MANAGEMENT SCIENCE
Vol. 23, No. 8, April, 1977
Printed in U.S.A.

DISCOUNTING, ERGODICITY AND CONVERGENCE FOR MARKOV DECISION PROCESSES*

THOMAS E. MORTON† AND WILLIAM E. WECKER‡

The rate at which Markov decision processes converge as the horizon length increases can be important for computations and judging the appropriateness of models. The convergence rate is commonly associated with the discount factor α. For example, the total value function for a broad set of problems is known to converge $0(\alpha^n)$, i.e., geometrically with the discount factor. But the rate at which the finite horizon optimal policies converge depends on the convergence of the *relative* value function. (Relative value at a given state is the difference between total value at that state and total value at some fixed reference state.) Relative value convergence in turn depends both on the discount factor and on ergodic properties of the underlying nonhomogeneous Markov chains. We show in particular that for the stationary finite state space compact action space Markov decision problem, the relative value function converges $0((\alpha\lambda)^n)$ for all $\lambda > r(P)$, the argument of the subdominant eigenvalue of the optimal infinite horizon policy (assumed unique). Easily obtained bounds for $r(P)$ are also given which are related to those of A. Brauer. Under additional restrictions, policy convergence is shown to be of the same order as relative value convergence, generalizing work of Shapiro, Schweitzer, and Odoni. The same result gives convergence properties for the undiscounted problem and for the case $\alpha > 1$. If $\alpha r(P) > 1$ the problem does not converge. As a by-product of the analysis, necessary conditions are given for the relative value function to converge $0((\alpha\lambda)^n)$, $0 < \alpha\lambda < 1$, for the nonstationary problem.

1. Introduction

We study the convergence of finite horizon (possibly nonstationary) Markov decision processes as the length of the horizon increases without bound. An understanding of convergence of stationary processes is important, for example, in evaluating the effectiveness of "value iteration" as a computational technique and in providing rules for stopping computations. Also, the convergence rate is related to the usefulness of applying a stationary model to a nonstationary problem and similar robustness issues.

Convergence of the total value function is well known to be geometric in α, the discount factor—that is, convergence is $0(\alpha^n)$[1] for a broad set of problems [1], [2], [4], [5], [8], [16], [23]. However, it is convergence of the relative value function (total value function less total value at a fixed state) which determines convergence of the policy[2] and gives meaning to the "infinite horizon" problem. We define the term "strong" convergence to describe the case where both the relative value function and policy converge $0((\alpha\lambda)^n)$,[1] $0 < \lambda < 1.0$. For the undiscounted, stationary, finite action, finite state space problem, White [27] has demonstrated strong convergence for a special case, while Schweitzer [24] has demonstrated $0(1/n)$ convergence for this case under rather general conditions. Iglehart [16] demonstrated $0(1/n)$ convergence for a certain continuous action and state space inventory problem. Morton [21] has recently shown

* Accepted by Edward J. Ignall; received February 1975. This paper has been with the authors 2 months, for 2 revisions.

† Carnegie-Mellon University.

‡ University of Chicago.

[1] Suppose $\lim_n A_n = B$: we say $\{A_n - B\}$ converges $0(f(n))$ and $A_n = B + 0(f(n))$ if

$$\limsup_{n\to\infty} |(A_n - B)/f(n)|$$

is finite.

[2] If there is a metric action space and finite state space, the distance between two policies is naturally defined by the max norm. Policy convergence is defined by this natural metric. (For the finite action space special case this reduces to simply attaining an optimal policy.)

strong convergence for a generalized version of this problem, together with an analytic expression for the convergence rate, while Hordijk and Tijms [12] have proved geometric convergence for a related problem.[3]

We develop a theory of strong convergence for a much larger class of problems. For clarity, we focus on the finite state space, compact action space problem, although the methodology is conjectured to be applicable to the upper semi-continuous return compact state space case. While the results for the stationary case are of primary importance, the necessity of developing machinery for evaluating products of nonidentical matrices produces nonstationary generalizations as a by-product. A central result is that relative value and policy convergence for the stationary problem is $0((\alpha\lambda)^n)$ for all $\lambda > r(P)$, where $r(P)$ is the argument of the subdominant eigenvalue[4] of the transition matrix P of the optimal infinite horizon policy (assumed to be unique).

In §2 the nonstationary finite horizon Markov decision model is presented including alternate definitions of the meaning of the infinite horizon problem when total rewards do not converge. In §3 the outer and inner ergodic separation of pairs of transition matrices is first defined. Theorem 1 relates outer separation to strong convergence, generalizing White's work. In §4, Theorem 2 relates inner separation and another concept called strong ergodicity to asymptotic strong convergence. Theorem 3 shows that strong convergence implies the limiting functions satisfy the infinite horizon equations. In §5, Theorems 4, 5 and 6 tie the policy convergence rate to the relative value convergence rate. Finally Theorem 7 relates convergence to strong ergodicity of the optimal limiting policies. In §6, conditions for strong ergodicity are investigated. First some special cases are investigated to gain insight. Next the Hajnal measure [10], [11] is discussed. Its multiplicative property is developed in Theorem 8. Finally the Hajnal and outer ergodic measures are related to the subdominant eigenvalue and combined with Theorem 7 to derive in Theorem 9 that convergence for a stationary problem implies that relative value and policy convergence is $0((\alpha\lambda)^n)$, all $\lambda > r(P)$. Bounds are given for $r(P)$ (for general P) and are related to those of Brauer [6]. Some examples are given. (Proofs deleted in this abridged version are available in the longer version on file in the TIMS office.)

2. The Model

In this section the nonstationary finite horizon Markov decision problem is defined, including the policy function and relative value function. The infinite horizon problem is then defined in terms of the limiting functions of the policy and relative value functions when the length of the finite horizon problem increases without bound.

At the beginning of each of a number of periods, a decision-maker observes the current state of the system and chooses an action. The time, state, and action determine an immediate expected reward and also a transition distribution to the next state. The rewards and transition distributions are assumed known for all periods and may differ from period to period.

Assumptions and definitions of the model include:

1. A countable sequence of (not necessarily equally spaced) periods labelled 1, 2, 3, . . . , t, ...

2. A finite set of states 1, 2, 3, . . . , i, . . . , N.

3. For each state i and period t, $z \in Z_t(i)$ represents a possible action. $Z_t(i)$ is

[3] An anonymous referee points out that P. J. Schweitzer and A. Federgruen have a paper in progress showing geometric convergence for the undiscounted, stationary, finite action, finite state space problem.

[4] The eigenvalue with the largest argument is always 1.0. Of the remaining eigenvalues, that with the largest argument is termed the "subdominant" eigenvalue (which may also be 1.0 in periodic or multichain cases).

assumed to be a compact subset of R^m. Let Z_t be the cross product of the $Z_t(i)$. Then $z \in Z_t$ represents a policy (a vector of actions), a decision $z_t(i)$ for state i at period t, for each i.

4. For each state i and period t, $q_{it}(z)$ and $P_{it}(z)$ are the one period expected reward and transition distribution, respectively. Both are assumed to be continuous on Z_t under the natural relative topology. (Note that if Z_t is finite, continuity is automatic.) $q_t(z_t)$ and $P_t(z_t)$ represent the vector of rewards and matrix of transitions associated with a policy z_t. We assume without loss of generality that $q_t(z_t) \geqslant 0$.

5. $|q_t - q_t(N)| \leqslant M$ all t, z_t. That is to say differences in rewards for different states are assumed not to grow more than geometrically fast over time. (The footnote explains why this is also a condition on the discount factor $0 < \alpha < \infty$.)[5]

6. The (t, n) finite horizon problem is defined as a problem starting in period t and ending at the end of period n, with the criterion of maximizing the total discounted expected reward. v_t^n is the optimal discounted expected reward vector for the (t, n) problem:

$$v_{n+1}^n = 0, \quad v_t^n = \max_{z \in Z_t} \{q_t(z) + \alpha P_t(z) v_{t+1}^n\},$$

which implicitly define at least one optimal policy z_t^n and associated action and transition matrix q_t^n, P_t^n which satisfy $v_t^n = q_t^n + \alpha P_t^n v_{t+1}^n$.

7. The associated relative value and reward functions are defined by:

$$\dot{v}_t^n \equiv v_t^n - v_t^n(N), \quad \dot{q}_t^n \equiv q_t^n - q_t^n(N).$$

8. The limiting relative value function and policy function are defined when they exist by

$$\dot{v}_t^* \equiv \lim_{n \to \infty} \dot{v}_t^n, \quad z_t^* \equiv \lim_{n \to \infty} z_t^n.$$

9. The relative value [policy] function converges *geometrically at rate* β if $(\dot{v}_t^* - \dot{v}_t^n)$ $[(z_t^* - z_t^n)]$ converges $0(\beta^n)$ for some β such that $0 < \beta < 1$. We say the process converges strongly if both the relative value and policy function converge geometrically at rate $\alpha\lambda$ and $0 < \lambda < 1$.

10. The infinite horizon (optimality) equations are defined by

$$\dot{v}_t + g_t = \max_{z_t \in Z_t} \{q_t(z_t) + \alpha P_t(z_t) \dot{v}_{t+1}\}$$

for $t = 1, 2, \ldots,$ where $\dot{v}_t(N) = 0$, and g_t is called the "gain for period t."

11. A sequence of functions $\dot{v}_1, \dot{v}_2, \dot{v}_3, \ldots, \dot{v}_t, \ldots$ is called a *regular solution* if (a) the sequence satisfies the system of infinite horizon equations, and (b) $0 \leqslant |\dot{v}_t| \leqslant M$ for all t (see footnote 2).

The authors argue that the existence and uniqueness of the limiting functions $\{\dot{v}_t^*, z_t^*\}$ give a perfectly adequate definition of the infinite horizon problem, since the only real use for the "infinite" horizon problem is as a common approximation to a number of different possible finite horizon problems. A more conventional definition is to show the existence and uniqueness of a set of solutions to the infinite horizon equations and then argue that $\{\dot{v}_t^*, z_t^*\}$ satisfies the system. Theorem 3 in §4 shows that geometric convergence to the limiting cost function does in fact imply, however, that they satisfy the infinite horizon equations and can be the only regular solutions to do so. In §5, Theorems 4, 5, and 6 develop similar results for the limiting policy function and tie the policy convergence rate to the relative value convergence rate.

[5] The discount factor α is unique up to the transformation $\tilde{q}_{t+k} = (\alpha\tilde{\alpha}^{-1})^k q_{t+k}$, which introduces the transformation $\alpha \to \tilde{\alpha}$, since if a reward V is given by $V = \sum_{k=0}^{\infty} \alpha^k q_{t+k}$, then V is also given by $V = \sum_{k=0}^{\infty} \tilde{\alpha}^k \{(\alpha\tilde{\alpha}^{-1})^k q_{t+k}\}$; that is, choosing α implies the definition of q_t to be used.

3. Ergodicity Measures and Strong Convergence

In this section we first define the outer and inner ergodic separation of pairs of transition matrices. Roughly, two matrices with low inner separation have similar stationary distributions while those with high separation have orthogonal stationary distributions. Low outer separation requires in addition that the two matrices be nearly identical. These lead to the concepts of outer and inner k-diameter of a set of matrices and outer and inner measure of a set of matrices. Theorem 1 shows that strong convergence of relative values in a Markov decision problem follows if the outer k-diameter of a certain set of transition matrices is less than 1.0.

DEFINITION 1. The outer ergodic separation of two transition matrices P_1 and P_2 is defined by

$$d^U(P_1, P_2) = 1 - \sum_{j=1}^{N} \min_{i_1, i_2} \{P_1(i_1, j), P_2(i_2, j)\},$$

where $P(i, j)$ represents the transition probability from state i to state j.

DEFINITION 2. The inner ergodic separation of two transition matrices is defined as the outer ergodic separation of their associated stationary matrices; that is, $d^L(P_1, P_2) = d^U(S_1, S_2)$, where as usual $S_i \equiv \lim_{n\to\infty} 1/n \sum_{k=0}^{n} (P_i)^k$.

REMARK. The inner separation is always less than or equal to the outer separation, and the inner and outer separations of two stationary matrices are equal. The inner separation of a matrix with itself is 0 if the matrix is unichain, 1.0 otherwise.

Now let V be an arbitrary set of transition matrices, and let P_1 and P_2 represent transition matrices which are the product of k matrices selected arbitrarily from V (repeated selections are allowed).

DEFINITION 3. The outer k-diameter of V is defined by $d_k^U(V) = \sup_{P_1, P_2 \in V^k} (d^U(P_1, P_2))^{1/k}$ for $k = 1, 2, \ldots,$ where V^k is the set of products of k matrices.

DEFINITION 4. The asymptotic outer diameter of V is defined by $d_\infty^U(V) = \lim \inf_{k\to\infty} d_k^U(V)$. The inner k-diameter and asymptotic inner diameter of V, d_k^L and d_∞^L are defined similarly. We shall occasionally need to set $P_1 = P_2$.

DEFINITION 5.

$$x^U(P) \equiv d^U(P, P),\ x_k^U(P) \equiv d_k^U(P, P) \quad \text{for } k = 1, 2, \ldots,\ x_\infty^U \equiv d_\infty^U(P, P), \quad \text{etc.}$$

In this case we term $x_k^U(V)$ the outer *measure* of the set of transition matrices rather than the *diameter*, and so forth.

EXAMPLE. Let V be the set of possible transition matrices in Howard's taxicab problem [14, p. 45]. There are 18 such matrices but only 8 distinct rows. Clearly $d^U(V) = 1.0 - \sum_{j=1}^{3} \min_{i,z}\{P(i, j, z)\} = \frac{7}{8}$. Also for the optimal policy

$$P = \begin{bmatrix} \frac{1}{16} & \frac{3}{4} & \frac{3}{16} \\ \frac{1}{16} & \frac{7}{8} & \frac{1}{16} \\ \frac{1}{8} & \frac{3}{4} & \frac{1}{8} \end{bmatrix}; \quad X^U(P) = d^U(P, P) = \tfrac{1}{8}.$$

By direct computation $x_2^U(P) = x_4^U(P) = \frac{1}{8}$ leading to the conjecture that $x_\infty^U(P) = \frac{1}{8}$, which is confirmed after Theorem 8.

DEFINITION 6. Suppose $\beta < 1$: we write $\{A_n\}$ converges $0^+(\beta^n)$, and say, "$\{A_n\}$ converges at high order β^n," if $\{A_n\}$ converges $0(\gamma^n)$ for every $\gamma > \beta$.

DEFINITION 7.

$$w_t^n \equiv v_t^{n+1} - v_t^n, \quad \dot{w}_t^n \equiv w_t^n - w_t^n(N),$$
$$\overline{w}_t^n \equiv \max_i w_t^n(i), \quad \underline{w}_t^n \equiv \min_i w_t^n(i).$$

We now derive the relationship between outer diameter and strong convergence.

THEOREM 1. *If there exists m_0 and t such that the set $V = \{P_T^n \mid n \geqslant T + m_0,\ T \geqslant t\}$ has outer diameter less than one, then the relative value function for period t converges $0^+((\alpha\beta)^n)$ where $\beta = d_\infty^U(V)$, assuming that $\alpha\beta < 1$.*

PROOF. Now in standard fashion [27] using model Definition 6, $\alpha P_t^{n+1} w_{t+1}^n \geqslant w_t^n \geqslant \alpha P_t^n w_{t+1}^n$. Thus defining $P_{t,k}^n \equiv P_t^n P_{t+1}^n \cdots P_{t+k-1}^n$ we have inductively that

$$\alpha^k P_{t,k}^{n+1} w_{t+k}^n \geqslant w_t^n \geqslant \alpha^k P_{t,k}^n w_{t+k}^n. \tag{1}$$

Also define a matrix Q by

$$Q(i,j) \equiv \min_{i_1, i_2} \left(P_{t,k}^{n+1}(i_1, j),\ P_{t,k}^n(i_2, j)\right). \tag{2}$$

Then Q is a matrix of nonnegative elements with identical rows which sum to $1 - q$, where $q \equiv d^U(P_{t,k}^{n+1}, P_{t,k}^n) \leqslant (d_k^U(V))^k$ [by Definition 3]. If $q > 0$, it is true that $\tilde{\tilde{P}} \equiv q^{-1}(P_{t,k}^{n+1} - Q)$ and $\tilde{P} \equiv q^{-1}(P_{t,k}^n - Q)$ are probability transition matrices. We have from (1) that $\alpha^k(P_{t,k}^{n+1} - Q)w_{t+k}^n \geqslant w_t^n - \alpha^k Q w_{t+k}^n \geqslant \alpha^k(P_{t,k}^n - Q)w_{t-k}^n$. Now $\alpha^k Q w_{t+k}^n$ is a column vector with identical entries that can be written $\alpha^k Q w_{t+k}^n \equiv m$, so we have $\alpha^k q(\tilde{\tilde{P}} w_{t+k}^n) \geqslant w_t^n - m \geqslant \alpha^k q(\tilde{P} w_{t+k}^n)$ and, therefore,

$$\alpha^k q\left(\overline{w}_{t+k}^n - \underline{w}_{t+k}^n\right) \geqslant \left(\overline{w}_t^n - \underline{w}_t^n\right) \geqslant 0 \quad \text{for all } t, n, (n \geqslant t + m_0),\ t + k \leqslant n. \tag{3}$$

Note (3) also holds for $q = 0$, by observing in this case that the outer diameter of a set of matrices can only be 0 if all are identical with identical rows. Hence $P_{t,k}^{n+1} = P_{t,k}^n \equiv S$. Then use (1) to show $w_t^n = c$ for some constant c.

But by Definition 3 we have $q \leqslant (\beta(k))^k$ where $\beta(k) \equiv d_k^U(V)$. Utilizing also the fact that $w_{n+1}^n = v_{n+1}^{n+1} - v_{n+1}^n = q_{n+1}^{n+1}$, then $\overline{w}_{n+1}^n - \underline{w}_{n+1}^n \leqslant 2M$ by model assumption 5. Using the fact that $\beta(j) \leqslant 1$ for all j, we easily get from (3) that for each $k \geqslant 1$ (using $|\ |$ as the max norm)

$$|\dot{v}_t^{n+1} - \dot{v}_t^n| \leqslant \overline{w}_t^n - \underline{w}_t^n \leqslant 2M\beta(k)^{-k-m_0}\left[\alpha\beta(k)\right]^{n+1-t}, \quad \text{for all } n \geqslant t + m_0 + k - 1. \tag{4}$$

Since $\alpha\beta < 1$, k can be chosen to make $\alpha\beta(k) < 1$. For such a k, then $|\dot{v}_t^{n+1} - \dot{v}_t^n| = 0((\alpha\beta(k))^n)$. Thus there is a geometric convergence to a limiting function $\dot{v}_t^*$ such that $\dot{v}_t^n = \dot{v}_t^* + 0((\alpha\beta(k))^n)$. But since k can be chosen to make $\beta(k)$ arbitrarily close to $\beta \equiv d_\infty^U(V)$, by *definition* $\dot{v}_t^n = \dot{v}_t^* + 0^+((\alpha\beta)^n)$. Q.E.D.

In White's [27] convergence theorem, the hypothesis requires a state accessed with some minimum probability from every other state in k transitions regardless of the sequence of decisions taken. Theorem 1 generalizes that theorem in several ways. First, it is now permissible to sum such probabilities over all such states. This is extremely important, since the results do not get weaker if states are subdivided, and hence results may be proved by taking limits for problems with continuous state spaces, etc. Secondly, rather than require that all possible k-products have the same state in common to which return occurs with positive probability, here the requirement is only that any two such products have such a state in common. Finally, the theorem generalizes White's result to nonstationary costs and transition matrices.

In some cases it is convenient to have a result like Theorem 1, but for inner diameters. We prove such a result in the following section.

4. Inner Diameter, Infinite Horizon Equations

DEFINITION 8. A set of matrices V is said to be strongly ergodic of order λ if for every possible product of k matrices $P_i \in V$, $i = 1, \ldots, k$, $\prod_{i=1}^k P_i = S_k + (\lambda)^k H_k$, $0 < \lambda < 1$, where S_k is the stationary component (with identical rows) of the product

matrix, and the H_k are uniformly bounded, that is $|H_k| \leqslant H$ (uniform both over k and choice of individual matrices).

THEOREM 2. *If the set* $V = \{P_T^n \mid n \geqslant T + m_0,\ T \geqslant t\}$ *has inner diameter* $d_\infty^L(V) < 1.0$ *and is strongly ergodic of order* λ, *then the relative cost function for period* t *converges* $0^+((\alpha\beta)^n)$ *where* $\beta(k) = \max(\lambda, d_k^L(V))$, $\beta = \lim\sup_{k\to\infty} \beta(k)$.

PROOF. We modify the proof for Theorem 1 where needed.

THEOREM 3. *If the conditions of either Theorem* 1 *or Theorem* 2 *are satisfied over all* t *with* m_0 *independent of* t, *and if in addition* $\alpha\beta < 1$, *then*

(a) *the limiting functions* $\dot{v}_t^*$, $t = 1, 2, 3, \ldots$, *of those theorems provide a regular solution to the infinite horizon equations*; (*see Definition* 11), *and*

(b) *this regular solution is unique.*

5. Policy Convergence and Strong Ergodicity

We next turn our attention to the question of policy convergence.

THEOREM 4. *If the finite horizon value functions for period* $t + 1$ *converge, and if the transition and reward functions for period* t *are continuous functions on a compact action space* Z_t, *then there exists at least one accumulation point of the sequence of optimal finite horizon policies, and any such policy is an optimal policy for the infinite horizon problem.*

REMARK. This result is related to Theorem 1(v) and Note 2 in Odoni [22].

REMARK. If under the hypothesis of Theorem 4, it is also true that $q_t(z) + \alpha P_t(z)\dot{v}_{t+1}^n$ and $q_t(z) + \alpha P_t(z)\dot{v}_{t+1}^*$ are strictly unimodal in z, then the sequence z_t^n is uniquely defined and has a unique limit z_t^* which maximizes the corresponding infinite horizon equation.

Concavity [2] is a special type of unimodality which may often be established by induction on the finite horizon problem and established for the infinite horizon problem by limiting arguments. Extension of this theorem to special nonunimodal cases (e.g., k-concavity [23]) should be possible.

THEOREM 5. *Suppose that*

(a) z_t^* *is an accumulation point of* z_t^n, *belonging to the interior of* z_t; *and*

(b) *the component functions* q_{jt} *and* P_{jt} *for each state* j *possess first and second derivatives* q'_{jt}, q''_{jt}, P'_{jt}, P''_{jt} *in closed neighborhoods about* $z_t^*(j)$, *the first derivatives being continuous on the interiors of the same intervals.*

(c) $q''_{jt}(z_t^*) + \alpha P''_{jt}(z_t^*)\dot{v}_{t+1}^* \neq 0$ *for any component. Then for the convergent subsequence to* z_t^* *it is true that* $|z_t^* - z_t^n| = 0(|\dot{v}_{t+1}^* - \dot{v}_{t+1}^n|)$.

Therefore $(\dot{v}_t^* - \dot{v}_{t+1}^n) = 0^+((\alpha\beta)^n)$ *implies* $(z_t^* - z_t^n) = 0^+((\alpha\beta)^n)$ *also.*

PROOF. By Theorem 4, $z_t^*(j)$ maximizes $q_{jt}(z) + \alpha P_{jt}(z)\dot{v}_{t-1}^*$ for each component j. Suppress j for clarity. Now $q'_t(z_t^*) + \alpha P'_t(z_t^*)\dot{v}_{t+1}^* = 0$, since we are at the interior maximum of a sufficiently smooth function. Taking the differential which is possible by assumption (b),

$$\{q''_t(z_t^*) + \alpha P''_t(z_t^*)\}dz + \alpha\{P'_t(z_t^*)\}dv = 0 \quad \text{or}$$

$$dz = -\alpha\{q''_t(z_t^*)\} + \{P''_t(z_t^*)\}^{-1}\{P'_t(z_t^*)\}dv$$

which cannot involve dividing by zero by assumption. Thus

$$z_t^* - z_t^n = 0(A \cdot (\dot{v}_{t+1}^* - \dot{v}_{t+1}^n)),$$

where A is a constant matrix which establishes the theorem. Q.E.D.

REMARK. Thus under reasonable conditions of local smoothness and convexity of values in policy space about the optimal policy, asymptotic policy convergence is identical to asymptotic relative value convergence.

For the finite action space, essentially the same proof mechanism can be used to derive the number of iterations needed to obtain an optimal policy; the regularity structure is no longer necessary. (This proof is related to the work of Shapiro [26], Schweitzer [24], and Odoni [22], and also to work in process by Hastings on the elimination of nonoptimal actions.)

THEOREM 6. *For the finite action space case there exists a constant K_t for each t such that the $(n-t)$ period optimal policy will be infinite horizon optimal for state j, provided*

$$\max_j \{|\dot{v}^n_{t+1}(j) - \dot{v}_{t+1}(j)| - k_t(j)/K_t\} < 0,$$

where $k_t(j)$ denotes the minimum (infinite horizon) opportunity cost of using a nonoptimal decision in state j at time t for one period.

PROOF. Follows method of Theorem 5. By hypothesis (suppress j) $(q^z_t - q^*_t) + \alpha(P^z_t - P^*_t)\dot{v}_{t+1} \leqslant -k_t < 0$ for all nonoptimal z. Hence

$$(q^z_t - q^*_t) + \alpha(P^z_t - P^*_t)\dot{v}^n_{t+1} \leqslant -k_t + \alpha(P^z_t - P^*_t)(\dot{v}^n_{t+1} - \dot{v}_{t+1}).$$

The proof follows, with $K_t = \alpha \max_{z,j} \sum_i |P^z_t(j, i) - P^*_t(j, i)|$.

COROLLARY. *If there exists $\bar{\bar{n}}_t$ such that $\max_j |\dot{v}^n_{t+1}(j) - \dot{v}_{t+1}(j)| \leqslant M_t(\alpha\lambda)^n$, for $n \geqslant \bar{\bar{n}}_t$, then there exist an integer $\bar{n}_t$ and integers $n_t(j)$ such that the policy for state j will be optimal for all*

$$n > n(j) = \log k_t(j)/|\log \alpha + \log \lambda| + \bar{n}_t.$$

We remark that the corollary may provide a termination criterion for obtaining the optimal policy in value iteration, since we conjecture that the various parameters can be estimated from the process itself with asymptotic exactness.

THEOREM 7. *Suppose, for each $T \geqslant t$, that the sequence $\dot{v}^n_T, \dot{v}^{n+1}_T, \ldots$ converges and that the conditions of Theorem 5 are satisfied for unique z^*_T so that the sequence $P^n_T, P^{n+1}_T, \ldots$ also converges. If the set $V = \{P^*_t, P^*_{t+1}, P^*_{t+2}, \ldots\}$ is strongly ergodic of order λ, then the process converges strongly $0^+(\alpha\lambda)^n$.*

PROOF. A slight modification of Theorem 2.

6. Strong Ergodicity and the Hajnal Measure

Theorem 7 shows that the asymptotic rate of convergence for a convergent process is closely related to strong ergodicity of the optimal infinite horizon transition matrices. In this section we investigate conditions under which strong ergodicity may be established. First we investigate some simple special cases to gain insight. Next the Hajnal measure [10], [11] of a transition matrix is discussed; its multiplicative property leads directly to simple sufficient conditions for strong ergodicity. Next the Hajnal k-measure of a set is developed. Its infimum over k is termed the Hajnal asymptotic measure of the set. The subradius of a transition matrix $r(P)$ is defined as the absolute value of the (complex) subdominant eigenvalue. It is shown that in the stationary problem, for the particular case of a single feasible policy, both the outer ergodic measure and the Hajnal asymptotic measure are equal to the subradius of the matrix. Thus the outer ergodic or Hajnal k measures (especially $k = 1$!) are each an upper bound on the subradius, a useful result in itself. Combined with Theorem 7, this leads to the result for the stationary problem that convergence implies that the process

converges strongly 0^+ $((\alpha r(P))^n)$, where $r(P)$ is the subradius of the optimal transition matrix (assumed unique).

OBSERVATION 1. The nth power of any primitive (unichain, aperiodic) transition matrix P may be expressed as $P^n = S + n^k r(P)^n H_n$, where

(a) $1 \leqslant k + 1 \leqslant$ the multiplicity of roots λ_i with $|\lambda_i| = r(P)$,

(b) $|H_n| \leqslant M$ all n,

(c) $\lim_{n\to\infty} H_n = H$ exists.

Observation 1 follows directly from considering the nth power of the Jordan canonical form of $(P - S)$. (See also Howard [14] for an alternative development.)

OBSERVATION 2. If at least one of the P_j is itself stationary, that is, $P_j = S_j$, then $\prod_{i=1}^{n} P_i = S_{i,\ldots,n} + (0)^n H_{1,\ldots,n}$.

OBSERVATION 3. If all P_j are identical, $P_j \equiv P$, then $\prod_{j=1}^{n} P_j = S_{1,\ldots,n} + \lambda^n H_{1,\ldots,n}$ for every $\lambda > r(P)$.

Note that if subdominant roots are not multiple, then $k_1 = 1$; we might then term the matrix *simply ergodic*. Note that for a simply ergodic matrix $P^n = S + r^n \cdot (H + 0(1/n))$.

DEFINITION 9. $S + r^n H$ is termed the asymptotic expansion of a simply ergodic matrix P.

DEFINITION 10. Matrices whose exact expansion is simply $P^n = S + r(P)^n H$ will be termed *asymptote matrices*.

OBSERVATION 4. The product of asymptote matrices is asymptote; the subradius of their products is the product of the subradii.

Observation 4 is suggestive and leads intuitively to the hypothesis that perhaps $r(P_1P_2) = r(P_1)r(P_2)$ in general. Hajnal [10], [11] proves this true for the case when all P_j commute with each other. Unfortunately Hajnal also constructs cases where $r(P_1) < 1.0$, $r(P_2) < 1.0$, yet $r(P_1P_2) = 1.0$. There is, however, a more conservative measure of closeness to stationarity which will preserve this multiplicative property. The definitions and results marked (H) are due to Hajnal [10], [11]. Recall that $(a)^+$ means $\max(a, 0)$.

DEFINITION 11. The Hajnal measure of a transition matrix $h(P)$ is defined by

$$h(P) \equiv \max_{i_1, i_2} \sum_j (P_{i_1 j} - P_{i_2 j})^+ = 1 - \min_{i_1, i_2} \sum_j \min(P_{i_1 j}, P_{i_2 j})$$

(since $(a - b)^+ = a - \min(a, b)$).

EXAMPLE. If

$$P = \begin{bmatrix} 0.0 & 0.1 & 0.2 & 0.7 \\ 0.1 & 0.3 & 0.6 & 0.0 \\ 0.2 & 0.0 & 0.1 & 0.7 \\ 0.1 & 0.2 & 0.0 & 0.7 \end{bmatrix}$$

then $h(P) = 0.8$. (Note that the outer 1-diameter of P is 1.0, so that the two concepts are not identical.)

EXAMPLE. In Howard's taxicab example if P_i represents the transition matrix of an optimal policy, then all $h(P_i) \leqslant \frac{13}{16}$.

DEFINITION 12.

$$g(P) = \max_j \max_{i_1, i_2} (P_{i_1 j} - P_{i_2 j}) = -\min_{i_1, i_2} \min_j (P_{i_1 j} - P_{i_2 j}).$$

In the first example above, $g(P) = 0.7$; in the second it is easily seen that $g(P_i) \leqslant \frac{3}{4}$ for every P_i. Recall the definition of $x^U(P)$ from Definition 5.

LEMMA. *For any transition matrices P, Q*

(a) $g(P) \leqslant h(P) \leqslant x^U(P)$,

(H) (b) $g(PQ) \leqslant h(P)g(Q) \leqslant h(P)h(Q)$,

(H) (c) $g(\prod_{i=1}^{n} P_i) \leqslant \{\prod_{i=1}^{n-1} h(P_i)\} g(P_n) \leqslant \prod_{i=1}^{n} h(P_i)$.

THEOREM 8. (a) *For any product of transition matrices* $\prod_{i=1}^{n} P_i = S_n + \{\prod_{i=1}^{n} h(P_i)\}H_n$ (S_n *stationary*, H_n *uniformly bounded*).

(b) *Any set of transition matrices* V *is strongly ergodic of order* $\lambda = \sup_{P \in V}\{h(P)\}$.

DEFINITION 13. The k-Hajnal measure of a set of matrices V is defined by $h_k(V) \equiv \sup_{P_i \in V}\{[h(\prod_{i=1}^{k} P_i)]^{1/k}\}$.

DEFINITION 14. The asymptotic Hajnal measure of a set of matrices V is defined by $h_\infty(V) \equiv \lim\inf_{k\to\infty} h_k(V)$.

THEOREM 9. *The asymptotic outer measure of a transition matrix and its asymptotic Hajnal measure are both equal to its subradius, that is*: $r(P) = x_\infty^U(P) = h_\infty(P)$.

COROLLARY 9.1.[6] *If a stationary Markov decision process converges, and the infinite horizon policy is unique, asymptotic value* (*with restrictions, also policy*) *convergence is* $0^+((\alpha r(P))^n)$.

PROOF. Direct combination of Theorems 5, 6, 7, 9. Q.E.D.

Note that for $\alpha = 1.0$ Schweitzer and others [24], [13] have recently shown convergence under rather general conditions, and thus Corollary 9.1 adds the information that convergence is geometric with rate $\alpha r(P)$. For $\alpha < 1$ convergence is improved from $0(\alpha^n)$ to $0^+((\alpha r(P))^n)$. Corollary 9.1 together with Theorems 5 and 6 gives a similar improvement for policy convergence.

COROLLARY 9.2. (a) $r(P) \leqslant h(P) \leqslant x^U(P)$.
(b) $r(P) \leqslant h_k(P)$.
(c) $r(P) \leqslant x_k^U(p)$.

Corollary 9.2 is of independent interest since (a) gives two relatively easily computed bounds on the subdominant eigenvalue, a quantity that is difficult to compute directly.

The fact that $r(P) \leqslant x^U(P)$ was pointed out by A. Brauer [6]. The relationship between the other bounds and those derived by Brauer will be discussed elsewhere.

EXAMPLE. In Howard's toymaker example [14, p. 4], there is only one transition matrix

$$P = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{2}{5} & \frac{3}{5} \end{bmatrix}.$$

We find immediately that $h(P) = x^U(P) = 0.1$, guaranteeing that convergence of the problem in horizon length is of order $(0.1)^n$. In [14, p. 10] via z-transform analysis, Howard shows that in fact $r(P) = 0.1$, so that in this case our bounds are exact. (This will always be the case for a two by two matrix.)

In Howard's taxicab example [14, p. 45], if V is the set of all possible transition matrices, we find immediately that $x^U(V) \leqslant \frac{14}{16}$ (simply take 1.0 minus the sum of the minimums over *all* rows to get a quick bound), and thus convergence of this problem is at least of order $(\frac{14}{16})^n$ by Theorem 1. In [14, pp. 47–48], Howard shows that the optimal infinite horizon transition matrix is

$$P = \begin{bmatrix} \frac{1}{16} & \frac{3}{4} & \frac{3}{16} \\ \frac{1}{16} & \frac{7}{8} & \frac{1}{16} \\ \frac{1}{8} & \frac{3}{4} & \frac{1}{8} \end{bmatrix}.$$

We see that $x^U(P) = \frac{2}{16}$, $h(P) = \frac{2}{16}$. Thus asymptotic convergence is at least of order

[6] Examination of the fixed policy case makes it at least intuitively clear that more powerful convergence results cannot be expected in general.

$(\frac{2}{16})^n$. Solving the characteristic equation

$$0 = \begin{vmatrix} 1 - 16\lambda & 12 & 3 \\ 1 & 14 - 16\lambda & 1 \\ 2 & 12 & 2 - 16\lambda \end{vmatrix}$$

we find that the eigenvalues are $\frac{16}{16}$, $\frac{2}{16}$, $-\frac{1}{16}$; thus $r(P) = \frac{2}{16}$ and asymptotic convergence is in fact $(\frac{2}{16})^n$.

By extending the results of this paper to the continuous state space case, geometric cost and policy convergence results similar to those obtained recently by Morton [21] and by Hordijk and Tijms [12] for undiscounted inventory problems would immediately be available for a variety of similar continuous problems.[7]

[7] The author wishes to acknowledge the associate editor and both referees who were especially helpful.

References

1. Arrow, K. J., Karlin, S. and Scarf, H., *Studies in the Mathematical Theory of Inventory and Production*, Stanford Press, Stanford, California, 1958.
2. Bellman, R. E., *Dynamic Programming*, Princeton Press, Princeton, N. J., 1957.
3. Blackwell, D., "Discrete Dynamic Programming," *Ann. Math. Stat.*, Vol. 33, pp. 719–726 (1962).
4. ———, "Discounted Dynamic Programming," *Ann. Math. Stat.*, Vol. 36, pp. 226–235 (1965).
5. Boylan, E. S., "Existence and Uniqueness Theorems for the Optimal Inventory Equation," *SIAM J. Appl. Math.*, Vol. 14, pp. 961–969 (1966).
6. Brauer, A., "Limits for the Characteristic Roots of a Matrix, IV: Applications to Stochastic Matrices," *Duke Math. J.*, Vol. 19, pp. 75–91 (1952).
7. Brown, B. W., "On the Iterative Method of Dynamic Programming on a Finite Space Discrete Time Markov Process," *Ann. Math. Stat.* , Vol. 36, pp. 1279–1285 (1965).
8. Denardo, E. V., "Contraction Mappings in the Theory Underlying Dynamic Programming," *SIAM Rev.*, Vol. 9, pp. 165–177 (1967).
9. Dirickx, Yvo, "Turnpike Theory in Deterministic Discrete Dynamic Programming with Discount Factor Greater Than One," *SIAM J. Appl. Math.*, Vol. 24, pp. 467–473 (1973).
10. Hajnal, J., "The Ergodic Properties of Nonhomogeneous Finite Markov Chains," *Proc. Cambridge Philosophical Soc.*, Vol. 52, pp. 67–77 (1956).
11. ———, "Weak Ergodicity in Nonhomogeneous Markov Chains," *Proc. Cambridge Philosophical Soc.*, Vol. 54, pp. 233–246 (1958).
12. Hordijk, A. and Tijms, H., "On a Conjecture of Iglehart," *Management Science*, Vol. 21, pp. 1342–1345 (1975).
13. ———, Schweitzer, P. and Tijms, H., "The Asymptotic Behavior of the Minimal Total Expected Cost for the Denumerable State Markov Decision Model," *J. Appl. Prob.*, Vol. 12, pp. 298–305 (1975).
14. Howard, R. A., *Dynamic Programming and Markov Processes*, The M.I.T. Press, Cambridge, Mass., 1960.
15. ———, *Dynamic Probabilistic Systems*, Vols. I and II, Wiley, New York, 1971.
16. Iglehart, D., "Dynamic Programming and Stationary Analysis of Inventory Problems," Chapter 1 in *Multistage Inventory Models and Techniques*, H. Scarf, D. Gilford and M. Shelly (eds.), Stanford Press, Stanford, California, 1963.
17. Karlin, S., "The Structure of Dynamic Programming Models," *Nav. Res. Log. Q.*, Vol. 2, pp. 285–294 (1955).
18. Kemeny, J. G. and Snell, J. L., *Finite Markov Chains*, Van Nostrand, New York, 1960.
19. Macqueen, J. B., "A Modified Dynamic Programming Method for Markovian Decision Problems," *J. Math. Anal. and Appl.*, Vol. 14, pp. 38–43 (1966).
20. Morton, T. E., "On the Asymptotic Convergence Rate of Cost Differences for Markovian Decision Processes," *Opns. Res.*, Vol. 19, pp. 244–248 (1971).
21 ———, "The Nonstationary Infinite Horizon Periodic Review Inventory Problem, I. and II.," Center for Mathematical Studies in Business and Economics, Reports 7332, 7333, University of Chicago (1973).
22. Odoni, A. R., "On Finding the Maximal Gain in Markov Decision Processes," *Opns. Res.*, Vol. 17, pp. 857–860 (1969).

23. SCARF, H., "The Optimality of (*s*, *S*) Policies in the Dynamic Inventory Problem," Chapter 13 in *Mathematical Methods in the Social Sciences*, K. J. Arrow, S. Karlin and P. Suppes (eds.), Stanford Press, Stanford, California, 1960.
24. SCHWEITZER, P. J., "Perturbation Theory and Markov Decision Processes," M.I.T. Operations Research Center Technical Report, No. 15 (June 1965).
25. ———, "Annotated Bibliography on Markov Decision Processes," IBM Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598.
26. SHAPIRO, J., "Turnpike Planning Horizons for a Markovian Decision Model," *Management Science*, Vol. 14, pp. 292–300 (1968).
27. WHITE, D. J., "Dynamic Programming, Markov Chains, and the Method of Successive Approximations," *J. Math. Anal. Appl.*, Vol. 6, pp. 373–376 (1963).
28. VEINOTT, A. F., JR., "Discrete Dynamic Programming with Sensitive Discount Optimality Criteria," *Ann. Math. Stat.*, Vol. 40, pp. 1635–1660 (1969).