Proc. Natl. Acad. Sci. USA
Vol. 90, pp. 1232–1236, February 1993
Statistics

General Gittins index processes in discrete time

(dynamic allocation/optimal stopping)

Nicole El Karoui† and Ioannis Karatzas‡

†Laboratoire de Probabilités, Université Pierre et Marie Curie, 4, place Jussieu–Tour 56, 75252 Paris Cedex 05, France; and
‡Department of Statistics, Columbia University, 619 Mathematics Building, New York, NY 10027

Communicated by Herbert Robbins, October 26, 1992

ABSTRACT We combine the formulation of Mandelbaum [Mandelbaum, A. (1986) *Probab. Theory Rel. Fields* 71, 129–147] with ideas from Whittle [Whittle, P. (1980) *J. R. Stat. Soc. B* 42, 143–149] to obtain a simple and constructive proof for the optimality of Gittins index processes in the general, non-markovian dynamic allocation (or "multi-armed bandit") problem. Our approach also provides an explicit expression for the value of this problem.

Section 1. Introduction

We consider in this paper the general, non-markovian formulation of the dynamic allocation (or "multi-armed bandit") problem in discrete time: there are *d* independent "projects" (or "arms"), only one of which may be engaged (pulled) at any given time, while the others remain "frozen." By engaging a particular project one receives a certain random reward, depending on time and on the history of the project that is being engaged. The objective is to maximize total expected discounted reward over an infinite horizon. How then is one optimally to schedule in time the pulling of the various arms?

Questions of this sort were essentially open for a long time, at least from the 1940s, until Gittins and his collaborators made fundamental contributions in the 1970s that amounted to a real breakthrough (cf. refs. 1 and 2). Working in a markovian framework for the evolution of the states of the independent projects, Gittins demonstrated that it is possible to assign to each project an "index function" of its state, computable (in principle) in terms of that project's dynamics only and such that the optimal policy takes the following form: at any given time, compute the indices of different projects and engage a project with maximal index. In refs. 3 and 4, Whittle then provided elegant, insightful, and mathematically concise demonstrations for the optimality of Gittins's rule; Whittle's work was later complemented by that of Tsitsiklis (5).

This work was extended to a general, non-markovian framework by Varaiya *et al.* (6) and by Mandelbaum (7), who also provided a formulation of the question as a control problem with time parameter in a multidimensional partially ordered set [in the manner of Mandelbaum and Vanderbei (8)]. The arguments and proofs in both these works, however, are quite lengthy and demanding.

We offer in this note an approach to the general discrete-time problem, which combines the powerful formulation of Mandelbaum (7) with the explicitness, brevity, and clarity of Whittle (3). It is based on a very detailed "martingale" study of individual optimal stopping problems and of the associated Gittins index sequences, in terms of which one is able to compute explicitly the value random field of the dynamic allocation problem, à là Whittle.

The paper is organized as follows: we start with the formulation and study of a family of optimal stopping problems (section 2), which allows the introduction of the Gittins index sequences and the study of their properties (section 3). The dynamic allocation problem is formulated in section 4 in the manner of Mandelbaum (7), and is solved in section 5 by extending to the non-markovian case the dynamic programming methodology of Whittle (3). Section 6 discusses a representation of the value that involves the so-called "index random field." All the proofs are collected in *Section 7.*

Section 2. A Family of Optimal Stopping Problems

Consider a probability space $(\Omega, \mathcal{F}, P)$ and on it a random sequence $H = \{h(t)\}_{t=1}^{\infty}$ that takes values in $[0, \infty)$, satisfies

$$E \sum_{t=0}^{\infty} \beta^{t} h(t+1) < \infty \qquad [2.1]$$

for some given *discount factor* $\beta \in (0, 1)$, and is predictable with respect to a given filtration $\{\mathcal{F}(t)\}_{t=0}^{\infty}$. We assume that $\mathcal{F}(0) = \{\phi, \Omega\}$, denote by $\mathcal{S}(\theta)$ the class of $\{\mathcal{F}(t)\}$-stopping times with values in $\{\theta, \theta + 1, \ldots\} \cup \{+\infty\}$ for any $\theta \in \mathbf{N}_0$, and consider the family of *optimal stopping* problems

$$V(t; m) \triangleq \operatorname{esssup}_{\tau \in \mathcal{S}(t)} E\left[\sum_{u=t}^{\tau-1} \beta^{u-t} h(u+1) + m\beta^{\tau-t} \,\middle|\, \mathcal{F}(t)\right], \qquad t \in \mathbf{N}_0 \qquad [2.2]$$

indexed by $m \in [0, \infty)$.

From standard theory on optimal stopping (e.g., chapter VI of ref. 9) we know, for any given $m \in [0, \infty)$, that the sequence

$$Z(t; m) \triangleq \beta^{t} V(t; m) + \sum_{u=0}^{t-1} \beta^{u} h(u+1), \qquad t \in \mathbf{N}_0 \quad [2.3]$$

is the *Snell envelope* of (i.e., the smallest nonnegative supermartingale that dominates) the sequence

$$Y(t; m) \triangleq \beta^{t} m + \sum_{u=0}^{t-1} \beta^{u} h(u+1), \qquad t \in \mathbf{N}_0; \qquad [2.4]$$

that the stopping time

$$\sigma_t(m) \equiv \sigma(t; m) \triangleq \inf\{\theta \geq t / V(\theta; m) = m\} \qquad [2.5]$$

is optimal for the problem of Eq. 2.2;
that the sequence

$$\{Z(\theta \wedge \sigma_t(m); m), \mathcal{F}(\theta)\}_{\theta=t}^{\infty} \qquad [2.6]$$

is a martingale; and that the *dynamic programming equation,*

$$V(t; m) = \max[m, h(t+1) + \beta E\{V(t+1; m) | \mathcal{F}(t)\}], \qquad [2.7]$$

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

holds a.s. for every $t \in \mathbf{N}_0$.

Based on this theory one can establish the following facts (proven in *Section 7*), for any given $t \in \mathbf{N}_0$.

LEMMA 2.1. *The mapping* $\mathrm{m} \mapsto \sigma(\mathrm{t};\, \mathrm{m})$ *is decreasing and right-continuous, a.s.*

LEMMA 2.2. *The mapping* $\mathrm{m} \mapsto \mathrm{V(t;\, m)}$ *is convex and increasing, with right-hand derivative*

$$\frac{\partial^+}{\partial \mathrm{m}} \mathrm{V(t;\, m)} \triangleq \mathit{lim}_{\delta \downarrow 0} \frac{1}{\delta}[\mathrm{V(t;\, m} + \delta) - \mathrm{V(t;\, m)}] \qquad \textbf{[2.8a]}$$

given, for any $\mathrm{m} \in [0, \infty)$, *by*

$$\frac{\partial^+}{\partial \mathrm{m}} \mathrm{V(t;\, m)} = \mathrm{E}[\beta^{\sigma(\mathrm{t;m})-\mathrm{t}} | \mathcal{F}(\mathrm{t})], \qquad \mathit{a.s.} \qquad \textbf{[2.8b]}$$

Section 3. Gittins Index Sequences

Let us denote by $\mathcal{M}(t)$ the class of all positive, $\mathcal{F}(t)$-measurable random variables. By analogy with Whittle (3), Varaiya *et al.* (6), and Mandelbaum (7), we define the *Gittins index* at time $t \in \mathbf{N}_0$ as

$$\begin{aligned} M(t) &\triangleq \operatorname{esssup}\{X \in \mathcal{M}(t)/V(t;\, X) > X, \text{ a.s.}\} \\ &= \operatorname{essinf}\{X \in \mathcal{M}(t)/V(t;\, X) = X, \text{ a.s.}\}. \end{aligned} \qquad \textbf{[3.1]}$$

It can be shown that this $\mathcal{F}(t)$-measurable random variable admits also the *forward induction* characterization

$$\begin{aligned} &(1-\beta)M(t) \\ &= \operatorname{esssup}_{\tau \in \mathcal{S}(t+1)} \frac{E\left[\sum_{u=t}^{\tau-1} \beta^u h(u+1) \middle| \mathcal{F}(t)\right]}{E\left[\sum_{u=t}^{\tau-1} \beta^u \middle| \mathcal{F}(t)\right]}, \qquad \text{a.s.} \end{aligned} \qquad \textbf{[3.2]}$$

as the maximum achievable "conditional expected discounted reward, over conditional expected discounted time," from period t onward. We shall not need this second characterization in the sequel. The notation

$$\underline{M}(t,\, \theta) \triangleq \min_{t \le u \le \theta} M(u), \qquad \underline{M}(\theta) \triangleq M(0,\, \theta) \qquad \textbf{[3.3]}$$

will be used throughout for the minimum to date of the *Gittins index random sequence* $M = \{\mathrm{M}(t), \mathcal{F}(t)\}_{t \in \mathbf{N}_0}$. The equivalences

$$\begin{aligned} &\sigma(t;\, m) > \theta \Leftrightarrow V(s;\, m) > m, \quad s = t, t+1, \ldots, \theta \\ &\qquad \Leftrightarrow m < \underline{M}(t,\, \theta), \end{aligned} \qquad \textbf{[3.4]}$$

for $\theta \ge t$, lead then to the following expressions for the optimal stopping time of identity **2.5** in terms of the index process and conversely:

$$\sigma(t;\, m) = \inf\{\theta \ge t/\underline{M}(t,\, \theta) \le m\} = \inf\{\theta \ge t/M(\theta) \le m\}, \quad \textbf{[3.5]}$$

$$\begin{aligned} \underline{M}(t,\, \theta) &= \inf\{m \ge 0/\sigma(t;\, m) \le \theta\} \\ &= \sup\{m \ge 0/\sigma(t;\, m) > \theta\}. \end{aligned} \qquad \textbf{[3.6]}$$

Remark 3.1: The observation **3.4** has the following interesting corollaries: for any $t \in \mathbf{N}_0$,

$$\begin{aligned} &E\left[\sum_{\theta=\sigma(t;m)}^{\infty} \beta^\theta h(\theta+1) \middle| \mathcal{F}(t)\right] \\ &\qquad = (1-\beta)\cdot E\left[\sum_{\theta=\sigma(t;m)}^{\infty} \beta^\theta \underline{M}(t,\, \theta) \middle| \mathcal{F}(t)\right], \qquad \text{a.s.} \end{aligned} \qquad \textbf{[3.7]}$$

$$\begin{aligned} \beta^t V(t,\, 0) &= E\left[\sum_{\theta=t}^{\infty} \beta^\theta h(\theta+1) \middle| \mathcal{F}(t)\right] \\ &= (1-\beta)\cdot E\left[\sum_{\theta=t}^{\infty} \beta^\theta \underline{M}(t,\, \theta) \middle| \mathcal{F}(t)\right], \qquad \text{a.s.} \end{aligned} \qquad \textbf{[3.8]}$$

$$\begin{aligned} &V(t;\, m) \\ &\quad = (1-\beta)\cdot E\left[\sum_{\theta=t}^{\infty} \beta^{\theta-t}(m \vee \underline{M}(t,\, \theta)) \middle| \mathcal{F}(t)\right], \qquad \text{a.s.} \end{aligned} \qquad \textbf{[3.9]}$$

The proofs are given in *Section 7*. Note the significance of Eq. **3.8**: in the computation of expected future total discounted reward, conditional on $\mathcal{F}(t)$, the reward sequence $\{h(\theta+1)\}_{\theta=t}^{\infty}$ may be replaced by the *decreasing* sequence $\{(1-\beta)\underline{M}(t,\, \theta)\}_{\theta=t}^{\infty}$.

Remark 3.2: For any given $t \in \mathbf{N}_0$, the function $m \mapsto (\partial^+/\partial m)V(t;\, m)$ is right-continuous and increasing; it is equal to zero at $m = 0$ and equal to one for $m \ge M(t)$, a.s. These properties follow readily from Eq. **3.4** and *Lemma 2.2*.

Section 4. The Dynamic Allocation Problem

Let us formulate now the dynamic allocation (or multi-armed bandit) problem, in the manner of Mandelbaum (7). Consider d *independent* filtrations $\mathbf{F}_i = \{\mathcal{F}_i(t)\}_{t \in \mathbf{N}_0}$, $i = 1, \ldots, d$, and construct a few filtration $\mathbf{F} = \{\mathcal{F}(\underset{\sim}{s})\}_{\underset{\sim}{s} \in S}$ by setting

$$\mathcal{F}(\underset{\sim}{s}) \triangleq \bigvee_{i=1}^{d} \mathcal{F}_i(s_i), \qquad \underset{\sim}{s} = (s_1, \ldots, s_d) \in S \qquad \textbf{[4.1]}$$

on the lattice $S = \mathbf{N}_0^d$ of nonnegative integers (endowed with the partial ordering $\underset{\sim}{r} \le \underset{\sim}{s} \Leftrightarrow r_i \le s_i,\ \forall\, i = 1, \ldots, d$). For any given $\underset{\sim}{s} \in S$, an *allocation strategy* (or optional increasing path) for $\underset{\sim}{s}$ is an S-valued random sequence $\underset{\sim}{T} = \{\underset{\sim}{T}(t)\}_{t \in \mathbf{N}_0}$ such that

$$\begin{aligned} &(i)\ \underset{\sim}{T}(0) = \underset{\sim}{s} \\ &(ii)\ \underset{\sim}{T}(t+1) = \underset{\sim}{T}(t) + \underset{\sim}{e}_i \quad \text{for some} \quad i = 1, \ldots, d, \end{aligned} \qquad \textbf{[4.2]}$$

and

$$\begin{aligned} &(iii)\ \{\underset{\sim}{T}(t+1) = \underset{\sim}{T}(t) + \underset{\sim}{e}_i,\ \underset{\sim}{T}(t) = \underset{\sim}{r}\} \in \mathcal{F}(\underset{\sim}{r}), \\ &\qquad\qquad \forall\, i = 1, \ldots, d, \qquad \forall\, \underset{\sim}{r} \in S. \end{aligned}$$

Here, $\underset{\sim}{e}_i$ denotes the ith unit vector in S. Note that, with $|\underset{\sim}{r}| \triangleq \Sigma_{i=1}^d r_i$ for $\underset{\sim}{r} \in S$, we have $|\underset{\sim}{T}(t)| = t + |\underset{\sim}{s}|$ for every such $\underset{\sim}{T}$ and $t \in \mathbf{N}_0$.

Intuitively, $T_i(t)$ counts "how many times the ith arm has been pulled up to time t." The requirements **4.2** express that $\underset{\sim}{T}$ "pulls one arm at a time," and the decision which arm to pull at time $t + 1$ is to be made on the basis of the information accumulated from the pulls of the various arms up to time t.

Associated with every arm (or project) $i = 1, \ldots, d$ is a random sequence $H_i = \{h_i(t)\}_{t \in \mathbf{N}}$ which represents the "reward" obtained by pulling that particular arm (or engaging that particular project); it is predictable with respect to the filtration $\mathbf{F}_i$ and takes values in $[0, K(1 - \beta)]$ for some fixed $K \in (0, \infty)$. The *dynamic allocation* (or multi-armed bandit) *problem* consists of maximizing the expected reward $E[\mathfrak{R}(\underset{\sim}{T})|\mathcal{F}(\underset{\sim}{s})]$ over all $\underset{\sim}{T} \in \mathcal{A}(\underset{\sim}{s})$, where

$$\mathfrak{R}(\underset{\sim}{T}) \triangleq \sum_{t=0}^{\infty} \sum_{i=1}^{d} \beta^t h_i(T_i(t + 1))[T_i(t + 1) - T_i(t)] \quad \textbf{[4.3]}$$

and $\mathcal{A}(\underset{\sim}{s})$ is the class of allocation strategies as in requirements **4.2**. We denote by

$$\Phi(\underset{\sim}{s}) \triangleq \operatorname{esssup}_{\underset{\sim}{T}\in\mathcal{A}(\underset{\sim}{s})} E[\mathfrak{R}(\underset{\sim}{T})|\mathcal{F}(\underset{\sim}{s})], \qquad \underset{\sim}{s} \in S \quad \textbf{[4.4]}$$

the *value random field* for this problem and expect it to satisfy

$$\Phi(\underset{\sim}{s}) = \max_{1\le j\le d} [h_j(s_j + 1) + \beta{\cdot}E\{\Phi(\underset{\sim}{s} + \underset{\sim}{e}_j)|\mathcal{F}(\underset{\sim}{s})\}], \underset{\sim}{s} \in S, \quad \textbf{[4.5]}$$

the formal *Bellman equation of dynamic programming* in this setup (cf. ref. 8). We shall verify this equation and produce a rather explicit representation for $\Phi(\underset{\sim}{s})$, which will in turn provide the structure of optimal allocation policies by following the method of Whittle (3) (cf. *Corollary 5.2* below). This method embeds the optimization problem of Eq. **4.4** into a family of problems of the same sort but with the additional option of "retiring" (i.e., abandoning all projects) and receiving a reward $M \ge 0$ (cf. Eq. **4.7**). This parametrization is the same as that in the family **2.2** of optimal stopping problems.

To carry out this embedding we shall need to extend the notion of an allocation strategy to accommodate a stopping time. We say that a measurable function $\underset{\sim}{\nu}{:}\Omega \to S$ is a *stopping point* of $\mathbf{F}$, if $\{\underset{\sim}{\nu} = \underset{\sim}{s}\} \in \mathcal{F}(\underset{\sim}{s})$, $\forall\, \underset{\sim}{s} \in S$, and for any such $\underset{\sim}{\nu}$ we consider the σ-field

$$\mathcal{F}(\underset{\sim}{\nu}) \triangleq \{A \in \mathcal{F} \;/\; A \cap \{\underset{\sim}{\nu} = \underset{\sim}{s}\} \in \mathcal{F}(\underset{\sim}{s}),\ \forall\ \underset{\sim}{s} \in S\}.$$

From the definition **4.2** of an allocation strategy, it follows that $\underset{\sim}{T}(t)$ is a stopping point for every $t \in \mathbf{N}_0$; thus $\mathcal{G}(t) = \mathcal{F}(\underset{\sim}{T}(t))$ is defined, and $\mathbf{G} = \{\mathcal{G}(t)\}_{t\in\mathbf{N}_0}$ is a filtration. A *policy* $\Pi = (\underset{\sim}{T}, \tau)$ is a pair consisting of an allocation strategy $\underset{\sim}{T} \in \mathcal{A}(\underset{\sim}{s})$ and a stopping time τ of $\mathbf{G}$; we denote by $\mathcal{P}(\underset{\sim}{s})$ the class of such policies and introduce the analogues of Eqs. **4.3–4.5**:

$$\mathfrak{R}(\Pi; M) \triangleq \sum_{t=0}^{\tau-1} \sum_{i=1}^{d} \beta^t h_i(T_i(t + 1))[T_i(t + 1) - T_i(t)] + M\beta^\tau \quad \textbf{[4.6]}$$

$$\Phi(\underset{\sim}{s}; M) \triangleq \operatorname{esssup}_{\Pi\in\mathcal{P}(\underset{\sim}{s})} E[\mathfrak{R}(\Pi; M)|\mathcal{F}(\underset{\sim}{s})], \qquad \underset{\sim}{s} \in S \quad \textbf{[4.7]}$$

$$\Phi(\underset{\sim}{s}; M) \quad \textbf{[4.8]}$$

$$= \max\left[M,\ \max_{1\le j\le d} \{h_j(s_j + 1) + \beta{\cdot}E[\Phi(\underset{\sim}{s} + \underset{\sim}{e}_j; M)|\mathcal{F}(\underset{\sim}{s})]\}\right].$$

Remark 4.1: For every $i = 1, \ldots, d$ we may consider the family of optimal stopping problems

$$V_i(t; m) = \operatorname{esssup}_{\tau\in\mathcal{S}_i(t)} E\left[\sum_{u=0}^{\tau-1} \beta^{u-t} h_i(u + 1) + m\beta^{\tau-t} \middle| \mathcal{F}_i(t)\right],$$

$$t \in \mathbf{N}_0 \quad \textbf{[4.9]}$$

for the filtration $\mathbf{F}_i$ (parametrized by $m \ge 0$), as well as the associated Gittins index process $\{M_i(t)\}_{t\in\mathbf{N}_0}$ of Eq. **3.1** (which now takes values in $[0, K]$). All the results of *Sections 2* and *3* are in force for these new objects.

The dynamic programming equation **4.8** extends Eq. **2.7** to the multidimensional case $d \ge 2$.

Section 5. The Whittle Reduction

Imitating Whittle (3), we make now the following Ansatz:

$$\frac{\partial^+}{\partial m} \Phi(\underset{\sim}{s}; m) = \prod_{i=1}^{d} \frac{\partial^+}{\partial m} V_i(s_i; m), \qquad \underset{\sim}{s} \in S, \quad \textbf{[5.1]}$$

which leads to the formula

$$\Phi(\underset{\sim}{s}; M) \quad \textbf{[5.2]}$$

$$= \left\{\begin{array}{lcr} \Phi(\underset{\sim}{s}) & ; & M = 0 \\ K - \displaystyle\int_M^K \left(\prod_{i=1}^{d} \frac{\partial^+}{\partial m} V_i(s_i; m)\right) dm & ; & 0 < M < K \\ M & ; & M \ge K \end{array}\right\},$$

$\underset{\sim}{s} \in S$, for the value random field of the dynamic allocation problem with retirement **4.7**. We shall vindicate this Ansatz in an indirect way, as follows:

THEOREM 5.1. *The random field*

$$\mathrm{F}(\underset{\sim}{s}; \mathrm{M}) \quad \textbf{[5.3]}$$

$$\triangleq \left\{\begin{array}{lcr} \mathrm{M} & ; & \mathrm{M} \ge \mathrm{K} \\ \mathrm{K} - \displaystyle\int_{\mathrm{M}}^{\mathrm{K}} \left(\prod_{i=1}^{d} \frac{\partial^+}{\partial \mathrm{m}} \mathrm{V_i(s_i; m)}\right) \mathrm{dm} & ; & 0 \le \mathrm{M} < \mathrm{K} \end{array}\right\},$$

satisfies the Bellman equation **4.8**; *more precisely, we have a.s.*

$$\mathrm{F}(\underset{\sim}{s}; \mathrm{M}) \ge \mathrm{M}; \qquad \forall\, \mathrm{M} \ge 0, \qquad \underset{\sim}{s} \in \mathrm{S}, \quad \textbf{[5.4]}$$

$$\mathrm{F}(\underset{\sim}{s}; \mathrm{M}) \ge \mathrm{h_j(s_j + 1)} + \beta{\cdot}\mathrm{E}[\mathrm{F}(\underset{\sim}{s} + \underset{\sim}{e}_j; \mathrm{M})|\mathcal{F}(\underset{\sim}{s})];$$

$$\forall\ M \ge 0, \qquad \underset{\sim}{s} \in \mathrm{S}, \quad \mathrm{j} = 1, \ldots\ldots, \mathrm{d}, \quad \textbf{[5.5]}$$

$$\mathrm{F}(\underset{\sim}{s}; \mathrm{M}) = \mathrm{M}, \quad \textit{on} \quad \left\{\mathrm{M} \ge \mathop{\mathit{max}}_{1\le j\le d} \mathrm{M_j(s_j)}\right\}; \qquad \forall\, \underset{\sim}{s} \in \mathrm{S}, \quad \textbf{[5.6]}$$

$$\mathrm{F}(\underset{\sim}{s}; \mathrm{M}) = \mathrm{h_i(s_i + 1)} + \beta{\cdot}\mathrm{E}[\mathrm{F}(\underset{\sim}{s} + \underset{\sim}{e}_i; \mathrm{M})|\mathrm{F}(\underset{\sim}{s})],$$

$$\textit{on} \left\{M < \mathrm{M_i(s_i)} = \mathop{\mathit{max}}_{1\le j\le d} \mathrm{M_j(s_j)}\right\}; \qquad \forall\, \underset{\sim}{s} \in \mathrm{S}. \quad \textbf{[5.7]}$$

COROLLARY 5.1. *The random field*

$$\mathrm{F}(\underset{\sim}{s}) \triangleq \mathrm{K} - \int_0^{\mathrm{K}} \left(\prod_{i=1}^{d} \frac{\partial^+}{\partial \mathrm{m}} \mathrm{V_i(s_i; m)}\right) \mathrm{dm}, \qquad \underset{\sim}{s} \in \mathrm{S} \quad \textbf{[5.8]}$$

satisfies the Bellman equation **4.5** *for the dynamic allocation problem* **4.4**; *in particular,* $\mathrm{F} \equiv \Phi$.

COROLLARY 5.2. *Every allocation strategy* $\underset{\sim}{\mathrm{T}}^*$ *of the index type—i.e., such that*

$$\mathrm{M_i(T_i^*(t))} = \mathop{\mathit{max}}_{1\le j\le d} \mathrm{M_j(T_j^*(t))} \quad \textbf{[5.9]}$$

$$\textit{on} \quad \{\underset{\sim}{\mathrm{T}}^*(\mathrm{t} + 1) = \underset{\sim}{\mathrm{T}}^*(\mathrm{t}) + \underset{\sim}{e}_i\}; \qquad \forall\, \mathrm{i} = 1, \ldots, \mathrm{d}, \qquad \mathrm{t} \in \mathbf{N}_0$$

is optimal for the problem **4.4**.

Remark 5.1: These results are proved in *Section 7*. The significance of statements **5.6** and **5.7** for the problem **4.7** is the following:

(*i*) According to statement **5.7**, it is optimal "always to pull an arm with maximal index, as long as this latter exceeds M."

(*ii*) According to statement **5.6**, "all arms should be abandoned once all indices have fallen below (or at) the retirement reward M."

Section 6. A Representation of the Value

Let us recall the notation $\underline{M}_i(\theta) = \min_{0 \le u \le \theta} M_i(u)$ of **3.2**, and introduce the *index random field*

$$\underline{M}(\underset{\sim}{s}) \triangleq \max_{1 \le j \le d} \underline{M}_j(s_j), \qquad \underset{\sim}{s} \in S. \qquad [6.1]$$

For an arbitrary allocation strategy $\underset{\sim}{I} \in \mathcal{A}(\underset{\sim}{0})$ of the index type, we have the following extensions of Eqs. **3.6** and **3.8** (with $t = 0$):

$$\underline{M}(\underset{\sim}{I}(\theta)) = \inf\left\{ m \ge 0 / \sum_{i=1}^{d} \sigma_i(0; m) \le \theta \right\}, \qquad \theta \in \mathbf{N}_0 \qquad [6.2]$$

$$\Phi(\underset{\sim}{0}) \equiv E \sum_{\theta=0}^{\infty} \beta^{\theta} \sum_{i=1}^{d} h_i(I_i(\theta+1))[I_i(\theta+1) - I_i(\theta)]$$

$$= (1-\beta) \cdot E \sum_{\theta=0}^{\infty} \beta^{\theta} \underline{M}(\underset{\sim}{I}(\theta)). \qquad [6.3]$$

The representation **6.3** for the value $\Phi(\underset{\sim}{0})$ of the dynamic allocation problem **4.4** (with $\underset{\sim}{s} = \underset{\sim}{0}$) is due to Mandelbaum (7), who derived it using a different methodology.

Section 7. Proofs

Proof of Lemma 2.1: The decrease follows readily from the decrease of the convex function $m \mapsto \varphi(t; m)$, given by

$$\varphi(t; m) \triangleq (V(t; m) - m)\beta^{t}$$

$$= \operatorname{esssup}_{\tau \in \mathcal{S}(t)} E\left[\sum_{u=t}^{\tau-1} \beta^{u}\{h(u+1) - m(1-\beta)\} \middle| \mathcal{F}(t) \right]. \qquad [7.1]$$

Hence, $\sigma_* \triangleq \lim_{\delta \downarrow 0} \sigma_t(m+\delta) \le \sigma_t(m)$, a.s. On the other hand, we have $\varphi(\sigma_t(m+1/k); m+\delta) = 0$ for $k > 1/\delta$, $k \in \mathbf{N}$. Letting $k \nearrow \infty$ we get $\varphi(\sigma_*; m+\delta) = 0$, $\forall\, \delta > 0$ and from the continuity of $m \mapsto \varphi(t; m)$ we obtain $\varphi(\sigma_*; m) = 0$, whence $\sigma_t(m) \le \sigma_*$, a.s.

Proof of Lemma 2.2: Only Eq. **2.8b** needs discussion, so we fix $m \in [0, \infty)$, recall

$$V(t; m) = E\left[\beta^{\sigma_t(m)-t} m + \sum_{u=t}^{\sigma_t(m)-1} \beta^{u-t} h(u+1) \middle| \mathcal{F}(t) \right], \qquad [7.2]$$

and observe that for $\delta > 0$ we have

$$Z(t; m) = E[Z(\sigma_t(m+\delta); m) | \mathcal{F}(t)], \qquad \text{a.s.} \qquad [7.3a]$$

from property **2.6**, *Lemma 2.1*, and optional sampling. This last equation yields, in conjunction with Eq. **7.2** (with m replaced by $m + \delta$):

$$V(t; m)$$

$$= E\left[\beta^{\sigma_t(m+\delta)-t} V(\sigma_t(m+\delta); m) + \sum_{u=t}^{\sigma_t(m+\delta)-1} \beta^{u-t} h(u+1) \middle| \mathcal{F}(t) \right]$$

$$\ge E[m\beta^{\sigma_t(m+\delta)-t} + V(t; m+\delta) - (m+\delta)\beta^{\sigma_t(m+\delta)-t} | \mathcal{F}(t)]. \qquad [7.3b]$$

Thanks to *Lemma 2.1*, we obtain from this

$$\overline{\lim}_{\delta \downarrow 0} \frac{V(t; m+\delta) - V(t; m)}{\delta} \le E[\beta^{\sigma_t(m)-t} | \mathcal{F}(t)], \qquad \text{a.s.} \qquad [7.4]$$

On the other hand, the supermartingale property of $Z(\cdot; m+\delta)$ gives $Z(t; m+\delta) \ge E[Z(\sigma_t(m); m+\delta) | \mathcal{F}(t)]$, a.s. Using this and Eq. **7.2**, we obtain

$$V(t; m+\delta)$$

$$\ge E\left[V(\sigma_t(m); m+\delta)\beta^{\sigma_t(m)-t} + \sum_{u=t}^{\sigma_t(m)-1} \beta^{u-t} h(u+1) \middle| \mathcal{F}(t) \right]$$

$$\ge E[(m+\delta)\beta^{\sigma_t(m)-t} + V(t; m) - m\beta^{\sigma_t(m)-t} | \mathcal{F}(t)], \qquad \text{a.s.,}$$

whence

$$\underline{\lim}_{\delta \downarrow 0} \frac{V(t; m+\delta) - V(t; m)}{\delta} \ge E[\beta^{\sigma_t(m)-t} | \mathcal{F}(t)], \qquad \text{a.s.} \qquad [7.5]$$

The conclusion follows from inequalities **7.4** and **7.5**.

Proof of Eqs. **3.7** *and* **3.8**: From the equivalence **3.4** we obtain $\beta^{\sigma_t(m)-t} = (1-\beta)\Sigma_{k=0}^{\infty} \beta^{k} 1_{\{\sigma_t(m) \le k+t\}} = (1-\beta)\Sigma_{\theta=t}^{\infty} \beta^{\theta-t} 1_{\{m \ge \underline{M}(t,\theta)\}}$, and thus from Eq. **2.8b** and Fubini, we obtain

$$V(t; m) - V(t; 0) = (1-\beta) \cdot E\left[\sum_{\theta=t}^{\infty} \beta^{\theta-t}(m - \underline{M}(t, \theta))^{+} \middle| \mathcal{F}(t) \right]$$

$$= (1-\beta) \cdot E\left[\sum_{\theta=\sigma_t(m)}^{\infty} \beta^{\theta-t}(m - \underline{M}(t, \theta)) \middle| \mathcal{F}(t) \right], \qquad \text{a.s.} \qquad [7.6]$$

On the other hand, we have from Eq. **7.2** and $\sigma_t(0) = \infty$:

$$V(t; m) - V(t; 0)$$

$$= E\left[m\beta^{\sigma_t(m)-t} - \sum_{\theta=\sigma_t(m)}^{\infty} \beta^{\theta-t} h(\theta+1) \middle| \mathcal{F}(t) \right], \qquad \text{a.s.}$$

We obtain Eq. **3.7** by comparing these two expressions; Eq. **3.8** follows then upon letting $m \uparrow \infty$.

Proof of Eq. **3.9**: From **2.2** and Eq. **3.7**, we have $V(t; 0) = E[\Sigma_{\theta=t}^{\infty} \beta^{\theta-t} h(\theta+1) | \mathcal{F}(t)] = (1-\beta) \cdot E[\Sigma_{\theta=t}^{\infty} \beta^{\theta-t} \underline{M}(t, \theta) | \mathcal{F}(t)]$, a.s. Now Eq. **3.9** follows from this and Eq. **7.6**.

Proof of properties **5.4** *and* **5.6**: From *Remark 3.2*, we have $0 \le \Pi_{i=1}^{d} (\partial^{+}/\partial m)\, V_i(s_i; m) \le 1$, which proves property **5.4**. From this same remark, $(\partial^{+}/\partial m)V_i(s_i; m) = 1$ for all $m \ge M$, $1 \le i \le d$ on the event $\{M \ge \max_{1 \le j \le d} M_j(s_j)\}$ and so we have $F(\underset{\sim}{s}; M) = M$ on this event.

Proof of property **5.5**: Fix $i \in \{1, \ldots, d\}$; for given $\underset{\sim}{s}^{(i)} \triangleq (s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_d)$ in $\mathbf{N}_0^{d-1}$, the function $m \mapsto P_i(\underset{\sim}{s}^{(i)}; m) \triangleq \Pi_{j \ne i}(\partial^{+}/\partial m)V_j(s_j; m)$ is increasing, and integrating by parts in Eq. **5.3** we obtain

$$F(\underset{\sim}{s}; M) \qquad [7.7]$$

$$= V_i(s_i; M)P_i(\underset{\sim}{s}^{(i)}; M) + \int_{M}^{K} V_i(s_i; m)\, d_m P_i(\underset{\sim}{s}^{(i)}; m).$$

Thus,

$$F(\underset{\sim}{s} + \underset{\sim}{e}_i; M)$$

$$= V_i(s_i + 1; M)P_i(\underset{\sim}{s}^{(i)}; M) + \int_{M}^{K} V_i(s_i + 1; m)\, d_m P_i(\underset{\sim}{s}^{(i)}; m),$$

$$E[F(\underset{\sim}{s} + \underset{\sim}{e}_i; M) | \mathcal{F}(\underset{\sim}{s})] = P_i(\underset{\sim}{s}^{(i)}; M) \cdot E[V_i(s_i + 1; M) | \mathcal{F}_i(s_i)]$$

$$+ \int_{M}^{K} E[V_i(s_i + 1; m) | \mathcal{F}_i(s_i)]\, d_m P_i(\underset{\sim}{s}^{(i)}; m),$$

whence

$$F(\underset{\sim}{s}; M) - h_i(s_i + 1) - \beta \cdot E[F(\underset{\sim}{s} + \underset{\sim}{e}_i; M)|\mathcal{F}(\underset{\sim}{s})]$$

$$= \varphi_i(s_i; M)P_i(\underset{\sim}{s}^{(i)}; M) + \int_M^K \varphi_i(s_i; m)\, d_m P_i(\underset{\sim}{s}^{(i)}; m), \qquad [\mathbf{7.8}]$$

with $\varphi_i(t; m) \triangleq V_i(t; m) - h_i(t + 1) - \beta \cdot E[V_i(t + 1; m)|\mathcal{F}_i(t)]$. From Eq. **2.7** this last expression in nonnegative, and thus the same is true for the right-hand side of Eq. **7.8**, a.s.

Proof of Eq. **5.7**: Consider arbitrary $m \in [M, K]$. On $\{m < M_i(s_i)\}$, we have $\varphi_i(s_i; m) = 0$, whereas on $\{m \geq M_i(s_i)\}$ we have from *Remark 3.2*, for every $j \neq i$,

$$m \geq M_j(s_j) \Leftrightarrow \frac{\partial^+}{\partial m} V_j(s_j; m) = 1$$

so that $P_i(\underset{\sim}{s}^{(i)}; m) = 1$. Thus, the right-hand side of Eq. **7.8** is then equal to zero.

Proof of Corollaries **5.1** and **5.2**: For any allocation strategy $\underset{\sim}{T} \in \mathcal{A}(\underset{\sim}{s})$, it is now easy to see that $Q(t; \underset{\sim}{T}) \triangleq \beta^t F(\underset{\sim}{T}(t)) + \Sigma_{u=0}^{t-1} \beta^u \Sigma_{i=1}^{d} h_i(T_i(u + 1))[T_i(u + 1) - T_i(u)]$ is a bounded, nonnegative, $\{\mathcal{G}(t)\}_{t \in \mathbf{N}_0}$-supermartingale and a martingale if $\underset{\sim}{T}$ is of index type. All claims follow readily.

A portion of this work was carried out during a brief visit by the second author at the University Pierre et Marie Curie (Paris VI) under the auspices of the National Science Foundation–INRIA Collaborative Research Effort in Stochastic Control (Grant NSF-INT-89-06965). The research of the second author is supported in part by the U.S. Army Research Office under Grant DAAL 03-91-G-0170.

1. Gittins, J. C. (1979) *J. R. Stat. Soc. B* **41,** 148–164.
2. Gittins, J. C. (1989) *Multi-Armed Bandits and Allocation Indices* (Wiley, Chichester, U.K.).
3. Whittle, P. (1980) *J. R. Stat. Soc. B* **42,** 143–149.
4. Whittle, P. (1982) *Optimization over Time: Dynamic Programming and Stochastic Control* (Wiley, New York).
5. Tsitsiklis, J. (1986) *IEEE Trans. Autom. Control* **AC-31,** 576–577.
6. Varaiya, P., Walrand, J. & Buyukkoc, C. (1985) *IEEE Trans. Autom. Control* **AC-30,** 426–439.
7. Mandelbaum, A. (1986) *Probab. Theory Rel. Fields* **71,** 129–147.
8. Mandelbaum, A. & Vanderbei, R. J. (1981) *Z. Wahrscheinlichkeitstheor. Verw. Geb.* **57,** 253–264.
9. Neveu, J. (1975) *Discrete-Parameter Martingales*, English transl. (North-Holand, Amsterdam).