Armed Bandit Allocation Indices*. John Wiley, Chichester, UK.

[10] Glazebrook, K. D. 1980. On stochastic scheduling with precedence relations and switching costs. *J. Appl. Probab.* **17** 1016–1024.

[11] Glazebrook, K. D., R. Garbe. 1999. Almost optimal policies for stochastic systems which almost satisfy conservation laws. *Ann. Oper. Res.* **99** 19–43.

[12] Glazebrook, K. D., C. Kirkbride. 2004. Index policies for the routing of background jobs. *Naval Res. Logist.* **52** 381–398.

[13] Glazebrook, K. D., C. Kirkbride. 2007. Dynamic routing to heterogeneous collections of unreliable servers. *Queueing Systems* **55** 9–25.

[14] Glazebrook, K. D., C. Kirkbride, D. Ruiz-Hernandez. 2006a. Spinning plates and squad systems—Policies for bi-directional restless bandits. *Adv. Appl. Probab.* **38** 95–115.

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

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

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

[18] Mahajan, A., D. Teneketzis. 2007. Multiarmed bandit problems. A. Hero, D. Castanon, D. Cochran, K. Kastella, eds. *Foundations and Applications of Sensor Management*.

[19] Nash, P. 1980. A generalized bandit problem. *J. Roy. Statist. Soc.* **B42** 165–169.

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

[21] Niño-Mora, J. 2002. Dynamic allocation indices for restless projects and queueing admission control: A polyhedral approach. *Math. Program.* **93** 361–413.

[22] Niño-Mora, J. 2007. A $(2/3)n^3$ fast-pivoting algorithm for the Gittins index and optimal stopping of a Markov chain. *INFORMS J. Comput.* **19** 596–606.

[23] Opp, M., K. D. Glazebrook, V. Kulkarni. 2005. Outsourcing warranty repairs—Dynamic allocation. *Naval Res. Logist.* **52** 381–398.

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

[25] Puterman, M. L. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley, New York.

[26] Reiman, M. I., L. M. Wein. 1998. Dynamic scheduling of a two-class queue with set-ups. *Oper. Res.* **46** 532–547.

[27] Robinson, D. R. 1982. Algorithms for evaluating the dynamic allocation index. *Oper. Res. Lett.* **1** 72–74.

[28] Van Oyen, M. P., D. Teneketzis. 1994. Optimal stochastic scheduling of forest networks with switching penalties. *Adv. Appl. Probab.* **26** 474–479.

[29] Weber, R. R., G. Weiss. 1990. On an index policy for restless bandits. *J. Appl. Probab.* **27** 637–648, (Addendum: *Adv. Appl. Probab.*, **23** 1991, 429–430).

[30] Whittle, P. 1988. Restless bandits: Activity allocation in a changing world. J. Gani, ed. *A Celebration of Applied Probability*. *J. Appl. Probab. Special Volume* **25**A 287–298. Probability Trust, Sheffield, UK.