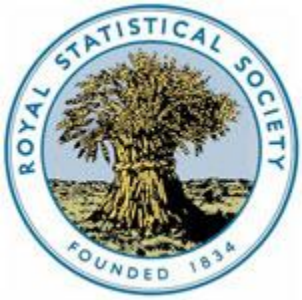

Multi-Armed Bandits and the Gittins Index
Author(s): P. Whittle
Source: *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 42, No. 2 (1980), pp. 143-149
Published by: Wiley for the Royal Statistical Society
Stable URL: http://www.jstor.org/stable/2984953
Accessed: 09-02-2018 21:44 UTC

J. R. Statist. Soc. B (1980),
42, *No.* 2, *pp.* 143–149

Multi-armed Bandits and the Gittins Index

By P. WHITTLE

Statistical Laboratory, University of Cambridge

[Received August 1979. Final revision December 1979]

SUMMARY

A plausible conjecture (*C*) has the implication that a relationship (12) holds between the maximal expected rewards for a multi-project process and for a one-project process (*F* and ϕ_i respectively), if the option of retirement with reward *M* is available. The validity of this relation and optimality of Gittins' index rule are verified simultaneously by dynamic programming methods. These results are partially extended to the case of so-called "bandit superprocesses".

Keywords: BANDIT PROCESSES; DYNAMIC ALLOCATION INDICES; TWO-ARMED BANDIT PROBLEM; OPTIMAL RESOURCE ALLOCATION

1. INTRODUCTION

THE multi-armed bandit problem (as it has become known) is important as one of the simplest non-trivial problems in which one must face the conflict between taking actions which yield immediate reward and taking actions (such as acquiring information, or preparing the ground) whose benefit will come only later. It has proved difficult enough to become a classic, and has now a large literature (for two recent contributions, apart from those of the Gittins school, to which I shall refer especially, see Wahrenberger, Antle and Klimko, 1977 and Rodman, 1978).

In its basic version one has N parallel projects or investigations, indexed $i = 1, 2, ..., N$ and at each instant of discrete time can work on only a single project. Let the state of project i at time t be denoted $x_i(t)$. If one works on project i at t then one acquires an immediate expected reward of $R_i(x_i(t))$. Rewards are additive, and discounted in time by factor β. The state value $x_i(t)$ changes to $x_i(t+1)$ by a Markov transition rule (which may depend upon i, but not upon t), while the state of the projects one has not touched remain unchanged: $x_j(t+1) = x_j(t)$ for $j \neq i$. The problem is how to allocate one's effort over projects sequentially in time so as to maximize expected total discounted reward over $t = 0, 1, 2, \ldots$.

The problem resisted essential solution until Gittins and his co-workers made fundamental progress in a series of papers which have yet to be recognized generally (see especially Gittins and Jones, 1974, Gittins and Glazebrook, 1977 and Gittins, 1979, 1980, but also the other papers listed for these authors in the references).

They proved that to each project i one could attach an index $v_i = v_i(x_i(t))$, this being a function of the project i and its current state $x_i(t)$ alone, such that the optimal action at time t is to work on that project for which the current index is greatest. The index v_i is calculated by solving the problem of allocating one's effort optimally between project i and a *standard project* which yields a constant reward (and so effectively has also a single state). Gittins' result thus reduces the case of general N to that of the case $N = 2$ (or to the case $N = 1\frac{1}{2}$, one might say, since one of the projects in the reduced problem is standard). In his 1979 paper Gittins gives a second illuminating interpretation of the index v_i in terms of a "forwards induction" rule, maximizing average yield up to an optimally chosen stopping time.

In his deduction of the index rule and his two characterizations of the index Gittins showed a rare intuition. Unfortunately, his proofs of the optimality of the index rule have been very difficult to follow, and this has doubtless been a reason why the full merits and point of this work have not yet been generally appreciated.

In this paper I give a simple proof of the optimality of the Gittins index procedure. The treatment begins from a conjecture, which implies an identity (12) between the maximal rewards for the cases of one and several projects. In the course of verifying this identity one both proves optimality of Gittins' rule, and extends the treatment to a version of a finite-horizon situation.

Nash (1973) produced an index result for the more general case of what Gittins terms a "bandit superprocess" for which, once one has decided to work on a given project, one can also choose between several procedures one might follow for pursuing that project. There is thus a second level of optimization. An index rule will be optimal for a superprocess only under certain conditions; in Section 5 we find that our treatment extends immediately to yield a new sufficient condition.

To submit a paper on this subject to a statistical journal is not inappropriate, since Gittins' own review (1979) of his work took the form of a paper read to the Royal Statistical Society. Furthermore, the "information-gathering" aspect of a project will often take the form of a sequential inference on unknown parameters.

Gittins refers to the index v_i as the "dynamic allocation index", abbreviated DAI. The term has an obvious rationale, but is clumsy to use. I find it both convenient and proper to refer to v_i as the "Gittins index".

2. Preliminaries

Let us denote the combined state vector $(x_1, x_2, ..., x_N)$ by x, or by $x(t)$ if we need to refer to its value at a particular time t. We shall assume the rewards uniformly bounded,

$$k(1-\beta) \leqslant R_i(x_i) \leqslant K(1-\beta), \tag{1}$$

where k, K are constants (which may be negative) and β is the discount factor. We assume strict discounting, so that $0 \leqslant \beta < 1$.

The reward at time t will be $R_{i(t)}(x_{i(t)})$ if $i(t)$ is the project engaged at time t; for simplicity we shall write this simply as $R(t)$. The total discounted reward $\Sigma \beta^t R(t)$ is then well defined, and bounded between k and K. Denote the maximal expected value of this reward, conditional on $x(0) = x$, by $\Phi(x)$. Then Φ is the unique bounded solution of the dynamic programming equation

$$\Phi = \max_i L_i \Phi, \tag{2}$$

where $L_i \Phi(x)$ is the expected reward if, in state x, one works on project i for one step and then, in the new state x' (differing from x only in the ith component), receives reward $\Phi(x')$. That is

$$L_i \Phi(x) = R_i(x_i) + \beta E[\Phi(x(t+1)) \,|\, x(t) = x, i(t) = i]. \tag{3}$$

Consider now the process when modified by addition of the option that one can "retire" (i.e. abandon all projects) at any time, for an immediate reward of M. Term this the "M-process", the original process the "continuing process", and let $F(x, M)$ be the analogue of $\Phi(x)$ for the M-process. F is then the unique bounded solution of

$$F = \max(M, \max L_i F). \tag{4}$$

Lemma 1. $F(x, M)$ is a non-decreasing convex function of M with

$$F(x, M) = \begin{cases} \Phi(x), & M \leqslant k, \\ M, & M \geqslant K. \end{cases} \tag{5}$$

For $M < k$ the optimal policies of the M-process and the continuing process are identical.

Proof. To increase M can only increase reward, so the non-decreasing character is evident. To retire is always optimal for $M \geqslant K$ and to continue is always optimal for $M \leqslant k$ whence (5) follows. To retire is never optimal if $M < k$, whence the last assertion.

Let V be the expected return from a policy whose prescription is independent of M. Then

$$V = V_c + ME(\beta^T), \tag{6}$$

where V_c is the expected reward before retirement (independent of M), and T is the time of retirement. All expectations in (6) are those determined by the policy, and are conditional on the observational history at $t = 0$. The event of non-retirement can be identified with the event $T = +\infty$ since, because $|\beta| < 1$, either convention will cause this contingency to yield zero contribution to $E(\beta^T)$.

Since F is the infimum over policies of expression (6), linear in M, it is convex in M. ■

Lemma 2. For almost all M

$$\frac{\partial F(x, M)}{\partial M} = E_M(\beta^T \mid x(0) = x), \tag{7}$$

where E_M denotes expectation under a policy optimal for the M-process.

Proof. Since in each state there are at most $N+1$ possible actions, optimal policies always exist. If one applies an M-optimal policy to the $(M+\delta)$-process one achieves an expected reward not exceeding $F(x, M+\delta)$; this assertion and relation (6) imply that for any M, δ

$$F(x, M+\delta) - F(x, M) \geqslant \delta E_M(\beta^T \mid x(0) = x). \tag{8}$$

(The expectation is conditional on initial history, but for an M-optimal policy the only effective conditioning variable is initial state.)

It follows from (8) that expression (7) is a sub-gradient to F (as a function of M), and so coincides with the gradient of F wherever this exists. But F, being convex, will have a gradient almost everywhere; hence we deduce the assertion of the Lemma. ■

If the optimal policy at M is non-unique then $E_M(\beta^T)$ may take more than one value. However, it follows from (8) that *all* such determinations of $E_M(\beta^T)$ are sub-gradients of F, and that the determination is in fact then unique at those values for which $\partial F/\partial M$ is defined.

Typically, a value of M at which $\partial F/\partial M$ does not exist, in that there is a discontinuity in its value, corresponds to a value of M at which there is a discontinuous change in the optimal policy. It is then understandable that there are at least two policies optimal at such a point.

3. A Conjecture, and its Implications

We shall now give a conjecture which implies, directly but not trivially, that F must have a particular form. In Section 4 we establish by direct arguments the validity of this form, the validity of the conjecture, and the optimality of the index rule.

It would be more satisfying to establish the truth of the conjecture by methods which do not involve simultaneously the other two verifications; this I have not been able to do. However, to present the conjecture first seems conceptually right, because it embodies the crucial idea.

We shall denote the conjecture by C, and shall distinguish all assertions whose validity is conditional on that of C (two lemmas of this section) also by a C.

Let us denote by $\mathscr{P}$ the class of policies for the bandit process for which project i is never used if its state lies in a some set S_i; and for which retirement takes place exactly when first $x_i \in S_i$ for all i.

That is, project i is "abandoned" when its state falls in a set S_i, and one retires when all projects are abandoned. Apart from this, there is no assumption that the projects are used in any particular order, or even that the policy is Markov.

Conjecture C. There is an M-optimal policy in $\mathscr{P}$.

If this is true, then the set S_i must be exactly the set of x_i for which one would retire if one only had the single project i, and so the only options were those of continuation or retirement. That is,

if we define $\phi_i(x_i, M)$ as the analogue of $F(x, M)$ for this single-project case, so that the analogue of (4) is

$$\phi_i = \max(M, L_i \phi_i) \tag{9}$$

then S_i is the set of x_i for which $\phi_i = M$.

Lemma 3. Let T be the time of retirement for a process with initial state vector $x = (x_1, x_2, ..., x_N)$ and τ_i the time of retirement for the process containing only project i with initial state x_i. If a policy in $\mathscr{P}$ is used then

$$E(\beta^T) = \prod_i E(\beta^{\tau_i}) \tag{10}$$

where the expectations are conditional on observational history up to $t = 0$.

Proof. We plainly have $T = \Sigma \tau_i'$, summing over i, where τ_i' is the "process time until abandonment" for project i : i.e. the number of times project i is operated. The assertion of the Lemma is equivalent to saying that the τ_i' are independently distributed, and that τ_i, τ_i' have the same distribution. But, since project evolution (in process time) is independent, the process time needed to take the state of project i into S_i is independent of the states of other projects, hence the assertion. ■

Although history before time $t = 0$ can conceivably affect the order in which projects are engaged, it cannot in fact affect the distribution of T, for a policy in $\mathscr{P}$.

The validity of (10) does not require that $T < \infty$. In the stochastic equivalence $T = \Sigma \tau_i$ the event $\tau_i = +\infty$ is understood to imply $T = +\infty$; if one project is never abandoned, retirement never occurs.

C. Lemma 4. For almost all M

$$\frac{\partial F(x, M)}{\partial M} = \prod_i \frac{\partial \phi_i(x_i, M)}{\partial M}. \tag{11}$$

Proof. The assertion follows from (7) and (10). But, of course, the appeal to (10) is justified only if there is an M-optimal policy in $\mathscr{P}$.

C. Lemma 5

$$F(x, M) = K - \int_M^K \prod_i \frac{\partial \phi_i(x_i, m)}{\partial m}\, dm. \tag{12}$$

Proof. The assertion follows from integration of (11), and determination of the constant of integration from (5). ■

The policies of $\mathscr{P}$ could be appropriately termed "write-off" policies, since project i is abandoned if and only if its state enters a "write-off" set S_i.

4. Verification of the Conjecture and of Optimality of the Index Rule

We shall verify the relation (12) in an apparently more general case. Let $F(x, s, M)$ denote the maximal expected reward for the M-process if the process time for project i may not exceed $s_i (i = 1, 2, ..., N)$, and let $\phi_i(x_i, s_i, M)$ denote the corresponding analogue of $\phi_i(x_i, M)$. If we define

$$\hat{F}(x, s, M) = K - \int_M^K \prod_i \frac{\partial \phi_i(x_i, s_i, m)}{\partial m}\, dm \tag{13}$$

then we shall establish that $F = \hat{F}$, and that this reward is achievable by a finite-horizon version of the Gittins index rule.

This is a finite-horizon version of the problem insofar as we set a finite horizon for each project: s_i for project i. The modification in a sense represents no real generalization, because (x_i, s_i) can just be viewed as a modified state-variable for project i. However, it does represent a useful enlargement of one's concept of the solution, and allows one to use inductive arguments.

Note that ϕ_i will now obey the generalized form of (9):

$$\phi_i(x_i, s_i, M) = \max[M, L_i \phi_i(x_i, s_i, M)], \tag{14}$$

where L_i is now the operator characterized by

$$L_i \theta(x, s, M) = R_i(x_i) + \beta E[\theta(x(t+1), D_i s, M) \,|\, x(t) = x, i(t) = i] \tag{15}$$

and the effect of D_i on s is to decrease s_i by one.

Let $M_i(x_i, s_i)$ denote the value of M which places (x_i, s_i) on the boundary of the continuation and retirement sets; i.e. it is the infimal value of M for which $\phi_i(x_i, s_i, M) = M$. We shall sometimes write this simply as M_i; with the (x_i, s_i)-dependence understood. The quantity $\nu_i = (1-\beta) M_i$ is just the Gittins index.

Theorem 1. The maximal reward $F(x, s, M)$ equals $\hat{F}(x, s, M)$ defined in (13). It is realized by a policy in which at (x, s) one engages a project i maximizing $M_i(x_i, s_i)$ if this maximal value exceeds M, and otherwise retires.

Proof. We shall establish identity of F and $\hat{F}$ if we can show that they agree at $s = 0$ and that $\hat{F}$ satisfies the dynamic programming equation

$$\hat{F} = \max(M, \max_i L_i \hat{F}), \tag{16}$$

where L_i now has the extended definition (15). Equation (16) is now recursive in s, in that the effect of an L-operator is to decrease Σs_i by one. We shall establish optimality of the index policy described in the Theorem if we can show that this policy chooses the maximizing option in (16). That is, that if the sequence $(M, M_1, M_2, \ldots)$ is maximal at its jth position then so is the sequence $(M, L_1 \hat{F}, L_2 \hat{F}, \ldots)$.

Suppose that $F(x, s, M) = \hat{F}(x, s, M)$ for $\Sigma_i s_i \leqslant r$. This assumption is certainly true for $r = 0$, because $F(x, 0, M) = M$, also $\phi_i(x_i, 0, M) = M$, and we see then from (13) that $\hat{F}(x, 0, M) = M$.

Define now

$$P_i(x, s, M) = \prod_{j \neq i} \frac{\partial \phi_j(x_j, s_j, M)}{\partial M}. \tag{17}$$

Regarded as a function of M this is non-negative, non-decreasing, and equal to unity for

$$M \geqslant M_{(i)} \triangleq \max_{j \neq i} M_j, \tag{18}$$

as is seen by appeal to Lemma 1 for each ϕ_i, and relation (9). It can then be regarded as a distribution function (in M) with all its support in $M \leqslant M_{(i)} \leqslant K$. We find by partial integration of (13) that

$$\hat{F}(x, s, M) = \phi_i(x_i, s_i, M) P_i(x, s, M) + \int_M^\infty \phi_i(x_i, s_i, m) \, d_m P_i(x, s, m). \tag{19}$$

For the rest of the proof we can suppress the (x, s) argument. We find from (19) that

$$\hat{F}(M) - L_i \hat{F}(M) = \delta(M) P_i(M) + \int_M^\infty \delta(m) \, d_m P_i(m), \tag{20}$$

where

$$\delta(M) = \phi_i(M) - L_i \phi_i(M).$$

But it follows from (14) that $\delta(M) \geqslant 0$, with equality if $M \leqslant M_i$. It then follows from (20) that

$$\hat{F} \geqslant L_i \hat{F}. \tag{21}$$

with equality if $M \leqslant M_i$ and $M_{(i)} \leqslant M_i$. This latter inequality is equivalent to $M_i = \max_j M_j$, so we deduce that $\hat{F}$ satisfies (16) with $\hat{F} = L_i \hat{F}$ if

$$M_i = \max_j M_j \geqslant M.$$

In the case $\max_j M_j \leqslant M$ we have $\phi_j(m) = m$ for all i if $m \geqslant M$, and (13) evaluates $\hat{F}$ as M. From this assertion and (21) we see that $\hat{F}$ satisfies (16) also in this case, with $\hat{F} = M$, corresponding to adoption of the retirement option.

Verification of (16) implies that if $F = \hat{F}$ for $\Sigma s_i \leqslant r$ then the assertion also holds for $\Sigma s_i \leqslant r+1$, so the induction is complete. We have then also identity of $F, \hat{F}$ and optimality of the index-rule in the infinite-horizon case, by known theorems for discounted processes (see Blackwell, 1965). ■

The index $M_i(x_i)$ evidently has the interpretation of an equitable surrender-value for project i when in state x_i.

5. SUPERPROCESSES

One can include an additional decision variable u_i whose value affects the reward, $R_i(x_i, u_i)$ and also the $x_i(t) \to x_i(t+1)$ transition rules, if project i is adopted at time t. So, at each instant of time one has first the decision of choosing which project i to work upon, and then of choosing which procedure u_i to follow. Gittins (1979) refers to a bandit process with this extra level of decision as a superprocess, and indicates that, at least under certain conditions, the index criterion may still yield an optimal policy, if either of the index characterizations are modified to incorporate a u-optimization as well as optimization of the retirement option.

One naturally asks whether solution (12) still holds if the definitions of F and ϕ_i are modified in the obvious way: to incorporate a u-optimization as well as the transfer optimizations demanded previously. However, the conjecture of Section 3 is now suspect, because it is not clear that the u-optimizations for different projects do not interact. Moreover, the proof of Theorem 1 now seems to fail; except in one case.

Theorem 2. Suppose that the quantities F, ϕ_i, Φ and M_i all have their definitions modified by incorporation of u-optimization as well as of the appropriate transfer optimizations. Then Theorem 1 still holds if, in the continuation regions for the one-project problems, the optimal u-decisions are independent of M.

Proof. For simplicity we shall consider only the infinite-horizon case, for which the argument s vanishes. The analogous results for the finite-horizon case are then implied, by the remarks at the beginning of Section 4.

The dynamic programming equation for ϕ_i is now

$$\phi_i = \max(M, \sup_u L_{iu} \phi_i), \tag{22}$$

where

$$L_{iu}\theta(x, M) = R_i(x_i, u) + \beta E[\theta(x(t+1), M) \,|\, x(t) = x, i(t) = i, u(t) = u].$$

Correspondingly, F is the unique bounded solution of

$$F = \max(M, \max_i \sup_u L_{iu} F).$$

By appeal to the form (19) of $\hat{F}$, as before, we find that

$$\hat{F}(M) - L_{iu}\hat{F}(M) = \delta(M) P_i(M) + \int_M^\infty \delta(m)\, d_m P_i(m),$$

where

$$\delta(M) = \phi_i(M) - L_{iu}\,\phi_i(M).$$

But it follows from (22) that $\delta(M) \geqslant 0$ with equality if both $M \leqslant M_i$ and u takes the value $h(x_i)$ which is optimal in the one-project case (for all M, by hypothesis).

The remainder of the argument then follows exactly as in Theorem 1. However, for this argument to hold we do have to be able to assert that $\delta(m) = 0$ for $M \leqslant m \leqslant M_i$, and so that there is a common optimal u-action for all retirement rewards in this interval. ■

REFERENCES

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

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

—— (1979). Bandit processes and dynamic allocation indices (with Discussion). *J. R. Statist. Soc.* B, **41**, 148–164.

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

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

—— —— (1974b). *A Dynamic Allocation Index for New-product Chemical Research.* Cambridge University Engineering Dept CUED/A–Mgt Stud/TR13.

GITTINS, J. C. and NASH, P. (1977). Scheduling, queues and dynamic allocation indices. *Proc. EMS, Prague 1974*, pp. 191–202. Prague : Czechoslovak Academy of Sciences.

GLAZEBROOK, K. D. (1976a). A profitability index for alternative research projects. *Omega*, **4**, 79–83.

—— (1976b). Stochastic scheduling with order constraints. *Int. J. Sys. Sci.*, **7**, 657–666.

—— (1978a). On a class of non-Markov decision processes. *J. Appl. Prob.*, **15**, 689–698.

—— (1978b). On the optimal allocation of two or more treatments in a controlled clinical trial. *Biometrika*, **65**, 335–340.

JONES, D. M. (1970). A sequential method for industrial chemical research. M.Sc. Thesis, University of Wales, Aberystwyth.

NASH, P. (1973). Optimal allocation of resources between research projects. Ph.D. Thesis, Cambridge University.

NASH, P. and GITTINS, J. C. (1977). A hamiltonian approach to optimal stochastic resource allocation. *Adv. Appl. Prob.*, **9**, 55–68.

RODMAN, L. (1978). On the many-armed bandit problem. *Ann. Prob.*, **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.