The Annals of Statistics
1981, Vol. 9, No. 5, 987–1001

MULTI-ARMED BANDITS WITH DISCOUNT FACTOR NEAR ONE: THE BERNOULLI CASE

BY F. P. KELLY

University of Cambridge

Each of n arms generates an infinite sequence of Bernoulli random variables. The parameters of the sequences are themselves random variables, and are independent with a common distribution satisfying a mild regularity condition. At each stage we must choose an arm to observe (or pull) based on past observations, and our aim is to maximize the expected discounted sum of the observations. In this paper it is shown that as the discount factor approaches one the optimal policy tends to the rule of least failures, defined as follows: pull the arm which has incurred the least number of failures, or if this does not define an arm uniquely select from amongst the set of arms which have incurred the least number of failures an arm with the largest number of successes.

1. Introduction. Each of n arms generates an infinite sequence of Bernoulli random variables. The parameter of the sequence generated by arm i, θ_i, is itself a random variable. The random variables $\theta_1, \theta_2, \cdots, \theta_n$ are independent with common distribution function $F(z)$. At times $t = 0, 1, \cdots$ we choose an arm to observe (or pull) based on past observations. A pull on an arm at time t makes known $R(t)$, the next Bernoulli random variable in the sequence associated with the arm pulled. We interpret $R(t) = 0$ (respectively 1) as a failure (respectively success) at time t. Our aim is to maximize the expected total discounted reward

$$E[\textstyle\sum_{t=0}^{\infty} \beta^t R(t)]$$

where the discount factor $\beta \in (0, 1)$. The main result of this paper is that, provided $1 - F(1 - z)$ varies regularly at the origin, as the discount factor approaches one the optimal policy tends to the least failures rule, defined as follows: pull the arm which has incurred the least number of failures, or, if this does not define an arm uniquely, select from amongst the set of arms which have incurred the least number of failures an arm with the largest number of successes. The least failures rule is a slight variation on the play-the-winner rule introduced by Robbins (1952) in connection with the two-armed bandit.

The proof proceeds by establishing asymptotic bounds on the Gittins index (see Gittins, 1979) associated with each arm. In Section 2 the key definitions and results of Gittins are described, and in Section 3 the bounds are obtained.

Robbins (1952) proposed another rule with better performance, as judged by the expected average reward criterion, than the play-the-winner rule. In Section 4 the structure of discount optimal policies is investigated and it is shown that for discount factors near one these policies perform well according to the expected average reward criterion, even though the limit rule obtained from them does not.

One practical application motivating work on multi-armed bandits is the design of sequential clinical trials, and in this context the play-the-winner rule has been investigated as an easily described procedure with some reasonably good properties. Zelen (1969) has

Received January 31, 1980; revised August 1, 1980.

AMS 1970 *Subject classifications.* Primary 90C40, 62L05; secondary 62L15.

Key words and phrases. Bernoulli bandit process, Markov decision process, multi-armed bandit, limit rule, play-the-winner rule, least failures rule, discount optimality, expected average reward optimality.

evaluated numerically the performance of a finite horizon policy which begins by using the play-the-winner rule and then, after a certain number of trials, chooses the arm achieving the highest ratio of successes to trials and plays this arm until the horizon. Sobel and Weiss (1972; see also the papers referenced there) compare the performance of the play-the-winner rule and a vector-at-a-time rule, which by time $t = nT$ has pulled each arm exactly T times. In a subsequent paper it will be shown that if rewards have a normal or Poisson, rather than a Bernoulli, distribution then the limit rule obtained from discount optimal policies may well be a vector-at-a-time rule.

Results have previously been obtained from multi-armed bandits by allowing the discount factor to approach one: see Nash (1973) or Gittins (1979) for a scheduling problem and Kelly (1979) for a search problem. These earlier results are of an apparently different kind: in both the scheduling and the search problems an appropriately defined expected total discounted reward approaches the same constant under every policy, and the rate of approach determines the optimal policy under a differently constructed expected total cost criterion.

There are two other papers in the recent literature on multi-armed bandits which relate to the work described here. Berry and Fristedt (1979) have obtained interesting results applicable in both finite horizon undiscounted and infinite horizon discounted frameworks, and certain of these results, to be discussed further in Section 3, can be interpreted as bounds on the Gittins index. Rothschild (1974), in his development of a model for market pricing, has investigated the long run behaviour of discount optimal policies. This work is referred to in Section 4.

The relationship between discount optimality and expected average reward optimality has previously been investigated within the broad context of Markov decision processes; see Blackwell (1962), Ross (1968) and Veinott (1974). However these authors have had to impose restrictive conditions not satisfied by the processes discussed in this paper. The multi-armed bandits considered here do not exhibit absorption or recurrence but rather a form of convergence.

2. Bandit processes. A bandit process is defined on a state space X, assumed to be a Borel subset of a complete separable metric space. At times $t = 0, 1, \cdots$ the bandit process can be either frozen or continued. If frozen, its state $x(\in X)$ remains the same and no reward is received. If continued, its state changes from x to a new state $y \in X$, selected in accordance with a transition mechanism $q(\cdot \mid x)$, and a reward $R(x, y)$ is received. The transition mechanism q is a regular conditional probability on X given X, and the reward function R is a Baire function on $X \times X$; for a detailed discussion of these, the usual, measurability conditions see Blackwell (1965), Strauch (1966) or Hinderer (1970). The reward function R is assumed to satisfy the condition

$$E[\textstyle\sum_{t=0}^{\infty} \beta^t \,|\, R(x(t), x(t+1))\,|\,] < \infty \qquad \forall \beta \in (0, 1), \forall x(0) \in X \tag{1}$$

where $(x(t), t = 0, 1, \cdots)$ is the Markov chain with initial state $x(0)$ generated by the transition mechanism q. Observe that $(x(t), t = 0, 1, \cdots)$ is the sequence of states obtained if the bandit process is continued at times $t = 0, 1, \cdots$. For a bandit process (X, q, R) in state x define the Gittins index $\nu_\beta(x)$ by

$$\nu_\beta(x) = \sup_{\tau > 0} \frac{E[\sum_{t=0}^{\tau-1} \beta^t R(x(t), x(t+1))]}{E[\sum_{t=0}^{\tau-1} \beta^t]} \tag{2}$$

where the supremum is taken over all positive stopping times for the Markov chain $(x(t), t = 0, 1, \cdots)$ with initial state $x(0) = x$.

LEMMA 2.1. *The Gittins index $\nu_\beta(x)$ is non-decreasing in β for each state x.*

PROOF. Suppose $0 < \beta_1 < \beta_2 < 1$. Let τ_1 be a positive stopping time. Define

$$\tau_2 = \min\{\tau_1, T\},$$

where T is a geometric random variable independent of $(x(t), t = 0, 1, \cdots)$ and τ_1 with distribution

$$P(T = l) = \left(1 - \frac{\beta_1}{\beta_2}\right)\left(\frac{\beta_1}{\beta_2}\right)^{l-1} \qquad l = 1, 2, \cdots.$$

Observe that τ_2 is a positive (randomized) stopping time, and that

$$\frac{E[\sum_{t=0}^{\tau_2-1} \beta_2^t R(x(t), x(t+1))]}{E[\sum_{t=0}^{\tau_2-1} \beta_2^t]} = \frac{E[\sum_{t=0}^{\tau_1-1} \beta_1^t R(x(t), x(t+1))]}{E[\sum_{t=0}^{\tau_1-1} \beta_1^t]}.$$

Thus, from definition (2),

$$\nu_{\beta_2}(x) \geq \nu_{\beta_1}(x) \qquad \forall\, x \in X. \qquad \square$$

A multi-armed bandit is a collection of n bandit processes $((X_i, q_i, R_i), 1 \leq i \leq n)$ with the constraint that at times $t = 0, 1, \cdots$ just one of them is chosen to be continued and the others are frozen. Let $\mathbf{x}(t) = (x_1(t_1), x_2(t_2), \cdots, x_n(t_n))$ denote the state of the multi-armed bandit at time t, where t_i is the number of times bandit process i has been continued and $t = t_1 + t_2 + \cdots + t_n$. We shall sometimes write $\mathbf{x}(t) = (x_1, x_2, \cdots, x_n)$. Let $a(t) \in \{1, 2, \cdots, n\}$ be the bandit process continued (or arm pulled) at time t. A policy π is a sequence $\pi_0, \pi_1, \cdots$, where π_t is a regular conditional probability on $\{1, 2, \cdots, n\}$ given the history of the system till time t, $h(t) = (\mathbf{x}(0), a(0), \mathbf{x}(1), \cdots, a(t-1), \mathbf{x}(t))$. Given $h(t)$ the policy π pulls arm i with probability $\pi_t(i \mid h(t))$. The expected total discounted reward associated with policy π is

$$V_\beta(\pi) = E[\textstyle\sum_{t=0}^{\infty} \beta^t R(t)]$$

where $R(t)$ is the reward received at time t under the policy π. The reward $V_\beta(\pi)$ is a function of the initial state $\mathbf{x}(0)$. Call a policy π^* β-optimal if

$$V_\beta(\pi^*) = \sup_\pi V_\beta(\pi).$$

THEOREM 2.2 (Gittins). *The following are equivalent*:

(i) π^* *is* β*-optimal*
(ii) $\pi_t^*(i \mid h(t)) = 0$ *if* $\nu_\beta(x_i) < \nu_\beta(x_j)$,

except possibly on a set of histories with probability zero under π^*.

The theorem shows that a β-optimal policy is obtained by choosing at each stage to pull the arm whose index is then maximal. If two or more arms tie for the maximal index, the arm to be pulled can be selected arbitrarily from amongst them.

REMARK 2.3. A standard arm λ is a bandit process with state space $X = \{\lambda\}$ and reward function $R(\lambda, \lambda) = \lambda$. The Gittins index of a standard arm λ is thus λ. Consider now a two-armed bandit comprising a standard arm λ and a bandit process in initial state x. Theorem 2.2 shows that a β-optimal policy could start by pulling either of the two arms if and only if $\nu_\beta(x) = \lambda$. This method of determining the Gittins index $\nu_\beta(x)$ is called comparison with a standard arm, and provides an alternative to a direct use of the definition (2).

REMARK 2.4. It is usual to replace the function $R(x, y)$ with a function of x alone by taking the expectation over the new state y in accordance with the transition mechanism $q(\cdot \mid x)$. We shall find it more convenient to use the actual reward received, $R(x, y)$. Observe that under any policy π

$$E[\textstyle\sum_{t=0}^{\infty} \beta^t |\, R(t)\,|] \leq \sum_{i=1}^{n} E[\beta^t |\, R_i(x_i(t), x_i(t+1))\,|]$$

and hence assumption (1) implies that $\sup_\pi V_\beta(\pi)$ is finite. In this paper all the reward functions considered will be uniformly bounded and hence assumption (1) will be satisfied.

3. Asymptotic bounds for Bernoulli bandits. The state of a Bernoulli bandit process is a distribution function G on [0, 1], interpreted as the distribution for the unknown parameter θ of the Bernoulli reward sequence generated by the bandit. Define operators σ, ϕ on a distribution function G as follows:

$$(\sigma G)(p) = \frac{\int_0^p z\, dG(z)}{\int_0^1 z\, dG(z)} \qquad 0 \leq p \leq 1$$

$$(\phi G)(p) = \frac{\int_0^p (1-z)\, dG(z)}{\int_0^1 (1-z)\, dG(z)} \qquad 0 \leq p \leq 1$$

These correspond to the posterior distributions for the parameter θ after observing a success or a failure respectively. Thus the operators commute and $\sigma^s\phi^f G$ is the posterior distribution after observing s successes and f failures. If the bandit process is continued from state G, there is a success with probability $\int_0^1 z\, dG(z)$ and a failure otherwise. A success means that the state becomes σG and unit reward is received. A failure means that the state becomes ϕG and zero reward is received. More formally, the state space, the transition mechanism and the reward function of a Bernoulli bandit process with initial state F are

$$X = \{\sigma^s\phi^f F : s, f \in \mathbb{Z}_+\},$$

$$q(\sigma G \,|\, G) = \int_0^1 z\, dG(z), \qquad q(\phi G \,|\, G) = \int_0^1 (1-z)\, dG(z),$$

$$R(G, \sigma G) = 1, \qquad R(G, \phi G) = 0.$$

Observe that if the process is continued indefinitely then with probability one the state will converge weakly to a degenerate distribution with unit mass concentrated at the initially unknown value of the parameter θ. Indeed this property could be taken as the definition of the variable θ.

Let $\nu_\beta(F)$ be the Gittins index for a Bernoulli bandit process with initial state F. Assume henceforth that $F(z) < 1$ for all $z \in (0, 1)$, and that $F(z) > 0$ for some $z \in (0, 1)$. The only interesting case that these assumptions exclude, when $\sup\{z : F(z) < 1\} < 1$, is discussed in Remark 4.14. In this section we establish upper and lower bounds for $\nu_\beta(F)$.

THEOREM 3.1. *There exists a function* $\lambda = \lambda_\beta(F)$ *such that*

$$\lambda_\beta(F) \geq \nu_\beta(F) \qquad \forall \beta \in (0,1)$$

and

$$\lim_{\beta \to 1} (1-\beta)^{-1} \int_\lambda^1 (z - \lambda)\, dF(z) = 1 - \int_0^1 z\, dF(z). \tag{3}$$

PROOF. Define an informative arm to be a bandit process which behaves as does a Bernoulli bandit process, except that after the first continuation the true value of the

parameter becomes known. More precisely, from the initial state F,

$$q((\sigma F, \theta)\,|\,F) = \theta\, dF(\theta), \qquad q((\phi F, \theta)\,|\,F) = (1-\theta)\, dF(\theta),$$
$$R(F, (\sigma F, \theta)) = 1, \qquad R(F,(\phi F, \theta)) = 0,$$

and from any subsequent state $(G, \theta) = (\sigma^s \phi^f F, \theta)$,

$$q((\sigma G, \theta)\,|\,(G, \theta)) = \theta, \qquad q((\phi G, \theta)\,|\,(G, \theta)) = 1-\theta,$$
$$R((G, \theta), (\sigma G, \theta)) = 1, \qquad R((G, \theta), (\phi G, \theta)) = 0.$$

Let $\lambda_\beta(F)$ be the Gittins index for the informative arm. Observe that each positive stopping time available for the Bernoulli bandit process corresponds to a positive stopping time for the informative arm based solely on the first component of its state, and hence from the definition (2)

$$\lambda_\beta(F) \geq \nu_\beta(F) \qquad \forall \beta \in (0, 1).$$

To calculate $\lambda_\beta(F)$ compare the informative arm with a standard arm λ. If a β-optimal policy for the resulting two-armed bandit could start by pulling the standard arm then by Theorem 2.2 it could continue pulling this arm indefinitely, resulting in a total reward $\lambda(1-\beta)^{-1}$. If a β-optimal policy starts by pulling the informative arm then the parameter θ becomes known: if $\theta > \lambda$ the β-optimal policy stays with the informative arm, while if $\theta < \lambda$ it reverts to the standard arm. Thus a β-optimal policy for the two-armed bandit could start by pulling either of the two arms if

$$\frac{\lambda}{1-\beta} = \int_0^1 z\, dF(z) + \frac{\beta}{1-\beta}\left[\int_\lambda^1 z\, dF(z) + \lambda \int_0^\lambda dF(z)\right].$$

Rearranging this equation for $\lambda = \lambda_\beta(F)$ we obtain

$$(1-\beta)^{-1} \int_\lambda^1 (z-\lambda)\, dF(z) = \int_0^\lambda (\lambda - z)\, dF(z). \tag{4}$$

The right hand side of equation (4) is bounded above and hence

$$\lim_{\beta \to 1} \int_\lambda^1 (z-\lambda)\, dF(z) = 0.$$

Thus as $\beta \uparrow 1$ the function $\lambda_\beta(F) \uparrow 1$, and so from equation (4)

$$\lim_{\beta \to 1} (1-\beta)^{-1} \int_\lambda^1 (z-\lambda)\, dF(z) = \int_0^1 (1-z)\, dF(z). \qquad \square$$

Define

$$F^*(z) = 1 - F(1-z), \qquad 0 \leq z \leq 1.$$

COROLLARY 3.2. *There exists a function* $\lambda = \lambda_\beta(F)$ *such that*

$$\lambda_\beta(F) \geq \nu_\beta(F) \qquad \forall \beta \in (0, 1)$$

and

$$\liminf_{\beta \to 1} (1-\beta)^{-1}(1-\lambda) F^*(1-\lambda) > 0.$$

PROOF. Observe that

$$\int_\lambda^1 (z-\lambda)\, dF(z) \leq (1-\lambda) F^*(1-\lambda) \qquad 0 \leq \lambda \leq 1$$

and hence the result follows from Theorem 3.1. Note that although the first expression is continuous the second expression may be only right-continuous. □

THEOREM 3.3. *For each* $\varepsilon \in (0, 1)$ *there exists a function* $\mu = \mu_\beta(F)$, *possibly depending on* ε, *such that*

$$\mu_\beta(F) \leq \nu_\beta(F) \qquad \forall \beta \in (0, 1)$$

and

$$\limsup_{\beta \to 1} (1-\beta)^{-1}(1-\mu)F^*((1-\mu)(1-\varepsilon)) < \infty. \tag{5}$$

PROOF. We shall find a lower bound for $\nu_\beta(F)$ by using a particular form of stopping time in definition (2). Identify the state space of the Bernoulli bandit process with $\mathbb{Z}_+^2$ using the bijection

$$(s, f) \leftrightarrow (\sigma^s \phi^f F).$$

Recall that $(x(t), t = 0, 1, \cdots)$ is the sequence of states obtained if the bandit process is continued indefinitely, and that $R(t) = R(x(t), x(t+1))$ is then the reward obtained at time t. For each $k > 0$, define the stopping set $D_k \subset \mathbb{Z}_+^2$ by

$$D_k = \{(s, f) : f > 0,\ s < (f+1)k\}$$

and the positive stopping time τ_k by

$$\begin{aligned} \tau_k &= \inf\{t : x(t) \in D_k\} && \text{if } \exists\ t : x(t) \in D_k, \\ &= \infty && \text{otherwise.} \end{aligned}$$

Let

$$\begin{aligned} \mathrm{T} &= \inf\{t : R(t) = 0\} && \text{if } \exists t : R(t) = 0, \\ &= \infty && \text{otherwise.} \end{aligned}$$

Observe that if $T \leq k$ then $\tau_k = T + 1$, and if $T \geq k$ then

$$\frac{1}{r}\sum_{t=0}^{r-1} R(t) \geq \frac{k}{k+1} \qquad 1 \leq r \leq \tau_k. \tag{6}$$

Where no confusion can arise we shall omit the subscript k and write τ for τ_k. Define

$$\begin{aligned} U(\tau) &= \sum_{t=0}^{\tau-1} \beta^t R(t) \qquad \text{and} \\ W(\tau) &= \sum_{t=0}^{\tau-1} \beta^t. \end{aligned}$$

From definition (2) it follows that a lower bound for $\nu_\beta(F)$ is given by $\mu' = \mu'_\beta(F)$ where

$$\begin{aligned} \mu' &= \frac{EU(\tau)}{EW(\tau)} \\ &= \frac{E[U(\tau)|\tau \leq k]P(\tau \leq k) + E[U(\tau)|k < \tau < \infty]P(k < \tau < \infty) + E[U(\tau)|\tau = \infty]P(\tau = \infty)}{E[W(\tau)|\tau \leq k]P(\tau \leq k) + E[W(\tau)|k < \tau < \infty]P(k < \tau < \infty) + E[W(\tau)|\tau = \infty]P(\tau = \infty)}. \end{aligned}$$

Let

$$\alpha_k = E[U(\tau)|\tau \leq k]P(\tau \leq k). \tag{7}$$

Then

$$\begin{aligned} E[W(\tau)|\tau \leq k]P(\tau \leq k) &= E[U(\tau) + \beta^{\tau-1}|\tau \leq k]P(\tau \leq k) \\ &\leq \alpha_k + 1. \end{aligned} \tag{8}$$

Observe also that

$$\frac{E[U(\tau)\,|\,k<\tau<\infty]}{E[W(\tau)\,|\,k<\tau<\infty]}\geq\frac{k}{k+1} \tag{9}$$

and

$$E[U(\tau)\,|\,\tau=\infty]\geq\frac{k}{k+1}(1-\beta)^{-1}, \tag{10}$$

since if $\tau > k$ it is possible to partition the set $\{R(0), R(1), \cdots R(\tau-1)\}$ into disjoint subsets in such a way that each subset contains a single failure and at least k successes all of which occurred before the failure (cf. inequality (6)). From relations (7), (8) and (10) we obtain

$$\mu'\geq\frac{\alpha_k+E[U(\tau)\,|\,k<\tau<\infty]P(k<\tau<\infty)+k(k+1)^{-1}(1-\beta)^{-1}P(\tau=\infty)}{\alpha_k+1+E[W(\tau)\,|\,k<\tau<\infty]P(k<\tau<\infty)+(1-\beta)^{-1}P(\tau=\infty)}.$$

Now $\alpha_k \leq k$, and so

$$\frac{\alpha_k}{\alpha_k+1}\leq\frac{k}{k+1}. \tag{11}$$

This together with inequality (9) implies that

$$\mu'\geq\frac{\alpha_k+k(k+1)^{-1}(1-\beta)^{-1}P(\tau=\infty)}{\alpha_k+1+(1-\beta)^{-1}P(\tau=\infty)}. \tag{12}$$

To proceed further we need more information on $P(\tau=\infty)$. Condition on θ, the underlying parameter of the sequence $R(0), R(1), \cdots$. Then

$$P(\tau=\infty\,|\,\theta)=\theta^k P(\textstyle\sum_{r=1}^{m} T_r\geq m(k+1),\ m=1,2,\cdots\,|\,\theta),$$

where conditional on θ the random variables $T_1, T_2, \cdots$ are independent and identically distributed with

$$P(T_1=l\,|\,\theta)=(1-\theta)\theta^{l-1} \qquad l=1,2,\cdots$$

Let $Y_1, Y_2, \cdots$ be independent exponentially distributed random variables with unit mean. Then $(-\log\theta)^{-1}Y_r$ is a random variable which is stochastically smaller than T_r and so

$$P(\tau=\infty\,|\,\theta)\geq\theta^k P\left(\frac{1}{m}\textstyle\sum_{r=1}^{m} Y_r\geq-(k+1)\log\theta,\ m=1,2,\cdots\right).$$

Choose $\delta>0$ and let

$$\theta=1-(k+1)^{-1}(1+2\delta)^{-1}.$$

Then

$$\lim_{k\to\infty}\theta^k=\exp[-(1+2\delta)^{-1}]$$

and

$$\lim_{k\to\infty}[-(k+1)\log\theta]=(1+2\delta)^{-1}.$$

Thus there exists K_δ such that for all $k\geq K_\delta$

$$P(\tau=\infty\,|\,\theta=1-(k+1)^{-1}(1+2\delta)^{-1})\geq\exp[-(1+2\delta)^{-1}]\cdot P\left(\frac{1}{m}\textstyle\sum_{r=1}^{m} Y_r\geq(1+\delta)^{-1},\ m=1,2,\cdots\right).$$

But by the strong law of large numbers

$$\frac{1}{m}\textstyle\sum_{r=1}^{m} Y_r\to 1 \quad \text{a.s.}$$

and so we can deduce that there exists a $P_\delta > 0$ such that

$$P\left(\frac{1}{m}\sum_{r=1}^{m} Y_r \geq (1+\delta)^{-1},\ m = 1, 2, \cdots\right) \geq P_\delta.$$

Thus there exists a $Q_\delta > 0$ such that for all $k \geq K_\delta$

$$P(\tau = \infty \mid \theta = 1 - (k+1)^{-1}(1+2\delta)^{-1}) \geq Q_\delta.$$

Now $P(\tau = \infty \mid \theta)$ is an increasing function of θ and hence

$$P(\tau = \infty) \geq Q_\delta F^*((k+1)^{-1}(1+2\delta)^{-1})$$

for $k \geq K_\delta$. Thus, from inequalities (11) and (12),

$$\mu' \geq \frac{\alpha_k + k(k+1)^{-1}(1-\beta)^{-1}Q_\delta F^*((k+1)^{-1}(1+2\delta)^{-1})}{\alpha_k + 1 + (1-\beta)^{-1}Q_\delta F^*((k+1)^{-1}(1+2\delta)^{-1})} \tag{13}$$

for $k \geq K_\delta$. Now choose k as a function of β so that

$$\begin{aligned}\delta Q_\delta (k+2)^{-1}F^*((k+2)^{-1}(1+2\delta)^{-1}) < 1-\beta \\ \leq \delta Q_\delta (k+1)^{-1}F^*((k+1)^{-1}(1+2\delta)^{-1}).\end{aligned} \tag{14}$$

Observe that as β approaches one, the chosen value of k approaches infinity, since $F^*(z) > 0$ for all $z > 0$. From inequalities (13) and (14) it follows that for β sufficiently close to one, or equivalently for k sufficiently large,

$$\mu' \geq 1 - (1+\delta)(k+1)^{-1}.$$

For β this close to one define $\mu = \mu_\beta(F)$ by

$$\mu = 1 - (1+\delta)(k+1)^{-1}$$

and for other values of β let $\mu_\beta(F) = \nu_\beta(F)$. Then $\mu_\beta(F) \leq \nu_\beta(F)$ and further, from inequality (14), for β sufficiently close to one

$$\delta Q_\delta (1-\mu)(2+\delta-\mu)^{-1}F^*((1-\mu)(2+\delta-\mu)^{-1}(1+2\delta)^{-1}) < 1-\beta.$$

Also

$$(2+\delta-\mu) \leq 1+2\delta$$

and so

$$\delta Q_\delta (1-\mu)(1+2\delta)^{-1}F^*((1-\mu)(1+2\delta)^{-2}) < 1-\beta.$$

Thus

$$\limsup_{\beta\to 1}(1-\beta)^{-1}F^*((1-\mu)(1+2\delta)^{-2}) \leq \frac{1+2\delta}{\delta Q_\delta}$$

from which assertion (5) follows. □

REMARK 3.4 The bounds obtained in Theorems 3.1 and 3.3 also apply to rather more general bandit processes. Consider, for example, a bandit process in which the rewards $(R(0), R(1), \cdots)$ are a sequence of independent bounded random variables with common, but unknown, distribution function H. Without loss of generality we can assume the rewards lie in the interval [0, 1]. Suppose that the initial state of the bandit process is a prior distribution for H, and let

$$F(z) = \text{Prob}\left\{\int_0^1 \theta\, dH(\theta) \leq z\right\}$$

where the probability is evaluated with respect to the prior distribution for H. Let ν_β be the Gittins index of the bandit process in its initial state. Then there exist upper and lower

bounds, $\lambda_\beta \geq \nu_\beta \geq \mu_\beta$, satisfying relations (3) and (5) respectively. To establish the upper bound consider an informative arm which reveals the distribution H after the first pull, and proceed as in the proof of Theorem 3.1. To establish the lower bound let $R'(t)$ be a Bernoulli random variable with mean $R(t)$, and observe that the lower bound can be achieved using stopping times based solely upon the sequence $(R'(0), R'(1), \cdots)$. Note that these bounds apply even if continuation of the bandit process at time t supplies more information about H than just $R(t)$. For example in a sequential clinical trial there might be more than just two possible responses to a treatment, and a patient's response might be just part of the information about the treatment acquired from observation of that patient. We shall not pursue the point further; the results to be obtained in the next section depend not just upon the bounds on the Gittins index, but also upon the way the Gittins index alters as new information becomes available.

REMARK 3.5. Upper and lower bounds on $\nu_\beta(F)$ follow from Theorems 5.1 and 5.2 of Berry and Fristedt (1979); see their Example 5.3. Their upper bound corresponds to the Gittins index of a bandit process which behaves as does a Bernoulli bandit process except that upon the occasion of the first failure the true value of the parameter becomes known. The upper bound of Berry and Fristedt is thus sharper than that obtained in Theorem 3.1. Our upper bound will however be sufficient for our purposes. It is determined by a slightly simpler implicit relation (viz. equation (4)) and generalizes to the situation described in Remark 3.4. The family of lower bounds derived by Berry and Fristedt correspond to stopping sets of the form

$$D_k = \{(s, f) : f > 0, s < k\}$$

and are not sharp enough for our purposes.

The function $F^*(z)$ varies regularly with exponent $\rho (0 \leq \rho < \infty)$ if it can be written in the form

$$F^*(z) = z^\rho L(z)$$

with L slowly varying at the origin (Feller, 1971). Thus if

$$\frac{dF(z)}{dz} = \frac{(a + b + 1)!}{a!b!} z^a (1 - z)^b$$

corresponding to a Beta distribution with parameters (a, b) then $F^*(z)$ varies regularly with exponent $b + 1$.

THEOREM 3.6. *If $F^*(z)$ varies regularly with exponent ρ then for any fixed $\delta > 0$ and all β sufficiently close to one*

$$(1 - \beta)^{\kappa+\delta} < 1 - \nu_\beta(F) < (1 - \beta)^{\kappa-\delta}$$

where

$$\kappa = (\rho + 1)^{-1}.$$

PROOF. Since $F^*(z)$ varies regularly with exponent ρ,

$$z^{\rho+\eta} < F^*(z) < z^{\rho-\eta}$$

for any fixed $\eta > 0$ and all z sufficiently small. Now from Corollary 3.2

$$\liminf_{\beta \to 1} (1 - \beta)^{-1}(1 - \nu) F^*(1 - \nu) > 0.$$

Thus

$$\liminf_{\beta \to 1} (1 - \beta)^{-1}(1 - \nu)^{\rho+1-\eta} > 0$$

for all fixed $\eta > 0$, and the first inequality of the theorem follows.

From Theorem 3.3

$$\limsup_{\beta\to 1}(1-\beta)^{-1}(1-\nu)F^*((1-\nu)(1-\varepsilon)) < \infty$$

for any fixed $\varepsilon \in (0, 1)$. Thus

$$(1-\varepsilon)^{\rho+\eta}\limsup_{\beta\to 1}(1-\beta)^{-1}(1-\nu)^{\rho+1+\eta} < \infty$$

for any fixed $\eta > 0$, and the second inequality of the theorem follows. □

4. The least failures rule. In this section we shall consider a multi-armed bandit in which each of the n constituent bandit processes is a Bernoulli bandit process with initial state F. We can write the state of the multi-armed bandit at time t as $\mathbf{x}(t) = (s_1, f_1, s_2, f_2, \cdots, s_n, f_n)$, where s_i (respectively f_i) is the number of successes (respectively failures) observed so far on arm i. Write $\nu_\beta(s, f)$ for the Gittins index of an arm in state $\sigma^s\phi^f F$.

LEMMA 4.1 (Bellman). *For all* $s, f \in \mathbb{Z}_+$

$$\nu_\beta(s, f) < \nu_\beta(s+1, f) \qquad \forall \beta \in (0, 1).$$

PROOF. See Bellman (1956, Theorem 2), or for a recent generalization Berry and Fristedt (1979, Theorem 4.1). □

REMARK 4.2. Lemma 4.1 in conjunction with Theorem 2.2 establishes that a β-optimal policy for the multi-armed bandit has the stay-on-a-winner property (cf. Berry, 1972) for all $\beta \in (0, 1)$.

LEMMA 4.3. *If* $F^*(z)$ *varies regularly then for all* $s, f \in \mathbb{Z}_+, n > 0$, *there exists* $B < 1$ *such that*

$$\nu_\beta(s+n, f+1) < \nu_\beta(s, f) \qquad \forall \beta \in (B, 1).$$

PROOF. If $F^*(z)$ varies regularly with exponent ρ then $(\sigma^s\phi^f F)^*(z)$ varies regularly with exponent $\rho + f$. Similarly $(\sigma^{s+n}\phi^{f+1}F)^*(z)$ varies regularly with exponent $\rho + f + 1$, whatever the value of n. The result thus follows from Theorem 3.6. □

REMARK 4.4. Observe that the constant B in Lemma 4.3 depends upon the values s, f and n. Indeed the rather coarse bound $\nu_\beta(F) \geq \int_0^1 z\, dF(z)$, obtained using the stopping time $\tau \equiv 1$ in definition (2), shows that for fixed s, f the constant B in Lemma 4.3 can be forced arbitrarily close to one by choosing n sufficiently large.

Call a policy π a symmetric Markov policy if $\pi_t(i \mid h(t))$ depends on t and $h(t)$ only through $\mathbf{x}(t)$, so that we can write $\pi_t(i \mid h(t)) = \pi(i \mid \mathbf{x}(t))$, and if

$$(s_i, f_i) = (s_j, f_j) \Rightarrow \pi(i \mid s_1, f_1, \cdots, f_n) = \pi(j \mid s_1, f_1, \cdots, f_n)$$

Theorem 2.2 shows that from amongst the symmetric Markov policies there exists one that is β-optimal: call it π_β. We shall no longer require the subscript t on π, and hence no confusion should arise. Recall that π_β is shorthand for the collection of distributions $(\pi_\beta(\mathbf{x}), \mathbf{x} \in X^n)$ where $\pi_\beta(\mathbf{x}) = (\pi_\beta(i \mid \mathbf{x}), 1 \leq i \leq n)$.

Let

$$S_1 = \{i : f_i = \min_{1\leq j\leq n} f_j\},$$

$$S_2 = \{i : i \in S_1, s_i = \max_{j\in S_1} s_j\}.$$

Let m be the cardinality of S_2. Define a symmetric Markov policy π_{LF}, called the least failures rule, as follows:

$$\begin{aligned} \pi_{\mathrm{LF}}(s_1, f_1, \cdots, f_n) &= m^{-1} && i \in S_2, \\ &= 0 && \text{otherwise.} \end{aligned}$$

The next theorem establishes that the β-optimal policy π_β tends to a limit rule as β approaches one, and that the limit rule is the least failures rule.

THEOREM 4.5. *If $F^*(z)$ varies regularly then*

$$\pi_\beta \to \pi_{\mathrm{LF}} \qquad as \qquad \beta \to 1$$

in the sense that for each $(s_1, f_1, \cdots, f_n) \in \mathbb{Z}_+^{2n}$ there exists $B < 1$ such that

$$\pi_\beta(s_1, f_1, \cdots, f_n) = \pi_{\mathrm{LF}}(s_1, f_1, \cdots, f_n) \qquad \forall \beta \in (B, 1).$$

PROOF. Consider a particular state $(s_1, f_1, \cdots, f_n)$. Theorem 2.2 together with Lemma 4.3 establish that for β sufficiently close to one the distribution $\pi_\beta(s_1, f_1, \cdots, f_n)$ is concentrated on the set S_1. Theorem 2.2 together with Lemma 4.1 then imply that $\pi_\beta(s_1, f_1, \cdots, f_n)$ is concentrated on the set S_2. The symmetry of π_β ensures $\pi_\beta(s_1, f_1, \cdots, f_n)$ is uniform on the set S_2, and hence identical to $\pi_{\mathrm{LF}}(s_1, f_1, \cdots, f_n)$.

REMARK 4.6. Remark 4.4 shows that the convergence obtained in Theorem 4.5 cannot be uniform over states $(s_1, f_1, \cdots, f_n)$. Define

$$C_\beta = \{(s_1, f_1, \cdots, f_n) : \pi_\beta(s_1, f_1, \cdots, f_n) = \pi_{\mathrm{LF}}(s_1, f_1, \cdots, f_n)\}.$$

Then the β-optimal policy π_β agrees with the least failures rule until the state of the multi-armed bandit leaves the set C_β. Let T_β^1 be the first stage at which this occurs. Since only a finite number of states are accessible within a finite number of stages, by making β sufficiently close to one it is possible to ensure that the minimum possible value for T_β^1 is arbitrarily large. This defines another sense in which the β-optimal symmetric Markov policies tend to the least failures rule as the discount factor approaches one.

REMARK 4.7. The least failures rule is a slight variation on the play-the-winner rule, introduced in connection with two-armed bandits by Robbins (1952). The play-the-winner rule is equally likely to pull arm 1 or 2 at time $t = 0$, and thereafter stays with arm i if a success has just been achieved with that arm and pulls the other arm otherwise. For an n-armed bandit the obvious generalization is the cyclic play-the-winner rule: this is equally likely to pull any of the n arms at time $t = 0$, and thereafter stays with arm i if a success has just been achieved with that arm and pulls arm $i + 1$ otherwise, where arm $n + 1$ is identified with arm 1. A variation of the cyclic play-the-winner rule is mentioned by Sobel and Weiss (1972). This labels the n arms in a random order at time $t = 0$, and reorders the n arms according to the number of successes (using randomization for ties) after every complete cycle of n failures, one from each population. This variation is precisely the least failures rule.

Let $R_i(t) = R_i(x_i(t), x_i(t + 1))$ be the reward received on pull $t + 1$ of arm i. To facilitate the simultaneous discussion of more than one policy we shall henceforth define all random events on a common probability space, with a sample point ω of this space determining the sequences $(x_i(t), t = 0, 1, \cdots)$, $1 \le i \le n$, and hence the sequences $(R_i(t), t = 0, 1, \cdots)$, $1 \le i \le n$. The behaviour of a multi-armed bandit can thus be regarded as a function of the sample point ω and the policy adopted. If $(R(t), t = 0, 1, \cdots)$ and $(R'(t), t = 0, 1, \cdots)$ are the sequences of rewards obtained from a multi-armed bandit under the least failures rule and the cyclic play-the-winner rule respectively then it is readily shown that

$$\left|\sum_{t=0}^{T-1} R(t) - \sum_{t=0}^{T-1} R'(t)\right| \le n - 1 \qquad T \ge 1$$

(for all ω). In this sense there is little difference between the least failures rule and the cyclic play-the-winner rule.

REMARK 4.8. Under the least failures rule the number of consecutive successes preceding each failure on arm i has a geometric distribution with mean $\theta_i(1 - \theta_i)^{-1}$. Hence,

from the strong law of large numbers,

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} R(t) = \left(\sum_{i=1}^{n} \frac{1}{1-\theta_i}\right) \sum_{i=1}^{n} \frac{\theta_i}{1-\theta_i} \quad \text{a.s.}$$

and so

$$\lim_{T\to\infty} E\left[\frac{1}{T}\sum_{t=0}^{T-1} R(t) \,\middle|\, \theta_1, \theta_2, \cdots, \theta_n\right] = \left(\sum_{i=1}^{n} \frac{1}{1-\theta_i}\right) \sum_{i=1}^{n} \frac{\theta_i}{1-\theta_i}. \tag{15}$$

The limit (15) is a weighted average of $\theta_1, \theta_2, \cdots, \theta_n$ and is less than $\max(\theta_1, \theta_2, \cdots, \theta_n)$ except when $\max(\theta_1, \theta_2, \cdots, \theta_n) = 1$ or $\theta_1 = \theta_2 = \cdots = \theta_n$. Despite this we shall see that as β approaches one, the performance of a β-optimal policy as judged by the expected average reward criterion approaches the best possible, namely $\max(\theta_1, \theta_2, \cdots, \theta_n)$. We shall establish this as a consequence of the structure of a β-optimal policy, which we now investigate.

The next lemma establishes the obvious result that if an arm is pulled often enough its Gittins index will converge to whatever the (fixed) value of β.

LEMMA 4.9. *Let $(x(t), t = 0, 1, \cdots)$ be the sequence of states obtained if a Bernoulli bandit process is continued at times $t = 0, 1, \cdots$. Then with probability one*

$$\nu_\beta(x(t)) \to \theta \qquad as \qquad t \to \infty,$$

where θ is the parameter of the Bernoulli reward sequence generated by the bandit process.

PROOF. A lower bound on the Gittins index is obtained using the stopping time $\tau \equiv 1$ in definition (2); this gives

$$\nu_\beta(G_t) \geq \int_0^1 z \, dG_t(z) = g_t, \tag{16}$$

say, where we identify the state $x(t)$ with a distribution G_t. An upper bound is provided by the informative arm used in the proof of Theorem 3.1: from equation (4)

$$\lambda_\beta(G_t) \geq \nu_\beta(G_t), \tag{17}$$

where $\lambda = \lambda_\beta(G_t)$ satisfies

$$(1-\beta)\left[\lambda - \int_0^1 z \, dG_t(z)\right] = \beta \int_\lambda^1 (z-\lambda) \, dG_t(z).$$

From inequalities (16) and (17)

$$g_t \leq \lambda$$

and hence

$$(1-\beta)(\lambda - g_t) \leq \beta \int_{g_t}^1 (z - g_t) \, dG_t(z). \tag{18}$$

Now with probability one G_t converges weakly to a degenerate distribution with unit mass concentrated at θ as $t \to \infty$. Thus

$$g_t \to \theta \qquad \text{a.s.}$$

and from inequality (18)

$$\lambda \to \theta \qquad \text{a.s.}$$

The desired result then follows from the bounds (16) and (17). □

REMARK 4.10. Consider again a multi-armed bandit in which each of the n constituent bandit processes is a Bernoulli bandit process with initial distribution F, and assume till the end of this remark that F has no atoms. This assumption ensures that $\theta_1, \theta_2, \cdots, \theta_n$ are distinct (with probability one, a phrase which will be omitted henceforth). It then follows from Lemma 4.9 that a β-optimal policy pulls just one arm infinitely often. This conclusion also follows from the important work of Rothschild (1974), at least when F has a continuous density (Rothschild's results also deal with the case where the arms may be dependent). Let

$$T_\beta^2 = \inf\{T : \exists I \text{ such that } \pi_\beta(I \mid \mathbf{x}(t)) = 1 \qquad \forall t \geq T\}.$$

Then $T_\beta^2 < \infty$, and, from time $t = T_\beta^2$ onwards, the β-optimal policy π_β pulls just one arm, say arm I. Let

$$P_\beta = P(\theta_I = \max(\theta_1, \theta_2, \cdots, \theta_n) \mid \theta_1, \theta_2, \cdots, \theta_n)$$

and let

$$\mathbf{x}(T_\beta^2) = (x_1(t_1), x_2(t_2), \cdots, x_n(t_n)),$$

so that t_i is the total number of times the policy π_β pulls arm i, for $i \neq I$. Assume now that $F^*(z)$ varies regularly, and consider the effect of letting β tend to one. From Remark 4.6 it follows that $T_\beta^1 \to \infty$, and hence $t_i \to \infty$. Thus, from Lemma 4.9 it follows that $P_\beta \to 1$. Now if $(R^\beta(t), t = 0, 1, \cdots)$ is the sequence of rewards obtained using the policy π_β

$$\lim_{T\to\infty} E\left[\frac{1}{T}\sum_{t=0}^{T-1} R^\beta(t) \,\middle|\, \theta_1, \theta_2, \cdots, \theta_n\right] \geq P_\beta \max(\theta_1, \theta_2, \cdots, \theta_n),$$

and so as β approaches one the performance of π_β as judged by the expected average reward criterion approaches the best possible.

The β-optimal policy π_β can be regarded as having three regimes. Between times $t = 0$ and $t = T_\beta^1$ it corresponds to the least failures rule. Between times $t = T_\beta^1$ and $t = T_\beta^2$ it moves further and further away from the least failures rule, and from time $t = T_\beta^2$ onwards it pulls just one arm, judged in some sense to be the best. Observe that although the end of the first regime T_β^1 is a stopping time for the Markov process $\{\mathbf{x}(t), t = 0, 1, \cdots\}$ generated by the multi-armed bandit under the policy π_β, the end of the second regime T_β^2 is not. Indeed at no time $t = T$ is it possible to determine from $\{\mathbf{x}(t), t = 0, 1, \cdots, T\}$ whether or not the third regime has started. This analysis makes clear that the least failures rule should not be expected to perform well under either a discounted or expected average reward criterion, since it corresponds to just the first 'information gathering' regime—the third, and most rewarding, regime is lost under the limiting process of allowing β to tend to one.

REMARK 4.11. Assume now that $F^*(z)$ varies regularly but allow the distribution $F(z)$ to have atoms. There is thus a positive probability that two or more of the parameters θ_1, $\theta_2, \cdots, \theta_n$ are equal. If $\theta_i \neq \theta_j$ then there is probability zero that the β-optimal policy will pull both arms i and j infinitely often. In the event that $\theta_i = \theta_j$ the β-optimal policy may pull both arms i and j infinitely often either with positive probability or with zero probability—examples can be constructed to illustrate each of these two possibilities. Thus the third regime of the policy may take a more complicated form, with a subset of the arms each being pulled infinitely often. Nevertheless it is still straightforward to show that as β approaches one the performance of π_β as judged by the expected average reward criterion approaches the best possible. Hence we have the following result.

THEOREM 4.12. *If $F^*(z)$ varies regularly then*

$$\lim_{\beta\to 1} \lim_{T\to\infty} E\left[\frac{1}{T}\sum_{t=0}^{T-1} R^\beta(t) \,\middle|\, \theta_1, \theta_2, \cdots, \theta_n\right] = \max(\theta_1, \theta_2, \cdots, \theta_n).$$

REMARK 4.13. Robbins (1952) proposed another rule for the two-armed bandit, a slight adaptation of which can be described as follows. Choose an increasing sequence of integers $0 = a_0, a_1, \cdots$ such that

$$a_{r+1} - a_r \geq n \qquad r = 0, 1, \cdots$$

and

$$\lim_{r\to\infty}\left(\frac{a_r}{r}\right) = \infty$$

At stages $t = a_r, a_r + 1, \cdots, a_r + n - 1$ pull arms $1, 2, \cdots, n$ respectively, for $r = 0, 1, \cdots$. At any other stage pull an arm for which the ratio of successes to trials is maximal. If $(R''(t), t = 0, 1, \cdots)$ is the sequence of rewards obtained from the multi-armed bandit using this policy then, following Robbins (1952),

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} R''(t) = \max(\theta_1, \theta_2, \cdots, \theta_n) \qquad \text{a.s.}$$

from the strong law of large numbers applied to each of the sequences $(R_i(t), t = 0, 1, \cdots)$, $1 \leq i \leq n$, and hence

$$\lim_{T\to\infty} E\left[\frac{1}{T}\sum_{t=0}^{T-1} R''(t) \,\middle|\, \theta_1, \theta_2, \cdots, \theta_n\right] = \max(\theta_1, \theta_2, \cdots, \theta_n).$$

Thus this rule performs optimally according to the expected average reward criterion.

REMARK 4.14. Throughout this paper it has been assumed that $F(z) < 1$ for all $z \in (0, 1)$, and many of the results depend upon the additional condition that $F^*(z)$ varies regularly. If

$$\sup\{z : F(z) < 1\} = \theta^* < 1$$

it is still possible to obtain bounds on $\nu_\beta(F)$ similar to those described in Section 3 but the least failures rule is no longer the limit rule for the multi-armed bandit. This follows from Theorem 3.1 of Berry and Fristedt (1979) since if

$$\frac{s}{s+f} > \theta^*$$

then $\sigma^s\phi^f F$ is strongly to the right of F and so

$$\nu_\beta(\sigma^s\phi^f F) \geq \nu_\beta(F) \qquad \forall \beta \in (0, 1).$$

In contrast, it is readily shown that the conclusion of Theorem 4.12 holds with no condition on the distribution function F.

REMARK 4.15. In his work on the finite horizon two-armed bandit Berry (1972) includes a conjecture which can be phrased as follows: if the number of failures differs between the two arms then for a distant enough horizon the optimal policy will choose the arm with the smaller number of failures. Despite the close relationship between Cesàro and Abel summability there seems to be no immediate link between Theorem 4.5 and Berry's conjecture. The observation contained in Remark 4.14 suggests however that some restriction on the function F may be necessary for Berry's conjecture to hold.

Acknowledgments. It is a pleasure to acknowledge the helpful conversations I have had with John Gittins, Andrew Barbour and Peter Nash.

REFERENCES

BELLMAN, R. (1956). A problem in the sequential design of experiments. *Sankhyā* **16** 221–229.
BERRY, D. A. (1972). A Bernoulli two-armed bandit. *Ann. Math. Statist.* **43** 871–897.
BERRY, D. A. and FRISTEDT, B. (1979). Bernoulli one-armed bandits—arbitrary discount sequences. *Ann. Statist.* **7** 1086–1105.
BLACKWELL, D. (1962). Discrete dynamic programming. *Ann. Math. Statist.* **33** 719–726.
BLACKWELL, D. (1965). Discounted dynamic programming. *Ann. Math. Statist.* **36** 226–235.
FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications* **2**, 2nd ed. Wiley, New York.
GITTINS, J. C. (1979). Bandit processes and dynamic allocation indices (with discussion). *J. Roy. Statist. Soc. Ser. B* **41** 148–177.
HINDERER, K. (1970). *Foundations of Non-stationary Dynamic Programming with Discrete Time Parameter.* Springer, Berlin.
KELLY, F. P. (1979). In discussion of Gittins (1979).
NASH, P. (1973). Optimal allocation of resources between research projects. Ph.D. thesis, Cambridge University.
ROBBINS, H. (1952). Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* **58** 527–535.
ROSS, S. M. (1968). Non-discounted denumerable Markovian decision models. *Ann. Math. Statist.* **39** 412–423.
ROTHSCHILD, M. (1974). A two-armed bandit theory of market pricing. *J. Econom. Theory* **9** 185–202.
SOBEL, M. and WEISS, G. H. (1972). Play-the-winner rule and inverse sampling for selecting the best of $k \geq 3$ binomial populations. *Ann. Math. Statist.* **43** 1808–1826.
STRAUCH, R. E. (1966). Negative dynamic programming. *Ann. Math. Statist.* **37** 871–890.
VEINOTT, JR., A. F. (1974). Markov decision chains. In *Studies of Mathematics 10 Studies in Optimization.* (G. B. Dantzig and B. C. Evans, eds). Mathematical Association of America.
ZELEN, M. (1969). Play-the-winner rule and the controlled clinical trial. *J. Amer. Statist. Assoc.* **64** 131–146.

STATISTICAL LABORATORY
UNIVERSITY OF CAMBRIDGE
16 MILL LANE
CAMBRIDGE, CB2 1SB
ENGLAND