EXTENSIONS OF THE MULTI-ARMED BANDIT PROBLEM

by

P. Varaiya, J. Walrand and C. Buyukkoc

Memorandum No. UCB/ERL M83/14

8 March 1983

EXTENSIONS OF THE MULTI-ARMED BANDIT PROBLEM

by

P. Varaiya, J. Walrand and C. Buyukkoc

Memorandum No. UCB/ERL M83/14

8 March 1983

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

Extensions of the Multi-armed Bandit Problem [1]

P. Varaiya, J. Walrand and C. Buyukkoc

Department of Electrical Engineering & Computer Sciences
and Electronics Research Laboratory
University of California, Berkeley CA 94720

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 when i is operated for the s^{th} time. At each time one must operate 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 variation of the reward structure for the bandit problem defines the more general tax problem and its index rule includes as a special case the well-known work of Klimov on waiting time problems. 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 pre-emptive and nonpre-emptive disciplines.

8 March 1983

[1] Research supported by Office of Naval Research Contract N00014-80-C-0507 and Department of Energy Contract DE-AC01-80RA50419.

Extensions of the Multi-armed Bandit Problem [1]

P. Varaiya, J. Walrand and C. Buyukkoc

Department of Electrical Engineering & Computer Sciences

and Electronics Research Laboratory

University of California, Berkeley CA 94720

1. Introduction

1.1. Background

In the basic version of the multi-armed bandit problem there are N independent machines. Let $x_i(t)$ be the state of machine i = 1, 2,.., N at time t = 1, 2,... . 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 value of the sequence of immediate rewards

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

where $0 < a \leq 1$ is a fixed discount factor. (It is generally assumed that $a < 1$. Such a restriction is not imposed here, and the case $a = 1$ may be interesting in some applications.)

[1] Research supported by Office of Naval Research Contract N00014-80-C-0507 and Department of Energy Contract DE-AC01-80RA50419.

This problem has received considerable attention since it was first formulated in the 1940s, 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.

This index result is extremely important since it converts the N dimensional bandit problem into N one dimensional problems.

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

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

where the maximization is over all stopping times $\tau > 1$. This was called the dynamic allocation index (DAI) and interpreted as the maximum expected reward per unit of discounted time. One other interpretation can also be given [10, 19].

Gittins and his co-workers 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 indices of the N machines.

1.2. Structure of the problem

Three features delimit the multi-armed 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 2, 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 DAI as shown in Section 4. In retrospect, it seems that the Markovian property led researchers to adopt a DP framework thereby obscuring the problem's simple structure.

1.3. The tax problem

Consider the problem in which 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 $C(t) := \sum_{j \neq i} C_j(x_j(t))$. The problem is to schedule the machines so as to minimize

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

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

At first sight property (ii) of the bandit problem appears to be violated here in a decisive way. We will show nevertheless that when $a < 1$ the two problems are equivalent. On the other hand, when $a = 1$ and $X^i(t) \geq 0$ for all i and t, the bandit problem is trivial since the present value (1.1) is essentially independent of the order in which machines are operated. This is not the case for the tax problem. In Section 2 it is shown that the optimal policy for the tax problem is also an index rule determined by the index

$$\gamma_i(x_i) := \max_{\tau > 1} \frac{E\{ a\, C_i(x_i) - a^\tau\, C_i(x_i(\tau)) \mid x_i(1) = x_i \}}{E\{ \sum_{1}^{\tau-1} a^t \mid x_i(1) = x_i \}}. \tag{1.4}$$

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

1.4. Extensions

Section 3 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 problem where time is continuous, but once a machine is operated it cannot be idled until a certain "phase" is completed. This corresponds to a nonpre-emptive discipline. Alternatively, one may view this as an extension of the discrete time Markov dynamics to the semi-Markov case.

Second, we treat the case where time is continuous and a machine may be idled at any time. This is the pre-emptive discipline.

Third, we consider the extension to what is called 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.

1.5. Computation of the index

As mentioned before, the results in Sections 2 and 3 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 4 and are simpler than others published in the literature.

1.6. Applications

There is an extensive literature showing that the multi-armed 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 1.3 is novel. It was suggested by the important work of Klimov [13, 14]. In Section 5 we show how the index rule for the tax problem provides an optimal policy for Klimov's problem. We also present an algorithm for computing the index which is simpler than the method proposed by Klimov.

2. Optimality of the index rule

2.1. Main idea illustrated

Since the simple essential idea of 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 and $a < 1$. If these were operated continually, they would respectively yield the sequences of immediate rewards

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

Following (1.2) we define the index (at time 1) of these machines as

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

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

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

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

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

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

and let T be the time when π operates machine X for the $(\tau-1)^{st}$ time, so that $Z(T) = X(\tau-1)$. The Z sequence takes the form

$$Y(1),\ldots, Y(k_1), X(1), Y(k_1+1),\ldots, Y(k_2), X(2),\ldots, Y(k_{\tau-1}), X(\tau-1), \\ Z(T+1), Z(T+2),\ldots \tag{2.5}$$

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

$$X(1),\ldots, X(\tau-1), Y(1),\ldots, Y(k_{\tau-1}), Z(T+1), Z(T+2), \ldots. \tag{2.6}$$

The present values of these policies are

$$V(\pi) := \sum_{1}^{k_1} a^t Y(t) + \ldots + a^{\tau-2} \sum_{k_{\tau-2}+1}^{k_{\tau-1}} a^t Y(t) + \sum_{1}^{\tau-1} a^{k_t+t} X(t) + \sum_{T+1}^{\infty} a^t Z(t),$$

$$V(\tilde{\pi}) := \sum_{1}^{\tau-1} a^t X(t) + a^{\tau-1} \sum_{1}^{k_{\tau-1}} a^t Y(t) + \sum_{T+1}^{\infty} a^t Z(t).$$

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

$$\Delta_X := \sum_{1}^{\tau-1} (1 - a^{k_t}) a^t X(t)$$

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

$$\geq \nu_X \sum_{1}^{\tau-1} (a^{k_{t-1}} - a^{k_t}) \sum_{n=t}^{\tau-1} a^n, \quad \text{by (2.3)}$$

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

$$\Delta_Y := (1 - a^{\tau-1}) \sum_{1}^{k_1} a^t\, Y(t) + (a - a^{\tau-1}) \sum_{k_1+1}^{k_2} a^t\, Y(t) + \ldots + (a^{\tau-2} - a^{\tau-1}) \sum_{k_{\tau-2}+1}^{k_{\tau-1}} a^t\, Y(t)$$

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

$$\leq \nu_X \sum_{1}^{\tau-1} (a^{t-1} - a^t) \sum_{n=1}^{k_t} a^n, \quad \text{by (2.4) and } \nu_X \geq \nu_Y$$

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

Hence $V(\tilde{\pi}) \geq V(\pi)$.

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 and the index rule.

2.2. Formulation of the bandit and tax problems

Machine i = 1, 2,.., N is characterized by the pair of sequences

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

$X^i(s)$ is the (random) reward obtained when i is operated for the s^{th} 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 $F^i := \bigvee_s F^i(s)$; ($X^i(s)$ need not be adapted to $F^i(s)$)

(ii) F^i and F^j are independent for $i \neq j$;

(iii) $E \sum_{1}^{\infty} a^t\, |X^i(t)| < \infty$, all i; here $0 < a \leq 1$ is a fixed discount factor.

At each time exactly one machine must be operated. Thus, $t = t^1 + \ldots + t^N$

where $t^i = t^i(t)$ is the number of times i is operated during 1, 2, . . , t. t^i or $t^i(t)$ is called the i^{th} machine time at time t.

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

$$F(t) := \bigvee_i F^i(t^i+1), \quad t = 1, 2, \ldots .$$

A policy is any sequence of decisions that satisfies this information constraint. The bandit problem is to find the policy π that maximizes

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

where i(t) is the machine operated at time t.

In the tax problem the data and assumptions are identical. The only difference is that X^i is interpreted as the tax that must be paid if machine i is idle. The tax problem is to find the policy π that minimizes

$$W(\pi) := E \{ \sum_{t=1}^\infty a^t [\sum_{i \neq i(t)} X^i(t^i(t)+1)] \mid F(1) \}. \tag{2.9}$$

2.3. Equivalence of the problems when a < 1

Suppose a < 1, and consider any policy π. Let $l_i(s)$ be the (process) time when π operates i for the s^{th} time. Then, for the bandit problem,

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

and for the tax problem,

$$W(\pi) = E \{ \sum_i \sum_{s=1}^\infty [a^{l_i(s-1)+1} + \ldots + a^{l_i(s)-1}] X^i(s) \mid F(1) \}, \quad l_i(0) := 0.$$

See Figures 1, 2.

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,

$$X^i(s) = Y^i(s) - a\, Y^i(s+1).$$

Some algebraic manipulation leads to the form

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

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

$$\min E \{ \sum_{i} \sum_{s} [a^{l_i(s-1)+1} + \ldots + a^{l_i(s)-1}]\, Y^i(s) \mid F(1) \}.$$

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

$$Z^i(s) := X^i(s) - a\, X^i(s+1).$$

Then one gets

$$\sum_{1}^{\infty} [a^{l_i(s-1)+1} + \ldots + a^{l_i(s)-1}]\, X^i(s) = (1-a)^{-1} [a\, X^i(1) - \sum_{1}^{\infty} a^{l_i(s)}\, Z^i(s)]$$

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

$$\max E \{ \sum_{i} \sum_{s} a^{l_i(s)}\, Z^i(s) \mid F(1) \}.$$

2.4. The index rules

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

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

where the maximization is over all stopping times $\infty \geq \tau > s$ of $\{ F^i(.) \}$.

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

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

One should observe that the indices 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, then (2.10) reduces to (1.2), while (2.11) reduces to (1.4).

The index rule for either problem is the policy that operates the machine with the largest current index.

2.5. Optimality of the index rule for a < 1

Because the two problems are equivalent only the bandit problem is considered. The optimality is based on the following simple proposition (cf (2.3)).

Lemma 2.1

Suppose τ is optimum in (2.10). Let $\sigma > s$ be any $\{ F^i(.) \}$ stopping time. Then

$$\frac{E\{ 1(\sigma<\tau) \sum_\sigma^{\tau-1} a^t X^i(t) \mid F^i(s) \}}{E\{ 1(\sigma<\tau) \sum_\sigma^{\tau-1} a^t \mid F^i(s) \}} \geq \nu_i(s) \text{ a.s.}$$

Proof

Clearly,

$$\begin{aligned} 0 &= E\{ \sum_s^{\tau-1} a^t [X^i(t) - \nu_i(s)] \mid F^i(s) \} \\ &= E\{ 1(\sigma<\tau) \sum_\sigma^{\tau-1} a^t [X^i(t) - \nu_i(s)] \mid F^i(s) \} \\ &+ E\{ 1(\sigma\geq\tau) \sum_s^{\tau-1} a^t [X^i(t) - \nu_i(s)] \mid F^i(s) \} \\ &+ E\{ 1(\sigma<\tau) \sum_s^{\sigma-1} a^t [X^i(t) - \nu_i(s)] \mid F^i(s) \}. \end{aligned}$$

Let $\delta := \min(\sigma, \tau)$. Then the sum of the last two terms equals

$$E\{\sum_{s}^{\delta-1} a^t [X^i(t) - \nu_i(s)] \mid F^i(s)\} \le 0, \text{ by (2.10).}$$

and the proof is concluded. ∎

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

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

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

$$X^i(1), X^i(2), \ldots \quad i = 1, \ldots, N.$$

Let $l_i(s)$ be the time when π operates machine i for the s^{th} time. Then

$$t^i(t) = \max\{ s \ge 0 \mid l_i(s) \le t \}, \; Z(l_i(s)) = X^i(s),$$
$$V(\pi) := E\{\sum_{1}^{\infty} a^t Z(t) \mid F(1)\} = E\{\sum_{i=1}^{N} \sum_{s=1}^{\infty} a^{l_i(s)} X^i(s) \mid F(1)\}.$$

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

$$\nu_1(1) \ge \nu_i(1), \text{ all i,} \tag{2.12}$$

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

$$T := l_1(\tau-1), \quad k_i := t^i(T), \text{ so that } k_1 = \tau-1.$$

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

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

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

(c) operate according to π at time T+1, T+2,....

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

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

Then $\tilde{Z}(t) = Z(t)$, $t > T$. Let $\tilde{l}_i(s)$ be the time when $\tilde{\pi}$ operates i for the s^{th} time. So $\tilde{l}_{1_A}^{(s)} = s$ for $s = 1,\ldots, \tau-1$. Then

$$\begin{aligned}
\Delta &:= V(\tilde{\pi}) - V(\pi) \\
&= E\{ \sum_{t=1}^{T} a^t [\tilde{Z}(t) - Z(t)] \mid F(1) \} = E\{ \sum_{i=1}^{N} \sum_{s=1}^{k_i} [a^{\tilde{l}_i(s)} - a^{l_i(s)}] X^i(s) \mid F(1) \} \\
&= E\{ \sum_{s=1}^{\tau-1} [a^s - a^{l_1(s)}] X^1(s) - \sum_{i\neq 1} \sum_{s=1}^{k_i} [a^{l_i(s)} - a^{\tilde{l}_i(s)}] X^i(s) \mid F(1) \} \\
&= E\{ \sum_{s=1}^{\tau-1} b_s^1 [a^s X^1(s) + \ldots + a^{\tau-1} X^1(\tau-1)] \mid F(1) \} \\
&\quad - E\{ \sum_{i\neq 1} \sum_{s=1}^{k_i} b_s^i [a X^i(1) + \ldots + a^s X^i(s)] \mid F(1) \} \qquad (2.13)
\end{aligned}$$

where

$$b_s^1 = a^{-s} [a^{l_1(s-1)+1} - a^{l_1(s)}] \geq 0,$$

since $l_1(s) \geq l_1(s-1) + 1$, and

$$b_s^i = a^{-s-1} [(a^{l_i(s)+1} - a^{l_i(s+1)}) - (a^{\tilde{l}_i(s)+1} - a^{\tilde{l}_i(s+1)})] \geq 0,$$

since $\tilde{l}_i(s) \geq l_i(s)$ and $\tilde{l}_i(s+1) - \tilde{l}_i(s) \leq l_i(s+1) - l_i(s)$.

Using Lemma 2.1 for the first term in (2.13) and (2.12) for the second term gives (with $\nu_1 := \nu_1(1)$)

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

Hence $\tilde{\pi}$ is better than π.

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

Theorem 2.1

If a < 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 $\nu_1(1), \ldots, \nu_N(1)$. Suppose $\nu_i(1)$ is the largest and let it be achieved at time $\tau_i > 1$. Operate machine i for time 1, 2,...., $\tau_i - 1$. 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 indices $\nu_1(S_k^1+1),\ldots, \nu_N(S_k^N+1)$. Suppose the j^{th} index is the largest and let it be achieved by the stopping time $\tau_j > S_k^j+1$. Operate machine j for time $T_k+1, \ldots, T_k+(\tau_j-1-S_k^j) := T_{k+1}$.

In words: at the end of each stage calculate all indices, 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 3.3.

2.6. Optimality of the index rule for a = 1

When a = 1 the bandit and tax problems seem not to be equivalent, and separate arguments appear necessary to prove optimality. However, since the two arguments are similar, only the tax problem is treated in detail.

The bandit problem for a = 1 is trivial if $X^i(s) \geq 0$ for all i, s. Indeed any policy which operates every machine infinitely often will then be optimal since it yields the maximum present value $E \sum_i \sum_s X^i(s)$. However, even for this trivial case, the index rule may be preferred since it will be close to optimal when a <

1 and close to 1. See Kelly [12] and the references therein. On the other hand, if $X^i(s) < 0$ for some i, s, there is no longer any obvious optimal policy since it may be advantageous to operate some machines only a finite number of times. Thus the bandit problem for a = 1 is of interest although this case has apparently been ignored in the literature.

We turn now to the tax problem. There are N machines $\{ X^i(s), F^i(s) \}$ and we seek a policy π to minimize (cf (2.9))

$$W(\pi) = E \{ \sum_{1}^{\infty} \sum_{i \neq j(t)} X^i(t^i(t)+1) \mid F(1) \}$$
$$= E \{ \sum_{i} \sum_{s} [l_i(s) - l_i(s-1) - 1] X^i(s) \mid F(1) \}$$

where $l_i(s)$ is the time when π operates i for the s^{th} time. See Figure 2.

We now prove the optimality of the index rule defined by the index (2.11). The next lemma is proved in a way similar to Lemma 2.1.

Lemma 2.2

Suppose τ is optimum for (2.11). Let $\sigma > s$ be any $\{ F^i(.) \}$ stopping time. Then

$$\frac{E \{ 1(\sigma<\tau) [X^i(\sigma) - X^i(\tau)] \mid F^i(s) \}}{E \{ 1(\sigma<\tau) [\tau - \sigma] \mid F^i(s) \}} \geq \gamma_i(s) \quad \text{a.s.}$$

Consider the effect of π from time t on. By modifying $F^i(1)$ we may suppose that t = 1. Let machine 1 have the largest index

$$\gamma_1(1) \geq \gamma_i(1), \quad \text{all i,} \tag{2.14}$$

and suppose it is achieved at time τ. Let

$$T := l_1(\tau-1), \quad k_i - 1 := t^i(T), \text{ so } k_1 - 1 = \tau - 1.$$

Define policy $\tilde{\pi}$ exactly as in the preceding section and let $\tilde{l}_i(s)$ be the time when $\tilde{\pi}$ operates i for the s^{th} time. Then

$$W(\tilde{\pi}) = E \{ \sum_{i} \sum_{s} [\tilde{l}_i(s) - \tilde{l}_i(s-1) - 1] X^i(s) \mid F(1) \}$$

and so

$$\begin{aligned}\Delta &:= W(\pi) - W(\tilde{\pi}) \\ &= E\{ \sum_{i=1}^{N} \sum_{s=1}^{k_i} [(l_i(s)-l_i(s-1)-1) - (\tilde{l}_i(s)-\tilde{l}_i(s-1)-1)]\, X^i(s) \mid F(1) \} \\ &= E\{ \sum_{s=1}^{\tau} [(l_1(s)-l_1(s-1)-1) - (\tilde{l}_1(s)-\tilde{l}_1(s-1)-1)]\, X^1(s) \mid F(1) \} \\ &\quad - E\{ \sum_{i\neq 1} \sum_{s=1}^{k_i} [(\tilde{l}_i(s)-\tilde{l}_i(s-1)-1) - (l_i(s)-l_i(s-1)-1)]\, X^i(s) \mid F(1) \}. \end{aligned} \tag{2.15}$$

Proposition

There exist $b_s^i(s) \geq 0$ such that

$$\Delta = E\{ \sum_{s=1}^{\tau} b_s^1 [X^1(s) - X^1(\tau)] - \sum_{i\neq 1} \sum_{s=1}^{k_i} b_s^i [X^i(1) - X^i(s)] \mid F(1) \}. \tag{2.16}$$

Proof

Let

$$b_\tau^1 := 0, \quad b_s^i := [l_i(s) - l_i(s-1)] - [\tilde{l}_i(s) - \tilde{l}_i(s-1)], \text{ otherwise.}$$

Then (2.15) and (2.16) are equal if

$$\sum_{s=1}^{k_i} [l_i(s) - l_i(s-1)] = \sum_{s=1}^{k_i} [\tilde{l}_i(s) - \tilde{l}_i(s-1)], \text{ for all } i \neq 1.$$

This reduces to $l_i(k_i) = \tilde{l}_i(k_i)$ which is certainly true since after time T the policies π and $\tilde{\pi}$ operate the same machines in the same order. (See Figure 2.) Also, for $(i, s) \neq (1, \tau)$, $b_s^i \geq 0$ since $\tilde{l}_i(s) \geq l_i(s)$ and $l_i(s) - l_i(s-1) \geq \tilde{l}_i(s) - \tilde{l}_i(s-1)$.

■

Using Lemma 2.2 for the first term in (2.16) and (2.14) for the second term gives, after some algebra,

$$\Delta \geq \gamma_1(1)\, E\{ \sum_{s=1}^{\tau} b_s^1 (\tau - s) - \sum_{i\neq 1} \sum_{s=1}^{k_i} b_s^i (s-1) \} = 0.$$

Hence $\tilde{\pi}$ is better than π.

Now $\tilde{\pi}$ coincides with the index rule over 1, 2,..., $\tau-1$. Since the initial time is arbitrary it follows that the index rule is optimal. A similar argument works for the bandit problem as well.

Theorem 2.2

If a = 1, the index rules defined by (2.10) and (2.11) are optimal.

3. Extensions

3.1. Continuous time, nonpre-emptive

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

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

$X^i(s)$ is the instantaneous reward (or tax) as before. If i is operated for the s^{th} 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), (ii), (iii) of Section 2.2 are maintained. 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,... 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 t^{th} period. Then the real (process) time at the end of period t is

$$\sigma(t) = \sigma^{i(1)}(t^{i(1)}(1)) + \ldots + \sigma^{i(t)}(t^{i(t)}(t)).$$

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

$$V(\pi) := E \{ \sum_{t=1}^{\infty} \int_{\sigma(t-1)}^{\sigma(t)} X^{i(t)}(t^{i(t)})\, a^r\, dr \mid F(1) \} \tag{3.3}$$

The integral gives the present value of rewards when i(t) is operated during the t^{th} period, discounted back to time 0. The case $\sigma^i(s) = 1$ reduces to the

standard bandit problem of Section 2.2.

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

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

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

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.

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))

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

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

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

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

3.2. Continuous time, pre-emptive

Machine i is now characterized by the continuous parameter process

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

$X^i(s)$ is the reward (or tax) process. $F^i(r) \subset F^i(s)$ for $r < s$. $F^i := \bigvee_s F^i(s)$, 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

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

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

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

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 most cases one can construct a proof as follows. Fix $\varepsilon > 0$, and restrict attention to policies π_ε which switch machines only at times 0, ε, 2ε, . . . This is a standard bandit problem of Section 2.2. Moreover

$$\sup V(\pi_\varepsilon) \leq \sup V(\pi_{\frac{1}{2}\varepsilon}) \leq \sup V(\pi).$$

A technical argument is now required to show that $\sup V(\pi) - \sup V(\pi_\varepsilon) \to 0$ as $\varepsilon \to 0$. In [11] the bandit problem for diffusions is analyzed 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 it has been operated for time s is given by

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

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

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

3.3. Superprocess

We consider the discrete time problems of Sections 2.2 with an additional degree of freedom: when a particular machine is operated one must also select a control action that affects both the immediate reward and the machine's "state transition". We call such a machine a "superprocess", following [6, 7, 19]. The control action is based on the available information, i.e., one selects a feedback law. Once a feedback law is chosen, this machine can be described as before by a pair of sequences of rewards and information σ–fields.

Thus, from an abstract point of view, the i^{th} superprocess is simply a collection $\underline{X}^i$ of standard machines $X^i = \{ X^i(s), F^i(s) \}$, with different machines in $\underline{X}^i$ corresponding to different feedback laws.

Suppose we are given N superprocesses. For each selection $X^i \in \underline{X}^i$, let $V^*(X^1, \ldots, X^N)$ be the maximum expected reward of the standard bandit problem associated with the machines $X^1, \ldots, X^N$. The bandit problem associated with the N superprocesses is to find $X^i \in \underline{X}^i$ to

$$\max_{\underline{X}^1 \times \ldots \times \underline{X}^N} V^*(X^1, \ldots, X^N).$$

It is easy to suspect that the selection of the optimal $(X^1, \ldots X^N)$ 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 i^{th} superprocess can be made independent of the selection of the machine in the j^{th} superprocess $j \neq i$. 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

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

where τ ranges over all stopping times of $\{ F^X(.) \}$. For later reference note that if τ_ν is the optimal stopping time for (3.7), and $\nu' \leq \nu$, then one can find an optimal stopping time $\tau_{\nu'}$ such that $\tau_{\nu'} \geq \tau_\nu$ a.s.

Observe also that if $\nu(1)$ is the index of machine X at time 1 given by (2.10), then

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

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

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

Theorem 3.4

Suppose $X^i \in \underline{X}^i$ is such that X^i dominates every $Y^i \in \underline{X}^i$. Then

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

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

The proof of Theorem 3.4 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 $\{ F^Z(.) \}$ stopping times $\sigma_1 < \sigma_2 < \ldots$ and a sequence of index values $\nu_1, \nu_2, \ldots$ 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 $(\sigma_{i+1}+1) > (\sigma_i+1)$.

Lemma 3.1

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

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

Proof

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

$$\begin{aligned} \sigma &= \sigma_i \text{ on } \{ \nu_{i+1} \leq \nu_i \} \\ &= \sigma_{i+1} \text{ on } \{ \nu_{i+1} > \nu_i \}. \end{aligned}$$

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

$$\begin{aligned} &E\{ \sum_{\sigma_{i-1}+1}^{\sigma} a^t Z(t) \mid F(\sigma_{i-1}+1) \} \\ &= E\{ \sum_{\sigma_{i-1}+1}^{\sigma_i} a^t Z(t) \mid F(\sigma_{i-1}+1) \} + E\{ 1\{\nu_{i+1}>\nu_i\} E\{ \sum_{\sigma_i+1}^{\sigma_{i+1}} a^t Z(t) \mid F(\sigma_i+1) \} \mid F(\sigma_{i-1}+1) \}\} \\ &= \nu_i E\{ \sum_{\sigma_{i-1}+1}^{\sigma_i} a^t \mid F(\sigma_{i-1}+1) \} + E\{ 1\{\nu_{i+1}>\nu_i\} \nu_{i+1} \times \\ &\qquad \times E\{ \sum_{\sigma_i+1}^{\sigma_{i+1}} a^t \mid F(\sigma_i+1) \mid F(\sigma_{i-1}+1) \}\} \\ &> \nu_i E\{ \sum_{\sigma_{i-1}+1}^{\sigma} a^t \mid F(\sigma_{i-1}+1) \} \quad \text{with positive probability,} \end{aligned}$$

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. If X dominates Y, then

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

Proof

First suppose $a < 1$.

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,

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

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

$$V^*(Y, Z) = E\{\sum_{1}^{\lambda_1} a^t Y(t) + a^{\sigma_1} \sum_{\lambda_1+1}^{\lambda_2} a^t Y(t) + \ldots\}$$
$$+ E\{a^{\lambda_1} \sum_{1}^{\sigma_1} a^t Z(t) + a^{\lambda_2} \sum_{\sigma_1+1}^{\sigma_2} a^t Z(t) + \ldots\}. \qquad (3.9)$$

According to Remark 2.1 we may assume that the interval $\sigma_i+1, \ldots, \sigma_{i+1}$ 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

$$E\{\sum_{\sigma_i+1}^{\sigma_{i+1}} a^t Z(t) \mid F^Z(\sigma_i+1)\} = \nu_{i+1} E\{\sum_{\sigma_i+1}^{\sigma_{i+1}} a^t \mid F^Z(\sigma_i+1)\} \qquad (3.10)$$

$\nu_{i+1} \leq \nu_i$ 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 (τ_1+1) of $\{F^X(.)\}$ such that

$$N(X, \nu_1) = E\{\sum_{1}^{\tau_1} a^t [X(t) - \nu_1]\}.$$

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

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

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

Because $\nu_{i+1} \le \nu_i$ a.s. we may assume that $\tau_{i+1} \ge \tau_i$ a.s. Operate machine X $(\tau_{i+1} - \tau_i)$ times. Then operate machine Z $(\sigma_{i+1} - \sigma_i)$ times.

This policy results in the sequence of immediate rewards

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

and so

$$\begin{aligned} V^*(X, Z) \ge E \{ & \sum_1^{\tau_1} a^t X(t) + a^{\sigma_1} \sum_{\tau_1+1}^{\tau_2} a^t X(t) + .. \} \\ & + E \{ a^{\tau_1} \sum_1^{\sigma_1} a^t Z(t) + a^{\tau_2} \sum_{\sigma_1+1}^{\sigma_2} a^t X(t) + .. \} \end{aligned} \tag{3.12}$$

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

$$\begin{aligned} \Delta_1 &= E \{ [\sum_1^{\tau_1} a^t X(t) - \sum_1^{\lambda_1} a^t Y(t)] + a^{\sigma_1} [\sum_{\tau_1+1}^{\tau_2} a^t X(t) - \sum_{\lambda_1+1}^{\lambda_2} a^t Y(t)] + .. \} \\ &= E \{ (1-a^{\sigma_1}) [\sum_1^{\tau_1} a^t X(t) - \sum_1^{\lambda_1} a^t Y(t)] \\ &\quad + (a^{\sigma_1} - a^{\sigma_2}) [\sum_1^{\tau_2} a^t X(t) - \sum_1^{\lambda_2} a^t Y(t)] + .. \} \end{aligned} \tag{3.13}$$

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

The typical term in (3.13) is

$$E \{ (a^{\sigma_{i-1}} - a^{\sigma_i}) E \{ \sum_1^{\tau_1} a^t X(t) - \sum_1^{\lambda_1} a^t Y(t) \mid F^Z(\sigma_i) \} \}$$

$$\ge E \{ (a^{\sigma_{i-1}} - a^{\sigma_i}) \nu_i (\sum_1^{\tau_1} a^t - \sum_1^{\lambda_1} a^t) \} = \frac{a}{1-a} E \{ (a^{\sigma_{i-1}} - a^{\sigma_i}) \nu_i (a^{\lambda_1} - a^{\tau_1}) \},$$

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

$$\frac{1-a}{a}\,\Delta_1 \geq E\,(a^{\lambda_1}-a^{\tau_1})\,(1-a^{\sigma_1})\,\nu_1 + E\,(a^{\lambda_2}-a^{\tau_2})\,(a^{\sigma_1}-a^{\sigma_2})\,\nu_2 + \ldots \qquad (3.15)$$

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

$$\frac{1-a}{a}\,\Delta_2 \leq E\,(a^{\lambda_1}-a^{\tau_1})\,(1-a^{\sigma_1})\,\nu_1 + E\,(a^{\lambda_2}-a^{\tau_2})\,(a^{\sigma_1}-a^{\sigma_2})\,\nu_2 + \ldots$$

which proves that $\Delta_1-\Delta_2 \geq 0$ as required. An analogous argument works for a = 1.

■

Corollary 3.1

Suppose X dominates Y. Then for any machines $Y^2, \ldots, Y^N$

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

Proof

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

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

where the sequence $\{ Z(s) \}$ is an interweaving of the reward sequences $\{ Y^2(s) \},\ldots,\{ Y^N(s) \}$. 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

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

Also $V^*(X, Z) \leq V^*(X, Y^2, \ldots, Y^N)$ 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 3.4

Repeated applications of the corollary above give

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

■

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

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

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

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

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

Theorem 3.5

Suppose $X^i \in \underline{X}^i$ is such that X^i dominates every $Y^i \in \underline{X}^i$. Then

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

3.4. Arm-acquiring bandits

We shall consider the discrete time bandit problem of Section 2.2 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, ... The i^{th} 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,.., n(t) is available. These are the machines which either were available at time 1 or arrived during 1,.., t-1. Let $t^i(t)$, i = 1,.., n(t) be the number of times that i was operated during time 1,.., t. Thus $t^i(t)$ is the i^{th} machine time at process time t. The decision at t+1 is 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 $i = 1,..., n(t+1)$. In addition to the assumptions (i)-(iii) imposed at the beginning of Section 2.2 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,..., 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 possiblity will be removed later.

We convert this problem into one involving N 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, ... 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 machine $X^{\pi} = \{ X^{\pi}(s), F^{\pi}(s) \}$; 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 $\underline{X}$. Of course $X \in \underline{X}$.

We want to show that $\underline{X}$ contains a dominating machine X^{π}.

The following proposition will be useful.

Lemma 3.3

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

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

Proof

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

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

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

$$N = E \{ 1(\sigma\leq\tau) \sum_{1}^{\tau-1} a^t Z(t) \} + E \{ 1(\sigma>\tau) \sum_{1}^{\sigma-1} a^t Z(t) \} - E \{ 1(\sigma>\tau) \sum_{\tau}^{\sigma-1} a^t Z(t) \}$$
$$= E \sum_{1}^{\lambda-1} a^t Z(t) - E \{ 1(\sigma>\tau) \sum_{\tau}^{\sigma-1} a^t Z(t) \}, \quad \lambda := \max (\sigma, \tau).$$

Since $E \sum_{1}^{\lambda-1} a^t Z(t) \leq N$, 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 $\{ F^\pi(.) \}$. Let

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

and let $\pi(\nu)$, $\tau(\nu)$ be optimal for (3.16).

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

Fix two numbers $\mu < \nu$.

Lemma 3.4

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

Proof

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

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

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

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

In the sequence (3.18) the Z(i) denote the rewards which explicitly appear in (3.17). Hence $k-1 \le \tau(\nu)-1$. By Lemma 3.3 and since $\mu < \nu$

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

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

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

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).

Next consider the policy π and stopping time $\tau := \tau(\mu)$ which gives the reward sequence

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

Assumption (iv) guarantees the feasability of π. Also π agrees with $\pi(\nu)$ during $1, \ldots, \tau(\nu)-1$. Since $N(\mu) = N(\pi(\mu), \mu)$,

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

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

$$- E\{ \sum_{j=1}^{\sigma_1} a^j [Y(j)-\mu] + \ldots + \sum_{j=\sigma_{k-1}+1}^{\sigma_k} a^{k-1+j} [Y(j)-\mu] + \sum_{i=1}^{k-1} a^{\sigma_1+i} [Z(i)-\mu] \}$$

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

$$- E\{ \sum_{j=1}^{\sigma_1} a^j [Y(j)-\nu] + \ldots + \sum_{j=\sigma_{k-1}+1}^{\sigma_k} a^{k-1+j} [Y(j)-\nu] + \sum_{i=1}^{k-1} a^{\sigma_1+i} [Z(i)-\nu] \}$$

$$= \sum_{i=1}^{k-1} (1-a^{\sigma_1}) a^i [Z(i)-\nu] - \sum_{i=1}^{k} \sum_{j=\sigma_{i-1}+1}^{\sigma_1} (a^{i-1}-a^{k-1}) a^j [Y(j)-\nu]$$

$$= \sum_{i=1}^{k-1} b_i^Z \sum_{s=i}^{k-1} a^s [Z(s)-\nu] - \sum_{i=1}^{\sigma_k} b_i^Y \sum_{s=1}^{i} a^s [Y(s)-\nu]$$

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

Exactly as in the proof of Theorem 2.1 one can show that $b_i^Z \geq 0$, $b_i^Y \geq 0$.

On the other hand,

$$N(\pi(\nu), \nu) = N(\pi, \nu) = \sum_{i=1}^{k-1} a^i [Z(i)-\nu] + \sum_{j=1}^{\sigma_k} a^{k-1+j} [Y(j)-\nu].$$

Hence, by Lemma 3.3

$$\sum_{s=i}^{k-1} a^s [Z(s)-\nu] \geq 0 \geq \sum_{s=1}^{i} a^{k-1+s} [Y(s)-\nu]$$

from which it follows that $\Delta_Z \geq 0 \geq \Delta_Y$, and so $N(\pi, \mu) = N(\pi(\mu), \mu)$. The proof is complete.

■

Theorem 3.6

There exists a policy π such that X^π dominates every machine in $\underline{X}$.

Proof

Let $\nu_1 > \nu_2 > \ldots \to -\infty$. By Lemma 3.4 there exist policies $\pi(\nu_i)$ and stopping times $\tau(\nu_i) \to \infty$ a.s. such that $\pi(\nu_{i+1})$ agrees with $\pi(\nu_i)$ during $1, \ldots, \tau(\nu_i)-1$. Then $\pi := \lim \pi(\nu_i)$ is the required policy.

■

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

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

At time t consider the i^{th} machine X^i, after it has been operated $s-1 = t^i(t)$times. This machine, together with the arrival process { A(.) }, defines a superprocess $\underline{X}^i(s)$. We define the index, $\nu_i(s)$ of X^i to be the index of the dominant machine in $\underline{X}^i$. More directly

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

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

Theorem 3.7

For the bandit problem with arrivals, it is optimal to operate at each time the available machine with the largest current index. (When a = 1, the index policy maximizes the average reward per unit time.)

Proof

At any process time one is faced with the superprocesses $\underline{X}^i$, i = 1,..., n(t). By Theorem 3.6 the dominant machine in $\underline{X}^i$ has index (3.20). By Theorem 3.3 it is sufficient to restrict attention to these dominant machines, but then Theorem 2.1 guarantees optimality of the index rule.

■

3.5. 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 2.2. The details are sufficiently

different to require a separate treatment.

Suppose initially that $a < 1$. The cost of any policy π is

$$W(\pi) = E\{ \sum_{t=1}^{\infty} a^t [\sum_{i \neq i(t)}^{n(t)} X^i(t^i(t)+1)] \}$$
$$= E\{ \sum_{i=1}^{\infty} \sum_{s=1}^{\infty} [a^{l_i(s-1)+1} + .. + a^{l_i(s)-1}] X^i(s) \}$$

where

$l_i(0) := 0$ if machine i is available at time 1
= the process time when machine i arrived, otherwise.

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

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

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

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

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 3.8

For the tax problem with arrivals an optimal policy is given by the index rule defined by (3.21). (When $a = 1$ the index policy minimizes the average tax per unit time.)

Remark 3.1

The indices 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 is important to remark that for both bandit and tax problems with arrivals and with a = 1, the index rules given by (3.20) and (3.21) give the same sequence of machine operations as the index rule which is calculated neglecting future arrivals. Thus the optimal policy can be very easily calculated when a = 1. To see this consider the bandit problem. The index rule π which neglects arrivals leads to an accumulation of expected rewards at the fastest possible rate. Hence π is a dominating policy.

Theorem 3.9

If a = 1, then the optimal index rule for the tax and bandit problem with arrivals is the same if the calculation of the index ignores future arrivals.

Remark 3.2

It should be clear that Theorems 3.6, 3.7 and 3.8 generalize in the obvious way to the situation where time is continuous and the discipline is pre-emptive or nonpre-emptive as in Sections 3.1, 3.2. Assumption (v) must now be read to mean that new machines arrive in a Poisson stream.

4. Calculating the index

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

4.1. Discrete time bandit problem

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

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

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

$$\nu_i = \max_{\tau>1} \frac{E_i \{ \sum_1^{\tau-1} a^t\, r(x(t)) \}}{E_i \{ \sum_s^{\tau-1} a^t \}} \tag{4.1}$$

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

Lemma 4.1

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

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

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

Lemma 4.2

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

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

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×K transition matrix of the chain $\{ x(t) \}$.

Theorem 4.1

Suppose $\nu_1 \geq \nu_2 \geq .. \geq \nu_{m-1}$ 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 $\alpha^m = (\alpha_1^m, .. , \alpha_K^m)^T$, $\beta^m = (\beta_1^m, .. , \beta_K^m)^T$ are given by

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

with

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

$$r := (r(1),..,r(K))^T, \ \underline{1} := (1, .. , 1)^T.$$

Proof

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

$$\nu_m = \frac{E_m \{ \sum_1^{\tau-1} a^t\, r(x(t)) \}}{E_m \{ \sum_1^{\tau-1} a^t \}}$$

with

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

Hence

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

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

which concludes the proof.

4.2. Continuous time, nonpre-emptive bandit problem

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

Let $\sigma(s) := T_s - T_{s-1}$. Let r(i) be the reward when $\psi(t) = i$. The nonpre-emptive discipline means that a machine must be operated until its next jump time. In terms of the notation of Section 3.1, 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 infor-

mation 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

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

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

As in the preceding section one obtains the following result.

Theorem 4.2

Suppose $\nu_1 \geq \ldots \geq \nu_{m-1}$. Let

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

Then

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

where

$$\alpha_i^m := E_i \sum_{1}^{\tau-1} a^{\sigma(1)+..+\sigma(s-1)} [1 - a^{\sigma(s)}] r(x(s)) = (1-b_i) r(i) + b_i \sum_{j<m} P_{ij} \alpha_j^m$$

$$\beta_i^m := E_i \sum_{1}^{\tau-1} a^{\sigma(1)+..+\sigma(s-1)} [1 - a^{\sigma(s)}] = (1-b_i) + b_i \sum_{j<m} P_{ij} \beta_j^m.$$

4.3. Discrete time tax problem

Since the equivalence of this problem to the discrete time bandit problem is established in Section 2.3 for $a < 1$, the index can be written easily as

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

under the same conditions as Sec.4.1 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.

For the case a = 1, a certain simplification is possible as seen in the next section.

5. An application

Consider a network of queues indexed i = 1,.., 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 nonpre-emptive. A job in queue i requires a random amount of service time $\sigma(i)$ whose mean is $\mu(i)^{-1}$. 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. Jobs arrive at the various queues from outside the network in independent Poisson streams.

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. Klimov [13, 14] considered the problem of assigning the single server to the jobs in such a way as to minimize the long run average waiting cost per unit time

$$\lim_{T\to\infty} \frac{1}{T} E \int_0^T \sum_i c(i)\, n_i(t)\, dt. \tag{5.1}$$

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,..., N }. If the job leaves the network after (s-1) service completions, let x(s) = 0. Let $F(s) := \sigma\{ x(1),..., x(s) \}$; 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)x(K+1) transition matrix P. One defines an index as in (3.5). If x(s) = i, the index is

$$\gamma_i = \max_{\tau>1} \frac{E_i \{ c(x(1)) - c(x(\tau)) \}}{E_i \{ \sigma(x(1)) + .. + \sigma(x(\tau-1)) \}}, \quad i = 1, .., K \qquad (5.2)$$

$$= 0, \quad i = 0.$$

where $E_i f := E \{ f \mid x(1) = i \}$ and $c(0) := 0$. Note that $\gamma_i > 0$ for $i > 0$. Theorem 3.9 now gives the following result first proved by Klimov.

Theorem 5.1

The index rule defined by the index (5.2) minimizes the long run average waiting cost (5.1).

Theorem 2.2 gives the following result not previously known.

Theorem 5.2

Suppose there are no arrivals. The index rule defined by (5.2) minimizes the total waiting cost

$$E \int_0^\infty \sum_i c(i)\, n_i(t)\, dt$$

for every initial condition { $n_i(t)$ }.

Klimov gives an algorithm for computing the index. That algorithm requires repeated solution of systems of linear equations of the same order as the number of queues, K. The algorithm given below is simpler. It finds in

sequence the queue or state with the highest index, removes it from the network and after updating certain parameters continues the process. Let "n" denote the step in the algorithm. Let $\Sigma := \{0, 1, .., K\}$.

Step 1 (Initializing)

Set $n = 1$, $P^1 = P$, $\alpha_i^1 = \mu(i)^{-1}$ for i in $\Sigma^1 = \Sigma$.

Step 2 (Calculation of n^{th} largest index)

Find

$$\gamma_n := \max_{i \in \Sigma^n} \frac{c(i) - \sum_{j \in \Sigma^n} P_{ij}^n \, c(j)}{\alpha_i^n}$$

and suppose the maximum is achieved at i_n. If $n = K$, stop.

Step 3 (Updating)

Let $\Sigma^{n+1} := \Sigma^n - \{i_n\}$,

$$P_{ij}^{n+1} := P_{ij}^n + P_{ii_n}^n \, P_{i_nj}^n, \quad i, j \in \Sigma^{n+1}.$$

$$\alpha_i^{n+1} := \alpha_i^n + P_{ii_n}^n \, \alpha_{i_n}^n, \quad i \in \Sigma^{n+1}.$$

Set $n = n+1$ and go to Step 2.

Thus $\{ P_{ij}^n \}$ is the transition matrix of the original chain $\{ x(s) \}$ watched when it is in Σ^n. And α_i^n is the expected (service) time needed by a customer who is in i to leave i and then to reenter a queue in Σ^n. With this interpretation one may prove the next result in the same way as Theorem 4.1.

Theorem 5.3

The indices calculated above satisfy (5.2).

6. Conclusions

The multi-armed bandit problem is perhaps the simplest non-trivial 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 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.

7. References

[1] Glazebrook K.D., "Scheduling tasks with exponential service times on parallel processors," *Journal of Applied Probability* Vol 16 (1979), 685-689.

[2] Glazebrook K.D., "Stoppable families of alternative bandit processes," *Journal of Applied Probability* Vol 16 (1979), 843-854.

[3] Glazebrook K.D., "On randomized dynamic allocation indices for sequential design of experiments," *Journal of the Royal Statistical Society* Vol 42 (1980), 342-346.

[4] Glazebrook K.D., "On stochastic scheduling with precedence relations and switching costs," *Journal of Applied Probability* Vol 17 (1980), 1016-1024.

[5] Glazebrook K.D., "On the evaluation of suboptimal policies for families of alternative bandit processes," *Journal of Applied Probability* Vol 19 (1982), 716-722.

[6] Glazebrook K.D., "On a sufficient condition for superprocesses due to Whittle," *Journal of Applied Probability* Vol 19 (1982), 99-110.

[7] Gittins J.C. "Bandit processes and dynamic allocation indices," *Journal of the Royal Statistical Society* Vol 41 (1979), 148-177.

[8] Gittins J.C. and K.D. Glazebrook, "On Bayesian models in stochastic scheduling," *Journal of Applied Probability* Vol 14 (1977), 556-565.

[9] Gittins J.C. and K.D. Glazebrook, "On single machine scheduling with precedence relation and linear or discounted costs," *Operations Research* Vol 29 (1981), 161-173.

[10] Gittins J.C. and D.M. Jones, "A dynamic allocation index for the sequential design of experiments," in J. Gani, K. Sarkadi and I. Vince (eds) *Progress in Statistics* European Meeting of Statisticians 1972, Vol 1, North-Holland (1974), 241-266.

[11] Karatzas, I., "Gittins indices in the dynamic allocation problem for diffusion processes," Columbia University, Department of Mathematical Statistics, Preprint (1982), 40 pp.

[12] Kelly, F.P., "Multi-armed bandits with discount factor near one: the Bernoulli case," *The Annals of Statistics* Vol 9 (1981), 987-1001.

[13] Klimov G.P., "Time sharing service systems I," *Theory of Probability and Applications* Vol 19 (1974), 532-551.

[14] Klimov G.P., "Time sharing service systems II," *Theory of Probability and Applications* Vol 23 (1978), 314-321.

[15] Nash P. and J.C. Gittins, "A Hamiltonian approach to optimal stochastic resource allocation," *Advances in Applied Probability* Vol 9 (1977), 55-68.

[16] Rodman L., "On the many-armed bandit problem," *The Annals of Probabilty* Vol 6 (1978), 491-498.

[17] Wahrenberger D., C. Antle and L. Klimko, "Bayesian rules for the two-armed bandit problem," *Biometrika* Vol 64 (1977), 172-174

[18] Weitzman M L., "Optimal search for the best alternative," *Econometrica* Vol 47 (1979), 641-654.

[19] Whittle P., "Multi-armed bandits and the Gittins index," *Journal of the Royal Statistical Society* Vol 42 (1980), 143-149.

[20] Whittle P., "Arm-acquiring bandits," *The Annals of Probability* Vol 9 (1981), 284-292.

[21] Whittle P., *Optimization over Time* Vol 1 (1982), John Wiley.

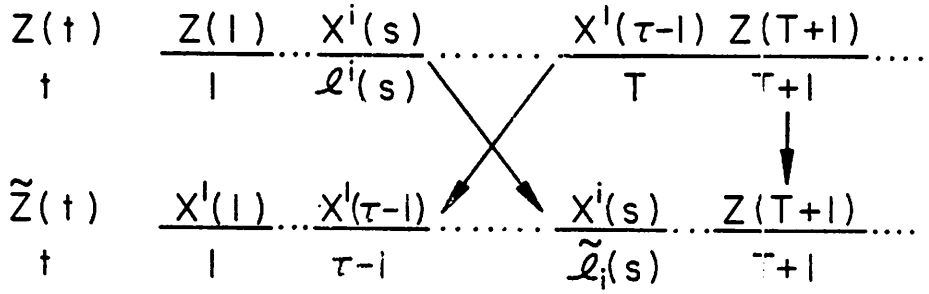

Fig. 1 Reward sequences Z, $\tilde{Z}$

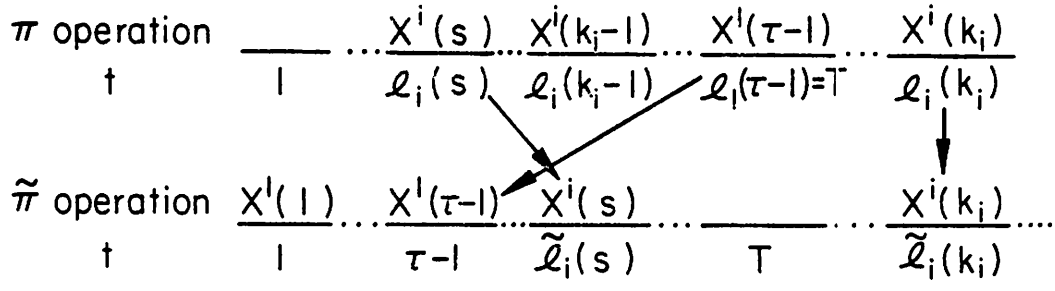

Fig. 2 Reward sequence due to π, $\tilde{\pi}$