Math Meth Oper Res (1999) 50:449–461

Mathematical Methods
of Operations Research

On the optimality of the Gittins index rule for multi-armed bandits with multiple plays*

Dimitrios G. Pandelis[1], **Demosthenis Teneketzis**[2]

[1] ERIM International, Inc., P.O. Box 134001, Ann Arbor, MI 48113-4001, USA
(e-mail: pandelis@erim-int.com)
[2] Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109-2122, USA (e-mail:teneket@eecs.umich.edu)

Abstract. We investigate the general multi-armed bandit problem with multiple servers. We determine a condition on the reward processes sufficient to guarantee the optimality of the strategy that operates at each instant of time the projects with the highest Gittins indices. We call this strategy the Gittins index rule for multi-armed bandits with multiple plays, or briefly the Gittins index rule. We show by examples that: (i) the aforementioned sufficient condition is not necessary for the optimality of the Gittins index rule; and (ii) when the sufficient condition is relaxed the Gittins index rule is not necessarily optimal. Finally, we present an application of the general results to the multiserver scheduling of parallel queues without arrivals.

Key words: Multi-armed bandits, Gittins index

1 Introduction

Models of dynamic allocation of scarce resources to competing projects have been widely used and are of great importance. The multi-armed bandit problem is concerned with the dynamic allocation of a single resource among several projects. The basic version of the stochastic multi-armed bandit problem, formulated in discrete time, is the following. There are n independent projects and one server. At each time t the server must work on exactly one project. Let $x_i(t)$, $i = 1, 2, \ldots, n$, be the state of project i at time t, and denote by $k(t)$ the project operated at time t. If $k(t) = i$, an immediate reward $R_i(x_i(t))$ is obtained, and the state of project i changes to $x_i(t+1)$ according to a

* Supported in part by NSF Grant NCR-9204419, AFOSR Grant F49620-96-1-0028, and ARO Grant DAAH04-96-1-0377.
Manuscript received: March 1999/final version received: July 1999

stationary Markov rule. The states of the idle projects remain frozen. The objective is to find a scheduling strategy that maximizes an infinite horizon expected discounted reward $\mathscr{R}$ given by

$$\mathscr{R} = E\left\{\sum_{t=0}^{\infty} \beta^t R_{k(t)}(x_{k(t)}(t))\right\},$$

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

This problem was first solved by Gittins (see Gittins and Jones (1974) and Gittins (1979)). Using a forward induction argument, he showed that an index policy is optimal. Specifically, at each time t project i, $i = 1, 2, \ldots, n$, is characterized by an index $v_i(x_i(t))$ that is a function only of its state. The optimal strategy, called the *Gittins index rule*, operates the project with the largest index. The Gittins index is given by

$$v_i(x_i) = \max_{\tau > 0} \frac{E\{\sum_{t=0}^{\tau-1} \beta^t R_i(x_i(t)) | x_i(0) = x_i\}}{E\{\sum_{t=0}^{\tau-1} \beta^t | x_i(0) = x_i\}}, \tag{1}$$

where the maximization is taken over all stopping times $\tau > 0$ of $\mathscr{F}_i(\cdot)$, where $\mathscr{F}_i(s)$ is the σ-field representing the information about project i after it has been operated s times. The Gittins index rule result is very important because it decomposes the n-dimensional problem into n one-dimensional problems. This is because the optimal policy is determined by n numbers, each depending only on the state of an individual project.

A different proof of the optimality of the Gittins index rule was provided by Whittle (1980). Gittins' original work has been extended in various directions such as superprocesses (Gittins (1979)), arm-acquiring bandits (Whittle (1981)), non-Markovian bandits (Varaiya et al. (1985)), and bandits with switching costs (Asawa and Teneketzis (1996)). The optimality of the Gittins index rule has also been shown for several variations of the multi-armed bandit problem (see Kelly (1981), Glazebrook (1982), Karatzas (1984), Mandelbaum (1986), Weber (1994), and Bertsimas and Nino-Mora (1996)).

It is known (see, for example, Ishikida (1992)) that the policy that operates the projects with the highest Gittins indices is not in general the optimal allocation rule for multi-armed bandits with multiple servers (or, equivalently, multiple plays) and an infinite horizon expected discounted reward criterion. The optimal solution of the aforementioned class of problems is not generally known. Anantharam et al. (1987) and Agrawal et al. (1990) have determined optimal allocation schemes for multi-armed bandits with multiple plays and the "learning loss" or "regret" criterion. Asymptotic (in the number of projects and servers) results for restless bandits with multiple plays appear in Whittle (1988) and Weber and Weiss (1990). In this paper we investigate, in discrete time, optimal strategies for the multi-armed bandit problem with multiple plays and an infinite horizon expected discounted reward criterion. For both the deterministic and stochastic multi-armed bandit problems with multiple plays we determine a condition on the reward processes sufficient to guarantee the optimality of the strategy that operates at each instant of time the projects with the highest Gittins indices. We call this strategy the *Gittins index rule for multi-armed bandits with multiple plays*, or briefly the *Gittins*

index rule. Furthermore, we show by examples that: (i) the aforementioned sufficient condition is not necessary for the optimality of the Gittins index rule; and (ii) when the sufficient condition is relaxed the Gittins index rule is not necessarily optimal.

The paper is organized as follows: In Section 2 the deterministic multi-armed bandit problem with multiple plays is formulated and analyzed. Its stochastic counterpart is presented in Section 3. An application of the general result to the multiserver scheduling of parallel queues without arrivals is presented in Section 4. We conclude in Section 5.

2 The deterministic multi-armed bandit problem with multiple plays

2.1 Problem formulation

In this section we formulate, in discrete time, the deterministic version of the multi-armed bandit problem with multiple plays. The problem, denoted by P1, is the following.

Problem P1. We have a collection of n projects ($n > 2$) and m processors ($1 < m < n$). Associated with each project i, $i = 1, 2, \ldots, n$, is a deterministic reward process $\{Z_i(l)\}_{l=0}^{\infty}$. At each time t each processor must work on exactly one project; no more than one processors can work on the same project at any time. We denote by $t^i(t)$, $i = 1, 2, \ldots, n$, the number of times project i has been operated during $0, 1, 2, \ldots, t-1$ ($t^i(0) := 0$, $i = 1, 2, \ldots, n$). If project i is operated at time t, a reward $Z_i(t^i(t))$ is received. Under the above conditions we seek to determine allocation schemes that maximize the total β-discounted reward.

As in the stochastic case, we can define the Gittins index of each project. Because the state of a project is determined by the number of times it has been operated, the Gittins index of project i, $i = 1, 2, \ldots, n$, at time t is a function of $t^i(t)$, i.e., a function of the number of times project i has been operated by time $t-1$. The Gittins index is defined by

$$v_i(l) = \max_{\tau \geq l+1} \frac{\sum_{s=l}^{\tau-1} \beta^s Z_i(s)}{\sum_{s=l}^{\tau-1} \beta^s}, \quad i = 1, 2, \ldots, n, \quad l = 0, 1, 2, \ldots, \tag{2}$$

where the maximizer in (2) is given by

$$\tau_i(l) = \inf\{s \geq l+1 : v_i(s) \leq v_i(l)\}. \tag{3}$$

In what follows we determine a condition on the reward processes under which the Gittins index rule is optimal for problem P1. We proceed in two steps. First, we define an auxiliary problem, called P2, that has the following characteristics: (i) for any allocation policy π, the corresponding total discounted reward for problem P2 upper-bounds the total discounted reward for problem P1 under the same policy; and (ii) under the Gittins index rule the total discounted rewards for problems P1 and P2 are equal. Second, under a certain condition on the reward processes we show that the Gittins index rule is optimal for problem P2. Consequently, it is optimal for problem P1.

2.2 Preliminaries

In this section we formulate an auxiliary problem, called P2, and establish its relation to problem P1. We also formulate another auxiliary problem, called P3, for which we determine an optimal strategy. Problem P3 is directly related to problem P2; its solution allows us to determine a condition under which the Gittins index rule is optimal for problem P2.

We begin by defining the concave envelopes of the reward processes $\{Z_i(l)\}_{l=0}^{\infty}$, $i = 1, 2, \ldots, n$, as in Mandelbaum (1986) and Ishikida (1992). The concave envelope $\{\bar{Z}_i(l)\}_{l=0}^{\infty}$ of $\{Z_i(l)\}_{l=0}^{\infty}$ is given by

$$\bar{Z}_i(l) = \min_{s \geq l} v_i(s), \quad l = 0, 1, 2, \ldots \tag{4}$$

Equivalently, the concave envelope is defined by

$$\bar{Z}_i(l) = v_i(\tau_i^k), \quad \tau_i^k \leq l < \tau_i^{k+1}, \tag{5}$$

where

$$\tau_i^0 = 0,$$

$$\tau_i^{k+1} = \inf\{s > \tau_i^k : v_i(s) \leq v_i(\tau_i^k)\}, \quad k = 0, 1, 2, \ldots \tag{6}$$

Problem P2 is the same as problem P1 with the reward processes replaced with their concave envelopes. We denote by $V_1(Z_1, Z_2, \ldots, Z_n; \pi)$ and $V_2(Z_1, Z_2, \ldots, Z_n; \pi)$ the total β-discounted rewards for problems P1 and P2, respectively, when strategy π is employed and the original reward processes are $Z_1, Z_2, \ldots, Z_n$. The following results hold.

Lemma 1. *The reward obtained under any policy π is not decreased when the original reward processes are replaced with their concave envelopes, i.e.,*

$$V_1(Z_1, Z_2, \ldots, Z_n; \pi) \leq V_2(Z_1, Z_2, \ldots, Z_n; \pi), \quad \textit{for any } \pi. \tag{7}$$

Lemma 2. *The reward obtained under the Gittins index rule does not change when the original reward processes are replaced with their concave envelopes, i.e., with π^* denoting the Gittins index rule, we have*

$$V_1(Z_1, Z_2, \ldots, Z_n; \pi^*) = V_2(Z_1, Z_2, \ldots, Z_n; \pi^*). \tag{8}$$

The proofs of Lemmas 1 and 2 are identical to the proofs of Claims 2 and 3 respectively in Ishikida (1992, p. 93), when only one project is to be operated at each time, and are omitted here.

We now proceed to formulate problem P3.

Problem P3. We have N families of jobs $(N > 2)$ and M processors $(1 < M < N)$ that operate in parallel. At each time instant a processor is allowed to work on at most one job. Moreover, no more than one processors are allowed to work on the same job at the same time. We denote jobs

in family i, $i = 1, 2, \ldots, N$, by $i_1, i_2, \ldots$. Job i_k requires X_{i_k} time units of processing; each time job i_k is processed a reward C_{i_k} is received. The following constraints are satisfied:

(S1) In family i, $i = 1, 2, \ldots, N$, job i_{k+1}, $k = 1, 2, \ldots$, must be processed after the processing of job i_k has been completed.
(S2) For jobs in family i, $i = 1, 2, \ldots, N$, $C_{i_{k+1}} < C_{i_k}$, $k = 1, 2, \ldots$.

The objective is to determine a scheduling strategy in the class of preemptive strategies that maximizes the total β-discounted reward.

Let $V_3(\pi)$ be the total β-discounted reward for problem P3 under policy π. The following lemma presents a condition under which it is possible to explicitly determine an optimal allocation policy for Problem P3.

Lemma 3. *Consider Problem P3. Suppose that the reward processes satisfy the following condition:*

(C1) For any two jobs k, l that belong to different families

$$C_k(1-\beta) \geq C_l \quad \text{whenever } C_k > C_l. \tag{9}$$

Then, at each instant of time it is optimal to process the jobs that yield the highest rewards.

Proof. Consider a policy π that satisfies precedence constraint (S1), but does not always give priority to the jobs with the highest rewards. Let t_1 be the first time instant where one of the jobs with the M highest rewards that are available for processing is not processed under π. Assume that this is job k_r. This implies that there exists time $t_2 > t_1$ such that job k_r is processed at t_2 under π. We consider two cases.

Case 1. Not all M processors are busy at time t_1. We construct policy π^1 to be identical to π except that it processes job k_r at time t_1 instead of t_2. Then

$$V_3(\pi^1) - V_3(\pi) = C_{k_r}\beta^{t_1} - C_{k_r}\beta^{t_2} > 0. \tag{10}$$

Case 2. All processors are busy at time t_1, i.e., some job l_j with $C_{l_j} < C_{k_r}$ is processed at t_1 under π ($l \neq k$ because of precedence constraint (S1)). Let $s_{j+1}, s_{j+2}, \ldots$ be the times jobs $l_{j+1}, l_{j+2}, \ldots$ are processed for the first time under π. We construct policy π^1 as follows: π^1 is identical to π except that it processes job k_r at time t_1 instead of t_2, job l_j at time s_{j+1} instead of t_1, and job l_p, $p = j+1, j+2, \ldots$ at time s_{p+1} instead of s_p. Then

$$V_3(\pi^1) - V_3(\pi) = C_{k_r}(\beta^{t_1} - \beta^{t_2}) - C_{l_j}(\beta^{t_1} - \beta^{s_{j+1}})$$

$$- \sum_{p=j+1}^{\infty} C_{l_p}(\beta^{s_p} - \beta^{s_{p+1}}). \tag{11}$$

Because of (11) and $C_{l_j} > C_{l_{j+1}} > \cdots$ (see reward constraint (S2)) we have

$$V_3(\pi^1) - V_3(\pi) \geq C_{k_r}(\beta^{t_1} - \beta^{t_2}) - C_{l_j}\beta^{t_1}. \tag{12}$$

Because of $t_2 > t_1$ and (9)

$$C_{k_r}(1 - \beta^{t_2 - t_1}) > C_{k_r}(1 - \beta) \geq C_{l_j}. \tag{13}$$

From (12) and (13) we get

$$V_3(\pi^1) > V_3(\pi). \tag{14}$$

Therefore, in both cases policy π^1 satisfies precedence constraint (S1) (by construction) and yields higher reward than π because of (10) and (14). We can now construct a modification π^2 of π^1 in the same way π^1 modifies π, i.e., the first time π^1 does not process one of the jobs with the highest rewards, π^2 processes that job and yields a higher reward than π^1. Repeating the construction of such modified policies we conclude that under condition (9) on the reward processes it is optimal to process at each instant of time the jobs with the highest rewards. ■

2.3 Optimality of the Gittins index rule

In this section we determine a condition sufficient to guarantee the optimality of the Gittins index rule for deterministic multi-armed bandits with multiple plays. We begin by noting that problem P2 can be formulated as a version of problem P3 as follows. We have n families of jobs and m processors that operate in parallel. The jobs in family i, $i = 1, 2, \ldots, n$, are denoted by i_0, $i_1, \ldots$. The processing time of job i_k is $\tau_i^{k+1} - \tau_i^k$. When processed, job i_k yields a reward $v_i(\tau_i^k)$ per unit time. The following constraints are satisfied:

(i) In family i, $i = 1, 2, \ldots, n$, job i_{k+1}, $k = 0, 1, \ldots$, must be processed after job i_k.
(ii) $v_i(\tau_i^{k+1}) < v_i(\tau_i^k)$, $i = 1, 2, \ldots, n$, $k = 0, 1, \ldots$.

Constraints (i) and (ii) correspond to constraints (S1) and (S2), respectively, in problem P3. Consider now the following condition:

(R1) For any $i \neq j$ and k, l such that $v_i(\tau_i^k) > v_j(\tau_j^l)$ we have

$$v_i(\tau_i^k)(1 - \beta) \geq v_j(\tau_j^l).$$

Note that (R1) corresponds to the reward condition given in (9). The main result of Section 2 is given in the following theorem.

Theorem 1. *Assume that the reward processes $\{Z_i(l)\}_{l=0}^{\infty}$, $i = 1, 2, \ldots, n$, are such that condition (R1) is satisfied. Then the Gittins index rule is optimal for problem P1, that is, at each instant of time it is optimal to operate the m projects with the highest Gittins indices.*

Proof. Let π^* be the policy that for problem P2 operates the m projects that yield the highest rewards. Note that, because of the way the concave envelopes of the original reward processes were defined, policy π^* is equivalent to the Gittins index rule. Then, because condition (R1) is satisfied, Lemma 3 implies that the Gittins index rule is optimal for problem P2, i.e., for any policy π

$$V_2(Z_1, Z_2, \ldots, Z_n; \pi) \leq V_2(Z_1, Z_2, \ldots, Z_n; \pi^*). \tag{15}$$

From (7), (8), and (15) we get that for any policy π

$$V_1(Z_1, Z_2, \ldots, Z_n; \pi) \leq V_1(Z_1, Z_2, \ldots, Z_n; \pi^*).$$

Therefore, under condition (R1) the Gittins index rule is optimal for problem P1. ■

When the rewards become identically equal to zero after a finite time for all arms, the result of Theorem 1 holds under a condition weaker than (R1). Let $l_i,\ i = 1, 2, \ldots, n$, be finite integers such that $Z_i(l) = 0$ for all $l > l_i$. From (2) and (4) we get that the concave envelopes become identically zero as well, that is, $\bar{Z}_i(l) = 0$ for all $l > l_i$, or equivalently (see (5)), there exist finite integers $k_i,\ i = 1, 2, \ldots, n$, such that $v_i(\tau_i^k) = 0$ for all $k > k_i$. Then, the Gittins index rule is optimal when the following condition is satisfied:

(R2) For any $i \neq j$ and k, l such that $v_i(\tau_i^k) > v_j(\tau_j^l)$ we have

$$v_i(\tau_i^k)(1 - \beta) \geq v_j(\tau_j^l)(1 - \beta^{\sum_{i=1}^{n} \tau_i^{k_i+1}}).$$

2.4 Discussion

The essence of Theorem 1 is the following: The solution of the multi-armed bandit problem with one server can be obtained by a forward induction argument because decisions made at any particular stage are not irrevocable. This, as pointed out in Gittins (1979), means that "any bandit process which is available for continuation at some stage, and which is not then chosen, may be continued at a later stage, and with exactly the same resulting sequence of rewards, apart from the discount factor. This (in turn) means there is no later advantage to compensate for the initial disadvantage of not choosing a forwards induction policy." Forward induction does not, in general, lead to optimal allocation decisions in multi-armed bandits with multiple servers because at each stage of the allocation process the optimal strategy does not allocate the servers one at a time, thus, the previous arguments do not hold. Consequently, the full effect of future rewards has to be taken into account in determining an optimal allocation strategy. However, if the Gittins indices of different arms are sufficiently separated, the dominant factors in determining an optimal allocation strategy become the reward-rate-maximizing portions of each bandit process starting from its current state. In such a situation, an optimal allocation strategy can be determined by forward induction and the Gittins index rule is an optimal allocation strategy. Conditions (R1) and (R2) present exactly a situation where there is enough separation among the Gittins indices to guarantee the optimality of the Gittins index rule.

We present two examples that highlight the nature of condition (R2). Specifically, the examples show that: (i) if condition (R2) is not satisfied, the Gittins index rule is not necessarily optimal for multi-armed bandits with multiple plays; and (ii) condition (R2) is sufficient but not necessary for the optimality of the Gittins index rule.

Example 1. We have $n = 3$ projects, $m = 2$ processors, and rewards discounted by a factor $\beta = 0.9$. The reward processes for each project are given by

$$Z_1(0) = 6, \quad Z_1(1) = Z_1(2) = 1, \quad Z_1(l) = 0, \quad l > 2,$$

$$Z_2(0) = 5, \quad Z_2(1) = Z_2(2) = Z_2(3) = 3, \quad Z_2(l) = 0, \quad l > 3, \quad \text{and}$$

$$Z_3(0) = Z_3(1) = Z_3(2) = Z_3(3) = 4, \quad Z_3(l) = 0, \quad l > 3.$$

Because the reward processes are decreasing, they are identical to their concave envelopes. Note that condition (R2) is not satisfied because $6 > 5$ but $0.6 = 6(1-\beta) < 5(1-\beta^{11}) \simeq 3.43$. Let π^* denote the Gittins index rule, which in this example is equivalent to the policy that operates at each instant of time the projects with the highest rewards. We have

$$V_1(Z_1, Z_2, Z_3; \pi^*) = (6+5) + (\beta+\beta^2+\beta^3)(4+3) + \beta^4(4+1) + \beta^5 1. \quad (16)$$

Consider now a policy π that is different from the Gittins index rule; it operates projects 2 and 3 at time 1, projects 1 and 3 at time 2, projects 3 and 2 at times 3 and 4, projects 3 and 1 at time 5, and project 1 at time 6. This policy yields a reward

$$V_1(Z_1, Z_2, Z_3; \pi) = (5+4) + \beta(6+4) + (\beta^2+\beta^3)(4+3) + \beta^4(3+1) + \beta^5 1. \quad (17)$$

From (16) and (17) we obtain

$$V_1(Z_1, Z_2, Z_3; \pi^*) - V_1(Z_1, Z_2, Z_3; \pi) = 2 - 3\beta + \beta^4 \simeq -0.04.$$

Therefore the Gittins index rule is not optimal.

Example 2. We have $n=3$ projects, $m=2$ processors, and $\beta=0.5$. The reward processes are given by

$$Z_1(0) = 4, \quad Z_1(1) = 2, \quad Z_1(l) = 0, \quad l > 1,$$

$$Z_2(0) = 4, \quad Z_2(1) = 2, \quad Z_2(l) = 0, \quad l > 1, \quad \text{and}$$

$$Z_3(0) = 3, \quad Z_3(l) = 0, \quad l > 0.$$

Condition (R2) is not satisfied because $4 > 3$ but $2 = 4(1-\beta) < 3(1-\beta^5) \simeq 2.9$. By computing the reward from all possible scheduling strategies it is straightforward to show that the Gittins index rule is optimal.

3 The stochastic multi-armed bandit problem with multiple plays

3.1 Problem formulation

In the stochastic multi-armed bandit problem with multiple plays, denoted by P1′, there are n projects ($n > 2$) and m processors ($1 < m < n$). Project i, $i = 1, 2, \ldots, n$, is characterized by the pair of sequences $\{Z_i(l), \mathscr{F}_i(l)\}_{l=0}^{\infty}$, where $Z_i(l)$ is the random reward obtained when project i is operated for the $(l+1)$th time and $\mathscr{F}_i(l)$ is the σ-field representing the information about project i after it has been operated l times. Let $\mathscr{F}^i := \bigvee_l \mathscr{F}_i(l)$, $i = 1, 2, \ldots, n$. We make the following assumptions:

(A1) $\mathscr{F}_i(l) \subset \mathscr{F}_i(l+1)$.
(A2) $\bigvee_l \sigma(Z_i(l)) \vee \mathscr{F}^i$, $i = 1, 2, \ldots, n$, are independent.
(A3) At each instant of time each processor must work exactly on one project; no more than one processors can work on the same project at any time.

Let $t^i(t)$, $i = 1, 2, \ldots, n$, denote the number of times project i has been operated during $0, 1, 2, \ldots, t-1$; $t^i(t)$ is called the ith project time at process time t. Denote by $k_1(t), k_2(t), \ldots, k_m(t)$ the projects operated at time t. When $k_1(t) = i_1, k_2(t) = i_2, \ldots, k_m(t) = i_m$, the states x_j of projects j, $j \neq i_1, i_2, \ldots, i_m$, remain frozen. The states x_{i_l}, $l = 1, 2, \ldots, m$, of projects $i_1, i_2, \ldots, i_m$ change according to the rule $P(x_{i_l}(t^{i_l}(t)+1) \in \tilde{A} | x_{i_l}(1), \ldots, x_{i_l}(t^{i_l}(t)))$, where $\tilde{A} \in \mathscr{F}^{i_l}(t^{i_l}(t)+1)$. Consider the decision at time, $t = 0, 1, 2, \ldots$. This decision is based on the available information

$$\mathscr{F}(t) = \bigvee_i \mathscr{F}_i(t^i(t)), \quad t = 0, 1, 2, \ldots.$$

$\mathscr{F}(t)$ is recursively defined as follows:

$$\mathscr{F}(t+1) = \mathscr{F}(t) \vee \mathscr{G}(t),$$

where $\mathscr{G}(t)$ is the σ-field generated by sets of the form $\{k_1(t) = i_1, k_2(t) = i_2, \ldots, k_m(t) = i_m\} \cap \{t^{i_1}(t) = s_{i_1}, t^{i_2}(t) = s_{i_2}, \ldots, t^{i_m}(t) = s_{i_m}\} \cap (A_{i_1} \times A_{i_2} \times \cdots \times A_{i_m})$, with $A_{i_l} \in \mathscr{F}^{i_l}(s_{i_l}+1)$. A policy is any sequence of decisions $\{\mathbf{u}(t), \mathbf{u}(t) = (k_1(t), k_2(t), \ldots, k_m(t)), t = 0, 1, 2, \ldots\}$, where $\mathbf{u}(t)$ is based only on $\mathscr{F}(t)$, and $\mathscr{F}(t)$ evolves according to the mechanism described above.

The multi-armed bandit problem with multiple plays is to find a policy π that maximizes

$$V(\pi) := E\left\{ \sum_{t=0}^{\infty} \beta^t \sum_{i=1}^{m} Z_{k_i(t)}(t^{k_i(t)}(t)) | \mathscr{F}(0) \right\}. \tag{18}$$

The Gittins index of project i after it has been operated l times is defined to be

$$v_i(l) = \max_{\tau \geq l+1} \frac{E\{\sum_{s=l}^{\tau-1} \beta^s Z_i(s) | \mathscr{F}_i(l)\}}{E\{\sum_{s=l}^{\tau-1} \beta^s | \mathscr{F}_i(l)\}}, \tag{19}$$

where the maximization is over all stopping times $\tau,\ l+1 \leq \tau < \infty$, of $\mathscr{F}_i(\cdot)$, and "max" in (19) is to be interpreted as "ess sup". Under assumptions (A1)–(A2) made in the problem formulation, there always exists a stopping time τ achieving the maximum in (19) (see Neveu (1975)).

Proceeding as in the deterministic multi-armed bandit problem with multiple plays we determine a condition on the reward processes under which the Gittins index rule, i.e., the policy that at each instant of time operates the m projects with the highest Gittins indices, is optimal for problem P1′.

3.2 Optimality of the Gittins index rule

In this section we establish a condition sufficient to guarantee the optimality of the Gittins index rule for stochastic multi-armed bandits with multiple plays. We begin by defining problem P2′ to be the same as problem P1′ with the reward processes replaced with their concave envelopes defined by equations (4)–(6) and (19). It can be shown (see Ishikida (1992)) that Lemmas 1 and 2 hold with $V_1(Z_1, Z_2, \ldots, Z_n; \pi)$, $V_2(Z_1, Z_2, \ldots, Z_n; \pi)$ denoting the expected total β-discounted rewards for problems P1′ and P2′, respectively, when strategy π is employed and the original reward processes are $Z_1, Z_2, \ldots, Z_n$. That is,

$$V_1(Z_1, Z_2, \ldots, Z_n; \pi) \leq V_2(Z_1, Z_2, \ldots, Z_n; \pi), \quad \text{for any } \pi, \tag{20}$$

$$V_1(Z_1, Z_2, \ldots, Z_n; \pi^*) = V_2(Z_1, Z_2, \ldots, Z_n; \pi^*), \tag{21}$$

where π^* denotes the Gittins index rule.

Problem P2′ can be formulated as a version of problem P3 as follows: For every realization ω of problem P2′ we have n families of jobs and m processors that operate in parallel. The jobs in family $i,\ i = 1, 2, \ldots, n$, are denoted by $i_0, i_1, \ldots$. The processing time of job i_k is $\tau_i^{k+1}(\omega) - \tau_i^k(\omega)$. When processed, job i_k yields a reward $v_i(\tau_i^k(\omega), \omega)$ per discounted unit time. The following constraints are satisfied:

(i) In family $i,\ i = 1, 2, \ldots, n$, job $i_{k+1},\ k = 0, 1, \ldots$, must be processed after job i_k.
(ii) $v_i(\tau_i^{k+1}(\omega), \omega) < v_i(\tau_i^k(\omega), \omega),\ \forall \omega,\ i = 1, 2, \ldots, n,\ k = 0, 1, \ldots$.

Consider now the following condition:

(R1′) For each realization ω of problem P2′, for any $i \neq j$, and k, l such that $v_i(\tau_i^k(\omega), \omega) > v_j(\tau_j^l(\omega), \omega)$ we have

$$v_i(\tau_i^k(\omega), \omega)(1 - \beta) \geq v_j(\tau_j^l(\omega), \omega).$$

Based on this condition we can prove the main result of Section 3.

Theorem 2. *Assume that* $\{Z_i(l), F_i(l)\}_{l=0}^{\infty},\ i = 1, 2, \ldots, n$, *are such that condition (R1′) is satisfied. Then the Gittins index rule is optimal for problem P1′, that is, at each instant of time it is optimal to operate the m projects with the highest Gittins indices.*

Proof. Because of condition (R1′) and Lemma 4 we get that the Gittins index rule, denoted by π^*, is optimal for problem P2′, i.e., for any policy π

$$V_2(Z_1, Z_2, \ldots, Z_n; \pi) \leq V_2(Z_1, Z_2, \ldots, Z_n; \pi^*). \tag{22}$$

The optimality of the Gittins index rule for problem P1′ follows from (20), (21), and (22).

4 An application: Multiserver scheduling of parallel queues without arrivals

As an application of the results of Section 3.2 we consider the dynamic multiserver scheduling problem of parallel queues without arrivals. We have, in discrete time, a system consisting of N parallel queues and m, $m < N$, identical servers. At each time each server must work on one queue, and no more than one servers can work on the same queue at any time. Queue j, $j = 1, 2, \ldots, N$, initially has q_j customers ($q_j < \infty$). The service times σ_j of customers in queue j, $j = 1, 2, \ldots, n$, are independent identically distributed (i.i.d.) random variables with non-decreasing hazard rate; furthermore, for all k, j, $k \neq j$, the random variables σ_k, σ_j are independent. Each customer present in queue j, $j = 1, 2, \ldots, N$, incurs an instantaneous holding cost h_j. The objective is to determine a scheduling policy that minimizes the total expected β-discounted ($0 < \beta < 1$) weighted flowtime of the customers initially present in the system, or, equivalently, to maximize the total expected β-discounted weighted reward, where rewards are obtained by customer service completions.

We can formulate the above scheduling problem as a multi-armed bandit with multiple plays as follows: Queue j, $j = 1, 2, \ldots, N$, is associated with bandit j where rewards are obtained only at customer completion epochs; time intervals between successive customer completion epochs in queue j, $j = 1, 2, \ldots, N$, are i.i.d. random variables σ_j with non-decreasing hazard rate; for all k, j, $k \neq j$, the random variables σ_j, σ_k are independent. The reward obtained from bandit j, $j = 1, 2, \ldots, N$, when the service of a customer is completed at time $t - 1$ is equal to $(\beta^t h_j)/(1 - \beta)$. Thus, we have N bandits with the reward structure described above, and m servers.

In the case of service times with non-decreasing hazard rates the Gittins index is achieved at the next completion epoch. The index for bandit j is

$$v_j(t) = h_j \frac{E\{\beta^{\sigma_j}\}}{1 - E\{\beta^{\sigma_j^t}\}}, \tag{23}$$

where t is the amount of service the current job has received and σ_j^t is the remaining service time. Because of the non-decreasing hazard rate assumption we have

$$v_j(0) \leq v_j(t), \quad \forall t. \tag{24}$$

Therefore, the concave envelope for bandit j (cf. (4)) is constant for all realizations of the service times and equal to

$$\bar{v}_j = v_j(0) = h_j S_j (1 - S_j)^{-1}, \tag{25}$$

where

$$S_j = E^{f_j}\{\beta^{\sigma_j}\} \tag{26}$$

and f_j is the probability mass function of σ_j. Suppose that the following condition holds.

(L1) Whenever

$$h_j S_j (1 - S_j)^{-1} > h_k S_k (1 - S_k)^{-1}, \tag{27}$$

then

$$(1 - \beta) h_j S_j (1 - S_j)^{-1} > h_k S_k (1 - S_k)^{-1}. \tag{28}$$

Under condition (L1), Theorem 2 implies that the optimal policy for the scheduling problem formulated above is described by the following rule: Serve the queues exhaustively in decreasing order of their indices, where by an exhaustive policy we mean one that serves a queue until there are no customers left in that queue.

5 Conclusions

We have presented a condition sufficient to guarantee the optimality of the Gittins index rule for the multi-armed bandit problem with multiple plays. The essence of this condition is the following: The requirement that the Gittins indices of different arms be sufficiently separated implies that the dominant factors in determining an optimal allocation strategy become the reward-rate-maximizing portions of each bandit process starting from its current state. This, in turn, implies that a forward induction argument that leads to the Gittins index rule (as defined in Section 1) is optimal.

We have shown by example that the aforementioned sufficient condition is not necessary to ensure the optimality of the Gittins index rule for the multi-armed bandit problem with multiple plays. The discovery of a condition that is both necessary and sufficient for the optimality of the Gittins index rule is currently an open problem.

References

[1] Agrawal R, Hegde M, Teneketzis D (1990) Multi-armed bandit problems with multiple plays and switching cost. Stochastics and Stochastic Reports 29:437–459

[2] Anantharam V, Varaiya P, Walrand J (1987) Asymptotically efficient allocation rules for multi-armed bandit problems with multiple plays. Part I: I.I.D. rewards, Part II: Markovian rewards. IEEE Transactions on Automatic Control 32:968-982

[3] Asawa M, Teneketzis D (1996) Multi-armed bandits with switching penalties. IEEE Transactions on Automatic Control 41:328–348

[4] Bertsimas D, Nino-Mora J (1996) Conservation laws, extended polymatroids and multi-armed bandit problems: a polyhedral approach to indexable systems. Mathematics of Operations Research 21:257–306

[5] Gittins JC, Jones D (1974) A dynamic allocation index for the sequential design of experiments. In: Gani J, Sarkadi K, Vince I (eds.) Progress in statistics. European Meeting of Statisticians 1972, vol. 1, pp. 241–266
[6] Gittins JC (1979) Bandit processes and dynamic allocation indices. Journal of the Royal Statistical Society 41:148–177
[7] Glazebrook KD (1982) On a sufficient condition on superprocesses due to Whittle. Journal of Applied Probability 19:99–110
[8] Ishikida T (1992) Informational aspects of decentralized resource allocation. Ph.D. Thesis, University of California, Berkeley.
[9] Karatzas I (1984) Gittins indices in the dynamic allocation problem for diffusion processes. The Annals of Probability 12:173–192
[10] Kelly FP (1981) Multi-armed bandits with discount factor near one: the Bernoulli case. The Annals of Statistics 9:987–1001
[11] Mandelbaum A (1986) Discrete multi-armed bandits and multi-parameter processes. Probability Theory 71:129–147
[12] Neveu J (1975) Discrete parameter martingales. North-Holland, New York
[13] Varaiya P, Walrand J, Buyukkoc C (1985) Extensions of the multi-armed bandit problem. IEEE Transactions on Automatic Control 30:426–439
[14] Weber RR, Weiss G (1990) On an index policy for restless bandits. Journal of Applied Probability 27:637–648
[15] Weber RR (1994) On the Gittins index for multi-armed bandits. The Annals of Applied Probability 2:1024–1033
[16] Whittle P (1980) Multi-armed bandits and the Gittins index. Journal of the Royal Statistical Society 42:143–149
[17] Whittle P (1981) Arm-acquiring bandits. The Annals of Probability 9:284–292
[18] Whittle P (1988) Restless bandits: activity allocation in a changing world. In: Gani J (ed.) A celebration of applied probability. Journal of Applied Probability 25A:287–298