A NEW OPTIMALITY CRITERION FOR NON-HOMOGENEOUS MARKOV DECISION PROCESSES

Wallace J. Hopp
James C. Bean
Robert L. Smith

Technical Report 84-24
Revised June, 1986
Revised December, 1986

A NEW OPTIMALITY CRITERION FOR NON-HOMOGENEOUS MARKOV DECISION PROCESSES [†]

Wallace J. Hopp
Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, Illinois 60201

James C. Bean
Robert L. Smith
Department of Industrial and Operations Engineering
The University of Michigan
Ann Arbor, Michigan 48109

ABSTRACT

We propose a new definition of optimality for non-homogeneous Markov decision processes called periodic forecast horizon (PFH) optimality. Using measures of discounting and ergodicity we establish conditions under which PFH optimal strategies exist and PFH optimality implies the more conventional notions of α-optimality and average optimality. Finally, we use this definition of optimality as the basis for a direct development of forecast horizon results for discounted and undiscounted non-homogeneous Markov decision processes.

A majority of the work on Markov decision processes deals with the stationary or homogeneous case. It is assumed that at each decision epoch the future process is stochastically identical to the process encountered at time 0. For many important applications, however, such as R & D modeling (Nelson and Winter [1982]), capacity expansion (Freidenfelds [1981], Luss [1982]), equipment replacement (Lohmann [1984]) and inventory control (Sobel [1971]) the appropriate models are Markovian but not stationary. At each decision epoch a different problem is encountered.

We consider a framework for solving non-homogeneous Markov decision processes. To solve a problem of this type requires a more general definition of an optimal solution and a more robust optimal solution seeking technique. Unlike the stationary case, in its most general form this problem can neither be stated in finite information nor solved in finite time.

Several optimality criteria have been proposed for non-homogeneous Markov decision processes. For problems where the value function converges for all feasible strategies, optimality is logically

[†] This material is based on work supported by the National Science Foundation under Grant No. ECS-8409682.

defined by maximum total reward, which defines strategy x^* to be *optimal* if the total reward for x^* is not less than the total reward for any other strategy in the strategy set X. This will be formally defined in Section 1.1.

Since these problems have infinite time horizons, often the rewards of the optimal strategies are infinite. Proposed optimality criteria for the divergent reward case include greatest average reward (Derman [1966], Ross [1968]), overtaking optimality (Denardo and Rothblum [1979]), 1–optimality (Blackwell [1962]), and average overtaking optimality (Veinott [1966]). Ideally, the definition of optimality should be neither overselective nor underselective. An *overselective* optimality criterion may define some intuitively optimal strategies to be suboptimal and can result in no optimal strategy existing. An *underselective* criterion may accept strategies as optimal that should clearly be considered suboptimal. In addition to being appropriately selective, we would like our optimality criterion to be consistent with practical methods for computing optimal strategies.

When reward functions diverge, the most commonly used definition of optimality is greatest average reward. A strategy x^* is defined to be *average optimal* if the limit as $N \to \infty$ of the average reward per period for x^* is not less than the limit of the average reward for any other strategy in X.

Average optimality is tail driven so that the quality of any strategy is determined independently of any finite leading segment of the strategy. For example, a strategy that produces rewards of -1000 in periods 0 through N and 1000 in every period beyond N, for any finite N, is equivalent under the average optimality criterion to a strategy which has a reward of 1000 in every period. Since average optimality does not distinguish between these two strategies, it is an underselective criterion. This underselectivity does not generally present difficulties in homogeneous problems since under a stationary strategy the expected reward will be the same in each period. That is, the tail is an identical replication of the original problem. In non–homogeneous problems, stationary strategies will not generally be optimal so it is possible to identify an average optimal strategy that performs poorly in early periods. Since in practical problems early performance is crucial, we need a stronger criterion than average optimality to rule out these strategies.

Overtaking optimality is often used as an alternative to average optimality. A strategy is overtaking optimal if there is some finite time, N^*, such that for any time horizon beyond N^* the strategy is optimal for that finite horizon problem. For example, suppose one strategy is optimal for all even time horizons and another strategy is optimal for all odd time horizons and that both strategies generate the same average reward per period, where average reward means expected reward over the horizon divided by the number of periods in the horizon. While it is reasonable to consider both strategies optimal in the infinite horizon problem, neither is overtaking optimal. Hence, a less stringent criterion is needed to guarantee existence of an optimal strategy.

The 1–optimality and average overtaking optimality criteria are more selective than average optimality because they emphasize returns in early periods. At the same time, they are less selective than overtaking optimality and therefore are not as likely to result in nonexistence of an optimal strategy. In the finite state homogeneous case these criteria can be shown to be equivalent (Denardo and Miller [1968]). While these criteria are consistent with computational techniques for homogeneous problems (Miller and Veinott [1969]), they do not lend themselves naturally to solution procedures for non–homogeneous problems. The optimality criterion introduced in this paper leads to straightforward solution techniques based on forecast horizons.

For homogeneous Markov decision problems, a variety of techniques, such as policy improvement, value iteration, and linear programming can be used to compute an optimal stationary strategy (Howard [1960], Manne [1960], Denardo [1967], Miller and Veinott [1969]). Since stationary strategies consist of a single policy repeated they are finitely expressible. Under certain conditions such as finite state and action spaces they are finitely computable.

In non–homogeneous problems attention cannot be restricted to stationary solutions. While many of the properties of homogeneous Markov decision processes can be generalized to the non–homogeneous case (Hinderer [1970]), this approach does not yield finite algorithms. Further, while a non–homogeneous problem can be converted into an equivalent homogeneous problem (Schäl [1975]), this can only be done at the expense of transforming a finite state space into a countably infinite one. Again, this cannot yield a finite algorithm.

Since an optimal solution to the infinite horizon non–homogeneous problem, however it is defined, cannot be found and stated in finite time, a surrogate for a finite optimal algorithm is necessary. Such a surrogate is suggested by the fact that only the first few decisions will be implemented immediately. The remainder of the strategy is found to evaluate future impacts of these first few decisions. This fact motivates the use of the forecast horizon approach commonly used in deterministic problems. A *forecast horizon* is a time horizon long enough to guarantee agreement in the optimal first decision for all longer horizons. We will show that, under certain regularity conditions, the first L optimal policy vectors, where L is any finite positive integer, can be found in finite time. Further, by recursively employing this fact in a rolling procedure, the entire optimal strategy can be recovered, though not in finite time.

The foundation for this work is Bean and Smith [1984] which proves the existence of forecast horizons for a wide class of deterministic sequential decision problems. The important concept in that paper is the assumed discounting of net cost flows. This discounting allows the present decision to become independent of decisions made sufficiently far into the future. Information beyond this future time can be discarded without effect on the initial decisions. This decreasing dependence will be referred to as *diminishing influence*. Bhaskaran and Sethi [1985] show that these results

can be simply extended to a stochastic case with discounting. Related observations concerning diminishing influence are also given in Morton [1979]. A much earlier paper by Shapiro [1968] used discounting to derive forecast horizon results for homogeneous Markov decision processes.

In this paper we generalize this concept of independence of the present and future decisions. The existence of a forecast horizon is seen to be an implication of a generalized type of ergodicity, which has precisely the same effect as that of discounting. The stochasticity present in a large class of Markov decision processes leads to this ergodicity. Consequently, forecast horizons can be shown to exist in the absence of discounting and even with negative discounting. Furthermore, ergodicity and discounting are seen to interact multiplicatively. The interaction of ergodicity and discounting was first noted in the context of Markov decision processes by Morton and Wecker [1977]. Their measure of ergodicity is primarily suited to homogeneous problems and is used to show convergence of the policies and relative value functions. This paper introduces a measure of ergodicity that applies to a much broader class of non-homogeneous problems and goes beyond convergence results to develop forecast horizon results.

Section one contains problem formulation, notation, and a discussion of PFH optimality. Section two includes results on weak ergodicity. The third section contains conditions for the existence of forecast horizons. Finally, section four includes extensions and conclusions.

1. Problem Formulation

1.1 Notation

We consider a process that is observed at discrete intervals to be in one of n states (some of which may be identical in order to ensure n states per stage). The decision-maker chooses a policy in stage k, x_k, by selecting actions, x_k^i, from finite sets, for states $i = 1, \ldots, n$. An infinite horizon strategy, x, consists of an infinite sequence of policies, one for each stage. The set of all feasible infinite horizon strategies is denoted by X.

Taking action x_k^i in state i of stage k results in an immediate reward, $r_k^i(x_k^i)$, and transition probabilities to all states in stage $k+1$, $p_k^{ij}(x_k^i)$, $j = 1, \ldots, n$. Let $\bar{R}$ denote an upper bound on the $r_k^i(x_k^i)$ which is assumed to be finite. We let $R_k(x_k)$ represent the $(n \times 1)$ column vector of rewards in stage k and $P_k(x_k)$ represent the $(n \times n)$ matrix of transition probabilities from all states in stage k to all states in stage $k+1$. Notice that both the rewards and transition probabilities may be stage dependent.

If strategy x is used and the one period discount factor is $0 \leq \alpha \leq 1$, the expected net present value at the beginning of stage k of the profit from period k through period $N, N > k$, is written $V_k(x_{(k)}; N)$, where $x_{(k)}$ represents the continuation of strategy x from stage k forward. By definition, $x_{(0)} = x$. The $V_k(\cdot)$ function maps into $\Re^n$ with the i^{th} element given by $V_k^i(x_k^i, x_{(k+1)}; N)$, which

represents the expected net present profit from stage k through stage N given the process is in state i of stage k. In general we are interested in the value function from stage zero onward, which can be written:

$$V_0(x;N) = \sum_{j=0}^{N} \alpha^j T_0^j(x) R_j(x_j)$$

where,

$$T_l^j(x) = \prod_{k=l}^{j-1} P_k(x_k),\ j > l$$
$$T_l^l(x) = I$$

In the finite horizon problem, optimization is equivalent to maximization of $V_0(x;N)$. Formally, we define x^* to be α*–optimal* if

$$\lim_{N\to\infty} V_0(x^*;N) - \lim_{N\to\infty} V_0(x;N) \geq 0,\ \forall x \in X.$$

This definition is valid as long as the limits exist. It is possible that $V_0(x;N)$ diverges with N, typically when $\alpha = 1$. We formally define x^* to be *average optimal* if

$$\liminf_{N\to\infty} \{ V_0(x^*;N) - V_0(x;N)\}/N \geq 0,\ \forall x \in X.$$

1.2 Periodic Forecast Horizon Optimality

If the optimal solutions to the finite horizon problems approach a particular infinite horizon strategy we would like to consider it as the optimal strategy. This idea has been discussed frequently in the forecast horizon literature (see Morton [1978]). We refer to such a solution as *forecast horizon optimal.* To formally define this notion we use a metric analogous to that presented in Bean and Smith. Let

$$\rho(x,\bar{x}) = \sum_{k=0}^{\infty} \Phi_k(x,\bar{x}) 2^{-(k+1)}$$

$$\text{where } \Phi_k(x,\bar{x}) = \begin{cases} 0 \text{ if } x_k^i = \bar{x}_k^i, i = 1,\ldots,n \\ 1 \text{ otherwise.} \end{cases}$$

The ρ metric has the property that any two strategies which agree in the first m policies are considered closer than any two strategies that do not. Any metric with this property would suffice equally well for the purposes of this paper.

Now we define x^* to be forecast horizon (FH) optimal if $x^*(N) \to x^*$ in the ρ metric as $N \to \infty$, where $x^*(N)$ is an α–optimal solution to the N stage problem. Unfortunately, $x^*(N)$ may not

converge due to the existence of multiple optima. Hence, we define here the weaker optimality criterion of *periodic forecast horizon* (PFH) optimality.

Definition: A strategy, x^* is *periodic forecast horizon* (PFH) optimal if there exists a subsequence of the integers, $\{N_m\}_{m=1}^{\infty}$, such that $x^*(N_m) \to x^*$ in the ρ metric as $m \to \infty$.

We assume that the strategy space, X, is compact in the metric space generated by ρ. This assumption precludes the possibility of a sequence of feasible strategies converging to an infeasible strategy. For further discussion of a related problem see Bean and Smith.

Theorem 1: A periodic forecast horizon optimal strategy exists for the non–homogeneous Markov decision process.

Proof: Compactness of X implies that the sequence $\{x^*(N)\}_{N=1}^{\infty}$ has a convergent subsequence. The limit of such a sequence is PFH optimal by definition.∎

We would like to have PFH optimality imply α–optimality if the value function is convergent and average optimality if the value function is divergent and an average optimal strategy exists. The first implication is a natural extension of the results of Bean and Smith and is stated as Theorem 2. The second implication requires the main theoretical result of this paper.

Theorem 2: If $\bar{R} < \infty$ and $\alpha < 1$ then PFH optimality implies α–optimality.

Proof: It is a simple matter to show that $V_0(x; N)$ is uniformly convergent on X as $N \to \infty$ (Hopp [1984]). Let x^* be PFH optimal. Choose a subsequence of integers, $\{N_m\}_{m=1}^{\infty}$, such that $x^*(N_m) \to x^*$ as $m \to \infty$. By definition,

$$V_0(x^*(N_m); N_m) \geq V_0(x; N_m), \ \forall x \in X.$$

Now

$$\lim_{m \to \infty} V_0(x^*(N_m); N_m) = \tilde{V}_0(x^*) < \infty,$$

where $\tilde{V}_0(x)$ is the expected net present value of strategy x over the infinite horizon. This follows from the uniform convergence of $V_0(x; N)$ over X. Hence,

$$\tilde{V}_0(x^*) \geq \tilde{V}_0(x), \ \forall x \in X,$$

so x^* is α–optimal.∎

Since we assume that $\bar{R} < \infty$, the profit functions can diverge only if $\alpha = 1$. In this case, we would like to show that PFH optimality implies average optimality. PFH optimality would then be stronger than average optimality since PFH optimality discourages leading sub–optimal policies. However, this implication is not true without additional conditions. We present a counterexample in the appendix where a PFH optimal strategy exists but is not average optimal. If this occurs, then

using the limiting strategy of a sequence of optimal finite horizon strategies might not maximize average returns.

In this counterexample the influence of the first policy extends infinitely far into the future. The same is true of the undiscounted "credit union" example given by Bean and Smith. Bean and Smith use discounting to resolve the difficulty presented by their counterexample. We will use the concept of *weak ergodicity* to develop conditions that rule out this type of behavior and ensure that PFH–optimal strategies are average optimal.

2. Diminishing Influence

To develop a measure of the diminishing influence that results from the stochasticity of non–homogeneous Markov decision processes, some results concerning forward products of stochastic matrices are needed.

If $\alpha = 1$ and strategy x is used, the expected reward in stage j is given by $T_0^j(x)R_j(x_j)$. To describe the diminishing influence of the first policy, x_0, on the reward in stage j we look at $T_0^j(x)$, the forward product of j stochastic matrices. The following results concerning $T_0^j(x)$ are taken from Seneta [1981].

2.1 Weak Ergodicity

Definition: A stochastic matrix is said to be *stable* if it has identical rows.

Definition: A scalar function, $\tau(\cdot)$, that is continuous on the set of $(n \times n)$ stochastic matrices (treated as points in $\Re^n \times \Re^n$) such that $0 \leq \tau(P) \leq 1$ for any stochastic matrix, P, is a *coefficient of ergodicity*. It is called *proper* if $\tau(P) = 0$ if and only if P is stable.

Definition: The Markov chain formed by strategy x achieves *weak ergodicity* if

$$\tau(T_l^j(x)) \to 0 \text{ as } j \to \infty, \forall l \geq 0,$$

where $\tau(\cdot)$ is a proper coefficient of ergodicity.

More informally, the forward product of any string of stochastic matricies in a weakly ergodic Markov chain will increasingly resemble a stable matrix as the number of matricies in the string increases. Note that unlike the conventional notion of ergodicity, called "strong" ergodicity for clarity, weak ergodicity does not require the forward product to converge to a limiting matrix. Indeed, the $T_l^j(x)$ may differ significantly as j increases due to the non–homogeneity of the problem. To be weakly ergodic, $T_l^j(x)$ need only have nearly identical rows for large j. In a homogeneous problem under a stationary strategy all transition matricies will be identical so that weak ergodicity and strong ergodicity are equivalent for this case.

We define the following coefficients of ergodicity for any $(n \times n)$ stochastic matrix.

$$\tau_1(P) = \max_{i,j}\{\sum_{s=1}^{n}(p_{is} - p_{js})/2\}$$
$$c(P) = 1 - \max_{s}(\min_{i} p_{is})$$

Lemma 3: For any stochastic matrix, P, and the infinite sequence of stochastic matrices defined by any strategy x, $\{P_k(x_k)\}_{k=0}^{\infty}$, the following are true:

(a) $\tau_1(\cdot)$ is a proper coefficient of ergodicity and $c(\cdot)$ is an improper coefficient of ergodicity

(b) $\tau_1(P) \leq c(P)$

(c) For any $l \geq 0$

$$\tau_1(T_l^j(x)) \leq \prod_{k=l}^{j-1}\tau_1(P_k(x_k)) \leq \prod_{k=l}^{j-1} c(P_k(x_k))$$

Proof: Seneta.

Theorem 4 (Seneta): Forward products of a sequence of stochastic matrices, $\{P_k(x_k)\}_{k=0}^{\infty}$, achieve weak ergodicity if

$$\sum_{k=0}^{\infty}\{1 - c[P_k(x_k)]\} = \infty.$$

vexit

Corollary 5: If $c(P_k(x_k)) \leq a_0 < 1$ for all $k = 0, 1, \ldots$, then weak ergodicity is achieved at a rate that is at least geometric with parameter a_0 (i.e., there exists an l such that $\tau_1(T_l^j(x)) \leq a_0^{j-l}$, $j \geq l$).

The probability distribution on the states in stage j starting from each of the n initial states is given by the $(n \times n)$ matrix $P_0(x_0)T_1^j(x)$. If $\tau_1(T_1^j(x)) \to 0$, so that $T_1^j(x)$ has increasingly identical rows as $j \to \infty$, then $P_0(x_0)T_1^j(x)$ approaches a matrix of identical rows regardless of what the $P_0(x_0)$ matrix is. *The probabilities on the states in stage j, and hence the expected rewards in stage j, become independent of the first policy as j grows large.*

Note that there are many other possible coefficients of ergodicity. Morton and Wecker define an alternate which has $a_0 = 1$ for most non-homogeneous problems. Hence, it does not lead to many of the results which follow.

2.2 Weakly Ergodic Markov Decision Processes

Theorem 4 and its corollary allow us to define a measure of the rate at which the Markov chain formed by a particular strategy achieves weak ergodicity. To provide a single quantitative measure of the minimum rate at which the Markov chain in a non-homogeneous Markov decision process achieves weak ergodicity, define a_0 as follows:

$$a_0 = \sup_{x \in X^*} \sup_{k} c(P_k(x_k)).$$

where X^* is the set of all potentially optimal x. In general, X^* is a subset of X that can be determined by additional problem structure. For example, if no such structure is available we take $X^* = X$. If $a_0 < 1$, then by Corollary 5 the Markov chain formed by any potentially optimal infinite horizon strategy achieves weak ergodicity at a rate that is at least geometric with parameter a_0. The condition $a_0 < 1$ essentially requires that under any potentially optimal strategy each transition matrix has a column with all entries greater than or equal to $1 - a_0 > 0$. This can be interpreted as requiring the existence of a uniformly accessible state in each stage such that the probability of reaching that state is at least $1 - a_0$ regardless of the state in the previous stage.

2.3 Assumptions

The following assumptions are made about the non-homogeneous Markov decision process:

(1) $\alpha a_0 < 1$

(2) At any decision point the choices available for each state are finite in number.

(3) $\bar{R} = \sup_{x \in X^*, k}\{\max_i |r_k^i(x_k^i)|\} < \infty$

(4) The strategy space, X, is compact in the metric space generated by ρ.

Assumption (1) requires either discounting, weak ergodicity, or both. Assumption (2) limits the problem to discrete, bounded decision variables. Assumption (3) requires an upper bound on the maximum reward obtainable at each stage.

2.4 Average Optimality

The result of Theorem 2 that PFH optimality implies α-optimality if $\alpha < 1$ is based on the uniform convergence of $V_0(x; N)$ on X. If $\alpha = 1$, then $V_0(x; N)$ is not necessarily convergent under assumptions (1) through (4). However, we can show that

$$f_k(x_k, \bar{x}_k, x_{(k+1)}; N) = V_k(x_k, x_{(k+1)}; N) - V_k(\bar{x}_k, x_{(k+1)}; N)$$

is uniformly convergent on the set of feasible $x_{(k+1)}$ and use this result to show that PFH optimality implies average optimality when $\alpha = 1$ under the assumptions in Section 2.3. The function $f_k(x_k, \bar{x}_k, x_{(k+1)}; N)$ represents the difference in the value function for stages k through N that results from using policies x_k versus $\bar{x}_k$ and then continuing with strategy $x_{(k+1)}$. Thus, f_k is a relative value function where the reference is determined by a fixed action, $\bar{x}_k$, rather than a fixed state as is usually done (Morton and Wecker, Ross).

Theorem 6: For any feasible x_k and $\bar{x}_k$ there exists a finite function, $\tilde{f}_k(x_k, \bar{x}_k, x_{(k+1)})$, such that

$$f_k(x_k, \bar{x}_k, x_{(k+1)}; N) = f_k(N) \to \tilde{f}_k = \tilde{f}_k(x_k, \bar{x}_k, x_{(k+1)})$$

uniformly on the set of feasible $x_{(k+1)}$ as $N \to \infty$.

Proof: We will proceed by establishing the Cauchy criterion for uniform convergence. Assume $N > k$. For any three stochastic matricies P_1, P_2 and T,

$$|(P_1 - P_2)TR_k(x_k)| \leq 2c(T)\bar{R}\ e,$$

where e is a column vector of ones (see Hopp [1984]). From Lemma 3 it can be shown that

$$c(T_{k+1}^{N+1}(x)) \leq a_0^{N-k}.$$

Combining these we get

$$|f_k(N+1) - f_k(N)| \leq (2\bar{R}\alpha(\alpha a_0)^{N-k})\ e.$$

This result can be extended additively for $k < M \leq N$ to

$$|f_k(M) - f_k(N)| < [\frac{2\bar{R}\alpha(\alpha a_0)^{N-k}}{(1-\alpha a_0)}]\ e.$$

By assumption, $(\alpha a_0) < 1$ so that this bound converges to zero with rate independent of x, $\bar{x}$, and $x_{(k+1)}$. Since the set of strategies is compact, this is sufficient to show that there exists a $\tilde{f}_k$ such that $f_k(N) \to \tilde{f}_k$ uniformly over all $x_{(k+1)}$.■

Intuitively, Theorem 6 implies that the effect on the reward in stage N from choosing x_k versus $\bar{x}_k$ in stage k becomes negligible as N gets large. Given this condition we can prove the main result, that PFH optimality implies average optimality.

Theorem 7: Under the assumptions (1) through (4), PFH optimality implies average optimality.

Proof: Let x^* be a PFH optimal strategy and x be any feasible strategy in X. Then there exists some infinite subsequence of the integers, $\{N_m\}_{m=1}^{\infty}$, such that $x^*(N_m) \to x^*$. By construction,

$$V_0(x^*(N_m); N_m) - V_0(x; N_m) \geq 0 \text{ for all } x \in X,$$

in particular, for $x = (x_0, x_1, \ldots, x_{k-1}, x_{(k)}^*)$. By structure of the metric ρ there exists some $\bar{m}$ such that $x_i^*(N_m) = x_i^*$ for all $m \geq \bar{m}$ and $i = 1, 2, \ldots, k-1$. Hence,

$$\lim_{m\to\infty} \{V_0(x_0^*, x_{(1)}^*(N_m); N_m) - V_0(x_0, x_{(1)}^*(N_m); N_m)\} \geq 0$$

for all feasible x_0. By Theorem 6, the expression inside the brackets is uniformly convergent and therefore,

$$\lim_{m\to\infty} \{V_0(x^*; N_m) - V_0(x_0, x_{(1)}^*; N_m)\}$$

exists and is non-negative. By the principle of optimality we also know that

$$\lim_{m\to\infty} \{V_1(x_{(1)}^*; N_m) - V_1(x_1, x_{(2)}^*; N_m)\} \geq 0.$$

Using this and the fact that

$$V_0(x; N_m) = R_0(x_0) + \alpha P_0(x_0)\, V_1(x_{(1)}; N_m)$$

inductively, we have that

$$\lim_{m\to\infty} \{ V_0(x^*; N_m) - V_0(x_0, x_1, \ldots, x_{k-1}, x^*_{(k)}; N_m)\} \geq 0.$$

for any finite k and $x \in X$. We can break this down to

$$\lim_{m\to\infty} \{ V_0(x^*; k-1) - V_0(x; k-1) + \alpha^k (T_0^k(x^*) - T_0^k(x)) \cdot V_k(x^*_{(k)}; N_m)\}$$

$$= V_0(x^*; k-1) - V_0(x; k-1) + \lim_{m\to\infty} \{\alpha^k (T_0^k(x^*) - T_0^k(x)) \cdot V_k(x^*_{(k)}; N_m)\} \geq 0.$$

Using the same uniform bound as in the proof of Theorem 6 this becomes

$$V_0(x^*; k-1) - V_0(x; k-1) + [\frac{2\bar{R}}{(1-\alpha a_0)}]e \geq 0.$$

Reparameterizing $N = k-1$, dividing by N, and taking the lim inf gives

$$\liminf_{N\to\infty} \frac{V_0(x^*; N) - V_0(x; N)}{N} \geq 0.$$

By definition, x^* is average optimal.∎

Theorems 2 and 7 establish the reasonableness of the PFH optimality criterion by verifying that PFH optimality implies α–optimality when $\alpha < 1$ and average optimality when $\alpha = 1$. The manner in which PFH optimality is defined makes the following forecast horizon results flow naturally.

Note that both discounting and weak ergodicity act geometrically to diminish the influence of the first policy on future rewards. For this reason, α and a_0 play analogous roles in the development of conditions for the existence of forecast horizons in the following section.

3. Conditions for the Existence of Forecast Horizons

This section develops results for non–homogeneous Markov decision processes parallel to those of Bean and Smith for deterministic problems. By using weak ergodicity, we obtain forecast horizon results for undiscounted, as well as discounted, problems. By use of forecast horizon results we can find PFH–optimal strategies when they are unique. This is an advantage relative to other criteria such as 1–optimality and average overtaking optimality.

Theorem 8: If the PFH optimal strategy is unique then $x^*(N) \to x^*$ in the ρ metric as $N \to \infty$.

Proof: Assume otherwise. Then there is an infinite subsequence of $\{x^*(N)\}_{N=1}^{\infty}$ that is bounded away from x^* in the ρ metric. Since X is compact this, in turn, has a convergent subsequence. By definition, its limit is also PFH optimal, a contradiction to the assumption of uniqueness.∎

Note that convergence of $x^*(N)$ to x^* implies that the optimal finite horizon strategy will agree with the optimal infinite strategy over an increasingly long initial sequence of policies. This fact follows from the way the ρ metric is defined and leads to the following forecast horizon results.

Theorem 9: If all PFH optimal strategies have the same first L policies then a forecast horizon exists leading to the optimal first L policies in the infinite horizon problem.

Proof: Similar to that of Theorem 8.

Corollary 10: If the PFH optimal strategy is unique, or all PFH optimal strategies have the same first policy, then a forecast horizon exists.

Theorems 8 and 9 and Corollary 10 show that in addition to assumptions (1) through (4), some form of uniqueness of the infinite horizon optimal policies is needed to yield sufficient conditions for the existence of a forecast horizon. Multiple optimal policies may cause cycling between the multiple optima as N increases. From the decision-maker's perspective it would be impossible to tell whether the cycling was due to multiple optima or to an insufficiently long time horizon. Hence, as in other optimization problems, degeneracy presents a serious theoretical problem in non-homogeneous Markov decision processes.

It is a simple matter to extend the results of Bean and Smith to show that if the set of potentially optimal strategies is countable then the PFH optimal strategy is unique for all α outside of a set of Lebesgue measure zero. In this case a forecast horizon exists.

The existence of forecast horizons allows computation of PFH-optimal strategies in a forward recursive manner. The first such technique is presented in Hopp [1985].

4. Extensions and Conclusions

We have developed sufficient conditions for the existence of forecast horizons in non-homogeneous Markov decision processes. These results form the stochastic counterparts to the forecast horizon results of Bean and Smith for deterministic problems and also represent the extension of Shapiro's work to the non-homogeneous and undiscounted problems. The difference between this and previous work is the role weak ergodicity plays in the existence of forecast horizons in non-homogeneous Markov decision processes. We have shown that weak ergodicity acts analogously to discounting in the existence conditions. Indeed, the discount factor interacts multiplicatively with an appropriately chosen numerical measure of the rate at which the Markov chain achieves weak ergodicity. This leads to the existence of forecast horizons in undiscounted as well as discounted problems.

Weak ergodicity yields insight into the underlying causes of forecast horizons that is not readily seen by investigating deterministic problems. We have shown that it is not discounting, per se, that generates forecast horizons, but rather, diminishing influence. If the influence of the first policy on

the reward in stage N decreases geometrically with N, then under a set of regularity conditions given in section 3, a forecast horizon will exist. Discounting produces this effect, as does weak ergodicity. Diminishing influence due to discounting occurs because the present value of the reward in stage N goes to zero geometrically with N, regardless of the choice of the first policy. Weak ergodicity causes diminishing influence because the probability distribution on the states in stage N becomes independent of the first policy as N grows large. Under the proper conditions, it does so geometrically.

The forecast horizon results of this paper are for a discrete time non–homogeneous Markov decision process with discrete decision variables. It is possible to relax the assumptions of discreteness of both time and the decision variables, although the results that are obtained are not as strong as those given above. A more detailed treatment is given in Hopp [1984]. The primary difference in the continuous time case is that the forecast horizons will be stated in terms of numbers of stages rather than numbers of periods. Since these stages have random lengths, the lengths of the forecast horizons are themselves random variables. In the continuous decision variable case forecast horizons leading to the optimal first policy no longer exits. However, horizons that yield first policies resulting in an error less that ϵ, for arbitrary ϵ, do exist for this case.

APPENDIX

This example presents a case where average optimal and forecast horizon optimal strategies exist but are not the same.

Suppose there are four states per stage, $\{1,2,3,4\}$. In all states in stage 0 there are four actions, $\{A,B,C,D\}$. In all subsequent stages action A is feasible in state 1, action B is feasible in state 2, action C is feasible in state 3, and action D is feasible in state 4. In addition, action B is feasible in state 1 of stage k for $k = 2^n - 1$, $n = 1,2,3,\ldots$ No other actions are feasible. Rewards and transitions are as follows.

$$r_0^i(x) = 0,\ i = 1,2,3,4,\ x = A,B,C,D$$

$$r_k^1(A) = r_k^1(B) = 1,\ k = 1,2,3,\ldots$$

$$r_k^2(B) = 2,\ k = 1,2,3,\ldots$$

$$r_k^3(C) = 1,\ k = 1,2,3,\ldots$$

$$r_k^4(D) = 5/4,\ k = 1,2,3,\ldots$$

$$p_0^{i1}(A) = p_0^{i2}(B) = p_0^{i3}(C) = p_0^{i4}(D) = 1,\ i = 1,2,3,4$$

$$p_k^{11}(A) = p_k^{33}(C) = p_k^{44}(D) = 1,\ k = 1,2,3,\ldots$$

$$p_k^{12}(B) = p_k^{23}(B) = 1,\ k = 2^n - 1,\ n = 1,2,3,\ldots$$

$$p_k^{22}(B) = 1,\ k \neq 2^n - 1,\ n = 1,2,3,\ldots$$

All other probabilities are zero. The problem is undiscounted so $\alpha = 1$.

Since transitions are deterministic, an α–optimal strategy in the problem from stage 0 through state N can be expressed as a simple sequence of $N+1$ actions. Since the four states in stage 0 are identical, $V_0^i(x^*(N);N)$ is the same for $i = 1,2,3,4$. α–optimal strategies and average reward for $i = 1,2,3,4$ are easily computed to be:

N	$x^*(N)$	$V_0^i(x^*(N);N)/N$
1	BB	2
2	ABB	3/2
3	$ABBB$	5/3
4	$ABBBC$	6/4
5	$AAABBB$	7/5
6	$AAABBBB$	9/6

⋮	⋮	⋮
10	$AAABBBBBCCC$	14/10
11	$AAAAAAABBBBB$	15/11

Continuing in this manner verifies that the strategy $\hat{x} = AAAA\ldots$ is forecast horizon optimal. However, it is also easily shown that

$$\liminf_{N\to\infty} \{ V_0(\bar{x}; N) - V_0(\hat{x}; N)\}/N = (1/4)e$$

where $\bar{x} = DDDD\ldots$ and e is the 4–vector of ones. Hence, $\bar{x}$ is preferred to $\hat{x}$ under the average optimality criterion. A unique forecast horizon optimal strategy exists but is not average optimal.

REFERENCES

Bean, J. and R. Smith [1984], "Conditions for the Existence of Planning Horizons," **Mathematics of Operations Research**, Vol. 9, pp. 391–401.

Bhaskaran, S. and S. Sethi [1985], "Conditions for the Existence of Decision Horizons for Discounted Problems in a Stochastic Environment: A Note," **Operations Research Letters**, Vol. 4, pp. 61–65.

Blackwell, D. [1962], "Discrete Dynamic Programming," **Annals of Mathematical Statistics**, Vol. 33, pp. 719–726.

Denardo, E. [1967], "Contraction Mappings in the Theory Underlying Dynamic Programming," **SIAM Review**, Vol. 9, pp. 165–177.

Denardo, E. and B. Miller [1968], "An Optimality Condition for Discrete Dynamic Programming with no Discounting," **The Annals of Mathematical Statistics**, Vol. 39, pp. 1220–1227.

Denardo, E. and U. Rothblum [1979], "Overtaking Optimality for Markov Decision Chains," **Mathematics of Operations Research**, Vol. 4, pp. 144–152.

Derman, C. [1966], "Denumerable State Markovian Decision Processes–Average Cost Criterion," **Annals of Mathematical Statistics**, Vol. 37, pp. 1545–1554.

Freidenfelds, J. [1981], **Capacity Expansion: Simple Models and Applications**, North–Holland.

Hinderer, K. [1970], **Foundations of Non–Stationary Dynamic Programming with Discrete Time Parameter**, Lecture Notes in Operations Research and Mathematical Systems, Springer–Verlag.

Hopp, W. [1984], "Non–Homogeneous Markov Decision Processes with Applications to R & D Planning," Unpublished Ph.D. dissertation, Department of Industrial and Operations Engineering, The University of Michigan, Ann Arbor, Michigan 48109.

Hopp, W. [1985], "Identifying Forecast Horizons in Non–Homogeneous Markov Decision Processes," Technical Report 85–5, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois, 60201.

Howard, R. [1960], **Dynamic Programming and Markov Processes**, Wiley.

Lohmann, J. [1984], "A Stochastic Replacement Economy Decision Model," Technical Report 84–11, Department of Industrial and Operations Engineering, The University of Michigan, Ann Arbor, Michigan 48109. To appear in **IIE Transactions**.

Luss, H. [1982] "Operations Research and Capacity Expansion Problems: A Survey," **Operations Research**, Vol. 30, pp. 907–947.

Manne, A. [1960], "Linear Programming and Sequential Decisions," **Management Science**, Vol. 6, pp. 259–267.

Miller, B. and A. Veinott [1969], "Discrete Dynamic Programming with a Small Interest Rate," **The Annals of Mathematical Statistics**, Vol. 40, pp. 366–370.

Morton, T. [1978], "The Non–Stationary Infinite Horizon Inventory Problem," **Management Science**, Vol. 24, pp. 1474–1482.

Morton, T. [1979], "Infinite Horizon Dynamic Programming—A Planning Horizon Formulation," **Operations Research**, Vol. 27, pp. 730–742.

Morton, T. and W. Wecker [1977], "Discounting, Ergodicity and Convergence for Markov Decision Processes," **Management Science**, Vol. 23, pp. 890–900.

Nelson, R. and S. Winter [1982], **An Evolutionary Theory of Economic Change**, Belknap Press.

Ross, S. [1968], "Non–Discounted Denumerable Markovian Decision Models," **Annals of Mathematical Statistics**, Vol. 39, pp. 412–423.

Ross, S. [1970], **Applied Probability Models with Optimization Applications**, San Francisco: Holden–Day.

Schäl, M. [1975], "Conditions for Optimality in Dynamic Programming and for the Limit of n–Stage Optimal Policies to be Optimal," **z. Wahrscheinlichkeitstheorie verw. Gebiete**, Vol. 32, pp. 179–196.

Seneta, E. [1981], **Non–negative Matricies and Markov Chains**, New York: Springer–Verlag.

Shapiro, J. [1968], "Turnpike Planning Horizons for a Markovian Decision Model," **Management Science**, Vol. 14, pp. 292–300.

Sobel, M. [1971], "Production Smoothing with Stochastic Demand II: Infinite Horizon Case," **Management Science**, Vol. 17, pp. 724–735.

Veinott, A. [1966], "On Finding Optimal Policies in Discrete Dynamic Programming with no Discounting," **Annals of Mathematical Statistics**, Vol. 37, pp. 1284–1294.