1991

Cost Rate Heuristics for Semi-Markov Decision Processes

Glazebrook, K.D.; Bailey, Michael P.; Whitaker, Lyn R.

Monterey, California. Naval Postgraduate School

NPS OR-91-17

COST-RATE HEURISTICS FOR SEMI-MARKOV DECISION PROCESSES

by

K. D. Glazebrook,
Department of Mathematics and Statistics,
University of Newcastle upon Tyne, U.K.

Michael P. Bailey,
Department of Operations Research,
Naval Postgraduate School,
Monterey, California 93943

Lyn R. Whitaker,
Department of Operations Research,
Naval Postgraduate School,
Monterey, California 93943

ABSTRACT: In response to the computational complexity of the dynamic programming/backwards induction approach to the development of optimal policies for semi-Markov decision processes, we propose a class of heuristics which result from an inductive process which proceeds forwards in time. These heuristics always choose actions in such a way as to maximize some measure of the current cost rate. We describe a procedure for calculating such cost-rate heuristics. The quality of the performance of such policies is related to the speed of evolution (in a cost sense) of the process. These ideas find natural expression in a class of Bayesian sequential decision problems. One such (a simple model of preventive maintenance) is described in detail. Cost-rate heuristics for this problem are calculated and assessed computationally.

KEY WORDS: Cost rate; dynamic programming; replacement policy; semi-Markov decision process.

Dr. Glazebrook was supported during this research by the National Research Council by means of a Senior Research Associateship at the Department of Operations Research, Naval Postgraduate School, Monterey, California. Dr. Bailey was supported by the Naval Weapons Support Center, Crane, IN. Dr. Whitaker was supported by the Naval Postgraduate School Research Foundation.

1. INTRODUCTION

Much research in discounted Markov and semi-Markov decision processes has centered around efficient implementations of value iteration (see Howard (1960)). Many authors (see Porteus (1980) for an overview) have studied refinements to the basic scheme. This large body of work is motivated, inter alia, by the inherent computational complexity of the dynamic programming/backwards induction approach. See Ross (1970) for an accessible account of iterative schemes for the solution of the semi-Markov decision processes of primary interest here.

Gittins (1989) describes an interestingly novel approach to the construction of policies for discounted semi-Markov decision processes. At time 0, a policy (π_1, say) and a stopping time on the process under π_1 (τ_1, say) are chosen to minimize a natural measure of cost rate incurred from the initial state at 0 up to the stopping time. The *forwards induction* policy constructed by this procedure then implements π_1 up to time τ_1. The state of the process at τ_1 ($X(\tau_1)$, say) is observed and a new policy/stopping time pair (π_2, τ_2, say) is chosen to minimize the cost rate from $X(\tau_1)$. Policy π_2 is then implemented during $[\tau_1, \tau_1 + \tau_2)$, and so on. Some strengths of this approach include the following:

(i) forward induction policies are optimal for a large class of models, especially in stochastic resource allocation. See Gittins (1989).

(ii) the on-line computation of such policies can often be performed in a way which offers considerable computational savings over conventional dynamic programming. See Katehakis and Veinott (1987) for a discussion.

(iii) the approach sometimes results in policies of simple structure (e.g., index-based). More generally it offers the prospect of relationships between model structure and policy structure which are theoretically accessible and (relatively) easily understood. See this illustrated in Glazebrook (1991).

We propose a general approach to the development of heuristics for discounted semi-Markov decision processes which uses cost-rates in a simpler fashion than in forwards induction, but which retains some of that procedure's strengths—especially those mentioned under (ii) and (iii) above. The approach is quasi-myopic and offers particular advantages in situations where to assume a fixed stationary model over an infinite horizon would be hazardous. In these heuristics, a simple choice for the stopping times τ_n, $n \geq 1$, is made a priori and cost-rate minimizations are over policies only. This class of cost-rate heuristics is introduced in Section 2 together with a procedure for their computation. Performance bounds for these heuristics are developed in Section 3 and are applied in Section 4 to the analysis of a class of Bayesian sequential decision problems. For this class of Bayesian problems, we are able to obtain results which elucidate the relationship between the performance of a cost-rate heuristic and (inter alia) the precision of initial beliefs about the unknown parameter as measured, for example, by the variance of a prior. These ideas are illustrated in Section 5 by means of computational results for a simple machine replacement problem. A cost-rate myopic policy is found to perform well much of the time.

(a) a state transition is observed, and

(b) a random amount of time elapses before the next decision epoch.

$P(G \mid x,a_j)$ is the probability that the state of the process at the next epoch lies in set $G \in F$ conditional upon the event $X(t) = x$. $F(H \mid x,y,a_j)$ is the probability that the time to the next decision epoch lies in Borel set H given that a transition from $x(=X(t))$ to y occurs. $P(G \mid \cdot,a_j):\Omega \to [0,1]$ is F-measurable and $F(H \mid \cdot,\cdot,a_j):\Omega \times \Omega \to [0,1]$ is $F \times F$-measurable. We shall denote by P^r, F^r the equivalent r-step measures—e.g., $P^r(G \mid x,\pi)$ is the probability that the state of the process at the r^{th} decision epoch after t lies in set G, given that $X(t) = x$ and that policy π (assumed not to depend upon the history of the process before t) is adopted. The first decision epoch is always assumed to be 0.

The following condition is standard in the study of semi-Markov decision processes (see, e.g., Ross (1970)). It guarantees (with probability 1) that we do not have an infinite number of decision epochs in finite time.

Condition 1. There exist positive ε, δ such that

$$\int_{\Omega} F\{(\delta,\infty) \mid x,y,a_j\} P(dy \mid x,a_j) > \varepsilon, \quad 1 \le j \le N, x \in \Omega$$

(v) **Optimal policies.** Denote by $C_r(\pi,x)$ the total expected cost incurred from the imposition of policy π from time 0 for r decision epochs when $X(0) = x$. If π is stationary $C_r(\pi,\cdot)$ may be recovered from the recursion:

$$C_0(\pi,x) = 0;$$

$$C_r(\pi,x) = c\{x,\pi(x)\} + \int_{\Omega} \int_{t=0}^{\infty} \alpha^t C_{r-1}(\pi,y) F\{dt|x,y,\pi(x)\} P\{dy|x,\pi(x)\}, r \geq 1.$$

We define

$$C(\pi,x) \equiv \lim_{r\to\infty} C_r(\pi,x) \tag{1}$$

as the total expected cost incurred by policy π when $X(0) = x$. The above assumptions (in particular the boundedness of costs and Condition 1) guarantee not only that the limit in (1) exists, but that the convergence is uniform over all policies π, for all $x \in \Omega$.

A policy π^* is optimal if

$$C(\pi^*,x) = \inf_{\pi} C(\pi,x) \equiv C(x), \quad x \in \Omega.$$

The general theory (see Blackwell (1965)) asserts the existence of an optimal policy π^* which is stationary and such that $C(\cdot)$ uniquely satisfies the recursion

$$C(x) = \min_{1 \leq j \leq N} \left\{ c(x,a_j) + \int_{\Omega} \int_{t=0}^{\infty} \alpha^t C(y) F(dt|x,y,a_j) P(dy|x,a_j) \right\}. \tag{2}$$

Procedures for determining $C(\cdot)$ and π^* include value iteration and policy iteration, as described by Ross (1970).

Now, write $\tau_r(\pi,x)$ for the random time of the r^{th} decision epoch after 0 when policy π is adopted and $X(0) = x$. We write $M_r(\pi,x) \equiv E\{\alpha^{\tau_r(\pi,x)}\}$. If π is stationary $M_r(\pi,\cdot)$ may be recovered from the recursion

$$M_0(\pi,x) = 1;$$

$$M_r(\pi,x) = \int_{\Omega} \int_{t=0}^{\infty} \alpha^t M_{r-1}(\pi,y) F\{dt|x,y,\pi(x)\} P\{dy|x,\pi(x)\}, r \geq 1.$$

Note that Condition 1 guarantees that for all π,x

$$1 > \left(1-\varepsilon+\varepsilon\alpha^{\delta}\right)^r \geq M_r(\pi,x), r \geq 1. \tag{3}$$

The notion expressed in Definition 1 is central to the ideas explored in the paper.

Definition 1. **The r-stage cost rate function for policy** π, $\Gamma_r(\pi,\cdot):\Omega \to \Re_{\geq 0}$ **is given by**

$$\Gamma_r(\pi,x) \equiv C_r(\pi,x)\{1-M_r(\pi,x)\}^{-1} \tag{4}$$

The rationale for calling $\Gamma_r(\pi,x)$ a cost rate emerges from the identity

$$\Gamma_r(\pi,x) \equiv C_r(\pi,x)\left[E\left\{\int_0^{\tau_r(\pi,x)} \alpha^t dt\right\}\right]^{-1} (-\ln\alpha)^{-1}, \tag{5}$$

in which the notion of averaging is an (appropriately) discounted one.

Definition 2. Policy $\hat{\pi}$ is (r,x)-optimal (r-stage cost rate optimal for state x) if

$$\Gamma_r(\hat{\pi},x) = \inf_{\pi} \Gamma_r(\pi,x) \tag{6}$$

In order to explore the properties of r-stage cost rates (Definition 1) and associated optimal policies (Definition 2) we introduce the mapping $T_r(x,\cdot):\Re_{\geq 0} \longrightarrow \Re_{\geq 0}$ defined by

$$T_r(x,u) = \inf_{\pi}\{C_r(\pi,x) + uM_r(\pi,x)\} \tag{7}$$

and its n-fold version $T_r^n(x,\cdot):\Re_{\geq 0} \to \Re_{\geq 0}$, where

$$T_r^n(x,u) = T_r\{x, T_r^{n-1}(x,u)\}, n \geq 1$$

Equation (7) defines a finite horizon dynamic program. We may assert the existence of a policy $\pi:\Omega\times\{1,2,...,r\} \longrightarrow \{a_1, a_2, ..., a_N\}$ attaining the infimum in (7). Here $\pi(x,s)$ is the action taken by policy π when in state $x \in \Omega$ at the s^{th} decision epoch. Call such a policy r-stage stationary.

Theorem 1. For each $x \in \Omega$, $r \geq 1$,

(a) $T_r(x,\cdot)$ is monotonic, non-decreasing;

(b) $T_r(x,\cdot)$ is a contraction mapping with respect to the L_1 norm;

(c) $\Gamma = \inf_{\pi} \Gamma_r(\pi,x)$ is the unique member of $\Re_{\geq 0}$ for which

$$T_r(x,\Gamma) = \Gamma;$$

(d) There exists an (r,x)-optimal policy which is r-stage stationary;

(e) For each $u \in \Re_{\geq 0}$

$$\lim_{n\to\infty} T_r^n(x,u) = \Gamma = \inf_{\pi} \Gamma_r(\pi,x),$$

this convergence being geometrical and uniform over x.

Proof.

(a) It is trivial from (7) that $u \geq v \Rightarrow T_r(x,u) \geq T_r(x,v)$.

(b) Suppose that u≥v. Write π(u) for an r-stage stationary policy attaining the infimum in (7). It is plain that

$$0 \le T_r(x,u) - T_r(x,v) \le M_r\{\pi(u),x\}(u-v) \le \left(1-\varepsilon+\varepsilon\alpha^{\delta}\right)^{r}(u-v),$$

from (3). This establishes (b).

(c) The contraction mapping fixed point theorem guarantees the existence of a unique fixed point for $T_r(x,\cdot)$. Call the fixed point γ. Write

$$\gamma = T_r(x,\gamma) = \inf_{\pi}\{C_r(\pi,x) + \gamma M_r(\pi,x)\} = C_r\{\pi(\gamma),x\} + \gamma M_r\{\pi(\gamma),x\} \qquad (8)$$

where we write π(γ) for a policy attaining the infimum in (8). It now follows that

$$\gamma = C_r\{\pi(\gamma),x\}\left[1-M_r\{\pi(\gamma),x\}\right]^{-1} = \Gamma_r\{\pi(\gamma),x\} \ge \Gamma.$$

Suppose that γ>Γ, and obtain a contradiction. We now have a policy $\tilde{\pi}$, say, such that

$$\gamma > C_r(\tilde{\pi},x)\{1-M_r(\tilde{\pi},x)\}^{-1}$$

from which it follows that

$$\begin{aligned}\gamma &> C_r(\tilde{\pi},x) + \gamma M_r(\tilde{\pi},x)\\ &\ge \inf_{\pi}\{C_r(\pi,x) + \gamma M_r(\pi,x)\} \Rightarrow \gamma > T_r(x,\gamma),\end{aligned}$$

from which we conclude that γ is not a fixed point of $T_r(x,\cdot)$, a contradiction. Hence γ=Γ, and we have established (c).

(d) It is now plain that any policy $\pi(\Gamma)$ attaining the infimum in (7) with $u=\Gamma$ is (r,x)-optimal. We have already noted that there is one such which is r-stage stationary. We have proved the result.

(e) This is a standard consequence of (b) and (c).

The above result plainly yields a value iteration approach to the computation of minimal cost rates and hence of (r,x)-optimal policies. We now describe the class of cost-rate heuristics for semi-Markov decision processes of primary interest to us. In Definition 3, $\underline{r} \equiv \left[\left\{r_n(\cdot):\Omega \rightarrow Z^+\right\}, n \in Z^+\right]$ is a sequence of F-measurable functions taking values in the positive integers.

Definition 3. A cost-rate heuristic determined by $\underline{r}$ is denoted $\hat{\pi}(\underline{r})$ and is a policy which operates as follows:

(a) If $X(0) = x$, $\hat{\pi}(\underline{r})$ takes the first $r_1(x)$ decisions according to an $\{r_1(x), x\}$-optimal policy;

(b) Suppose that the state of the process following the first $\sum_{m=1}^{n} r_m(X_{m-1})$ decisions and transitions (i.e., follow the first *n stages*) under policy $\hat{\pi}(\underline{r})$ is X_n, $n \geq 1$, where $X_0 \equiv X(0)$. Policy $\hat{\pi}(\underline{r})$ takes the next $r_{n+1}(X_n)$ decisions according to an $\{r_{n+1}(X_n), X_n\}$-optimal policy, $n \geq 1$.

Comments.

1. Hence policy $\hat{\pi}(\underline{r})$ implements an $\{r_1(x),x\}$-optimal policy from time 0 when $X(0)=x$ as a procedure for determining the first $r_1(x)$ decisions. The state

is then updated to X_1. The number of decisions to be taken in the second stage if $r_2(X_1)$ and is allowed to depend upon X_1. An $\{r_2(X_1), X_1\}$–optimal policy is computed and implemented from state X_1, and so on.

2. Apart from any possibility there might be of obtaining (r,x)-optimal policies of special structure, a major opportunity for cost-rate heuristics to reduce computational requirements (as compared with the application of standard dynamic programming) arises from the fact that value iteration for (r,x)-optimal policies based on Theorem 1 only needs to look at states which are accessible in r steps from state x. In the Bayesian sequential problems to which these ideas will be especially applied, considerable savings are often possible. Another instance is where state variable x is enhanced to include (for example) the number of decisions taken to date as a means of accommodating non-stationarity.

3. If each function $r_n(\cdot)$ is a constant (i.e., the number of decisions in each stage is fixed at the outset), $\hat{\pi}(\underline{r})$ is called a **fixed sequence cost-rate heuristic.** We shall often be interested in fixed sequence policies for which $r_n(\cdot)\equiv 1$, $n\geq 2$. In relation to such a choice note that (1,x)-optimal policies are often trivial to compute. Cost-rate heuristics for which $r_n(\cdot)\equiv 1$, $n\geq 1$, will be called **cost-rate myopic.**

We now explore further the rationale for considering such heuristics.

3. GENERAL PERFORMANCE BOUNDS FOR COST-RATE HEURISTICS

Write

$$\Delta(x,y) = C(y)-C(x) \equiv C(\pi^*,x)-C(\pi^*,y)$$

for the change in minimal costs which occurs upon a transition from x to y. As before, write $\tau_r(\pi,x)$ for the random time of the r^{th} decision epoch after 0 when policy π is adopted and $X(0)=x$. The subscript in the notation E_π indicates that an expectation is to be taken over realisations of the system conditional upon implementation of the policy π.

$$C\{\hat{\pi}(\underline{r}),t\}-C(\pi^*,t) \le E_{\hat{\pi}(\underline{r})}\left(\sum_{n=0}^{\infty}\left[\phi_{r_{n+1}}(\hat{\pi}_{n+1},U_n)-\psi_{r_{n+1}}(\pi^*,U_n)\right.\right.$$
$$\left.\left.\times\{1-M_{r_{n+1}}(\hat{\pi}_{n+1},U_n)\}\{1-M_{r_{n+1}}(\pi^*,U_n)\}^{-1}\right]\alpha^{\sum_{m=1}^{N(n)}Y_m}\Big|T_1=t\right)$$

Definition 4. The r-decision speed function for policy π, $\Delta_r(\pi,\cdot):\Omega\to\Re$ is given by

$$\Delta_r(\pi,x) \equiv E_\pi\left\{\alpha^{\tau_r(\pi,x)}\Delta[x,X\{\tau_r(\pi,x)\}]\right\}\{1-M_r(\pi,x)\}^{-1}$$
$$=\left[\left\{\int_{\Omega}\int_{t=0}^{\infty}\alpha^t C(y)F^r(dt|x,y,\pi)P^r(dy|x,\pi)\right\}-M_r(\pi,x)C(x)\right]\{1-M_r(\pi,x)\}^{-1} \quad (9)$$

See (5). $\Delta_r(\pi,x)$ represents a (discounted) rate at which future prospects (as measured by $C(\cdot)$) change during an r-decision implementation of policy π. It will emerge that we can go some way toward analysing policies in terms of a combination of cost rate and speed functions. The following result is an example.

Lemma 2. For each $x\in\Omega$, $r\ge 1$,

$$C(x) = \Gamma_r(\pi^*,x) + \Delta_r(\pi^*,x)$$

Proof.

By standard results, C(·) satisfies the recursion

$$C(x) = C(\pi^*,x) = C_r(\pi^*,x) + \int_{\Omega} \int_{t=0}^{\infty} \alpha^t C(y) F^r(dt|x,y,\pi^*) P^r(dy|x,\pi^*)$$
$$= \Gamma_r(\pi^*,x)\{1 - M_r(\pi^*,x)\} + \Delta_r(\pi^*,x)\{1 - M_r(\pi^*,x)\} + M_r(\pi^*,x)C(x),$$

from (4) and (9). Invoking (3), the result follows trivially.

Lemma 3. For each π and $x \in \Omega$

$$\lim_{r \to \infty} \Delta_r(\pi,x) = 0 \tag{10}$$

the convergence in (10) being uniform over all policies π and states x.

Proof.

From (3) and (9)

$$|\Delta_r(\pi,x)| \le \left\{\sup_{x \in \Omega} C(x)\right\}\left(1 - \varepsilon + \varepsilon\alpha^{\delta}\right)^r \left\{1 - \left(1 - \varepsilon + \varepsilon\alpha^{\delta}\right)^r\right\}^{-1}. \tag{11}$$

The result follows trivially.

Lemmas 2 and 3 create the expectation that (crudely speaking) should a decision process have uniformly small r-decision speed functions then an analysis in terms of r-decision cost rates could be successful. Lemma 3 tells us that we can always force the speed functions to be small by choosing r large enough. However, we note that the larger r is, the more computationally demanding is the development of (r,x)-optimal policies. We make these ideas more explicit as follows: Suppose that $\hat{\pi}$ is an (r,x)-optimal policy (see Theorem 1(d)). Write

$$C^r(x) = C^r(\hat{\pi},x) + \int_{\Omega} \int_{t=0}^{\infty} \alpha^t C(y) F^r(dt|x,y,\hat{\pi}) P^r(dy|x,\hat{\pi}) \tag{12}$$

for the total expected cost from implementing $\hat{\pi}$ for r decisions, thereafter followed by an optimal policy. Theorem 4 bounds how much is lost by pursuing $\hat{\pi}$ instead of an optimal policy for these first r decisions.

Theorem 4. For each $x \in \Omega, r \geq 1$,

$$C^r(x) - C(x) \leq \{\Delta_r(\hat{\pi},x) - \Delta_r(\pi^*,x)\}\{1 - M_r(\hat{\pi},x)\} \to 0, \quad \text{as } r \to \infty, \tag{13}$$

uniformly over all states x.

Proof.

From (9) and (12),

$$\begin{aligned} C^r(x) &= C_r(\hat{\pi},x) + \int_{\Omega} \int_{t=0}^{\infty} \alpha^t \left[\Delta(x,y) + \{C(x) - C^r(x)\} + C^r(x)\right] F^r(dt|x,y,\hat{\pi}) P^r(dy|x,\hat{\pi}) \\ &= C_r(\hat{\pi},x) + \Delta_r(\hat{\pi},x)\{1 - M_r(\hat{\pi},x)\} + \{C(x) - C^r(x)\} M_r(\hat{\pi},x) + C^r(x) M_r(\hat{\pi},x). \end{aligned}$$

Hence we deduce that

$$C^r(x) = \Gamma_r(\hat{\pi},x) + \Delta_r(\hat{\pi},x) + \{C(x) - C^r(x)\} M_r(\hat{\pi},x)\{1 - M_r(\hat{\pi},x)\}^{-1}.$$

Now, from Lemma 2

$$\begin{aligned} C^r(x) - C(x) &= \{\Gamma_r(\hat{\pi},x) - \Gamma_r(\pi^*,x)\} + \{\Delta_r(\hat{\pi},x) - \Delta_r(\pi^*,x)\} \\ &\quad + \{C(x) - C^r(x)\} M_r(\hat{\pi},x)\{1 - M_r(\hat{\pi},x)\}^{-1} \\ &\leq \Delta_r(\hat{\pi},x) - \Delta_r(\pi^*,x) + \{C(x) - C^r(x)\} M_r(\hat{\pi},x)\{1 - M_r(\hat{\pi},x)\}^{-1}, \end{aligned}$$

since $\hat{\pi}$ is (r,x)-optimal and so $\Gamma_r(\hat{\pi},x) \le \Gamma_r(\pi^*,x)$. Inequality (13) now follows trivially. The convergence result is a simple consequence of Lemma 3.

From Theorem 4, we may deduce a bound on the suboptimality of cost-rate heuristic $\hat{\pi}(\underline{r})$ expressed in terms of speed functions. Recall the notation X_n in Definition 3(b) for the state of the process following the first $\sum_{m=1}^{n} r_m(X_{m-1})$ decisions and transitions under policy $\hat{\pi}(\underline{r})$. We also write $\hat{\pi}_{n+1}$ for the $\{r_{n+1}(X_n), X_n\}$-optimal policy adopted at that stage. For notational simplicity, $r_{n+1}(X_n)$ is abbreviated to r_{n+1} in the statement and proof of Corollary 5.

Corollary 5. For any $\hat{\pi}(\underline{r})$ and $x \in \Omega$

$$C\{\hat{\pi}(\underline{r}),x\} - C(x) \le E_{\hat{\pi}(\underline{r})}\left[\sum_{n=0}^{\infty}\left\{\Delta_{r_{n+1}}(\hat{\pi}_{n+1},X_n) - \Delta_{r_{n+1}}(\pi^*,X_n)\right\}\right.$$

$$\left. \times \alpha^{\sum_{m=1}^{n}\tau_{r_m}(\hat{\pi}_m,X_{m-1})}\left\{1 - M_{r_{n+1}}(\hat{\pi}_{n+1},X_n)\right\} \middle| X(0) = x\right] \qquad (14)$$

$\to 0$ as $r_1(x) \to \infty$ (with other $r_n(\cdot), n \ge 2$, fixed). For fixed sequence cost-rate heuristics this convergence is uniform over all states x.

Proof.

Denote by $\hat{\pi}(\underline{r},n)$ a policy which follows $\hat{\pi}(\underline{r})$ for the first $\sum_{m=1}^{n} r_m(X_{m-1})$ decisions and which thereafter chooses actions optimally. We may think of $\hat{\pi}(\underline{r},n)$ as a cost-rate heuristic determined by a sequence which agrees with $\underline{r}$ up to the n^{th} term and which chooses $r_{n+1} = \infty$. By a simple argument conditioning upon X_n and the time of completion of the first n stages under $\hat{\pi}(\underline{r})$ we deduce from Theorem 4 that

$$C\{\hat{\pi}(\underline{r},n+1),x\} - C\{\hat{\pi}(\underline{r},n),x\} \le$$
$$E_{\hat{\pi}(\underline{r})}\left[\alpha^{\sum_{m=1}^{n} \tau_{r_m}(\hat{\pi}_m, X_{m-1})}\left\{\Delta_{r_{n+1}}(\hat{\pi}_{n+1},X_n) - \Delta_{r_{n+1}}(\pi^*,X_n)\right\}\left\{1 - M_{r_{n+1}}(\hat{\pi}_{n+1},X_n)\right\}\middle| X(0) = x\right] \quad (15)$$

To obtain (14) we now take $\sum_{n=0}^{\infty}$ over both sides in (15) and note that $\hat{\pi}(\underline{r},0) \equiv \pi^*$, an optimal policy, and

$$\lim_{n\to\infty} C\{\hat{\pi}(\underline{r},n),x\} = C\{\hat{\pi}(\underline{r}),x\}, \quad (16)$$

the (uniform) convergence in (16) being guaranteed by the boundedness of costs and Condition 1.

To consider the convergence of the right-hand side of (14) we easily derive from (3)

$$\left\{\Delta_{r_1}(\hat{\pi}_1,x) - \Delta_{r_1}(\pi^*,x)\right\}\left\{1 - M_{r_1}(\hat{\pi}_1,x)\right\}$$
$$+2M_{r_1}(\hat{\pi}_1,x)\left\{\varepsilon\left(1-\alpha^{\delta}\right)\right\}^{-1}\sup_{r,x,\pi}\left|\Delta_r(\pi,x)\right| \quad (17)$$

as an upper bound for it. We now invoke (3), Lemma 3 and (11) to deduce that the expression (17) converges to 0 as $r_1(x)\to\infty$. This convergence is plainly uniform for a fixed-sequence policy with $r_1(x) \equiv r_1$.

Comment

1. Consider Comment 3 at the conclusion of Section 2. If we make the computationally simple choice $r_n(\cdot) \equiv 1, n \ge 2$, we know (from Corollary 5) that

for any given $\gamma>0$ we can choose $r_1(x)$ large enough to ensure that the cost-rate heuristic $\hat{\pi}(\underline{r})$ is γ-optimal. The question of interest (from the computational complexity point of view) concerns what is the smallest value of $r_1(x)$ to achieve this?

2. From Corollary 5, it is not difficult to show that an alternative way of guaranteeing γ-optimality is to choose $r_n(\cdot)\equiv r(\cdot), n\geq 1$, in the heuristic $\hat{\pi}(\underline{r})$ where $r(\cdot)$ is such that

$$\left|\sup_{\pi} \Delta_{r(x)}(\pi,x)\right| \leq \gamma\varepsilon\left(1-\alpha^{\delta}\right)\left\{2\left(1-\varepsilon+\varepsilon\alpha^{\delta}\right)\right\}^{-1}, x\in\Omega.$$

Lemma 3 guarantees that this is achievable for any $\gamma>0$.

3. Plainly, in order to implement the suggestions contained in the previous two comments, we need to be able to characterize and/or obtain bounds on the speed functions of concern to us. To that end, in Section 4 we consider a class of problems where some progress is possible.

4. A CLASS OF BAYESIAN SEQUENTIAL DECISION PROBLEMS

Bayesian sequential decision problems seem natural candidates for the application of cost-rate heuristics. Suppose that in such a problem the current posterior distribution for the unknown parameter (or some summary of it) is the state of the process. It would seem intuitive that speed functions for policies should be related to the spread (loosely defined) of the current posterior. In particular the posterior distribution with a unit atom of probability at one parameter value (i.e., the case of known parameter) will have all speed functions equal to zero.

The following class of Bayesian sequential decision problems include the replacement problem to be considered in Section 5 as a special case. The elements of each decision problem are as follows:

(i) $X_1, X_2, \ldots$, a sequence of independent and identically distributed $\Re^d$-valued random variables with distribution F_θ, known apart from the value of parameter $\theta \in \Theta$. The support of F_θ does not depend upon θ.

(ii) Θ^*, a space of probability distributions over a fixed σ-algebra of subsets of Θ. $G \in \Theta^*$ is the prior distribution for θ.

(iii) $a_1, a_2, \ldots, a_N$, a set of actions available at each decision epoch.

(iv) $Y_1, Y_2, \ldots$, a sequence of $\Re^+$-valued random variables available to the decision-maker for observation. Should action a_j be taken at the n^{th} decision epoch then $Y_n = \Phi(X_n, a_j)$, where Φ is a measurable function. Y_n would then have distribution F_θ^j.

(v) $T_1, T_2, \ldots$, a sufficient sequence for θ (see Ferguson (1967)). For each $n \geq 1$, T_n is sufficient for θ based on $Y_1, Y_2, \ldots, Y_{n-1}$ and the actions taken at the first $n-1$ decision epochs. The posterior distribution at the n^{th} decision epoch is written $G_n \equiv G(\cdot \mid T_n)$, $n \geq 1$. Should action a_j be taken at the n^{th} decision epoch then:

(a) $Y_n = \Phi(X_n, a_j)$ is observed;

(b) a (discounted) bounded non-negative cost $\hat{c}(Y_n, a_j)$ is incurred. Taking an expectation with respect to the current posterior for θ we write the expected cost incurred as

$$c(T_n, a_j) \equiv \int_{\Theta} \int_{\Re^+} \hat{c}(y, a_j) F_\theta^j(dy) G(d\theta | T_n);$$

(c) the time between the n^{th} and $(n+1)^{st}$ decision epochs is Y_n. We shall ensure that Condition 1 holds by requiring that

$$F_\theta^j\{(\delta, \infty)\} > \varepsilon, \quad 1 \le j \le N, \theta \in \Theta,$$

for some choice of positive ε,δ. We shall also suppose that if we take T_n for the state of the process at the n^{th} decision epoch, the measurability requirements described in Section 2 are met. Our goal is to develop Bayes optimal (and good Bayes suboptimal) decision rules. If we suppose that π_n is the action taken by policy π at the n^{th} decision epoch, we write the Bayes cost for π from initial state t (a value of T_1) as

$$C(\pi, t) \equiv E_t\{\hat{C}(\pi, \theta)\} \tag{18}$$

where

$$\hat{C}(\pi, \theta) \equiv E_{\pi,\theta}\left\{ \sum_{n=1}^{\infty} \alpha^{\sum_{m=1}^{n-1} Y_m} \hat{c}(Y_n, \pi_n) \right\} \tag{19}$$

In (18) E_t denotes an expectation taken over Θ with respect to the prior $G(\cdot | t)$ and in (19) $E_{\pi,\theta}$ is, for fixed $\theta \in \Theta$, an expectation taken over realisations of the system conditional upon implementation of the policy π. An optimal policy π^* satisfies

$$C(\pi^*, t) = \inf_{\pi} C(\pi, t) \equiv C(t)$$

for all choices of t. Denote specifically by π_t^* a policy which chooses actions in an identical fashion to an optimal policy beginning in state t—i.e., for assumed prior G(· | t).

We have here a semi-Markov decision process to which the results of Sections 2 and 3 apply. In order to evaluate cost-rate heuristics we shall develop bounds on the r-stage speed functions as follows:

Theorem 6. For any r≥1, policy π and initial state t

$$\Delta_r(\pi,t)\{1-M_r(\pi,t)\} \le \sqrt{\mathrm{var}_{t,\pi,\theta}\left(\alpha^{\sum_{m=1}^{r} Y_m}\right)}\sqrt{\mathrm{var}_t\left\{\hat{C}\left(\pi_t^*,\theta\right)\right\}} \equiv \phi_r(\pi,t)$$

Proof.

For any two states t, t′

$$\begin{aligned}\Delta(t,t') = C(t') - C(t) &\le C\left(\pi_t^*,t'\right) - C\left(\pi_t^*,t\right)\\ &= E_{t'}\left\{\hat{C}\left(\pi_t^*,\theta\right)\right\} - E_t\left\{\hat{C}\left(\pi_t^*,\theta\right)\right\}\end{aligned} \tag{20}$$

since π_t^* is a suboptimal policy from initial state t′. Now, by Definition 4

$$\Delta_r(\pi,t)\{1-M_r(\pi,t)\} = E_{t,\pi,\theta}\left\{\alpha^{\sum_{m=1}^{r} Y_m}\Delta(t,T_{r+1})\right\}$$

$$\le E_{t,\pi,\theta}\left(\alpha^{\sum_{m=1}^{r} Y_m}\left[E_{T_{r+1}}\left\{\hat{C}\left(\pi_t^*,\theta\right)\right\} - E_t\left\{\hat{C}\left(\pi_t^*,\theta\right)\right\}\right]\right) \tag{21}$$

$$= E_{t,\pi,\theta}\left\{\alpha^{\sum_{m=1}^{r} Y_m}\hat{C}\left(\pi_t^*,\theta\right)\right\} - E_{t,\pi,\theta}\left(\alpha^{\sum_{m=1}^{r} Y_m}\right)E_t\left\{\hat{C}\left(\pi_t^*,\theta\right)\right\} \tag{22}$$

Inequality (21) is a consequence of (20) while (22) follows from standard results on conditional expectation. The result now follows from the fact that the correlation between $\alpha^{\sum_{m=1} Y_m}$ and $\hat{C}\left(\pi_t^*, \theta\right)$ cannot exceed one.

Theorem 7. For any r≥1, policy π and initial state t

$$\begin{aligned} \Delta_r(\pi,t)\{1-M_r(\pi,t)\} &\geq E_{t,\pi,\theta}\left(\alpha^{\sum_{m=1}^{r} Y_m}\right) E_{t,\pi,\theta}\left\{\hat{C}\left(\pi_{T_{r+1}}^*,\theta\right)-C\left(\pi_{T_{r+1}}^*,t\right)\right\} \\ &\quad -\sqrt{\operatorname{var}_{t,\pi,\theta}\left(\alpha^{\sum_{m=1}^{r} Y_m}\right)} \sqrt{\operatorname{var}_{t,\pi,\theta}\left\{\hat{C}\left(\pi_{T_{r+1}}^*,\theta\right) - C\left(\pi_{T_{r+1}}^*,t\right)\right\}} \\ &\equiv \psi_r(\pi,t) \end{aligned}$$

Proof.

For any two states t, t′

$$\Delta(t,t') \geq C\left(\pi_{t'}^*,t'\right)-C\left(\pi_{t'}^*,t\right)$$

since $\pi_{t'}^*$ is a suboptimal policy from initial state t. We now proceed along the lines of the proof of Theorem 6.

Comments.

Observe from Theorems 6 and 7 that each of the terms of the expressions for $\phi_r(\pi,t)$ and $\psi_r(\pi,t)$ is in itself a product of two quantities. The first of these is either the expectation or standard deviation of $\alpha^{\sum_{m=1}^{r} Y_m}$ and relates to the amount of discounting from the implementation of policy π for r stages. Each of these terms must converge to 0 as r→∞ uniformly over π and t which in turn ensures the same for both $\phi_r(\pi,t)$ and $\psi_r(\pi,t)$.

The second quantity in each term relates to the spread of $\hat{C}(\bar{\pi},\theta)$ for some policy $\bar{\pi}$ where θ is sampled from $G(\cdot\,|\,t)$. It is reasonably clear that such quantities will usually be related to the spread of $G(\cdot\,|\,t)$ itself. Consider now two special cases:

Case 1. $G \in \Theta^*$ **is a two-point prior.** This property must be shared by each posterior $G(\cdot\,|\,t)$. We write

$$G(\theta_1|t) = p_t = 1 - G(\theta_2|t) \quad \text{where } \theta_1, \theta_2 \in \Theta.$$

If $\Theta \subseteq \Re$ it is well known that the variance of this posterior is $(\theta_1-\theta_2)^2 p_t(1-p_t)$. For the second quantity in the expression for $\phi_r(\pi,t)$ it is easy to show that

$$\sqrt{\text{var}_t\left\{\hat{C}\left(\pi_t^*,\theta\right)\right\}} = \left|\hat{C}\left(\pi_t^*,\theta_1\right) - \hat{C}\left(\pi_t^*,\theta_2\right)\right|\sqrt{p_t(1-p_t)},$$

which is hence proportional to the standard deviation of the posterior.

It is not difficult to show that the second quantities in the two terms in $\psi_r(\pi,t)$ are proportional to $p_t(1-p_t)$ and $\sqrt{p_t(1-p_t)}$ respectively.

Case 2. We shall now assume that $\Theta \subseteq \Re$ together with sufficient regularity so that we can

(a) expand $\hat{C}(\bar{\pi},\theta)$ as a Taylor series in θ about the mean of the posterior distribution $G(\cdot\,|\,t)$ for appropriately chosen policies $\bar{\pi}$;

(b) take expectations term by term in the series.

Denoting the mean of $G(\cdot\,|\,t)$ by μ_t, we write

$$\hat{C}\left(\pi_t^*,\theta\right) = \hat{C}\left(\pi_t^*,\mu_t\right) + \sum_{n=1}^{\infty} C_n(t)(\theta-\mu_t)^n.$$

Inserting this expression into (22) and taking expectations term by term we deduce that

$$\Delta_r(\pi,t)\{1-M_r(\pi,t)\} \le \sum_{n=1}^{\infty} C_n(t) \sqrt{\mathrm{var}_{t,\pi,\theta}\left(\alpha^{\sum_{m=1}^{r} Y_m}\right)} \sqrt{\mathrm{var}_t\left\{\left(\theta - \mu_t\right)^n\right\}} \equiv \hat{\phi}_r(\pi,t).$$

In the expression for $\hat{\phi}_r(\pi,t)$ the dependence upon the spread of G(· | t) is now explicit. (Note that, if $\hat{C}\left(\pi_t^*, \theta\right)$ is close to linear at μ_t then the n = 1 term in $\hat{\phi}_r(\pi,t)$ may be an approximate upper bound for $\Delta_r(\pi,t)\{1-M_r(\pi,t)\}$.)

The equivalent analysis applied to Theorem 7 yields

$$\Delta_r(\pi,t)\{1-M_r(\pi,t)\} \ge \sum_{n=1}^{\infty} \sqrt{\mathrm{var}_{t,\pi,\theta}\left(C_n(T_{r+1})\alpha^{\sum_{m=1}^{r} Y_m}\right)} \sqrt{\mathrm{var}_t\left\{\left(\theta - \mu_t\right)^n\right\}} \equiv \hat{\psi}_r(\pi,t).$$

and similar comments apply.

Now we draw together Corollary 5 with Theorems 6 and 7 to yield an evaluation of cost-rate heuristic $\hat{\pi}(\underline{r})$ in terms of the functions ϕ_r and ψ_r. Before doing so, we write U_n as the state of the process following $\sum_{m=1}^{n} r_m$ decisions and transitions under policy $\hat{\pi}(\underline{r})$-i.e.,

$$U_n = T_{N(n)+1}, \text{where } N(n) = \sum_{m=1}^{n} r_m.$$

Otherwise, the notation is as established in Sections 2 and 3.

Theorem 8. For each sequence $\underline{r}$ and initial state t

$$C\{\hat{\pi}(\underline{r}),t\}-C(\pi^*,t)\le E_{\hat{\pi}(\underline{r})}\left(\sum_{n=0}^{\infty}\left[\phi_{r_{n+1}}(\hat{\pi}_{n+1},U_n)-\psi_{r_{n+1}}(\pi^*,U_n)\right.\right.$$

$$\left.\left.\times\left\{1-M_{r_{n+1}}(\hat{\pi}_{n+1},U_n)\right\}\left\{1-M_{r_{n+1}}(\pi^*,U_n)\right\}^{-1}\right]\alpha^{\sum_{m=1}^{N(n)}Y_m}\,\Big|\,T_1=t\right)$$

Recall Theorem 4, Corollary 5 and the comments thereafter. Making the computationally simple choice $r_n = 1$, $n \ge 2$, the question raised there concerned how large r_1 needed to be for cost-rate heuristic $\hat{\pi}(\underline{r})$ to be close to optimal for some initial state t. In view of Theorems 6 and 7 and the above comments, it is clear that for the class of Bayesian sequential decision problems under discussion the answer to that question will be related to the spread of $G(\cdot\,|\,t)$. The proof of Theorem 9 (which asserts the asymptotic optimality of all fixed sequence cost-rate heuristics as the variance of $G(\cdot\,|\,t)$ goes to zero) contains calculations which shed light on such matters.

Theorem 9. If

(i) $\Theta \subseteq \Re$;

(ii) $X_1, X_2, \ldots$ have density $f(x,\theta)$ such that $\frac{\partial}{\partial\theta} f(x,\theta)$ exists and is continuous everywhere, and

(iii) $E_\theta\left(\left[\frac{\partial}{\partial\theta}\{\ln f(X_1,\theta)\}\right]^2\right)$ is bounded for $\theta \in \Theta$, then for any fixed sequence cost-rate heuristic $\hat{\pi}(\underline{r})$

$$C\{\hat{\pi}(\underline{r}),t\}-C(\pi^*,t)\to 0, \text{ as } \mathrm{var}_t(\theta)\to 0.$$

Proof.

See Theorem 8. Under the stated conditions we will show that the first term in the upper bound given there is

$$E_{\hat{\pi}(\underline{r})}\left(\sum_{n=0}^{\infty}\phi_{r_{n+1}}(\hat{\pi}_{n+1},U_n)\left\{1-M_{r_{n+1}}(\hat{\pi}_{n+1},U_n)\right\}\left\{1-M_{r_{n+1}}(\pi^*,U_n)\right\}^{-1}\times\alpha^{\sum_{m=1}^{N(n)}Y_m}\,|T_1=t\right)$$

$$\equiv\phi\left\{\hat{\pi}(\underline{r}),t\right\} \quad (23)$$

goes to zero as $\text{var}_t(\theta)$ goes to zero. The analysis for the second term is very similar.

Utilizing the definition of $\phi_r(\pi,t)$ in the statement of Theorem 6 we deduce that

$$\phi\left\{\hat{\pi}(\underline{r}),t\right\}=E_{\hat{\pi}(\underline{r})}\left[\sum_{n=0}^{\infty}\sqrt{\text{var}_{U_n,\hat{\pi}_{n+1},\theta}\left(\alpha^{\sum_{m=N(n)+1}^{N(n+1)}Y_m}\right)}\sqrt{\text{var}_{U_n}\left\{\hat{C}\left(\pi^*_{U_n},\theta\right)\right\}}\right.$$

$$\left.\times\alpha^{\sum_{m=1}^{N(n)}Y_m}\left\{1-M_{r_{n+1}}(\hat{\pi}_{n+1},U_n)\right\}\left\{1-M_{r_{n+1}}(\pi^*,U_n)\right\}^{-1}|T_1=t\right]$$

$$\le k_1E_{\hat{\pi}(\underline{r})}\left[\sum_{n=0}^{\infty}\sqrt{\text{var}_{U_n}\left\{\hat{C}\left(\pi^*_{U_n},\theta\right)\right\}}\cdot\alpha^{\sum_{m=1}^{N(n)}Y_m}\cdot\left\{1-\varepsilon+\varepsilon\alpha^{\delta}\right\}^{\frac{N(n+1)-N(n)}{2}}|T_1=t\right] \quad (24)$$

where k_1 depends neither upon $\underline{r}$ nor t. To achieve inequality (24) we simply note that

$$\operatorname{var}_{U_n,\hat{\pi}_{n+1},\theta}\left(\alpha^{\sum_{m=N(n)+1}^{N(n+1)} Y_m}\right) \le \left\{1-\varepsilon+\varepsilon\alpha^{\delta}\right\}^{N(n+1)-N(n)}$$

and, from (3) that for all choices of r, π and x

$$M_r(\pi,x) \le \left(1-\varepsilon+\varepsilon\alpha^{\delta}\right) < 1.$$

In order to bound (24) we shall require a Taylor series expansion for $\hat{C}(\pi,\theta)$ for suitably chosen π. To that end we note from (19) that for general π,θ we can write

$$\hat{C}(\pi,\theta) - E_{\pi,\theta}\left\{\sum_{n=1}^{\infty} \alpha^{\sum_{m=1}^{n-1} Y_m} \hat{c}(Y_n,\pi_n)\right\}$$

$$= \sum_{n=1}^{\infty} \int \cdots \int \alpha^{\sum_{m=1}^{n-1} \Phi(x_m,\pi_m)} \hat{c}\{\Phi(x_n,\pi_n),\pi_n\} \prod_{i=1}^{n} f(x_i,\theta) \underline{dx},$$

invoking the boundaries of costs. Upon making use of assumption (ii) above and standard arguments we deduce that

$$\frac{\partial}{\partial\theta}\hat{C}(\pi,\theta) = \sum_{n=1}^{\infty} \int \cdots \int \alpha^{\sum_{m=1}^{n-1} \Phi(x_m,\pi_m)} \hat{c}\{\Phi(x_n,\pi_n),\pi_n\} \times \sum_{i=1}^{n} \frac{\frac{\partial}{\partial\theta} f(x_i,\theta)}{f(x_i,\theta)} \left\{\prod_{j=1}^{n} f(x_j,\theta)\right\} \underline{dx}$$

from which it follows immediately that

$$\left|\frac{\partial}{\partial\theta}\hat{C}(\pi,\theta)\right| \le \sum_{n=1}^{\infty} E_{\pi,\theta}\left\{\alpha^{\sum_{m=1}^{n-1} Y_m} \hat{c}(Y_n,\pi_n) \sum_{i=1}^{n} \left|\frac{\partial}{\partial\theta} f(X_i,\theta)\right|\right\}$$

and therefore, utilizing the boundedness of costs we infer that

$$\left|\frac{\partial}{\partial\theta}\hat{C}(\pi,\theta)\right| \le k_2 \sum_{n=1}^{\infty} E_\theta\left\{\alpha^{\sum_{m=1}^{n-1} Z_m} \sum_{i=1}^{n}\left|\frac{\partial}{\partial\theta}\ln f(X_i,\theta)\right|\right\} \qquad (25)$$

where $Z_m = \min_{1\le j\le N} \Phi\left(X_m, a_j\right)$ and k_2 depends upon neither π and θ.

It follows from the independence of the X_i's that Z_i is independent of Z_j and X_j, $i \ne j$, and hence we have that

$$E_\theta\left\{\alpha^{\sum_{m=1}^{n-1} Z_m} \sum_{i=1}^{n}\left|\frac{\partial}{\partial\theta}\ln f(X_i,\theta)\right|\right\} =$$

$$E_\theta\left(\alpha^{\sum_{m=1}^{n-1} Z_m}\right) E_\theta\left\{\left|\frac{\partial}{\partial\theta}\ln f(X_i,\theta)\right|\right\} + \sum_{i=1}^{n-1} E_\theta\left(\alpha^{\sum_{m=1, m\ne i}^{n-1} Z_m}\right) E_\theta\left\{\alpha^{Z_i}\left|\frac{\partial}{\partial\theta}\ln f(X_i,\theta)\right|\right\}$$

$$= \left(E_\theta\alpha^{Z_1}\right)^{n-1} E_\theta\left\{\left|\frac{\partial}{\partial\theta}\ln f(X_i,\theta)\right|\right\} + (n-1)\left(E_\theta\alpha^{Z_1}\right)^{n-2} E_\theta\left\{\alpha^{Z_i}\left|\frac{\partial}{\partial\theta}\ln f(X_i,\theta)\right|\right\}$$

$$\le n\left(E_\theta\alpha^{Z_1}\right)^{n-2}\left[\left(E_\theta\alpha^{Z_1}\right) E_\theta\left\{\left|\frac{\partial}{\partial\theta}\ln f(X_1,\theta)\right|\right\} + E_\theta\left\{\alpha^{Z_1}\left|\frac{\partial}{\partial\theta}\ln f(X_1,\theta)\right|\right\}\right]$$

line these up ⟵

$$\le n\left(E_\theta\alpha^{Z_1}\right)^{n-2}\left\{\sqrt{E_\theta\left(\alpha^{2Z_1}\right)}\sqrt{E_\theta\left[\left\{\frac{\partial}{\partial\theta}\ln f(X_1,\theta)\right\}^2\right]}\right\}$$

by Cauchy-Schwarz. We now recall assumption (iii) and conclude upon substitution into (25) that

$$\left|\frac{\partial}{\partial\theta}\hat{C}(\pi,\theta)\right| \le k_3$$

where k_3 depends neither upon π nor θ.

Recall (24). Taking a Taylor series expansion of $\hat{C}\left(\pi^*_{U_n},\theta\right)$ about μ_{U_n}, the mean of $G(\cdot \mid U_n)$, we obtain

$$\sqrt{\mathrm{var}_{U_n}\left\{\hat{C}\left(\pi^*_{U_n},\theta\right)\right\}} = \sqrt{\mathrm{var}_{U_n}\left(\theta-\mu_{U_n}\right)\left\{\frac{\partial}{\partial\theta}\hat{C}(\pi,\theta)\bigg|_{\theta=\tilde{\theta}}\right\}}$$

$$\leq \sqrt{E_{U_n}\left[\left(\theta-\mu_{U_n}\right)\left\{\frac{\partial}{\partial\theta}\hat{C}(\pi,\theta)\bigg|_{\theta=\tilde{\theta}}\right\}\right]} \leq k_3\sqrt{\mathrm{var}_{U_n}(\theta)}$$

Since in the above $\tilde{\theta}$ always has been θ and μ_{U_n}. Upon substitution into (24) and use of standard arguments we deduce that

$$\phi\{\hat{\pi}(\underline{r}),t\} \leq k_4 E_{\hat{\pi}(\underline{r})}\left[\sum_{n=0}^{\infty}\sqrt{\mathrm{var}_{U_n}(\theta)}\cdot\alpha^{\sum_{m=1}^{N(n)}Y_m}\cdot\left\{1-\varepsilon+\varepsilon\alpha^{\delta}\right\}^{\frac{N(n+1)-N(n)}{2}}\bigg|T_1=t\right]$$

$$\leq k_4\sum_{n=0}^{\infty}\sqrt{E_{\hat{\pi}(\underline{r})}\left[\left\{\mathrm{var}_{U_n}(\theta)\right\}\big|T_1=t\right]}\cdot\sqrt{E_{\hat{\pi}(\underline{r})}\left\{\alpha^{\sum_{m=1}^{N(n)}Y_m}\big|T_1=t\right\}}$$

$$\times\left\{1-\varepsilon+\varepsilon\alpha^{\delta}\right\}^{\frac{N(n+1)-N(n)}{2}}$$

$$\leq k_4\sum_{n=0}^{\infty}\sqrt{\mathrm{var}_t(\theta)}\cdot\left\{1-\varepsilon+\varepsilon\alpha^{\delta}\right\}^{\frac{N(n+1)}{2}} \tag{26}$$

These should be lined up

$\rightarrow 0$, as $\mathrm{v\grave{a}r}_t(\theta) \rightarrow 0$ since k_4 depends upon neither $\underline{r}$ nor t. Please note that inequality (26) is obtained by means of standard conditioning arguments.

A similar argument for the second term in the upper bound of Theorem 8 completes the proof.

Comments.

In part answer to the question raised in the paragraph preceding the statement of Theorem 9 concerning the size of r_1 needed (when, say, $r_n = 1, n \geq 2$) for $\hat{\pi}(\underline{r})$ to perform well, consider the bound (26). Since in this case $N(n+1) = r_1 + n, n \geq 0$, we obtain a bound of the form

$$k\left\{1-\varepsilon+\varepsilon\alpha^{\delta}\right\}^{r_1/2}\sqrt{\mathrm{var}_t(\theta)}$$

where k depends upon neither r_1 nor t. Plainly the larger the value of $\mathrm{var}_t(\theta)$ the larger the value of r_1 needed to make this expression small. In general we may regard the bound obtained at the end of the above proof as representing a trade-off between the amount of discounting available from the choice of sequence $\underline{r}$ and the amount of prior information about θ.

5. A SIMPLE MODEL OF PREVENTATIVE MAINTENANCE

A system is subject to random deterioration and failure. A new system is installed at time 0 and (in the absence of intervention) its time to failure has distribution F_θ where $\theta \in \Theta$ is unknown. Replacing a failed system is expensive. At time t the cost is $\alpha^t c_1$ where as usual $\alpha \in [0,1)$ is a discount rate. Alternatively, a (less expensive) planned replacement can be made in advance of system failure — here the cost at t is $\alpha^t c_2$.

Hence at time 0, one of N possible (planned) replacement times $0 < a_1 < a_2 < ... < a_N$ must be chosen. Note that we might have $a_N = \infty$, i.e., the choice of such an a_N implies that the system is left to fail with no planned replacement in anticipation of failure. We have $X_1, X_2, ...$ a sequence of i.i.d. system failure times with $X_i \sim F_\theta$. If action a_i is taken at 0, a planned replacement occurs at a_i if

$X_1 > a_i$ and otherwise the system is replaced at failure. At time $Y_1 = \min(X_1, a_i)$ one of the N replacement times $\{a_j, 1 \le j \le N\}$ is chosen for the new system. We proceed in this fashion. Choosing replacement times which are too small incurs unnecessary costs from a surfeit of planned replacements. Replacement times which are too large carry the risk of large numbers of expensive replacements upon failure of the system. We suppose that θ has a prior distribution G and look for a Bayes sequential decision rule for this problem.

Our replacement problem is a simple instance of the class discussed in the previous section. We shall assume (i) – (v) of Section 4 along with the additional measurability requirements of Section 2, together with Condition 1. This problem also (in common with, say, bandit problems) presents in a simple way the tension between taking decisions whose prime purpose is to gain information (and hence improve the quality of future decisions) and taking decisions which exploit the information already available.

More elaborate versions of this problem are discussed for models with known stochastic structure (i.e., known θ) by Aven (1983) and Chen and Savits (1988). For example, Aven (1983) studies a system whose failure rate is a nonnegative, progressively measurable stochastic process. Further, all costs are random variables. Now, for our model with θ known it is clear that cost rate myopic policies are optimal. To see this, take $r = 1$ in Theorem 4 and note that all speed functions are zero. Hence an optimal policy for known θ always chooses a_i to minimize

$$\left[\int_0^{a_i} \alpha^t c_1 F_\theta(dt) + \alpha^{a_i} c_2 F_\theta\{[a_i,\infty]\}\right]\left(1 - \int_0^{a_i} \alpha^t F_\theta(dt) - \alpha^{a_i} F_\theta\{[a_i,\infty]\}\right)^{-1} \quad (27)$$

Indeed both Aven (result R1, 1983) and Chen and Savits (Theorem 3.9, 1988) analyze their systems according to cost rates. Aven is able to proceed to recover opimal policies of simple structure. A later paper by Aven and Bergman (1986) presents some results which draw together the discounted cost case with that incorporating average cost per unit time.

Attempts at learning about such a system have usually been structured according to partially observable Markov Decision Processes. See Albright (1978) and White (1979) for important contributions along these lines. In models with the average cost per unit time criterion, Bather (1977), Frees and Ruppert (1985) and Aras and Whitaker (1990) have taken non-Bayesian and nonparametric approaches to learning about the underlying system.

In our Bayesian model a cost rate myopic policy will no longer usually take a single fixed action at all decision epochs, in contrast to (27). If the current posterior for θ is $\bar{G}$ a cost rate myopic policy chooses a_i to minimize

$$\left[\int_{\Theta}\left(\int_0^{a_i}\alpha^t c_1 F_\theta(dt)+\alpha^{a_i}c_2F_\theta\{[a_i,\infty)\}\right)\bar{G}(d\theta)\right]$$
$$\times\left(1-\int_{\Theta}\left(\int_0^{a_i}\alpha^t c_1 F_\theta(dt)+\alpha^{a_i}c_2F_\theta\{[a_i,\infty)\}\right)\bar{G}(d\theta)\right)^{-1}. \qquad (28)$$

Hence cost rate myopic policies are adaptive, depending as they do upon the current posterior for θ. Executing the minimization in (28) is usually computationally trivial, rendering this class of policies attractive as heuristics.

In the Bayesian context cost rate myopic policies are no longer optimal in general. Results in Section 4 give us guidance concerning when they are guaranteed to perform well. In particular this happens when the spread of prior G is small and/or when substantial discounting takes place from one decision to the next. When these conditions are not satisfied we may need to consider a cost-rate heuristic $\hat{\pi}(\underline{r})$ with $r_1 > 1$, $r_n = 1$, $n \geq 2$. Corollary 5 assures us that this class is rich enough. We now present some computational results bearing upon these phenomena.

Consider a replacement problem with $c_1 = 10$, $c_2 = 1$ and $\alpha = 0.99$. Failure times are assumed to be independent Weibull $(n, 0.4)$ random variables, i.e., having density

$$f(x; n, \lambda) = \lambda n x^{n-1} \exp(-\lambda x^n), \quad x > 0$$

with $\lambda = 0.4$. G is a two point prior with

$$G(n_1) = p = 1 - G(n_2) \tag{29}$$

where $n_1 = 1$ and $n_2 = 8$. At each decision epoch we are faced with a choice between $N = 50$ planned replacement times given by

$$a_j = 1.0 + (j-1)0.04, \quad 1 \leq j \leq 50.$$

We restrict discussion to fixed sequence cost-rate heuristics $\hat{\pi}(\underline{r})$ with $r_n = 1$, $n \geq 2$. The discussion following Theorems 6 and 7 in Section 4 (see especially Comment 1, Case 1) leads us to expect that most is to be gained by choosing a heuristic with large r_1 when the prior variance is large.

For simplicitiy of notation, denote by $C(p)$ the Bayes cost incurred when adopting an optimal policy with prior distribution (29) and $C_m(p)$ the equivalent cost from adopting $\hat{\pi}(\underline{r})$ with $r_1 = m$; $r_n = 1$, $n \geq 2$. The (m, p)-optimal policy which constitutes the first stage of $\hat{\pi}(\underline{r})$ is calculated according to the computational procedure derived from Theorem 1. It may be of interest to note that in this procedure the number of calculations per iteration grows linearly in m. The computation of $(1, p)$-optimal policies is trivial. The costs $C(p)$, $C_m(p)$ are computed by value iteration or some simple variant of it.

In Tables 1 and 2 find values of the absolute differences $C_m(p) - C(p)$, $m = 1, 2, 3$ and the relative differences $\{C_m(p) - C(p)\}\ \{C(p)\}^{-1}$, for $m = 1, 2, 3$, and $p = 0(0.1)1$. Figures 1 and 2 present these data graphically.

TABLE 1. ABSOLUTE DIFFERENCES BETWEEN THE COST FROM HEURISTIC $\hat{\pi}(r)$ AND AN OPTIMAL POLICY

p	$C_1(p) - C(p)$	$C_2(p) - C(p)$	$C_3(p) - C(p)$
0.0	0.000	0.000	0.000
0.1	1.097	0.693	0.404
0.2	1.941	1.225	0.716
0.3	2.650	1.672	0.978
0.4	3.232	2.037	1.195
0.5	3.714	2.338	1.376
0.6	4.023	2.536	1.487
0.7	4.073	2.579	1.494
0.8	3.772	2.396	1.376
0.9	2.686	1.736	0.950
1.0	0.000	0.000	0.000

TABLE 2. RELATIVE PERCENTAGE DIFFERENCES BETWEEN THE COST FROM HEURISTIC $\hat{\pi}(r)$ AND AN OPTIMAL POLICY

p	$100\{C_1(p) - C(p)\}\{Cp)\}^{-1}$	$100\{C_2(p) - C(p)\}\{Cp)\}^{-1}$	$100\{C_3(p) - C(p)\}\{Cp)\}^{-1}$
0.0	0.000	0.000	0.000
0.1	1.022	0.646	0.377
0.2	1.396	0.881	0.515
0.3	1.551	0.978	0.572
0.4	1.593	1.004	0.589
0.5	1.581	0.995	0.586
0.6	1.506	0.949	0.557
0.7	1.360	0.861	0.499
0.8	1.136	0.721	0.414
0.9	0.735	0.475	0.260
1.0	0.000	0.000	0.000

Figure 1. Absolute Differences between the Cost from Heuristic $\hat{\pi}(r)$ and an Optimal Policy

Figure 2. Relative Differences between the Cost from Heuristic $\hat{\pi}(\underline{r})$ and an Optimal Policy

If, for example, we wished to choose a heuristic $\hat{\pi}(\underline{r})$ whose Bayes' cost is within 1% of the optimum then, from Table 2, choosing $r_1 = 1$ would suffice for $p = 0, 0.9, 1.0$; choosing $r_1 = 2$ would suffice for $p = 0.1, 0.2, 0.3, 0.5, 0.6, 0.7$ and 0.8 but we would need $r_1 = 3$ to attain this level of performance when $p = 0.4$. This pattern of behavior is what Section 4 would lead us to expect. One striking feature of our numerical study of this replacement problem is the consistently strong performance of the cost-rate myopic policy with $r_1 = 1$. In Figures 3 and 4 find values of $C_1(p) - C(p)$ for the problem described above but with discount rate now taken to be $\alpha = 0.95$ and a range of repair costs $c_1 = 5(1)10$. Figure 3 is for a case with small prior variance ($p = 0.1$) and Figure 4 for large prior variance ($p = 0.5$). It seems that the simple cost-rate myopic policy will deliver adequate performance for our replacement problem much of the time.

Figure 3. Absolute differences between the cost from the cost-rate myopic policy and an optimal policy when $p = 0.1$

Figure 4. Absolute differences between the cost from the cost-rate myopic policy and an optimal policy when $p = 0.5$

REFERENCES

Albright, S. C. (1978) Structural results for partially observable Markov decision processes, *Opns. Res.*, **27**, 1041-1053.

Aras, G. and Whitaker, L. R. (1990) *Sequential Nonparametric Estimation of an Age Replacement Policy*, Technical Report, Department of Operations Research, Naval Postgraduate School, Monterey, California.

Aven, T. (1983) Optimal replacement under a minimal repair strategy — a general failure model, *Adv. Appl. Prob.*, **15**, 198-211.

Aven, T. and Bergman, B. (1986) Optimal replacement times — a general set-up, *J. Appl. Prob.*, **23**, 432-442.

Bather, J. A. (1977) On the sequential construction of an optimal age replacement policy, *Bull. Inst. Int. Stat.*, **47**, 253-266.

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

Chen, C. S. and Savits, T. H. (1988) A discounted cost relationship, *J. Mult. Anal.*, **27**, 105-115.

Ferguson, T. S. (1967) *Mathematical statistics—a decision-theoretic approach*, Academic Press.

Frees, E. and Ruppert, D. (1985) Sequential nonparametric age replacement policies, *Ann. Stat.*, **13**, 650-662.

Gittins, J. C. (1989) *Multi-armed bandit allocation indices*, Wiley.

Glazebrook, K. D. (1991) Strategy evaluation for stochastic scheduling problems with order constraints, *Adv. Appl. Prob.* (to appear)

Howard, R. (1960) *Dynamic programming and Markov processes*, M.I.T. Press.

Katehakis, M. N. and Veinott, A. F. (1987) The multi-armed bandit problem: decomposition and computation, *Math. Opns. Res.*, **12**, 262-268.

Porteus, E. (1980) Overview of iterative methods for discounted finite Markov and semi-Markov decision chains, in *Recent developments in Markov decision processes*, R. Hartley, L. C. Thomas and D. J. White (eds.), 1-20.

Ross, S. M. (1970) *Applied probability models with optimization applications*, Holden-Day.

White, C. C. (1979) Bounds on optimal cost for a replacement problem with partial observations, *Nav. Res. Logist. Qu.*, 26, 415-422.