J. R. Statist. Soc. B (1979),
41, *No.* 2, *pp.* 148–177

Bandit Processes and Dynamic Allocation Indices

By J. C. GITTINS

Keble College, Oxford

[Read before the ROYAL STATISTICAL SOCIETY at a meeting organized by the RESEARCH SECTION on Wednesday, February 14th, 1979, the Chairman Professor J. F. C. KINGMAN in the Chair]

SUMMARY

The paper aims to give a unified account of the central concepts in recent work on bandit processes and dynamic allocation indices; to show how these reduce some previously intractable problems to the problem of calculating such indices; and to describe how these calculations may be carried out. Applications to stochastic scheduling, sequential clinical trials and a class of search problems are discussed.

Keywords: BANDIT PROCESSES; DYNAMIC ALLOCATION INDICES; TWO-ARMED BANDIT PROBLEM; MARKOV DECISION PROCESSES; OPTIMAL RESOURCE ALLOCATION; SEQUENTIAL RANDOM SAMPLING; CHEMICAL RESEARCH; CLINICAL TRIALS; SEARCH

1. INTRODUCTION

A scheduling problem

There are n jobs to be carried out by a single machine. The times taken to process the jobs are independent integer-valued random variables. The jobs must be processed one at a time. At the beginning of each time unit any job may be selected for processing, whether or not the job processed during the preceding time unit has been completed, and there is no penalty or delay involved in switching from one job to another. The probability that $t+1$ time units are required to complete the processing of job i, conditional on more than t time units being needed, is $p_i(t)$ $(i = 1, 2, \ldots, n;\ t \in \mathbf{Z})$. The reward for finishing job i at time s is $a^s V_i$ $(0 < a < 1;\ V_i > 0, i = 1, 2, \ldots, n)$, and there are no other rewards or costs. The problem is to decide which job to process next at each stage so as to maximize the total expected reward.

A multi-armed bandit problem

There are n arms which may be pulled repeatedly in any order. Each pull takes one time unit and only one arm may be pulled at a time. A pull may result in either a success or a failure. The sequence of successes and failures which result from pulling arm i forms a Bernoulli process with an unknown success probability θ_i $(i = 1, 2, \ldots, n)$. A successful pull on any arm at time t yields a reward a^t $(0 < a < 1)$, whilst an unsuccessful pull yields a zero reward. At time zero θ_i has the probability density

$$(\alpha_i(0)+\beta_i(0)+1)!\,(\alpha_i(0)!\,\beta_i(0)!)^{-1}\,\theta_i^{\alpha_i(0)}(1-\theta_i)^{\beta_i(0)},$$

i.e. a beta distribution with parameters $(\alpha_i(0), \beta_i(0))$, and these distributions are independent for the different arms. The problem is to decide which arm to pull next at each stage so as to maximize the total expected reward from an infinite sequence of pulls.

From Bayes' theorem it follows that at every stage θ_i has a beta distribution, but with parameters which change at each pull on arm i. If in the first t pulls there are r successes, the new values of the parameters, which we denote by $(\alpha_i(t), \beta_i(t))$, are $(\alpha_i(0)+r, \beta_i(0)+t-r)$. If the $(t+1)$st pull on arm i takes place at time s, the expected reward, conditional on the record of successes and failures up to then, is a^s times the expected value of a beta variate with parameters $(\alpha_i(t), \beta_i(t))$, which is $(\alpha_i(t)+1)/(\alpha_i(t)+\beta_i(t)+2)$.

Both the problems described above involve a sequence of decisions, each of which is based on more information than its predecessors, and thus both problems may be tackled by dynamic

programming (see Bellman, 1957). This is a computational algorithm based on the principle that, "an optimal policy has the property that whatever the initial state and initial decision, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision". This observation means that if the optimal policy from a certain stage (or time) onwards is known, then it is relatively easy to extend this policy so as to give an optimal policy starting one stage earlier. Repetition of this procedure is the basis of an algorithm for solving such problems, which is often described as a process of *backwards induction.*

A simpler procedure than backwards induction is at each stage to make that decision which maximizes the expected reward before the next decision time. This procedure will be termed a *one-step look-ahead policy*, following the terminology used by Ross (1970) for stopping problems. The idea is that each decision is based on what may happen in just one further time unit or step.

The notion of a one-step look-ahead policy may be extended in the obvious way to form *s-step look-ahead policies.* In general such policies perform better as s increases and approach optimality as s tends to infinity, whilst the algorithms to which they lead become progressively more complex as s increases.

As a further extension of an s-step look-ahead policy we may allow the number of steps τ which we look ahead at each stage to depend in an arbitrary manner on what happens whilst those steps are taking place, so that τ is a random variable. Given any rule for taking our sequence of decisions, τ may be chosen so as in some sense to maximize the expected rate of reward per step for the next τ steps. A second maximization with respect to decision rules selects a decision rule. Our extended look-ahead policy starts by following the decision rule just described for the random number of steps τ. The process of finding a decision rule, and a corresponding random number of further steps τ', is then repeated with respect to the state reached after the first τ steps. The new rule is followed for the next τ' steps, and the process may be repeated indefinitely. In this way a rule is defined which specifies the decision to be made at every stage. Such a rule will be termed a *forwards induction policy*, in contrast with the backwards induction of dynamic programming. A formal definition is given in Section 3.

Forwards induction policies are optimal for a class of problems, which includes the two problems described above, in which effort is allocated in a sequential manner between a number of competing candidates for that effort, a result which will be described as the *forwards induction theorem.* These candidates will be described as alternative *bandit processes.* From the optimality of forwards induction policies it follows that a *dynamic allocation index* (DAI) may be defined on the state space of each bandit process, with the property that an optimal policy must at each stage allocate effort to one of those bandit processes with the largest DAI value. This result will be described as the DAI *theorem* and the policy as a DAI *policy.* The proofs of these results will be published separately (Gittins, 1979).

The existence of a function with this property, and the fact that it may be written in the form used here, were proved in earlier papers (Gittins and Jones, 1974a; Gittins and Glazebrook, 1977) without using the concept of a forwards induction policy, and the particular cases discussed in the present paper depend only on these results. The approach via the forwards induction theorem has the advantage that it is intuitively plausible that such a result should hold, and it leads naturally, as we shall see, to the general functional form of the dynamic allocation index. Moreover, the forwards induction theorem continues to hold under appropriate conditions, and essentially the same proof works, if bandit processes arrive in a random manner, or are subject to precedence constraints. This leads to results analogous to the DAI theorem in the theories of priority queues and of more complex stochastic scheduling situations. Some of these applications have been described by Nash (1973) and Glazebrook (1976a, b), respectively. A more complete account, using the simplifying concept of a forwards induction policy, will be published in due course. Sometimes, too, as shown by Glazebrook (1978a), a decision problem may be simplified by expressing just part of the problem in terms of bandit processes.

In the present paper these extensions are mentioned only in passing. The aims are: (i) to give a unified account, in the context of Markov decision processes and without detailed proofs, of the central concepts in recent work on bandit processes and DAIs; (ii) to show how these concepts reduce some previously intractable problems to the problem of calculating DAIs; and (iii) to describe how these calculations may be carried out.

A bandit process is defined in Section 2, and the main theorems are formally stated and discussed in Section 3. In Section 4 the general functional form of the DAI is examined more closely, and Section 5 shows how this simplifies under certain conditions. Formulae for the DAI function for the scheduling problem are derived in Section 6. Possible applications include the scheduling of jobs on a computer and the allocation of effort between competing research projects. A method of calculating, and the general form of, the DAI function for the multi-armed bandit are described in Section 7. Section 8 describes a method of calculating the DAI function for any bandit process. The main possibility of applying the results of Section 7 is in clinical trials. DAI functions for similar, and sometimes more realistic, problems for which the result of each trial is a normally distributed random variable are discussed in Section 9. In Section 10 a variant of the multi-armed bandit problem is considered in which the object is to minimize the expected number of trials up to the first success, rather than to maximize the expected value of an infinite stream of successes and failures. Once again a version of the problem for which the distribution of scores on each trial is normally distributed is of interest, as well as the Bernoulli trials version. This problem has possibilities of application to the screening of chemicals in pharmaceutical research.

For the sake of simplicity attention is restricted to discrete-time bandit processes. Every result mentioned here also has a continuous-time counterpart, which may be obtained by letting the discrete-time quantum tend to zero in an appropriate fashion. For example, Nash and Gittins (1977) establish the continuous-time version of the optimal policy for the scheduling problem, though using a different method.

2. Bandit Processes

All the processes considered are indexed by a time variable whose value set is the non-negative integers, which we denote by $\mathbf{Z}$. They are also stationary, i.e. their properties involve no explicit time-dependence, and are particular types of Markov decision process. It may be noted that the assumption of stationarity rules out versions of the allocation problems considered with finite time horizons. The reason for the restriction (see Gittins, 1975, and Gittins and Nash, 1977) is that DAI policies are not in general optimal in such cases.

A *Markov decision process* is defined on a state-space Θ, together with a σ-algebra $\mathscr{X}$ of subsets of Θ which includes every subset consisting of just one element of Θ. When the process is in state x the set of controls which may be applied is $\Omega(x)$. $P(A|x,u)$ is the probability that the state y of the process at time $t+1$ belongs to A ($\in\mathscr{X}$), given that at time t the process is in state x and control $u(\in\Omega(x))$ is applied. Application of control u at time t with the process in state x yields a reward $a^t R(x,u)$ $(0<a<1)$. The functions $P(A|\cdot,u)$ and $R(\cdot,u)$ are $\mathscr{X}$-measurable.

A *policy* for a Markov decision process is any rule, including randomized rules, which for all t specifies the control to be applied at time t as a function of t, the states at times $0,1,2,\ldots,t$, and the controls applied at times $0,1,2,\ldots,t-1$; we shall describe this by saying that the control at time t is *sequentially determined. Deterministic* policies are those which involve no randomization. *Stationary* policies are those which involve no explicit time-dependence. *Markov* policies are those for which the control chosen at time t is independent of the states and the controls applied at times $0,1,2,\ldots,t-1$.

Blackwell (1965) has shown that if the control set $\Omega(x)$ is finite and the same for all x then there is a deterministic stationary Markov policy for which, for any initial state, the total expected reward is the supremum of the total expected rewards for the class of all policies. We shall refer to this result as *Blackwell's theorem*, and to a policy which achieves the supremum

just mentioned as an *optimal* policy. It is assumed throughout the paper that $\Omega(x)$ is finite for all x and that the supremum of the total expected reward is finite. To a large extent, therefore, attention may be restricted to deterministic stationary Markov policies. Such a policy is defined by an $\mathscr{X}$-measurable function g on Θ such that $g(x) \in \Omega(x)$, $\forall x$.

A *bandit process* is a Markov decision process for which $\Omega(x) = \{0, 1\}$, $\forall x$. The control 0 *freezes* the process in the sense that $P(\{x\}|x, 0) = 1$ and $R(x, 0) = 0$, $\forall x$. Control 1 is termed the *continuation* control. No restriction is placed on the transition probabilities and rewards if control 1 is applied. The number of times control 1 has been applied to a bandit process is termed the *process time*. The state at process time t is denoted by $x(t)$. The reward between times t and $t+1$ if control 1 is applied at each stage, so that process time coincides with real time, is $\alpha^t R(x(t), 1)$, which we abbreviate to $\alpha^t R(t)$. A *standard* bandit process is a bandit process for which, for some λ, $R(x, 1) = \lambda$, $\forall x$.

An arbitrary policy for a bandit process is termed a *freezing rule*. Given any freezing rule f the random variables $f(t)$, $t \in \mathbf{Z}$, are sequentially determined, where $f(t)$ ($\geqslant f(t-1)$) is the number of times control 0 is applied before the $(t+1)$st application of control 1. Deterministic stationary Markov policies divide the state space Θ into a *stopping set*, on which control 0 is applied, and a *continuation set*, on which control 1 is applied. They are clearly such that $f(t) = 0$, $\forall t < \tau$, and $f(\tau) = \infty$, for some sequentially determined random variable τ, which may take the value infinity with positive probability. These properties define a *stopping rule*, and τ is the associated *stopping time*. Stopping rules have been extensively studied, for the most part in the context of stopping problems (e.g. see Chow *et al.*, 1971), which may be regarded as being defined by bandit processes for which $R(x, 0) \neq 0$. Frequent reference will be made to stopping times. It should be noted that the definition is as above, and there is no implication that the process concerned actually does stop at such a time.

The following notation will be used in conjunction with an arbitrary bandit process D. $R_f(D)$ denotes the expected total reward under the freezing rule f. Thus

$$R_f(D) = E\sum_{t=0}^{\infty} \alpha^{t+f(t)} R(t), \quad R'(D) = \sup_f R_f(D).$$

Also

$$W_f(D) = E\sum_{t=0}^{\infty} \alpha^{t+f(t)}, \quad \nu_f(D) = R_f(D)/W_f(D), \quad \text{and} \quad \nu'(D) = \sup_{\{f: f(0)=0\}} \nu_f(D).$$

Similarly, for stopping rules,

$$R_\tau(D) = E\sum_{t=0}^{\tau-1} \alpha^t R(t), \quad R(D) = \sup_\tau R_\tau(D), \quad W_\tau(D) = E\sum_{t=0}^{\tau-1} \alpha^t,$$

$$\nu_\tau(D) = R_\tau(D)/W_\tau(D) \quad \text{and} \quad \nu(D) = \sup_{\tau>0} \nu_\tau(D).$$

From Blackwell's theorem it follows that $R'(D) = R(D)$, that $\nu'(D) = \nu(D)$ (though this is less obvious) and is an $\mathscr{X}$-measurable function of x, and that stopping times exist for which the respective suprema are attained. All these quantities naturally depend on the initial state $x(0)$ of the bandit process D. When necessary $R_f(D, x)$ and $W_f(D, x)$, for example, will be used to indicate the values of $R_f(D)$ and $W_f(D)$ when $x(0) = x$.

The quantities $\nu_f(D)$ and $\nu_\tau(D)$ are thus expected rewards per unit of discounted time under f and τ respectively. The conditions $f(0) = 0$ and $\tau > 0$ in the definitions of $\nu'(D)$ and $\nu(D)$ mean that the policies considered are all such that at time zero control 1 is applied. This restriction is required to rule out zero denominators $W_f(D)$ or $W_\tau(D)$. In the case of $\nu'(D)$ it also has the effect of removing a common factor from the numerator and denominator of $\nu_f(D)$ for those f for which $f(0) \neq 0$, and otherwise implies no loss of generality. The class of stopping times $\{\tau > 0\}$ is stationary from time 1 onwards, rather than from time 0.

For reasons which will become apparent in the next section, in which the forwards induction theorem and the DAI theorem are formally stated, $\nu(D,x)$ is defined to be the *dynamic allocation index* for the bandit process D when it is in state x.

3. The Main Theorems

We begin with some further terminology and notation.

Given any Markov decision process $\mathscr{M}$, together with a deterministic stationary Markov policy g, a bandit process may be defined by introducing the freeze control 0 with the usual properties, and requiring that at each time t either the control 0 or the control given by g be applied. This bandit process is termed the *superprocess* $(\mathscr{M},g)$. Thus application of the continuation control 1 to $(\mathscr{M},g)$, when $\mathscr{M}$ is in state x, is equivalent to applying control $g(x)$ to $\mathscr{M}$. The idea of a superprocess is due to Nash (1973), who used it to show that the DAI theorem may be extended to cover the case when new bandit processes arrive in a Poisson process.

The following notation extends that already set up for a bandit process:

$$R_{g\tau}(\mathscr{M}) = R_\tau((\mathscr{M},g)), \quad W_{g\tau}(\mathscr{M}) = W_\tau((\mathscr{M},g)), \nu_{g\tau}(\mathscr{M}) = R_{g\tau}(\mathscr{M})/W_{g\tau}(\mathscr{M}),$$
$$\nu_g(\mathscr{M}) = \sup_{\tau>0} \nu_{g\tau}(\mathscr{M}), \quad \nu(\mathscr{M}) = \sup_g \nu_g(\mathscr{M}).$$

Since for an arbitrary bandit process D there is a stopping time τ for which the supremum is attained in the definition of $\nu(D)$ it follows that the same is true of $\nu_g(\mathscr{M})$. Also if the control set $\Omega(x)$ is finite for all x then Blackwell's theorem may be extended to show that, for some g, $\nu_g(\mathscr{M}) = \nu(\mathscr{M})$, and $\nu(\mathscr{M})$ is unaltered if g is allowed to range over the entire set of policies for $\mathscr{M}$. As for bandit processes, $\nu(\mathscr{M},x)$, for example, denotes the value of $\nu(\mathscr{M})$ when $\mathscr{M}$ is initially in state x.

With this notation we are now in a position to give a formal definition of a forwards induction policy for the Markov decision process $\mathscr{M}$, whose state at time zero we denote by x_0. The first step is to find a policy γ_1 and a stopping time σ_1 such that the discounted average reward per unit time of the superprocess $(\mathscr{M},g)$ up to the stopping time $\tau\ (>0)$ is maximized over all g and τ by setting $(g,\tau) = (\gamma_1,\sigma_1)$. Thus $\nu_{\gamma_1\sigma_1}(\mathscr{M}) = \nu(\mathscr{M})\ (= \nu(\mathscr{M},x_0))$.

Let x_1 be the (random) state of the superprocess $(\mathscr{M},\gamma_1)$ at time σ_1. We now define the policy γ_2 and the stopping time σ_2 to be such that $\nu_{\gamma_2\sigma_2}(\mathscr{M},x_1) = \nu(\mathscr{M},x_1)$. In general γ_2 and σ_2 depend on x_1, and are such that the discounted average reward per unit time of $(\mathscr{M},g)$ up to $\tau\ (>0)$ is maximized when $(g,\tau) = (\gamma_2,\sigma_2)$ if $\mathscr{M}$ is initially in state x_1.

A forwards induction policy for $\mathscr{M}$ starts by applying policy γ_1 up to time σ_1, and then applies policy γ_2 up to time $\sigma_1+\sigma_2$. Let x_2 be the state of $\mathscr{M}$ at this stage, and define γ_3 and σ_3 to be such that $\nu_{\gamma_3\sigma_3}(\mathscr{M},x_2) = \nu(\mathscr{M},x_2)$. A forwards induction policy continues by applying policy γ_3 between times $\sigma_1+\sigma_2$ and $\sigma_1+\sigma_2+\sigma_3$. Let x_3 be the state of $\mathscr{M}$ at this stage, and define γ_4 and σ_4 to be such that $\nu_{\gamma_4\sigma_4}(\mathscr{M},x_3) = \nu(\mathscr{M},x_3)$. A forwards induction policy applies γ_4 between times $\sigma_1+\sigma_2+\sigma_3$ and $\sigma_1+\sigma_2+\sigma_3+\sigma_4$.

This process may obviously be continued indefinitely, thus defining the class of *forwards induction policies* for the Markov decision process $\mathscr{M}$. There may be more than one such policy for the same x_0 since there may, for example, be more than one γ_1 and σ_1 such that $\nu_{\gamma_1\sigma_1}(\mathscr{M}) = \nu(\mathscr{M})$.

The term *forwards induction policy* is in contrast to a backwards induction policy derived from the dynamic programming optimality principle quoted in Section 1. This principle leads to a recurrence relation which goes backwards in time (equations (13), (14) and (16) are examples), from which an optimal policy may be determined by backwards induction. With a forwards induction policy, at each successive stopping time the expected reward per unit of discounted time up to the next stopping time is maximized, so the policy is defined by a sequence of steps proceeding forwards in time. The step length is the sequentially determined time $\sigma_r\ (r = 1,2,\ldots)$.

Forwards induction policies are often easier to determine than backwards induction policies. However, unless suitable restrictions are put on $\mathscr{M}$ they are not optimal. Fortunately there is one quite large class of Markov decision processes for which forwards induction policies *are* optimal, as well as being relatively simple to determine. These are the processes which may be regarded as simple families of alternative bandit processes.

A *family of alternative bandit processes* is formed by bringing together a set of n bandit processes, with the constraint that control 1 must be applied to just one bandit process at a time, so that control 0 is applied to the other $n-1$ bandit processes. The reward at time t is the reward yielded by the bandit process to which control 1 is applied at time t. Thus at each stage the bandit processes are alternative candidates for continuation. We shall suppose that there are no constraints restricting the set of bandit processes which may be chosen for continuation at any time. In the absence of such constraints a family of alternative bandit processes will be described as *simple*.

We may now state the following theorem.

The Forwards Induction Theorem. *For a simple family of alternative bandit processes a policy is optimal if and only if it coincides almost always with a forwards induction policy.*

In order to gain some feeling for why it is that a forwards induction policy is optimal for simple families of alternative processes, but not for all Markov decision processes, consider the problem of choosing a route for a journey by car. Suppose there are several different possible routes all of the same length which intersect at various points, and the object is to choose that route which minimizes the time taken. The problem may be modelled as a Markov decision process by interpreting the distance so far covered as the "time" variable, the time taken to cover each successive mile as minus the reward, position as the state, and choosing a value just less than one for the discount factor a. The control set $\Omega(x)$ has more than one element when the state x corresponds to a cross-roads, the different controls representing the various possible exits.

For this problem the first stage in a forwards induction policy is to find a route γ_1, and a distance σ_1 along γ_1 from the starting point, such that the average speed in travelling the distance σ_1 along γ_1 is maximized. Thus a forwards induction policy might very well start with a short stretch of motorway, which then must be followed by a very slow section, in preference to a trunk road which permits a good steady average speed. The trouble is that irrevocable decisions have to be taken at each cross-roads in the sense that those exits which are not chosen are not available later on.

The distinctive property of a simple family of alternative bandit processes is that decisions are not in this sense irrevocable, since any bandit process which is available for continuation at some stage, and which is not then chosen, may be continued at any later stage, and with exactly the same resulting sequence of rewards, apart from the discount factor. This means there is no later advantage to compensate for the initial disadvantage of not choosing a forwards induction policy.

The first stage of a forwards induction policy is such that the expected reward per unit of discounted time up to an arbitrary stopping time is maximized. For a simple family of alternative bandit processes it is intuitively plausible, and it can be rigorously shown, that this maximum is attainable by a policy under which just one of the alternative bandit processes is continued up to the stopping time in question. The reason is that if more than one bandit process were to be continued during the first stage, then the expected reward per unit of discounted time during the first stage would be a weighted average of the expected rewards per unit of discounted time for each of the bandit processes to be continued. Since a weighted average is never larger than the largest of the quantities averaged it follows that there is no point in averaging over more than one quantity, i.e. no point in continuing more than one bandit process during the first stage. This observation may be developed as a formal proof.

Now for any single bandit process D in state x the maximum expected reward per unit of discounted time up to an arbitrary stopping time is by definition the DAI, $\nu(D,x)$. In the light of the previous paragraph it thus follows that at time zero one of the bandit processes whose DAI is then maximal should be continued. This leads to

The DAI Theorem. *For a simple family of alternative bandit processes a policy is optimal if and only if at each stage the bandit process selected for continuation is almost always one of those whose dynamic allocation index is then maximal.*

4. A More Precise Characterization of the DAI

The final result of this section leads to the algorithm described in Section 7 for calculating the DAI for the multi-armed bandit problem. The proofs indicate the kind of argument required in proving the two main theorems.

As mentioned in Section 2, the DAI for a bandit process D in state x may be written as

$$\nu(D,x) = \sup_{(f:f(0)=0)} \left[E\left\{ \sum_{t=0}^{\infty} a^{t+f(t)} R(x(t),1) \,|\, x(0)=x \right\} \Big/ E\left\{ \sum_{t=0}^{\infty} a^{t+f(t)} \,|\, x(0)=x \right\} \right] \tag{1}$$

$$= \sup_{\tau>0} \nu_\tau(D,x) = \sup_{\tau>0} E\left[\left\{ \sum_{t=0}^{\tau-1} a^t R(x(t),1) \,|\, x(0)=x \right\} \Big/ E\left\{ \sum_{t=0}^{\tau-1} a^t \,|\, x(0)=x \right\}. \right] \tag{2}$$

The expression (2) uses the fact that the set of freezing rules over which the supremum is taken in (1) may be restricted to those which, from process time 1 onwards, are determined by a stopping set Θ_0 and a complementary continuation set Θ_1. We now proceed to prove the following lemma.

Lemma. The supremum in (2) is attained by setting

$$\Theta_0 = \{y \in \Theta : \nu(D,y) < \nu(D,x)\}.$$

Proof. Dropping the condition $x(0)=x$ from the notation, we have, for any non-random $s \in \mathbf{Z}^+$ and for any stopping time τ,

$$\nu_\tau(D,x) = \left[E\sum_{t=0}^{\sigma-1} a^t R(x(t),1) + E\left\{ E\sum_{t=s}^{\tau-1} a^t R(x(t),1) \,|\, x(s) \right\} \right] \Big/ \left[E\sum_{t=0}^{\sigma-1} a^t + E\left\{ E\sum_{t=s}^{\tau-1} a^t \,|\, x(s) \right\} \right], \tag{3}$$

where $\sigma = \min(s,\tau)$, and the inner expectations in both numerator and denominator are conditional on the value taken by the random variable $x(s)$. Now if $\tau > s$ then

$$E\left\{ \sum_{t=s}^{\tau-1} a^t R(x(t),1) \,|\, x(s) \right\} \Big/ E\left\{ \sum_{t=s}^{\tau-1} a^t \,|\, x(s) \right\} = \nu_{\tau-s}(D, x(s)) \leqslant \nu(D, x(s)). \tag{4}$$

From (3) and (4) it follows that if the probability of the event $E_s = \{\tau > s \cap \nu(D,x(s)) < \nu_\tau(D,x)\}$ is positive, and the random integer ρ is defined to take the value s when E_s occurs and otherwise to equal τ, then $\nu_\rho(D,x) > \nu_\tau(D,x)$. Thus if τ is such that the supremum is attained in (2) we must have $P(\bigcup_{s=1}^{\infty} E_s) = 0$. This is equivalent to saying that the probability that, starting in state x, the bandit process D passes through a state which belongs to the set defined in the statement of the lemma before process time τ is zero. Thus the stopping set Θ_0 which defines τ must include the given set, except perhaps for a subset which is reached before τ with probability zero.

A similar argument shows that $P\{\nu(D,x(\tau)) > \nu(D,x) \,|\, x(0)=x\} = 0$, since otherwise $\nu_\tau(D,x)$ could be increased by increasing τ in an appropriate fashion for those realizations of D for which $\nu(D,x(\tau)) > \nu(D,x)$. A further similar argument shows that $\nu_\tau(D,x)$ is unaffected by the inclusion or exclusion from Θ_0 of states belonging to the set $\{y \in \Theta : \nu(D,y) = \nu(D,x)\}$.

From the three preceding observations and since (i) for some τ, $\nu_\tau(D,x) = \nu(D,x)$, (ii) $\nu_\tau(D,x)$ is unchanged by changes in Θ_0 on sets which are reached by time τ with probability zero, and (iii) $\nu(D,\cdot)$ is an $\mathscr{X}$-measurable function, it follows that Θ_0 may be chosen as the lemma states. This completes the proof.

A point to be noted in the proof is that, unlike τ, the random time ρ is not necessarily defined by a freezing rule which is stationary or Markov from time 1 onwards. However, this does not invalidate the proof, since ρ is defined by *some* freezing rule, and the freezing rules in (1) are not restricted to be stationary or Markov.

For the purposes of the algorithm described in Section 7 we need to consider what happens when the set of stopping times $\{\tau > 0\}$ is modified by allowing the stopping set Θ_0 to depend on the process time t, and by imposing the restriction $\tau \leqslant M$, where M is a non-random integer. This new set of stopping times will be denoted by $\{0 < \tau \leqslant M\}$ and we define

$$\nu^M(D,x) = \sup_{0<\tau\leqslant M} \nu_\tau(D,x). \tag{5}$$

The lemma leads to the following corollaries.

Corollary 1. The supremum in (5) is attained by setting

$$\Theta_0(t) = \{y \in \Theta\colon \nu^{M-t}(D,y) < \nu^M(D,x)\}, \quad t = 1,2,\ldots,M-1.$$

Corollary 2. The right-hand side of (5) is unaltered if the stopping sets $\Theta_0(t)$ defining the stopping times τ are restricted to be of the form $\Theta_0(t) = \{y \in \Theta\colon \nu^{M-t}(D,y) < \mu\}$, $t = 1,2,\ldots,M-1$, for some non-random μ.

Proof. Define the bandit process D^* as follows. The state $y(t)$ of D^* at process time t is $(x(t),t)$. The rewards from D^* are identical to those from D up to process time M, after which they take very large negative values. It is easy to show that $\nu^{M-t}(D,x(t)) = \nu(D^*,y(t))$ for all $x(t) \in \Theta$ and for $0 \leqslant t < M$. Corollary 1 then follows by applying the lemma to D^*. Corollary 2 is an immediate consequence.

5. Improving and Deteriorating DAIs

In this section we describe two cases for which the definition of the DAI leads directly to an expression from which particular values may be determined in a straightforward manner. Consider first any bandit process D, an arbitrary state of which is denoted by x. Dropping D from the notation, we have

$$\nu(x) = \sup_{\tau>0} \nu_\tau(x) = \sup_{\tau>0} \frac{R_\tau(x)}{W_\tau(x)} = \sup_{\sigma\geqslant 0} \frac{R(x,1) + aE\{R_\sigma(x(1)) \mid x(0) = x\}}{1 + aE\{W_\sigma(x(1)) \mid x(0) = x\}}, \tag{6}$$

where τ and σ are stopping times, and τ is restricted to be positive.

Case 1 (*the deteriorating case*): $P\{\nu(x(1)) \leqslant \nu(x(0)) \mid x(0) = x\} = 1$

Since $\nu(x(1)) = \sup_{\sigma>0}\{R_\sigma(x(1))/W_\sigma(x(1))\}$ it follows immediately from (6) that $\nu(x) = R(x,1)$.

For Case 1 our conclusion, then, is particularly simple. The process for which $R(x,1)$ is largest at any particular time is the process which yields the largest immediate reward if it is continued, and the DAI theorem tells us that this is the process which should be continued. Thus the one-step look-ahead policy is optimal. Since Case 1 covers a situation in which the future prospects of gain from a process are bound to deteriorate when it is continued, such a conclusion is not unexpected.

The deteriorating case may be compared with the monotone case in the study of stopping problems, which is discussed by Chow *et al.* (1971). Here too, and for similar reasons, the solution is particularly simple.

It is easy to see that a sufficient condition for $P\{\nu(x(1)) \leqslant \nu(x(0)) \,|\, x(0) = x\} = 1$ is that $P\{R(x(t+1), 1) \leqslant R(x(t), 1) \,|\, x(t) = y\} = 1$, for all states y which may be reached from x in any number of steps.

Case 2 (*the improving case*): $P\{\nu(x(s)) \geqslant \nu(x(s-1)) \geqslant \ldots \geqslant \nu(x(1)) \geqslant \nu(x(0)) \,|\, x(0) = x\} = 1$, *for some non-random integer* s

From (6) it follows that for this case

$$\nu(x) = \sup_{\sigma > 0} \left[E\left\{ \sum_{t=0}^{s-1} a^t R(x(t), 1) \,|\, x(0) = x \right\} + a^s E\{R_\sigma(x(s)) \,|\, x(0) = x\} \right.$$
$$\left. \times [1 + a + \ldots + a^{s-1} + a^s E\{W_\sigma(x(s)) \,|\, x(0) = x\}]^{-1} \right],$$

which simplifies if the defining condition holds for all s and if we set $s = \infty$. This will, for example, be so if the defining condition holds for $s = 1$ and for all x.

Cases 1 and 2 are illustrated by the scheduling problem.

6. The Scheduling Problem

Let D be a bandit process such that $\Theta = \{C\} \cup \mathbf{Z}$, $P(\{C\} \,|\, C, 1) = 1$, $P(\{C\} \,|\, t, 1) = p(t)$, $P(\{t+1\} \,|\, t, 1) = 1 - p(t)$, $R(C, 1) = 0$ and $R(t, 1) = p(t)\, V$, $\forall\, t \geqslant 0$. A bandit process with these properties corresponds to one of the jobs in the scheduling problem described in Section 1. If the bandit process is in state C this signifies that the job has been completed. Thus unless the job has reached state C its state coincides with the process time if $x(0) = 0$. Also, it is true generally that $a^t R(x, 1)$ may be taken to be the *expected* reward if control 1 is applied at time t with the process in state x, and this device has been used here. It may be noted that, unlike the multi-armed bandit problem, the scheduling problem does not involve probability distributions with unknown parameters. However, it is a simple matter (see Gittins and Glazebrook, 1977) to extend the discussion which follows to include this possibility.

If $p(t)$ is a non-increasing function of t then D is a deteriorating bandit process, since the sufficient condition for Case 1 holds for all x. Thus with jobs of the above type the job to be continued at any time is one of those for which $p_i(t_i)\, V_i$ is largest, where i runs over the set of uncompleted jobs.

A job for which $p(t)$ is a non-decreasing function of t provides an example of a modification of Case 2. We now have

$$P\{\nu(x(s)) \geqslant \nu(x(s-1)) \geqslant \ldots \geqslant \nu(x(1)) \geqslant \nu(x(0)) \geqslant 0 \,|\, x(0) = x, x(s) \neq C\} = 1, \quad s \in \mathbf{Z},$$

and $\nu(C) = 0$. It thus follows from (6) that

$$\nu(x) = E\left\{ \sum_{t=0}^{\tau-1} a^t R(x(t), 1) \,|\, x(0) = x \right\} \Big/ E\{1 + a + \ldots + a^{\tau-1} \,|\, x(0) = x\}, \tag{7}$$

where $\tau = \min\{s\colon x(s) = C\}$. Equation (7) may be rewritten in the form

$$\nu(x) = V(1-a)\, E(a^{\tau-1} \,|\, x(0) = x) / \{1 - E(a^\tau \,|\, x(0) = x)\}.$$

For an arbitrary job, with no restriction on the function $p(t)$, it is easy to see that the τ for which the supremum in (6) is attained is no greater than the time taken to complete the job. For uncompleted jobs the state coincides with process time, so that the stopping set which defines τ must be reached at some non-random (and possibly infinite) time r. Thus τ is of the form $\min\{r, \min[s\colon x(s) = C]\}$. It follows that $\nu(C) = 0$, and

$$\nu(x) = \sup_{r > 0} \frac{V(1-a)\{Ea^{\rho-1} - P(\rho > r)\, E(a^{\rho-1} \,|\, \rho > r)\}}{1 - Ea^\rho + P(\rho > r)\{E(a^\rho \,|\, \rho > r) - a^r\}}, \tag{8}$$

where $\rho = \min\{s: x(s) = C\}$ and the expectations and probability are all conditioned by the event $x(0) = x \neq C$.

It is interesting to see what happens to a problem involving jobs of this type as a tends to one. The reward Va^t on completion of the job at time t may be expressed as

$$V\{1-t(1-a)\}+o(1-a).$$

Thus, if $1-a$ is small, the largest contribution to the reward which depends on t, and thereby on the policy for allocating effort to the job, is given by the term $-Vt(1-a)$. Not surprisingly, therefore, given a set of jobs for which the penalties for delays in completion are proportional to the extent of the delays, the DAI policies defined by letting a tend to one in (8), and setting V equal to the cost c of unit delay for the particular job, are optimal. The limit of (8) as a tends to one may be written

$$\nu(x) = \sup_{r>0}\left\{c\sum_{i=0}^{r-1} p(x+i)\Big/\sum_{i=0}^{r-1} P(\rho > x+i \mid x(0) = x)\right\}. \tag{9}$$

The optimality of the DAI policy based on this expression for the scheduling problem with penalties proportional to the delays was first demonstrated, using a different method, by Sevcik (1972), whose primary interest was the scheduling of jobs on a computer. Models of this type are also applicable to the planning of industrial chemical research (e.g. see Gittins, 1973). Nash (1973) has shown that the DAI policy remains optimal for the case with random arrivals. This result is an important contribution to the theory of priority queues, as is made clear by Simonovits (1973), and is perhaps the most striking consequence of the DAI theorem which has so far been obtained.

7. The Multi-armed Bandit Problem

Let D be a bandit process whose states are a class of probability distributions for a random variable θ defined on $[0, 1]$. Continuation of D at process time t is defined as observing the $(t+1)$st member of a sequence of independently distributed random variables $X_1, X_2, \ldots$, each of which is equal to 1 with probability θ and equal to 0 with probability $1-\theta$. If $x(t)$ has a continuous density $\pi(\theta)$, then $x(t+1)$ has a density proportional to $\pi(\theta)\,\theta^{X_{t+1}}(1-\theta)^{1-X_{t+1}}$, as follows from Bayes' theorem. If $X_{t-1} = 1$ a reward a^s accrues, where s is the time at which X_{t+1} is observed, and a zero reward if $X_{t+1} = 0$. As in Case 1, $a^s R(x, 1)$ is taken to be the *expected* reward which accrues at time s if D is then in state x and is continued. Thus

$$R(x, 1) = E(E_\theta X_{s+1}) = E(\theta) = \int_0^1 \theta\pi(\theta)\,d\theta. \tag{10}$$

Clearly the multi-armed bandit problem described in Section 1 amounts to finding an optimal policy for a simple family of alternative bandit processes of this type.

This problem owes its picturesque name to its resemblance to the situation facing a gambler with a choice between several one-armed bandits (or just one multi-armed bandit). It is an intriguing problem, on which a considerable number of papers have been written, recent examples being those by Wahrenberger *et al.* (1977) and Rodman (1978). This is probably because it is the simplest worthwhile problem in the sequential design of experiments. Its chief practical significance is in the context of clinical trials. Bellman (1956) gave the first Bayesian formulation and obtained some properties of the optimal policy and maximum expected reward for the case when there are two "arms" (i.e. bandit processes), one of them a standard process.

As in Section 1 we shall suppose that $x(0)$, and therefore $x(t)$, $t \in \mathbf{Z}^+$, is a beta distribution. As pointed out by Raiffa and Schlaifer (1961), this greatly simplifies any calculations, whilst the two parameters $\alpha(0)$ and $\beta(0)$ allow an arbitrary specification of the mean and variance

of the prior distribution, which for many purposes is quite adequate. Thus an arbitrary state of D may be represented by the corresponding parameter values (α, β).

Applying Corollary 2 of Section 4 (and using the notation of that section), we have, if α, β and N are non-negative integers with $N > \alpha+\beta$,

$$\nu^{N-\alpha-\beta}(\alpha,\beta) = \sup_{\mu} \frac{R(\alpha,\beta;1) + \sum_{m=1}^{N-\alpha-\beta-1} a^m \sum_{r=0}^{m} Q(r,\alpha,\beta,m,\mu)\, R(\alpha+r,\beta+m-r;1)}{1 + \sum_{m=1}^{N-\alpha-\beta-1} a^m \sum_{r=0}^{m} Q(r,\alpha,\beta,m,\mu)}. \tag{11}$$

Here

$$Q(r,\alpha,\beta,m,\mu) = P\{(\alpha(m),\beta(m)) = (\alpha+r,\beta+m-r) \cap \nu^{N-\alpha(t)-\beta(t)}(\alpha(t),\beta(t)) \geqslant \mu, \\ 1 \leqslant t \leqslant m \,|\, (\alpha(0),\beta(0)) = (\alpha,\beta)\}.$$

The expression (11) leads to the following algorithm for calculating the function $\nu^{N-\alpha-\beta}(\alpha,\beta)$ for a given value of N.

(1) If $\alpha+\beta = N-1$, the stopping time τ in the definition of $\nu^{N-\alpha-\beta}(\alpha,\beta)$ must be equal to one. Thus, using equation (10),

$$\nu^{N-\alpha-\beta}(\alpha,\beta) = R(\alpha,\beta;1) = (\alpha+1)/(\alpha+\beta+2). \tag{12}$$

(2) Equation (12) enables us to calculate the function $Q(r,\alpha,\beta,m,\mu)$ for $\alpha+\beta = N-2$, $m = 1$ and $r = 0, 1$. We have

$$Q(1,\alpha,\beta,1,\mu) = P\{X_1 = 1 \,|\, (\alpha(0),\beta(0)) = (\alpha,\beta)\} \times \begin{cases} 1 \text{ if } \nu^{N-\alpha-\beta-1}(\alpha+1,\beta) \geqslant \mu, \\ 0 \text{ if } \nu^{N-\alpha-\beta-1}(\alpha+1,\beta) < \mu, \end{cases}$$

$$Q(0,\alpha,\beta,1,\mu) = P\{X_1 = 0 \,|\, (\alpha(0),\beta(0)) = (\alpha,\beta)\} \times \begin{cases} 1 \text{ if } \nu^{N-\alpha-\beta-1}(\alpha,\beta+1) \geqslant \mu, \\ 0 \text{ if } \nu^{N-\alpha-\beta-1}(\alpha,\beta+1) < \mu, \end{cases}$$

and

$$P\{X_1 = 1 \,|\, (\alpha(0),\beta(0)) = (\alpha,\beta)\} = (\alpha+1)/(\alpha+\beta+2).$$

Values of $\nu^{N-\alpha-\beta}(\alpha,\beta)$ for $\alpha+\beta = N-2$ may now be calculated from equation (11) by substituting the above quantities and using equation (12).

(3) Now knowing the function $\nu^{N-\alpha-\beta}(\alpha,\beta)$ for $\alpha+\beta = N-1$ and $\alpha+\beta = N-2$, calculations similar to those described in stage (2) of the algorithm give values of $Q(r,\alpha,\beta,m,\mu)$ for $\alpha+\beta = N-3$, $m = 1, 2$ and $r = 0, 1, \ldots, m$. These may now be substituted into equation (11) to give values of $\nu^{N-\alpha-\beta}(\alpha,\beta)$ for $\alpha+\beta = N-3$, again using equation (12).

(4) Similar calculations give in turn values of $\nu^{N-\alpha-\beta}(\alpha,\beta)$ for $\alpha+\beta = N-4$, $N-5$, and so on, the final quantity to be calculated being $\nu^N(0,0)$.

Now clearly $\nu^T(\alpha,\beta)$ is increasing in T and tends to $\nu(\alpha,\beta)$ as T tends to infinity, so for any integer-valued α and β the above algorithm provides arbitrarily close approximations to $\nu(\alpha,\beta)$ by increasing the value of N. Some calculations along these lines have been carried out by Gittins and Jones (1979). The general form of the results is shown in Fig. 1.

The origin of the axes drawn on the figure is at the point $(\alpha,\beta) = (-1,-1)$. This means (see equation (12)) that $R(\alpha,\beta;1)$ is constant on straight lines through the origin. Using $R(\alpha,\beta;1)$ as an allocation index in place of $\nu(\alpha,\beta)$ is, of course, equivalent to adopting a one-step look-ahead rule. Each curve of constant $\nu(\alpha,\beta)$, or iso-DAI, is asymptotic to a straight line which is parallel to the corresponding line of constant $R(\alpha,\beta;1)$ as $\alpha+\beta$ tends to infinity. This is not surprising since large values of α and β mean that the probability is high that the unknown probability θ of success is close to $R(\alpha,\beta;1)$; if θ were actually known we should have a standard bandit process with the parameter, and therefore the DAI, equal to θ. For finite values of α and β, $\nu(\alpha,\beta) > R(\alpha,\beta;1)$, as is obvious from the definition of a DAI. This corresponds to the possibility that θ may be greater than $R(\alpha,\beta;1)$.

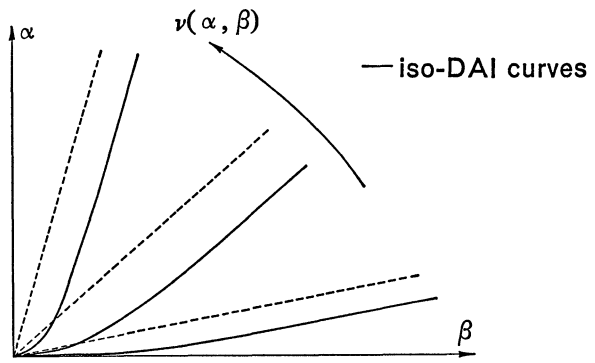

FIG. 1. The dynamic allocation index for the multi-armed bandit problem.

The extent to which the iso-DAIS curve away from their asymptotes for small values of α and β increases with the discounting parameter a. This is another way of saying that $\nu(\alpha, \beta)$ increases with a for any values of α and β. It reflects the fact that we may expect to find that optimal policies differ most from one-step look-ahead rules when what happens in the more distant future is comparable in importance with what happens in the immediate future, in other words when a is close to one.

8. A GENERAL METHOD FOR CALCULATING DAIs

The determination of DAIS for the scheduling problem and for the multi-armed bandit problem using the methods described in the previous two sections depends on certain special features of the bandit processes involved. A good general method when the problem does not simplify in some such fashion is to use the standard bandit processes as a calibration device.

Consider the simple family of alternative bandit processes $\{D, \lambda\}$ formed by an arbitrary bandit process D together with a standard bandit process with the parameter λ. Optimal policies for $\{D, \lambda\}$ are DAI policies, and therefore start by continuing D if $\nu(D) > \lambda$, and by continuing the standard process if $\nu(D) < \lambda$. If $\nu(D) = \lambda$, and only if this is so, an optimal policy may start in either of these ways. Our calibration procedure consists of finding a value of λ such that an optimal policy for $\{D, \lambda\}$ can start either by continuing D or by continuing the standard process. It then follows that $\nu(D) = \lambda$.

As shown by Blackwell (1965), the maximum total expected reward for any Markov decision process satisfies a dynamic programming functional equation. For the family $\{D, \lambda\}$ this equation may be written as

$$R(\{D, \lambda\}, x) = \max [\lambda/(1-a), R(x, 1) + aE_x R(\{D, \lambda\}, y)]. \tag{13}$$

Here $R(x, 1)$ is the reward resulting from continuing D when it is in state x at time zero. E_x denotes the expectation with respect to the state y of D at time one, given that D is in state x at time zero, when control 1 is applied. It may be noted that the standard bandit process has just one state, so that the state of D also defines the state of $\{D, \lambda\}$. Also for this reason the state of $\{D, \lambda\}$ does not change if the standard process is continued, and it follows that a deterministic stationary Markov policy must continue the standard process for all time after it has done so for one time unit. If this happens at time zero, the total expected reward is $\lambda/(1-a)$, the first term on the right-hand side of equation (13).

Blackwell also shows that, provided the maximum total expected reward is bounded over all initial states, equations of the form (13) may be solved either by one of the policy improvement algorithms available for the purpose, or by starting with an approximate function, substituting in the right-hand side and thus obtaining a second approximation, and so on. From the DAI theorem it follows that, for any x, $\nu(D,x)$ is the unique value of λ for which the maximum on the right-hand side of equation (13) occurs both for the first and second terms in square brackets. Thus $\nu(D,x)$ may be determined by solving (13) for a succession of values of λ in the neighbourhood of $\nu(D,x)$.

At this point the reader may wonder what is the point of calculating $\nu(D,x)$ in this way, since for any family $\mathscr{F}$ of alternative bandit processes the optimal policy and maximum total expected reward may always be calculated directly from the equation,

$$R(\mathscr{F},x) = \max_{u\in\Omega(x)} \{R(\mathscr{F},x,u)+aE_{x,u}R(\mathscr{F},y)\}, \tag{14}$$

which is rather simpler than (13). Here $R(\mathscr{F},x,u)$ is the reward resulting from applying u to the family $\mathscr{F}$ in state x at time zero. $E_{x,u}$ denotes the expectation with respect to the state y of $\mathscr{F}$ at time one, given that $\mathscr{F}$ is in state x at time zero, when control u is applied. The answer is that the state-space for $\mathscr{F}$ is the product of the state-spaces for its constituent bandit processes. In general this means that the states x for which (13) is solved are of lower dimensionality than those involved in (14), and this frequently brings an otherwise intractable problem within the bounds of computational feasiblity, as illustrated by the example described in the next Section.

9. A Multi-armed Bandit with Normally Distributed Rewards

Let D be a bandit process whose states are the set of $N(\xi,m^{-1})$ (i.e. normal with mean ξ and variance m^{-1}) distributions for a random variable θ. Continuation of D at process time t is defined as observing the $(t+1)$st member of a sequence of independently distributed $N(\theta,\sigma^2)$ random variables $X_1, X_2, \ldots$, where σ^2 is known. Changes of state occur according to Bayes' theorem, so that (see Raiffa and Schlaifer, 1961)

$$\xi(t) = \frac{m(0)\,\xi(0)+t\bar{X}_t\sigma^{-2}}{m(0)+t\sigma^{-2}} \quad \text{and} \quad m(t) = m(0)+t\sigma^{-2},$$

where $\xi(t)$ and $m(t)$ are the parameters which define the state of the bandit process at process time t and $\bar{X}_t = t^{-1}(X_1+X_2+\ldots+X_t)$. Thus, as for the ordinary multi-armed bandit, we have chosen a family of distributions for θ which is closed under sampling, a restriction which is virtually essential in the ensuing calculations.

The reward at the $(t+1)$st observation if this occurs at time s is $a^s X_{t+1}$. As before, $a^s R(x,1)$ is the *expected* reward if D is continued in state x at time s. Thus if $x=(\xi,m)$ then $R(x,1)=\xi$.

A simple family of alternative bandit processes of this type forms a natural extension of the multi-armed bandit problem. A model of this type might well be appropriate in clinical trials if a number of treatments are to be compared whose object is to control some variable which is measured on a continuous scale.

It is convenient to include in the notation the dependence on σ of the various quantities which arise. Thus, for example, $\nu(\xi,m,\sigma)$ denotes the DAI for D in the state (ξ,m).

It may be shown that

$$\nu(\xi,m,\sigma) = \xi+\sigma\nu(0,m,1). \tag{15}$$

The proof is in two stages, proceeding roughly as follows. Firstly, if a constant is added to any set of numbers then the effect is to add the same constant to any weighted average of those numbers. If follows that $\nu(\xi,m,\sigma) = \xi+\nu(0,m,\sigma)$. Secondly, if any set of numbers is multiplied by a constant then the effect is to multiply any weighted average by the same constant, so that $\nu(0,m,\sigma) = \sigma\nu(0,m,1)$.

The most convenient method of calculating the DAI function in this case is to combine equation (15) with the procedure described in Section 8. Equation (13) becomes

$$R(\lambda, \xi, m, \sigma) = \max\left[\lambda/(1-a); \int_{-\infty}^{\infty}\left\{y + aR\left(\lambda, \frac{m\xi + y\sigma^{-2}}{m+\sigma^{-2}}, m+\sigma^{-2}, \sigma\right)\right\} dG(y)\right]. \tag{16}$$

Here y denotes the value of the next observation when D is in the state (ξ, m), and G denotes its distribution function, which may be shown to be $N(\xi, m^{-1}+\sigma^2)$.

Now in view of equation (15) we need only solve equation (16) for $\xi = 0$ and $\sigma = 1$. Also, arguing along similar lines to the first part of the proof of (15), we have

$$R(\lambda, \xi, m, \sigma) = \xi/(1-a) + R(\lambda - \xi, 0, m, \sigma).$$

Thus to determine the function $\nu(\xi, m, \sigma)$ we need to solve the equation

$$R(\lambda, 0, m, 1) = \max\left\{\lambda/(1-a); a\int_{-\infty}^{\infty} R\left(\lambda - \frac{y}{m+1}, 0, m+1, 1\right) dG(y)\right\}, \tag{17}$$

where G is $N(0, m^{-1}+1)$. This may be done by substituting a reasonable approximation to the function $R(\cdot, 0, M, 1)$, for a moderately large value of M, into the right hand side of (17), setting $m+1 = M$, and hence finding an approximation to $R(\cdot, 0, M-1, 1)$, then substituting this in the right-hand side of (17), and so on.

It should be noted that these iterations involve functions of a single real variable. Any calculations based on equation (14) involve iterations with functions of $2n$ real variables, and are quite impracticable for n greater than 2.

By choosing M to be sufficiently large, arbitrarily close approximations to the DAI function may be obtained. Moreover, a large value of M corresponds to a high probability that if D is in the state (ξ, M) then θ is close to ξ. This means that D is hardly distinguishable from a standard bandit process with the parameter ξ, leading to an obvious close approximation to the function $R(\cdot, 0, M, 1)$.

Calculations along these lines have been carried out and will be reported separately. The function $\nu(0, m, 1)$ turns out to have the general form shown in Fig. 2. This is because a bandit process in the state $(0, m)$ with m large is very similar to a standard bandit process with the parameter zero, whilst the probability that θ is substantially greater than zero increases as m decreases.

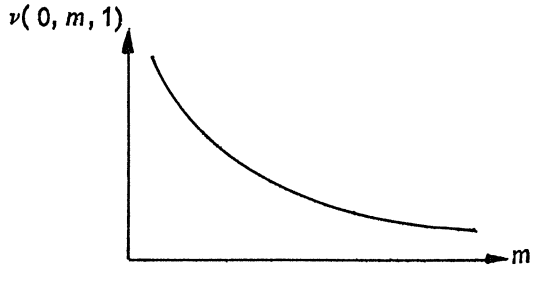

FIG. 2. The dynamic allocation index for the multi-armed bandit with normally distributed rewards.

Robbins and Siegmund (1974) have proposed a heuristic allocation rule for sequential probability ratio tests between two treatments, which is designed to cut down the number of tests with the inferior treatment. It would be interesting to compare the characteristics of their rule, which is designed for the case of normal distributions with known variance, with a DAI policy using the function $\nu(\xi, m, \sigma)$.

10. A Class of Search Problems

As a first example, consider the modification of the multi-armed bandit problem for which the total reward arises entirely from the first successful pull, and is equal to a^s if this is the sth pull to be made. The Markov decision process formed in this way might be a suitable model for a situation in which a number of different populations are being searched with the aim of finding as soon as possible an individual with some rare characteristic, at which point the search stops. However, at first sight the problem is not one which can be modelled by a simple family of alternative bandit processes, since once a success has been obtained on a pull of one arm no further non-zero rewards may be obtained from any of the arms. This difficulty may be overcome as follows.

Consider the bandit process described in Section 7 with the following modifications. If $X_{t+1} = 1$ a zero reward accrues. If $X_{t+1} = 0$ a zero reward accrues if at least one of $X_1, X_2, \ldots, X_t$ is equal to one; otherwise a reward equal to $-a^s$ accrues, where s is the time at which X_{t+1} is observed. The state-space may be defined by adding to the state-space for an arm of a multi-armed bandit a state C, indicating that a success has occurred.

For a bandit process of this type the DAI is negative until the first success occurs, and thereafter equal to zero. Consequently an optimal policy for a simple family $\mathscr{F}$ of alternative bandit processes of this type will always select for continuation a bandit process in state C if there is one available. If none of the bandit processes in $\mathscr{F}$ is initially in state C and the first success occurs at the sth trial, then all subsequent rewards are equal to zero and the total reward is $(a^s-1)/(1-a)$. An optimal policy for $\mathscr{F}$ is therefore one which maximizes the expectation of a^s, and is an optimal policy for our modified multi-armed bandit problem. Thus the optimal policies for our search problem are those given by the DAI theorem for the corresponding $\mathscr{F}$.

It may be shown, and indeed it is fairly obvious, that $\nu(x(1)) < \nu(x(0))$ unless $x(1) = C$. It follows, using an argument similar to those used in Section 5 and assuming that $x(0)$ is a beta distribution as in Section 7, that $\nu_\tau(\alpha, \beta) = \nu(\alpha, \beta)$ if

$$\tau = \begin{cases} 1 & \text{if } X_1 = 0, \\ \infty & \text{if } X_1 = 1. \end{cases}$$

Thus

$$\nu(\alpha, \beta) = \frac{-1 \times P\{X_1 = 0 \,|\, x(0) = (\alpha, \beta)\}}{1 + a(1-a)^{-1} P\{X_1 = 1 \,|\, x(0) = (\alpha, \beta)\}}.$$

This is a strictly increasing function of $P\{X_1 = 1 \,|\, x(0) = (\alpha, \beta)\}$. It is therefore optimal to use this probability, which is equal to $(\alpha+1)/(\alpha+\beta+2)$, as a DAI. This means that a one-step look-ahead policy is optimal for our search problem.

The bandit process described in Section 9 may also be modified so as to model a search problem. Suppose that if X_{t+1} belongs to some measurable subset B of the real line then a zero reward accrues; and if $X_{t+1} \notin B$ then a zero reward accrues if at least one of $(X_1, X_2, \ldots, X_t) \in B$, and otherwise a reward of $-a^s$ accrues, where s is the time at which X_{t+1} is observed. Again we add a state C, indicating that an observation belonging to B has been made, to the state space for the multi-armed bandit with normally distributed rewards. This time it turns out that a one-step look-ahead policy is not in general optimal.

A simple family of alternative bandit processes of this type might be a suitable model if the aim is to find as soon as possible an individual belonging to B from any one of a number of populations. The DAI function may be calculated as described in Section 8. Some results for the case $B = [0, \infty)$ are described by Jones (1970).

Clearly a range of different search problems (and corresponding multi-armed bandit problems) may be modelled by considering distributions other than 0–1 and normal for the observations $X_1, X_2, \ldots$. For example, Gittins and Jones (1974b) have prepared a set of

tables based on negative exponential distributions with an added probability atom at zero. These are designed for use in the screening of chemicals in new-product chemical research. Glazebrook (1978b) considers a multi-armed bandit problem in which several different outcomes, rather than just two, are possible at each trial.

11. Possible Further Developments

The examples which have been described show that there is considerable scope for applying the notions of forwards induction policies and dynamic allocation indices, using the theorems of Section 3. However, at this stage the story is incomplete. Later instalments may touch on the following points.

(i) There may well be types of Markov decision process other than families of alternative bandit processes for which forwards induction policies are optimal. A simple characterization of the class of Markov decision processes with this property would be useful, since in many cases forwards induction policies are relatively easy to determine.

One example of a Markov decision process for which forwards induction policies are known to be optimal, and which is not a family of alternative bandit processes, is described by Black (1965). This is a search problem for which an object is hidden in one of a number of boxes. For each box there is a detection probability on searching, if it contains the object, and a cost. The aim is to find the object at minimum cost.

(ii) At present one is much more aware of the above-mentioned scope for practical applications than of such applications actually being made. We hope this situation will change.

Acknowledgements

I am very grateful to Mr A. G. Baker of Unilever Research, Port Sunlight, for his encouragement over several years, and for naming the dynamic allocation index. I should also like to thank Drs K. D. Glazebrook, D. M. Jones and P. Nash for many enjoyable and stimulating discussions, and the referees, whose comments on earlier drafts have led to a much improved paper.

References

Bellman, R. E. (1956). A problem in the sequential design of experiments. *Sankhyā A*, **16**, 221–229.

—— (1957). *Dynamic Programming*. Princeton: Princeton University Press.

Black, W. L. (1965). Discrete sequential search. *Information and Control*, **8**, 159–162.

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

Chow, Y. S., Robbins, H. and Siegmund, S. (1971). *Great Expectations, the Theory of Optimal Stopping*. New York: Houghton Mifflin.

Davies, D. G. S. (1970). Research planning diagrams. *R and D Management*, **1**, 22–29.

Gittins, J. C. (1973). How many eggs in a basket? *R and D Management*, **3**, 73–81.

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

—— (1979). Two theorems on bandit processes. Submitted for publication.

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. *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.

—— (1979). A dynamic allocation index for the discounted multi-armed bandit problem. *Biometrika* (to appear).

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.

—— (1978 b). 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.

RAIFFA, H. and SCHLAIFER, R. (1961). *Applied Statistical Decision Theory.* Boston: Harvard Business School.

ROBBINS, H. and SIEGMUND, D. O. (1974). Sequential tests involving two populations. *J. Amer. Statist. Ass.*, **69**, 132–139.

RODMAN, L. (1978). On the many-armed bandit problem. *Ann. Prob.*, **6**, 491–498.

ROSS, S. M. (1970). *Applied Probability Models with Optimisation Applications.* San Francisco: Holden-Day.

SEVCIK, K. C. (1972). The use of service-time distributions in scheduling. Technical Report CSRG–14, University of Toronto.

SIMONOVITS, A. (1973). Direct comparison of different priority queueing disciplines. *Studia Scientiarum Mathematicarum Hungarica*, **8**, 225–243.

WAHRENBERGER, D. L., ANTLE, C. E. and KLIMKO, L. A. (1977). Bayesian rules for the two-armed bandit problem. *Biometrika*, **64**, 172–174.

DISCUSSION OF DR GITTINS' PAPER

Professor J. A. BATHER (University of Sussex): I shall restrict my comments to the multi-armed bandit problem described in Sections 1 and 7 of Dr Gittins' paper. He remarks that "its chief practical significance is in the context of clinical trials". This is true, but I would like to spend a few minutes considering why, after many years of study, there has been so little effect on the conduct of sequential medical trials.

In the notation of Section 7, $\theta_1, \theta_2, \ldots, \theta_n$ are the unknown probabilities of success in n different sequences of Bernoulli trials or, alternatively, we can think of a single sequence of patients and n possible treatments for any one of them. The problem is to find a rule for allocating a treatment to each patient so that the number of successful treatments is maximized, in some sense. Suppose that, after a total of t trials, we have observed r_i successes in m_i trials with treatment i. The proportion of successes achieved so far is r/t, where $r = \Sigma r_i$ and $t = \Sigma m_i$, summing over i from 1 to n. We need a rule which tells us which treatment should be given to the next patient in the sequence.

The optimization problem is not well defined without further assumptions, which Dr Gittens expresses in the choice of a prior distribution and a discount factor $a<1$. Even then, there are genuine difficulties: his result that the optimal policy can always be expressed in terms of dynamic allocation indices is a very impressive reduction of the problem, but the procedure described in Section 7 is still very complicated (see also Fabius and Van Zwet, 1970). I would like to ask Dr Gittins about the sensitivity of the optimal policy to changes in the prior distribution and in the discount factor, particularly as $a \uparrow 1$ which is the most important special case. It seems to me that we might do well to consider something less than exact optimality; I think the best may be the enemy of the good.

I will conclude with a suggestion which I hope is constructive. Consider a family of sequential decision procedures depending on a randomized allocation index. The randomization is useful even though it is not a direct consequence of any particular optimality criterion. Let $\{\lambda_m\}$ be a sequence of positive numbers such that $\lambda_m \to 0$ as $m \to \infty$ and let X_{it}, $i = 1, 2, \ldots, n$, $t = 1, 2, \ldots$, be i.i.d non-negative random variables with a distribution which is unbounded. Given the record of successes and failures in the first t trials, the next treatment is chosen according to

$$\max_{1 \leqslant i \leqslant n} \{r_i/m_i + \lambda_{m_i} X_{it}\}.$$

In other words, the next treatment must be one of the current "favourites" according to an index made up of the observed proportion of successes and a positive bias. The idea is that the random terms will tend to favour those treatments which have so far had relatively few trials.

Any such decision procedure is asymptotically optimal in the following sense. Suppose that $\theta_1 > \theta_2 \geqslant \theta_3 \geqslant \ldots \geqslant \theta_n$. Then the random variables $r_i(t)$ and $m_i(t)$ have the property that, with probability 1, $m_1(t)/t \to 1$ and $\Sigma r_i(t)/t \to \theta_1$ as $t \to \infty$, so the observed proportion of successes in all the trials converges to $\max(\theta_1, \theta_2, \ldots, \theta_n)$. This result is a consequence of the strong law of large numbers. As Robbins pointed out (1952), it is easy to construct decision procedures which are

asymptotically optimal, but not all of them are "good'. I claim that some of the randomized allocation procedures obtained by defining $\lambda_m = 1/m$ perform well for any values of the unknown probabilities and over short as well as long sequences of trials. However, the evidence for this is not by any means complete.

Dr Gittins has certainly provided us with plenty of food for thought and I hope my introduction of a rival index will not confuse matters nor delay still further the time when the theory of sequential decisions is translated into practice. I have much pleasure in proposing a vote of thanks.

Professor P. WHITTLE (Statistical Laboratory, Cambridge): We should recognize the magnitude of Dr Gittins' achievement. He has taken a classic and difficult problem, that of the multi-armed bandit, and essentially solved it by reducing it to the case of comparison of a single arm with a standard arm. In this paper he brings a number of further insights. Giving words their everyday rather than their technical usage, I would say that my admiration for this piece of work is unbounded, meaning, of course, very great.

Despite the fact that Dr Gittins proved his basic results some seven years ago, the magnitude of his advance has not been generally recognized and I hope that one result of tonight's meeting will be that the strength of his contribution, its nature and its significance will be apparent to all.

As I said, the problem is a classic one; it was formulated during the war, and efforts to solve it so sapped the energies and minds of Allied analysts that the suggestion was made that the problem be dropped over Germany, as the ultimate instrument of intellectual sabotage. In the event, it seems to have landed on Cardiff Arms Park. And there is justice now, for if a Welsh Rugby pack scrumming down is not a multi-armed bandit, then what is?

And the name of DAI seems then also well chosen. But what is surprising is the hedonistic origin of the DAI concept, and of the forward induction principle. To someone brought up on the conventional backwards induction principle, like myself, the notion of a terminal reward or a terminal cost is an ingrained one, expressing as it does the consequences in the hereafter of one's actions in the present. But DAI has no consciousness of the hereafter, he behaves literally like there was no tomorrow, grabs what he can while it lasts, and then opts out. It is still somewhat unclear to me how it is that an optimal strategy can ignore the future to this degree; it must be, as Dr Gittins says, because the bandit formulation allows one to postpone certain courses of action without prejudice.

Dr Gittins has given the interpretation of Section 8 in other papers (i.e. the calculation of DAI by calibration against a standard arm) but the interpretation of Sections 2 and 3 is new to me. This is the characterization of DAI as the maximal reward rate up to some stopping time. This is reminiscent of the characterization of average cost optimality by the maximization of reward rate up to a stopping time defined by recurrence to the initial state. However, again there is a contrast: this latter criterion shows the awareness of moral principles, of which DAI is so lamentably negligent, in that it observes the precept "leave things as you found them".

I really have no contribution of substance to make. Obviously there are many questions one could ask, and generalizations one could suggest, but it seems most appropriate at the moment to congratulate Dr Gittins warmly on having developed a powerful optimization technique of great practical and conceptual significance.

[*A further comment added in writing after the meeting*]: An index result which I might mention concerns sequential choice of experiment (types of experiment being indexed by u) for optimal discrimination between two simple hypotheses. The criterion for choice of u given in Theorem 4 of Whittle (1965) can be more simply expressed: choose the u for which $\gamma_u P_1 + \delta_u P_2$ is minimal. Here P_1 and P_2 are the probabilities of the two hypotheses conditional on current information, and γ_u, δ_u are the quantities defined in the paper quoted; essentially ratios of cost of experiment to Kullback–Liebler number for experiment u. The rule is optimal to within a no-overshoot approximation—I should be interested to know if it could also be derived by the methods of Dr Gittins' paper.

The vote of thanks was passed by acclamation.

Mr D. G. S. DAVIES: I should like to speak from the standpoint of a research planning man rather than a statistician. I should also like to congratulate Dr Gittins and to draw attention to two features of his work which I think are important.

First, the idea of a forwards induction policy is important. I know that many decision problems can be solved—perhaps all of them—by a backwards induction policy but, as Dr Gittins has pointed

out, this is often prohibitively difficult to calculate. In the world of research it is extremely difficult to get research workers to come up with data, and particularly to place any credence in long and involved computer calculations based upon the data which they have produced. If we can develop figures of merit and indices which are soundly based, and which can be used for allocation of effort in a forwards sequential manner, and if it is simply a matter of looking these things up in the tables, provided that the model is appropriate, I am sure that this is something which the research worker at the bench would be prepared to contemplate. However, if it is a matter of doing a large-scale modelling exercise on his project, then sending it away for computer analysis, he is much more reluctant about it—I speak from bitter experience.

Secondly, Dr Gittins has emphasized the distinction between the DAI and the probability of success for the different routes. If we take the very simple model of the bandit, basically all we do is to carry out trials. If they succeed, that is fine; if they do not succeed, we do another trial. Dr Gittins has emphasized that we are gaining information as we do the trials, which gives us a potential way of re-evaluating which route to take, based on the way the trials are done. In the rather restricted range of applications in research where the DAI can be applied as it stands, information is gained simply by gaining an enhanced view of the frequency of occurrence of successes in any chosen route. However, this is only an example of a more general phenomenon, that in research generally there is always a conflict between going for immediate exploitation and going for information.

Very often either we can do a trial straightaway, in which case we may succeed immediately, or we can do some background work instead which we hope will give a greater chance of success when the trial finally is made. This is the conflict—also mentioned by Professor Whittle—and there is a contribution here in the DAI in which some of these considerations are incorporated into the index itself.

One *caveat* is that this is a very limited model, with limited application in research and development. In research and development we like to projectize our work—by "project", I mean a piece of work such that we can tell when it is finished. This particular method is applicable to a lifetime's work where we are continually doing trials—in the expectation, it is true, that they will come to fruition. But, as Dr Gittins said, it is a method with an unlimited time horizon. We like to be able to set finite time horizons in research and development. We hope that there is a learning curve superimposed on the work that is going on, so that we are not simply pulling the arm of the bandit all the time but also modifying that bandit as it goes along. I feel sure that this concept can be incorporated, but at present I am not absolutely certain how to do it.

I should like to hope that we can go further and obtain more indices of this kind that are applicable in a forward induction sense—let us not worry too much about them being optimal because that does not matter as long as they are useful. There are many precedents for this. For example, if we are scheduling a critical path network under resource constraints, this cannot be done optimally because we are up against completely prohibitive combinatorial problems if any non-trivial plan is attempted, if we try to do it optimally. We can, however, still develop useful heuristic rules which will take us forward in a powerful way.

Professor B. FRISTEDT (University of Liverpool): A big assumption is that the discount sequence, denoted by $(a^t: t = 0, 1, \ldots)$ by Dr Gittins, is geometric. That one wants there to be stationary policies that are optimal is not the only reason for this assumption. As Dr Gittins (1975) has indicated, without some such assumption the principle is not valid that multi-armed bandit problems may be solved by comparing each bandit to a standard bandit.

It is not clear that arbitrarily good approximations of $R(\lambda, 0, 1, 1)$ can be obtained via equation (17). Conceivably, if M is chosen so large that $R(\cdot, 0, \infty, 1)$ is a good approximation of $R(\cdot, 0, M, 1)$, then the small initial error may grow through $M-1$ iterations into a substantial error.

Suppose, in Section 2, one defines $E\Sigma a^t R(x(t), u(t))$ to equal $-\infty$ when according to the usual conventions it does not exist, even as $+\infty$ or $-\infty$. Does Blackwell's Theorem then hold with no assumptions, other than measurability, on R? I believe it does.

Equation (11) does not depend on θ having a beta distribution, since an arbitrary state that may occur can, for any initial distribution with or without a density, be expressed in terms of the numbers α (successes) and β (failures).

In many situations I think that the only good alternative to a Bayesian approach is a minimax approach involving a risk function. See Fabius and van Zwet (1970). In case one feels compelled

to avoid a Bayesian outlook I think it is unrealistic to do so by regarding a first certain number of trials as merely experimental and the remaining trials as having no aspect of the experimental in them. Real-world problems do arise in which experiment, decisions, and acts based on those decisions are inherently interwoven.

Mr A. G. BAKER (Unilever Research Laboratory, Wirral): Arising from the discussions in 1966 at the OR Conference in Edinburgh, may I add my thanks and congratulations to Dr Gittins for the progress made by him and his colleagues since then.

I should like, though, to bring out the implications of this work, as I see them, to a practising statistician—which is slightly related to Mr Davies' comments. There are two ways in which this work may be used: first, in the formal mathematical sense. For that, we would always be dependent on the theory being developed.

Secondly, there are other aspects of this work which a practising statistician can already use. He can use the arguments, and the mental approach suggested by Dr Gittins' work in his debate with research colleagues on how to tackle a programme of work. This is important; the fact that there are theoretical justifications for looking at how to proceed from the approach of the theorem on DAI, in particular the concept that it sometimes pays to buy information. Mr Davies referred to this as "background work", which is not a term I would use because it really is buying information, whereas background work is more a matter of basic research.

Those two points are the ones I should like to stress. Dr Gittins' work has given the practising statistician a basis for arguing on buying information, and the importance of doing so and, secondly, the importance of proceeding by using the DAI theorem.

Dr F. P. KELLY (University of Cambridge): Today's paper reviews an extremely important advance in the theory of Markov decision processes whose ramifications are widespread and still not fully explored. To illustrate this I shall discuss two relatively old problems in the field where the DAI theorem can be used to extend the best known results, recently obtained by Kadane and Simon (1977). The first is the search problem referred to by Dr Gittins in the final section of his paper, which can be described as follows. An object is hidden in one of n boxes. Initially the probability that the object is in box i is $P(i)$. The jth look in box i costs $c(i,j)$ and detects the object, given that it is in the box, with probability $d(i,j)$. A policy is an infinite sequence $b_1\, b_2 \ldots$, where b_t is the box to be looked in at time t if the object has not been found before then, and the aim is to minimize the expected cost incurred until the object is found. I shall deal first with the case $c(i,j) = 1$, where the aim is to minimize the expected time till the object is found. Consider the related discounted decision process in which no costs are incurred, a reward a^t is obtained if the object is found at time t, and the searcher is not told whether or not he has yet found the object. A policy is again an infinite sequence $b_1\, b_2 \ldots$. If this policy requires that at time t box i be looked in for the jth time, then the expected reward at time t is $a^t R(i,j)$, where

$$R(i,j) = P(i)\Big\{\prod_{k=1}^{j-1}(1-d(i,k))\Big\}\, d(i,j),$$

the unconditional probability the object is found on the jth look in box i. The discounted decision process is thus a family of alternative bandit processes. Let T be the time at which the object is found. Provided ET is finite

$$E(a^T) = 1-(1-a)\, ET + o(1-a),$$

and the policy minimizing ET can be deduced from the optimal policy for the discounted decision process. The original problem in which the $c(i,j)$ are not all equal can also be recast as a family of alternative bandit processes provided $\Sigma\, c(i,j)$ diverges for each i (summing over j from 1 to ∞); we just let $c(i,j)$ be the time it takes to look in box i for the jth time. The conclusion is that if $\nu(i) = \max_{t>0}\{\Sigma\, R(i,j)/\Sigma\, c(i,j)\}$ (where the summations are over j from 1 to t) then the optimal policy for the original problem begins by looking in that box i for which $\nu(i)$ is a maximum.

The second problem I shall discuss is the gold-mining problem first formulated by Bellman (1957). A man owns n gold mines and a delicate gold-mining machine. Each day the man must assign the machine to one of his mines. When the machine is assigned to mine i for the jth time there is a probability $p(i,j)$ that it extracts an amount of gold $r(i,j)$ and remains in working order, and a probability $1-p(i,j)$ that it extracts no gold and breaks down irreparably. The man's aim is to

maximize the expected amount of gold extracted before the machine breaks down. Let $s(i,j) = -\log p(i,j)$, and interpret $s(i,j)$ as the "time" it takes to work mine i on the jth occasion the machine is assigned to it. With respect to this standard time scale the machine remains in working order for an exponentially distributed period independent of the policy adopted, provided $\Pi_{j=1}^{\infty} p(i,j) = 0$ for each i. The man's problem thus corresponds to the related decision process in which the machine works for ever, but an amount of gold $r(i,j)$ extracted at standard time s is worth $e^{-s} r(i,j)$. This decision process is a family of alternative bandit processes, and so the optimal policy begins by looking in that mine i for which

$$\nu(i) = \sup_{t>0}\left[\sum_{j=1}^{t} r(i,j) \prod_{k=1}^{j} p(i,k) \Big/ \left\{1 - \prod_{j=1}^{t} p(i,j)\right\}\right]$$

is a maximum.

The results just described have been established using a different method by Kadane and Simon (1977), who also consider the problems under arbitrary precedence constraints. Observe though that both problems are essentially deterministic: the optimal policy does not have to adapt to information becoming available with time. The advantage of formulating the problems as families of alternative bandit processes is that this allows the results to be generalized to the case where the characteristics of box or mine i are not certain but have probability distributions which alter as box or mine i is investigated. As a simple example suppose that in the search problem the jth look in box i is informative with probability $D(i,j)$ and uninformative otherwise. An informative look determines whether or not the box contains the object, and an uninformative look yields no indication either way. Put more precisely this is equivalent to the assumption that the detection probabilities are independent Bernoulli random variables with $E\{d(i,j)\} = D(i,j)$, and that $d(i,j)$ becomes known after the jth look in box i. If

$$\nu(i) = P(i) \sup_{t>0}\left[\left[\sum_{j=1}^{t}\left\{\prod_{k=1}^{j-1}(1 - D(i,k))\right\} D(i,j)\right] \Big/ \left[\sum_{j=1}^{t}\left\{\prod_{k=1}^{j-1}(1 - D(i,k))\right\} c(i,j)\right]\right],$$

then the optimal policy begins by looking in that box i for which $\nu(i)$ is a maximum.

Dr D. M. ROBERTS (Ministry of Defence): My first comment on Dr Gittins' paper concerns the practical significance of the concept of a DAI. One area in which I have recently been looking at this is new product chemical research. Specifically, one is confronted with a number of research projects all competing for a limited amount of effort. The way in which each project is characterized tends to be complex. For in order to be realistic, account must be taken of such factors as the way in which the effectiveness of research effort varies with time, the chances of success as a function of useful work done, as well as various financial parameters. Thus a casual look at the possibilities gives little indication of where effort should be applied and at what levels.

However, it is possible to write a computer program which does two things. First, for any planned allocation, it shows the expected profitability of such an allocation. And, second, for each project, it calculates the DAI. A comparison of indices suggests ways in which effort might profitably be reallocated between projects, either as a modification of the initial allocation or, since the indices are functions of time, at an appropriate time within the forecast period.

I have just completed the development of such a program and runs carried out so far tend to indicate that, in spite of the complexity of detail surrounding each project (which means that the DAI Theorem is not directly applicable), the DAI provides us with an effective single measure for comparing projects.

My second observation on Dr Gittins' paper concerns his reference to the search problem where an object is hidden with known occupation probability distribution in one of a number of boxes. It has been shown that, to minimize the expected cost of the search, one should look in the box where the product of two terms—the probability of the object being there and the detection probability—divided by the cost is greatest. Although this principle is generally demonstrated using a dynamic programming approach, the optimal strategy is actually a forwards induction policy, and it is interesting to note that Ross (1970) is able to derive its form solely by considering two-step look-ahead policies. Inevitably therefore, one is left wondering whether the Forwards Induction Theorem can be extended to cover this situation.

Dr K. D. GLAZEBROOK (Newcastle University): I should like to put on record my thanks to Dr Gittins, not only for his interesting paper but also for being an immensely helpful and stimulating

supervisor and colleague. I feel, too, that after some years of familiarity with these results one is inclined to be blasé about them and forget how demanding these problems have been to solve. Perhaps I could make just three points:

(i) As Dr Gittins indicated, the policy which maximizes the total expected reward earned by a family of N alternative bandit processes during $[0, M]$, M fixed, is not in general a forwards induction policy. Suppose, though, that we consider the problem of maximizing the expected reward earned during $[0, \tau]$, τ an integer-valued stopping time, and ask for what τ is there a forwards induction policy which is optimal? Two important examples where this is the case are:

$$\tau = \inf_{t \geqslant 0} \{t;\ x_i(t) \in C_i,\ i = 1, \ldots, N\} \tag{1}$$

and

$$\tau = \inf_{t \geqslant 0} \{t;\ x_i(t) \in C_i \text{ for some } i\}, \tag{2}$$

where $x_i(t)$ is the state of bandit process i at time t and C_i is some subset of the state space of bandit process i. The scheduling problem discussed by Dr Gittins in Section 6 is an example of (1) and the search problem in Section 10 an example of (2).

(ii) We might want to make stopping part of our decision structure; this could well be so in problems relating to research planning and clinical trials. We could model this by having a choice of $2N$ actions at each decision-epoch instead of N as previously. These actions would be "continue bandit process i", $i = 1, \ldots, N$, and "stop and decide in favour of bandit process i", $i = 1, \ldots, N$. I have obtained some optimal policies for such problems as these (Glazebrook, 1979).

(iii) Many of the continuous-time analogues of the discrete-time decision processes discussed here will be controlled jump processes with the discounted cost criterion. Suppose that such a process is in state i at time 0, is subject to control u until its first transition, and is subject to an optimal control (if any such exists) thereafter. Let $R[i, u]$ be the expected return from such a policy and let V_α be the optimal return function under discount rate $\alpha > 0$. Under appropriate conditions we have that

$$V_\alpha(i) = \inf_u \{R[i, u]\}, \tag{3}$$

the infimum being over all admissible controls u. For a wide range of decision problems in research planning, stochastic scheduling and queueing (and indeed many continuous-time analogues of the problems discussed today), the optimal control problem stated in (3) looks very similar to a problem solved by Nash and Gittins (1977). Indeed so much so that I feel it may well be worthwhile defining a class of controlled jump processes which reflect the rather strange property that they may be solved by the techniques discussed there.

Dr M. A. H. DEMPSTER (Balliol College, University of Oxford): I should like to make a few brief remarks concerning an important area of practical application—scheduling problems in a stochastic environment. As pointed out elsewhere by Dr Gittins and his associates stochastic scheduling problems arise in computer scheduling, reliability and R and D management as well as in factory scheduling. However, it is in the latter area where my own interest and these remarks are centred. (I am currently involved in a collaborative effort in this field with Fisher, Lageweg, J. K. Lenstra and Rinnooy Kan, cf. Dempster, 1979.)

In manufacturing job shops, a three-level hierarchy of planning decisions may be outlined in terms of increasingly finer time units. The first two levels can currently be handled by known deterministic linear programming and combinatorial permutation procedures, but the third—concerning the sequencing of jobs through a single machine centre—is directly related to Dr Gittins' paper. At this level practical production scheduling involves a stochastic m-machine problem whose natural setting is in continuous time.

Very recently Dr Gittins and his co-workers have obtained results for discrete time problems which show that DAI policies are optimal for the m-machine scheduling problem with a fixed queue of jobs j whose processing times t_j are independent random variables. The discrete distributions F_{t_j} involved are either exponential, i.e. constant *completion rate* (cf. failure rate in reliability theory), monotone completion rate—either increasing or decreasing—and identical in the sense that they are all conditional distributions of the same distribution after arbitrary amounts of processing (Weber and Nash, 1978; Weber, 1979) or non-overlapping completion rate in the sense that the original monotone ordering of job processing time completion rates $f_{t_j}(0)/(1 - F_{t_j}(0))$ is not changed

by subsequent processing (Gittins, 1979). It does not appear to be an entirely trivial technical matter to extend these results to continuous time. Although the optimal m-machine sequencing policies to minimize respectively expected makespan and expected flowtime—essentially *longest* and *shortest expected processing time* first (LEPT and SEPT)—are DAI policies, the methods used appear particular. It would be interesting to investigate how the general approach of Dr Gittins' paper could be utilized to obtain continuous time results.

In this regard I should like to call attention to the work of Dr Weiss, who is currently visiting Birmingham University. Following recent work of Bruno and Downey and Fredrickson, he has shown with Pinedo (1978) that the above results regarding suitable variants of the LEPT and SEPT DAI policies are optimal for the problem of sequencing jobs with exponential processing times on m machines of differing speeds. From the point of view of practical operations research this is an extremely important result which we might hope to obtain more generally for continuous time stochastic scheduling problems using bandit process theory.

There has recently been a considerable, deep and detailed combinatorial study of deterministic scheduling problems (in continuous time) as to their computational complexity (see Graham *et al.*, 1977). In layman's terms the simple question addressed is whether or not it is possible to find a computational algorithm for a deterministic scheduling problem that is polynomial in the problem parameters (*easy*) or whether the parameter dependency must be effectively exponential (*NP-hard*). For even the two-machine problem of minimizing makespan with no pre-emption of running jobs, the problem is known to be NP-hard in the deterministic case. On the other hand, a LEPT (DAI) policy is often used to sequence jobs in a practical m-machine problem—such as for a bank of lathes in a machine shop. The current theoretical operations research view, based on deterministic analysis, would say that such a policy is a suboptimal heuristic (cf. Graham *et al.*). The interesting property of the Weiss–Pinedo result is that this policy is indeed optimal as soon as specific random processing times are allowed. If extensions of these results could be found for different distributions (as in the discrete time case) and in more complex scheduling problems involving release and due dates (which are closer to those in the real world), we would have the extremely important result that heuristics which have been derived from practical experience can be proved optimal when we have the right *model*—namely one involving random variables.

Finally, going considerably further, a problem arising in understanding of real job shops involves the analysis of a network of m-machine problems. There the work of Dr Kelly and his associates at Cambridge on networks of queues, and related work in the U.S. and Europe, will hopefully soon be relevant to stochastic production scheduling. Each node of the appropriate network would be not simply a single server but rather a scheduled m-machine system, so that input and output processes would be considerably more complicated than we have so far seen. Nevertheless, there is some hope that the elegant theory of Walrand and Varaiya (1978), developed for queueing networks, could be applied more generally.

This is a big programme, but I must emphasize that there is much of practical importance in it for operations research—both regarding computer networks and for factory scheduling.

Dr J. POLONIECKI: Dr Gittins' proposed solution to the infinite horizon multi-armed bandit problem has a very surprising feature. The method consists of looking at a function of the data (r successes, n trials) on each of the arms at a time; and then deciding for the next step to use that arm for which this function has the largest value. One-step ahead horizon optimal solutions can clearly be expressed in this way. The two-step ahead horizon optimal solution cannot.

In view of this surprising feature of the solution, the name "DAI" does not do justice to its appeal. A statistician knows not to look to the observed average rate of success of the arm (r/n) for an optimal decision, nor to the expected rate of success $\{(r+1)/(n+2)\}$, nor to the expected waiting time to the next success (cf. $r/(n+1)$). The DAI tells us to look at the "maximum expected rate of return", and choose the arm for which this is the largest.

For practical application, we need a set of tables (one table per discount factor). These tables are not yet available, although Glazebrook (1978b) shows how they would be used. It is not clear, however, what happens as the working boundary for their calculation is extended. For clinical trial work there is needed, in addition, some reappraisal of the decision-making role of clinical trials.

Is the "maximum expected rate of return" policy as optimal as Dr Gittins suggests? It is based on comparing an unknown process with a standard process, and we are told that the optimal

procedure here must have the property that once the known process has been used it will be used thereafter. Clearly such a policy is not asymptotically optimal, in the sense that there is a positive probability that the known process will be used an overwhelming proportion of the time, when the probability that it is superior is not equal to one. Having been told that the "optimal" procedure is not asymptotically optimal, it is disturbing that there exist procedures which are. The existence of asymptotically optimal or "convergent" procedures has been shown under fairly general conditions (Poloniecki, 1978).

Dr G. WEISS (Birmingham University): I want to congratulate the author for pinpointing two important theorems, the forward induction and the DAI theorem and for showing how they underlie the scheduling and the multi-armed bandit problems. It seems likely that the theorems continue to hold when the definition of a bandit process is extended to be a semi-Markov decision process, where the continuation control is associated with a transition to another state, a reward and, in addition, a random time that passes until the next decision. In the scheduling context this formulation includes the scheduling problem when no pre-emptions are allowed. Harrison (1975) has calculated DAI's for that case. A further generalization of the bandit process is to allow the random emergence of new bandit processes when the continuation control is applied. This allows the treatment of arrivals as well as more complex feedback situations (see Meilijson and Weiss, 1977).

On Professor Whittle's question concerning the validity of DAI's when other arms can change state when one arm is pulled, Meilijson (1975, private communication) worked out a counter-example.

The following contributions were received in writing, after the meeting.

Professor E. M. L. BEALE (Scicon): Dr Gittins is to be congratulated on a clear exposition of a unifying approach to a narrow but significant class of problems. This approach is presented as an alternative to Dynamic Programming, but the algorithm for computing the DAI can equally be regarded as an application of Dynamic Programming. This can be seen most clearly when there is only a finite number of possible states.

The DAI ν is defined as the maximum value of the expected discounted net reward per unit of expected discounted time, when we have the option of giving up at any time after the first stage. It is natural to compute this by iteration in policy space, i.e. by iterative improvement in the set C of states from which we continue.

Let R_i be the reward for continuing when in state i, p_{ij} the transition probability from state i to state j, and i_0 the initial state. Let C_k denote the set of states from which we continue under the kth trial policy, and let $x_i^{(k)}$ and $w_i^{(k)}$ denote the discounted expected further reward and further duration respectively, when in state i. Then $x_i^{(k)} = w_i^{(k)} = 0$ if $i \notin C_k$, and otherwise

$$x_i^{(k)} = R_i + a\sum_j p_{ij}\, x_j^{(k)}, \tag{1}$$

$$w_i^{(k)} = 1 + a\sum_j p_{ij}\, w_j^{(k)} \tag{2}$$

These equations can be solved for $x_i^{(k)}$ and $w_i^{(k)}$, and ν_k can then be computed as

$$\nu_k = (R_{i_0} + a\sum_j p_{i_0 j}\, x_j^{(k)})/(1 + a\sum_j p_{i_0 j}\, w_j^{(k)}). \tag{3}$$

A new continuation set C_{k+1} can then be defined by the condition that $i \in C_{k+1}$ if and only if

$$R_i + a\sum_j p_{ij}\, x_j^{(k)} > \nu_k(1 + a\sum_j p_{ij}\, w_j^{(k)}). \tag{4}$$

With this algorithm $\nu_{k+1} \geqslant \nu_k$ and $\nu_k = \nu$ if $C_{k+1} = C_k$.

The algorithm can be streamlined by writing $y_i^{(k)} = x_i^{(k)} - \nu_k\, w_i^{(k)}$. Then from (1) and (2) we deduce that

$$y_i^{(k)} = R_i - \nu_k + a\sum_j p_{ij}\, y_j^{(k)} \quad \text{if } i \in C_k, \tag{5}$$

while from (3) we deduce that

$$R_{i_0} - \nu_k + a\sum_j p_{i_0 j}\, y_j^{(k)} = 0. \tag{6}$$

Here (5) and (6) are a set of linear simultaneous equations from which the $y_j^{(k)}$ and ν_k can be computed. Condition (4) can then be written: $i \in C_{k+1}$ if and only if

$$R_i - \nu_k + a \sum_j p_{ij} y_j^{(k)} > 0.$$

Whether or not ν is computed this way, we can write the equations defining the final values of ν and y_j in the form

$$y_i = \max (0, R_i - \nu + a \sum_j p_{ij} y_j), \tag{7}$$

$$R_{i_0} - \nu + a \sum p_{i_0 j} y_j = 0. \tag{8}$$

These equations can be justified from first principles. They apply whether the number of states is finite or infinite, and can be solved by other means that nevertheless fall within the scope of Dynamic Programming.

Professor R. BELLMAN (University of Southern California): These are important problems.

It is interesting to note that the two-arm bandit problem can be used as an example of learning. See Bellman (1971, 1978); Dreyfus and Law (1977). There is much work to be done in the area of adaptive processes.

Miss J. M. CAULDWELL: One of our chemists recently calculated that there was a total of 8×10^{14} chemical structures in a series which he was screening. At the present rate of progress it would take $1{\cdot}3 \times 10^{12}$ years to test them here. If he could persuade the total population of the world to help, this time could be reduced to about 6000 years. Decision making in the early stages of the screening process is therefore vital.

At some stage in the screening process someone has to make a decision about which compound type is showing no response and is not worthy of further investigation, as opposed to one or more compound types which are showing the potential of reaching the test target. Establishing the best line of follow-up when a series of compounds is being investigated is clearly a recurrent problem which has proved difficult to solve.

Dr Gittins has recently applied the DAI theory to a set of data which had been collected by some of our chemists working on a particular research project. The results of the statistical analysis were of considerable interest to the chemists concerned because, although the project in question had been completed, the DAI theory picked out those groups which the chemists themselves had felt to be the most promising. The theory gave statistical support to what they felt were perhaps slightly woolly reasons for following up certain groups and abandoning other groups. Furthermore, our chemists recognized the potential of the theory in assisting with decision-making at an earlier stage in the screening process, with the advantage of having some indication of the number of compounds that would have to be tested before finding one that reached the test target.

Dr P. W. JONES (University of Keele): I have two comments to make. If the optimal policy may be obtained by using DAI's for the situation where bandit processes arrive randomly in time, then presumably this approach may now be used for the optimal solution of the problem of varietal selection where varieties may be introduced at any stage in the selection procedure.

In a note, Jones (1975), the Bernoulli two-armed bandit with finite horizon, no discounting and independent beta priors was considered. The numerical work presented concerned the performance of two suboptimal policies. The one-step look ahead policy was found to be in excess of 99 per cent efficient compared with the optimal design. Using DAI's in this case would give an efficiency which is at least as large as this. This seems to suggest that the considerable computational effort required using Dynamic Programming to obtain the optimal policy is not worthwhile. The play the winner rule was also used and this had an efficiency of over 90 per cent for all the cases, this is rather surprising since this rule depends only on the previous observation. In Freeman (1970) the Bayesian estimation, under quadratic loss of the median effective dose for up to three dose levels was considered. This is, of course, a multi-armed bandit problem. It was found that the up-and-down method of allocation, which is closely related to the play the winner rule, was in excess of 90 per cent efficient in most cases.

In practice one would accept a slightly suboptimal rule which was easy to use. Therefore it would be interesting to investigate the efficiency of simple rules analogous to play the winner or

up-and-down rules for the Bernoulli multi-armed bandit problem with finite horizon or infinite horizon with discounting. The play the winner rule could be used to switch from the current arm to a randomly chosen arm or alternatively the play the winner rule could be used to switch from the current arm to that arm with the largest expected return for the next trial. To reduce the computational complexity, perhaps a mechanism for the early rejection of arms could be incorporated. Is there any evidence to suggest that part of the optimal policy for the multi-armed bandit behaves in a relatively simple way?

Dr P. NASH (Churchill College, Cambridge): The approach to the DAI theorem via forwards induction can be characterized as a branch-and-bound calculation. In deterministic dynamic programming, branch-and-bound methods attempt to overcome the curse of dimensionality by replacing the optimal remaining reward function by an estimate of it which is an upper bound (in a maximizing problem). At each stage, the total remaining reward given any particular initial decision is estimated as the sum of the one-step cost given this decision and the upper bound on the remaining reward in the state reached. The decision tree is evaluated by starting at the initial point and taking at each stage the decision for which the estimated total reward (including all the one-step costs for branches already traversed) is greatest. At any stage, attention centres on that node of the tree for which the sum of the estimated further reward and the one-step rewards obtained in reaching that node from the initial point is greatest. Eventually, this node is a final decision point, and then the upper bound calculations imply that the path leading to this node has higher total reward than any other. For a good enough upper bound, this occurs long before all paths have been evaluated. In contrast, the backwards induction of DP always evaluates all paths.

For a family F of alternative bandit processes, an upper bound is (extending the notation of the paper)

$$B(0) = \sup_{D \in F} \{\nu(D, x(0))\}/(1-a)$$

Consider the sequence of decisions which fixes (D_1, τ_1), (D_2, τ_2), ... One can show from the definition of $\nu(D)$ that if the first decision is (D, τ) and $r(D, \tau)$ is the maximum expected reward given this initial decision, then

$$r(D, \tau) \leqslant R_\tau(D) + B(0)\, E\{a^\tau\}.$$

The first step of a branch-and-bound calculation fixes a particular choice of D_1 and τ_1, the particular choice being that which maximizes

$$R_{\tau_1}(D_1) + B(0)\, E\{a^{\tau_1}\}$$

This means choosing the process whose DAI is equal to $(1-a)\, B(0)$, and τ_1 as the stopping time which yields the supremum in the definition of the DAI. The force of the forwards induction theorem is then that no decision path whose first branch does not coincide with this one need ever be investigated as we continue to branch and bound. This would seem to reinforce the hope that forwards induction policies can be proved optimal in more general circumstances, since for that particular initial decision to be optimal, we only require that paths which do not start with it will *ultimately* be abandoned in the branch-and-bound procedure. This is a weaker property than that by which the DAI theorem is proved.

Professor D. O. SIEGMUND (Stanford University): Typically dynamic programming problems are well understood qualitatively but difficult to implement computationally. In this paper Dr Gittins has described an interesting class of problems in which a simple but ingenious trick reduces these computational difficulties to manageable proportions. A given problem is replaced by a family of optimal stopping problems, which are much easier to solve. This produces a "splitting" of the given problem into independent components, the individual solutions to which may be glued together to solve the original.

The key technical idea is that of the DAI. Given a bandit process, the DAI is intuitively that value λ which makes one indifferent between accepting an immediate reward of λ and optimally stopping the bandit process with a residual reward of λ discounted by a^t if stopping occurs at time t.

The following example seems instructive. Let arm one of a MAB return 1 or 0 independently on each trial with known probability p. Let arm two return only ones with probability π_1 and only zeros with probability π_0. Then the DAI for arm two satisfies $\lambda = \pi_1/(1-a) + \pi_0 a\lambda$, and if $\lambda < p/(1-a)$, one should always continue arm one. Suppose now there are N arms stochastically

identical to arm two (but independent). For N large it is practically certain that at least one of these arms is better than arm one; but searching for it will bring a reward of

$$\pi_1/(1-a)+\pi_0 a\pi_1/(1-a)+\ldots+\pi_0^{N-1}a^{N-1}\pi_1/(1-a)<\pi_1/(1-a)(1-\pi_0 a) = \lambda,$$

so arm one remains optimal.

The class of problems for which the methods of this paper are applicable is special, albeit important. It would be interesting to know how well this class might serve to approximate other problems. For example, the results do not appear to apply directly to multi-armed bandits with correlated prior distributions. However, for discount factors close to one and a relatively small number of arms, perhaps not too much is lost. Are there analogous results for an average return criterion, which by the relation of Cesàro to Abelian summability is related to the discounted return criterion?

The potential applications to clinical trials are very thought-provoking, but their acceptance in practice may hinge on considerations apparently not amenable to systematic decision theoretic treatment (e.g. the desirability for randomization as an (the ?) important aspect of experimental design).

Finally, the reader stimulated to study the proof of the DAI theorem (Gittins and Jones, 1974a) should be warned that Lemma 2 of that paper appears to have a crucial inequality reversed.

Professor B. W. TURNBULL (Cornell University): I have been acquainted with the author's work on DAI's for some time and this paper gives a very readable account of what is an interesting and significant contribution to the theory of sequential decision processes and sequential design of experiments. I wonder whether the theory can be adapted, as in Section 10 perhaps, to handle the problem where, at each stage, one option is to freeze eternally all the rival bandit processes and take a terminal reward which depends on a termminal decision to be taken then. If so, it would be of interest to compare the DAI rules with the asymptotically optimal procedures of Bessler (1960) who took a sequential game theoretic approach. Unlike the DAI procedure, Bessler's rules have the property of being randomized which is an advantage in clinical trials because of the problem of selection bias. Of course, in other applications, non-randomized rules may be preferable.

In referring to Robbins and Siegmund (1974), the author alludes to the selection formulation of the n-armed bandit problem where it is desired to find a procedure that maximizes cumulative one-stage rewards from among that class of rules that eventually stop and select the best treatment with prescribed error probabilities. Also of interest here are the asymptotically optimal rules of Louis (1975, 1977). These papers all deal only with the case $n = 2$; for $n \geqslant 3$, similar methods can be used but there are some difficulties (Turnbull *et al.*, 1978).

The proposed use of adaptive sampling in medical trials in practice has been much criticized recently (Bailer, 1976; Simon, 1977). Two objections given are:

(A) Although adaptive sampling can lead to fewer expected number of patients on inferior treatments (ITN), it increases the total expected sample size (ASN) compared to a non-adaptive method. This delays conclusion of the trial and perhaps adversely affects patients not part of the trial.

(B) Adaptive sampling rules are too complicated.

In response, it should be noted that (A) is only true for $n = 2$; for $n \geqslant 3$ substantial savings in both ASN and ITN can be achieved simultaneously by use of adaptive sampling. This is demonstrated in Turnbull *et al.* (1978) and is intuitively clear because non-contending treatments can now be dropped from consideration early. In response to (B), it might be noted that adaptive allocation of patients to treatments based on previous responses need not be much more complicated than the adaptive allocation rules, based on prognostic variables, designed to maintain balance in a stratified study. Yet the latter type of adaptive procedure is gaining acceptance in practice, e.g. in multi-clinic trials. Finally, since ASN as well as ITN can be reduced, adaptive sampling might be applicable in animal experiments where statistical considerations can play a greater role in the design and conduct of the study.

The AUTHOR replied later, in writing, as follows.

For me at any rate the discussion has been most interesting, and I should like to begin by thanking the proposer and seconder of the vote of thanks, and indeed all the participants, for their contributions and for their kind words.

Professor Bather raises the question of the sensitivity of the solution to the multi-armed bandit problem to changes in the prior distribution and the discount factor. The simplest generalization (Gittins and Jones, 1979) is that the gap between any iso-DAI, and the line through the origin to which the iso-DAI is asymptotically parallel, is always less than 4·0 for discount factors which are not greater than 0·99. This means that, except for small values of α and β, the optimal policy is well approximated by one which always selects the arm for which the posterior expectation $(\alpha+1)/(\alpha+\beta+2)$ of the unknown success probability is largest. Call this policy A. To this approximation, then, the solution is robust to changes in the discount factor. The effects of changes in the prior distribution on policy A are measurable as constant changes in α and β for the arm concerned. As for most Bayesian procedures, the precise choice of prior distribution is not crucial, but priors which differ by assigning high probabilities to different regions of the parameter space (these correspond to high initial values of α and β) lead to substantially different procedures. The calculations reported by Jones (1975) and Wahrenberger *et al.* (1977) show that for the finite horizon undiscounted problem policy A again does well, and is not unduly sensitive to changes in the prior distribution.

The randomized allocation indices proposed by Professor Bather are variations of policy A, and their good performance is thus not altogether surprising. The device of randomization leads to the asymptotic optimality property which he describes, and which, as Dr Poloniecki points out, the Bayes policy based on DAI's does not have. A thoroughgoing Bayesian would not, of course, regard this as a particularly strong objection. However, it would be interesting to examine the performance policies obtained by adding a random component to the DAI, rather than the proportion of successes, for each arm. In this way it might be possible to have the best of both worlds. Extensive calculations of the DAI function have been carried out for various values of the discount factor, and are described by Gittins and Jones (1979).

It can actually be shown that for any values of α and β the DAI tends to one as the discount factor tends to one. This means that the gap between an iso-DAI and the asymptotically parallel line through the origin must tend to infinity, despite the above-mentioned unremarkable behaviour for discount factors up to 0·99. The behaviour of the iso-DAI's in the limit is an intriguing open question, as (Berry, 1972) is the nature of the optimal policy for the undiscounted case as the horizon tends to infinity, though the practical significance may not be particularly great in either case.

As Professor Bather says, there is a noticeable lack of enthusiasm among medical statisticians for allocation rules designed to reduce the number of patients given inferior treatments in clinical trials. My impression, like that of Professor Siegmund, is that this is largely attributable to the importance attached to randomization as a means of removing bias. However, pressure from medical practitioners and from governments may lead to a change of attitude. Professor Turnbull also makes some interesting comments on this point.

The result mentioned by Professor Whittle is intuitively appealing. The quantities γ_u and δ_u are natural measures of the cost of progress towards a terminal decision, under H_1 and H_2 respectively, when experiment u is used. Thus one might hope to find an elementary derivation. However, I have been unable to find an interpretation of this rule as a forwards induction policy, and would be inclined to look for for an appropriate generalization of Wald's equation.

The remarks of Mr Davies, Mr Baker and Miss Cauldwell refer primarily to the set of DAI tables prepared by Gittins and Jones (1974b) as an aid in new-product chemical research. It is encouraging to hear from them of scope for practical application. I am in the process of analysing several sets of compound screening data provided by pharmaceutical companies with the help of these tables. The findings of this exercise will be reported in due course.

Mr Davies also raises the question of what to do if a bandit process improves as a result of the research team's increased skill in selecting new compounds. My feeling is that such changes can best be taken into account by calculating the DAI on the basis of recent results only, rather than by modelling the learning process itself. Of course, as Professor Bellman remarks, the models do incorporate an aspect of learning, but this is not the one to which Mr Davies refers.

The computer-based procedure mentioned by Dr Roberts is also designed as a learning model, this time for the purpose of dividing resources between different new-product chemical research projects. I have been following its development with interest.

I should like to congratulate Dr Kelly on finding two ingenious new applications of the DAI theorem in the form which allows the time between successive decision points to depend on the state

of the bandit process which is currently being continued. As Dr Weiss surmises, the theorem still holds if this time is also allowed to be random, a result which is given in a slightly different form by Gittins and Nash (1977). The striking feature of Dr Kelly's two examples is his use of this variable time to represent first a cost, and then a probability, thereby establishing results for situations where the things which look most like bandit processes do not function independently.

For his first example of a hidden object, with no costs, but a reward of a^t if it is found at time t, each box taking one unit of time to search, the expected reward under an arbitrary deterministic policy is

$$\sum_{i=1}^{n} P(i) \sum_{j=1}^{\infty} \left\{ \prod_{k=1}^{j-1} (1 - d(i,k)) \right\} d(i,j)\, a^{t(i,j)}. \qquad (*)$$

Here the time at which the jth search of box i takes place, if the object has not been found before then, is denoted as $t(i,j)$. We note that the expression (*) is also the expected total reward for a family of n alternative bandit processes for each of which the state coincides with the process time, under a policy which, for all i and j, continues bandit process i for the jth time at time $t(i,j)$. To make this interpretation we must let the undiscounted reward from continuing bandit process i when it is in state j be

$$P(i) \prod_{k=1}^{j-1} (1 - d(i,k))\, d(i,j).$$

The optimal policy for both problems is therefore expressible in terms of DAI's. For the case when the time taken by the jth search of box i is $c(i,j)$ we simply replace $t(i,j)$ by $\Sigma_{k=1}^{j} c(i,k)$ in (*), and make the corresponding change in the expression for the DAI. Letting a tend to one in this expression leads to the index $\nu(i)$ given by Dr Kelly. Thus a policy based on this index must be such as to minimize the term of order $1-a$ as a tends to 1 in (*), and this is precisely what is required for the original undiscounted search problem for which $c(i,j)$ is the cost of the jth search of box i. Indeed a neat piece of work, and I hope Dr Kelly will not mind my filling in these details.

Professor Fristedt and Dr Poloniecki draw attention to the possibility that the iterative methods of calculation which I have described may lead to unacceptable accumulation of errors. This is an important consideration, and checks must be incorporated in any set of calculations to ensure that this does not happen.

Dr Glazebrook indicates three areas of current and prospective research interest His paper on stoppable families of alternative bandit processes provides a partial answer to a question raised by Professor Turnbull. It extends the DAI theorem to stoppable families under a certain condition, which includes monotonicity conditions as special cases.

Dr Glazebrook suggests using a hamiltonian approach to solve continuous-time sequential allocation problems, along the lines of Nash and Gittins (1977). This is certainly a line worth pursuing, and the account given by Nash (1973) is still worth reading, not least for its discussion (Section 2.4) of the multi-server problems to which Dr Dempster refers. It seems to me, however, that we could also do with a general theorem for translating discrete-time results into their obvious continuous-time analogues. The entire theory of Markov decision processes has a gap at this point.

Dr Dempster gives a useful outline of current work on multi-processor scheduling problems. As he says, this is an exciting area in which much remains to be done. I conjecture, for example, that conditions which ensure that the policy which minimizes expected average flow-time is expressible in terms of a DAI, when no new jobs arrive, will also ensure this when the arrivals of new jobs form a Poisson process. For the single processor case this has already been established by Nash (1973), as a consequence of the DAI theorem, and independently by Meilijson and Weiss (1977), who used a neat, and entirely different, inductive argument. As Dr Jones says, there is a possible application in varietal selection, though here the calculation of DAI's may present serious problems.

It is interesting to note that for this result the criterion is the average return per unit time, and was established by Nash from the corresponding discounted return problem by letting the discount factor tend to one. I share Professor Siegmund's view that there must be more general results of this type waiting to be proved.

Professor Beale presents an attractive algorithm for carrying out the calculations outlined in Section 8, which I am sure will prove useful. His equation (7) is equivalent to equation (13) of the paper, and his equation (8) is implicit in the text of the following paragraph. As he says, this is all dynamic programming, but I stand by my assertion that "forwards induction policies are often

easier to determine than backwards induction policies", and that it is therefore worth knowing when forwards induction policies are optimal. Dr Nash's characterization of forwards induction as a branch-and-bound calculation supports this view. This is a connection which itself warrants further investigation.

Finally, I should like to thank Professors Fristedt and Siegmund for clarifying several points.

References in the Discussion

BALLAR, J. (1976). Patient assignment algorithms: an overview. *Proc. 9th Int. Biometric Conf.*, **I**, 189–203.

BELLMAN, R. (1971). *Introduction to the Mathematical Theory of Control Processes, Vol. II.* New York: Academic Press.

—— (1978). *An Introduction to Artificial Intelligence: Can Computers Think?* San Francisco: Boyd and Fraser.

BERRY, D. A. (1972). A Bernoulli two armed bandit. *Ann. Math. Statist.*, **43**, 871–897.

BESSLER, S. (1960). Theory and applications of the sequential design of experiments, k actions and infinitely many experiments. Technical Reports Nos. 55, 56, Dept. of Statistics, Stanford University.

DEMPSTER, M. A. H. (1979). Some programs in dynamic stochastic scheduling. Presented at 2nd Int. Conf. on Stochastic Programming, Oberwolfach, Germany, February 2nd, 1979.

DREYFUS, S. and LAW, A. M. (1977). *The Art and Theory of Dynamic Programming.* New York: Academic Press.

FABIUS, J. and VAN ZWET, W. R. (1970). Some remarks on the two-armed bandit. *Ann. Math. Statist.*, **41**, 1906–1916.

FREEMAN, P. R. (1970). Optimal Bayesian sequential estimation of the median effective dose. *Biometrika*, **57**, 79–89.

GITTINS, J. C. (1979). Sequential stochastic scheduling with more than one server. *Math. Operationsforschung* (in press).

GLAZEBROOK, K. D. (1979). Stoppable families of alternative bandit processes. *J. Appl. Prob.* (in press).

GRAHAM, R. L., LAWLER, E. L., LENSTRA, J. K. and RINNOOY KAN, A. H. G. (1977). Optimization and approximation in deterministic sequencing and scheduling: a survey. Technical Report BW 82/77, Mathematisch Centrum, Amsterdam. To appear in *Proc. of Discrete Optimization* 1977, August 8th–12th, 1977.

HARRISON, J. M. (1975). Dynamic scheduling of a multiclass queue: discount optimality. *Oper. Res.*, **23**, 270–282.

JONES, P. W. (1975). The two armed bandit. *Biometrika*, **62**, 523–524.

KADANE, J. B. and SIMON, H. A. (1977). Optimal strategies for a class of constrained sequential problems. *Ann. Statist.*, **5**, 237–255.

LOUIS, T. A. (1975). Optimal allocation in sequential tests comparing the means of two Gaussian populations.

—— (1977). Sequential allocation in clinical trials comparing two exponential survival curves. *Biometrics*, **33**, 627–634.

MEILIJSON, I. and WEISS, G. (1977). Multiple feedback at a single server station. *Stoch. Proc. Applic's*, **5**, 195–205.

POLONIECKI, J. D. (1978). The two armed bandit and the controlled clinical trial. *Statistician*, **2**, 97–102.

ROBBINS, H. (1952). Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.*, **68**, 527–535.

SIMON, R. (1977). Adaptive treatment assignment methods and clinical trials. *Biometrics*, **33**, 743–749.

TURNBULL, B. W., KASPI, H. and SMITH, R. L. (1978). Adaptive sequential procedures for selecting the best of several normal populations. *J. Statist. Comput. Simul.*, **7**, 133–150.

WALRAND, J. and VARAIYA, P. (1978). The output of Jacksonian networks are Poissonian. Memo ERL-M78/60. Electronics Research Laboratory, University of California at Berkeley.

WEBER, R. R. (1979). Scheduling stochastic jobs on parallel machines. Cambridge University.

WEBER, R. R. and NASH, P. (1978). An optimal strategy in multi-server stochastic scheduling. *J. R. Statist. Soc.* B, **40**, 322–327.

WEISS, G. and PINEDO, M. (1978). Scheduling tasks with exponential service times on non-identical processors to minimize various cost functions. Technical Report ORC 78–16. Operations Research Centre, University of California at Berkeley. (To appear in *J. Appl. Prob.*)

WHITTLE, P. (1965). Some general results in sequential design. *J. R. Statist. Soc.* B, **27**, 371–394.