J. Appl. Prob. **27**, 637–648 (1990)
Printed in Israel

ON AN INDEX POLICY FOR RESTLESS BANDITS

RICHARD R. WEBER,* *University of Cambridge*
GIDEON WEISS,** *Georgia Institute of Technology*

Abstract

We investigate the optimal allocation of effort to a collection of n projects. The projects are 'restless' in that the state of a project evolves in time, whether or not it is allocated effort. The evolution of the state of each project follows a Markov rule, but transitions and rewards depend on whether or not the project receives effort. The objective is to maximize the expected time-average reward under a constraint that exactly m of the n projects receive effort at any one time. We show that as m and n tend to ∞ with m/n fixed, the per-project reward of the optimal policy is asymptotically the same as that achieved by a policy which operates under the relaxed constraint that an average of m projects be active. The relaxed constraint was considered by Whittle (1988) who described how to use a Lagrangian multiplier approach to assign indices to the projects. He conjectured that the policy of allocating effort to the m projects of greatest index is asymptotically optimal as m and n tend to ∞. We show that the conjecture is true if the differential equation describing the fluid approximation to the index policy has a globally stable equilibrium point. This need not be the case, and we present an example for which the index policy is not asymptotically optimal. However, numerical work suggests that such counterexamples are extremely rare and that the size of the suboptimality which one might expect is minuscule.

FLUID APPROXIMATIONS; GITTINS INDEX; LARGE DEVIATION THEORY; MULTI-ARMED BANDIT PROBLEM; STOCHASTIC SCHEDULING

1. Restless bandits

Whittle (1988) has recently studied an interesting generalization of the classical multi-armed bandit problem. The classical problem concerns n projects, the state of project i at time t being denoted $x_i(t)$. At each time t just one project is to be operated. If project i is operated then a reward $g_i(x_i(t))$ is received and the transition $x_i(t) \to x_i(t+1)$ follows a Markov rule specific to project i. The $n-1$ projects which are not operated produce no reward, and their states do not change. One thinks of a gambler who at each turn can pull exactly one of the n arms of a multi-armed bandit, or slot-machine, and who desires to maximize his time-average reward by an optimal sequence of pulls.

Received 14 July 1989; revision received 29 September 1989.

* Postal address: Cambridge University Engineering Department, Management Studies Group, Mill Lane, Cambridge CB2 1RX, UK.

** Postal address: School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332–0205, USA.

Gittins (1970) (see Gittins and Jones (1974)) showed that an *index policy* is optimal for this problem. The *Gittins index*, denoted by $v_i(x_i)$, can be calculated for each project as a function of the label i and state x_i alone; the optimal policy is simply to operate the project of greatest index.

Whittle has studied a variation in which two generalizations are introduced. Firstly, at each moment exactly m of the projects are to be operated. Secondly, those $n - m$ which are not operated may nevertheless contribute reward and change their state. A project is said to be *active* or *passive* depending upon whether or not it is operated. It is because passive projects may change state that they are called *restless bandits* (since we are now thinking of a multi-armed bandit machine for which even those arms which are not pulled change state). Whittle gave several examples of problems which are nicely modelled as restless bandits, for example the exhaustion and recuperation modes of n workers, where m must always be active. Here a change of state corresponds to a change of a worker's physical condition. A change of state may also correspond to a change of information. One might gain different information when making a project active than when making it passive. For example, projects might deteriorate in an unobserved manner when not made active. We might think of m helicopters trying to keep track of the positions of n submarines.

For simplicity of exposition we suppose that all projects are of the same type: they have the same finite state space, with states labelled $\{1, \cdots, k\}$, and change their state according to an identical Markov rule. It is convenient to formulate the model in continuous time. We suppose that following entry to state i the next *potential* change of state occurs at a time which is exponentially distributed with mean 1. At that time the project moves to state j with probability $P_{ij}(a)$, where action $a = 1$ or 2 denote respectively that the active or passive action was being applied to the project just prior to its potential change of state. Note that the diagonal entries of $P(a)$ are not necessarily 0, so the project may not actually change state. This *uniformization device* is a standard idea in the study of Markov processes and simplifies the analysis. We shall also suppose that the transition matrices are such that the states form a single closed class regardless of the policy employed. Reward is earned at a rate $g(i, a)$ whenever the project is in state i and action a is taken.

If one were attempting to maximize average reward over the infinite horizon for a single project, the solution would be found from the optimality equation

$$\gamma + f(i) = \max_{a=1,2} \left\{ g(i, a) + \sum_{j=1}^{k} P_{ij}(a) f(j) \right\},$$

and γ would be the time-average optimal reward. Consider now the optimality equation obtained if the reward received when taking the passive action is subsidised by an amount v:

$$(1) \qquad \gamma(v) + f(i) = \max \left\{ g(i, 1) + \sum_{j=1}^{k} P_{ij}(1) f(j),\ v + g(i, 2) + \sum_{j=1}^{k} P_{ij}(2) f(j) \right\}.$$

One thinks of v as a subsidy which is paid for taking the passive action. It is intuitive that as v increases from $-\infty$ to $+\infty$ the proportion of time for which it is optimal to take the

passive action increases. One can interpret ν as the Lagrangian multiplier associated with a constraint that the passive action be taken for a proportion of the time $1-\alpha$, $0 \leqq \alpha \leqq 1$. This leads us to consider a problem in which the constraint is relaxed to demanding only that the number of projects to be made active satisfies $Em(t) = \alpha n$ in time average. It is clear that the policy which maximizes the average reward subject to this constraint satisfies (1) when ν is pitched at the right level. A *subsidy policy* is defined by Whittle as a policy induced by (1) for some value of ν and which resolves any tie between the terms within the maximization operator by choosing the passive action. However, except for a finite number of values of α, for which the non-randomizing ν-subsidy policies are optimal, the appropriate subsidy ν will be such that the active and passive actions are equally attractive for some state i. By appropriately randomizing the choice of action in this state a stationary policy is obtained for which the constraint is satisfied. Denote by $r(\alpha)$ the maximum average reward that can be achieved under the relaxed constraint; this can be expressed as (Whittle (1988), Proposition 1)

$$r(\alpha) = \inf_{\nu} \{\gamma(\nu) - \nu(1-\alpha)\}. \tag{2}$$

Clearly $nr(\alpha)$ is an upper bound for $R_{\text{opt}}^{(n)}(\alpha)$, the maximal average reward to be obtained from n projects subject to the more demanding constraint that exactly $m = \alpha n$ are to be made active at all times. The policy which achieves the value $nr(\alpha)$ simply applies the same policy to each project independently, making each project active, passive, or perhaps randomizing between active and passive actions, on the basis of the state of the project alone. We call this policy, which is optimal under the relaxed constraint, the *relaxed-constraint optimal policy*, or more briefly the *relaxed policy*, and denote it σ_{rel}.

Remark. If α and n are such that αn is not an integer, then we interpret the constraint $m = \alpha n$ as demanding that $\lfloor \alpha n \rfloor$ projects be made active, $n - \lceil \alpha n \rceil$ projects be made passive, and the remaining project be made active with probability $\alpha n - \lfloor \alpha n \rfloor$. This is consistent with the idea that the resource available for applying active actions is continuously divisible. This interpretation also avoids technical issues that are not central to our discussion.

Whittle defines the index $\nu(i)$ for a project in state i as the least value of the subsidy ν for which it could be optimal in (1) to make the project passive in state i. It turns out that this index reduces to the usual Gittins index in the case that the passive projects do not change state and yield no rewards. If indexing is to be meaningful it should induce a consistent ordering, meaning that if it is optimal to make a project passive when the subsidy is ν then it is also optimal to make it passive for all $\nu' > \nu$. The concept is formalized in Whittle's definition of *indexability*.

Definition. *Let $D(\nu)$ be the set of states for which a project would be made passive under a ν-subsidy policy. The project is indexable if $D(\nu)$ increases monotonically from* 0 *to* $\{1, \cdots, k\}$ *as ν increases from* $-\infty$ *to* $+\infty$.

Whittle's *index policy*, denoted σ_{ind}, is the one which at all times makes active the m $(= \alpha n)$ projects of greatest index (where this is interpreted according to the remark

above). For this policy we denote the time-average reward for a problem with n projects by $R_{\text{ind}}^{(n)}(\alpha)$. Clearly

$$R_{\text{ind}}^{(n)}(\alpha) \leqq R_{\text{opt}}^{(n)}(\alpha) \leqq nr(\alpha). \tag{3}$$

One expects that under the index policy the reward per project will be close to $r(\alpha)$ and the policy will be nearly optimal. This is Whittle's conjecture, which can be stated as follows.

Conjecture (Whittle (1988)).

$$R_{\text{ind}}^{(n)}(\alpha)/n \to r(\alpha) \quad \text{as } n, m \to \infty, m = \alpha n. \tag{4}$$

Because it is possible to compute the $v(i)$'s, the index policy is easy to implement. The truth of the conjecture would make the index policy an attractive policy for the constrained problem, since we could be sure that for large n it nearly achieves a reward of $nr(\alpha)$; this is also a quantity that can be computed.

The conjecture is plausible since, as n increases, one expects a weaker coupling between the states of distinct projects. If the relaxed policy is applied to n projects then the equilibrium number in state i will be binomially distributed as $B(n\pi_i, n\pi_i(1-\pi_i))$, where π_i is the proportion of time a single project spends in state i. If the initial distribution is the relaxed policy's equilibrium distribution and one starts to apply the index policy then, at least initially, the relaxed and index policies will differ in the actions they take on a number of states whose expected value is only $O(\sqrt{n})$, and the expected difference in reward per project between the policies will be $O(1/\sqrt{n})$. Since actions do not differ very much, the equilibrium distribution of the index policy will also have about $n\pi_i \pm O(\sqrt{n})$ projects in state i.

However, we have discovered that conjecture (4) is false and we show this in Section 4 using ideas from the theory of large deviations and a specific counterexample. The counterexample was by no means easy to find. We still believe the conjecture to be true in most circumstances. Section 3 describes an analysis of the index policy using the theory of large deviations. In it we give a sufficient condition for the truth of (4). Section 2 begins the paper with a proof of a positive result: that the second inequality in (3) is an equality to within $O(\sqrt{n})$. Thus for large n imposition of the more demanding constraint does not lead to substantial reduction of reward per project.

2. The asymptotic reward of the optimal policy

We begin by showing that asymptotically nothing is gained by the relaxation of the constraint. Asymptotically the optimal reward per project is the same for the constrained and relaxed problems.

Theorem 1.

$$R_{\text{opt}}^{(n)}(\alpha)/n \to r(\alpha) \quad \text{as } n, m \to \infty, m = \alpha n. \tag{5}$$

Proof. Let π be the equilibrium distribution for the state of a single project when the relaxed policy is employed. Consider the expected-average-cost optimality equation for the constrained problem:

$$R_{\text{opt}}^{(n)}(\alpha)/n + f(x) = \max_a \left\{(1/n) \sum_{i=1}^{n} g(x_i, a_i) + E_a f(X)\right\}.$$

Here x_i denotes the state of project i, a_i denotes the action taken for that project, and X_i denotes the state of project i subsequent to the next potential transition (which occurs after time exponentially distributed with mean $1/n$), x, $X \in \{1, \cdots, k\}^n$, $a \in \{1, 2\}^n$. Assume data in the problem is such that π is rational and n is such that $n\pi$ is a vector of integers. Let $n_i(x)$ denote the number of projects in state i. Suppose the state x is such that the number of projects in state i is exactly $n_i(x) = n\pi_i$. Then, application of the relaxed policy satisfies the constraint that exactly m projects will be made active. Suppose now that the relaxed policy is applied to every project; denote by a_i^* the action applied to project i. Moreover, suppose that for a time δ the constant action a_i^* is applied to project i even if that project changes state. The expected number of potential state changes which occur during this interval of length δ is $n\delta$. The expected reward obtained during the interval is bounded below by $\delta r(\alpha)n - n\delta^2 G$, where $G = 2 \max_{i,a} |g(i, a)|$. Clearly the policy is suboptimal, so

$$\delta R_{\text{opt}}^{(n)}(\alpha) + f(x) \geqq \{\delta r(\alpha)n - n\delta^2 G + E_{a^*} f(X^\delta)\}. \tag{6}$$

Here X_i^δ is the state of the project i after time δ. Define, for any two states x and y, the distance $d(x, y)$ as the minimal number of components in which x and y can be made to differ by permuting the components of y: this is $d(x, y) = (1/2)\Sigma_i |n_i(x) - n_i(y)|$. Note that we can write $n_i(X^\delta) = Y_1 + \cdots + Y_k$ where $Y_1, \cdots, Y_k$ are independent random variables, Y_j has a binomial distribution $B(n_j(x), P_{ji}(a_j, \delta))$ and $P_{ji}(a_j, \delta) = e^{-\delta}(\delta_{ji} + \delta P_{ji}(a_j) + \delta^2 P_{ji}^2(a_j)/2! + \cdots)$ denotes the probability with which a project in state j is in state i after time δ given that the fixed action a_j is applied. From this, and the fact that $n_i(x) = n\pi_i$, it follows that the expected value and variance of $n_i(X^\delta)$ are $n_i(x) + no(\delta)$ and $2\delta n_i(x)\{1 - P_{ii}(a_i)\} + no(\delta)$. By the central limit theorem and the fact that $d(\cdot)$ satisfies a triangle inequality, it follows that $Ed(x, X^\delta) \leqq no(\delta) + A\sqrt{n\delta}$ for some A.

Suppose we could show that there exists a $B > 0$ such that $f(y) - f(x) \geqq -Bd(x, y)$, for any x and y. Then from (6) we would have

$$r(\alpha) \geqq R_{\text{opt}}^{(n)}(\alpha)/n \geqq r(\alpha) - \delta G - B\{o(\delta)/\delta + A/\sqrt{n\delta}\}. \tag{7}$$

By taking δ sufficiently small and then letting $n \to \infty$ (through a subsequence for which $n\pi \in Z^k$) the right-hand side of (7) has a limit greater than $r(\alpha) - \varepsilon$, for any $\varepsilon > 0$. This would prove the theorem.

Since $d(\cdot)$ obeys a triangle inequality, it suffices to prove the claim of the above paragraph to consider x, y such that $d(x, y) = 1$. Suppose this is the case and that x and y differ on project i. We use a coupling argument: suppose that in state y we apply to each project exactly the same action which σ_{opt} applies to that project in state x. We continue

to do this until there is a potential change of state for project i. This occurs after a time which is exponentially distributed with mean 1 and we deduce from the optimality equation

$$f(x) - f(y) \geqq -G + E_a\{f(X) - f(Y)\}, \tag{8}$$

where X and Y denote the states of the projects at the time of the first potential change of state for project i. Now $d(X, Y) \leqq 1$. Moreover, there is always some probability, say at least $\omega > 0$, that $X_i = Y_i$ and therefore that $d(X, Y) = 0$. Thus from (8), we have

$$\min_{x,y:\, d(x,y)=1} \{f(x) - f(y)\} \geqq -G + (1-\omega) \min_{x,y:\, d(x,y)=1} \{f(x) - f(y)\},$$

which implies $f(x) - f(y) \geqq -G/\omega$, completing the proof of the theorem.

3. A sufficient condition for asymptotic optimality of the index policy

In this section we consider the index policy, σ_{ind}, applied to a collection of n projects. Let $z_n(t) = (z_{n1}(t), \cdots, z_{nk}(t))$ be a state for the system, where $z_{ni}(t)$ denotes the proportion of projects in state i. Possible transitions of the process are of the form $z_n \to z_n + (1/n)e_{ij}$, where e_{ij} is a vector with -1 in the ith component, $+1$ in the jth component and 0's in all other components. Consider a policy in which the $m = \alpha n$ projects of greatest index are made active. We extend the definition of the index policy by stating that when αn is not an integer then $\lfloor \alpha n \rfloor$ projects of greatest index are made active and then one further project, with a greatest index amongst those remaining, is made active with probability $\alpha n - \lfloor \alpha n \rfloor$. Let q_{ji}^1 and q_{ji}^2 denote the transition rates from state i to j under the active and passive actions respectively. For convenience, we have chosen to write the transition rate matrices so that (q_{ji}^1), and (q_{ji}^2) have columns summing to 0 (which is contrary to the usual convention for Markov processes, but convenient in this exposition). The remainder of the paper does not employ the uniformization, but works directly with the transition rates. Define for any numbers a^1 and a^2, and $1 \leqq i \leqq k$,

$$u_i(z) = \min\left\{z_i, \max\left\{0, \alpha - \sum_{h=i+1}^{k} z_h\right\}\right\} \Big/ z_i,$$

$$1 - u_i(z) = \min\left\{z_i, \max\left\{0, 1 - \alpha - \sum_{h=1}^{i-1} z_h\right\}\right\} \Big/ z_i.$$

$$\phi_i(z, a^1, a^2) = u_i(z)a^1 + (1 - u_i(z))a^2.$$

Here $u_i(z)$ and $1 - u_i(z)$ are the probabilities that a project which is selected at random from amongst those which are presently in state i will be made active or passive. Under the index policy the transition rate associated with e_{ij} is $nq_{ji}(z_n)z_{ni}$, where

$$q_{ji}(z) = \phi_i(z, q_{ji}^1, q_{ji}^2) = \{u_i(z)q_{ji}^1 + (1 - u_i(z))q_{ji}^2\}. \tag{9}$$

Define the path $z(t)$ starting at $z(0)$, $\Sigma\, z_i(0) = 1$, by the differential equation

$$dz/dt = \sum_{i,j} q_{ji}(z)z_i e_{ij} = Q(z)z, \tag{10}$$

or equivalently, $dz_i/dt = \Sigma_{j\neq i} q_{ij}(z)z_j - \Sigma_{j\neq i} q_{ji}(z)z_i$. This is the *fluid approximation* for $z_n(t)$.

A sufficient condition for the asymptotic optimality of the index policy can be stated as follows.

Theorem 2. *Let π be the equilibrium distribution of a single project operated under the relaxed-constraint optimal policy. Suppose that the differential equation* (10) *has no limit cycles, nor do its solutions behave chaotically. Then if the projects are indexable,* (10) *has the unique fixed point π and $z(t)\to\pi$ for all $z(0)$. Furthermore conjecture* (4) *is true: $R^{(n)}_{\text{ind}}(\alpha)/n \to r(\alpha)$, as n and m increase to infinity with $m = \alpha n$.*

Firstly, we prove a lemma which shows that indexability implies that π is a unique fixed point. Consider a single project. Suppose it is indexable and that, without loss of generality, $v(1) < \cdots < v(k)$. Let $\sigma(i, \theta)$ be the policy which takes the passive action in states $1, \cdots, i-1$, the active action in states $i+1, \cdots, k$, and which takes the passive and active actions in state i with probabilities θ and $1-\theta$ respectively, $0 \leqq \theta \leqq 1$. Let $\alpha(i, \theta)$ denote the time-average proportion of time that the active action is taken using policy $\sigma(i, \theta)$.

Lemma 1. *Suppose the project is strictly indexable, such that $v(1) < \cdots < v(k)$. Then $\alpha(i, \theta)$ is a strictly decreasing function of θ.*

Proof. $\gamma(v)$ is a convex, piecewise-linear, increasing function of v. This follows from the fact that in the set S of stationary non-randomizing policies, a policy $\sigma \in S$ has a reward function, say $\gamma_\sigma(v)$, which is linear in v and $\gamma(v) = \max_{\sigma\in S}\{\gamma_\sigma(v)\}$. Also, $\gamma(v)$ is strictly increasing for $v > v(1)$. By indexability and the definitions of $v(i-1)$ and $v(i)$ we have that

$$\gamma_{\sigma(i,0)}(v) = \gamma(v(i-1)) + (v - v(i-1))(1 - \alpha(i, 0))$$

and

$$\gamma_{\sigma(i,1)}(v) = \gamma(v(i)) + (v - v(i))(1 - \alpha(i, 1))$$

are both subgradients to $\gamma(v)$ at $v = v(i)$. Hence

$$\gamma(v(i)) = \gamma(v(i-1)) + (v(i) - v(i-1))(1 - \alpha(i, 0)). \tag{11}$$

Now since $v(i-1) < v(i)$,

$$\gamma_{\sigma(i,1)}(v(i-1)) = \gamma(v(i)) + (v(i-1) - v(i))(1 - \alpha(i, 1)) < \gamma(v(i-1)). \tag{12}$$

So (11) and (12) imply $\alpha(i, 0) > \alpha(i, 1)$. Now suppose that π^0 and π^1 are the equilibrium distributions of policies $\sigma(i, 0)$ and $\sigma(i, 1)$. The equilibrium distribution induced by $\sigma(i, \theta)$ is a linear combination of π^0 and π^1, namely $\pi^\theta = (1-\rho)\pi^0 + \rho\pi^1$, where $\rho = \theta\pi^0_i/\{\theta\pi^0_i + (1-\theta)\pi^1_i\}$. Note that $\pi^\theta_i = \pi^0_i\pi^1_i/\{\theta\pi^0_i + (1-\theta)\pi^1_i\}$. Thus $\alpha(i, \theta) = 1 - (\pi^\theta_1 + \cdots + \pi^\theta_{i-1} + \theta\pi^\theta_i)$ is the ratio of two linear functions of θ and is therefore monotone as θ goes from 0 to 1. Since $\alpha(i, 0) > \alpha(i, 1)$ it must be strictly decreasing. This proves the lemma.

We shall also make use of the following result which states that on average z_n does not much differ from a fixed distribution ζ. In the context of diffusion processes this result

was proved by Freidlin and Ventsel (1984). It has been reformulated by Mitra and Weiss (1988) in a form which is appropriate to a Markov jump process. In fact, their statement of the proposition also allows state-dependent, exogenous, Lipschitz-continuous arrival and departure rates.

Proposition (see Mitra and Weiss (1988), Theorem 2). *Suppose there exists a probability distribution ζ such that for every initial probability distribution $z(0)$ the fluid approximation $dz/dt = Q(z)z$ has $z(t) \to \zeta$, and the transition rates $q_{ij}(z)$ are bounded and Lipschitz-continuous. Then for every $\varepsilon > 0$ there exist positive constants c_1 and c_2 such that for any initial state $z_n(0)$*

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t P(\| z_n(u) - \zeta \|_2 > \varepsilon) du \leqq c_1 \exp(-nc_2). \tag{13}$$

Proof of Theorem 2. Note that our $q_{ij}(z)$ are indeed Lipschitz-continuous. Observe also that in state $z_n(t)$ the index policy has instantaneous reward per project of $\Sigma_i \phi_i(z_n(t), g(i, 1), g(i, 2))$, where we use the function ϕ_i that was defined at the start of this section to evaluate the reward obtained by the mixing of active and passive actions. Similarly, if π is the equilibrium distribution of the state of a single project under the relaxed policy then $r(\alpha) = \Sigma_i \phi_i(\pi, g(i, 1), g(i, 2))$. Clearly $\phi_i(z_n(t), g(i, 1), g(i, 2))$ is a continuous function of z over a compact region. Suppose for the moment that (10) has the unique equilibrium π. Since we assume (10) has no limit cycles or chaotic behaviour it must be that for any initial distributions $z(0)$, $z(t) \to \pi$ as $t \to \infty$. Taking $\zeta = \pi$, the proposition above implies that the difference between $r(\alpha)$ and $R_{\text{ind}}^{(n)}(\alpha)/n$ can bounded by

$$2 \max_{i,a} |g(i, a)| c_1 \exp(-nc_2)$$

$$+ \sup_{z: \| z-\pi \|_2 < \varepsilon} \sum_i |\phi_i(z, g(i, 1), g(i, 2)) - \phi_i(\pi, g(i, 1), g(i, 2))|.$$

This may be made smaller than any arbitrary η, by first choosing ε such that the second term is less than $\eta/2$ and then taking n large enough that the first term is also less than $\eta/2$.

The proof of the theorem is completed by showing that $Q(z)z$ has the unique zero $Q(\pi)\pi = 0$. Note first that $Q(\pi)\pi = 0$, since (from (10) and the following line) this is simply a statement of the balance equations for the equilibrium distribution of the relaxed policy. Similarly, if $\zeta \neq \pi$ were a second probability vector such that $Q(\zeta)\zeta = 0$, then ζ would have to be the stationary distribution of some other policy of the form $\sigma(i, \theta)$, such that $\alpha(i, \theta) = \alpha$. But this would contradict Lemma 1. Hence $Q(z)z = 0$ has the unique root $z = \pi$. This completes the proof of Theorem 2.

4. Suboptimality of the index policy

It has been established that Whittle's conjecture is true if the differential equation for the fluid approximation has an equilibrium point which is globally asymptotically stable within the $(k-1)$-dimensional space of probability vectors. Indexability was used as

sufficient condition to guarantee uniqueness of the equilibrium point. In this section we begin by showing that even if it were known by some other argument that the equilibrium point is unique, indexability is a necessary condition for the stability of that point. Although Lemma 2 is interesting in itself, the main reason for its presentation is in order to explain how the question of the stability of the equilibrium point of (10), which is apparently a non-linear differential equation, reduces to a question of the stability of a linear system. In the second part of this section we use these ideas to explain a counterexample to the conjecture that occurs because the equilibrium point is unstable.

Lemma 2. *Suppose that for a given* α *the stable point of* (10) *is the equilibrium distribution* π, *and* i *is a state such that*, $0 < u_i(\pi) < 1$; $u_j(\pi) = 0, j < i$; $u_j(\pi) = 1, j > i$. *Suppose that* π^0 *and* π^1 *are the equilibrium distributions of policies* $\sigma(i, 0)$ *and* $\sigma(i, 1)$. *Recall that* $\sigma(i, 0)$ *is the non-randomizing policy which takes the active action in states* $i, \cdots, k$, *and the passive action in states* $1, \cdots, i-1$. $\sigma(i, 1)$ *is the non-randomizing policy which takes the active action in states* $i+1, \cdots, k$, *and the passive action in states* $1, \cdots, i$. *Then in order that* $z(t) \to \pi$ *for all* $z(0)$ *it is necessary that* $\alpha(i, 1) < \alpha(i, 0)$. *Equivalently this is*

$$\alpha(i, 1) = \pi^1_{i+1} + \cdots + \pi^1_k < \pi^0_i + \cdots + \pi^0_k = \alpha(i, 0). \tag{14}$$

Note that (4) might be true for some values of α and untrue for others. Condition (14) must hold if (4) is true for any α between $\alpha(i, 1)$ and $\alpha(i, 0)$. If (4) is true for all α then indexability is required.

Proof. Let q^k_j be the jth column of the matrix (q^k_{ij}). In a region C_i, defined as the closure of the set $\{z : 0 < u_i(z) < 1, \Sigma z_i = 1\}$, Equation (10) can be written

$$dz(t)/dt = A_i z(t) + b,$$

where

$$b = (1-\alpha)q^2_i + \alpha q^1_i,$$

$$A_i = (q^2_1 - q^2_i \mid \cdots \mid q^2_{i-1} - q^2_i \mid 0 \mid q^1_{i+1} - q^1_i \mid \cdots \mid q^1_k - q^1_i), \tag{15}$$

and A_i is a matrix partitioned by columns. Interestingly, (10) is just a linear differential equation in region C_i. Indeed dz/dt is linear in z in each of k regions, $C_1, \cdots, C_k$. We can eliminate z_i from the right-hand side of (10). (This is a consequence of the fact that z is constrained to the $(k-1)$-dimensional subspace of probability vectors.) Let $\tilde{A}_i$ be the $(k-1) \times (k-1)$ matrix formed from A_i by deleting the ith row and column and let $\tilde{q}^1_j$, $\tilde{q}^2_j$, $\tilde{z}$ and $\tilde{\pi}$ be $(k-1)$-dimensional vectors formed from q^1_j, q^2_j, z and π respectively by deleting the ith component. From $d(\tilde{z} - \tilde{\pi})/dt = \tilde{A}_i(\tilde{z} - \tilde{\pi})$ we see that if $z(t) \to \pi$ for all $z(0)$ then A_i must be a stability matrix. Thus the eigenvalues of A_i must have negative real parts. Using the fact that

$$-q^k_i = \left(\sum_{j<i} q^2_j \pi^{k-1}_j + \sum_{j>i} q^1_j \pi^{k-1}_j \right) \Big/ \pi^{k-1}_i, \qquad k = 1, 2,$$

we can eliminate q^1_i and q^2_i from (15), to get $\tilde{A}_i = \tilde{Q}_i B_i$, where

$$\tilde{Q}_i = (\tilde{q}^2_1 \mid \cdots \mid \tilde{q}^2_{i-1} \mid \tilde{q}^1_{i+1} \mid \cdots \mid \tilde{q}^1_k),$$

$$B_i = I_{k-1} + (\tilde{\pi}^1/\pi_i^1 \mid \cdots \mid \tilde{\pi}^1/\pi_i^1 \mid \tilde{\pi}^0/\pi_i^0 \mid \cdots \mid \tilde{\pi}^0/\pi_i^0).$$

Here B_i is the sum of the identity matrix and a rank-2 matrix having all its first $i-1$ and last $k-i$ columns identical. $\tilde{Q}_i$ is a Metzler matrix (of negative diagonal and non-negative off-diagonal entries) and has non-positive column sums. Therefore its Perron–Frobenius eigenvalue is non-positive. In fact, this eigenvalue cannot be 0, for if $\xi > 0$ were the corresponding eigenvector we would have

$$Q(\pi)(\xi_1, \cdots, \xi_{i-1}, 0, \xi_{i+1}, \cdots, \xi_k)^{\mathrm{T}} = 0,$$

which contradicts the irreducibility of $Q(\pi)$. Hence the real parts of the eigenvalues of $\tilde{Q}_i$ must be negative and $\tilde{Q}_i$ is a stability matrix. A necessary condition for a $d \times d$ matrix to be a stability matrix is that its determinant have the same sign as $(-1)^d$ (since this has the sign of the product of the real eigenvalues when all are negative). Therefore $\det(B_i)$ is necessarily positive. It is not hard to show that

$$\det(B_i) = (1 - \pi_1^0 - \cdots - \pi_{i-1}^0 - \pi_{i+1}^1 - \cdots - \pi_k^1)/\pi_i^0\pi_i^1,$$

which is positive if and only if (14) holds. This completes the proof of the lemma.

Consider now the problem described by the following data in which $k = 4$. All the following matrix calculations were carried out using the software PC-MATLAB 3.10.

$$(q_{ji}^1) = \begin{bmatrix} -2.5 & 0.0025 & 0 & 1.0 \\ 0.5 & -0.2825 & 0 & 0 \\ 1.0 & 0.28 & -2.0 & 1.0 \\ 1.0 & 0 & 2.0 & -2.0 \end{bmatrix}, \quad g(j,1) = \begin{bmatrix} 0 \\ 10 \\ 10 \\ 10 \end{bmatrix}$$

$$(q_{ji}^2) = \begin{bmatrix} -2.5 & 0.5 & 0 & 1.0 \\ 0.5 & -56.5 & 0 & 0 \\ 1.0 & 56.0 & -2.0 & 1.0 \\ 1.0 & 0 & 2.0 & -2.0 \end{bmatrix}, \quad g(j,2) = \begin{bmatrix} 10 \\ 10 \\ 1 \\ 0 \end{bmatrix}.$$

In states 1, 3 and 4 the transition rates are the same for both passive and active actions. In state 2 the passive rates q_{2i}^2 are chosen to be 200 times the active ones q_{2i}^1. The reader can check that projects are indexable and that as the subsidy for passivity increases from $-\infty$ through the values -10, 0, 9 and 10, the set of states which ought to be made passive, $D(\nu)$, increases monotonically by the addition of states 1, 2, 3 and 4 in that order. Suppose α is chosen so that the relaxed-constraint optimal policy makes state 1 passive, states 3 and 4 active, and state 2 passive or active with some probabilities θ and $1-\theta$. Then

$$\tilde{A}_2 = \begin{bmatrix} -3.0 & -0.0025 & 0.9975 \\ -55.0 & -2.28 & 0.72 \\ 1.0 & 2.0 & -2.0 \end{bmatrix}.$$

$\tilde{A}_2$ is not a stability matrix since its eigenvalues are -7.4037 and $0.0618 \pm 3.9670i$. This leads to a counterexample to conjecture (4). Suppose we take $\alpha = 0.835$. For this value of α the relaxed policy is passive on state 1, active on states 3 and 4, and on state 2 it is active and passive with probabilities 0.9938 and 0.0062 respectively. The equilibrium is $\pi = (0.1644, 0.0973, 0.3281, 0.4102)$.

For this value of α the solution to (10) does not tend to π as $t \to \infty$. Numerical integration of (10) shows that the fluid flow approximation for the index policy actually tends to a limit cycle of period 1.6384498 in which the state for which $0 < u_i(z) < 1$ alternates between state 1 and state 2. So, in contrast to the relaxed policy, projects in state 1 are sometimes made active. The proof which Weiss gives for the proposition in Section 3 can be adapted in an obvious way to a version in which the assumption that $z(t)$ tends to a unique equilibrium, $Q(\zeta)\zeta = 0$, is replaced by the assumption that $z(t)$ tends to a unique limit cycle for all initial probability vectors $z(0)$. It can then be shown using arguments similar to those in Section 3 that the asymptotic average reward per project of the index policy is the reward obtained by averaging the reward function around the path of the limit cycle. For our data this integral comes to $10 - 1.26577 \times 10^{-4}$. Clearly the relaxed policy achieves an average reward of 10. Thus the index policy is asymptotically suboptimal by the tiny amount of 0.00126577%.

5. Conclusions

For the data of the counterexample in Section 4 there is a heuristic policy which is asymptotically optimal. Suppose that $m/n = \alpha$ and the relaxed policy takes the active and passive actions in state 2 with probabilities $1 - \theta$ and θ. Suppose there are n_2 projects in state 2. Consider the policy which makes $\min\{(1-\theta)n_2, m\}$ of these projects active, and then makes enough of the remaining projects in states 1 and 3 active, to bring the total number of active projects to exactly $m = \alpha n$. (Which projects in states 1 and 3 are made active does not matter since the transition rates do not depend on the action taken in these states). The resulting Markov process is a migration process and satisfies detailed balance equations. So one can find expressions for the equilbrium distribution and show that it has asymptotically the same proportions of projects in each state as the relaxed policy.

Conjecture (4) is always true when $k = 2$. In this case the regions C_1 and C_2 have a single point as their boundary. The trajectories of dz/dt must enter one or the other of these regions and never leave it. The only behaviour consistent with this is $z(t) \to \pi$, so the conditions of Theorem 2 are met. In fact, for the case $k = 2$ we have derived expressions for the equilibrium distribution of the index policy. One can give a direct proof of the truth of conjecture (4). It turns out that the asymptotic difference between $R_{\text{ind}}^{(n)}(\alpha)/n$ and $r(\alpha)$ is even less than $O(1/\sqrt{n})$.

Our counterexample had $k = 4$ and it is not clear whether there might be a counterexample with $k = 3$. Certainly it will be harder to discover such an example, since we can show that when $k = 3$ indexability always implies the stability of $\tilde{A}_i$. Thus the equilibrium point of (10) is at least locally asymptotically stable.

By randomly generating values for the active and passive matrices, Q^1 and Q^2, we have found a number of counterexamples for $k = 4$ and $k = 5$. In our earliest experiments we generated the off-diagonal entries in these matrices as uniform random variables in [0, 10] and then multiplied the columns by random factors. Roughly 90% of the test problems were indexable, but in a sample of over 20000 test problems no counterexample to the conjecture was found. Counterexamples were finally discovered by restricting attention to test matrices for which Q^1 and Q^2 differed in just one column, as in the example of Section 4. Our thinking was that for this very specialised case a proof of the conjecture or a counterexample might more easily be discovered. The experiments were rewarded with counterexamples. While we did not try to accurately estimate their frequency, our impression is that counterexamples were produced for less than 1 in 1000 test problems. The size of the asymptotic suboptimality of the index policy was no more than 0.002% in any example. Of course one should not place too much emphasis on results which depend on the way test problems are generated. We may be missing a class of examples for which the degree of suboptimality is greater. A better understanding might lead to more dramatic counterexamples, but the reasoning that led to the counterexample in Section 4 does not seem to help. Nonetheless, the evidence so far is that counterexamples to the conjecture are rare and that the degree of suboptimality is very small. It appears that in most cases the index policy is a very good heuristic.

Acknowledgements

We wish to thank Dr Alan Weiss, of AT&T Bell Laboratories, who explained to us his interesting work on large deviation theory and Markov jump processes, and Professor Keith Glover, of Cambridge University Engineering Department's Control Group, who produced the numbers for the matrices (q_{ji}^1) and (q_{ji}^2) in the counterexample of Section 4. These provided a more attractive counterexample than those found by randomly generating test problems.

Our research has been sponsored by National Science Foundation grant #ECS-8712798. The second author acknowledges additional support from the FAW (Forschungs Institut für Anwendungsorientierte Wissensverarbeitung), Ulm, West Germany, for support during the summer of 1988.

References

FREIDLIN, M. I. AND VENTSEL, A. D. (1984) *Random Perturbations of Dynamical Systems*. Springer-Verlag, New York.

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

MITRA, D. AND WEISS, A. (1988) A fluid limit of a closed queueing network with applications to data networks.

WHITTLE, P. (1988) Restless bandits: activity allocation in a changing world. In *A Celebration of Applied Probability*, ed. J. Gani, *J. Appl. Prob.* **25A**, 287–298.