JOURNAL OF MATHEMATICAL ANALYSIS AND APPLICATIONS 179, 525–536 (1993)

Optimal Average Value Convergence in Nonhomogeneous Markov Decision Processes*

YUNSUN PARK

Department of Industrial Engineering, Myungji University, Yongin-Eup Yongin-Kun, Kyungkee-Do, Seoul, South Korea

AND

JAMES C. BEAN AND ROBERT L. SMITH

Department of Industrial and Operations Engineering, The University of Michigan, Ann Arbor, Michigan 48109-2117

Submitted by E. Stanley Lee

Received January 10, 1992

We address the undiscounted nonhomogeneous Markov decision process with average reward criterion and prove two structural results. First, we establish equivalence of this problem to a discounted Markov decision process by means of an ergodic coefficient embedded in the original problem. Second, we prove, for the original problem, that the optimal finite horizon average values converge to the infinite horizon optimal average value under an ergodic condition. © 1993 Academic Press, Inc.

1. INTRODUCTION

Many problems can be modeled as Markov decision processes, but are not necessarily homogeneous. That is, rewards or transitions may be time dependent. Examples include R & D modelling (Nelson and Winter [12]), capacity expansion (Freidenfelds [6], Luss [11]), equipment replacement (Lohmann [10]), and inventory control (Sobel [16]). In some of these applications average reward criteria are more appropriate than discounted objectives.

* This work was supported in part by the National Science Foundation under Grants ECS-8700836, DDM-9018515, and DDM-9214894 to the University of Michigan.

0022-247X/93 $5.00

In this paper we address the nonhomogeneous Markov decision process with the objective to maximize average reward. This analysis is complicated by the fact that the average reward criterion is tail driven. That is, whatever is done during any finite leading solution segment is irrelevant to the final objective value. In many homogeneous problems this is not a concern since the tail is an exact replica of the original problem. Such is not the case in nonhomogeneous problems. To further complicate the issue, in nonhomogeneous problems we are often interested only in a leading strategy segment since only it must be implemented now.

One approach is to transform the Markov decision process to an equivalent discounted problem. To accomplish this we generalize results of Ross [15] that consider homogeneous problems, and develop a more efficient transformation. We then adapt results from Alden and Smith [1] that were developed for finite state, primarily discounted problems.

We also prove convergence of the finite horizon average optimal values to the infinite horizon average optimal value. This facilitates planning or solution horizon approaches. Our goal is to establish the mathematical framework necessary to create algorithms to solve these problems.

Section 2 introduces the notation and definitions necessary for our discussion. In Section 3, we describe the relationship between three common ergodic coefficients. Section 4 proves equivalence between the nonhomogeneous Markov decision process to maximize average reward, and a discounted problem. In Section 5 we prove average optimal value convergence for the nonhomogeneous Markov decision process. Section 6 summarizes the results of this paper.

2. Notation and Definitions

We generalize the notation of Bean, Smith, and Lasserre [2] for undiscounted nonhomogeneous Markov decision processes.

We observe a process at time points $k = 0, 1, \ldots$ to be in one of a countable number of states $i = 1\ 2, \ldots$. The decision maker chooses a policy in stage k, $x_k \in X_k$, by selecting any actions, x_k^i, from finite sets, X_k^i, for states $i = 1, 2, \ldots$. An infinite horizon feasible strategy, x, is an infinite sequence of policies. It resides in the set of feasible strategies, X.

A finite horizon strategy, $x(k, N)$, is a sequence of policies from time k through time $N-1$. Even though a finite horizon feasible strategy consists of a finite number of policies, we require that $x(k, N) \in X$ by allocating arbitrary policies before time k and after time N. Also, if $k = 0$, denote $x(N) \equiv x(0, N)$. We use an asterisk to represent the optimality of an action, policy, or strategy in the minimum class to which it belongs. For example, $x^*(N)$ is an N-horizon optimal strategy.

The set of all feasible strategies, X, is compact in the metric topology presented in Bean, Smith, and Lasserre. The metric, ρ, is defined

$$\rho(x, \bar{x}) = \sum_{k=0}^{\infty} \sum_{i=1}^{\infty} \Phi_k^i(x, \bar{x})\, 2^{-(k+i)} \qquad \text{for all} \quad x, \bar{x} \in X,$$

and

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

Using this metric we define algorithmic optimality, first introduced by Hopp, Bean, and Smith [9].

DEFINITION. Under this topology, an infinite horizon strategy, $\hat{x}$, is called *algorithmically optimal* if, for some sequence of integers, $\{N_m\}_{m=1}^{\infty}$,

$$x^*(N_m) \to \hat{x} \qquad \text{in } \rho\text{-metric as } m \to \infty.$$

If we take action x_k^i in state i at time k, then, independent of past actions, two things happen:

(1) We gain a reward $r_k^i(x_k^i)$. The vector of such rewards is $R_k(x_k)$.

(2) We transit to states, j, at time $k+1$ according to the probability transition matrix $\{p_k^{ij}(x_k^i)\} \equiv P_k(x_k)$.

Note that both the rewards and transition probabilities may be stage dependent.

The basis for many optimality criteria is the finite horizon reward function. Given an infinite horizon strategy, x, and a one period discount factor, $0 < \alpha \leqslant 1$, the expected net present value of the total rewards from time k to time N, $N > k$, at the beginning of stage k, is written $V_k(x; N)$. Note that in evaluating $V_k(x; N)$, the first k policies of x are ignored. Let $V_k(x; N)$ map into $\mathfrak{R}^{\infty}$ with the ith element given by $V_k^i(x; N)$ which represents the expected net present profit from state i in stage k through stage N under strategy x. Note that $V_k^i(x^*(N); N) = V_k^i(x^*(k, N); N)$ for all $k = 0, 1, \ldots$ by the principle of optimality.

The value function from stage k to stage N is written as

$$V_k(x; N) = \sum_{n=k}^{N-1} \alpha^n T_k^n(x)\, R_n(x_n),$$

where

$$T_k^n(x) = \prod_{l=k}^{n-1} P_l(x_l), \qquad n > k \geqslant 0$$

$$T_k^n(x) = I, \qquad n \leqslant k.$$

Throughout the paper we make the following assumptions:

Assumptions. (1) The state space, I, is countable.

(2) The number of decisions available is finite for all states, i.e.,

$$|X_k^i| < \infty, \qquad \text{for all } i \text{ and } k,$$

where $|A|$ is the cardinality of set A.

(3) Rewards are uniformly bounded for all states and decisions, i.e., for some $\bar{R} < \infty$,

$$|r_k^i(x_k^i)| \leqslant \bar{R}, \qquad \text{for all} \quad i \in I,\ k = 0, 1, \ldots, \qquad \text{and} \qquad x_k^i \in X_k^i.$$

(4) From each state, i, at stage k, under strategy x, the set of reachable states, $\{j \mid p_k^{ij}(x_k^i) > 0\}$, is finite. That is, only a finite set of states is reachable in one transition from any state, under any action. Further, $\max\{j \mid p_k^{ij}(x_k^i) > 0\}$ is uniformly bounded over $x_k^i \in X_k^i$ for each stage, k.

In the infinite horizon problem, with discount factor α, $0 < \alpha \leqslant 1$, define x^* to be an α-optimal strategy if

$$V_0(x^*) \geqslant V_0(x), \qquad \text{for all} \quad x \in X,$$

where

$$V_0(x) = \lim_{N \to \infty} V_0(x; N).$$

This definition is valid if the limit exists. By Assumption (3) it exists whenever $\alpha < 1$. However, the primary interest of this paper is the case $\alpha = 1$. In this case it is possible that $V_0(x; N)$ diverges with N. Then we define x^* to be average optimal if

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

Assumption (3) implies that the lim inf is always finite.

3. Weak Ergodicity

In this section, we formally define weak ergodicity and the corresponding ergodic coefficients.

Let

$$\mathscr{P}_{n,N}(x) = \mathscr{P}_0 P_n(x_n)\, P_{n+1}(x_{n+1}) \cdots P_{N-1}(x_{N-1}),$$

where $\mathscr{P}_0$ is a starting vector (initial distribution). Let $\mathscr{Q}_{n,N}(x)$ be the same forward product with starting vector $\mathscr{Q}_0$. If $\mathscr{P} = (p_j)$ is a vector, we define the norm of $\mathscr{P}$ to be

$$\|\mathscr{P}\| = \sum_{j=1}^{\infty} |p_j|.$$

If $P = (p_{ij})$ is a square matrix, we define the norm of P to be

$$\|P\| = \sup_i \sum_{j=1}^{\infty} |p_{ij}|.$$

DEFINITION. A nonhomogeneous Markov decision process is called *weakly ergodic* if and only if, for all n,

$$\lim_{N \to \infty} \sup_{\mathscr{P}_0, \mathscr{Q}_0} \|\mathscr{P}_{n,N}(x) - \mathscr{Q}_{n,N}(x)\| = 0 \qquad \text{for all} \quad x \in X,$$

and is called *strongly ergodic* if and only if there exists a vector $q(x) = (q_1(x), q_2(x), \ldots)$, with $\|q(x)\| = 1$ and $q_i(x) \geqslant 0$ for all $i = 1, 2, \ldots$ such that for all n

$$\lim_{N \to 0} \sup_{\mathscr{P}_0} \|\mathscr{P}_{n,N}(x) - q(x)\| = 0 \qquad \text{for all} \quad x \in X.$$

That is, a nonhomogeneous Markov decision process is weakly ergodic if and only if it eventually loses memory of the starting vector. For a problem to be strongly ergodic, the process not only must lose memory, but also converge to a fixed probability vector.

It is often difficult to determine if any specific problem satisfies these definitions. To facilitate the identification of weak ergodicity, we define several ergodic coefficients: the *Ross* coefficient (a_0), the *Doeblin* coefficient (β), and the *Hajnal* coefficient (γ).

DEFINITION. Ergodic coefficients:

- Ross coefficient,

$$a_0 = \sup_k \sup_{x_k \in X_k} a_0(P_k(x_k)),$$

where $a_0(P_k(x_k)) = 1 - \sup_j \inf_i p_k^{ij}(x_k)$.

- Doeblin coefficient,

$$\beta = \sup_k \sup_{x_k \in X_k} \beta(P_k(x_k)),$$

where $\beta(P_k(x_k)) = 1 - \sum_{j=1}^{\infty} \inf_i p_k^{ij}(x_k)$.

- Hajnal coefficient,

$$\gamma = \sup_k \sup_{x_k \in X_k} \gamma(P_k(x_k)),$$

where $\gamma(P_k(x_k)) = 1 - \inf_{i_1, i_2} \sum_{j=1}^{\infty} \min(p_k^{i_1 j}(x_k), p_k^{i_2 j}(x_k))$.

We call a_0 the *Ross coefficient* since the homogeneous version was used by Ross [15]. For the nonhomogeneous case, Hopp, Bean, and Smith used this coefficient to prove the average optimality of an algorithmically optimal strategy. Alden and Smith used the *Doeblin coefficient* to show that the error between a rolling horizon strategy and a discounted optimal strategy goes to zero as the horizon is lengthed when $\beta < 1$. The *Hajnal* coefficient was first introduced by Dobrushin [4], followed by several papers and books such as Hajnal [7] and Paz [13, 14]. For applications of this coefficient, see Hopp [8].

The following lemma describes the relationship between the coefficients and the property of weak ergodicity. The proofs are straightforward and omitted.

LEMMA 1. (a) $a_0 < 1$ *if and only if* $\beta < 1$.

(b) $a_0 \geq \beta$.

(c) *If* $\beta < 1$ *then* $\gamma < 1$.

(d) *If any of* $a_0 < 1$, $\beta < 1$, *or* $\gamma < 1$, *then the nonhomogeneous Markov decision process is weakly ergodic.*

Even though we know from Lemma 1 that the *Hajnal* condition ($\gamma < 1$) is the weakest of the three implying weak ergodicity, we will use the *Doeblin* coefficient to show may of the results in Section 4. The advantage of the *Doeblin* coefficient is that we can transform the undiscounted Markov decision process into an equivalent *discounted* Markov decision process by exploiting β as a discount factor. We can also transform using a_0, but since $a_0 \geq \beta$, the *Doeblin* coefficient can lead to faster convergence when we solve the equivalent discounted problem.

4. A DISCOUNTED EQUIVALENT PROBLEM

The traditional transformation from a nonhomogeneous problem to a homogeneous problem defines states in the homogeneous problem as a (time, state) pair. If the original problem has a countable number of states, then so does the transformed problem. However, even if the nonhomogeneous problem satisfies the conditions for weak ergodicity in Lemma 1, the transformed, homogeneous problem may not. For example,

begin with a finite state problem where transitions in an even numbered stage occur with transition probability matrix P_{even} and in odd numbered stages follow P_{odd}. An equivalent homogeneous problem would have transition matrix.

$$P=\begin{bmatrix} \mathbf{0} & P_{\text{even}} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & P_{\text{odd}} & \mathbf{0} & \cdots \\ \vdots & \vdots & \vdots & & \end{bmatrix}.$$

Since each of the columns of P contains predominantly zeroes, none of the sufficient ergodic conditions, of which we are aware, are satisfied (see Federgruen and Tijms [5]).

We now present an improved transformation that preserves the Doeblin condition for weak ergodicity. Alden and Smith proved the finite state version of the following theorem. The extension to the countable state case is straightforward and omitted.

THEOREM 1. *Every one step probability transition matrix can be expressed as a convex combination of another stochastic matrix and a stable matrix (stochastic matrix with identical rows), using the ergodic coefficient β as a multiplier. That is, for all k and $x_k \in X_k$,*

$$P_k(x_k)=\beta\tilde{P}_k(x_k)+(1-\beta)L_k,$$

where $\tilde{P}_k(x_k)$ is a stochastic matrix, L_k is a stable matrix independent of x_k, and $0 \leqslant \beta \leqslant 1$.

Solving for $\tilde{P}_k(x_k)$ in the case $\beta > 0$ we have

$$\tilde{P}_k(x_k)=\frac{P_k(x_k)-(1-\beta)L_k}{\beta}, \qquad \text{for each } k, \text{ and for each } x_k.$$

Let (P) be the original problem defined in Section 2. Based on the above theorem, define another class of nonhomogeneous Markov decision processes, $(\tilde{P})$.

$(\tilde{P})$: The β-discounted nonhomogeneous Markov decision process with probability transition matrix $\tilde{P}_k(x_k)$, reward $R_k(x_k)$, value function $\tilde{V}_k(\cdot)$, infinite horizon optimal strategies $\tilde{x}^*$, and finite horizon optimal strategies $\tilde{x}^*(k, N)$.

We transform the original undiscounted nonhomogeneos Markov decision process, (P), into the β-discounted nonhomogeneous Markov decision process, $(\tilde{P})$, using the ergodic coefficient β. This generalizes an approach by Ross since it considers nonhomogeneous problems and uses

the more efficient Doeblin coefficient. The following lemma shows that the finite horizon optimal value of (P) can be obtained from $(\tilde{P})$ and the set of finite optimal solutions of (P) is equal to that of $(\tilde{P})$. It generalizes a result in Alden and Smith from finite states to countable states. It is similar in structure to the result in Alden and Smith, but the lemma below expresses $V_k^i(x^*(k, N); N)$ in terms of $\tilde{V}$ whereas the earlier paper expresses it in terms of V and $\tilde{V}$. We denote an element of the matrix L_k as L_k^{ij}. However, since L_k is stable, L_k^{ij} is independent of i.

LEMMA 2. *Under the condition that $\beta < 1$, we can represent the finite horizon optimal value of (P) as a function of the finite horizon optimal value of $(\tilde{P})$, i.e.,*

$$V_k^i(x^*(k, N); N) = \tilde{V}_k^i(\tilde{x}^*(k, N); N) + (1-\beta) \sum_{l=k+1}^{N} \sum_{j=1}^{\infty} L_{l-1}^{ij} \tilde{V}_l^j(\tilde{x}^*(l, N); N),$$

for all states, i; $k = 0, ..., N-1$.

Moreover, the finite horizon optimal strategy set of (P) is equal to that of $(\tilde{P})$, i.e.,

$$x^*(k, N) = \tilde{x}^*(k, N), \qquad \text{for all} \quad k = 0, ..., N-1.$$

Proof. Let $V_k^i(N) = V_k^i(x^*(N); N) = V_k^i(x^*(k, N); N)$.
We will prove the result by induction on k. For $k = N-1$,

$$\tilde{V}_{N-1}^i(N) = \max_{x_{N-1}^i} \{r_{N-1}^i(x_{N-1}^i)\} = V_{N-1}^i(N), \qquad \text{thus} \quad \tilde{x}^*(N-1, N) = x^*(N-1, N).$$

Now fix k in $0 \leqslant k \leqslant N-2$ and assume that the result holds from period $k+1$. If $\beta = 0$ then the result holds as above. If $\beta > 0$ then

$$\begin{aligned}
\tilde{V}_k^i(N) &= \max_{x_k^i} \left\{ r_k^i(x_k^i) + \beta \sum_{j=1}^{\infty} \tilde{p}_k^{ij} \tilde{V}_{k+1}^j(N) \right\} \\
&= \max_{x_k^i} \left\{ r_k^i(x_k^i) + \beta \sum_{j=1}^{\infty} \left[\frac{p_k^{ij}(x_k^i) - (1-\beta) L_k^{ij}}{\beta} \right] \tilde{V}_{k+1}^j(N) \right\} \\
&= \max_{x_k^i} \left\{ r_k^i(x_k^i) + \sum_{j=1}^{\infty} p_k^{ij}(x_k^i) \tilde{V}_{k+1}^j(N) \right. \\
&\qquad \left. - (1-\beta) \sum_{j=1}^{\infty} L_k^{ij} \tilde{V}_{k+1}^j(N) \right\}
\end{aligned}$$

$$
\begin{aligned}
&= \max_{x_k^i} \left\{ r_k^i(x_k^i) + \sum_{j=1}^{\infty} p_k^{ij}(x_k^i) \right. \\
&\quad \times \left[V_{k+1}^j(N) - (1-\beta) \sum_{l=k+2}^{N} \sum_{m=1}^{\infty} L_{l-1}^{jm} \tilde{V}_l^m(N) \right] \\
&\quad \left. - (1-\beta) \sum_{j=1}^{\infty} L_k^{ij} \tilde{V}_{k+1}^j(N) \right\} \\
&= \max_{x_k^i} \left\{ r_k^i(x_k^i) + \sum_{j=1}^{\infty} p_k^{ij}(x_k^i)\, V_{k+1}^j(N) \right\} \\
&\quad - (1-\beta) \sum_{l=k+1}^{N} \sum_{j=1}^{\infty} L_{l-1}^{ij} \tilde{V}_l^j(N) \\
&= V_k^i(N) - (1-\beta) \sum_{l=k+1}^{N} \sum_{j=1}^{\infty} L_{l-1}^{ij} \tilde{V}_l^j(N),
\end{aligned}
$$

since $L_{l-1}^{jm} = L_{l-1}^{im}$ for all i, j, which is the desired result. Also from the second to last equation, we can see the equivalence of the solution set, since the last term of that equation is independent of x_k^i. ▮

The above lemma is interesting since both the *finite* horizon optimal strategy and value of an original undiscounted nonhomogeneous Markov decision process problem, (P), can be obtained by solving the β-discounted nonhomogeneous Markov decision process problem, $(\tilde{P})$.

Now, we prove the main theorem of this section.

THEOREM 2. *Under Assumptions* (1)–(4) *and the condition that* $\beta < 1$, *any algorithmically optimal strategy for* $(\tilde{P})$ *is an average optimal strategy for* (P).

Proof. From Lemma 2, any algorithmically optimal strategy of $(\tilde{P})$ is an algorithmically optimal strategy of (P). Bean, Smith, and Lasserre [2, Theorem 5] shows that an algorithmically optimal solution is an average optimal solution under these hypotheses. Hence, the result follows. ▮

Now a traditional transformation to a discounted, homogeneous problem can be carried out.

5. AVERAGE VALUE CONVERGENCE

In a direct forecast horizon approach to the nonhomogeneous Markov decision process we seek to find the optimal policy for the first stage since we must implement that policy now. We proceed, as in Bean, Smith, and

Lasserre, by truncating the infinite horizon problem at some finite horizon. We then solve the finite horizon problem and test to see if the policy for the first stage is optimal for the infinite horizon problem.

Such approaches require convergence of the finite horizon optimal values to the infinite horizon optimal value. See the Appendix of Hopp, Bean, and Smith [9] for an example where this property fails. This convergence implies that the average value obtained by solving a sufficiently long finite horizon problem is within any chosen ε of the infinite horizon average optimal value. Average optimal value convergence justifies the truncation of an infinite horizon problem to a sufficiently long finite horizon problem; an approach commonly used for real world problems.

In this section, we establish conditions under which we can prove that optimal average values converge. If $\alpha < 1$ the question is trivial. Below we assume that $\alpha = 1$.

Mathematically, average optimal value convergence is

$$\liminf_{N\to\infty} \frac{V_0^i(x^*(N); N)}{N} = \liminf_{N\to\infty} \frac{V_0^i(x^*; N)}{N}, \qquad \text{for all} \quad i \in I.$$

We begin with a technical lemma and then prove the main theorem.

LEMMA 3. *Let $\hat{x}$ be an algorithmically optimal strategy so that for some $\{N_m\}_{m=1}^{\infty}$, $x^*(N_m) \to \hat{x}$. Then, under Assumptions* (1)–(4) *and the condition that $\beta < 1$, for any times, $k < N$, and state, i, $V_k^i(x^*(N); N) - V_k^i(\hat{x}; N) \leqslant 2\bar{R}/(1-\beta)$.*

Proof. By Assumption (4) and the definition of the ρ metric, for m sufficiently large, $\hat{x}$ and $x^*(N_m)$ agree in action for all states, j, and for all times through N. Hence,

$$V_k(x^*(N_m); N_m) = V_k(\hat{x}; N) + T_k^N(\hat{x})\, V_N(x^*(N_m); N_m). \tag{1}$$

Define $\bar{x}$ as the concatenation of policies from $x^*(N)$ for times 0 to N with policies from $x^*(N_m)$ for N and beyond. Then

$$V_k(\bar{x}; N_m) = V_k(x^*(N); N) + T_k^N(x^*(N))\, V_N(x^*(N_m); \geqslant N_m). \tag{2}$$

Subtracting (2) from (1) and recognizing the superiority of $x^*(N_m)$ to $\bar{x}$ over the horizon N_m we get

$$\begin{aligned} \mathbf{0} \leqslant\ & V_k(\hat{x}; N) - V_k(x^*(N); N) \\ & + [T_k^N(\hat{x}) - T_k^N(x^*(N))]\, V_N(x^*(N_m); N_m), \end{aligned} \tag{3}$$

where **0** is a vector of zeros. The last term in (3) is equal to

$$\sum_{n=N}^{N_m-1} \alpha^n [T_k^N(\hat{x}) - T_k^N(x^*(N))]\, T_N^n(x^*(N_m))\, R_n(x^*(N_m)).$$

By Lemma 4 of Bean, Smith, and Lasserre [2] this is bounded above by $2\bar{R}/(1-\beta)$. Hence, the result follows. ∎

THEOREM 3. *Under the Assumptions* (1)–(4) *and the condition that* $\beta < 1$, *the optimal average values of the finite horizon problems converge to the optimal value of the infinite horizon problem, that is,*

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

Proof. Let $\hat{x}$ be an algorithmically optimal strategy. The existence of such a strategy is proved in Theorem 1 of Bean, Smith, and Laserre [2]. For any given N, by Lemma 3,

$$V_0^i(x^*(N); N) - V_0^i(\hat{x}; N) \leqslant \frac{2\bar{R}}{(1-\beta)}.$$

Dividing by N and taking the lim inf gives

$$\liminf_{N\to\infty} \frac{V_0^i(x^*(N); N)}{N} = \liminf_{N\to\infty} \frac{V_0^i(\hat{x}; N)}{N}.$$

By Theorem 5 of Bean, Smith, and Lasserre [2]

$$\liminf_{N\to\infty} \frac{V_0^i(\hat{x}; N)}{N} = \liminf_{N\to\infty} \frac{V_0^i(x^*; N)}{N}.$$

Hence, the result follows. ∎

Theorem 3 suggests a conceptual, tail value algorithm similar to that in Bès and Lasserre [3] for the discounted problem.

(1) Choose $\varepsilon > 0$ and, by Theorem 3, choose N such that the average value of the finite horizon optimum is guaranteed to be within ε of the infinite horizon optimal average value.

(2) Find all strategies with finite horizon average value within 2ε of the optimal finite horizon value. By Theorem 3, no strategy outside of this set can be optimal. If all strategies in the set begin with the same policy at time 0, it must be optimal to the infinite horizon problem. Else, decrease ε and go to Step 1.

6. Summary

This paper presents several structural results for an infinite state nonhomogeneous Markov decision process with average reward criterion. First, under the Doeblin condition, the problem is shown to be equivalent to a discounted problem. Under the same condition, we show that the optimal finite horizon optimal average values converge to the infinite horizon optimal average value.

References

1. J. Alden and R. Smith, Rolling horizon procedures in nonhomogeneous Markov decision processes, *Oper. Res.* **40**, No. 2 (1992).
2. J. Bean, R. Smith, and J. Lasserre, Denumerable state nonhomogeneous Markov decision processes, *J. Math. Anal. Appl.* **153** (1990), 64–77.
3. C. Bès and J. Lasserre, An on-line procedure in discounted infinite horizon stochastic optimal control, *J. Optim. Theory Appl.* **50** (1986), 61–67.
4. R. Dobrushin, Central limit theorems for non-stationary Markov chains, II, *Theory Probab. Appl.* **1** (1956), 329–383.
5. A. Federgruen and H. Tijms, The optimality equation in average cost denumerable state semi-Markov decision problems, recurrency conditions and algorithms, *J. Appl. Probab.* **15** (1978), 356–373.
6. J. Freidenfelds, "Capacity Expansion: Simple Models and Applications," North-Holland, Amsterdam, 1981.
7. J. Hajnal, Weak ergodicity in nonhomogeneous Markov chains, *Math. Proc. Cambridge Philos. Soc.* **52** (1958), 67–77.
8. W. Hopp, Identifying forecast horizons in nonhomogeneous Markov decision processes, *Oper. Res.* **37** (1989), 339–343.
9. W. Hopp, J. Bean, and R. Smith, A new optimality criterion for nonhomogeneous Markov decision processes, *Oper. Res.* **35** (1987), 875–883.
10. J. Lohmann, "A Stochastic Replacement Economy Decision Model," Tech. Rep. No. 84–11, Department of Industrial and Operations Engineering, the University of Mchigan, Ann Arbor, MI 48109, 1984.
11. H. Luss, Operations research and capacity expansion problems: A survey, *Oper. Res.* **30** (1982), 907–947.
12. R. Nelson and S. Winter, "An Evolutionary Change of Economic Change," Belknap Press, 1982.
13. A. Paz, Graph-theoretic and algebraic characterizations of some Markov chains, *Israel J. Math.* **3** (1963), 169–180.
14. A. Paz, "Introduction to Probabilistic Automata," Academic Press, New York, 1971.
15. S. Ross, Non-discounted denumerable Markov decision models, *Ann. Math. Statist.* **39** (1968), 412–423.
16. N. Sobel, Production smoothing with stochastic demand. II. Infinite horizon case, *Management Sci.* **17** (1971), 724–735.