UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 07-42
Statistics and Econometrics Series 10
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 II: Switching Delays*

José Niño-Mora [1]

Abstract

This paper addresses the multi-armed bandit problem with switching penalties including both costs and delays, extending results of the companion paper [J. Niño-Mora. ``Two-Stage Index Computation for Bandits with Switching Penalties I: Switching Costs." Conditionally accepted at INFORMS J. Comp.], which addressed the no switching delays case. Asawa and Teneketzis (1996) introduced an index for bandits with delays that partly characterizes optimal policies, attaching to each bandit state a ``continuation index" (its Gittins index) and a ``switching index," yet gave no algorithm for it. This paper presents an 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 $(5/2)n^3 + O(n)$ arithmetic operations for an n-state bandit. The paper exploits the fact that the Asawa and Teneketzis index is the Whittle, or marginal productivity, index of a classic bandit with switching penalties in its semi-Markov restless reformulation, by deploying work-reward analysis and LP-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 a benchmark index policy across a wide instance range.

Keywords: Dynamic programming, semi-Markov, finite state; bandits; switching delays; index policy; Whittle index; hysteresis; work-reward analysis; LP-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 II: Switching Delays

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

This paper addresses the multi-armed bandit problem with switching penalties including both costs and delays, extending results of the companion paper [J. Niño-Mora. "Two-Stage Index Computation for Bandits with Switching Penalties I: Switching Costs." Conditionally accepted at INFORMS J. Comp.], which addressed the no switching delays case. Asawa and Teneketzis (1996) introduced an index for bandits with delays that partly characterizes optimal policies, attaching to each bandit state a "continuation index" (its Gittins index) and a "switching index," yet gave no algorithm for it. This paper presents an 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 $(5/2)n^2 + O(n)$ arithmetic operations for an n-state bandit. The paper exploits the fact that the Asawa and Teneketzis index is the Whittle, or marginal productivity, index of a classic bandit with switching penalties in its semi-Markov restless reformulation, by deploying work-reward analysis and LP-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 a benchmark index policy across a wide instance range.

Key words: Dynamic programming, semi-Markov, finite state; bandits; switching delays; index policy; Whittle index; hysteresis; work-reward analysis; LP-indexability; analysis of algorithms
History: submitted May 8, 2007

---

1. Introduction

This paper addresses the *multi-armed bandit problem with switching penalties* (MABSP) — see, e.g., Jun (2004) for an extensive survey — which incorporates both switching costs and delays, extending results of the companion (Part I) paper Niño-Mora (2006c), which addressed the simpler case with no switching delays. While this paper also deploys the work-reward analysis approach to restless bandit indexation used in Part I, we will see that incorporation of switching delays warrants a separate treatment, as the previous analysis does not directly extend to the present case. We thus start by pointing out the key differences that, we argue, justify the present paper: (i) in Part I, classic bandits with switching costs were formulated as

Markovian restless bandits, whereas incorporation of switching delays requires a semi-Markovian formulation; (ii) the analysis in Part I held under the assumption that the sum of startup and shutdown costs be nonnegative, whereas here the extra assumption is required that bandit rewards be nonnegative — as pointed out in (Asawa and Teneketzis, 1996, Sec. IV.C); (iii) in Part I, a fast index algorithm was given that substanstially improved upon that proposed in Asawa and Teneketzis (1996) for the case of switching costs only, yet no index algorithm is given in their paper for the case of switching delays; (iv) the complexity of the switching-index algorithm in Part I is of at most $n^2 + O(n)$ arithmetic operations for an n-state bandit, whereas incorporation of switching delays requires an algorithm with increased complexity of (at most) $(5/2)n^2 + O(n)$ operations; and (v) more importantly, in Part I, the indexability analysis was based on establishing that the restless bandits of concern satisfied the PCL-indexability conditions we had introduced in earlier work; yet, the analyses below reveal that incorporation of switching delays yields restless bandits that need not be PCL-indexable, and hence require a different approach to establish their indexability; in this paper we successfully deploy for such a purpose the more powerful *LP-indexability* conditions recently introduced in Niño-Mora (2007).

The present paper follows the form and structure of its Part I counterpart as closely as possible, even using vertatim sentences from it when a variation would add nothing of substance, with the intent that a reader of both papers can more easily appreciate their similarities and differences.

To extend the initial example given in Part I, 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 must incur an upfront lump-sum *startup cost*, as well as a *startup delay*, 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*, as well as a *shutdown delay*. It can then 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.

In this and many other applications switching delays play a fundamental role, and should thus be incorporated into corresponding system models. Thus, startup delays may represent, e.g., time to lay up the groundwork or to build up infrastructure, as well as training or learning time for workers. Similarly, shutdown delays may arise, e.g., when dismantling installed infrastructure.

The problem is cast as a *semi-Markov decision process* (SMDP) by modeling projects as *bandits*, i.e., binary-action (active/passive) SMDPs that can only change state while active. In the no switching penalties case, one thus obtains the *multi-armed bandit problem* (MABP), which is optimally solved by the *Gittins index* policy. See Gittins (1979).

The optimal index solution for the MABP prompted investigation of *priority-index policies* for the MABPSP. 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 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." To be consistent which such a *hysteretic* property, the indices must satisfy

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

Though Banks and Sundaram (1994) proved that such policies are not generally optimal in the presence of switching costs, Asawa and Teneketzis (1996) introduced an intuitively appealing index for the MABSP, which we will refer to henceforth as the *AT index*, both for the case of only switching costs and for that of only switching delays, and showed that it partly characterizes optimal policies. Their continuation index is the bandit's 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 — including the initial delay.

In Asawa and Teneketzis (1996), an index computation method is presented to jointly compute both indices in the case of only switching costs. Yet, no algorithm is given in there to compute the index under switching delays. This raises the need to develop an efficient index computation method for bandits with switching delays, which is the prime goal of this paper, while the second goal is to investigate empirically the performance of the resulting AT index policy.

We will address such goals in the setting of an extended model that allows state-dependent startup and shutdown costs and delays for each bandit, which we will reduce to the case of no shutdown penalties, through a seemingly indirect route: by exploiting the natural reformulation of a classic bandit with switching penalties as a *semi-Markov restless bandit* — one that can change state while passive — *without* switching penalties, through which the MABSP is cast as a *semi-Markov multi-armed restless bandit problem* (SMARBP).

Such a reformulation will allow us to deploy the powerful indexation theory available 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, 2006b, 2007), where the *Whittle index* and extensions are shown to measure trade-off (reward vs. work) rates, whence our terming it *marginal productivity index* (MPI).

Of most relevance to this paper is Niño-Mora (2007), where the tractable class of *LP-indexable* bandits — as they are based on *linear programming* (LP) analyses — is introduced, for which the MPI is efficiently

computed by an *adaptive-greedy algorithm*. The scope of such an algorithm is thus extended from the class of *PCL-indexable* bandits in the author's earlier work to the larger class of LP-indexable bandits. Such an extension will play a crucial role in this paper, as the restless bandits of concern will be shown to be LP-indexable, yet are not necessarily PCL-indexable.

We deploy here such a theory, by proving and exploiting the fact that the AT index of a bandit with switching costs and delays is precisely the bandit's Whittle index/MPI in its semi-Markov restless reformulation. We will establish that such restless bandits are LP-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 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 introduced in Niño-Mora (2006a). For the second stage, we will present here a fast switching-index algorithm that performs *at most* $(5/2)n^2 + O(n)$ arithmetic operations, thus achieving an order of magnitude improvement 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 complexity translates into dramatic runtime savings. Such a 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 penalties case, defines the AT index, and gives the SMARBP reformulation. Section 3 reviews the indexation theory to be deployed. Section 4 carries out a work-reward analysis of reformulated restless bandits. Section 5 draws on such an analysis to develop the new decoupled index algorithm. Section 6 discusses dependence of the index on switching penalties. Section 7 reports the computational study's results. Section 8 concludes.

2. Model, AT index and Restless-Bandit Reformulation

2.1. The MABPSP

Consider a collection of M finite-state bandits, one of which must be engaged (*active*) at each discrete *decision period* $\tau_k \in \mathbb{Z}_+$, with $0 \leq \tau_k \nearrow \infty$ as $k \to \infty$, while the others are rested (*passive*). Switching bandits is costly, involving startup and shutdown costs and delays. We assume that a freshly set up bandit must be *worked on* for at least one period, and will say that a bandit is *engaged* if it is either being worked on, or is

undergoing a startup or a shutdown delay.

A rested bandit m occupying state i_m — belonging in its state space N_m — accrues no rewards, i.e., $R_m^0(i_m) \equiv 0$, and its state remains frozen. When freshly engaged, it incurs startup cost $c_m(i_m)$, followed by a discrete random startup delay $\xi_m(i_m) \in \mathbb{Z}_+$ having *z-transform* $\phi_m(z;i_m) \triangleq \mathbb{E}\left[z^{\xi_m(i_m)}\right]$, during which no rewards accrue. When the startup is completed, the bandit must be worked on, yielding an active reward $R_m^1(i_m) = R_m(i_m)$ and changing state at the following period to j_m with probability $p_m(i_m, j_m)$. After one or more periods at which the bandit is worked on, it may be rested. If this happens in state j_m, shutdown cost $d_m(j_m)$ is incurred, followed by a random shutdown delay $\eta_m \in \mathbb{Z}_+$ having z-transform $\psi_m(z) \triangleq \mathbb{E}\left[z^{\eta_m}\right]$, during which no rewards accrue. Then, the bandit must be rested for at least one period. Note that we allow startup delay distributions to be state-dependent, while shutdown delay's are constant — due to results in Section 2.2. Rewards and costs are time-discounted with factor $0 < \beta < 1$. We will find it convenient to write $\phi_m(\beta; i_m)$ and $\psi_m(\beta)$ as $\phi_m(i_m)$ and ψ_m.

Note that such a model can readily accomodate the case where switching costs are instead incurred at rates $C_m(i_m)$ and $D_m(i_m)$ per period during the startup and shutdown delays, respectively. Clearly, one should then use the equivalent lump-sum switching costs

$$c_m(i_m) \triangleq \frac{1-\phi_m(i_m)}{1-\beta} C_m(i_m) \quad \text{and} \quad d_m(i_m) \triangleq \frac{1-\psi_m}{1-\beta} D_m(i_m).$$

Actions are chosen by adoption of 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. Focus on such a version, instead of on that where *at most* one bandit can be engaged, is without loss of generality. The MABPSP is to find an admissible policy that maximizes the expected total discounted value of rewards earned minus switching costs incurred.

We will denote by $X_m(t)$ and $a_m(t) \in \{0,1\}$ the prevailing state and action for bandit m at period t, respectively, where $a_m(t) = 1$ (resp. $a_m(t) = 0$) means that the bandit is engaged (resp. rested). 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 $\widehat{\mathbf{X}}(t) \triangleq \left(\widehat{X}_m(t)\right)_{m=1}^M$, and the *joint action process* is $\mathbf{a}(t) \triangleq \left(a_m(t)\right)_{m=1}^M$.

2.2. Reduction to the Normalized No Shutdown Penalties Case

We show in this section that it suffices to restrict attention to the no shutdown penalties case, without loss of generality. 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. Let us drop the bandit label m, and denote by $\mathbf{R} = (R_j)$,

$\mathbf{c} = (c_j)$ and $\mathbf{d} = (d_j)$ the bandit's state-dependent active reward, startup and shutdown cost vectors. Let us further denote by $\phi = (\phi_j)$ the bandit's state-dependent startup z-transform vector, evaluated at $z = \beta$, and let ψ denote the corresponding constant shutdown z-transform value. We can thus write the expected discounted net reward earned on the bandit during such a time span, starting at $X(0) = i$, as

$$f_i^\tau(\mathbf{R},\mathbf{c},\mathbf{d},\phi,\psi) \triangleq \mathbb{E}_i^\tau\left[-c_i + \beta^{\xi_i}\sum_{t=0}^{\tau-1} R_{X(t)}\beta^t - d_{X(\tau)}\beta^{\xi_i+\tau}\right], \tag{2}$$

where ξ_i is the random startup delay starting at i. The corresponding discounted amount of *work* expended on the bandit is

$$g_i^\tau(\phi,\psi) \triangleq \mathbb{E}_i^\tau\left[\frac{1-\beta^{\xi_i}}{1-\beta} + \beta^{\xi_i}\sum_{t=0}^{\tau-1}\beta^t + \frac{1-\beta^{\eta}}{1-\beta}\beta^{\xi_i+\tau}\right], \tag{3}$$

where, as mentioned above, both the startup and shutdown delays ξ_i and η are counted as "work."

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

(a) $f_i^\tau(\mathbf{R},\mathbf{c},\mathbf{d},\phi,\psi) = f_i^\tau\Big(\frac{1}{\psi}\{\mathbf{R}+(\mathbf{I}-\beta\mathbf{P})\mathbf{d}\},(c_j+\phi_j d_j)_{j\in N},\mathbf{0},\psi\phi,1\Big)$.

(b) $g_i^\tau(\phi,\psi) = g_i^\tau(\psi\phi,1)$.

Proof. (a) 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

$$\begin{aligned}
f_i^\tau(\mathbf{R},\mathbf{c},\mathbf{d},\phi,\psi) &\triangleq \mathbb{E}_i^\tau\left[-c_i + \beta^{\xi_i}\sum_{t=0}^{\tau-1} R_{X(t)}\beta^t - d_{X(\tau)}\beta^{\xi_i+\tau}\right] \\
&= -c_i + \phi_i\mathbb{E}_i^\tau\left[\sum_{t=0}^{\tau-1} R_{X(t)}\beta^t - d_{X(\tau)}\beta^\tau\right] \\
&= -c_i + \phi_i\left\{-d_i + \mathbb{E}_i^\tau\left[\sum_{t=0}^{\tau-1}\{R_{X(t)} + d_{X(t)} - \beta d_{X(t+1)}\}\beta^t\right]\right\} \\
&= -c_i - \phi_i d_i + \phi_i\mathbb{E}_i^\tau\left[\sum_{t=0}^{\tau-1}\{R_{X(t)} + d_{X(t)} - \beta d_{X(t+1)}\}\beta^t\right] \\
&= -c_i - \phi_i d_i + \phi_i\psi\mathbb{E}_i^\tau\left[\sum_{t=0}^{\tau-1}\frac{R_{X(t)} + d_{X(t)} - \beta d_{X(t+1)}}{\psi}\beta^t\right] \\
&= f_i^\tau\Big(\frac{1}{\psi}\{\mathbf{R}+(\mathbf{I}-\beta\mathbf{P})\mathbf{d}\},(c_j+\phi_j d_j)_{j\in N},\mathbf{0},\psi\phi,1\Big).
\end{aligned}$$

(b) This part follows by writing

$$
\begin{aligned}
g_i^\tau &\triangleq \mathbb{E}_i^\tau\left[\frac{1-\beta^{\xi_i}}{1-\beta}+\beta^{\xi_i}\sum_{t=0}^{\tau-1}\beta^t+\frac{1-\beta^{\eta}}{1-\beta}\beta^{\xi_i+\tau}\right]=\frac{1-\phi_i}{1-\beta}+\phi_i\mathbb{E}_i^\tau\left[\sum_{t=0}^{\tau-1}\beta^t+\frac{1-\psi}{1-\beta}\beta^{\tau}\right]\\
&=\frac{1-\phi_i}{1-\beta}+\phi_i\mathbb{E}_i^\tau\left[\sum_{t=0}^{\tau-1}\beta^t+\frac{1-\psi}{1-\beta}\big\{1-(1-\beta)\sum_{t=0}^{\tau-1}\beta^t\big\}\right]\\
&=\frac{1-\phi_i\psi}{1-\beta}+\phi_i\mathbb{E}_i^\tau\left[\sum_{t=0}^{\tau-1}\big\{1-(1-\psi)\big\}\beta^t\right]=\frac{1-\phi_i\psi}{1-\beta}+\phi_i\psi\mathbb{E}_i^\tau\left[\sum_{t=0}^{\tau-1}\beta^t\right]\\
&=g_i^\tau\big(\psi\phi,1\big).
\end{aligned}
$$

□

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

$$
\widetilde{c}_j\triangleq c_j+\phi_j d_j,\quad \widetilde{\phi}_j\triangleq\psi\phi_j,\quad\text{and}\quad \widetilde{\mathbf{R}}\triangleq\frac{1}{\psi}\big\{\mathbf{R}+(\mathbf{I}-\beta\mathbf{P})\mathbf{d}\big\}. \tag{4}
$$

Note that, in the case $c_j\equiv c$ and $d_j\equiv d$, one obtains $\widetilde{c}_j\equiv c+d\phi_j$ and $\widetilde{R}_j=\{R_j+(1-\beta)d\}/\psi$.

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

2.3. The AT Index

We next define the AT index for a bandit, whose label m we drop from the notation, extending the definitions in Asawa and Teneketzis (1996) to the present setting. The continuation AT index is

$$
\nu_{(1,i)}^{\mathrm{AT}}\triangleq\max_{\tau>0}\frac{\mathbb{E}_i^\tau\left[\sum_{t=0}^{\tau-1}R_{X(t)}\beta^t\right]}{\mathbb{E}_i^\tau\left[\sum_{t=0}^{\tau-1}\beta^t\right]}, \tag{5}
$$

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

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

where now τ is a stopping time/rule that engages a bandit starting at i which needs to be set up.

Notice that, writing $g_i^\tau=\mathbb{E}_i^\tau\left[\sum_{t=0}^{\tau-1}\beta^t\right]$ and $f_i^\tau=\mathbb{E}_i^\tau\left[\sum_{t=0}^{\tau-1}R_{X(t)}\beta^t\right]$, we have that

$$
\frac{f_i^\tau}{g_i^\tau}-\frac{-c_i+\phi_i f_i^\tau}{\frac{1-\phi_i}{1-\beta}+\phi_i g_i^\tau}=\frac{1}{g_i^\tau}\frac{(1-\beta)c_i g_i^\tau+(1-\phi_i)f_i^\tau}{1-\phi_i+(1-\beta)\phi_i g_i^\tau}\geq 0,
$$

provided that $c_j \geq 0$ and $R_j \geq 0$, for $j \in N$. In such a case, on which we will focus our analyses, it follows from the above that $\nu^{\mathrm{AT}}_{(1,i)} \geq \nu^{\mathrm{AT}}_{(0,i)}$, consistently with (1).

2.4. Semi-Markov Restless-Bandit Reformulation

Taking $\widehat{X}_m(t)$ as the state of each bandit m yields a reformulation of the MABPSP as a SMARBP *without* switching penalties, having joint state and action processes $\widehat{\mathbf{X}}(t)$ and $\mathbf{a}(t)$, where actions can only be taken at the sequence τ_k of decision periods discussed above. The rewards and dynamics for restless bandit m in such a reformulation are as follows. If at period τ_k the bandit occupies (augmented) state $(1, i_m)$ and is engaged, the active reward $\widehat{R}^1_m(1, i_m) \triangleq R_m(i_m)$ is earned, and the state moves at the next decision period $\tau_{k+1} = \tau_k + 1$ to $(1, j_m)$ with active transition probability $\widehat{p}^1_m\big((1, i_m), (1, j_m)\big) \triangleq p_m(i_m, j_m)$. If the bandit is instead rested, no passive reward is earned, i.e., $\widehat{R}^0_m(1, i_m) \equiv 0$, and the state moves at the next decision period $\tau_{k+1} = \tau_k + 1$ to $(0, i_m)$ with a unity passive transition probability, i.e., $\widehat{p}^0_m\big((1, i_m), (0, i_m)\big) \equiv 1$.

If the restless bandit occupies at τ_k state $(0, i_m)$ and is engaged, the expected active reward

$$\widehat{R}^1_m(0, i_m) \triangleq \mathbb{E}[-c_m(i_m) + \beta^{\xi_m(i_m)} R_m(i_m)] = -c_m(i_m) + \phi_m(i_m) R_m(i_m) \tag{7}$$

accrues up to the next decision period $\tau_{k+1} = \tau_k + \xi_m(i_m) + 1$, at which its state moves to $(1, j_m)$ with active transition probability $\widehat{p}^1_m\big((0, i_m), (1, j_m)\big) \triangleq p_m(i_m, j_m)$. If the bandit is instead rested, no passive reward accrues, i.e., $\widehat{R}^0_m(0, i_m) \equiv 0$, and the state remains frozen at the next decision period $\tau_{k+1} = \tau_k + 1$, i.e., $\widehat{p}^0_m\big((0, i_m), (0, i_m)\big) \equiv 1$.

We can thus formulate the MABPSP as the SMARBP

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

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

3. Restless Bandit Indexation: Theory and Computation

We discuss in this section the semi-Markov restless bandit indexation theory referred to in Section 1, as it applies to a single bandit m as above — in its restless reformulation. We hence drop again the bandit label m 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, where such a notation distinguishes them from their boldface counterparts used in the multi-bandit setting above. We will assume that (normalized) startup costs and active rewards are nonnegative.

Assumption 3.1 *For $i \in N$*:

(i) $c_i \geq 0$*; and*

(ii) $R_i \geq 0$.

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*

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

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

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

which gives 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. $\nu^*_{(a^-,i)} \leq \nu$). Note that Whittle (1988)'s original definition of indexability was stated in an equivalent form in terms of optimal passive sets.

To establish indexability and compute the MPI, we have developed in Niño-Mora (2001, 2002, 2006b, 2007) 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. An Illustrative Example

To help the reader unfamiliar with the above concepts to grasp them, we present next an illustrative example. Consider the 3-state normalized (no shutdown penalties) bandit instance with no startup cost, startup delay given by its z-transform value $\phi = \phi(\beta)$,

$$\beta = 0.95, \quad \mathbf{R} = \begin{bmatrix} 0.0250 \\ 0.4242 \\ 0.0338 \end{bmatrix}, \quad \text{and} \quad \mathbf{P} = \begin{bmatrix} 0.6635 & 0.0285 & 0.3080 \\ 0.6345 & 0.3583 & 0.0072 \\ 0.4868 & 0.0530 & 0.4602 \end{bmatrix}.$$

Work and reward measures g^π and f^π are evaluated assuming that the initial state is uniformly drawn. The left pane in Figure 1 shows the *achievable work-reward performance region* in the classic no startup delay ($\phi = 1$) case. The four points displayed, which determine the region's upper boundary, are the work-reward performance points corresponding to the policies having active sets, from left to right, $\emptyset$, $\{2\}$, $\{2,3\}$, and $\{2,3,1\}$. The work-reward trade-off slopes/rates between such points are the bandit's Gittins index values:

$$\nu^*_2 = 0.4242 > \nu^*_3 = 0.061487 > \nu^*_1 = 0.048002.$$

The right pane in Figure 1 shows a corresponding plot for the case with $\phi = 0.98$. The upper work-reward boundary is determined by the seven points displayed, which are the work-reward performance points corresponding, from left to right, to the policies having active sets $\emptyset \oplus \emptyset$, $\emptyset \oplus \{2\}$, $\{2\} \oplus \{2\}$, $\{2\} \oplus \{2,3\}$, $\{2,3\} \oplus \{2,3\}$, $\{2,3\} \oplus \{2,3,1\}$ and $\{2,3,1\} \oplus \{2,3,1\}$. The work-reward trade-off slopes between such points give the MPI values:

$$\nu^*_{(1,2)} = 0.424 > \nu^*_{(0,2)} = 0.334 > \nu^*_{(1,3)} = 0.061 > \nu^*_{(0,3)} = 0.051 > \nu^*_{(1,1)} = 0.048 > \nu^*_{(0,1)} = 0.047.$$

The plot represents the right end-points giving a continuation index value by a black circle, and those giving a switching index value by a white square. Note further that the continuation index matches the Gittins index of the previous case.

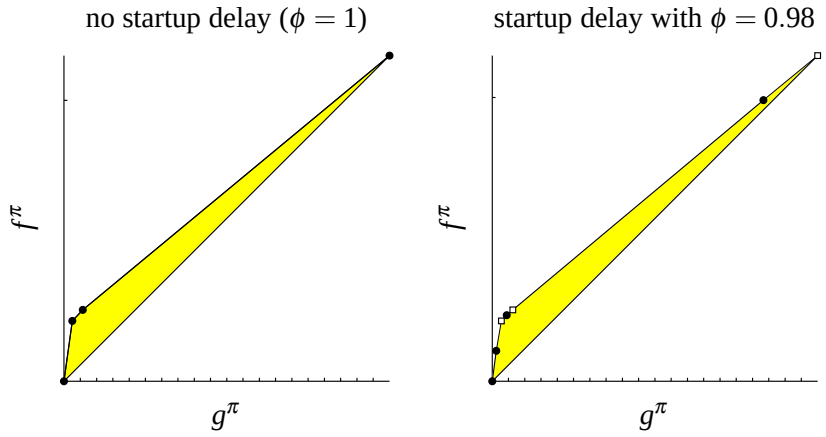

Figure 1: Achievable Work-Reward Performance Regions and Structure of Upper Boundaries.

The left pane of Figure 2 shows the achievable work-reward performance region for the case $\phi = 0.8$. Now, the seven points displayed, which characterize the upper boundary, correspond, from left to right, to the policies having active sets $\emptyset \oplus \emptyset$, $\emptyset \oplus \{2\}$, $\{2\} \oplus \{2\}$, $\{2\} \oplus \{2,3\}$, $\{2\} \oplus \{2,3,1\}$, $\{2,3\} \oplus \{2,3,1\}$ and $\{2,3,1\} \oplus \{2,3,1\}$. The work-reward trade-off slopes/rates between such points give the bandit's MPI values:

$$\nu^*_{(1,2)} = 0.424 > \nu^*_{(0,2)} = 0.099 > \nu^*_{(1,3)} = 0.061 > \nu^*_{(1,1)} = 0.048 > \nu^*_{(0,3)} = 0.039 > \nu^*_{(0,1)} = 0.038.$$

Finally, the right pane of Figure 2 shows the corresponding plot for the case $\phi = 0.5$. The seven points characterizing the region's upper boundary correspond, from left to right, to the policies having active sets $\emptyset \oplus \emptyset$, $\emptyset \oplus \{2\}$, $\emptyset \oplus \{2,3\}$, $\emptyset \oplus \{2,3,1\}$, $\{2\} \oplus \{2,3,1\}$, $\{2,3\} \oplus \{2,3,1\}$, and $\{2,3,1\} \oplus \{2,3,1\}$. The resulting MPI values given by the successive slopes are

$$\nu^*_{(1,2)} = 0.424 > \nu^*_{(1,3)} = 0.061 > \nu^*_{(1,1)} = 0.048 > \nu^*_{(0,2)} = 0.038 > \nu^*_{(0,3)} = 0.025 > \nu^*_{(0,1)} = 0.024.$$

Note that in each case the continuation index value $\nu^*_{(1,i)}$ matches the Gittins index value ν^*_i. Further, the successive active sets $S_0 \oplus S_1$ characterizing the efficient frontiers belong in the active-set family $\widehat{\mathscr{F}}$ in (10). Also, the continuation index value $\nu^*_{(1,i)}$ is larger than the corresponding switching index value $\nu^*_{(0,i)}$ value, consistently with (1).

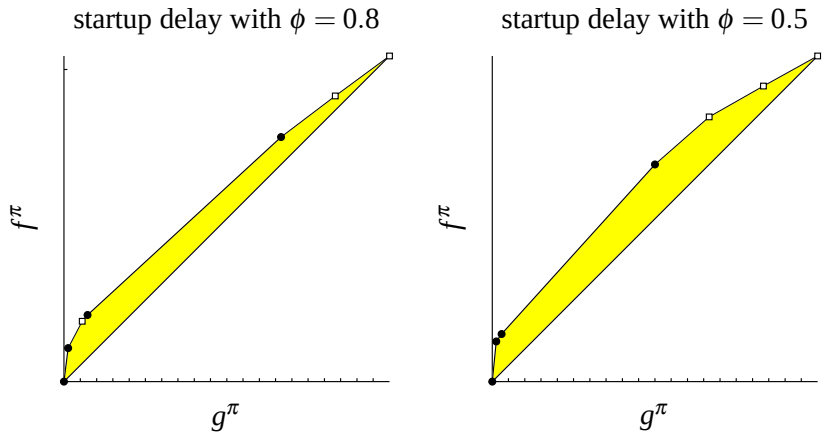

Figure 2: Achievable Work-Reward Performance Regions and Structure of Upper Boundaries.

3.3. LP-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 LP-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 initially takes action a 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, when $w^{S_0 \oplus S_1}_{(a^-,i)} \neq 0$, the *marginal productivity measure*

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

We will deploy the LP-indexability approach to indexation introduced in Niño-Mora (2007), which extends the earlier PCL-indexability approach introduced and developed in Niño-Mora (2001, 2002, 2006b).

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.

We will further write below

$$\underline{r}^{\widehat{S}} \triangleq \max_{(a^-,j) \in \widehat{S}^c, w^{\widehat{S}}_{(a^-,j)} = 0} r^{\widehat{S}}_{(a^-,j)} \quad \text{and} \quad \overline{r}^{\widehat{S}} \triangleq \min_{(a^-,j) \in \widehat{S}, w^{\widehat{S}}_{(a^-,j)} = 0} r^{\widehat{S}}_{(a^-,j)},$$

adopting the convention that the maximum (resp. minimum) over an empty set is $-\infty$ (resp. $+\infty$).

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

(i) $w^{\emptyset}_{(a^-,i)}, w^{\widehat{N}}_{(a^-,i)} \geq 0$ for $(a^-,i) \in \widehat{N}$, and $\underline{r}^{\emptyset} \leq 0 \leq \overline{r}^{\widehat{N}}$;

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

(iii) 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 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 in (Niño-Mora, 2007, Th. 5.4), which refers to a generic restless bandit and active-set family F.

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

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

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

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

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

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

Theorem 3.2 *An LP($\mathscr{F}$)-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., $v^{S_0^{k_0-1} \oplus S_1^{k_1-1}}_{(0,j)}$ as $v^{(k_0-1,k_1-1)}_{(0,j)}$. 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 $k_0 \le k_1$. 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 $1 \le k_0 < k_1 \le n$, consistently with (10).

3.4. 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):

$$\begin{aligned}
\vartheta^*_{(1,i)}(\nu) &= \max\left\{\beta\,\vartheta^*_{(0,i)}(\nu), R_i - \nu + \beta \sum_{j\in N} p_{ij}\vartheta^*_{(1,j)}(\nu)\right\} \\
\vartheta^*_{(0,i)}(\nu) &= \max\left\{\beta\,\vartheta^*_{(0,i)}(\nu), -c_i - \frac{1-\phi_i}{1-\beta}\nu + \phi_i\{R_i - \nu + \beta \sum_{j\in N} p_{ij}\vartheta^*_{(1,j)}(\nu)\}\right\}.
\end{aligned} \tag{16}$$

Proposition 3.3 *For every wage $\nu \in \mathbb{R}$ there exists an optimal active set $\widehat{S} \in \widehat{\mathscr{F}}$ for (9), 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

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

We want to show that this implies that it is optimal to rest it in state $(0,i)$, i.e.,

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

Suppose first that $\nu < 0$. In such a case, it suffices to draw on classic bandit theory, noting that once the bandit is active it is optimal to keep it active in states i for which $\nu \leq \nu^*_i$, where ν^*_i is the bandit's Gittins index. Now, Assumption 3.1(ii) ensures that $\nu^*_i \geq 0$ for every state i, and hence it will never be optimal to rest the bandit, once engaged, if $\nu < 0$.

Consider now the case $\nu \geq 0$. In such case, we have the inequalities

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

where the second inequality follows immediately by reformulating it as

$$(1-\phi_i)\big\{R_i + \beta \sum_{j\in N} p_{ij}\vartheta^*_{(1,j)}(\nu)\big\} \geq -c_i - \beta\frac{1-\phi}{1-\beta}\nu,$$

and noting that Assumption 3.1(ii) ensures that the left-hand side in the latter inequality is nonnegative, whereas Assumption 3.1(i) and $\nu \geq 0$ ensure that its right-hand side is nonpositive. □

Note that Proposition 3.3 establishes LP($\widehat{\mathscr{F}_T}$)-indexability condition (iii) above. In order to further establish the remaining conditions (i, ii) 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 LP-Indexability Proof

We set out in this section to carry out a work-reward analysis of a single bandit with startup penalties as above, in its semi-Markov restless bandit reformulation, and to establish its LP-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 semi-Markov 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

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

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

(b) $g_{(1,i)}^{S_0\oplus S_1} = g_i^{S_1}$, *for* $i \in S_1$.

(c) $g_{(0,i)}^{S_0\oplus S_1} = (1-\phi_i)/(1-\beta) + \phi_i g_i^{S_1}$, *for* $i \in S_0$.

(d) $g_{(0,i)}^{S_0\oplus S_1} = 0$, *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

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

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

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

$$g_{(0,i)}^{S_0\oplus S_1} = \frac{1-\phi_i}{1-\beta} + \phi_i g_{(1,i)}^{S_0\oplus S_1} = \frac{1-\phi_i}{1-\beta} + \phi_i g_i^{S_1},$$

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

$$\begin{aligned} w_{(1,i)}^{S_0\oplus S_1} &= 1 + \beta \sum_{j\in N} p_{ij} g_{(1,j)}^{S_0\oplus S_1} - \beta g_{(0,i)}^{S_0\oplus S_1} \\ w_{(0,i)}^{S_0\oplus S_1} &= \frac{1-\phi_i}{1-\beta} + \phi_i \big\{ 1 + \beta \sum_{j\in N} p_{ij} g_{(1,j)}^{S_0\oplus S_1} \big\} - \beta g_{(0,i)}^{S_0\oplus S_1}. \end{aligned} \tag{21}$$

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

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

(a) $w_{(1,i)}^{S_0\oplus S_1} = w_i^{S_1}$, *for* $i \in S_1^c$.

(b) $w_{(0,i)}^{S_0\oplus S_1} = \dfrac{1-\phi_i}{1-\beta} + w_i^{S_1}$, *for* $i \in S_1^c$.

(c) $w_{(1,i)}^{S_0\oplus S_1} = \dfrac{1-\beta\phi_i}{1-\beta} \big\{ w_i^{S_1} - \beta \dfrac{1-\phi_i}{1-\beta\phi_i} \big\}$, *for* $i \in S_0$.

(d) $w_{(0,i)}^{S_0\oplus S_1} = 1 - \phi_i + \phi_i w_i^{S_1}$, *for* $i \in S_0$.

(e) $w_{(1,i)}^{S_0\oplus S_1} = \dfrac{w_i^{S_1}}{1-\beta}$, *for* $i \in S_1 \setminus S_0$.

(f) $w_{(0,i)}^{S_0\oplus S_1} = \dfrac{1-\phi_i}{1-\beta} + \dfrac{\phi_i}{1-\beta} w_i^{S_1}$, *for* $i \in S_1 \setminus S_0$.

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

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

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

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

$$\begin{aligned} w_{(0,i)}^{S_0\oplus S_1} &= \frac{1-\phi_i}{1-\beta} + \phi_i \big\{ 1 + \beta \sum_{j\in N} p_{ij} g_{(1,j)}^{S_0\oplus S_1} \big\} - \beta g_{(0,j)}^{S_0\oplus S_1} \\ &= \frac{1-\phi_i}{1-\beta} + \phi_i \big\{ 1 + \beta \sum_{j\in S_1} p_{ij} g_j^{S_1} \big\} = \frac{1-\phi_i}{1-\beta} + \phi_i w_i^{S_1}, \end{aligned}$$

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

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

$$\begin{aligned} w_{(1,i)}^{S_0\oplus S_1} &= g_{(1,i)}^{S_0\oplus S_1} - \beta g_{(0,i)}^{S_0\oplus S_1} = g_i^{S_1} - \beta \big\{ \frac{1-\phi_i}{1-\beta} + \phi_i g_i^{S_1} \big\} \\ &= (1-\beta\phi_i) g_i^{S_1} - \beta \frac{1-\phi_i}{1-\beta} = \frac{1-\beta\phi_i}{1-\beta} \big\{ w_i^{S_1} - \beta \frac{1-\phi_i}{1-\beta\phi_i} \big\}, \end{aligned}$$

where we have used (21), $S_0 \subseteq S_1$, Lemma 4.1, Lemma 4.2(b, c), and (19).

(d) We have, for $i \in S_0$,

$$\begin{aligned} w_{(0,i)}^{S_0\oplus S_1} &= \frac{1-\phi_i}{1-\beta} + \phi_i g_{(1,i)}^{S_0\oplus S_1} - \beta g_{(0,i)}^{S_0\oplus S_1} = \frac{1-\phi_i}{1-\beta} + \phi_i g_i^{S_1} - \beta \big\{ \frac{1-\phi_i}{1-\beta} + \phi_i g_i^{S_1} \big\} \\ &= 1 - \phi_i + \phi_i (1-\beta) g_i^{S_1} = 1 - \phi_i + \phi_i w_i^{S_1}, \end{aligned}$$

where we have used Lemma 4.1, $S_0 \subseteq S_1$, Lemma 4.2(b, c), and (19).

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

$$w_{(1,i)}^{S_0\oplus S_1} = g_{(1,i)}^{S_0\oplus S_1} - \beta g_{(0,i)}^{S_0\oplus S_1} = g_i^{S_1} = \frac{w_i^{S_1}}{1-\beta},$$

where we have used (21), Lemma 4.1, Lemma 4.2(d), and (19).

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

$$w_{(0,i)}^{S_0\oplus S_1} = \frac{1-\phi_i}{1-\beta} + \phi_i g_{(1,i)}^{S_0\oplus S_1} = \frac{1-\phi_i}{1-\beta} + \phi_i g_i^{S_1} = \frac{1-\phi_i}{1-\beta} + \frac{\phi_i}{1-\beta} w_i^{S_1},$$

where we have used (21), Lemma 4.1, Lemma 4.2(b), and (19). □

Note that, at this point in the corresponding analysis in Niño-Mora (2006c) — for the no startup delay case $\phi_i \equiv 1$ — we could immediately establish positivity of marginal workloads, i.e., $w^{\widehat{S}}_{(a^-,i)} > 0$, for $(a^-,i) \in \widehat{N},\widehat{S} \in \widehat{\mathscr{F}}$, which is a prerequisite for PCL-indexability. In the present setting, however, it is clear from Lemma 4.3(c) that $w^{S_0\oplus S_1}_{(1,i)}$, for $i \in S_0$, can become negative if $w^{S_1}_i < \beta$ and ϕ_i is close enough to zero. This is why we cannot use here the same argument in that paper to establish indexability, and use instead the more powerful LP-indexability conditions.

4.2. Reward and Marginal Reward Measures

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

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

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

while the marginal reward measure r^S_i is given by

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

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

Lemma 4.4

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

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

Lemma 4.5 *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) $f_{(0,i)}^{S_0 \oplus S_1} = -c_i + \phi_i f_i^{S_1}$, *for* $i \in S_0$.

(d) $f_{(0,i)}^{S_0 \oplus S_1} = 0 = f_i^{S_0}$, *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) We can write, for $i \in S_1$,

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

where we have used Lemma 4.4 and part (a). Hence, the $f_{(1,i)}^{S_0 \oplus S_1}$'s, for $i \in S_1$, satisfy the evaluation equations in (22) for corresponding terms $f_i^{S_1}$, which yields the result.

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

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

where we have used Lemma 4.4, (22), and parts (a, b).

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

Regarding marginal reward measure $r_{(a^-,i)}^{S_0 \oplus S_1}$, we obtain from (12) and Lemma 4.4 that

$$\begin{aligned} r_{(1,i)}^{S_0 \oplus S_1} &= R_i + \beta \sum_{j \in N} p_{ij} f_{(1,j)}^{S_0 \oplus S_1} - \beta f_{(0,i)}^{S_0 \oplus S_1} \\ r_{(0,i)}^{S_0 \oplus S_1} &= -c_i + \phi_i \{ R_i + \beta \sum_{j \in N} p_{ij} f_{(1,j)}^{S_0 \oplus S_1} \} - \beta f_{(0,i)}^{S_0 \oplus S_1}. \end{aligned} \tag{24}$$

The following result represents marginal reward $r_{(a^-,i)}^{S_0 \oplus S_1}$ in terms of the r_i^S's.

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

(a) $r_{(1,i)}^{S_0 \oplus S_1} = r_i^{S_1}$, *for* $i \in S_1^c$.

(b) $r_{(0,i)}^{S_0 \oplus S_1} = -c_i + r_i^{S_1}$, *for* $i \in S_1^c$.

(c) $r_{(1,i)}^{S_0 \oplus S_1} = \beta c_i + \dfrac{1 - \beta \phi_i}{1 - \beta} r_i^{S_1}$, *for* $i \in S_0$.

(d) $r_{(0,i)}^{S_0 \oplus S_1} = -(1 - \beta) c_i + \phi_i r_i^{S_1}$, *for* $i \in S_0$.

(e) $r_{(1,i)}^{S_0 \oplus S_1} = \dfrac{r_i^{S_1}}{1 - \beta}$, *for* $i \in S_1 \setminus S_0$.

(f) $r_{(0,i)}^{S_0 \oplus S_1} = -c_i + \phi_i \dfrac{r_i^{S_1}}{1 - \beta}$, *for* $i \in S_1 \setminus S_0$.

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

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

where we have used (24), Lemma 4.4, Lemma 4.5(a, b), (22) and (23).

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

$$\begin{aligned} r_{(0,i)}^{S_0\oplus S_1} &= -c_i + \phi_i\big\{1 + \beta \sum_{j\in N} p_{ij} f_{(1,j)}^{S_0\oplus S_1}\big\} - \beta f_{(0,j)}^{S_0\oplus S_1} \\ &= -c_i + \phi_i\big\{1 + \beta \sum_{j\in S_1} p_{ij} f_j^{S_1}\big\} = -c_i + \phi_i r_i^{S_1}, \end{aligned}$$

where we have used (24), Lemma 4.5(a, b) and (23).

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

$$\begin{aligned} r_{(1,i)}^{S_0\oplus S_1} &= f_{(1,i)}^{S_0\oplus S_1} - \beta f_{(0,i)}^{S_0\oplus S_1} = f_i^{S_1} - \beta\big\{-c_i + \phi_i f_i^{S_1}\big\} \\ &= \beta c_i + (1-\beta\phi_i) f_i^{S_1} = \beta c_i + \frac{1-\beta\phi_i}{1-\beta} r_i^{S_1}, \end{aligned}$$

here we have used (24), $S_0 \subseteq S_1$, Lemma 4.4, Lemma 4.5(b, c) and (23).

(d) We have, for $i \in S_0$,

$$\begin{aligned} r_{(0,i)}^{S_0\oplus S_1} &= -c_i + \phi_i f_{(1,i)}^{S_0\oplus S_1} - \beta f_{(0,i)}^{S_0\oplus S_1} = -c_i + \phi_i f_i^{S_1} - \beta\big\{-c_i + \phi_i f_i^{S_1}\big\} \\ &= -(1-\beta)c_i + \phi_i(1-\beta) f_i^{S_1} = -(1-\beta)c_i + \phi_i r_i^{S_1}, \end{aligned}$$

where we have used Lemma 4.4, $S_0 \subseteq S_1$, Lemma 4.5(b, c) and (23).

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

$$r_{(1,i)}^{S_0\oplus S_1} = f_{(1,i)}^{S_0\oplus S_1} - \beta f_{(0,i)}^{S_0\oplus S_1} = f_i^{S_1} = \frac{r_i^{S_1}}{1-\beta},$$

where we have used (24), Lemma 4.4, Lemma 4.5(d) and (23).

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

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

where we have used (24), Lemma 4.4, Lemma 4.5(b), and (23). This completes the proof. □

4.3. Marginal Productivity Measures

We continue by addressing calculation of the 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 nonrestless 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 $\nu_{(a^-,i)}^{S_0\oplus S_1}$ in terms of the ν_i^S's.

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

(a) $v_{(1,i)}^{S_0\oplus S_1} = v_i^{S_1}$, *for* $i \in S_1^c$.

(b) $v_{(0,i)}^{S_0\oplus S_1} = \dfrac{-c_i + r_i^{S_1}}{\frac{1-\phi_i}{1-\beta} + w_i^{S_1}} = \dfrac{w_i^{S_1}}{\frac{1-\phi_i}{1-\beta} + w_i^{S_1}} \big\{ v_i^{S_1} - \dfrac{c_i}{w_i^{S_1}} \big\}$, *for* $i \in S_1^c$.

(c) $v_{(1,i)}^{S_0\oplus S_1} = \dfrac{\beta c_i + \frac{1-\beta\phi_i}{1-\beta} r_i^{S_1}}{\frac{1-\beta\phi_i}{1-\beta}\big\{w_i^{S_1} - \beta\frac{1-\phi_i}{1-\beta\phi_i}\big\}} = \dfrac{w_i^{S_1}}{w_i^{S_1} - \beta\frac{1-\phi_i}{1-\beta\phi_i}} \big\{ v_i^{S_1} + \dfrac{\beta(1-\beta)}{1-\beta\phi_i} \dfrac{c_i}{w_i^{S_1}} \big\}$, *for* $i \in S_0$ *such that* $w_i^{S_1} \neq \beta\frac{1-\phi_i}{1-\beta\phi_i}$.

(d) $v_{(0,i)}^{S_0\oplus S_1} = \dfrac{-(1-\beta)c_i + \phi_i r_i^{S_1}}{1-\phi_i+\phi_i w_i^{S_1}} = \dfrac{-(1-\beta)c_i + \phi_i w_i^{S_1} v_i^{S_1}}{1-\phi_i+\phi_i w_i^{S_1}}$, *for* $i \in S_0$.

(e) $v_{(1,i)}^{S_0\oplus S_1} = v_i^{S_1}$, *for* $i \in S_1 \setminus S_0$.

(f) $v_{(0,i)}^{S_0\oplus S_1} = v_i^{S_1} - \dfrac{\big(1-\beta\big)c_i + \big(1-\phi_i\big)v_i^{S_1}}{1-\phi_i+\phi_i w_i^{S_1}}$, $i \in S_1 \setminus S_0$.

Proof. All parts follow immediately from (13), (25), Lemma 4.3 and Lemma 4.6. □

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

We next draw on the above results to establish that the restless bandits of concern are LP($\widehat{\mathscr{F}}$)-indexable, which ensures the validity of index algorithm $\mathrm{AG}_{\widehat{\mathscr{F}}}$ via Theorem 3.2. See Section 3.3.

Theorem 4.8 *Under Assumption* 3.1, *the restless reformulation of a bandit with switching penalties is LP($\widehat{\mathscr{F}}$)-indexable. Hence, it is indexable, and algorithm* $\mathrm{AG}_{\widehat{\mathscr{F}}}$ *computes its MPI.*

Proof. The defining LP($\widehat{\mathscr{F}}$)-indexability condition (iii) in Section 3.3 was established in Proposition 3.3. As for condition (i), it follows by noting that, for $i \in N$,

$$w_{(1,i)}^{\emptyset\oplus\emptyset} = w_i^{\emptyset} = 1 > 0, \quad w_{(0,i)}^{\emptyset\oplus\emptyset} = \frac{1-\phi_i}{1-\beta} + w_i^{\emptyset} = \frac{1-\phi_i}{1-\beta} + 1 > 0$$
$$w_{(1,i)}^{N\oplus N} = \frac{1-\beta\phi_i}{1-\beta}\big\{w_i^N - \beta\frac{1-\phi_i}{1-\beta\phi_i}\big\} = 1 > 0, \quad w_{(0,i)}^{N\oplus N} = 1-\phi_i+\phi_i w_i^N = 1 > 0,$$

where we have used $w_i^{\emptyset} = w_i^N = 1$ along with emma 4.3(a)–(d), respectively.

Regarding condition (ii), consider an active set $\widehat{S} = S_0 \oplus S_1 \in \widehat{\mathscr{F}}$. Then, we have

$$w_{(a^-,i)}^{S_0\oplus S_1} < 0 \Longrightarrow a^- = 1 \text{ and } i \in S_0 \Longrightarrow (1,i) \notin \partial^{\mathrm{in}}_{\widehat{\mathscr{F}}}\widehat{S} \text{ and } (1,i) \notin \partial^{\mathrm{out}}_{\widehat{\mathscr{F}}}\widehat{S},$$

where we have used Lemma 4.3, (15), (14) and (10). Hence condition (ii) holds.

The proof is now 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 $\mathrm{AG}_{\widehat{\mathscr{F}}}$ into the *Version 3* shown in Table 3. In the latter, we use Lemma 4.7(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 4.7(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 $\mathrm{AG}_{\widehat{\mathscr{F}}}$.

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

$S_0^0:=\emptyset;\ \ S_1^0:=\emptyset;\ \ k_0:=1;\ \ k_1:=1$
while $k_0+k_1\le 2n+2$ **do**
  if $k_1\le n$ **pick** $j_1^{\max}\in\arg\max\{\nu_j^{(k_1-1)}:\ j\in S_1^{c,k_1-1}\}$
  $\nu^{(0,k_1-1)}_{(0,j)}:=\nu_j^{(k_1-1)}-\dfrac{(1-\beta)c_j+(1-\phi_j)\nu_j^{(k_1-1)}}{1-\phi_j+\phi_j w_j^{(k_1-1)}},\ j\in S_1^{k_1-1}\setminus S_0^{k_0-1}$
  if $k_0<k_1$ **pick** $j_0^{\max}\in\arg\max\{\nu^{(0,k_1-1)}_{(0,j)}:\ j\in S_1^{k_1-1}\setminus S_0^{k_0-1}\}$
  if $k_1=n+1$ **or** $\{k_0<k_1\le n$ **and** $\nu_{j_1^{\max}}^{(k_1-1)}<\nu^{(0,k_1-1)}_{j_0^{\max}}\}$
    $i_0^{k_0}:=j_0^{\max};\ \ \nu^*_{(0,i_0^{k_0})}:=\nu^{(0,k_1-1)}_{(0,i_0^{k_0})};\ \ S_0^{k_0}:=S_0^{k_0-1}\cup\{(0,i_0^{k_0})\};\ \ k_0:=k_0+1$
  else
    $i_1^{k_1}:=j_1^{\max};\ \ \nu^*_{(1,i_1^{k_1})}:=\nu^{(k_1-1)}_{(1,i_1^{k_1})};\ \ S_1^{k_1}:=S_1^{k_1-1}\cup\{(1,i_1^{k_1})\};\ \ k_1:=k_1+1$
  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 (5)–(6) 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

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

and

$$\nu^{\mathrm{AT}}_{(0,i)}\triangleq\max_{i\in S\subseteq N}\frac{-c_i+\phi_i f_i^S}{\dfrac{1-\phi_i}{1-\beta}+\phi_i g_i^S}.\tag{27}$$

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

Proposition 4.9 *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

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

where we have used the result that the bandit is $\widehat{\mathscr{F}}$-indexable, and hence if it is optimal to rest it at $(1,i)$ then it is also optimal to rest it at $(0,i)$, along with Lemma 4.2(b) and Lemma 4.5(b).

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

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

where we have used that the bandit is $\widehat{\mathscr{F}}$-indexable (cf. Proposition 4.8), along with Lemma 4.2(c) and Lemma 4.5(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.

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 (2006a). For actual

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

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* (TD) fashion, i.e., from highest to lowest. Table 5 shows its *bottom up* (BU) version: algorithm $\mathrm{AG}^0_{\mathrm{BU}}$. Notice that we have formulated such algorithms in a form that applies to the case where the startup delay is positive at every state j, so that $\phi_j < 1$.

The following is the main result of this paper.

Theorem 5.1 *Algorithms* $\mathrm{AG}^0_{\mathrm{TD}}$ *and* $\mathrm{AG}^0_{\mathrm{BU}}$ *compute the switching index* $\nu^*_{(0,i)}$.

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

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

Proposition 5.2 *Algorithms* $\mathrm{AG}^0_{\mathrm{TD}}$ *and* $\mathrm{AG}^0_{\mathrm{BU}}$ *perform at most* $(5/2)n^2 + O(n)$ *operations each.*

Proof. The operation count is dominated by the statement

$$\nu^{(0,k_1)}_{(0,j)} := \nu_j^{(k_1-1)} - \frac{\widehat{c}_j + \nu_j^{(k_1-1)}}{1 + z_j w_j^{(k_1-1)}}, \quad j \in S_1^{k_1} \setminus S_0^{k_0},$$

Table 5: Switching-Index Algorithm $\mathrm{AG}^0_{\mathrm{TD}}$: Top-Down Version.

ALGORITHM $\mathrm{AG}^0_{\mathrm{TD}}$:
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: $\{i_0^{k_0}\}_{k_0=1}^n$, $\{v^*_{(0,j)}:\ j\in N\}$

$\widehat{c}_j := \dfrac{1-\beta}{1-\phi_j}c_j,\ j\in N;\ \ z_j=\phi_j/(1-\phi_j);\ \ S_0^0:=\emptyset;\ \ S_1^0:=\emptyset;\ \ k_0:=0$
for $k_1:=1$ **to** n **do**
  $S_1^{k_1}:=S_1^{k_1-1}\cup\{i_1^{k_1}\};\ \ \mathrm{AUGMENT}_1:=\texttt{false}$
  $v^{(0,k_1)}_{(0,j)}:=v_j^{(k_1-1)}-\dfrac{\widehat{c}_j+v_j^{(k_1-1)}}{1+z_jw_j^{(k_1-1)}},\ j\in S_1^{k_1}\setminus S_0^{k_0}$
  while $k_0<k_1$ **and not**$(\mathrm{AUGMENT}_1)$ **do**
    pick $j_0^{\max}\in\arg\max\{v^{(0,k_1)}_{(0,j)}:\ j\in S_1^{k_1}\setminus S_0^{k_0}\}$
    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
      $\mathrm{AUGMENT}_1:=\texttt{true}$
    end { if }
  end { while }
end { for }

in algorithm $\mathrm{AG}^0_{\mathrm{TD}}$, and in the statement

$$v^{(0,k_1)}_{(0,j)}:=v_j^{(k_1-1)}-\frac{\widehat{c}_j+v_j^{(k_1-1)}}{1+z_jw_j^{(k_1-1)}},\quad j\in S_0^{k_0},$$

in algorithm $\mathrm{AG}^0_{\mathrm{BU}}$, for $2\le k_1\le n+1$. In each such statement, at most $5k_1$ arithmetic operations are performed, which yields the stated maximum total count. □

6. Dependence of the Index on Switching Penalties

We next present and discuss some insightful properties on how the index depends on switching penalties, focusing on the case $c_i\equiv c$, $d_i\equiv d$ and $\phi_i\equiv\phi$, for $i\in N$ We will make explicit in the notation below the prevailing switching costs, writing the continuation index as $v^*_{(1,i)}(d,\psi)$ — as it does not depend on c nor on ϕ, and the switching index as $v^*_{(1,i)}(c,d,\phi,\psi)$.

We further denote by $v_i^*\ge 0$ and by $f_i^S\ge 0$ the Gittins index and the reward measure of the underlying

Table 6: Switching-Index Algorithm $\mathrm{AG}^0_{\mathrm{BU}}$: Bottom-Up Version.

ALGORITHM $\mathrm{AG}^0_{\mathrm{BU}}$:
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: $\{i_0^{k_0}\}_{k_0=1}^n$, $\{v^*_{(0,j)}:\ j\in N\}$

$\widehat{c}_j := \dfrac{1-\beta}{1-\phi_j}c_j;\ \ z_j=\phi_j/(1-\phi_j);\ \ v^{(0,n)}_{(0,j)} := v_j^{(n)} - \dfrac{\widehat{c}_j + v_j^{(n)}}{1+z_j w_j^{(n)}},\ j\in N$
$S_0^n := N;\ \ S_1^n := N;\ k_1 := n$
for $k_0 := n$ **down to** 1 **do**
  $\mathrm{SHRINK}_0 := \texttt{false}$
  while $k_0 \le k_1$ **and not**(SHRINK_0) **do**
    pick $j_0^{\min} \in \arg\min\{v^{(0,k_1)}_{(0,j)}:\ j\in S_0^{k_0}\}$
    if $k_0 = k_1$ **or** $v^{(0,k_1)}_{(0,j_0^{\min})} \le v^*_{i_1^{k_1}}$
      $i_0^{k_0} := j_0^{\min};\ \ v^*_{(0,i_0^{k_0})} := v^{(0,k_1)}_{(0,i_0^{k_0})}$
      $S_0^{k_0-1} := S_0^{k_0}\setminus\{i_0^{k_0}\};\ \ \mathrm{SHRINK}_0 := \texttt{true}$
    else
      $S_1^{k_1-1} := S_1^{k_1}\setminus\{i_1^{k_1}\};\ \ k_1 := k_1-1$
      $v^{(0,k_1)}_{(0,j)} := v_j^{(k_1-1)} - \dfrac{\widehat{c}_j + v_j^{(k_1-1)}}{1+z_j w_j^{(k_1-1)}},\ j\in S_0^{k_0}$
    end { if }
  end { while }
end { for }

bandit with no switching penalties. We will use the switching-index expression

$$v^*_{(0,i)}(c,d,\phi,\psi) = \max_{i\in S\subseteq N} H\big(c,d,\phi,\psi,f_i^S,g_i^S\big), \tag{28}$$

where

$$H(c,d,\phi,\psi,f,g) \triangleq \frac{-(c+\phi d)+\phi\big(f+(1-\beta)dg\big)}{\dfrac{1-\phi\psi}{1-\beta}+\phi\psi g}.$$

Notice that identity (28) draws on the transformation discussed in Section 2.2 along with the switching-index representation in (27), where we have used that the bandit's reward measure with modified rewards $\widetilde{R}_j = \big\{R_j+(1-\beta)d\big\}/\psi$, for $j\in N$, is given by $\widetilde{f}_i^S = \big\{f_i^S+(1-\beta)dg_i^S\big\}/\psi$.

We will use the following preliminary result.

Lemma 6.1

(a) *If $S\subset S'\subseteq N$, then $f_i^S \le f_i^{S'}$ and $g_i^S \le g_i^{S'}$.*

(b) *If* $d+\psi c \geq \phi\psi f_i^N$, *then* $H(c,d,\phi,\psi,f,g)$ *is increasing in* f *and in* g, *for* $0 \leq f \leq f_i^N$ *and* $0 \leq g \leq g_i^N = 1/(1-\beta)$.

Proof. (a) This part follows immediately from the interpretation of reward and work measures, using Assumption 3.1 for the former.

(b) The result follows immediately from the following expressions:

$$\frac{\partial}{\partial f}H(c,d,\phi,\psi,f,g) = \frac{\phi}{\frac{1-\phi\psi}{1-\beta}+\phi\psi g} > 0 \quad \text{and} \quad \frac{\partial}{\partial g}H(c,d,\phi,\psi,f,g) = \phi\frac{d+\psi c-\phi\psi f}{\left(\frac{1-\phi\psi}{1-\beta}+\phi\psi g\right)^2} > 0.$$

□

Proposition 6.2

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

(b) *If* $d+\psi c \geq \phi\psi f_i^N$, *then* $\nu^*_{(0,i)} = \phi\nu_i^N - (1-\beta)c$.

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

(d) *For* $d+\psi c \geq \phi\psi f_i^N$, *or for* $c,d \geq 0$ *small enough and* $R_i > 0$, *or for* $c = d = 0$, $\nu^*_{(0,i)}(c,d,\phi,\psi)$ *is nondecreasing convex in* ϕ *and in* ψ.

(e) $\lim_{\phi\searrow 0}\nu^*_{(0,i)}(c,d,\phi,\psi) = -(1-\beta)c$.

(f) $\nu^*_{(0,i)}(c,d,\phi,\psi) = \phi\nu_i^N - (1-\beta)c + O(\psi^2)$, *as* $\psi \searrow 0$.

Proof. (a) This part follows immediately from the fact that $\nu^*_{(1,i)}(d,\psi)$ is the Gittins index of the bandit with modified active rewards $\widetilde{R}_j = \{R_j + (1-\beta)d\}/\psi$ (cf. Section 2.2), which is related to the Gittins index ν_i^* of the bandit with unmodified rewards R_j by the given expression.

(b) Use Lemma 6.1(b) and $\nu_i^N = (1-\beta)f_i^N$ to write

$$\nu^*_{(0,i)}(c,d,\phi,\psi) = \max_{(f,g)\in[0,f_i^N]\times[0,g_i^N]} H(c,d,\phi,\psi,f,g) = H\big(c,d,\phi,\psi,f_i^N,g_i^N\big) = \phi\nu_i^N - (1-\beta)c.$$

(c) This part follows by noting that (28) represents $\nu^*_{(0,i)}(c,d,\phi,\psi)$ as the maximum of linear functions in (c,d) that are decreasing in c and nonincreasing in d.

(d) Regarding dependence on ϕ, in the case $d+\psi c \geq \phi\psi f_i^N$ the result follows by part (b). Further, we can write

$$\frac{\partial}{\partial\phi}H\big(c,d,\phi,\psi,f_i^S,g_i^S\big) = \big(1-\beta\big)\frac{f_i^S - \big(1-(1-\beta)g_i^S\big)(d+\psi c)}{\big\{1-\phi\psi\big(1-(1-\beta)g_i^S\big)\big\}^2} \geq 0$$

$$\frac{\partial^2}{\partial\phi^2}H\big(c,d,\phi,\psi,f_i^S,g_i^S\big) = \frac{2(1-\beta)\big(1-(1-\beta)g_i^S\big)\psi}{\big\{1-\phi\psi\big(1-(1-\beta)g_i^S\big)\big\}^3}\big\{f_i^S - \big(1-(1-\beta)g_i^S\big)(d+\psi c)\big\} \geq 0,$$

where the inequalities are easily shown to hold for c, d small enough, using that $R_i > 0$ to ensure that $f_i^S > 0$, and for $c = d = 0$. Thus, $\nu^*_{(0,i)}(c,d,\phi,\psi)$ is the maximum of nondecreasing convex functions, which is also nondecreasing convex.

The same line of argument applies to the dependence on ψ, noting that

$$\frac{\partial}{\partial \psi} H\big(c,d,\phi,\psi,f_i^S,g_i^S\big) = \frac{(1-\beta)\big(1-(1-\beta)g_i^S\big)\phi}{\big\{1-\phi\psi\big(1-(1-\beta)g_i^S\big)\big\}^2}\big\{\phi f_i^S - c - \big(1-\big(1-\beta\big)g_i^S\big)\phi d\big\}$$

$$\frac{\partial^2}{\partial \psi^2} H\big(c,d,\phi,\psi,f_i^S,g_i^S\big) = \frac{2(1-\beta)\big(1-(1-\beta)g_i^S\big)^2\phi^2}{\big\{1-\phi\psi\big(1-(1-\beta)g_i^S\big)\big\}^3}\big\{\phi f_i^S - c - \big(1-\big(1-\beta\big)g_i^S\big)\phi d\big\}.$$

Parts (e) and (f) follow by straightforward algebra. □

We conjecture that the Lemma 6.2(c) should hold without the stated qualifications.

We next give two examples to illustrate the above results. The first concerns the 3-state bandit instance with no shutdown penalties nor startup costs, startup delay transform's value ϕ, $\beta = 0.95$,

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

Figure 3 plots the bandit's switching index for each state vs. $1-\phi$. Notice that the plot is indeed consistent with Proposition 6.2(d, e). It further illustrates that the relative state ordering induced by the switching index can change as ϕ varies.

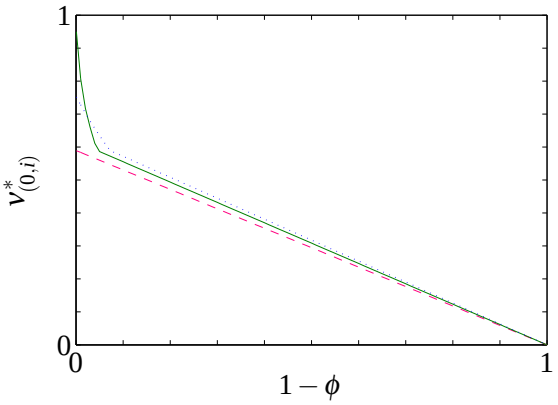

Figure 3: Dependence of Switching Index on Startup Delay Transform.

The next example concerns the same base 3-state bandit, but with no startup delay and shutdown delay transform ψ. Figure 4 plots the continuation and switching indices for each state vs. $1-\psi$. The plots are consistent with Proposition 6.2(a, d, f). Notice that, in particular, the continuation index $\nu^*_{(1,i)}(d,\psi)$ grows to infinity as ψ approaches 0, reflecting that the incentive to stay in a bandit grows steeply as the shutdown delay gets large. Further, the plot for the switching index shows that the relative state ordering induced by it can change as ψ varies.

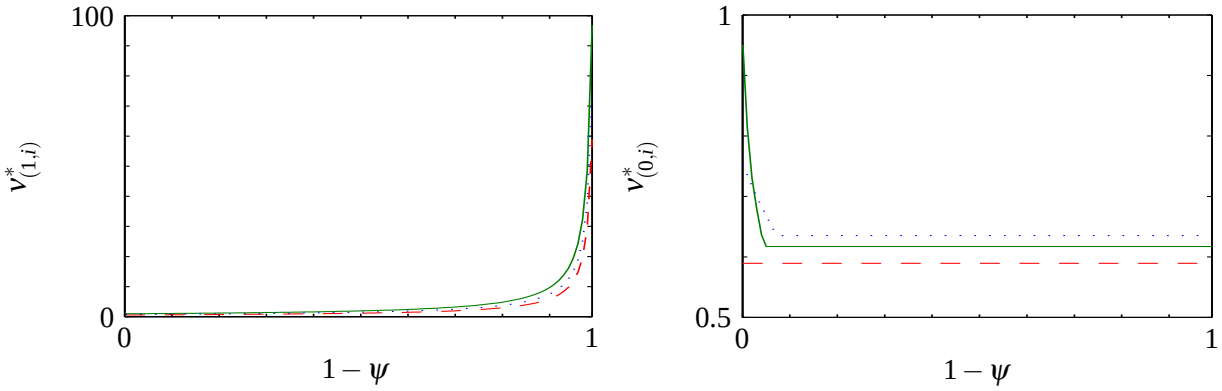

Figure 4: Dependence of Continuation and Switching Indices on Shutdown Delay Transform.

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 (2006a), the time to compute the switching MPI by algorithms $\mathrm{AG}^0_{\mathrm{TD}}$ and $\mathrm{AG}^0_{\mathrm{BU}}$, and the time to jointly compute both indices using algorithm FPAG in (Niño-Mora, 2006a, Sec. 6.3), which is a fast-pivoting implementation of the algorithmic scheme $\mathrm{AG}_{\widehat{\mathscr{F}}}$ discussed herein. This experiment was run under MATLAB R2006b 64-bit on Windows XP x64, on an HP xw9300 2.8 GHz AMD Opteron workstation with 4GB of memory.

The results are displayed in Figure 5. The left pane shows total runtimes, in hours, for computing both indices vs. n, along with curves obtained by cubic least-squares fit, which are consistent with the theoretical $O(n^3)$ complexity. Squares correspond to the $\mathrm{AG}_{\widehat{\mathscr{F}}}$ scheme, while circles correspond to our two-stage scheme. The results show that the two-stage method consistently achieved about a 4-fold speedup over the single-stage method.

The right pane shows runtimes, in *seconds*, for the switching index algorithms vs. n, along with curves obtained by quadratic least-squares fit, which are consistent with the theoretical $O(n^2)$ complexity. Now, squares (resp. circles) correspond to the top-down (resp. bottom-up) algorithm $\mathrm{AG}^0_{\mathrm{TD}}$ (resp. $\mathrm{AG}^0_{\mathrm{BU}}$). The change of timescale from hours to seconds demonstrates the order-of-magnitude runtime improvement achieved. Further, the bottom-up algorithm consistently outperformed the top-down one, though the difference is negligible, given the small runtimes.

We further investigated how the switching index algorithms' relative performance depends on startup

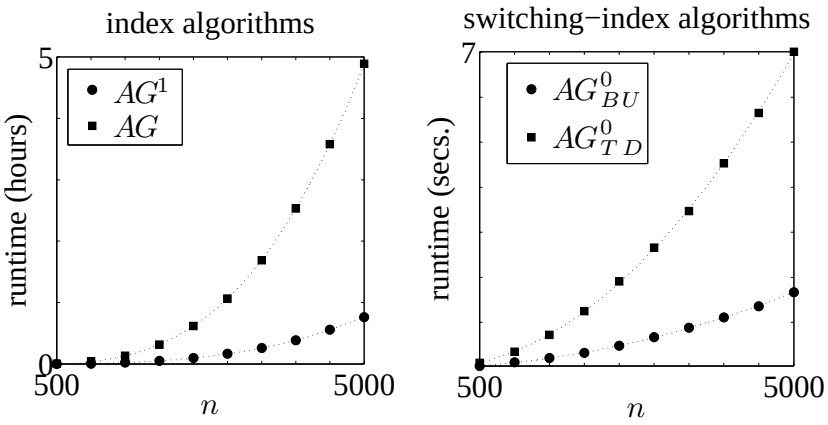

Figure 5: Exp. 1(a):Runtimes of Index Algorithms.

delays. Figure 6 plots the average arithmetic operation count (aoc) for each algorithm over 100 random instances of sizes $n = 500, 1000, \ldots, 5000$, vs. ϕ. The top-down algorithm is better for ϕ small enough (longer startup delays), while the bottom-up one is better for ϕ large enough (shorter delays), which agrees with intuition. Remarkably, the critical ϕ value remains invariant as n varies. The curves shown are obtained by quadratic least-squares fit.

The following experiments 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-delay transform's value ϕ and discount factor — there are no shutdown penalties. A sample of 100 instances (with 10-state bandits) was randomly generated with MATLAB. In every 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]). For each instance $k = 1, \ldots, 100$ and startup cost-discount factor combination in the range $(\phi, \beta) \in [0.5, 0.99] \times [0.5, 0.95]$ — using a 0.1 grid — the optimal objective value $\vartheta^{(k),\text{opt}}$ and the objective values of the MPI ($\vartheta^{(k),\text{MPI}}$) and the benchmark ($\vartheta^{(k),\text{bench}}$) policies

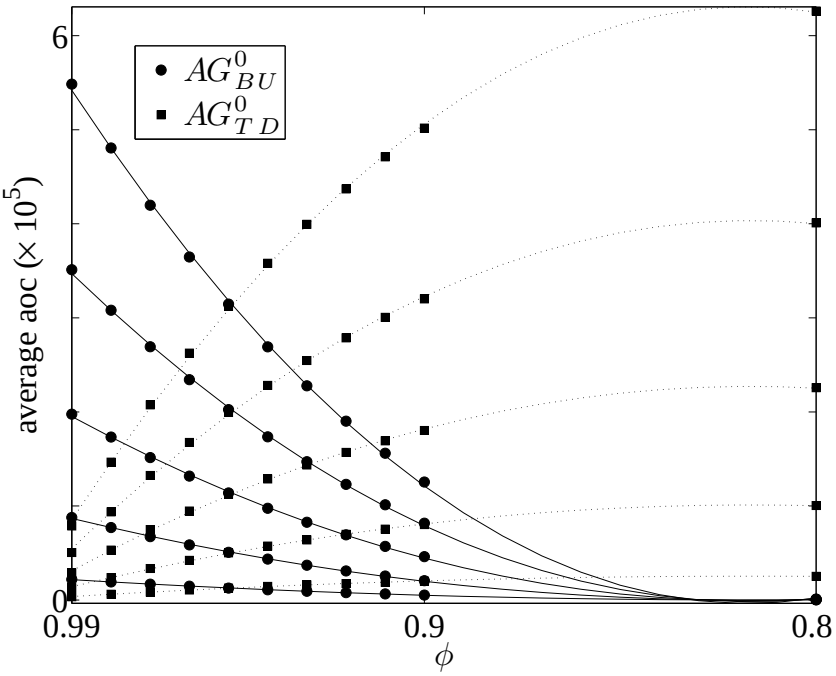

Figure 6: Exp. 1(b): Arithmetic Operation Count of Switching-Index Algorithms vs. ϕ.

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 7 plots Δ^{MPI} vs. the ϕ — notice the inverted ϕ-axis we use throughout — for multiple discount factors β, using cubic interpolation for smoothing. Such a gap starts at 0 as ϕ approaches 1 (as the optimal policy is then recovered), then increases up to a maximum value, which is less than 0.18%, and then decreases to 0 as ϕ gets smaller. Such a pattern is consistent with intuition: for small enough ϕ, 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} is not

monotonic in β.

Figure 8 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 45% of that for the benchmark policy. Such a ratio takes the value 0 for ϕ small enough, as the MPI policy is then optimal. Finally, the ratio increases with β.

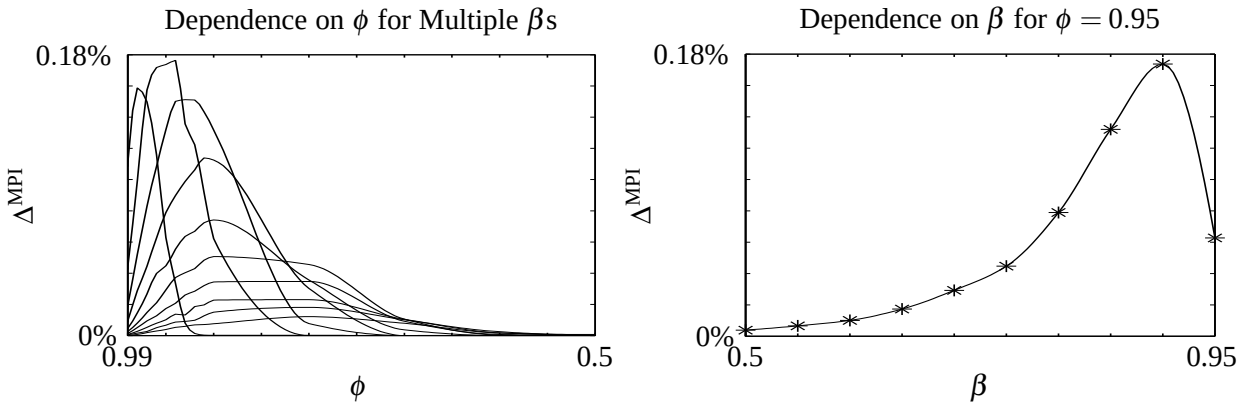

Figure 7: Exp. 2: Average Relative Suboptimality Gap of MPI Policy.

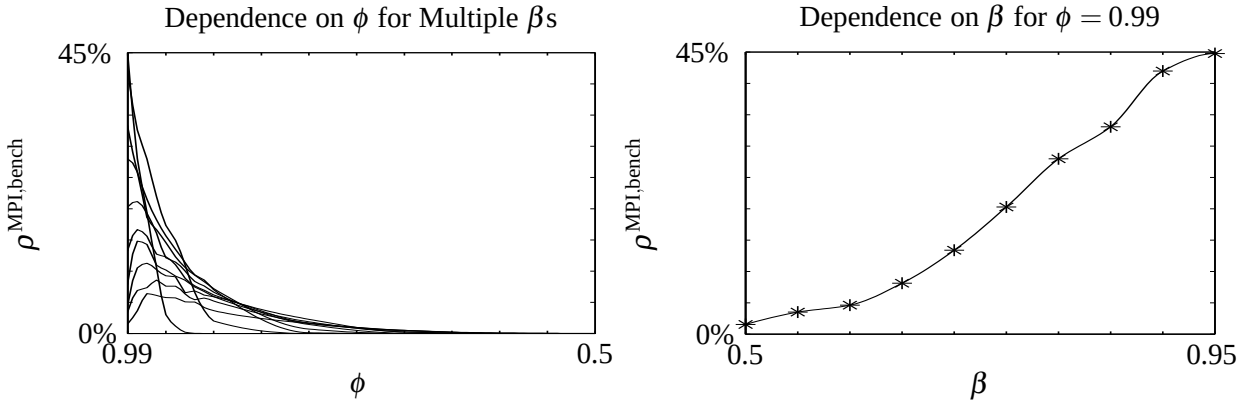

Figure 8: Exp. 2: Average Suboptimality-Gap Ratio of MPI over Benchmark Policy.

The third experiment was setup as the previous one, but considering a constant startup delay T for each bandit, so that $\phi = \beta^T$. Figures 9 and 10 display the results, showing that the MPI policy was optimal for $T \geq 2$, had a relative suboptimality gap of no more than 0.06%, and improved substantially on the benchmark Gittins-index policy, as the suboptimality-gap ratio remains below 2%.

The fourth experiment investigated the effect of asymmetric constant startup delay transform values, as these vary over the range $(\phi_1, \phi_2) \in [0.8, 0.99]^2$, in two-bandit instances with $\beta = 0.9$. The left contour plot in Figure 11 shows that the average relative suboptimality gap of the MPI policy, Δ^{MPI}, reaches a maximum value of about 0.14%, vanishing as both ϕ_1 and ϕ_2 approach unity, and as either gets small enough. The right

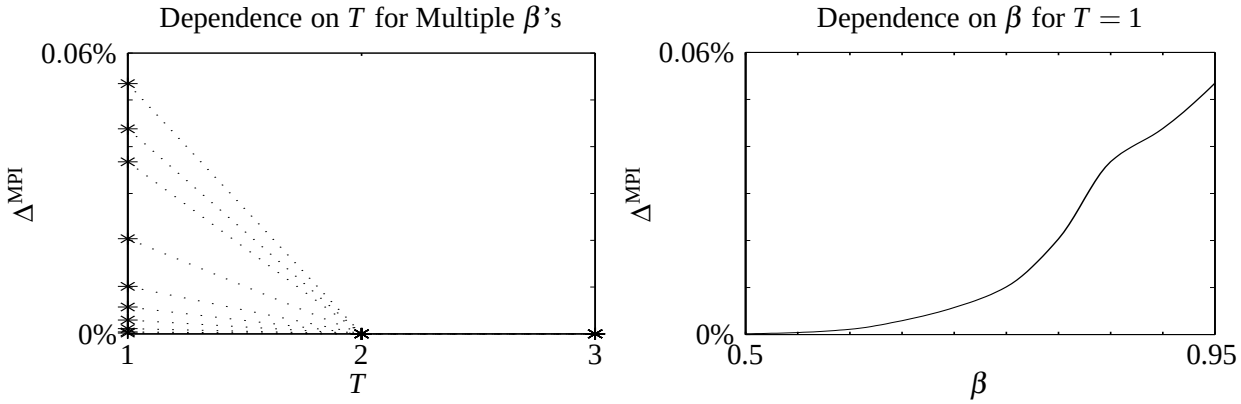

Figure 9: Exp. 3: Average Suboptimality Gap of MPI Policy.

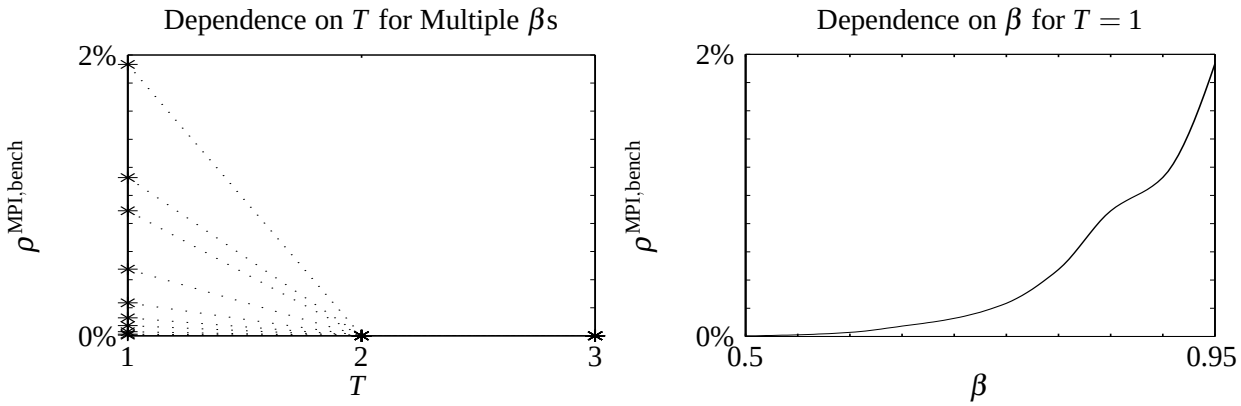

Figure 10: Exp. 3: Average Suboptimality-Gap Ratio of MPI over Benchmark Policy.

contour plot shows that the suboptimality-gap ratio ρ^{MPI} reaches maximum values of about 50%, vanishing as either ϕ_1 or ϕ_2 gets small enough.

The fifth experiment evaluated the effect of state-dependent startup delay parameters ϕ_i, as the discount factor varies. Uniform[0.9, 1] i.i.d. state-dependent startup costs were randomly generated for each instance. The left pane in Figure 12 plots the average relative suboptimality gap vs. the discount factor, which shows that such a gap remains below 0.14%. The right pane shows that the average suboptimality-gap ratio $\rho^{\mathrm{MPI,bench}}$ remains below 20%.

The sixth and last experiment evaluated the relative performance of the MPI policy on three-bandit instances as a function of a common startup delay parameter ϕ 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 $(\phi, \beta) \in [0.5, 0.99] \times [0.5, 0.95]$. The results are shown in Figures 13 and 14, which are the counterparts of experiment 2's Figure 7 and 8. Comparison of Figures 7 and 13 reveals a slight performance degradation of the MPI policy's performance in the latter, though the average gap Δ^{MPI} remains quite small, below 0.25%. Comparison of Figures 8 and 14 reveals similar values for the

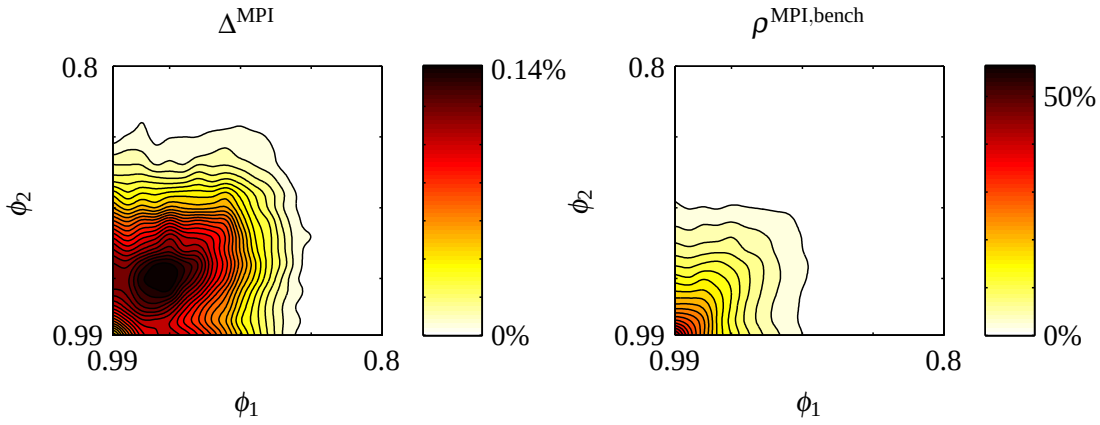

Figure 11: Exp. 4: Average Relative Performance of MPI Policy vs. (ϕ_1, ϕ_2), for $\beta = 0.9$.

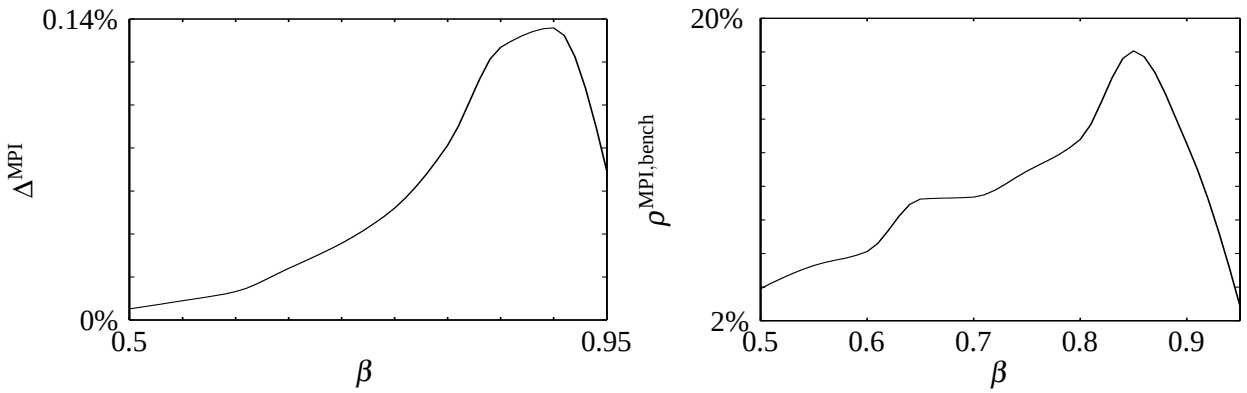

Figure 12: Exp. 5: Average Performance of MPI Policy for State-Dependent Startup Delays.

ratio $\rho^{\mathrm{MPI,bench}}$.

8. Concluding Remarks

We have addressed the important extension of the classic multi-armed bandit problem that incorporates both costs and delays for switching bandits. The paper has demonstrated the practical applicability of the index policy based on the index introduced by Asawa and Teneketzis (1996), by introducing an efficient index algorithm and providing experimental evidence of the near optimality of such a policy. 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. Thus, the Asawa and Teneketzis index has been shown to be precisely the Whittle index of the bandits of concern in their natural restless reformulation. To establish indexability and compute the index we have deployed the LP-indexability approach recently introduced in Niño-Mora (2007), which extends the earlier PCL-indexability approach in the author's earlier work. This paper demonstrates the relevance of such an extension, since the restless bandits analyzed herein

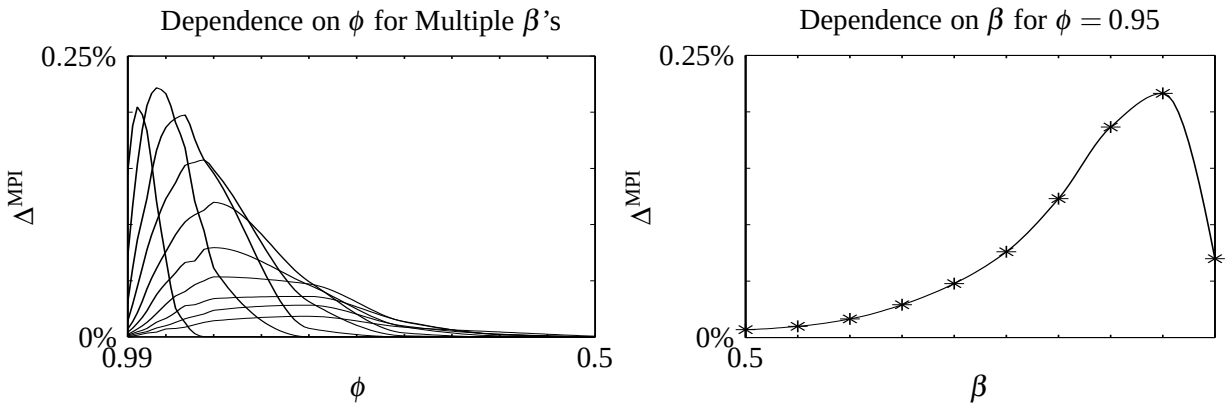

Figure 13: Exp. 6: Counterpart of Figure 7 for Three-Bandit Instances.

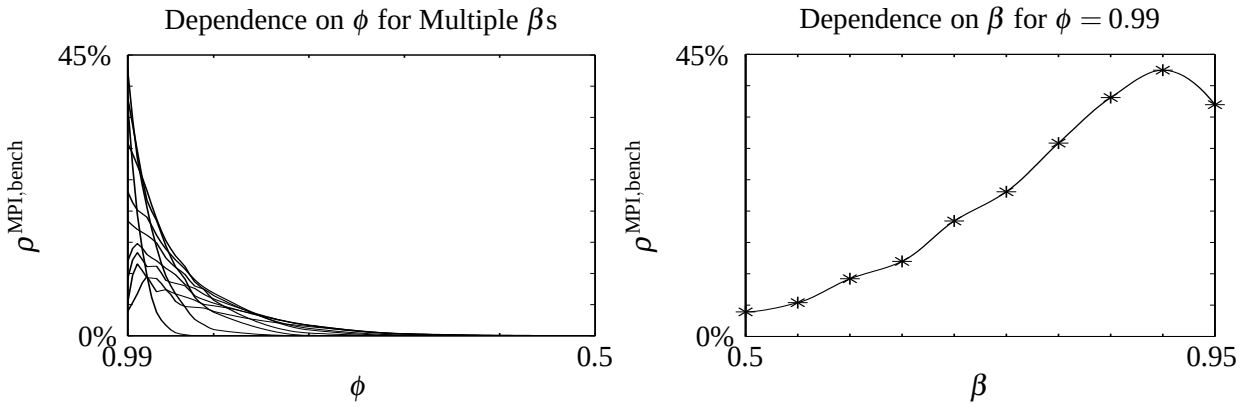

Figure 14: Exp. 6: Counterpart of Figure 8 for Three-Bandit Instances.

have been proven to be LP-indexable, yet are not PCL-indexable.

Acknowledgments

This work was 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 and Euro-FGI, and by the Autonomous Community of Madrid-UC3M under grants UC3M-MTM-05-075 and CCG06-UC3M/ESP-0767.

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. 1979. Bandit processes and dynamic allocation indices. *J. Roy. Statist. Soc. Ser. B* **41** 148–177. With discussion.

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. 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. 2006b. 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. 2006c. Two-stage index computation for bandits with switching penalties I: switching costs. Working Paper 07-41, Statistics and Econometrics Series 09, `http://halweb.uc3m.es/jnino/eng/public2.html`, Univ. Carlos III de Madrid, Spain. Conditionally accepted at INFORMS J. Comput.

Niño-Mora, J. 2007. Characterization and computation of restless bandit marginal productivity indices. Working Paper 07-43, Statistics and Econometrics Series 11, `http://halweb.uc3m.es/jnino/eng/public2.html`, 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.