Extensions of the Multiarmed Bandit Problem: The Discounted Case

PRAVIN P. VARAIYA, FELLOW, IEEE, JEAN C. WALRAND, MEMBER, IEEE, AND CAGATAY BUYUKKOC

Abstract—**There are N independent machines. Machine i is described by a sequence $\{X^i(s), F^i(s)\}$ where $X^i(s)$ is the immediate reward and $F^i(s)$ is the information available before i is operated for the sth time. At each time one operates exactly one machine; idle machines remain frozen. The problem is to schedule the operation of the machines so as to maximize the expected total discounted sequence of rewards. An elementary proof shows that to each machine is associated an index, and the optimal policy operates the machine with the largest current index. When the machines are completely observed Markov chains, this coincides with the well-known Gittins index rule, and new algorithms are given for calculating the index. A reformulation of the bandit problem yields the tax problem, which includes, as a special case, Klimov's waiting time problem. Using the concept of superprocess, an index rule is derived for the case where new machines arrive randomly. Finally, continuous time versions of these problems are considered for both preemptive and nonpreemptive disciplines.**

I. Introduction

A. Background

In the basic version of the multiarmed bandit problem, there are N independent machines. Let $x_i(t)$ be the state of machine $i = 1, 2, \cdots, N$ at time $t = 1, 2, \cdots$. At each t one must operate exactly one machine. If machine i is selected, one gets an immediate reward $R(t) = R_i(x_i(t))$ and its state changes to $x_i(t+1)$ according to a stationary Markov transition rule; the states of the idle machines remain frozen, $x_j(t+1) = x_j(t)$, $j \neq i$. The states of all machines are observed, and the problem is to schedule the order in which the machines are operated so as to maximize the expected present values of the sequence of immediate rewards

$$E\sum_{t=1}^{\infty} a^t R(t) \tag{1.1}$$

where $0 < a < 1$ is a fixed discount factor. (In a subsequent paper the case $a = 1$ and the case of average reward per unit time will be considered.)

This problem has received considerable attention since it was first formulated in the 1940's, dynamic programming (DP) being the preferred framework for its analysis. The essential breakthrough came only in 1972 when Gittins and Jones [10] showed that to each machine i is attached an index $\nu_i(x_i(t))$ which is a function only of its state, and that the optimal policy operates the machine with the largest current index. Call this the index rule.

Manuscript received March 22, 1983; revised April 16, 1984. Paper recommended by Past Associate Editor, P. R. Kumar. This work was supported by the Office of Naval Research under Contract N00014-80-C-0507, by the Department of Energy under Contract DE-AC01-80RA50419, and by the National Science Foundation under Grant ECS-7903879-A02.

P. P. Varaiya and J. C. Walrand are with the Department of Electrical Engineering and Computer Sciences and the Electronics Research Laboratory, University of California, Berkeley, CA 94720.

C. Buyukkoc is with the AT&T Bell Laboratories,

This index result is significant, since it decomposes the N-dimensional bandit problem into N one-dimensional problems.

The index was subsequently [7], [8] shown to be

$$\nu_i(x_i) = \max_{\tau>1} \frac{E\left\{\sum_{t=1}^{\tau-1} a^t R_i(x_i(t)) \,|\, x_i(1) = x_i\right\}}{E\left\{\sum_{t=1}^{\tau-1} a^t \,|\, x_i(1) = x_i\right\}} \tag{1.2}$$

where the maximization is over the set of all stopping times $\tau > 1$. Gittins called this the dynamic allocation index (DAI) and interpreted it as the maximum expected reward per unit of expected discounted time. (In recognition of Gittins' contribution, we follow Whittle and call this the Gittins index.) One other interpretation of the Gittins index can also be given [10], [19].

Gittins and his associates did not use DP in their study. "Unfortunately," as Whittle [19] wrote, "[Gittins'] proofs of the optimality of the index rule have been difficult to follow, and this has doubtless been the reason why the full merits and point of his work have not yet been generally appreciated." Whittle then proceeded to supply an elegant proof using DP, and revealed the intimate connection between the (optimal) value function and the indexes of the N machines.

B. Structure of the Problem

Three features delimit the multiarmed bandit problem within the general class of stochastic control problems:

i) idle machines are frozen;
ii) frozen machines contribute no reward; and
iii) machine dynamics are Markovian.

As will be seen in Section II, properties i) and ii) almost trivially imply the optimality of the index rule. The Markovian property iii) is useful only in that it permits a simple calculation of the Gittins index as shown in Section IV. In retrospect, it seems that the Markovian property led researchers to adopt a DP framework, thereby obscuring the problem's simple structure.

C. The Tax Problem

In waiting time problems (see Section V), the reward structure is the "reverse" of the bandit problem. As before, exactly one of N machines can be operated at a time and the idle machines remain frozen. If i is operated at t, then one is charged a tax on the idle machines $C(t) := \Sigma_{j\neq i} C_j(x_j(t))$. The problem is to schedule the machines so as to minimize

$$E\sum_{t=1}^{\infty} a^t C(t) \tag{1.3}$$

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

At first sight, property ii) of the bandit problem appears to be violated. We will show nevertheless that the two problems are equivalent. In Section II it is shown that the optimal policy for the

0018-9286/85/0500-0426$01.00

tax problem is also an index rule determined by the index

$$\gamma_i(x_i) := \max_{\tau>1} \frac{E\{aC_i(x_i) - a^\tau C_i(x_i(\tau)) \mid x_i(1) = x_i\}}{E\left\{\sum_1^{\tau-1} a^t \mid x_i(1) = x_i\right\}}. \quad (1.4)$$

This version of the Gittins index can be interpreted as the maximum expected decrease in taxes per unit of expected discounted time.

D. Extensions

Section III is devoted to several extensions of the bandit and tax problems. In each case the optimal policy turns out to be an index rule, although the form of the index varies.

First, we consider the continuous time problem where, once a machine is operated, it cannot be idled until a certain "phase" is completed. This corresponds to a nonpreemptive discipline. Alternatively, one may view this as a natural extension of the discrete time Markov dynamics to the semi-Markov case treated in [21].

Second, we treat the continuous time problem where a machine may be idled at any time. This is the preemptive discipline.

Third, we consider the more significant extension to a superprocess [6], [7], [19]. Here, in addition to selecting the machine to be operated, one also chooses a control action. Under fairly restrictive conditions, similar to those given by Whittle [19], an index rule is shown to be optimal.

Finally, we consider the situation where new machines are being made available: if time is discrete, the new machines must form an i.i.d. sequence; if time is continuous, they must form a Poisson process. This situation is analyzed using the results on superprocesses.

E. Computation of the Index

As mentioned before, the results in Sections II and III on the optimality of the index rule do not require Markovian dynamics. In this general setting it is not easy to compute the index. However, when the machines evolve according to a finite state Markov chain, one can give algorithms to compute the index. Such algorithms are described in Section IV and are simpler than others published in the literature.

F. Applications

There is an extensive literature showing that the multiarmed bandit and its variants can be used to model the decision problems in job scheduling, resource allocation, sequential random sampling, clinical trials, investment in new products, random search, etc. See [1]–[4], [7], [15]–[18], [21] and the references listed therein. There is no need to review these applications here. It may be worth noting that since we do not assume Markovian dynamics, some new applications may be possible.

On the other hand, the tax problem formulated in Section I-C is novel. It was suggested by the important work of Klimov [13], [14]. In Section V we show how the index rule for the tax problem provides an optimal policy for a version of Klimov's problem.

II. Optimality of the Index Rule

A. Main Idea Illustrated

Since the simple idea underlying the proof might be obscured in the general case by the cumbersome notation, we illustrate it by the example of a deterministic two-armed bandit problem.

Suppose there are two deterministic machines, X and Y. If these were operated continually, they would respectively yield the sequences of immediate rewards

$$X(1),\ X(2),\ X(3),\ \cdots \quad \text{and} \quad Y(1),\ Y(2),\ Y(3),\ \cdots. \quad (2.1)$$

In analogy with (1.2), we define the index (at time 1) of these machines as

$$\nu_X := \max_{\tau>1} \frac{\sum_{t=1}^{\tau-1} a^t X(t)}{\sum_{t=1}^{\tau-1} a^t}, \quad \nu_Y := \max_{\tau>1} \frac{\sum_{t=1}^{\tau-1} a^t Y(t)}{\sum_{t=1}^{\tau-1} a^t}. \quad (2.2)$$

Suppose ν_X is realized at τ and that $\nu_X \geq \nu_Y$. It is easy to check that (2.2) implies

$$\sum_{\sigma}^{\tau-1} a^t X(t) \geq \nu_X \sum_{\sigma}^{\tau-1} a^t, \qquad 1 \leq \sigma < \tau, \quad (2.3)$$

$$\sum_{1}^{\sigma} a^t Y(t) \leq \nu_Y \sum_{1}^{\sigma} a^t, \qquad \sigma \geq 1. \quad (2.4)$$

Consider the sequence of immediate rewards obtained by using an arbitrary policy π. This sequence will be an interweaving of the two sequences in (2.1). Call it

$$Z(1),\ Z(2),\ Z(3),\ \cdots$$

and let T be the time when π operates machine X for the $(\tau - 1)$st time, so that $Z(T) = X(\tau - 1)$. (If π operates X fewer than $\tau - 1$ times, then $T = \infty$.) The Z sequence must be of the form

$$\begin{aligned} &Y(1),\ \cdots,\ Y(k_1),\ X(1),\ Y(k_1+1),\ \cdots,\ Y(k_2),\\ &X(2),\ \cdots,\ Y(k_{\tau-1}),\ X(\tau-1),\ Z(T+1),\ Z(T+2),\ \cdots. \end{aligned} \quad (2.5)$$

Next consider the policy $\tilde{\pi}$ which first operates the X machine $(\tau - 1)$ times, machine Y for $k_{\tau-1}$ times, and then follows policy π to yield the sequence

$$\begin{aligned} &X(1),\ \cdots,\ X(\tau-1),\ Y(1),\ \cdots,\ Y(k_{\tau-1}),\\ &Z(T+1),\ Z(T+2),\ \cdots. \end{aligned} \quad (2.6)$$

The present values of these policies are

$$\begin{aligned} V(\pi) :=& \sum_1^{k_1} a^t Y(t) + \cdots + a^{\tau-2} \sum_{k_{\tau-2}+1}^{k_{\tau-1}} a^t Y(t)\\ &+ \sum_1^{\tau-1} a^{k_t+t} X(t) + \sum_{T+1}^{\infty} a^t Z(t)\\ V(\tilde{\pi}) :=& \sum_1^{\tau-1} a^t X(t) + a^{\tau-1} \sum_1^{k_{\tau-1}} a^t Y(t) + \sum_{T+1}^{\infty} a^t Z(t). \end{aligned}$$

Then $V(\tilde{\pi}) - V(\pi) = \Delta_X - \Delta_Y$, where (with $k_0 := 0$)

$$\begin{aligned} \Delta_X :=& \sum_1^{\tau-1} (1 - a^{k_t}) a^t X(t)\\ =& \sum_1^{\tau-1} (a^{k_{t-1}} - a^{k_t}) \sum_{n=t}^{\tau-1} a^n X(n)\\ \geq& \nu_X \sum_1^{\tau-1} (a^{k_{t-1}} - a^{k_t}) \sum_{n=t}^{\tau-1} a^n, \qquad \text{by (2.3)}\\ =& \nu_X \sum_1^{\tau-1} (1 - a^{k_t}) a^t \end{aligned}$$

$$\Delta_Y := (1-a^{\tau-1})\sum_{1}^{k_1} a^t Y(t) + (a-a^{\tau-1})\sum_{k_1+1}^{k_2} a^t Y(t)$$

$$+ \cdots + (a^{\tau-2}-a^{\tau-1})\sum_{k_{\tau-2}+1}^{k_{\tau-1}} a^t Y(t)$$

$$= \sum_{1}^{\tau-1} (a^{t-1}-a^t)\sum_{n=1}^{k_t} a^n Y(n)$$

$$\le \nu_X \sum_{1}^{\tau-1} (a^{t-1}-a^t)\sum_{n=1}^{k_t} a^n, \qquad \text{by (2.4) and } \nu_X \ge \nu_Y$$

$$= \nu_X \sum_{1}^{\tau-1} (1-a^{k_t})a^t.$$

Hence, $V(\tilde{\pi}) \ge V(\pi)$.

Thus, it is better to follow the index rule until time $\tau - 1$. The argument can now be repeated starting at time τ. This proves the optimality of the index rule. Observe that the freezing property is needed to guarantee that the sequence (2.6) is feasible; property ii) (idle machines yield no reward) is used to compare the rewards obtained by any policy with the rewards due to the index rule.

B. Formulation of the Bandit and Tax Problems

Machine $i = 1, 2, \cdots, N$ is characterized by the pair of sequences

$$\{X^i(s), F^i(s)\}, \qquad s=1, 2, \cdots. \tag{2.7}$$

$X^i(s)$ is the (random) reward obtained when i is operated for the sth time. $F^i(s)$ is the σ-field representing the information about machine i gathered after it has been operated $(s-1)$ times. It is assumed that

i) $F^i(s) \subset F^i(s+1)$; let $F^i := V_s F^i(s)$ ($X^i(s)$ need not be adapted to $F^i(s)$);
ii) $V_s \sigma(X^i(s)) V F^i$, $i = 1, \cdots, N$, are independent;
iii) $E\sum_1^\infty a^t |X^i(t)| < \infty$, all i; here $0 < a < 1$ is a fixed discount factor.

At each time exactly one machine must be operated. Thus, $t = t^1 + \cdots + t^N$ where $t^i = t^i(t)$ is the number of times i is operated during $1, 2, \cdots, t$. t^i or $t^i(t)$ is called the ith machine time at process time t.

Consider the decision at time $t+1$. This must be based on the available information

$$F(t+1) := V_i F^i(t^i+1), \qquad t=0, 1, \cdots.$$

(Here and subsequently, $V_i F^i(t^i + 1)$ is shorthand for the following inductive definition. Let $i(t)$ be the machine operated at time t. Then $F(t+1) = F(t) V G(t)$, where $G(t)$ is the σ-field generated by sets of the form $\{i(t) = i\} \cap \{t^i(t) = s\} \cap A$, with $A \in F^i(s+1)$.) A policy is any sequence of decisions that satisfies this information constraint. The bandit problem is to find the policy π that maximizes

$$V(\pi) := E\left\{\sum_1^\infty a^t X^{i(t)}(t^{i(t)}(t)) \,|\, F(1)\right\} \tag{2.8}$$

where $i(t)$ is the machine operated at time t. The conditioning with respect to $F(1)$ will prove convenient later when a change of time origin will be used.

In the tax problem the data and assumptions are identical. The only difference is that $X^i(s)$ is interpreted as the tax that must be paid at time t if, after being operated $(s-1)$ times, machine i is idled at time t. The tax problem is to find the policy π that minimizes

$$W(\pi) := E\left\{\sum_{t=1}^\infty a^t \left[\sum_{i \ne i(t)} X^i(t^i(t)+1)\right] \,|\, F(1)\right\}. \tag{2.9}$$

C. Equivalence of the Problems

Consider any policy π. Let $l_i(s)$ be the process time when π operates i for the sth time; $l_i(0) := 0$. Then, for the bandit problem

$$V(\pi) = E\left\{\sum_i \sum_{s=1}^\infty a^{l_i(s)} X^i(s) \,|\, F(1)\right\}$$

and for the tax problem

$$W(\pi) = E\left\{\sum_i \sum_{s=1}^\infty [a^{l_i(s-1)+1} + \cdots + a^{l_i(s)-1}] X^i(s) \,|\, F(1)\right\}.$$

Suppose we wish to maximize $V(\pi)$. Define machines $\{Y^i(s), F^i(s)\}$ by

$$Y^i(s) := \sum_{r=0}^\infty a^r X^i(s+r).$$

Then,

$$X^i(s) = Y^i(s) - aY^i(s+1).$$

Simple algebraic manipulation leads to the form

$$\sum_{s=1}^\infty a^{l_i(s)} X^i(s) = aY^i(1) - (1-a)\sum_1^\infty [a^{l_i(s-1)+1} + \cdots + a^{l_i(s)-1}] Y^i(s).$$

Since $Y^i(1)$ fixed, it follows that maximization of $V(\pi)$ is equivalent to the tax problem:

$$\min\ E\left\{\sum_i \sum_s [a^{l_i(s-1)+1} + \cdots + a^{l_i(s)-1}] Y^i(s) \,|\, F(1)\right\}.$$

On the other hand, suppose we wish to minimize $W(\pi)$. Define machine $\{Z^i(s), F^i(s)\}$ by

$$Z^i(s) := X^i(s) - aX^i(s+1).$$

Then one gets

$$\sum_1^\infty [a^{l_i(s-1)+1} + \cdots + a^{l_i(s)-1}] X^i(s) = (1-a)^{-1}\left[aX^i(1) - \sum_1^\infty a^{l_i(s)} Z^i(s)\right]$$

and so the tax problem is equivalent to the bandit problem:

$$\max\ E\left\{\sum_i \sum_s a^{l_i(s)} Z^i(s) \,|\, F(1)\right\}.$$

D. The Index Rules

For the bandit problem, the index of machine i after it has been operated $(s-1)$ times is defined as

$$\nu_i(s) := \max_{\tau > s} \frac{E\left\{\sum_{t=s}^{\tau-1} a^t X^i(t) \,|\, F^i(s)\right\}}{E\left\{\sum_{t=s}^{\tau-1} a^t \,|\, F^i(s)\right\}} \tag{2.10}$$

where the maximization is over all stopping times $\infty \geq \tau > s$ of $\{F^i(\cdot)\}$. (Here, and in the sequel, "max" is to be interpreted as the more cumbersome phrase "ess sup" used in Appendix A.)

For the tax problem, the index of i after it has been operated $(s-1)$ times is defined as

$$\gamma_i(s) := \max_{\tau > s} \frac{E\{a^s X^i(s) - a^\tau X^i(\tau) \mid F^i(s)\}}{E\left\{\sum_{t=s}^{\tau-1} a^t \mid F^i(s)\right\}}. \tag{2.11}$$

The indexes in (2.10) and (2.11) are in conformity with the equivalence transformations introduced in the preceding section. Note also that if the machine dynamics are Markovian and stationary, then (2.10) reduces to (1.2), while (2.11) reduces to (1.4).

In Appendix A it is shown that there always exists τ achieving the maximum in (2.10). The *index rule* for either problem is the policy that operates the machine with the largest current index.

E. Optimality of the Index Rule

Because the two problems are equivalent, only the bandit problem is considered. The optimality is based on the following lemma, whose proof is relegated to Appendix B.

Let $X(t)$, $t = 1, 2, \cdots$, be a sequence of random variables on a probability space (Ω, F, P). Let $\{F(t)\}$ be an increasing family of sub-σ-fields of F, and suppose that $E\sum_1^\infty |X(t)| < \infty$.

Lemma 2.1: Suppose that

$$\max_{\tau > 1} E\left\{\sum_{t=1}^{\tau-1} X(t) \mid F(1)\right\} = 0$$

where the maximization is over all $\{F(\cdot)\}$ stopping times and suppose that τ^* achieves the maximum.

a) Then

$$E\left\{\sum_{t=1}^{\infty} \alpha(t) X(t) \mid F(1)\right\} \leq 0 \qquad \text{a.s.}$$

for all $\{F(\cdot)\}$-adapted random sequences $\{\alpha(t)\}$ such that

$$1 \geq \alpha(t) \geq \alpha(t+1) \geq 0 \qquad \text{a.s. for all } t.$$

b) Moreover,

$$E\left\{\sum_{t=1}^{\tau^*-1} \beta(t) X(t) \mid F(1)\right\} \geq 0 \qquad \text{a.s.}$$

for all $\{F(\cdot)\}$-adapted random sequences $\{\beta(t)\}$ such that

$$0 \leq \beta(t) \leq \beta(t+1) \leq 1 \qquad \text{a.s. for all } t.$$

Clearly, the results of the lemma will continue to hold if $\{\alpha(t)\}$ and $\{\beta(t)\}$ are adapted to the sequence $\{F(\cdot)VG\}$, where G is any field independent of F.

Before proving optimality of the index rule, we give two corollaries of Lemma 2.1. These corollaries are of independent interest; they are not used in the sequel.

Corollary 2.1: The index is nondecreasing in the discount rate.

Proof: Let $\{X(t), F(t)\}$, $t = 1, 2, \cdots$, be as in Lemma 2.1 and define for $a \in (0, 1]$

$$\nu(a) := \max_{\tau > 1} \frac{E\left\{\sum_{t=1}^{\tau-1} a^t X(t) \mid F(1)\right\}}{E\left\{\sum_{t=1}^{\tau-1} a^t \mid F(1)\right\}}.$$

Let $b \in (0, a)$. We show that $\nu(b) \leq \nu(a)$. One has

$$\max_{\tau > 1} E\left\{\sum_{t=1}^{\tau-1} a^t [X(t) - \nu(a)] \mid F(1)\right\} = 0.$$

Now,

$$E\left\{\sum_{t=1}^{\tau-1} b^t [X(t) - \nu(a)] \mid F(1)\right\}$$

$$= E\left\{\sum_{t=1}^{\tau-1} \left(\frac{b}{a}\right)^t a^t [X(t) - \nu(a)] \mid F(1)\right\}$$

$$= E\left\{\sum_{t=1}^{\infty} \alpha(t) a^t [X(t) - \nu(a)] \mid F(1)\right\} \leq 0$$

since $\alpha(t) := (b/a)^t 1\ (\tau > t)$ is such that $1 \geq \alpha(t) \geq \alpha(t+1) \geq 0$. Since τ is arbitrary, it follows that $\nu(b) \leq \nu(a)$. ■

Corollary 2.2: Let $\{\alpha(t), X(t), F(t)\}$, $t = 1, 2, \cdots$, be as in Lemma 2.1 and define for $a \in [0, 1]$

$$\nu_x = \max_{\tau > 1} \frac{E\left\{\sum_1^{\tau-1} a^t X(t) \mid F(1)\right\}}{E\left\{\sum_1^{\tau-1} a^t \mid F(1)\right\}}$$

$$\nu_{\alpha x} = \frac{E\left\{\sum_1^{\tau-1} a^t \alpha(t) X(t) \mid F(1)\right\}}{E\left\{\sum_1^{\tau-1} a^t \mid F(1)\right\}}.$$

If $\nu_x \geq 0$, then $\nu_x \geq \nu_{\alpha x}$.

Proof: One has $\max_{\tau > 1} E\{\sum_1^{\tau-1} a^t [X(t) - \nu_x] | F(1)\} = 0$. Therefore, by Lemma 2.1, for all $\tau > 1$,

$$E\left\{\sum_1^{\tau-1} a^t \alpha(t) [X(t) - \nu_x] \mid F(1)\right\} \leq 0,$$

so that

$$E\left\{\sum_1^{\tau-1} a^t \alpha(t) X(t) \mid F(1)\right\} \leq \nu_x E\left\{\sum_1^{\tau-1} \alpha(t) a^t \mid F(1)\right\}$$

$$\leq \nu_x E\left\{\sum_1^{\tau-1} a^t \mid F(1)\right\}.$$

Hence, $\nu_x \geq \nu_{\alpha x}$. ■

We now prove the optimality of the index rule. The main difficulty is one of notation. Consider the effect of any policy π from time t on. By a change of time origin we can set $t = 1$, so long as the information available from operating the machines up to time $t - 1$ is incorporated in the initial σ-fields $F^i(1)$. Let

$$Z(1), Z(2), \cdots$$

be the sequence of immediate rewards resulting from π. This sequence is an interweaving of the N sequences

$$X^i(1), X^i(2), \cdots, \qquad i = 1, \cdots, N.$$

Let $l_i(s)$ be the time when π operates machine i for the sth time. (If π operates machine i fewer than $s-1$ times, then $l_i(s) := \infty$.)

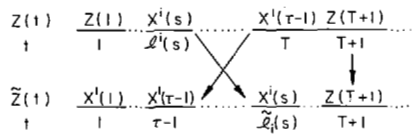

Fig. 1. Reward sequences Z, $\tilde{Z}$.

Then

$$t^i(t) = \max\ \{s \geq 0 \mid l_i(s) \leq t\}, \qquad Z(l_i(s)) = X^i(s),$$

$$V(\pi) := E\left\{\sum_1^\infty a^t Z(t) \mid F(1)\right\} = E\left\{\sum_{i=1}^N \sum_{s=1}^\infty a^{l_i(s)} X^i(s) \mid F(1)\right\}.$$

Suppose without loss of generality that machine 1 has the largest index,

$$\nu_1 := \nu_1(1) \geq \nu_i(1), \qquad \text{all } i, \tag{2.12}$$

and let it be achieved at the stopping time τ of $\{F^1(\cdot)\}$. Let

$$T := l_1(\tau - 1), \quad k_i := t^i(T),$$

so that $k_1 = \tau - 1$.

Let $\tilde{\pi}$ be the policy defined as follows:

a) operate machine 1 at time 1, 2, $\cdots$, $\tau - 1$,

b) operate machines $i \neq 1$ at time $\tau, \cdots, T$ in the same order as π, and

c) operate according to π at time $T + 1$, $T + 2$, $\cdots$.

See Fig. 1. It is readily seen that $\tilde{\pi}$ is a (feasible) policy. Let the resulting sequence of immediate rewards be

$$\tilde{Z}(1), \tilde{Z}(2), \cdots.$$

Then $\tilde{Z}(t) = Z(t)$, $t > T$. Let $\tilde{l}_i(s)$ be the time when $\tilde{\pi}$ operates i for the sth time. So $\tilde{l}_1(s) = s$ for $s = 1, \cdots, \tau - 1$. Then

$$\begin{aligned}
\Delta &:= V(\tilde{\pi}) - V(\pi) \\
&= E\left\{\sum_{t=1}^T a^t [\tilde{Z}(t) - Z(t)] \mid F(1)\right\} \\
&= E\left\{\sum_{i=1}^N \sum_{s=1}^{k_i} [a^{\tilde{l}_i(s)} - a^{l_i(s)}] X^i(s) \mid F(1)\right\} \\
&= E\left\{\sum_{s=1}^{\tau-1} [a^s - a^{l_i(s)}] X^1(s) - \sum_{j \neq 1} \sum_{s=1}^{k_i} [a^{l_i(s)} - a^{\tilde{l}_i(s)}] X^i(s) \mid F(1)\right\} \\
&= E\left\{\sum_{s=1}^{\tau-1} \beta_s a^s X^1(s) - \sum_{i \neq 1} \sum_{s=1}^{k_i} \alpha_s^i a^s X^i(s) \mid F(1)\right\}
\end{aligned}$$

where $\beta_s := 1 - a^{l_i(s)-s}$ and $\alpha_s^i := a^{l_i(s)-s} - a^{\tilde{l}_i(s)-s}$. Since $l_i(s + 1) \geq l_i(s) + 1$ and for $i \neq 1$, $\tilde{l}_i(s) \geq l_i(s)$ and $\tilde{l}_i(s + 1) - \tilde{l}_i(s) \leq l_i(s + 1) - l_i(s)$, one has a.s.

$$0 \leq \beta_s \leq \beta_{s+1} \leq 1, \quad 1 \geq \alpha_s^i \geq \alpha_{s+1}^i \geq 0.$$

Now,

$$\begin{aligned}
\Delta \geq\ & E\left\{\sum_{s=1}^{\tau-1} \beta_s a^s [X^1(s) - \nu_1] \mid F(1)\right\} \\
& - E\left\{\sum_{i \neq 1} \sum_{s=1}^{k_i} \alpha_s^i a^s [X^i(s) - \nu_i] \mid F(1)\right\} \\
& + \nu_1 E\left\{\sum_{s=1}^{\tau-1} \beta_s a^s - \sum_{i \neq 1} \sum_{s=1}^{k_i} \alpha_s^i a^s \mid F(1)\right\}. \qquad (2.13)
\end{aligned}$$

The last term in (2.13) is

$$\begin{aligned}
& \nu_1 E\left\{\sum_{s=1}^{\tau-1} [a^s - a^{l_1(s)}] - \sum_{i \neq 1} \sum_{s=1}^{k_i} [a^{l_i(s)} - a^{\tilde{l}_i(s)}] \mid F(1)\right\} \\
&= \nu_1 E\left\{\sum_{i=1}^N \sum_{s=1}^{k_i} [a^{\tilde{l}_i(s)} - a^{l_i(s)}] \mid F(1)\right\} \\
&= \nu_1 E\left\{\sum_{t=1}^T a^t - \sum_{t=1}^T a^t \mid F(1)\right\} = 0.
\end{aligned}$$

By Lemma 2.1b (with $X(t) := a^t[X^1(t) - \nu]$), the first term of (2.13) is nonnegative; by Lemma 2.1a (with $X(t) := a^t[X^i(t) - \nu_i]$), the second term of (2.13) is nonpositive. Therefore, $\Delta \geq 0$ a.s. Hence, $\tilde{\pi}$ is better than π.

Now $\tilde{\pi}$ coincides with the index rule over 1, 2, $\cdots$, $\tau - 1$. Since the initial time was arbitrary, Theorem 2.1 is proved.

Theorem 2.1: The index rules defined by (2.10) and (2.11) are optimal.

Remark 2.1: From the proof of Theorem 2.1, one can see that the index rule proceeds in "stages" as follows.

Stage 1: Calculate $\nu_1(1), \cdots, \nu_N(1)$. Suppose $\nu_i(1)$ is the largest and let it be achieved at time $\tau_i > 1$. Operate machine i for time 1, 2, $\cdots$, $\tau_i - 1$. At the end of stage 1, the process time is $T_1 := \tau_i - 1$.

Stage $k + 1$: Suppose T_k is the process time at the end of stage k and let the corresponding machine times be $S_k^i := t^i(T_k)$. Calculate the indexes $\nu_1(S_k^1 + 1), \cdots, \nu_N(S_k^N + 1)$. Suppose the jth index is the largest and let it be achieved at the stopping time $\tau_j > S_k^j + 1$. Operate machine j for time

$$T_k + 1, \cdots, T_k + (\tau_j - 1 - S_k^j) := T_{k+1}.$$

In words: at the end of each stage calculate all indexes, and operate the machine with the largest index for a time given by the corresponding optimal stopping time. This alternative construction of the index rule will be used in Section III-C.

III. Extensions

A. Continuous Time, Nonpreemptive

The data are slightly different. Machine $i = 1, \cdots, N$ is described by the triple

$$\{X^i(s),\ \sigma^i(s),\ F^i(s)\}, \qquad s = 1, 2, \cdots. \tag{3.1}$$

$X^i(s)$ is the instantaneous reward (or tax) as before. If i is operated for the sth time, it must be operated without interruption for the (random) time interval $\sigma^i(s)$. $F^i(s)$ is, as before, the information obtained after i has been operated $(s - 1)$ times. Assumptions i) and iii) of Section II-B are maintained. Assumption ii) is replaced by: the N σ-fields generated by $\{F^i, X^i(s), \sigma^i(s), s = 1, 2, \cdots\}$, $i = 1, \cdots, N$, are independent. It is not assumed that $\sigma^i(s)$ is adapted to $F^i(s)$ or $F^i(s + 1)$.

The discrete parameter $t = 1, 2, \cdots$ now denotes the process period number and $t^i = t^i(t)$ is the number of times i is operated during the first t periods. Let $i(t)$ be the machine operated during the tth period. Then the real (process) time at the end of period t is

$$\sigma(t) = \sigma^{i(1)}(t^{i(1)}(1)) + \cdots + \sigma^{i(t)}(t^{i(t)}(t)) \tag{3.2}$$

with $\sigma(0) := 0$.

With this additional notation the present value of rewards for the bandit problem is [cf. (2.8)]

$$V(\pi) := E\left\{\sum_{t=1}^\infty \int_{\sigma(t-1)}^{\sigma(t)} X^{i(t)}(t^{i(t)}) a^\tau\, d\tau \mid F(1)\right\}. \tag{3.3}$$

The integral gives the present value of rewards when $i(t)$ is operated during the tth period, discounted back to time 0. The case $\sigma^i(s) = 1$ reduces to the standard bandit problem of Section II-B.

The index of i after it has been operated $(s-1)$ times is now defined as [cf. (2.10)]

$$\nu_i(s) := \max_{\tau > s} \frac{E\left\{\sum_{t=s}^{\tau-1} a^{\sigma^i(s)+\cdots+\sigma^i(t-1)} X^i(t) \int_0^{\sigma^i(t)} a^r \, dr \,|\, F^i(s)\right\}}{E\left\{\int_0^{\sigma^i(s)+\cdots+\sigma^i(\tau-1)} a^r \, dr \,|\, F^i(s)\right\}} \tag{3.4}$$

where τ is any stopping time of $\{F^i(\cdot)\}$.

At the end of each period, the index rule operates the machine with the largest current index and for the associated period σ. The proof of the next result requires obvious changes in the proof of Theorem 2.1. For the case of discrete time semi-Markov processes, this result was obtained by Whittle [21].

Theorem 3.1: The present value given by (3.3) is maximized by the index rule defined by the index (3.4).

A similar result holds for the tax problem. The present value of the tax stream resulting from policy π is [cf. (2.9)]

$$W(\pi) := E\left\{\sum_{t=1}^{\infty} \int_{\sigma(t-1)}^{\sigma(t)} \sum_{i \neq i(t)} X^i(t^i(t)+1) a^r \, dr \,|\, F(1)\right\}.$$

The index of i after it has been operated $(s-1)$ times is now defined as [cf. (2.11)]

$$\gamma_i(s) := \max_{\tau > s} \frac{E\{X^i(s) - a^{\sigma^i(s)+\cdots+\sigma^i(\tau-1)} X^i(\tau) \,|\, F^i(s)\}}{E\left\{\int_0^{\sigma^i(s)+\cdots+\sigma^i(\tau-1)} a^r \, dr \,|\, F^i(s)\right\}}. \tag{3.5}$$

One can then show that the index rule defined by this index is optimal for the tax problem.

B. Continuous Time, Preemptive

Machine i is now characterized by the continuous parameter process

$$\{X^i(s), F^i(s)\}, \qquad s \geq 0.$$

$X^i(s)$ is the reward (or tax) process. $F^i(r) \subset F^i(s)$ for $r < s$. $F^i := V_s F^i(s)$, F^i and F^j are independent for $i \neq j$.

At any (process) time t, any machine may be operated. Let $t^i = t^i(t)$ denote the Lebesgue measure of the process time during which i is operated. Then the present value of a policy π is

$$V(\pi) := E\left\{\int_0^{\infty} a^t X^{i(t)}(t^{i(t)}(t)) \, dt \,|\, F(0)\right\}.$$

The index for machine i after it has been operated for time s is defined by

$$\nu_i(s) := \sup_{\tau > s} \frac{E\left\{\int_s^{\tau} a^t X^i(t) \, dt \,|\, F^i(s)\right\}}{E\left\{\int_s^{\tau} a^t \, dt \,|\, F^i(s)\right\}}. \tag{3.6}$$

The index rule is to operate at each t the machine with the largest current index.

To prove the optimality of the index rule, various additional technical assumptions must be made so that $i(t)$, $t^i(t)$, and (3.6) are well defined. In many cases one can construct a proof as follows. Fix $\epsilon > 0$, and restrict attention to policies π_ϵ which switch machines only at times $0, \epsilon, 2\epsilon, \cdots$. This is a standard bandit problem of Section II-B. Moreover,

$$\sup \, V(\pi_\epsilon) \leq \sup \, V(\pi_{\epsilon/2}) \leq \sup \, V(\pi).$$

A technical argument is now required to show that $\sup V(\pi) - \sup V(\pi_\epsilon) \to 0$ as $\epsilon \to 0$. The bandit problem for diffusions is analyzed in [11] by extending Whittle's dynamic programming argument.

An index rule for the continuous time tax problem can be derived in a similar way. The index for machine i after it has been operated for time s is given by

$$\gamma_i(s) := \sup_{\tau > s} \frac{E\{a^s X^i(s) - a^\tau X^i(\tau) \,|\, F^i(s)\}}{E\left\{\int_s^{\tau} a^t \, dt \,|\, F^i(s)\right\}}.$$

The index rule defined by this index minimizes the present value of taxes

$$W(\pi) := E\left\{\int_0^{\infty} \sum_{i \neq i(t)} a^t X^i(t^i(t)) \, dt \,|\, F(0)\right\}.$$

C. Superprocess

A superprocess is a collection of independent machines. For $i = 1, 2, \cdots, N$ let the ith superprocess be the collection X^i of standard machines $X^i = \{X^i(s), F^i(s)\}$. (The filtration $\{F^i(s)\}$ is machine-dependent.) For each selection $X^i \in X^i$, let $V^*(X^1, \cdots, X^N)$ be the maximum expected reward of the standard bandit problem associated with the machines $X^1, \cdots, X^N$. The bandit problem associated with the N superprocesses is to find $X^i \in X^i$ to

$$\max_{X^1 \cdots X^N} V^*(X^1, \cdots, X^N).$$

It is natural to expect that the selection of the optimal $(X^1, \cdots, X^N)$ will usually have to be jointly determined. Our aim is to give a condition which implies that the selection of the best machine in the ith superprocess can be made independently of the selection of the machine in the jth superprocess $j \neq 1$. The condition is a generalization of that given by Whittle [19] and involves the concept of machine domination which is introduced next.

For any machine $X = \{X(s), F^X(s)\}$ and $\nu \in R$, let

$$N(X, \nu) := \max_{\tau > 0} \; E\left\{\sum_{1}^{\tau-1} a^t [X(t) - \nu]\right\} \tag{3.7}$$

where τ ranges over all stopping times of $\{F^X(\cdot)\}$. Observe that $N(X, \nu) \geq 0$, since the sum in (3.7) vanishes for $\tau \equiv 1$ a.s.

Remark 3.1: Note that if τ_ν is the optimal stopping time for (3.7), and $\mu \leq \nu$, then one can find an optimal stopping time τ_μ such that $\tau_\mu \geq \tau_\nu$ a.s. (See the proof of Lemma 3.4 for a similar result.)

Observe also that if $\nu(1)$ is the index of machine X at time 1 and if $\tau^* > 1$ is the corresponding stopping time given by (2.10), then

$$N(X, \nu(1)) = 0$$

and this is achieved by τ^*.

Say that machine $X = \{X(s), F^X(s)\}$ *dominates* machine $Y = \{Y(s), F^Y(s)\}$ (for the bandit problem) if

$$N(X, \nu) \geq N(Y, \nu) \qquad \text{for all } \nu. \tag{3.8}$$

The interpretation of domination is as follows. Say that one may operate one of the machines X and Y up to some random time, after which one retires and receives the constant reward ν. Then X

dominates Y if and only if it is optimal to choose X over Y for any given value of ν. (See Whittle [19].)

Remark 3.2: If X dominates Y, then

$$E\left\{\sum_1^\infty a^t X(t)\right\} \geq E\left\{\sum_1^\infty a^t Y(t)\right\}.$$

This can be verified by letting $\nu \to -\infty$ in (3.7), (3.8).

Theorem 3.2: Suppose $X^i \in \mathcal{X}^i$ dominates every other $Y^i \in \mathcal{X}^i$. Then

$$V^*(X^1, \cdots, X^N) = \max_{Y^i \in \mathcal{X}^i} V^*(Y^1, \cdots, Y^N).$$

Thus, if each superprocess contains a dominating machine, then making the joint optimal selection over $\mathcal{X}^1 \times \cdots \times \mathcal{X}^N$ reduces to N decoupled optimization problems. The condition that there exists a dominating machine is quite restrictive.

The proof of Theorem 3.2 depends upon the crucial Lemma 3.2, which in turn requires the next instructive result.

For a machine $Z = \{Z(s), F^Z(s)\}$ define a sequence of $\{F^Z(\cdot)\}$ stopping times $\sigma_1 < \sigma_2 < \cdots$ and a sequence of index values ν_1, ν_2, $\cdots$ as follows.

Stage 1: Let $\nu_1 := \nu^Z(1)$ and suppose the index is attained by the stopping time $\sigma_1 + 1 > 1$. See (2.10); here ν^Z is the index of machine Z.

Stage $i+1$: Let $\nu_{i+1} := \nu^Z(\sigma_i + 1)$ and suppose it is attained at time $(\sigma_{i+1} + 1) > (\sigma_i + 1)$ (when finite).

Lemma 3.1: ν_i is measurable with respect to $F^Z(\sigma_{i-1} + 1)$($\sigma_0 := 0$). Also

$$\nu_i \geq \nu_{i+1}, \qquad \text{a.s.}$$

Proof: The first assertion is immediate from definition (2.10). Suppose $P\{\nu_{i+1} > \nu_i\} > 0$. Define

$$\sigma = \sigma_i \qquad \text{on } \{\nu_{i+1} \leq \nu_i\}$$
$$= \sigma_{i+1} \qquad \text{on } \{\nu_{i+1} > \nu_i\}.$$

It is easily seen that $\sigma + 1$ is a stopping time since $\{\nu_{i+1} > \nu_i\} \in F^Z(\sigma_i + 1)$. Moreover,

$$E\left\{\sum_{\sigma_{i-1}+1}^{\sigma} a^t Z(t) \,|\, F(\sigma_{i-1}+1)\right\}$$
$$= E\left\{\sum_{\sigma_{i-1}+1}^{\sigma_i} a^t Z(t) \,|\, F(\sigma_{i-1}+1)\right\} + E\left\{1\{\nu_{i+1} > \nu_i\} \cdot E\left\{\sum_{\sigma_i+1}^{\sigma_{i+1}} a^t Z(t) \,|\, F(\sigma_i+1)\right\} \,|\, F(\sigma_{i-1}+1)\right\}$$
$$= \nu_i E\left\{\sum_{\sigma_{i-1}+1}^{\sigma_i} a^t \,|\, F(\sigma_{i-1}+1)\right\} + E\left\{1\{\nu_{i+1} > \nu_i\}\nu_{i+1} \times E\left\{\sum_{\sigma_i+1}^{\sigma_{i+1}} a^t \,|\, F(\sigma_i+1)\right\} \,|\, F(\sigma_{i-1}+1)\right\}$$
$$> \nu_i E\left\{\sum_{\sigma_{i-1}+1}^{\sigma} a^t \,|\, F(\sigma_{i-1}+1)\right\}$$

with positive probability, which contradicts (2.10), and the proof is complete. ∎

Lemma 3.2: Let $X = \{X(s), F^X(s)\}$, $Y = \{Y(s), F^Y(s)\}$, $Z = \{Z(s), F^Z(s)\}$ be three machines. Suppose $\sigma(X)$, $\sigma(Y)$, $\sigma(Z)$ are independent where $\sigma(X) := V_s\sigma(X(s))V_s F^X(s)$, and $\sigma(Y)$, $\sigma(Z)$ are defined similarly. If X dominates Y, then

$$V^*(X, Z) \geq V^*(Y, Z).$$

Proof: By Theorem 2.1 $V^*(Y, Z)$ is attained by the corresponding index rule. Suppose the index rule leads to the sequence of immediate rewards,

$$Y(1), \cdots, Y(\lambda_1), Z(1), \cdots, Z(\sigma_1), Y(\lambda_1+1), \cdots, Y(\lambda_2),$$
$$Z(\sigma_1+1), \cdots, Z(\sigma_2), \cdots$$

where $\lambda_{i+1} \geq \lambda_i$ and $\sigma_{i+1} \geq \sigma_i$ are stopping times of $\{F^Y(\cdot)\}$ and $F^Z(\cdot)\}$, respectively. Then

$$V^*(Y, Z) = E\left\{\sum_1^{\lambda_1} a^t Y(t) + a^{\sigma_1} \sum_{\lambda_1+1}^{\lambda_2} a^t Y(t) + \cdots\right\}$$
$$+ E\left\{a^{\lambda_1} \sum_1^{\sigma_1} a^t Z(t) + a^{\lambda_2} \sum_{\sigma_1+1}^{\sigma_2} a^t Z(t) + \cdots\right\}. \tag{3.9}$$

According to Remark 2.1, we may assume that the interval $\sigma_i + 1, \cdots, \sigma_{i+1}$ is a stage in the implementation of the index rule. Let $\nu_i := \nu^Z(\sigma_{i-1} + 1)$. Then (2.10) and Lemma 3.1 respectively imply

$$E\left\{\sum_{\sigma_i+1}^{\sigma_i+1} a^t Z(t) \,|\, F^Z(\sigma_i+1)\right\} = \nu_{i+1} E\left\{\sum_{\sigma_i+1}^{\sigma_{i+1}} a^t \,|\, F^Z(\sigma_i+1)\right\} \tag{3.10}$$

$$\nu_{i+1} \leq \nu_i \qquad \text{a.s.}$$

We now specify in stages a policy for the bandit problem involving the two machines X, Z.

Stage 1: Calculate ν_1. Find the stopping time $(\tau + 1)$ of $\{F^X(\cdot)\}$ such that

$$N(X, \nu_1) = E\left\{\sum_1^{\tau_1} a^t[X(t) - \nu_1]\right\}.$$

Operate machine X for τ_1 times. Then operate machine Z for σ_1 times.

Stage $i+1$: Calculate ν_{i+1}. Find the stopping time $(\tau_{i+1} + 1)$ of $\{F^X(\cdot)\}$ such that

$$N(X, \nu_{i+1}) = E\left\{\sum_1^{\tau_{i+1}} a^t[X(t) - \nu_{i+1}]\right\}. \tag{3.11}$$

Because $\nu_{i+1} \leq \nu_i$ a.s., we may assume that $\tau_{i+1} \geq \tau_i$ a.s.; see Remark 3.1. Operate machine X for $(\tau_{i+1} - \tau_i)$ times. Then operate machine Z for $(\sigma_{i+1} - \sigma_i)$ times.

This policy results in the sequence of immediate rewards

$$X(1), \cdots, X(\tau_1), Z(1), \cdots, Z(\sigma_1), X(\tau_1+1), \cdots, X(\tau_2),$$
$$Z(\sigma_1+1), \cdots, Z(\sigma_2), \cdots$$

and so

$$V^*(X, Z) \geq E\left\{\sum_1^{\tau_1} a^t X(t) + a^{\sigma_1} \sum_{\tau_1+1}^{\tau_2} a^t X(t) + \cdots\right\}$$
$$+ E\left\{a^{\tau_1} \sum_1^{\sigma_1} a^t Z(t) + a^{\tau_2} \sum_{\sigma_1+1}^{\sigma_2} a^t Z(t) + \cdots\right\} \tag{3.12}$$

which will be compared with (3.9). We have $V^*(X, Z) - V^*(Y, Z) \geq \Delta_1 - \Delta_2$ where

$$\Delta_1 = E\left\{\left[\sum_1^{\tau_1} a^t X(t) - \sum_1^{\lambda_1} a^t Y(t)\right] + a^{\sigma_1}\left[\sum_{\tau_1+1}^{\tau_2} a^t X(t) - \sum_{\lambda_1+1}^{\lambda_2} a^t Y(t)\right] + \cdots\right\}$$

$$= E\left\{(1-a^{\sigma_1})\left[\sum_1^{\tau_1} a^t X(t) - \sum_1^{\lambda_1} a^t Y(t)\right] + (a^{\sigma_1} - a^{\sigma_2})\left[\sum_1^{\tau_2} a^t X(t) - \sum_1^{\lambda_2} a^t Y(t)\right] + \cdots\right\} \quad (3.13)$$

$$\Delta_2 = E\left\{(a^{\lambda_1} - a^{\tau_1})\sum_1^{\sigma_1} a^t Z(t)\right\} + E\left\{(a^{\lambda_2} - a^{\tau_2})\sum_{\sigma_1+1}^{\sigma_2} a^t Z(t) + \cdots\right\}. \quad (3.14)$$

The typical term in (3.13) is

$$E\left\{(a^{\sigma_{i-1}} - a^{\sigma_i})\, E\left\{\sum_1^{\tau_i} a^t X(t) - \sum_1^{\lambda_i} a^t Y(t) \mid F^Z(\sigma_i + 1)\right\}\right\}$$

$$\geq E\left\{(a^{\sigma_{i-1}} - a^{\sigma_i}) v_i \left(\sum_1^{\tau_i} a^t - \sum_1^{\lambda_i} a^t\right)\right\}$$

$$= \frac{a}{1-a} E\left\{(a^{\sigma_{i-1}} - a^{\sigma_i}) v_i (a^{\lambda_i} - a^{\tau_i})\right\}$$

using (3.11), and the hypothesis that X dominates Y; we also used the identity $a + \cdots + a^t = a(1-a)^{-1}(1-a^t)$. Hence,

$$\frac{1-a}{a}\Delta_1 \geq E(a^{\lambda_1} - a^{\tau_1})(1-a^{\sigma_1})v_1 + E(a^{\lambda_2} - a^{\tau_2})(a^{\sigma_1} - a^{\sigma_2})v_2 + \cdots. \quad (3.15)$$

Similarly, using (3.10) in (3.14), one finds

$$\frac{1-a}{a}\Delta_2 \leq E(a^{\lambda_1} - a^{\tau_1})(1-a^{\sigma_1})v_1 + E(a^{\lambda_2} - a^{\tau_2})(a^{\sigma_1} - a^{\sigma_2})v_2 + \cdots$$

which proves that $\Delta_1 - \Delta_2 \geq 0$ as required. ■

Corollary 3.1: Suppose X and Y are in X^1 and X dominates Y. Then for any machines $Y^2 \in X^2, \cdots, Y^N \in X^N$

$$V^*(X, Y^2, \cdots, Y^N) \geq V^*(Y, Y^2, \cdots, Y^N).$$

Proof: Consider any policy that attains $V^*(Y, Y^2, \cdots, Y^N)$ and let the corresponding sequence of immediate rewards be

$$Y(1), \cdots, Y(\lambda_1), Z(1), \cdots, Z(\sigma_1), Y(\lambda_1+1), \cdots, Y(\lambda_2), Z(\sigma_1+1), \cdots, Z(\sigma_2), \cdots$$

where $\{Z(s)\}$ is an interweaving of the reward sequences $\{Y^2(s)\}, \cdots, \{Y^N(s)\}$. We can certainly construct a machine $Z = \{Z(s), F^Z(s)\}$ where $Z(s)$ is as above and $F^Z(s)$ is the corresponding information σ-field. Then

$$V^*(Y, Y^2, \cdots, Y^N) = V^*(Y, Z).$$

Also $V^*(X, Z) \leq V^*(X, Y^2, \cdots, Y^N)$ since operating Z is more restrictive. By Lemma 3.2, $V^*(X, Z) \geq V^*(Y, Z)$ and the result is proved. ■

Proof of Theorem 3.2: Repeated applications of the corollary above give

$$V^*(Y^1, \cdots, Y^N) \leq V^*(X^1, Y^2, \cdots, Y^N) \leq \cdots \leq V^*(X^1, \cdots, X^N).$$ ■

Corollary 3.2: Suppose $X^i \in X^i$ dominates every $Y^i \in X^i$. Let $X = \{X(s), F(s)\}$ be the machine corresponding to $\{X^1, \cdots, X^N\}$ operated according to the index rule, let $Y = \{Y(s), F(s)\}$ be the machine corresponding to $\{Y^1, \cdots, Y^N\}$ with an arbitrary strategy. Then X dominates Y.

Proof: Five $\nu \in R$ and let Z be a machine always giving the reward 0. The maximum reward for the multiarmed bandit problem corresponding to the $N+1$ machines $\{X^1 - \nu, \cdots, X^N - \nu, Z\}$, with $X^i - \nu := \{X^i(s) - \nu, F^i(s)\}$ is $N(X, \nu)$. Similarly, $N(Y, \nu)$ is less than the maximum reward corresponding to $\{Y^1 - \nu, \cdots, Y^N - \nu, Z\}$. Now, if X^i dominates Y^i, then it is clear that $X^i - \nu$ dominates $Y^i - \nu$. Therefore, by Theorem 3.2, $N(X, \nu) \geq N(Y, \nu)$ and the proof is complete. ■

For the tax problem there is an analogous result, except that the definition of domination is different.

We say that $X = \{X(s), F^X(s)\}$ dominates $Y = \{Y(s), F^Y(s)\}$ for the tax problem if

$$\Gamma(X, \gamma) \geq \Gamma(Y, \gamma) \qquad \text{for all } \gamma$$

where, for a machine $Z = \{Z(s), F^Z(s)\}$,

$$\Gamma(Z, \gamma) := \max_{\tau > 1} E\left\{aZ(1) - a^t Z(\tau) - \gamma \sum_1^{\tau-1} a^t\right\}.$$

For the tax problem with machines $X^1, \cdots, X^N$, let $W^*(X^1, \cdots, W^N)$ be the minimum expected cost.

Theorem 3.3: Suppose $X^i \in X^i$ is such that X^i dominates every $Y^i \in X^i$. Then

$$W^*(X^1, \cdots, X^N) = \min_{X^1 \cdots X^N} W^*(Y^1, \cdots, Y^N).$$

Superprocesses arise in multiarmed bandit problems that involve controlled machines. Consider the "standard" discrete time problems of Section II-B with an additional degree of freedom: when a particular machine is operated, one must select also a control action that affects both the immediate reward and the machine "state transition." The control action is based on the available information about the machine and also about the others.

Consider each machine with all its possible "local" feedback laws as a superprocess $X^i = \{Y^i\}$. By "local" feedback law it is meant that the control actions are based only on the available information about that machine. It is shown below that if every superprocess X^i contains a dominating machine X^i, then the optimal strategy for the controlled multiarmed bandit problem is to operate these dominating machines according to the index rule. In particular, "local" feedback laws are optimal. It can be shown that this is not the case, in general, if there is no dominating machine for every X^i. To prove this optimality, we consider the case of two superprocesses $X^1 = \{Y^1\}$ and $X^2 = \{Y^2\}$. Thus, every $Y^i = \{Y^i(s), F^i(s)\}$ corresponds to a process with a control law $\gamma^i = \{\gamma_1^i, \gamma_2^i, \cdots\}$ that is $F^i(s)$ adapted. Fix an arbitrary strategy for the controlled multiarmed bandit process. At every time s the strategy defines whether to operate X^1 or X^2 and also which control action to take on the basis of the available information. Notice that the choice of the control action is equivalent to the choice of a local feedback law for the machine that is operated at time s on the basis of the information available about the other machine. For arbitrary feedback laws γ^1 for X^1 and γ^2 for X^2 and an arbitrary "switching rule" σ for selecting X^1 or X^2, denote by $A_n(\gamma^1, \gamma^2, \sigma)$ the event that the fixed strategy

for the controlled multiarmed bandit problem has been using γ^1, γ^2, and σ at least up to time n. To simplify the notation, let $\vartheta = (\gamma^1, \gamma^2, \sigma)$ and denote by Θ the collection of all possible ϑ's. It can be assumed that $\Sigma_{\vartheta\in\Theta} 1\{\omega \in A_n(\vartheta)\} = 1$ for all n. That is, policies that agree up to time n are identified up to time n. Then

$$Z_n = \sum_{\vartheta} 1\{\omega\in A_n(\vartheta)\}Z_n^{\vartheta}$$

where $\{Z_n^{\vartheta}\}$ would be the sequence of rewards corresponding to $(\gamma^1, \gamma^2, \sigma) = \vartheta$. Notice also that $A_{n+1}(\vartheta) \subset A_n(\vartheta) \in F^{Z^{\vartheta}}(n)$ for all ϑ, n. The assumption that there are dominating local feedback laws, say γ^{1*} for X^1 and γ^{2*} for X^2, implies by Corollary 3.2 that $\vartheta^* = (\gamma^{1*}, \gamma^{2*}$, index rule) is such that $\{Z_n^{\vartheta^*}\}$ dominates all the $\{Z_n^{\vartheta}\}$ for $\vartheta \in \Theta$. It then remains only to prove the following.

Lemma 3.3: Z^{ϑ^*} dominates Z.

Proof: Fix $\nu \in R$ and define

$$b^* = \max_{\tau>1} E\left\{\sum_1^{\tau-1} a^t[Z_t^{\vartheta^*} - \nu]\right\}$$

$$b^{\vartheta} = \max_{\tau>1} E\left\{\sum_1^{\tau-1} a^t[Z_t^{\vartheta} - \nu]\right\}.$$

Then $b^* \geq b^{\vartheta}$, $\vartheta \in \Theta$. Define $\tilde{Z}_n^{\vartheta} = Z_n^{\vartheta} - a^{-1}b^{\vartheta}1\{n = 1\}$, $\vartheta \in \Theta$. Then

$$\max_{\tau>1} E\left\{\sum_1^{\tau-1} a^t[\tilde{Z}_t^{\vartheta} - \nu]\right\} = 0.$$

Therefore, by Lemma 2.1,

$$\max_{\tau>1} E\left\{\sum_1^{\tau-1} a^t 1\{\omega\in A_t(\vartheta)\}[\tilde{Z}_t^{\vartheta} - \nu]\right\} \leq 0$$

so that

$$\max_{\tau>1} E\left\{\sum_1^{\tau-1} a^t 1\{\omega\in A_t(\vartheta)\}[Z_t^{\vartheta} - \nu] - b^*1\{\omega\in A_1(\vartheta)\}\right\} \leq 0$$

$$\max_{\tau>1} E\left\{\sum_1^{\tau-1} a^t 1\{\omega\in A_t(\vartheta)\}[Z_t - \nu] - b^*1\{\omega\in A_1(\vartheta)\}\right\} \leq 0.$$

Therefore, by summing over ϑ and using $\Sigma_{\vartheta} 1\{\omega \in A_t(\vartheta)\} = 1$,

$$\max_{\tau-1} E\left\{\sum_1^{\tau-1} a^t[Z_t - \nu]\right\} \leq b^*.$$

Hence, Z^{ϑ^*} dominates Z, as was to be shown. ■

Lemma 3.3 is valid for an arbitrary number of superprocesses.

D. Arm-Acquiring Bandits

We shall consider the discrete time bandit problem of Section II-B but, in addition we permit the arrival of new machines. Whittle [20] calls this an arm-acquiring bandit. To describe the model, the previous notation must be extended as follows.

There is now a potentially infinite number of machines $i = 1, 2, \cdots$. The ith machine $X^i = \{X^i(s), F^i(s)\}$ is described exactly as before. At time t, only a finite number of machines $i = 1, 2, \cdots, n(t)$ is available. These are the machines which either were available at time 1 or arrived during $1, \cdots, t-1$. Let $t^i(t)$, $i = 1, \cdots, n(t)$, be the number of times that i was operated during time $1, \cdots, t$. Thus, $t^i(t)$ is the ith machine time at process time t. The decision at $t+1$ is to be based on

$$F(t) := \bigvee_{i=1}^{n(t)} F^i(t_i(t)+1).$$

At time t a set $A(t)$ of new machines arrive. These are "new" in the sense that at t their machine times are zero. Let $|A(t)|$ denote the number of machines in $A(t)$. Then

$$n(t+1) = n(t) + |A(t)|$$

and at $t+1$ one may operate any machine $i = 1, \cdots, n(t+1)$. Here $n(1)$ is the number of machines available at time 1. In addition to the assumptions i)–iii) imposed at the beginning of Section II-B, we make the following assumption.

iv) For each t the set of random arrivals $A(t)$ is independent of the control actions taken during $1, \cdots, t$.

The assumption means essentially that the number and type of machines arriving in the future cannot be affected by the order in which machines were operated in the past. The assumption permits future arrivals to be dependent on past arrivals. This possibility will be removed later.

We convert this problem into one involving superprocesses.

To begin, suppose only one machine $X = \{X(s), F(s)\}$ is available at time 1. The arrival of new machines is described by the random sequence $\{A(t)\}$, $t = 1, 2, \cdots$. A policy π prescribes at each time t whether to operate machine X or to operate one of the machines that arrived before t. Each such policy will determine a sequence of immediate rewards and an associated sequence of information fields. We may regard this pair of sequences as a "standard" machine $X^{\pi} = \{X^{\pi}(s), F^{\pi}(s)\}$ of the type introduced in Section II-B; different policies will be associated with different machines. The set of all (feasible) policies can, in this way, equivalently be regarded as a set of possible machines, in other words, as a superprocess, say X. Note that the original machine X corresponds to a policy that does not operate any newly arrived machine. Hence, $X \in X$.

We want to show that X contains a dominating machine X^{π}. The following proposition will be useful.

Lemma 3.4: Let $Z = Z(s), F(s)\}$ be a machine. Consider

$$\max_{\tau>1} E\sum_1^{\tau-1} a^t Z(t)$$

and let τ be optimal. Let $\sigma > 1$ be any stopping time. Then

$$E\left\{1(\sigma<\tau)\sum_{\sigma}^{\tau-1} a^t Z(t)\right\} \geq 0 \geq E\left\{1(\sigma>\tau)\sum_{\tau}^{\sigma-1} a^t Z(t)\right\}.$$

The result continued to hold if σ is allowed to be a stopping time of $F(s)VG$, where G is any σ-field independent of F.

Proof: Let $N = E\Sigma_1^{\tau-1} a^t Z(t)$. Then

$$N = E\left\{1(\sigma>\tau)\sum_{\sigma}^{\tau-1} a^t Z(t)\right\} + E\left\{1(\sigma\geq\tau)\sum_{1}^{\tau-1} a^t Z(t)\right\}$$

$$+ E\left\{1(\sigma<\tau)\sum_{1}^{\sigma-1} a^t Z(t)\right\}$$

$$= E\left\{1(\sigma<\tau)\sum_{\sigma}^{\tau-1} a^t Z(t)\right\}$$

$$+ E\sum_{1}^{\delta-1} a^t Z(t), \qquad \delta := \min(\sigma, \tau).$$

Since $E\Sigma_1^{\delta-1} a^t Z(t) \leq N$, the first inequality is proved. Also,

$$N = E\left\{1(\sigma \leq \tau) \sum_1^{\tau-1} a^t Z(t)\right\} + E\left\{1(\sigma > \tau) \sum_1^{\sigma-1} a^t Z(t)\right\}$$

$$- E\left\{1(\sigma > \tau) \sum_\tau^{\sigma-1} a^t Z(t)\right\}$$

$$= E \sum_1^{\lambda-1} a^t Z(t) - E\left\{1(\sigma > \tau) \sum_\tau^{\sigma-1} a^t Z(t)\right\},$$

$$\lambda := \max\ (\sigma, \tau).$$

Since $E\Sigma_1^{\lambda-1} a^t Z(t) \leq N$, the second inequality is proved. ■

For any policy π and number ν, let

$$N(\pi, \nu) := \max_{\tau > 1}\ E \sum_1^{\tau-1} a^t [X^\pi(t) - \nu]$$

where τ is a stopping time of $\{F^\pi(\cdot)\}$. Let

$$N(\nu) := \max_\pi\ N(\pi, \nu) = \max_\pi\ \max_{\tau > 1}\ E \sum_1^{\tau-1} a^t [X^\pi(t) - \nu] \tag{3.16}$$

and let $\pi(\nu)$, $\tau(\nu)$ be optimal for (3.16). We assume that $\pi(\nu)$ exists.

Then X^π dominates every machine in X if $N(\pi, \nu) = N(\nu)$ for all ν [see (3.8)].

Fix two numbers $\mu < \nu$.

Lemma 3.5: There exists a policy π which agrees with $\pi(\nu)$ during $1, \cdots, \tau(\nu) - 1$ such that $N(\pi, \mu) = N(\pi(\mu), \mu) = N(\mu)$.

Proof: Denote the reward sequence during $1, \cdots, \tau(\nu) - 1$ corresponding to $\pi(\nu)$ by

$$Z(1) - \nu,\ Z(2) - \nu,\ \cdots,\ Z(\tau(\nu) - 1) - \nu \tag{3.17}$$

and the reward sequence during $1, \cdots, \tau(\mu) - 1$ corresponding to $\pi(\mu)$ by

$$Y(1) - \mu,\ \cdots,\ Y(\sigma_1) - \mu,\ Z(1) - \mu,\ Y(\sigma_1 + 1) - \mu,\ \cdots,\ Y(\sigma_2) - \mu,$$
$$Z(2) - \mu,\ Y(\sigma_2 + 1) - \mu,\ \cdots,\ Z(k-1) - \mu,$$
$$Y(\sigma_{k-1} + 1) - \mu,\ \cdots,\ Y(\sigma_k) - \mu. \tag{3.18}$$

In the sequence (3.18) the $Z(i)$ denote the rewards which explicitly appear in (3.17). Hence, $k - 1 \leq \tau(\nu) - 1$. By Lemma 3.4, and since $\mu < \nu$,

$$0 \leq E \sum_k^{\tau(\nu)-1} a^t [Z(t) - \nu] < E \sum_k^{\tau(\nu)-1} a^t [Z(t) - \mu].$$

Hence, if $k \neq \tau(\nu)$, the policy which gives the reward sequence

$$Y(1) - \mu,\ \cdots,\ Y(\sigma_1) - \mu,\ Z(1) - \mu,\ \cdots,\ Y(\sigma_k) - \mu,$$
$$Z(k) - \mu,\ \cdots,\ Z(\tau(\nu) - 1) - \mu$$

will give a larger reward than $\pi(\mu)$, which is not possible since $\pi(\mu)$ is optimal. Hence, we may assume that $k = \tau(\nu)$ in (3.18), and in particular $\tau(\mu) \geq \tau(\nu)$.

Next, consider the policy ρ which up to the stopping time $\tau := \tau(\mu)$ gives the reward sequence

$$Z(1) - \mu,\ \cdots,\ Z(k-1) - \mu,\ Y(1) - \mu,\ \cdots,\ Y(\sigma_k) - \mu \tag{3.19}$$

and after that it agrees with the policy $\pi(\mu)$. Assumption iv) guarantees the feasibility of π. Also π agrees with $\pi(\nu)$ during $1, \cdots, \tau(\nu) - 1$. Since $N(\mu) = N(\pi(\mu), \mu)$,

$$0 \geq N(\pi, \mu) - N(\pi(\mu), \mu)$$

$$= E\left\{\sum_{i=1}^{k-1} a^i [Z(i) - \mu] + \sum_{j=1}^{\sigma_k} a^{k-1+j} [Y(j) - \mu]\right\}$$

$$- E\left\{\sum_{j=1}^{\sigma_1} a^j [Y(j) - \mu] + \cdots + \sum_{j=\sigma_{k-1}+1}^{\sigma_k} a^{k-1+j} [Y(j) - \mu]\right.$$

$$\left. + \sum_{i=1}^{k-1} a^{\sigma_i + i} [Z(i) - \mu]\right\}$$

$$= E\left\{\sum_{i=1}^{k-1} a^i [Z(i) - \nu] + \sum_{j=1}^{\sigma_k} a^{k-1+j} [Y(j) - \nu]\right\}$$

$$- E\left\{\sum_{j=1}^{\sigma_1} a^j [Y(j) - \nu] + \cdots + \sum_{j=\sigma_{k-1}+1}^{\sigma_k} a^{k-1+j} [Y(j) - \nu]\right.$$

$$\left. + \sum_{i=1}^{k-1} a^{\sigma_i + i} [Z(i) - \nu]\right\}$$

$$= E\left\{\sum_{i=1}^{k-1} (1 - a^{\sigma_i}) a^i [Z(i) - \nu]\right\}$$

$$- E\left\{\sum_{i=1}^{k-1} \sum_{j=\sigma_{i-1}+1}^{\sigma_i} (a^{i-1} - a^{k-1}) a^j [Y(j) - \nu]\right\}$$

$$=: \Delta_Z - \Delta_Y.$$

By Lemma 3.4,

$$E\left\{\sum_{i=1}^{k-1} a^i [Z(i) - \nu]\right\} = 0 = \max_\tau\ E\left\{\sum_{i=1}^{\tau-1} a^i [Z(i) - \nu]\right\}.$$

Using this in Lemma 2.1b (with $\beta_i = 1 - a^{\sigma_i}$) shows $\Delta_Z \geq 0$. Also, by Lemma 3.4,

$$E\left\{\sum_{i=1}^{k-1} \sum_{j=\sigma_{i-1}+1}^{\sigma_i} a^j [Y(j) - \nu]\right\} = E \sum_{i=1}^{\sigma_{k-1}} a^i [Y(i) - \nu] \leq 0.$$

Using this in Lemma 2.1a shows $\Delta_Y \leq 0$. The proof is complete. ■

Theorem 3.4: There exists a policy π such that X^π dominates every machine in X.

Proof: Let $\nu_1 > \nu_2 > \cdots \to -\infty$. By Lemma 3.4 there exist policies $\pi(\nu_i)$ and stopping times $\tau(\nu_i) \to \infty$ a.s. (since the terms in 3.7 become positive) such that $\pi(\nu_{i+1})$ agrees with $\pi(\nu_i)$ during $1, \cdots, \tau(\nu_i) - 1$. Then the unique strategy π which extends all the $\pi(\nu_i)$ is the required policy. ■

We now return to the bandit problem with arrivals introduced at the beginning of this section. In addition to assumption i)–iv), we impose the following.

v) $A(t)$, $t = 1, 2, \cdots$, is a sequence of i.i.d. random variables.

At time t consider the ith machine X^i, after it has been operated $s - 1 = t^i(t)$ times. (X^i may be the machine available at time 1 or any machine that arrived before t.) This machine, together with the arrival process $\{A(\cdot)\}$, defines a superprocess $X^i(s)$. We define the index $\nu_i(s)$ of X^i to be the index of the dominant

machine in $X^i(s)$. More directly,

$$\nu_i(s) := \max_{\pi} \max_{\tau > s} \frac{E\left\{\sum_{s}^{\tau-1} a^t X^\pi(t) \mid F^i(s)\right\}}{E\left\{\sum_{s}^{\tau-1} a^t \mid F^i(s)\right\}}. \tag{3.20}$$

Assumption v) guarantees that the index depends only on the machine type i, the machine time s, and the law of the arriving process, not on the process time t.

Theorem 3.5: For the bandit problem with arrivals, it is optimal to operate at each time the available machine with the largest current index.

Proof: At any process time one is faced with the superprocesses X^i, $i = 1, \cdots, n(t)$. By Theorem 3.4 the dominant machine in X^i has index (3.20). By Theorem 3.3 it is sufficient to restrict attention to these dominant machines, but then Theorem 2.1 guarantees optmality of the index rule. ■

E. Tax Problem with Arrivals

The setup is exactly as in the arm-acquiring bandit problem, except that $X^i(s)$ is the tax when machine i is idle. We study this by transforming it into an equivalent bandit problem as in Section II-B. The details are sufficiently different to require a separate treatment.

The cost of a policy π is

$$W(\pi) = E\left\{\sum_{t=1}^{\infty} a^t \left[\sum_{i \neq i(t)}^{n(t)} X^i(t^i(t)+1)\right]\right\}$$

$$= E\left\{\sum_{i=1}^{\infty}\sum_{s=1}^{\infty} [a^{l_i(s-1)+1} + \cdots + a^{l_i(s)-1}] X^i(s)\right\}$$

where

$$l_i(s) := 0 \quad \text{if machine } i \text{ is available at time 1}$$

$$= \text{the process time when machine } i \text{ arrived, otherwise.}$$

Define new machines Z^i by $Z^i(s) = X^i(s) - aX^i(s+1)$, in terms of which

$$W(\pi) = (1-a)^{-1} E\left\{\sum_{i=1}^{\infty} a^{l_i(0)+1} X^i(1) - \sum_{i=1}^{\infty}\sum_{s=1}^{\infty} a^{l_i(s)} Z^i(s)\right\}$$

so that the tax problem is equivalent to the arm-acquiring bandit problem with the machines Z^i.

Thus, the index for machine X^i in the tax problem after it has been operated $s - 1 = t^i(t)$ times is

$$\gamma_i(s) := \max_{\pi} \max_{\tau > s} \frac{E\left\{\sum_{s}^{\tau-1} a^t Z^\pi(t) \mid F^i(s)\right\}}{E\left\{\sum_{s}^{\tau-1} a^t \mid F^i(s)\right\}} \tag{3.21}$$

where

$$Z^\pi(t) := X^i(t^i(t)+1) - aX^i(t^i(t)+2).$$

Note that, since π may operate different machines, the sum in the numerator in (3.21) does not collapse as in (2.11).

Theorem 3.6: For the tax problem with arrivals, an optimal policy is given by the index rule defined by (3.21).

Remark 3.2: The indexes given by (3.20) and (3.21) are much more difficult to compute than those given by (2.10) and (2.11), where no arrivals are considered.

It should be clear that Theorems 3.4, 3.5, and 3.6 generalize in the obvious way to the situation where time is continuous and the discipline is preemptive or nonpreemptive, as in Sections III-A and III-B. Assumption v) must now be read to mean that new machines arrive in a Poisson stream.

IV. Calculating the Index

In this section we develop algorithms for calculating the various indexes in the case where the machine is described by a finite state Markov chain.

A. Discrete Time Bandit Problem

Let $x(s)$, $s = 1, 2, \cdots$, be a Markov chain with state space $\{1, 2, \cdots, K\}$. Let $r(i)$ be the reward when $x(t) = i$. Suppose the state is observed. Then one has the standard machine $\{X(s), F(s)\}$ where

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

From (2.10) we see that if $x(s) = i$, then the corresponding index $\nu(s) = \nu_i$ where

$$\nu_i = \max_{\tau > 1} \frac{E_i\left\{\sum_{t=1}^{\tau-1} a^t r(x(t))\right\}}{E_i\left\{\sum_{t=1}^{\tau-1} a^t\right\}}. \tag{4.1}$$

Here $E_i f := E\{f \mid x(1) = i\}$, and τ ranges over all stopping times of $\{x(\cdot)\}$. We wish to calculate ν_i, $i = 1, 2, \cdots, K$.

Lemma 4.1: Suppose $\nu_1 \geq \nu_2 \geq \cdots \geq \nu_K$. Then an optimal stopping time for (4.1) is

$$\tau_i = \min\ \{t > 1 \mid x(t) \notin \{1, \cdots, i\}\}.$$

For a direct proof see Gittins [7, p. 154]; alternatively one can give a slight modification of the proof of Lemma 3.1. The same arguments also give the following.

Lemma 4.2: Suppose $\nu_1 \geq \nu_2 \geq \cdots \geq \nu_K$. Then an optimal stopping time for (4.1) is

$$\tau_i = \min\ \{t > 1 \mid x(t) \notin \{1, \cdots, i-1\}\}.$$

We use these results to find in sequence the state with the largest, second largest, third largest index, etc. Let $P = \{P_{ij}\}$ denote the $K \times K$ transition matrix of the chain $\{x(t)\}$.

Theorem 4.1: Suppose $\nu_1 \geq \nu_2 \geq \cdots \geq \nu_{m-1}$ for some m. Then

$$\nu_{i*} = \max_{i \geq m} \nu_i = \max \frac{\alpha_i^m}{\beta_i^m} = \frac{\alpha_{i*}^m}{\beta_{i*}^m}$$

where $\alpha^m = (\alpha_i^m, \cdots, \alpha_K^m)^T$, $\beta^m = (\beta_i^m, \cdots, \beta_K^m)^T$ are given by

$$\alpha^m := a[I - aP^m]^{-1} r, \quad \beta^m := a[I - aP^m]^{-1}\mathbf{1}$$

with

$$P_{ij}^m = \begin{cases} P_{ij} & j < m \\ 0 & j \geq m \end{cases}$$

$$r := (r(1), \cdots, r(K))^T, \quad \mathbf{1} := (1, \cdots, 1)^T.$$

Proof: Suppose $\nu_m = \max_{i \geq m} \nu_i$. By Lemma 4.2

$$\nu_m = \frac{E_m\left\{\sum_{1}^{\tau-1} a^t r(x(t))\right\}}{E_m\left\{\sum_{1}^{\tau-1} a^t\right\}}$$

with

$$\tau = \min\ \{t > 1 \mid x(t) \notin \{1, \cdots, m-1\}\}.$$

Hence,

$$\alpha_i^m := E_i \sum_1^{\tau-1} a^t r(x(t)) = ar(i) + a \sum_{j<m} P_{ij}\alpha_j^m$$

$$\beta_i^m := E_i \sum_1^{\tau-1} a^t = a + a \sum_{j<m} P_{ij}\beta_j^m$$

which concludes the proof. ∎

B. Continuous Time, Nonpreemptive Bandit Problem

Let $\psi(t)$, $t \geq 0$, be a continuous parameter, right-continuous pure jump process with jump times $0 = T_0 < T_1 < \cdots$ such that $\{x(s) := \psi(T_{s-1})\}$, $s = 1, 2, \cdots$, is a Markov chain with values in $\{1, \cdots, K\}$ and $K \times K$ probability transition matrix P.

Let $\sigma(s) := T_s - T_{s-1}$. Let $r(i)$ be the reward when $\psi(t) = i$. The nonpreemptive discipline means that a machine must be operated until its next jump time. In terms of the notation of Section III-A, this gives an abstract machine $\{X(s), \sigma(s), F(s)\}$ where $X(s) := r(x(s))$, $F(s) := \sigma\{x(i), \sigma(i-1); i \leq s\}$ is the information available after the machine has been operated for $(s-1)$ periods.

Finally, it is assumed that the conditional distribution of $\sigma(s)$ given $F(s)$ depends only on $x(s)$. In other words, $\psi(t)$ is a semi-Markov process. Let

$$b_i := E\{a^{\sigma(s)} \mid x(s) = i\}.$$

From (3.4) we see after evaluating the integrals that if $x(s) = i$, the corresponding index $\nu(s) = \nu_i$ where

$$\nu_i := \max_{\tau>1} \frac{E_i\left\{\sum_{s=1}^{\tau-1} a^{\sigma(1)+\cdots+\sigma(s-1)}[1 - a^{\sigma(s)}]r(x(s))\right\}}{E_i\left\{\sum_{s=1}^{\tau-1} a^{\sigma(1)+\cdots+\sigma(s-1)}[1 - a^{\sigma(s)}]\right\}}$$

where $E_i f := E\{f \mid x(1) = i\}$.

As in the preceding section, one obtains the following result.

Theorem 4.2: Suppose $\nu_1 \geq \cdots \geq \nu_{m-1}$. Let

$$\tau := \min\ \{s > 1 \mid x(s) \notin \{1, \cdots, m-1\}\}.$$

Then

$$\max_{i \geq m} \nu_i = \max_{i \geq m} \frac{\alpha_i^m}{\beta_i^m}$$

where

$$\alpha_i^m := E_i \sum_1^{\tau-1} a^{\sigma(1)+\cdots+\sigma(s-1)}[1 - a^{\sigma(s)}]r(x(s))$$

$$= (1 - b_i)r(i) + b_i \sum_{j<m} P_{ij}\alpha_j^m$$

$$\beta_i^m := E_i \sum_1^{\tau-1} a^{\sigma(1)+\cdots+\sigma(s-1)}[1 - a^{\sigma(s)}]$$

$$= (1 - b_i) + b_i \sum_{j<m} P_{ij}\beta_j^m.$$

C. Discrete Time Tax Problem

Since the equivalence of this problem to the discrete time bandit problem is established in Section II-C, the index can be written easily as

$$\gamma_i := \max_{\tau>1} \frac{E_i\left\{\sum_{t=1}^{\tau-1} a^t(c(x(t)) - ac(x(t+1))\right\}}{E_i\left\{\sum_{t=1}^{\tau-1} a^t\right\}}$$

under the same conditions as Section IV-A, except that $c(i)$ now denotes the cost per unit time when $x(t) = i$. The algorithms developed in the preceding section apply to this case with obvious modifications.

V. An Application

Consider a network of queues indexed $i = 1, \cdots, K$. A single server is to be assigned to service jobs in any queue. If this server is allocated to a job in queue i, that job must be completed before the server may be reassigned. In other words, the service discipline is nonpreemptive. A job in queue i requires a random amount of service time $\sigma(i)$. All service times are independent, and service times for jobs in the same queue are identically distributed.

Once a job in a queue i is completed, then with a fixed "routing" probability P_{ij} the job joins queue j, and with probability P_{i0} it leaves the network.

Let $n_i(t)$ be the number of customers waiting in queue i at time t. (The job being serviced is not counted in the n_i.) Let $c(i) > 0$ be constants. Consider the problem of assigning the single server to the jobs in such a way as to minimize the waiting cost

$$E \int_0^\infty a^t \sum_i c(i)n_i(t)\ dt. \tag{5.1}$$

This semi-Markov decision problem can readily be recast as a tax problem. One associates to each job a machine $X = \{X(s), \sigma(s), F(s)\}$ in the following manner. Suppose that after $(s-1)$ service completions the job is in queue $x(s) \in \{1, \cdots, N\}$. If the job leaves the network after $(s-1)$ service completions, let $x(s) = 0$. Let $F(s) := \sigma\{x(1), \cdots, x(s)\}$; let $\sigma(s)$ have the same distribution as $\sigma(x(s))$ if $x(s) > 0$, $\sigma(0) = 0$ if $x(s) = 0$. The reformulation as a tax problem is complete if one interprets assignment of the single server to a job as the operation of the corresponding machine.

Observe that $\{x(s)\}$ is a Markov chain with absorbing state 0 and $(K+1) \times (K+1)$ transition matrix P. One defines an index as in (3.5). If $x(s) = i$, the index is

$$\gamma_i = \max_{\tau>1} \frac{E_i\{c(x(1)) - a^{\sigma(x(1))+\cdots+\sigma(x(\tau-1))}c(x(\tau))\}}{E_i\left\{\int_0^{\sigma(x(1))+\cdots+\sigma(x(\tau-1))} a^r\ dr\right\}}, \quad i = 1, \cdots, K$$

$$= 0, \qquad i = 0 \tag{5.2}$$

where $E_i f := E\{f \mid x(1) = i\}$ and $c(0) := 0$. Note that $\gamma_i > 0$ for $i > 0$. (The index (5.2) can be calculated using the algorithm given in the preceding section.) Theorem 3.9 now gives the following result.

Theorem 5.1: The index rule defined by the index (5.2) minimizes the waiting cost (5.1).

The problem discussed above was motivated by the work of Klimov [13], [14]. Klimov permits Poisson arrivals, and he minimizes the average waiting cost per unit time.

VI. Conclusions

The multiarmed bandit problem is perhaps the simplest nontrivial problem in stochastic control for which a reasonably complete analysis is available. Most previous investigations of this problem were conducted within the framework of dynamic

programming. That framework has tended to hide the essential structure of the problem. In this paper the problem was studied using what, following Gittins [7], might be called a "forwards induction" argument. That argument has allowed us to dispense with the restrictions to Markovian dynamics and to complete state observations. Removal of these restrictions may increase the range of applications.

The paper also proposes a more general formulation of superprocesses. These are bandit problems in which a control variable is present. Further study of superprocesses may reveal an interesting class of applications.

Finally, the paper formulates a new class of problems which we have called the tax problem. In the discounted case considered here, the tax and bandit problems are equivalent; they are not equivalent when there is no discount. In situations involving allocation of a single resource where waiting costs are significant, the tax problem appears to provide a more convenient model.

Appendix A

Let $\{X(s), F(s)\}$ be a machine with

$$E\sum_{1}^{\infty} a^t |X(t)| < \infty. \tag{7.1}$$

Let T be the set of all stopping times τ, $s < \tau \leq \infty$. We will show that there exists τ^* in T that achieves

$$\gamma(s) := \operatorname{ess\,sup}_{\tau \in T} \frac{E\left\{\sum_{t=s}^{\tau-1} a^t(X) \mid F(s)\right\}}{E\left\{\sum_{t=2}^{\tau-1} a^t \mid F(s)\right\}}. \tag{7.2}$$

Let $Y(t) := X(t) - \gamma(s)$. Then,

$$\operatorname{ess\,sup}_{\tau \in T} E\left\{\sum_{t=s}^{\tau-1} a^t Y(t) \mid F(s)\right\} = 0 \quad \text{a.s.}$$

and the problem is equivalent to finding τ^* in T such that

$$E\left\{\sum_{t=s}^{\tau^*-1} a^t Y(t) \mid F(s)\right\} = 0 \quad \text{a.s.} \tag{7.3}$$

But this problem is very similar to Snell's problem discussed in Neveu [22]. Our problem is simpler, since

1) the associated martingale is uniformly integrable because of (7.1);
2) we permit τ^* to be infinite.

The following result can be proved in a way parallel to the proof of Proposition VI-1-3 in [22].

Theorem 7.1: The stopping rule τ^* defined below is optimal for (7.2) and (7.3):

$$\tau^* := \min\left\{t > s \mid \operatorname{ess\,sup}_{\tau \geq t} E\left\{\sum_{r=t}^{\tau-1} a^r[X(r) - \gamma(s)] \mid F(t)\right\} \geq 0\right\}$$

$$+\infty \text{ if } \operatorname{ess\,sup}_{\tau \geq t} E\left\{\sum_{r=t}^{\tau-1} a^r[X(r) - \gamma(s)] \mid F(t)\right\} < 0$$

for all $t > s$.

Appendix B

The proof of Lemma 2.1 is given below.

Define for $n = 1, 2, \cdots$,

$$A(n) := \left\{\omega \mid E\left[\sum_{n}^{\infty} \alpha(t)X(t) \mid F(n)\right] > 0\right\}.$$

Set $\alpha(n)/\alpha(n-1) := 0$ on $\{\alpha(n-1) = 0\}$. Notice that since $\alpha(n)/\alpha(n-1) \in [0, 1]$, one has

$$\frac{\alpha(n)}{\alpha(n-1)} E\left[\sum_{n}^{\infty} \frac{\alpha(t)}{\alpha(n)} X(t) \mid F(n)\right]$$

$$\leq 1(A(n)) E\left[\sum_{n}^{\infty} \frac{\alpha(t)}{\alpha(n)} X(t) \mid F(n)\right] \tag{7.4}$$

where $1(A(n))$ is the indicator of $A(n)$.

From (7.4) one finds for $n > 1$, and using $Z(n) := 1(A(1) \cap \cdots \cap A(n))$,

$$\varphi(n-1) := E\left[\sum_{t=1}^{n-1} Z(t)X(t) + Z(n-1)\sum_{t=n}^{\infty} \frac{\alpha(t)}{\alpha(n-1)} X(t) \mid F(1)\right]$$

$$= E\left\{\sum_{t=1}^{n-1} Z(t)X(t) + Z(n-1)\frac{\alpha(n)}{\alpha(n-1)} \cdot E\left[\sum_{t=n}^{\infty} \frac{\alpha(t)}{\alpha(n)} X(t) \mid F(n)\right] \mid F(1)\right\}$$

$$\leq E\left\{\sum_{t=1}^{n-1} Z(t)X(t) + Z(n-1)1(A(n)) \cdot E\left[\sum_{t=n}^{\infty} \frac{\alpha(t)}{\alpha(n)} X(t) \mid F(n)\right] \mid F(1)\right\}$$

$$= E\left[\sum_{t=1}^{n-1} Z(t)X(t) + Z(n)X(n) + Z(n)\sum_{t=n+1}^{\infty} \frac{\alpha(t)}{\alpha(n)} X(t) \mid F(1)\right]$$

$$= \varphi(n).$$

Therefore,

$$\varphi(1) = E\left[\sum_{t=1}^{\infty} Z(1)\frac{\alpha(t)}{\alpha(1)} X(t) \mid F(1)\right\}$$

$$\leq \lim_{n \to \infty} \varphi(n) = E\left[\sum_{t=1}^{\infty} 1(A(1) \cap \cdots \cap A(t))X(t) \mid F(1)\right]$$

Assume now that $E[\Sigma_1^{\infty} \alpha(t)X(t)|F(1)] > 0$ on some set of positive measure. Then, on that set, $\varphi(1) > 0$, which implies that

$$0 < E\left[\sum_{t=1}^{\infty} 1(A(1) \cap \cdots \cap A(t))X(t) \mid F(1)\right]$$

$$= E\left[\sum_{t=1}^{\tau-1} X(t) \mid F(1)\right] \tag{7.5}$$

where τ is the $\{F(\cdot)\}$-stopping time defined by

$$\tau = n \text{ on } A(1) \cap \cdots \cap A(n-1) \cap A^c(n).$$

But (7.5) contradicts the assumption of Lemma 2.1. This proves part a).

Part b) follows by observing that

$$0 = E\left[\sum_{t=1}^{\tau^*-1} X(t) \mid F(1)\right] = E\left[\sum_{t=1}^{\tau^*-1} (1-\beta(t))X(t) \mid F(1)\right] + E\left[\sum_{t=1}^{\tau^*-1} \beta(t)X(t) \mid F(1)\right]$$

and that by part a), the first term on the right-hand side of this equality must be nonpositive.

ACKNOWLEDGMENT

The authors are grateful to anonymous reviewers for pointing out previous errors, and for many excellent suggestions for improving this paper.

REFERENCES

[1] K. D. Glazebrook, "Scheduling tasks with exponential service times on parallel processors," *J. Appl. Prob.*, vol. 16, pp. 685–689, 1979.
[2] ——, "Stoppable families of alternative bandit processes," *J. Appl. Prob.*, vol. 16, pp. 843–854, 1979.
[3] ——, "On randomized dynamic allocation indices for sequential design of experiments," *J. Roy. Statis. Soc.*, vol. 42, pp. 342–346, 1980.
[4] ——, "On stochastic scheduling with precedence relations and switching costs," *J. Appl. Prob.*, vol. 17, pp. 1016–1024, 1980.
[5] ——, "On the evaluation of suboptimal policies for families of alternative bandit processes," *J. Appl. Prob.*, vol. 19, pp. 716–722, 1982.
[6] ——, "On a sufficient condition for superprocesses due to Whittle," *J. Appl. Prob.*, vol. 19, pp. 99–110, 1982.
[7] J. C. Gittins, "Bandit processes and dynamic allocation indices," *J. Roy. Statist. Soc.*, vol. 41, pp. 148–177, 1979.
[8] J. C. Gittins and K. D. Glazebrook, "On Bayesian models in stochastic scheduling," *J. Appl. Prob.*, vol. 14, pp. 556–565, 1977.
[9] ——, "On single machine scheduling with precedence relation and linear or discounted costs," *Oper. Res.*, vol. 29, pp. 161–173, 1981.
[10] J. C. Gittins and D. M. Jones, "A dynamic allocation index for the sequential design of experiments," in *Progress in Statistics, Euro. Meet. Statis.*, 1972, vol. 1, J. Gani, K. Sarkadi, and I. Vince, Eds. New York: North-Holland, 1974, pp. 241–266.
[11] I. Karatzas, "Gittins indices in the dynamic allocation problem for diffusion processes," Dep. Math. Statis., Columbia Univ., New York, NY, preprint, 1982.
[12] F. P. Kelly, "Multi-armed bandits with discount factor near one: The Bernoulli case," *Ann. Statis.*, vol. 9, pp. 987–1001, 1981.
[13] G. P. Klimov, "Time sharing service systems I," *Theory Prob. Appl. (USSR)*, vol. 19, pp. 532–551, 1974.
[14] ——, "Time sharing service systems II," *Theory Prob. Appl.*, vol. 23, pp. 314–321, 1978.
[15] P. Nash and J. C. Gittins, "A Hamiltonian approach to optimal stochastic resource allocation," *Adv. Appl. Prob.*, vol. 9, pp. 55–68, 1977.
[16] L. Rodman, "On the many-armed bandit problem," *Ann. Prob.*, vol. 6, pp. 491–498, 1978.
[17] D. Wahrenberger, C. Antle, and L. Klimko, "Bayesian rules for the two-armed bandit problem," *Biometrika*, vol. 64, pp. 172–174, 1977.
[18] M. L. Weitzman, "Optimal search for the best alternative," *Econometrica*, vol. 47, pp. 641–654, 1979.
[19] P. Whittle, "Multi-armed bandits and the Gittins index," *J. Roy. Statis. Soc.*, vol. 42, pp. 143–149, 1980.
[20] ——, "Arm-acquiring bandits," *Ann. Prob.*, vol. 9, pp. 284–292, 1981.
[21] ——, *Optimization over Time*, vol. 1 New York: Wiley, 1982.
[22] J. Neveu, *Discrete-Parameter Martingales*. New York: North-Holland, 1975.
[23] K. D. Glazebrook, "Optimal strategies for families of alternative bandit processes," *IEEE Trans. Automat. Contr.*, vol. 28, pp. 858–861, 1983.

Pravin P. Varaiya (M'68–SM'78–F'80) received the B.S. degree in electrical engineering from the University of Bombay, Bombay, India, and the Ph.D. degree from the University of California, Berkeley.

He is currently a Professor of Electrical Engineering and Economics at University of California, Berkeley. He was a Guggenheim Fellow in 1972 and a Miller Research Professor in 1978. His research interests include stochastic control, queueing networks, power systems, and urban economics.

Jean C. Walrand (S'71–M'74) was born in Belgium in 1951. He received the Ingenieur degree in electronics from the Université de Liège, Liège, Belgium, in 1974, and the Ph.D. degree in electrical engineering from the University of California, Berkeley, in 1979.

He was with the School of Electrical Engineering, Cornell University, Ithaca, NY, from 1979 to 1981. He is now with the Department of Electrical Engineering and Computer Sciences, Berkeley. His research interests are in communication networks, stochastic control, and decentralized systems.

Cagatay Buyukkoc was born in Turkey on April 23, 1954. He received the B.S. and M.S. degrees in electrical engineering from the Middle East Technical University, Ankara, Turkey, and the Ph.D. degree in electrical engineering and computer science from the University of California, Berkeley, in 1984.

He joined the Performance Analysis Department of AT&T Bell Laboratories in November 1984. His research interests are in queueing systems and stochastic control theory.