UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 07-41
Statistics and Econometrics Series 09
May 2007

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34-91) 6249849

Two-Stage Index Computation for Bandits with Switching Penalties I: Switching Costs*

José Niño-Mora [1]

Abstract

This paper addresses the multi-armed bandit problem with switching costs. Asawa and Teneketzis (1996) introduced an index that partly characterizes optimal policies, attaching to each bandit state a ``continuation index" (its Gittins index) and a ``switching index." They proposed to jointly compute both as the Gittins index of a bandit having $2n$ states — when the original bandit has n states — which results in an eight-fold increase in $O(n^3)$ arithmetic operations relative to those to compute the continuation index alone. This paper presents a more efficient, decoupled computation method, which in a first stage computes the continuation index and then, in a second stage, computes the switching index an order of magnitude faster in at most $n^2 + O(n)$ arithmetic operations. The paper exploits the fact that the Asawa and Teneketzis index is the Whittle, or marginal productivity, index of a classic bandit with switching costs in its restless reformulation, by deploying work-reward analysis and PCL-indexability methods introduced by the author. A computational study demonstrates the dramatic runtime savings achieved by the new algorithm, the near-optimality of the index policy, and its substantial gains against the benchmark Gittins index policy across a wide range of instances.

Keywords: Dynamic programming, Markov, finite state; bandits; switching costs; index policy; Whittle index; hysteresis; work-reward analysis; PCL-indexability; analysis of algorithms
JEL Classification: C61, C63

[1] Niño-Mora, Departamento de Estadística, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe (Madrid), e-mail: jose.nino@uc3m.es. Supported in part by the Spanish Ministry of Education & Science under grant MTM2004-02334 and a Ramón y Cajal Investigator Award, by the EU's Networks of Excellence Euro-NGI/FGI, and by the Autonomous Community of Madrid-UC3M's grants UC3M-MTM-05-075 and CCG06-UC3M/ESP-0767.

Two-Stage Index Computation for Bandits with Switching Penalties I: Switching Costs

José Niño-Mora
Department of Statistics, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe (Madrid), Spain,
jnimora@alum.mit.edu

This paper addresses the multi-armed bandit problem with switching costs. Asawa and Teneketzis (1996) introduced an index that partly characterizes optimal policies, attaching to each bandit state a "continuation index" (its Gittins index) and a "switching index." They proposed to jointly compute both as the Gittins index of a bandit having $2n$ states — when the original bandit has n states — which results in an 8-fold increase in $O(n^3)$ arithmetic operations relative to those to compute the continuation index alone. This paper presents a more efficient, decoupled computation method, which in a first stage computes the continuation index and then, in a second stage, computes the switching index an order of magnitude faster in at most $n^2 + O(n)$ arithmetic operations. The paper exploits the fact that the Asawa and Teneketzis index is the Whittle, or marginal productivity, index of a classic bandit with switching costs in its restless reformulation, by deploying work-reward analysis and PCL-indexability methods introduced by the author. A computational study demonstrates the dramatic runtime savings achieved by the new algorithm, the near-optimality of the index policy, and its substantial gains against the benchmark Gittins index policy across a wide range of instances.

Key words: Dynamic programming, Markov, finite state; bandits; switching costs; index policy; Whittle index; hysteresis; work-reward analysis; PCL-indexability; analysis of algorithms;
History: submitted March 2006; revised September 2006; May 8, 2007

---

1. Introduction

Imagine a firm owning a portfolio of dynamic and stochastic projects, of which it can engage one at a time. To (re)start a project, the firm incurs an upfront lump-sum *startup cost*, after which it accrues rewards and operating expenses. The firm can decide at any time to abandon the project currently in operation, incurring a lump-sum *shutdown cost*, to switch to another project. Such a firm faces the problem of designing a dynamic project selection policy that maximizes the expected total discounted value of its net earnings.

The problem is cast as a *Markov decision process* (MDP) by modeling projects as discrete-time and -state *bandits*: binary-action (active/passive) MDPs that can only change state while active. In the no switching costs case, one thus obtains the classic *multi-armed bandit problem* (MABP).

In a celebrated result, Gittins and Jones (1974) furnished an elegant and efficient solution to the MABP: there exists an *index* attached to each bandit, which is a function of its state, such that the resulting *priority-index policy*, which engages at each time a bandit of largest index, is optimal.

Yet, as pointed out in Banks and Sundaram (1994), "it is difficult to imagine a relevant economic decision problem in which the decision-maker may costlessly move between alternatives." Incorporation of bandit startup/shutdown costs into the MABP yields a form of the *multi-armed bandit problem with switching costs* (MABPSC), which is extensively surveyed in Jun (2004).

The MABP's optimal index solution motivates the investigation of good index policies for the MABPSC. As discussed in Banks and Sundaram (1994), such policies attach an index $\nu_m(a_m^-, i_m)$ to each bandit m, which is a function of its previous action a_m^- and current state i_m, thus decoupling into a "continuation index" $\nu_m(1, i_m)$ and a "switching index" $\nu_m(0, i_m)$. They further observed that "it is obvious that in comparing two otherwise identical arms, one of which was used in the previous period, the one which was in use must necessarily be more attractive than the one which was idle." Thus, to be consistent which such a *hysteretic* property, the indices must satisfy

$$\nu_m(1, i_m) \geq \nu_m(0, i_m). \tag{1}$$

While Banks and Sundaram proved that such policies are not generally optimal, Asawa and Teneketzis (1996) introduced an intuitive index, which we will refer to henceforth as the *AT index*, and showed that it partly characterizes optimal policies. Their continuation index is the Gittins index, while their switching index is the maximum rate, achievable by stopping rules that engage an initially passive bandit, of expected discounted reward earned minus initial startup cost incurred per unit of expected discounted time. Though they focused on the case where each bandit has a constant startup cost and no shutdown cost, their results extend to bandits having a constant shutdown cost, using the transformation given in Banks and Sundaram (1994, Sec. 3).

Asawa and Teneketzis proposed to jointly compute both indices by: (i) formulating an augmented bandit *without* switching costs, yet having *twice* the number of states — the (a_m^-, i_m)'s —; and (ii) computing the Gittins index of the latter bandit. Since computing the Gittins index requires $O(n^3)$ arithmetic operations for an n-state bandit, such a scheme yields an 8-fold increase relative to the effort to compute the continuation index alone. Hence, computing the switching AT

index in such a fashion involves steep computational costs for large-scale models, which hinders applicability of such an index.

Motivated by such considerations, this paper sets out to establish the practical viability and usefulness of the AT index for the MABSC, by seeking to develop a significantly more efficient computation method. While that is the prime goal of this paper, the second goal is to investigate empirically the relative performance of the resulting AT index policy.

We will pursue such goals in the setting of an extended model with state-dependent switching costs, via a seemingly indirect route: by exploiting the reformulation of a classic bandit with switching costs as a *restless bandit* — one that can change state while passive — *without* switching costs, under which the MABSC becomes a *multi-armed restless bandit problem* (MARBP).

Such a reformulation allows us to deploy the powerful indexation theory for restless bandits. This was introduced by Whittle (1988), who first realized that the Gittins-index definition via calibration also yields an index for restless bandits, albeit only for the limited range of so-called *indexable* instances. He proposed to use the resulting index policy as a heuristic for the MARBP, which is generally suboptimal. The theory has been developed in Niño-Mora (2001, 2002, 2006a, 2007a). Such work has identified two tractable classes of indexable bandits, termed *PCL-indexable* — after their satisfaction of *partial conservation laws* (PCLs) — and *LP-indexable*, for which the *Whittle index* and extensions are efficiently computed by an *adaptive-greedy algorithm*. The index measures trade-off (reward vs. work) rates, whence our terming it *marginal productivity index* (MPI).

This paper deploys such a theory, by proving and exploiting the fact that the AT index of a bandit with switching costs is precisely the bandit's Whittle index/MPI in its restless formulation. We will show that such restless bandits are PCL-indexable, relative to the family of hysteretic policies consistent with (1), which will allow us to compute the index using the adaptive-greedy algorithm referred to above. A work-reward analysis will then reveal that such an algorithm naturally decouples into two stages: a first stage that computes the Gittins index and required extra quantities; and a second stage, which is fed the first-stage's output, that computes the switching index.

To implement such a scheme, one can use for the first stage any of several $O(n^3)$ algorithms in Niño-Mora (2006b). For the second stage, we introduce here a fast switching-index algorithm that performs *at most* $n^2 + O(n)$ arithmetic operations, thus achieving an order of magnitude improvement in complexity that renders negligible the marginal effort to compute the switching index. Such an algorithm is the main contribution of this paper.

The paper further reports on a computational study demonstrating that such an improved com-

plexity translates into dramatic runtime savings. The study is complemented by a set of experiments that demonstrate the near-optimality of the index policy and its substantial gains against the benchmark Gittins index policy across an extensive range of two- and three-bandit instances.

Section 2 describes the model, shows how to reduce it to the normalized no shutdown costs case, defines the AT index, and gives the MARBP reformulation. Section 3 reviews the indexation theory to be deployed. Section 4 carries out a work-reward analysis of the bandits of concern, reformulated as restless bandits, and establishes their PCL-indexability. Section 5 draws on such an analysis to develop the new decoupled index algorithm. Section 6 discusses dependence of the index on switching costs. Section 7 reports the computational study's results. Section 8 concludes.

In the companion paper Niño-Mora (2007b), the results herein are extended to the case where bandits incorporate both switching costs and delays.

2. Model, AT Index and Restless-Bandit Reformulation

2.1. The MABPSC

Consider a collection of M finite-state bandits, one of which must be engaged (*active*) at each discrete time period $t \geq 0$ over an infinite horizon, while the others are rested (*passive*). When bandit m occupies state i_m — belonging in its state space N_m — and is engaged, it yields an *active reward* $R_m^1(i_m) = R_m(i_m)$ and its state moves to j_m with probability $p_m(i_m, j_m)$. If the bandit is rested, it yields a zero *passive reward* $R_m^0(i_m) \equiv 0$ and its state does not change.

Switching bandits is costly. When bandit m occupies state i_m and is freshly engaged (resp. rested), a *startup cost* $c_m(i_m)$ (resp. *shutdown cost* $d_m(i_m)$) is incurred, which satisfy $c_m(i_m) + d_m(i_m) \geq 0$. Rewards and costs are time-discounted with factor $0 < \beta < 1$.

Actions are chosen by adopting a *scheduling policy* π, drawn from the class Π of *admissible policies*, which are nonanticipative relative to the history of states and actions, and engage one bandit at a time. Our focus on such a problem version, instead of on that where *at most* one bandit can be engaged, is without loss of generality. The MABPSC is to find an admissible policy maximizing the expected total discounted value of rewards earned minus switching costs incurred.

We will denote by $X_m(t) \in N_m$ and $a_m(t) \in \{0,1\}$ the state and action for bandit m at period t, respectively, and use the notation

$$a_m^-(t) \triangleq a_m(t-1), \quad \bar{a}_m(t) \triangleq 1 - a_m(t), \quad \text{and} \quad \bar{a}_m^-(t) \triangleq \bar{a}_m(t-1). \tag{2}$$

Since it must be specified whether each bandit m is initially set up, we denote such status by

$a_m^-(0)$. We define the bandit's *augmented state* to be $\widehat{X}_m(t) \triangleq (a_m^-(t), X_m(t))$, which moves over the *augmented state space* $\widehat{N}_m \triangleq \{0,1\} \times N_m$. The *joint augmented state* is thus $\widehat{\mathbf{X}}(t) \triangleq \big(\widehat{X}_m(t)\big)_{m=1}^M$, and the *joint action process* is $\mathbf{a}(t) \triangleq \big(a_m(t)\big)_{m=1}^M$. We can thus formulate the MABPSC as

$$\max_{\pi\in\Pi} \mathbb{E}_{\widehat{\imath}}^{\pi}\left[\sum_{m=1}^{M}\sum_{t=0}^{\infty}\big\{R_m^{a_m(t)}\big(X_m(t)\big) - c_m\big(X_m(t)\big)\overline{a_m^-}(t)a_m(t) - d_m\big(X_m(t)\big)a_m^-(t)\overline{a}_m(t)\big\}\beta^t\right], \tag{3}$$

where $\mathbb{E}_{\widehat{\imath}}^{\pi}[\cdot]$ denotes conditional expectation relative to initial joint state $\widehat{\mathbf{X}}(0) = \widehat{\imath}$ under π.

2.2. Reduction to the Normalized No Shutdown Costs Case

This section shows that it suffices to restrict attention to the no shutdown costs case. Suppose that, at a certain time, which we take to be $t = 0$, a bandit is freshly engaged for a random duration given by a stopping time/rule τ. Dropping the bandit label m, and denoting by $\mathbf{R} = (R_j)$, $\mathbf{c} = (c_j)$ and $\mathbf{d} = (d_j)$ its state-dependent active reward, startup and shutdown cost vectors, we can write the expected discounted net earnings during such a time span, starting at $X(0) = i$, as

$$f_i^{\tau}(\mathbf{R},\mathbf{c},\mathbf{d}) \triangleq \mathbb{E}_i^{\tau}\left[-c_i + \sum_{t=0}^{\tau-1} R_{X(t)}\beta^t - d_{X(\tau)}\beta^{\tau}\right]. \tag{4}$$

We have the following result, where $\mathbf{I}$ is the identity matrix indexed by the state space N, $\mathbf{P} = (p_{ij})_{i,j\in N}$ is the transition probability matrix, and $\mathbf{0}$ is a vector of zeros.

Lemma 2.1 $f_i^{\tau}(\mathbf{R},\mathbf{c},\mathbf{d}) = f_i^{\tau}\big(\mathbf{R} + (\mathbf{I}-\beta\mathbf{P})\mathbf{d}, \mathbf{c}+\mathbf{d}, \mathbf{0}\big)$.

Proof. Use the elementary identity

$$d_{X(\tau)}\beta^{\tau} = d_i - \sum_{t=0}^{\tau-1}\{d_{X(t)} - \beta d_{X(t+1)}\}\beta^t$$

to obtain

$$\begin{aligned} f_i^{\tau}(\mathbf{R},\mathbf{c},\mathbf{d}) &\triangleq -c_i + \mathbb{E}_i^{\tau}\left[\sum_{t=0}^{\tau-1} R_{X(t)}\beta^t - d_{X(\tau)}\beta^{\tau}\right] \\ &= -c_i - d_i + \mathbb{E}_i^{\tau}\left[\sum_{t=0}^{\tau-1}\big\{R_{X(t)} + d_{X(t)} - \beta d_{X(t+1)}\big\}\beta^t\right] = f_i^{\tau}\big(\mathbf{R} + (\mathbf{I}-\beta\mathbf{P})\mathbf{d}, \mathbf{c}+\mathbf{d}, \mathbf{0}\big). \end{aligned}$$

□

Lemma 2.1 shows how to eliminate shutdown costs: one need simply incorporate them into modified startup costs and active rewards given by the transformations

$$\widetilde{\mathbf{c}} \triangleq \mathbf{c}+\mathbf{d} \quad \text{and} \quad \widetilde{\mathbf{R}} \triangleq \mathbf{R}+(\mathbf{I}-\beta\mathbf{P})\mathbf{d}. \tag{5}$$

Note that, in the case $c_j \equiv c$ and $d_j \equiv d$ discussed in Banks and Sundaram (1994), such transformations reduce to $\widetilde{c}_j \equiv c+d$ and $\widetilde{R}_j = R_j+(1-\beta)d$, in agreement with their results.

We will hence focus our discussion henceforth in the *normalized* no shutdown costs case.

2.3. The AT Index

The continuation AT index for a bandit, whose label m we drop from the notation, is

$$\nu^{\mathrm{AT}}_{(1,i)} \triangleq \max_{\tau>0} \frac{\mathbb{E}^{\tau}_i\left[\sum_{t=0}^{\tau-1} R_{X(t)}\beta^t\right]}{\mathbb{E}^{\tau}_i\left[\sum_{t=0}^{\tau-1} \beta^t\right]}, \tag{6}$$

where τ is a stopping time/rule that engages a bandit needing no setup starting at state i; hence, $\nu^{\mathrm{AT}}_{(1,i)}$ is precisely the bandit's Gittins index. The switching AT index is

$$\nu^{\mathrm{AT}}_{(0,i)} \triangleq \max_{\tau>0} \frac{\mathbb{E}^{\tau}_i\left[\sum_{t=0}^{\tau-1} R_{X(t)}\beta^t\right]-c_i}{\mathbb{E}^{\tau}_i\left[\sum_{t=0}^{\tau-1} \beta^t\right]}, \tag{7}$$

where now τ is a stopping time/rule engaging a bandit starting at i that needs to be set up. Notice that $\nu^{\mathrm{AT}}_{(1,i)} \geq \nu^{\mathrm{AT}}_{(0,i)}$ if $c_i \geq 0$, consistently with (1).

2.4. Restless-Bandit Reformulation

Taking $\widehat{X}_m(t)$ as the state of bandit m yields a reformulation of (3) as an MARBP *without* switching costs. The bandit's rewards and transition probabilities in such a reformulation are as follows. If it occupies state (a^-_m, i_m) and is engaged, the active reward $\widehat{R}^1_m(a^-_m, i_m) \triangleq R^1_m(i_m) - c_m(i_m)(1-a^-_m)$ accrues and the state moves to $(1, j_m)$ with active transition probability $\widehat{p}^1_m\big((a^-_m, i_m),(1, j_m)\big) \triangleq p_m(i_m, j_m)$; if rested, the one-period passive reward $\widehat{R}^0_m(a^-_m, i_m) \equiv 0$ accrues, and the state moves to $(0, i_m)$ with a unity passive transition probability, i.e., $\widehat{p}^0_m\big((a^-_m, i_m),(0, i_m)\big) \equiv 1$.

We can thus reformulate (3) as the MARBP

$$\max_{\pi\in\Pi} \mathbb{E}^{\pi}_{\widehat{i}}\left[\sum_{m=1}^{M}\sum_{t=0}^{\infty} \widehat{R}^{a_m(t)}_m\big(\widehat{X}_m(t)\big)\beta^t\right]. \tag{8}$$

3. Restless Bandit Indexation: Theory and Computation

We discuss in this section the restless bandit indexation theory referred to in Section 1, as it applies to a single bandit as above — in its restless reformulation. We hence drop again the bandit label henceforth. so that, e.g., N and $\widehat{N} \triangleq \{0,1\} \times N$ denote the bandit's original and augmented state spaces. We will denote by Π the space of admissible bandit operating policies π. Notice that such a notation distinguishes the latter from their boldface counterparts used in the multi-bandit setting above. We assume that (normalized) startup costs are nonnegative.

Assumption 3.1 *$c_i \geq 0$, for $i \in N$.*

3.1. Indexability and the MPI

We use two criteria to evaluate a policy π, relative to an initial state (a_0^-, i_0): the *reward measure*

$$f^{\pi}_{(a_0^-,i_0)} \triangleq \mathbb{E}^{\pi}_{(a_0^-,i_0)}\left[\sum_{t=0}^{\infty} \widehat{R}\big(\widehat{X}(t)\big)\beta^t\right],$$

giving the expected total discounted value of *net rewards* — net of switching costs — that accrue on the bandit; and the *work measure*

$$g^{\pi}_{(a_0^-,i_0)} \triangleq \mathbb{E}^{\pi}_{(a_0^-,i_0)}\left[\sum_{t=0}^{\infty} a(t)\beta^t\right],$$

giving the corresponding expected total discounted amount of *work* expended. We will actually consider the average measures f^{π} and g^{π} obtained by drawing the initial state from a positive probability mass function $p_{(a^-,i)} > 0$ for $(a^-,i) \in \widehat{N}$.

Imagining that work is paid for at *wage* rate ν leads us to consider the *ν-wage problem*

$$\max_{\pi \in \Pi} f^{\pi} - \nu g^{\pi}, \tag{9}$$

which is to find an admissible bandit operating policy achieving the maximum value of net rewards earned minus labor costs incurred. We will use (9) to *calibrate* the *marginal value of work* at each state, by analyzing the structure of optimal policies as ν varies.

MDP theory ensures that for every wage $\nu \in \mathbb{R}$ there exists an optimal policy that is stationary deterministic and independent of the initial state. Any such a policy is characterized by its *active set*, or subset of states where it prescribes to engage the bandit. We will write active sets as

$$S_0 \oplus S_1 \triangleq \{0\} \times S_0 \cup \{1\} \times S_1, \quad S_0, S_1 \subseteq N.$$

Thus, the policy that we denote by $S_0 \oplus S_1$ engages the bandit when it was previously rested (resp. engaged) if the original state $X(t)$ lies in S_0 (resp. in S_1).

Hence, to any wage ν there corresponds a unique *maximal optimal active set* $S_0^*(\nu) \oplus S_1^*(\nu) \subseteq \widehat{N}$, which is the union of all optimal active sets. Now, we say that the bandit is *indexable* if there exists an *index* $\nu^*_{(a^-,i)}$ for $(a^-,i) \in \widehat{N}$ such that

$$S_0^*(\nu) = \big\{(0,i)\colon \nu^*_{(0,i)} \geq \nu\big\} \quad \text{and} \quad S_1^*(\nu) = \big\{(1,i)\colon \nu^*_{(1,i)} \geq \nu\big\}, \quad \nu \in \mathbb{R}.$$

We then say that $\nu^*_{(a^-,i)}$ is the bandit's *marginal productivity index* (MPI), or *Whittle index*, terming $\nu^*_{(1,i)}$ the *continuation MPI*, and $\nu^*_{(0,i)}$ the *switching MPI*.

Thus, the bandit is indexable with MPI $\nu^*_{(a^-,i)}$ if it is optimal in (9), to engage (resp. rest) the bandit when it occupies state (a^-,i) iff $\nu^*_{(a^-,i)} \geq \nu$ (resp. $\nu^*_{(a^-,i)} \leq \nu$).

To establish indexability and compute the MPI, we developed in Niño-Mora (2001, 2002, 2006a, 2007a) an approach based on positing and then establishing the structure of optimal active sets, as an *active-set family* $\widehat{\mathscr{F}} \subseteq 2^{\widehat{N}}$ that *contains* all sets $S_0^*(\nu) \oplus S_1^*(\nu)$ as ν varies, under a possibly restricted range of reward/cost parameters. The intuition that, if startup costs satisfy Assumption 3.1, optimal policies should have the hysteretic property that, if it is optimal to engage a bandit when it was previously rested, then, other things being equal, it should be optimal to engage it when it was previously active, leads us to guess that the right choice of $\widehat{\mathscr{F}}$ should be

$$\widehat{\mathscr{F}} \triangleq \big\{S_0 \oplus S_1\colon S_0 \subseteq S_1 \subseteq N\big\}. \tag{10}$$

Notice that $\widehat{\mathscr{F}}$ represents a family of policies consistent with (1), which we posit to contain the optimal policies for (9). When $S_0 \neq S_1$, such policies present the *hysteresis region* $S_1 \setminus S_0$, on which bandit dynamics depend on the previous action. We will thus aim to establish indexability relative to such a family, meaning that the bandit is indexable and $S_0^*(\nu) \oplus S_1^*(\nu) \in \widehat{\mathscr{F}}$ for $\nu \in \mathbb{R}$.

3.2. PCL-Indexability and Adaptive-Greedy Index Algorithm

We next discuss the approach we will deploy to establish indexability and compute the MPI of the restless bandits of concern herein, based on showing that they are PCL-indexable relative to $\widehat{\mathscr{F}}$, and using the adaptive-greedy index algorithm that is valid for such bandits.

Given an action $a \in \{0,1\}$ and an active set $S_0 \oplus S_1 \in \widehat{\mathscr{F}}$, denote by $\langle a, S_0 \oplus S_1\rangle$ the policy that takes action a in the initial period and adopts the $S_0 \oplus S_1$-*active policy* thereafter. Now, for an augmented state (a^-,i) and an active set $S_0 \oplus S_1 \in \widehat{\mathscr{F}}$, define the *marginal work measure*

$$w^{S_0 \oplus S_1}_{(a^-,i)} \triangleq g^{\langle 1, S_0 \oplus S_1\rangle}_{(a^-,i)} - g^{\langle 0, S_0 \oplus S_1\rangle}_{(a^-,i)}, \tag{11}$$

along with the *marginal reward measure*

$$r^{S_0\oplus S_1}_{(a^-,i)} \triangleq f^{\langle 1,S_0\oplus S_1\rangle}_{(a^-,i)} - f^{\langle 0,S_0\oplus S_1\rangle}_{(a^-,i)} \tag{12}$$

and the *marginal productivity measure*

$$\nu^{S_0\oplus S_1}_{(a^-,i)} \triangleq \frac{r^{S_0\oplus S_1}_{(a^-,i)}}{w^{S_0\oplus S_1}_{(a^-,i)}}. \tag{13}$$

As will see (cf. Proposition 4.4), the latter measure is well defined, as its denominator is positive.

We will deploy the PCL-indexability approach to indexation in Niño-Mora (2007a), which revises that introduced and developed in Niño-Mora (2001, 2002, 2006a). For an active set $\widehat{S} = S_0 \oplus S_1 \in \widehat{\mathscr{F}}$, let

$$\partial^{\text{out}}_{\widehat{\mathscr{F}}}\widehat{S} \triangleq \big\{(a^-,i) \in \widehat{S}^c \colon \widehat{S}\cup\{(a^-,i)\} \in \widehat{\mathscr{F}}\big\} = \big\{(0,i)\colon i \in S_1 \setminus S_0\big\} \cup \big\{(1,i)\colon i \in S_1^c\big\}, \tag{14}$$

where $\widehat{S}^c \triangleq \widehat{N} \setminus \widehat{S}$ and $S_1^c \triangleq N \setminus S_1$, be the *outer boundary of* $\widehat{S}$ *relative to* $\widehat{\mathscr{F}}$; and let

$$\partial^{\text{in}}_{\widehat{\mathscr{F}}}\widehat{S} \triangleq \big\{(a^-,i) \in \widehat{S} \colon \widehat{S}\setminus\{(a^-,i)\} \in \widehat{\mathscr{F}}\big\} = \big\{(1,i)\colon i \in S_1 \setminus S_0\big\} \cup \big\{(0,i)\colon i \in S_0\big\} \tag{15}$$

be the corresponding *inner boundary*. Note that the right-most identities in (14)–(15) follow from (10). Now, we require that *set system* $(\widehat{N},\widehat{\mathscr{F}})$ be *monotonically connected*, which in the present setting means that:

(i) $\emptyset, \widehat{N} \in \widehat{\mathscr{F}}$;

(ii) for every $\widehat{S},\widehat{S}' \in \mathscr{F}$ with $\widehat{S} \subset \widehat{S}'$ there exist $(a,j) \in \partial^{\text{out}}_{\widehat{\mathscr{F}}}\widehat{S}$ and $(a',j') \in \partial^{\text{in}}_{\widehat{\mathscr{F}}}\widehat{S}'$ such that $\widehat{S} \subset \widehat{S}\cup\{(a,j)\} \subseteq \widehat{S}'$ and $\widehat{S} \subseteq \widehat{S}' \setminus \{(a',j')\} \subset \widehat{S}'$;

(iii) for any $\widehat{S},\widehat{S}' \in \widehat{\mathscr{F}}$ with $\widehat{S} \neq \widehat{S}'$, it holds that $\widehat{S}\cup\widehat{S}' \in \widehat{\mathscr{F}}$,

As the reader can immediately verify, the $\widehat{\mathscr{F}}$ defined in (10) satisfies indeed such conditions.

Now, we will say that the bandit is *PCL-indexable* relative to $\widehat{\mathscr{F}}$, or *PCL*$(\widehat{\mathscr{F}})$*-indexable*, if:

(i) for each active set $\widehat{S} \in \widehat{\mathscr{F}}$, $w^{\widehat{S}}_{(a^-,i)} > 0$ for $(a^-,i) \in \widehat{N}$; and

(ii) for every wage $\nu \in \mathbb{R}$ there exists an optimal policy for (9) with active set $\widehat{S} \in \widehat{\mathscr{F}}$.

We will further refer to the *adaptive-greedy algorithmic scheme* $\text{AG}_{\widehat{\mathscr{F}}}$ shown in Table 1, where $n \triangleq |N|$ denotes the number of bandit states in the original (nonrestless) formulation. The algorithm

Table 1: Version 1 of Adaptive-Greedy Algorithmic Scheme $\mathrm{AG}_{\widehat{\mathscr{F}}}$.

ALGORITHM $\mathrm{AG}_{\widehat{\mathscr{F}}}$:
Output: $\left\{(a_k^-, i_k), \nu^*_{(a_k^-,i_k)}\right\}_{k=1}^{2n}$

$\widehat{S}^0 := \emptyset \oplus \emptyset$
for $k := 1$ **to** $2n$ **do**
  pick $(a_k^-, i_k) \in \arg\max \left\{\nu^{\widehat{S}^{k-1}}_{(a^-,i)} : (a^-, i) \in \partial^{\mathrm{out}}_{\widehat{\mathscr{F}}} \widehat{S}^{k-1}\right\}$
  $\nu^*_{(a_k^-,i_k)} := \nu^{\widehat{S}^{k-1}}_{(a_k^-,i_k)}$; $\widehat{S}^k := \widehat{S}^{k-1} \cup \{(a_k^-, i_k)\}$
end { for }

produces an output consisting of a string $\{(a_k^-, i_k)\}_{k=1}^{2n}$ of distinct augmented states spanning $\widehat{N}$, with $\widehat{S}^k \triangleq \{(a_1^-, i_1), \ldots, (a_k^-, i_k)\} \in \widehat{\mathscr{F}}$, for $1 \leq k \leq 2n$, along with corresponding index values $\{\nu^*_{(a_k^-,i_k)}\}_{k=1}^{2n}$. Ties for picking the (a_k^-, i_k)'s are broken arbitrarily. We use the term *algorithmic scheme* as it is not yet specified how to compute the required marginal productivity rates.

We will later invoke the following key result, introduced and developed in Niño-Mora (2001, 2002, 2006a, 2007a), which refers to a generic restless bandit and active-set family F.

Theorem 3.2 *A PCL($\mathscr{F}$)-indexable bandit is indexable and algorithm* $\mathrm{AG}_{\mathscr{F}}$ *computes its MPI.*

Using the definition of $\widehat{\mathscr{F}}$ in (10) yields the more explicit *Version 2* of the algorithm shown in Table 2, where the output is decoupled. We use in this and later versions a more algorithm-like notation, writing, e.g., $\nu^{S_0^{k_0-1} \oplus S_1^{k_1-1}}_{(0,j)}$ as $\nu^{(k_0-1,k_1-1)}_{(0,j)}$. Notice that the active sets constructed in both versions are related by $\widehat{S}^{k-1} \triangleq S_0^{k_0-1} \oplus S_1^{k_1-1}$, with $k = k_0 + k_1 - 1$ and $k_0 \leq k_1$. Version 2 draws on the fact that, at each step, the algorithm augments the current active set by a state that can be of the form $(1,i)$ or $(0,i)$. Sets $S_0^{k_0}$ and $S_1^{k_1}$ in the algorithm are $S_0^{k_0} = \{i_0^1, \ldots, i_0^{k_0}\}$ and $S_1^{k_1} = \{i_1^1, \ldots, i_1^{k_1}\}$, and satisfy that $S_0^{k_0} \subset S_1^{k_1}$, for $1 \leq k_0 < k_1 \leq n$, consistently with (10).

3.3. Optimality of Hysteretic $\widehat{\mathscr{F}}$-Policies

We proceed to show that PCL($\widehat{\mathscr{F}}$)-indexability condition (ii) above holds for the model of concern, namely that $\widehat{\mathscr{F}}$*-policies*, i.e., those with active sets $\widehat{S} \in \widehat{\mathscr{F}}$, solve (9). For such a purpose we will use the *Bellman equations* characterizing the value function $\vartheta^*_{(a^-,i)}(\nu)$ for (9) starting at (a^-, i):

$$\vartheta^*_{(a^-,i)}(\nu) = \max\left\{\beta\vartheta^*_{(0,i)}(\nu), R_i - (1-a^-)c_i - \nu + \beta\sum_{j\in N} p_{ij}\vartheta^*_{(1,j)}(\nu)\right\},\ (a^-,i) \in \widehat{N}. \qquad (16)$$

Table 2: Version 2 of Algorithmic Scheme $\mathrm{AG}_{\widehat{\mathscr{F}}}$.

ALGORITHM $\mathrm{AG}_{\widehat{\mathscr{F}}}$:
Output: $\left\{(0,i_0^{k_0}),\nu^*_{(0,i_0^{k_0})}\right\}_{k_0=1}^{n}$, $\left\{(1,i_1^{k_1}),\nu^*_{(1,i_1^{k_1})}\right\}_{k_1=1}^{n}$

$S_0^0:=\emptyset$; $S_1^0:=\emptyset$; $k_0:=1$; $k_1:=1$
while $k_0+k_1\le 2n+1$ **do**
  if $k_1\le n$ **pick** $j_1^{\max}\in\arg\max\left\{\nu^{(k_0-1,k_1-1)}_{(1,j)}:\ j\in N\setminus S_1^{k_1-1}\right\}$
  if $k_0<k_1$ **pick** $j_0^{\max}\in\arg\max\left\{\nu^{(k_0-1,k_1-1)}_{(0,j)}:\ j\in S_1^{k_1-1}\setminus S_0^{k_0-1}\right\}$
  if $k_1=n+1$ **or** $\left\{k_0<k_1\le n\right.$ **and** $\left.\nu^{(k_0-1,k_1-1)}_{(1,j_1^{\max})}<\nu^{(k_0-1,k_1-1)}_{(0,j_0^{\max})}\right\}$
    $i_0^{k_0}:=j_0^{\max}$; $\nu^*_{(0,i_1^{k_0})}:=\nu^{(k_0-1,k_1-1)}_{(0,i_1^{k_0})}$; $S_0^{k_0}:=S_0^{k_0-1}\cup\{i_0^{k_0}\}$; $k_0:=k_0+1$
  else
    $i_1^{k_1}:=j_1^{\max}$; $\nu^*_{(1,i_1^{k_1})}:=\nu^{(k_0-1,k_1-1)}_{(1,i_1^{k_1})}$; $S_1^{k_1}:=S_1^{k_1-1}\cup\{i_1^{k_1}\}$; $k_1:=k_1+1$
  end { if }
end { while }

Proposition 3.3 *For every wage $\nu\in\mathbb{R}$ there exists an optimal policy for (9) with active set $\widehat{S}\in\widehat{\mathscr{F}}$, i.e., if it is optimal to rest the bandit in state $(1,i)$ then it is optimal to rest it in $(0,i)$.*

Proof. Fix ν. Formulate the assumption that it is optimal to rest the bandit in $(1,i)$ as

$$\beta\,\vartheta^*_{(0,i)}(\nu)\ge R_i-\nu+\beta\sum_{j\in N}p_{ij}\vartheta^*_{(1,j)}(\nu). \tag{17}$$

We want to show that this implies the optimality of resting it in state $(0,i)$, i.e.,

$$\beta\,\vartheta^*_{(0,i)}(\nu)\ge R_i-c_i-\nu+\beta\sum_{j\in N}p_{ij}\vartheta^*_{(1,j)}(\nu).$$

But this follows immediately, by writing

$$\beta\,\vartheta^*_{(0,i)}(\nu)\ge R_i-\nu+\beta\sum_{j\in N}p_{ij}\vartheta^*_{(1,j)}(\nu)\ge R_i-c_i-\nu+\beta\sum_{j\in N}p_{ij}\vartheta^*_{(1,j)}(\nu),$$

where we have used (17) and Assumption 3.1. □

Note that Proposition 3.3 establishes PCL($\widehat{\mathscr{F}_T}$)-indexability condition (ii) above. In order to further establish the remaining condition (i) and to simplify the index algorithm we will have to draw on the work-reward analysis carried out in the next section.

4. Work-Reward Analysis and PCL($\widehat{\mathscr{F}}$)-Indexability Proof

We set out in this section to carry out a work-reward analysis of a single bandit with startup costs as above, in its restless reformulation, and to establish its PCL($\widehat{\mathscr{F}}$)-indexability.

4.1. Work and Marginal Work Measures

We start by addressing calculation of work and marginal work measures $g_{(a^-,i)}^{S_0\oplus S_1}$ and $w_{(a^-,i)}^{S_0\oplus S_1}$. We will show that they are closely related to their counterparts g_i^S and w_i^S for the underlying nonrestless bandit, where stationary deterministic policies are represented by their active sets $S \subseteq N$.

For each $S \subseteq N$, work measures g_i^S are characterized by the evaluation equations

$$g_i^S = \begin{cases} 1+\beta \sum_{j\in S} p_{ij} g_j^S & \text{if } i \in S \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

Notice that the solution to (18) is unique, since matrix $\mathbf{I}_S - \beta \mathbf{P}_{SS}$ is invertible, as $\mathbf{P}_{SS}$ is a substochastic matrix and $0 < \beta < 1$, where $\mathbf{I}_S$ is the identity matrix indexed by S and $\mathbf{P}_{SS} \triangleq (p_{ij})_{i,j\in S}$.

Further, the marginal work measure w_i^S is evaluated by

$$w_i^S \triangleq g_i^{\langle 1,S\rangle} - g_i^{\langle 0,S\rangle} = 1+\beta \sum_{j\in N} p_{ij} g_j^S - \beta g_i^S = \begin{cases} (1-\beta) g_i^S & \text{if } i \in S \\ 1+\beta \sum_{j\in S} p_{ij} g_j^S & \text{otherwise.} \end{cases} \tag{19}$$

Notice that (18) and (19) imply that

$$w_i^S > 0, \quad i \in N. \tag{20}$$

We now return to the bandit's restless reformulation. The following result gives the evaluation equations for work measure $g_{(a^-,i)}^{S_0\oplus S_1}$, for a given active set $S_0 \oplus S_1 \in \widehat{\mathscr{F}}$.

Lemma 4.1

$$g_{(a^-,i)}^{S_0\oplus S_1} = \begin{cases} 1+\beta \sum_{j\in N} p_{ij} g_{(1,j)}^{S_0\oplus S_1} & \textit{if } i \in S_{a^-} \\ \beta g_{(0,i)}^{S_0\oplus S_1} & \textit{otherwise.} \end{cases}$$

The next result represents work measure $g_{(a^-,i)}^{S_0\oplus S_1}$ in terms of the g_i^S's.

Lemma 4.2 *For $S_0 \oplus S_1 \in \widehat{\mathscr{F}}$:*

(a) $g^{S_0\oplus S_1}_{(a^-,i)} = g^{S_1}_i = 0$, *for* $a^- \in \{0,1\}, i \in S_1^c$.

(b) $g^{S_0\oplus S_1}_{(1,i)} = g^{S_1}_i$, *for* $i \in S_1$.

(c) $g^{S_0\oplus S_1}_{(0,i)} = g^{S_1}_i$, *for* $i \in S_0$.

(d) $g^{S_0\oplus S_1}_{(0,i)} = 0$, *for* $i \in S_1 \setminus S_0$.

Proof. (a) This part follows immediately from the definition of policy $S_0 \oplus S_1$.

(b) For $i \in S_1$, we can write

$$g^{S_0\oplus S_1}_{(1,i)} = 1 + \beta \sum_{j\in S_1} p_{ij} g^{S_0\oplus S_1}_{(1,j)} + \beta \sum_{j\in S_1^c} p_{ij} g^{S_0\oplus S_1}_{(1,j)} = 1 + \beta \sum_{j\in S_1} p_{ij} g^{S_0\oplus S_1}_{(1,j)},$$

where we have used Lemma 4.1 and part (a). Hence, the $g^{S_0\oplus S_1}_{(1,i)}$'s satisfy the evaluation equations in (18) characterizing the $g^{S_1}_i$'s, for $i \in S_1$, which yields the result.

(c) We have, for $i \in S_0$, that

$$g^{S_0\oplus S_1}_{(0,i)} = 1 + \beta \sum_{j\in S_1} p_{i,j} g^{S_0\oplus S_1}_{(1,j)} + \beta \sum_{j\in S_1^c} p_{ij} g^{S_0\oplus S_1}_{(1,j)} = g^{S_0\oplus S_1}_{(1,i)} = g^{S_1}_i,$$

where we have used Lemma 4.1, the relation $S_0 \subseteq S_1$ and parts (a, b).

(d) This part follows immediately from the definition of policy $S_0 \oplus S_1$. □

Regarding $w^{S_0\oplus S_1}_{(a^-,i)}$, we readily obtain from (11) and Lemma 4.1 that

$$w^{S_0\oplus S_1}_{(0,i)} = w^{S_0\oplus S_1}_{(1,i)} = 1 + \beta \sum_{j\in N} p_{ij} g^{S_0\oplus S_1}_{(1,j)} - \beta g^{S_0\oplus S_1}_{(0,i)}. \tag{21}$$

The following result represents marginal workloads $w^{S_0\oplus S_1}_{(a^-,i)}$ in terms of the w^S_i's.

Lemma 4.3 *For* $a^- \in \{0,1\}, S_0 \oplus S_1 \in \widehat{\mathscr{F}}$:

(a) $w^{S_0\oplus S_1}_{(a^-,i)} = w^{S_1}_i$, *for* $i \in S_0 \cup S_1^c$.

(b) $w^{S_0\oplus S_1}_{(a^-,i)} = w^{S_1}_i/(1-\beta)$, *for* $i \in S_1 \setminus S_0$.

Proof. (a) We can write, for $i \in S_0 \cup S_1^c$,

$$w^{S_0\oplus S_1}_{(a^-,i)} = 1 + \beta \sum_{j\in N} p_{ij} g^{S_0\oplus S_1}_{(1,j)} - \beta g^{S_0\oplus S_1}_{(0,i)} = 1 + \beta \sum_{j\in S_1} p_{ij} g^{S_1}_j - \beta g^{S_1}_i = w^{S_1}_i,$$

where we have used (21), Lemma 4.2(a, b, c) and (19).

(b) We have, for $i \in S_1 \setminus S_0$,

$$w_{(a^-,i)}^{S_0 \oplus S_1} = 1 + \beta \sum_{j \in N} p_{ij} g_{(1,j)}^{S_0 \oplus S_1} - \beta g_{(0,i)}^{S_0 \oplus S_1} = 1 + \beta \sum_{j \in S_1} p_{ij} g_j^{S_1} = w_i^{S_1} + \beta g_i^{S_1} = \frac{w_i^{S_1}}{1-\beta},$$

where we have used (21), Lemma 4.2(a, b, d) and (19). □

From the above, we obtain the required positivity or marginal workloads.

Proposition 4.4 $w_{(a^-,i)}^{S_0 \oplus S_1} > 0$, *for* $(a^-, i) \in \widehat{N}, S_0 \oplus S_1 \in \widehat{\mathscr{F}}$.

Proof. The result follows immediately from (20) via Lemma 4.3. □

4.2. Reward and Marginal Reward Measures

We continue by addressing calculation of required reward and marginal reward measures $f_{(a^-,i)}^{S_0 \oplus S_1}$ and $r_{(a^-,i)}^{S_0 \oplus S_1}$. Again, we will show that they are closely related to their counterparts f_i^S and r_i^S for the underlying nonrestless bandit wit no startup costs.

For each active set $S \subseteq N$, the reward measure f_i^S is characterized by the evaluation equations

$$f_i^S = \begin{cases} R_i + \beta \sum_{j \in S} p_{ij} f_j^S & \text{if } i \in S \\ \\ 0 & \text{otherwise,} \end{cases} \tag{22}$$

while the marginal reward measure r_i^S is given by

$$r_i^S \triangleq f_i^{\langle 1,S \rangle} - f_i^{\langle 0,S \rangle} = R_i + \beta \sum_{j \in S} p_{ij} f_j^S - \beta f_i^S = \begin{cases} (1-\beta) f_i^S & \text{if } i \in S \\ \\ R_i + \beta \sum_{j \in S} p_{ij} f_j^S & \text{otherwise.} \end{cases} \tag{23}$$

Returning to the restless formulation, the next result gives the evaluation equations for reward measures $f_{(a^-,i)}^{S_0 \oplus S_1}$, for a given active set $S_0 \oplus S_1 \in \widehat{\mathscr{F}}$. Recall the notation in (2).

Lemma 4.5

$$f_{(a^-,i)}^{S_0 \oplus S_1} = \begin{cases} R_i - (1-a^-)c_i + \beta \sum_{j \in N} p_{ij} f_{(1,j)}^{S_0 \oplus S_1} & \textit{if } i \in S_{a^-} \\ \\ \beta f_{(0,i)}^{S_0 \oplus S_1} & \textit{otherwise.} \end{cases}$$

The next result represents reward measure $f_{(a^-,i)}^{S_0 \oplus S_1}$ in terms of the f_i^S's.

Lemma 4.6 *For* $S_0 \oplus S_1 \in \widehat{\mathscr{F}}$:

(a) $f^{S_0\oplus S_1}_{(a^-,i)} = 0 = f^{S_1}_i$, *for* $a^- \in \{0,1\}, i \in S_1^c$.

(b) $f^{S_0\oplus S_1}_{(1,i)} = f^{S_1}_i$, *for* $i \in S_1$.

(c) $f^{S_0\oplus S_1}_{(0,i)} = f^{S_1}_i - c_i$, *for* $i \in S_0$.

(d) $f^{S_0\oplus S_1}_{(0,i)} = 0 = f^{S_0}_i$, *for* $i \in S_1 \setminus S_0$.

Proof. (a) This part is straightforward.

(b) We can write, for $i \in S_1$,

$$f^{S_0\oplus S_1}_{(1,i)} = R_i + \beta \sum_{j\in S_1} p_{ij} f^{S_0\oplus S_1}_{(1,j)} + \beta \sum_{j\in S_1^c} p_{ij} f^{S_0\oplus S_1}_{(1,j)} = R_i + \beta \sum_{j\in S_1} p_{ij} f^{S_0\oplus S_1}_{(1,j)},$$

where we have used Lemma 4.5 and part (a). Hence, the $f^{S_0\oplus S_1}_{(1,i)}$'s, for $i \in S_1$, satisfy the evaluation equations in (22) for corresponding terms $f^{S_1}_i$, which yields the result.

(c) We can write, for $i \in S_0$,

$$f^{S_0\oplus S_1}_{(0,i)} = R_i - c_i + \beta \sum_{j\in S_1} p_{ij} f^{S_0\oplus S_1}_{(1,j)} + \beta \sum_{j\in S_1^c} p_{ij} f^{S_0\oplus S_1}(1,j) = f^{S_0\oplus S_1}_{(1,i)} - c_i = f^{S_1}_i - c_i,$$

where we have used that $S_0 \subseteq S_1$ along with parts (a, b).

(d) This part is straightforward. □

Regarding marginal reward measure $r^{S_0\oplus S_1}_{(a^-,i)}$, we obtain from (12) and Lemma 4.5 that

$$r^{S_0\oplus S_1}_{(a^-,i)} = R_i - (1-a^-)c_i + \beta \sum_{j\in N} p_{ij} f^{S_0\oplus S_1}_{(1,j)} - \beta f^{S_0\oplus S_1}_{(0,i)}. \tag{24}$$

The following result represents $r^{S_0\oplus S_1}_{(a^-,i)}$ in terms of the r^S_i's.

Lemma 4.7 *For* $S_0 \oplus S_1 \in \widehat{\mathscr{F}}$:

(a) $r^{S_0\oplus S_1}_{(0,i)} = r^{S_0\oplus S_1}_{(1,i)} - c_i$, *for* $i \in N$.

(b) $r^{S_0\oplus S_1}_{(1,i)} = r^{S_1}_i$, *for* $i \in S_1^c$.

(c) $r^{S_0\oplus S_1}_{(1,i)} = r^{S_1}_i + \beta c_i$, *for* $i \in S_0$.

(d) $r^{S_0\oplus S_1}_{(1,i)} = r^{S_1}_i/(1-\beta)$, *for* $i \in S_1 \setminus S_0$.

Proof. (a) This part follows immediately from (24).

(b) We can write, for $i \in S_1^c$,

$$r_{(1,i)}^{S_0 \oplus S_1} = R_i + \beta \sum_{j \in N} p_{ij} f_{(1,j)}^{S_0 \oplus S_1} - f_{(1,i)}^{S_0 \oplus S_1} = R_i + \beta \sum_{j \in S_1} p_{ij} f_j^{S_1} - f_i^{S_1} = r_i^{S_1},$$

where we have used (24), Lemma 4.6(a, b), and (23).

(c) We can write, for $i \in S_0$,

$$r_{(1,i)}^{S_0 \oplus S_1} = R_i + \beta \sum_{j \in N} p_{ij} f_{(1,j)}^{S_0 \oplus S_1} - \beta f_{(0,i)}^{S_0 \oplus S_1} = R_i + \beta \sum_{j \in S_1} p_{ij} f_j^{S_1} - \beta (f_i^{S_1} - c_i) = r_i^{S_1} + \beta c_i,$$

where we have used $S_0 \subseteq S_1$, (24), Lemma 4.6(a, b, c) and (23).

(d) We can write, for $i \in S_1 \setminus S_0$,

$$r_{(1,i)}^{S_0 \oplus S_1} = R_i + \beta \sum_{j \in N} p_{ij} f_{(1,j)}^{S_0 \oplus S_1} - \beta f_{(0,i)}^{S_0 \oplus S_1} = f_i^{S_1} = \frac{r_i^{S_1}}{1-\beta},$$

where we have used (24), Lemma 4.6(a, b, d), (22) and (23). This completes the proof. □

4.3. Marginal Productivity Measures

We continue by addressing calculation of required marginal productivity measures $\nu_{(a^-,i)}^{S_0 \oplus S_1}$ in (13). Again, we will show that they are closely related to their counterparts ν_i^S for the underlying non-restless bandit without startup costs, given by

$$\nu_i^S \triangleq \frac{r_i^S}{w_i^S}, \quad i \in N, S \subseteq N. \tag{25}$$

The next result represents the required $\nu_{(a^-,i)}^{S_0 \oplus S_1}$'s in terms of the ν_i^S's.

Lemma 4.8 *For $S_0 \oplus S_1 \in \widehat{\mathscr{F}}$:*

(a) $\nu_{(0,i)}^{S_0 \oplus S_1} = \nu_{(1,i)}^{S_0 \oplus S_1} - c_i / w_{(1,i)}^{S_0 \oplus S_1}$, *for* $i \in N$.

(b) $\nu_{(1,i)}^{S_0 \oplus S_1} = \nu_i^{S_1} = \nu_{(1,i)}^{\emptyset \oplus S_1}$, *for* $i \in S_0^c$.

(c) $\nu_{(1,i)}^{S_0 \oplus S_1} = \nu_i^{S_1} + \beta c_i / w_i^{S_1}$, *for* $i \in S_0$.

(d) $\nu_{(0,i)}^{S_0 \oplus S_1} = \nu_i^{S_1} - (1-\beta) c_i / w_i^{S_1} = \nu_{(0,i)}^{\emptyset \oplus S_1}$, $i \in S_1 \setminus S_0$.

Proof. All parts follow immediately from (13), (25), Lemma 4.3 and Lemma 4.7. □

4.4. Proof of PCL($\widehat{\mathscr{F}}$)-Indexability

We next draw on the above results to establish that the restless bandits of concern are PCL($\widehat{\mathscr{F}}$)-indexable, which ensures the validity of index algorithm AG$_{\widehat{\mathscr{F}}}$ via Theorem 3.2. See Section 3.2.

Theorem 4.9 *Under Assumption* 3.1, *the restless reformulation of a bandit with switching costs is PCL($\widehat{\mathscr{F}}$)-indexable. Hence, it is indexable, and algorithm* AG$_{\widehat{\mathscr{F}}}$ *computes its MPI.*

Proof. The defining PCL($\widehat{\mathscr{F}}$)-indexability conditions (i) and (ii) in Section 3.2 were established in Propositions 4.4 and 3.3, respectively. The proof is completed by invoking Theorem 3.2. □

4.5. Further Simplification of the Index Algorithm

The above results allow us to further simplify Version 2 of index algorithm AG$_{\widehat{\mathscr{F}}}$ into the *Version 3* shown in Table 3. In the latter, we use Lemma 4.8(b, d) to represent required marginal productivity rates $\nu^{S_0\oplus S_1}_{(a^-,i)}$ in terms of the ν^S_i's. Notice that in Version 3 we use $\nu^{(0,k_1-1)}_{(0,j)}$ (which denotes $\nu^{S_0\oplus S_{k_1-1}}_{(0,j)}$) in place of $\nu^{(k_0-1,k_1-1)}_{(0,j)}$, drawing on Lemma 4.8(d). We do so for computational reasons, as storage of quantities $\nu^{(0,k_1-1)}_{(0,j)}$ requires one less dimension than storage of the $\nu^{(k_0-1,k_1-1)}_{(0,j)}$'s.

Table 3: Version 3 of Algorithmic Scheme AG$_{\widehat{\mathscr{F}}}$.

ALGORITHM AG$_{\widehat{\mathscr{F}}}$:
Output: $\big\{(0,i_0^{k_0}),\nu^*_{(0,i_0^{k_0})}\big\}_{k_0=1}^n$, $\big\{(1,i_1^{k_1}),\nu^*_{(1,i_1^{k_1})}\big\}_{k_1=1}^n$

$S_0^0:=\emptyset$; $S_1^0:=\emptyset$; $k_0:=1$; $k_1:=1$
while $k_0+k_1\le 2n+2$ **do**
  if $k_1\le n$ **pick** $j_1^{\max}\in\arg\max\big\{\nu_j^{(k_1-1)}:\ j\in N\setminus S_1^{k_1-1}\big\}$
  $\nu^{(0,k_1-1)}_{(0,j)}:=\nu_j^{(k_1-1)}-(1-\beta)c_j/w_j^{(k_1-1)}$, $j\in S_1^{k_1-1}\setminus S_0^{k_0-1}$
  if $k_0<k_1$ **pick** $j_0^{\max}\in\arg\max\big\{\nu^{(0,k_1-1)}_{(0,j)}:\ j\in S_1^{k_1-1}\setminus S_0^{k_0-1}\big\}$
  if $k_1=n+1$ **or** $\big\{k_0<k_1\le n$ **and** $\nu_{j_1^{\max}}^{(k_1-1)}<\nu^{(0,k_1-1)}_{(0,j_0^{\max})}\big\}$
    $i_0^{k_0}:=j_0^{\max}$; $\nu^*_{(0,i_0^{k_0})}:=\nu^{(0,k_1-1)}_{(0,i_0^{k_0})}$; $S_0^{k_0}:=S_0^{k_0-1}\cup\{i_0^{k_0}\}$; $k_0:=k_0+1$
  else
    $i_1^{k_1}:=j_1^{\max}$; $\nu^*_{i_1^{k_1}}:=\nu^{(k_1-1)}_{i_1^{k_1}}$; $S_1^{k_1}:=S_1^{k_1-1}\cup\{i_1^{k_1}\}$; $k_1:=k_1+1$
  end { if }
end { while }

4.6. The MPI is the AT Index

We next establish the identity between the MPI and the AT index for the bandits of concern in this paper. We will find it convenient to reformulate the expressions for the AT index, given in (6)–(7) in terms of stopping times, using instead active sets $S \subseteq N$ to represent the latter — as it suffices to consider stationary deterministic policies. In the above notation, we can thus formulate the continuation and switching AT indices as

$$\nu^{\mathrm{AT}}_{(1,i)} \triangleq \max_{i \in S \subseteq N} \frac{f^S_i}{g^S_i}, \tag{26}$$

and

$$\nu^{\mathrm{AT}}_{(0,i)} \triangleq \max_{i \in S \subseteq N} \frac{f^S_i - c_i}{g^S_i}. \tag{27}$$

Recall that we denote the MPI by $\nu^*_{(a^-,i)}$.

Proposition 4.10 *Under Assumption* 3.1, $\nu^*_{(1,i)} = \nu^{\mathrm{AT}}_{(1,i)}$ *and* $\nu^*_{(0,i)} = \nu^{\mathrm{AT}}_{(0,i)}$, *for* $i \in N$.

Proof. We first show that $\nu^*_{(1,i)} = \nu^{\mathrm{AT}}_{(1,i)}$, through the equivalences

$$\begin{aligned}
\nu \geq \nu^*_{(1,i)} &\iff \text{it is optimal in (9) to rest the bandit at } (1,i) \\
&\iff 0 \geq \max_{S_0 \subseteq S_1 \subseteq N:\ i \in S_1} f^{S_0 \oplus S_1}_{(1,i)} - \nu g^{S_0 \oplus S_1}_{(1,i)} \\
&\iff \nu \geq \max_{S_0 \subseteq S_1 \subseteq N:\ i \in S_1} \frac{f^{S_0 \oplus S_1}_{(1,i)}}{g^{S_0 \oplus S_1}_{(1,i)}} \\
&\iff \nu \geq \max_{i \in S_1 \subseteq N} \frac{f^{S_1}_i}{g^{S_1}_i} = \nu^{\mathrm{AT}}_{(1,i)},
\end{aligned}$$

where we have used Proposition 3.3 along with Lemmas 4.2(b) and 4.6(b).

Now, we show that $\nu^*_{(0,i)} = \nu^{\mathrm{AT}}_{(0,i)}$, through the equivalences

$$\begin{aligned}
\nu \geq \nu^*_{(0,i)} &\iff \text{it is optimal in (9) to rest the bandit at } (0,i) \\
&\iff 0 \geq \max_{S_0 \subseteq S_1 \subseteq N:\ i \in S_0} f^{S_0 \oplus S_1}_{(0,i)} - \nu g^{S_0 \oplus S_1}_{(0,i)} \\
&\iff \nu \geq \max_{S_0 \subseteq S_1 \subseteq N:\ i \in S_0} \frac{f^{S_0 \oplus S_1}_{(0,i)}}{g^{S_0 \oplus S_1}_{(0,i)}} \\
&\iff \nu \geq \max_{S_1 \subseteq N:\ i \in S_1} \frac{f^{S_1}_i - c_i}{g^{S_1}_i} = \nu^{\mathrm{AT}}_{(0,i)},
\end{aligned}$$

where we have used Proposition 3.3, and Lemmas 4.2(c) and 4.6(c). This completes the proof. □

5. Two-Stage Index Computation

In this section we further simplify Version 3 of the index algorithm, by *decoupling* computation of the continuation and the switching index into a two-stage scheme.

5.1. First Stage: Computing the Continuation Index

We start with continuation index $\nu^*_{(1,i)}$, which is the Gittins index ν^*_i of the bandit. We will need further quantities as input for the second-stage algorithm to be discussed later.

Table 4: Gittins-Index Algorithmic Scheme AG^1.

ALGORITHM AG^1:
Output: $\{i_1^{k_1}\}_{k_1=1}^n$, $\{\nu^*_j : j \in N\}$, $\{(w_j^{(k_1)}, \nu_j^{(k_1)}) : j \in S_1^{k_1}\}_{k_1=1}^n$

set $S_1^0 := \emptyset$; **compute** $\{(w_i^{(0)}, \nu_i^{(0)}) : i \in N\}$
for $k_1 := 1$ **to** n **do**
  pick $i_1^{k_1} \in \arg\max \{\nu_i^{(k_1-1)} : i \in N \setminus S_1^{k_1-1}\}$
  $\nu^*_{i_1^{k_1}} := \nu^{(k_1-1)}_{i_1^{k_1}}$; $S_1^{k_1} := S_1^{k_1-1} \cup \{i_1^{k_1}\}$
  compute $\{(w_i^{(k_1)}, \nu_i^{(k_1)}) : i \in N\}$
end

To compute such an index and extra quantities, we refer to the algorithmic scheme AG^1 in Table 4. This is a variant of the algorithm of Varaiya et al. (1985), reformulated as in Niño-Mora (2006b). For actual implementations, one can use several algorithms in the latter paper, such as the *Fast-Pivoting* algorithm with extended output FP(1), performing $(4/3)n^3 + O(n^2)$ arithmetic operations; or the *Complete-Pivoting* (CP) algorithm, performing $2n^3 + O(n^2)$ operations.

5.2. Second Stage: Computing the Switching Index

We next address computation of the switching index, *after* having computed the Gittins index and required extra quantities. Consider the algorithm $\mathrm{AG}^0_{\mathrm{TD}}$ in Table 5, which is fed as input the output of AG^1, and produces a sequence of states $i_0^{k_0}$ spanning N, along with corresponding index values $\nu^*_{(0,i_0^{k_0})}$, computed in a *top down* fashion, i.e., from highest to lowest.

The following is the main result of this paper.

Theorem 5.1 *Algorithm* $\mathrm{AG}^0_{\mathrm{TD}}$ *computes the switching index* $\nu^*_{(0,i)}$.

Proof. The result follows by noticing that algorithm AG^0 is obtained from Version 3 of index algorithm AG$_{\widehat{\mathscr{F}}}$ in Table 3 by decoupling the computation of the $v^*_{(0,i)}$'s and the v^*_i's. □

Table 5: Switching-Index Algorithm AG^0.

ALGORITHM AG^0:
Input: $\{i_1^{k_1}\}_{k_1=1}^n$, $\{v_j^*:\ j\in N\}$, $\{(w_j^{(k_1)}, v_j^{(k_1)}):\ j\in S_1^{k_1}\}_{k_1=1}^n$
Output: $\{i_0^{k_0}\}_{k_0=1}^n$, $\{v^*_{(0,j)}:\ j\in N\}$

$\hat{c}_j := (1-\beta)c_j$, $j\in N$; $S_0^0 := \emptyset$; $S_1^0 := \emptyset$; $k_0 := 0$
for $k_1 := 1$ **to** n **do**
  $S_1^{k_1} := S_1^{k_1-1} \cup \{i_1^{k_1}\}$; AUGMENT$_1$:= `false`
  $v_{(0,j)}^{(0,k_1)} := v_j^{(k_1)} - \hat{c}_j / w_j^{(k_1)}$, $j\in S_1^{k_1}\setminus S_0^{k_0}$
  while $k_0 < k_1$ **and not**(AUGMENT$_1$) **do**
    pick $j_0^{\max} \in \arg\max\{v_{(0,j)}^{(0,k_1)}:\ j\in S_1^{k_1}\setminus S_0^{k_0}\}$
    if $k_1 = n$ **or** $v^*_{i_1^{k_1}} < v^{(0,k_1)}_{(0,j_0^{\max})}$
      $i_0^{k_0+1} := j_0^{\max}$; $v^*_{(0,i_0^{k_0+1})} := v^{(0,k_1)}_{(0,i_0^{k_0+1})}$
      $S_0^{k_0+1} := S_0^{k_0} \cup \{i_0^{k_0+1}\}$; $k_0 := k_0+1$
    else
      AUGMENT$_1$:= `true`
    end { if }
  end { while }
end { for }

We next assess the arithmetic operation count of the switching index algorithm.

Proposition 5.2 *Algorithm* AG^0 *performs at most* $n^2 + O(n)$ *operations.*

Proof. The operation count is dominated by the statements

$$v_{(0,j)}^{(0,k_1)} := v_j^{(k_1)} - \hat{c}_j / w_j^{(k_1)},\ j\in S_1^{k_1}\setminus S_0^{k_0},$$

where at most $2k_1$ arithmetic operations are performed. Adding up over k_1 yields the stated maximum total count. □

6. Dependence of the Index on Switching Costs

We discuss next some insightful properties on the index's dependence on switching cost, focusing on the case $c_i \equiv c$ and $d_i \equiv d$. We will make explicit such costs in the notation, writing the continuation index as $\nu^*_{(1,i)}(d)$ — as it does not depend on c — and the switching index as $\nu^*_{(1,i)}(c,d)$. We further denote by ν^*_i the Gittins index of the underlying bandit with no switching costs.

Proposition 6.1

(a) $\nu^*_{(1,i)}(d) = \nu^*_i + (1-\beta)d$.

(b) *For large enough* $c+d$, $\nu^*_{(0,i)} = \nu^N_i - (1-\beta)c$.

(c) $\nu^*_{(0,i)}(c,d)$ *is piecewise linear convex in* (c,d), *decreasing in* c *and nonincreasing in* d.

Proof. (a) This part follows from the fact that $\nu^*_{(1,i)}(d)$ is the Gittins index of a bandit with modified rewards $\widetilde{R}_j = R_j + (1-\beta)d$ (cf. Section 2.2). The effect of such an addition of a constant term to rewards is to increment the Gittins index by the same constant, which yields the result.

(b) The second identity in (28) implies that, for $c+d$ large enough, term $(c+d)/g^S_i$ becomes dominant, and hence the maximum value of the given expression is attained by maximizing the denominator: g^S_i. Given the latter's interpretation, its maximum value is achieved by $S = N$, for which $g^N_i = 1/(1-\beta)$. Since $\nu^N_i = r^N_i / w^N_i = f^N_i / g^N_i$, this yields the result.

(c) Using the transformation in Section 2.2 along with the index representation in (27) it is readily verified that the latter yields the expression

$$\nu^*_{(0,i)}(c,d) = \max_{i \in S \subseteq N} \frac{f^S_i - c - \{1-(1-\beta)g^S_i\}d}{g^S_i} = (1-\beta)d + \max_{i \in S \subseteq N} \frac{f^S_i - (c+d)}{g^S_i}, \qquad (28)$$

where f^S_i is the reward measure of the underlying nonrestless bandit with rewards R_j — note that the corresponding reward measure with modified rewards $\widetilde{R}_j$ as above is $f^S_i(d) = f^S_i + (1-\beta)dg^S_i$. Now, the first identity in (28) represents $\nu^*_{(0,i)}(c,d)$ as the maximum of linear functions in (c,d) that are decreasing in c and nonincreasing in d, which implies the result. □

Note that Proposition 6.1(a) shows that the incentive to stay on an active bandit increases linearly with its shutdown cost, but decreases as the discount factor approaches unity.

We next give two examples to illustrate the above results. The first concerns the 3-state bandit instance with startup cost c, no shutdown cost, $\beta = 0.95$,

$$\mathbf{R} = \begin{bmatrix} 0.7221 \\ 0.9685 \\ 0.1557 \end{bmatrix} \quad \text{and} \quad \mathbf{P} = \begin{bmatrix} 0.8061 & 0.1574 & 0.0365 \\ 0.1957 & 0.0067 & 0.7976 \\ 0.1378 & 0.5959 & 0.2663 \end{bmatrix}.$$

Figure 1 plots the bandit's switching index for each state vs. the startup cost. Notice that the plot is consistent with Proposition 6.1(b, c). It further illustrates that the relative state ordering induced by the switching index can change as the startup cost varies.

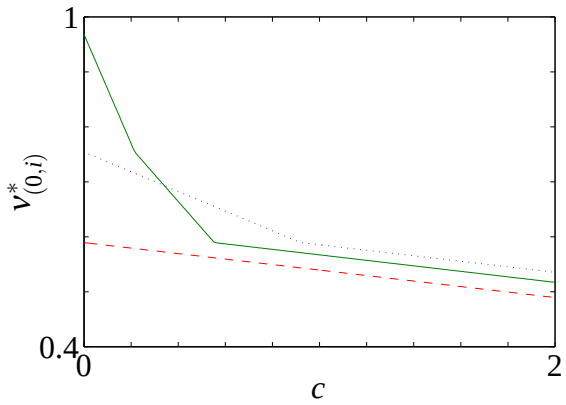

Figure 1: Dependence of Switching Index on Startup Cost.

The next example concerns the same base instance but now with shutdown cost d and no startup cost. Figure 2 plots the continuation and switching index for each state vs. the shutdown cost. The plots are consistent with Proposition 6.1. Further, the plot for the switching index shows that the relative state ordering induced by it can change as the shutdown cost varies.

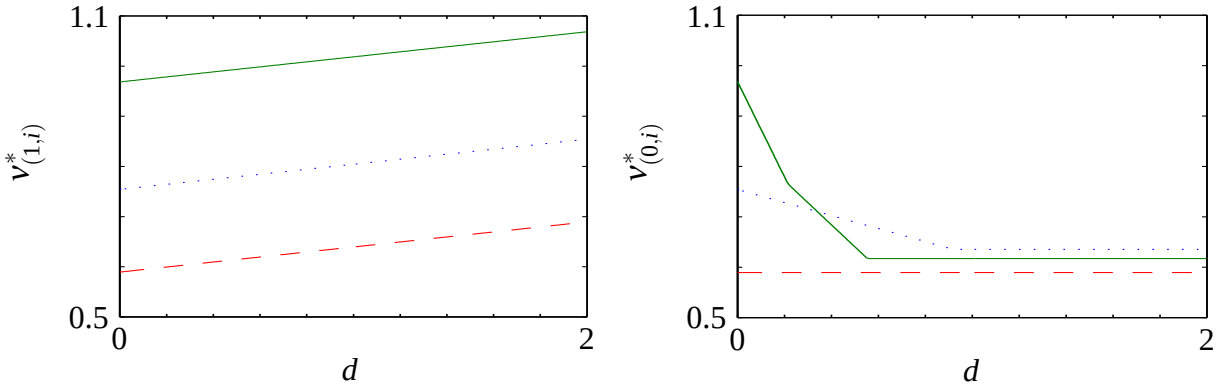

Figure 2: Dependence of Continuation and Switching Indices on Shutdown Cost.

7. Computational Experiments

This section reports the results of a computational study, based on the author's MATLAB implementations of the algorithms described herein.

The first experiment investigated the runtime performance of the decoupled index computation method. We made MATLAB generate a random bandit instance with startup costs for each of the

state-space sizes $n = 500, 1000, \ldots, 5000$. For each n, MATLAB recorded the time to compute the continuation index and required extra quantities with algorithm FP(1) in Niño-Mora (2006b), the time to compute the switching index by algorithms AG^0 and $\mathrm{AG}^0_{\mathrm{BU}}$, and the time to jointly compute both indices as in Asawa and Teneketzis (1996), using the Gittins-index algorithm FP(0) in Niño-Mora (2006b). This experiment was run on MATLAB R2006b under Windows XP x64, in an HP xw9300 254 (2.8 GHz) Opteron workstation with 4GB of memory.

The results are displayed in Figure 3. The left pane shows total runtimes, in hours, for computing both indices vs. n, along with curves obtained by cubic least-squares (LS) fit, which are consistent with the theoretical $O(n^3)$ complexity. The dotted line corresponds to the Asawa and Teneketzis (AT) scheme, while the solid line corresponds to the two-stage method herein. The results show that the two-stage method consistently achieved about a 4-fold speedup over the AT method.

The right pane shows runtimes, in *seconds*, for the switching-index algorithm vs. n, along with a curve obtained by quadratic least-squares fit, which is consistent with the theoretical $O(n^2)$ complexity. The change of timescale from hours to seconds demonstrates the order-of-magnitude runtime improvement achieved.

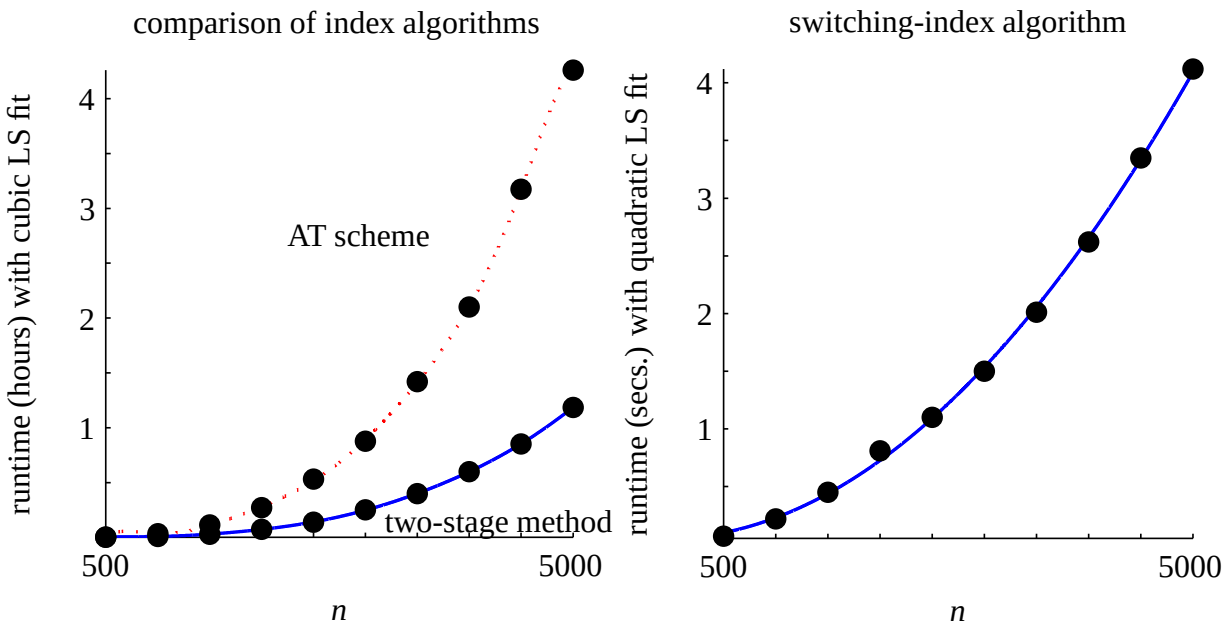

Figure 3: Exp. 1(a):Runtimes of Index Algorithms.

The following experiments were designed to assess the average relative performance of the MPI policy in random samples of two- and three-bandit instances, both against the optimal policy,

and against the benchmark Gittins index policy. For each instance, the optimal performance was computed by solving the LP formulation of the Bellman equations using the CPLEX LP solver, interfaced with MATLAB via TOMLAB. The MPI and benchmark policies were evaluated by solving with MATLAB the corresponding linear evaluation equations.

The second experiment assessed how the relative performance of the MPI policy on two-bandit instances depends on a common constant startup cost and discount factor — shutdown costs are zero. A sample of 100 instances (with 10-state bandits) was randomly generated with MATLAB. In each instance, parameter values for each bandit were independently generated: transition probabilities (obtained by scaling a matrix with Uniform$[0,1]$ entries, dividing each row by its sum) and active rewards (Uniform$[0,1]$). Passive rewards were set to zero. For each instance $k = 1,\ldots,100$ and startup cost-discount factor combination in the range $(c,\beta) \in [0,1] \times [0.2,0.9]$ — using a 0.1 grid — the optimal objective value $\vartheta^{(k),\mathrm{opt}}$ and the objective values of the MPI ($\vartheta^{(k),\mathrm{MPI}}$) and the benchmark ($\vartheta^{(k),\mathrm{bench}}$) policies were computed, along with the corresponding relative suboptimality gap of the MPI policy $\Delta^{(k),\mathrm{MPI}} \triangleq 100(\vartheta^{(k),\mathrm{opt}} - \vartheta^{(k),\mathrm{MPI}})/|\vartheta^{(k),\mathrm{opt}}|$, and the suboptimality-gap ratio of the MPI over the benchmark policy $\rho^{(k),\mathrm{MPI,bench}} \triangleq 100(\vartheta^{(k),\mathrm{MPI}} - \vartheta^{(k),\mathrm{opt}})/(\vartheta^{(k),\mathrm{bench}} - \vartheta^{(k),\mathrm{opt}})$ — scaled as percentages. The latter were then averaged over the 100 instances for each (c,β) pair, to obtain the average values Δ^{MPI} and $\rho^{\mathrm{MPI,bench}}$.

Ojective values $\vartheta^{(k),\mathrm{opt}}$, $\vartheta^{(k),\mathrm{MPI}}$ and $\vartheta^{(k),\mathrm{bench}}$ were evaluated as follows. First, the corresponding *value functions* $\vartheta^{(k),\mathrm{opt}}_{((a_1^-,i_1),(a_2^-,i_2))}$, $\vartheta^{(k),\mathrm{MPI}}_{((a_1^-,i_1),(a_2^-,i_2))}$ and $\vartheta^{(k),\mathrm{bench}}_{((a_1^-,i_1),(a_2^-,i_2))}$ were computed as mentioned above. Then, the objective values were evaluated as

$$\vartheta^{(k),\pi} \triangleq \frac{1}{n^2} \sum_{i_1,i_2 \in N} \vartheta^{(k),\pi}_{((0,i_1),(0,i_2))}, \quad \pi \in \{\mathrm{opt},\mathrm{MPI},\mathrm{bench}\}, \tag{29}$$

where each bandit has state space $N = \{1,\ldots,n\}$, with $n = 10$. Notice that (29) corresponds to assuming that both bandits are initially passive.

Figure 4 plots Δ^{MPI} vs. the startup cost c for multiple discount factors β, using cubic interpolation for smoothing. Such a gap starts at 0 for $c = 0$ (as the optimal policy is then recovered), then increases up to a maximum value, which is less than 0.25%, at about $c \approx 0.3$, and then decreases, hitting again a value of 0 at about $c \approx 0.9$ and staying there for larger values of c. Such a pattern is consistent with intuition: for large enough c, both the optimal and the MPI policies will initially pick a bandit and stay on it thereafter. Since the best bandit can be determined through single-bandit evaluations, the MPI policy will identify it. Notice also that Δ^{MPI} increases with β.

Figure 5 shows corresponding plots for the suboptimality-gap ratio $\rho^{\mathrm{MPI,bench}}$ of the MPI over the benchmark policy. They show that the average suboptimality gap for the MPI policy is in each

case less than 40% of that for the benchmark policy. Such a ratio takes the value 0 for $c = 0$ and for c large enough, as the MPI policy is then optimal. Finally, the ratio increases with β.

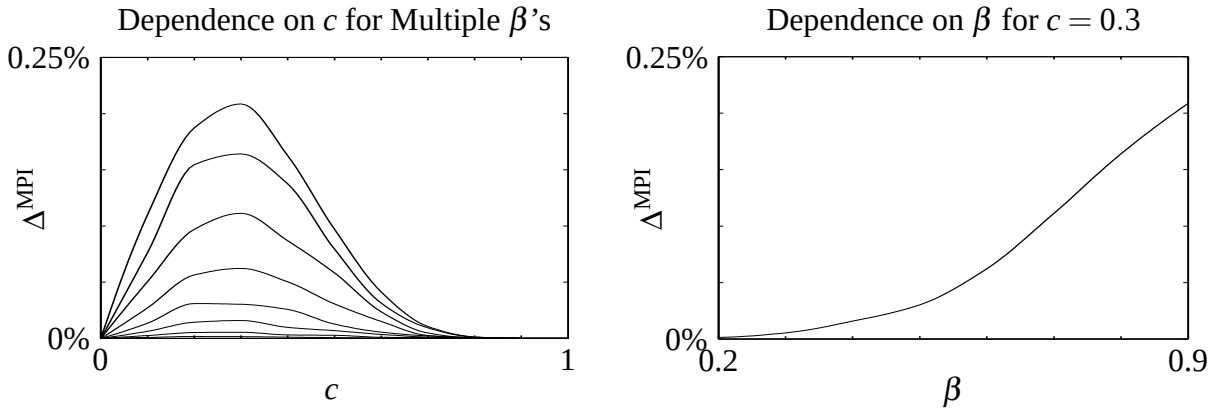

Figure 4: Exp. 2: Average Relative Suboptimality Gap of MPI Policy.

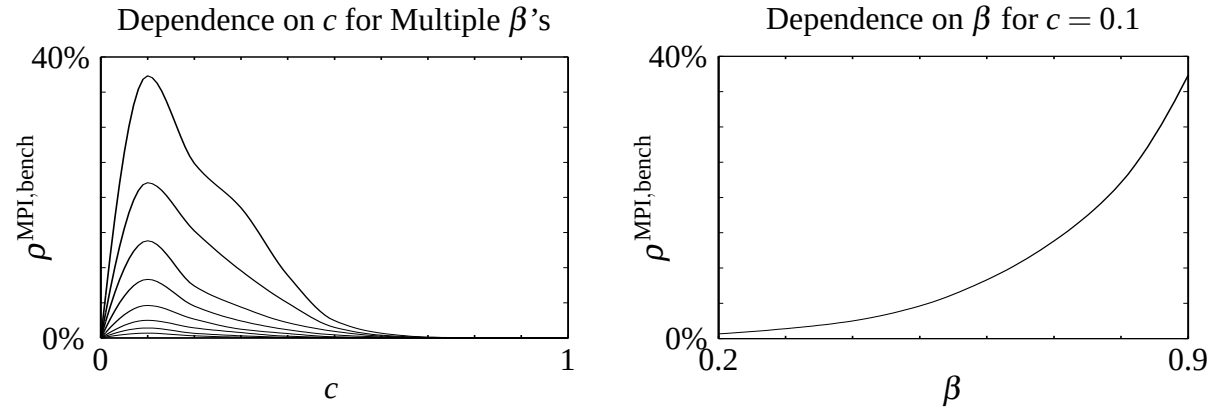

Figure 5: Exp. 2: Average Suboptimality-Gap Ratio of MPI over Benchmark Policy.

We also carried out the counterpart of experiment 2 for a common shutdown cost d. Since the patterns obtained are very close to those above, we do not present them here.

The third experiment investigated the effect of asymmetric constant startup costs, as these vary over the range $(c_1, c_2) \in [0,1]^2$, in two-bandit instances with no shutdown costs and $\beta = 0.9$. The left contour plot in Figure 6 shows that the average relative suboptimality gap of the MPI policy, Δ^{MPI}, reaches a maximum value of about 0.2% for $(c_1, c_2) \approx (0.3, 0.3)$. It further vanishes as both startup costs approach zero, and as either grows large enough. The right contour plot in the Figure shows that the average suboptimality-gap ratio ρ^{MPI} reaches maximum values of about 35%, and vanishes as either startup cost grows large. Figure 7 zooms the latter plot over the range $(c_1, c_2) \in [0, 0.3]^2$, showing that ρ^{MPI} also vanishes as both startup costs approach zero.

The fourth experiment evaluated the effect of state-dependent startup costs in two-bandit instances with no shutdown costs, as the discount factor varies. Uniform[0, 1] i.i.d. state-dependent

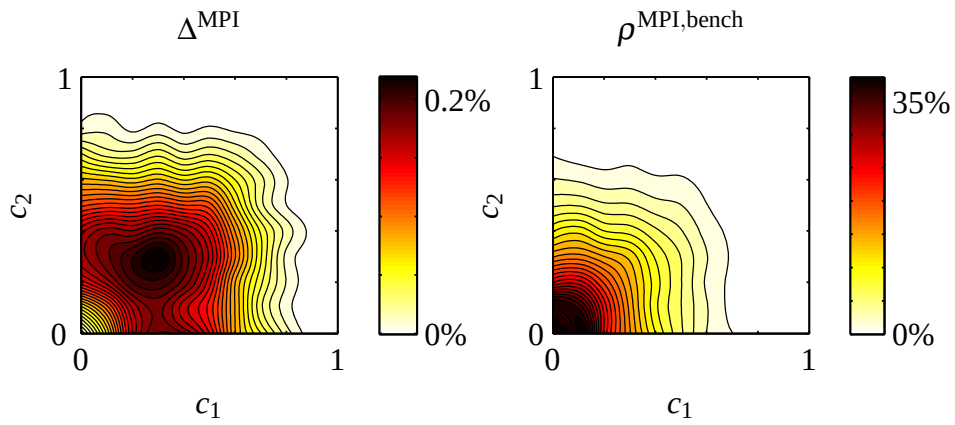

Figure 6: Exp. 3: Average Relative Performance of MPI Policy vs. (c_1, c_2), for $\beta = 0.9$.

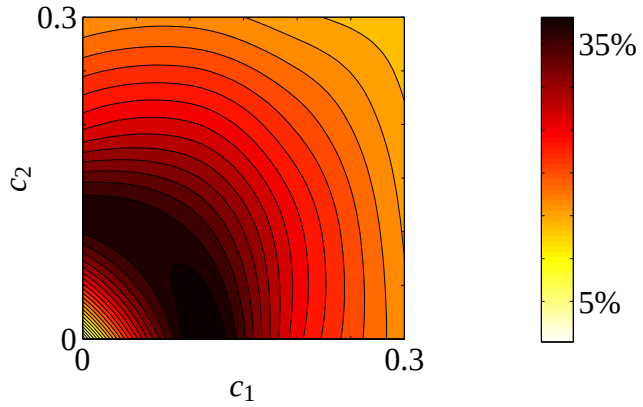

Figure 7: Exp. 3: Zoom of the Right Plot in Figure 6.

startup costs were randomly generated for each instance. Figure 8 plots the average relative suboptimality gap vs. the discount factor, which shows that such a gap tends to increase, leveling off at values near one, while remaining well below 0.5%. The figure shows that both Δ^{MPI} and $\rho^{\text{MPI,bench}}$ increase with β, with the former remaining below 0.14%, and the latter below 4%.

The fifth and last experiment evaluated the relative performance of the MPI policy on three-bandit instances as a function of a common startup cost and discount factor, based on a random sample of 100 instances of three 8-state bandits each. For each instance, the startup cost-discount factor combination was varied over the range $(c, \beta) \in [0, 1] \times [0.2, 0.9]$. The results are shown in Figures 9 and 10, which are the counterparts of experiment 2's Figure 4 and 5. Comparison of Figures 4 and 9 reveals a slight performance degradation of the MPI policy's performance in the latter, though the average gap Δ^{MPI} remains quite small, below 0.3%. Comparison of Figures 5 and 10 reveals similar values for the ratio $\rho^{\text{MPI,bench}}$.

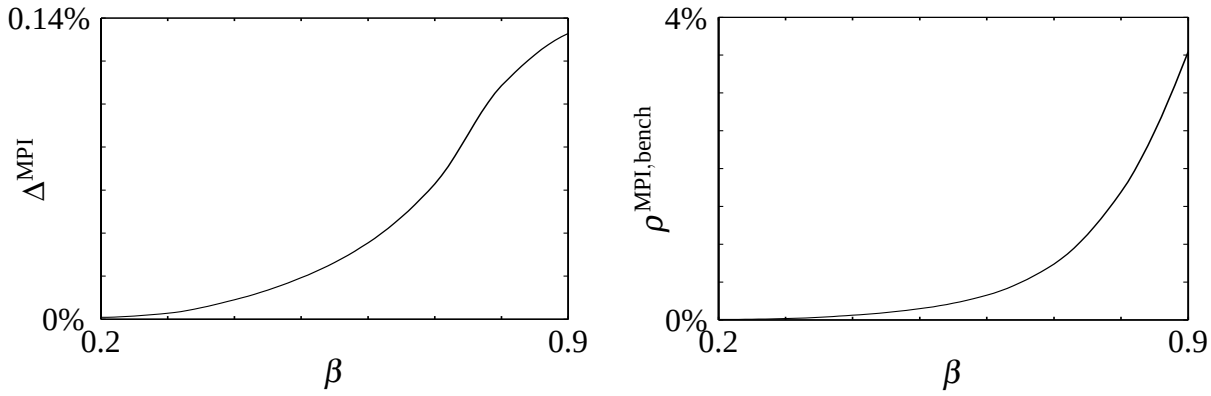

Figure 8: Exp. 4: Average Performance of MPI Policy for State-Dependent Startup Costs.

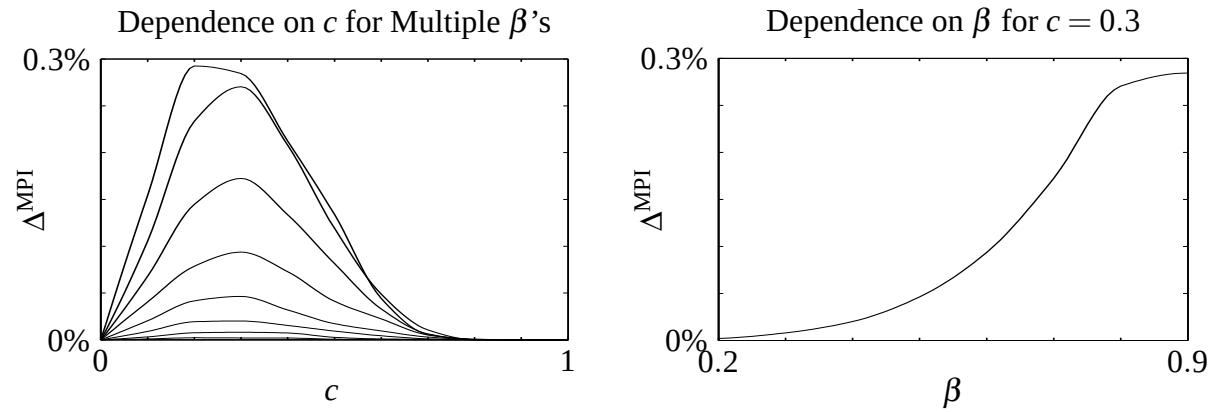

Figure 9: Exp. 5: Counterpart of Figure 4 for Three-Bandit Instances.

8. Concluding Remarks

We have addressed the important extension of the classic multi-armed bandit problem that incorporates costs for switching bandits. The paper has demonstrated the tractability and usefulness of the index policy based on the index introduced by Asawa and Teneketzis (1996). The mode of analysis has been based on deploying the powerful indexation theory for restless bandits introduced by Whittle (1988) and developed by the author in recent work. In the companion paper Niño-Mora (2007b) the approach and results herein are extended to the case where bandit switching penalties involve both costs and delays. The analyses herein extend only in part to such a case, as the restless reformulation then yields semi-Markov bandits that need not be PCL-indexable.

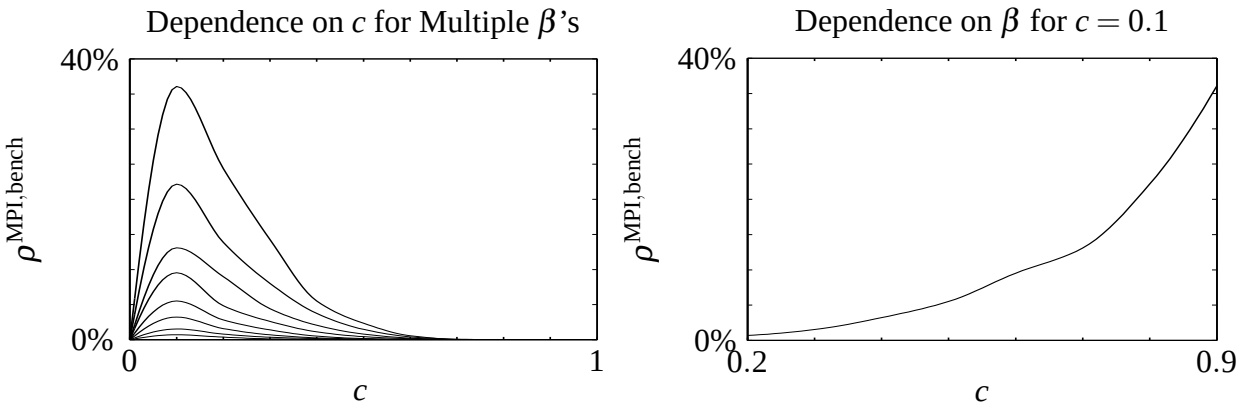

Figure 10: Exp. 5: Counterpart of Figure 5 for Three-Bandit Instances.

Acknowledgments

The author thanks the anonymous Associate Editor and two reviewers for valuable suggestions that helped improve the paper. This work, which the author started while he was at Univ. Pompeu Fabra in Barcelona, was supported in part by the Spanish Ministry of Education & Science under grants BEC2000-1027, MTM2004-02334 and a Ramón y Cajal Investigator Award, by NATO grant PST.CLG.976568, by the EU's Networks of Excellence Euro-NGI and Euro-FGI, and by the Autonomous Community of Madrid-UC3M under grants UC3M-MTM-05-075 and CCG06-UC3M/ESP-0767. Part of this work was presented at the Schloss Dagstuhl Seminar on Algorithms for Optimization with Incomplete Information (Wadern, Germany, 2005).

References

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

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

Gittins, J. C., D. M. Jones. 1974. A dynamic allocation index for the sequential design of experiments. J. Gani, K. Sarkadi, I. Vincze, eds., *Progress in Statistics (European Meeting of Statisticians, Budapest, 1972)*. North-Holland, Amsterdam, The Netherlands. 241–266.

Jun, T. 2004. A survey on the bandit problem with switching costs. *De Economist* **152** 513–541.

Niño-Mora, J. 2001. Restless bandits, partial conservation laws and indexability. *Adv. in Appl. Probab.* **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.

Niño-Mora, J. 2006a. Restless bandit marginal productivity indices, diminishing returns and optimal control of make-to-order/make-to-stock $M/G/1$ queues. *Math. Oper. Res.* **31** 50–84.

Niño-Mora, J. 2006b. A $(2/3)n^3$ fast-pivoting algorithm for the Gittins index and optimal stopping of a Markov chain. *INFORMS J. Comput.* In press.

Niño-Mora, J. 2007a. Characterization and computation of restless bandit marginal productivity indices. Working Paper 07-43, Statistics and Econometrics Series 11, Univ. Carlos III de Madrid, Spain. Submitted.

Niño-Mora, J. 2007b. Two-stage index computation for bandits with switching penalties II: switching delays. Working Paper 07-42, Statistics and Econometrics Series 10, Univ. Carlos III de Madrid, Spain. Submitted.

Varaiya, P. P., J. C. Walrand, C. Buyukkoc. 1985. Extensions of the multiarmed bandit problem: the discounted case. *IEEE Trans. Automat. Control* **30** 426–439.

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