Denumerable State Nonhomogeneous Markov Decision Processes

James C. Bean
and
Robert L. Smith
Department of Industrial and Operations Engineering
The University of Michigan
Ann Arbor, MI 48109-2117

Jean B. Lasserre
LAAS
7 avenue du Colonel Roche
31077
Toulouse, France

Technical Report 89-15

April 1989

DENUMERABLE STATE NONHOMOGENEOUS MARKOV DECISION PROCESSES

James C. Bean[†], Jean B. Lasserre[‡], Robert L. Smith[†]

April 19, 1989

ABSTRACT

We consider denumerable state nonhomogeneous Markov decision processes and extend results from both denumerable state homogeneous and finite state nonhomogeneous problems. We show that, under weak ergodicity, accumulation points of finite horizon optimal (termed algorithmic optima) are average cost optimal. We also establish the existence of solution horizons. Finally, an algorithm is presented to solve problems of this class for the case where there is a unique algorithmic optimum.

Research on Markov decision problems has concentrated on homogeneous problems. Of most interest here are problems with denumerable state spaces such as in Federgruen, Schweitzer and Tijms [1978], Cavazos-Cadena [1986], and Hernandez-Lerma and Lasserre [1988]. The usual optimality criterion employed is average optimality. Traditional solution procedures include policy iteration or fixed point methods.

Much less work has been done on the nonhomogeneous version of the problem. Almost all of this work has considered finite state spaces. Recent works include Hopp, Bean, and Smith [1987] and Hopp [1985]. The primary optimality criteria used is algorithmic optimality (i.e., solutions which are accumulation points of finite horizon optima) since we no longer have the optimality equations for average optimality. Traditional solution procedures use solution horizons with cost based stopping rules. A solution horizon is a

[†] Department of Industrial and Operations Engineering, The University of Michigan, Ann Arbor, Michigan 48109. The work of these authors was supported in part by the National Science Foundation under Grant ECS-8700836 to The University of Michigan and by a Rackham Grant for International Research from The University of Michigan.

[‡] LAAS, 7 avenue du Colonel Roche, 31077, Toulouse, France

time such that solving a finite horizon problem to that horizon or beyond gives a first period policy in agreement with the infinite horizon optimal first policy.

Despite these differences, the problems have a great deal in common. Any finite state nonhomogeneous Markov decision process can be reformulated as a denumerable state homogeneous problem. States are relabeled to include time period and state designation in the new formulation. However, the converse is not always true. Some denumerable homogeneous problems cannot be transformed into finite nonhomogeneous problems.

It is interesting that the finite nonhomogeneous problem is a special case of the denumerable homogeneous problem since the former is generally considered harder. We attempt, in this paper, to investigate this relationship and expand the techniques and theory available to each problem. As a mechanism we consider the general denumerable state nonhomogeneous problem which subsumes both of the traditional problems.

We show that, under weak ergodicity, an algorithmically optimal strategy is also average optimal. We also provide a necessary and sufficient condition for a first period decision to be average optimal. Based on this result we propose a modified value iteration algorithm. At each step of the value iteration procedure a stopping rule permits elimination of some nonoptimal decisions until only one remains. For a given initial state, the optimal first period decision is obtained in a finite number of steps of the value iteration procedure. If we assume that, at each state, there are a finite number of one-step neighbors with strictly positive probability, then the modified value iteration procedure is implementable despite the countable state space.

Section 1 formally states the problem under consideration. Section 2 generalizes the solution horizon theory of Hopp, Bean, and Smith. Section 3 presents the stopping rule and the modified value iteration procedure. Finally, Section 4 includes a summary and conclusions.

1. Problem Statement

1.1 Notation

We generalize the notation of Hopp, Bean, and Smith. The decision maker chooses a policy in stage k, x_k, by selecting actions, x_k^i, from finite sets, for states $i = 1, 2, 3, \ldots$. An infinite horizon strategy, x, is an infinite sequence of policies. The set of all feasible strategies is denoted by X.

Taking action x_k^i in state i of stage k gains a reward, $r_k^i(x_k)$, and transition probabilities to all states in stage $k+1$, $p_k^{ij}(x_k)$, $j = 1, 2, 3, \ldots$. Let $\bar{R} < \infty$ denote an upper bound on $|r_k^i(x_k)|$. Let $R_k(x_k)$ represent the column vector of rewards in stage k and $P_k(x_k)$ represent the matrix of transition probabilities from all states in stage k to all states in stage $k+1$. The vector of transition probabilities from state i at stage k is denoted $P_k(x_k; i)$. Note that both the rewards and transition probabilities may be stage dependent, i.e., nonhomogeneous.

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 stage k to stage $N, N > k$, is written $V_k(x; N)$. In evaluating $V_k(x; N)$, the first $k-1$ policies of x are irrelevant. The $V_k(\cdot)$ function maps into $\Re^\infty$ with the i^{th} element given by $V_k^i(x; N)$, which represents the expected net present profit from stage k to stage N given that the process is in state i at stage k. In general we are interested in the value function from stage zero onward, which is written:

$$V_0(x; N) = \sum_{n=0}^{N-1} \alpha^n T_0^n(x) R_n(x_n),$$

where

$$T_l^n(x) = \prod_{k=l}^{n-1} P_k(x_k), \; n > l \quad \text{and}$$

$$T_l^l(x) = I.$$

In the finite horizon problem, with discount factor α, optimization is equivalent to maximization of $V_0(x; N)$ over $x \in X$. In the infinite horizon problem define x^* to be ***α-optimal***, for $0 < \alpha < 1$, if

$$\lim_{N\to\infty} V_0(x^*; N) \geq \lim_{N\to\infty} V_0(x; N), \text{ for all } x \in X.$$

If $\alpha = 1$, so that in general $V_0(x; N)$ diverges with N, we define x^* to be *average optimal* when

$$\liminf_{N\to\infty} \frac{V_0(x^*; N)}{N} \geq \liminf_{N\to\infty} \frac{V_0(x; N)}{N}, \text{ for all } x \in X.$$

Our assumption that $\bar{R} < \infty$ implies that the liminf are always finite. While this measure is the most commonly used, it is less than satisfactory since it allows inclusion of suboptimal leading policies (see Hopp, Bean and Smith). The following section presents a more restrictive measure of optimality.

1.2 Algorithmic Optimality

If optimal solutions to the finite horizon problems, $\{x^*(N)\}_{N=1}^{\infty} \subseteq \Re^{\infty \times \infty}$, approach a particular infinite horizon strategy we would like to consider it as the optimal strategy. The most useful implementation of this idea is algorithmic optimality defined in Hopp, Bean, and Smith (called periodic forecast horizon optimality there). The term algorithmically optimal was first used in Schochetman and Smith [1989] and arises since precisely these solutions can be discovered by a forward looking algorithm. To formally define this notion we use a metric analogous to that presented in Bean and Smith [1984]. Let

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

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

The topology induced by this metric is the product of product topologies. In particular, a sequence $x(n) \to x$ as $n \to \infty$ if and only if, for all $I, K < \infty$, there is an $N(I, K)$ such that $x_k^i(n) = x_k^i$ for $i = 1, 2, \ldots, I$ and $k = 1, 2, \ldots, K$ for all $n \geq N(I < K)$.

Definition: A strategy, x^*, is *algorithmically 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 $m \to \infty$. Note that $x^*(N_m)$ is composed of the finite optimal solution to the N_m problem extended arbitrarily to cover the infinite horizon.

The following two theorems generalize results from Hopp, Bean, and Smith. They are stated here without the proofs, since they are obvious extensions.

Theorem 1: If X is compact in the topology generated by ρ, an algorithmically optimal strategy exists for the denumerable nonhomogeneous Markov decision process.

As in the finite state problem, algorithmic optimality implies α–optimality if the value function is convergent. This implication is a natural extension of the results of Hopp, Bean and Smith and is stated as Theorem 2.

Theorem 2: If Assumptions (2)-(4) (Section 2.2) hold and $\alpha < 1$ then algorithmic optimality implies α–optimality.

Remark: The converse of Theorem 2 is false. For a counterexample see Ryan, Bean and Smith [1987].

If we assume that $\bar{R} < \infty$, the profit functions can diverge only if $\alpha = 1$. In this case, we would like to show that algorithmic optimality implies average optimality. Algorithmic optimality would then be stronger than average optimality since algorithmic optimality discourages leading suboptimal policies. However, this implication is not true without additional conditions.

2. Solution Horizons and Average Optimality

We will show that in most denumerable nonhomogeneous problems the algorithmically optimal solutions are also average optimal, and that solution horizons exist leading to discovery of algorithmically optimal solutions. However, there are cases where these desirable characteristics fail. Consider the following example from Ross [1968].

We have a deterministic problem with state space $\{1, 2, \ldots\}$. At each state, i, there are two choices: remain in place and receive reward i, or move up one state and receive reward 1. The infinite horizon average reward optimal solution is to go up one state to i, remain there for i transitions, then repeat. This is not an algorithmically optimal solution.

We seek conditions to eliminate the possibility of such behavior. As in the finite nonhomogeneous case, the important characteristic is ***weak ergodicity***.

2.1 Weak Ergodicity

A stochastic matrix is ***stable*** if it has identical rows. A scalar function, $\tau(\cdot)$, that is continuous in an appropriate topology on the set of doubly infinite stochastic matrices (treated as points in $\Re^{\infty\times\infty}$) such that $0 \le \tau(P) \le 1$ 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. See Iosifescu [1972] for a full discussion.

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

$$\lim_{n\to\infty} \tau(T_l^n(x)) = 0, \quad \text{for all } l \ge 0,$$

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

The most commonly discussed coefficient of ergodicity is the Hajnal coefficient,

$$\tau_1(P) = \sup_{i,j}\{\sum_{s=1}^{\infty} |p_{is} - p_{js}|/2\}.$$

For example, see Iosifescu; Hopp, Bean and Smith.

Lemma 3: For any 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

(b) For any $l \ge 0$

$$\tau_1(T_l^n(x)) \le \prod_{k=l}^{n-1} \tau_1(P_k(x_k))$$

Proof: Iosifescu. ∎

Lemma 4: For any three stochastic matrices P_1, P_2, T, vector R with $|R_k| \le \bar{R}$, and proper coefficient of ergodicity, $\tau(T)$,

$$(P_1 - P_2)TR \le (2\tau(T)\bar{R})e,$$

where e is a vector of ones.

Proof: Omitted. ∎

The probability distribution on the states in stage n starting from each of the initial states is given by the matrix $P_0(x_0)T_1^n(x)$. If $\tau_1(T_1^n(x)) \to 0$, so that $T_1^n(x)$ has asymptotically identical rows as $n \to \infty$, then $P_0(x_0)T_1^n(x)$ also approaches a matrix of identical rows irrespective of the $P_0(x_0)$ matrix. The probabilities on the states in stage n, and hence the expected rewards in stage n and subsequent stages, become independent of x_0 as n increases.

2.2 Assumptions

The following assumptions are invoked in subsequent results:

(1) $\sum_{n=k}^{\infty} \tau_1(T_k^n(x_l)) \leq A_k < \infty$, uniformly over X, for all $k = 1, 2, \ldots$, where $A_k/k \to 0$ as $k \to \infty$.

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

(3) $|r_k^i(x_k)| \leq \bar{R} < \infty$ for all stages, k, states, i, and strategies, x.

(4) From each state, i, at stage k, under strategy x, there exists a finite set $\{j | P_k^{ij}(x) > 0\}$. That is, only a finite set of states is reachable in one transition from any state, under any strategy. Further, $\max\{j | P_k^{ij}(x) > 0\}$ is uniform over $x \in X$ for each stage k.

Assumption (1) requires slightly more than weak ergodicity since the series must converge rather than just its terms going to zero. This is weaker than the equivalent condition in Hopp, Bean and Smith which requires that A_k be independent of k. All results requiring this assumption can be shown with $\alpha < 1$ substituting for Assumption (1).

Assumption (2) limits the problem to discrete, bounded decision variables. Assumption (3) requires an upper bound on the maximum reward obtainable at each stage.

Given a known starting state, Assumption (4) requires that the accessible states, for any finite time, are finite.

Remark: Since any sequence of policies is feasible and, by Assumption (2), there are a finite number of choices at each state, X is compact in the topology generated by ρ (Tychonoff Theorem).

2.3 Average Optimality

Theorem 2, stating that algorithmic 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, the uniform convergence of $\sum_n \tau_1(T_k^n(x))$ over X will suffice.

Theorem 5: Under the Assumptions (1) through (4), algorithmic optimality implies average optimality.

Proof: Let x^* be algorithmically optimal. By definition there exists a sequence $\{N_m\}_{m=1}^{\infty}$ such that $x^*(N_m) \to x^*$. So, by Assumption (4), for all k there exists M such that $m \geq M$ implies that $x_n^{*i}(N_m) = x_n^{*i}$, $n = 1, 2, \ldots, k$, state i reachable in k steps. Note that $T_0^k(x)$ is identically zero for columns beyond the maximum reachable state in k steps. Fix k and choose $x \in X$ and m sufficiently large. Since $x^*(N_m)$ is optimal for horizon N_m,

$$V_0(x^*(N_m);N_m) \geq V_0(x;k) + T_0^k(x)V_k(x^*(N_m);N_m)$$

which implies that

$$V_0(x^*;k) + T_0^k(x^*)V_k(x^*(N_m);N_m) - V_0(x;k) - T_0^k(x)V_k(x^*(N_m);N_m) \geq 0.$$

since $V_0(x^*;k) = V_0(x^*(N_m);k)$ and $T_0^k(x^*) = T_0^k(x^*(N_m))$. Hence,

$$V_0(x^*;k) - V_0(x;k) + [T_0^k(x^*) - T_0^k(x)]V_k(x^*(N_m);N_m) \geq 0.$$

Now

$$[T_0^k(x^*) - T_0^k(x)]V_k(x^*(N_m);N_m) = \sum_{n=k}^{N_m} \alpha^n [T_0^k(x^*) - T_0^k(x)]T_k^n(x^*(N_m))R_n(x^*(N_m))$$

$$\leq \sum_{n=k}^{N_m} \left[\alpha^n 2\tau_1(T_k^n(x^*(N_m)))\bar{R}\right] e$$

by Lemma 4. By Assumption (1), this is not greater than $2A_k\bar{R}e$. Hence

$$V_0(x^*;k) \geq V_0(x;k) - 2A_k\bar{R}e.$$

Divide both sides by k and take the liminf of both sides to conclude that x^* is superior to x in average reward, and hence, is average optimal.∎

2.4 Existence of Solution Horizons

This section extends results for nonhomogeneous Markov decision processes from those of Hopp, Bean and Smith; and Schochetman and Smith. By use of solution horizon results we can find algorithmically optimal strategies when they are unique.

Theorem 6: Under Assumption (2), $x^*(N) \to x^*$ in the ρ metric as $N \to \infty$ for all choices $x^*(N)$ at each N if and only if the algorithmically optimal strategy is unique.

Proof: (*if*) 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 by Assumption (2), this, in turn, has a convergent subsequence. By definition, its limit is also algorithmically optimal, a contradiction to the assumption of uniqueness.

(*only if*) If the algorithmically optimal strategy is not unique, then there exist at least two distinct strategies which are limit points of finite horizon optima. Then the limit of finite horizon strategies cannot exist.∎

Corollary 7: Under Assumption (2), if all algorithmically optimal strategies have the same first L policies then a solution horizon exists leading to the optimal first L policies in the infinite horizon problem.

Note that convergence of $x^*(N)$ to x^* implies that the optimal finite horizon strategy will agree with the optimal infinite strategy over an increasing range of states and stages. This fact follows from the definition of the metric, ρ.

The existence of solution horizons allows computation of algorithmically optimal strategies in a forward recursive manner. The following section develops such an algorithm.

3. A Cost Based Algorithm

3.1 Preliminary Results

In this section we prove that Assumptions (1) and (3) imply that the difference between future values from any two states converges. In this case, the problem can be transformed into a finite value problem as in Hopp. This characteristic, termed *coherence*, is necessary for the main theorem of Section 3.

Definition: A nonhomogeneous Markov decision process is *coherent* if, for all k, and state pairs, i, j, $V_k^i(x;N) - V_k^j(x;N)$ is uniformly convergent as $N \to \infty$, over i, j and $x \in X$.

Theorem 8: Under Assumptions (1) and (3) the denumerable nonhomogeneous Markov decision problem is coherent.

Proof: From Assumption (1), for any fixed k, for all $\epsilon > 0$, there exists N independent of x, such that

$$\sum_{l=N}^{\infty} \tau_1(T_k^l(x)) < \epsilon.$$

Let $T_k^l(x;i)$ be the ith row of the matrix $T_k^l(x)$. By definition of the Hajnal coefficient, $\tau_1(\cdot)$, for any states, i, j, and strategy, x,

$$|(T_k^l(x;i) - T_k^l(x;j)) \cdot e| \leq 2\tau_1(T_k^l(x)).$$

Hence, by Assumption (3),

$$|(T_k^l(x;i) - T_k^l(x;j)) \cdot R_l(x_l)| \leq 2\tau_1(T_k^l(x))\bar{R}.$$

Then

$$\sum_{l=N}^{\infty} [T_k^l(x;i)R_l(x_l) - T_k^l(x;j)R_l(x_l)] \leq 2\bar{R} \sum_{l=N}^{\infty} \tau_1(T_k^l(x)) < 2\bar{R}\epsilon,$$

uniformly over $x \in X$. Hence, the series defining $V_k^i(x) - V_k^j(x)$ is uniformly convergent for i, j and $x \in X$ and the problem is coherent.■

If a problem is coherent, it can be transformed, without loss of optimality, to a finite value problem. Define the infinite horizon coherent value for state i as

$$\bar{V}_1^{*i} = \lim_{N \to \infty} [V_1^i(x^*;N) - V_1^1(x^*;N)]$$

, which is well defined by Theorem 8. Similarly, the finite horizon coherent value is $\bar{V}_1^{*i}(N) = V_1^i(x^*(N);N) - V_1^1(x^*(N);N)$. For the finite state problem, Hopp shows that if x_0^* uniquely maximizes $[R_0(x_0) + P_0(x_0)\bar{V}_1^*]$, it is the first policy of an algorithmically optimal strategy. That is, we can transform the infinite valued problem into a traditional finite valued problem, even without discounting. Note that an x_0^* exists by Theorem 1. We extend this to the denumerable problem.

The following lemma demonstrates convergence of the finite horizon coherent values to the infinite horizon coherent value. It is used to prove the validity of the stopping rule. We use the notation

$$\epsilon_k(N) = \sup_{x \in X} \sum_{n=N}^{\infty} r_1(T_k^n(x))R_n(x).$$

Under Assumption (1), $\epsilon_k(N) \to 0$ as $N \to \infty$ for all k.

Lemma 9: (Coherent Value Convergence) Under Assumptions (1) and (3), $|\bar{V}_1^{*i}(N) - \bar{V}_1^{*i}| \leq 2\epsilon_2(N)\bar{R}$ for all $i = 1, 2, \ldots$.

Proof: Let $V_1^*(N) = V_1(x^*(N); N)$ with element i denoted by $V_1^{*i}(N)$. For any two integers, N, m,

$$V_1^{*i}(N+m) = r_1^i(x^*(N+m)) + P_1(x_1^*(N+m); i)V_2^*(N+m) \qquad (1)$$

$$V_1^{*1}(N+m) \geq r_1^1(x^*(N)) + P_1(x_1^*(N); 1)V_2^*(N+m) \qquad (2)$$

$$V_1^{*i}(N) \geq r_1^i(x^*(N+m)) + P_1(x_1^*(N+m); i)V_2^*(N) \qquad (3)$$

$$V_1^{*1}(N) = r_1^1(x^*(N)) + P_1(x_1^*(N); 1)V_2^*(N). \qquad (4)$$

Taking equations $[(1)-(2)]-[(3)-(4)]$ gives

$$\bar{V}_1^{*i}(N+m) - \bar{V}_1^{*i}(N) \leq [P_1(x_1^*(N+m); i) - P_1(x_1^*(N); 1)][V_2^*(N+m) - V_2^*(N)]. \quad (5)$$

Taking $(1)-(3)$ gives

$$V_1^*(N+m) - V_1^*(N) \leq P_1(x_1^*(N+m))[V_2^*(N+m) - V_2^*(N)].$$

Similarly, for all $k \leq N$,

$$V_k^*(N+m) - V_k^*(N) \leq P_k(x_k^*(N+m))[V_{k+1}^*(N+m) - V_{k+1}^*(N)].$$

Using this fact recursively we can refine (5) to $\bar{V}_1^{*i}(N+m) - \bar{V}_1^{*i}(N)$

$$\leq [P_1(x_1^*(N+m);i) - P_1(x_1^*(N);1)] \left[\prod_{n=2}^{N-1} P_n(x_n^*(N+m))\right] [V_N^*(N+m) - V_N^*(N)].$$

Noting that $V_N^*(N) = 0$, the right hand side equals

$$[P_1(x_1^*(N+m);i) - P_1(x_1^*(N);1)] \sum_{n=N}^{N+m-1} T_2^n(x^*(N_m))R_n(x^*(N_m) \leq 2\epsilon_2(N)\bar{R},$$

by Assumption (1) and Lemma 4.

A similar analysis for the opposite inequality allows strengthening of the result to

$$|\bar{V}_1^{*i}(N+m) - \bar{V}_1^{*i}(N)| \leq 2\epsilon_2(N)\bar{R}. \qquad (6)$$

Since $\epsilon_2(N) \rightarrow 0$ in N, the sequence $\{\bar{V}_1^{*i}(N)\}$ is Cauchy and converges. In the product topology, vector convergence is equivalent to pointwise convergence so $\{\bar{V}_1^*(N)\}$ converges. It must converge to $\bar{V}_1^*$ or any limit point of the sequence $\{x^*(N)\}$ would have greater value than x^*. Since $\lim_{m\rightarrow\infty} \bar{V}_1^*(N+m) = \bar{V}_1^*$, and (6) holds for all m, we conclude that $|\bar{V}_1^*(N) - \bar{V}_1^*| \leq 2\epsilon_2(N)\bar{R}$.■

Recall that $\bar{V}_1^* = \lim_{N\rightarrow\infty} V_1(x^*;N)$, where x^* is algorithmically optimal. An immediate corollary of Lemma 9 is that $\bar{V}_1^*$ takes the same values for any algorithmically optimal x^*.

For ease of expression we introduce the notation

$$v^i(x_0) = r_0^i(x_0) + P_0(x_0;i)\bar{V}_1^*;$$

and

$$v^i(x_0;N) = r_0^i(x_0) + P_0(x_0;i)\bar{V}_1^*(N).$$

Lemma 10: Under Assumptions (1) through (3), if x_0^* is the first policy of an algorithmically optimal strategy, it maximizes $v^{i_0}(x_0)$ for the starting state i_0.

Proof: By definition of algorithmic optimality, there exists $\{x^{*i_0}(N_m)\}$ beginning with $x_0^{*i_0}$. That is, for any x_0,

$$r_0^{i_0}(x_0^*) + P_0(x_0^*; i_0)V_1^*(N_m) \geq r_0^{i_0}(x_0) + P_0(x_0; i_0)V_1^*(N_m).$$

For any x_0,

$$\begin{aligned} v^{i_0}(x_0; N_m) =& r_0^{i_0}(x_0) + P_0(x_0; i_0)[V_1^*(N_m) - (V_1^{*1}(N_m))e] \\ =& r_0^{i_0}(x_0) + P_0(x_0; i_0)V_1^*(N_m) - V_1^{*1}(N_m)P_0(x_0; i_0)e \end{aligned}$$

since the series are convergent over the finite horizon (note that $V_1^{*1}(N_m)$ is a scalar). This last term is a constant over all x_0 since P_0 is stochastic. Hence $v^{i_0}(x_0^*; N_m) \geq v^{i_0}(x_0; N_m)$. From Lemma 9 and the definition of $v^{i_0}(x_0; N_m)$, $v^{i_0}(x_0; N_m) \to v^{i_0}(x_0)$. Hence, $v^{i_0}(x_0^*) \geq v^{i_0}(x_0)$. ∎

Assumption (5): The maximizer of $v^{i_0}(x_0)$ is unique.

Corollary 11: Under Assumptions (1) through (5), x_0^* maximizes $v^{i_0}(x_0)$ if and only if it is the first decision of an algorithmically optimal strategy.

We now propose a stopping rule for a value iteration procedure.

3.2 A Stopping Rule

In this section we propose a stopping rule inspired by that presented in Bès and Lasserre [1986] in the discounted cost case and Hernandez-Lerma and Lasserre in the (homogeneous) average cost case. At each step of the value iteration procedure we use a test which permits us to eliminate actions which are not optimal for the average cost criterion. Tests in previous works such as Shapiro [1968] or Hinderer and Hubner [1974], include parameters which are difficult to obtain since they require knowledge of the optimal average cost. We only use known parameters. The most difficult values to obtain are the $\epsilon_2(N)$. In problems with a geometrically decreasing coefficient of ergodicity, these values are easily calculated from problem structure.

The test is based on the following theorem.

Theorem 12: Under Assumptions (1) through (5), an initial action for state i, x_0^i, is *not* algorithmically optimal if and only if there exists a time N such that

$$[r_0^i(x_0^*(N)) - r_0^i(x_0)] + [P_0(x_0^*(N);i) - P_0(x_0;i)]\bar{V}_1^*(N) > 2\epsilon_2(N)\bar{R}.$$

That is, if and only if $v^i(x_0^*(N)) - v^i(x_0) > 2\epsilon_2(N)\bar{R}$.

Proof: *(if)* Let x_0^* be the first policy of an algorithmically optimal strategy. By Corollary 11, $v^i(x_0^*) \geq v^i(x_0^*(N))$. By Lemma 9, for any i, $|\bar{V}_1^{*i}(N) - \bar{V}_1^{*i}| \leq 2\epsilon_2(N)\bar{R}$. For any x_0, a simple algebraic manipulation extends this to

$$|v^i(x_0;N) - v^i(x_0)| \leq 2\epsilon_2(N)\bar{R}. \tag{7}$$

From (7), we have $v^i(x_0^*(N)) \geq v^i(x_0^*(N);N) - 2\epsilon_2(N)\bar{R}$. Combining gives

$$v^i(x_0^*) \geq v^i(x_0^*(N);N) - 2\epsilon_2(N)\bar{R}.$$

Since $x_0^*(N)$ maximizes $v^i(x_0;N)$, we have $v^i(x_0^*(N);N) \geq v^i(x_0^*;N)$. From (7), $v^i(x_0^*) \leq v^i(x_0^*;N) + 2\epsilon_2(N)\bar{R}$ for any i. Combining gives

$$v^i(x_0^*) \leq v^i(x_0^*(N);N) + 2\epsilon_2(N)\bar{R}, \quad \text{for all } i.$$

Summarizing,

$$|v^i(x_0^*) - v^i(x_0^*(N);N)| \leq 2\epsilon_2(N)\bar{R}, \quad \text{for all } i.$$

This proves the contrapositive of the *(if)* direction of the theorem.

(only if) Again by contrapositive, if, for all N, $0 \leq v^i(x_0^*(N);N) - v^i(x_0;N) \leq 2\epsilon_2(N)\bar{R}$, by letting N go to infinity we have that $v^i(x_0) = v^i(x_0^*)$ and, by Corollary 11, x_0 is algorithmically optimal. ∎

3.3 Algorithm Statement

Given an initial state i_0, we propose an algorithm to obtain, the optimal decision x^{*i_0} which begins the algorithmically optimal (also average optimal) strategy. It is the classical value iteration procedure plus a test based on the previous theorem which permits us to eliminate nonoptimal actions without computing an optimal strategy.

Step 0: Initialize $U(i_0)$, the (finite) set of potentially optimal actions in state i_0 at time 0.

Step 1: For all $x_0^{i_0} \in U(i_0)$, if $[r_0^{i_0}(x_0^*(N)) - r_0^{i_0}(x_0)] + [P_0(x_0^*(N); i_0) - P_0(x_0; i_0)]\bar{V}_1^*(N) > 2\epsilon_2(N)\bar{R}$, then eliminate $x_0^{i_0}$ from $U(i_0)$. If $U(i_0)$ is a singleton then stop. Else, set $N := N + 1$ and go to Step 1.

Corollary 13: Under Assumptions (1) through (5), the algorithm is guaranteed to stop in a finite number of steps with $U(i_0) = \{x_0^{*i_0}\}$.

Proof: By Assumption (5) there is a unique maximizer of $v^{i_0}(x_0)$. The algorithm will eliminate any other x_0 by Theorem 12. ∎

To carry out this algorithm we know all necessary data except $\epsilon_2(N)$, which depends on special structure. For example, if weak ergodicity is indeed geometric, i.e., $\tau_1(P_k(x)) \leq \beta < 1$ for all x and k, then $\epsilon_2(N)$ can be computed from the tail of the geometric series as in Bès and Sethi.

Note that this algorithm permits us to obtain on-line an optimal average reward strategy. At time 0, in state i_0, run the preceding algorithm until $x_0^{*i_0}$ is obtained. Implement this decision and observe the new state i_1 at time 1. Rerun the above algorithm where the index 0 is replaced by 1, 1 by 2, and the initial state is i_1. In the same manner, compute $x_1^{*i_1}$ and iterate. At time n we observe the state i_n after we have implemented the sequence of actions $x_0^{*i_0}, x_1^{*i_1}, \ldots, x_{n-1}^{*i_{n-1}}$. Computation of an average optimal policy would be impossible since we have an infinite number of states. Here, since we are interested in the optimal action at state i_0 at time 0, we need only evaluate the value functions $V_k(x(N); N)$ $k = 0, \ldots, N-1$, at states accessible in k steps from i_0. They are finite in number by Assumption (4).

4. Summary and Conclusions

We have shown that the analytical framework developed in Hopp, Bean and Smith for the finite state space nonhomogeneous Markov decision process can be generalized to the denumerable state case. Solution horizons exist if the problem is weakly ergodic.

The algorithm of Hernandez-Lerma and Lasserre is generalized to this problem. These results and algorithms can also be applied to the well studied denumerable stationary problem as a special case.

REFERENCES

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

Bès, C. and J. B. Lasserre [1986], "An On-line Procedure in Discounted Infinite Horizon Stochastic Optimal Control," **Journal of Optimization Theory and Applications**, Vol. 50, pp. 61-67.

Cavazos-Cadena, R. [1986], "Existence of Optimal Stationary Policies in Average Reward Markov Decision Processes with a Recurrent State," Submitted to **Applied Mathematics and Optimization.**

Federgruen, A., P. Schweitzer and H. Tijms [1978], "Contraction Mappings Underlying Undiscounted Markov Decision Problems," **Journal of Mathematical Analysis and Applications**, Vol. 65, pp. 711-730.

Hernandez-Lerma, O. and J. B. Lasserre [1988], "A Forecast Horizon and a Stopping Rule for General Markov Decision Processes," **Journal of Mathematical Analysis** and **Applications**, Vol. 132, pp. 388-400.

Hinderer, K. and G. Hubner [1974], "An Improvement of Shapiro's Turnpike Theorem for the Horizon of Finite Stage Discrete Dynamic Programs," **Trans. 7th Prague** Conf. **on Information Theory, Statistical Decision Functions, Random** Processes.

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. To appear in **Operations Research**

Hopp, W., J. Bean and R. Smith [1987], "A New Optimality Criteria for Non–Homogeneous Markov Decision Processes," **Operations Research**, Vol. 35, pp. 875-883.

Iosifescu, M. [1972], "On Two Recent Papers on Ergodicity in Nonhomogeneous Markov Chains," **The Annals of Mathematical Statistics**, Vol. 43, pp. 1732-1736.

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

Ryan, S., J. Bean and R. Smith [1987], "A Tie-Breaking Algorithm for Discrete Infinite Horizon Optimization," Technical Report 87-24, Department of Industrial and Operations Engineering, The University of Michigan, Ann Arbor, Michigan 48109.

Schochetman, I. and R. Smith [1989], "Infinite Horizon Optimization," **Mathematics of Operations Research**, Vol. 14, pp. 1-16.

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