Journal of the Korean
OR/ MS Society
Vol. 17, No. 2, August 1992

Equivalent Transformations of Undiscounted Nonhomogeneous Markov Decision Processes

Yunsun Park*

Abstract

Even though nonhomogeneous Markov Decision Processes subsume homogeneous Markov Decision Processes and are more practical in the real world, there are not many results for them. In this paper we address the nonhomogeneous Markov Decision Process with objective to maximize average reward. By extending works of Ross [17]in the homogeneous case adopting the result of Bean and Smith [3] for the discounted deterministic problem, we first transform the original problem into the discounted nonhomogeneous Markov Decision Process. Then, secondly, we transform into the discounted deterministic problem. This approach not only shows the interrelationships between various problems but also attacks the solution method of the undiscounted nonhomogeneous Markov Decision Process.

1. Introduction

Many problems can be modelled as Markov Decision Processes, but are not necessarily homogeneous. That is, rewards and transitions are time dependent. Examples include R&D modelling [15], capacity expansion [8, 14], equipment replacement [13], and inventory control [20]. In some of these applications average reward criteria are more appropriate than descounting.

In this paper we address the nonhomogeneous Markov Decision Process with objective to maximize average reward. This analysis is complicated by the facts that nonhomogeneous problems do not have average value functions and that the average reward criteria is tail driven (see Appendix). That is, whatever is done during any finite leading strategy segment is irrelevant (a *strategy* is an infinite sequence of state dependent decisions). In homogeneous problems, under certain ergodic conditions, this is not a concern since the tail is an exact

* Department of Industrial Engineering, Myongji University

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. Nonetheless, under certain ergodic conditions, some average optimal strategies can be found and shown to make sense. This paper will establish these conditions, substantiate the reasonability of the solutions and suggest approaches for finding them. Our goal in this paper is to establish the mathematical framework necessary to create algorithms to solve these problems.

The traditional approach to a nonhomogeneous problem transforms it to a homogeneous problem. Standard transformations are available, but result in a homogeneous problem lacking the necessary ergodic conditions. Even if each transition in the nonhomogeneous problem is well behaved, the homogeneous equivalent is not. Section 4 includes an example.

In this paper we consider two approaches to the nonhomogeneous Markov Deciosion Process with average reward criteria : an improved transformation that preserves necessary conditions, and a direct approach based on forecast horizons. The former extends work of Ross [17] and Alden and Smith [1]. The latter extends work of Hopp, Bean and Smith [11] and Bean, Smith and Lasserre [2].

Section 2 introduces the notation and definitions necessary for our discussion. In Section 3, we define weak ergodicity, present three ergodic coefficients, and describe the relationship between these coefficients. Section 4 discusses transformations using the Doeblin coefficient for undiscounted non homogeneous Markov Decision Processes. Section 6 summarizes this paper. Finally, Appendix shows the abnormal behavior of average value function.

2. Notation and Definitions

We consider the undiscounted nonhomogeneous Markov Decision Processes, and generalize the notation of Bean, Smith and Lasserre [2].

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

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 finite number of policies, we assume 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 is denoted by X which is compact in the metric topology introduced in Bean, Smith and Lasserre [2], where the metric, ρ, is defined.

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

and

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

Using this metric we define a concept of optimality that will be cited frequently in this work.

Definition. Under this topology, an infinite horixpn strategy, $\hat{x}$, is called *algorithmically optimal if*, for some sequence of integers, $\{N_m\}_{m=0}^{\infty}$,

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

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

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

The basis for many optimality criteria is the finite reward function. Given an infinite horizon strategy, x, and a one period discount factor, $0 \leqq \alpha \leqq 1$, the expected net present value of the total rewards from time k through 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. Then $V_k(x\,;N)$ maps into R^{∞} with the i^{th} 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 $k=0, 1, \cdots$, by the principle of optimality.

In general, we are interested in the value function from stage 0 onward, which is written :

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

where

$$T_l^n(x) = \prod_{k=l}^{n-l} P_k(x_k), \quad n > l \geqq 0$$
$$T_0^0(x) = I.$$

Throughout the paper we make the following assumptions :

Assumptions

1. The state space, I, is countable.
2. The number of decisions available in each state is finite for all states, i.e.,

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

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

$$r_k^i \leqq \bar{R} \text{ , for all } i\in I \text{ and } x_k^i \in X_k^i.$$

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

$$V_\alpha(x^*)-V_\alpha(x)\geqq 0, \text{ for all } x\in X,$$

where

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

This definition is valid if the limits exist. 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 an average optimal strategy if

$$\liminf_{N\to\infty}\frac{V_0(x^*;N)}{N}-\liminf_{N\to\infty}\frac{V_0(x;N)}{N}\geqq 0, \text{ for all } x\in X.$$

Assumption 3 implies that this lim inf always exists.

3. Weak Ergodicity

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

Let

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

where $\mathcal{P}_0$ is a starting vector (initial distribution). Similarly,

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

where $\mathcal{Q}_0$ is a starting vector (initial distribution). If $\mathcal{P}=(p_j)$ is a vector, we define the norm of $\mathcal{P}$ to be

$$\|\mathcal{P}\| = \sum_{j=1}^{x} |p_j|.$$

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

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

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

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

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

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

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

It is difficult to determine if any specific problem satisfies this definition. To facilitate the identification of weak (strong) ergodicity, we define several ergodic coefficients : 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 $\beta(P_k(x_k)) = 1 - \sum_{j=1}^{x} \inf_i P_k^{ij}(x_k)$.

- Hajnal coefficient :

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

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

We call a_0 Ross' coefficient since the homogeneous version was used in Ross [17] to show the existence of a stationary optimal strategy. For the nonhomogeneous case, Hopp, Bean,

and Smith [11] used this coefficient to prove the average optimality of an algorithmically optimal strategy. Alden and Smith [1] used the *Doeblin coefficient* to show that the error between the rolling horizon strategy and the (average) optimal strategy goes zero when the *Doeblin* coefficient, β, is less than 1. The *Hajnal* coefficient was first introduced by Dobrushin [6], followed by several papers and books such as Hajnal [9], Paz [16]. For applications of this coefficient, see Hopp [10].

Now, we state some well known results on identification of weak ergodicity through ergodic coefficients.

Lemma 1. *a)* (Seneta [19]) *If P and Q are stochastic matrices,*

$$\gamma(QP) \leqq \gamma(Q)\gamma(P).$$

b) (Isacson and Madsen [12]) *A nonhomogeneous Markov decision process is weakly ergodic if and only if, for all n,* $\gamma(T_n^N(x)) \to 0$ *as* $N \to \infty$ *for any feasible strategy x.*

c) (from a) and b)) A nonhomogeneous Markov Decision Process is weakly ergodic if $\gamma < 1$.

The following lemma describes the relationship between the coefficients. Proofs are straightforward and omitted.

Lemma 2. *a)* $a_0 < 1$ *if and only if* $\beta < 1$.

b) $a_0 \geqq \beta$.

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

d) if $a_0 < 1$, *if* $\beta < 1$, *or if* $\gamma < 1$.

Even though we know from Lemma 2 that the *Hajnal* condition ($\gamma < 1$) is the weakest of the three, we will use the *Doeblin* coefficient to show many results in this Section. The advantage of the *Doeblin* coefficient is that we can transform the undiscounted Markov decision process into an equivalent *discounted* Markov decision process exploiting β as a real discount factor. We can also transform using $a_0 \geqq \beta$, the *Doeblin* coefficient may lead to faster convergence when we solve the transformed discounted problem.

4. Transformation into Equivalent problems

The traditional transformation for a nonhomogeneous problem to a homogeneous problem defines states in the homogeneous problem as a (time, state) pair. If the original problem has countable states, then so does the resultant problem. However, even if the nonhomogeneous problem satisfies any of the conditions for weak ergodicity, the transformed problem may not. For example, begin with a finith state problem where transitions in an even numbered stage occur with transition probability matrix $\mathbb{P}1$ and in odd numbered stages follow $\mathbb{P}2$. An equivalent homogeneous problem would have transition matrix

$$\mathbf{P}=\begin{matrix} \mathbb{0} & \mathbb{P}1 & \mathbb{0} & \mathbb{0} & \cdots \\ \mathbb{0} & \mathbb{0} & \mathbb{P}2 & \mathbb{0} & \cdots \\ \vdots & \vdots & \vdots & & \end{matrix}$$

Since each of the columns of $\mathbf{P}$ contains predominantly zeroes, none of the conditions for reasonable behavior, of which we are aware, are satisfied (see Federgruen and Tijms [7]).

We now present an improved transformation that preserves the Doeblin condition for weak ergodicity.

We define two Markov Decision Processes using the Doeblin coefficient :

PN 1 The undiscounted nonhomogeneous Markov decision process with probaility transition matrix $P_k(x_k)$, reward $R_k(x_k)$, value function $V_k(\cdot)$, average value function $A_k(\cdot)$, infinite horizon average optimal solution x^* and finite horizon optimal strategy from time k through time N, $x^*(k, N)$.

Alden and Smith [1] proposed the following theorem.

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

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

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

Solving for $\tilde{P}_k(x_k)$ we have

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

Based on the above theorem, define another class of nonhomogeneous Markov Decision Processes, **PN2**.

PN2 The β−discounted nonhomogeneous Markov decision process with probability transition matrix $\tilde{P}_k(x_x)$, reward $R_k(x_x)$, value function $\tilde{V}_k(\cdot)$, infinite horizon optimal strategy $\tilde{x}^*$ and finite horizon optimal strategy from time k through time N, $\tilde{x}^*(k,N)$.

This section will demonstrate solution methodologies for the undisconted nonhomogeneous Markov Decision Process, **PN1**. The *Doeblin* coefficient will also be used to show solution procedures for **PN1** by transformations. Throughout this section the Assumptions of Section 2 will be in effect.

As defined earlier, an infinite horizon optimal strategy is called an algorithmically optimal if it is the cluster (accumulation) point of the finite horizon optimal strategies. Bean, Smith and Lasserre [2] show that an algorithmically optimal strategy exists for the undiscounted nonhomogeneous Markov Decision Process, and that when the problem is weakly ergodic, an algorithmically optimal solution is average optimal. With the above fact and definition in mind, we transform the original undiscounted nonhomogeneous Markov Decision Process, **PN1** into the β−discounted nonhomogeneous Markov Decision Process, **PN2**, using the ergodic coefficient β. This generalizes an approach by Ross [17] since it considers nonhomogeneous problems and uses the slightly more effcient Doeblin coefficient. The following lemma shows that the finite horizon optimal value of **PN1** can be obtained from **PN2** and the set of finite optimal solutions of **PN1** is equal to that of **PN2**.

Lemma 3. *Under the condition that* $\beta<1$, *we can represent the finite horizon optimal value of* **PN1** *as a function of the finite horizon optimal value of* **PN2**, *i.e.*,

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

for all $i\in I$, $k=0, \cdots, N-1$.

Moreover, the finite optimal strategy set of **PN1** *is equivalent to that of* **PN2**, *i.e.*,

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

Proof. First, for convenience, let $V_k^i(N)$represent the optimal expected total rewards gained from stage k through N starting at state i in stage k, i.e., $V_k^i(N)=V_k^i(x^*(N);N)$.

We will prove the result by induction on k. For $k=N-1$,

$$\tilde{V}_{N-1}^i(N)=\max\{r_k^i(x_k^i)\}=V_{N-1}^i(N), \text{ thus } \tilde{x}^*(N-1,N)=x^*(N-1,N).$$

Now assume that the result holds from period $k+1$ to period $N-1$ for $1 \leqq k \leqq N-2$. Then,

$$
\begin{aligned}
\tilde{V}_k^i(N) &= \max_{x_k^i} \{r_k^i(x_k^i) + \beta \sum_{j=1}^{\alpha} \tilde{p}_k^{ij} \tilde{V}_{k+1}^j(N)\} \\
&= \max_{x_k^i} \{r_k^i(x_k^i) + \beta \sum_{j=1}^{\alpha} \left(\frac{p_k^{ij}(x_k^i) - (1-\beta) L_k^{ij}}{\beta}\right) \tilde{V}_{k+1}^j(N)\} \\
&= \max_{x_k^i} \{r_k^i(x_k^i) + \sum_{j=1}^{\alpha} p_k^{ij}(x_k^i) \tilde{V}_{k+1}^j(N) - (1-\beta) \sum_{j=1}^{\alpha} L_k^{ij} \tilde{V}_{k+1}^j(N)\} \\
&= \max_{x_k^i} \{r_k^i(x_k^i) + \sum_{j=1}^{\alpha} p_k^{ij}(x_k^i) \tilde{V}_{k+1}^j(N) - (1-\beta) \sum_{l=k+2}^{N} \sum_{m=1}^{\alpha} L_k^{im} \tilde{V}_l^m(N)) - (1-\beta) \sum_{j=1}^{\alpha} L_k^{ij} \tilde{V}_{k+1}^N(N)\} \\
&= \max_{x_k^i} \{r_k^i(x_k^i) + \sum_{j=1}^{\alpha} p_k^{ij}(x_k^i) V_{k+1}^j(N)\} - (1-\beta) \sum_{l=k+1}^{N} \sum_{j=1}^{\alpha} L_k^{ij} \tilde{V}_e^j(N) \\
&= V_k^i(N) - (1-\beta) \sum_{l=k+1}^{N} \sum_{j=1}^{\alpha} L_k^{ij} \tilde{V}_i^j(N),
\end{aligned}
$$

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 solution and value of an original undiscounted nonhomogeneous Markov Decision Process problem, PN1, can be obtained by solving the β−discounted nonhomogeneous Markov Decision Process problem, PN2.

Now, we prove the main theorem of this section which shows the equivalence between the average optimal strategies of PN1, and PN2.

Theorem 2. *Under the condition that $\beta<1$, any algorithmically optimal strategy to PN2 is an average optimal strategy to PN1.*

Proof. From Lemma 3, we can conclude that any algorithmically optimal strategy of PN1 is an algorithmically optimal strategy of PN2, by the definition of an algorithmically optimal strategy. Hopp, Bean, and Smith [11] showed that an algorithmically optimal solution is an average optimal solution when the ergodic coefficient, $a_0<1$, which means $\beta<1$ (recall Lemma 2). Hence, the theorem is justified. ■

Thus, to obtain an average optimal strategy of the original undiscounted nonhomogeneous Markov Decision Process, PN1, we can obtain an algorithmically optimal strategy of the transformed β−discounted nonhomogeneous Markov Decision Progess, PN2. Bean and Smith [2] developed an algorithm to find an algorithmically optimal strategy when the problem is discounted and deterministic, and when the optimal strategy of the discounted deterministic problem is unique.

Unfortunately, **PN2** is not a deterministic problem, although it is β−discounted. The next theorem shows that we can represent **PN2** deterministically by taking probability destributions as states. Then, by solving the deterministic version of **PN2**, we can obtain the algorithmically optimal strategy of **PN2**.

Theorem 3. *The discounted denumerable nonhomogeneous Markov Decision Process problem,* **PN2**, *can be represented as a deterministic problem, adopting probability distributions as states, and can be solved by applying the deterministic solution algorithm developed by Bean and Smith* [*2*].

Proof. Consider a β−discounted nonhomogeneous Markov Decision Process, **PN2**. Let $\mathcal{P}_k=(p_k^j, j\in I)$ be a probability distribution over states 1, 2, ⋯ at time , and let

$f_{\mathcal{P}k}$ = an optimal expected value of beginning at stage 0 with state distribution P_k

$= \max\{\sum_{i\in I} r_k^i(x_k^i)p_k^i+\beta f p_{k+1}\}$, $p_{k+1}\in\Phi_{k+1}$ for $k=0, 1, 2, \cdots$,

where

$$p_{k+1}^j=\sum_{i\in I} p^{ij}p_k^i,\quad j\in I$$

Φ_k=set of feasible distributions at stage k.

There is nothing stochastic in the above functional equation, since if we know $\mathcal{P}_0$, we can have all $\mathcal{P}_k$'s recursively. So we have represented **PN2** as *a deterministic problem*, which can be solved by the deterministic algorithm developed by Bean and Smith [2]. ■

The next corollary summarizes the content of this section.

Corollary 1. *Under the condition that* $\beta<1$, *and when the algorithmically optimal strategy of* **PN1** *is unique, we can find an average optimal strategy of* **PN1** *by transforming it into* **PN2** *and solving deterministically.*

Proof. From Theorem 2, Theorem 3, and the uniqueness theorem (Theorem 6) in Bean and Smith [3]. ■

Thus, to solve **PN1**, we perform the following transformations :

- **PN1** ⇒ **PN2**
- **PN2** ⇒ an equivalent β−discounted deterministic problem.

Then, we can apply the deterministic solution algorithm to the β−discounted deterministic problem to solve the undiscounted nonhomogeneous Markov Decision Process.

5. Conclusion

This paper presents an approach for solving the nonhomogeneous Markov Decision Process with average reward criterion. First, under the *Doeblin* condition, we transform the original problem into an equivalent undiscounted homogeneous Markov Decision Process. Then, we transform again into an equivalent discounted deterministic problem, taking, the *Doeblen* coefficient as a discount factor. Since we know how to solve the discounted homogeneous problem, we can solve the original problem.

APPENDIX : Behavior of the Average Value Function

The following examples show the abnormal (with respect to the discounted value function) behavior of the average value function.

Question 1. (*the deterministic case*) Is the average value function continuous? That is, when $x^N \to x$ in the $\rho-$metric as $N\to\infty$ and $x^N,\ x\in X$ for all N, is it true that

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

(**counterexample**) Let

$$x^1=(a,\ b,\ b,\ b,\ \cdots)$$
$$x^2=(a,\ a,\ b,\ b,\ b,\ \cdots)$$
$$\vdots$$
$$x^N=(\underbrace{a,\ \cdots,\ a}_{N},\ b,\ b,\ b,\ \cdots)$$
$$\vdots$$

and a reward from taking decision a decision b be 0.

Then,

$$x=(a,\ a,\ a,\ \cdots)$$

However, since the average value function is tail-driven,

$$\liminf_{N\to\infty}\frac{V_0(x^N;N)}{N}=0,\ \text{and}$$

$$\liminf_{N\to\infty}\frac{V_0(x^N;N)}{N}=1$$

Thus, in general

$$\liminf_{N\to\infty}\frac{V_0(x^N;N)}{N}\neq\liminf_{N\to\infty}\frac{V_0(x;N)}{N}.\quad\blacksquare$$

Question 2. (*the Markov Decision Process case*) When a Markov Decision Process is weakly ergodic, is the average value function is continuous? That is, when the *Doeblin* coefficient, β, is less than 1 and $x^N \to x$ in the $\rho-$metric as $N \to \infty$ and x^N, $x \in X$ for all N, is it true that

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

(counterexample) We have a two state Markov Decision Process. The policy a consists of decisions, a_1 and a_2 and the policy b consists of decisions, b_1 and b_2, i.e.,

$$a=(a_1,\ a_2)^T \text{ and } b=(b_1,\ b_2)^T.$$

By taking the decision a_1 or a_2, a reward of 1 can be obtained. By taking the decision b_1 or b_2, no reward (0) can be obtained, i.e.,

$$R_N(a)=(1,1)^T \text{ and } R_N(b)=(0,0)^T.$$

The probability transition matrices for the policy a and the policy b are as the following.

$$R_N(a)=R_N(b)=\begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$$

Then, the Markov Decision Process is weakly ergodic with an ergodic coefficient, a_0, equal to 0.5.

Let

$$x^1=(a,\ b,\ b,\ b,\ \cdots)$$
$$x^2=(a,\ a,\ b,\ b,\ b,\ \cdots)$$
$$\vdots$$
$$x^N=(\underbrace{a,\ \cdots,\ a}_{N},\ b,\ b,\ b,\ \cdots)$$
$$\vdots$$

Then,

$$x=(a,\ a,\ a,\ \cdots)$$

However, since the average value function is tail-driven,

$$\liminf_{N\to\infty} \frac{V_0(x^N;N)}{N} = (0,\ 0)^T, \text{ and}$$

$$\liminf_{N\to\infty} \frac{V_0(x;N)}{N} = (1,\ 1)^T.$$

Thus, in general

$$\liminf_{N\to\infty}\frac{V_0(x^N;N)}{N} \neq \liminf_{N\to\infty}\frac{V_0(x;N)}{N}. \quad \square$$

References

[1] Alden, J. and R. Smith, "Rolling Horizon Procedures in Nonhomogeneous Markov Decision Processes," Technical Report 87-25, Department of Industrial and Operations Engineering, The University of Michigan, Ann Arbor, MI, 1987.

[2] Bean, J., R. Smith, and J. Lasserre, "Denumerable State Nonhomogeneous Markov Decision Processes," *Journal of Mathematical Analysis and Applications*, Vol. 153(1990), pp.64-77.

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

[4] Bean, J. and R. Smith, "Conditions for the Discovery of Solution Horizons," Technical Report 86-23, Department of Industrial and Operations Engineering, The University of Michigan, Ann Arbor, Michigan, 1986.

[5] Denardo, E., *Dynamic Programming: Models and Application*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1982.

[6] Dobushin, R, "Central Limit Theorems for Non-stationary Markov Chains II," *Theory of Probability and Its Applications*, Vol. 1(1956), pp.329-383.

[7] Federgruen, A. and H. Tijms, "The Optimality Equation in Average Cost Denumerable State Semi-Markov Decision Problems, Recurrency Conditions and Algorithms," *Journal of Applied Probability*, Vol. 15(1978), pp.356-373.

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

[9] Hajnal, J., "Weak Ergodicity in Nonhomogeneous Markov Chains," *Proceedings of the Cambridge Philosophical Society*, Vol. 52(1958), pp.67-77.

[10] Hopp, W., "Identifying Forecast Horizons in Nonhomogeneous Markov Decision Processes, *Operations Research*, Vol. 37(1989), pp.339-343.

[11] Hopp, W., J. Bean and R. Smith, "A New Optimality Criterion for Nonhomogeneous Markov Decision Processes," *Operations Research*, Vol. 35(1987), pp.875-883.

[12] Isaacson, L. and R. Madsen, *Markov Chains: Theory and Applications*, M.I.T. Press, 1976.

[13] Lohmann, J., "A Stochastic Replacement Economy Decision Model," Technical Report 84-11, Department of Industial and Operations Engineering, The University of Michigan, Ann Arbor, MI, 1984.

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

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

[16] Paz, A., *Introduction to Probabilistic Automata*, Academic Press Inc., New York, New York, 1971.

[17] Ross, S., "Non-Discounted Denumerable Markov Decision Models," *The Annals of Mathematical Statistics*, Vol. 39(1968), pp.412-423.

[18] Schochetman, I. and R. Smith, "Infinith Horizon Optimization," *Mathematics of Operations Research*, Vol. 14(1989), pp.559-574.

[19] Seneta, E., *Non-negative Matrices*, Halsted Press, New York, New York, 1973.

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