STUDIES ON A CLASS OF DIVERGENT MARKOV DECISION PROCESSES

A DISSERTATION SUBMITTED TO THE
INDIAN INSTITUTE OF TECHNOLOGY, KHARAGPUR
FOR THE AWARD OF THE DEGREE

OF

DOCTOR OF PHILOSOPHY
IN
SCIENCE

BY

KALAVENDI JAGADEESA SARMA

DEPARTMENT OF MATHEMATICS
INDIAN INSTITUTE OF TECHNOLOGY
KHARAGPUR - 721 302, INDIA
1980

CHAPTER I

GENERAL INTRODUCTION

CHAPTER I

GENERAL INTRODUCTION

1.1. Introduction

A discrete decision process is a process in which a finite or infinite sequence of decisions is executed on a system, serially at different points of time (epochs) and each decision earns a return or reward (cost may be regarded as a negative reward). There is also usually a system of discounting which inflates or deflates the returns from future decisions. In a general set up the decisions, rewards and discount factors at any epoch may depend on the epoch, the current state and the history of the process. The number of decisions to be taken may be known from beforehand. If altogether T-decisions have to be taken the process is called a T-horizon process. If T is infinity (or if the number of decisions to be taken are indefinite), the process may be called an infinite horizon process. The objective is to find the 'best' sequence of decisions with respect to a specific optimality criterion. For rigorous definition of a general discrete decision process in mathematical terms one may see, e.g., Porteus (1975b), Hindemer (1970), and Wang (1974).

In a Markov decision process the decisions, rewards and discount factors at any epoch depend only on the epoch and on the current state of the system and not on the past history.

A detailed definition of a general Markov decision process can be found in Kreps (1975), Strauch (1966), Blackwell (1976) and also in Porteus (1975b).

1.2. The Model of the Present Thesis

The class of Markov decision processes with which the present thesis is concerned can be outlined as follows.

A system can be in any one of a finite set $S = \{1, 2, \ldots\ldots N\}$ of states at any epoch n, $n = 0, 1, 2, \ldots,$. At state $i \in S$ a finite (non empty) set of actions or decisions $k \in A(i)$, $A(i) = \{1, 2, \ldots\ldots K(i)\}$ is open to the decision maker; the set $\bigcup_{i \in S} A(i) = A$ is called the action space. If at any epoch n the system is in state i, $i \in S$ and an action k, $k \in A(i)$ is implemented, then the system has a trnsition to state j, $j \in S$ with a probability p_{ij}^{k} ;

$$p_{ij}^{k} \geq 0, \quad \sum_{j=1}^{N} p_{ij}^{k} = 1, \quad i, j \in S, \quad k \in A(i) \; ; \qquad (1.2.1)$$

the triplet $(i, k, j) \in S \times A \times S$ is referred to as a transition.

An N-dimensional column vector

$f = (f(1), f(2), \ldots\ldots f(N))'$, $f(i) \in A(i)$, $i = 1, 2, \ldots,$ N , is called a decision function or policy function (policy), which is a map from S into A. The set of policy functions

is denoted by F and is called the policy space. The policy space is isomorphic to $\underset{i=1}{\overset{N}{X}} A(i)$ and is finite.

The $N \times N$ matrix $((p_{ij}^{f(i)})) = P(f)$, (1.2.2)

is called the transition probability matrix corresponding to f, $f \in F$.

It is assumed that a decision function is executed at each epoch n, $n = 0,1,2,\ldots\ldots,$. The interval $(n, n+1)$ is called the $(n+1)$th stage. The policy followed at the nth epoch is denoted by

$$f_{n+1} = (f_{n+1}(1),\ f_{n+1}(2),\ f_{n+1}(3),\ \ldots\ldots\ f_{n+1}(N))' \ ;$$

$$f_n \in F,\ n = 0,1,2,\ \ldots\ldots.$$

If altogether a finite number T of decision functions are to be implemented on the system the sequence $\{f_1, f_2, \ldots\ f_T\} = \pi_T$, called a T-horizon strategy; and T is referred to as the length of the planning horizon. When the planning horizon is very large or does not have any definite or forseeable end, we may say that we have an infinite planning horizon (cf. Koopmans, 1967). The corresponding strategy is called an infinite horizon strategy and is denoted by $\pi = (f_1, f_2, \ldots\ f_T, \ldots.)$. The classes of all T-horizon and infinite horizon strategies are denoted by Δ_T and Δ respectively. If an identical policy f is followed at each epoch, the strategies π_T and π are called stationary

and are denoted by f^T and f^∞ respectively. The class of all stationary T-horizon and infinite horizon strategies are denoted by ∇_T and ∇ respectively. Obviously $\nabla_T \subset \Delta_T$ and $\nabla \subset \Delta$.

It is assumed that there is an immediate stage-wise reward r_{ij}^k ; $0 \leq r_{ij}^k < \infty$ [≠] associated with every transition $(i,k,j) \in S \times A \times S$. If a decision function f is executed at any epoch a transition $(i, f(i), j)$ takes place with probability $p_{ij}^{f(i)}$ so that the expected immediate reward at state i is

$$r(i, f(i)) = r(f)(i) = \sum_{j=1}^{N} p_{ij}^{f(i)} \; r_{ij}^{f(i)}, (i=1,2,.....N). \qquad (1.2.3)$$

The N-dimensional column vector

$$r(f) = (r(1,f(1)), ..., r(N,f(N))' , \qquad (1.2.4)$$

is called the expected immediate reward vector associated with policy $f \in F$.

There is also a stagewise discount factor α_{ij}^k ; $0 \leq \alpha_{ij}^k < \infty$ associated with every transition $(i,k,j) \in S \times A \times S$, which is a sort of accumulator or multiplier (see Sec. 1.6) used to determine the present values of rewards earned after this transition. For example, suppose f and g are the

---

≠ More generally, r_{ij}^k may take finite negative values also. However, in the present dissertation, non-negativity of r_{ij}^k is assumed.

policy functions implemented on the system at the nth and the (n+1)th epochs respectively and we are interested in obtaining the present value vectors of r(f) and r(g) at the nth epoch. The present value vector of r(f) is r(f) itself, but the present value vector of r(g) is

$$\bar{P}(f)\, r(g)\,, \tag{1.2.5}$$

where

$$\bar{P}(f) = ((p_{ij}^{f(i)}\ \alpha_{ij}^{f(i)})) \tag{1.2.6}$$

so that the ith element of the present value vector is given by

$$\sum_{j=1}^{N} p_{ij}^{f(i)} \left\{ \alpha_{ij}^{f(i)}\, r(j,\, g(j)) \right\}.$$

This is because the present value vector of r(g) depends on the probabilities and discount factors involved in the transitions associated with the preceding decision function f : for a transition (i, f(i), j) at the nth epoch, the jth element r(j, g(j)) of r(g) will be discounted by $\alpha_{ij}^{f(i)}$. So the present value of r(j, g(j)) due to transition (i, f(i), j) is $\alpha_{ij}^{f(i)}\, r(j, g(j))$. Since this transition has a probability $p_{ij}^{f(i)}$ the present value vector of r(g) has the form given by (1.2.5). The N x N matrix $\bar{P}(f)$ given by (1.2.6) is called the Markov matrix or the transition matrix associated with $f \in F$ and its (i,j)th element, the transition measure (cf. Hinderer and Hubner, 1978) associated with the transition (i, f(i), j). Elements of $\bar{P}(f)$ are bounded

since $p_{ij}^{f(i)}$ and $\alpha_{ij}^{f(i)}$ are both bounded for every $(i, f(i), j) \in S \times A_X \times S$. So the total present value at nth epoch for executing f and g at nth and (n+1)th epochs respectively is

$$r(f) + \bar{P}(f)\, r(g) .$$

Now consider a given T-horizon ($T < \infty$) strategy $\pi_T = \{f_1, f_2, \ldots f_T\}$. The expected immediate reward vector of f_n ; $n = 1,2,\ldots T$ is $r(f_n)$ and its present value vector (expected discounted reward vector) (cf. Iwamoto, 1974; Schal, 1975; Serfazo, 1976) at 0th epoch is

$$\bar{P}(f_1)\ \bar{P}(f_2) \ldots \bar{P}(f_{n-1})\, r(f_n) = \bar{P}^{n-1}(\pi_T)\, r(f_n) ,$$

$$n = 1,2, \ldots T, \text{ say} \qquad (1.2.7)$$

where $$\bar{P}^{n}(\pi_T) = \prod_{t=1}^{n} \bar{P}(f_t) , \quad n \geq 1 \qquad (1.2.8)$$

is an n step transition matrix,

and $$\bar{P}^{0}(\pi_T) = I , \quad \text{for } n = 0 \qquad (1.2.9)$$

is an identity matrix. Hence the total present value vector or the total expected discounted return vector for executing the sequence $f_1, f_2, \ldots f_n$ of decision functions for $n = 1,2,\ldots T-1$, is

$$r^n(\pi_T) = r(f_1) + \bar{P}(f_1)\, r(f_2) + \ldots\ldots + \bar{P}(f_1) \ldots \bar{P}(f_{n-1})\, r(f_n)$$

$$= \sum_{t=0}^{n-1} \bar{P}^t(\pi_T)\, r(f_{t+1}) \qquad (1.2.10)$$

where $\bar{P}^t(\pi_T)$ and $\bar{P}^0(\pi_T)$ are of the form given in (1.2.8) and (1.2.9) .

If the planning horizon is of length T , it is assumed that with epoch T a terminal reward or serap value vector (Howard, Vol.II, 1970), $C = (c_1, c_2, \ldots\ldots c_N)' \geq 0$ is associated. Its present value at 0th epoch is $\bar{P}(f_1)\ \bar{P}(f_2) \ldots\ldots \bar{P}(f_T)\, C$, which may be called the salvage term (cf. Wagner, 1970). So the total expected discounted return vector associated with π_T is

$$r^T(\pi_T) = r(f_1) + \bar{P}(f_1)\, r(f_2) + \ldots\ldots + \bar{P}(f_1) \ldots \bar{P}(f_{n-1})\, r(f_n)$$

$$+ \ldots\ldots + \bar{P}(f_1) \ldots \bar{P}(f_{T-1})\, r(f_T)$$

$$+ \bar{P}(f_1) \ldots\ldots \bar{P}(f_T)\, C \qquad (1.2.11)$$

$$= \sum_{n=0}^{T-1} \bar{P}^n(\pi_T)\, r(f_{n+1}) + \bar{P}^T(\pi_T)\, C \qquad (1.2.12)$$

In particular if the starting state is i, $i \in S$ the total reward earned is given by the ith element $r^T(\pi_T)(i)$ of $r^T(\pi_T)$. Thus

$$r^{T}(\pi_T)\ (i) =$$

$$r(i, f_1(i)) + \sum_{j=1}^{N} p_{ij}^{f_1(i)}\ \alpha_{ij}^{f_1(i)}\ r(j,\ f_2(j))$$

$$+ \sum_{j,k=1}^{N} p_{ij}^{f_1(i)}\ p_{jk}^{f_2(j)}\ \alpha_{ij}^{f_1(i)}\ \alpha_{jk}^{f_2(j)}\ r(k,\ f_3(k))$$

$$+ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$+ \sum_{j,k,\ldots h,l=1}^{N} p_{ij}^{f_1(i)}\ p_{jk}^{f_2(j)}\ \ldots p_{hl}^{f_{n-1}(h)}\ \alpha_{ij}^{f_1(i)}\ \alpha_{jk}^{f_2(j)}\ \ldots\ \alpha_{hl}^{f_{n-1}(h)}\ r(l, f_n(l))$$

$$+ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$+ \sum_{j,k,\ldots h,l,\ldots mm'=1}^{N} p_{ij}^{f_1(i)} \ldots p_{mm'}^{f_T(m)}\ \alpha_{ij}^{f_1(i)} \ldots \alpha_{mm'}^{f_T(m)}\ c_{m'};$$

$$i = 1, 2, \ldots\ N. \qquad (1.2.13)$$

Thus $r^{n}(\pi_T)$ is the total expected discounted return associated with the part $(f_1, f_2, \ldots\ f_n)$, $n \leq T-1$ of $\pi_T = \{f_1, f_2, \ldots, f_T\}$ and $r^{T}(\pi_T)$ is the total expected discounted return associated with the whole strategy π_T.

When π_T is stationary, that is $\pi_T = f^T$, $f \in F$, (1.2.10) and (1.2.12) take the forms

$$r^{n}(f^T) = \sum_{t=0}^{n-1} \bar{P}^{t}(f^T)\ r(f)$$

and $$r^{T}(f^T) = \sum_{n=0}^{T-1} \bar{P}^{n}(f^T)\ r(f) + \bar{P}^{T}(f^T)\ c \qquad (1.2.14)$$

respectively, where $\bar{P}^n(f^T) = \left[\bar{P}(f)\right]^n$, $n = 0,1,2,\ldots\ldots T$.

When $\alpha_{ij}^k = \alpha$, a constant for every $(i,k,j) \in S \times A \times S$ (1.2.5) and (1.2.8) take the forms

$$\bar{P}(f) = ((\alpha\, p_{ij}^{f(i)})) = \alpha\, P(f) \qquad (1.2.15)$$

and $$\bar{P}^n(\pi_T) = \prod_{t=1}^{n} \bar{P}(f_t) = \prod_{t=1}^{n} \alpha^n\, P(f_t) = \alpha^n\, P^n(\pi_T) \text{ (say)} \qquad (1.2.16)$$

where $$P^n(\pi_T) = P(f_1) \ldots\ldots P(f_n),\ n \geq 1 , \qquad (1.2.17)$$

the (i,j)th element of $P^n(\pi_T)$ denotes the probability of the system to be in state j at the nth epoch given that it started in state i at 0th epoch. In this case (1.2.12), (1.2.13) and (1.2.14) respectively take the forms

$$\begin{aligned} r^T(\pi_T) &= r(f_1) + \alpha\, P(f_1)\, r(f_2) + \ldots\ldots\ldots\ldots \\ &\quad + \alpha^{n-1} P(f_1) \ldots\ldots P(f_{n-1})\, r(f_n) + \ldots\ldots\ldots \\ &\quad + \alpha^{T-1} P(f_1)\, P(f_2) \ldots P(f_{T-1})\, r(f_T) \\ &\quad + \alpha^T\, P(f_1) \ldots\ldots P(f_T)\, C \\ &= \sum_{n=0}^{T-1} \alpha^n\, P^n(\pi_T)\, r(f_{n+1}) + \alpha^T\, P^T(\pi_T)\, C , \end{aligned} \qquad (1.2.18)$$

$$r^T(\pi_T)(i) = r(i, f_1(i)) + \alpha \sum_{j=1}^{N} p_{ij}^{f_1(i)} r(j, f_2(j))$$

$$+ \alpha^2 \sum_{j,k=1}^{N} p_{ij}^{f_1(i)} p_{jk}^{f_2(j)} r(k, f_3(k))$$

$$+ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$+ \alpha^{n-1} \sum_{j,k\ldots h,l=1}^{N} p_{ij}^{f_1(i)} p_{jk}^{f_2(i)} \cdots p_{hl}^{f_{n-1}(h)} r(l, f_n(l))$$

$$+ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$+ \alpha^T \sum_{j,k..h,l..m,m'=1}^{N} p_{ij}^{f_1(i)} .. p_{hl}^{f_{n-1}(h)} .. p_{mm'}^{f_T(m)} c_{m'} ,$$

$$\text{for } i, 1, 2, \ldots\ldots N. \qquad (1.2.19)$$

and

$$r^T(f^T) = \sum_{n=0}^{T-1} \alpha^n P^n(f^T) + \alpha^T P^T(f^T) C \qquad (1.2.20)$$

where $P^n(f^T)$ is an n step transition probability matrix under the policy function f .

Now we consider the limiting case, when the horizon length T tends to ∞. In this case the strategy $\pi = \{f_1, f_2, \ldots\ldots f_n \ldots\} \in \Delta$ and the stationary strategy $\pi = \{f, f, \ldots\ldots f\} = f^\infty \in \forall \subset \Delta$. We shall use the notation $r^T(\pi)$ to denote the total expected discounted return of the first T-horizon part $\{f_1, f_2, \ldots\ldots f_T\}$ of π . Thus

$$r^T(\pi) = \sum_{n=0}^{T-1} \bar{P}^n(\pi)\, r(f_{n+1}) \qquad (1.2.21)$$

where $\bar{P}^{n}(\pi) = \prod_{t=1}^{n} \bar{P}(f_t)$. (1.2.22)

As will be seen later $\lim_{T\to\infty} r^{T}(\pi)$ may or may not exist. In case it exists and is finite for every $\pi \in \Delta$ one may say that the problem is convergent, otherwise the problem is said to be divergent. For convergent processes

$$\lim_{T\to\infty} r^{T}(\pi) = r(\pi) \qquad (1.2.23)$$

may be regarded as the total expected discounted return of an infinite horizon strategy $\pi = \{f_1, f_2, \ldots f_n, \ldots\}$.

Conditions for the existence and finiteness of (1.2.23) are discussed in section 1.4A. As will be seen there the most general sufficient condition has been provided by Veinott (1969) in terms of the dominant eigenvalues $\sigma(\bar{P}(f))$ of the matrices $\bar{P}(f)$, $f \in F$. He shows that when the rewards are finite, i.e.,

$$\max_{\substack{i \in S, \\ f(i) \in A(i)}} \left\{ r(i,f(i)), c_i \right\} \leqq M < \infty \ , \qquad (1.2.24)$$

and $$\sup_{f \in F} \left\{ \sigma(\bar{P}(f)) \right\} < 1 \ , \qquad (1.2.25)$$

the process converges. We shall use the notation

$$\chi^{*} = \sup_{f \in F} \left\{ \sigma(\bar{P}(f)) \right\} \qquad (1.2.26)$$

to denote the supremum of the dominant eigenvalues.

In particular for the Markov decision process with constant discount factor α, $\sigma(\bar{P}(f)) = \alpha$, for every $f \in F$ (since $P(f)$ is stochastic). Thus in this case the Markov decision process is convergent if $\alpha < 1$.

It will be seen in Chapter IV that the process does not converge if $\lambda^* \geq 1$ i.e., there exists at least one strategy π such that $\lim_{T \to \infty} r^T(\pi)$ tends to infinity. Then the process is called a divergent process.

It is thus seen that the Markov decision process under consideration* is completely characterised by the six-tuple

$$\{ S,\ A,\ p,\ r,\ \eta,\ C \} \tag{1.2.27}$$

where S is a finite state space;

$$A = \bigcup_{i \in S} A(i) \quad , \tag{1.2.28}$$

a finite action space;

$$p = \left\{ p_{ij}^{f(i)} \ ;\ p_{ij}^{f(i)} \geq 0,\ \sum_{j=1}^{N} p_{ij}^{f(i)} = 1,\ i, j \in S,\ f(i) \in A \right\} , \tag{1.2.29}$$

a set of transition probabilities;

---

* The Markov decision process considered above is in fact a stationary Marcov decision process in the sense that the rewards, discount factors and transition matrices remain invariant with time (epoch). We shall have occasions to consider a sort of nonstationary Markov decision processes in Chapters III and IV (the equivalent process and the asymptotically equivalent process) in which the rewards are time dependent. For such processes $\lim_{T \to \infty} \sup r^T(\pi)$ and $\lim_{T \to \infty} \inf r^T(\pi)$ exist but do not necessarily have identical values, so that $\lim r^T(\pi)$ does not necessarily exists. Such processes may also be regarded as convergent processes.

$$r = \left\{ r_{ij}^{f(i)} \; ; \; 0 \leq r_{ij}^{f(i)} < \infty, \; i,j \in S, \quad f(i) \in A \right\} \qquad (1.2.30)$$

a set of stagewise immediate rewards;

$$\eta = \left\{ \alpha_{ij}^{f(i)} \; ; \; 0 \leq \alpha_{ij}^{f(i)} < \infty, \; i,j \in S, \quad f(i) \in A \right\} \qquad (1.2.31)$$

a set of stagewise discount factors ; and

$$C = (c_1, c_2, \ldots . c_N)', \qquad (1.2.32)$$

a terminal reward vector.

In particular when $\alpha_{ij}^{k} = \alpha$ for all $(i,k,j) \in S \times A \times S$ a constant discount factor of the process, the Markov decision process is characterized by the six-tuple

$$\left\{ S, \; A, \; p, \; r, \; \alpha, \; C \right\} . \qquad (1.2.33)$$

Somewhat more general Markov decision processes can however be defined in which a policy specifies the probabilities with which different actions are taken at each state and each action leads to transitions to different states with given sets of probabilities. Such policies may be called randomized policies and the corresponding Markov decision process is called a Markov decision process with randomized plans [see Derman, 1970; Iwamoto, 1975 and Silver and Moore, 1976].

1.3. The Optimality Criterion

It is well known (e.g., Eilon, 1969) that a decision problem is concerned with choosing the 'best' alternative out of several available ones : -- 'best' with respect to a specific optimality criterion. The most commonly used criterion for choosing an optimal strategy is that of maximal total expected discounted return #.

According to this criterion, when T is finite, the objective is to maximize the total expected discounted return vector (component wise), given in (1.2.12) . The optimal T-horizon strategy π_T^* is the one which satisfies the relation

$$r^T(\pi_T)(i) \lesssim r^T(\pi_T^*)(i) \; ; \text{ for every } i \in S, \; \pi_T \in \Delta_T \tag{1.3.1}$$

so that

$$r^T(\pi_T) \lesssim r^T(\pi_T^*) \; ; \text{ for any } \pi_T \in \Delta_T \tag{1.3.2}$$

A backward recurrence relation## of dynamic programming based on Bellman's principle of optimality (cf. Derman, 1970; Bellman, 1957; Mine and Osaki, 1970) can be made use of to find such a strategy. The recurrence relations are given by

---

See Section 1.4C, for other criteria proposed.

An analogous recursion formula may be used for non-stationary Markov decision processes (cf. Porteus, 1975b; Howard, 1970 and Chapters III and IV).

$$v_i^n = \max_{f(i) \in A(i)} \left\{ \sum_{j=1}^{N} p_{ij}^{f(i)} \left[r_{ij}^{f(i)} + \alpha_{ij}^{f(i)} v_j^{n-1} \right] \right\} \tag{1.3.3}$$

$$i = 1,2,\ldots\ldots N ;$$

$$n = 0,1,2,\ldots\ldots T$$

$$= \max_{f(i) \in A(i)} \left\{ r(i,f(i)) + \sum_{j=1}^{N} \alpha_{ij}^{f(i)} p_{ij}^{f(i)} v_j^{n-1} \right\} \tag{1.3.4}$$

$$i = 1,2,\ldots\ldots N ;$$

$$n = 0,1,2,\ldots\ldots T$$

and $v_i^0 = c_i$; $i = 1,2,\ldots\ldots N$. (1.3.5)

The above system can also be written in vectorial form as

$$v^n = \max_{f \in F} \left\{ r(f) + \bar{P}(f) v^{n-1} \right\} , \tag{1.3.6}$$

$$n = 0,1,2, \ldots\ldots T$$

and $v^0 = C$ (1.3.7)

Again when $\alpha_{ij}^k = \alpha$ for $(i,k,j) \in S \times A \times S$ so that $\bar{P}(f) = \alpha P(f)$ for any $f \in F$ then the iteration formulae (1.3.3) and (1.3.6) reduce to

$$v_i^n = \max_{f(i) \in A(i)} \left\{ r(i,f(i)) + \alpha \sum_{j=1}^{N} p_{ij}^{f(i)} v_j^{n-1} \right\} ; \tag{1.3.8}$$

$$i = 1,2,\ldots\ldots N ,$$

$$n = 0,1,2,\ldots\ldots T$$

and

$$v^n = \max_{f \in F} \left\{ r(f) + \alpha P(f) v^{n-1} \right\} \qquad n = 1,2,\ldots\ldots T \qquad (1.3.9)$$

respectively.[ø] By using the above recursion formulae one can determine an optimal policy for every stage n in a sequential fashion, and hence an optimal T-horizon strategy π_T^* and the corresponding maximal total expected discounted return vector v^T. Obviously

$$v^T = r^T(\pi_T^*) \quad \text{for any} \quad T < \infty \,, \qquad (1.3.10)$$

and if v^T corresponds to g_T^*, v^{T-1} to g_{T-1}^* and so on lastly v^1 to g_1^* then

$$\pi_T^* = \left\{ g_T^*, g_{T-1}^*, \ldots\ldots g_1^* \right\} = \left\{ f_1^*, f_2^* \ldots\ldots f_T^* \right\} \qquad (1.3.11)$$

where $f_i^* = g_{T-i+1}^*$ is the optimal T-horizon strategy which is to be executed in practice. To avoid confusion we shall always use the convention that the optimal policy corresponding to v^n will be denoted by g_n^* and the optimal policy to be followed at the nth epoch will be denoted by f_n^*. It may be noted that an optimal T-horizon strategy obtained by the

---

ø It may be noted that v^n defined in (1.3.6) satisfy $v^n \leq v^{n+1}$ for $n = 0,1,2,\ldots\ldots$, (see e.g. Bellman, 1957; Mine and Osaki, 1970) and v^n defined by (1.3.9) is continuous in α (Hordijk and Tijms, 1974).

above formulae in general depends on the terminal reward vector. This procedure of finding the optimal T-horizon strategy is also called the 'method of successive approximations' (MSA) (cf. Derman, 1970).

As long as T is small the procedure remains computationally feasible, particularly when the state space S is not very large. But for large T , the procedure becomes quite cumbersome even when S is small.

When T is large, it is reasonable to consider an **infinite horizon process**[*] (the process which continues indefinitely) as an approximation to the T-horizon process (cf. Bellman, 1957; Hinderer and Hubner, 1978; Morton and Wecker, 1977), provided an optimal solution to the infinite horizon process exists, that is, the process is convergent in the sense of section 1.2 ($\alpha^* < 1$) . In this case, an optimal infinite horizon strategy π^* satisfies the relation

$$r(\pi^*)(i) \geq r(\pi)(i) \; ; \quad \text{for all } i \in S \text{ and } \pi \in \Delta \tag{1.3.12}$$

where $r(\pi)(i)$ is the ith element of

$$r(\pi) = \lim_{T \to \infty} r^T(\pi) \; .$$

---

[*] Of course an actual infinite horizon process does not exist in the real world. It is taken into account only to investigate the structure of optimal strategies for sufficiently long horizons and for the development of an analytical theory concerning the problem. Thus Bellman (1957) says that 'Although an unbounded process is always a physical fiction, as a mathematical process it has many attractive features'.

It is well known (cf. Veinott, 1969; Mine and Osaki, 1970) that there is an optimal strategy that is stationary and is independent of the terminal reward vector C.

When $\alpha_{ij}^{k} = \alpha = 1$ for $(i,k,j) \in S \times A \times S$, $\lambda^{*} = 1$, so that the Markov decision process becomes divergent. In this case $r^{T}(\pi_T)$ given by (1.2.18) takes the form

$$r^{T}(\pi_T) = \sum_{n=0}^{T-1} P^{n}(\pi_T)\, r(f_{n+1}) + P^{T}(\pi_T)\, C \qquad (1.3.13)$$

where $P^{n}(\pi_T)$ is given by (1.2.17) and (1.2.9) for $n \geq 1$ and $n = 0$ respectively. This case has been studied extensively through the use of the expected average return of (ERAR) criterion (cf. Derman, 1970). The expected average return of π_T is given by

$$R^{T}(\pi_T) = r^{T}(\pi_T)/T \qquad (1.3.14)$$

It can be shown that when $r^{T}(\pi_T)$ is of the form (1.3.13), $R^{T}(\pi_T)$ remains bounded as $T \longrightarrow \infty$ for $\pi_T \in \Delta_T$ so that

$$\liminf_{T \longrightarrow \infty} R^{T}(\pi) \qquad (1.3.15)$$

and $\limsup_{T \longrightarrow \infty} R^{T}(\pi)$

exist for $\pi \in \Delta$, but do not necessarily have the same value. In particular if π is stationary, they have identical values and

$$R(\pi) = \lim_{T \to \infty} R^{T}(\pi) \qquad (1.3.16)$$

exists (cf. Derman, 1970).

Now consider $\pi = f^{\infty}$, where $P(f)$ is irreducible, (1.3.16) exists and is equal to

$$R(\pi) = \lim_{T\to\infty} \frac{r^T(f^T)}{T} = \lim_{T\to\infty}\left\{\frac{\sum_{n=0}^{T-1} P^n(f^T)\, r(f) + P^T(f^T)\, C}{T}\right\},$$

where $P^n(f^T) = (P(f))^n$

for $n = 0, 1, \ldots\ldots T$

$$= P^* r(f)$$

$$= a\,\underline{1}\ , \text{ (say)}$$

where $a > 0$ is a constant and $\underline{1} = (1, 1, \ldots\ldots,)'$, an N-dimensional vector of ones, since P^* is a stationary stochastic matrix. When $P(f)$ is not irreducible and contains ℓ ergodic classes and possibly some transient states then the elements of $R(\pi)$ will be of the form

$$\sum_{i=1}^{\ell} a_i\, \underline{1}_i^0$$

where $a_i > 0$ and $\underline{1}_i^0$ is an N-dimensional column vector with elements unity corresponding to each state belonging to ergodic class i and zero elements elsewhere.

Let $\pi^* = f^{*\infty}$ denote an optimal stationary strategy and $r_\alpha(\pi^*)$ the corresponding total expected discounted return function when the discount factor is α, $\alpha \in (0,1)$ then

$$\lim_{\alpha \to 1^-} (1-\alpha)\, r_\alpha(\pi^*) = \lim_{T \to \infty} \frac{1}{T}\, r^T(\pi^*) = R(\pi^*) \qquad (1.3.17)$$

(see Derman, 1970; Mine and Osaki, 1970; and Blackwell, 1962). The left hand side of (1.3.17) is concerned with the behaviour of α near enough to but less than one. Thus by a limiting process it can be established that for the undiscounted case there exists a stationary strategy which is optimal on the expected average return criterion.

It is thus seen that a stationary optimal infinite horizon strategy exists when $\lambda^* < 1$ on the total expected discounted return criterion. But when $\alpha_{ij}^k = \alpha = 1$, $r^T(\pi) \longrightarrow \infty$ as $T \longrightarrow \infty$ so that no well defined optimal infinite horizon strategy on total expected discounted return criterion exists; however on the expected average return criterion there is an optimal infinite horizon strategy which is stationary. In other situations no optimal infinite horizon strategy exists on either of the above criteria.

When an optimal infinite horizon strategy exists on the total expected discounted return criterion and T is sufficiently large one can make use of Shapiro (1968b)'s turnpike theory to obtain an optimal strategy for the T-horizon problem. Shapiro shows that there exists a T^* such that when $T \geq T^*$, an optimal policy for each of the first $T-T^*$ epochs is provided by the corresponding policy of the optimal infinite horizon strategy. So in practice the optimal infinite horizon strategy can be implemented during the first $T-T^*$ epochs and

for the remaining T^* epochs the policy functions given by the iteration formula (1.3.3) can be used. $T-T^*$ is called the optimal turnpike. Unfortunately for the undiscounted case ($\alpha = 1$) turnpike results are not as precise as in the convergent case ($\lambda^* < 1$). Reference may be made to the work of Romanovskii (1970), Federgruen and Schweitzer (1977) and Hinderer and Hubner (1978) in this connection.

1.4. A Review of Previous Work

In the last three decades a considerable amount of work has been reported on Markov decision processes and allied problems. Among the initiators of the subject mention may be made of Bellman and LaSalle (1949), Bellman and Blackwell (1949) and Shapley (1953). While these investigators view the Markov decision process in the context of two person dynamic games, Bellman (1957a) has given an explicit formulation of the problem outside the game theoretic context; he has also given the outlines of the computational technique of successive approximations(see Section 1.3) for the evaluation of an optimal T-horizon strategy in a general Markov decision process. Since then there has been a good amount of theoretical and methodological development. For a bibliography of the work concerning Markov decision processes upto 1973 see Schweitzer (1973). Teugels's (1976) more recent bibliography on semi-Markov processes includes the work on more general decision models. Some more recent work has been included in the bibliography section of the present thesis.

In what follows we present the broad details of previous work related to the present thesis.

1.4A. Convergent Markov Decision Processes

Convergent Markov decision processes with constant discount factor α, $\alpha \in [0,1)$ have been studied extensively by a good number of authors (e.g. Howerd, 1960; Blackwell, 1962, 1965; Derman, 1970; Mine and Osaki, 1970; Van Nunen, 1976; Veinott, 1966, 1969, 1973).

Howard (1960) has developed a computational technique called the policy improvement algorithm (PIA) under the assumption that there exists an optimal infinite horizon strategy that is stationary for Markov decision process with $\alpha \in [0,1)$ on the basis of total expected discounted return criterion. Later on Blackwell (1962) has established the existence of an optimal stationary infinite horizon strategy. He has also established that this optimal infinite horizon stationary strategy is independent of the terminal reward vector. So the search for an optimal infinite horizon strategy can be confined to the class of stationary infinite horizon strategies only. An algorithm based on the linear programming approach for finding the optimal infinite horizon stationary strategy for the Markov decision process with $\alpha \in [0,1)$ has been given by D'Epenoux (1963). Later De Gellinck and Eppen (1967) and Derman (1970) have developed the theory based on this approach for the randomized Markov

decision process and have found that the linear programming formulation also finds a strategy which is pure (independent of action probabilities, Derman, 1970), stationary and independent of the state where the process is started initially. The relation between linear programming and policy improvement algorithms can be found in Mine and Osaki (1970).

An aspect of problem of Markov decision process is its so called sensitivity analysis. This is concerned with examining how far the optimal infinite horizon stationary strategy remains invariant as α increases from 0 to 1. Studies in these lines have been made by Howard (1960) and Smallwood (1966).

For variable α_{ij}^{k} case ($\alpha_{ij}^{k} \in [0,1)$ for $i,j \in S$ and $k \in A$), Iwamoto (1974, 1975) has found analogous results as in $\alpha \in [0,1)$ case under a slightly more general set up.

The problem of convergence of Markov decision processes has been studied from the point of view of contraction mappings by several authors. An operator H is a contraction mapping on E^N with contraction coefficient c ($c < 1$) if for two vectors $u, v \in E^N$ and for any $f \in F$,

$$\| Hu - Hv \| \leq c \| u - v \| \qquad (1.4.1)$$

where $\| \cdot \|$ stands for the norm†.

---

† Here the norm of a vector $u = (u_1, u_2, \ldots u_N)'$ is taken to be equal to $\max_i |u_i|$ (Wilkinson, 1965) and the norm of the matrix $A = ((a_{ij}))$ is equal to $\max_i \sum_{j=1}^{N} |a_{ij}|$ that is, the maximal sum of the absolute values of the elements of a row (Wilkinson, 1966).

Thus H is a contraction operator if $\| H \| < 1$ (see e.g., Denardo, 1967; Iwamoto, 1974 and Porteus, 1975b). Denardo (1967) has used this approach in a general dynamic programming problem from which it follows that the operator $L(f)$ defined by

$$L(f)u = r(f) + \alpha P(f)u, \quad \text{for } f \in F \text{ and } u \in E^N \tag{1.4.2}$$

is a contraction operator with contraction coefficient α if the constant discount facor $\alpha < 1$. For the variable discount factors case Iwamoto (1974) has shown that if $\max_{(i,k,j) \in SxAxS} \{\alpha_{ij}^k\} = c' < 1$, then the operator $L(f)$ defined by

$$L(f)u = r(f) + \bar{P}(f)u \tag{1.4.3}$$

is a contraction operator with contraction coefficient at most equal to c'. If the contraction coefficient is less than one for every $L(f)$, $f \in F$ then the total expected discounted return of any infinite horizon strategy $\pi \in \Delta$ (either in the constant discount factor case or in the variable discount factor case) converges (cf. Iwamoto, 1974; Eaves, 1977; Blackwell, 1962). Thus the corresponding Markov decision processes are convergent.

If $u^* \in E^N$ exists such that for an operator A

$$Au^* = u^* ,$$

then u^* is called a fixed point of A. Sharpening and

generalizing Shapley (1953)'s and Denardo (1967)'s results Veinott has shown that if $\sigma(\bar{P}(f)) < 1$ for $f \in F$ (i.e. λ^* defined in (1.2.26) is less than unity) the maximization operator $\mathcal{R}$ defined by

$$\mathcal{R}u \equiv \max_{f \in F} \left\{ r(f) + \bar{P}(f)u \right\}, \quad u \in E^N \tag{1.4.4}$$

is an n-stage contraction operator (that is $\mathcal{R}^n$ is a contraction operator for some positive interger n) and possesses a unique fixed point.

It therefore follows that the given Markov decision process converges[1] if $\lambda^* < 1$, in the sense that every strategy $\pi \in \Delta$ has a finite total expected discounted return. An equivalent condition for the convergence of the Markov decision process is that

$$\sum_{n=0}^{\infty} \bar{P}^n(f^\infty) < \infty \tag{1.4.5}$$

for every $f \in F$ (see, Veinott, 1969, 1973). Veinott proves that when condition (1.4.5) holds then

$$\sum_{n=0}^{\infty} \bar{P}^n(\pi) < \infty \tag{1.4.6}$$

for every $\pi \in \Delta$, whene $\bar{P}^n(\pi)$ is given by (1.2.22).

---

[1] Convergent processes are called transient processes by Veinott (1969, 1973), terminating processes (cf. Shapley, 1953; Aggarwal et.al., 1977; Mine and Osaki, 1970) and discounted processes (Porteus, 1975a) have analogous properties.

So for the convergence of the Markov decision process, n-stage contraction ($n \geq 1$) is sufficient. This happens for example when (1) the transition probability matrix of every $f \in F$ is substochastic, so that $\sigma(\bar{P}(f)) < 1$, $f \in F$; or (2) the constant discount factor α, $\alpha \in (0, \infty)$ is such that $\sigma(\alpha P(f)) < 1$, $f \in F$ or (3) variable discount factors α_{ij}^{k}, $\alpha_{ij}^{k} \in (0, \infty)$ are such that $\sigma(\bar{P}(f)) < 1$ for $f \in F$.

Veinott has shown that for any infinite horizon strategy π

$$\min_{f \in F} \left\{ \sigma(\bar{P}(f)) \right\} \leq \left\{ \sigma(\bar{P}^{n}(\pi)) \right\}^{1/n} \leq \max_{f \in F} \left\{ \sigma(\bar{P}(f)) \right\} \tag{1.4.7}$$

for any $n \geq 1$, so that when $\pi = f^{\infty}$

$$\sigma(\bar{P}^{n}(\pi)) = \left\{ \sigma(\bar{P}(f)) \right\}^{n} \tag{1.4.8}$$

It may be noted that when $\alpha_{ij}^{k} = \alpha < 1$, the total expected discounted return of any infinite horizon strategy converges geometrically at the rate of α (cf. Morton and Wecker, 1977).

For some recent studies dealing with contraction mappings in Markov decision processes reference may be made to Van Nunen (1976) and Eames (1977). Using the approach of monotone mappings in dynamic programming which is some what analogous to the contraction mappings approach to Markov decision processes, Bertsekas (1977) obtained several results relating mostly to the convergence of the dynamic programming algorithm and the existence of optimal stationary strategies,

then he applied the theory to several problems regarding optimal control theory. See also Serfozo (1976) and Sladky (1974) in this connection.

1.4B. Divergent Markov Decision Processes.

Thus divergence in a Markov decision process due to discounting occurs when $\lambda^* \geq 1$. We briefly review the literature concerning the cases $\lambda^* = 1$ and $\lambda^* > 1$ separately. Among divergent problems only Markov decision processes with $\alpha_{ij}^k = \alpha = 1$ for $i,j = 1,2, \ldots\ldots N$ and $k \in A$ (also called the undiscounted Markov decision processes) have been studied extensively. In the game theoretic (cf. Hoffman and Karp, 1966; Gillette, 1957) context, they are called nonterminating stochastic games. The Markov decision process with $\alpha = 1$ has been studied both as a limiting case ($\alpha \longrightarrow 1^-$) (see Blackwell, 1962; Mine and Osaki, 1970; Derman, 1970) and independently on the basis of the expected average reward per transition criterion (see Howard, 1960; Blackwell, 1962; Mine and Osaki, 1970; Derman, 1970). Both linear programming and policy improvement algorithms (Manne, 1960; Howard, 1960; Veinott, 1966; Brown, 1965; Mine and Osaki, 1970; Derman, 1970; Hordijk and Kallenberg, 1979) are available for finding the optimal infinite horizon strategy, which is stationary and independent of α and of the state where the process has started. These algorithms are available for more general models also (see, e.g., Jewell, 1963).

Among the few investigators on other types of divergent problems mention may be made of the work of Morton (1971, 1976) and Morton and Wecker (1977). They have given some convergence properties of the Markov decision process when the constant discount factor $\alpha > 1$. They have shown that in this case if there is a stationary optimal infinite horizon strategy f^{∞} then the relative value function (total value function less total value at a fixed or reference state N, White, 1963, 1978) converges if $\alpha \lambda'' < 1$, where λ'' is the modulus of the subdominant eigenvalue of $P(f)$. They have also shown that the asymptotic convergence of relative value function implies asymptotic policy convergence under certain broad 'regularity' conditions. The theory thus slightly relaxes the restriction that α should be less than one, but of course even the relative values diverge when α is so large that $\alpha \lambda'' \geq 1$.

Hinderer (1976, 1977) and Hinderer and Hübner (1978) have studied the case of $\alpha > 1$ by the 'method of extrapolation', wherein he has given a procedure for finding the optimal strategies of sufficiently large finite horizons by using the optimal strategies of small horizons.

Bellman (1957) has given the formulation of the multiplicative Markov decision process which is a special case of (1.2.27), when $r(f_n) = 0$ (or $\overline{P}^{n-1}(\pi)\, r(f_n) = 0$) for $n = 1,2,\ldots\ldots$, $\overline{P}^{n}(\pi)\, C \geq 0$ for every $\pi \in \Delta$ and $\lambda^{*} \in (0,\infty)$.

He showed that the total expected discounted return of an n-horizon stationary strategy which is optimal in the class of stationary strategies (see Rothblum, 1974) 'grows like the nth power' of λ^*, for large values of n under the assumptions that the policy function f having $\sigma(\bar{P}(f)) = \lambda^*$ is unique and $\bar{P}(f) > 0$ for every $f \in F$.

Rothblum (1974) has also considered multiplicative Markov decision chains. His model is somewhat more general than Bellman's in the sense that $\bar{P}(f)$ for any $f \in F$ can be reducible and λ^* may be equal to zero (the 'nilpotent' case). His attention has not been restricted to stationary strategies as in Bellman (1957) and he has used a new criterion called the 'cummulative optimality criterion' which is based on Cesaro means of higher orders (cf. Hardy, (1949). He has observed that the optimal strategies are usually nonstationary and has given a constructive proof of the existence of such optimal strategies. He has also given various algorithms which find stationary strategies which are infinite cummulative optimal. Other works on multiplicative Markov decision processes include the work of Mandl (1967), Mandl and Seneta (1969) and Howard and Matheson (1972).

Porteus (1975b), Schal (1975) and Serfozo (1976) have posed a general nonstationary decision process in which discount factors are dependent on time, states and actions. and belongs to the $[0, \infty)$ region. Under the framework of multistage decision processes Porteus (1975b) has given

sufficient conditions for the general processes to converge. He also has given the assumptions required for an ϵ-optimal ($\epsilon > 0$) (Blackwell, 1965) stationary strategy and optimal stationary strategy to exist. Schal (1975) has given sufficient conditions for the existence of an optimal strategy and has interrelated the optimal total expected rewards as well as the optimal actions of the model with infinite horizon and those of the model with finite horizon T, as T tends to infinity. His analysis is based on some results on set valued mappings, upper semi continuous functions measurable selections and topologies on spaces of probability measures. Serfozo (1976) has extended Schal's work to monotone optimal policies for Markov decision processes and illustrated them with controlled random walks and machine maintenance examples.

Wang (1974) has considered a nonstationary non-Markovian general model with Borel state and action spaces and discount factors $\alpha_t(H_t) \in (0, \infty)$, where H_t is the history of the system upto t+1 epochs starting from 0th epoch. His study is based on the combination of additive and multiplicative reward systems. His model is considered to be a generalization of that of Howard and Matheson (1972) and Jaquette (1976) who have analysed the Markov decision process in the presence of the exponential utility criterion. Wang has also presented sufficient conditions for the existence of nonrandomized optimal strategies.

Dirickx (1971) analysed the deterministic dynamic programming with $\alpha > 1$ under a new criterion called Modified equivalent average return which avoids the difficulty arising from divergence of the present value of an infinite horizon strategy. His optimization procedure is based on linear programming and works successfully in the deterministic case but since in the stochastic case the optimal infinite horizon strategy is nonstationary, no finite algorithm for its determination appears to be available.

1.4C. Some Other Optimality Criteria.

The most commonly used criteria for choosing an optimal strategy in a Markov decision process are the total expected discounted return (total present value) criterion for convergent processes and the expected average return criterion for undiscounted processes. Some other alternative criteria have also been proposed by several authors. They are the relative value criterion (White, 1963, 1978; Morton and Wecker, 1977); ϵ-optimality, p-optimality and (p, ϵ)-optimality (Blackwell, 1965; Strauch,1966) criteria; modified equivelent average return criterion (Dirickx 1971); equivalent average return criterion (cf. Wagner, 1969; Lippman, 1969; Dirickx, 1971), sensitive discount optimality criterion (Veinott, 1969; Rothblum, 1974, 1977), moment optimality criterion (cf. Jaquette, 1974, 1977; Eagle, 1975; Denardo, 1971; Goldwerger, 1977), cummulative optimality criterion (Rothblum, 1974) and

average overtaking and overtaking optimality criteria (Veinott, 1966, 1977; Rothblum and Veinott, 1976).

Kreps (1975) and Porteus (1975b) have explicitely defined some different types of utility functions like additive, multiplicative, exponential, separable and separated utilities and have used an expected utility criterion with respect to the particular type of utility function used (see also Howard, 1968 and Howard and Matheson, 1972 in this connection).

Some of the above criteria are variants of the total expected discounted return when the latter cannot be applied directly. Reference may be made to the bibliography of the thesis for some other criteria used for more general processes (including semi-Markovian and non-Markovian processes).

Variants of the Markov decision process considered in section 1.2 and their applications (both practical and theoretical), have received quite a good deal of attention. Since these are not directly related to the work incorporated in the thesis, no review of such investigations is presented here, though some of them are of very great interest, and are included in the bibliography.

1.5. Motivation of the Present Study.

In a decision process discount factors come in the picture when rewards (or more generally returns or utilities) from future decisions are to be related to the present epoch. The exact way in which they affect the total reward earned from future decisions is explained in section 1.2. Several interpretations are available about their physical meaning or significance.

For the constant discount factor case the first interpretation is economic. It is well known that the money value of an investment does not remain invariant because it may earn interest or may suffer or gain in real terms due to inflation, deflation, devaluation, revaluation and so on. Likewise the value of the reward at a future instant may not be the same as its present worth. So it is customary to use a discount factor to reflect possible changes in the value of a reward earned at a future epoch. For example if net interest (interest rate - inflation rate) is to be taken into account $\alpha = 1/(1+\rho)$ where ρ, $-1 < \rho < \infty$ represents the net interest rate, may be used (Wagner, 1970; Keeny and Raiffa, 1974). It may be noted that negative values of ρ indicate that the rate of inflation is higher than the rate of interest, (a very common situation now a days!), in which case α takes a value greater than unity.

Again α may be regarded as an indicator of the phenomenon of 'time preference of a decision maker' which can be

described as the 'greed impatience trade off' (Howard, 1968). $\alpha > 1$ corresponds to the case where the decision maker prefers to delay decisions with higher rewards (cf. Dirickx, 1971; Koopmans, 1966), 'in order to have more later or less now'.

The case $\alpha \leq 1$ may be interpreted as an indicator of the impatience of the decision maker towards taking such decisions (cf. Koopmans, 1967; Fishburn, 1970; Dirickx, 1971). The case $\alpha = 1$ indicates that the feeling of the decision maker is neutral about all decisions. For further details regarding this approach to the discount factor see Koopmans (1967), Howard (1968), Keeny and Raiffa (1974) and Fishburn (1970).

In some problems relating to the deterministic networks, a discount factor refers to a multiplier or gain rate of a particular arc (cf. Jewell, 1962; Dirickx, 1974; Frank and Frisch, 1971; Jensen and Bhowmik, 1977). This sort of gain rate is usually present in large electrical networks with transformers and in transportation networks (cf. Frank and Frisch, 1971). When the flow passes through a node it is either amplified or diminished depending on whether its value is less than or greater than one.

Dirickx (1971, 1973) remarks that the case of $\alpha > 1$ has been considered as a possibility by Koopmans (1967) in the context of economic growth theory. He also cites an example about the optimization of an intertemporal consumption over

an unbounded horizon given by Manne (1970), wherein a particular term plays the role of a discount factor. Though Manne has assumed that it lies between 0 and 1 for the sake of convenience, Dirickx (1971) points out that this term can actually be greater than one.

Finally $\alpha < 1$ is interpreted as the probability that the process will continue to earn rewards after the next transition (Howard, 1960). So $1 - \alpha$ is the probability that the process will stop at its present stage thus the process with discounting ($\alpha < 1$) may be interpreted as one of indefinite duration.

It is thus seen that the $\alpha > 1$ case is physically meaningful in many practical situations though comparitively few investigations are concerned with it. This may be because of the mathematical complexities produced by divergence. To see how far these mathematical difficulties can be overcome is one of the chief motivations of the present thesis.

That the discount factors can vary with time and with the transitions involved has been realised by several authors (cf. Iwamoto, 1974; Schal, 1975; Porteus, 1975b; Bellman, 1957; Serfazo, 1976; and Sladky, 1976). So in general one should take into account stagewise discount factors of the form $\alpha_{ij}^{k}(t)$, $-\infty < \alpha_{ij}^{k}(t) < \infty$; $i, j \in S$, $k \in A$, $t = 0,1,2,\ldots,$ for the transition (i,k,j) at epoch t.

The present thesis considers that the discount factors are non-negative and may vary with transitions, but remains invariant with time. Perhaps time-invariance is not a very realistic assumption to make. But since in the basic model of the thesis factors like state space, action space and transition probabilities are all regarded as stationary, this limitation with regard to discount factors had to be maintained. Even then divergence poses certain mathematically interesting problems and provides a motivation for a detailed investigation.

The same reason has led to an attempt to use the total expected discounted return criterion throughout the thesis. Though the total expected discounted returns of some infinite horizon strategies diverge our attempt has been to answer the following question. Does there exist a strategy π_T^* which has a larger total expected discounted return than any other strategy π_T, for all T sufficiently large ?.

1.6. Investigations Reported in the Present Study.

The present thesis is devoted to a study of a class of divergent Markov decision processes, in which the total expected discounted return vector of an infinite horizon strategy diverges because the transition matrices $\bar{P}(f)$ corresponding to some of the policies f, $f \in F$ have spectral radii $\sigma(\bar{P}(f)) \geq 1$. Such a situation arises when some of the

discount factors are too large. Our main interest is to investigate the structure of an optimal T-horizon strategy of such processes for sufficiently large values of T . The criterion of optimality used is the total expected discounted return.

Chapter II presents the mathematical preliminaries required in the later chapters (Chapters III and IV) of the thesis. In the first few sections (Sections 2.2 - 2.7) of this chapter some of the known concepts and results concerning non-negative matrices are recapitulated. In the later sections (Sections 2.8 - 2.12) a number of new results are presented. First a theorem on a necessary and sufficient condition for strong ergodicity of a sequence of stochastic matrices is established, based on the concept of fixed points. Second it is established that if two non-negative irreducible matrices A and B of a set have the same spectral radius λ^* then they will have a common extremal eigenvector z , provided, the matrices (assumed irreducible) generated by interchanging any row of A by the corresponding row of B have spectral radii not greater than λ^*. Finally some fixed point algorithms regarding the convergence of products of a sequence of stochastic matrices and a sequence of irreducible non-negative matrices each having spectral radius unity are presented.

Chapter III is concerned with the study of a divergent Markov decision process in which the constant discount factor

α is greater than one. The order of the total expected discounted return function of a T-horizon strategy is shown to be equal to α^T. Two processes related to the original process (the divergent Markov decision process, $\alpha > 1$), called the equivalent process and the asymptotically equivalent process, are proposed and the stage by stage equivalence of the original process and the equivalent process and the asymptotic equivalence of the equivalent process and the asymptotically equivalent process are established. The structure of an optimal infinite horizon strategy of the asymptotically equivalent process is then investigated. This enables one to determine the structure of an optimal infinite horizon strategy of the equivalent process. From this the structure of an optimal T-horizon strategy of the original process and corresponding total expected discounted return functions can be found, for sufficiently large values of T. Finally an iterative procedure to obtain an optimal T-horizon strategy of the original process is suggested. The optimal strategy found is usually non-stationary.

Chapter IV is concerned with the study of divergent Markov decision processes with variable (state, action dependent) discount factors. Throughout this chapter it is assumed that the transition matrix corresponding to each policy function is irreducible. It is first shown that when the expected immediate rewards are bounded and λ^*, the supremum of the spectral radii is greater than or equal to one then the Markov

decision process diverges. For the case of $\chi^* = 1$ it is established that on the expected average return criterion a stationary optimal infinite horizon strategy exists and is independent of the terminal reward vector.

For the case $\chi^* > 1$ it is established that the order of the total expected discounted return function of any T-horizon strategy, when T is large, does not exceed χ^{*T} so that the total expected discounted return of an optimal T-horizon strategy is of order χ^{*T}. In order to study the structure of such an optimal strategy two processes called the equivalent process and the asymptotically equivalent process are proposed (just as in the previous chapter) and the stage by stage equivalence of the original process and the equivalent process and the asymptotic equivalence of the equivalent process and the asymptotically equivalent process are established. The structure of an optimal infinite horizon strategy of the asymptotically equivalent process is then determined. This enables one to determine the structure of an optimal infinite horizon strategy of the equivalent process. From this the structure of an optimal T-horizon strategy of the original process and the corresponding total expected discounted return functions can be found. Finally two iterative procedures for finding an optimal T-horizon strategy of the original process are suggested. As in the constant discount factor $\alpha > 1$ case an optimal strategy is non-stationary, in general.