Adv. Appl. Prob. **38,** 95–115 (2006)
Printed in Northern Ireland

SPINNING PLATES AND SQUAD SYSTEMS: POLICIES FOR BI-DIRECTIONAL RESTLESS BANDITS

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

Abstract

This paper concerns two families of Markov decision problem that fall within the family of (bi-directional) restless bandits, an intractable class of decision processes introduced by Whittle. The *spinning plates problem* concerns the optimal management of a portfolio of reward-generating assets whose yields grow with investment but otherwise tend to decline. In the model of asset exploitation called the *squad system*, the yield from an asset tends to decline when it is used but will recover when the asset is at rest. In all cases, simply stated conditions are given that guarantee indexability of the problem, together with conditions necessary and sufficient for its strict indexability. The index heuristics for asset activation that emerge from the analysis are assessed numerically and found to perform very strongly.

Keywords: Index policy; Lagrangian method; Markov decision problem; restless bandit; stochastic dynamic programming

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

1. Introduction

In entertainment shows of a certain vintage, a popular act featured a performer keeping a large number of plates spinning on the top of flexible poles. The audience would express dismay when one of the plates started to wobble badly, prompting urgent attention to prevent it from falling from its stick. The performer's problem (of keeping the plates spinning) is a vivid metaphor for that facing a manager responsible for a collection of reward-generating assets, each of whose (reward) performance is enhanced in time by an active (investment) intervention, but otherwise tends to deteriorate. The crucial issue arises of how such interventions should be organized to maximize the overall reward yield from an entire asset portfolio. In Section 2, the performer's/manager's problem is formulated as a Markov decision problem with the average reward criterion. In honour of the frivolous application cited above, we call this the *spinning plates problem*.

In contrast to the above are situations in which a manager has a large number of reward-generating assets at her disposal, a fixed number of which need to be deployed/exploited at all times. Deployment of an asset activates its reward stream, but erodes over time its (reward) performance. Resting (not deploying) an asset allows it to recover. The key issue here concerns

Received 28 April 2005; revision received 14 October 2005.
* Postal address: Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, UK. Email address: k.glazebrook@lancaster.ac.uk
** Postal address: Department of Management Science, Lancaster University, Lancaster LA1 4YX, UK.
*** Postal address: Department of Economics and Business, Universitat Pompeu Fabra, Barcelona, E-08005, Spain.

how assets should be deployed so as to maximize the rewards earned from them over time. In Section 2 this problem is also formulated as a Markov decision problem with the average reward criterion. In honour of the similar problem faced by coaches in professional sports, we call this problem the *squad system*. To the authors' knowledge, the spinning plates problem, as formulated in Section 2, is new and there is no previous literature on it. Whittle (1988) gave a brief discussion of a particular case of the squad system that had a linear structure for both rewards and stochastic dynamics. He called this the *Ehrenfest project*. Niño-Mora (2001a) discussed a discounted reward version of the squad system using polyhedral methods, but was not able to employ this analysis to give an account of the system with the average reward criterion.

The Markov decision processes concerned are formulated and presented in Section 2 and all fall within the class of so-called *restless bandit problems* introduced by Whittle (1988). These form a class of decision process that generalizes the *multiarmed bandit* of Gittins (1979), (1989) by allowing passive evolution. Complexity analyses due to Papadimitriou and Tsitsiklis (1999) imply that restless bandit problems are almost certainly intractable. Whittle (1988) proposed a class of *index heuristics* which extend the *Gittins index policies* that are optimal for multiarmed bandits. Under Whittle's proposal, each asset has a *calibrating index*, which is a function of its state, and his heuristic activates that asset (or those assets) whose current index value is largest. However, Whittle's proposed asset index may not exist (this is the issue of *indexability*) and the resulting policy will not in general be optimal, even for indexable problems. That said, Weber and Weiss (1990), (1991) have demonstrated a form of asymptotic optimality of Whittle's heuristic under certain conditions. Furthermore, Glazebrook *et al.* (2002) have explored the closeness to optimality of the index policy for a class of discounted restless bandit problems of simple structure. Applications of Whittle's ideas to the control of multiclass queueing systems have provided empirical evidence of outstanding performance of the index heuristics concerned. See, for example, Ansell *et al.* (2003) and Glazebrook *et al.* (2003). In general, the issue of whether a restless bandit problem is indexable is complex. Niño-Mora (2001b), (2001a), (2002) has used a polyhedral approach to express conditions on model parameters that guarantee indexability.

In Sections 3 and 5 we give simple and direct accounts of the index structure of, respectively, the spinning plates problem and the squad system. In both cases we give simply stated conditions that guarantee the models' indexability. Furthermore, we present algorithms that yield the indices. *Strict indexability* means that not only is the problem concerned indexable, but also that all index functions are one-to-one (namely, that distinct states of an asset have distinct index values). Our analysis yields conditions necessary and sufficient for strict indexability in both models, together with formulae for the indices in closed form. The authors believe this to be the first time that simply stated conditions equivalent to strict indexability have been formulated for any restless bandit model for which strict indexability is not guaranteed. Numerical results testify to the very strong performance of the index heuristic for both models. In 800 instances of the spinning plates problem, the index heuristic was never more than 0.024% suboptimal. Section 4 contains a somewhat shorter discussion of the index structure of a version of the spinning plates problem with the discounted reward criterion. The equivalent material for the squad system is to be found at the conclusion of Section 5.

In addition to the intrinsic interest of the theoretical results in Sections 3 and 5, the authors believe that the approach adopted will be applicable to a wide range of restless bandit problems with the average reward criterion. Investigation of an asset's index structure involves study of the so-called *W-subsidy problem*. The latter is a decision problem defined for the asset of interest in which a subsidy W is paid for every unit of time for which the asset is passive.

Indexability of the asset is related simply to the fact that the value of the W-subsidy problem is increasing, piecewise linear, and convex in W (see also Niño-Mora (2002)). For strictly indexable cases the number of pieces in the piecewise-linear function is greater by 1 than the number of asset states.

2. Two families of bi-directional restless bandit

Each of the families of restless bandit considered here is a class of Markov decision process with the average reward criterion. In each case, J projects (assets) are available for investment/exploitation. Resource constraints mean that only M assets ($1 \leq M < J$) may be active at any time. The decision problem concerns how assets should be optimally chosen for activation at each decision epoch of the system to maximize the reward rate earned over an infinite horizon.

Definition 1. (*Family 1: Spinning plates (investment in assets).*) A typical member of this family is as follows.

(i) Each of the assets evolves stochastically through time $t \in \mathbb{R}^+$. We write $X_j(t)$ for the *state* of the asset j at time $t \in \mathbb{R}^+$, $1 \leq j \leq J$, and $\boldsymbol{X}(t) = \{X_1(t), X_2(t), \ldots, X_J(t)\}$ for the corresponding *system state*. The state of asset j is an integer in the range $[\underline{K}_j, \overline{K}_j] \equiv \{\underline{K}_j, \underline{K}_j + 1, \ldots, \overline{K}_j\}$, and for most of the development (and until stated otherwise) we shall suppose that $-\infty < \underline{K}_j < \overline{K}_j < \infty$, $1 \leq j \leq J$.

(ii) Time 0 together with the times of every state transition of the process constitute the set of decision epochs for the system. In each system state there are $\binom{J}{M}$ possible actions, one corresponding to each subset of $\{1, 2, \ldots, J\}$ of size M. If S is one such subset, then $A(S)$ denotes the action that chooses both an active regime (the active action, denoted a) for the assets whose identifiers are in S, and an inactive regime (the passive action, denoted b) for the remaining assets. Under action $A(S)$ applied in state $\boldsymbol{x}$, the time to the next system transition is exponentially distributed with rate

$$\sum_{j \in S} \lambda_j(x_j) + \sum_{j \notin S} \mu_j(x_j) =: \Delta(S, \boldsymbol{x}).$$

If $\Delta(S, \boldsymbol{x}) > 0$ then the state immediately following this transition will be $\boldsymbol{x} + \boldsymbol{e}_j$, $j \in S$, with probability $\lambda_j(x_j)\{\Delta(S, \boldsymbol{x})\}^{-1}$, and will be $\boldsymbol{x} - \boldsymbol{e}_j$, $j \notin S$, with probability $\mu_j(x_j)\{\Delta(S, \boldsymbol{x})\}^{-1}$. Note that $\boldsymbol{e}_j$ is a J-vector whose jth component is 1, with 0s elsewhere. Equivalently, the J assets evolve independently under the action applied (a or b). If project j should be active (a) then it evolves from state x_j to $x_j + 1$ at rate $\lambda_j(x_j)$, while under the passive action (b) it evolves from state x_j to $x_j - 1$ at rate $\mu_j(x_j)$, $x_j \in [\underline{K}_j, \overline{K}_j]$, $1 \leq j \leq J$. The transition rates λ_j and μ_j, $1 \leq j \leq J$, satisfy

$$\lambda_j(\overline{K}_j) = \mu_j(\underline{K}_j) = 0$$

but are otherwise strictly positive. If $\Delta(S, \boldsymbol{x}) = 0$ then the state $\boldsymbol{x}$ is absorbing under action $A(S)$.

(iii) The system earns rewards at rate $\sum_j R_j(x_j)$ while in state $\boldsymbol{x}$. Each reward rate function $R_j : [\underline{K}_j, \overline{K}_j] \to \mathbb{R}^+$ is (weak-sense) increasing. The goal of the analysis is to determine a policy (a rule for taking actions) that maximizes the average system reward rate earned over an infinite horizon, or comes close to doing so.

Definition 2. (*Family 2: The squad system (exploitation of assets).*) A typical member of this family is as follows.

(i) The states of the system are as in Definition 1(i).

(ii) The available actions in each state are as in Definition 1(ii). Now, however, under action $A(S)$ applied in state $\boldsymbol{x}$, the time to the next system transition is exponential with rate

$$\sum_{j\in S} \nu_j(x_j) + \sum_{j\notin S} \rho_j(x_j) =: \Upsilon(S,\boldsymbol{x}).$$

If $\Upsilon(S,\boldsymbol{x}) > 0$ then the state immediately following this transition will be $\boldsymbol{x} - \boldsymbol{e}_j$, $j \in S$, with probability $\nu_j(x_j)\{\Upsilon(S,\boldsymbol{x})\}^{-1}$, and will be $\boldsymbol{x} + \boldsymbol{e}_j$, $j \notin S$, with probability $\rho_j(x_j)\{\Upsilon(S,\boldsymbol{x})\}^{-1}$. Equivalently, the J assets evolve independently under the action applied (a or b). If project j should be active (a) then it evolves from x_j to $x_j - 1$ at rate $\nu_j(x_j)$, while under the passive action (b) it evolves from x_j to $x_j + 1$ at rate $\rho_j(x_j)$, $x_j \in [\underline{K}_j, \overline{K}_j]$, $1 \le j \le J$. The transition rates satisfy $\nu_j(\underline{K}_j) = \rho_j(\overline{K}_j) = 0$, $1 \le j \le J$, but are otherwise strictly positive. If $\Upsilon(s,\boldsymbol{x}) = 0$ then the state $\boldsymbol{x}$ is absorbing under action $A(S)$.

(iii) If the system is in state $\boldsymbol{x}$ and action $A(S)$ is current, then the system earns rewards at rate $\sum_{j\in S} R_j(x_j)$, where each function $R_j\colon [\underline{K}_j, \overline{K}_j] \to \mathbb{R}^+$ is (weak-sense) increasing. The goal of the analysis is to determine a policy that maximizes the average system reward rate earned over an infinite horizon, or comes close to doing so.

Remarks 1. 1. In family 1, the active action applied to an asset enhances its reward-earning capacity. Hence, plant and machinery are maintained and updated, employees are trained, and products are improved and/or advertised – in short, activity represents a positive investment decision taken with regard to an asset. In the absence of such investment decisions (i.e. under the passive action) the reward-earning capacity of an asset tends to decline. Note from Definition 1(iii) that in family 1 assets earn rewards (at a higher or lower rate) all the time and not only when in receipt of investment.

2. In family 2, the active action represents the use or exploitation of an asset. As the asset is used it becomes 'tired' or depleted and loses some of its reward-earning capacity. Under the passive action the asset recovers its potential to earn high returns. Note from Definition 2(iii) that in family 2 assets only earn rewards when they are used (i.e. under the active action).

3. The reward structures in Definitions 1(iii) and 2(iii) are natural to the envisaged applications. Note that the modification of family 1 in which assets only earn rewards under the active action has a trivial solution: always apply the active action to those M assets with the largest associated values of $R_j(\overline{K}_j)$, $1 \le j \le J$. Also, the version of family 2 in which assets earn rewards whether activated or not is of little interest. No policy can do better in reward rate terms than an application of the passive action to all assets always.

4. In both cases, the theory of stochastic dynamic programming guarantees the existence of an optimal policy that is stationary, deterministic, and Markov (Puterman (1994, pp. 353–361)). The above families fall within the class of intractable restless bandit problem, introduced by Whittle (1988) and demonstrated to be PSPACE-hard by Papadimitriou and Tsitsiklis (1999). Whittle (1988) advocated the deployment of index heuristics, such policies emerging from the formulation and solution of Lagrangian relaxations of the original optimization problem. We

sketch the essentials of this approach before exploring its implications for the above families in detail in the following sections.

Write $\mathcal{U}$ for the class of stationary, deterministic, Markov policies for an identified member of either family 1 or family 2, and $u \in \mathcal{U}$ for an individual policy. We use $r_j(u)$ for the average reward rate earned by asset j under policy u. The optimization problem of interest is expressed as

$$r^{\text{opt}} = \max_{u \in \mathcal{U}} \left\{ \sum_{j=1}^{J} r_j(u) \right\}. \tag{1}$$

We now relax the optimization problem in (1) by considering schemes that activate *any number of assets* at each decision epoch (i.e. any number between 0 and J, not necessarily M) and use $\mathcal{U}'$ to denote the class of policies that do this in a stationary, deterministic, Markov way. Our interest will reside in those members of $\mathcal{U}'$ that activate M assets (or, equivalently, fail to activate $J - M$ assets) *on average* over an infinite horizon. To formulate the corresponding optimization problem, write $I_j(u)$ for the proportion of time for which asset j is passive under $u \in \mathcal{U}'$. Hence, we relax (1) to

$$\bar{r}^{\text{opt}} = \max_{u \in \mathcal{U}'} \left\{ \sum_{j=1}^{J} r_j(u) \right\} \tag{2}$$

subject to

$$\sum_{j=1}^{J} I_j(u) = J - M. \tag{3}$$

Plainly, the relaxation yields increased optimal rewards (compared to the original problem), and, hence, $\bar{r}^{\text{opt}} \geq r^{\text{opt}}$.

We now incorporate constraint (3) in a Lagrangian fashion. We write

$$r(W) = \max_{u \in \mathcal{U}'} \left\{ \sum_{j=1}^{J} \{ r_j(u) + W I_j(u) \} - W(J - M) \right\} \tag{4}$$

$$= \sum_{j=1}^{J} \left[\max_{u_j \in \mathcal{U}'_j} \{ r_j(u) + W I_j(u) \} \right] - W(J - M), \tag{5}$$

where in (4) and (5) W is a Lagrange multiplier that has an economic interpretation as a *subsidy for passivity*. The additive nature of the objective in (4) together with the character of policy set $\mathcal{U}'$ means that the optimal activation scheme for the entire set of assets is achieved by concatenating optimal activation schemes for the individual assets. The additive decomposition in expression (5) is the consequence. In (5), $\mathcal{U}'_j$ is the set of stationary, deterministic, Markov policies that choose between actions a and b for asset j (alone), $1 \leq j \leq J$. The optimization problem

$$r_j(W) = \max_{u_j \in \mathcal{U}'_j} \{ r_j(u) + W I_j(u) \} \tag{6}$$

is called the *W-subsidy problem* for asset j and aims to choose a policy for activating j to maximize its overall return from rewards earned and passive subsidies received. Since expressions (2) and (4) are equal when constraint (3) is satisfied, it is plain that $r(W) \geq \bar{r}^{\text{opt}} \geq r^{\text{opt}}$ for all W.

An issue that arises in consideration of the W-subsidy problem in (6) is possible non-uniqueness of the policy or policies achieving the maximum. We resolve any such non-uniqueness in two steps. First, we demonstrate (see Lemmas 1 and 3) that for both families 1 and 2 there exist optimal policies for the W-subsidy problems of interest which have *monotone structure*. Hence, we restrict the analysis to policies from the appropriate monotone classes in each case. Second, should more than one monotone policy achieve the maximum in (6), then we choose the policy with the largest *passive set* (i.e. the largest set of states in which the corresponding policy chooses the passive action). Use $\pi_j(W)$ to denote the resulting policy, $b_j(W)$ to denote its passive set, and $\pi(W)$ to denote the policy for the entire system that applies $\pi_j(W)$ to each asset j, $1 \leq j \leq J$. Policy $\pi(W)$ solves the optimization problem (4). The following definition expresses a natural requirement on (optimal) policy structure.

Definition 3. Asset j is *indexable* if there exist $\underline{W}_j$ and $\overline{W}_j$ such that $-\infty < \underline{W}_j < \overline{W}_j < \infty$, $b_j(W) = \varnothing$ for $W < \underline{W}_j$, and $b_j(W) = [\underline{K}_j, \overline{K}_j]$ for $W \geq \overline{W}_j$, with $b_j : [\underline{W}_j, \overline{W}_j] \to 2^{[\underline{K}_j, \overline{K}_j]}$ increasing.

The above decision problems are indexable when all constituent projects are.

Should an asset be indexable then a natural calibration in the form of a *fair subsidy for passivity* may be defined.

Definition 4. If asset j is indexable then its index $W_j : [\underline{K}_j, \overline{K}_j] \to \mathbb{R}$ is defined by

$$W_j(x) = \inf\{W,\ x \in b_j(W)\}, \qquad x \in [\underline{K}_j, \overline{K}_j].$$

It now follows that $\pi(W)$ will choose to activate in system state $\boldsymbol{x}$ those assets for which $W_j(x_j) > W$, and apply the passive action to the remainder. Furthermore, if there exists a W^* such that $\pi(W^*)$ satisfies (3) then it must follow that

$$r(W^*) = \inf_W r(W) = \bar{r}^{\text{opt}} \geq r^{\text{opt}}$$

and $\pi(W^*)$ solves the relaxed optimization problem in (2) and (3). A natural index heuristic for the original optimization problem emerges from the above discussion. The heuristic chooses in state $\boldsymbol{x}$ to activate M assets with maximal index values $W_j(x_j)$, $1 \leq j \leq J$, with ties resolved in some arbitrary manner.

In Sections 3 and 5 we shall study families 1 and 2 in turn. In each case we shall give conditions sufficient for the indexability of the decision problems together with algorithms that yield the resulting indices. We further give conditions necessary and sufficient for the strict indexability of each asset, namely that the index function be one-to-one. Under strict indexability, the indices are available in closed form. Section 4 contains a discussion of the index structure of the spinning plates problem with the discounted reward criterion.

3. Family 1 (spinning plates): a model for optimal investment in assets

We now drop the asset suffix and consider the W-subsidy problem in (6) for a single asset drawn from a decision problem in family 1 whose associated parameters are $\overline{K}$, $\underline{K}$, $\lambda(\cdot)$, $\mu(\cdot)$, and $R(\cdot)$. From Definitions 1(ii) and 1(iii), recall that, under the application of the active action a in state x, the asset evolves to state $x+1$ at rate $\lambda(x)$ and earns rewards at rate $R(x)$ while doing so. Under application of the passive action b in state x, the asset evolves to state $x-1$ at rate $\mu(x)$ and (in the W-subsidy problem) earns rewards at rate $R(x) + W$ while doing so.

The intermediate goal of our analysis is the identification of policies that maximize the average reward rate earned by the asset over an infinite horizon.

Without loss of generality, we restrict to the class of stationary, deterministic, Markov policies $\pi : [\underline{K}, \overline{K}] \to \{a, b\}$ and highlight the class B of *monotone policies* for which

$$\pi(x) = b \Leftrightarrow x \geq y \quad \text{for some } y \in [\underline{K}, \overline{K} + 1]. \tag{7}$$

We shall denote the policy in (7) by $(y),\ y \in [\underline{K}, \overline{K} + 1]$. Note that $(\overline{K} + 1)$ chooses the active action a in all states while $(\underline{K})$ chooses the passive action b in all states.

Lemma 1. *For all $W \in \mathbb{R}$ there exists an optimal policy for the W-subsidy problem in B.*

Proof. Fix a W and an initial asset state $\hat{x}$. Consider asset evolution under a general stationary, deterministic, Markov policy π. The average reward rate earned under π will always be matched by that earned by some member of B from *any* initial state.

Suppose, for example, that $\pi(\hat{x}) = a$ and that $\pi(x) = b$ for some $x,\ \overline{K} \geq x > \hat{x}$. Write

$$\bar{x} = \min\{x : x > \hat{x} \text{ and } \pi(x) = b\}.$$

Under π, the asset reaches state $\bar{x}$ in finite time almost surely, and thereafter has alternating sojourns in states $\bar{x}$ and $\bar{x} - 1$. The associated average reward rate for the W-subsidy problem is

$$[\{W + R(\bar{x})\}\lambda(\bar{x} - 1) + R(\bar{x} - 1)\mu(\bar{x})]\{\lambda(\bar{x} - 1) + \mu(\bar{x})\}^{-1}.$$

This is also the average reward rate achieved by policy $(\bar{x})$ from *any* initial state. The remaining cases, that is,

- $\pi(x) = a,\ \overline{K} \geq x \geq \hat{x}$,
- $\pi(x) = b,\ \hat{x} \geq x \geq \underline{K}$,
- $\pi(\hat{x}) = b$ and $\pi(x) = a$ for some $\hat{x} > x \geq \underline{K}$,

are dealt with similarly. The required result follows.

From Lemma 1, policy $(x(W))$, where

$$x(W) \in \underset{\underline{K} \leq x \leq \overline{K}+1}{\operatorname{argmax}} \{[\{W + R(x)\}\lambda(x - 1) + R(x - 1)\mu(x)]\{\lambda(x - 1) + \mu(x)\}^{-1}\}, \tag{8}$$

solves the W-subsidy problem. Note that in the event of more than a single x-value achieving the maximum on the right-hand side of (8), $x(W)$ is taken to be the smallest such value. Also note that, in (8), $\lambda(\underline{K} - 1)$ and $\mu(\overline{K} + 1)$ are assigned arbitrary positive values.

From Definition 3, in order to establish the asset's indexability it will be enough to show that there exist finite $\underline{W}$ and $\overline{W}$, $\underline{W} < \overline{W}$, such that $x(W) = \overline{K} + 1$ for $W < \underline{W}$, $x(W) = \underline{K}$ for $W \geq \overline{W}$, and $x(\cdot) : [\underline{W}, \overline{W}] \to [\underline{K}, \overline{K} + 1]$ is decreasing. For an indexable asset, the index in state x will be given by

$$W(x) = \inf\{W : x(W) \leq x\}, \qquad \underline{K} \leq x \leq \overline{K}. \tag{9}$$

We now introduce the function $\phi : [\underline{K}, \overline{K} + 1] \to [0, 1]$, defined by

$$\phi(x) = \lambda(x - 1)\{\lambda(x - 1) + \mu(x)\}^{-1}, \qquad \underline{K} \leq x \leq \overline{K} + 1.$$

Note that $\phi(\overline{K}+1)=0$ and $\phi(\underline{K})=1$, and observe that, from (8), the average reward rate achieved by policy (x) for the W-subsidy problem is written

$$\{W+R(x)\}\phi(x)+R(x-1)\{1-\phi(x)\}, \qquad \underline{K}\le x\le \overline{K}+1.$$

Theorem 1. *If ϕ is decreasing then the asset is indexable.*

Proof. Define, for $W\ge 0$,

$$V(W)=\max_{\underline{K}\le x\le \overline{K}+1}\{\{W+R(x)\}\phi(x)+R(x-1)\{1-\phi(x)\}\}. \tag{10}$$

From the discussion following (8), $x(W)$ is the smallest maximizer of the right-hand side of (10). It is straightforward to show that, since $0<\phi(x)<1$ for $\underline{K}+1\le x\le \overline{K}$, we have

$$x(W)=\overline{K}+1, \qquad W<0, \tag{11}$$

and

$$x(W)=\underline{K}, \qquad W\ge \hat{W}, \tag{12}$$

for some sufficiently large $\hat{W}$. Furthermore, if the subset $\{x_1, x_1+1,\dots,x_2\}\subseteq[\underline{K},\overline{K}+1]$ is such that $\phi(x_1)=\phi(x_1+1)=\cdots=\phi(x_2)$, and is maximal in this regard, then it is straightforward to show that the range of $x(W)$ contains at most a single value from this subset.

By standard theory, $V:\mathbb{R}^+\to\mathbb{R}^+$ is convex and piecewise linear, and is easily seen to be increasing. Suppose now that $0\le W_1<W_2$. Since clearly $\phi(x(W))$ is the right gradient of V for every $W\ge 0$, it immediately follows from the convexity of V, the hypothesis of the theorem, and the foregoing discussion that

$$x(W_1)\ge x(W_2). \tag{13}$$

The result now follows from (10)–(13) and the discussion around (8).

We now seek to understand the asset's index structure under the hypothesis of Theorem 1. Suppose that there are $L\ge 0$ points at which the gradient of V is discontinuous. List the corresponding W-values as

$$0<W^1<\cdots<W^L,$$

where plainly $W^L\le\hat{W}$, from (12). Write $W^0=0$. We use x^l, $0\le l\le L-1$, for the integers for which $x(W)=x^l$, $W\in(W^l,W^{l+1})$, and which satisfy

$$\phi(x^l)=\{V(W^{l+1})-V(W^l)\}(W^{l+1}-W^l)^{-1}, \qquad 0\le l\le L-1.$$

Also, write $x^L=\underline{K}$. The convexity of V and the decreasing nature of ϕ imply that

$$\overline{K}+1\ge x^0>x^1>\cdots>x^L=\underline{K}.$$

We now complete the description of $x(\cdot)$ as follows:

$$x(W)=\begin{cases}\overline{K}+1, & W<0,\\ x^l, & W\in[W^l,W^{l+1}),\ 0\le l\le L-1,\\ \underline{K}, & W\ge W^L.\end{cases} \tag{14}$$

Our next result is an immediate consequence of (9) and (14).

Theorem 2. *If ϕ is decreasing then the index $W\colon [\underline{K}, \overline{K}] \to \mathbb{R}^+$ is given by*

$$W(x) = \begin{cases} 0, & x^0 \le x \le \overline{K}, \\ W^l, & x^l \le x < x^{l-1},\ 1 \le l \le L. \end{cases}$$

Remarks 2. 1. Note from Theorem 2 that the index is decreasing in the state. Hence, in the spinning plates problem it is assets which are achieving low returns (wobbly plates) that are a high priority for activation.

2. Also note that the sufficient condition of Theorems 1 and 2 – that ϕ be decreasing – is equivalent to the requirement that the ratio between the active and passive rates for moving between $x-1$ and x, i.e. $\lambda(x-1)/\mu(x)$, be decreasing in x.

Lemma 2 gives an inductive specification of the key integers $\{x^l,\ 0 \le l \le L\}$ and the reals $\{W^l,\ 0 \le l \le L\}$. To state the result we require additional notation. Suppose x to be an integer such that $\underline{K} < x \le \overline{K} + 1$. We write

$$m(x) = \max\{y \in \mathbb{Z}\colon \underline{K} \le y < x \text{ and } \phi(y) > \phi(x)\}.$$

Lemma 2. *The collections $\{x^l,\ 0 \le l \le L\}$ and $\{W^l,\ 0 \le l \le L\}$ are as follows.*

(i) *The initial values are $x^0 = x(0)$ and $W^0 = 0$.*

(ii) *If $k \in \mathbb{N}$ and $x^k > \underline{K}$ then*

$$W^{k+1} = \min_{\underline{K} \le x \le m(x^k)} \{\{[R(x^k)\phi(x^k) + R(x^k - 1)\{1 - \phi(x^k)\}] - [R(x)\phi(x) + R(x-1)\{1-\phi(x)\}]\}\{\phi(x) - \phi(x^k)\}^{-1}\}, \quad (15)$$

with x^{k+1} the smallest minimizer in (15). If $x^{k+1} = \underline{K}$ then $L = k+1$.

Proof. Part (i) is immediate from the above. For part (ii), note that, from (14), if $x^k > \underline{K}$ then W^{k+1} can be characterized as

$$W^{k+1} = \inf\{W > W^k\colon x(W) < x^k\}.$$

If $W \in (W^k, W^{k+1})$ then x^k is the smallest maximizer of $V(W)$ and, in particular, attains a value strictly greater than that attained by any x for which $x \le m(x^k)$. Hence,

$$\{W + R(x)\}\phi(x) + R(x-1)\{1-\phi(x)\} < \{W + R(x^k)\}\phi(x^k) + R(x^k - 1)\{1 - \phi(x^k)\}, \qquad \underline{K} \le x \le m(x^k),$$

from which we deduce that

$$W < \{[R(x^k)\phi(x^k) + R(x^k - 1)\{1 - \phi(x^k)\}] - [R(x)\phi(x) + R(x-1)\{1-\phi(x)\}]\}\{\phi(x) - \phi(x^k)\}^{-1}, \qquad \underline{K} \le x \le m(x^k). \quad (16)$$

It immediately follows from (16) that

$$W^{k+1} \le \min_{\underline{K} \le x \le m(x^k)} \{\{[R(x^k)\phi(x^k) + R(x^k - 1)\{1 - \phi(x^k)\}] - [R(x)\phi(x) + R(x-1)\{1-\phi(x)\}]\}\{\phi(x) - \phi(x^k)\}^{-1}\}. \quad (17)$$

However, the fact that $x(W^{k+1}) < x^k$ must mean that $x(W^{k+1}) \leq m(x^k)$ and, hence, that there are maximizers of $V(W^{k+1})$ in the range $\underline{K} \leq x \leq m(x^k)$. For any such maximizer, $\hat{x}$ say, we have

$$\{W^{k+1} + R(\hat{x})\}\phi(\hat{x}) + R(\hat{x}-1)\{1-\phi(\hat{x})\}$$
$$\geq \{W^{k+1} + R(x^k)\}\phi(x^k) + R(x^k-1)\{1-\phi(x^k)\}$$

and, hence,

$$\begin{aligned} W^{k+1} &\geq \{[R(x^k)\phi(x^k) + R(x^k-1)\{1-\phi(x^k)\}] \\ &\quad - [R(\hat{x})\phi(\hat{x}) + R(\hat{x}-1)\{1-\phi(\hat{x})\}]\}\{\phi(\hat{x}) - \phi(x^k)\}^{-1} \\ &\geq \min_{\underline{K} \leq x \leq m(x^k)} \{\{[R(x^k)\phi(x^k) + R(x^k-1)\{1-\phi(x^k)\}] \\ &\quad - [R(x)\phi(x) + R(x-1)\{1-\phi(x)\}]\}\{\phi(x) - \phi(x^k)\}^{-1}\}. \end{aligned} \tag{18}$$

Equation (15) follows from (17) and (18).

By a modest extension of the above calculations, it follows that the minimizers on the right-hand side of (15) are precisely the maximizers of $V(W^{k+1})$. By definition, $x(W^{k+1}) = x^{k+1}$ is the smallest of these. This completes the proof.

Important special cases occur in which all of the states in an indexable asset have distinct indices. When this happens we say that the asset is strictly indexable. The next result gives a condition necessary and sufficient for strict indexability.

Theorem 3. (i) *The following assertions are equivalent.*

(a) *The asset is strictly indexable.*

(b) *Both $\phi(x)$ and*

$$\begin{aligned} \overline{W}(x) &:= \{[R(x+1)\phi(x+1) + R(x)\{1-\phi(x+1)\}] \\ &\quad - [R(x)\phi(x) + R(x-1)\{1-\phi(x)\}]\}\{\phi(x) - \phi(x+1)\}^{-1} \end{aligned} \tag{19}$$

are strictly decreasing over the range $\underline{K} \leq x \leq \overline{K}$.

(ii) *Under the conditions of part (i)(b) we have $W(x) = \overline{W}(x)$, $\underline{K} \leq x \leq \overline{K}$, and the index is strictly decreasing in the state.*

Proof. (i) To prove that assertion (a) follows from assertion (b), note that if the hypotheses of (b) hold then it is straightforward to show that the minimum in (15) will be attained uniquely by $x^{k+1} = x^k - 1$. We will then have $W^{k+1} = \overline{W}(x^k - 1)$, and the inference of strict indexability will follow simply from Theorem 2.

Now assume that assertion (a) holds. If the asset is strictly indexable, it must follow that there exists a function $\tilde{W}(x)$, $\underline{K} \leq x \leq \overline{K}$, strictly decreasing in x, such that

$$x(W) = \begin{cases} \overline{K}+1, & W < \tilde{W}(\overline{K}), \\ x, & W \in [\tilde{W}(x), \tilde{W}(x-1)),\ \underline{K}+1 \leq x \leq \overline{K}, \\ \underline{K}, & W \geq \tilde{W}(\underline{K}). \end{cases}$$

Plainly, when $W = \tilde{W}(x)$, $\underline{K} \le x \le \overline{K}$, both x and $x+1$ attain the maximum in $V(W)$. It follows that

$$\{\tilde{W}(x) + R(x)\}\phi(x) + R(x-1)\{1-\phi(x)\}$$
$$= \{\tilde{W}(x) + R(x+1)\}\phi(x+1) + R(x)\{1-\phi(x+1)\}. \tag{20}$$

Moreover, if $W \in [\tilde{W}(x+1), \tilde{W}(x))$ then $x(W) = x+1$ and, so, $(x+1)$ must strictly outperform (x) in this range. Hence, for $W \in [\tilde{W}(x+1), \tilde{W}(x))$,

$$\{W + R(x)\}\phi(x) + R(x-1)\{1-\phi(x)\}$$
$$< \{W + R(x+1)\}\phi(x+1) + R(x)\{1-\phi(x+1)\}. \tag{21}$$

It must then follow from (20) and (21) that $\phi(x) > \phi(x+1)$, $\underline{K} \le x \le \overline{K} - 1$, and, hence, that $\phi(x)$ is strictly decreasing for $\underline{K} \le x \le \overline{K}$. Furthermore, from (20),

$$\begin{aligned}\tilde{W}(x) &= \{[R(x+1)\phi(x+1) + R(x)\{1-\phi(x+1)\}] \\ &\quad - [R(x)\phi(x) + R(x-1)\{1-\phi(x)\}]\}\{\phi(x) - \phi(x+1)\}^{-1} \\ &= \overline{W}(x), \qquad \underline{K} \le x \le \overline{K}.\end{aligned}$$

We conclude that $\overline{W}(x)$ is strictly decreasing in x for $\underline{K} \le x \le \overline{K}$. This concludes the proof of part (i).

Part (ii) follows trivially from the above analysis.

Remark 3. The reader should note that the index for state x in (19) involves quantities evaluated at $x-1$, x, and $x+1$. The index may be understood as a quantity that weighs the benefits of the positive reward enhancement achieved by the active action taken in x (the positive term) against the effects of reward deterioration experienced when the asset is passive (the negative term).

Example 1. Suppose that the reward rate is linear in the state in such a way that

$$R(x) = r(x - \underline{K}), \qquad \underline{K} \le x \le \overline{K},$$

for some $r > 0$. The function $\overline{W}$ in (19) then becomes

$$\overline{W}(x) = r[1 - \{\phi(x) - \phi(x+1)\}]\{\phi(x) - \phi(x+1)\}^{-1}, \qquad \underline{K} \le x \le \overline{K}, \tag{22}$$

and will be strictly decreasing when $\phi(x) - \phi(x+1)$ is strictly increasing over the range $x \in [\underline{K}, \overline{K}]$, i.e. when ϕ is (strictly) decreasing concave there. From Theorem 3, (22) gives the index in this case.

Example 2. It is in fact possible, by natural extension of the above material, to develop indexable assets with semi-infinite state spaces of the form $(-\infty, \overline{K}]$. Consider such an example for which $\overline{K} = 0$. We suppose that the reward rates are given by

$$R(x) = re^{\eta x}, \qquad x \le 0,$$

where $r > 0$ and $\eta > 0$. Furthermore, suppose that

$$\phi(x) = 1 - e^{\theta(x-1)}, \qquad x \le 0,$$

where $\theta > 0$ guarantees that ϕ is strictly decreasing. The function $\overline{W}$ in (19) then becomes

$$\overline{W}(x) = r(\mathrm{e}^{\eta} - 1)(1 - \mathrm{e}^{-\theta})^{-1}\{\mathrm{e}^{(\eta-\theta)x} - \mathrm{e}^{\eta x}(1 - \mathrm{e}^{-\eta}\mathrm{e}^{-\theta})\}, \qquad x \leq 0, \tag{23}$$

and will be strictly decreasing when $0 < \eta < \theta$. From a suitable extension of Theorem 3, (23) gives the index in this case.

3.1. Numerical results

In Table 1 we present some results of an extensive numerical investigation into the quality of performance of the index heuristics developed in this section. Each problem studied has $J = 4$ and $M = 1$; that is, at each decision epoch one of four possible assets must be chosen for activation. Each constituent asset is structured as in Example 1 above, with $\underline{K} = 0$, $\overline{K} = 10$, and

$$\phi(x) = 1 - \left(\frac{x}{11}\right)^{\alpha}, \qquad 0 \leq x \leq 10,$$

for some $\alpha > 1$. The function ϕ is indeed decreasing concave and results from the choices

$$\lambda(x) = \{(11)^{\alpha} - (x + 1)^{\alpha}\}(x + 1)^{-\alpha+1}$$

and $\mu(x) = x$, $0 \leq x \leq 10$. In this model, each of the four assets is characterized by the parameter pair (r, α). In all cases the αs are chosen by sampling from a continuous uniform distribution. The rs are chosen either from a uniform$(0, 5)$ distribution or from a uniform$(5, 10)$ distribution. Table 1 presents results for $800(= 4 \times 2 \times 100)$ randomly generated problems.

Four policies were applied to each problem generated. These are as follows.

- OPT: An optimal policy and its corresponding average reward rate r^{opt} were computed by dynamic programming value iteration. See, for example, Tijms (1994).
- IND: This is the index policy developed in the current section. At every decision epoch it activates the asset with currently maximal index.
- MYO: This is a myopic heuristic that attaches the index $r\lambda(x)$ to an asset in state x and activates the asset of largest index. This index may be understood as the rate at which the asset's reward-earning capacity may be enhanced by activation.
- SMA: This is the policy that always activates the asset of smallest state, with ties broken at random.

For each problem generated the average reward rates r^{ind}, r^{myo}, and r^{sma} were computed by dynamic programming value iteration, yielding the *percentage suboptimalities*

$$\Delta(H) := 100(r^{\mathrm{opt}} - r^{H})(r^{\mathrm{opt}})^{-1}, \qquad H = \mathrm{ind}, \mathrm{myo}, \mathrm{sma}.$$

Table 1 summarizes the collections of percentage suboptimalities (each collection of size 100) arising from the application of each of the IND, MYO, and SMA policies to each of the eight problem configurations. Each collection is summarized by the order statistics MIN (minimum), LQ (lower quartile), MED (median), UQ (upper quartile), and MAX (maximum). For example, from the top left-hand corner of Table 1, we see that when MYO is applied to the 100 problems with $\alpha \sim$ uniform$(1.1, 1.4)$ and $r \sim$ uniform$(0, 5)$ we obtain a median percentage suboptimality of 3.097 and a worst case which is 15.674% suboptimal.

The dominant feature of Table 1 is the outstanding performance of the index policy. In its worst performance in the 800 randomly generated problems analysed, it was just 0.024%

TABLE 1: Comparative performance of the index policy (IND), a myopic heuristic (MYO), and the 'smallest state' policy (SMA) for problems structured as in Example 1. First section: $r \sim$ uniform(0, 5) and $\alpha \sim$ uniform(1.1, 1.4) (*left*), $\alpha \sim$ uniform(1.4, 1.7) (*right*); second section: $r \sim$ uniform(5, 10) and $\alpha \sim$ uniform(1.1, 1.4) (*left*), $\alpha \sim$ uniform(1.4, 1.7) (*right*); third section: $r \sim$ uniform(0, 5) and $\alpha \sim$ uniform(1.7, 2.0) (*left*), $\alpha \sim$ uniform(1.01, 1.31) (*right*); fourth section: $r \sim$ uniform(5, 10) and $\alpha \sim$ uniform(1.7, 2.0) (*left*), $\alpha \sim$ uniform(1.01, 1.31) (*right*).

	IND	MYO	SMA	IND	MYO	SMA
MIN	0.000	0.099	1.015	0.000	0.021	0.462
LQ	0.000	1.572	13.677	0.000	0.655	8.143
MED	0.000	3.097	21.459	0.001	1.353	14.003
UQ	0.004	6.732	30.166	0.004	2.830	20.745
MAX	0.018	15.674	50.783	0.022	9.173	37.961
MIN	0.000	0.128	0.280	0.000	0.005	0.226
LQ	0.000	0.818	2.264	0.000	0.232	1.016
MED	0.002	1.407	4.387	0.001	0.484	2.134
UQ	0.004	2.382	6.100	0.003	0.721	3.115
MAX	0.024	7.784	13.899	0.010	2.236	6.210
MIN	0.000	0.035	0.404	0.000	0.325	2.112
LQ	0.000	0.370	5.635	0.000	2.451	17.224
MED	0.001	0.873	10.612	0.000	4.891	25.204
UQ	0.004	1.732	15.590	0.000	9.531	34.495
MAX	0.015	5.823	31.341	0.019	19.318	56.117
MIN	0.000	0.005	0.182	0.000	0.179	0.384
LQ	0.000	0.117	0.660	0.000	1.386	3.440
MED	0.000	0.224	1.490	0.000	2.333	6.369
UQ	0.001	0.351	2.068	0.003	4.676	8.800
MAX	0.007	0.948	4.251	0.019	14.593	20.683

suboptimal. For each of the other heuristics (MYO and SMA) problematic instances arose in which they performed poorly. In particular, there is clear evidence of a deterioration of performance of both as the generated αs approach 1 from above.

4. The spinning plates problem with discounted rewards

Now consider discounted reward versions of the Markov decision processes introduced in Section 2. Family 1 is defined as in Definitions 1(i) and 1(ii), but here rewards are accumulated at rate $\mathrm{e}^{-\alpha t} \sum_j R_j(X_j(t))$ at time $t \in \mathbb{R}^+$, where $\alpha > 0$. Family 2 is defined as in Definitions 2(i) and 2(ii), but now if action $A(S)$ is operative at time $t \in \mathbb{R}^+$, rewards are then accumulated at rate $\mathrm{e}^{-\alpha t} \sum_{j \in S} R_j(X_j(t))$. An appropriate version of the development of (1)–(6) above again yields a decomposition of a Lagrangian relaxation of the optimization problem into a collection of J W-subsidy problems, one for each asset. Under given conditions it will emerge for both families that there exist optimal policies for the W-subsidy problems which are monotone, as in (7) above. We continue to use the convention established prior to Definition 3 that, in the event of more than one monotone policy being optimal, we choose the one with the largest passive set. We have indexability when this passive set is increasing in W. The corresponding indices are as in Definition 4.

For the remainder of this section, we focus on family 1 and write $V_\alpha(\cdot, W)\colon [\underline{K}, \overline{K}] \to \mathbb{R}$ for the *value function* of the W-subsidy problem defined for an asset whose associated parameters

are $\overline{K}$, $\underline{K}$, $\lambda(\cdot)$, $\mu(\cdot)$, and $R(\cdot)$. Hence, $V_\alpha(x, W)$ is the total expected reward earned over an infinite horizon when an optimal policy is applied and x is the state of the asset at time 0. Standard theory (see, for example, Puterman (1994, pp. 142–156)) guarantees the existence of a stationary, deterministic, Markov policy whose value function satisfies the dynamic programming optimality equations. In the case of family 1 these are

$$\begin{aligned} V_\alpha(\underline{K}, W) &= \max\{[R(\underline{K}) + \lambda(\underline{K})V_\alpha(\underline{K}+1, W)]\{\lambda(\underline{K})+\alpha\}^{-1}, [W + R(\underline{K})]\alpha^{-1}\}, \\ V_\alpha(x, W) &= \max\{[R(x) + \lambda(x)V_\alpha(x+1, W)]\{\lambda(x)+\alpha\}^{-1}, \\ &\qquad [W + R(x) + \mu(x)V_\alpha(x-1, W)]\{\mu(x)+\alpha\}^{-1}\}, \qquad (24) \\ &\qquad\qquad \underline{K}+1 \le x \le \overline{K}-1, \\ V_\alpha(\overline{K}, W) &= \max\{R(\overline{K})\alpha^{-1}, [W + R(\overline{K}) + \mu(\overline{K})V_\alpha(\overline{K}-1, W)]\{\mu(\overline{K})+\alpha\}^{-1}\}. \end{aligned}$$

Throughout (24), the first quantity on the right-hand side is the total expected reward earned when choosing action a in the current state and thereafter proceeding optimally. The second quantity is the total expected reward earned when choosing action b in the current state and then proceeding optimally.

We introduce $V_\alpha^x(\hat{x}, W)$ as the value function for monotone policy (x) evaluated at initial state $\hat{x} \in [\underline{K}, \overline{K}]$ (see (7)). By direct calculation we have

$$\begin{aligned} V_\alpha^x(\hat{x}, W) &= \sum_{y=0}^{\hat{x}-x-1}\left[\prod_{u=0}^{y-1}\mu(\hat{x}-u)\{\mu(\hat{x}-u)+\alpha\}^{-1}\right]\{W + R(\hat{x}-y)\}\{\mu(\hat{x}-y)+\alpha\}^{-1} \\ &\quad + \left[\prod_{y=0}^{\hat{x}-x-1}\mu(\hat{x}-y)\{\mu(\hat{x}-y)+\alpha\}^{-1}\right]V_\alpha^x(x, W), \qquad \underline{K} \le x \le \hat{x}, \\ V_\alpha^x(\hat{x}, W) &= \sum_{y=0}^{x-\hat{x}-1}\left[\prod_{u=0}^{y-1}\lambda(\hat{x}+u)\{\lambda(\hat{x}+u)+\alpha\}^{-1}\right]R(\hat{x}+y)\{\lambda(\hat{x}+y)+\alpha\}^{-1} \\ &\quad + \left[\prod_{y=0}^{x-\hat{x}-1}\lambda(\hat{x}+y)\{\lambda(\hat{x}+y)+\alpha\}^{-1}\right]V_\alpha^x(x, W), \qquad \hat{x} < x \le \overline{K}, \qquad (25) \\ V_\alpha^{\overline{K}+1}(\hat{x}, W) &= \sum_{y=0}^{\overline{K}-\hat{x}-1}\left[\prod_{u=0}^{y}\lambda(\hat{x}+u)\{\lambda(\hat{x}+u)+\alpha\}^{-1}\right]R(\hat{x}+y)\{\lambda(\hat{x}+y)+\alpha\}^{-1} \\ &\quad + \left[\prod_{y=0}^{\overline{K}-\hat{x}-1}\lambda(\hat{x}+y)\{\lambda(\hat{x}+y)+\alpha\}^{-1}\right]R(\overline{K})\alpha^{-1}. \end{aligned}$$

We also observe that if $X(0) = x$, $\underline{K}+1 \le x \le \overline{K}$, then, under policy (x), the asset has an initial passive sojourn in x followed by an active sojourn in $x-1$ that is terminated by a return to x. It then follows that

$$\begin{aligned} V_\alpha^x(x, W) &= \{W + R(x)\}\{\mu(x)+\alpha\}^{-1} + R(x-1)\mu(x)\{\mu(x)+\alpha\}^{-1}\{\lambda(x-1)+\alpha\}^{-1} \\ &\quad + \mu(x)\lambda(x-1)\{\mu(x)+\alpha\}^{-1}\{\lambda(x-1)+\alpha\}^{-1}V_\alpha^x(x, W), \end{aligned}$$

from which we infer that

$$V_\alpha^x(x, W) = [\{W + R(x)\}\{\lambda(x-1)+\alpha\} + R(x-1)\mu(x)]\{\alpha\lambda(x-1) + \alpha\mu(x) + \alpha^2\}^{-1}, \quad \underline{K} \le x \le \overline{K}. \tag{26}$$

We are now ready to proceed to our main result.

Theorem 4. (Family 1: discounted rewards.) *If*

$$\{\lambda(x-1)+\alpha\}\{\mu(x+1)+\alpha\} > \lambda(x)\mu(x), \qquad x \in [\underline{K}, \overline{K}], \tag{27}$$

and

$$\begin{aligned}\overline{W}_\alpha(x) := [&R(x+1)\lambda(x)\{\lambda(x-1) + \mu(x) + \alpha\} \\ &+ R(x)[\mu(x)\{\mu(x+1)+\alpha\} - \lambda(x)\{\lambda(x-1)+\alpha\}] \\ &- R(x-1)\mu(x)\{\lambda(x) + \mu(x+1) + \alpha\}] \\ &\times [\{\lambda(x-1)+\alpha\}\{\mu(x+1)+\alpha\} - \lambda(x)\mu(x)]^{-1}\end{aligned} \tag{28}$$

is strictly decreasing for $\underline{K} \le x \le \overline{K}$, then the asset is strictly indexable and $\overline{W}_\alpha(x)$ is the index for state $x \in [\underline{K}, \overline{K}]$.

Proof. Fix an initial state $\hat{x}$. By an argument akin to that in the proof of Lemma 1, the expected total reward earned by the asset over an infinite horizon under any stationary, deterministic, Markov policy from $\hat{x}$ will be exactly matched by that earned by some member of B. It follows that the value function for the W-subsidy problem evaluated at $\hat{x}$ is the expected reward achieved by the best monotone policy from this class.

From (25)–(27) we deduce the following (in)equalities via straightforward algebra:

$$V_\alpha^x(\hat{x}, W) > V_\alpha^{x+1}(\hat{x}, W), \qquad W > \overline{W}_\alpha(x),\ x \in [\underline{K}, \overline{K}], \tag{29}$$

$$V_\alpha^x(\hat{x}, W) = V_\alpha^{x+1}(\hat{x}, W), \qquad W = \overline{W}_\alpha(x),\ x \in [\underline{K}, \overline{K}], \tag{30}$$

$$V_\alpha^x(\hat{x}, W) < V_\alpha^{x+1}(\hat{x}, W), \qquad W < \overline{W}_\alpha(x),\ x \in [\underline{K}, \overline{K}]. \tag{31}$$

Using the fact that $\overline{W}_\alpha(x)$ is strictly decreasing we can now infer from (29)–(31) that

$$\max_{\underline{K} \le y \le \overline{K}+1} V_\alpha^y(\hat{x}, W)$$

is achieved at $y = \overline{K} + 1$ for $W < \overline{W}_\alpha(\overline{K})$, at $y = x$ for $W \in [\overline{W}_\alpha(x), \overline{W}_\alpha(x-1))$, $\underline{K}+1 \le x \le \overline{K}$, and at $y = \underline{K}$ for $W \ge \overline{W}_\alpha(\underline{K})$. We thus deduce that

$$V_\alpha(\hat{x}, W) = \begin{cases} V_\alpha^{\overline{K}+1}(\hat{x}, W), & W < \overline{W}_\alpha(\overline{K}), \\ V_\alpha^x(\hat{x}, W), & W \in [\overline{W}_\alpha(x), \overline{W}_\alpha(x-1)),\ \underline{K}+1 \le x \le \overline{K}, \\ V_\alpha^{\underline{K}}(\hat{x}, W), & W \ge \overline{W}_\alpha(\underline{K}). \end{cases} \tag{32}$$

However, the initial state $\hat{x}$ was chosen arbitrarily in the above. We thus infer from (32) that, for *all* initial states, monotone policy $(\overline{K}+1)$ is optimal for the W-subsidy problem for $W < \overline{W}_\alpha(\overline{K})$, policy (x) is optimal for $\overline{W}_\alpha(x) \le W < \overline{W}_\alpha(x-1)$, $\underline{K}+1 \le x \le \overline{K}$, and

policy $(\underline{K})$ is optimal for $W \geq \overline{W}_\alpha(\underline{K})$. Denoting by $b(W)$ the maximal optimal passive set, we infer that

$$b(W) = \begin{cases} \varnothing, & W < \overline{W}_\alpha(\overline{K}), \\ [x, \overline{K}], & W \in [\overline{W}_\alpha(x), \overline{W}_\alpha(x-1)), \\ [\underline{K}, \overline{K}], & W \geq \overline{W}_\alpha(\underline{K}). \end{cases} \tag{33}$$

From Definitions 3 and 4, (strict) indexability follows from (33), and $\overline{W}_\alpha(x)$ is the index for state $x \in [\underline{K}, \overline{K}]$. This concludes the proof.

Remarks 4. 1. If we set $\alpha = 0$ in (27) and (28), we recover the conditions expressed in Theorem 3(i)(b). It follows that any asset which meets the (necessary and sufficient) conditions of Theorem 3 will also meet the conditions of Theorem 4 for sufficiently small α.

2. It is not difficult to show that, if an a priori restriction to monotone policies for the W-subsidy problem is imposed, then the conditions expressed in (27) and (28) are necessary and sufficient for strict indexability.

5. Family 2 (the squad system): a model for the optimal exploitation of assets

We now consider family 2 with the average reward criterion, as described in Section 2. Since there are points of similarity between the theoretical development of the indexability analysis for family 2 and that for family 1, given in Section 3, we highlight the main features of the former only and omit proofs. Further details are available from the authors.

The asset suffix is again dropped and the W-subsidy problem considered for a single asset with associated parameters $\overline{K}$, $\underline{K}$, $\nu(\cdot)$, $\rho(\cdot)$, and $R(\cdot)$. From Definitions 2(ii) and 2(iii) recall that, under the application of active action a in state x, the asset evolves to state $x-1$ at rate $\nu(x)$ and earns rewards at rate $R(x)$ while doing so. Under application of the passive action b in state x, the asset evolves to state $x+1$ at rate $\rho(x)$ and (in the W-subsidy problem) earns rewards at rate W while doing so. The intermediate goal of our analysis is to identify policies that maximize the average reward rate earned over an infinite horizon.

We identify the class A of monotone policies for which

$$\pi(x) = a \Leftrightarrow x \geq y \quad \text{for some } y \in [\underline{K}, \overline{K}+1], \tag{34}$$

and write $[y]$ for the policy in (34). Hence, policy $[\overline{K}+1]$ chooses action b in all states while policy $[\underline{K}]$ chooses action a in all states.

Lemma 3. *For all $W \in \mathbb{R}$ there exists an optimal policy for the W-subsidy problem in A.*

The proof of Lemma 3 follows along lines similar to that of Lemma 1.

The average reward rate for the W-subsidy problem under policy $[y]$ is given by

$$W\psi(y) + R(y)\{1 - \psi(y)\},$$

where

$$\psi(y) = \nu(y)\{\nu(y) + \rho(y-1)\}^{-1}, \qquad \underline{K} \leq y \leq \overline{K}+1.$$

Note that $\psi(\overline{K}+1) = 1$ and $\psi(\underline{K}) = 0$. Now write $[y(W)]$ for the policy with maximal passive set solving the the W-subsidy problem. From the above, we have

$$y(W) \in \operatorname*{argmax}_{\underline{K} \leq y \leq \overline{K}+1} \{W\psi(y) + R(y)\{1 - \psi(y)\}\}. \tag{35}$$

Should more than a single y-value attain the maximum on the right-hand side of (35) then $y(W)$ is chosen to be the largest.

From Definition 3, in order to establish the asset's indexability it will be enough to show that there exist finite real numbers $\underline{W}$ and $\overline{W}$, $\underline{W} < \overline{W}$, such that $y(W) = \underline{K}$ for $W < \underline{W}$, $y(W) = \overline{K} + 1$ for $W \geq \overline{W}$, and $y(\cdot)\colon [\underline{W}, \overline{W}] \to [\underline{K}, \overline{K} + 1]$ is increasing. For an indexable asset, the index in state y will be given by

$$W(y) = \inf\{W\colon y(W) \geq y + 1\}. \tag{36}$$

Theorem 5. *If ψ is increasing then the asset is indexable.*

The proof of Theorem 5 is similar to that of Theorem 1.

In order to describe the asset's index structure, we develop a collection consisting of a positive integer L, a set of $L + 1$ integers $\{y^l,\ 0 \leq l \leq L\}$ such that

$$\underline{K} = y^0 < y^1 < \cdots < y^L = \overline{K} + 1,$$

and an accompanying set of L reals $\{W^l,\ 1 \leq l \leq L\}$ such that

$$-\infty < W^1 < \cdots < W^L < \infty.$$

To initiate the inductive specification of these we require additional notation. Suppose y to be an integer such that $\underline{K} \leq y \leq \overline{K}$. We write

$$n(y) = \min\{x \in \mathbb{Z}\colon \overline{K} + 1 \geq x > y \text{ and } \psi(x) > \psi(y)\}.$$

We now define

$$W^1 = \min_{\underline{K}+1 \leq y \leq \overline{K}+1} \{[R(\underline{K}) - R(y)\{1 - \psi(y)\}]\{\psi(y)\}^{-1}\} \tag{37}$$

and denote by y^1 the largest minimizer in (37). If $y^1 = \overline{K} + 1$ then we set $L = 1$ and stop. If $y^1 < \overline{K} + 1$ then we define

$$W^2 = \min_{n(y^1) \leq y \leq \overline{K}+1} \{[R(y^1)\{1 - \psi(y^1)\} - R(y)\{1 - \psi(y)\}]\{\psi(y) - \psi(y^1)\}^{-1}\} \tag{38}$$

and denote by y^2 the largest minimizer in (38). In general, if $y^k < \overline{K} + 1$ then we define

$$W^{k+1} = \min_{n(y^k) \leq y \leq \overline{K}+1} \{[R(y^k)\{1 - \psi(y^k)\} - R(y)\{1 - \psi(y)\}]\{\psi(y) - \psi(y^k)\}^{-1}\} \tag{39}$$

and denote by y^{k+1} the largest minimizer in (39). This continues until we find $\overline{K} + 1$ as the largest minimizer. Should this happen at step $k + 1$ (i.e. in the calculation of W^{k+1}) we have $L = k + 1$.

From a discussion along the lines of the previous section, we find that if ψ is increasing then

$$y(W) = \begin{cases} \underline{K}, & W < W^1, \\ y^k, & W \in [W^k, W^{k+1}),\ 1 \leq k \leq L - 1, \\ \overline{K} + 1, & W \geq W^L. \end{cases} \tag{40}$$

The index structure of the asset now follows from (36) and (40) and is described in Theorem 6, as follows.

Theorem 6. *If ψ is increasing then the index $W\colon [\underline{K}, \overline{K}] \to \mathbb{R}$ is given by*

$$W(y) = W^k, \qquad y^{k-1} \le y \le y^k - 1,\ 1 \le k \le L.$$

Remark 5. Note from Theorem 6 that the index is increasing in the state. Hence, in the squad system it is assets which are achieving high rewards that are a high priority for activation. Note also that, unlike in the spinning plates problem, indices can now be negative. This raises the question of whether idling may be preferable to asset deployment.

As before, important special cases occur in which all states have distinct indices. Theorem 7 gives a condition necessary and sufficient for strict indexability.

Theorem 7. (i) *The following assertions are equivalent.*

(a) *The asset is strictly indexable.*

(b) *Both $\psi(y)$ and*

$$\overline{W}(y) = [R(y)\{1 - \psi(y)\} - R(y+1)\{1 - \psi(y+1)\}]\{\psi(y+1) - \psi(y)\}^{-1} \tag{41}$$

are strictly increasing over the range $\underline{K} \le y \le \overline{K}$.

(ii) *Under the conditions in part (i)(b), we have $W(y) = \overline{W}(y)$, $\underline{K} \le y \le \overline{K}$, and the index is strictly increasing in the state.*

Example 3. Suppose that the reward is linear in the state and, hence, that

$$R(y) = r(y - \underline{K}), \qquad \underline{K} \le y \le \overline{K}.$$

Moreover, suppose that the transition rates are also linear, i.e.

$$\nu(y) = \overline{\nu}(y - \underline{K}), \qquad \underline{K} \le y \le \overline{K}, \tag{42}$$

$$\rho(y) = \overline{\rho}(\overline{K} - y), \qquad \underline{K} \le y \le \overline{K}, \tag{43}$$

where r, $\overline{\nu}$, and $\overline{\rho}$ are all positive constants. It follows trivially from (42) and (43) that ψ is strictly increasing. By direct computation we find from (41) that, in this case,

$$\overline{W}(y) = r\left\{(y - \underline{K})(y + 1 - \underline{K}) - \frac{\bar{\rho}}{\bar{\nu}}(\overline{K} - y)(\overline{K} - y + 1)\right\}(\overline{K} - \underline{K} + 1)^{-1},$$
$$\underline{K} \le y \le \overline{K}, \tag{44}$$

which is strictly increasing. From Theorem 7, (44) gives the index in this case. This is the example referred to by Whittle (1988) as the Ehrenfest project. He used a heuristic argument to develop an index which approximates that in (44).

Example 4. It is possible to develop indexable assets with semi-infinite state spaces of the form $[\underline{K}, \infty)$. Consider such an example for which $\underline{K} = 0$. We suppose that reward rates are given by

$$R(y) = r[1 - (y+1)^{-\alpha}], \qquad y \ge 0, \tag{45}$$

and, furthermore, that

$$\psi(y) = 1 - (y+1)^{-\beta}, \qquad y \ge 0. \tag{46}$$

In (45) and (46), r, α, and β are positive constants. The function in (41) now becomes

$$\overline{W}(y) = r\left(1 - \frac{(y+1)^{-(\alpha+\beta)} - (y+2)^{-(\alpha+\beta)}}{(y+1)^{-\beta} - (y+2)^{-\beta}}\right), \qquad y \geq 0, \tag{47}$$

which is strictly increasing. From a suitable extension of Theorem 7, (47) gives the index in this case.

Example 5. We can generalize Example 4 as follows. Let $\xi : \mathbb{N} \to [1, \infty)$ be a strictly increasing function with $\xi(0) = 1$ and $\xi(y) \to \infty$ as $y \to \infty$, and let r, α, and β be positive constants. We suppose that $R(y) = r[1 - \{\xi(y)\}^{-\alpha}]$, $y \geq 0$, and that $\psi(y) = 1 - \xi(y)^{-\beta}$, $y \geq 0$. The function in (41) then becomes

$$\overline{W}(y) = r\left(1 - \frac{\xi(y)^{-(\alpha+\beta)} - \xi(y+1)^{-(\alpha+\beta)}}{\xi(y)^{-\beta} - \xi(y+1)^{-\beta}}\right), \qquad y \geq 0, \tag{48}$$

which is strictly increasing. From a suitable extension of Theorem 7, (48) gives the index in this case.

5.1. Numerical results

In Table 2 we present some results derived from an extensive, numerically based assessment of the quality of performance of the index heuristics developed in this section. For the cases in the first and third sections of Table 2, we have $J = 4$ and $M = 1$. In the second and fourth sections these cases are embellished by the inclusion of an *idling option*, to be thought of as a zero-reward asset whose state space is a singleton. Such an asset is trivially indexable, with index always zero. In all cases, the remaining constituent assets are structured as in Example 4 above with the function $\psi(y) = 1 - (y+1)^{-\beta}$, $y \geq 0$, which arises from the choices

$$\nu(y) = 1(y \geq 1) \tag{49}$$

and

$$\rho(y) = \{1 - (y+2)^{-\beta}\}^{-1} - 1, \qquad y \geq 0.$$

In (49), $1(\cdot)$ is the indicator function. In this model each of the four assets (excepting the idling option in the second and fourth sections) is characterized by the parameter triple (r, α, β). In all cases the αs are chosen by sampling from a continuous uniform(0.5, 1.0) distribution and the rs are drawn from uniform(2, 4). The βs are also chosen by sampling from uniform distributions, as indicated in the caption of Table 2. Table 2 presents results for $800 (= 4 \times 2 \times 100)$ randomly generated problems.

Four policies were applied to each problem generated. Policies OPT and IND are equivalent to the corresponding ones described in the numerical study in Section 3. Note now, however, that when it is present the idling option will be taken by IND only when all four of the conventional assets have negative indices. The remaining two policies are as follows.

- MYO: This is a myopic heuristic that activates an asset with largest current reward rate $r(y)$.
- LAR: This is the policy that always activates whichever of the four conventional assets is in the largest state. Ties are broken at random.

Note that MYO and LAR never choose the idling option (when it is available).

TABLE 2: Comparative performance of the index policy, a myopic heuristic, and the 'largest state' policy for problems structured as in Example 4. First section: $J = 4$, $M = 1$, and $\beta \sim$ uniform(0.5, 1.0) (*left*), $\beta \sim$ uniform(1.0, 2.0) (*right*); second section: $J = 4$, $M = 1$ (plus idling option), and $\beta \sim$ uniform(0.5, 1.0) (*left*), $\beta \sim$ uniform(1.0, 2.0) (*right*); third section: $J = 4$, $M = 1$, and $\beta \sim$ uniform(2.0, 3.0) (*left*), $\beta \sim$ uniform(0.5, 3.0) (*right*); fourth section: $J = 4$, $M = 1$ (plus idling option), and $\beta \sim$ uniform(2.0, 3.0) (*left*), $\beta \sim$ uniform(0.5, 3.0) (*right*).

	IND	MYO	LAR	IND	MYO	LAR
MIN	0.006	0.191	0.046	0.028	0.073	0.171
LQ	0.056	1.349	1.105	0.165	0.411	1.657
MED	0.101	2.519	1.993	0.357	0.702	2.672
UQ	0.227	3.675	2.961	0.635	1.152	4.432
MAX	1.199	7.874	9.128	1.049	3.175	7.951
MIN	0.007	0.191	0.046	0.026	0.116	0.311
LQ	0.056	1.349	1.105	0.135	0.944	2.149
MED	0.101	2.519	1.994	0.257	1.230	3.418
UQ	0.227	3.675	2.961	0.464	1.944	5.037
MAX	1.199	7.874	9.128	0.795	4.802	8.912
MIN	0.070	0.070	0.627	0.010	0.086	0.347
LQ	0.265	0.685	2.510	0.202	1.244	2.262
MED	0.409	3.301	4.189	0.355	1.881	3.708
UQ	0.552	4.958	6.628	0.662	3.565	6.300
MAX	1.031	16.650	17.032	1.114	16.209	16.974
MIN	0.061	4.670	6.330	0.012	0.510	0.395
LQ	0.217	8.233	9.606	0.191	1.887	2.824
MED	0.337	9.704	10.794	0.357	2.955	5.094
UQ	0.446	12.121	13.008	0.601	5.608	8.128
MAX	0.889	21.827	22.185	1.052	20.239	20.967

As in the previous numerical study, percentage suboptimalities are presented in Table 2 for the heuristics IND, MYO, and LAR for collections containing 100 problems of common structure. The index policy continues to perform strongly, with a worst case of 1.199% suboptimality among the 800 problems generated. There is evidence of modestly enhanced performance following inclusion of the idling option. For each of the other heuristics (MYO and LAR) problematic instances arose in which they performed poorly. There is clear evidence of deterioration in the performance of these policies as the βs increase.

We conclude by remarking that an analysis of a discounted reward version of the squad system, similar to that given in Section 4 for family 1, yields the following theorem. Similar comments to those following Theorem 4 apply.

Theorem 8. (Family 2: discounted rewards.) *If*

$$\{\nu(y+1)+\alpha\}\{\rho(y-1)+\alpha\} > \nu(y)\rho(y), \qquad y \in [\underline{K}, \overline{K}],$$

and

$$\begin{aligned}\overline{W}_\alpha(y) := [&R(y)\{\nu(y+1)+\rho(y)+\alpha\}\{\rho(y-1)+\alpha\} \\ &- R(y+1)\rho(y)\{\nu(y)+\rho(y-1)+\alpha\}] \\ &\times [\{\nu(y+1)+\alpha\}\{\rho(y-1)+\alpha\} - \nu(y)\rho(y)]^{-1}\end{aligned}$$

is strictly increasing over the range $\underline{K} \le y \le \overline{K}$, then the asset is strictly indexable and $\overline{W}_\alpha(y)$ is the index for state $y \in [\underline{K}, \overline{K}]$.

Acknowledgements

The first two authors acknowledge the support of the Engineering and Physical Sciences Research Council through the award of grant no. GR/S45188/01. The third author acknowledges the support of the Servei de Recerca of Universitat Pompeu Fabra through the award of grant no. EBES-REI2645.

References

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.

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

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

Glazebrook, K. D., Lumley, R. R. and Ansell, P. S. (2003). Index heuristics for multi-class $M/G/1$ systems with non-preemptive service and convex holding costs. *Queueing Systems* **45,** 81–111.

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.

Niño-Mora, J. (2001a). PCL-indexable restless bandits: diminishing marginal returns, optimal marginal reward rate index characterization, and a tiring–recovery model. Unpublished manuscript.

Niño-Mora, J. (2001b). 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 queueing network control. *Math. Operat. Res.* **24,** 293–305.

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

Tijms, H. C. (1994). *Stochastic Models: An Algorithmic Approach*. John Wiley, New York.

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

Weber, R. R. and Weiss, G. (1991). Addendum to 'On an index policy for restless bandits'. *Adv. Appl. Prob.* **23,** 429–430.

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