Multi-Armed Bandits with Switching Penalties

Manjari Asawa and Demosthenis Teneketzis, *Member, IEEE*

Abstract— **The multi-armed bandit problem with switching penalties (switching cost and switching delays) is investigated. It is shown that under an optimal policy, decisions about the processor allocation need to be made only at stopping times that achieve an appropriate index, the well-known "Gittins index" or a "switching index" that is defined for switching cost and switching delays. An algorithm for the computation of the "switching index" is presented. Furthermore, sufficient conditions for optimality of allocation strategies, based on limited look-ahead techniques, are established. These conditions together with the above-mentioned feature of optimal scheduling policies simplify the search for an optimal allocation policy.**

For a special class of multi-armed bandits (scheduling of parallel queues with switching penalties and no arrivals), it is shown that the aforementioned property of optimal policies is sufficient to determine an optimal allocation strategy. In general, the determination of optimal allocation policies remains a difficult and challenging task.

I. Introduction

MODELS of dynamic allocation of a scarce resource to competing projects have been widely used and are of great importance. The multi-armed bandit problem is concerned with the question of how to dynamically allocate a single resource among several alternative projects. It is important because it has found applications in several disciplines, such as machine scheduling in manufacturing (see, for example, [1]–[3] and references therein), job search and labor market analysis in economics [4], search problems in oil exploration [5], target tracking [1], [6], resource allocation problems in communication networks [7], industrial research under budget constraint [1], job selection [8], [9], clinical trials [10], etc. Furthermore, it is important from a theoretical point of view because it is one of the simplest nontrivial problems in the area of stochastic control where one must face the conflict between taking actions which yield an immediate reward and taking actions (such as learning about the system or preparing for the future) the benefit of which will come later [11]. It is a classical problem in stochastic control that has witnessed major advances since 1972 when Gittins [12] first solved the problem.

In the basic version of the stochastic multi-armed bandit problem, there are N independent machines or projects (in the rest of this paper, machines, bandits, and projects are used interchangeably) and one server. Let $x_i(t)$ denote the state of machine i at time t, where $i = 1, 2, \cdots, N$ and $t = 1, 2, \cdots$. At each time instant t, the server must select exactly one machine for operation (or service). Denote this machine by $m(t)$. If $m(t) = i$, i.e., machine i is selected for operation at time t, an immediate reward $R(t) := R_i(x_i(t))$ is obtained, and the state of machine i changes to $x_i(t+1)$ according to a stationary Markov transition rule. The states of the idle machines remain frozen, i.e., $x_j(t+1) = x_j(t)$, $j \neq i$. The states of all machines are perfectly observed, and the objective is to schedule the order in which machines are to be operated so as to maximize an infinite horizon expected discounted reward $\mathcal{R}$, given by

$$\mathcal{R} = E\left\{\sum_{t=0}^{\infty} \beta^t R(t)\right\}$$

where $0 < \beta < 1$ is a fixed discount factor.

This problem has received considerable attention since it was first formulated in the 1940's, but no substantial progress toward its solution was made until 1972 when Gittins and Jones [12], using a forward induction argument, showed that the optimal dynamic allocation policy for the aforementioned bandit problem is described by the following rule: To each machine i attach an index $\nu_i(\cdot)$ that is a function only of machine i's state and the information concerning machine i; at each instant of time operate the machine with the largest current index. This strategy is called the Gittins index rule. The Gittins index rule result is very significant because it converts the N-dimensional bandit problem into N one-dimensional problems. Hence, the implementation of the optimal policy involves only finding the maximum of N numbers and each of these numbers can be calculated individually by the corresponding machine.

The index $\nu_i(\cdot)$ (called the Gittins index or Dynamic Allocation index) was subsequently [13] shown to be

$$\nu_i(x_i) = \max_{\tau > 1} \frac{E\left\{\sum_{t=1}^{\tau-1} \beta^t R_i(x_i(t)) \mid x_i(1) = x_i\right\}}{E\left\{\sum_{t=1}^{\tau-1} \beta^t \mid x_i(1) = x_i\right\}} \tag{1.1}$$

where the maximization is over the set of all stopping times τ which are greater than one. Gittins interpreted $\nu_i(\cdot)$ as the maximum expected discounted reward per unit of expected discounted time. In the remainder of this paper, we shall use the term expected discounted reward rate to refer to expected discounted reward per unit of expected discounted time.

In 1980, Whittle [14] came up with an elegant proof of the optimality of the index rule using dynamic programming (DP). He introduced the "retirement benefit option" (or "retirement reward") and using that, he was able to decouple the original N-machine problem into N 1-machine problems, each one

Manuscript received April 15, 1994; revised February 26, 1995. Recommended by Associate Editor, T. S. Chang. This research was supported in part by the National Science Foundation under Grant NCR-9204419. The work of Manjari Asawa was supported in part by a Fellowship from the University of MI.

The authors are with the Department of Electrical Engineering and Computer Science at the University of Michigan, Ann Arbor, Michigan 48109 USA.

Publisher Item Identifier S 0018-9286(96)02098-3.

concerned with the optimal operation of an individual machine. The idea of "retirement reward" is very intimately related to the Gittins index. Subsequently, Gittins' original work was also extended in various directions such as "superprocesses" [13], arm-acquiring bandits [15], non-Markovian bandits [16], and correlated bandits [17], [18]. Several variations of the multi-armed bandit problem in discrete or continuous time were formulated, and the optimality of an index policy was reported in [19]–[25]. A considerable amount of effort was also made to calculate the index ([1], [10], [16], and [26]) and to compute good suboptimal policies ([1], [27], and [28] and references therein).

An assumption maintained in all of the above extensions is that the server can switch instantaneously from one machine to another. In reality, when the server switches between different machines, a new setup may be needed, and a delay and/or a cost is incurred. During the switching period no machine in the system is served, and this lack of service can be modeled either by a switching cost or a switching delay. Although it is often realistic to include a penalty for each switch made from one project to another, the inclusion of switching penalty drastically changes the nature of the bandit problem. As shown in [29], the optimal policy is not given by an index rule anymore. The resulting problem is difficult, and to the best of our knowledge only a very limited number of results are available so far. Agrawal *et al.* [30], [31] have solved the bandit problem with switching cost and a performance criterion that is described by the "learning loss" or "regret." Van Oyen *et al.* [3] showed that for a system of parallel queues with switching penalties and no arrivals, optimal scheduling policies are characterized by an index-like rule. Glazebrook [32] showed that for stochastic scheduling with precedence relations and switching costs, it is enough to search for optimal policies among the class of nonpreemptive policies.

The nature of optimal strategies for the general multi-armed bandit problem with switching penalty is not known. Optimal strategies can no longer be obtained by a forward induction argument, and it may be difficult to explicitly determine them. However, knowledge of properties of optimal policies can guide the search for an optimal policy. Hence, it becomes useful to discover some of the qualitative properties of optimal policies. The main contributions of this paper are i) the development of qualitative properties of optimal strategies for the multi-armed bandit problem with switching penalties and ii) the establishment of sufficient conditions for optimality of allocation strategies in multi-armed bandits with switching cost, based on limited look-ahead techniques. These results simplify the search for optimal allocation policies.

This paper is organized as follows: In Section II, the deterministic bandit problem with switching cost is investigated. Attention is initially restricted to this problem for two reasons: i) to convey the main arguments used to derive qualitative properties of optimal policies and ii) to highlight how the properties of optimal policies combined with simple limited look-ahead techniques can simplify the search for optimal allocation strategies. Stochastic bandits with switching cost are discussed in Section III. In Section IV, three possible extensions to multi-armed bandits with switching cost are discussed. These extensions include bandit problems with switching delays. A summary of the paper's main results and conclusions appear in Section VI.

II. The Deterministic Two-Armed Bandit Problem with Switching Cost

A. Problem Formulation

There are two-deterministic machines, X and Y, and one server. Machine $X[Y]$ is characterized by its reward sequence $\{X(s), s = 0, 1, 2, \cdots\}$ $[\{Y(s), s = 0, 1, 2, \cdots\}]$, where $X(s)[Y(s)]$ represents the reward obtained when machine $X[Y]$ is operated for the $(s+1)$th time. We assume that $\sum_{t=0}^{\infty} \beta^t |X(t)| < \infty$, where $0 < \beta < 1$ is a fixed discount factor, and that a similar condition holds for machine Y. At each time instant exactly one machine must be operated. Thus $t = t^X + t^Y$, where $t^i := t^i(t)$, $i = X, Y$ is the number of times machine i is operated during time $0, 1, \cdots, t-1$. Denote by $m(t)$ the machine that is operated at time t. If $m(t) = i$, i.e., machine i is selected for operation at time t, an immediate reward of $R(t) = i(t^i(t))$ is obtained. The idle machine remains frozen and does not yield any reward. If $m(t) \neq m(t-1)$, a switching cost C is incurred at time t.

The deterministic two-armed bandit problem with switching cost is to determine the order of operation of the machines to maximize

$$\sum_{t=0}^{\infty} \beta^t (R(t) - I(m(t) \neq m(t-1))C) \tag{2.1}$$

where $I(m(t) \neq m(t-1))$ is the indicator function of the event $\{m(t) \neq m(t-1)\}$. A switching cost at $t = 0$ may or may not incur.

The Gittins index rule is the policy that assigns the server to the machine with the highest Gittins index with ties broken arbitrarily. Unlike the problem without switching cost, the Gittins index rule is no longer optimal for the problem with switching cost. To see this, consider the following example.

The two machines X and Y are characterized by reward sequences $(20, 18, 0, 0, \cdots)$ and $(19, 17, 0, 0, \cdots)$, respectively. Let $C = 3$ and $\beta = 0.5$. Then, the Gittins index policy operates machines in the order X, Y, X, Y and yields a reward of $(20-3)+(19-3)\beta+(18-3)\beta^2+(17-3)\beta^3 = 30.5$, whereas the policy that operates machines in the order X, X, Y, Y yields a reward of 32.125.

As pointed out in the previous section, the solution to the deterministic two-armed bandit problem with switching cost is not presently known. In this section, we develop properties of optimal scheduling policies that guide the search for an optimal solution.

B. Analysis

To proceed with the analysis, define for machine X after it has been operated t times the following quantities: the Gittins index

$$\nu_{gx}(t) = \max_{\tau > t} \frac{\sum_{l=t}^{\tau-1} \beta^l X(l)}{\sum_{l=t}^{\tau-1} \beta^l} \tag{2.2}$$

and the "switching cost index"

$$\nu_{cx}(t) = \max_{\tau > t} \frac{\sum_{l=t}^{\tau-1} \beta^l X(l) - C\beta^t}{\sum_{l=t}^{\tau-1} \beta^l}. \tag{2.3}$$

Suppose the maximum in (2.2) and (2.3) is achieved at time $\tau_{gx}(t)$ and $\tau_{cx}(t)$, respectively (in case of more than one maximizer, choose the smaller). Quantities similar to (2.2) and (2.3) are also defined for machine Y. These quantities have the following intuitive interpretation:

i) The Gittins index $\nu_{gx}(t)$ represents the maximum discounted reward rate that can be obtained from machine X after it has been operated t times, when no switching cost is incurred at the time instant X is operated for the $(t+1)$th time.

ii) The "switching cost index" $\nu_{cx}(t)$ represents the maximum discounted reward rate that can be obtained from machine X after it has been operated t times, when a switching cost C is incurred at the time instant X is operated for the $(t+1)$th time.

According to [24], the Gittins and switching indexes can also be interpreted as follows; suppose an operator has to pay a fixed charge each time it operates a machine and there are two possibilities: i) at the beginning of the operation, the operator has to pay an extra fee C to obtain the right to operate the machine; ii) the operator does not incur any extra fee at the beginning of the operation. The Gittins index $\nu_{gx}(t)$ represents the "fair charge" the operator has to pay to operate the machine with reward sequence $\{X(t), X(t+1), \cdots\}$ in case ii). The switching index $\nu_{cx}(t)$ represents the fair charge the operator has to pay to operate the same machine in case i). The Gittins and switching indexes are related as follows.

Lemma 2.1: For a given t

$$\nu_{gx}(t) > \nu_{cx}(t). \tag{2.4}$$

Proof: Let $\tau_{gx}(t)$ and $\tau_{cx}(t)$ be the stopping times that achieve the Gittins and switching indexes $\nu_{gx}(t)$ and $\nu_{cx}(t)$, respectively. Then, by the definition of $\nu_{gx}(t)$, $\nu_{cx}(t)$, $\tau_{gx}(t)$, and $\tau_{cx}(t)$, we have

$$\begin{aligned}\nu_{gx}(t) &\geq \frac{\sum_{l=t}^{\tau_{cx}(t)-1} \beta^l X(l)}{\sum_{l=t}^{\tau_{cx}(t)-1} \beta^l} > \frac{\sum_{l=t}^{\tau_{cx}-1} \beta^l X(l) - C\beta^t}{\sum_{l=t}^{\tau_{cx}(t)-1} \beta^l} \\ &= \nu_{cx}(t). \end{aligned}$$

□

The following lemma characterizes properties of discounted reward rates received by the operation of individual machines and relates them to their Gittins and switching cost indexes and the stopping times that achieve these indexes.

Lemma 2.2: For a given t

$$\frac{\sum_{l=\sigma}^{\tau_{cx}(t)-1} \beta^l X(l)}{\sum_{l=\sigma}^{\tau_{cx}(t)-1} \beta^l} \geq \nu_{cx}(t) \geq \frac{\sum_{l=t}^{t_1-1} \beta^l X(l) - C\beta^t}{\sum_{l=t}^{t_1-1} \beta^l} \quad \forall t_1 > t,\ t \leq \sigma < \tau_{cx}(t) \tag{2.5}$$

and

$$\frac{\sum_{l=\sigma}^{\tau_{gx}(t)-1} \beta^l X(l)}{\sum_{l=\sigma}^{\tau_{gx}(t)-1} \beta^l} \geq \nu_{gx}(t) \geq \frac{\sum_{l=t}^{t_1-1} \beta^l X(l)}{\sum_{l=t}^{t_1-1} \beta^l} \quad \forall t_1 > t,\ t \leq \sigma < \tau_{gx}(t). \tag{2.6}$$

Proof: Inequality (2.6) is a standard result in multi-armed bandits [16]. Inequality (2.5) can be proved by similar arguments. □

Furthermore, the stopping times $\tau_{gx}(t)$ and $\tau_{cx}(t)$ defined above are related as follows.

Lemma 2.3: For all t, $\tau_{cx}(t) \geq \tau_{gx}(t)$.

Proof: Suppose the statement of the lemma is not true; then for some t, we have $\tau_{cx}(t) < \tau_{gx}(t)$. By the definition of $\tau_{gx}(\cdot)$

$$\frac{\sum_{l=t}^{\tau_{cx}(t)-1} \beta^l X(l)}{\sum_{l=t}^{\tau_{cx}(t)-1} \beta^l} \leq \frac{\sum_{l=t}^{\tau_{gx}(t)-1} \beta^l X(l)}{\sum_{l=t}^{\tau_{gx}(t)-1} \beta^l}. \tag{2.7}$$

As $\beta > 0$, $C > 0$, and $\tau_{cx}(t) < \tau_{gx}(t)$

$$\frac{C\beta^t}{\sum_{l=t}^{\tau_{cx}(t)-1} \beta^l} > \frac{C\beta^t}{\sum_{l=t}^{\tau_{gx}(t)-1} \beta^l}. \tag{2.8}$$

From (2.7) and (2.8) it follows that

$$\frac{\sum_{l=t}^{\tau_{cx}(t)-1} \beta^l X(l) - C\beta^t}{\sum_{l=t}^{\tau_{cx}(t)-1} \beta^l} < \frac{\sum_{l=t}^{\tau_{gx}(t)-1} \beta^l X(l) - C\beta^t}{\sum_{l=t}^{\tau_{gx}(t)-1} \beta^l}. \tag{2.9}$$

Also, by the definition of $\tau_{cx}(\cdot)$

$$\frac{\sum_{l=t}^{\tau_{cx}(t)-1} \beta^l X(l) - C\beta^t}{\sum_{l=t}^{\tau_{cx}(t)-1} \beta^l} = \max_{\tau > t} \frac{\sum_{l=t}^{\tau-1} \beta^l X(l) - C\beta^t}{\sum_{l=t}^{\tau-1} \beta^l}. \tag{2.10}$$

Equations (2.9) and (2.10) are contradictory, hence $\tau_{cx}(t) < \tau_{gx}(t)$ is not possible. Consequently, $\tau_{cx}(t) \geq \tau_{gx}(t)$, $\forall t$. □

To determine the qualitative properties of an optimal policy, we proceed via a series of lemmas. Suppose, that a policy π which plays machines in the following order

$$\begin{aligned} &X(0), X(1) \cdots X(t_1 - 1), Y(0), \cdots Y(s_1 - 1), X(t_1), \cdots \\ &\quad X(t_2 - 1), Y(s_1), \cdots Y(s_2 - 1), \cdots, X(t_{k-1}), \cdots \\ &\quad X(t_k - 1), Y(s_{k-1}), \cdots Y(s_k - 1), X(t_k), \cdots, \\ &\quad X(\tau_x(0) - 1), \cdots X(t_{k+1} - 1), Z(1), Z(2), \cdots \end{aligned}$$

is optimal, where $Z(1), Z(2), \cdots$ are an interleaving of reward sequences from X and Y from time $t_{k+1} + s_k$ onwards and $\tau_x(0) = \tau_{cx}(0)$ or $\tau_{gx}(0)$ depending on whether or not the server has to pay a switching cost at $t = 0$. The time instants t_k and t_{k+1} are such that $t_k < \tau_x(0) \leq t_{k+1}$. Let $V(\pi)$ denote the infinite horizon discounted reward obtained from policy π.

Lemma 2.4: When π is optimal, we must have:

i) a) If at $t = 0$ a switching cost is always incurred or never incurred, then

$$\frac{\sum_{l=0}^{t_1-1} \beta^l X(l) - C}{\sum_{l=0}^{t_1-1} \beta^l} \geq \frac{\sum_{l=0}^{s_1-1} \beta^l Y(l) - C}{\sum_{l=0}^{s_1-1} \beta^l}. \tag{2.11}$$

b) If at $t = 0$ machine X does not incur a switching cost whereas machine Y does, then

$$\frac{\sum_{l=0}^{t_1-1} \beta^l X(l)}{\sum_{l=0}^{t_1-1} \beta^l} \geq \frac{\sum_{l=0}^{s_1-1} \beta^l Y(l) - C - C\beta^{s_1}}{\sum_{l=0}^{s_1-1} \beta^l}. \tag{2.12}$$

c) If at $t = 0$ machine X incurs a switching cost whereas machine Y does not, then

$$\frac{\sum_{l=0}^{t_1-1}\beta^l X(l) - C - C\beta^{t_1}}{\sum_{l=0}^{t_1-1}\beta^l} \geq \frac{\sum_{l=0}^{s_1-1}\beta^l Y(l)}{\sum_{l=0}^{s_1-1}\beta^l}. \quad (2.13)$$

ii) At times $s_i + t_i$, $i = 1, 2, \cdots, k-1$

$$\frac{\sum_{l=t_i}^{t_{i+1}-1}\beta^l X(l) - C\beta^{t_i} - C\beta^{t_{i+1}}}{\sum_{l=t_i}^{t_{i+1}-1}\beta^l} \geq \frac{\sum_{l=s_i}^{s_{i+1}-1}\beta^l Y(l)}{\sum_{l=s_i}^{s_{i+1}-1}\beta^l}. \quad (2.14)$$

iii) At times $s_i + t_{i+1}$, $i = 0, 1, 2, \cdots, k-1$

$$\frac{\sum_{l=s_i}^{s_{i+1}-1}\beta^l Y(l) - C\beta^{s_i} - C\beta^{s_{i+1}}}{\sum_{l=s_i}^{s_{i+1}-1}\beta^l} \geq \frac{\sum_{l=t_{i+1}}^{t_{i+2}-1}\beta^l X(l)}{\sum_{l=t_{i+1}}^{t_{i+2}-1}\beta^l}. \quad (2.15)$$

Proof:

i) Consider policy $\tilde{\pi}$ obtained from π as follows: policy $\tilde{\pi}$ operates machine Y for s_1 units of time, then it operates machine X for t_1 units of time and follows π from time $t_1 + s_1$ onwards. Assume that at $t = 0$ a switching cost is always incurred (the argument is the same when a switching cost is never incurred). Then, the rewards obtained from policies π and $\tilde{\pi}$ are

$$V(\pi) = -C + \sum_{l=0}^{t_1-1}\beta^l X(l) - C\beta^{t_1} + \beta^{t_1}\sum_{l=0}^{s_1-1}\beta^l Y(l) - C\beta^{t_1+s_1} + \beta^{s_1}\sum_{l=t_1}^{t_2-1}\beta^l X(l) + \Delta$$

where Δ is the discounted reward earned from time $s_1 + t_2$ onwards, and

$$V(\tilde{\pi}) = -C + \sum_{l=0}^{s_1-1}\beta^l Y(l) - C\beta^{s_1} + \beta^{s_1}\sum_{l=0}^{t_2-1}\beta^l X(l) + \Delta$$

respectively. Consequently

$$\frac{(V(\pi) - V(\tilde{\pi}))(1-\beta)}{(1-\beta^{s_1})(1-\beta^{t_1})} = \frac{\sum_{l=0}^{t_1-1}\beta^l X(l) - C}{\sum_{l=0}^{t_1-1}\beta^l} - \frac{\sum_{l=0}^{s_1-1}\beta^l Y(l) - C}{\sum_{l=0}^{s_1-1}\beta^l} - \frac{C\beta^{t_1+s_1}(1-\beta)}{(1-\beta^{s_1})(1-\beta^{t_1})}.$$

Since, π is assumed to be optimal, we must have $V(\pi) - V(\tilde{\pi}) \geq 0$, therefore

$$\frac{\sum_{l=0}^{t_1-1}\beta^l X(l) - C}{\sum_{l=0}^{t_1-1}\beta^l} - \frac{\sum_{l=0}^{s_1-1}\beta^l Y(l) - C}{\sum_{l=0}^{s_1-1}\beta^l} \geq \frac{C\beta^{t_1+s_1}(1-\beta)}{(1-\beta^{s_1})(1-\beta^{t_1})} \geq 0$$

and the proof of part i-a) is complete. The proof of parts i-b) and i-c) follows from arguments similar to the above, hence it is omitted.

ii) Consider policy π' obtained from π as follows: policy π' follows π up to time $t_i + s_i - 1$, at time $t_i + s_i$ it operates the machine Y for $s_{i+1} - s_i$ units of time, then it operates machine X for $t_{i+1} - t_i$ units of time and follows π again from time $t_{i+1} + s_{i+1}$ onwards.

Let α denote the discounted reward obtained until time $t_i + s_i - 1$ from policy π. Then, the discounted rewards obtained from policies π and π' are

$$V(\pi) = \alpha - C\beta^{s_i+t_i} + \beta^{s_i}\sum_{l=t_i}^{t_{i+1}-1}\beta^l X(l) - C\beta^{s_i+t_{i+1}} + \beta^{t_{i+1}}\sum_{l=s_i}^{s_{i+1}-1}\beta^l Y(l) - C\beta^{s_{i+1}+t_{i+1}} + \Lambda$$

where Λ is the reward earned from time $s_{i+1} + t_{i+1}$ onwards, and

$$V(\pi') = \alpha + \beta^{t_i}\sum_{l=s_i}^{s_{i+1}-1}\beta^l Y(l) - C\beta^{s_{i+1}+t_i} + \beta^{s_{i+1}}\sum_{l=t_i}^{t_{i+1}-1}\beta^l X(l) + \Lambda$$

respectively. Therefore

$$\frac{(V(\pi) - V(\pi'))(1-\beta)}{(\beta^{s_i} - \beta^{s_{i+1}})(\beta^{t_i} - \beta^{t_{i+1}})} = \frac{\sum_{l=t_i}^{t_{i+1}-1}\beta^l X(l) - C\beta^{t_i} - C\beta^{t_{i+1}}}{\sum_{l=t_i}^{t_{i+1}-1}\beta^l} - \frac{\sum_{l=s_i}^{s_{i+1}-1}\beta^l Y(l)}{\sum_{l=s_i}^{s_{i+1}-1}\beta^l} - \frac{2C\beta^{s_{i+1}+t_{i+1}}(1-\beta)}{(\beta^{s_i} - \beta^{s_{i+1}})(\beta^{t_i} - \beta^{t_{i+1}})}.$$

Since π is assumed to be optimal we must have, $V(\pi) - V(\pi') \geq 0$. Consequently

$$\frac{\sum_{l=t_i}^{t_{i+1}-1}\beta^l X(l) - C\beta^{t_i} - C\beta^{t_{i+1}}}{\sum_{l=t_i}^{t_{i+1}-1}\beta^l} - \frac{\sum_{l=s_i}^{s_{i+1}-1}\beta^l Y(l)}{\sum_{l=s_i}^{s_{i+1}-1}\beta^l} \geq \frac{2C\beta^{t_{i+1}+s_{i+1}}(1-\beta)}{(\beta^{s_i} - \beta^{s_{i+1}})(\beta^{t_i} - \beta^{t_{i+1}})} \geq 0$$

should be true.

iii) The proof is very similar to that of part ii) and is therefore omitted. □

Lemma 2.4 has the following intuitive interpretation: Suppose that along an optimal policy machine X's (Y's) operation is interrupted at some time instant and machine Y (X) is operated for l units of time. Then the discounted reward rate received by the operation of machine Y (X) must exceed the discounted reward rate received by the next operation of machine X (Y) at least by a specific amount that depends on the switching cost C, l and the duration of the next operation of machine X (Y).

Based on Lemmas 2.2 and 2.4, we can prove the following result.

Lemma 2.5: Consider the policy π that is described at the beginning of this section and is assumed to be optimal. Construct a policy $\hat{\pi}$ from π as follows: policy $\hat{\pi}$ follows π until time $s_{k-1} + t_k - 1$, at time $s_{k-1} + t_k$ it continues the operation of machine X for $\tau_x(0) - t_k$ units of time (recall $t_k < \tau_x(0) \leq t_{k+1}$), then it operates machine Y for $s_k - s_{k-1}$ units of time, and again follows π from time $\tau_x(0) + s_k$ onwards. Then, $\hat{\pi}$ is an optimal policy.

Proof: By assumption $t_k < \tau_x(0) \leq t_{k+1}$. Let γ denote the discounted reward obtained until time $t_{k-1} + s_{k-1} - 1$. Then, the discounted rewards obtained by policies π and $\hat{\pi}$ are

$$V(\pi) = \gamma - C\beta^{s_{k-1}+t_{k-1}} + \beta^{s_{k-1}} \sum_{l=t_{k-1}}^{t_k-1} \beta^l X(l) - C\beta^{s_{k-1}+t_k} + \beta^{t_k} \sum_{l=s_{k-1}}^{s_k-1} \beta^l Y(l) - C\beta^{s_k+t_k} + \beta^{s_k} \sum_{l=t_k}^{t_{k+1}-1} \beta^l X(l) + \delta$$

where δ is the reward earned from time s_k+t_{k+1} onwards, and

$$V(\hat{\pi}) \geq \gamma - C\beta^{s_{k-1}+t_{k-1}} + \beta^{s_{k-1}} \sum_{l=t_{k-1}}^{\tau_x(0)-1} \beta^l X(l) - C\beta^{s_{k-1}+\tau_x(0)} + \beta^{\tau_x(0)} \sum_{l=s_{k-1}}^{s_k-1} \beta^l Y(l) - C\beta^{\tau_x(0)+s_k} + \beta^{s_k} \sum_{l=\tau_x(0)}^{t_{k+1}-1} \beta^l X(l) + \delta$$

respectively. The above relation is a strict inequality only when $\tau_x(0) = t_{k+1}$; otherwise it is an equality. From the above two inequalities we get

$$\frac{(V(\hat{\pi}) - V(\pi))(1-\beta)}{(\beta^{s_{k-1}} - \beta^{s_k})(\beta^{t_k} - \beta^{\tau_x(0)})} \geq \frac{\sum_{l=t_k}^{\tau_x(0)-1} \beta^l X(l)}{\sum_{l=t_k}^{\tau_x(0)-1} \beta^l} - \frac{\sum_{l=s_{k-1}}^{s_k-1} \beta^l Y(l) - C\beta^{s_{k-1}} - C\beta^{s_k}}{\sum_{l=s_{k-1}}^{s_k-1} \beta^l}. \quad (2.16)$$

Since it is optimal to operate machine X at time $s_{k-1}+t_{k-1}$ and a switch is incurred at that time, by Lemma 2.4 [part ii)]

$$\frac{\sum_{l=t_{k-1}}^{t_k-1} \beta^l X(l) - C\beta^{t_{k-1}} - C\beta^{t_k}}{\sum_{l=t_{k-1}}^{t_k-1} \beta^l} \geq \frac{\sum_{l=s_{k-1}}^{s_k-1} \beta^l Y(l)}{\sum_{l=s_{k-1}}^{s_k-1} \beta^l}. \quad (2.17)$$

There are two possibilities:

i) A switching cost is incurred by the operation of machine X at $t = 0$. In this case, $\tau_x(0) = \tau_{cx}(0)$. Because it is optimal to operate machine X at $t = 0$, we have from Lemma 2.4 [part i-a) or i-c)] depending on whether or not a switching cost is incurred by the operation of machine Y that

$$\frac{\sum_{l=0}^{t_1-1} \beta^l X(l) - C}{\sum_{l=0}^{t_1-1} \beta^l} \geq \frac{\sum_{l=0}^{s_1-1} \beta^l Y(l) - C}{\sum_{l=0}^{s_1-1} \beta^l}. \quad (2.18)$$

Moreover, since it is optimal to operate machine Y at $t = t_1$, from Lemma 2.4 (part iii)

$$\frac{\sum_{l=0}^{s_1-1} \beta^l Y(l) - C - C\beta^{s_1}}{\sum_{l=0}^{s_1-1} \beta^l} \geq \frac{\sum_{l=t_1}^{t_2-1} \beta^l X(l)}{\sum_{l=t_1}^{t_2-1} \beta^l}. \quad (2.19)$$

By a series of arguments similar to those leading to (2.18) and (2.19) applied at times $t = t_i+s_{i-1}$, $i = 2, \cdots, k-1$, and $t_i + s_i$, $i = 1, 2, \cdots, k-1$, we obtain the series of inequalities

$$\frac{\sum_{l=0}^{t_1-1} \beta^l X(l) - C}{\sum_{l=0}^{t_1-1} \beta^l} \geq \frac{\sum_{l=t_1}^{t_2-1} \beta^l X(l)}{\sum_{l=t_1}^{t_2-1} \beta^l} \geq \cdots \geq \frac{\sum_{l=t_{k-1}}^{t_k-1} \beta^l X(l)}{\sum_{l=t_{k-1}}^{t_k-1} \beta^l}. \quad (2.20)$$

From Lemma 2.2

$$\frac{\sum_{l=t_k}^{\tau_{cx}-1} \beta^l X(l)}{\sum_{l=t_k}^{\tau_{cx}-1} \beta^l} \geq \frac{\sum_{l=0}^{t_1-1} \beta^l X(l) - C}{\sum_{l=0}^{t_1-1} \beta^l}. \quad (2.21)$$

Combining (2.17), (2.20), and (2.21) we obtain

$$\frac{\sum_{l=t_k}^{\tau_{cx}-1} \beta^l X(l)}{\sum_{l=t_k}^{\tau_{cx}-1} \beta^l} \geq \frac{\sum_{l=s_{k-1}}^{s_k-1} \beta^l Y(l)}{\sum_{l=s_{k-1}}^{s_k-1} \beta^l}. \quad (2.22)$$

Because of (2.16) and (2.22), we conclude that $V(\hat{\pi}) \geq V(\pi)$. Since π was assumed to be an optimal policy, $\hat{\pi}$ is also an optimal policy.

ii) No switching cost is incurred by the operation of machine X at time $t = 0$. In this case $\tau_x(0) = \tau_{gx}(0)$ and by arguments similar to those of case i), it can be shown that $\hat{\pi}$ is an optimal policy. □

The preceding lemmas can be used to prove the main result of Section II-B.

Theorem 2.1: An optimal scheduling policy for the deterministic two-armed bandit problem with switching cost has the following property: Decisions about the processor allocation need to be made only at those time instants where the appropriate index (the Gittins index or the switching cost index) of the machine that is operated is achieved.

Remarks:

1) In effect, the theorem states that after a machine is selected for operation, it is operated at least until the time instant that maximizes the discounted reward rate received by the operation.
2) Based on the Gittins index solution, the result of Theorem 2.1 has the following interpretation: In the multi-armed bandit problem with switching cost, under an optimal policy, switches between machines occur only at times during which the Gittins or switching cost index of the currently active machine falls to a value below any it has achieved thus far (cf. Section III-C and [16, Section 4]).

Proof: Without any loss of generality, assume that it is optimal to operate machine X at time $t = 0$ (an argument similar to the one that follows applies if at $t = 0$ it is optimal to operate machine Y). Consider a policy π of the form

$$X(0), X(1) \cdots X(t'_1 - 1), Y(0), \cdots Y(s'_1 - 1), X(t'_1), \cdots X(t'_2 - 1), Y(s'_1), \cdots Y(s'_2 - 1), \cdots, X(t'_{k-1}), \cdots X(t'_k - 1), Y(s'_{k-1}), \cdots Y(s'_k - 1), X(t'_k), \cdots X(t'_{k+1} - 1), Z(1), Z(2), \cdots$$

where $Z(1), Z(2), \cdots$ are an interleaving of reward sequences obtained from machines X and Y from time $t'_{k+1} + s'_k$ onwards. At time 0 there are the following three possibilities:

i) There exists a $k \geq 1$, such that the maximum discounted reward rate for machine X is achieved between time instants $t'_k + s'_k$ and $t'_{k+1} + s'_k - 1$. In this case, repeated application of Lemma 2.5 at times $t'_k + s'_{k-1}$, $t'_{k-1} + s'_{k-2}, \cdots, t'_1$ proves that it is optimal to operate X up until $\tau_{cx}(0) - 1$ (or $\tau_{gx}(0) - 1$); ii) The maximum discounted reward rate for machine X is achieved before t'_1. In this case, we serve X up until $\tau_{cx}(0) - 1$ (or $\tau_{gx}(0) - 1$). At $\tau_{cx}(0)$ [or $\tau_{gx}(0)$] we are faced with the same problem as at time $t = 0$; thus, it is optimal to serve the machine selected at $\tau_{cx}(0)$ (or $\tau_{gx}(0)$) until the maximum discounted reward rate obtained by its operation is achieved; iii) The maximum discounted reward rate for machine X is achieved at $\tau_x(0) = \infty$, and along policy π we have $t_k,\ s_k < \infty, \forall k \geq 1$. In this case, we construct a policy $\tilde{\pi}$ from π as follows: policy $\tilde{\pi}$ follows π until some time $t_k + s_{k-1} - 1$ for some finite k, $k \geq 1$; from time $t_k + s_{k-1}$, $\tilde{\pi}$ continues the operation of machine X until its index is achieved. Denote by $V(\tilde{\pi})$ and $V(\pi)$ the discounted rewards obtained by policies $\tilde{\pi}$ and π, respectively. Let α denote the discounted reward (including switching costs) obtained until time $t_k + s_{k-1} - 1$. Then

$$V(\tilde{\pi}) = \alpha + \beta^{s_{k-1}} \sum_{l=t_k}^{\infty} \beta^l X(l) \tag{2.23}$$

and

$$V(\pi) = \alpha - C\beta^{s_{k-1}+t_k} + \beta^{t_k} \sum_{l=s_{k-1}}^{s_k-1} \beta^l Y(l) - C\beta^{s_k+t_k} + \delta \tag{2.24}$$

where δ denotes the discounted reward obtained from policy π from time $s_k + t_k$ onwards. Using Lemma 2.4 to upper bound the right-hand side of (2.24), we get

$$V(\pi) \leq \alpha + \beta^{t_k} \left[\frac{\sum_{l=s_{k-1}}^{s_k-1} \beta^l Y(l) - C\beta^{s_{k-1}} - C\beta^{s_k}}{\sum_{l=s_{k-1}}^{s_k-1} \beta^l} \right] \cdot \sum_{l=s_{k-1}}^{\infty} \beta^l. \tag{2.25}$$

Subtracting (2.25) from (2.23) yields, after some algebra

$$\frac{(V(\tilde{\pi}) - V(\pi))}{\sum_{l=s_{k-1}+t_k}^{\infty} \beta^l} \geq \frac{\sum_{l=t_k}^{\infty} \beta^l X(l)}{\sum_{l=t_k}^{\infty} \beta^l} - \frac{\sum_{l=s_{k-1}}^{s_k-1} \beta^l Y(l) - C\beta^{s_{k-1}} - C\beta^{s_k}}{\sum_{l=s_{k-1}}^{s_k-1} \beta^l}. \tag{2.26}$$

Arguments similar to those leading from (2.16) to (2.22) in Lemma 2.5 give

$$\text{Right-Hand Side (R.H.S.) of (2.26)} \geq 0.$$

Hence, $V(\tilde{\pi}) \geq V(\pi)$. We can now modify policy $\tilde{\pi}$ according to case i) and show that it is optimal to operate machine X up until $\tau_{cx}(0) - 1$ (or $\tau_{gx}(0) - 1$).

Repetition of the above arguments prove that along an optimal policy, decisions about the processor allocation need to be made only at time instants where the discounted reward rates, resulting from the operation of the machines, are maximized. □

The result of Theorem 2.1 can be intuitively explained as follows: If at a certain instant of time it is optimal to allocate the processor to a certain machine X, then it should be optimal to maximize the reward rate obtained from the operation of X during this allocation. Consequently, X must be operated until the time where the appropriate index is achieved, and at that time the next allocation must be decided.

Theorem 2.1 does not specify how decisions are made at time instants where the discounted reward rates, resulting from the operation of the machines, are maximized. Thus, the determination of optimal decisions remains an open problem. The question that naturally arises is the following: Is the policy that operates the machines according to the highest appropriate index (the Gittins index or the switching cost index) optimal? If such a policy were optimal for the two-armed deterministic bandit problem with switching cost, the solution to this problem would be of the same level of complexity as the solution to the standard deterministic two-armed bandit problem. Unfortunately, the answer to the above question is negative as the following example demonstrates.

Consider two machines X and Y with reward sequences 20, 16, 0, 0, $\cdots$ and 19, 18, 0, 0, $\cdots$, respectively. Let $\beta = 0.5$ and $C = 3$, and assume that at $t = 0$ a switching (setup) cost is always incurred. Then, the policy that operates according to the highest appropriate index plays the machines in the order X, Y, Y, X, and yields a reward of 31.125. The policy that operates the machines in the order Y, Y, X, X, yields a reward of 31.25.

The point illustrated by the above example is in agreement with the result of [29] where it is shown that it is not possible to define indexes which have the property that the resulting index strategy is optimal on the domain of all multi-armed bandit problems with switching cost.

However, the notion of indexes depending only on one machine and the knowledge of the arm that was operated at the previous instant is not entirely useless for the multi-armed bandits with switching costs. Theorem 2.1 demonstrates that along an optimal policy, an arm, once selected, is operated at least until its appropriate index is achieved. In the next subsection we show that if at a particular decision epoch the difference between (appropriate) indexes is greater than some amount, then it is possible to determine the machine that is optimal to operate at that epoch.

C. Computational Considerations

Theorem 2.1 reduces the search for an optimal policy to the determination of the optimal decisions at stopping times that achieve appropriate indexes (the "Gittins index" or the "switching cost index"). Still, the task of finding an optimal policy remains formidable. In this section, we present sufficient conditions under which it is possible to further simplify the search for an optimal strategy. These conditions are described in Lemmas 2.6–2.8 that follow.

Lemma 2.6: Consider a decision instant t along an optimal policy for the two-armed bandit problem with switching cost, and assume that a switching cost C is incurred at t if we decide

to operate either machine.[1] Let τ and s denote the amount of time machines X and Y have been operated, respectively, before time t. Let $\tau_{c1}+\tau$ and $s_{c1}+s$ denote the times that achieve the switching cost indexes $\nu_{cx}(\tau)$ and $\nu_{cy}(s)$ for machines X and Y, respectively. If

$$\frac{\sum_{l=0}^{\tau_{c1}-1}\beta^l X(l+\tau)-C}{\sum_{l=0}^{\tau_{c1}-1}\beta^l}-\frac{\sum_{l=0}^{s_{c1}-1}\beta^l Y(l+s)-C}{\sum_{l=0}^{s_{c1}-1}\beta^l} \geq \frac{C\beta^{\tau_{c1}+s_{c1}}(1-\beta)}{(1-\beta^{s_{c1}})(1-\beta^{\tau_{c1}})} \quad (2.27)$$

then at t it is optimal to operate machine X.

Proof: By Theorem 2.1 it is sufficient to restrict attention to the policies that possibly switch between machines only at stopping times that achieve the appropriate indexes.

Consider a policy that proceeds with machine Y at t, operates Y for $s_{c1}+s_{g2}+\cdots+s_{gk}$, $k=1,2,\cdots$ units of time (where $s+s_{c1}+\cdots+s_{gi}$ achieves $\nu_{gy}(s+s_{c1}+\cdots+s_{g(i-1)})$, $i=2,3,\cdots,k$,) then operates machine X for τ_{c1} units of time and proceeds optimally afterwards. Call this policy $\tilde{\pi}$. Construct the following policy called π'. Policy π' operates machine X at time t for τ_{c1} units of time, then switches to machine Y, operates it for $s_{c1}+s_{g2}+\cdots+s_{gk}$ units of time, and follows policy $\tilde{\pi}$ from time $t+\tau_{c1}+s_{c1}+s_{g2}+\cdots+s_{gk}$ onwards. Denote by $V(\tilde{\pi})$ and $V(\pi')$ the expected discounted rewards obtained by policies $\tilde{\pi}$ and π', respectively. Then

$$V(\tilde{\pi})=\alpha+\beta^t\left[-C+\sum_{l=0}^{s_c'-1}\beta^l Y(l+s) - C\beta^{s_c'}\sum_{l=0}^{\tau_{c1}-1}\beta^l X(l+\tau)\right]+\Lambda \quad (2.28)$$

where α denotes the discounted reward obtained until time t (including switching costs until time t), $s_c'=s_{c1}+s_{g2}+\cdots+s_{gk}$, and Λ is the reward earned from time $t+s_c'+\tau_{c1}$ onwards; and

$$V(\pi')\geq\alpha+\beta^t\left[-C+\sum_{l=0}^{\tau_{c1}-1}\beta^l X(l+\tau)-C\beta^{\tau_{c1}} + \beta^{\tau_{c1}}\sum_{l=0}^{s_c'-1}\beta^l Y(l+s)-C\beta^{s_c'+\tau_{c1}}\right]+\Lambda \quad (2.29)$$

respectively. Subtracting (2.28) from (2.29) yields

$$\frac{(V(\pi')-V(\tilde{\pi}))(1-\beta)}{\beta^t(1-\beta^{s_c'})(1-\beta^{\tau_{c1}})}\geq\frac{\sum_{l=0}^{\tau_{c1}-1}\beta^l X(l+\tau)-C}{\sum_{l=0}^{\tau_{c1}-1}\beta^l} - \frac{\sum_{l=0}^{s_c'-1}\beta^l Y(l+s)-C}{\sum_{l=0}^{s_c'-1}\beta^l}-\frac{C\beta^{s_c'+\tau_{c1}}(1-\beta)}{(1-\beta^{s_c'})(1-\beta^{\tau_{c1}})}. \quad (2.30)$$

By the definition of stopping time s_{c1}, we have

$$\frac{\sum_{l=0}^{s_c'-1}\beta^l Y(l+s)-C}{\sum_{l=0}^{s_c'-1}\beta^l}\leq\frac{\sum_{l=0}^{s_{c1}-1}\beta^l Y(l+s)-C}{\sum_{l=0}^{s_{c1}-1}\beta^l}. \quad (2.31)$$

[1]For the two-armed bandit problem, the situation where a switching cost C is incurred if we decide to operate either machine may arise only at $t=0$. However, the result of Lemma 2.6 is stated for general t because such a situation arises in the N-armed bandit problem ($N>2$) (see [33]) in pairwise comparisons of machines not being operated at time $t-1$.

Since $0<\beta<1$ and $s_c'\geq s_{c1}$, from (2.31) and (2.27) we obtain

$$\text{R.H.S. of (2.30)}\geq 0.$$

Therefore $V(\pi')\geq V(\tilde{\pi})$. Thus, for any policy $\tilde{\pi}$ that proceeds at t with machine Y (and plays it according to Theorem 2.1), it is possible to find another policy π' that proceeds at t with machine X and does better than $\tilde{\pi}$. Consequently, under the assumption given by inequality (2.27), it is optimal to operate machine X at t. □

Lemma 2.7: Consider a decision instant t along an optimal policy for the two-armed bandit problem with switching cost, and assume that at $t-1$ the server operates machine X. Let τ and s denote the amount of time machines X and Y have been operated before time t, respectively. Let $\tau_{g1}+\tau$ and $\tau_{c1}+\tau$ denote the times that achieve the Gittins index $\nu_{gx}(\tau)$ and the switching cost index $\nu_{cx}(\tau)$ for machine X, respectively. Further, let $s_{c1}+s$ denote the time that achieves the switching cost index $\nu_{cy}(s)$ for machine Y. If

$$\frac{\sum_{l=0}^{\tau_{g1}-1}\beta^l X(l+\tau)}{\sum_{l=0}^{\tau_{g1}-1}\beta^l}\geq\frac{\sum_{l=0}^{s_{c1}-1}\beta^l Y(l+s)-C}{\sum_{l=0}^{s_{c1}-1}\beta^l} \quad (2.32)$$

then at t it is optimal to operate machine X.

Proof: By Theorem 2.1, we only need to consider policies along which switches between machines can possibly occur only at stopping times that achieve appropriate indexes. Consider any such policy $\tilde{\pi}$ that proceeds with machine Y at t, operates Y for $s_{c1}+s_{g2}+\cdots+s_{gk}$, ($k=1,2,\cdots$) units of time (where $s+s_{c1}+\cdots+s_i$ achieves $\nu_{gy}(s+s_{c1}+\cdots+s_{g(i-1)})$, $i=2,3,\cdots,k$), then operates machine X for τ_{c1} units of time and proceeds optimally afterwards. Next, construct a modified policy π' that operates machine X at time t for τ_{g1} units of time, switches to machine Y, operates it for $s_{c1}+s_{g2}+\cdots+s_{gk}$ units of time, then switches back to X and operates it for time $\tau_{c1}-\tau_{g1}$ ($\tau_{c1}\geq\tau_{g1}$ by Lemma 2.1) and follows policy $\tilde{\pi}$ from time $t+\tau_{c1}+s_{c1}+s_{g2}+\cdots+s_{gk}$ onwards.

Let α denote the discounted reward (including switching costs) obtained until time t, $s_c':=s_{c1}+s_{g2}+\cdots+s_{gk}$ and Λ be the reward earned from time $t+s_c'+\tau_{c1}$ onwards. Then, the discounted rewards obtained by policies $\tilde{\pi}$ and π' are given by

$$V(\tilde{\pi})=\alpha+\beta^t\left[-C+\sum_{l=0}^{s_c'-1}\beta^l Y(l+s)-C\beta^{s_c'} + \beta^{s_c'}\sum_{l=0}^{\tau_{c1}-1}\beta^l X(l+\tau)\right]+\Lambda \quad (2.33)$$

and

$$V(\pi')\geq\alpha+\beta^t\left[\sum_{l=0}^{\tau_{g1}-1}\beta^l X(l+\tau)-C\beta^{\tau_{g1}} + \beta^{\tau_{g1}}\sum_{l=0}^{s_c'-1}\beta^l Y(l+s)-C\beta^{s_c'+\tau_{g1}} + \beta^{s_c'}\sum_{l=\tau_{g1}}^{\tau_{c1}-1}\beta^l X(l+\tau)\right]+\Lambda \quad (2.34)$$

respectively. The inequality in (2.34) can be explained as follows: Consider the case where $\tau_{c1} = \tau_{g1}$ and at time $t+s'_c+\tau_{c1}$, policy $\tilde{\pi}$ continues optimally by operating machine Y. Then, at time $t + s'_c + \tau_{c1}$ a switching penalty is incurred according to $\tilde{\pi}$, whereas π' does not incur any switching penalty. In this case $V(\pi')$ is greater than the right-hand side of (2.34). In all other cases, (2.34) is valid with equality. From (2.33) and (2.34), we obtain

$$\frac{(V(\pi')-V(\tilde{\pi}))(1-\beta)}{\beta^t(1-\beta^{s'_c})(1-\beta^{\tau_{c1}})} \geq \frac{\sum_{l=0}^{\tau_{g1}-1}\beta^l X(l+\tau)}{\sum_{l=0}^{\tau_{g1}-1}\beta^l} - \frac{\sum_{l=0}^{s'_c-1}\beta^l Y(l+s)-C}{\sum_{l=0}^{s'_c-1}\beta^l} + \frac{C\beta^{s'_c}}{\sum_{l=0}^{s'_c-1}\beta^l}. \quad (2.35)$$

By the definition of stopping time s_{c1} (see 2.31) along with (2.32) and (2.35), we get

$$\text{R.H.S. of (2.35)} \geq 0.$$

Therefore $V(\pi') \geq V(\tilde{\pi})$, and under (2.32) it is optimal to continue operation of machine X at time t. □

Lemma 2.8: Consider a decision instant t along an optimal policy for the two-armed bandit problem with switching cost, and assume that at $t-1$ the server operates machine Y. Let τ and s denote the amount of time machines X and Y have been operated before time t, respectively. Let $\tau_{c1}+\tau$ and $s_{g1}+s$ denote the stopping times that achieve the switching cost index $\nu_{cx}(\tau)$ and the Gittins index $\nu_{gy}(s)$ for machines X and Y, respectively. If

$$\frac{\sum_{l=0}^{\tau_{c1}-1}\beta^l X(l+\tau)-C-C\beta^{\tau_{c1}}}{\sum_{l=0}^{\tau_{c1}-1}\beta^l} - \frac{\sum_{l=0}^{s_{g1}-1}\beta^l Y(l+s)}{\sum_{l=0}^{s_{g1}-1}\beta^l} \geq \frac{2C\beta^{\tau_{c1}+s_{g1}}(1-\beta)}{(1-\beta^{\tau_{c1}})(1-\beta^{s_{g1}})} \quad (2.36)$$

then at t it is optimal to operate machine X.

Proof: Again, by Theorem 2.1, it is sufficient to restrict attention to policies that possibly switch among machines only at stopping times that achieve an appropriate index. Consider any such policy $\tilde{\pi}$ that proceeds with machine Y at t, operates Y for $s_{g1}+s_{g2}+\cdots+s_{gk}$ units of time (where $s+s_{g1}+\cdots+s_{gi}$ achieves $\nu_{gy}(s+s_{g1}+\cdots+s_{g(i-1)})$, $i=2,3,\cdots,k$), then operates machine X for τ_{c1} units of time and proceeds optimally afterwards. Now, modify this policy to obtain policy π', which operates machine X at time t for τ_{c1} units of time, switches to machine Y, and operates it for $s_{g1}+s_{g2}+\cdots+s_{gk}$ units of time and follows policy $\tilde{\pi}$ from time $t+\tau_{c1}+s_{g1}+s_{g2}+\cdots+s_{gk}$ onwards. The policy that starts with machine X at time t and proceeds optimally from $t+\tau_{c1}$ onwards, will do better than the policy π', hence if π' does better than $\tilde{\pi}$, then it is optimal to play machine X at time t.

By calculations similar to those of Lemmas 2.6 and 2.7, we obtain

$$\frac{(V(\pi')-V(\tilde{\pi}))(1-\beta)}{\beta^t(1-\beta^{s'_g})(1-\beta^{\tau_{c1}})} \geq \frac{\sum_{l=0}^{\tau_{c1}-1}\beta^l X(l+\tau)-C-C\beta^{\tau_{c1}}}{\sum_{l=0}^{\tau_{c1}-1}\beta^l} - \frac{\sum_{l=0}^{s'_g-1}\beta^l Y(l+s)}{\sum_{l=0}^{s'_g-1}\beta^l} - \frac{2C\beta^{\tau_{c1}+s'_g}(1-\beta)}{(1-\beta^{\tau_{c1}})(1-\beta^{s'_g})} \quad (2.37)$$

where $s'_g := s_{g1}+s_{g2}+\cdots+s_{gk}$. Further, by the definition of stopping time s_{g1}

$$\frac{\sum_{l=0}^{s'_g-1}\beta^l Y(l+s)}{\sum_{l=0}^{s'_g-1}\beta^l} \leq \frac{\sum_{l=0}^{s_{g1}-1}\beta^l Y(l+s)}{\sum_{l=0}^{s_{g1}-1}\beta^l}. \quad (2.38)$$

Since $0<\beta<1$ and $s'_g \geq s_{g1}$, (2.38) and (2.36) give that

$$\text{R.H.S. of (2.37)} \geq 0.$$

Therefore $V(\pi') \geq V(\tilde{\pi})$. Thus, when (2.36) is true, it is optimal to operate machine X at time t. □

To illustrate how Theorem 2.1 and the above lemmas reduce the complexity of the search for an optimal policy, refer to Figs. 1–3. Originally, without the result of Theorem 2.1, we have to make a decision at every instant of time, and hence we have to search for an optimal path in the dense binary tree of Fig. 1. Because of Theorem 2.1, we only have to find an optimal path in the tree of Fig. 2; this tree may be considerably sparser than that of Fig. 1. We may be able to further prune the tree of Fig. 2 by using the results of Lemmas 2.6–2.8 (see, for example, Fig. 3, where condition (2.36) is satisfied at n_b). Thus, we may eventually be able to significantly reduce the computation required to determine an optimal allocation strategy. Nevertheless, determination of an optimal strategy still remains a difficult and challenging problem.

The sufficient conditions provided by Lemmas 2.6–2.8 are simple. At any decision epoch, one has to compute the appropriate index and corresponding stopping time of each machine and determine whether the indexes satisfy a condition such as (2.27), (2.32), or (2.36) (depending on the situation). Thus, the search for an optimal strategy may be simplified by just "looking one branch ahead" (i.e., looking, at any decision instant t, as far ahead as the stopping time that achieves the next appropriate index) in the tree of Fig. 2.

If the sufficient conditions of Lemmas 2.6–2.8 are not satisfied, then one has to look further into the future to simplify the search for an optimal strategy. Lemmas 2.9–2.11 that follow present sufficient conditions for optimality by "looking two branches ahead" in the tree of Fig. 3. The general idea behind these lemmas is the following. Consider the tree of Fig. 2, move forward in the tree, and prune this tree using the results of Lemmas 2.6–2.8 to obtain the tree of Fig. 3. Suppose that: i) at a certain node n_a of the tree of Fig. 3 the conditions of Lemmas 2.6–2.8 are not satisfied; ii) at the next node n_b, after we operate machine $Y(X)$ the conditions of Lemmas 2.6–2.8 are satisfied, and it is optimal to operate machine $X(Y)$ at n_b. In this case, by "looking two branches ahead" in the tree of Fig. 3, it may be possible to obtain sufficient conditions that are satisfied at node n_a and allow further pruning of the three of Fig. 3 at node n_a. These sufficient conditions are stated in Lemmas 2.9–2.11 and are tighter than the sufficient conditions of Lemmas 2.6–2.8. Thus, the combination of Lemmas 2.6–2.11 allows us to obtain conditions that simplify the search for an optimal scheduling policy by first employing a "one branch look-ahead" technique and following it by a "two branch look-ahead" method as described above. It is possible to extend the philosophy of Lemmas 2.6–2.11 to further simplify the search for an optimal

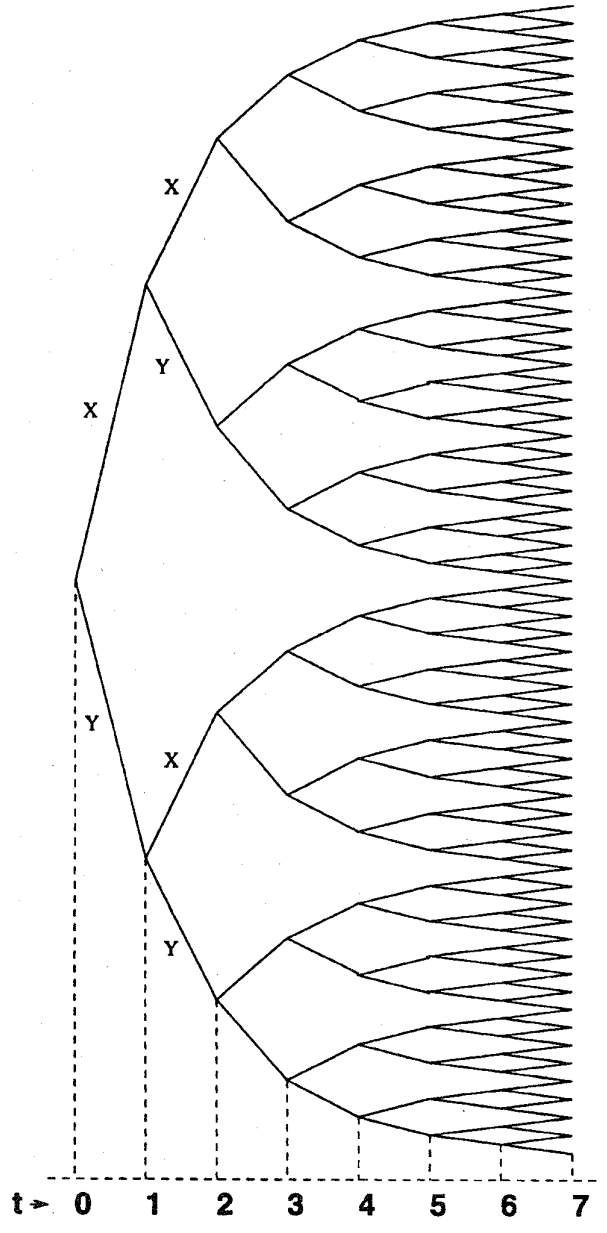

Fig. 1. Original decision tree.

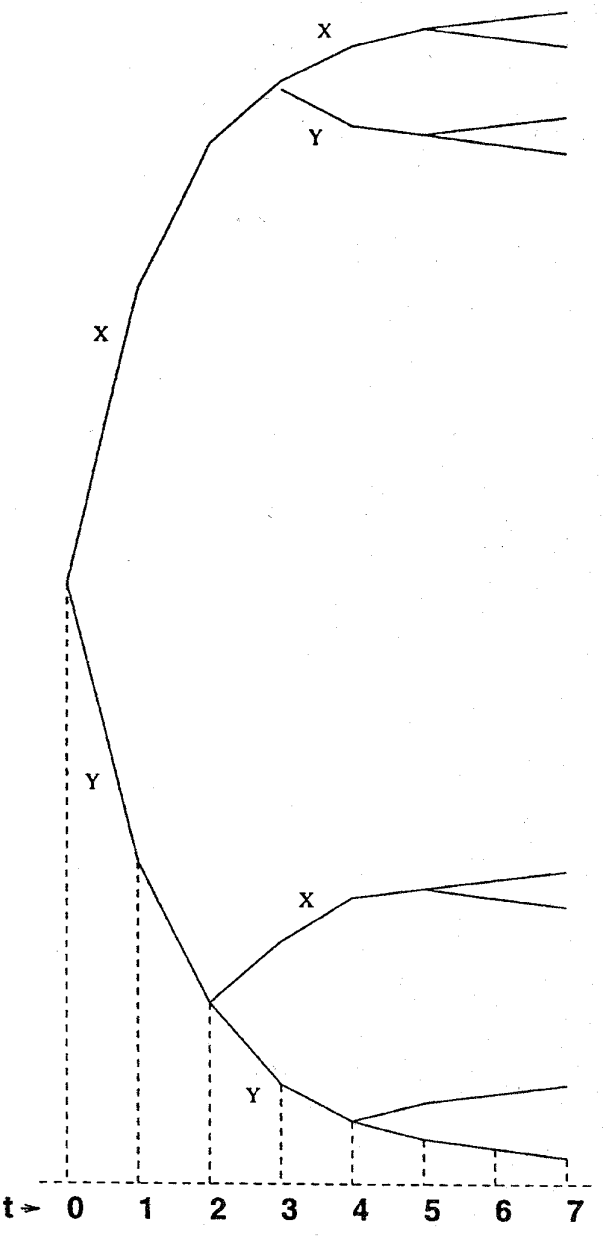

Fig. 2. Decision tree after applying Theorem 2.1 ($\tau_{cx}(0) = 3$, $\tau_{cy}(0) = 2$, $\tau_{gx}(3) = \tau_{cx}(3) = 5$, $\tau_{gy}(2) = 4$, $\tau_{cy}(2) = 5$, $\tau_{gx}(5) = 7$, $\tau_{gy}(4) = 8$).

policy by employing an "n branch look-ahead" ($n > 2$) technique.

Lemma 2.9: Consider a decision instant t along an optimal policy for the two-armed bandit problem with switching cost, and assume that a switching cost C is incurred at t if we decide to operate either machine. Let τ and s denote the amount of time machines X and Y have been operated, respectively, before time t. Let $\tau_{c1} + \tau$ and $s_{c1} + s$ denote the stopping times that achieve the switching cost indexes $\nu_{cx}(\tau)$ and $\nu_{cy}(s)$ for machines X and Y, respectively. Further, let $\tau_{g2} + \tau_{c1} + \tau$ denote the stopping time that achieves the Gittins index $\nu_{gx}(\tau + \tau_{c1})$ for machine X.

Suppose that:

i) The condition of Lemma 2.6 is not satisfied at time t for either machine;

ii) If we operate machine Y at time t, then at time $t + s_{c1}$ it is optimal to operate machine X;

iii)

$$\frac{\sum_{l=0}^{\tau_{c1}-1} \beta^l X(l+\tau) - C + \frac{C\beta^{\tau_{c1}+s_{c1}}}{(1-\beta^{s_{c1}})}}{\sum_{l=0}^{\tau_{c1}-1} \beta^l} \geq \frac{\sum_{l=0}^{s_{c1}-1} \beta^l Y(l+s) - C}{\sum_{l=0}^{s_{c1}-1} \beta^l} \quad (2.39)$$

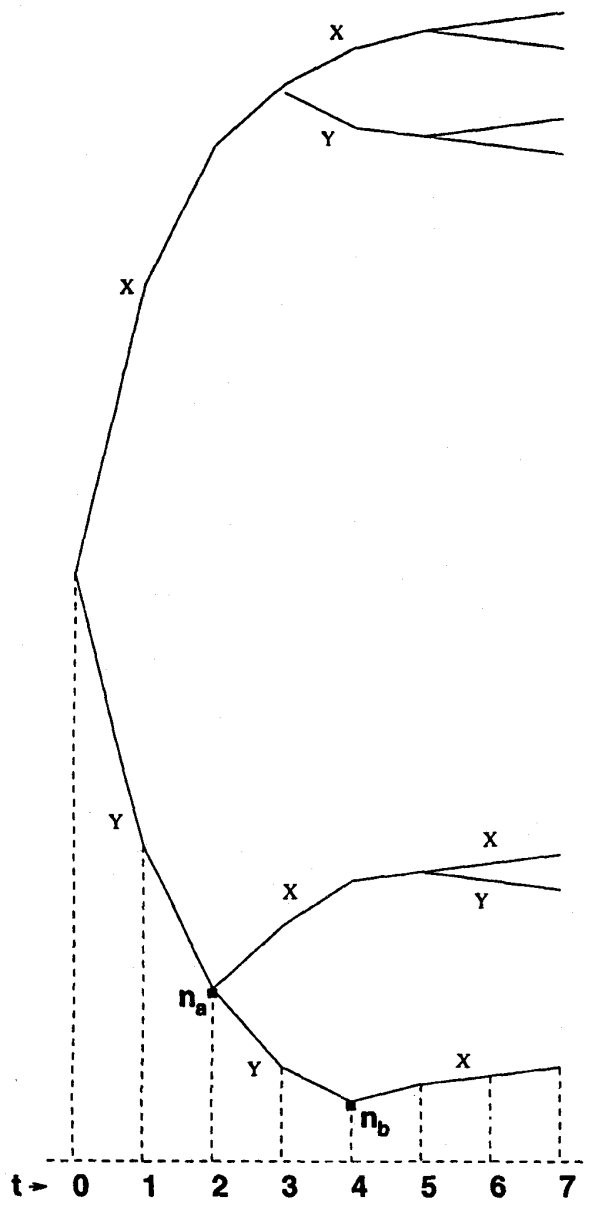

Fig. 3. Decision tree after applying Lemmas 2.6–2.8 to the tree of Fig. 2.

iv)

$$\frac{\sum_{l=0}^{\tau_{c1}+\tau_{g2}-1}\beta^l X(l+\tau)-C}{\sum_{l=0}^{\tau_{c1}+\tau_{g2}-1}\beta^l}-\frac{\sum_{l=0}^{s_{c1}-1}\beta^l Y(l+s)-C}{\sum_{l=0}^{s_{c1}-1}\beta^l}\geq\frac{C\beta^{\tau_{c1}+\tau_{g2}+s_{c1}}(1-\beta)}{(1-\beta^{s_{c1}})(1-\beta^{\tau_{c1}+\tau_{g2}})}. \quad (2.40)$$

Then, it is optimal to operate machine X at time t.

Proof: We can restrict attention to policies that possibly switch between machines only at stopping times that achieve the appropriate index. Let $s_{c2}+s_{c1}+s$ denote the stopping time that achieves the switching cost index $\nu_{cy}(s+s_{c1})$ for machine Y. Because of condition ii), for any policy that proceeds with the operation of machine Y at time t and is a candidate for being optimal, there are only the following two possibilities:

a) Operate machine Y for s_{c1} units of time, then operate machine X for τ_{c1} units of time, switch back to machine Y and operate it for s_{c2} units of time, and then proceed optimally from time $t+s_{c1}+s_{c2}+\tau_{c1}$ onwards. Call this policy $\tilde{\pi}_{yxy}$.

b) Operate machine Y for s_{c1} units of time, then operate machine X for $\tau_{c1}+\tau_{g2}$ units of time and proceed optimally from time $t+s_{c1}+\tau_{c1}+\tau_{g2}$ onwards. Call this policy $\tilde{\pi}_{yxx}$.

Now construct the following two policies called π_{xyy} and π_{xxy}:

a') Policy π_{xyy} operates machine X for τ_{c1} units of time, switches to machine Y, operates it for $s_{c1}+s_{c2}$ units of time, and follows policy $\tilde{\pi}_{yxy}$ from time $t+s_{c1}+s_{c2}+\tau_{c1}$ onwards.

b') Policy π_{xxy} operates machine X for $\tau_{c1}+\tau_{g2}$ units of time, switches to machine Y, operates it for s_{c1} units of time and follows policy $\tilde{\pi}_{yxx}$ from time $t+s_{c1}+\tau_{c1}+\tau_{g2}$ onwards.

Then

$$\begin{aligned}&\frac{(V(\pi_{xyy})-V(\tilde{\pi}_{yxy}))(1-\beta)}{\beta^t(1-\beta^{s_{c1}})(1-\beta^{\tau_{c1}})}\\&=\frac{\sum_{l=0}^{\tau_{c1}-1}\beta^l X(l+\tau)-C}{\sum_{l=0}^{\tau_{c1}-1}\beta^l}\\&\quad-\frac{\sum_{l=0}^{s_{c1}-1}\beta^l Y(l+s)-C}{\sum_{l=0}^{s_{c1}-1}\beta^l}\\&\quad+\frac{C\beta^{s_{c1}+\tau_{c1}}(1-\beta)}{(1-\beta^{s_{c1}})(1-\beta^{\tau_{c1}})}.\end{aligned} \quad (2.41)$$

From condition iii), (2.39), we have

$$\text{R.H.S. of (2.41)} \geq 0.$$

Hence, $V(\pi_{xyy})\geq V(\tilde{\pi}_{yxy})$. Furthermore

$$\begin{aligned}&\frac{(V(\pi_{xxy})-V(\tilde{\pi}_{yxx}))(1-\beta)}{\beta^t(1-\beta^{s_{c1}})(1-\beta^{\tau_{c1}+\tau_{g2}})}\\&\geq\frac{\sum_{l=0}^{\tau_{c1}+\tau_{g2}-1}\beta^l X(l+\tau)-C}{\sum_{l=0}^{\tau_{c1}+\tau_{g2}-1}\beta^l}\\&\quad-\frac{\sum_{l=0}^{s_{c1}-1}\beta^l Y(l+s)-C}{\sum_{l=0}^{s_{c1}-1}\beta^l}\\&\quad-\frac{C\beta^{\tau_{c1}+\tau_{g2}+s_{c1}}(1-\beta)}{(1-\beta^{s_{c1}})(1-\beta^{\tau_{c1}+\tau_{g2}})}\end{aligned} \quad (2.42)$$

and using condition iv), (2.40), we get

$$\text{R.H.S. of (2.42)} \geq 0.$$

Hence, $V(\pi_{xxy})\geq V(\tilde{\pi}_{yxx})$.

Therefore, under conditions i)–iv) of the lemma, for any policy $\tilde{\pi}$ that proceeds at t with the operation of machine Y, it is possible to find another policy π that proceeds at t with machine X and does better than $\tilde{\pi}$. Consequently, it is optimal to operate machine X at time t. □

Remark: If at t it is known that policy $\tilde{\pi}_{yxy}$ is better than policy $\tilde{\pi}_{yxx}$, then (2.39) along with conditions i) and ii) of Lemma 2.9 are sufficient to guarantee that it is optimal to operate machine X at time t. This shows how knowledge of future optimal decisions along a subtree can relax the sufficient conditions that determine the optimal decision at t. Such knowledge of future optimal decisions along a subtree may be possible at any time t, if one initially uses Lemmas 2.6–2.8 to prune the tree of Fig. 2.

Lemma 2.10: Consider a decision instant t along an optimal policy for the two-armed bandit problem with switching cost, and assume that at $t-1$ the server operates machine X. Let τ and s denote the amount of time machines X and Y have been operated before time t, respectively. Let $\tau_{g1}+\tau$ and $\tau_{c1}+\tau$ denote the stopping times that achieve the Gittins index $\nu_{gx}(\tau)$ and the switching cost index $\nu_{cx}(\tau)$ for machine X, respectively. Further, let $s_{c1}+s$ denote the stopping time that achieves the switching cost index $\nu_{cy}(s)$ for machine Y. Assume the following at t:

i) The conditions of Lemmas 2.7 and 2.8 are not satisfied at time t for either machine.

ii) If we operate machine Y at time t, then at time $t+s_{c1}$, it is optimal to operate machine X.

iii)

$$\frac{\sum_{l=0}^{\tau_{g1}-1}\beta^l X(l+\tau)}{\sum_{l=0}^{\tau_{g1}-1}\beta^l} \geq \frac{\sum_{l=0}^{s_{c1}-1}\beta^l Y(l+s)-C-C\beta^{s_{c1}}}{\sum_{l=0}^{s_{c1}-1}\beta^l}. \tag{2.43}$$

Then it is optimal to continue operation of machine X at time t.

Proof: We consider only those policies that possibly switch between machines only at stopping times that achieve appropriate indexes.

Because of condition ii) of the lemma, any policy that starts with machine Y at time t and is a candidate for being an optimal policy must operate machine Y for s_{c1} units of time, must switch to machine X at time $t+s_{c1}$, operate it for τ_{c1} units of time, and proceed optimally from time $t+s_{c1}+\tau_{c1}$. Call this policy $\tilde{\pi}_{yx}$. Now construct the following policy called π_{xy}. Policy π_{xy} operates machine X at time t for τ_{g1} units of time, then switches to machine Y, operates it for s_{c1} units of time, then switches back to machine X,, and operates it for time $\tau_{c1}-\tau_{g1}(\tau_{c1}\geq\tau_{g1}$ by Lemma 2.1) and follows policy $\tilde{\pi}_{yx}$ from time $t+\tau_{c1}+s_{c1}$ onwards. By arguments similar to those of Lemma 2.7, we get

$$\frac{(V(\pi_{xy})-V(\tilde{\pi}_{yx}))(1-\beta)}{\beta^t(1-\beta^{s_{c1}})(1-\beta^{\tau_{g1}})} \geq \frac{\sum_{l=0}^{\tau_{g1}-1}\beta^l X(l+\tau)}{\sum_{l=0}^{\tau_{g1}-1}\beta^l} - \frac{\sum_{l=0}^{s_{c1}-1}\beta^l Y(l+s)-C-C\beta^{s_{c1}}}{\sum_{l=0}^{s_{c1}-1}\beta^l}. \tag{2.44}$$

By (2.43), we obtain

$$\text{R.H.S. (2.44)} \geq 0.$$

Therefore $V(\pi_{xy})\geq V(\tilde{\pi}_{yx})$. Thus, under conditions i)–iii), it is optimal to operate machine X at t. □

Lemma 2.11: Consider a decision instant t along an optimal policy for the two-armed bandit problem with switching cost, and assume that at $t-1$ the server operates machine Y. Let τ and s denote the amount of time machines X and Y have been operated before time t, respectively. Let $\tau_{c1}+t$ and $\tau_{g2}+\tau_{c1}+t$ denote the stopping times that achieve the switching cost index $\nu_{cx}(\tau)$ and the Gittins index $\nu_{gx}(\tau+\tau_{c1})$ for machine X, respectively. Further, let $s_{g1}+s$ denote the stopping time that achieves the Gittins index $\nu_{gy}(s)$ for machine Y.

Suppose that:

i) The conditions of Lemmas 2.7 and 2.8 are not satisfied at time t for either machine.

ii) If we operate machine Y at t, then at time $t+s_{g1}$ it is optimal to operate machine X.

iii)

$$\frac{\sum_{l=0}^{\tau_{c1}-1}\beta^l X(l+\tau)-C-C\beta^{\tau_{c1}}}{\sum_{l=0}^{\tau_{c1}-1}\beta^l} \geq \frac{\sum_{l=0}^{s_{g1}-1}\beta^l Y(l+s)}{\sum_{l=0}^{s_{g1}-1}\beta^l}. \tag{2.45}$$

iv)

$$\frac{\sum_{l=0}^{\tau_{c1}+\tau_{g2}-1}\beta^l X(l+\tau)-C-C\beta^{\tau_{c1}+\tau_{g2}}}{\sum_{l=0}^{\tau_{c1}+\tau_{g2}-1}\beta^l} - \frac{\sum_{l=0}^{s_{g1}-1}\beta^l Y(l+s)}{\sum_{l=0}^{s_{g1}-1}\beta^l} \geq \frac{2C\beta^{\tau_{c1}+\tau_{g2}+s_{g1}}(1-\beta)}{(1-\beta^{\tau_{c1}+\tau_{g2}})(1-\beta^{s_{g1}})}. \tag{2.46}$$

Then, it is optimal to operate machine X at time t.

Proof: Again by Theorem 2.1, it is enough to consider only those policies which possibly switch between machines only at stopping times that achieve the appropriate indexes. Let $s_{c2}+s_{g1}+s$ denote the stopping time that achieves the switching cost index $\nu_{cy}(s+s_{g1})$ for machine Y. Because of condition ii), for any policy that proceeds with the operation of machine Y at t and is a candidate for being optimal, there are only the following two possibilities:

a) Operate machine Y for s_{g1} units of time, then operate machine X for τ_{c1} units of time, switch back to machine Y and operate it for s_{c2} units of time, and then proceed optimally from time $t+s_{g1}+s_{c2}+\tau_{c1}$ onwards. Call this policy $\tilde{\pi}_{yxy}$.

b) Operate machine Y for s_{g1} units of time, then operate machine X for $\tau_{c1}+\tau_{g2}$ units of time and proceed optimally from time $t+s_{g1}+\tau_{c1}+\tau_{g2}$ onwards. Call this policy $\tilde{\pi}_{yxx}$.

Now construct the following two policies called π_{xyy} and π_{xxy}:

a') Policy π_{xyy} operates machine X for τ_{c1} units of time, then switches to machine Y, operates it for $s_{g1}+s_{c2}$ units of time, and follows policy $\tilde{\pi}_{yxy}$ from time $t+s_{g1}+s_{c2}+\tau_{c1}$ onwards.

b') Policy π_{xxy} operates machine X for $\tau_{c1}+\tau_{g2}$ units of time, switches to machine Y and operates it for s_{g1} units of time, and follows policy $\tilde{\pi}_{yxx}$ from time $t+s_{g1}+\tau_{c1}+\tau_{g2}$ onwards.

By calculations similar to those of Lemma 2.9, we can show that under condition iii), $V(\pi_{xyy}) \geq V(\tilde{\pi}_{yxy})$ and under condition iv) $V(\pi_{xxy}) \geq V(\tilde{\pi}_{yxx})$. Therefore, under the conditions of the lemma, it is optimal to operate machine X at time t. □

The results of Lemmas 2.6–2.11 can be used to do pairwise comparisons of machines and simplify the search for optimal policies in the deterministic N-armed ($N > 2$) bandit problem.

III. The Stochastic Multi-Armed Bandit Problem with Switching Cost

A. Problem Formulation

In the stochastic multi-armed bandit problem, there are N machines $X^1, X^2, \cdots, X^N$ and one server. Machines are characterized by the pair of sequences $\{X^i(s), F^i(s)\}$, $s = 0, 1, 2, \cdots$, where $X^i(s)$ is the (random) reward obtained when X^i is operated for the $(s+1)$th time and $F^i(s)$ is the σ-field representing the information about machine X^i gathered after it has been operated s times. Let $F^i = \vee_s F^i(s)$, where $i \in 1, \cdots, N$. We make the following assumptions:

A1) $F^i(s) \subset F^i(s+1)$.

A2) $\vee_s \sigma(X^i(s)) \vee F^i$, $i = 1, 2, \cdots N$, are independent.

A3) $E\sum_{t=0}^{\infty} \beta^t |X^i(t)| < \infty$; where $1 \leq i \leq N$, and $0 < \beta < 1$ is a fixed discount factor.

At each time instant t, exactly one machine must be operated. Thus, $t = t^1 + t^2 + \cdots + t^N$, where $t^i := t^i(t)$ is the number of times machine X^i is operated during $0, 1, 2, \cdots, t-1$. t^i is called the X^ith machine time at process time t. Let $m(t)$ denote the machine operated at time t. If $m(t) \neq m(t-1)$, $t = 1, 2, \cdots$, then a switching cost C is incurred at time t; this cost may or may not incur at $t = 0$.

Consider the decision at time t, $t = 0, 1, \cdots$. This decision must be based on the available information which is $F(t) = \vee_{i=1}^N F^i(t^i)$, and $m(t-1)$. $F(t) = F(t-1) \vee G(t-1)$ where $G(t-1)$ is the σ-field generated by the sets of the form $\{m(t-1) = i\} \cap \{t^i(t-1) = s\} \cap A$, with $A \in F^i(s+1)$. If $m(t) = X^i$, then the states x_j of machines X^j, $j \neq i$, remain frozen, and the state x_i of machine X^i changes according to the transition rule $P(x_i(t^i(t)+1) \in \tilde{A} \mid x_i(0), \cdots x_i(t^i(t)))$, $\tilde{A} \in F^i(t^i(t)+1)$. A policy is any sequence of decisions $\{u(t) : u(t) = m(t), t = 0, 1, 2, \cdots\}$, where $u(t)$ is based only on $F(t)$ and $m(t-1)$, and $F(t)$ evolves according to the mechanism described above.

The bandit problem with switching cost is to find a policy π that maximizes

$$V(\pi) := E\left\{\sum_{t=0}^{\infty} \beta^t [X^{m(t)}(t^{m(t)}(t)) - 1(m(t) \neq m(t-1))C] \,\middle|\, F(0)\right\}. \quad (3.1)$$

B. Analysis

To proceed with the analysis, for each t, $t = 0, 1, \cdots$ and $1 \leq i \leq N$, we define $\hat{F}^i(s) := \bigvee_{j=1, j\neq i}^N F^j(t^j)$, where $s := \sum_{j=1, j\neq i}^N t^j$, and we denote by $T^i(t)$ the set of all stopping times τ, $t < \tau \leq \infty$, of $\{F^i(\cdot)\}$. For each machine X^i, after it has been operated t times we define the following quantities: the Gittins index [13]

$$\nu_{gi}(t) := \max_{\tau > t} \frac{E\left\{\sum_{l=t}^{\tau-1} \beta^l X^i(l) \mid F^i(t)\right\}}{E\left\{\sum_{l=t}^{\tau-1} \beta^l \mid F^i(t)\right\}} \quad (3.2)$$

and the "switching cost index"

$$\nu_{ci}(t) := \max_{\tau > t} \frac{E\left\{\sum_{l=t}^{\tau-1} \beta^l X^i(l) - C\beta^t \mid F^i(t)\right\}}{E\left\{\sum_{l=t}^{\tau-1} \beta^l \mid F^i(t)\right\}} \quad (3.3)$$

where the maximization is over all stopping times $\tau \in T^i(t)$, and "max" in (3.2) and (3.3) as well as in the sequel is to be interpreted as "ess sup." Under Assumptions A1)–A3), made in the problem formulation, there always exist stopping times τ and τ' achieving the maximum in (3.2) and (3.3), respectively (see [34]).

Define by $\tau_{gi}(t) \in T^i(t)$ and $\tau_{ci}(t) \in T^i(t)$ the stopping times that achieve $\nu_{gi}(t)$ and $\nu_{ci}(t)$, respectively. Then, the indexes $\nu_{ci}(t)$, $\nu_{gi}(t)$ and the stopping times $\tau_{ci}(t)$ and $\tau_{gi}(t)$ as defined above satisfy the following.

Lemma 3.1: For any stopping times $\sigma \in T^i(t)$

$$\frac{E\left\{1(\sigma < \tau_{gi}(t)) \sum_{l=\sigma}^{\tau_{gi}(t)-1} \beta^l X(l) \mid F^i(t)\right\}}{E\left\{1(\sigma < \tau_{gi}(t)) \sum_{l=\sigma}^{\tau_{gi}(t)-1} \beta^l \mid F^i(t)\right\}} \geq \nu_{gi}(t) \geq \frac{E\left\{\sum_{l=t}^{t_1-1} \beta^l X(l) \mid F^i(t)\right\}}{E\left\{\sum_{l=t}^{t_1-1} \beta^l \mid F^i(t)\right\}} \quad \text{a.s. } \forall t_1 > t \quad (3.4)$$

and

$$\frac{E\left\{1(\sigma < \tau_{ci}(t)) \sum_{l=\sigma}^{\tau_{ci}(t)-1} \beta^l X(l) \mid F^i(t)\right\}}{E\left\{1(\sigma < \tau_{ci}(t)) \sum_{l=\sigma}^{\tau_{ci}(t)-1} \beta^l \mid F^i(t)\right\}} \geq \nu_{ci}(t) \geq \frac{E\left\{\sum_{l=t}^{t_1-1} \beta^l X(l) - C\beta^t \mid F^i(t)\right\}}{E\left\{\sum_{l=t}^{t_1-1} \beta^l \mid F^i(t)\right\}} \quad \text{a.s. } \forall t_1 > t. \quad (3.5)$$

We do not prove this lemma, as part i) is a special case of [16, Lemma 2.1(b)], and part ii) follows from arguments similar to part i). The stopping times $\tau_{gi}(t)$ and $\tau_{ci}(t)$ are related as follows.

Lemma 3.2: For all t, $\tau_{ci}(t) \geq \tau_{gi}(t)$ a.s. $[F^i(t)]$.

Proof: Suppose the statement of the lemma is not true. Then there exists a t, such that $\tau_{ci}(t) < \tau_{gi}(t)$ on a set $\mathcal{A}$ s.t. $P(\mathcal{A} \mid F^i(t)) > 0$ a.s. We will show that in this case $\tau_{ci}(t)$ does not achieve the maximum in the right-hand side of (3.3), thus reaching a contradiction.

Define $\tilde{\tau}_{ci}(t)$ as follows:

$$\tilde{\tau}_{ci}(t) = \begin{cases} \tau_{ci}(t) & \text{on } \mathcal{A}^c \\ \tau_{gi}(t) & \text{on } \mathcal{A}. \end{cases} \quad (3.6)$$

Then $\tilde{\tau}_{ci}(t) \in T^i(t)$, $\tilde{\tau}_{ci}(t) \geq \tau_{ci}(t)$ a.s. and $\tilde{\tau}_{ci}(t) > \tau_{ci}(t)$ on $\mathcal{A}$. To complete the proof of the lemma we only need to show that

$$\frac{E\left\{\sum_{l=t}^{\tilde{\tau}_{ci}(t)-1} \beta^l X^i(l) - C\beta^t \mid F^i(t)\right\}}{E\left\{\sum_{l=t}^{\tilde{\tau}_{ci}(t)-1} \beta^l \mid F^i(t)\right\}} > \frac{E\left\{\sum_{l=t}^{\tau_{ci}(t)-1} \beta^l X^i(l) - C\beta^t \mid F^i(t)\right\}}{E\left\{\sum_{l=t}^{\tau_{ci}(t)-1} \beta^l \mid F^i(t)\right\}} \quad \text{a.s.} \quad (3.7)$$

By the definition of $\tilde{\tau}_{ci}(t)$, $\nu_{gi}(t)$ and Lemma 3.1-i) we have

$$\frac{E\left\{1_{\mathcal{A}}\sum_{l=\tau_{ci}(t)}^{\tau_{gi}(t)-1} \beta^l X^i(l) \mid F^i(t)\right\}}{E\left\{1_{\mathcal{A}}\sum_{l=\tau_{ci}(t)}^{\tau_{gi}(t)-1} \beta^l \mid F^i(t)\right\}} \geq \nu_{gi}(t) \geq \frac{E\left\{\sum_{l=t}^{\tau_{ci}(t)-1} \beta^l X^i(l) \mid F^i(t)\right\}}{E\left\{\sum_{l=t}^{\tau_{ci}(t)-1} \beta^l \mid F^i(t)\right\}} \quad \text{a.s.} \quad (3.8)$$

Combining the two inequalities in (3.8) we obtain

$$E\left\{1_{\mathcal{A}} \sum_{l=\tau_{ci}(t)}^{\tau_{gi}(t)-1} \beta^l X^i(l) \,\middle|\, F^i(t)\right\} E\left\{\sum_{l=t}^{\tau_{ci}(t)-1} \beta^l \,\middle|\, F^i(t)\right\} \geq E\left\{\sum_{l=t}^{\tau_{ci}(t)-1} \beta^l X^i(l) \,\middle|\, F^i(t)\right\} \cdot E\left\{1_{\mathcal{A}} \sum_{l=\tau_{ci}(t)}^{\tau_{gi}(t)-1} \beta^l \,\middle|\, F^i(t)\right\} \quad \text{a.s.} \quad (3.9)$$

or equivalently

$$\frac{E\left\{\sum_{l=t}^{\tilde{\tau}_{ci}(t)-1} \beta^l X^i(l) \mid F^i(t)\right\}}{E\left\{\sum_{l=t}^{\tilde{\tau}_{ci}(t)-1} \beta^l \mid F^i(t)\right\}} \geq \frac{E\left\{\sum_{l=t}^{\tau_{ci}(t)-1} \beta^l X^i(l) \mid F^i(t)\right\}}{E\left\{\sum_{l=t}^{\tau_{ci}(t)-1} \beta^l \mid F^i(t)\right\}} \quad \text{a.s.} \quad (3.10)$$

Since $\tilde{\tau}_{ci}(t) \geq \tau_{ci}(t)$, a.s. $\tilde{\tau}_{ci}(t) > \tau_{ci}(t)$ on $\mathcal{A}$, $P(\mathcal{A} \mid F^i(t)) > 0$, $\beta > 0$, and $C > 0$, it follows that

$$\frac{C\beta^t}{E\left\{\sum_{l=t}^{\tilde{\tau}_{ci}(t)-1} \beta^l \mid F^i(t)\right\}} < \frac{C\beta^t}{E\left\{\sum_{l=t}^{\tau_{ci}(t)-1} \beta^l \mid F^i(t)\right\}} \quad \text{a.s.} \quad (3.11)$$

Combining (3.10) and (3.11) we obtain (3.7). Q.E.D.

To determine the qualitative properties of an optimal policy we proceed, as in Section II, via a series of lemmas. Suppose that a policy π which plays the machines in the order

$$\begin{aligned} &X^1(0,\omega), X^1(1,\omega)\cdots X^1(t_1(\omega)-1,\omega), Z(0,\omega),\cdots \\ &\quad Z(s_1(\omega)-1,\omega), X^1(t_1(\omega),\omega),\cdots, \\ &\quad X^1(t_2(\omega)-1,\omega), Z(s_1(\omega),\omega),\cdots, \\ &\quad Z(s_2(\omega)-1,\omega),\cdots, \\ &\quad X^1(t_{k(\omega)-1,(\omega)},\omega),\cdots,X^1(t_{k(\omega)}(\omega)-1,\omega), \\ &\quad Z(s_{k(\omega)-1}(\omega),\omega),\cdots, \\ &\quad Z(s_{k(\omega)}(\omega)-1,\omega), X^1(t_{k(\omega)}(\omega),\omega),\cdots, \\ &\quad X^1(\tau_1(0,\omega)-1,\omega),\cdots,X^1(t_{k(\omega)+1}(\omega)-1,\omega), \\ &\quad Y(t^y(t_{k(\omega)+1}(\omega)+s_{k(\omega)}(\omega)),\omega), \\ &\quad Y(t^y(t_{k(\omega)+1}(\omega)+s_{k(\omega)}(\omega)+1),\omega),\cdots \end{aligned}$$

for sample path ω, is optimal, where (by some abuse of notation) $Z(0,\omega)$, $Z(1,\omega)$, $Z(2,\omega),\cdots,Z(s_i(\omega),\omega),\cdots$, $Z(s_{k(\omega)}(\omega)-1,\omega)$ are an interleaving of reward sequences from machines X^2, $X^3,\cdots,X^N$ (or a subset of it) and $Y(t^y(t_{k(\omega)+1}(\omega)+s_{k(\omega)}(\omega)),\omega)$, $Y(t^y(t_{k(\omega)+1}(\omega)+s_{k(\omega)}(\omega)+1),\omega),\cdots$ represents an interleaving of reward sequences from machines X^1, $X^2,\cdots,X^N$ from time $t_{k(\omega)+1}(\omega)+s_{k(\omega)}(\omega)$ onwards. Suppose the time instants $t_{k(\omega)}$ and $t_{k(\omega)+1}$ are such that $t_{k(\omega)} < \tau_1(0,\omega) \leq t_{k(\omega)+1}$, where $\tau_1(0,\omega) = \tau_{c1}(0,\omega)$ or $\tau_{g1}(0,\omega)$, depending on whether or not a switching cost is incurred at time $t=0$.

Lemma 3.3: Denote by A_i, $i=1,2,\cdots,k$, the discounted reward obtained in the time interval $[t_i+s_{i-1}, t_i+s_i-1]$, $(s_0:=0)$ excluding the switching cost incurred at t_i+s_{i-1}.

i) a) If at $t=0$ a switching cost is always incurred or never incurred, then we must have

$$\frac{E\left\{\sum_{l=0}^{t_1-1} \beta^l X^1(l) - C \mid F^1(0)\right\}}{E\left\{\sum_{l=0}^{t_1-1} \beta^l \mid F^1(0)\right\}} \geq \frac{E\{A_1 - C \mid \hat{F}^1(0)\}}{E\left\{\sum_{l=0}^{s_1-1} \beta^l \mid \hat{F}^1(0)\right\}} \quad \text{a.s.} \quad (3.12)$$

b) If at $t=0$ a switching cost is incurred when machines X^2 or X^3 or $\cdots X^N$ are operated, but no switching cost is incurred when machine X^1 is operated, then

$$\frac{E\left\{\sum_{l=0}^{t_1-1} \beta^l X^1(l) \mid F^1(0)\right\}}{E\left\{\sum_{l=0}^{t_1-1} \beta^l \mid F^1(0)\right\}} \geq \frac{E\{A_1 - C - C\beta^{s_1} \mid \hat{F}^1(0)\}}{E\left\{\sum_{l=0}^{s_1-1} \beta^l \mid \hat{F}^1(0)\right\}} \quad \text{a.s.} \quad (3.13)$$

c) Let X^j be the machine operated at time t_1 according to policy π, and suppose that no switching cost would be incurred at $t=0$ if machine X^j were operated at time 0. If a switching cost is incurred according to policy π at time $t=0$, then

$$\frac{E\left\{\sum_{l=0}^{t_1-1} \beta^l X^1(l) - C - C\beta^{t_1} \mid F^1(0)\right\}}{E\left\{\sum_{l=0}^{t_1-1} \beta^l \mid F^1(0)\right\}} \geq \frac{E\{A_1 \mid \hat{F}^1(0)\}}{E\left\{\sum_{l=0}^{s_1-1} \beta^l \mid \hat{F}^1(0)\right\}} \quad \text{a.s.} \quad (3.14)$$

ii) At $t = s_i + t_{i+1}$, we must have

$$\frac{E\{\beta^{s_i} A_{i+1} - C\beta^{s_i} - C\beta^{s_{i+1}} \mid \hat{F}^1(s_i)\}}{E\left\{\sum_{l=s_i}^{s_{i+1}-1} \beta^l \mid \hat{F}^1(s_i)\right\}} \geq \frac{E\left\{\sum_{l=t_{i+1}}^{t_{i+2}-1} \beta^l X^1(l) \mid F^1(t_{i+1})\right\}}{E\left\{\sum_{l=t_{i+1}}^{t_{i+2}-1} \beta^l \mid F^1(t_{i+1})\right\}} \quad \text{a.s.} \quad (3.15)$$

iii) If (according to π) the machine operated at $t_i + s_i - 1$ is different from the machine operated at time $t_{i+1} + s_i$, then

$$\frac{E\left\{\sum_{l=t_i}^{t_{i+1}-1} \beta^l X^1(l) - C\beta^{t_i} \mid F^1(t_i)\right\}}{E\left\{\sum_{l=t_i}^{t_{i+1}-1} \beta^l \mid F^1(t_i)\right\}} \geq \frac{E\{\beta^{s_i} A_{i+1} - C\beta^{s_i} \mid \hat{F}^1(s_i)\}}{E\left\{\sum_{l=s_i}^{s_{i+1}-1} \beta^l \mid \hat{F}^1(s_i)\right\}} \quad \text{a.s.} \quad (3.16)$$

On the other hand, if (according to π) the machine operated at $t_i + s_i - 1$ is same as the machine operated at $t_{i+1} + s_i$, then

$$\frac{E\left\{\sum_{l=t_i}^{t_{i+1}-1} \beta^l X^1(l) - C\beta^{t_i} - C\beta^{t_{i+1}} \mid F^1(t_i)\right\}}{E\left\{\sum_{l=t_i}^{t_{i+1}-1} \beta^l \mid F^1(t_i)\right\}} \geq \frac{E\{\beta^{s_i} A_{i+1} \mid \hat{F}^1(s_i)\}}{E\left\{\sum_{l=s_i}^{s_{i+1}-1} \beta^l \mid \hat{F}^1(s_i)\right\}} \quad \text{a.s.} \quad (3.17)$$

Proof:

i) Consider policy $\tilde{\pi}$ obtained from π as follows: Policy $\tilde{\pi}$ initially operates machines $X^2, \cdots, X^N$ for s_1 units of time in the same order as policy π does starting at time t_1, then it operates machine X^1 for t_1 units of time and follows π from time $t_1 + s_1$ onwards. Assume that at $t = 0$ a switching cost is always incurred. Then the expected rewards obtained by policies π and $\tilde{\pi}$ are

$$V(\pi) = E\left\{-C + \sum_{l=0}^{t_1-1} \beta^l X^1(l) - C\beta^{t_1} + \beta^{t_1} A_1 - C\beta^{t_1+s_1} + \beta^{s_1} \sum_{l=t_1}^{t_2-1} \beta^l X^1(l) + \Delta \;\middle|\; F(0)\right\} \quad \text{a.s.}$$

where Δ is the reward earned from time $s_1 + t_2$ onwards, and

$$V(\tilde{\pi}) = E\left\{-C + A_1 - C\beta^{s_1} + \beta^{s_1} \sum_{l=0}^{t_2-1} \beta^l X^1(l) + \Delta \;\middle|\; F(0)\right\} \quad \text{a.s.}$$

respectively. Assumption A2) along with some algebra yields

$$\begin{aligned} V(\pi) - V(\tilde{\pi}) &= \frac{E\{(1-\beta^{s_1})|\hat{F}^1(0)\}E\{(1-\beta^{t_1})|F^1(0)\}}{(1-\beta)} \cdot \left\{ \frac{E\left\{\sum_{l=0}^{t_1-1} \beta^l X^1(l) - C \mid F^1(0)\right\}}{E\left\{\sum_{l=0}^{t_1-1} \beta^l \mid F^1(0)\right\}} - \frac{E\{A_1 - C \mid \hat{F}^1(0)\}}{E\left\{\sum_{l=0}^{s_1-1} \beta^l \mid \hat{F}^1(0)\right\}} \right. \\ &\quad \left. - \frac{CE\{\beta^{t_1} \mid F^1(0)\}E\{\beta^{s_1} \mid \hat{F}^1(0)\}(1-\beta)}{E\{(1-\beta^{s_1}) \mid \hat{F}^1(0)\}E\{(1-\beta^{t_1}) \mid F^1(0)\}} \right\} \quad \text{a.s.} \end{aligned}$$

Since π is assumed to be optimal, we must have $V(\pi) - V(\tilde{\pi}) \geq 0$ a.s. Therefore

$$\begin{aligned} &\frac{E\left\{\sum_{l=0}^{t_1-1} \beta^l X^1(l) - C \mid F^1(0)\right\}}{E\left\{\sum_{l=0}^{t_1-1} \beta^l \mid F^1(0)\right\}} - \frac{E\{A_1 - C \mid \hat{F}^1(0)\}}{E\left\{\sum_{l=0}^{s_1-1} \beta^l \mid \hat{F}^1(0)\right\}} \\ &\geq \frac{CE\{\beta^{t_1} \mid F^1(0)\}E\{\beta^{s_1} \mid \hat{F}^1(0)\}(1-\beta)}{E\{(1-\beta^{s_1}) \mid \hat{F}^1(0)\}E\{(1-\beta^{t_1}) \mid F^1(0)\}} \quad \text{a.s.} \end{aligned} \quad (3.18)$$

as $t_1 > 0$ a.s., $s_1 > 0$ a.s. Since

$$\text{R.H.S. of (3.18)} \geq 0 \quad \text{a.s.}$$

the proof of part i-a), when a switching cost is always incurred, is complete. The proof of the remaining parts of i) proceeds along similar arguments.

ii) Consider policy $\hat{\pi}$, obtained from π as follows: Policy $\hat{\pi}$ follows π up to time $t := t_{i+1} + s_i - 1$; at time $t_{i+1} + s_i$ it operates machine X^1 for $t_{i+2} - t_{i+1}$ units of time; afterward it operates machines $X^2, \cdots, X^N$ for $s_{i+1} - s_i$ units of time in the same order as policy π does, starting at time $t_{i+1} + s_i$; and follows π again from time $t_{i+2} + s_{i+1}$ onwards.

Let $V_t(\pi)[V_t(\hat{\pi})]$ denote the expected discounted reward obtained from policy $\pi[\hat{\pi}]$ from time t onwards, conditioned on the information available until time t, i.e., $F(t)$. Then

$$V_t(\pi) = E\left\{-C\beta^{s_i+t_{i+1}} + \beta^{s_i+t_{i+1}} A_{i+1} - C\beta^{s_{i+1}+t_{i+1}} + \beta^{s_{i+1}} \sum_{l=t_{i+1}}^{t_{i+2}-1} \beta^l X^1(l) - C\beta^{s_{i+1}+t_{i+2}} + \Lambda \;\middle|\; F(t)\right\} \quad \text{a.s.}$$

where Λ is the reward earned from time $s_{i+1} + t_{i+2}$

onwards (excluding the switching cost at $s_{i+1}+t_{i+2}$), and

$$V_t(\hat{\pi}) \geq E\Bigg\{\beta^{s_i}\sum_{l=t_{i+1}}^{t_{i+2}-1}\beta^l X^1(l) - C\beta^{s_i+t_{i+2}} + \beta^{s_i+t_{i+2}}A_{i+1} - C\beta^{s_{i+1}+t_{i+2}} + \Lambda \,\Big|\, F(t)\Bigg\} \quad \text{a.s.}$$

respectively. Hence, by using some algebra and the Assumption A2) and $F(t) = \hat{F}^1(s_i) \vee F^1(t_{i+1})$, we have the equations shown at the bottom of the page. Since policies π and $\hat{\pi}$ are the same until time $t := t_{i+1}+s_i-1$, and policy π is assumed to be optimal, we must have $V_t(\pi) - V_t(\hat{\pi}) \geq 0$ a.s. Therefore

$$\frac{E\{\beta^{s_i}A_{i+1} - C\beta^{s_i} - C\beta^{s_{i+1}} \mid \hat{F}^1(s_i)\}}{E\Big\{\sum_{l=s_i}^{s_{i+1}-1}\beta^l \mid \hat{F}^1(s_i)\Big\}} \geq \frac{E\Big\{\sum_{l=t_{i+1}}^{t_{i+2}-1}\beta^l X^1(l) \mid F^1(t_{i+1})\Big\}}{E\Big\{\sum_{l=t_{i+1}}^{t_{i+2}-1}\beta^l \mid F^1(t_{i+1})\Big\}} \quad \text{a.s.}$$

iii) Modify policy π to obtain policy π' as follows: policy π' follows π up to time $t_i + s_i - 1$; at time $t_i + s_i$ π' operates machines $X^2, \cdots, X^N$ for $s_{i+1} - s_i$ units of time in the same order as policy π does starting at time $t_{i+1} + s_i$; then, it operates machine X^1 for $t_{i+1} - t_i$ units of time; finally π' follows π again from time $t_{i+1} + s_{i+1}$ onwards.

As before, let $V_t(\pi)$ denote the expected discounted reward obtained from policy π from time t onwards, conditioned on $F(t)$. Then

$$V_t(\pi) = E\Bigg\{-C\beta^{s_i+t_i} + \beta^{s_i}\sum_{l=t_i}^{t_{i+1}-1}\beta^l X^1(l) - C\beta^{s_i+t_{i+1}} + \beta^{s_i+t_{i+1}}A_{i+1} - C\beta^{s_{i+1}+t_{i+1}} + \Lambda \,\Big|\, F(t)\Bigg\} \quad \text{a.s.}$$

where Λ is the reward earned from time $s_{i+1}+t_{i+1}$ onwards, excluding the switching cost incurred at time $s_{i+1}+t_{i+1}$. To compute $V_t(\pi')$, we consider the following two possibilities:

a) According to policy π, the machine operated at t_i+s_i-1 is different from the machine operated at time $t_{i+1}+s_i$.

Then

$$V_t(\pi') = E\Bigg\{-C\beta^{s_i+t_i} + \beta^{s_i+t_i}A_{i+1} - C\beta^{s_{i+1}+t_i} + \beta^{s_{i+1}}\sum_{l=t_i}^{t_{i+1}-1}\beta^l X^1(l) + \Lambda \,\Big|\, F(t)\Bigg\} \quad \text{a.s.}$$

Because of Assumption (A2) and $F(t) = \hat{F}^1(s_i) \vee F^1(t_i)$, we obtain

$$\begin{aligned}&\frac{(V_t(\pi) - V_t(\pi'))(1-\beta)}{E\{(\beta^{s_i} - \beta^{s_{i+1}}) \mid \hat{F}^1(s_i)\}E\{(\beta^{t_i} - \beta^{t_{i+1}}) \mid F^1(t_i)\}} \\ &= \frac{E\Big\{\sum_{l=t_i}^{t_{i+1}-1}\beta^l X^1(l) - C\beta^{t_i} \mid F^1(t_i)\Big\}}{E\Big\{\sum_{l=t_i}^{t_{i+1}-1}\beta^l \mid F^1(t_i)\Big\}} \\ &\quad - \frac{E\{\beta^{s_i}A_{i+1} - C\beta^{s_i} \mid \hat{F}^1(s_i)\}}{E\Big\{\sum_{l=s_i}^{s_{i+1}-1}\beta^l \mid \hat{F}^1(s_i)\Big\}} \\ &\quad - \frac{C(1-\beta)E\{\beta^{s_{i+1}} \mid \hat{F}^1(s_i)\}E\{\beta^{t_{i+1}} \mid F^1(t_i)\}}{E\{(\beta^{s_i} - \beta^{s_{i+1}}) \mid \hat{F}^1(s_i)\}E\{(\beta^{t_i} - \beta^{t_{i+1}}) \mid F^1(t_i)\}} \quad \text{a.s.}\end{aligned}$$

As π is assumed to be optimal, we must have $V_t(\pi) - V_t(\hat{\pi}) \geq 0$ a.s., hence

$$\frac{E\Big\{\sum_{l=t_i}^{t_{i+1}-1}\beta^l X^1(l) - C\beta^{t_i} \mid F^1(t_i)\Big\}}{E\Big\{\sum_{l=t_i}^{t_{i+1}-1}\beta^l \mid F^1(t_i)\Big\}} \geq \frac{E\{\beta^{s_i}A_{i+1} - C\beta^{s_i} \mid \hat{F}^1(s_i)\}}{E\Big\{\sum_{l=s_i}^{s_{i+1}-1}\beta^l \mid \hat{F}^1(s_i)\Big\}} \quad \text{a.s.}$$

b) According to π, the machine operated at $t_i + s_i - 1$ is the same as the machine operated at $t_{i+1} + s_i$. Then

$$V_t(\pi') = E\Bigg\{\beta^{s_i+t_i}A_{i+1} - C\beta^{s_{i+1}+t_i} + \beta^{s_{i+1}}\sum_{l=t_i}^{t_{i+1}-1}\beta^l X^1(l) + \Lambda \,\Big|\, F(t)\Bigg\} \quad \text{a.s.}$$

By Assumption (A2) and $F(t) = \hat{F}^1(s_i) \vee F^1(t_i)$, we

---

$$\frac{(V_t(\pi) - V_t(\hat{\pi}))(1-\beta)}{E\{(\beta^{s_i} - \beta^{s_{i+1}}) \mid \hat{F}^1(s_i)\}E\{(\beta^{t_{i+1}} - \beta^{t_{i+2}}) \mid F^1(t_{i+1})\}} \leq \frac{E\{\beta^{s_i}A_{i+1} - C\beta^{s_i} - C\beta^{s_{i+1}} \mid \hat{F}^1(s_i)\}}{E\Big\{\sum_{l=s_i}^{s_{i+1}-1}\beta^l \mid \hat{F}^1(s_i)\Big\}}$$
$$- \frac{E\Big\{\sum_{l=t_{i+1}}^{t_{i+2}-1}\beta^l X^1(l) \mid F^1(t_{i+1})\Big\}}{E\Big\{\sum_{l=t_{i+1}}^{t_{i+2}-1}\beta^l \mid F^1(t_{i+1})\Big\}} - \frac{CE\{\beta^{t_{i+2}} \mid F^1(t_{i+1})\}E\{\beta^{s_{i+1}} \mid \hat{F}^1(s_i)\}(1-\beta)}{E\{(\beta^{s_i} - \beta^{s_{i+1}}) \mid \hat{F}^1(s_i)\}E\{(\beta^{t_{i+1}} - \beta^{t_{i+2}}) \mid F^1(t_{i+1})\}}$$

get

$$\frac{(V_t(\pi)-V_t(\pi'))(1-\beta)}{E\{(\beta^{s_i}-\beta^{s_{i+1}})\mid \hat{F}^1(s_i)\}E\{(\beta^{t_i}-\beta^{t_{i+1}})\mid F^1(t_i)\}}$$
$$= \frac{E\left\{\sum_{l=t_i}^{t_{i+1}-1}\beta^l X^1(l) - C\beta^{t_i} - C\beta^{t_{i+1}} \mid F^1(t_i)\right\}}{E\{\sum_{l=t_i}^{t_{i+1}-1}\beta^l \mid F^1(t_i)\}}$$
$$- \frac{E\{\beta^{s_i}A_{i+1}\mid \hat{F}^1(s_i)\}}{E\{\sum_{l=s_i}^{s_{i+1}-1}\beta^l \mid \hat{F}^1(s_i)\}}$$
$$- \frac{2C(1-\beta)E\{\beta^{s_{i+1}}\mid \hat{F}^1(s_i)\}E\{\beta^{t_{i+1}}\mid F^1(t_i)\}}{E\{(\beta^{s_i}-\beta^{s_{i+1}})\mid \hat{F}^1(s_i)\}E\{(\beta^{t_i}-\beta^{t_{i+1}})\mid F^1(t_i)\}} \quad \text{a.s.}$$

By the optimality of π we must have

$$\frac{E\left\{\sum_{l=t_i}^{t_{i+1}-1}\beta^l X^1(l) - C\beta^{t_i} - C\beta^{t_{i+1}} \mid F^1(t_i)\right\}}{E\left\{\sum_{l=t_i}^{t_{i+1}-1}\beta^l \mid F^1(t_i)\right\}} \geq \frac{E\{\beta^{s_i}A_{i+1}\mid \hat{F}^1(s_i)\}}{E\left\{\sum_{l=s_i}^{s_{i+1}-1}\beta^l \mid \hat{F}^1(s_i)\right\}} \quad \text{a.s.} \quad \square$$

Lemma 3.3 has an interpretation similar to that of Lemma 2.4. The following corollary is an immediate consequence of the above lemma.

Corollary 3.1: Suppose π is optimal. Then

i) At $t = s_i + t_{i+1}$, we must have

$$\frac{E\{\beta^{s_i}A_{i+1} - C\beta^{s_i} - C\beta^{s_{i+1}}\mid \hat{F}^1(0)\}}{E\left\{\sum_{l=s_i}^{s_{i+1}-1}\beta^l \mid \hat{F}^1(0)\right\}} \geq \frac{E\left\{\sum_{l=t_{i+1}}^{t_{i+2}-1}\beta^l X^1(l)\mid F^1(0)\right\}}{E\left\{\sum_{l=t_{i+1}}^{t_{i+2}-1}\beta^l \mid F^1(0)\right\}} \quad \text{a.s.} \quad (3.19)$$

ii) If (according to π) the machine operated at $t_i + s_i - 1$ is different from the machine operated at time $t_{i+1}+s_i$, then

$$\frac{E\left\{\sum_{l=t_i}^{t_{i+1}-1}\beta^l X^1(l) - C\beta^{t_i}\mid F^1(0)\right\}}{E\left\{\sum_{l=t_i}^{t_{i+1}-1}\beta^l \mid F^1(0)\right\}} \geq \frac{E\{\beta^{s_i}A_{i+1} - C\beta^{s_i}\mid \hat{F}^1(0)\}}{E\left\{\sum_{l=s_i}^{s_{i+1}-1}\beta^l \mid \hat{F}^1(0)\right\}} \quad \text{a.s.} \quad (3.20)$$

On the other hand, if (according to π) the machine operated at t_i+s_i-1 is the same as the machine operated at $t_{i+1}+s_i$, then

$$\frac{E\left\{\sum_{l=t_i}^{t_{i+1}-1}\beta^l X^1(l) - C\beta^{t_i} - C\beta^{t_{i+1}}\mid F^1(0)\right\}}{E\left\{\sum_{l=t_i}^{t_{i+1}-1}\beta^l \mid F^1(0)\right\}} \geq \frac{E\{\beta^{s_i}A_{i+1}\mid \hat{F}^1(0)\}}{E\left\{\sum_{l=s_i}^{s_{i+1}-1}\beta^l \mid \hat{F}^1(0)\right\}} \quad \text{a.s.} \quad (3.21)$$

Proof: The corollary follows from (3.15)–(3.17) together with Assumptions A1)–A3) and [35, Proposition 4.20, part (ii)]. $\square$

Lemma 3.4: Consider the policy π that is described at the beginning of this section and is assumed to be optimal. Construct a policy $\hat{\pi}$ from π as follows:

i) $\hat{\pi}$ follows π until time $s_{k-1}+t_k-1$.

ii) At time $s_{k-1}+t_k$, $\hat{\pi}$ operates machine X^1 for $\tau_1(0)-t_k$ units of time.

iii) Afterwards, $\hat{\pi}$ operates machine $X^2, X^3, \cdots, X^N$ for $s_k - s_{k-1}$ units of time in the same order as policy π does, starting at time $t_k + s_{k-1}$.

iv) $\hat{\pi}$ follows π from time $\tau_1(0)+s_k$ onwards.

Then $\hat{\pi}$ is an optimal policy.

Proof: Let γ denote the discounted reward obtained until time $t = t_{k-1}+s_{k-1}-1$. Then, the expected discounted rewards obtained by policies π and $\hat{\pi}$ are

$$V(\pi) = E\Bigg\{\gamma - C\beta^{s_{k-1}+t_{k-1}} + \beta^{s_{k-1}}\sum_{l=t_{k-1}}^{t_k-1}\beta^l X^1(l) - C\beta^{s_{k-1}+t_k} + \beta^{s_{k-1}+t_k}A_k - C\beta^{s_k+t_k} + \beta^{s_k}\sum_{l=t_k}^{t_{k+1}-1}\beta^l X^1(l) + \delta \,\Big|\, F(0)\Bigg\} \quad \text{a.s.}$$

where δ is the reward earned from time s_k+t_{k+1} onwards, and

$$V(\hat{\pi}) \geq E\Bigg\{\gamma - C\beta^{s_{k-1}+t_{k-1}} + \beta^{s_{k-1}}\sum_{l=t_{k-1}}^{\tau_1(0)-1}\beta^l X^1(l) - C\beta^{s_{k-1}+\tau_1(0)} + \beta^{s_{k-1}+\tau_1(0)}A_k - C\beta^{\tau_1(0)+s_k} + \beta^{s_k}\sum_{l=\tau_1(0)}^{t_{k+1}-1}\beta^l X^1(l) + \delta \,\Big|\, F(0)\Bigg\} \quad \text{a.s.}$$

respectively. The inequality above is strict only when $\tau_1(0) = t_{k+1}$. Therefore, because of Assumption A2)

$$\frac{(V(\hat{\pi})-V(\pi))(1-\beta)}{E\{(\beta^{s_{k-1}}-\beta^{s_k})\mid \hat{F}^1(0)\}E\{(\beta^{t_k}-\beta^{\tau_1(0)})\mid F^1(0)\}} \geq \frac{E\left\{\sum_{l=t_k}^{\tau_1(0)-1}\beta^l X^1(l)\mid F^1(0)\right\}}{E\left\{\sum_{l=t_k}^{\tau_1(0)-1}\beta^l \mid F(0)\right\}} - \frac{E\{\beta^{s_{k-1}}A_k - C\beta^{s_{k-1}} - C\beta^{s_k}\mid \hat{F}^1(0)\}}{E\left\{\sum_{l=s_{k-1}}^{s_k-1}\beta^l \mid \hat{F}^1(0)\right\}} \quad \text{a.s.} \quad (3.22)$$

As it is optimal to operate machine X^1 at time $s_{k-1}+t_{k-1}$, and a switch is incurred at that time, by Corollary 3.1, part ii)

$$\frac{E\left\{\sum_{l=t_{k-1}}^{t_k-1}\beta^l X^1(l) - C\beta^{t_{k-1}}\mid F^1(0)\right\}}{E\left\{\sum_{l=t_{k-1}}^{t_k}\beta^l \mid F^1(0)\right\}} \geq \frac{E\{\beta^{s_{k-1}}A_k - C\beta^{s_{k-1}}\mid \hat{F}^1(0)\}}{E\left\{\sum_{l=s_{k-1}}^{s_k-1}\beta^l \mid \hat{F}^1(0)\right\}} \quad \text{a.s.} \quad (3.23)$$

In terms of the switching cost incurred at $t = 0$, there are only the following two possibilities:

i) A switching cost is incurred at $t = 0$. In this case, $\tau_1(0) = \tau_{c1}(0)$. As it is optimal to operate machine X^1

at $t = 0$, we have from Lemma 3.3 [part i-a) or i-c)], depending on the machine operated at time t_1) that

$$\frac{E\left\{\sum_{l=0}^{t_1-1}\beta^l X^1(l) - C \mid F^1(0)\right\}}{E\left\{\sum_{l=0}^{t_1-1}\beta^l \mid F^1(0)\right\}} \geq \frac{E\{A_1 - C \mid \hat{F}^1(0)\}}{E\left\{\sum_{l=0}^{s_1-1}\beta^l \mid \hat{F}^1(0)\right\}} \quad \text{a.s.} \quad (3.24)$$

Furthermore, by Corollary 3.1 (part i) at $t = t_1$

$$\frac{E\{A_1 - C - C\beta^{s_1} \mid \hat{F}^1(0)\}}{E\left\{\sum_{l=0}^{s_1-1}\beta^l \mid \hat{F}^1(0)\right\}} \geq \frac{E\left\{\sum_{l=t_1}^{t_2-1}\beta^l X^1(l) \mid F^1(0)\right\}}{E\left\{\sum_{l=t_1}^{t_2-1}\beta^l \mid F^1(0)\right\}} \quad \text{a.s.} \quad (3.25)$$

By repeating the arguments resulting in (3.24) and (3.25) at times $t = t_{i+1} + s_i$ and $t_i + s_i$, $i = 1, 2, \cdots k-1$, we obtain

$$\frac{E\left\{\sum_{l=0}^{t_1-1}\beta^l X^1(l) - C \mid F^1(0)\right\}}{E\left\{\sum_{l=0}^{t_1-1}\beta^l \mid F^1(0)\right\}} \geq \frac{E\left\{\sum_{l=t_1}^{t_2-1}\beta^l X^1(l) \mid F^1(0)\right\}}{E\left\{\sum_{l=t_1}^{t_2-1}\beta^l \mid F^1(0)\right\}} \geq \cdots \geq \frac{E\left\{\sum_{l=t_{k-1}}^{t_k-1}\beta^l \mid F^1(0)\right\}}{E\left\{\sum_{l=t_{k-1}}^{t_k-1}\beta^l \mid F^1(0)\right\}} \quad \text{a.s.} \quad (3.26)$$

Furthermore, by Lemma 3.1

$$\frac{E\left\{\sum_{l=t_k}^{\tau_1(0)-1}\beta^l X^1(l) \mid F^1(0)\right\}}{E\left\{\sum_{l=t_k}^{\tau_1(0)-1}\beta^l \mid F^1(0)\right\}} \geq \frac{E\left\{\sum_{l=0}^{t_1-1}\beta^l X^1(l) - C \mid F^1(0)\right\}}{\{E\sum_{l=0}^{t_1-1}\beta^l \mid F^1(0)\}} \quad \text{a.s.} \quad (3.27)$$

Combining (3.23), (3.26), and (3.27) we obtain

$$\frac{E\left\{\sum_{l=t_k}^{\tau_1(0)-1}\beta^l X^1(l) \mid F^1(0)\right\}}{E\left\{\sum_{l=t_k}^{\tau_1(0)-1}\beta^l \mid F^1(0)\right\}} \geq \frac{E\{\beta^{s_{k-1}}A_k - C\beta^{s_{k-1}} \mid \hat{F}^1(0)\}}{E\{\sum_{l=s_{k-1}}^{s_k-1}\beta^l \mid \hat{F}^1(0)\}} \quad \text{a.s.} \quad (3.28)$$

From (3.22) and (3.28) we conclude that $V(\hat{\pi}) \geq V(\pi)$ a.s. Since π was assumed to be an optimal policy, $\hat{\pi}$ is also an optimal policy.

ii) No switching cost is incurred at time $t = 0$. In this case $\tau_1(0) = \tau_{gi}(0)$ and arguments similar to those of case i) show that $\hat{\pi}$ is an optimal policy. □

Based on Lemmas 3.1, 3.3, and 3.4, we can now obtain the stochastic analogue of Theorem 2.1 for N-armed bandits.

Theorem 3.1: An optimal scheduling policy π for the stochastic N-armed bandit problem with switching cost has the following property: Decisions about the processor allocation are made only at those $F(\cdot)$ stopping times that achieve an appropriate index (the Gittins index or the switching cost index).

Proof: Without any loss of generality, assume that it is optimal to operate machine X^1 at $t = 0$. Consider a policy π of the form

$$\begin{aligned} &X^1(0,\omega), X^1(1,\omega)\cdots X^1(t_1'(\omega)-1,\omega), Z(0,\omega),\cdots \\ &\quad Z(s_1'(\omega)-1,\omega), X^1(t_1'(\omega),\omega),\cdots, \\ &\quad X^1(t_2'(\omega)-1,\omega), Z(s_1'(\omega),\omega),\cdots, \\ &\quad Z(s_2'(\omega)-1,\omega),\cdots X^1(t'_{k(\omega)-1,}(\omega),\omega),\cdots, \\ &\quad X^1(t'_{k(\omega)}(\omega)-1,\omega), Z(s'_{k(\omega)-1}(\omega),\omega),\cdots, \\ &\quad Z(s'_{k(\omega)}(\omega)-1,\omega), X^1(t'_{k(\omega)}(\omega),\omega),\cdots, \\ &\quad X^1(\tau_1(0,\omega)-1,\omega),\cdots, X^1(t'_{k(\omega)+1}(\omega)-1,\omega), \\ &\quad Y(t^y(t'_{k(\omega)+1}(\omega)+s'_{k(\omega)}(\omega)),\omega), \\ &\quad Y(t^y(t'_{k(\omega)+1}(\omega)+s'_{k(\omega)}(\omega)+1),\omega),\cdots \end{aligned}$$

where (by some abuse of notation) $Z(0,\omega)$, $Z(1,\omega)$, $Z(2,\omega),\cdots, Z(s_1',\omega),\cdots, Z(s'_{k(\omega)}-1,\omega)$ are an interleaving of reward sequences from machines X^2, $X^3,\cdots, X^N$ (or a subset of these machines), $Y(t^y(t'_{k(\omega)+1}+s'_{k(\omega)}),\omega)$, $Y(t^y(t'_{k(\omega)+1}+s'_{k(\omega)}+1),\omega),\cdots$ represents an interleaving of reward sequences from machines X^1, $X^2,\cdots, X^N$ from time $t'_{k(\omega)+1}+s'_{k(\omega)}-1$ onwards, and $\tau_1(0,\omega) = \tau_{c1}(0,\omega)$ or $\tau_{g1}(0,\omega)$, depending on whether a switching cost is incurred at $t = 0$ or not. When the index for machine X^1 is achieved between time instants $t_k'+s_k'$ and $t_{k+1}'+s_k'-1$, then repeated application of Lemma 3.4 at times $t_k'+s_{k-1}'$, $t_{k-1}'+s_{k-2}',\cdots t_1'$ proves that it is optimal to operate X^1 up until $\tau_{c1}(0)-1$ (or $\tau_{g1}(0)-1$). When the index for machine X^1 is achieved before t_1' we serve X^1 until $\tau_{c1}(0)-1$ (or $\tau_{g1}(0)-1$). At $\tau_{c1}(0)$ [or $\tau_{g1}(0)$] we are faced with the same problem as at time $t = 0$; thus it is optimal to serve the machine selected at $\tau_{c1}(0)$ (or $\tau_{g1}(0)$) until its index is achieved.

Repetition of the above argument proves that along an optimal policy, decisions about the processor allocation need to be made only at $F(\cdot)$ stopping times that achieve a Gittins index or a switching cost index (as appropriate). □

As in the case of the two-armed deterministic bandit, Theorem 3.1 guarantees that the search for an optimal scheduling strategy can be reduced to the set of policies which possibly switch only at stopping times that achieve an appropriate index. Furthermore, by the counterexample following Theorem 2.1, the policy that operates machines according to the highest appropriate index is not necessarily optimal. Thus, determining an optimal allocation strategy for the multi-armed bandit problem with switching cost remains a difficult and challenging problem. However, there is at least one situation where an index policy is optimal. This is the case where along each sample path of positive probability the reward process associated with each machine is identically zero after a finite time (depending on the sample path) and the stopping time $\tau_{ci}(0)$ is such that $X^i(t) = 0$ a.s. for all $i \in \{1,\cdots, N\}$ and

for all t such that $t \geq \tau_{ci}(0)$ a.s. In such a case, Theorem 3.1 implies that the optimal policy is an exhaustive[2] service policy. Then, a simple interchange argument shows that it is optimal to serve the machines in decreasing order of indexes. This situation includes the scheduling problem with switching cost studied in [3]. Indeed, we show that in this case the "switching cost index" defined by (3.3) is identical to the index defined in [3]. This is done as follows: The problem analyzed in [3] deals with the dynamic scheduling of parallel queues with switching cost and no arrivals. The reward index of queue i, $i = 1, 2, \cdots, N$ (in discrete time) is defined (in [3]) as

$$\nu_i = h_i S_i (1 - S_i)^{-1} - (1 - \beta) C (1 - S_i^{q_i})^{-1} \tag{3.29}$$

where

h_i	The instantaneous holding cost per customer for queue i, $i = 1, 2, \cdots, N$.
$S_i = E\{\beta^{\sigma_i(1)}\}$	
$\sigma_i(j)$, $j = 1, \cdots, q_i$	The length of the jth service period in queue i, and service times of jobs in queue i are independent and identically distributed (i.i.d.).
q_i	The initial length of queue i.
C	The switching cost.
β	The discount factor.

In the context of this scheduling problem, we can interpret our results as follows: each bandit i is associated with a queue i of length q_i and instantaneous holding cost h_i; rewards are obtained only at job completion epochs; the reward obtained from bandit i when a job is completed at time $(t-1)$ in queue i is $\beta^t h_i/(1-\beta)$. Consequently, according to (3.3), the "switching cost index" of bandit i is

$$\nu_{ci} = \frac{E\left\{\sum_{k=1}^{q_i} \beta^{f_{i,k}} \frac{h_i}{1-\beta} - C\right\}}{E\left\{\sum_{k=0}^{f_{i,q_i}-1} \beta^k\right\}} \tag{3.30}$$

where

$$f_{i,k} = \sum_{j=1}^{k} \sigma_i(j).$$

Since successive service times are i.i.d.

$$E\{\beta^{f_{i,k}}\} = S_i^k.$$

Therefore

$$\nu_{ci} = \frac{\frac{h_i}{1-\beta} \frac{S_i(1-S_i^{q_i})}{1-S_i} - C}{\frac{1-S_i^{q_i}}{1-\beta}} \tag{3.31}$$

and by simplification we get

$$\nu_{ci} = h_i S_i (1 - S_i)^{-1} - (1 - \beta) C (1 - S_i^{q_i})^{-1} \tag{3.32}$$

which is identical to (3.29).

[2]By an exhaustive policy, we mean one that serves a machine until the machine's reward sequence becomes identically zero.

C. Calculation of the Indexes

In this section we present a method for computing the switching indexes when each machine is described by a finite state Markov chain; the Gittins index for finite state Markovian bandits has been computed in [16, Section 4].

Let $x(s)$, $s = 0, 1, 2, \cdots$ be a Markov chain with state space $\Theta \doteq \{1, 2, \cdots, M\}$ and transition probability matrix $P = \{P_{ij}\}$. Let $\mathcal{R}(i)$ be the reward when $x(t) = i$. Suppose the state is perfectly observed. Then, we have the standard machine $\{X(s), F(s)\}$ where

$$X(s) = R(x(s)),\ F(s) = \sigma\{x(0), x(1), x(2), \cdots x(s)\}.$$

From (3.2) and (3.3) it follows that if $x(t) = i$, the Gittins and switching indexes are given by

$$\nu_{gi}(t) := \max_{\tau > t} \frac{E_i\left\{\sum_{l=t}^{\tau-1} \beta^l X^i(l)\right\}}{E_i\left\{\sum_{l=t}^{\tau-1} \beta^l\right\}} \tag{3.33}$$

and

$$\nu_{ci}(t) := \max_{\tau > t} \frac{E_i\left\{\sum_{l=t}^{\tau-1} \beta^l X^i(l) - C\beta^t\right\}}{E_i\left\{\sum_{l=t}^{\tau-1} \beta^l\right\}} \tag{3.34}$$

respectively, where $E_i f \doteq E\{f \mid x(t) = i\}$ and τ ranges over all stopping times of $\{x(\cdot)\}$. To calculate ν_{ci}, $i = 1, 2, \cdots, M$ we proceed as follows: We construct a new Markov chain $\hat{x}(s)$, $s = 0, 1, 2, \cdots$ with state space $\Theta = \{1, 2, \cdots M, 1', 2', \cdots, M'\}$ and transition probability matrix $\hat{P} = \{\hat{P}_{lk}\}$, $l, k \in \hat{\Theta}$, given by

$$\begin{aligned}
&\hat{P}_{ij} = P_{ij} \quad \forall i, j \in \{1, 2, \cdots, M\} \\
&\hat{P}_{ij'} = 0 \quad \forall i \in \{1, 2, \cdots, M\}, \forall j' \in \{1', 2', \cdots, M'\} \\
&\hat{P}_{i'j} = P_{ij} \quad \forall i' \in \{1', 2', \cdots, M'\}, \forall j \in \{1, 2, \cdots, M\} \\
&\hat{P}_{i'j'} = 0 \quad \forall i', j' \in \{1', 2', \cdots, M'\}.
\end{aligned} \tag{3.35}$$

We take the reward $\hat{\mathcal{R}}(\cdot)$, associated with this chain, to be

$$\begin{aligned}
&\hat{R}(j) = R(j) \quad \forall j \in \{1, 2, \cdots, M\}, \\
&\hat{R}(j') = R(j) - C \quad \forall j' \in \{1', 2', \cdots, M'\}.
\end{aligned} \tag{3.36}$$

Then we have the standard machine $\{\hat{X}(s), \hat{F}(s)\}$ where

$$\hat{X}(s) = \hat{R}(\hat{x}(s)),\ \hat{F}(s) = \sigma\{\hat{x}(0), \hat{x}(1), \cdots, \hat{x}(s)\}.$$

Using the algorithms presented in [16, Section 4], we compute the Gittins indexes $\hat{\nu}_{gi'}(t)$ for the above machine. From (3.33) and (3.34) and the specification of the standard machine $\{\hat{X}(s), \hat{F}(s)\}$, it follows directly that

$$\begin{aligned}
&\hat{\nu}_{gj}(t) = \nu_{gj}(t) \quad \forall j \in \{1, 2, \cdots, M\} \\
&\hat{\nu}_{gj'}(t) = \nu_{cj}(t) \quad \forall j' \in \{1', 2', \cdots, M'\}.
\end{aligned} \tag{3.37}$$

To determine the stopping time that achieves the switching index, we define for each state $j' \in \{1', 2', \cdots, M'\}$ a stopping set $\Theta_{oc}(j') \subset \Theta$ and its complementary (with respect to Θ) continuation set $\Theta_{1c}(j')$. The set $\Theta_{oc}(j')$ has the following interpretation: if the standard machine $\{\hat{X}(s), \hat{F}(s)\}$ is in state j' when its operation begins, it is operated until the first moment its state enters $\Theta_{oc}(j')$. Then, from (3.37) and a result of Gittins [13, p. 154], we obtain the following.

Lemma 3.5: The maximum in (3.34) is obtained by setting

$$\Theta_{oc}(j') = \{i \in \Theta : \nu_{gi}(\cdot) < \hat{\nu}_{gj'}(\cdot)\}.$$

IV. Extensions to Multi-Armed Bandits with Switching Cost

In this section, we briefly discuss three possible extensions of the problem presented in Section III:

i) a bandit problem with a single-server and machine-dependent switching cost (i.e., the setup cost for a particular machine depends on that machine);
ii) multi-armed bandits with multiple servers and switching cost; and
iii) multi-armed bandits with switching delays.

We find that optimal scheduling policies for the first and third problems have the qualitative properties described by Theorem 3.1. On the contrary, the optimal scheduling policy for the second problem does not, in general, have the qualitative properties described by the above-mentioned theorem.

A. Machine Dependent Switching Cost

Let the setup cost for machine X^i be denoted by C_i, $i \in \{1, \cdots, N\}$. The bandit problem with machine dependent switching cost is to find the policy π that maximizes

$$V(\pi) := E\left\{\sum_{t=0}^{\infty} \beta^t [X^{m(t)}(t^{m(t)}(t)) - 1(m(t) \neq m(t-1))C_{m(t)}] \,\middle|\, F(0)\right\}. \quad (4.1)$$

The "switching cost index" is defined as

$$\nu_{ci}(t) := \max_{\tau > t} \frac{E\left\{\sum_{l=t}^{\tau-1} \beta^l X^i(l) - C_i\beta^t \mid F^i(t)\right\}}{E\left\{\sum_{l=t}^{\tau-1} \beta^l \mid F^i(t)\right\}} \quad (4.2)$$

where, as before, the maximization is over all stopping times $\tau(t < \tau \leq \infty)$ of $\{F^i(\cdot)\}$.

The basic results of Sections II and III (Theorems 2.1 and 3.1) remain valid for the machine-dependent switching-cost problem with the switching index defined as above. The proofs of the counterparts of Theorems 2.1 and 3.1 are omitted because they are based on the same arguments as the case of machine-independent switching cost. The counterparts of the sufficient conditions of Lemmas 2.6–2.8 [(2.27), (2.32), (2.36)] are

$$\frac{\sum_{l=0}^{\tau_{c1}-1} \beta^l X(l+\tau) - C_x}{\sum_{l=0}^{\tau_{c1}-1} \beta^l} - \frac{\sum_{l=0}^{s_{c1}-1} \beta^l Y(l+s) - C_y}{\sum_{l=0}^{s_{c1}-1} \beta^l} \geq \frac{C_x \beta^{\tau_{c1}+s_{c1}}(1-\beta)}{(1-\beta^{s_{c1}})(1-\beta^{\tau_{c1}})} \quad (4.3)$$

$$\frac{\sum_{l=0}^{\tau_{g1}-1} \beta^l X(l+\tau)}{\sum_{l=0}^{\tau_{g1}-1} \beta^l} \geq \frac{\sum_{l=0}^{s_{c1}-1} \beta^l Y(l+s) - C_y}{\sum_{l=0}^{s_{c1}-1} \beta^l} \quad (4.4)$$

and

$$\frac{\sum_{l=0}^{\tau_{c1}-1} \beta^l X(l+\tau) - C_x - C_y\beta^{\tau_{c1}}}{\sum_{l=0}^{\tau_{c1}-1} \beta^l} - \frac{\sum_{l=0}^{s_{g1}-1} \beta^l Y(l+s)}{\sum_{l=0}^{s_{g1}-1} \beta^l} \geq \frac{(C_x + C_y)\beta^{\tau_{c1}+s_{g1}}(1-\beta)}{(1-\beta^{\tau_{c1}})(1-\beta^{s_{g1}})} \quad (4.5)$$

respectively. Lemmas 2.9–2.11 remain valid too with their conditions modified appropriately.

B. Multiple Servers with Switching Cost

Another important extension is that of multiple servers attending the N-projects. Multiserver problems are considerably more difficult to analyze than single-server ones, and many of the structure of single-server problems is lost when one considers their multiserver counterparts. The qualitative properties of the optimal strategies of single-server bandits developed in Sections II and III no longer hold for the multiserver, multi-armed bandit problem with switching cost, as the following counterexample illustrates.

Counterexample: Consider two servers 1, 2 and three machines X, Y, and Z. Machines X and Y are identical. The reward sequences for machines X, and Z are 15, 15, 15, 15, 15, 16, 0, 0, $\cdots$ and 8, 8, 8, 8, 8, 8, 8, 8, 9, 0, 0, $\cdots$, respectively. Let $\beta = 0.9$ and $C = 1$. At $t = 0$, either a switching cost is incurred for all machines or no switching cost is incurred. Note that in this case both the switching and the Gittins indexes are achieved at the last time instant before the reward becomes zero. Then, according to Theorem 3.1, the optimal policy for the single-server problem with switching cost is an exhaustive policy. For the two-server problem the following two exhaustive policies are possible:

i) At $t = 0$, start with machines X and Y and serve them until $t = 5$, i.e., until their reward becomes zero; from $t = 6$ until $t = 14$, serve machine Z with one of the two servers. Call this policy π_A.
ii) At $t = 0$ start with machines X and Z. At $t = 6$, i.e., when the reward from machine X is zero, switch the server of machine X to machine Y (and continue with the other server at Z). Serve machines Z and Y until their reward becomes zero. Call this policy π_B.

Note that any other possible exhaustive policy will be equivalent to π_A or π_B in terms of total discounted reward incurred. Now, construct the following policy called π_C: Begin by serving machines X and Y at $t = 0$. At $t = 3$ continue the operation of machine X at one server and switch the second server to machine Z. At $t = 6$, switch the server for machine X to machine Y and serve Y until its reward becomes zero. Simple calculations show that $V(\pi_C) - V(\pi_A) = 0.87 > 0$ and $V(\pi_C) - V(\pi_B) = 10.84 > 0$; hence, operating the machines until appropriate indexes are achieved is not optimal in the case of multiple servers. □

The above result is not surprising because it is known [7] that even in multi-armed bandit problems with multiple servers and no switching cost, the Gittins index strategy is not optimal. One of the important factors in determining optimal schedules in multiple-server problems is efficient server-utilization, and this issue destroys the optimality of

the policies described by Theorems 2.1 and 3.1 as one can see from the above counterexample. This counterexample illustrates the difficulties encountered even in the case of deterministic projects, machine-independent switching cost, and two servers.

C. The Multi-Armed Bandits with Switching Delay

Although we have emphasized the case of switching cost, we also address the case of switching delay which we deem to be at least as important. The multi-armed bandit problem with switching delay is the same as the problem with switching cost, except that a switching (setup) delay D (instead of a switching cost) is incurred when the server moves from one project to another and rewards are nonnegative. No reward is incurred during the switching interval. We assume that the delay D is a nonnegative integer random variable with a given distribution such that $0 < E[D] < \infty$ and is independent of machine dynamics. The multi-armed bandit problem with switching delay is to find a policy π that maximizes

$$V(\pi) := E\left\{\sum_{t=0}^{\infty} \beta^t X^{m(t)}(t^{m(t)}(t)) \;\middle|\; F(0)\right\}. \tag{4.6}$$

As in the case of switching cost, the Gittins index rule is not optimal for the problem with switching delay. However, a result similar to that of Theorem 3.1 is true. Define the switching delay index $\nu_{di}(t)$, after machine X^i has been operated for t units of time, as follows

$$\nu_{di}(t) := \max_{\tau > t} \frac{E\left\{\beta^D \sum_{l=t}^{\tau-1} \beta^l X^i(l) \mid F^i(t)\right\}}{E\left\{\sum_{l=t}^{\tau+D-1} \beta^l \mid F^i(t)\right\}} \tag{4.7}$$

where the maximization is over all stopping times $t < \tau \leq \infty$ of $\{F^i(\cdot)\}$. By arguments similar to those of Section III, we can prove the following result.

Theorem 4.1: An optimal scheduling policy π for the N-armed stochastic bandit problem with switching delay has the following property: Decisions about the processor allocation are made only at those $F(\cdot)$ stopping times that achieve an appropriate index (the Gittins index or the switching delay index).

The details of the analysis of multi-armed bandits with switching delays can be found in [33].

V. Conclusions

The presence of switching penalties complicates considerably the nature of the multi-armed bandit problem. Operating the machine with the highest appropriate index (the Gittins index or the switching cost index or the switching delay index) is not necessarily optimal. We have shown that along optimal policies, decisions about the processor allocation need to be made only at stopping times that achieve an appropriate index (a Gittins index or a switching index). This feature of optimal scheduling policies is intuitively pleasing: If at a certain time instant it is optimal to allocate the processor to a certain machine X, then it should be optimal to maximize the reward rate obtained from the operation of X during this allocation; hence X must be operated on until the appropriate index is achieved, and at that point the next processor allocation should be decided.

The above feature of optimal policies, together with the sufficient conditions of Lemmas 2.6–2.11 simplify the search for optimal allocation strategies. The conditions of Lemmas 2.6–2.11 are derived by looking one or two branches ahead in the decision tree that results after using the property of optimal allocation strategies expressed by Theorem 2.1. Additional conditions that may result in further simplification of the search for an optimal policy can be derived by looking three or more branches ahead in the aforementioned tree.

We have identified one class of problems (parallel queues with switching penalties and no arrivals) where the qualitative properties of optimal policies described by Theorem 3.1 are sufficient to determine an optimal allocation strategy. In general, determination of optimal allocation strategies remains a difficult and challenging task.

Acknowledgment

The authors are grateful to the anonymous reviewers whose excellent suggestions have improved this paper.

References

[1] J. C. Gittins, *Multi-Armed Bandit Allocation Indices.* New York: Wiley, 1989.
[2] J. C. Walrand, *An Introduction to Queueing Networks.* Englewood Cliffs, NJ: Prentice-Hall, 1988.
[3] M. P. Van Oyen, D. G. Pandelis, and D. Teneketzis, "Optimality of index policies for stochastic scheduling with switching penalties," *J. Appl. Probab.*, vol. 29, pp. 957–966, 1992.
[4] D. T. Mortensen, "Job search and labor market analysis," in *Handbook of Labor Economics*, O. Ashenfelter and R. Layard, Eds, 1986, pp. 849–919.
[5] K. D. Glazebrook and R. J. Boys, "A class of Bayesian models for optimal exploration," manuscript, Dept. Mathematics Statistics, Univ. Newcastle upon Tyne, 1994.
[6] M. H. DeGroot, *Optimal Statistical Decisions.* New York: McGraw-Hill, 1970.
[7] T. Ishikida, "Informational aspects of decentralized resource allocation," Ph.D. dissertation, Univ. Calif., Berkeley, 1992.
[8] J. S. Banks and R. K. Sundaram, "Denumerable-armed bandits," *Econometrica,* vol. 60, pp. 1071–1096, Sept. 1992.
[9] M. L. Weitzman, "Optimal search for the best alternative," *Econometrica,* vol. 47, pp. 641–654, 1979.
[10] M. N. Katehakis and C. Derman, "Computing optimal sequential allocation rules in clinical trials," *Inst. Math. Stat. Lecture Note Series: Adaptive Allocation Procedures and Related Topics,* vol. 8, pp. 29–39, 1986.
[11] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification and Adaptive Control.* Englewood Cliffs, NJ: Prentice-Hall, 1986.
[12] J. C. Gittins and D. Jones, "A dynamic allocation index for the sequential design of experiments," in *Progress in Statistics, European Meeting of Statisticians*, vol. 1, J. Gani, K. Sarkadi, and I. Vince, Eds. Amsterdam: North Holland, 1974, pp. 241–266.
[13] J. C. Gittins, "Bandit processes and dynamic allocation indices," *J. Royal Statistical Soc.*, vol. 41, pp. 148–177, 1979.
[14] P. Whittle, "Multi-armed bandits and the Gittins index," *J. Royal Statistical Soc.*, vol. 42, pp. 143–149, 1980.
[15] ——, "Arm acquiring bandits," *Annals Probab.*, vol. 9, pp. 284–292, 1981.
[16] P. Varaiya, J. Walrand, and C. Buyukkoc, "Extensions of the multi-armed bandit problem," *IEEE Trans. Automat. Contr.*, vol. AC-30, pp. 426–439, 1985.
[17] R. Agrawal, D. Teneketzis, and V. Anantharam, "Asymptotically efficient adaptive allocation schemes for controlled i.i.d. processes: Finite

parameter space," *IEEE Trans. Automat. Contr.*, vol. 34, pp. 258–267, Mar. 1989.
[18] ——, "Asymptotically efficient adaptive allocation schemes for controlled Markov chains: Finite parameter space," *IEEE Trans. Automat. Contr.*, vol. 34, pp. 1249–1259, Dec. 1989.
[19] F. P. Kelly, "Multi-armed bandits with discount factor near one: the Bernoulli case," *Ann. Statistics*, vol. 9, pp. 987–1001, 1981.
[20] K. D. Glazebrook, "On a sufficient condition for superprocesses due to Whittle," *J. Appl. Probab.*, vol. 19, pp. 99–110, 1982.
[21] A. Mandelbaum, "Discrete multi-armed bandits and multi-parameter processes," *Probab. Theory*, vol. 71, pp. 129–147, 1986.
[22] I. Karatzas, "Gittins indices in the dynamic allocation problem for diffusion processes," *Ann. Probab.*, vol. 12, pp. 173–192, 1984.
[23] J. N. Tsitsiklis, "A lemma on the multi-armed bandit problem," *IEEE Trans. Automat. Contr.*, vol. AC-31, pp. 576–577, 1986.
[24] R. R. Weber, "On the Gittins index for multi–armed bandits," *Ann. Appl. Probab.*, vol. 2, pp. 1024–1033, 1994.
[25] D. Bertsimas and J. Nino-Mora, "Conservation laws, extended polymatroid and multi-armed bandit problems: A unified approach to indexable systems," MIT, Tech. Rep., 1994.
[26] M. N. Katehakis and A. F. Veinott, "The multi-armed bandit problem: decomposition and computation," *Mathematics Operations Res.*, vol. 12, pp. 262–268, 1987.
[27] K. D. Glazebrook, "On the evaluation of suboptimal policies for families of alternative bandit processes," *J. Appl. Probab.*, vol. 19, pp. 716–722, 1982.
[28] ——, "Methods for the evaluation of permutations as strategies in stochastic scheduling," *Mgmt. Sci.*, vol. 29, pp. 1142–1155, 1983.
[29] J. S. Banks and R. K. Sundaram, "Switching costs and the Gittins index," *Econometrica*, vol. 62, pp. 687–694, May 1994.
[30] R. Agrawal, M. Hegde, and D. Teneketzis, "Asymptotically efficient adaptive allocation rules for the multi-armed bandit problem with switching cost," *IEEE Trans. Automat. Contr.*, vol. AC-33, pp. 899–906, October 1988.
[31] ——, "Multi-armed bandit problems with multiple plays and switching cost," *Stochastics Stochastic Rep.*, vol. 29, pp. 437–459, 1990.
[32] K. D. Glazebrook, "On stochastic scheduling with precedence relations and switching costs," *J. Appl. Probab.*, vol. 17, pp. 1016–1024, 1980.
[33] M. Asawa and D. Teneketzis, "Multi-armed bandits with switching penalties," Department of EECS, Univ. of Mich., Ann Arbor, MI, Control Group Rep. No. 94-01, Feb. 1994.
[34] J. Neveu, *Discrete-Parameter Martingales*. Amsterdam: North-Holland, 1975.
[35] L. Breiman, *Probability*. Philadelphia: SIAM, 1992.

Manjari Asawa received the B.E. degree in electronics and telecommunication engineering from Jabalpur University, India, in 1988, the M.E. degree in electrical communication engineering from the Indian Institute of Science, Bangalore, India in 1990, and the Ph.D. degree in electrical engineering systems from the University of Michigan, Ann Arbor, MI in 1995.

Since May 1995, she has been with the PATH program at the University of California, Berkeley. Her research interests include the modeling, performance analysis, and control of stochastic systems with emphasis on applications to wireless communication systems and high-speed networks.

Demosthenis Teneketzis (M'87) received the B.S. degree in electrical engineering from the University of Patras, Greece, in 1974 and the M.S., E.E., and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 1976, 1977, and 1979, respectively.

From 1979 to 1980, he worked for Systems Control Inc., Palo Alto, CA, and from 1980 to 1984 he was with Alphatech Inc., Burlington, MA. Since September, 1984, he has been with the University of Michigan, Ann Arbor, where he is a Professor of Electrical Engineering and Computer Science. In 1992, he was a Visiting Professor at the Institute for Signal and Information Processing of the Swiss Federal Institute of Technology (ETH), Zurich, Switzerland. His current research interests include stochastic control, decentralized systems, queueing and communication networks, stochastic scheduling and resource allocation problems, and discrete-event systems.