The Annals of Probability
1981, Vol. 9, No. 2, 284–292

ARM-ACQUIRING BANDITS

BY P. WHITTLE

University of Cambridge

We consider the problem of allocating effort between projects at different stages of development when new projects are also continually appearing. An expression (14) is derived for the expected reward yielded by the Gittins index policy. This is shown to satisfy the dynamic programming equation for the problem, so confirming optimality of the policy.

1. Introduction. The classic multiarmed bandit problem can be posed in specialised or varied forms; we shall understand it as follows. At each stage in time one has the option of working on exactly one of N projects. To do so causes the "state" of that project to change, in a Markov fashion; the states of remaining projects do not change. One receives a reward depending on the project and its state. Rewards are discounted and added over time. One wishes to deduce an optimal policy; i.e., a sequential rule for determining which project to work on at each instant of time in order to maximise the expected total discounted reward.

In the gambling-machine version of the problem the projects are the "arms" of the machine. However, the problem describes a range of much more significant practical situations; e.g., sequential selection trials in medicine and agriculture, the evaluation and programming of evolving projects generally. It is also important because it embodies a universal problem: the conflict between taking those actions which yield immediate reward, and those (such as acquiring information, or preparing the ground) whose benefits manifest themselves only later.

In medical, agricultural and technological applications one can expect that new projects will be added as time goes on, as new compounds, technical possibilities, etc., become available for investigation. This variant of the problem has been considered by Nash (1973), and is the subject of this paper. In seeking a name for it one realises the inappropriateness of "multiarmed bandit" as a technical term. "Arm-acquiring bandit" maintains the analogy, although a term such as "open sequential allocation" would be preferable. We shall certainly find it convenient to refer to the cases of fixed and increasing N as the "closed" and "open" cases respectively.

At this point one should enter a disclaimer. A rational approach to project selection cannot allow in any real sense for the possibility of fundamental scientific or technological advance. This it could do no more than a rational approach to hypothesis testing could allow for the possibility that additional hypotheses might be added in the course of time, of natures not even formulable initially, because they lie beyond the investigator's initial conception and insight. So, we shall regard the "new projects" as being very much variants as the old ones, occurring in a statistically homogeneous stream. For example, one might think of an industrial chemist routinely testing the efficacy of a large number of compounds as adhesives, or of an agriculturalist routinely testing a large number of wheat varieties. Such "research" is exploratory rather than innovative. But, by its nature, creative research cannot be formalised.

The multiarmed bandit problem is classic because of its difficulty as well as its importance. The most fundamental contribution has been made by Gittins and his coauthors (see especially Gittins and Jones (1974), Gittins and Glazebrook (1977) and Gittins (1979)). Gittins shows that to each project can be attached an index ν, which is a

Received August 27, 1979.
AMS 1970 *subject classifications.* 42C99, 62C99.
Key words and phrases. Multiarmed bandit, dynamic programming, allocation index.

function only of the project and its current state, such that the optimal policy is to work on a project whose index is currently greatest. We shall refer to ν as the "Gittins index"; it obviously supplies a formal project evaluation. Gittins shows that it is determined by solving the problem of choosing between the project in question and a "standard project" (of fixed state and reward) and so effectively reduces the case of general N to that of $N = 2$ (or $N = 1\frac{1}{2}$, one might say, since one of the projects is standard).

Gittins' contribution has not been appreciated at its true value, possibly largely because the proofs he and his coauthors give are difficult to follow. However, such a fundamental result must have a natural derivation. In a recent paper (Whittle (1980)) I gave what I believe to be such a derivation, by a constructive argument which leads to an explicit expression for the maximal expected reward for the multiproject case in terms of those for the single (versus standard) project case.

The present paper extends this result to the open case. This is a nontrivial matter, because a one-project situation does not remain so, and one is forced to employ rather different methods.

Nash used Hamiltonian and dynamic programming methods in his 1973 work, and did not obtain solutions of the relatively explicit form derived here. Workers other than those of the Gittins school (see, e.g., Berry (1972), Wahrenberger, et al. (1977), Rodman (1978) and Berry and Fristedt (1979)), have tended to exploit special properties of special versions of the problem.

2. Formulation. A project is usually described by two variables: a label for the project itself, and a "state" variable, indicating the stage of development of the project. For an open process we shall have indefinitely many projects, so it is better to use a single project state variable x, which indicates both the type of the project and its stage. For example, for the industrial chemist, "type" might label particular classes of compound being investigated. If a project cannot change type then "type" labels ergodic classes, and "stage" labels state within a class. For simplicity of exposition we shall assume that the set of values X which a project may adopt is finite, although this assumption could almost certainly be relaxed.

Let $n_t(x)$ be the number of projects in state x at time t. We shall set the problem in integral time, so that t takes only integral values. We shall regard $n_t = \{n_t(x); x \in X\}$ as the state of the decision process at time t, and shall denote a generic value of this state variable by $n = \{n(x)\}$. By $n > 0$ we shall understand that $n(x) \geqslant 0$ for all x, with strict inequality for some x.

Let $e(u) = \{e(x, u)\}$ denote the value of the vector n which is zero except for a unit in the uth place:

$$e(x, u) = \begin{cases} 1, & x = u \\ 0, & x \neq u. \end{cases}$$

Suppose that at time t one works on a project of state u (so that necessarily $n_t \geqslant e(u)$). Then we shall suppose that n_t undergoes the transition

$$n_{t+1} = n_t - e(u) + e(u') + W_{t+1} \tag{1}$$

where $W_{t+1}(x)$ is the number of new projects of state x which enter the process at time $t + 1$, and u' is the state which the project (state u) engaged at time t acquires at $t + 1$. Implicit in relation (1) is the statement that no other project changes state. The quantities W_{t+1} and u' are random variables. We shall suppose that the process is stochastically time-homogeneous; that, conditional on u, the new state u' is independent of $\{n_s, W_s; s \leq t\}$ and of allocation decisions before time t; and that W_{t+1} is independent of these variables and also of u, u'.

We shall suppose that if one works on a project in state u at time t then one receives an expected reward $R(u)$, which is uniformly bounded

$$|R(u)| \leq K(1 - \beta) < \infty.$$

Here β is the discount factor, taking a value in [0, 1). Again, the assumption of boundedness is convenient, but probably not necessary. We shall allow the option of inactivity at any stage (i.e., of pursuing no project at all) so that a standard project of zero reward is effectively always available. Being always available, it need not be listed in n.

Suppose that at time t one operates a project of state u and then terminates at $t + 1$ with reward $\psi(n_{t+1})$. The expected total reward conditional on $n_t = n$ would then be

$$L_u\psi(n) = R(u) + \beta E[\psi(n - e(u) + e(u') + W) \mid u], \tag{2}$$

where the expectation $E[\cdot \mid u]$ is over u' and W, but the u-conditioning of course affects only u'. In the actual process operations will continue indefinitely, but the one-stage operator L_u plays an essential role in the later discussion.

We are now interested in maximising the total discounted expected reward $E\sum_{r=0}^{\infty}\beta^t R_t$ from time $t = 0$, where we have used R_t to denote the reward received at time t. If the expectation is made conditional on process history up to time $t = 0$ then, in virtue of the discounted Markov character of the decision process, the maximal expected reward will be a function of initial state alone, $\Phi(n_0)$, say, and will be the unique bounded solution of the dynamic programming equation

$$\Phi = \sup_{u\in U(n)} L_u\Phi \tag{3}$$

(see Blackwell (1965); Bertsekas (1976) page 229). Here $U(n)$ is the set of u for which $n(u) > 0$. We shall refer to $\Phi(n)$ simply as the "reward function".

3. The retirement option. We now modify the process by assuming that one has the additional option of retiring at any time with reward M. Let us refer to process thus modified as the "M-process", and to the unmodified process as the "continuing process". If $F(n, M)$ is the maximal expected reward for this modified process then it satisfies the modified dynamic programming equation

$$F = \max[M, \sup_{u\in U(n)} L_uF]. \tag{4}$$

LEMMA 1. *$F(n, M)$ is nondecreasing function of M and n for which*

$$F(n, M) = \begin{cases} M, & M \geqslant K \\ \Phi(n), & M \leqslant \Phi(0) \end{cases}$$

and the optimal policies for the continuing process and the M-process are identical for $M < \Phi(0)$.

PROOF. Increase in M plainly cannot decrease F. Increase in n increases the range of options (because inactivity is permissible) and so also cannot decrease F. Since in the continuing process one has the option of inactivity, one also has the option of effectively discarding all projects currently available. To take this option is to settle for an expected reward of $\Phi(0)$, so it is as if one retired with a reward of $\Phi(0)$. For the M-process this option will be equally as attractive as retirement if $M = \Phi(0)$, and retirement will be an option never exercised if $M < \Phi(0)$. On the other hand, it will be an option as attractive as any if $M \geqslant K$, for to continue for s steps and then retire cannot yield a reward exceeding $(1 - \beta^s)K + \beta^s M \leq M$. The assertions of the lemma then follow. □

4. Write-off policies. Consider a policy in which a project is considered written-off (i.e., permanently abandoned) as soon as its state enters a *write-off set* $\mathscr{G}$, a subset of X. The policy is further such that one never uses a written-off project, one continues as long as there are projects not written-off, and retires when all projects are written-off. "Inactivity" is a project of constant state value, which may or may not belong to $\mathscr{G}$.

We shall term such a policy a *write-off policy*. Note that there is no prescription of the order in which projects are operated before retirement, and the policy need be neither Markov nor stationary, except insofar as $\mathscr{G}$ is held fixed.

Let H_0 be the observational history at time $t = 0$, which includes knowledge of n_0. We shall assume that n_0 takes the value n; a generic value of the vector of project numbers. The expected reward $E[\sum_0^\infty \beta^t R_t \mid H_0]$ for a given policy will then be a function $V(H_0)$ of H_0.

LEMMA 2. *For a prescribed policy, independent of M,*

$$\frac{\partial V}{\partial M} = E(\beta^T \mid H_0) \tag{5}$$

where T is the moment of retirement. For a prescribed write-off policy

$$\frac{\partial V}{\partial M} = E(\beta^T \mid n). \tag{6}$$

PROOF. Relation (5) follows from the fact that

$$V = V_c + ME(\beta^T \mid H_0) \tag{7}$$

where V_c is the expected reward before retirement, independent of M. In the case of a write-off policy, the quantity T is the time needed to bring all projects currently available (i.e., those initially available plus those which have entered the process) into $\mathscr{G}$. History before $t = 0$ may affect the order in which these projects are operated, but cannot affect the value of T, so (5) reduces to (6). □

Note that there is no presumption that $T < \infty$; the event of nonretirement can be identified with the event $T = +\infty$ because, since $|\beta| < 1$, neither contingency will contribute to $E(\beta^T)$.

Although V is in general a function of H_0, in the case of a write-off policy we shall sometimes write $\partial V(H_0)/\partial M$ as $\partial V(n)/\partial M$, since the derivative does indeed depend on H_0 only through $n_0 = n$.

Now let σ_t be the work-load in the system at time t, i.e., the time needed to take all projects currently in the system to $\mathscr{G}$. The quantity σ_t is a random variable, determined by events after time t. However, since projects do not interact, we can regard each project in the system as carrying a sealed label with a random "time needed for completion" noted on it. The quantity σ_t can then be regarded as defined at t, even if not then observable. It is a nonnegative integer and obeys the recursion

$$\sigma_{t+1} = \sigma_t - 1 + w_{t+1}, \qquad \sigma_t > 0 \tag{8}$$

where w_{t+1} is the additional work-load brought by the bundle W_{t+1} of projects entering the system at time $t + 1$. By the assumptions on W of Section 2, w_{t+1} is independent of $\{\sigma_s, w_s; s \leq t\}$ and has a distribution independent of t. Provided there is work in the system (i.e., $\sigma_t > 0$) then one unit of load will be worked off in the passage from t to $t + 1$, hence the -1 in (8). The termination time T is the smallest nonnegative value of t for which $\sigma_t = 0$.

LEMMA 3. *For a write-off policy*

$$\frac{\partial V}{\partial M} = E(\beta^T \mid n) = E(\zeta^{\sigma_0}/n) \tag{9}$$

where ζ is the smaller real root of

$$\zeta = \beta A(\zeta) \tag{10}$$

and $A(Z)$ is the probability generating function of w.

PROOF. We can regard the σ-process as a random walk on the integers with an increment whose probability generating function is $A(Z)/Z$. The state $\sigma = 0$ is absorbing,

and T is the absorption time. It follows then by standard arguments that

$$E(\beta^T \mid \sigma_0) = \zeta^{\sigma_0}. \tag{11}$$

(An appeal to Wald's identity suffices, or see formula (67) on page 53 of Cox and Miller (1969) which makes the domain of validity of Wald's identity quite plain.) The result (11) is exact (because $\sigma_{t+1} - \sigma_t \geqslant -1$) and again holds even if $T = +\infty$ with positive probability (because $|\beta| < 1$).

The second equation of (9) now follows from (11). □

THEOREM 1. *For a write-off policy*

$$\frac{\partial V(n)}{\partial M} = \prod_x \left[\frac{\partial V(e(x))}{\partial M}\right]^{n(x)} \tag{12}$$

PROOF. Because the work-loads generated by distinct projects are independent, then

$$E(\zeta^{\sigma_0}/n) = \prod_x [E(\zeta^{\sigma_0} \mid e(x))]^{n(x)} \tag{13}$$

Relation (12) then follows from (9), (13). □

Theorem 1 states the key property of a write-off policy. It turns out to be shared by the optimal policy.

5. The reward function for the Gittins index policy. The Gittins index policy is a Markov policy, whose expected reward for the M-process we shall denote by $V(n, M)$. It is characterised by the fact that at any stage one works on an available project of largest index, provided this exceeds M, where the index $M(x)$ of a project in state x is defined as the infimal value of m for which $V(e(x), m) = m$. If no available project has index exceeding M, then one retires. The policy is then a write-off policy, with $\mathscr{G} = \{x : M(x) \leqslant M\}$. It is to be noted that $\mathscr{G}$ depends upon M; let us denote it $\mathscr{G}_M$. Let us also denote $V(e(x), M)$ by $\phi(x, M)$.

THEOREM 2(i). *The reward $V(n, M)$ for the Gittins index policy is a nondecreasing convex piece-wise linear function of M.*

(ii) *$V(n, M)$ can be expressed in terms of the one-project rewards $\phi(x, M)$ by*

$$V(n, M) = K - \int_M^K \prod_x \left[\frac{\partial \phi(x, m)}{\partial m}\right]^{n(x)} dm. \tag{14}$$

(iii) *The one-project rewards are the unique solutions of the recursions*

$$\phi(u, M) = \max\left\{M, R(u) + E[\phi(u', M) \mid u]\zeta(M) + \int_M^\infty E[\phi(u', m) \mid u]\, d\zeta(m)\right\} \tag{15}$$

satisfying $\phi(u, M) = M$ $(M \geqslant K)$, where

$$\zeta(M) = \beta E\left\{\prod_x \left[\frac{\partial \phi(x, M)}{\partial M}\right]^{W(x)}\right\}. \tag{16}$$

PROOF. We shall prove the assertions recursively for decreasing M. All assertions hold trivially for $M \geqslant K$, when $V(n, M) = M$. Suppose they hold for $M \geqslant \mu$.

Now, although the index policy is a write-off policy, we cannot necessarily deduce from (12) that

$$\frac{\partial V(n, M)}{\partial M} = \prod_x \left[\frac{\partial \phi(x, M)}{\partial M}\right]^{n(x)}, \tag{17}$$

because (12) was deduced under the assumption that the write-off set $\mathcal{G}$ was independent of M, whereas now we have an M-dependent write-off set $\mathcal{G}_M$. However, if $\mathcal{G}_M$ does not change for M in some interval then $V(n, M)$ will vary linearly with M in that interval, and (17) will hold in that interval (with the derivatives interpreted as directional derivatives into the interior at the ends of the interval).

If $\mathcal{G}_M$ does not change as M decreases from μ then (14) will continue to hold by integration of (17). We require that the one-project rewards continue to satisfy

$$\phi(u, M) = \max[M, L_u\phi(u, M)] \tag{18}$$

where L_u is the operator defined in (2). Partial integration of expression (14) shows that

$$V(W + e(u'), M) = \phi(u', M)\pi(W, M) + \int_M^\infty \phi(u', m)\, d_m\pi(W, m) \tag{19}$$

where

$$\pi(W, M) = \prod_x \left[\frac{\partial\phi(x, M)}{\partial M}\right]^{W(x)},$$

a nondecreasing function of M which is certainly constant for $M > K$. From (18), (19) we deduce (15).

Note that, since $\partial\phi(x, M)/\partial M = 1$ for x in $\mathcal{G}_M$, then it follows from (14) that

$$V(n + r, M) = V(n, M) \tag{20}$$

if $r(x) > 0$ only for x in $\mathcal{G}_M$. That is, addition of written-off projects does not change V.

All assertions of the theorem continue to hold, as M is decreased from μ, as long as the maximising option in (18) does not change for any u, and so $\mathcal{G}_M$ does not change. However, at some point the maximising option will change, and $\mathcal{G}_M$ will change, and we must establish that the assertions of the theorem continue to hold as M decreases further, despite the discontinuous change in policy.

Suppose that at the current value of M we have $\phi(u, M) = L_u\phi(u, M) > M$, so that $u \notin \mathcal{G}_M$. Since $\partial\phi/\partial M \leqslant 1$ then the inequality $\phi > M$ continues to hold as M decreases, and the maximising option in (18) does not change.

Suppose that at the current value of M we have $\phi(u, M) = M \geqslant L_u\phi(u, M)$, so that $u \in \mathcal{G}_M$. As M decreases past some value (which is just $M(u)$) it may be that this inequality is violated, so that $M < L_u\phi_\mu(u, M)$, where $\phi_\mu(u, M)$ is the value of $\phi(u, M)$ calculated on the hitherto constant evaluation $\mathcal{G}_\mu$ of $\mathcal{G}_M$. By relation (20) this implies also that $M < L_u V_\mu(n, M)$ if $n \geqslant e(u)$, and $n(x) > 0$ only for x in $\mathcal{G}_\mu$. That is, that the policy hitherto employed would be improved if, when left with projects with states in $\mathcal{G}_\mu$ one operated a project of state u, if available. (The inequality implies that to employ this procedure for one step and then revert to the previous policy constitutes an improvement. It follows then from the Howard improvement theorem (see Blackwell (1962) page 720) that indefinite application of this modification will provide an improvement).

For the modified policy at $M = M(u)$ the write-off set $\mathcal{G}_M$ has decreased from $\mathcal{G}_\mu$ by deletion of state u. Note, however, that projects of state u will be used only when these are the only unabandoned projects available. That is, when no projects of greater index are available. The policy thus recursively constructed is just the Gittins index policy.

It may be that several states are deleted simultaneously from the write-off set in this way. However, the effect is always that $\mathcal{G}_M$ decreases as M decreases.

As $\mathcal{G}_M$ decreases, so the time T required to bring all projects to $\mathcal{G}_M$ increases, and the value of $E(\beta^T/n) = \partial V(n, M)/\partial M$ appropriate to the smaller $\mathcal{G}_M$ is smaller: That is, $V(n, M)$ is indeed convex in M, and linear in those intervals of M for which $\mathcal{G}_M$ is constant.

The relations of the theorem continue to hold as M decreases through the value $M(u)$ at which the policy changes discontinuously, because the manner of change in policy (with

u-projects given lowest priority of the unabandoned projects) means that there is no discontinuity in expected reward. One can thus continue to integrate (17) to deduce (14), (15).

The course of the proof (inductive in M for decreasing M) shows that the $\phi(u, M)$ are indeed uniquely determined by relations (15), (16) and the condition $\phi(u, M) = M$ ($M \geqslant K$): $V(n, m)$ is then uniquely determined from (14). □

The use of $\zeta(M)$ to denote expression (16) is consistent with the notation introduced in Lemma 3. We have

$$\frac{\partial \phi(x, M)}{\partial M} = E(\zeta^{T(x)})$$

where $T(x)$ is the work-load generated by a project in state x (i.e., the process time needed to bring the state of that project to $\mathscr{G}_M$), and ζ is the smaller solution of (10). The right-hand member of (16) is then $\beta A(\zeta)$, and so equal to ζ, by (10).

At the points $M(x)$ of discontinuity of policy, where $\partial V/\partial M$ is undefined, we shall henceforth make the convention of identifying its value with the right-derivative. This is consistent with the convention that, if $L_x\phi(x, M) = M$, then we assign x to $\mathscr{G}_M$.

6. Optimality of the Gittins index policy.

LEMMA 4. *Suppose that* $n \geqslant e(u)$. *Define* $\Delta_u(n, M) = V(n, M) - L_u V(n, M)$ *and* $\delta_u(M) = \Delta_u(e(u), M) = \phi(u, M) - L_u\phi(u, M)$. *Then*

$$\Delta_u(n, M) = \delta_u(M)P_u(n, M) + \int_M^\infty \delta_u(m)\, d_m P_u(n, m) \tag{21}$$

where

$$P_u(n, M) \triangleq \prod_x \left[\frac{\partial \phi(x, M)}{\partial M}\right]^{n(x)-e(x,u)} \tag{22}$$

can be regarded as a right-continuous distribution function in M, *with all its mass in* $[\Phi(0), K)$.

PROOF. The function P_u plainly has the properties asserted, being nonnegative, non-decreasing, and equal to unity for $M \geqslant K$. One finds from (17) that

$$\frac{\partial \Delta_u}{\partial M} = P_u \frac{\partial \delta_u}{\partial M}$$

whence (21) follows by integration, if one recalls that $\Delta_u = \delta_u = (1 - \beta)M - R(u)$ and $P_u = 1$ for $M \geqslant K$. □

THEOREM 3. *The Gittins index policy is optimal.*

PROOF. If we can show that $V(n, M)$ satisfies the dynamic programming equation (4) then this will imply that $V = F$, and that the Gittins index policy is optimal.

Let us define

$$\mu(n) = \sup_{U(n)} M(u).$$

It follows then from (14) that

$$V(n, M) \geqslant M \tag{23}$$

with equality for $M \geqslant \mu(n)$.

We know that $\delta_u(M) \geqslant 0$ with equality for $M \leqslant M(u)$, and that $0 \leqslant P_u(n, M) \leqslant 1$ with

equality in the second inequality for $M \geqslant \mu(n - e(u))$. It then follows from (21) that $\Delta_u \geqslant 0$, i.e.,

$$V(n, M) \geqslant L_u V(n, M) \qquad u \in U(n) \tag{24}$$

with equality if $M(u) \geqslant \mu(n - e(u))$ and $M(u) \geqslant M$, i.e., if $M(u) = \mu(n) \geqslant M$.

From inequalities (23), (24) and the cases of equality characterised after them, we see that V satisfies (4). □

7. An example. The index result constitutes a powerful reduction of the problem. However, one still has the task of calculating the index function $M(x)$. Even in the closed case, for which a one-project problem remains a one-project problem, and relations (15) simplify considerably, there are as yet few explicit solutions or analytic results. In the proof of Theorem 2 we indicate a natural approach for numerical solution, but the problem of analytic solution remains.

So, while it would be satisfying to give a substantial example, we must content ourselves for the moment with a case for which a one-step look-ahead rule is fairly obviously optimal. This is what Gittins (1979) refers to as the *deteriorating case*, for which, if one works on a project of state $x(t)$ at time t, then $P(x(t + 1) = x' \mid x(t) = x) > 0$ implies that $R(x') < R(x)$.

THEOREM 4. *For the deteriorating case an optimal policy is to work on an available project u for which $R(u)$ is greatest.*

PROOF. If $M = M(x)$, so that $x \in \mathscr{G}_M$, then also $x' \in \mathscr{G}_M$, and $F(W + e(x'), M) = F(W, M)$, by (20). The relations $M(x) = M = \phi(x, M) = L_x\phi(x, M)$ thus imply that

$$\begin{aligned} M &= R(x) + \beta EF(W, M(x)) \\ &= R(x) + \theta(M(x)) \end{aligned}$$

say. Now the function $\theta(M)$ is nondecreasing, convex, and is equal to βM for $M \geqslant K$, so that

$$H(M) = M - \theta(M)$$

increases, strictly and continuously, from $-\infty$ to $+\infty$ with M. Hence

$$M(x) = H^{-1}(R(x))$$

is a strictly increasing function of $R(x)$, whence the assertion follows. □

REFERENCES

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.

BERTSEKAS, D. P. (1976). *Dynamic Programming and Stochastic Control.* Academic.

BLACKWELL, D. (1962). Discrete dynamic programming. *Ann. Math. Statist.* **33** 719–726.

BLACKWELL, D. (1965). Discounted dynamic programming. *Ann. Math. Statist.* **36** 226–235.

COX, D. R. and MILLER, H. D. (1965). *The Theory of Stochastic Processes.* Methuen.

FELDMAN, D. (1962). Contributions to the two-armed bandit problem. *Ann. Math. Statist* **33** 847–856.

GITTINS, J. C. (1975). The two-armed bandit problem: variations on a conjecture by H. Chernoff. *Sankhyā Ser. A* **37** 287–291.

GITTINS, J. C. (1979). Bandit processes and dynamic allocation indices. *J. Roy. Statist. Soc. Ser. B* **41** 148–164.

GITTINS, J. C. and GLAZEBROOK, K. D. (1977). On Bayesian models in stochastic scheduling. *J. Appl. Probability* **14** 556–565.

GITTINS, J. C. and JONES, D. M. (1974). A dynamic allocation index for the sequential design of experiments. In *Progress in Statistics.* (J. Gani, ed.) 241–266. North Holland, Amsterdam.

GITTINS, J. C. and NASH, P. (1977). Scheduling, queues, and dynamic allocation indices. In *Proc. EMS, Prague.* 191–202. Czechoslovak Academy of Sciences, Prague.

GLAZEBROOK, K. D. (1976a). A profitability index for alternative research projects. *Omega* **4** 79–83.
GLAZEBROOK, K. D. (1976b). Stochastic scheduling with order constraints. *Int. J. Sys. Sci.* **7** 657–666.
GLAZEBROOK, K. D. (1978a). On a class of non-Markov decision processes. *J. Appl. Probability* **15** 689–698.
GLAZEBROOK, K. D. (1978b). On the optimal allocation of two or more treatments in a controlled clinical trial. *Biometrika* **65** 335–340.
NASH, P. (1973). Optimal allocation of resources between research projects. Ph.D. thesis, Cambridge Univ.
NASH, P. and GITTINS, J. C. (1977). A Hamiltonian approach to optimal stochastic resource allocation. *Adv. Appl. Probability* **9** 55–68.
RODMAN, L. (1978). On the many-armed bandit problem. *Ann. Probability* **6** 491–498.
WAHRENBERGER, D. L., ANTLE, C. E. and KLIMKO, L. A. (1977). Bayesian rules for the two-armed bandit problem. *Biometrika* **64** 172–174.
WHITTLE, P. (1980). Multi-armed bandits and the Gittins index. *J. Roy. Statist. Soc., Ser. B* **42** 143–149.

UNIVERSITY OF CAMBRIDGE
DEPARTMENT OF PURE MATHEMATICS
AND MATHEMATICAL STATISTICS
STATISTICAL LABORATORY
16 MILL LANE
CAMBRIDGE CB2 1SB
ENGLAND