MATHEMATICAL CENTRE TRACTS 97

MARKOVIAN CONTROL PROBLEMS

FUNCTIONAL EQUATIONS AND ALGORITHMS

A. FEDERGRUEN

MATHEMATISCH CENTRUM AMSTERDAM 1983

1980 Mathematics subject classification: 90C40, 90D15, 93E05, 90C45, 90C47

ISBN 90 6196 165 3

PREFACE

This volume is a reprint of the doctoral dissertation of A. Federgruen. In 1978, at the time of its original publication, Dr. Federgruen was a member of the Department of Operations Research of the Mathematical Centre, He is currently at the Graduate School of Business, Columbia University, New York.

Thanks are due to all those at the Centre for Mathematics and Computer Science who have contributed to this publication.

CONTENTS

INTRODUCTION

Most of the charms and challenges of life seem to find their origin in our desires to steer the partially controllable and foreseeable evolution of the systems we feel concerned with.

These systems may be thought of as being in a state which changes through time by adopting one of a generally large set of potential values. It is the *uncertainty* of the future evolution of the state of the world that tickles our phantasies, our incentives for taking risks, and our desires to predict the future.

Dynamism is the second important aspect that makes these systems variegated enough to fascinate our attention. Finally the (partial) controllability of the future creates complex optimization problems that keep challenging our intellectual capacities.

In the process of modelling these dynamic systems a major breakthrough was originated by the Russian mathematician Andrei Andreivich Markov (1856-1922). Previously uncertainty had been treated as a sequence of independent trials which provides on adequate description only for simple situations like parlor games as roulette or black jack.

As a major extension Markov incorporated the possibility of the current state of the system, to depend in a probabilistic sense upon its previous value. It is remarkable that this Markov property fits most of the discrete-time systems we encounter (albeit after a possible respecification of the information to be embodied into the state concept).

What is now called the Markovian model of dependence, has incited a tremendous development in probability theory and the theory of stochastic processes with several areas (Markov chains, Markov processes) bearing the name of their founding father.

Within the field of optimization theory or Operations Research it has inspired Richard Bellman, Ronald Howard and Lloyd Shapley in the fifties and early sixties to create the area of Markov Decision Theory.

Here we generally assume that a decision maker derives rewards and costs out of a given system, which he can control by choosing, in dependence on the observed state an alternative out of a set of feasible options.

His choice influences both the expectation of his current reward, and the evolution of the state of the system in the future. As a consequence his objective is to determine a complete strategy for his entire planning

period, i.e. a sequence of rules to be applied in each of the decision epochs. The above described framework has proven to be extremely useful when analyzing a wide range of problems, like determining policies for the control of inventory and production systems, the operation of water or raw material resevoirs, the regulation of traffic and telecommunication systems as arising e.g. in computer networks; as well as the determination of optimal harvesting policies.

Quite frequently the control over a particular system is in the hands of several parties which tend to have conflicting interests. As Luce and Raiffa, in their book, "Games and Decisions", Wiley & Sons (1957) have put it: "In all of man's written record there has been preoccupation with conflict of interest; possibly only the topics of God, love and inner struggle have received comparable attention". And obviously, even inner struggle could or should be modelled as a game between several centers of the human mind, thereby abandoning its representation as a monolithic entity, and no games seem to be more fascinating than those originating from *love*, the players' interests only partially and occasionally concurring. Finally, even theological discussions center around the question whether and if so, in what way and to what extent a divine Ruler sets boundaries to our freedom of *controlling* the world and points out strategies to act upon.

The extension of Markov Decision Theory to the case where *several* players control the system simultaneously has led to the area of stochastic games. Part II (chapters 6, 7 and 8) is devoted to the latter.

Solving these *Markovian control problems* generally amounts to deriving (a system of) functional (or optimality) equations. Next, one analyzes the properties of the solution space of these equations, and develops algorithms for finding a particular solution.

In some cases, like in *finite* Markov Decision Problems (cf. chapter 1) exact and finite algorithms can be obtained, based upon techniques like Policy Iteration of Linear Programming. However, for large systems, these exact methods become infeasible, because at each step of the algorithm, they tend to require the solution of a large set of equations. We therefore try to concentrate upon *successive approximation methods* which in general can tackle much larger problems, and in some models are the only alternative available.

In chapter 1, we review various successive approximation methods for Markov Decision Problems (MDP's) with finite state and action spaces. We deal both with the *discounted* version (where the present value of a stream

of rewards is the criterion to be considered) and with the *undiscounted* problem (where the long-run average return per unit time is to be optimized). Both the convergence conditions of these methods and their *rates* of convergence are analyzed. In addition we dwell on the computation of upper and lower bounds on various quantities of interest, elimination schemes for non-optimal actions and data-transformations to ensure or accelerate the convergence of our algorithms. As a second topic we present a number of turnpike properties, which show the relation between the finite-horizon and infinite-horizon models.

From a mathematical point of view, most of the problems considered in chapter 1 center around the properties of the so-called *value-iteration operator* which turns the total expected reward function for a planning period of n epochs into the corresponding function for a planning period of n + 1 epochs. Whereas this operator is a contraction mapping in the *discounted* model, it fails to satisfy this property in the *undiscounted* version. However in some cases a *reduction* to a contraction operator can be achieved. In chapter 2, we derive both necessary and sufficient conditions for this reducibility. This "reduced contraction-property" has important consequences for the (geometric) convergence rate of the value-iteration method, which is the most minutely discussed algorithm in chapter 1. Moreover it can be exploited to obtain lower and upper bounds, variational characterizations for the fixed points of the optimality equations, as well as tests for eliminating suboptimal actions.

Chapter 3 is devoted to the case where the parameters of the MDP-model can only be obtained via approximating schemes, or where it is computationally preferable to approximate the parameters rather than employing exact algorithms for their computation. This situation occurs e.g. when the one-step rewards appear as the optimal values of underlying optimization problems or when one faces the *combined* problem of simultaneously having to determine the *design* of a particular system, as well as a policy for its day-to-day operation.

Finally a third example occurs when trying to solve *nested* sequences of (piecewise linear) functional equations, where each functional equation has the structure of the optimality equation in undiscounted MDP's or Markov Renewal Programs (cf. section 1.9).

Nested sequences of functional equations of the above described type, occur e.g. when considering next to the average return per unit time criterion, a set of more selective criteria. It is the objective of chapter 4,

to derive a successive approximation method for solving these systems of functional equations.

We next consider the case where the state space of our problem is *denumerable* instead of finite. Here optimal policies, as well as a solution to the average return optimality equation may fail to exist. In chapter 5, we present a number of recurrency conditions on the underlying laws of motion, under which the optimality equation has a bounded solution. Such a solution yields, in fact, a policy which is optimal for a strong version of the average return optimality criterion. Besides the existence of a bounded solution to this equation, we will show that both the value-iteration and the policy-iteration method can be used to determine such a solution.

Chapter 6 in Part II of this book is devoted to the *stochastic games-model* where a finite number (say N) players control the system simultaneously. Here the objective is to find an *equilibrium* tuple of rules, i.e. an N-tuple of policies with the property that no player is able to better himself while the other players tie themselves down to their respective policies. The state space is again taken to be denumerable, and we consider both the discounted and the undiscounted version of the game.

Finally, the last two chapters deal with the *finite* two-person model where we have a closed system, i.e. everything player 1 wins must be lost by player 2, and vice versa. We consider once again, the average return and a number of more selective equilibrium criteria. In chapter 7 we discuss the functional equations that arise in this model, and in chapter 8 we obtain two successive approximation methods for solving the undiscounted model.

We assume the reader of this book to be familiar with the basic principles of mathematical optimization, linear algebra, calculus and Markov Chain theory. Some elementary knowledge of game theory could help the understanding of part II.

Concerning the numbering of formal statements, theorem 7.2.3 is the third theorem of section 7.2 in chapter 7. Equations, lemmas, propositions, corollaries and definitions are numbered in the same way. The symbol □ signifies the end of a proof.

Part I. Markov Decision Theory

CHAPTER 1

Value iteration in finite Markov Decision Problems

1.1. INTRODUCTION

We first describe the frequently studied model of Markov Decision Problems (MDPs) (cf. e.g. [5],[63]). A system is observed at equally spaced epochs numbered 0,1,2,... . At each epoch the system is observed to occupy one of N states, which are numbered 1 through N. Let $\Omega = \{1,\ldots,N\}$ denote the state space of the problem. Each state i has associated with it a finite non-empty decision set K(i). Whenever state i is observed, some decision $k \in K(i)$ must be chosen, after which a one-step expected reward q_i^k is earned immediately, whereas the probability that state j is to be observed at the next epoch, is given by P_{ij}^k ($P_{ij}^k \geq 0$; $\sum_{j=1}^N P_{ij}^k = 1$; $i, j \in \Omega$, $k \in K(i)$).

This introductory chapter surveys both older and recent results on the asymptotic behaviour of the value-iteration scheme

$$(1.1.1) \qquad v(n+1)_i = \max_{k\in K(i)}[q_i^k + \beta \sum_{j=1}^N P_{ij}^k \, v(n)_j], \quad 1 \leq i \leq N;\ n = 0,1,\ldots$$

with $0 \leq \beta \leq 1$ and where the starting point v(0) (scrap value vector) is arbitrary and $v(n)_i$ denotes the maximum possible expected n-period reward starting from state i (cf. DERMAN [25]).

Parts of this chapter have been distilled from survey papers by FEDERGRUEN & SCHWEITZER [34] and FEDERGRUEN, SCHWEITZER & TIJMS [44].

Asymptotic results are of interest because they show the relation between the finite-horizon and infinite-horizon models where use of the latter case is justified if the planning horizon is large, although possibly not (exactly) known. Two types of asymptotic results are presented. One type involves the asymptotic behaviour of the value function, i.e.

(1) v(n) if the discount factor β satisfies $0 \leq \beta < 1$, or

(2) $v(n) - ng^*$ where g^* is the maximal gain rate vector in the *undiscounted* case where $\beta = 1$.

The other type of asymptotic result concerns the behaviour of the sequence of the sets of optimizing policies S(n), where

$$(1.1.2) \qquad S(n) = X_{i=1}^{N} K(n,i); \quad n = 1,2,3,\ldots$$

with

$$K(n,i) = \{k \in K(i) \mid v(n)_i = q_i^k + \beta \sum_{j=1}^{N} P_{ij}^k \, v(n-1)_j\}; \; i = 1,\ldots,N$$

as well as the existence of so-called initially stationary or periodic optimal or ε-optimal strategies (see below).

The following notation will be employed. We let $S = X_{i=1}^{N} K(i)$ denote the finite set of *policies*. We use the notation $f = (f(1),\ldots,f(N))$ where $f(i) \in K(i)$ denotes the alternative used in state i, $i \in \Omega$.

A strategy $\pi = (\ldots,f^{(\ell)},\ldots,f^{(1)})$ is an infinite sequence of policies where applying strategy π means using policy $f^{(\ell)}$ when there are ℓ periods to go.

A strategy is said to be *stationary* if it uses the same policy at each period, i.e. if $f^{(\ell)} = f$ for all $\ell = 1,2,\ldots$. Note that each policy specifies a stationary strategy. Likewise, a strategy $\pi = (\ldots,f^{(\ell)},\ldots,f^{(1)})$ is called *initially* stationary if there exists an integer $n_0 \geq 1$ and a policy f such that $f^{(\ell)} = f$ for all $\ell \geq n_0$.

Finally, a strategy is *optimal* (or ε-optimal for some $\varepsilon > 0$) if for each $n = 1,2,\ldots$ the total expected reward when there are n-periods to go equals (comes within ε of) the maximal vector $v(n)$, for every possible starting state $i \in \Omega$.

Observe that a strategy $\pi = (\ldots,f^{(\ell)},\ldots,f^{(1)})$ is optimal if and only if $f^{(\ell)} \in S(\ell)$ for all $\ell = 1,2,\ldots$. For each $\varepsilon > 0$, and $n = 1,2,\ldots$ we define $S(n,\varepsilon)$:

$$S(n,\varepsilon) = \{f \in S \mid q_i^{f(i)} + \beta \textstyle\sum_j P_{ij}^{f(i)} \, v(n-1)_j \geq v(n)_i - \varepsilon, \; i = 1,\ldots,N\}$$

Associated with each policy $f = (f(1),f(2),\ldots,f(N)) \in S$ are the reward vector $q(f) = [q_i^{f(i)}]$ and transition probability matrix (tpm) $P(f) = [P_{ij}^{f(i)}]$. Thus (1.1.1) and (1.1.2) may be rewritten as:

$$(1.1.3) \qquad v(n+1) = Qv(n) = Q^{n+1}v(0), \quad n = 0,1,2,3,\ldots$$

where the operator Q: $E^N \to E^N$ is defined by:

(1.1.4) $\quad Qx_i = \max_{k \in K(i)} [q_i^k + \beta \sum_{j=1}^{N} p_{ij}^k x_j], \quad i = 1,\ldots,N.$

Separate treatment will be given for the discounted and undiscounted cases. In both models, the geometric rate of convergence of the value function (i.e. of $v(n)$ or of $v(n) - ng^*$) plays a central role.

In section 2 and 3 we deal with the discounted case. In section 4, we first give the notation and preliminary results that will be needed in the following chapters.

In section 5, a historical review will be given of the study of convergence conditions for undiscounted value-iteration. In section 6, we discuss the *rate* of convergence for this undiscounted model.

The behaviour of the sequence of sets of optimizing policies $S(n)$ is discussed in section 7, whereas section 8 presents some algorithms and data-transformations that may be applied in the undiscounted case. Section 9 finally gives an introduction to Markov Renewal Programs (MRPs).

1.2. DISCOUNTED CASE: ASYMPTOTIC BEHAVIOUR OF $v(n)$

The discounted case possesses an elegant treatment because the Q-operator defined by (1.1.4) is a contraction operator with contraction modulus less than or equal to $\beta < 1$ when we use the norm $\|x\| = \max_i |x_i|$

(1.2.1) $\quad \|Qx - Qy\| \leq \beta \|x-y\|, \quad$ all $x,y \in E^N$.

The classical theory of contraction operators summarized for example in DENARDO [20], may be brought to bear, with the following immediate results:

First let Q^n denote the n-fold application of the Q-operator, i.e. $Q^n x = Q(Q^{n-1}x)$ for $n \geq 1$ with $Q^0 x = x$.

(1.2.2) $\quad Q$ has a unique fixed point $v^* = Qv^*$

(1.2.3) $\quad$ for any starting point x, $Q^n x$ converges geometrically to the fixed point:

(1.2.4) $\quad \|Q^n x - v^*\| \leq \beta^n \|x - v^*\|; \quad n = 1,2,3,\ldots$

(1.2.5) $\quad$ an upperbound on the distance between $Q^n x$ and v^* can be computed after just one iteration of Q via

(1.2.6) $\quad \|Q^n x - v^*\| \leq \beta^n \|Qx - x\|/(1-\beta); \quad n = 1,2,\ldots$

which is fairly sharp provided β is not too close to unity.

Additional properties follow from the fact that Q is a monotone operator ($x \geq y$ implies $Qx \geq Qy$). E.g.

$$(1.2.7) \qquad v_i^* = \max_{f \in S} v(f)_i, \quad 1 \leq i \leq N$$

where $v(f)$ is the total expected discounted return vector associated with policy f:

$$(1.2.8) \qquad v(f) = \sum_{n=0}^{\infty} \beta^n P(f)^n q(f) = [I - \beta P(f)]^{-1} q(f).$$

Observe that both v^* and $v(f)$, $f \in S$, are independent of the scrap value vector $v(0) \in E^N$.

As a consequence the unique fixed point of the Q-operator coincides with the maximal total discounted return vector. Moreover,

$$(1.2.9) \qquad x \geq v^* (x \leq v^*) \text{ implies } Q^n x \downarrow v^* (Q^n x \uparrow v^*)$$

Some of these results have been modified by using instead of (1.2.1):

$$(1.2.10) \qquad \beta(x-y)_{min} \leq (Qx-Qy)_{min} \leq (Qx-Qy)_{max} \leq \beta(x-y)_{max}$$

where $x_{min} = \min_i x_i$ and $x_{max} = \max_i x_i$.
Thus (1.2.6) is replaced by

LEMMA 1.2.1. *For all* $n \geq 1$; $x \in E^N$ *and* $i \in \Omega$:

$$(1.2.11) \qquad x_i + \frac{(Qx-x)_{min}}{(1-\beta)} \leq Qx_i + \frac{\beta(Qx-x)_{min}}{(1-\beta)} \leq \ldots \leq Q^n x_i + \frac{\beta^n (Qx-x)_{min}}{(1-\beta)}$$

$$\leq v(f^{(n)})_i \leq v_i^* \leq$$

$$\leq Q^n x_i + \frac{\beta^n}{(1-\beta)} \frac{(Qx-x)_{max}}{(1-\beta)} \leq \ldots \leq Qx_i + \frac{\beta}{1-\beta} \frac{(Qx-x)_{max}}{(1-\beta)} \leq x_i + \frac{(Qx-x)_{max}}{(1-\beta)}$$

where $f^{(n)} \in S(n)$, *with* $v(0) = x$ *and* $v(f^{(n)})$ *is the associated total return vector.*

PROOF. Note by a repeated application of (1.2.10) that

$$Q^{n+1} x_i - Q^n x_i \geq \beta^n [Qx-x]_{min} = \frac{\beta^n}{1-\beta} [Qx-x]_{min} - \frac{\beta^{n+1}}{1-\beta} [Qx-x]_{min}$$

and hence $Q^{n+1}x_i + \frac{\beta^{n+1}}{1-\beta}[Qx-x]_{min} \geq Q^n x_i + \frac{\beta^n}{1-\beta}[Qx-x]_{min}$, for all $n \geq 0$ and $i \in \Omega$, which proves that the sequence of lower bounds $\{Q^m x_i + \frac{\beta^m}{(1-\beta)}[Qx-x]_{min}\}_{m=1}^{\infty}$ is monotonically non-decreasing towards v_i^*. In a similar way one verifies that the sequence of upper bounds $\{Q^m x_i + \frac{\beta^m}{(1-\beta)}[Qx-x]_{max}\}_{m=1}^{\infty}$ is monotonically non-increasing towards v_i^*. The inequality $v_i^* \geq v(f^{(n)})_i$ is immediate from (1.2.7) which leaves us with the proof of

$$Q^n x_i + \frac{\beta^n}{(1-\beta)}[Qx-x]_{min} \leq v(f^{(n)})_i, \quad n \geq 1 \text{ and } i \in \Omega.$$

Let $H: E^N \to E^N: x \to q(f^{(n)}) + \beta P(f^{(n)})x$ and note as a special case of (1.2.4) that $\lim_{m\to\infty} H^m y = v(f^{(n)})$ for all $y \in E^N$. Then

$$[v(f^{(n)}) - Q^n x]_{min} = [\lim_{m\to\infty} H^m Q^{n-1}x - Q^n x]_{min} =$$

$$\lim_{m\to\infty}[H^m Q^{n-1}x - Q^n x]_{min} = \lim_{m\to\infty}[\sum_{\ell=1}^{m-1}(H^{\ell+1}Q^{n-1}x - H^{\ell}Q^{n-1}x)]$$

$$\geq \sum_{\ell=1}^{\infty}[H^{\ell}Q^n x - H^{\ell}Q^{n-1}x]_{min} \geq \sum_{\ell=1}^{\infty}\beta^{\ell+n-1}[Qx-x]_{min} =$$

$$\beta^n(1-\beta)^{-1}[Qx-x]_{min},$$

where the third equality follows from $HQ^{n-1}x = Q^n x$ in view of $f^{(n)} \in S(n)$ and where the last inequality follows from a repeated application of (1.2.10). □

Note that the bounds in (1.2.11) are invariant to adding a constant c to each component of x. In addition we recall that the bounds in (1.2.11) were originally derived for n = 1 by PORTEUS [94] who sharpened MacQueen's ([81]) original bounds (cf. (1.2.11) with n = 1):

$$x_i + (1-\beta)^{-1}(Qx-x)_{min} \leq v(f^{(1)})_i \leq v_i^* \leq x_i + (1-\beta)^{-1}(Qx-x)_{max}$$

Additional improvements on the bounds as well as on the rate of convergence can be based upon data transformations ([95], [107], [38], [108]) or Gauss-Seidel variants of the iterative scheme ([52], [73], [107]), extrapolation and over-relaxation techniques ([96], [103], [125]) as well as by removal of self transitions. These transformations obviously destroy the interpretation of v(n).

In terms of the original value-iteration scheme $v(n) = Q^n x$ where $x = v(0)$, the above results have been useful in at least four ways:

(a) v(n) is shown to approach v^* geometrically fast

(b) the n = 0 or 1 versions of (1.2.6) or (1.2.11) get computable bounds on the error between the fixed point v^* and the current best guess (x or Qx)

(c) eliminations via the bounds of alternatives which are not optimal for the ∞-horizon problem, cf. MACQUEEN [81], HASTINGS and MELLO [55] and GRINOLD [50]

(d) prior estimation of how many *additional* iterations n(x) are required given that the current estimate of v^* is x, until the new estimate $Q^n x$ lies within ε of v^* or until a policy $f^{(n)} \in S(n)$ found at the end of these n iterations has a return vector $v(f^{(n)})$ which lies within ε of v^*. Bounds on n(x) are obtained by setting (cf. (1.2.6))

$$(1.2.12) \quad \|Q^n x - v^*\| \leq \frac{\beta^n \|Qx - x\|}{1-\beta} \leq \varepsilon$$

or cf. FINKBEINER and RUNGALDIER [45]:

$$(1.2.13) \quad 0 \leq v^* - v(f^{(n)}) \leq \frac{2\beta^n \|Qx - x\|}{1-\beta} \underline{1} \leq \varepsilon \underline{1}$$

with the result that at most

$$(1.2.14) \quad n(x) \leq \ln\left[\frac{(1 \text{ or } 2)\|Qx - x\|}{\varepsilon(1-\beta)}\right] / \left|\ln(\beta)\right|$$

additional iterations are required. This has the property $n(Qx) \leq n(x)-1$, so that the number of remaining iterations to get accuracy ε decreases by at least unity with each iteration; hence the termination criterion will be met after a finite number of steps.

Unfortunately n(x) can be large if β is close to unity or if the initial guess x is far from v^*. An encouraging feature is that n(x) varies only logarithmically with ε so that it is practical to achieve high precisions as long as β is not too close to unity.

We finally note that using (1.2.11) the upperbound for n(x) in (1.2.14) may be replaced by:

$$(1.2.15) \quad n(x) \leq \ln\left[\frac{sp[Qx - x]}{\varepsilon(1-\beta)}\right] / \left|\ln(\beta)\right|$$

where

$$(1.2.16) \quad sp[x] = x_{max} - x_{min}$$

denotes the span of x (cf. BATHER [3]).

1.3. DISCOUNTED CASE; ASYMPTOTIC BEHAVIOUR OF S(n) AND THE EXISTENCE OF INITIALLY STATIONARY ε-OPTIMAL STRATEGIES

The main question of interest is the relation of the sets S(n) to the set S^* of policies which are optimal for the infinite horizon problem.

$$(1.3.1) \qquad S^* = \{f \in S \mid v^* = q(f) + \beta P(f) v^*\}.$$

Note that by (1.2.2) S^* is uniquely determined and has a Cartesian product structure.

It follows directly from (1.2.13) and (1.2.14) that for each starting point $x = v(0)$, we find $S(n) \subseteq S^*$, for sufficiently large n, say $n \geq n_1(x)$. As a choice of $n_1(x)$ one may evaluate (1.2.14) with

$$(1.3.2) \qquad \varepsilon > \varepsilon_0 = \begin{cases} \min\{v_i^* - v(f)_i \mid f \in S \text{ and } 1 \leq i \leq N \text{ such that } v_i^* > v(f)_i\} & \text{if } S^* \neq S \\ \infty & \text{, if } S^* = S. \end{cases}$$

Thus value-iteration eventually settles upon optimal policies. Unfortunately this result can not be used in general while performing calculations because the lack of prior knowledge about v^* - and the resulting inability to evaluate ε_0 - makes it impossible to calculate $n_1(x)$ a priori. Estimation of $n_1(x)$ remains an outstanding problem. Until the problem is resolved, no ways are available to deduce whether a policy in S(n) lies in S^*, except by elimination of suboptimal actions. That is, a policy can appear during the first (say) 50 iterative steps yet fail to be optimal for the infinite horizon-model. Furthermore a policy from S^* might appear in say S(1), *not* appear in S(2) and reappear in S(4) (or never reappear); so that a policy which has "dropped out" of S(n) cannot be eliminated as suboptimal (cf. [114]).

In the special case where v^* is known, ε_0 may be estimated (cf. SHAPIRO [114]) from:

$$v_i^* - v(f)_i = [I - \beta P(f)]^{-1} \Gamma(f)_i = \sum_{n=0}^{\infty} \sum_{j=1}^{N} (\beta P(f))_{ij}^n \Gamma(f)_j, \quad i \in \Omega$$

where

$$\Gamma(f) = [v^* - q(f) - \beta P(f) v^*] \geq 0.$$

Namely assuming that S^* is a proper subset of S, i.e. $\varepsilon_0 < \infty$ we can pick a pair (i,f) which achieve ε_0 in (1.3.2) and a state j and an integer $n \leq N$, such that $(\beta P(f))_{ij}^n > 0$ and $\Gamma(f)_j > 0$. We thus find:

$$\varepsilon_0 \geq (\beta\alpha)^n \delta_0$$

where

$$\alpha = \min\{P_{rs}^k \mid \text{all } 1 \leq r,s \leq N \text{ and } k \in K(r) \text{ with } P_{rs}^k > 0\}$$

$$\delta_0 = \min\{\Gamma(f)_j \mid \text{all } f \in S, 1 \leq j \leq N \text{ with } \Gamma(f)_j > 0\} > 0$$

the last inequality following from the assumption $S^* \neq S$. Hence it suffices to take $\varepsilon = (\beta\alpha)^N \delta_0$ when computing $n_1(x)$ via (1.2.14).

The following properties are known regarding convergence of S(n) for large n

(a) if S^* is a singleton, S(n) must reduce to S^* for large enough n (i.e. for $n \geq n_1(x)$)

(b) if S^* is not a singleton, S(n) does not need to possess a limit as n tends to infinity. SHAPIRO [114] has constructed a 2-state example where S(n) oscillates with period 2 between the two members of S^*. Both his example, and an example in BROWN [13] suggest that the set S(n) exhibits at least an ultimately *periodic* behaviour. However, an example which is similar to the one given in BATHER [2] for the *undiscounted* case (see below) shows that the worst behaviour of S(n) will be *non-periodic* oscillations.

(c) Since S(n), for large n, may oscillate or contain only a *proper* subset of S^*, the individual S(n)'s do not by themselves determine S^*. However, one may find the entire set S^* from ε-optimal policies, i.e. from

$$(1.3.3) \qquad S^* = \lim_{n\to\infty} S(n,\varepsilon_n)$$

where $\{\varepsilon_n\}_{n=1}^{\infty}$ may be taken as an arbitrary sequence of positive numbers approaching 0, provided that the rate of convergence of $\{\varepsilon_n\}_{n=1}^{\infty}$ is slower than the one $\{v(n)\}_{n=1}^{\infty}$ exhibits, i.e. whenever $\varepsilon_n \beta^{-n} \to \infty$, as $n \to \infty$. One choice is $\varepsilon_n = n^{-1}$ and more generally, take ε_n^{-1} as a positive polynomial in n. To confirm (1.3.3) note that for all $f \in S(n,\varepsilon_n)$ $- \varepsilon_n \leq q(f) + \beta P(f) v(n-1) - v(n) = q(f) + \beta P(f) v^* - v^* + O(\beta^n)$.

In view of $\lim_{n\to\infty} \varepsilon_n = 0$, this implies that for all n sufficiently large, $q(f) + \beta P(f) v^* = v^*$, i.e. $S(n,\varepsilon_n) \subseteq S^*$ for all n sufficiently large. To prove the reversed inclusion, note that for $f \in S^*$, $q(f) + \beta P(f) v(n-1) - v(n) = q(f) + \beta P(f) v^* - v^* + O(\beta^n) = O(\beta^n) \geq - \varepsilon_n$ for all n sufficiently large, as a consequence of $\varepsilon_n \beta^{-n} \to \infty$ as $n \to \infty$.

We finally turn to the issue of determining initially stationary optimal strategies. We observed before that an optimal strategy must lie in $X_{n=1}^{\infty} S(n)$. SHAPIRO's example (cf. [114]) shows that in general there may be no (optimal) policy which is contained within all of the sets S(n) for all n large enough. That is, $\lim\inf_{n\to\infty} S(n)$ may be empty, or, none of the sequences of policies that may be generated by value-iteration needs to converge. So in general, no initially stationary optimal strategy may exist and the adaptation of example 1 in BATHER [2], mentioned above, shows that in general no initially *periodic* optimal strategy needs to exist either.

Only in the case where S^* is a singleton ($S^* = \{f^*\}$), do we know that $S(n) = \{f^*\}$ for all $n \geq n_1(x)$, so that in this case every optimal strategy is initially stationary. Or, in other words, f^* is the best choice of current policy if the planning horizon is *at least* $n_1(x)$ *additional* periods and this choice is optimal without knowing the exact length of the planning horizon.

We observe however that every policy in S^* comes closer and closer to being optimal at the n^{th} stage, as n tends to infinity. This may be verified from

$$\|v(n) - q(f) - \beta P(f) v(n-1)\| = \|v(n) - v^* - \beta P(f)[v(n-1)-v^*]\|$$
$$\leq 2\beta^n \|Qx-x\|/(1-\beta), \quad f \in S^*$$

using (1.2.6).

This in turn implies for every $\varepsilon > 0$, the existence of an initially stationary strategy that is ε-*optimal*. In addition we point out the following two properties:

(1) Any policy in S^* may be used in the initially stationary part of the ε-optimal strategy; i.e. the initially stationary part does not depend upon the scrap-value vector v(0).

(2) An upperbound for the length of the non-stationary tail of the ε-optimal strategy is given by (cf. [34])

$$m(x) \leq \ln\left[\frac{2\|Qx-x\|}{\varepsilon(1-\beta)^2}\right] / \left|\ln(\beta)\right|$$

which varies again logarithmically with the precision ε.

1.4. UNDISCOUNTED CASE: NOTATION AND PRELIMINARIES

In the undiscounted case, $\beta = 1$ and Q is a non-expansive operator:

(1.4.1) $(x-y)_{min} \leq (Qx-Qy)_{min} \leq (Qx-Qy)_{max} \leq (x-y)_{max}$; all $x,y \in E^N$.

In addition the Q operator has the property

(1.4.2) $Q(x+c\underline{1}) = Qx + c\underline{1}$ for all $x \in E^N$ and scalars c.

Note as a consequence of (1.4.2) that the Q operator never has a unique fixed point and hence is never a contraction operator on E^N (and neither is any of its powers). Both (1.4.1) and (1.4.2) suggest choosing (cf. (1.2.16))

$$sp[x] = x_{max} - x_{min}$$

as a quasi-norm (cf. BATHER [3]). However, example 2 in chapter 2 shows that Q (or any of its powers) is not necessarily *contracting* with respect to the sp-norm either. That is, only under special conditions with respect to the (chain- and periodicity) structure of the problem, (cf. chapter 2) does there exist a number $0 \leq \alpha < 1$, and an integer $n \geq 1$ such that

$$sp[Q^n x - Q^n y] \leq \alpha \, sp[x-y]; \text{ for all } x,y \in E^N.$$

As a consequence the asymptotic behaviour of $\{v(n)\}_{n=1}^{\infty}$ requires an entirely different and more complicated analysis in the undiscounted case.

Randomized policies turn out to play an indispensible role in the study of the asymptotic behaviour of $\{v(n)\}_{n=1}^{\infty}$. We therefore define a (stationary) randomized policy f as a tableau $[f_{ik}]$ satisfying $f_{ik} \geq 0$ and $\sum_{k \in K(i)} f_{ik} = 1$, where f_{ik} is the probability that the k-th alternative is chosen when entering state i. We therefore distinguish between S_R, the set of all randomized policies, and $S_P \subseteq S_R$ the set of all pure (non-randomized) policies (i.e. each $f_{ik} = 0$ or 1, for $f \in S_P$).

We associate again, with each $f \in S_R$, a N-component reward vector q(f) and N×N-matrix P(f):

(1.4.3) $q(f)_i = \sum_{k \in K(i)} f_{ik} q_i^k$; $P(f)_{ij} = \sum_{k \in K(i)} f_{ik} P_{ij}^k$, $1 \leq i, j \leq N$.

Note that P(f) is a stochastic matrix. For any $f \in S_R$, define the stochastic matrix $\Pi(f)$ as the Cesaro limit of the sequence $\{P^n(f)\}_{n=1}^{\infty}$ and define the *fundamental* matrix

(1.4.4) $Z(f) = [I - P(f) + \Pi(f)]^{-1}$.

These matrices always exist and have the following properties (cf. [10], [71]):

$$(1.4.5)\quad \Pi(f) = P(f)\Pi(f) = \Pi(f)P(f) = \Pi(f)^2 = \Pi(f)Z(f) = Z(f)\Pi(f)$$

$$(1.4.6)\quad [I-P(f)]Z(f) = Z(f)[I-P(f)] = I-\Pi(f)$$

$$(1.4.7)\quad Z(f) = I + \lim_{a\uparrow 1} \sum_{n=1}^{\infty} a^n[P(f)^n - \Pi(f)].$$

Denote by $n(f)$ the number of subchains (closed, irreducible sets of states) for $P(f)$. Then

$$(1.4.8)\quad \Pi(f)_{ij} = \sum_{m=1}^{n(f)} \Phi_i^m(f)\, \pi_j^m(f), \qquad 1 \le i, j \le N$$

where $\pi^m(f)$ is the unique equilibrium distribution of $P(f)$ on the m^{th} subchain $C^m(f)$, and $\Phi_i^m(f)$ is the probability of absorption in $C^m(f)$, starting from state i (cf. [22] and [105]). Observe $\sum_i \pi_i^m(f) = 1$ and $\pi^m(f)P(f) = \pi^m(f)$, as well as $\Phi^m(f) = P(f)\Phi^m(f)$, $m = 1,\ldots,n(f)$.

Finally, let

$$(1.4.9)\quad R(f) = \{j \in \Omega \mid \Pi(f)_{jj} > 0\}$$

i.e. $R(f)$ is the set of *recurrent* states for $P(f)$, with $\Omega\backslash R(f)$ the set of transient states.

A policy $f \in S_R$ is said to be *aperiodic* in case the stochastic matrix $P(f)$ is aperiodic; otherwise, f is said to be *periodic*. For each $f \in S_R$, we define the gain rate vector $g(f) = \Pi(f)q(f)$, such that $g(f)_i$ represents the long run average expected return per unit time, when the initial state is i, and policy f is used. Next, define the maximal gain rate vector g^* by

$$(1.4.10)\quad g_i^* = \sup_{f\in S_R} g(f)_i, \qquad i = 1,\ldots,N.$$

We know from DERMAN [25] that there exist pure policies f which attain the N suprema in (1.4.10) simultaneously. As a consequence we define:

$$(1.4.11)\quad S_{PMG} = \{f \in S_P \mid g(f) = g^*\};\ S_{RMG} = \{f \in S_R \mid g(f) = g^*\},$$

as the set of all pure and the set of all randomized maximal gain policies. Next, define R^* as the set of states that are recurrent under some maximal gain policy

$$(1.4.12)\quad R^* = \{i \in \Omega \mid i \in R(f) \text{ for some } f \in S_{RMG}\} = \{i \in \Omega \mid i \in R(f) \text{ for some } f \in S_{PMG}\}$$

where the second equality in (1.4.12) was shown in th.3.2 part(a) of [109]. Likewise we define $\hat{R}$ as the set of states that are recurrent under (some) arbitrary policy:

$$(1.4.13)\quad \hat{R} = \{i \in \Omega \mid i \in R(f) \text{ for some } f \in S_R\} = \{i \in \Omega \mid i \in R(f) \text{ for some } f \in S_P\}$$

where the second equality in (1.4.13) is a special case of the second equality in (1.4.12) by taking every $q_i^k = 0$. Note that $R^* \subseteq \hat{R}$.

The following lemma was proven in th. 3.2 part(b) of [109]:

LEMMA 1.4.1.

(a) *There exist policies* $f \in S_{RMG}$, *with* $R(f) = R^*$.

(b) *There exist policies* $f \in S_R$, *with* $R(f) = \hat{R}$.

Note that randomization is essential for this result; in general no pure (maximal gain) policies need to exist, with a maximal set of recurrent states.

Finally we consider the well-known pair of optimality equations for the average return per unit time criterion:

$$(1.4.14)\quad g_i = \max_{k \in K(i)} \sum_j p_{ij}^k g_j, \quad i \in \Omega$$

$$(1.4.15)\quad v_i + g_i = \max_{k \in L(i)} \left\{ q_i^k + \sum_j p_{ij}^k v_j \right\}, \quad i \in \Omega$$

where

$$(1.4.16)\quad L(i) = \{k \in K(i) \mid g_i = \sum_j p_{ij}^k g_j\}.$$

We recall (cf. e.g. [109] th.3.1) that there always exists a solution pair (g,v) to (1.4.14) and (1.4.15). In addition any solution pair (g,v) to (1.4.14) and (1.4.15) has $g = g^*$, which implies that the g-part of the solution and hence each of the sets L(i) are uniquely determined. Finally let

$$(1.4.17)\quad V = \{v \in E^N \mid (g^*, v) \text{ satisfies } (1.4.15)\}.$$

For any $v \in E^N$, define

$$(1.4.18)\quad b(v)_i^k = q_i^k - g_i^* + \sum_{j=1}^N p_{ij}^k v_j - v_i, \quad i \in \Omega,\ k \in K(i).$$

Note that

(1.4.19) $v \in V \iff \max_{k \in L(i)} b(v)_i^k = 0.$

Note that unlike the discounted case, v is not uniquely determined by (1.4.15). Observe e.g. that if $v \in V$, then so does $v+c\underline{1}$ for any scalar c. A characterization of the set V is given in [109], and is rather complex. As an example , we merely state that the necessary and sufficient condition for $v \in V$ to be *unique up to a multiple of* $\underline{1}$ is given by:

(1.4.20) (UNI): There exists a policy $f \in S_{RMG}$, which has R^* as its single subchain.

The condition (UNI) will turn out to play an important role in the subsequent analysis.

We finally define for each $v \in V$, the sets of alternatives L(i,v) by:

(1.4.21) $L(i,v) = \{\ell \in L(i) \mid b(v)_i^\ell = 0 = \max_{k \in L(i)} b(v)_i^k\}, \quad i \in \Omega$

and let

(1.4.22) $S^*(v) = X_{i=1}^N L(i,v)$

denote the Cartesian product set of policies achieving the maxima in (1.4.15) for the particular solution $v \in V$.

Lemma 1.4.2 concludes this section by giving a characterization of the set of maximal gain policies.

LEMMA 1.4.2. (Properties of maximal gain policies)

a) $f \in S_{RMG}$ *if and only if* $g^* = P(f)g^*$ *and* $\Pi(f)[q(f)-g^*] = 0$.

b) *Let* $f \in S_R$:

(1) *Suppose that* $k \in L(i)$ *for each* (i,k) *with* $f_{ik} > 0$ *and that for some* $v \in V$, $b(v)_i^k = 0$ *for each* (i,k) *with* $f_{ik} > 0$, *and* $i \in R(f)$. *Then* $f \in S_{RMG}$.

(2) *Conversely, if* $f \in S_{RMG}$, *then for each* $i = 1,\ldots,N$: $f_{ik} > 0$ *implies* $k \in L(i)$ *and for* $i \in R(f)$, $f_{ik} > 0$ *implies* $b(v)_i^k = 0$ *for all* $v \in V$.

As to the proof of this lemma, we refer to [109], th.3.1, part (a) and (e).

1.5. UNDISCOUNTED CASE: ASYMPTOTIC BEHAVIOUR OF $\{v(n)\}_{n=1}^{\infty}$

The first asymptotic property of the sequence $\{v(n)\}_{n=1}^{\infty}$, is due to BELLMAN [5] who showed that if every one-step transition probability P_{ij}^{k} is strictly positive:

$$(1.5.1) \qquad \lim_{n\to\infty} \frac{v(n)_i}{n} = g^*, \quad \text{for all } i \in \Omega$$

where g^* is the maximal gain rate. Note that Bellman's assumption is the strongest possible, one can make with respect to the chain- and periodicity structure of the problem. HOWARD [63] conjectured that there generally exist two N-vectors g^* and v^* such that

$$(1.5.2) \qquad \lim_{n\to\infty} v(n) - ng^* - v^* = 0.$$

However, (1.5.2) may clearly fail to hold, if some of the transition probability matrices (tpm's) are periodic, as is illustrated by the two-state Markov process which has $P_{12} = P_{21} = 1$ and $q_1 = q_2 = 0$ (Take e.g. $v(0) = [1,0]$ and note that $\{v(n)\}_{n=1}^{\infty}$ alternates between the two limit points [1,0] and [0,1]). BROWN [13] showed in all generality that $\{v(n) - ng^*\}_{n=1}^{\infty}$ is bounded in n, permitting the interpretation of $g_i^* = \lim_{n\to\infty} v(n)_i/n$ as the maximal expected return per unit time starting from state i.
Two cases can be distuinguished.

In the first case $\{v(n) - ng^*\}_{n=1}^{\infty}$ has a limit for *any* choice of v(0). This corresponds roughly to the situation in the discounted process. In the second case, $\{v(n) - ng^*\}_{n=1}^{\infty}$ has a limit for some, but not all choices of v(0). It is possible to show that for each Markov Decision Process there exist $v(0) \in E^N$ such that $\lim_{n\to\infty} [v(n)-ng^*]$ exists, namely $v(0) = v^* + ag^*$ where $a >> 0$ and v^* satisfies the optimality equation (1.4.15) above (cf. lemma 2.2 in [111]).

It is also possible to construct MDP's in which case 2 holds, namely when certain tpm's have periodic states. For example consider a four-state MDP with only one policy f having

$$q(f) = \begin{vmatrix} 0 \\ 0 \\ 0 \\ 0 \end{vmatrix}, \qquad P(f) = \begin{vmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{vmatrix}, \qquad g^* = \begin{vmatrix} 0 \\ 0 \\ 0 \\ 0 \end{vmatrix}$$

(1.5.3) $\lim_{n\to\infty} v(n)$ exists if and only if $v(0) = (b,b,b,b)$

(1.5.4) $\lim_{n\to\infty} v(2n)$ exists whereas $\{v(n)\}_{n=1}^{\infty}$ has two distinct limit points, if $v(0) = (b,c,b,c)$ with $b \neq c$

(1.5.5) $\lim_{n\to\infty} v(4n)$ always exists, whereas $\{v(n)\}_{n=1}^{\infty}$ has four distinct limit points, if $v(0) = (b,c,d,e)$ with b,c,d,e distinct.

Conditions determining the existence of $\lim_{n\to\infty} [v(n)-ng^*]$ are of importance for at least the following reasons:

(1) If $v(0)$ is such that $\lim_{n\to\infty} [v(n)-ng^*]$ exists, then $v(n)-v(n-1)$ converges to g^*, and $nv(n-1)-(n-1)v(n)$ converges to a solution $v \in V$. That is, both the maximal gain rate vector and a solution to the optimality equation (1.4.15) can be computed.

(2) Convergence of $\{v(n)-ng^*\}_{n=1}^{\infty}$ guarantees that $S(n) \subseteq S_{PMG}$ for all n large enough (cf. ODONI [89]), hence value-iteration may be used to identify maximal gain policies. However if $v(0)$ is such that $\lim_{n\to\infty} [v(n)-ng^*]$ does not exist then $S(n) \subseteq S_{PMG}$ is *not* guaranteed to hold for all large n: LANERY [74] has given an example where $S(n) \subseteq S \backslash S_{PMG}$ for infinitely many n, and FEDERGRUEN & SCHWEITZER have given an example (cf. [35]) where $S(n) \subseteq S \backslash S_{PMG}$ for every n. In such cases value-iteration will not settle on maximal gain policies. In section 7 a more detailed analysis of the asymptotic behaviour of $\{S(n)\}_{n=1}^{\infty}$ will be given, both for the case where $\{v(n)-ng^*\}_{n=1}^{\infty}$ converges and for the one where it fails to converge.

Since value-iteration is the only practical computational method for finding maximal gain policies when $N >> 1$, it is desirable to check whether $\lim_{n\to\infty} [v(n)-ng^*]$ is guaranteed to exist, or whether a data transformation should be performed (cf. section 8) on the original data so as to enforce convergence.

(3) Convergence of $\{v(n)-ng^*\}_{n=1}^{\infty}$ guarantees the existence of initially stationary ε-optimal strategies for any positive ε.

Conversely, MDP's may be constructed in which for some choices of the scrap value vector $v(0)$ for which $\{v(n)-ng^*\}_{n=1}^{\infty}$ fails to converge, no initially stationary strategy can be found which is ε-optimal for ε sufficiently small (see section 7 below).

Sufficient conditions for the convergence of $\{v(n)-ng^*\}_{n=1}^{\infty}$ were obtained by WHITE [131], SCHWEITZER [104] and others. BROWN [13] and LANERY

[74] both obtained, albeit with faulty proofs, that there exists a positive integer J^*, such that

$$\lim_{n\to\infty}[v(nJ^*+r)-(nJ^*+r)g^*] \text{ exists for any } v(0) \text{ and any } r = 0,\ldots,J^*-1.$$

A new proof was provided by SCHWEITZER and FEDERGRUEN who obtained the following generalizations (cf. [110]):

(a) there exists an integer $J^* \geq 1$ such that $\lim_{n\to\infty}[v(nJ+r)-(nJ+r)g^*]$ exists for every $v(0) \in E^N$ and $r = 0,\ldots,J-1$ *if and only if* J is a multiple of J^*

(b) for any given $v(0) \in E^N$, there exists an integer $J^0 \geq 1$ which depends upon $v(0)$ such that

$$\lim_{n\to\infty}[v(nJ+r)-(nJ+r)g^*] \text{ exists for } \textit{some } r = 0,\ldots,J-1$$

if and only if

(1.5.6) J is a multiple of J^0.

In addition, if (1.5.6) holds then

(1.5.7) $\lim_{n\to\infty}[v(nJ+r)-(nJ+r)g^*]$ exists for *all* $r = 0,\ldots,J-1$.

As an illustration of part (b), (1.5.3)-(1.5.5) show $J^0 = 1,2,4$ depending upon $v(0)$. Note also that J^0 divides J^*, which is 4 in this example, and that for some $v(0)$, J^0 equals J^*.

The above results require a detailed investigation of the chain- and periodicity structure of the set of maximal gain policies, including the *randomized* ones. In fact J^* can be computed using a *finite* algorithm, and can be expressed as a function of the periods (and the chain structure) of the policies in S_{PMG}.

The consequence of (a) is that $\lim_{n\to\infty}[v(n)-ng^*]$ exists for all $v(0)$ if and only if $J^* = 1$, and the following theorem gives a number of equivalent statements of the necessary and sufficient condition for *global* convergence of $\{v(n)-ng^*\}_{n=1}^{\infty}$, i.e. convergence of $\{v(n)-ng^*\}_{n=1}^{\infty}$ for all $v(0) \in E^N$:

THEOREM 1.5.1. (cf. th.5.4 of [110])

The following three statements are (equivalent) necessary and sufficient conditions for the convergence of $\{v(n)-ng^*\}_{n=1}^{\infty}$ *for all* $v(0) \in E^N$:

(GC) (I) $J^* = 1$.

(GC) (II) *There exists an aperiodic randomized maximal gain policy* f, *with* $R(f) = R^*$.

(GC) (III) *Each state* $i \in R^*$ *lies within an aperiodic subchain of some randomized maximal gain policy.*

Example 1 below emphasizes the fact that the adjective "randomized" in conditions (II), and (III) cannot be replaced by (the more restrictive) "pure".

EXAMPLE 1.

$N = 4$; $K(1) = K(3) = K(4) = \{1\}$; $K(2) = \{1,2\}$;
$P^1_{12} = P^1_{34} = P^1_{42} = P^1_{21} = P^2_{23} = 1$; all $q^k_i = 0$, i.e.

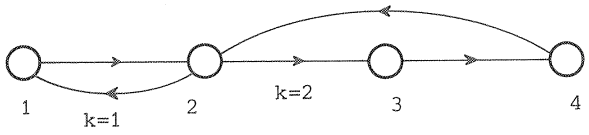


Figure 1.

Note that the two policies in S_P (and $S_{PMG} = S_P$) are both periodic with periods 2 and 3; however a *randomized* policy which uses both alternatives in state 2, is aperiodic and has $R(f) = R^* = \Omega$, and as a consequence $J^* = 1$. Note that neither (GC) (II) nor (GC) (III) holds when replacing "randomized" by "pure" (cf. also the examples in [110]). Example 1 shows that condition (I)-(III) contain the possibility that all of the *pure* policies are *periodic*; on the other hand, the existence of an aperiodic maximal gain policy f is only sufficient for global convergence, when $R(f) = R^*$.

We conclude this section by enumerating a number of conditions that are sufficient for the existence of $\lim_{n\to\infty} v(n) - ng^*$ for *all* possible choices of $v(0) \in E^N$.

We have seen that for arbitrary $J \geq 1$ and some *fixed* $v(0)$ the sequences $\{v(nJ+r)-(nJ+r)g^*\}_{n=1}^{\infty}$ may fail to converge for some (or all) $i \in \Omega$ and for some (or all) $r \in \{0,1,\ldots,J-1\}$. We refer to section 5 of [110] for an investigation of the various ways in which the convergence of these sequences interdepends.

THEOREM 1.5.2. (cf. th.5.5 of [110])

The following conditions are sufficient for the existence of $\lim_{n\to\infty} v(n)-ng^*$

for all $v(0) \in E^N$:

(I) *All of the transition probabilities are strictly positive:* $P_{ij}^k > 0$ *for all* $i,j \in \Omega$ *and* $k \in K(i)$ (cf. BELLMAN [5], BROWN [13]).

(II) *For all* $v(0) \in E^N$ *there exists an aperiodic* $f \in S_P$ *and an integer* n_0 *such that*
$v(n+1) = q(f)+P(f)v(n)$, *for all* $n \geq n_0$ (cf. MORTON and WECKER [86]).

(III) *There exists a state* s *and an integer* $\nu \geq 1$, *such that*
$P(f^1)\ldots P(f^\nu)_{is} > 0$ *for all* $f^1,f^2,\ldots,f^\nu \in S_P$; $i \in \Omega$ (cf. WHITE [131]).

(IV) *Every* $f \in S_P$ *is aperiodic* (cf. SCHWEITZER [104] & [106]).

(V) *Every* $f \in S_{PMG^*}$ *is aperiodic* (cf. SCHWEITZER [104] & [106]).

(VI) *For each* $i \in R^*$ *there exists a pure maximal gain policy* f, *such that state* i *is recurrent and aperiodic for* P(f).

(VII) *Every pure maximal gain policy has a unichained* tpm *and at least one of them is aperiodic.*

1.6. THE RATE OF CONVERGENCE OF UNDISCOUNTED VALUE-ITERATION

Whereas section 5 settles the issue if one demands global convergence, we recall that there always exists a (non-empty) closed subset $W \subseteq E^N$ of scrap-value vectors for which $\{v(n)-ng^*\}$ converges.

For any $x \in W$ we define

$$L(x) = \lim_{n\to\infty} Q^n x - ng^*$$

and recall the following easily verifiable properties of the $L(.)$-function:

(1.6.1) (a) $L(x) \in V$ (cf. [111], lemma 2.2 part(g))

(b) $L(Q^n x) = L(x) + ng^*$ (cf. [111] lemma 2.1 part(f)).

In this part we turn to the topic of the *rate* of convergence. As a major result, it can be shown (cf. [111]) that whenever $\{v(n)-ng^*\}_{n=1}^{\infty}$ converges, the approach to the limit is ultimately geometric in the sense that there exist numbers C and λ, with $0 \leq \lambda < 1$ such that

$$sp[v(n)-ng^* - L(v(0))] \leq C\lambda^n, \quad \text{for all } n = 1,2,\ldots$$

Applying the same analysis to so-called multi-step policies (cf. section 7), this result may be generalized in the sense that for all $v(0) \in E^N$ and all

$r = 1,\ldots,J^0(v(0))$

$v(nJ^0+r) - (nJ^0+r)g^*$ approaches its limit *geometrically* fast,

where J^0 was defined above.

As a consequence various successive approximation methods which are based on the value-iteration scheme (1.1.1) exhibit a geometric rate of convergence as well (cf. section 8). We observe that this generalization of

(1) what is known to be the case in a simple Markov Process, i.e. in a MDP with a single policy (cf. [113]), and
(2) White's result [131]

holds in all generality with no restrictions imposed on either the chain-, periodicity- or reward structure of the problem. In addition, the result is to some extent surprising, since we noted that the value-iteration operator Q, is in general not a (J step) contraction mapping for any J = 1,2,... on E^N (cf. chapter 2); nor is there in general an obvious way of reducing it to such a mapping on some subspace of E^N. To be more specific, we mentioned earlier (cf. section 4) that Q does not even need to be (J-step) contracting with respect to the (quasi) sp-norm, defined by (1.2.16), unless some very restrictive conditions on the chain- and periodicity structure of the problem are satisfied (cf. chapter 2).

Example 2 in chapter 2 shows e.g. that the combination of the (UNI)- and the (GC)- condition is in itself insufficient. The geometric convergence result of $\{v(n)-ng^*\}_{n=1}^{\infty}$ is obtained by analyzing the solution of the Q- operator in $\{Q^n x\}_{n=1}^{\infty}$ for any $x \in W$.

The evolution of the Q-operator

First of all we recall from lemma 2.2 in [111] or from BROWN [13] that for all $x \in E^N$ there exists an integer $n_1(x)$ such that

$$(1.6.2) \qquad Q^n x = T(Q^{n-1}x) = T^{n-n_1}(Q^{n_1}x) \quad \text{for all } n \geq n_1(x)$$

where the T-operator is defined by:

$$(1.6.3) \qquad Tx_i = \max_{k \in L(i)} \{q_i^k + \Sigma_j p_{ij}^k x_j\}; \quad x \in E^N.$$

This is due to the fact that, after a finite number of iterations, only

alternatives $k \in L(i)$ attain the maximum in the value-iteration equation (1.1.1).

Note that the T-operator has the additional properties:

(1.6.4) $T(x+cg^*) = Tx + cg^*$ for all $x \in E^N$; $c \in E^1$

and

(1.6.5) $sp[Tx-g^*-v] = sp[Tx-Tv] \leq sp[x-v]$, for all $x \in E^N$ and all $v \in V$.

In other words, after $n_1(x)$ iterations the "distance" between $Q^n x - ng^*$ and any $v \in V$, as measured by the sp[ ]-norm is monotonically *non-increasing*. Next, define for $x \in W$:

$$e(n,x) = Q^n x - ng^* - L(x).$$

We note that by using definition (1.4.18), it follows that $\{e(n,x)\}_{n=1}^{\infty}$ satisfies the recursion equation:

(1.6.6) $$e(n+1,x)_i = \max_{k \in L(i)} \{b(L(x))_i^k + \Sigma_j p_{ij}^k e(n,x)_j\}, \quad n \geq n_1(x).$$

Since $\lim_{n\to\infty} e(n,x) = 0$ for all $x \in W$, it follows that after a still larger number of (say after $n_2(x)$) iterations, only alternatives $k \in L(i)$ attain the maximum in the value-iteration equation (1.1.1), for which $b(L(x))_i^k = 0$. More specifically, for any $v \in V$ let

(1.6.7) $$\delta(v) = \min\{|b(v)_i^k| \mid i \in \Omega,\ k \in L(i),\ b(v)_i^k < 0\}.$$

Next, for any $x \in W$, let $n_2(x) = \inf\{n \mid n \geq n_1(x);\ sp[e(n,x)] < \delta(L(x))\} < \infty$. Then for all $x \in W$ and $n \geq n_2(x)$:

(1.6.8) $$e(n+1,x) = U(L(x))e(n,x)$$

where for any $v \in V$, the $U(v)$ - operator is defined by (cf. (1.4.21)):

(1.6.9) $$U(v)x_i = \max_{k \in L(i,v)} [\Sigma_j p_{ij}^k x_j], \quad i \in \Omega;\ x \in E^N.$$

To verify (1.6.8) assume to the contrary that for some $n > n_2(x)$, and $i \in \Omega$ there exists a $k \in L(i) \backslash L(i,L(x))$ which attains the maximum in (1.6.6).

Then $$e(n+1,x)_i \leq -\delta(L(x)) + e(n,x)_{min} + \sum_j p_{ij}^k [e(n,x)_j - e(n,x)_{min}]$$
$$\leq -\delta(L(x)) + e(n,x)_{min} + sp[e(n,x)] < e(n,x)_{min}$$

which contradicts $e(n+1,x)_{min} \geq U(L(x))e(n,x)_{min} \geq e(n,x)_{min}$ (cf. (1.4.1) and (1.6.6)). Observe that in spite of V being an infinite subset of E^N, only finitely many U(v)-operators occur since there are only finitely many subsets of $X_i L(i)$.

Note in addition that the U(v)-operators have, on top of the properties (1.4.1) and (1.4.2) of Q and the property (1.6.4) of T, the extremely useful characteristic of being *positively homogeneous* i.e.:

(1.6.10) $U(v)[ax] = aU(v)x$ for all $x \in E^N$ and $a \geq 0$.

As a consequence the convergence of $\{Q^n x - ng^*\}_{n=1}^{\infty}$ for any $x \in W$ occurs in three phases. The first $n_1(x)$ iterations constitute the *first* phase and the second phase terminates after the $n_2(x)$-th iteration, and is followed by the third phase from thereon.

We conclude this subsection by a short description of the behaviour of the Q-operator during the first phase. We first observe that this phase is void, if K(i) = L(i) for all $i \in \Omega$, which is e.g. the case when $g_i^* = \langle g^* \rangle$, $i \in \Omega$, i.e. when the maximal gain rate is independent of the initial state of the system. On the other hand, $n_1(x)$ may be unbounded in $x \in E^N$ or $x \in W$. In fact in the worst case the length of the first phase may be linear in sp[x] as is proven in [111], th.3.1. This is why the first phase is said to have a *finite though linear* type of convergence.

The following example illustrates this:

EXAMPLE 2.

$\Omega = \{1,2,3\}$; $K(1) = K(3) = \{1\}$; $K(2) = \{1,2\}$; $q_1^1 = q_2^1 = q_2^2 = 0$; $q_3^1 = -1$

$$P_{11}^1 = P_{21}^1 = P_{23}^2 = P_{33}^1 = 1.$$

k = 1 k = 2

Note that $g^* = (0,0,-1)$ and that $L(2) = \{1\}$

Let $x = [0,0,X]$ with $X >> 1$ and verify that $Q^n x = [0, \max(0,X-n+1), X-n]$ such that $n_1(x) = sp[x] = \|x\| = X$.

The behaviour of $sp[Q^n x - ng^* - L(x)] = sp[e(n,x)]$ during the first phase may be capricious. E.g. $\{sp[e(n,x)]\}_{n=1}^{\infty}$ may be alternatingly increasing and decreasing such that the first phase is not necessarily terminated as soon

as $[Q^n x - ng^*]$ starts coming closer in sp[ ]-norm to the limit $L(x)$ (cf. example 1 in [111]). In the second phase the Q-operator essentially reduces to the T-operator. Let

$$\tilde{W} = \{x \in E^N \mid \lim_{n\to\infty} T^n x - ng^* \text{ exists}\} \text{ and } \tilde{L}(x) = \lim_{n\to\infty} T^n x - ng^*, \text{ for } x \in \tilde{W}.$$

Note that by (1.6.4) we have $T^n v = v + ng^*$ for any $v \in V$ and $n \geq 1$, and so $V \subseteq \tilde{W}$. In analogy to $e(n,x)$ define for $n = 1,2,\ldots$ and $x \in \tilde{W}$:

$$\tilde{e}(n,x) = T^n x - ng^* - \tilde{L}(x). \tag{1.6.11}$$

It follows from (1.6.1) that for all $x \in W$, and with $n_1(x)$ defined by (1.6.2):

$$\tilde{L}(Q^{n_1}x) = \lim_{n\to\infty} T^n(Q^{n_1}x) - ng^* =$$

$$\lim_{n\to\infty} Q^{n+n_1(x)}x - (n+n_1(x))g^* + n_1(x)g^* = L(x) + n_1(x)g^*. \tag{1.6.12}$$

In other words for all $x \in W$, $Q^{n_1}x \in \tilde{W}$. As a consequence studying the convergence of $\{Q^n x - ng^*\}_{n=1}^{\infty}$ in the second and third phase amounts to characterizing the behaviour of T on $\tilde{W}$.

The Second and Third Phase. Geometric Convergence

First we define for all $x \in \tilde{W}$ and all $n = 1,2,\ldots$ the n-step contraction factor $f_n(x)$ by:

$$f_n(x) = \begin{cases} \dfrac{sp[\tilde{e}(n,x)]}{sp[\tilde{e}(0,x)]} = \dfrac{sp[T^n x - ng^* - \tilde{L}(x)]}{sp[x - \tilde{L}(x)]} = \dfrac{sp[T^n x - T^n\tilde{L}(x)]}{sp[x - \tilde{L}(x)]}, & \text{if } x \notin V \\ 0 & \text{otherwise} \end{cases} \tag{1.6.13}$$

since $sp[x - \tilde{L}(x)] = 0$ can be shown to occur only if $x \in V$ (cf. [111], lemma 2.2 part (h)) and where the equality in (1.6.13) follows from a repeated application of (1.6.4).

Note that for all $x \in \tilde{W}$, and $n = 1,2,\ldots, f_n(x) \leq 1$ and that $\{f_n(x)\}_{n=1}^{\infty}$ is monotonically non-increasing towards 0, such that there exists an integer $M(x) \geq 1$ with:

$$f_n(x) < 1 \quad \text{for all } n \geq M(x). \tag{1.6.14}$$

Next the *key* result in the geometric convergence proof is provided by (cf. [110], th. 4.1).

THEOREM 1.6.1. *There exists an integer* M^* *such that for all* $x \in \tilde{W}$:

$$f_{M^*}(x) < 1. \quad \square \tag{1.6.15}$$

Thus th. 1.6.1 expresses that M(x) the number of steps needed for contraction is bounded in $x \in \tilde{W}$.

For each $m = 1,2,\ldots$ and $x \in \tilde{W}$, let

$$h_m(x) = \sup_{n=1,2,\ldots} f_m(T^n x - ng^*) \leq 1 \tag{1.6.16}$$

where the inequality follows from $f_m(y) \leq 1$, $y \in \tilde{W}$, since by (1.6.4) we have $T^n x - ng^* \in \tilde{W}$ for all $n \geq 1$. The second part of the geometric convergence proof consists of showing that for all $x \in \tilde{W}$:

$$h_{M^*}(x) < 1. \tag{1.6.17}$$

We note that (1.6.17) is obtained by a detailed analysis of the U(v)-operator appearing in the third phase of the process. Further it was shown in [111] that for any $n \geq 1$ and $x \in \tilde{W}$:

$$\begin{aligned} sp[\tilde{e}(nM^*+r,x)] \leq sp[\tilde{e}(nM^*,x)] &\leq f_{M^*}(T^{(n-1)M^*}x-(n-1)M^*g^*)\, sp[\tilde{e}((n-1)M^*,x)] \\ &\leq h_{M^*}(x)\, sp[\tilde{e}((n-1)M^*,x)]; \quad r = 0,\ldots,M-1. \end{aligned} \tag{1.6.18}$$

Finally, some further analysis leads to the main result (cf. [111], th.4.2).

THEOREM 1.6.2. (Geometric convergence)
For all $x \in W$, *there exists a number* K(x) *such that*

$$\begin{aligned} &\|Q^n x - ng^* - L(x)\| \leq K(x)\{h_{M^*}(x)\}^{\lfloor n/M^* \rfloor} \\ &sp[Q^n x - ng^* - L(x)] \leq K(x)\{h_{M^*}(x)\}^{\lfloor n/M^* \rfloor} \end{aligned} \tag{1.6.19}$$

where $\lfloor x \rfloor$ *indicates the largest integer less than or equal to* x.

We observe that $h_{M^*}(x)$ does *not* represent the *ultimate convergence rate* or *ultimate average contraction factor per step*, which is defined by:

$$(1.6.20)\qquad \begin{cases} \lim\limits_{n\to\infty} f_n(x)^{1/n} = \lim\limits_{n\to\infty} \left\{ \dfrac{sp[\tilde{e}(n,x)]}{sp[\tilde{e}(0,x)]} \right\}^{1/n} & \text{for } x \in \tilde{W}\backslash V \\ \\ 0 & \text{for } x \in V. \end{cases}$$

It was shown in [110], section 6, that for all $x \in \tilde{W}$, the ultimate convergence rate may be bounded by

$$(1.6.21)\qquad \lambda^* \stackrel{def}{=} \max_{v\in V} \sup \left\{ \frac{sp[U(v)^{M^*}y]}{sp[y]} \;\middle|\; \lim_{n\to\infty} U(v)^n y = 0 \right\} < 1.$$

Observe that on the right hand side of (1.6.21) the maximum is taken over a *finite* number of distinct U(v)-operators. Note in addition that in the case of a single policy this reduces to the well-known fact that the convergence rate is bounded by the subdominant eigenvalue of the associated transition probability matrix (cf. also MORTON and WECKER [86], who found the same result in the special case of policy convergence, i.e. when there exists an integer $n_0(x)$ and a policy $f \in S_p$ such that:

$$(1.6.22)\qquad Q^n x = q(f) + P(f)Q^{n-1}x \quad \text{for all } n \geq n_0(x).$$

Whereas the ultimate convergence rate is bounded on $\tilde{W}$, the same does not necessarily hold for the *n-step contraction factor* $f_n(x)$ whatever the choice for $n = 1,2,\ldots$. That is, we may have:

$$(1.6.23)\qquad \sup_{x\in\tilde{W}} f_n(x) = 1 \quad \text{for all } n = 1,2,\ldots$$

as is illustrated by example 2 in chapter 2.

The problem of finding conditions which in all generality are both necessary and sufficient for the existence of a uniform n-step contraction factor for *some* $n = 1,2,\ldots$, has not been solved yet. However, under (UNI) the following necessary and sufficient condition was obtained in [111]:

(1.6.24) (UR) There exists a *randomized* policy $f \in S_R$ which has $\hat{R}$ as its single subchain.

Another topic of interest is the dependence of M^* on the size of the problem. Again, under (UNI) it was shown in [111], th. 5.2 that:

$$(1.6.25)\qquad M^* \leq N^2 - 2N + 2.$$

The upperbound was obtained by a combinatorial proof and is sharp up to a term of O(N) (cf. example 2 in [111]). The quadratic upperbound obviously represents the worst case behaviour, and contrasts with the fact that computational experience as reported e.g. in SU and DEININGER [120] and TIJMS [122] shows that (in most cases) $M^* = 1$ or 2.

1.7. UNDISCOUNTED CASE; ASYMPTOTIC BEHAVIOUR OF S(n) AND THE EXISTENCE OF INITIALLY STATIONARY OR PERIODIC ε-OPTIMAL STRATEGIES

As discussed earlier, separate treatment is given to

a) the case where $\lim_{n\to\infty} [v(n)-ng^*]$ exists and
b) the one where the sequence fails to converge.

We mentioned earlier that for the latter case an example was constructed in [35], in which S(n) lies outside S_{PMG} for every n. In this case one merely knows (cf. BROWN [13]) that for large n, $S(n) \subseteq X_i L(i)$. In the case where $v^* = L(v(0)) = \lim_{n\to\infty} v(n)-ng^*$, exists then for large n:

(1.7.1) $\quad S(n) \subseteq S^*(v^*) \subseteq S_{PMG} \subseteq X_i L(i)$.

Thus (1.7.1) shows that value-iteration settles upon maximal gain policies provided that convergence is guaranteed.

The explanation of this discrepancy with respect to the behaviour of S(n) between the case where $[v(n)-ng^*]$ converges and the one where it fails to converge, requires the notion of *multistep policies* and *periodic strategies*.

For each integer $J \geq 1$, a J-step policy is a J-tuple of policies $(f^{*(1)},\ldots,f^{*(J)})$ and specifies a J-periodic strategy

(1.7.2) $\quad \pi = (\ldots,f^{(\ell)},\ldots,f^{(1)})$

$$f^{(nJ+r)} = f^{*(r)} \qquad \text{for all } n = 0,1,\ldots \text{ and } r = 1,\ldots,J;$$

so, a J-step policy is called maximal gain, if the long run average return vector of the associated J-periodic strategy equals g^*. (1.7.1) holds for the special case where $\lim_{n\to\infty} v(n)-ng^*$ exists, i.e. the case where $J^0(v(0)) = 1$, and the following generalization for $J^0 \geq 2$ may be obtained (cf. [35]), (with J^0 having been defined by (1.5.6)):

(1.7.3) For all n large enough, each J^0-tuple of policies in $S(n+1) \times\ldots\times S(n+J^0)$ is maximal gain as a J^0-step policy.

Apparently a multistep-policy may be maximal gain, with each of the component-policies being *non*-maximal gain. Indeed a close investigation shows that the necessary and sufficient conditions for a multistep-policy to be maximal gain, reduce to the actions, prescribed by the component policies, being required to satisfy the optimality equations (1.4.14) and (1.4.15) only in a very special subset of Ω (cf. [35]).

The aforementioned example in [2] shows that even in the case where $v(n)-ng^*$ converges (in fact even in the case where each policy is unichained and aperiodic), $S(n)$ may have a very irregular behaviour, the worst case of which exhibits nonperiodic oscillations.

As a consequence we are only guaranteed to have an initially stationary (or periodic) strategy if $S^*(v^*)$ is a singleton where $v^* = \lim_{n\to\infty} v(n)-ng^*$. Using the geometric convergence result as discussed in section 6, we obtain however (cf. [35]) that for all $\varepsilon > 0$, there exists an initially periodic strategy which is ε-optimal. In fact, the (initial) period of this strategy may be taken to be equal to $J^0(v(0))$.

In particular, we see that in case $J^0 = 1$, i.e. when $v(n)-ng^*$ exists, an initially stationary ε-optimal strategy exists for all $\varepsilon > 0$, and in addition $S^*(v^*)$ represents the set of policies which can be used in the initially stationary part of the strategy. This generalizes LANERY [75] who established the above result for all $\varepsilon \geq$ (some) ε^*. When $J^0 \geq 2$, a similar characterization may be given for the set of J^0-step policies which can be used in the initially periodic part of any ε-optimal strategy. In addition, MDP's can be constructed in which there exist choices of $v(0)$ for which every initially J-periodic ε-optimal strategy (with ε small enough) has J as a multiple of J^0 (this result obviously doesn't hold for *every* MDP with $J^* \geq 2$ as is illustrated by the case where S contains a single periodic policy. Observe that unless condition (UNI) is met, and unlike the discounted case the best (or ε-best) choice for a current policy depends upon the terminal reward vector $v(0)$, whatever the length of the planning horizon.

Since this terminal reward vector may not be known (exactly) in advance, and since $S^*(v^*)$ may depend discontinuously upon $v(0)$, it would be desirable to choose a policy which lies in the intersection of the sets $\{S^*(v^*) \mid v^* \in V\}$. However, $\bigcap_{v\in V} S^*(v)$, which may be written as a *finite* intersection, *may be empty*.

In [109] it was shown that *convexity of* V is the necessary and sufficient condition for $\bigcap_{v\in V} S^*(v) \neq \emptyset$, i.e. for the existence of a policy which can be used in the initially stationary part of the ε-optimal strategy, *in*

complete independence of v(0). Moreover an example was provided in which convexity of V fails to hold. Sufficient conditions for convexity of V are given by (cf. [109], th. 4.3):
(1) $R^* = \Omega$; (2) K(i) is a singleton for all $i \in \Omega \backslash R^*$; (3) (UNI).

It is worthwhile observing that in some cases a (Blackwell-) optimal policy, i.e. a policy which is optimal in the discounted model for all β sufficiently close to 1 (cf. BLACKWELL [10]) cannot be used in the initially stationary part of the ε-optimal policy (cf. [35]).

In the unichained case, i.e. when all policies are unichained, an explicit upperbound may be derived for m(v(0)), the length of the non-stationary (or non-periodic) tail of the ε-optimal strategy; the latter being due to the existence of bounds for the distance between v^* and the relative value vector of a policy in S(n) (cf. section 8 and chapter 2).

However, in the general multichain case and unlike the discounted model no bounds have been obtained as yet for m(v(0)). In analogy with the discounted case, m(v(0)) can however *in all generality* be shown to vary logarithmically with the precision ε. For the case of continuous time Markov Decision Problems, in which no periodicity problems arise, some of the above results were obtained by LEMBERSKY [76] and [77].

Finally, several difficulties appear when trying to find the set S_{PMG}. First for all $v \in V$, $S^*(v)$ can be a *strict* subset of S_{PMG} so that value-iteration fails to yield *all* maximal gain policies. Indeed even $U_{v^* \in V} S^*(v^*)$ can be a strict subset of S_{PMG} so that varying the starting point v(0) of value-iteration will fail to identify all maximal-gain policies. The explanation is that a maximal gain policy f is merely required to choose its actions within L(i), for those states that are transient under P(f) (cf. lemma 1.4.2).

The second difficulty is provided by the irregular behaviour of the sets $\{S(n)\}_{n=1}^{\infty}$ as described above. This difficulty can however be overcome in a way similar to the one employed for the discounted model. Let $\{\varepsilon_n\}_{n=1}^{\infty}$ be a sequence of positive numbers approaching 0, at a slower rate than the (geometric) convergence rate of $[v(n)-ng^*]$. That is, let $\lim_{n\to\infty} \varepsilon_n/\lambda^n = \infty$, e.g. by taking $\varepsilon_n = n^{-1}$.

THEOREM 1.7.1. *Assume* $v^* = \lim_{n\to\infty} v(n)-ng^*$ *exists.*

(a) $\lim_{n\to\infty} S(n,\varepsilon_n) = S^*(v^*)$.

(b) *Let* $g(n) = v(n) - v(n-1)$, *and define for all* $\varepsilon > 0$:
$G(n,\varepsilon) = \{f \in S \mid P(f)g(n) \geq g(n+1) - \varepsilon \underline{1}\}$.

Then, $\lim_{n\to\infty} G(n,\varepsilon_n) = X_i L(i)$.

PROOF.

(a) Use $\lim_{n\to\infty} \varepsilon_n = 0$, as well as (1.6.8) to verify that $S(n,\varepsilon_n) \subseteq S^*(v^*)$ for all n sufficiently large. To prove the reversed inclusion, fix $f \in S^*(v^*)$ and note that $q(f) + P(f)v(n) = v^* + (n+1)g^* + P(f)e(n) = v(n+1) - e(n+1) + P(f)e(n) = v(n+1) + 0(\lambda^n) \geq v(n+1) - \varepsilon_n 1$ for all n sufficiently large, where $e(n) = v(n) - ng^* - v^*$ and where we use $g^* = P(f)g^*$ and $g^* + v^* = q(f) + P(f)v^*$.

(b) Use $\lim_{n\to\infty} \varepsilon_n = 0$, as well as $\lim_{n\to\infty} g(n) = g^*$ to conclude that $G(n,\varepsilon_n) \subseteq X_i L(i)$ for all n sufficiently large. To prove the reversed inclusion, fix $f \in X_i L(i)$ and note that $P(f)g(n) = P(f)[g(n)-g^*] + g(n+1) + [g^* - g(n+1)] = g(n+1) + 0(\lambda^n) \geq g(n+1) - \varepsilon_n 1$, for all n sufficiently large.

Table 1 concludes this section by summarizing some of the main results of section 3 and section 7.

Table 1.

I	II	III	
< 1	1	1	β
-	1	≥ 2	$J^0(v(0))$
$\lim_{n\to\infty} v(n) = v^*$	$\lim_{n\to\infty} v(n)-ng^* = v \in V$	$\lim_{n\to\infty} v(nJ+r)-(nJ+r)g^*$ exists, iff J is a multiple of $J^0(v(0))$	asymptotic behaviour of $\{v(n)\}_{n=1}^{\infty}$
geometric	geometric	geometric	rate of convergence
yes	yes	no, only J^0-tuples of consecutively generated policies need to be maximal gain as multi-step policies	policies generated need to be optimal after finite number of iterations
no	no	no	stationary or periodic, optimal strategies need to exist
yes	yes	no, only ε-optimal J^0-periodic strategies need to exist	ε-optimal stationary strategies need to exist.

1.8. UNDISCOUNTED CASE; ALGORITHMS AND SOME DATA-TRANSFORMATIONS

In this section we show which successive approximation methods can be used in order to find maximal gain policies and the maximal gain rate vector. For the schemes that are based upon pure value-iteration the convergence results obviously follow from the study of the asymptotic behaviour of $\{v(n)\}_{n=1}^{\infty}$. In sections 5 and 7, we observed that only in case $\{v(n)-ng^*\}_{n=1}^{\infty}$ converges, will value-iteration be guaranteed to ultimately settle upon maximal gain policies and only then, can sequences be derived from $\{v(n)\}_{n=1}^{\infty}$ which converge to g^* and some $v \in V$.

In the case where $\{v(n)-ng^*\}_{n=1}^{\infty}$ may fail to converge for some $v(0) \in E^N$ i.e. whenever $J^* = 1$ is not guaranteed by the structure of the problem, the following alternatives can be used:

A) Elimination of the periodicities using the following data-transformation:

$$(1.8.1) \qquad \tilde{q}_i^k = \sigma\, q_i^k; \quad i \in \Omega,\ k \in K(i)$$

$$(1.8.2) \qquad \tilde{p}_{ij}^k = \tau(p_{ij}^k - \delta_{ij}) + \delta_{ij}, \quad 1 \le i, j \le N \text{ and } k \in K(i)$$

where $\sigma > 0$ and $0 < \tau < 1$, and with δ_{ij} denoting the Kronecker delta function, i.e. $\delta_{ij} = 0$ for $i \neq j$, and $\delta_{ij} = 1$ for $i = j$. This transformation makes all of the diagonal elements of all of the tpm's strictly positive such that in the transformed model all of the policies are aperiodic.

Moreover, the transformation turns the MDP into an *equivalent* one, in the sense that it has the same state- and policy space and that each policy f has $\tilde{g}(f) = \sigma\, g(f)$ as its gain rate vector and $\tilde{V} = \{v \in E^N \mid \sigma^{-1}\tau\, v \in V\}$ as the set of solutions to the corresponding optimality equation (1.4.15), as is shown in the next lemma. (cf. also (EQUI) in (1.9.2)):

LEMMA 1.8.1. (cf. SCHWEITZER [108])

(a) $\tilde{V} = \{v \in E^N \mid \sigma^{-1}\tau\, v \in V\}$.

(b) *For all* $f \in S_R$, $\tilde{g}(f) = \sigma\, g(f)$; *hence* $\tilde{g}^* = \sigma\, g^*$, $\tilde{S}_{RMG} = S_{RMG}$ *and* $\tilde{S}_{PMG} = S_{PMG}$ *represent the maximal gain rate vector, and the sets of randomized and pure maximal gain policies in the transformed model.*

PROOF: Rewrite the optimality equations (1.4.14) and (1.4.15) in a homogeneous way, i.e.

$$(1.8.3) \quad 0 = \max_{k\in K(i)}[\sum_j (P_{ij}^k - \delta_{ij})g_j], \quad i \in \Omega$$

$$(1.8.4) \quad 0 = \max_{k\in L(i)}[q_i^k - g_i + \sum_j (P_{ij}^k - \delta_{ij})v_j], \quad i \in \Omega.$$

Note that the solution space of the optimality equations remains unaltered when multiplying the expressions within brackets in (1.8.3) and (1.8.4) by $\sigma\tau > 0$ and $\sigma > 0$ resp. Hence,

$$(1.8.5) \quad 0 = \max_{k\in K(i)}[\sum_j \tau(P_{ij}^k - \delta_{ij})(\sigma g_j)], \quad i \in \Omega$$

$$(1.8.6) \quad 0 = \max_{k\in L(i)}[\sigma q_i^k - \sigma g_i + \sum_j \tau(P_{ij}^k - \delta_{ij})(\sigma\tau^{-1}v_j)], \quad i \in \Omega$$

such that in case (g,v) satisfies the optimality equations of the original model, then $(\sigma g, \sigma\tau^{-1}v)$ will satisfy the corresponding equations in the transformed model; and vice versa in case $(\tilde{g},\tilde{v})$ satisfies (1.8.5) and (1.8.6), then $(\sigma^{-1}\tilde{g}, \sigma^{-1}\tau\tilde{v})$ is a solution pair to the optimality equations in the original model.

(b) Apply the proof of part (a) to the system of equations $g = P(f)g$; $v = q(f) - g + P(f)v$, to verify that if (g,v) is a solution to this system, then $(\sigma g, \sigma\tau^{-1}v)$ will satisfy the corresponding system in which $P(f)$ and $q(f)$ are replaced by $\tilde{P}(f)$ and $\tilde{q}(f)$. Since the "g-" part of the solution to these systems is uniquely determined by the gain rate vector, (cf. lemma 1 in [22]) it follows that $\tilde{g}(f) = \sigma g(f)$ represents the gain rate vector of f in the transformed model. □

The above presented transformation will play an important role throughout this entire thesis, especially so for the choices $\sigma = 1$ and $\sigma = \tau$, since the first case has $\tilde{g}^* = g^*$ and the second one has $\tilde{V} = V$.

Due to the obtained aperiodicities $\{\tilde{v}(n) - n\tilde{g}^*\}_{n=1}^{\infty}$ converges (geometrically) in the transformed model for whatever choice of $\tilde{v}(0) \in E^N$ (cf. th.1.5.2, condition (IV)).

B) The *modified value-iteration technique* by HORDIJK and TIJMS [60]. This scheme has the discount factor β in (1.1.1) depending upon the index of the iteration stage, and tending to one as the index tends to infinity:

$$(1.8.7) \quad u(n+1)_i = \max_{k\in K(i)}\{q_i^k + \beta_n \sum_j P_{ij}^k\, u(n)_j\}, \quad i \in \Omega$$

where $u(0)$ is a given N-vector.

The scheme can only be used when

(1.8.8) $g^* = \langle g^* \rangle \underline{1}$.

In this case

(1.8.9) $u(n) - \gamma_n g^* \to z^* \in V \quad \text{as } n \to \infty$

where $\{\gamma_n\}_{n=1}^{\infty}$ is obtained recursively by

$$\gamma_{n+1} = 1 + \beta_n \gamma_n \quad \text{for } n \geq 0 \text{ with } \gamma_0 = 0$$

provided that

(a) $\beta_n \beta_{n-1} \cdots \beta_1 \to 0$

(b) $\sum_{j=1}^{n} \beta_n \cdots \beta_{j+1} |\beta_j - \beta_{j-1}| \to 0$

(a) and (b) essentially express that $\{\beta_n\}_{n=1}^{\infty}$ should increase to one at a low enough rate, and a computationally tractible choice is provided by

(1.8.10) $\beta_n = 1 - n^{-b} \quad \text{with } 0 < b \leq 1.$

The analysis of the behaviour of this scheme uses the Laurent series expansion of the total maximal discounted return vector in powers of $(1-\beta)$ for discountfactors β that are close enough to one (cf. MILLER and VEINOTT [85]).

The scheme eventually settles upon maximal gain policies, and with the choice (1.8.10) it can be shown that the ultimate convergence rate is $0(n^{-b} ln\ n)$ which is substantially slower than the geometric convergence rate we obtained for the ordinary value-iteration scheme (cf. also th.4.3.3 in chapter 4, where a generalization of this scheme is given).

However the scheme has two very nice characteristics:

(1) convergence occurs regardless of the chain- and periodicity structure of the problem.

(2) For every starting point $u(0) \in E^N$ the scheme converges to the same limit vector z^* which has the following very important interpretation:

(1.8.11) $z_i^* = \max_{f \in S_{PMG}} z(f)_i, \quad i \in \Omega$

where $z(f) = Z(f)[q(f) - g^*]$.

That is, z^* is the *optimal bias-vector*, where the biasvector $z(f)$ of a policy $f \in S_p$ is the *second* term in the Laurent series expansion of the total discounted return vector $V(f,\beta)$:

$$(1.8.12) \quad V(f,\beta) = \frac{g(f)}{1-\beta} + z(f) + O(1-\beta), \quad \beta \to 1$$

(cf. BLACKWELL [10] and MILLER and VEINOTT [85]).

The HORDIJK-TIJMS scheme, however, does not necessarily settle upon bias optimal policies i.e. policies which attain the N maxima in (1.8.11) simultaneously (cf. chapter 4).

We next review the bounds that have been derived for the maximal gain rate vector g^*. In the case where $g^* = \langle g^* \rangle \underline{1}$, which holds e.g. if all of the policies are unichained, these were obtained by ODONI [89] and HASTINGS [53] namely:

$$(1.8.13) \quad [Qx-x]_{min} \leq g(f)_i \leq \langle g^* \rangle \leq [Qx-x]_{max}$$

for all $x \in E^N$ and $i = 1,\ldots,N$ and f achieving Qx.

Moreover both bounds are sharp when $x \in V$. In the context of value-iteration (1.8.13) becomes

$$(1.8.14) \quad [v(n+1)-v(n)]_{min} \leq \langle g^* \rangle \leq [v(n+1)-v(n)]_{max}.$$

The bounds move inward (monotonically) as n increases and if $\lim_{n\to\infty} v(n)-ng^*$ exists, the bounds both converge geometrically fast to $\langle g^* \rangle$.

In the context of the approach under B), the bounds on g^* have to be altered as follows (cf. [60]): For $\ell = 1,2,\ldots$

$$(1.8.15) \quad \min_i \{w(\ell)_i - \beta_\ell w(\ell-1)_i\} \leq g(f_\ell)_i \leq g^*_i \leq \max_i \{w(\ell)_i - \beta_\ell w(\ell-1)_i\}$$

where f_ℓ is any policy which attains the N maxima at the ℓ-th iteration stage of (1.8.7). Again, whenever $g^*_i = \langle g^* \rangle$, $i \in \Omega$, will the outer bounds in (1.8.15) converge to $\langle g^* \rangle$.

Under (UNI) the bounds on the scalar gain rate $\langle g^* \rangle$ have been unaccompanied by corresponding bounds on the deviation of the current vector x from $v^* \in V$, which in this case is unique up to a multiple of $\underline{1}$. In view of the latter, this bound should be invariant to a replacement of x by $x + a\underline{1}$ for some scalar a. The existence of such bounds is also useful for demonstrating *convergence* of this or related types of value-iteration

schemes. Specifically, ZANGWILL [133] has shown that an iterative scheme $x(n+1) = Ax(n)$ will converge to x^* if the continuous operator A and a continuous (Lyapunov) function $\phi(x)$ satisfy:

(1.8.16) (a) $\phi(x) \geq 0$ all $x \in E^N$

(b) $\phi(x) = 0$ if and only if $x = x^*$

(c) $\phi(Ax) \leq \phi(x)$ all $x \in E^N$

(d) for some integer $m \geq 1$, $\phi(A^m x) < \phi(x)$, for all x with $\phi(x) > 0$.

One choice of a Lyapunov function, not computable until v^* is known, is

$$\phi_1(x) = sp[x-v^*], \text{ with}$$

(1.8.17) $$Ax = Qx-(Qx)_N\underline{1}$$
$$x^* = v^* - v^*_N\underline{1}.$$

Condition (1.8.16) (d) may be verified as the scalar gain rate version of (1.6.19), with $m = N^2 - 2N + 2$ (see above), assuming that $J^* = 1$.

Another choice of Lyapunov function which may be computed while in the midst of the value-iteration process is

(1.8.18) $$\phi_2(x) = sp[Qx-x]$$

with the same choice of A and x^*. The conditions (1.8.16) (a)-(c) are easily verified while (1.8.16) (d) holds e.g. when every policy in S_{PMG} is unichained, and assuming that the data-transformation (1.8.1) and (1.8.2) has been applied so as to ensure that $J^* = 1$ (cf. [39]).

The important *new* property is that the deviation of v^* from x may be deduced from $\phi_2(x)$ just as (1.2.6) and (1.2.11) were used in the *discounted* case. Specifically, under (UNI) there exists a constant $\rho \geq 0$ such that

(1.8.19) $\frac{1}{2}\phi_2(x) \leq sp[x-v^*] \leq \rho\phi_2(x)$ for all x if and only if (UR) holds (cf. (1.6.24))

Under (UNI) a unique representation v^* of $v \in V$ can be obtained by requiring that $v^*_N = 0$. So far, bounds for each of the components of v^* have only been obtained for the case where every policy is unichained. The bounds will be derived in chapter 2, and arise by showing that the MDP can be transformed into an equivalent one, in which the operator $\hat{Q}$, defined by $\hat{Q}x = Qx - [Qx]_N\underline{1}$ is a N-step contraction operator, on $\hat{E}^N = \{x \in E^N \mid x_N = 0\}$.

The bounds are of the same type as in (1.2.11) and allow for the derivation (although not for the actual computation *prior* to solving the MDP) of upper-bounds on the number of iterations needed to have

(1) $\hat{Q}^n x$ within ε of v^*,

(2) $S(n) \leq S_{PMG}$,

(3) v^* as a relative value vector $v(f)$ for every policy f in $S(n)$, i.e. $S(n) \leq S^*(v^*)$, as well as on

(4) the length of the tail of an initially stationary (periodic) ε-optimal strategy.

All of the bounds in (1)-(4) vary logarithmically with ε^{-1} $sp[Qx-x]$ where in (2), ε has to be taken $\leq \min\{sp[g^*-g(f)] \mid f \in S_p,\ g^* > g(f)\}$ and in (3), ε has to be taken $\leq \min\{sp[v^*-v(f)] \mid f \in S_p,\ sp[v^*-v(f)] > 0\}$. Except for the case where every policy is unichained, (cf. chapter 2) and due to the lack of bounds on $v \in V$, no tests have been proposed for *permanent* elimination of non-optimal actions. However, a device for *temporary* elimition was recently obtained in HASTINGS [54].

Another open question is obtaining a computationally tractible estimate of the size of λ^*. Nothing is known with the exception of the above mentioned case where S_{PMG} is a singleton and the cases studied by WHITE [131] and ANTHONISSE and TIJMS [1] where a n-step generalization of the ergodic- (or scrambling-) coefficient provides an upperbound for λ^*.

A further problem arises both in approach A) and approach B) due to the fact that the sequences generated ($\{v(n)\}_{n=1}^{\infty}$ and $\{u(n)\}_{n=1}^{\infty}$) diverge linearly with n. That is, one has to do computations with numbers that grow linearly with the number of stages needed to come within the required precision.

In case $g^* = \langle g^* \rangle \underline{1}$ the problem can be eliminated using White's procedure, i.e. in approach A) we generate

$$(1.8.20) \quad \hat{v}(n)_i = v(n)_i - v(n)_N = Qv(n-1)_i - Qv(n-1)_N.$$

Then $\hat{v}(n) \to L(v(0)) - \langle L(v(0))_N \rangle \underline{1} \in V$, and $Q\hat{v}(n)_N \to \langle g^* \rangle$, as $n \to \infty$.

In the general multichain case where $g^* = \langle g^* \rangle \underline{1}$ fails to hold, only approach A) needs to be considered. The only thing that comes to mind when trying to eliminate the above mentioned difficulty is the following: Write $v(n) = ng(n) + y(n)$, with

$$(1.8.21) \quad g(n) = v(n) - v(n-1)$$

(1.8.22) $y(n) = nv(n-1) - (n-1)v(n)$.

Observe that the sequence $\{g(n)\}_{n=1}^{\infty}$ and $\{y(n)\}_{n=1}^{\infty}$ converge to g^* and $L(v(0))$ whenever $L(v(0)) = \lim_{n\to\infty} v(n) - ng^*$ exists. Note in addition that $g(n)$ and $y(n)$ can be generated from the schemes:

$$(1.8.23)\quad g(n+1)_i = \max_{k\in K(i)} \{q_i^k + n \sum_j (P_{ij}^k - \delta_{ij}) g(n)_j + \sum_j (P_{ij}^k - \delta_{ij}) y(n)_j\}$$

$$(1.8.24)\quad y(n+1)_i = y(n)_i + n[g(n)_i - \max_{k\in K(i)} \{q_i^k + n \sum_j (P_{ij}^k - \delta_{ij})(y(n)_j + ng(n)_j)\}], \quad i \in \Omega.$$

By generating (1.8.23) and (1.8.24) only two *bounded* sequences of numbers have to be *stored*. Unfortunately, however, this solves our numerical difficulty only partially, since it is still necessary to do *computations* with unbounded terms when determining the right hand sides of (1.8.23) and (1.8.24).

In some cases one may be interested in obtaining (as large as possible a subset of) the entire set S_{PMG}, so as to make further selections on the basis of additional criteria.
In section 7 we discussed the irregularities that may appear in the sequences of policies generated by the value iteration method (and which are identical both in approach A) and B)).
Th.1.7.1. showed that $S^*(L(v(0)))$ can be obtained by keeping track of the sets $\{S(n,\varepsilon_n)\}_{n=1}^{\infty}$, and in approach B), $S^*(z^*)$ can be computed in exactly the same way, provided $\{\varepsilon_n\}_{n=1}^{\infty}$ is chosen to decrease to 0 at a rate which is slower than the convergence rate of $\{u(n)\}_{n=1}^{\infty}$. That is, with the choice (1.8.10) for $\{\beta_n\}_{n=1}^{\infty}$, choose

$$(1.8.25)\quad \lim_{n\to\infty} (\ln n)^{-1} n^b \varepsilon_n \to \infty$$

i.e. take e.g. $\varepsilon_n = n^{-b/2}$.

1.9. MARKOV RENEWAL PROGRAMS

In this section, we consider the more general class of Markov Renewal Programs (cf. [23], [69]) in which the times between two successive transitions of state are random variables, whose distributions depend both on

the current state and the action chosen. Let $\tau_{ij}^k \geq 0$ for $i,j \in \Omega$; $k \in K(i)$ denote the *conditional* expected holding time in state i, given the action $k \in K(i)$ is chosen and that state j is the next state to be observed. We assume that the *unconditional* expected holding times:

$$(1.9.1) \quad T_i^k = \Sigma_j P_{ij}^k \tau_{ij}^k > 0 \quad (i \in \Omega;\ k \in K(i))$$

For each policy $f \in S_R$, q(f) and P(f) are defined as in section 2, whereas $g(f)_i$ denotes again the long run average return per unit time, when starting in state i. We recall that g(f) is given by (cf. e.g. lemma 1 in DENARDO [22]):

$$(1.9.2) \quad g(f)_i = \sum_{m=1}^{n(f)} \Phi_i^m(f)\, g^m(f), \quad i \in \Omega$$

with

$$g^m(f) = <\pi^m(f), q(f)> / <\pi^m(f), T(f)>.$$

Next, we define for each policy $f \in S_R$, the holding time vector T(f):

$$(1.9.3) \quad T(f)_i = \sum_{k \in K(i)} f_{ik}\, T_i^k.$$

Finally we call to mind that in this model the optimality equations (1.4.14) and (1.4.15) have to be altered as follows:

$$(1.9.4) \quad g_i = \max_{k \in K(i)} \sum_j P_{ij}^k\, g_j, \quad i \in \Omega$$

$$(1.9.5) \quad v_i = \max_{k \in L(i)} \{q_i^k - \sum_j P_{ij}^k\, \tau_{ij}^k\, g_j + \sum_j P_{ij}^k\, v_j\}, \quad i \in \Omega$$

with L(i) defined as in (1.4.16). In addition the vector g^* and the sets S_{PMG} and S_{RMG} are defined as in section 4, where the non-emptyness of these sets in the MRP-model was shown in [69]. Again (1.9.4) and (1.9.5) always have a solution pair, and again each solution pair (g,v) has $g = g^*$, the maximal gain rate vector (cf. [23], and [109]). Redefine $V = \{v \in E^N \mid (g^*, v)$ satisfy (1.9.4) and (1.9.5)$\}$. The properties of V, and the correspondence between S_{RMG} and V, as mentioned in section 4 hold unaltered for the general MRP case.

Both the Policy Iteration Algorithm and the Linear Programming Approaches which were originally developed for MDP's have been adapted for the more general MRP-model (cf. e.g. [23]). To obtain a successive approximation method for undiscounted MRP's, *two* data-transformations have to be

applied:

We first observe from (1.9.2) that the gain rate vectors $g(f)$ depend on the quantities τ_{ij}^k only through the *unconditional* holding times T_i^k. As a consequence, we conclude that every MRP is transformed into an equivalent one, by replacing $\tilde{\tau}_{ij}^k = T_i^k$ $(i,j \in \Omega;\ k \in K(i))$. In this context we define two undiscounted MRPs to be equivalent, if

(1.9.6) (EQUI) they have the same state and action spaces, and if the gain rate vector of any policy (and hence the maximal gain rate vector) in the two models merely differs by a multiplicative constant.

Note that two equivalent MRPs share the same set of maximal gain policies. Carrying out the above tranformation, we obtain the following pair of optimality equations:

$$(1.9.7)\qquad g_i = \max_{k\in K(i)} \sum_j p_{ij}^k\, g_j, \qquad i \in \Omega$$

$$(1.9.8)\qquad v_i = \max_{k\in L(i)} \{q_i^k - T_i^k g_i + \sum_j p_{ij}^k\, v_j\}, \qquad i \in \Omega.$$

Let $\hat{V}$ be the set of solutions to (1.9.8).

Next, we recall the following generalization of the data-transformation (1.8.1) and (1.8.2) (cf. [108], [38]):

$$(1.9.9)\qquad \begin{aligned} \tilde{q}_i^k &= \sigma\, q_i^k / T_i^k, && i \in \Omega,\ k \in K(i)\\ \tilde{p}_{ij}^k &= \delta_{ij} + \tau(p_{ij}^k - \delta_{ij})/T_i^k, && i \in \Omega,\ k \in K(i)\\ \tilde{T}_i^k &= 1, && i \in \Omega\ \ k \in K(i) \end{aligned}$$

where $\sigma > 0$ and τ has to be chosen such that

$$(1.9.10)\qquad 0 < \tau \le \min\{T_i^k/(1-p_{ii}^k) \,|\, (i,k) \text{ with } p_{ii}^k < 1\}.$$

Using the proof of lemma 1.8.1, one verifies that again $\tilde{V} = \{v \in E^N \,|\, \sigma^{-1}\tau\, v \in \hat{V}\}$ is the set of solutions to the optimality equation (1.4.15) in the transformed MDP, and that every policy f has $\tilde{g}(f) = \sigma g(f)$ as its gain rate vector in the transformed model, i.e. the original MRP and the transformed MDP are equivalent (cf. (1.9.6)). Hence, the choice $\sigma = \tau$ leads again to $\tilde{V} = \hat{V}$ and $\tilde{g}^* = \tau g^*$ and the choice $\sigma = 1$ leads to $\tilde{V} = \{v \in E^N \,|\, \tau\, v \in V\}$ and $\tilde{g}^* = g^*$. By choosing τ strictly less than the upperbound in (1.9.10) the same

transformation ensures, that every policy in the transformed MDP is *aperiodic*, such that the value-iteration method is guaranteed to converge for any starting point, with all of the nice consequences that were exhibited above.

As a consequence, applying value-iteration to the transformed model will yield us the maximal gain rate vector, maximal gain policies as well as a solution to the optimality equations (1.9.7) and (1.9.8). However, it won't be possible in the general multichain case to find a solution to the optimality equations (1.9.4) and (1.9.5) of the original MRP-model, using a *single* successive approximation scheme.

This is due to the fact that there does not need to exist a clear relationship between the sets V and $\hat{V}$ (as opposed to the one pointed out between $\hat{V}$ and $\tilde{V}$). Only in case $g^* = \langle g^* \rangle \underline{1}$, do the pairs of equations (1.9.4)-(1.9.5) and (1.9.7)-(1.9.8) and hence the sets V and $\hat{V}$ need to coincide (cf. (1.9.1)), as is pointed out by the following example:

Example 3.

i	k	p^k_{i1}	p^k_{i2}	p^k_{i3}	q^k_i
1	1	1	0	0	-1
2	1	5	0	.5	0
	2	0	1	0	0
3	1	0	0	1	1

$\tau^k_{ij} = 1$, except for $\tau^1_{21} \neq 1 \neq \tau^1_{23}$ with $.5(\tau^1_{21}+\tau^3_{23}) = 1$.

Note that $g^* = (-1,0,1)$; $K(i) = L(i)$ for all $i \in \Omega$ and $R^* = \Omega$. Verify that V and $\hat{V}$ are given by the half spaces:

$$V = \{v \in E^3 \mid v_2 \geq 0.5(v_1+v_3)\},$$
$$\hat{V} = \{v \in E^3 \mid v_2 \geq 0.5(v_1+v_3) + 0.5\tau^1_{21} - 0.5\tau^1_{23}\}.$$

In case $\tau^1_{21} < \tau^1_{23}$ V is a strict subset of $\hat{V}$ and vice versa for the case $\tau^1_{21} > \tau^1_{23}$. Note that V and $\hat{V}$ do not even need to coincide in the components that lie within R^*, as state 2 belongs to R^*.

Whereas in general, no method has been obtained to find a solution $v \in V$ via a *single* successive approximation scheme, it will be shown in chapter 4 that this objective can be achieved, employing a *pair* of simultaneously generated schemes.

CHAPTER 2

Contraction mappings underlying undiscounted Markov decision problems

2.1. INTRODUCTION AND SUMMARY

We pointed out in chapter 1, that the value-iteration operator Q in undiscounted MDPs, is *non-expansive* (cf. (1.4.1)) and in addition has the property (1.4.2):

(2.1.1) $\quad Q(x+c\underline{1}) = Qx+c\underline{1}, \quad$ for all $x \in E^N$ and $c \in E^1$.

In view of (2.1.1) Q can never be a contraction mapping on E^N, or a J-step contraction mapping for some integer $J \geq 1$, where the latter is defined as follows (cf. e.g. DENARDO [20]).

(2.1.2) Let X be a normed vector space; an operator $A: X \to X$ is a J-step contraction operator, if and only if there exists a scalar ρ, $0 < \rho \leq 1$ such that for all $x, y \in X$: $|A^J x - A^J y| \leq (1-\rho)|x-y|$, where $|\ \ |$ is the norm on X.

The fact that Q can never be (J-step) contracting on E^N (for some $J \geq 1$) may e.g. be verified by noting that the operator never has a *unique fixed point* in view of (2.1.1).

This constrasts with what is known to be the case in the substochastic or discounted case where $\sum_j P_{ij}^k < 1$ ($i \in \Omega$, $k \in K(i)$) (cf. section 2 of chapter 1, as well as DENARDO [20]).

It should be pointed out that the fact whether an operator A, as defined in (2.1.2) is J-step contracting for some $J = 1,2,\ldots$, is independent of the norm chosen on X as may easily be verified using the fact that any two norms $|x|$ and $|x|'$ are equivalent in the sense that there exist finite constants K and K' such that $|x| \leq K|x|'$ and $|x|' \leq K'|x|$ for all $x \in E^N$ (cf. COLLATZ [15], §9.2). (2.1.1) suggests considering the following equivalence relation on the N-dimensional Euclidean space E^N:

(2.1.3) $x \sim y \iff$ there exists a scalar c such that $x = y+c\underline{1}$.

Let $\bar{E}^N$ be the quotient space which is generated by this equivalence relation, and note that $\bar{E}^N$ is a (N-1-dimensional) vector space, with the conventional addition and scalar multiplication. Note that the sp[ ]-norm, defined by (1.2.16) which is a quasi-norm on E^N, is a real norm on $\bar{E}^N$. As a consequence we endow $\bar{E}^N$ with this sp[ ]-norm.

Let $\bar{Q}: E^N \to \bar{E}^N$ denote the reduced value-iteration operator i.e. $\bar{Q}x$ denotes the (unique) representation of Qx within $\bar{E}^N$. Example 1 below shows that $\bar{Q}$ is sometimes a contraction mapping on $\bar{E}^N$, or in other words Q may be contracting with respect to the sp[ ]-"norm". On the other hand, example 2 shows that *the combination of the (UNI)-condition* (which is the necessary and sufficient condition for $v \in V$ to be unique up to a multiple of $\underline{1}$) *and the (GC)-condition* (which is the necessary and sufficient condition for $\{Q^n x - ng^*\}_{n=1}^{\infty}$ to converge for *all* $x \in E^N$) is in itself insufficient for $\bar{Q}$ to be (J-step) contracting (for some $J \geq 1$).

We first define for any N×N-matrix A:

(2.1.4) $\|A\| = \max_i \sum_j |A_{ij}|$.

Also, for any real number a, define $a^+ = \max(a,0)$ and $a^- = \min(a,0)$, (with $a^+ \geq 0$, $a^- \leq 0$ and $a^+ + a^- = a$ and $a^+ - a^- = |a|$).

EXAMPLE 1. Let $S = \{f\}$ where P(f) is unichained and aperiodic. Verify that

$$sp[Q^n x - Q^n y] = sp[P(f)^n (x-y)] = sp[(P(f)^n - \Pi(f))(x-y)] \leq$$

$\|P(f)^n - \Pi(f)\| sp[x-y] \leq K\lambda^n sp[x-y]$, for some $K > 0$ and $0 \leq \lambda < 1$. The second equality follows from $\Pi(f)(x-y)$ being a multiple of $\underline{1}$ and the second inequality may e.g. be found on p.131 in [67], whereas the first inequality follows from the property

(2.1.5) $sp[Ax] \leq \|A\| sp[x]$, for any matrix $A = [a_{ij}]$ with $A\underline{1} = \underline{0}$; $x \in E^N$

To verify (2.1.5) note, using the identity $\sum_j a_{ij}^+ = -\sum_j a_{ij}^-$, $i \in \Omega$ which follows from $A\underline{1} = 0$, that:

$$\begin{aligned} sp[Ax] &= \max_i\{\textstyle\sum_j a_{ij}^+ x_j + \sum_j a_{ij}^- x_j\} - \min_i\{\sum_j a_{ij}^+ x_j + \sum_j a_{ij}^- x_j\} \\ &\leq \max_i\{\textstyle\sum_j a_{ij}^+ x_{max} + \sum_j a_{ij}^- x_{min}\} - \min_i\{\sum_j a_{ij}^+ x_{min} + \sum_j a_{ij}^- x_{max}\} \\ &= sp[x]\{\max_i \textstyle\sum_j a_{ij}^+ - \min_i \sum_j a_{ij}^-\} = 2sp[x] \max_i \sum_j a_{ij}^+. \end{aligned}$$

Let $\sum_j a^+_{kj} = \max_i \sum_j a^+_{ij}$. Then, $2 \max_i \sum_j a^+_{ij} = \sum_j a^+_{kj} - \sum_j a^-_{kj} = \sum_j (a^+_{kj} - a^-_{kj}) = \sum_j |a_{kj}| \leq \|A\|$ (cf. (2.1.4)) which completes the proof of (2.1.5).

As an even sharper result, one can show that in this case $\bar{Q}$ is J-step contracting for some $J \leq \frac{1}{2}N(N-1)$ as a result of $P(f)^n$ being *scrambling* for all $n \geq \frac{1}{2}N(N-1)$ (cf. th.4.4 on p.89 in SENETA [113]), where the scrambling notion and its implications will be discussed in section 3.

EXAMPLE 2.

i	k	p^k_{i1}	p^k_{i2}	q^k_i
1	1	1	0	0
2	1	1	0	0
	2	0	1	-1

$g^* = [0,0]$, hence $K(i) = L(i)$ for all $i \in \Omega$.

Note that (GC) (cf. th.1.5.1) is satisfied in view of every policy being aperiodic (cf. th.1.5.2 cond. (IV)). In addition, it is directly verified that $V = \{c\underline{1} \mid c \in E^1\}$ which implies that (UNI) is satisfied as well (cf. (1.4.20)). Take $x = [0,X]$ and $y = 0$. Note that

$$Q^n x = [0, \max(0,X-n)] \text{ and } Q^n y = 0 \quad \text{for all } n = 0,1,2,\ldots \text{ i.e.}$$

$$1 \geq \sup\left\{\frac{sp[Q^n u - Q^n v]}{sp[u-v]} \;\middle|\; u,v \in E^N,\ sp[u-v] > 0\right\} \geq$$

$$\geq \lim_{X\to\infty} \frac{sp[Q^n x - Q^n y]}{sp[x-y]} = \lim_{X\to\infty} \frac{\max(0,X-n)}{X} = 1 \quad \text{for all } n = 1,2,\ldots$$

(cf. also section 7 of [111]).

In this chapter we give (both necessary and sufficient) conditions for the $\bar{Q}$-operator to be a J-step contraction mapping for some $J = 1,2,\ldots$. The identification of these conditions is of particular importance since with $\bar{Q}$ being contracting, the geometric convergence result of value-iteration, as discussed in section 6 of chapter 1, and which in the *general* case requires a complicated analysis, is straightforward (cf. theorem 2.2.1), and in addition the contraction-property may be exploited in order to obtain:

(1) a lower bound for the convergence rate of the value iteration method.

(2) Upper and lower bounds, as well as variational characterizations for the fixed point v^* of the functional equation (1.4.15) which in this

case is unique up to a multiple of $\underline{1}$ (i.e. its representation in $\bar{E}^N$ is unique).

(3) A test for eliminating suboptimal actions in the value-iteration method.

As necessary conditions we obtain some important charaterizations with respect to the chain- and periodicity structure of the problem. In addition we present a general *sufficient* condition of a "scrambling" type (cf. [1], [51]) which encompasses a number of important and easily checkable conditions. We note that in [86] a special case of this "scrambling-type" condition was used to prove the convergence of the relative cost differences. In section 9 of the previous chapter, a data-transformation (cf. (1.9.9)) was presented which turns every undiscounted Markov Renewal Program into an "equivalent" undiscounted MDP. In addition the transformed problem has every policy aperiodic so that the (geometric) convergence of $\{Q^n x - ng^*\}_{n=1}^{\infty}$ is guaranteed for all $x \in E^N$, i.e. $J^* = 1$ or (GC) is satisfied. In section 3, we show that for unichained MRPs, this data-transformation has the considerably stronger property of turning the MRP into an equivalent MDP, in which the value iteration-operator is at least N-step contracting with all of the nice consequences mentioned above. These results are obtained by showing that the transformed problem satisfies the above "scrambling-type" condition. The results in this chapter are based upon FEDERGRUEN, SCHWEITZER and TIJMS [43].

2.2. NECESSARY AND SUFFICIENT CONDITIONS FOR $\bar{Q}$ TO BE A (J-STEP) CONTRACTION MAPPING, AND SOME OF ITS IMPLICATIONS

Before studying necessary and sufficient conditions for $\bar{Q}$ to be a J-step contraction mapping for some $J = 1,2,\ldots$, we first show that the geometric convergence of the sequence $\{Q^n x - ng^*\}_{n=1}^{\infty}$ for all $x \in E^N$ is straightforward when $\bar{Q}^J$ is a contraction mapping. We first formulate and prove this result with respect to the T-operator (cf. (1.6.3)). The corresponding property for the Q-operator then follows from corollary 2.2.3 below.

THEOREM 2.2.1. (Geometric convergence of value-iteration)
Let $\bar{T}$ be a J-step contraction operator on $\bar{E}^N$, for some $J = 1,2,\ldots$ and some contraction factor $0 < \rho \leq 1$ (cf. (2.1.2)). Then, for each $x \in E^N$, there exists a $v^0 = v^0(x) \in V$ such that for all $i \in \Omega$,

$$(2.2.1) \qquad |T^{nJ+r}x_i - (nJ+r)g_i^* - v_i^0| \leq (1-\rho)^n sp[x-v^0]; \; n = 1,2,\ldots; \; r = 0,\ldots,J-1.$$

PROOF. Fix $x \in E^N$ and $v \in V$. For all $n = 0,1,2,\ldots$, let $e(n,x) = T^n x - ng^* - v$. Then $e(n,x) = T^n x - T^n v$ as follows from a repeated application of (1.6.4) and (1.4.15). From the non-expansiveness of the Q- and T operator (cf. (1.4.1)), follows for all $n \geq 1$:

$$(2.2.2) \quad (x-v)_{\min} \leq e(n,x)_{\min} \leq e(n+1,x)_{\min} \leq e(n+1,x)_{\max} \leq e(n,x)_{\max} \leq (x-v)_{\max}$$

hence $\{e(n,x)_{\max}\}_{n=1}^{\infty}$ [$\{e(n,x)\}_{\min}$] is monotonically non-increasing [non-decreasing] to some limit $t(x)^+$ [$t(x)^-$]. But

$$0 \leq sp[e(nJ,x)] = sp[T^{nJ}x - T^{nJ}v] = sp[\bar{T}^{nJ}x - \bar{T}^{nJ}v] \leq (1-\rho)^n sp[x-v], \text{ as } n\to\infty$$

hence $t(x)^+ - t(x)^- = \lim_{n\to\infty} sp[e(nJ,x)] = 0$, so $t(x)^+ = t(x)^- = t(x)$. Thus, $\lim_{n\to\infty} e(n,x) = t(x)\underline{1}$ and $\lim_{n\to\infty} T^n x - ng^* = v + t(x)\underline{1} = v^0(x) \in V$.

Finally use the fact that for all $n = 1,2,\ldots$ and $r = 0,\ldots,J-1$:

$$[T^{nJ+r}x - (nJ+r)g^* - v^0(x)]_{\min} = e(nJ+r,x)_{\min} - t(x) \leq 0 \leq$$

$$\leq e(nJ+r,x)_{\max} - t(x) = [T^{nJ+r}x - (nJ+r)g^* - v^0(x)]_{\max}$$

just as the fact that $x_{\min} \leq 0 \leq x_{\max}$ implies $\|x\| \leq sp[x]$, to obtain

$$\|T^{nJ+r}x - (nJ+r)g^* - v^0(x)\| \leq sp[T^{nJ+r}x - (nJ+r)g^* - v^0(x)] = sp[T^{nJ+r}x - T^{nJ+r}v^0] \leq (1-\rho)^n sp[x-v^0]. \quad \square$$

We next introduce two conditions with respect to the chain- and periodicity structure, both of which appear as *necessary* conditions for $\bar{Q}^J$ or $\bar{T}^J$ to be a contraction operator (for some $J = 1,2,\ldots$).

A_1: There exists a *randomized aperiodic* policy $f \in S_{RMG}$, whcih has R^* as its *single* subchain.

A_2: There exists a *randomized aperiodic* policy $f \in S_R$, which has $\hat{R}$ as its *single* subchain.

Note that A_1 and A_2 strenghten the conditions (UNI) and (UR) that were introduced in sections 4 and 6 of the previous chapter by requiring the policy f to have the additional property of aperiodicity. The following statements are equivalent formulations for both A_1 and A_2, which are expressed in terms of the structure of the finite set of *pure* (maximal gain) policies only (cf. corollary 3.3 in [109] and th.3.1 part (c) in [110], and observe that S_R

appears as the set of all maximal gain policies, when taking all $q_i^k = 0$):

A_1': Let $\mathcal{C}^* = \{C \subseteq \Omega \mid C$ is a subchain for $P(f)$, for some $f \in S_{PMG}\}$

Then (a) for any pair $C, C' \in \mathcal{C}^*$, there exists $\{C^{(1)} = C, C^{(2)}, \ldots, C^{(n)} = C'\}$ with $C^{(i)} \in \mathcal{C}^*$ and $C^{(i)} \cap C^{(i+1)} \neq \emptyset$ $(i = 1, \ldots, n-1)$

(b) the integers which appear as the period of some subchain of some policy in S_{PMG}, are relatively prime.

A_2': Let $\hat{\mathcal{C}} = \{C \subseteq \Omega \mid C$ is a subchain for $P(f)$, for some $f \in S_P\}$

Then (a) for any pair $C, C' \in \hat{\mathcal{C}}$, there exists $\{C^{(1)} = C, C^{(2)}, \ldots, C^{(n)} = C'\}$ with $C^{(i)} \in \hat{\mathcal{C}}$ and $C^{(i)} \cap C^{(i+1)} \neq \emptyset$ $(i = 1, \ldots, n)$

(b) the integers which appear as the period of some subchain of some policy in S_P, are relatively prime.

We note that whereas part (b) of A_1' implies part (b) of A_2' the parts (a) of A_1' and A_2' are mutually independent. In addition, we remark that more *efficient* procedures have been established to verify A_1 and A_2 (or alternatively A_1' and A_2'). (cf. [109] and [110]).

THEOREM 2.2.2. (Necessary conditions for $\bar{T}$ to be a contraction mapping). *Let $\bar{T}$ be a J-step contraction mapping on $\bar{E}^N$ for some $J = 1,2,\ldots$* (cf. (2.1.2)). *Then*

(1) $v \in V$ *is unique up to a multiple of* $\underline{1}$, *i.e. (UNI) holds*

(2) $g_i^* = g^*$ *for all* $i \in \Omega$; *hence* $L(i) = K(i)$, *for all* $i \in \Omega$ *and* $Qx = Tx$ *for all* $x \in E^N$.

(3) A_1 *and* A_2 *hold.*

PROOF. Let $v^0, v^{00} \in V$. By a repeated application of (1.4.15), we obtain, using (1.6.4):

$$T^J v^0 = v^0 + Jg^* \quad \text{and} \quad T^J v^{00} = v^{00} + Jg^*.$$

hence

$$sp[v^0 - v^{00}] = sp[T^J v^0 - T^J v^{00}] \leq (1-\rho)\, sp[v^0 - v^{00}],$$

which implies $sp[v^0 - v^{00}] = 0$, or $v \in V$ is unique up to a multiple of $\underline{1}$. This condition in turn, is equivalent with the existence of a policy $f \in S_{RMG}$, which has R^* as its *single* subchain i.e. with (UNI) (cf. section 4 of chapter 1).

Condition A_1, i.e. the fact that even *aperiodic* policies can be found with this property, then follows from the convergence of $\{T^n x - ng^*\}_{n=1}^{\infty}$

for all $x \in E^N$ (cf. theorem 2.2.1, using th.5.4 part (b) and th.3.1 part (f) of [110]). The existence of a *unichained* maximal gain policy in turn implies part (2) of the theorem.

Next, assume to the contrary that A_2 does not hold. State i is said to *reach* state j, if there exists a policy $f \in S_p$, and some integer $r \geq 0$, such that $P(f)^r_{ij} > 0$. Let f^* be any randomized policy which has $f^*_{ik} > 0$ for all $i \in \Omega$, $k \in K(i)$. We claim

(2.2.3) there exists a pair of states j_1, $j_2 \in \hat{R}$ such that j_2 does not reach j_1.

For assuming the contrary of (2.2.3) would imply that all states in $\hat{R}$ communicate with each other under $P(f^*)$, i.e. $R(f^*) = \hat{R}$, in view of $R(f^*) \subseteq \hat{R}$, following from the definition (1.4.13). This contradicts our assumption that A_2 does not hold, since in this case $P(f^*)$ has to be aperiodic. To verify the latter, fix an aperiodic maximal gain policy h, the existence of which was shown above. Let $i \in R(h) \subseteq \hat{R}$ and note that $1 \leq$ greatest common divisor (g.c.d.)$\{n|P(f^*)^n_{ii} > 0\} \leq$ g.c.d.$\{n|P(h)^n_{ii} > 0\} = 1$, where the last equality follows from the aperiodicity of h (cf. [71]), and where the second inequality is due to $\{n|P(f^*)^n_{ii} > 0\} \supseteq \{n|P(h)^n_{ii} > 0\}$, the latter following from the definition of f^*.

Hence, $P(f^*)$ is aperiodic, thus completing the proof of (2.2.3). Next, fix a policy $f_1 \in S_p$ with $j_1 \in R(f_1)$ and let C be the subchain of $P(f_1)$ which contains j_1. Obviously j_2 does not reach any one of the states in C. Next choose $x \in E^N$ such that $x_i = \lambda >> 1$ for $i \in C$ and $x_i = 0(1)$ otherwise, where $0(1)$ denotes any bounded term in λ. Let v^o satisfy (1.4.15). Since

$$T^J x_i \geq [P(f_1)^J x]_i + \sum_{\ell=0}^{J-1} [P(f_1)^\ell q(f_1)]_i,$$

and since C is a subchain of $P(f_1)$, we have for each $J \geq 1$,

$$T^J x_i = \lambda + 0(1), \quad \text{for } i \in C.$$

Since j_2 cannot reach C, we have $(Tx)_{j_2} = 0(1)$ and so $(T^J x)_{j_2} = 0(1)$. Finally observing that $T^J v^o = 0(1)$, we have

$$sp[T^J x - T^J v^o] = \lambda + 0(1),$$

whereas

$$sp[x - v^o] = \lambda + 0(1)$$

as well. Conclude that for each $J \geq 1$,

$$1 \geq \sup\left\{\frac{sp[T^J u - T^J v]}{sp[u-v]} \;\middle|\; u,v \in E^N \text{ with } sp[u-v] > 0\right\} \geq$$

$$\geq \lim_{\lambda \to \infty} \frac{sp[T^J x - T^J v^o]}{sp[x-v^o]} = 1,$$

thus contradicting the fact that $\bar{T}$ is a contraction mapping. This proves A_2 by contradiction. □

COROLLARY 2.2.3. *Fix* $J = 1,2,\ldots$

(1) $\bar{Q}$ *is a* J*-step contraction operator on* $\bar{E}^N$, *for some contraction factor* $\rho > 0$ (cf. (2.1.2)) *if and only if*

(2) $\bar{T}$ *is a* J*-step contraction operator on* $\bar{E}^N$, *for some contraction factor* $\rho > 0$.

In addition both (1) *and* (2) *imply that the* Q*- and* T*-operators coincide.*

PROOF.

(2) ==> (1): follows from theorem 2.2.2 since condition (2) implies $Q = T$.

(1) ==> (2): we recall that the Q operator reduces to the T operator as follows:

for each $x \in E^N$ there exists a scalar $t_0(x)$, such that $Q^n(x+tg^*) = T^n(x+tg^*)$ for $n = 1,2,\ldots$ and $t \geq t_0(x)$

the proof of which is easy and may be found in lemma 2.2, part (g) of [111]. Next, assume to the contrary, that there exist two vectors $x,y \in E^N$, such that

$$sp[T^J x - T^J y] > (1-\rho)\, sp[x-y].$$

Let $t \geq \max\{t_0(x), t_0(y)\}$ and observe, using (1.6.4), that

$$sp[Q^J(x+tg^*) - Q^J(y+tg^*)] = sp[T^J(x+tg^*) - T^J(y+tg^*)] =$$

$$= sp[T^J x - T^J y] > (1-\rho)\, sp[(x+tg^*) - (y+tg^*)],$$

thus contradicting condition (1). □

REMARK 1. If $\bar{Q}$ (or $\bar{T}$) is a J-step contraction operator on $\bar{E}^N$, with contraction factor ρ, then in the geometric convergence result achieved in theorem

2.2.1, an upper bound may be obtained for the number of steps J needed for contraction, i.e. there exists an integer $M \leq N^2 - 2N + 2$ and a number λ, with $0 \leq \lambda \leq (1-\rho)^{M/J}$ such that for each $x \in E^N$, there exists a $v^o = v^o(x) \in V$ with

$$\| Q^{nM+r} x_i - (nM+r) g_i^* - v_i^o \| < \lambda^n \, sp[x - v^o]$$

$$n = 1,2,\ldots; \quad r = 0,\ldots,M-1; \quad i \in \Omega.$$

The upperbound on M holds whenever condition A1 is satisfied, as was pointed out in (1.6.25) and we know from th.2.2.2 that A1 holds whenever $\bar{Q}$ is a (J-step) contraction operator.

In addition the upperbound on M is at least sharp up to a term of the order O(N) as has been demonstrated by example 2 in [111] (cf. also section 6 of chapter 1). One may verify that in this example, the $\bar{Q}$-operator is a contraction operator.

We next introduce a general "scrambling type" recurrency condition under which the $\bar{Q}$-operator will be shown to be a contraction operator (cf. also [1], [51]):

(S): there exists an integer $J \geq 1$, such that for every pair of J-tuples of pure policies $(f_1,\ldots,f_J)$ and $(h_1,\ldots,h_J)$:

$$(2.2.4) \qquad \sum_{j=1}^{N} \min[P(f_J)\ldots P(f_1)_{i_1 j}; \; P(h_J)\ldots P(h_1)_{i_2 j}] > 0; \text{ for all } i_1 \neq i_2 \in \Omega.$$

Note that if (2.2.4) holds for some integer $J \geq 1$, then it equally holds for any integer $m \geq J$ (cf. e.g. the proof of lemma V. 2.3 in [67], or lemma 3.3.2 part (b)).

Theorem 2.2.4 below shows that this condition (S), encompasses a number of important and easily checkable conditions.

THEOREM 2.2.4. The following conditions are special cases of condition (S):

(1) $\sum_j \min(P^{k_1}_{i_1 j}, P^{k_2}_{i_2 j}) > 0$ for all $i_1 \neq i_2$ and $k_1 \in K(i_1)$, $k_2 \in K(i_2)$.

(2) There exists a state s and an integer $\nu \geq 1$, such that $P(f^1)\ldots P(f^\nu)_{is} > 0$ for all $f^1, f^2, \ldots, f^\nu \in S_p$; $i \in \Omega$ (cf. WHITE [131]).

(3) Every policy is unichained; there exists a state $s \in \Omega$ which is recurrent under every policy, and $P^k_{ss_k} > 0$ for all $k \in K(s)$.

(4) Every policy is unichained and $P^k_{ii} > 0$ for all $i \in \Omega$, $k \in K(i)$.

PROOF. (1) $\Longrightarrow$ (S) with J = 1; (2) $\Longrightarrow$ (S) with J = ν, was shown in [131]; (3) $\Longrightarrow$ (2) with ν = N - 1, was shown in [1] th.2. (4) $\not\Longrightarrow$ (S): Fix two sequences of policies $(f_N,\ldots,f_1)$ and $(h_N,\ldots,h_1)$ and i_1, $i_2 \in \Omega$ with $i_1 \neq i_2$. Let

$$S(n) = \{j \mid P(f_1)\ldots P(f_n)_{i_1 j} > 0\} \text{ and } W(n) = \{j \mid P(h_1)\ldots P(h_n)_{i_2 j} > 0\}.$$

Note that, in view of $P^k_{ii} > 0$ for all $i \in \Omega$, $k \in K(i)$:

(2.2.5) $S(n+1) \supseteq S(n)$, $W(n+1) \supseteq W(n)$ $\quad n = 1,2,\ldots$

Thus assuming to the contrary that $S(N) \cap W(N) = \emptyset$, it follows that $S(m) \cap W(m) = \emptyset$, for all $0 \leq m \leq N$. Together this result and the fact that $S(k) \cup W(k)$ is nondecreasing in k, implies for some $m < N$, $S(m+1) = S(m)$ and $W(m+1) = W(m)$. Letting f^* be any policy such that $f^*(i) = f_{m+1}(i)$ for $i \in S(m)$ and $f^*(i) = h_{m+1}(i)$ for $i \in W(m)$, we then have that both S(m) and W(m) are closed sets of states for $P(f^*)$. This contradicts the unichainedness of $P(f^*)$ and so $S(N) \cap W(N)$ is non-empty.

REMARK 2. Observe that condition (1) requires each P(f), $f \in S_P$, to be scrambling (cf. e.g. [51]). In addition we note that conditions (1), (2) and (4) are mutually independent. To verify that (2) $\not\Longrightarrow$ (1), and (2) $\not\Longrightarrow$ (4), consider an example in which $S_P = \{f\}$, with

$$P(f) = \begin{vmatrix} 0 & * & 0 \\ 0 & 0 & * \\ 0 & 0 & * \end{vmatrix}$$

which satisfies (2) with $\nu = 2$ (where a * indicates a positive entry). Next, the example in which $S_P = \{f_1, f_2\}$ with

$$P(f_1) = \begin{vmatrix} * & 0 & * \\ * & 0 & * \\ * & 0 & 0 \end{vmatrix} \quad \text{and} \quad P(f_2) = \begin{vmatrix} * & 0 & * \\ * & 0 & * \\ 0 & 0 & * \end{vmatrix}$$

satisfies (1) but not White's condition, nor (4). Finally, the example with $S_P = \{f\}$ and

$$P(f) = \begin{vmatrix} * & * & 0 \\ 0 & * & * \\ 0 & 0 & * \end{vmatrix}$$

shows (4) $\not\Longrightarrow$ (1), whereas (4) $\not\Longrightarrow$ (2) follows from the fact that (4) includes cases where no state is recurrent under every policy. Finally

observe that condition (S) requires each policy to have a unichained and aperiodic tpm.

Finally we note that (S) is not necessarily satisfied, in case there is an integer $J \geq 1$ such that for all $f_J,\ldots,f_1 \in S_P$ the stochastic matrix $P(f_J)\ldots P(f_1)$ is scrambling, where a stochastic matrix $P = (P_{ij})$; $i,j \in \Omega$ is said to be *scrambling* if:

$$\sum_{j=1}^{N} \min[P_{i_1 j}; P_{i_2 j}] > 0 \quad \text{for all } i_1, i_2 \in \Omega.$$

This can be verified from the example in which $S_P = \{f_1, f_2\}$ with

$$P(f_1) = \begin{vmatrix} 0 & * & 0 \\ 0 & * & 0 \\ * & 0 & 0 \end{vmatrix}, \quad \text{and} \quad P(f_2) = \begin{vmatrix} 0 & * & 0 \\ * & 0 & * \\ * & 0 & 0 \end{vmatrix}.$$

Verify that, the product matrix of the tpm's of any *triple* of policies is scrambling. However for any $h \geq 1$,

$$P(f_{i_1})\ldots P(f_{i_h})P(f_1)^2 P(f_1) = \begin{vmatrix} 0 & * & 0 \\ 0 & * & 0 \\ 0 & * & 0 \end{vmatrix}, \quad \text{and}$$

$$P(f_{i_1})\ldots P(f_{i_h})P(f_1)^2 P(f_2) = \begin{vmatrix} * & 0 & * \\ * & 0 & * \\ * & 0 & * \end{vmatrix}$$

which shows that (2.2.4) can't be satisfied for any $J \geq 1$, i.e. (S) does not hold.

Theorem 2.2.5 below shows that condition (S) is sufficient for $\bar{Q}$ to be a (J-step) contraction operator:

THEOREM 2.2.5. *Condition* (S) *is a sufficient condition for* $\bar{Q}$ *to be a (J-step) contraction operator on* $\bar{E}^N$.

PROOF. The proof of this theorem is related to the one of th.1 in [1]. First, define

$$\alpha = \min\{\sum_j \min[P(f_J)\ldots P(f_1)_{i_1 j};\ P(h_J)\ldots P(h_1)_{i_2 j}] \mid \tag{2.2.6}$$

$$i_1, i_2 \text{ with } i_1 \neq i_2,\ f_k,\ h_k \ (1 \leq k \leq J)\},$$

where $\alpha > 0$ follows from (2.2.4) and the fact that in (2.2.6) the minimum is taken over a finite number of combinations. We shall prove that:

(2.2.7) $\quad (Q^J x - Q^J y)_i - (Q^J x - Q^J y)_\ell \leq (1-\alpha)\ sp[x-y] \quad$ for all $i, \ell \in \Omega$.

The theorem clearly follows from (2.2.7). The inequality in (2.2.7) trivially holds when $i = \ell$. Fix now $i \neq \ell$, and let

$$Q^J x_i = q(f_J)_i + \sum_{k=1}^{J-1} P(f_J)\ldots P(f_{J-k+1}) q(f_{J-k})_i + P(f_J)\ldots P(f_1) x_i,$$

and

$$Q^J y_\ell = q(h_J)_\ell + \sum_{k=1}^{J-1} P(h_J)\ldots P(h_{J-k+1}) q(h_{J-k})_\ell + P(h_J)\ldots P(h_1) y_\ell.$$

Next introduce the shorthand notation,

$$\beta_j = P(f_J)\ldots P(f_1)_{ij} \quad \text{and} \quad \gamma_j = P(h_J)\ldots P(h_1)_{\ell j}.$$

Using the fact that

$$\sum_j a_j^+ = -\sum_j a_j^-, \quad \text{if} \quad \sum_j a_j = 0,$$

as well as the fact that $(a-b)^+ = a - \min(a,b)$, we obtain

$$(Q^J x - Q^J y)_i - (Q^J x - Q^J y)_\ell \leq \sum_j \beta_j (x-y)_j - \sum_j \gamma_j (x-y)_j =$$

$$= \sum_j [\beta_j - \gamma_j]^+ (x_j - y_j) + \sum_j [\beta_j - \gamma_j]^- (x-y)_j \leq (x-y)_{max} \sum_j [\beta_j - \gamma_j]^+$$

$$+ (x-y)_{min} \sum_j [\beta_j - \gamma_j]^- = \sum_j [\beta_j - \gamma_j]^+ sp[x-y] =$$

$$= [1 - \sum_j \min(\beta_j, \gamma_j)]\ sp[x-y] \leq (1-\alpha)\ sp[x-y]. \quad \square$$

2.3. ON TRANSFORMING UNICHAINED MARKOV RENEWAL PROGRAMS INTO EQUIVALENT AND CONTRACTING MARKOV DECISION PROBLEMS

In this section, we consider the more general class of Markov Renewal Programs, as introduced in section 9 of chapter 1. It was pointed out in section 9 of chapter 1 that by replacing $\hat{\tau}_{ij}^k = T_i^k$ $(i,j \in \Omega,\ k \in K(i))$, and by applying the data-transformation (1.9.9) with the choice $\sigma = 1$, our MRP-model will be transformed into an *equivalent* undiscounted MDP-problem where the equivalency between two undiscounted MRPs was defined by (EQUI) in (1.9.6).

Let $\tilde{Q}$ be the value-iteration operator in this transformed MDP. It was pointed out that, by taking τ in (1.9.9) strictly smaller than the upper-bound given in (1.9.10), $\{\tilde{Q}^n x - ng^*\}_{n=1}^{\infty}$ converges *geometrically* fast to a

solution $v \in \tilde{V}$, i.e. for each $x \in E^N$, there exists a vector $L(x) \in \tilde{V}$, and numbers $K = K(x)$, $0 \leq \lambda < 1$, such that:

$$\| \tilde{Q}^n x - ng^* - L(x) \| < K \lambda^n, \quad n = 0,1,2,\ldots$$

This shows that, by applying the above data-transformation, and by subsequently doing value-iteration with respect to the transformed MDP, we find sequences which approach g^* and some $v \in \tilde{V}$; moreover, it follows from a generalization of ODONI [89] and from the fact that the original MRP and the transformed MDP are equivalent, that any policy which is generated by the value-iteration scheme after a large enough number of iterations is maximal gain (in the original MRP).

We henceforth assume condition (U) to hold

(2.3.1) (U): every pure policy in the MRP is unichained, i.e. $n(f) = 1$, $f \in S_p$.

We note that under (U), $g^* = \langle g^* \rangle \underline{1}$ and as a consequence $V = \hat{V}$ (cf. section 1.9). It thus follows that under (U), $\tau L(x) \in V$, for all $x \in E^N$, so that the above described value iteration method will yield us in addition a sequence converging at a geometric rate towards a solution of the optimality equation in the *original* MRP.

We next make the important observation that, with τ *chosen strictly less than the upperbound in* (1.9.10), *the* $\tilde{Q}$*-operator satisfies condition* (4) *of* th.2.2.4, *and as a consequence has the considerably stronger property of being* J*-step contracting for some* $J \leq N$ (cf. th.2.2.5), *where the fact that* J *can be chosen less than or equal to* N *follows from the proof of* th.2.2.4.

Note that since the $\tilde{Q}$-operator is contracting under (U), $v \in \tilde{V}$ is unique up to a multiple of $\underline{1}$ (cf. th.2.2.2), and in view of $V = \hat{V} = \{v \in E^N \mid \tau^{-1} v \in \tilde{V}\}$, we have that $v \in V$ is unique up to a multiple of $\underline{1}$ as well. As a consequence the representation $\bar{v}^*$ of $v \in V$, in $\bar{E}^N$ is unique. In the remainder of this chapter, we will show that for unichained MRP's the above data-transformation and the resulting contraction property of the operator $\tilde{Q}$ in the transformed MDP may be exploited, in order to

(a) find lower and upper bounds for $\bar{v}^*$

(b) derive variational characterizations (extremal principles) for $\bar{v}^*$

(c) derive a test for eliminating nonoptimal actions.

To our knowledge, these bounds for $\bar{v}^*$ are the first one obtained in undiscounted MRPs.

We will use the following representation of $\bar{E}^N$ (cf. section 2.1): $\bar{E}^N = \{x \in E^N \mid x_N = 0\}$ so that the reprsentation of a vector $x \in E^N$ in $\bar{E}^N$ is given by $\bar{x}$, with $\bar{x}_i = x_i - x_N$, $i \in \Omega$. Note that since $\bar{x}_{min} \leq 0 \leq \bar{x}_{max}$, for all $x \in E^N$:

$$(2.3.2) \qquad \|\bar{x}\| \leq sp[\bar{x}] = sp[x].$$

For ease of presentation, we first discuss the above topics for the discrete time MDP as considered in section 2.1.

THEOREM 2.3.1. *Consider the MDP value iteration operator Q. Define $\bar{Q}$ as the reduction of the operator Q to $\bar{E}^N$, i.e. $\bar{Q}: E^N \to \bar{E}^N$ with $\bar{Q}x = Qx - [Qx]_N \underline{1}$. Assume that $\bar{Q}$ is a (J-step) contraction operator (for some $J \geq 1$) on $\bar{E}^N$, with contraction factor $\rho > 0$ (cf. (2.1.2)).*

Finally, let $\bar{v}^*$ be the unique fixed point of $\bar{Q}$ on $\bar{E}^N$, i.e. let $\bar{v}^*$ be the unique representation of $v \in V$ in $\bar{E}^N$.

(a) *(Upper and lower bounds)*

For all $x \in E^N$, $n \geq 0$, $i \in \Omega$ and $0 \leq r \leq J-1$:

$$\bar{Q}^{nJ+r}x_i - \rho^{-1}(1-\rho)^n sp[Q^J x - x] \leq v_i^* \leq \bar{Q}^{nJ+r}x_i + \rho^{-1}(1-\rho)^n sp[Q^J x - x]$$

and, for $v \in V$

$$sp[Q^{nJ+r}x - v^*] \leq \rho^{-1}(1-\rho)^n sp[Q^J x - x]$$

(b) *(Alternative elimination)*

If for some $x \in E^N$, some state $i \in \Omega$, and some action $k \in K(i)$

$$(2.3.3) \qquad q_i^k + \sum_j p_{ij}^k x_j - x_i < (Q^J x - Q^{J-1}x)_{min} - \rho^{-1} sp[Q^J x - x],$$

then k does not satisfy the maximum in the optimality equation (1.4.15) i.e. k is non-optimal.

PROOF. The proof of part (a) goes along lines with the one given for lemma 1.2.1.

(a) Using the continuity of the sp[x]-norm on E^N, the fact that $g^* = \langle g^* \rangle \underline{1}$, as well as (2.3.2) we obtain:

$$|\bar{Q}^{nJ+r}x_i - v_i^*| \leq sp[Q^{nJ+r}x - \lim_{m\to\infty}\{Q^{mJ+r}x - (mJ+r)g^*\}] =$$

$$= \lim_{m\to\infty} sp[Q^{mJ+r}x - Q^{nJ+r}x] = \lim_{m\to\infty} sp[\sum_{\ell=n}^{m-1}(Q^{(\ell+1)J+r}x - Q^{\ell J+r}x)] \leq$$

$$\leq \sum_{\ell=n}^{\infty} sp[Q^{(\ell+1)J+r}x - Q^{\ell J+r}x] \leq \sum_{\ell=n}^{\infty}(1-\rho)^{\ell}\, sp[Q^{J+r}x - Q^{r}x]$$

$$\leq \rho^{-1}(1-\rho)^{n}\, sp[Q^{J}x - x]$$

where the last inequality uses (1.4.1). This verifies part (a).

(b) It follows from the proof of theorem 1 of [89] that $\langle g^*\rangle \geq (Q^Jx-Q^{J-1}x)_{min}$. Suppose alternative $k \in K(i)$ which satisfies (2.3.3), attains the maximum in the optimality equation (1.4.15). Note from corollary 2.2.3 that the Q-operator and T-operator coincide. Then, using part (a) and the fact that $v^* \in V$, we have

$$q_i^k + \sum_j P_{ij}^k x_j - x_i = q_i^k - g^* + \sum_j P_{ij}^k \bar{v}_j^* - \bar{v}_i^* + \sum_j P_{ij}^k (x_j - \bar{v}_j^*) -$$

$$- (x_i - \bar{v}_i^*) + g^* \geq (x-\bar{v}^*)_{min} - (x-\bar{v}^*)_{max} + g^* = -sp[x-\bar{v}^*] + g^* \geq$$

$$\geq -\rho^{-1}\, sp[Q^Jx-x] + (Q^Jx-Q^{J-1}x)_{min}.$$

REMARK 3. The reduction of the Q-operator to $\bar{E}^N$, was first used in White [131], in order to ensure the boundedness of his value-iteration scheme. The lower- and upper bounds for $\bar{v}^*$ are in fact generalizations of the lower- and upper bounds obtained by MACQUEEN [81] and PORTEUS [94] for discounted MDP's (cf. lemma 1.2.1). Note that our bounds with n = 0 coincide with the analogue of MacQueen's bounds, whereas the analogue of Porteus' bounds is obtained by taking n = 1.

We now return to the general MRP-model. By using the above data-transformation, and by applying th.2.3.1 to the transformed MDP, we obtain upper- and lower bounds as well as variational characterizations for each of the components of $\bar{v}^*$, and in addition a test for eliminating *non-optimal actions*.

COROLLARY 2.3.2. *Consider a unichained* MRP. *Fix* $0<\tau<\min\{T_i^k/(1-P_{ii}^k) \mid (i,k)$ *with* $P_{ii}^k < 1\}$. *Let* $\tilde{Q}$ *be the value-iteration operator in the transformed* MDP *(cf. section* 1.9*). Next, let* $\bar{Q}$ *be the reduction of* $\tilde{Q}$ *to* $\bar{E}^N$, *i.e.* $\bar{Q}x=\tilde{Q}x-[\tilde{Q}x]_N 1$ *for all* $x \in E^N$. *Finally, let* ρ *be the (N-step) contraction factor of the operator* $\bar{Q}$ *(cf.* (2.1.2) *and th.2.2.4). Then, the unique* $\bar{v}^* \in V$, *with* $\bar{v}_N^* = 0$, *satisfies*

(a) $\bar{Q}^{nN+r}x_i - \rho^{-1}(1-\rho)^n sp[\tilde{Q}^N x - x] \le \tau^{-1}\bar{v}_i^* \le \bar{Q}^{nN+r}x_i + \rho^{-1}(1-\rho)^n sp[\tilde{Q}^N x - x]$

for all $x \in E^N$, *and* $n = 0,1,\ldots;\ r = 0,\ldots,N-1;\ i \in \Omega$

(b) $v_i^* = \tau \max_{x\in E^N}\{\bar{Q}^{nN+r}x_i - \rho^{-1}(1-\rho)^n sp[\tilde{Q}^n x - x]\}$

$= \tau \min_{x\in E^N}\{\bar{Q}^{nN+r}x_i + \rho^{-1}(1-\rho)^n sp[\tilde{Q}^N x - x]\}$

$i \in \Omega;\ n = 0,1,\ldots;\ r = 0,\ldots,n-1.$

(c) *If for some* $x \in E^N$, *some state* $i \in \Omega$, *and some action* $k \in K(i)$

$\tilde{q}_i^k + \sum_j \tilde{P}_{ij}^k x_j - x_i < (\tilde{Q}^N x - \tilde{Q}^{N-1}x)_{min} - \rho^{-1} sp[\tilde{Q}^N x - x]$

then k *does not satisfy the maximum in the optimality equation* (1.9.5).

The variational characterizations in part (b) follow from part (a) by taking $x = v \in V$ and using the fact that $\bar{Q}v = \bar{v}^*$ for all $v \in V$. Variational characterizations for g^* were recently obtained in [112]. One might use both lower and upper bounds for $\bar{v}^*$, and the test for eliminating suboptimal actions (cf. part (a)), in the course of the following value-iteration scheme for finding g^*, $\bar{v}^*$ and some maximal gain policy.

(2.3.4) $y(n)_i = \bar{Q}y(n-1)_i = \max_{k\in K(i)}\{\tilde{q}_i^k + \sum_j \tilde{P}_{ij}^k y(n-1)_j\}$

$- \max_{k\in K(N)}\{\tilde{q}_N^k + \sum_j \tilde{P}_{Nj}^k y(n-1)_j\},\quad i \in \Omega$

with $y(0) \in E^N$ chosen arbitrarily.

Let f_n be a policy which achieves the N maxima in (2.3.4). Define

$$\theta_L(n) = [\tilde{Q}y(n-1) - y(n-1)]_{min};\ \theta_U(n) = [\tilde{Q}y(n-1) - y(n-1)]_{max}.$$

The sequence $\{y(n)\}_{n=1}^{\infty}$ has the following, easily verified and previously discussed properties.

(a) $\tau y(n) \to \bar{v}^*$

(b) $\theta_L(n) \le g(f_n) \le g^* \le \theta_U(n)$ (cf. HASTINGS [53] and ODONI [89])

with $\lim_{n\to\infty} \theta_L(n) = g^* = \lim_{n\to\infty} \theta_U(n)$

(c) f_n is maximal gain, for all n sufficiently large (cf. ODONI [89]).

E.g. whenever at some stage n, i.e. for $x = y(n)$, the test in part (c) of cor.2.3.2 is met for some $i \in \Omega$, and $k \in K(i)$, k may be deleted permanently from K(i) thus reducing the number of calculations in the following iterations. However, both the application of the bounds for $\bar{v}^*$ as the use of the elimination test require the computation of at least some lower bound of the (N-step) contraction factor of the operator $\bar{Q}$.

PROPOSITION 2.3.3. *Define* $\tilde{\alpha}$ *by the right hand side of* (2.2.6) *with* P(f) *replaced by* $\tilde{P}(f)$, *and define*

$$(2.3.5) \qquad \hat{\rho} = \min\{\tilde{P}(f_N)\ldots\tilde{P}(f_1)_{ij} \mid \tilde{P}(f_N)\ldots\tilde{P}(f_1)_{ij} > 0;\ i \in \Omega;\ f_1,\ldots,f_N \in S_P\}.$$

Then

(2.3.6) $0 < \hat{\rho} \leq \tilde{\alpha}$, i.e. $\hat{\rho}$ may be used as a lower bound on the (N-step) contraction factor of $\bar{Q}$.

PROOF. By the proof of th.2.2.5 we can take the scrambling coefficient $\tilde{\alpha}$ as (N-step) contraction factor for the operator $\bar{Q}$. We shall now verify (2.3.6). $\hat{\rho} > 0$ is immediate from the fact that in (2.3.5) the minimum is taken over a *finite* number of positive numbers. Next, let the minimum in (2.2.6) (with P(f) replaced by $\tilde{P}(f)$) be attained for $s,t \in \Omega$; $f_k^*, h_k^* \in S_P$ $(1 \leq k \leq N)$ and let γ be such that

$$\beta = \min[\tilde{P}(f_N^*)\ldots\tilde{P}(f_1^*)_{s\gamma}; \tilde{P}(h_N^*)\ldots\tilde{P}(h_1^*)_{t\gamma}] > 0.$$

Then $\tilde{\alpha} \geq \tilde{\beta} \geq \hat{\rho}$. □

The lower bound $\hat{\rho}$ may be computed as follows. Let x^o be defined by

$$x_i^o = \min\{\tilde{P}_{ij}^k > 0 \mid j \in \Omega,\ k \in K(i)\}, \quad i \in \Omega.$$

Then, $\hat{\rho} = [U^N x^o]_{\min}$, where the operator U is defined by:

$$(2.3.7) \qquad Ux_i = \min_{k\in K(i)} \textstyle\sum_j \tilde{P}_{ij}^k x_j, \quad i \in \Omega;\ x \in E^N.$$

Observe from the analogue of (1.4.1) that

$$\hat{\rho} = [U^N x^o]_{\min} \geq [U^{N-1} x^o]_{\min} \geq \ldots \geq x^o{}_{\min}$$

so that

$$\hat{\rho} = \min\{\tilde{P}_{ij}^k \mid \tilde{P}_{ij}^k > 0;\ i,j \in \Omega,\ k \in K(i)\}$$

is another lower bound on $\tilde{\alpha}$ (it may however be worthwhile to do a number of iterations with the U-operator on x^o, in order to obtain a better approximation of $\tilde{\alpha}$).

If the employed approximation for $\tilde{\alpha} << 1$, then the bounds of cor.2.3.2 part (a) will not be sharp, and the test of part (c) will not be met unless sp[x-v] is very close to zero, namely when $x = y(n)$ and $n >> 1$. Hence, if $\hat{\rho} << 1$, the bounds and the test will only be important near the very end of the calculations. In addition one should observe that N represents the worst case behaviour for the number of steps needed for contraction, which is enormously high, compared with the empirical fact that in most cases $J = 1$ or 2 (cf. e.g. [120] and [122]).

Alternatively, one might want to use the test of part (c) in combination with a device, given recently by HASTINGS [54] in order to eliminate actions on a provisional rather than on a permanent basis.

REMARK 4. Hastings' test works as follows. Let

$$g(n,i,k) = \tilde{Q}y(n-1) - \tilde{q}_i^k - \sum_j \tilde{P}_{ij}^k \, y(n-1)_j \geq 0; \; \phi(n) = \theta_U(n) - \theta_L(n),$$

and

$$H(m,n,i,k) = g(n,i,k) - \sum_{c=n}^{m-1} \phi(c), \; m > n.$$

Then, action $k \in K(i)$ is non-optimal at value iteration stage m, if $H(m,n,i,k) > 0$ (for some $n < m$).

We observe that theorem 2 of [54] holds unconditionally, for every (multichain) MDP, i.e. there is a stage after which no nonoptimal action will pass the above test. This is an immediate consequence of the geometric convergence result in section 6 of chapter 1. However, whereas the *identification* of non-optimal actions is possible in the unichain case, using the above value-iteration scheme and cor.2.3.2 part (c), this is (so far) infeasible for the general *multichain* case.

CHAPTER 3

Nonstationary Markov Decision Problems with converging parameters

3.1. INTRODUCTION AND SUMMARY

In chapter 1, while discussing value-iteration in discounted and undiscounted MDPs, we assumed that all of the parameters of the model were known, perfectly and in advance.

In a large number of applications, however, these parameters can only be obtained via approximating schemes, or otherwise it is computationally preferable to approximate the parameters rather than employing exact algorithms for their computation.

In this chapter we distinguish between the set $K(i)$ representing the *finite* set of *all* (feasible and non-feasible) alternatives in state i ($i \in \Omega$), and the set $K(i) \subseteq K(i)$, the set of all *feasible* alternatives.

So, as a basic assumption, we will suppose throughout this chapter that the parameters q_i^k, p_{ij}^k and the sets $K(i)$ ($i \in \Omega$, $k \in K(i)$) are unknown in advance, but that instead one can compute sequences

$$(3.1.1) \quad \{K(i,n)\}_{n=1}^{\infty} \to K(i);\ i \in \Omega \text{ where } K(i,n) \subseteq K(i),\ i \in \Omega;\ n \geq 1$$

$$(3.1.2) \quad \{q_i^k(n)\}_{n=1}^{\infty} \to q_i^k;\ i \in \Omega;\ k \in K(i)$$

$$(3.1.3) \quad \{p_{ij}^k(n)\}_{n=1}^{\infty} \to p_{ij}^k;\ i,j \in \Omega,\ k \in K(i), \text{ where}$$

$$p_{ij}^k(n) \geq 0 \text{ and } \sum_j p_{ij}^k(n) = 1;\ i,j \in \Omega;\ k \in K(i);\ n \geq 1.$$

The following three examples illustrate that this situation occurs in a large number of applications:

EXAMPLE 1. MDPs in which e.g. the one-step rewards q_i^k appear as the optimal values of underlying optimization problems. As an example, consider a resource or inventory system which serves to supply (say) n simultaneous users.

At each period of time, one has to decide upon the amount to be withdrawn from the system, as well as upon the optimal way to allocate this amount among the n users. With i representing the inventory level (in the resource system) and k the amount to be withdrawn from the latter, the one-step net benefit q_i^k may be obtained by subtracting a holding cost function h(i) and a transfer cost T(k) from the benefit to the entire system that is associated with an optimal allocation of k units among the users. The latter may e.g. be computed by solving a mathematical program so that q_i^k could e.g. have the following structure

$$(3.1.4) \qquad q_i^k = - h(i) - T(k) + \max c(x)$$
$$\text{s.t. } x \in X$$
$$f(x) \leq k$$
$$x \geq 0$$

where x_i (i = 1,...,n) represents the amount allocated to the i-th user, and where the constraints, $x \in X$, describe the restrictions imposed by the other resources and by the technological structure.

There are various reasons for avoiding the computation of all of the q_i^k ($i \in \Omega$, $k \in K(i)$) *prior* to solving the MDP:

(a) in many applications, exact solution methods for the mathematical program in (3.1.4) are either non-available or hardly feasible, i.e. one needs or prefers to employ an approximation method, like a Lagrangean technique, a gradient projection method, or a reduced gradient method. Rather than first solving the $\sum_{i=1}^{N} \|K(i)\|$ mathematical programs with these approximation methods and next using ε- approximations for the q_i^k when solving the MDP - in case a good stopping criterion for the algorithms that solve the mathematical programs is at all available - one would prefer to use the approximating schemes for the q_i^k, in a method which *simultaneously* solves the MDP.

(b) For the actions that turn out to be suboptimal, which in general represents the vast majority of the total number of $\sum_{i=1}^{N} \|K(i)\|$, there is no need to do the computational effort of calculating the associated one-step expected rewards precisely.

In any method which generates approximating schemes for the numbers $\{q_i^k \mid i \in \Omega, k \in K(i)\}$ and *simultaneously* solves the MDP, one could stop the schemes associated with those actions that a test procedure detects to be suboptimal.

We recall from chapter 1 that suboptimality tests of this kind have been derived in connection with the value-iteration method, both for the discounted and undiscounted version of the model. With respect to the former we referred to GRINOLD [50], HASTINGS and MELLO [55], MACQUEEN [81] and PORTEUS [94]; and as far as the latter is concerned, we recalled that a device for *temporary* elimination of suboptimal actions was proposed by HASTINGS [54], which, although originally stated for the unichained case, may be applied to the general multichain model (cf. remark 4 in chapter 2).

In addition, for the unichained undiscounted case, a test for *permanent* elimination of actions was derived in section 3 of chapter 2. All of these elimination procedures can be adapted straightforwardly for the case, where rather than applying value-iteration to a MDP with *exact* knowledge of the expected rewards, one would use upper and lower bounds that ultimately converge to the latter.

Note that most of the approximation techniques, mentioned above for solving the mathematical programs in (3.1.4) have the special feature that, whenever convergence occurs, the *rate* of convergence is at least *geometric*, where a sequence $\{x(n)\}_{n=1}^{\infty}$ is said to *converge to* x^* *geometrically* if there exist numbers $K > 0$, and $0 \leq \lambda < 1$ such that

$$(3.1.5) \qquad \|x(n) - x^*\| \leq K \lambda^n, \quad n = 0,1,2,\ldots$$

(cf. e.g. sections 11.5 and 11.7 in LUENBERGER [80], as well as a recent survey on the subject by GOFFIN [49]).

As examples of the above described model, we refer to RUSSEL [102], VERKHOVSKY [128] and VERKHOVSKY and SPIVAK [129].

EXAMPLE 2. MDPs are generally used for describing dynamic systems which have to be controlled on a periodic basis and the design of which is assumed to be given. In many applications, however, one faces the problem of simultaneously having to make a one-time decision with respect to one or more design parameters, as well as finding an optimal policy for operating the system, once having been constructed. Usually both the laws of motion and the operating costs of the system are heavily affected by the choice of the design parameters. In mathematical terms, the problem amounts to solving:

$$(3.1.6) \qquad \min_{\alpha_0 \leq \alpha \leq \alpha_1} [V_i(\alpha) + \phi(\alpha)]$$

where α represents a scalar or vector of design parameters. In the discounted version of the model, $V_i(\alpha^*)$ would represent the minimal expected total discounted operating costs, when the initial state of the system is i, and $\phi(\alpha^*)$ the design costs, when choosing $\alpha = \alpha^*$. Similarly, in the undiscounted version of the model, $V_i(\alpha^*)$ would denote the minimal long run average operating cost when starting in state i, and $\phi(\alpha^*)$ the depreciation and interest costs of the investment that is needed to implement the design parameters α^*. Note that the one-step rewards and transition probabilities in the MDP depend upon α, i.e.

$$(3.1.7) \qquad q_i^k \overset{\text{def}}{=} q_i^k(\alpha); \; p_{ij}^k \overset{\text{def}}{=} p_{ij}^k(\alpha); \qquad i,j \in \Omega; \; k \in K(i),$$

(3.1.6) may be considered as an unconstrained optimization problem with respect to α, which is well-defined under obvious continuity assumptions. Note that the optimal value of a MDP is not necessarily differentiable with respect to its parameters, and even it it is, the derivatives are extremely hard to compute.

As a consequence, one will have to confine oneself to *direct search methods*, like the Fibonacci method or the simplex method (cf. MURRAY [87]). Note that each evaluation of the objective function in (3.1.6) requires the solution of a MDP which is extremely expensive. On the other hand, in most direct search methods, one is, at each step of the algorithm merely interested in the *relative* order of the values of the objective function in a number of points, i.e. one can quit calculating the component $V_i(\alpha)$ for some trial point α, as soon as it becomes clear that α is suboptimal. We recall that when solving the MDP via value-iteration, both in the discounted model (cf. MACQUEEN [81], PORTEUS [94]) and in the undiscounted unichain case (cf. ODONI [89]) an upper bound on $V_i(\alpha)$ may be calculated that converges to $V_i(\alpha)$ as the number of iterations tends to infinity. Hence, suboptimality of any point α may be detected after a finite number of steps, after which the search procedure may be continued by starting the evaluation of the objective function in (3.1.6) for a different choice of α.

The above considerations lead to a proposal for solving the entire problem (3.1.6) by a single value-iteration scheme in which the parameters $q_i^k(.)$ and $p_{ij}^k(.)$ are adapted in accordance with the search procedure and ultimately converge to the parameter values corresponding with the optimal value of α.

Note that most direct search methods have the property of locating the optimum at a *geometric* rate, such that in general the approximations

for the parameters $q_i^k(.)$ and $P_{ij}^k(.)$ will converge to the desired values at a geometric rate as well (cf. the proposition on p.130 in LUENBERGER [80]).

For a more detailed description of the proposed method we refer to the appendix in this chapter.

EXAMPLE 3. Solving nested sequences of (piecewise linear) functional equations where each functional (vector)-equation has the structure of the optimality equation of an undiscounted MDP or Markov Renewal Program (cf. (1.4.15)):

$$(3.1.8)\qquad \begin{array}{l} x(0)_i = \max\limits_{k\in K^0(i)} [a_i^k(0) + \sum_j P_{ij}^k\, x(0)_j], \quad i \in \Omega \\ \vdots \\ x(m)_i = \max\limits_{k\in K^m(i)} [a_i^k(m) + \sum_j P_{ij}^k\, x(m)_j], \quad i \in \Omega \\ \vdots \\ x(r)_i = \max\limits_{k\in K^r(i)} [a_i^k(r) + \sum_j P_{ij}^k\, x(r)_j], \quad i \in \Omega \end{array}$$

where $K^r(i) \subseteq \ldots \subseteq K^m(i) \subseteq \ldots \subseteq K^0(i)$, $1 \leq m \leq r-1$, and where the quantities $a_i^k(m)$ and the sets $K^m(i)$ both depend upon $x(0),\ldots,x(m-1)$ i.e. upon the solution of the first m functional equations in the sequence (3.1.8). A sequence of nested equations of this type occurs e.g. when trying to find the maximal gain rate vector or some of the higher terms in the Laurent series expansion of the maximal total discounted return vector in powers of the interest rate; and accordingly, when trying to locate maximal gain policies, or policies that are optimal under more selective (sensitive discounted or average overtaking) optimality criteria (cf. VEINOTT [127], MILLER and VEINOTT [85], DENARDO [22]). For a more detailed specification of the sequence (3.1.8) and for a characterization of the solution set, we refer to chapter 4.

In view of the dependence of the sets $K^m(i)$ and the quantities $a_i^k(m)$ on the solution to the previous m equations in (3.1.8) one conceivable way of solving the m+1-st equation is by computing these sets and quantities beforehand with the help of an *exact* solution method (Linear Programming, or the Policy Iteration Algorithm, cf. DENARDO [22] and VEINOTT [127]). However, when the state space becomes large, exact solution methods become infeasible, and a successive approximation method is needed to solve the entire system; moreover, even when exact methods can still be applied, their use may none the less invoke numerical instability problems (cf. chapter 4). Such a successive approximation method will be developed in chapter 4,

where a sequence of value-iteration schemes is simultaneously generated in order to solve the entire system of equations (3.1.8). The schemes that aim at finding a solution to the m+1-st equation, have $a_i^k(m)$ and the sets $K^m(i)$ replaced by approximating sequences $\{a_i^k(m)[n]\}_{n=1}^{\infty}$ and $\{K^m(i)[n]\}_{n=1}^{\infty}$ which are distilled from the schemes that aim at finding a solution to the previous equations, and which have the property of converging to the correct quantities and sets.

All of the schemes involved may be interpreted as value-iteration schemes for undiscounted MDPs, *the parameters of which are replaced by approximating sequences.*

Moreover, here again, the sequences $\{a_i^k(m)[n]\}_{n=1}^{\infty}$ may be constructed in such a way that

(3.1.9) $a_i^k(m)[n] \to a_i^k(m)$, *geometrically* as $n \to \infty$; $i \in \Omega$, $k \in K^o(i)$, $m = 0,\ldots,r$

and the successive approximation method will be shown to converge to a solution of the entire system (3.1.8) at a geometric rate as well.

In this chapter we study, the working of value-iteration for the case where the parameters of the MDP have to be approximated, i.e. where at the n-th stage, they are substituted by the currently available approximations $q_i^k(n)$, $P_{ij}^k(n)$ and $K(i,n)$.

For the discounted version of the model, geometric convergence can easily be obtained in the general multichain case, as is briefly shown in section 2. No assumptions are made with respect to the type of convergence in (3.1.2) and (3.1.3).

For the undiscounted version, we henceforth assume:

(3.1.10) (GEO) $\{q_i^k(n)\}_{n=1}^{\infty} \to q_i^k$, *geometrically;* $i \in \Omega$, $k \in K(i)$

$\{P_{ij}^k(n)\}_{n=1}^{\infty} \to P_{ij}^k$, *geometrically;* $i,j \in \Omega$, $k \in K(i)$

which was satisfied in all of our examples.

A modified value-iteration method is shown to exhibit geometric convergence for the general multichain model, in case only the rewards and actions sets have to be approximated.

If the transition probabilities are to be approximated as well, more care is required since in this case the study of non-stationary Markov

chains is involved. So far, this topic has only been studied under the unichainedness assumption (for a survey, cf. PAZ [91], SENETA [113] and ISAACSON and MADSEN [67]). Under this assumption we establish geometric convergence of our value-iteration method as well. The undiscounted model is dealt with in section 4, and in the appendix we specify our algorithm for the models, mentioned in example 2.

First however, we derive in section 3, a new result on non-stationary Markov Chains, which will be needed in the subsequent analysis. The results in this chapter have been distilled from FEDERGRUEN and SCHWEITZER [36], but for section 3 which is distilled from FEDERGRUEN [33].

3.2. THE DISCOUNTED MODEL

In the discounted version of the model with discount factor $0 \leq \beta < 1$, we consider the following iterative scheme:

$$(3.2.1) \qquad v(n+1)_i = Q(n)v(n)_i, \quad i \in \Omega$$

where $v(0) \in E^N$ may be chosen arbitrarily, and where the $Q(n)$-operators are defined by:

$$(3.2.2) \qquad Q(n)x_i = \max_{k \in K(i,n)} [q_i^k(n) + \beta \sum_j p_{ij}^k(n) x_j], \quad i \in \Omega.$$

One easily verifies that the $Q(n)$-operators satisfy the property (cf. (1.2.10)):

$$(3.2.3) \qquad \text{(a)} \quad \beta[x-y]_{min} \leq [Q(n)x-Q(n)y]_{min} \leq [Q(n)x-Q(n)y]_{max} \leq \beta[x-y]_{max},$$

so that

$$\text{(b)} \quad \|Q(n)x - Q(n)y\| \leq \beta\|x-y\|.$$

Finally define $Q^{(n)}$ by $Q^{(n)}x = Q(n)\ldots Q(1)x; \; x \in E^N$.

THEOREM 3.2.1. $v(n) \to v^*$, *geometrically, where* v^* *is the maximal total discounted return vector, i.e.* v^* *is the unique solution to* $v = Qv$ (cf. (1.2.2)).

PROOF. Let M be such that $|q_i^k(n)| \leq M$ for all $i \in \Omega$, $k \in K(i)$, $n = 1,2,\ldots$ where $M < \infty$ follows from (3.1.2). Verify that $\|v(n)\| \leq M \sum_{\ell=0}^{\infty} \beta^{\ell} = M/(1-\beta)$ for all $n \geq 1$, and conclude that $\{v(n)\}_{n=1}^{\infty}$ is a bounded sequence. Let $\{v(n_k)\}_{k=1}^{\infty}$ and $\{v(m_k)\}_{k=1}^{\infty}$ be two convergent subsequences with limit vectors

v^o and v^{oo} respectively. It is no restriction to assume that $n_k > m_k$ for all $k \geq 1$. Apply (3.2.3) repeatedly to conclude that

$$(3.2.4) \quad \|v(n_k) - v(m_k)\| \leq \beta^{m_k} \|v(n_k - m_k) - v(0)\|$$

and let k tend to infinity in order to verify that $\|v^o - v^{oo}\| = 0$ in view of the second factor to the right of (3.2.4) being bounded. Hence $\{v(n)\}_{n=1}^{\infty}$ converges and its limit vector satisfies the optimality eqaution $v = Qv$, which implies $\lim_{n\to\infty} v(n) = v^*$. Finally to show that the rate of convergence is geometric, replace m_k in (3.2.4) by a fixed integer m, and let k tend to infinity so as to conclude that

$$(3.2.5) \quad \|v(m) - v^*\| \leq \beta^m \|v^* - v(0)\|. \qquad \square$$

The set of all optimal policies can be obtained in the same way as in the stationary model (cf. chapter 1 section 3). In case the parameters q_i^k and P_{ij}^k are both approached from below and from above, all of the bounds on v^*, stopping criteria for ε-approximations or ε-optimal policies, as well as tests for eliminating suboptimal actions, that were found for the stationary model, can be adapted in a straighforward manner.

3.3. ON NON-STATIONARY MARKOV CHAINS WITH CONVERGING TRANSITION MATRICES

In the subsequent analysis for the undiscounted version of this model, a characterization will be needed of the asymptotic behaviour of backwards matrix products of the type:

$$(3.3.1) \quad P(m+n) \ldots P(m),$$

as both n, and m tend to infinity. Here $\{P(m)\}_{m=1}^{\infty}$ is a non-stationary N-state Markov chain, with

$$(3.3.2) \quad \lim_{m\to\infty} P(m) = P$$

Matrix products of the type (3.3.1) are strongly related to the *forward* products, known as inhomogeneous Markov chains, and studied in an extensive literature that started with the papers by HAJNAL [51] (cf.[68], [91], [113] for a survey of the present state of the art). Other than in the context of this chapter, the backward matrix products arise e.g.

(a) in estimate modification processes, where n individuals each of whom has an estimate of some unknown quantity, enter information exchanges which lead them to readapt their estimates in an (infinite) sequence of iterations (cf. DE GROOT [17], and CHATTERJEE and SENETA [14] and DALKEY [16])

(b) in non-stationary Markov Decision Processes when analyzing the total reward in a planning period of n epochs as n tends to infinity (cf. MORTON and WECKER [86], and BOWERMAN [12]).

Let U(r,k) be the stochastic matrix defined by

$$(3.3.3) \qquad U(r,k) = P(r+k)\ldots P(r+1), \quad k = 1,2,\ldots;\ r = 0,1,\ldots$$

The sequence $\{P(k)\}_{k=1}^{\infty}$ is said to be ergodic (in a backwards direction) if

$$(3.3.4) \qquad \lim_{k\to\infty} U(r,k) = \underline{1}d(r)', \qquad r \geq 0$$

where d(r) is obviously a probability vector, i.e. $d(r) \geq 0$ and $\sum_i d(r)_i = 1$. Ergodicity of $\{P(k)\}_{k=1}^{\infty}$ was shown in CHATTERJEE and SENETA [14] (th.5 and corollary) for the case where P is aperiodic and unichained, and can equally be obtained by a mere adaptation of the proof of th.1 in ANTHONISSE and TIJMS [1]. Also in these papers the convergence in (3.3.4) was shown to be geometrical. Hence we have:

LEMMA 3.3.1. *Assume that* P *is unichained and aperiodic. Then* $\lim_{k\to\infty} U(r,k) = \underline{1}\, d(r)'$, *geometrically.*

Note that the rate of convergence of $\{U(r,k)\}_{k=1}^{\infty}$ is independent of the rate at which $\{P(k)\}_{k=1}^{\infty}$ approaches P. In this section we characterize the asymptotic behaviour of $\{d(r)\}_{r=1}^{\infty}$. First, however, example 4 shows that d(r) may heavily depend upon P(r), the first matrix in the product.

For any N × N-stochastic matrix Q and for j = 1,...,N let

$$(3.3.5) \qquad M_j(Q) = \max_i Q_{ij} \quad \text{and} \quad m_j(Q) = \min_i Q_{ij}$$

and note from the identity $Q(2)Q(1)_{ij} = \sum_k Q(2)_{ik}\, Q(1)_{kj}$, that for any pair Q(1), Q(2) of stochastic matrices:

$$(3.3.6) \qquad M_j(Q(2)Q(1)) \leq M_j(Q(1)) \quad \text{and} \quad m_j(Q(2)Q(1)) \geq m_j(Q(1));\ j = 1,\ldots,N.$$

A matrix is said to be strictly positive, if all of its entries are strictly

positive.

EXAMPLE 4. In this example we show that $d(r)$ is strictly positive whenever $P(r)$ is. In other words, whenever $P(r) > 0$ and P has transient states, $d(r) \neq \pi$ where π is the (unique) stationary probability distribution associated with the matrix P.

To verify the implication $P(r) > 0 \Rightarrow d(r) > 0$, note from (3.3.6) that $m_j(P(r)) \leq m_j(U(r,k)) \leq M_j(U(r,k)) \leq M_j(P(r))$ for $r,k = 1,2,\ldots$ and $j = 1,\ldots,N$. Conclude that for all $i = 1,\ldots,N$: $d(r)_j = \lim_{k\to\infty} U(r,k)_{ij} \geq m_j(P(r)) > 0$.

Before characterizing the asymptotic behaviour of $\{d(r)\}_{r=1}^{\infty}$, we first need to introduce the following notions. First for any matrix $A = [A_{ij}]$, let its norm be given by

$$(3.3.7) \qquad \|A\| = \max_i \sum_j |A_{ij}|$$

and define its delta coefficient $\delta(A)$ by

$$(3.3.8) \qquad \delta(A) = 1 - \min_{i,k} \sum_{j=1}^{N} \min(A_{ij}, A_{kj})$$

(which is one minus the ergodic coefficient, cf. e.g. [67], p.144). The following lemma recalls a number of elementary properties of the norm and delta - coefficient:

LEMMA 3.3.2.

(a) *For any pair of matrices,* $\|AB\| \leq \|A\| \, \|B\|$.

(b) *If* A *and* B *are stochastic matrices, then* $\delta(AB) \leq \delta(A)\delta(B)$.

(c) *If* A *is any matrix with* $A\underline{1} = \underline{0}$, *and* B *is a stochastic matrix,* $\|AB\| \leq \|A\| \, \delta(B)$.

(d) *For any aperiodic and unichained (stochastic) matrix* A,

(1) $0 \leq \delta(A^n) < 1$ *for all* $n \geq \frac{1}{2}N(N+1)$

(2) $\sum_{\ell=0}^{\infty} \delta(A^{\ell}) < \infty$.

PROOF.

(a) Note that $\sum_j |a_{ij} \, b_{ij}| \leq (\sum_j |a_{ij}|)(\sum_j |b_{ij}|)$,

(b) cf. lemma V. 2.3 in [67];

(c) Cf. lemma V. 2.4 in [67]; (d)(1) immediate from the combination of lemma 4.1 and th.4.4 in chapter II of [91]. To verify (d)(2) note that in view of parts (b) and (d)(1):

$$\sum_{\ell=0}^{\infty} \delta(A^{\ell}) \leq \sum_{\ell=0}^{\infty} \delta(A^{\frac{1}{2}N(N+1)})^{\lfloor 2\ell N^{-1}(N+1)^{-1} \rfloor} < 1$$

where $[x]$ indicates the largest integer less than or equal to x. □

The next theorem shows that $\underline{1}\pi'$ appears as the limit matrix when both r and k tend to infinity in the matrix product $U(r,k)$; in addition the rate of convergence is specified. Related results for forward products were recently obtained in HUANG, ISAACSON & VINOGRADE [65].

Let $\{\varepsilon_k\}_{k=1}^{\infty}$ be a non-increasing sequence of positive numbers such that $\|P(k)-P\| \le \varepsilon_k$, $k = 1,2,\ldots$.

THEOREM 3.3.3.

$$\lim_{r\to\infty}\lim_{k\to\infty} U(r,k) = \lim_{r\to\infty} \underline{1}d(r)' = \underline{1}\pi', \tag{3.3.9}$$

and

$$\|d(r)-\pi\| = O(\varepsilon_r). \tag{3.3.10}$$

PROOF. We first prove that there exists a scalar $K > 0$, such that

$$\|U(r,k)-P^k\| \le K\varepsilon_r \quad \text{for all } r,k = 1,2,\ldots, \text{ i.e.} \tag{3.3.11}$$

$$\|U(r,k)-P^k\| = O(\varepsilon_r) \text{ as } r,k \to \infty.$$

In view of lemma 3.3.2, part (d)(2) it is sufficient to show:

$$\|U(r,k)-P^k\| \le \varepsilon_r\left(1 + \sum_{\ell=1}^{k-1} \delta(P^{\ell})\right) \text{ for all } r,k = 1,2,\ldots \tag{3.3.12}$$

Fix $r \ge 1$. Note that (3.3.12) holds for $k = 1$ and assume it holds for some k. Then,

$$\|U(r,k+1)-P^{k+1}\| = \|P(r+k+1)U(r,k)-P(r+k+1)P^k + P(r+k+1)P^k-P^{k+1}\|$$

$$\le \|P(r+k+1)\|\,\|U(r,k)-P^k\| + \|P(r+k+1)-P\|\,\delta(P^k)$$

$$\le \varepsilon_r\left(1 + \sum_{\ell=1}^{k-1} \delta(P^{\ell})\right) + \varepsilon_{r+k+1}\delta(P^k) \le \varepsilon_r\left(1 + \sum_{\ell=1}^{k} \delta(P^{\ell})\right)$$

where the first inequality follows from part (a) and (c) of lemma 3.3.2. This proves (3.3.12) by complete induction with respect to k.

Fix $j = 1,\ldots,N$ and $\delta > 0$ and recall from the aperiodicity and unichainedness of P that there exists an integer $n \ge 1$ such that

$$p_{ij}^n - \delta \le \pi_j \le p_{ij}^n + \delta; \quad i = 1,\ldots,N.$$

Hence,

$$M_j(P^n) - \delta \le \pi_j \le m_j(P^n) + \delta. \tag{3.3.13}$$

Use (3.3.11) with $k = n_o$ and the fact that both $M_j(.)$ and $m_j(.)$ are Lipschitz continuous functions on the set of all N × N-matrices, to conclude that,

$$(3.3.14)\quad |M_j(U(r,n_o)) - M_j(P^{n_o})| = O(\varepsilon_r);\ |m_j(U(r,n_o)) - m_j(P^{n_o})| = O(\varepsilon_r).$$

Insert (3.3.14) into (3.3.13) to conclude that

$$(3.3.15)\quad M_j(U(r,n_o)) - O(\varepsilon_r) - \delta \leq \pi_j \leq m_j(U(r,n_o)) + O(\varepsilon_r) + \delta.$$

Next one verifies by a repeated application of (3.3.6) and in view of the fact that $M_j(.)$ and $m_j(.)$ are Lipschitz-continuous, that for all $r = 1,2,\ldots$ $\{M_j(U(r,k))\}_{k=1}^{\infty}$ and $\{m_j(U(r,k))\}_{k=1}^{\infty}$ are resp. monotonically non-increasing and non-decreasing towards $M_j(\underline{1}d(r)') = m_j(\underline{1}d(r)') = d(r)_j$. In particular we have for all $r = 1,2,\ldots$:

$$(3.3.16)\quad m_j(U(r,n_o)) \leq d(r)_j \leq M_j(U(r,n_o))$$

and insert (3.3.16) into (3.3.15) to conclude that for all $\delta > 0$

$$(3.3.17)\quad d(r)_j - O(\varepsilon_r) - \delta \leq \pi_j \leq d(r)_j + O(\varepsilon_r) + \delta$$

and hence

$$|d(r)_j - \pi_j| = O(\varepsilon_r). \qquad \square$$

Finally, example 5 below shows that the upperbound for the rate of convergence of $\{d(r)\}_{r=1}^{\infty}$ towards π is the sharpest possible one:

EXAMPLE 5. Let

$$P(k) = \begin{bmatrix} \tfrac{1}{2} + \alpha_k & \tfrac{1}{2} - \alpha_k \\ \tfrac{1}{2} + \alpha_k & \tfrac{1}{2} - \alpha_k \end{bmatrix} \text{ where } \{\alpha_k\}_{k=1}^{\infty} \downarrow 0.$$

Verify that $U(r,k) = P(r)$ for all $k = 1,2,\ldots$, such that $\underline{1}d(r)' = \lim_{k\to\infty} U(r,k) = P(r)$. Conclude that $\{d(r)\}_{r=1}^{\infty}$ approaches π at the same rate as is exhibited by the convergence of $\{P(k)\}_{k=1}^{\infty}$ towards P (or alternatively by the rate of convergence of $\{\alpha_k\}_{k=1}^{\infty}$ towards zero).

3.4. THE UNDISCOUNTED MODEL

In chapter 1 we showed that in the stationary case, where all parameters are known and available in advance, the following value-iteration

scheme is used to locate maximal gain policies

$$(3.4.1) \qquad v(n+1)_i = \max_{k\in K(i)}[q_i^k + \sum_j P_{ij}^k\, v(n)_j], \quad i \in \Omega.$$

Assuming henceforth that all policies have aperiodic tpm's, $\{v(n)-ng^*\}_{n=1}^{\infty}$ was shown to converge geometrically for any choice of the scrap value vector $v(0) \in E^N$.

This *aperiodicity assumption* may be made without loss of generality since in section 8 of chapter 1, the data-transformation (1.8.1) and (1.8.2), with the choice $\sigma = 1$, was exhibited to turn every MDP into an *equivalent one* in which every policy is *aperiodic*. In this context, two undiscounted MDPs were defined to be equivalent, if they have the same state - and action spaces and if every policy has the same gain rate vector, such that the two problems have the same maximal gain rate vector and the same set of maximal gain policies. In addition it was pointed out that V, the set of solutions to the average return optimality equation (1.4.15) in the original model, and $\tilde{V}$, the corresponding set in the transformed model, satisfy the simple correspondence $\tilde{V} = \{v \in E^N | \tau\, v \in V\}$ (cf. section 8 of chapter 1).

Finally we recall from (1.8.21) and (1.8.22):

$$(3.4.2) \qquad g(n) = v(n) - v(n-1) \to g^*, \text{ geometrically}$$

$$y(n) = nv(n-1) - (n-1)v(n) \to v \in V, \text{ geometrically.}$$

In case only approximations of the parameters are available it seems natural to consider the following iterative scheme:

$$(3.4.3) \qquad x(n+1)_i = \max_{k\in K(i,n)}[q_i^k(n) + \sum_j P_{ij}^k(n)\, x(n)_j], \quad i \in \Omega$$

with $x(0) \in E^N$ arbitrarily chosen.

That is, we modify the classical value iteration method, merely in the sense that at each iteration, the unknown data of the problem are substituted by their current guesses.

With each policy $f \in S_p$ we associate the approximating tpm's $P(f;n)$ and reward vectors $q(f;n)$, $n \geq 1$:

$$(3.4.4) \qquad P(f;n)_{ij} = P_{ij}^{f(i)}(n); \quad i,j \in \Omega; \quad n = 1,2,\ldots$$

$$q(f;n)_i = q_i^{f(i)}(n); \quad i \in \Omega; \quad n = 1,2,\ldots .$$

A unified analysis for both the case

(I) the general multichain model where only the rewards and action sets have to be approximated, and

(II) the unichained model, where *all* of the parameters of the MDP have to be approximated,

is possible in view of

(3.4.5) $\lim_{m\to\infty} \lim_{n\to\infty} P(f;n+m)\ldots P(f;m) \to \Pi(f), \quad f \in S_P$

Note that (3.4.5) holds for case I), in view of $\lim_{n\to\infty} P^n(f) = \Pi(f)$, for aperiodic matrices $P(f)$, and for case II) in view of th.3.3.3.

We next present the main result of this chapter:

THEOREM 3.4.1. *Assume condition* (GEO) *as well as* (3.4.5) *to hold for every policy* f. *Then* (*with* $\{x(n)\}_{n=1}^{\infty}$ *satisfying* (3.4.3)):

(3.4.6) $\{x(n)-ng^*\}_{n=1}^{\infty} \to v \in V$, *geometrically*. □

The proof of this theorem is provided by the following lemmas. First, fix $v^o \in V$ and let $e(n) = x(n)-ng^*-v^o$. Choose numbers $K > 0$, and $0 \leq \lambda < 1$, such that:

(3.4.7) $$|q_i^k(n) - q_i^k| \leq K\lambda^n; \quad i \in \Omega,\ k \in K(i);\ n = 1,2,\ldots$$

$$|P_{ij}^k(n) - P_{ij}^k| \leq K\lambda^n; \quad i,j \in \Omega,\ k \in K(i);\ n = 1,2,\ldots .$$

LEMMA 3.4.2. $\{x(n)-ng^*\}_{n=1}^{\infty}$ *is bounded.*

PROOF. The proof of this lemma is related to the one given in th.5.1 of [110]. Fix $f \in X_i\ L(i,v^o)$ (cf. (1.4.16) and (1.4.21)). Then, in view of $L(i,v^o) \subseteq L(i)$, $i \in \Omega$:

$$x(n+1)_i-(n+1)g_i^*-v_i^o \geq q(f;n)_i + \textstyle\sum_j P(f;n)_{ij}\{x(n)_j-ng_j^*-v_j^o\}$$

$$- q(f)_i + \textstyle\sum_j [P(f;n)_{ij}-P(f)_{ij}][ng_j^*+v_j^o],$$

i.e.

(3.4.8) $$e(n+1)_i \geq - K\lambda^n - NK\lambda^n(n\|g^*\|+\|v^o\|) + \textstyle\sum_j P(f;n)_{ij}e(n)_j$$

$$\geq - K\lambda^n - NK\lambda^n(n\|g^*\|+\|v^o\|) + e(n)_{\min}$$

By iterating (3.4.8) n times, we obtain for all $i \in \Omega$:

$$(3.4.9)\quad e(n+1)_{min} \geq -K(1+N\|v^o\|)\sum_{\ell=0}^{n}\lambda^n - KN\|g^*\|\sum_{\ell=0}^{n}\ell\lambda^{\ell} + e(0)_{min}$$

$$\geq -\frac{K(1+N\|v^o\|)}{(1-\lambda)} - \frac{KN\|g^*\|\lambda}{(1-\lambda)^2} + e(0)_{min}.$$

To show that $\{e(n)\}_{n=1}^{\infty}$ is bounded from above as well, let $a_i^k = \sum_j P_{ij}^k g_j^* - g_i^*$, $i \in \Omega$, $k \in K(i)$. Use (1.4.18) and (3.4.7) to note (cf. also remark 1 in [111]):

$$(3.4.10)\quad e(n+1)_i = \max_{k\in K(i)}\{na_i^k + b(v^o)_i^k + \sum_j P_{ij}^k e(n)_j$$

$$\sum_j [P_{ij}^k(n) - P_{ij}^k]x_j + q_i^k(n) - q_i^k\}$$

$$\leq \max_{k\in K(i)}\{na_i^k + b(v^o)_i^k\} + e(n)_{max} + NK\lambda^n\|x\| + K\lambda^n$$

Next use $a_i^k < 0$ for $k \in K(i)\backslash L(i)$, $i \in \Omega$ to conclude that there exists an integer $n_0 \geq 1$ such that for all $n \geq n_0$ the first term to the right of the inequality (3.4.10) is achieved for $k \in L(i)$ and hence vanishes (cf. (1.4.15) and (1.4.18)). By iterating (3.4.10) one concludes that for all $n \geq n_0$:

$$(3.4.11)\quad e(n+1)_{max} \leq e(n_0)_{max} + K(N\|x\|+1)\sum_{\ell=n_0}^{n}\lambda^{\ell}$$

$$\leq e(n_0)_{max} + K(N\|x\|+1)/(1-\lambda)$$

(3.4.11) together with (3.4.9) establish the lemma. □

LEMMA 3.4.3. $\{x(n)-ng^*\}_{n=1}^{\infty} \to v^o \in V$.

PROOF. The proof of lemma 3.4.3 has the same structure as the one given for th.5.1 in [110]. Define $x_i = \liminf_{n\to\infty}[x(n)-ng^*]_i$ and $X_i = \limsup_{n\to\infty}[x(n)-ng^*]_i$, $i \in \Omega$. Let f_n, satisfy the N maxima in (3.4.3). From lemma 3.4.2, it follows that $-\infty < x_i \leq X_i < \infty$ for all $i \in \Omega$. Recall the equality part in (3.4.10):

$$e(n+1)_i = \max_{k\in K(i)}\{na_i^k + b(v^o)_i^k + \sum_j P_{ij}^k e(n)_j + \sum_j [P_{ij}^k(n) - P_{ij}^k]x_j + q_i^k(n) - q_i^k\}.$$

Observe using lemma 3.4.2 and (3.1.3) that the left side of this equality is bounded, just like all of the terms to its right side, with the possible exception of the first one. Conclude that for all n sufficiently large, only alternatives $k \in L(i)$ achieve the maximum in (3.4.3), or

(3.4.12) $f_n \in X_i\ L(i)$, for all n sufficiently large

whereas in addition,

$$(3.4.13)\quad x(n+1)_i-(n+1)g_i^* = \max_{k\in L(i)}\{q_i^k(n)-g_i^* + \sum_j P_{ij}^k(n)[x(n)_j-ng_j^*] + n\sum_j(P_{ij}^k(n)-P_{ij}^k)g_j^*\},\quad i\in\Omega.$$

Fix $i \in \Omega$, take (sub)sequences $\{n_k\}_{k=1}^{\infty}$ (with $\lim_{k\to\infty} n_k = \infty$), such that $\lim_{k\to\infty}[x(n_k)-n_kg^*]$ exists and $\lim_{k\to\infty}[x(n_k+1)-(n_k+1)g^*]_i = x_i$ (or X_i resp.) which is feasible in view of lemma 3.4.2. Replace n by n_k in (3.4.13), let k tend to infinity and note that $\lim_{n\to\infty} n(P_{ij}^k(n) - P_{ij}^k) = 0$ in view of assumption (GEO), in order to conclude

$$(3.4.14)\quad x_i \geq \max_{k\in L(i)}[q_i^k-g_i^* + \sum_j P_{ij}^k\, x_j],\quad i\in\Omega$$

$$(3.4.15)\quad X_i \leq \max_{k\in L(i)}[q_i^k-g_i^* + \sum_j P_{ij}^k\, X_j],\quad i\in\Omega.$$

We next show that

$$(3.4.16)\quad x_i = X_i,\quad \text{for all } i\in R^*.$$

In order to verify that $x_i = X_i$ for all $i \in \Omega\backslash R^*$ as well, i.e. in order to prove convergence of $\{x(n)-ng^*\}_{n=1}^{\infty}$ in *each* of its components, one can next apply the proof of part (d) of th.5.1 in [110] starting with equations (5.4) and (5.5).

Finally let n tend to infinity in (3.4.13) to verify that $v = \lim_{n\to\infty} x(n)-ng^*$ satisfies the optimality equation (1.4.15), i.e. $v \in V$. To prove (3.4.16), fix $i \in R^*$, i.e. let $i \in R(f)$ for some $f \in X_j\ L(j,v^O)$ (cf. (1.4.21)). Iterate (3.4.8) to get:

$$(3.4.17)\quad e(n+m)_i \geq -K\sum_{\ell=m}^{n+m-1}\lambda^{\ell} - NK\sum_{\ell=m}^{n+m-1}\lambda^{\ell}(\ell\|g^*\|+\|v^O\|) + P(f;n+m-1)\ldots P(f;m)e(m)_i,\quad i\in\Omega.$$

Since $\{e(n)\}_{n=1}^{\infty}$ is bounded, it has at least one limit point. Let y, z be two limit points of the sequence. Take sequences $\{n_k\}_{k=1}^{\infty}$, and $\{m_k\}_{k=1}^{\infty}$ with $\lim_{k\to\infty} n_k = \lim_{k\to\infty} m_k = \infty$, such that $\lim_{k\to\infty} e(m_k) = y$ and $\lim_{k\to\infty} e(n_k+m_k) = z$. Replace n and m in (3.4.17) by n_k and m_k, let k tend to infinity, and use (3.4.5) to conclude:

$$(3.4.18)\quad z \geq \lim_{k\to\infty}(-K\sum_{\ell=m_k}^{\infty}\lambda^{\ell} - NK\sum_{\ell=m_k}^{\infty}\lambda^{\ell}(\ell\|g^*\|+\|v^O\|)) + \Pi(f)y = \Pi(f)y$$

where the equality in (3.4.18) follows from $\lim_{n\to\infty}\sum_{\ell=n}^{\infty} a_\ell = \sum_{\ell=1}^{\infty} a_\ell - \lim_{n\to\infty}\sum_{\ell=1}^{n} a_\ell = 0$, for any converging series $\sum_\ell a_\ell$.

In a similar way, we have $y \geq \Pi(f)y$, i.e. $y_i = \sum_j \Pi(f)_{ij} y_j$, $i \in R(f)$, as may be verified by multiplying the inequality by $\Pi(f) \geq 0$ and using $\Pi(f)^2 = \Pi(f)$ (cf. (1.4.5)).

We conclude that

$z_i \geq y_i$, $i \in R(f)$, and in a similar way,
$y_i \geq z_i$, $i \in R(f)$, such that
$y_i = z_i$, $i \in R(f)$, which proves convergence of $\{x(n)-ng^*\}_{n=1}^{\infty}$ on R^*,

and hence (3.4.16). □

We conclude the proof of theorem 3.4.1 by establishing the *geometric* rate of convergence of $\{x(n)-ng^*\}_{n=1}^{\infty}$:

Proof of theorem 3.4.1. Let $v = \lim_{n\to\infty} x(n)-ng^* \in V$ (cf. lemma 3.4.3) and define $e(n) = x(n) - ng^* - v$; $n \geq 1$. It follows from (3.4.12) that for all n sufficiently large, and all $i \in \Omega$:

$$(3.4.19)\quad e(n+1)_i = \max_{k\in L(i)}\{q_i^k(n)-q_i^k + \sum_j P_{ij}^k(n)[x(n)_j - ng_j^* - v_j] + n\sum_j (P_{ij}^k(n)-P_{ij}^k)g_j^* + \sum_j (P_{ij}^k(n)-P_{ij}^k)v_j + b(v)_i^k\}.$$

Note that with the possible exception of the last term in (3.4.19), all of the terms converge to zero, as n tends to infinity. Hence there exists an integer $n_0 \geq 1$, such that for all $n \geq n_0$, only alternatives $k \in L(i,v)$ attain the maxima in (3.4.19). That is, using (1.6.9) and the equality $\sum_j P_{ij}^k(n)e(n)_j = \sum_j P_{ij}^k e(n)_j + \sum_j (P_{ij}^k(n)-P_{ij}^k)e(n)_j$, one easily verifies:

$$(3.4.20)\quad \|e(n+1) - U(v)e(n)\| \leq K(1+N\|v\|)\lambda^n + KN\|g^*\|n\lambda^n + KN\lambda^n\|e(n)\|.$$

We conclude in view of lemma 3.4.2, that there exist numbers $A,B > 0$ with

$$(3.4.21)\quad \|e(n+1)-U(v)e(n)\| \leq (A+Bn)\lambda^n, \quad n \geq n_0.$$

Verify next, by complete induction with respect to m, that:

$$(3.4.22)\quad \|e(n+m)-U(v)^m e(n)\| \leq \sum_{\ell=n}^{n+m-1}(A+B\ell)\lambda^\ell, \quad n \geq n_0.$$

For, note that if (3.4.22) holds for all $n \geq n_0$ and some $m \geq 1$, then

$$\|e(n+m+1) - U(v)^{m+1}e(n)\| = \|e(n+m+1) - U(v)^{m}e(n+1)\|$$

$$+ \|U(v)^{m}e(n+1) - U(v)^{m+1}e(n)\|$$

$$\leq \sum_{\ell=n+1}^{n+m}(A+B\ell)\lambda^{\ell} + \|e(n+1) - U(v)e(n)\| \leq \sum_{\ell=n}^{n+m}(A+B\ell)\lambda^{\ell}$$

in view of the property $\|U(v)x - U(v)y\| \leq \|x-y\|$ for all $x,y \in E^N$ (cf. (1.4.1)). Next, define $U(v)^{\infty}x = \lim_{n\to\infty} U(v)^{n}x$ and note that the existence of $U(v)^{\infty}x$, for any $x \in E^N$, follows as a special case of lemma 3.4.3 with $q_i^k(n) = 0$ and $P_{ij}^k(n) = P_{ij}^k$ for all $i,j \in \Omega$, $k \in K(i)$ (cf. also section 6 of chapter 1). Let m tend to infinity in (3.4.22) and conclude that for all n sufficiently large:

$$(3.4.23)\quad \|U(v)^{\infty}e(n)\| \leq \sum_{\ell=n}^{\infty}(A+B\ell)\lambda^{\ell} \leq \frac{A\lambda^{n}}{1-\lambda} + \frac{2Bn\lambda^{n}}{|\ell n\lambda|}$$

where the second inequality follows from $x\lambda^{x}$ being monotonically non-increasing for $x > -(\ell n\ \lambda)^{-1}$, such that $\sum_{\ell=n}^{\infty}\ell\lambda^{\ell} \leq \int_{n-1}^{\infty}x\lambda^{x}dx = \frac{(n-1)\lambda^{n-1}}{|\ell n\ \lambda|} + \frac{\lambda^{n-1}}{(\ell n\ \lambda)^{2}}$, for n sufficiently large.

We next recall from [111] that there exists a *contraction factor* Γ, with $0 \leq \Gamma < 1$ and a number $M > 0$ such that for all $x \in E^N$:

$$(3.4.24)\quad \|U(v)^{n}x - U(v)^{\infty}x\| \leq M\Gamma^{n}, \quad n = 1,2,\ldots .$$

Finally, use (3.4.22) with m = n, and conclude:

$$-\sum_{\ell=n}^{2n-1}(A+B\ell)\lambda^{\ell} + [U(v)^{n}e(n) - U(v)^{\infty}e(n)] + U(v)^{\infty}e(n) \leq e(2n) \leq$$

$$+\sum_{\ell=n}^{2n-1}(A+B\ell)\lambda^{\ell} + [U(v)^{n}e(n) - U(v)^{\infty}e(n)] + U(v)^{\infty}e(n)$$

whence we obtain, using (3.4.23) and (3.4.24) that for all n sufficiently large

$$(3.4.25)\quad \|e(2n)\| \leq \sum_{\ell=n}^{\infty}(A+B\ell)\lambda^{\ell} + \|U(v)^{\infty}e(n)\| + \|U(v)^{n}e(n) - U(v)^{\infty}e(n)\|$$

$$\leq \frac{2A(\sqrt{\lambda})^{2n}}{1-\lambda} + \frac{4Bn(\sqrt{\lambda})^{2n}}{|\ell n\lambda|} + M(\sqrt{\Gamma})^{2n};$$

whereas a similar *geometric* bound may be obtained for the odd members of the sequence $\{e(n)\}_{n=1}^{\infty}$. □

APPENDIX

In this appendix we describe the algorithm, we propose for solving the models mentioned in example 2. Assuming that the functions $q_i^k(\alpha)$ and $P_{ij}^k(\alpha)$ and $\phi(\alpha)$ are continuous in α (cf. (3.1.6) and (3.1.7)), the function to be minimized in (3.1.6) is guaranteed to be continuous in α in the discounted version, whereas in the undiscounted version some additional requirements on the chain structure of the tpm's of the policies in S_p have to be imposed (continuity is e.g. guaranteed in the unichain case; cf. SCHWEITZER [105]). In the absence of these requirements on the chain structure, $V_i(\alpha)$ can still be shown to be piecewise continuous, with a finite number of discontinuities, and an obvious modification of the below described algorithm can be employed:

step 0: Initialize Max := $-\infty$ and $x \in E^N$. Fix $\alpha^{best} := \alpha^{new} := \alpha^{old} \in [\alpha_0, \alpha_1]$ and $\varepsilon > 0$

Step 1: $x := \max_{k \in K(i)} [q_i^k(\alpha^{new}) + \beta \sum_j P_{ij}^k(\alpha^{new}) x_j]$, $i \in \Omega$
and compute lower and upper bounds on $V(\alpha^{new})$ as a function of x:
$L(\alpha^{new}) \leq V(\alpha^{new}) \leq U(\alpha^{new})$

step 2: "If" $U(\alpha^{new}) + \phi(\alpha^{new}) <$ MAX, "then" {α^{new} is suboptimal; $\alpha^{old} := \alpha^{new}$ and choose α^{new} according to a specifically chosen unconstrained search procedure; go to step 5}

step 3: "If" $U(\alpha^{new}) - L(\alpha^{new}) < \varepsilon$, "then" {MAX := $L(\alpha^{new}) + \phi(\alpha^{new})$; $\alpha^{best} := \alpha^{new}$, i.e. α^{new} is the best parameter choice so far; $\alpha^{old} := \alpha^{new}$; choose α^{new} according to the unconstrained search procedure; go to step 5}

step 4: go back to step 1, and execute the next iteration

step 5: "if" $\|\alpha^{old} - \alpha^{new}\| < \varepsilon$ "then" go to "END" "else" go back to step 1, and execute the next iteration with the adapted parameters

"END" Use α^{best} as an ε-optimal parameter choice, and $\frac{1}{2}(U(\alpha^{best}) + L(\alpha^{best}))$ as an ε-approximation of the value of the entire problem.

CHAPTER 4

Successive approximation methods for solving nested functional equations in Markov Decision Theory

4.1. INTRODUCTION AND SUMMARY

This chapter is concerned with sequences of *nested* functional equations of the following structure:

$$
\begin{aligned}
(4.1.1)\quad x_i^{(0)} &= \max_{k\in K^0(i)}\Big[a_i^k(0) + \textstyle\sum_{j=1}^N p_{ij}^k\, x_j^{(0)}\Big], \quad i \in \Omega\\
x_i^{(1)} &= \max_{k\in K^1(i)}\Big[a_i^k(1) - \textstyle\sum_j H_{ij}^k(1)\, x_j^{(0)} + \sum_j p_{ij}^k\, x_j^{(1)}\Big], \quad i \in \Omega\\
&\vdots\\
x_i^{(m)} &= \max_{k\in K^m(i)}\Big[a_i^k(m) - \textstyle\sum_j H_{ij}^k(m)\, x_j^{(m-1)} + \sum_j p_{ij}^k\, x_j^{(m)}\Big], \quad i \in \Omega\\
&\vdots\\
x_i^{(n)} &= \max_{k\in K^n(i)}\Big[a_i^k(n) - \textstyle\sum_j H_{ij}^k(n)\, x_j^{(n-1)} + \sum_j p_{ij}^k\, x_j^{(n)}\Big], \quad i \in \Omega.
\end{aligned}
$$

$\Omega = \{1,\ldots,N\}$ denotes the finite *state space* of the decision problem. For all $i \in \Omega$, $K^0(i)$ is a given finite set of *alternatives* in state i, and for fixed i, the sets $K^m(i)$ are *nested* subsets of $K^0(i)$, i.e. $K^m(i) \subseteq K^{m-1}(i)$, $m = 1,\ldots,n$. The numbers p_{ij}^k, $H_{ij}^k(m)$ $(i,j \in \Omega;\ k \in K^0(i);\ m = 1,\ldots,n)$ are assumed to be nonnegative:

$$(4.1.2)\quad p_{ij}^k \geq 0;\ H_{ij}^k(m) \geq 0 \quad (i,j \in \Omega;\ k \in K^0(i);\ m = 1,\ldots,n)$$

where in addition

$$(4.1.3)\quad \sum_{j=1}^N p_{ij}^k = 1, \quad (i \in \Omega;\ k \in K^0(i))$$

and where the numbers $H_{ij}^k(m)$ satisfy condition (CLO) to be stated below. For all $i \in \Omega$, and $k \in K^0(i)$, the quantities $a_i^k(0)$ and $a_i^k(1)$ are given constants. For $m \geq 2$, the $a_i^k(m)$ are given *affine* functions of $x^{(0)},\ldots,x^{(m-2)}$.

$$(4.1.4)\quad a_i^k(m) = \beta_i^k(m) + \sum_{\ell=0}^{m-2} \langle b_i^{\ell;k}(m);\ x^{(\ell)}\rangle, \quad m \geq 2$$

with $\beta_i^k(m) \in E^1$ and $b_i^{\ell;k}(m)$ a given N-component row vector.

Finally, $K^m(i) \subseteq K^{m-1}(i)$ represents the set of alternatives which attain the maxima in the first m-1 functional equations. (For a more precise definition we refer to section 2).

Sequences of nested functional equations of this type arise e.g. in Markov Renewal Programs (MRPs) in which one wants to find the maximal gain rate vector g^*, the maximal bias vector z^*, or any of the higher order terms of the Laurent series expansion in powers of the interest rate of the maximal total discounted return vector (cf. MILLER and VEINOTT [85], DENARDO [22]) as well as policies that are maximal gain, bias-optimal or optimal under more selective discounted- or average overtaking optimality criteria (cf. VEINOTT [127], DENARDO [22] and SLADKÝ [116]).

In the three cases a system of resp. two, three and four or more equations arises, i.e. the three cases correspond with $n = 1$, $n = 2$ and $n \geq 3$ resp.

For finding the maximal gain rate vector in multichain MRPs, three methods exist:

(1) the Policy Iteration Algorithm (PIA) (cf. HOWARD [63], JEWELL [69])
(2) Linear Programming (LP) formulations: (cf. MANNE [84], DENARDO & FOX [23])
(3) successive approximation methods (cf. chapter 1).

For finding the maximal bias vector or more generally to solve a set of n+1 nested functional equations, only the first two of the three above mentioned methods have been generalized: the former by VEINOTT [127] and the latter by DENARDO [22] who proposed a decomposition of the problem into a sequence of Linear Programs combined with a number of search procedures. These two methods are impractical for large problems and it would be desirable to fill the hiatus, by generalizing the third method of successive approximations, since this method is the only practical one. This objective is precisely the purpose of this chapter.

As a special case we present a successive approximations scheme to find the optimal bias vector and a bias-optimal policy. HORDIJK and TIJMS [60] established a scheme which finds z^*-though not a bias-optimal policy- for the special class of discrete-time Markov Decision Problems (MDP's) with a maximal gain rate that is independent of the initial state of the system. Our scheme finds both the optimal vector and the optimal policy for all multichain MRP's and generalizes to the higher order functional

equations.

In addition it is worth mentioning that a wide range of stochastic decision models varying from Markov Decision Chains with multiplicative utilities (cf. HOWARD and MATHEMSON [64]) to controlled branching processes (cf. MANDL [83] and PLISKA [92]) may be formulated as so-called Multiplicative Markov Decision Chains (MMDC's), as was pointed out by ROTHBLUM [99]. The latter generalize the ordinary Markov Decision Problems in the sense that the transition matrices are merely required to be nonnegative, rather than (sub)stochastic (cf. (4.1.3)).

ROTHBLUM [99] showed that in these Multiplicative Markov Decision Chains, even when restricting attention to a "first order" criterion, a sequence of up to N nested functional equations arises, which satisfies the above described structure perfectly but for the (sub)stochasticity assumption (4.1.3).

Although the specific successive approximation method presented in this paper uses the entire structure as given by (4.1.1) up to (4.1.4), including the (sub)stochasticity assumption, the basic ideas underlying our approach will be needed when establishing an approximation for the general MMDC-case.

In section 2 we give some notation and preliminaries. In section 3 we summarize the properties of "single-equation" value-iteration schemes, as established in chapter 1 and 3 and as far as needed in the remainder.

In section 4, we first show why any approximation method has to solve the entire sequence simultaneously rather than each of the equations successively; next, we present our method for solving a *pair* of *consecutive* functional equations in the sequence (4.1.1), i.e. a pair of equations of the structure:

$$(4.1.5)\qquad x_i^* = \max_{k\in K(i)}\left[b_i^k + \sum_{j=1}^{N} P_{ij}^k\, x_j^*\right], \quad i \in \Omega$$

$$(4.1.6)\qquad y_i^* = \max_{k\in M(i,x^*)}\left[c_i^k - \sum_j H_{ij}^k\, x_j^* + \sum_{j=1}^{N} P_{ij}^k\, y_j^*\right], \quad i \in \Omega$$

where $K(i) \subset K^o(i)$, $i \in \Omega$ and where, for each solution x of the equation (4.1.5), the set $M(i,x)$ is defined as:

$$(4.1.7)\qquad M(i,x) = \{k \in K(i) \mid k \text{ attains the maximum in (4.1.5) for the solution } x\}.$$

Finally, in section 5 it is shown how this method may be generalized to solve the entire sequence and next it is pointed out to which (simplified) algorithms, this method reduces in a number of special cases. The results in this chapter have been distilled from FEDERGRUEN and SCHWEITZER [37].

4.2. NOTATION AND PRELIMINARIES

The following notation will be employed. For any policy $f = (f(1),f(2),\ldots,f(N)) \in X_{i=1}^{N} K^{O}(i)$ and $m = 0,\ldots,n$ we define N-vectors $a(f;m) = [a_i^{f(i)}(m)]$; $\beta(f;m) = [\beta_i^{f(i)}(m)]$; $b(f) = [b_i^{f(i)}]$ and $c(f) = [c_i^{f(i)}]$, as well as the N×N matrices $b^{\ell}(f;m) = [b_i^{\ell;f(i)}(m)]$; $H(f)$ and $H(f;m)$ $(m = 1,\ldots,n)$:

$$(4.2.1)\qquad H(f)_{ij} = H_{ij}^{f(i)};\quad i,j \in \Omega$$

$$H(f;m)_{ij} = H_{ij}^{f(i)}(m);\quad i,j \in \Omega;\quad m = 1,\ldots,n.$$

We assume that the numbers $H_{ij}^{k} \geq 0$ and $H_{ij}^{k}(m) \geq 0$ $(m = 1,\ldots,n)$ satisfy the closedness assumption (CLO)

$$(4.2.2)\qquad \text{(CLO)(a)}\quad T_i^k = \sum_{j=1}^{N} H_{ij}^{k} > 0;\quad i \in \Omega,\ k \in K^{O}(i)$$

$$\sum_{j=1}^{N} H_{ij}^{k}(m) > 0;\quad i \in \Omega,\ k \in K^{O}(i),\ m = 1,\ldots,n$$

(b) For any policy $f \in X_{i=1}^{N} K^{O}(i)$ and any subchain C of the tpm $P(f)$, C is closed for $H(f)$ and $H(f;m)$ $(m = 1,\ldots,n)$, i.e. if $i \in C$ then $H(f)_{ij} = 0$, and $H(f;m)_{ij} = 0$ if $j \notin C$

Assumption (CLO) is satisfied in all MRPs where $H_{ij}^{k}(m) = H_{ij}^{k}$; $i,j \in \Omega$; $k \in K^{O}(i)$; $m = 1,\ldots,n$ and where T_i^k represents the expected holding time in state i, when using alternative $k \in K^{O}(i)$, and where either $H_{ij}^{k} = \delta_{ij}$ or $H_{ij}^{k} = P_{ij}^{k}\,\tau_{ij}^{k}$ with $\tau_{ij}^{k} \geq 0$, denoting the expected *conditional* holding time in state i, when using alternative $k \in K^{O}(i)$ and given state j is the next state to be observed. We finally define the N-vector $T(f) = [T_i^{f(i)}]_{i=1}^{N}$, for any $f \in X_j K^{O}(j)$ (cf. chapter 1, section 9).

The following theorem recalls the basic characterization of the solution set to the pair of functional equations (4.1.5) and (4.1.6) (cf. [112]).

THEOREM 4.2.1.

(a) *The* 2N *functional equations* (4.1.5) *and* (4.1.6) *have a solution pair* $\{x^*,y^*\}$ *if and only if*

$$(4.2.3)\quad \max_{f\in X_{j=1}^{N} K^o(j)} \Pi(f)b(f)_i = 0;\quad i \in \Omega.$$

If this condition is met then $S_{MG} = \{f \in X_j K^o(j) \mid \Pi(f)b(f) = 0\}$ *is non-empty and* x^* *is unique and given by*

$$(4.2.4)\quad x_i^* = \max_{f\in S_{MG}} x(f)_i,\quad i \in \Omega$$

where

$$(4.2.5)\quad x(f)_i = Z(f)b(f)_i + \sum_{m=1}^{n(f)} \Phi_i^m(f)\, \frac{\langle \pi^m(f), c(f)-H(f)Z(f)b(f)\rangle}{\langle \pi^m(f), T(f)\rangle},\quad i \in \Omega.$$

Moreover there exist policies $f \in S_{MG}$ *which attain the* N *maxima in* (4.2.4) *simultaneously, i.e. with* $x(f) = x^*$

(b) *if* (4.2.3) *is met then the* y^**-part of the solution pair is not unique, e.g. if* y^* *satisfies* (4.1.6), *then so does* $y^*+d\underline{1}$ *for any scalar* d. □

We observe that (4.2.3) is the necessary and sufficient condition for the existence of a solution to the single (vector-) equation (4.1.5) as well. So in other words, th.4.2.1 expresses that a solution of the pair of equations (4.1.5) and (4.1.6) exists if and only if the single (vector-) equation (4.1.5) has a solution (cf. [112]).

Part (b) of the above theorem shows that the set $Y = \{y^* \in E^N \mid (x^*,y^*)$ satisfy (4.1.5) and (4.1.6)$\}$ is unbounded. For a more detailed characterization of this set we refer to [109]. We next return to the system (4.1.1). Henceforth assuming that

$$(4.2.5)\quad \max_{f\in X_j K^o(j)} \Pi(f)a(f;0)_i = 0,\quad i \in \Omega$$

let $S_{MG}^{(0)} = \{f \in X_j K^o(j) \mid \Pi(f)a(f;0) = 0\}$. For any policy $f \in S_{MG}^{(0)}$, consider the system of linear equations:

$$(4.2.6)\quad \begin{cases} x(f;0) = a(f;0) + P(f)x(f;0) \\ x(f;m) = a(f;m) - H(f;m)x(f,m-1) + P(f)x(f;m);\ m = 1,\ldots,n. \end{cases}$$

COROLLARY 4.2.2. *For all* $f \in S_{MG}^{(0)}$, *the system* (4.2.6) *has a solution. Moreover, all the vectors* $x(f;0),\ldots,x(f;n-1)$ *are uniquely determined, whereas only the last and* n+1-st *equation in* (4.2.6) *has an unbounded solution set.*

PROOF. We prove the corollary by complete induction with respect to n. For n = 1 the assertion follows as a special case of th. 4.2.1; next assume it holds for some value of n, and extend the system (4.2.6) with a n+2-nd vector-equation: $x(f;n+1) = a(f;n+1)-H(f;n+1)x(f;n)+P(f)x(f;n+1)$. Consider the *subsystem* constituted by the n+1-st and n+2-nd vector-equation. In view of the first n+1 vector-equations in (4.2.6) having a solution, it follows by multiplying both sides of the n+1-st equation with $\Pi(f)$, that

$$\Pi(f)[a(f;n)-H(f;n)x(f;n-1)] = 0. \tag{4.2.7}$$

Using (4.2.7) as well as the fact that a(f;n) and a(f;n+1) are uniquely determined in the *extended* system- which follows from (4.1.4) and the induction assumption- we conclude by applying th.4.2.1 to the above mentioned (single policy) subsystem that the *extended* system has a (n+2)-tuple of solution vectors in which $x(f;0),\ldots,x(f;n)$ are uniquely determined. Note finally that the vector x(f;n+1) which only appears in the last vector-equation of the system, is *not* uniquely determined, since any multiple of $\underline{1}$ can be added to it. □

We next define recursively:

$$x_i^{*(\ell)} = \max_{f\in S_{MG}^{(\ell)}} x(f;\ell)_i; \quad i \in \Omega; \quad \ell = 0,\ldots,n-1 \tag{4.2.8}$$

with

$$S_{MG}^{(\ell)} = \{f \in S_{MG}^{(\ell-1)} \mid x(f;\ell-1) = x^{*(\ell-1)}\} \quad \text{for } \ell = 1,\ldots,n-1.$$

The following theorem, the proof of which goes along lines with that of corollary 4.2.2, extends the basic results obtained in th.4.2.1 to the system of functional equations (4.1.1).

THEOREM 4.2.3. *The system* (4.1.1) *has* n+1*-tuples of solution vectors; moreover* $x^{(0)},\ldots,x^{(n-1)}$ *are uniquely determined by* $x^{(\ell)} = x^{*(\ell)}$ *for* $\ell = 0,\ldots,n-1$ *whereas the last component vector* $x^{(n)}$ *is not unique (if* $[x^{*(0)},\ldots x^{*(n-1)},x^{(n)}]$ *satisfies* (4.1.1), *then so does* $[x^{*(0)},\ldots,x^{*(n-1)},x^{(n)}+c\underline{1}]$ *for any scalar* c). □

Th.4.2.3 enables us to give a precise definition of the sets $K^m(i)$ for m = 1,...,n

$$K^m(i) = \{k \in K^o(i) \mid k \text{ attains the maxima on the right-hand side of the first m functional equations in (4.1.1) for the (uniquely determined) solutions } x^{*(0)},\ldots,x^{*(m-1)}\}. \tag{4.2.9}$$

Lemma 4.2.4 finally gives a basic characterization of the sets $S_{MG}^{(\ell)}$.

LEMMA 4.2.4:

(4.2.10) $S_{MG}^{(\ell)} = \{f \in X_j K^{\ell}(j) | f(i)$ satisfies the $\ell+1$-st vector equation in (4.1.1) for all $i \in R(f)$ and any solution $[x^{(0)},\ldots,x^{(\ell-1)},x^{(\ell)}]$ of the system (4.1.1), $\ell = 0,\ldots,n\}$.

PROOF. Fix $f \in S_{MG}^{(\ell)}$ $(0 \leq \ell \leq n-1)$ and a solution $[x^{(0)},\ldots,x^{(n)}]$ to the system (4.1.1). Note from (4.2.6) and the definition of $S_{MG}^{(\ell)}$ that $f \in X_{j=1}^{N} K^{\ell-1}(j)$. Multiply both sides in (4.2.6) by $\Pi(f)$ to verify that, $\Pi(f)[a(f;\ell)-H(f;\ell)x^{*(\ell-1)}]=0$ and conclude from $f \in X_{j=1}^{N} K^{\ell-1}(j)$, and the inequality $x^{(\ell)} \geq a(f;\ell) - H(f;\ell)x^{(\ell-1)} + P(f)x^{(\ell)}$ that $f(i)$ satisfies the $\ell+1$-st vector equation in (4.1.1) for all $i \in R(f)$. This proves that $S_{MG}^{(\ell)}$ is included within the policy set to the right of (4.2.10). To verify the reversed inclusion fix a policy within the latter set and apply corollary 4.2.2 to the first ℓ equations in (4.2.6) to conclude that $f \in S_{MG}^{(\ell-1)}$. Note in addition that both $x^{*(\ell-1)}$ and $x(f;\ell-1)$ satisfy the equations $y = a(f;\ell-1) - H(f;\ell-1)x^{*(\ell-2)} + P(f)y$, and $\Pi(f)[a(f;\ell) - H(f;\ell)y] = 0$, with the additional convention that $x^{*(-1)} = 0$. It then follows from lemma 1 in [23] that $x^{*(\ell-1)} = x(f;\ell-1)$ i.e. $f \in S_{MG}^{(\ell)}$. □

4.3. SINGLE EQUATION VALUE-ITERATION; A REVIEW

Our successive approximation method consists of a sequence of iterative schemes which are generated simultaneously. Some of these schemes aim at finding either a vector that may be interpreted as the maximal gain rate vector of some Markov Renewal Program, or a solution to the corresponding optimality equation (cf. chapter 1, section 9).

In addition these schemes face in general the additional difficulty that

(a) some (or all) of its one-step expected rewards, and
(b) some (or all) of its action sets,

are unknown in advance since depending upon quantities that have to be approximated simultaneously.

Thus for any Cartesian product space of policies $X_{i=1}^{N} K(i) \subseteq X_{i=1}^{N} K^{o}(i)$, consider the MRP which has

(1) $X_{i=1}^{N} K(i)$ as its policy-space; (2) numbers $\{q_i^k \mid i \in \Omega,\ k \in K(i)\}$ as its one-step expected rewards and (3) the numbers P_{ij}^k, T_i^k $(i,j \in \Omega;\ k \in K(i))$ as the transition probability to state j and the expected holding time when choosing alternative k in state i.

Finally, suppose we wish to obtain an iterative scheme which approximates the maximal gain rate vector g^* in this MRP, or a solution to the corresponding optimality equation (cf. (1.9.8)):

$$(4.3.1) \qquad v_i^* = \max_{k \in L(i)} [q_i^k - g_i^* T_i^k + \sum_j P_{ij}^k v_j^*],\ i \in \Omega$$

where $L(i) = \{k \in K(i) \mid g_i^* = \sum_j P_{ij}^k g_j^*\}$, $i \in \Omega$ (cf. (1.4.16)).

Let $\hat{V}$ denote the solution space of the optimality equation (4.3.1) (cf. (1.9.8)) and for any $v \in \hat{V}$, let

$$(4.3.2) \qquad S(i,v) = \{k \in L(i) \mid k \text{ attains the maximum on the right hand side of (4.3.1) for the solution } v \in \hat{V}\};\ i \in \Omega.$$

Assume in addition that rather than having exact knowledge of the quantities $\{q_i^k \mid i \in \Omega,\ k \in K(i)\}$ and the sets $\{K(i) \mid i \in \Omega\}$ all we have, are:

(1) sequences $\{q_i^k(n)\}_{n=1}^{\infty} \to q_i^k$, *geometrically* as $n \to \infty$; $i \in \Omega$; $k \in K(i)$

(2) sequences $\{K(i,n)\}_{n=1}^{\infty} \to K(i)$, as $n \to \infty$, i.e.

$K(i,n) = K(i)$ for n sufficiently large, say $n \geq n_0$, where n_0 is unknown in advance.

The definition of $\{x(n)\}_{n=1}^{\infty} \to x^*$, *geometrically* was given in (3.1.5). In chapter 3, we derived a value iteration scheme which solves the special case of undiscounted *discrete-time* MDPs with (geometrically) converging parameters, whereas in addition it was pointed out that all of the quantities of interest can be approached at a *geometric* rate. Theorem 4.3.1 below combines this method with the data-transformation (1.9.9), with $\sigma = \tau$, which turns every undiscounted MRP into an *equivalent* undiscounted MDP in which all of the policies are aperiodic. We recall that the aforementioned equivalence notion was defined by (EQUI) (cf. (1.9.6)).

So, th.4.3.1 below considers the scheme

$$(4.3.3) \qquad v(n+1)_i = \max_{k \in K(i,n)} \left[\frac{\tau q_i^k(n)}{T_i^k} + \sum_j [\delta_{ij} + \frac{\tau}{T_i^k} (P_{ij}^k - \delta_{ij})] v(n)_j \right],\ i \in \Omega$$

with

$$(4.3.4) \qquad 0 < \tau < \min\{T_i^k / (1 - P_{ii}^k) \mid (i,k) \text{ with } P_{ii}^k < 1\}$$

and where the starting point $v(0) \in E^N$ may be chosen arbitrarily.

First, we convene that hereafter, $\{\varepsilon_n\}_{n=1}^{\infty}$ will indicate any sequence of positive numbers approaching zero in such a way that $\lambda^n/\varepsilon_n \to 0$ for any λ, $0 \leq \lambda < 1$, e.g. take $\varepsilon_n = n^{-1}$ or the reciprocal of any positive polynomial in n.

THEOREM 4.3.1.

(a) $g(n) = \tau^{-1}\{v(n)-v(n-1)\} \to g^*$, *geometrically, as* $n \to \infty$

(b) $w(n) = nv(n-1)-(n-1)v(n) \to v^o$, *geometrically, as* $n \to \infty$ *where* $v^o \in \hat{V}$.

(c) *For all* $i \in \Omega$, $n = 1,2,\ldots$ *and* $\varepsilon > 0$ *let*

$$A(i,n,\varepsilon) = \{k \in K(i,n) \mid \textstyle\sum_j P_{ij}^k \, g(n)_j \geq g(n+1)_i - \varepsilon\}, \text{ and}$$

$$B(i,n,\varepsilon) = \{k \in K(i,n) \mid \frac{\tau q_i^k(n)}{T_i^k} + \textstyle\sum_j [\delta_{ij} + \frac{\tau(P_{ij}^k - \delta_{ij})}{T_i^k}] v(n)_j \geq v(n+1)_i - \varepsilon\}.$$

Then, $A(i,n,\varepsilon_n) \to L(i)$, *as* $n \to \infty$, $i \in \Omega$ *and*

$B(i,n,\varepsilon_n) \to S(i,v^o)$ *as* $n \to \infty$, $i \in \Omega$

i.e. for all n *sufficiently large,* $A(i,n,\varepsilon_n) = L(i)$ *and* $B(i,n,\varepsilon_n) = S(i,v^o)$ *for all* $i \in \Omega$.

PROOF. Consider the discrete-time MDP with the same state- and policy space and $\tilde{q}_i^k = \tau q_i^k / T_i^k$; $i \in \Omega$, $k \in K(i)$ and $\tilde{P}_{ij}^k = \frac{\tau}{T_i^k}[P_{ij}^k - \delta_{ij}] + \delta_{ij}$; $i,j \in \Omega$ and $k \in K(i)$ which corresponds with the transformation formula (1.9.9) with the choice $\sigma = \tau$. It was pointed out in section 9 of chapter 1, that this transformation turns the MRP into a discrete-time MDP which is equivalent in the (EQUI) sense (cf. (1.9.6)). In addition by choosing τ strictly less than the upper bound in (4.3.4), all of the policies in the transformed MDP are aperiodic (cf. section 1.9). Next, one easily verifies that (4.3.3) corresponds to the scheme (3.4.3) in this transformed MDP. Hence, applying th.3.4.1 and using the equivalence between the original MRP and the transformed MDP, one concludes that for any starting point $v(0) \in E^N$:

(4.3.5) $\quad v(n) - n\tau g^* \to v^o$, *geometrically* where $v^o \in \hat{V}$.

The limit results in part (a) and (b) then follow as in (1.8.21) and (1.8.22) and the results in part (c) are immediate from th.1.7.1. □

We note that in some cases the parameter τ in (4.3.3) may be taken to be *equal* to the upperbound in (4.3.4). In particular, for discrete time MDP's the choice $\tau = 1$ is sometimes allowed for any starting point $v(0) \in E^N$.

We refer to th.1.5.1 for the necessary and sufficient condition of the latter.

Keeping track, at each stage of the iterative scheme (4.3.3), of the sets of ε_n-optimal actions, rather than of the sets of optimal actions, is in general inevitable since it was pointed out in section 1.7 that the sequences $\{A(i,n,0)\}_{n=1}^{\infty}$ and $\{B(i,n,0)\}_{n=1}^{\infty}$, $i \in \Omega$, may have a very irregular behaviour.

We finally recall from section 1.8 (cf. (1.8.20)) that in case $g^* = \langle g^* \rangle \underline{1}$ a simpler and numerically preferable scheme was established by WHITE [131] to separate g^* and v^o.

An important special case of the latter occurs when computing a solution to a functional equation of the type

$$(4.3.6) \qquad v_i^* = \max_{k \in K(i)} \{q_i^k + \sum_j p_{ij}^k v_j^*\}$$

the necessary and sufficient condition for the existence of which is given by $\max_{f \in X_j K(j)} \Pi(f) q(f) = 0$ (cf. th.4.2.1). Note that under this condition, (4.3.6) may be interpreted as the optimality equation of a discrete time MDP with $g^* = 0$, and the following iterative scheme may be applied:

$$(4.3.7) \qquad y(n+1)_i = \max_{k \in K(i,n)} \{\tau q_i^k(n) + \sum_j [\delta_{ij} + \tau(p_{ij}^k - \delta_{ij})] y(n)_j\}, \quad i \in \Omega$$

with $0 < \tau < 1$. Verify that

$$(4.3.8) \qquad \{y(n)\}_{n=1}^{\infty} \to v^o \text{ } \textit{geometrically}, \text{ with } v^o \text{ satisfying (4.3.6).}$$

For discrete-time MDP's in which the maximal gain rate is independent of the initial state of the system, the following adapted version of the "modified value-iteration method" of HORDIJK and TIJMS [60] (cf. section 1.8, in particular (1.8.7)-(1.8.10)) may be used as an alternative to the scheme (4.3.3). In addition this algorithm has the special property of converging to the *optimal bias vector* z^*, defined by (1.8.11), albeit at a considerably slower rate than (4.3.3) exhibits. Since the scheme will merely be needed in the case where $g^* = \underline{0}$, a simplified convergence proof will be given for this special case. The proof goes along the lines of the one given in HORDIJK and TIJMS [60] and the proof for the more general case where $g^* = \langle g^* \rangle \underline{1}$ can be obtained along the lines of [60] as well:

THEOREM 4.3.3. *Let* $T_i^k = 1$; $i \in \Omega$ *and* $k \in K(i)$ *and* $g^* = \underline{0}$. *Assume* $\{q_i^k(n)\}_{n=1}^{\infty} \to q_i^k$, *geometrically* ($i \in \Omega$; $k \in K(i)$). *Consider the scheme*

(4.3.9) $z(n+1)_i = \max_{k\in K(i,n)}[q_i^k(n) + \beta_n \sum_j P_{ij}^k z(n)_j], \quad i \in \Omega$

where $\beta_n = 1-n^{-b}$, *for some* $0 < b \leq 1$. *Then* $\{z(n)\}_{n=1}^{\infty} \to z^*$, *where*

(4.3.10) $\|z(n) - z^*\| = 0(n^{-b}\ln n)$.

PROOF. Let $V(\beta)$ denote the total maximal expected β-discounted reward vector, satisfying the optimality equation:

(4.3.11) $V(\beta)_i = \max_{k\in K(i)}[q_i^k + \beta \sum_j P_{ij}^k V(\beta)_j], \quad i \in \Omega.$

Note from MILLER and VEINOTT [85] that there exists a constant $K > 0$, such that

(4.3.12) $\|V(\beta) - z^*\| \leq K(1-\beta)$, and

(4.3.13) $\|V(\beta_1)-V(\beta_2)\| \leq K|\beta_1-\beta_2|$, for β_1,β_2 close enough to one.

Let $C > 0$ and $0 \leq \lambda < 1$ be such that

(4.3.14) $|q_i^k(n) - q_i^k| \leq C\lambda^n; \; i \in \Omega, \; k \in K(i); \; n = 1,2,\ldots$

Finally let f_n be a policy satisfying the N maxima in (4.3.9) and let g_n be a β_n-optimal policy. Then,

(4.3.15) $L(n)_i \leq z(n+1)_i - V(\beta_n)_i \leq U(n)_i, \quad i \in \Omega$

where

$$L(n)_i = q(g_n;n)_i - q(g_n)_i + \beta_n \sum_j P(g_n)_{ij}[z(n)_j - V(\beta_n)_j]$$

$$U(n)_i = q(f_n;n)_i - q(f_n)_i + \beta_n \sum_j P(f_n)_{ij}[z(n)_j - V(\beta_n)_j].$$

Hence,

$$\|z(n+1) - V(\beta_n)\| \leq C\lambda^n + \beta_n\|z(n)-V(\beta_n)\| \leq C\lambda^n + \beta_n\|z(n)-V(\beta_{n-1})\|$$

$$+ \beta_n\|V(\beta_n)-V(\beta_{n-1})\| \leq C\lambda^n+\beta_n\|z(n) - V(\beta_{n-1})\| + K\beta_n|\beta_n-\beta_{n-1}|$$

for all $n \geq n_0$ (say). Iterating this inequality we obtain for all $m \geq 0$:

$$\|z(n_0+m+1) - V(\beta_{n_0+m})\| \leq C\sum_{j=n_0+1}^{n_0+m}[\beta_{j+1}\cdots\beta_{n_0+m}]\lambda^j$$

$$+ (\beta_{n_0+1}\cdots\beta_{n_0+m})\|z(n_0+1)-V(\beta_{n_0})\| + K\sum_{j=n_0+1}^{n_0+m}[\beta_j\cdots\beta_{n_0+m}]|\beta_j-\beta_{j-1}|$$

with the convention that $\prod_{r=L}^{U} a_r = 1$ if $L > U$. Hence,

$$(4.3.16)\quad \|z(n_0+m+1)-z^*\| \le K(1-\beta_{n_0+m}) + C\sum_{j=n_0+1}^{n_0+m}[\beta_{j+1}\cdots\beta_{n_0+m}]\lambda^j$$

$$+ (\beta_{n_0+1}\cdots\beta_{n_0+m})\|z(n_0+1)-V(\beta_{n_0})\| + K\sum_{j=n_0+1}^{n_0+m}\beta_j\cdots\beta_{n_0+m}|\beta_j-\beta_{j-1}|.$$

This implies that for the choice $\beta_n = 1-n^{-b}$, $0 < b \le 1$ convergence of $\{z(n)\}_{n=1}^{\infty}$ is determined by the last two terms to the right of (4.3.16). To verify convergence as well as its rate, note, using the mean value theorem that $n^b-(n-1)^b \le 1$ for all $n = 1,2,\ldots$ and use this inequality in order to verify that (cf. also chapter 8):

$$\beta_n\cdots\beta_2 = \frac{(n^b-1)}{n^b}\,\frac{((n-1)^b-1)}{(n-1)^b}\cdots\frac{(2^b-1)}{2^b} \le \frac{2^b-1}{n^b}.$$

Next we apply the mean value theorem to verify that

$$\sum_{j=2}^{n}(\beta_j\cdots\beta_n)|\beta_j-\beta_{j-1}| \le$$

$$b\sum_{j=2}^{n}\frac{(n^b-1)}{n^b}\cdots\frac{(j^b-1)}{j^b}(j-1)^{-b-1} \le bn^{-b}\sum_{j=1}^{n-1}j^{-1} = O(n^{-b}\ln n)$$

which determines the convergence rate of $\{z(n)\}_{n=1}^{\infty}$. □

4.4. SOLVING TWO COUPLED FUNCTIONAL EQUATIONS

In this section we wish to solve, by successive approximations, the two coupled vector equations (4.1.5) and (4.1.6). Here b_i^k and c_i^k ($i\in\Omega$, $k\in K(i)$) are assumed to be independent of x^* and y^*. We first treat the case where all of the sets $K(i)$, and all of the parameters b_i^k, c_i^k, H_{ij}^k and P_{ij}^k ($i,j\in\Omega; k\in K(i)$) are known exactly. Later on, we treat the case where instead of knowing the sets $K(i)$ and the parameters b_i^k and c_i^k, one has approximations:

(1) $\{K(i,n)\}_{n=1}^{\infty} \to K(i)$; (2) $\{b_i^k(n)\}_{n=1}^{\infty} \to b_i^k$, geometrically;
(3) $\{c_i^k(n)\}_{n=1}^{\infty} \to c_i^k$, geometrically.

Theorem 4.2.1 shows us that x^*, if it exists, is unique. Our successive approximation scheme decomposes x^* into *three* components which are approximated simultaneously. The first component represents an arbitrary solution ξ^* to (4.1.5) alone. Its computation is accomplished in accordance with the scheme (4.3.7) (cf. th. 4.3.1).

In addition, the difference vector $w^* = x^* - \xi^*$ is decomposed into two components $\bar{g}^*$ and v^*, i.e. $x^* = \xi^* + \bar{g}^* + v^*$. The vector $\bar{g}^*$ can be interpreted as the maximal gain rate vector of a certain MRP and can be found in accordance with a scheme of the type (4.3.3) (cf. th. 4.3.1). Finally the vector v^* satisfies the inequalities $v_i^* \geq 0$, $i \in \Omega$, with strict equality in a non-empty subset of the state space, to be specified, below.

This vector can be interpreted as the solution to an optimal stopping problem, hence it is the minimal solution to a certain functional equation arising in a transient MDP. It will be found by a successive approximation scheme in accordance with (4.3.6) and (4.3.7).

In summary the proposed method generates simultaneously three value-iteration schemes with *geometric* convergence to the three vectors $\xi^*, \bar{g}^*$ and v^*. Then $x^* = \xi^* + \bar{g}^* + v^*$ is part of the desired solution to (4.1.5) and (4.1.6). The non-unique vector y^* in (4.1.6) may finally be determined by interpreting (4.1.6) as a special case of (4.3.6) with $q_i^k = c_i^k - \sum_j H_{ij}^k x_j^*$ and K(i) replaced by $M(i,x^*)$, and by applying th.4.3.1, while geometric estimates of x^* and $M(i,x^*)$ ($i \in \Omega$) are generated.

We finally note that the above decomposition is similar to the one employed by DENARDO [21]. The temptation exists to devise a set of successive approximation schemes which solve the functional equations one after the other; e.g. the first scheme creates a sequence $\{x^{(n)}\}_{n=1}^{\infty}$ approaching some solution $x^{(\infty)}$ of (4.1.5) and sequences $\{M^{(n)}(i)\}_{n=1}^{\infty}$ that approach the action sets $M(i,x^{(\infty)})$. One might try to terminate these sequences after a *finite* number of steps and replace x^* and $M(i,x^*)$, $i \in \Omega$ by the currently available approximations when starting a *second* iterative scheme to solve the *second* functional equation.

There are *three* reasons why this method cannot work. First, th.4.2.1 shows that not any solution $x^{(\infty)}$ of (4.1.5) is acceptable, but only the unique solution x^* which satisfies the second equation (4.1.6) for some $y^* \in E^N$ as well. Second, even in case $\{x^{(n)}\}_{n=1}^{\infty}$ converges to the required solution x^*, it is still possible for $\max_{f \in X_j M(j,x^*)} \Pi(f)[c(f) - H(f)x^{(n)}]$ to be non-vanishing for all $n = 1,2,\ldots$ and as a consequence any successive approximation scheme for solving (4.1.6) in which x^* is replaced by the currently available approximation $x^{(n)}$, will explode. Finally, one doesn't know when - if at all - the sets $M(i,x^*)$ can be replaced by the sets $M^{(n)}(i)$.

We conclude that the equations (4.1.5) and (4.1.6) must be solved by simultaneously computed successive approximation schemes, where the scheme

for y^* uses an *ever-improving* rather than a fixed approximation of x^*. The above problems seemingly did not arise in Denardo's LP-decomposition method (cf. [21]), where one assumes to have solved the first functional equation *exactly* before tackling the next one. In the presence of numerical errors, however, the second equation (4.1.6) may be *insolvable*, since, with $\hat{x}^*$ representing the computed value of x^*, $\max_{f\in X_j M(j,x^*)} \Pi(f)[c(f)-H(f)\hat{x}^*] = 0$ won't hold precisely.

Henceforth assuming that condition (4.2.3) holds, we recall that the first equation (4.1.5) may be considered as an optimality equation of an undiscounted MDP of the special type (4.3.6). We conclude that with the help of the single equation methods presented in the previous section (cf. (4.3.7)), sequences $\{\xi(n)\}_{n=1}^{\infty}$ and $\{\Xi^n(i)\}_{n=1}^{\infty}$, $i \in \Omega$, may be generated with the properties

(4.4.1) $\{\xi(n)\}_{n=1}^{\infty} \to \xi^*$, geometrically, with ξ^* a solution to (4.1.5)
$\{\Xi^{(n)}(i)\}_{n=1}^{\infty} \to M(i,\xi^*)$ as $n \to \infty$.

LEMMA 4.4.1. *Pick* $\xi(0) \in E^N$ *arbitrarily and* $0 < \tau < 1$. *Then the iterative scheme*

$$(4.4.2)\qquad \xi(n+1)_i = \max_{k\in K(i)}\{\tau b_i^k + \sum_{j=1}^N [\tau(P_{ij}^k - \delta_{ij}) + \delta_{ij}]\xi(n)_j\},\ i \in \Omega$$

has the property $\{\xi(n)\}_{n=1}^{\infty} \to \xi^*$, *geometrically, where* ξ^* *satisfies* (4.1.5). *In addition, let*

$$\Xi^{(n)}(i) = \{k \in K(i) \mid b_i^k + \sum_{j=1}^N [\tau(P_{ij}^k - \delta_{ij}) + \delta_{ij}]\xi(n)_j \geq \xi(n+1)_i - \varepsilon_n\},$$
$$i \in \Omega;\ n = 1,2,\ldots$$

Then $\lim_{n\to\infty} \Xi^{(n)}(i) = M(i,\xi^*)$, $i \in \Omega$. □

The next step is finding an expression for $w^* = x^* - \xi^*$. Insert $x^* = \xi^* + w^*$ into (4.1.5) and (4.1.6) to get the new coupled equations:

$$(4.4.3)\qquad w_i^* = \max_{k\in K(i)}\{\bar{b}_i^k + \sum_j P_{ij}^k w_j^*\},\quad i \in \Omega$$

$$(4.4.4)\qquad y_i^* = \max_{k\in M(i;w^*+\xi^*)}\{\bar{c}_i^k - \sum_j H_{ij}^k w_j^* + \sum_j P_{ij}^k y_j^*\}$$

with

$$(4.4.5)\qquad \bar{b}_i^k = b_i^k + \sum_j P_{ij}^k \xi_j^* - \xi_i^*;\quad i \in \Omega,\ k \in K(i)$$

$$\bar{c}_i^k = c_i^k - \sum_j H_{ij}^k \xi_j^*;\quad i \in \Omega,\ k \in K(i).$$

Observe that (4.4.3) and (4.4.4) have the same structure as (4.1.5) and (4.1.6). Note in addition, that

$$\max_{f\in X_j K(j)} \Pi(f)\bar{b}(f) = \max_{f\in X_j K(j)} \Pi(f)[b(f)+(P(f)-I)\xi^*] =$$

$$\max_{f\in X_j K(j)} \Pi(f)b(f) = 0, \text{ with } \bar{S}_{MG} = \{f|\Pi(f)\bar{b}(f) = 0\} = S_{MG}$$

and apply theorem 4.2.1 to (4.4.3) and (4.4.4) to conclude that

$$(4.4.6)\quad w_i^* = \max_{f\in S_{MG}} w(f)_i;\quad i \in \Omega \text{ with}$$

$$w(f)_i = Z(f)\bar{b}(f)_i + \sum_{m=1}^{n(f)} \Phi_i^m(f)\, \frac{\langle \pi^m(f), \bar{c}(f)-H(f)Z(f)\bar{b}(f)\rangle}{\langle \pi^m(f), T(f)\rangle},\ i \in \Omega.$$

For $f \in X_j\ M(j,\xi^*)$, we have $\bar{b}(f) = 0$. This motivates us to define

$$(4.4.7)\quad \bar{g}_i^* = \max_{f\in X_j M(j,\xi^*)} w(f)_i,\quad i \in \Omega$$

with the following properties. First, let

$$(4.4.8)\quad S_{opt} = \{f \in S_{MG} | w(f) = w^*\}.$$

THEOREM 4.4.2.

(a) *For all* $f \in S_{MG}$, $w(f)_i = Z(f)\bar{b}(f)_i + \sum_{m=1}^{n(f)} \Phi_i^m(f)\, \frac{\langle \pi^m(f), \bar{c}(f)\rangle}{\langle \pi^m(f), T(f)\rangle}$, $i \in \Omega$ *where* $Z(f)\bar{b}(f)_i = 0$ *for all* $i \in R(f)$

(b) $\bar{g}_i^* = \max_{f\in X_j M(j,\xi^*)} \sum_{m=1}^{n(f)} \Phi_i^m(f)\, \frac{\langle \pi^m(f), \bar{c}(f)\rangle}{\langle \pi^m(f), T(f)\rangle}$, $i \in \Omega$

(c) $w^* \geq \bar{g}^*$ and $w_i^* = \bar{g}_i^*$ *for all* $i \in R_{opt} = \{j \in \Omega | j \in R(f),\ f \in S_{opt}\}$. □

PROOF.

(a) Fix $f \in S_{MG}$. Verify that $\xi^* \geq b(f)+P(f)\xi^*$, since ξ^* satisfies (4.1.5) and multiply this inequality by $\Pi(f) \geq 0$ to conclude that: $0 = \Pi(f)b(f) = \Pi(f)[\xi^*-b(f)-P(f)\xi^*] \geq 0$, the equality part following from $f \in S_{MG}$. Hence,

$$(4.4.9)\quad \bar{b}(f)_i = 0,\quad i \in R(f)$$

Next, note from assumption (CLO), (cf. (4.4.2)), and (4.4.7) that for all $i \in R(f)$, $H(f)_{ij} > 0$ or $Z(f)_{ij} > 0$ only if $j \in R(f)$, and conclude to part (a) using (4.4.6).

(b) follows from part (a) and the equality $b(f) = 0$ for $f \in X_j\ M(j,\xi^*)$

Part (c): The inequality $w^* \geq \bar{g}^*$ follows from $S_{MG} \supseteq X_j M(j,\xi^*)$. We next recall from th.4.2.1, part (a) that $R_{opt} \neq \emptyset$. To verify that the reversed inequality $w_i^* \leq \bar{g}_i^*$ holds at least for the components $i \in R_{opt}$ as well, fix $f \in S_{MG}$, with $w^* = w(f)$, as well as a state $i \in C^m(f)$ $(1 \leq m \leq n(f))$. Define a policy h such that

$$h(i) = \begin{cases} f(i), \text{ for all } i \in R(f) \\ \in M(i,\xi^*) \text{ otherwise} \end{cases},$$

and note from (4.4.9) that $h \in X_j M(j,\xi^*)$. Conclude that,

$$w_i^* = w(f)_i = <\pi^m(f),\bar{c}(f)>/<\pi^m(f),T(f)>$$
$$= <\pi^m(h),\bar{c}(h)> <\pi^m(h),T(h)> \leq \bar{g}_i^*$$

the second equality following from part (a) and the third one from the fact that R(f) is closed under P(h), thus proving part (c). □

The key observation now, is that theorem 4.4.2, part (b) represents $\bar{g}^*$ as the maximal gain rate vector of a MRP with the Cartesian policy space $X_j M(j,\xi^*)$ and one-step expected rewards $\bar{c}_i^k = c_i^k - \sum_j H_{ij}^k \xi_j^*$, $i \in \Omega$ and $k \in K(i)$. These are not known in advance, because the sets $M(i,\xi^*)$, $i \in \Omega$, and the vector ξ^* are unknown in advance. However both the policy space and the one-step expected rewards can be approximated by resp. (cf. (4.4.1)):

$$(4.4.10) \quad \{\Xi^n(i)\}_{n=1}^{\infty} \text{ and } \{\bar{c}_i^k(n)\}_{n=1}^{\infty} = \{c_i^k - \sum_j H_{ij}^k \xi(n)_j\}_{n=1}^{\infty}, \; i \in \Omega, \; k \in K(i).$$

We now invoke th.4.3.1 to get the following successive approximation method which converges geometrically to $\bar{g}^*$.

THEOREM 4.4.3. *Pick* $h(0) \in E^N$ *and* $0 < \bar{\tau} < \min\{T_i^k/(1-P_{ii}^k) \mid (i,k) \text{ with } P_{ii}^k < 1\}$. *Then the iterative scheme*

$$(4.4.11) \quad \begin{cases} h(n+1)_i = \max_{k \in \Xi(n)(i)} \{\bar{\tau} \dfrac{\bar{c}_i^k(n)}{T_i^k} + \sum_j [\dfrac{\bar{\tau}(P_{ij}^k - \delta_{ij})}{T_i^k} + \delta_{ij}] h(n)_j\}, i \in \Omega \\ \bar{g}(n+1)_i = \bar{\tau}^{-1}\{h(n+1)_i - h(n)_i\}, \quad i \in \Omega \end{cases}$$

has the property $\{\bar{g}(n)\}_{n=1}^{\infty} \to \bar{g}^*$, *geometrically*. □

We finally turn to the evaluation of v^*. Insert $w^* = \bar{g}^* + v^*$ and the inequality $w^* \geq \bar{g}^*$ (cf. th.4.4.2 part (b)) into (4.4.3) to get

$$(4.4.12)\qquad \begin{cases} v_i^* = \max_{k\in K(i)}\{r_i^k + \sum_j p_{ij}^k v_j^*\}, & i \in \Omega \\ v_i^* \geq 0, & i \in \Omega \end{cases}$$

where $r_i^k = \bar{b}_i^k + \sum_j p_{ij}^k \bar{g}_j^* - \bar{g}_i^* = b_i^k + \sum_j p_{ij}^k (\xi_j^* + \bar{g}_j^*) - \xi_i^* - \bar{g}_i^*$. Note that $v_i^* = 0$, for all $i \in R_{opt}$.

As a consequence v^* satisfies the functional equation

$$(4.4.13)\qquad v_i = \max_{k\in K(i)}\{0;\ r_i^k + \sum_j p_{ij}^k v_j\},\ i \in \Omega.$$

The following theorem shows as a key observation that v^* is in fact the *smallest* solution to (4.4.13):

THEOREM 4.4.4. *v^* is the smallest solution to* (4.4.13) *i.e. every solution* v *of* (4.4.13) *has* $v \geq v^*$.

PROOF. Fix a solution $\tilde{v}$ to (4.4.13) and define $z_i = \min\{v_i^*, \tilde{v}_i\}$. Fix $f \in S_{opt}$. Define the vector $r(f)$ by $r(f)_i = r_i^{f(i)}$. Note that both,

$$v_i^* \geq r(f)_i + P(f)v_i^* \geq r(f)_i + P(f)z_i,\quad i \in \Omega,\ \text{and}$$

$$\tilde{v}_i \geq r(f)_i + P(f)\tilde{v}_i \geq r(f)_i + P(f)z_i,\quad i \in \Omega.\ \text{Hence,}$$

$$(4.4.14)\qquad z - r(f) - P(f)z \geq 0.$$

Multiplying (4.4.14) by $\Pi(f) \geq 0$, yields $0 = \Pi(f)([I-P(f)]z - r(f))$, in view of $0 = \Pi(f)r(f) = \Pi(f)b(f) = 0$ (cf. assumption (4.2.3) and note that $f \in S_{opt} \subseteq S_{MG}$). This implies that (4.4.14) is a strict equality for components $i \in R(f)$. Using this and the fact that as a result of (1.4.7) for $j \notin R(f)$, $Z(f)_{ij} \geq 0$ for all i, with $Z(f)_{ij} = 0$ when $i \in R(f)$, we get

$$(4.4.15)\qquad z \geq Z(f)r(f) + \Pi(f)z$$

by multiplying (4.4.14) by $Z(f)$ and invoking (1.4.6). Next note from part (b) of th.4.4.2 that for all $i \in R_{opt}$, hence especially for all $i \in R(f)$, $z_i \geq 0 = v_i^*$. As a consequence, we have $z_i = v_i^*$ for all $i \in R(f)$ and combine this with (4.4.15) to obtain:

$$\begin{aligned} z_i &\geq Z(f)r(f)_i + \Pi(f)v_i^* = Z(f)r(f)_i + \Pi(f)[w(f)_i - \bar{g}_i^*] \\ &= Z(f)r(f)_i + \Pi(f)Z(f)\bar{b}(f)_i + \sum_{m=1}^{n(f)} \Phi_i^m(f)\,\frac{\langle \pi^m(f), \bar{c}(f)\rangle}{\langle \pi^m(f), T(f)\rangle} - \Pi(f)\bar{g}_i^* \\ &= Z(f)\bar{b}(f)_i + \Pi(f)\bar{g}_i^* - \bar{g}_i^* + \Pi(f)\bar{b}(f)_i - \Pi(f)\bar{g}_i^* + \end{aligned}$$

$$+ \sum_{m=1}^{n(f)} \Phi_i^m(f) \frac{<\pi^m(f),\bar{c}(f)>}{<\pi^m(f),T(f)>}$$

$$= w(f)_i - \bar{g}_i^* = w_i^* - \bar{g}_i^* = v_i^*, \quad \text{for } \textit{all } i \in \Omega$$

in view of $\Pi(f)\bar{b}(f) = 0$. Conclude that $\tilde{v} \geq v^*$. □

Functional equations of the type (4.4.13), with a smallest solution arise inter alia in Leontief Substitution Systems (cf. KOEHLER et al. [72]). The equation (4.4.13) may be interpreted as an optimal stopping problem (cf. DERMAN [25], ch.8) where the decision to stop earns a reward 0 from thereon. More formally, append a new stopping state 0 i.e.

(4.4.16) $\bar{\Omega} = \Omega \cup \{0\}$; $\bar{K}(i) = K(i) \cup \{0\}$, $i \in \Omega$ and $\bar{K}(0) = \{0\}$

$$\bar{r}_i^k = \begin{cases} r_i^k & ; i \in \Omega \text{ and } k \in K(i) \\ 0 & ; i \in \bar{\Omega} \text{ and } k = 0 \end{cases}$$

$$p_{ij}^k = \begin{cases} p_{ij}^k & ; k \in K(i); \quad i,j \in \Omega \\ \delta_{j0} & ; k = 0; \quad i,j \in \bar{\Omega} \end{cases}$$

and rewrite (4.4.13) as

(4.4.17) $v_i = \max_{k \in \bar{K}(i)} [\bar{r}_i^k + \sum_{j \in \bar{\Omega}} \bar{p}_{ij}^k v_j]$, $i \in \bar{\Omega}$, with

(4.4.18) $v_0 = 0$.

This is a MDP with $\underline{0}$ as the maximal gain rate vector, and state 0 a trapping state under each policy $f \in S_{MG}$. The following theorem shows that $[v^*,0]$ may be interpreted as the optimal bias vector of this MDP such that th.4.3.2 may be invoked in order to obtain an iterative scheme that approaches v^* ultimately from above.

THEOREM 4.4.5.

(a) $[v^*,0]$ *is the smallest solution of* (4.4.17) *and* (4.4.18)

(b) $[v^*,0]$ *is the optimal bias vector in the* MDP, *specified by* (4.4.16)

(c) *Consider the scheme*

(4.4.19) $z(n+1)_i = \max_{k \in \bar{K}(i)} [\bar{r}_i^k(n) + \beta_n \sum_{j \in \bar{\Omega}} \bar{p}_{ij}^k z(n)_j]$

where $\beta_n = 1 - n^{-b}$, *for some* $0 < b \leq 1$, *and*

(4.4.20) $\bar{r}_i^k(n) = \begin{cases} b_i^k + \sum_j p_{ij}^k (\xi(n)_j + \bar{g}(n)_j) - \xi(n)_i - \bar{g}(n)_i & ; i \in \Omega; k \in K(i) \\ 0 & ; \text{otherwise} \end{cases}$

and where $z(0) \in E^{N+1}$ *has* $z(0)_0 = 0$.

Let $\{\eta_n\}_{n=1}^{\infty}$ *be a sequence of non-increasing positive numbers, such that* $\eta_n = o(n^{-b}\ln n)$; *e.g. take* $\eta_n = n^{-b/2}$; $n \geq 1$. *Then*

(4.4.21) $\{z(n)\}_{n=1}^{\infty} \to v^*$ *and* $\{z(n)+\eta_{n-1}\}_{n=1}^{\infty}$ *approaches* v^* *from above.*

PROOF.

(a) Note that for any solution v of (4.4.13) the vector [v,0] satisfies (4.4.17) and (4.4.18) and vice versa. Next invoke theorem 4.4.4.

(b) Let $\bar{\bar{S}}_{MG}$ be the set of maximal gain policies in the above considered MDP. For each policy f, let $\bar{r}(f)$, $\bar{P}(f)$, $\bar{Z}(f)$ and $\bar{\Pi}(f)$ denote the associated reward vector, transition probability, fundamental and invariant probability matrix. Let d be the least common multiple of the periods of the policies in $\bar{\bar{S}}_{MG}$, and let $\bar{z}^*$ represent the optimal bias vector. Then, using (1.4.7):

$$\bar{z}^*_i = \max_{f\in\bar{\bar{S}}_{MG}} \bar{Z}(f)\bar{r}(f)_i =$$

$$\max_{f\in\bar{\bar{S}}_{MG}} \lim_{a\uparrow 1} \lim_{n\to\infty} \sum_{\ell=0}^{nd} a^{\ell}\bar{P}^{\ell}(f)\bar{r}(f)_i$$

$$= \max_{f\in\bar{\bar{S}}_{MG}} \lim_{n\to\infty} \sum_{\ell=0}^{nd} P^{-\ell}(f)\bar{r}(f)_i; \quad i \in \Omega$$

where the interchange of the limit and summation operator is justified by

$$\Big\|\sum_{\ell=0}^{nd} a^{\ell}\bar{P}^{\ell}(f)\bar{r}(f)\Big\| \leq K(1-\lambda^n), \quad \text{for some } K > 0,\ 0 \leq \lambda < 1, \text{ and } \textit{all } 0 \leq a \leq 1 \text{ and } n \geq 1.$$

and the well-known fact that the limit function of a uniformly convergent sequence of continuous functions on a closed interval, is a continuous function itself. Next, take a solution v of (4.4.17) and (4.4.18) and fix $f \in \bar{\bar{S}}_{MG}$. Then,

$$v \geq \bar{r}(f)+\bar{P}(f)v \geq \sum_{n=1}^{M} \sum_{\ell=(n-1)d}^{nd-1} \bar{P}^{\ell}(f)\bar{r}(f) + \bar{P}^{Md+1}(f)v$$

$$\geq \sum_{n=1}^{M} \sum_{\ell=(n-1)d}^{nd-1} \bar{P}^{\ell}(f)\bar{r}(f), \quad \text{in view of } v \geq 0 \text{ (cf. (4.4.13))}$$

Hence, $v \geq \sum_{n=1}^{\infty}\sum_{\ell=(n-1)d}^{nd-1} \bar{P}^{\ell}(f)\bar{r}(f)$, or $v \geq \max_{f\in\bar{\bar{S}}_{MG}} \sum_{n=1}^{\infty}\sum_{\ell=(n-1)d}^{nd-1} \bar{P}^{\ell}(f)\bar{r}(f) = \bar{z}^*$. Observe on the other hand, that $\bar{z}^*$, as the optimal bias vector, is itself a solution to (4.4.17), with $\bar{z}^*_0 = 0$, hence satisfying (4.4.18) as well, and conclude that $\bar{z}^*$ is the smallest solution of (4.4.17) and (4.4.18) i.e. $\bar{z}^* = v^*$.

(c) Part (c) follows from th.4.3.2, in combination with lemma 4.4.1 and th.4.4.3. □

Although (4.4.20) provides an approximation method for the third and last component in the decomposition $x^* = \xi^* + \bar{g}^* + v^*$, one should observe that its rate of convergence is far from geometric. A geometric type of convergence will, however, be needed in any attempt to solve the second equation in the pair (4.1.5) and (4.1.6). Fortunately, the latter may be achieved by observing that v^* is the smallest vector v, with $v_0 = 0$, that satisfies the average return optimality equation (4.4.17). The method, presented in th.4.4.6 below exploits both this property, and the availability of an asymptotically converging upper bound for v^* (cf. part (c) of the previous theorem).

THEOREM 4.4.6. *Consider the scheme:*

step 0: *Pick a sequence of positive numbers* $\{\gamma_m\}_{m=1}^{\infty}$, *with* $\gamma_m \uparrow 1$. *Initialize* $m := 0$; $v := 0$; $n := 0$. Fix $0 < \tau < 1$.

step 1: $v_i := \max_{k \in \bar{K}(i)} \{\tau \hat{r}_i^k(n) + \sum_j [\delta_{ij} + \tau(\bar{P}_{ij}^k - \delta_{ij})] v_j\}$, $i \in \bar{\Omega}$ *where* $\hat{r}_i^k(n) = \bar{r}_i^k(n) - (\gamma_m)^n$; $i \in \bar{\Omega}$; $k \in \bar{K}(i)$ *and with* $\bar{r}_i^k(n)$ *defined by* (4.4.20).

step 2: "if" $v_i > z(n+1)_i + \eta_{n+1}$, *for some* $i \in \bar{\Omega}$ "then"
(a) $v := 0$; (b) $m := m+1$

step 3: $n := n+1$; *go back to step* 1.

Let $\{v(n)\}_{n=1}^{\infty}$ denote the sequence of values adopted by the vector variable v as $n = 1,2,\ldots$. Then,

(a) the test in step 2, can only be met for a finite number of times
(b) $\{v(n)\} \to v^*$, geometrically.

PROOF.

(a) Note that $|\hat{r}_i^k(n) - \bar{r}_i^k| \leq C\lambda^n$ for some $C > 0$, and $0 \leq \lambda < 1$. Since $\{\gamma_m\}_{m=1}^{\infty} \uparrow 1$, and in view of part (c) of the previous theorem there is an integer $m_0 \geq 1$, such that

(4.4.22)(a) $C\lambda^n < (\gamma_{m_0})^n$ i.e. $\hat{r}_i^k(n) < r_i^k$ for all $n \geq m_0$; $i \in \bar{\Omega}$; $k \in \bar{K}(i)$

(b) $z(n)_i + \eta_n > v_i^*$ for all $n \geq m_0$; $i \in \bar{\Omega}$

Next, we show that m cannot be incremented above m_0+1; i.e. the test in step 2 can be met for at most m_0+1 times. For assume that m adapts

the value m_0+1. Note that the current value n^* of $n \geq m_0+1$, since m can only be incremented after incrementing n by at least one. Observe furthermore that $v(n^*+1) = 0$. Then,

$$v(n^*+2)_i = \max_{k\in\bar{K}(i)}[\hat{r}_i^k(n^*+1) + \sum_j \bar{P}_{ij}^k\, v(n^*+1)]$$

$$\leq \max_{k\in\bar{K}(i)}[\bar{r}_i^k + \sum_j \bar{P}_{ij}^k\, v_j^*] = v_i^*, \quad i \in \Omega$$

where the inequality part follows from (4.4.22) and $v(n^*+1) = 0 \leq v^*$ (cf. (4.4.13)). Proceed by induction to verify that for all $n \geq n^*+1$ $v(n)_i \leq v_i^* \leq z(n)_i + \eta_n$, $i \in \bar{\Omega}$, such that the test in step 2 is never met again.

(b) We conclude from part (a), that after the test in step 2 has been met for a finite number of times, the vector v is reinitialized at 0, and from thereon, the algorithm behaves exactly like the scheme (4.3.7) with $\{\hat{r}_i^k(n)\}_{n=1}^{\infty} \to \bar{r}_i^k$, *geometrically*. Note in addition that $v(n)_0 = 0$ for all $n \geq 1$ and apply th.4.3.1 to conclude that $\{v(n)\}_{n=1}^{\infty} \to v$, geometrically, where v satisfies (4.4.17) and (4.4.18). Moreover, since $v(n)_i \leq z(n)_i + \eta_n$, $i \in \bar{\Omega}$, it follows that $v \leq v^*$ by letting n tend to infinity, and invoking part (c) of the previous theorem. Conclude that, $v = v^*$, in view of v^* being the smallest solution to (4.4.17) and (4.4.18) (cf. part (a) of th.4.4.5). Finally, the *geometric* rate of convergence follows from th.4.3.1. □

We conclude this section by putting all the pieces together. We present our algorithm for solving the pair of coupled equations (4.1.5) and (4.1.6), for the more general case, where instead of knowing the parameters in advance, only *geometric* approximations are available, i.e. we assume to have:

(4.4.23) $\{K(i,n)\}_{n=1}^{\infty} \to K(i), \quad i \in \Omega$

$\{b_i^k(n)\}^{\infty} \to b_i^k$, geometrically, $i \in \Omega$, $k \in K(i)$

$\{c_i^k(n)\}^{\infty} \to c_i^k$, geometrically, $i \in \Omega$, $k \in K(i)$.

It is easily verified that all of the results in theorem 4.4.2 to 4.4.6 go through when replacing the parameters in the pair of equations (4.1.5) and (4.1.6) by their approximations.

THEOREM 4.4.7. *(Main Result).*

(a) *To find the unique solution* x^* *to* (4.1.5) *and* (4.1.6), *given approximations for its parameters, as specified by* (4.4.23) *construct the sequences* $\{\xi(n)\}_{n=1}^{\infty} \to \xi^*$, $\{\bar{g}(n)\}_{n=1}^{\infty} \to \bar{g}^*$, $\{v(n)\}_{n=1}^{\infty} \to v^*$, *by generating the schemes* (4.4.2), (4.4.11) *as well as the scheme in th.4.4.6, in which* $\{K(i) \mid i \in \Omega\}$, $\{b_i^k \mid i \in \Omega,\ k \in K(i)\}$ *and* $\{c_i^k \mid i \in \Omega,\ k \in K(i)\}$ *have been replaced by their approximations. Lemma* 4.4.1, *and theorem* 4.4.3 *and* 4.4.6 *show that the three sequences exhibit a geometric rate of convergence. Note that the construction of* $\{v(n)\}_{n=1}^{\infty}$ *requires the generation of a fourth sequence* $\{z(n)\}_{n=1}^{\infty}$, *via the scheme* (4.4.19) *in which the parameters of the problem are again replaced by their approximations. Then*

(4.4.24) $\{x(n)\}_{n=1}^{\infty} = \{\xi(n)+\bar{g}(n)+v(n)\}_{n=1}^{\infty} \to x^*$, *geometrically*

(b) *Let* $F(i,n,\varepsilon) = \{k \in K(i,n) \mid x(n+1)_i \leq b_i^k(n) + \sum_j P_{ij}^k x(n)_j - \varepsilon\}$; $i \in \Omega; n \geq 1; \varepsilon > 0$. *Then,* $\lim_{n\to\infty} F(i,n,\varepsilon_n) = M(i,x^*)$; $i \in \Omega$.

(c) *To find a (non-unique)* y^* *satisfying* (4.1.6) *generate the sequence*

(4.4.25) $y(n+1)_i = \max_{k \in F(i,n,\varepsilon_n)} \{\tau(c_i^k(n) - \sum_j H_{ij}^k x(n)_j) + \sum_j [\delta_{ij} + \tau(P_{ij}^k - \delta_{ij})] y(n)_j\}$

$i \in \Omega$; $n = 1,2,\ldots$

with $0 < \tau < 1$ *and* $y(0) \in E^N$ *arbitrarily chosen. Then* $y(n) \to y^*$, *geometrically.* □

PROOF. Only parts (b) and (c) need to be proved. Part (b) follows from the proof of th.1.7.1. part (a). Part (c) follows by invoking th.4.3.1 and by interpreting (4.1.6) as the average return optimality equation of a MDP, with $M(i,x^*)$ as the action set in state i and $\{c_i^k - \sum_j H_{ij}^k x_j^* \mid i \in \Omega; k \in M(i,x^*)\}$ as the set of one-step expected rewards. □

REMARK 1. All of the sequences employed in the above algorithm are bounded, but for $\{h(n)\}_{n=1}^{\infty}$ which is generated in (4.4.11) and which diverges linearly with n. We refer to section 1.8 ((1.8.20)-(1.8.24)) for a discussion of possible methods to eliminate this numerical difficulty.

4.5. THE n+1 NESTED FUNCTIONAL EQUATIONS

We showed in theorem 4.4.7 how to solve *two* coupled equations by successive approximations. For two equations, we saw that 4 successive

approximation schemes (cf. th.4.4.7 part (a)) suffice to find the first unknown and one additional scheme (cf. th.4.4.7 part (c)) will find a solution to the second equation. Since (4.1.1) has a similar structure, similar methods will work for n+1 sets of equations with n arbitrarily large. The n unique vectors $\{x^{*(0)},\ldots,x^{*(n-1)}\}$ (cf. th.4.2.3) are found by 4 n successive approximation schemes and the (non-unique) vector $x^{*(n)}$ is found by one additional scheme.

To avoid excessive notation, the procedure will only be described verbally. First construct 4 schemes (given by applying th.4.4.7 part (a) to the first 2 equations) to find $x^{*(0)}$. Here the action sets $K^{o}(i)$, $i \in \Omega$, and the rewards $a_i^k(0)$ and $a_i^k(1)$, $i \in \Omega$ and $k \in K^{o}(i)$, are known exactly. The result of these 4 schemes is a sequence $\{x^{(0)}(n)\}_{n=1}^{\infty}$ which converges geometrically to $x^{*(0)}$ and sequences of sets $\{K^1(i,n)\}_{n=1}^{\infty}$, $i \in \Omega$, which converge to $K^1(i)$. Next construct 4 more schemes (given by applying th.4.4.7 to the second and third equation in (4.1.1)) to find $x^{*(1)}$. Here the action sets $K^1(i)$, $i \in \Omega$, and the rewards $a_i^k(1) - \sum_j H_{ij}^k(1)\, x_j^{*(0)}$ and $a_i^k(2)$, $i \in \Omega$; $k \in K^{o}(i)$, are not known in advance; instead we employ the approximations available from the earlier schemes:

$$K^1(i,n) \quad \text{for } K^1(i); \quad i \in \Omega$$

$$a_i^k(1) - \sum_j H_{ij}^k(1)\, x^{(0)}(n)_j \quad \text{for} \quad a_i^k(1) - \sum_j H_{ij}^k(1)\, x_j^{*(0)}; \; i \in \Omega; \; k \in K^{o}(i).$$

$$\beta_i^k(2) + \langle b_i^{o;k}(2), x^{(0)}(n) \rangle \quad \text{for} \quad a_i^k(2); \; i \in \Omega; \; k \in K^{o}(i).$$

The result is a scheme giving a sequence $\{x^{(1)}(n)\}_{n=1}^{\infty}$ which converges geometrically to $x^{*(1)}$ and sets $K^2(i;n)$ which converge to $K^2(i)$, $i \in \Omega$. Continuing this way, we get 4n+1 *simultaneous* successive approximation schemes, each dependent upon its predecessors and each converging geometrically.

Note the feature that the geometric rate of convergence propagates from one pair of functional equations to the next; this is crucial for guaranteeing convergence of the whole set of 4n+1 *successive approximation schemes.*

We conclude this chapter by considering the following special cases:

A) $a_i^k(0) = 0$, $i \in \Omega$, $k \in K^{o}(i)$. This case occurs in all MRPs in which one wants to find policies that are maximal gain or optimal under more selective discounted or average overtaking optimality criteria (cf. [22], [127] and section 4.1). For notational simplicity assume, as is actually the case in the MRP model, that $H_{ij}^k(m) = H_{ij}^k$ for all $m = 1,\ldots,n$ and $i,j \in \Omega$ and $k \in K^{o}(i)$. A solution to the first *two* equations in (4.4.1) may be obtained by generating a system of only *two* (rather than four, as in the general case of th.

4.4.7) schemes. To verify the latter, note from th.4.2.1 that in this particular case, $S_{MG} = X_j K^o(j)$ such that

$$x_i^{*(0)} = \max_{f \in X_j K(j)} \sum_{m=1}^{n(f)} \Phi_i^m(f) \frac{<\pi^m(f), a(f;1)>}{<\pi^m(f), T(f)>},$$

i.e. we rederive that $x^{*(0)}$ itself represents a "maximal gain rate" vector and may be approximated with a single equation scheme (cf. section 1.9 and section 4.3).

We recall having noticed in section 1.9, that whereas the maximal gain rate and maximal gain policies in an undiscounted MRP, can be found via a single iteration scheme, the computation of a solution to the optimality equation 1.9.5 requires *two* (simultaneously generated) schemes. The above described procedure clarifies how this objective may be met.

In the special case where either (1) $x^{*(0)}$ has identical components, or (2) $H_{ij}^k = T_i^k \delta_{ij}$ or (3) $H_{ij}^k = T_i^k P_{ij}^k$, the second equation in (4.1.1) may be rewritten as

$$(4.5.1) \quad x_i^{*(1)} = \max_{k \in K^1(i)} [a_i^k(1) - x_i^{*(0)} T_i^k + \sum_j P_{ij}^k x_j^{*(1)}], \quad i \in \Omega$$

which coincides with the simplified optimality equation, (1.9.8) associated with the maximal gain rate vector $x^{*(0)}$. We conclude that in this case, a sequence approaching some solution y^0 to (4.5.1) may be derived from the same scheme that is needed to obtain $x^{*(0)}$ (cf. th.4.3.1 part (b)), i.e. a *single* equation scheme suffices to solve the first two functional equations.

B) *Finding bias-optimal policies in* MRPs. This problem reduces to solving the following triple of nested equations.

$$(4.5.2) \quad g_i^* = \max_{k \in K^o(i)} [\sum_j P_{ij}^k g_j^*], \quad i \in \Omega$$

$$z_i^* = \max_{k \in K^1(i)} [q_i^k - \sum_j H_{ij}^k g_j^* + \sum_j P_{ij}^k z_j^*], \quad i \in \Omega$$

$$u_i^* = \max_{k \in K^2(i)} [a_i^k(2) - \sum_j H_{ij}^k z_j^* + \sum_j P_{ij}^k u_j^*], \quad i \in \Omega.$$

Since (4.5.2) satisfies A) only two schemes are needed to solve the first two equations, and an additional quadruple of schemes has to be added to solve the entire problem.

The algorithm simplifies, however, under condition (UNI) (cf. (1.4.20)).

LEMMA 4.5.1. *Assume condition (UNI) to hold. Fix* $v^o \in V$. *Let the vector* u *and the scalar* d *satisfy*

$$(4.5.3) \quad u_i = \max_{k \in S(i,v^o)} \{[a_i^k(2) - \sum_j H_{ij}^k v_j^o] - d\,T_i^k + \sum_j p_{ij}^k u_j\}, \; i \in \Omega.$$

Then, $(v^o + d\underline{1}, u)$ *satisfy the last two equations in* (4.5.2), *i.e.* $(\langle g^* \rangle \underline{1}, v^o + d\underline{1}, u)$ *solves the entire system* (4.5.2).

PROOF. The fact that $g^* = \langle g^* \rangle \underline{1}$ follows immediately from (1.4.20). Note that (UNI) implies that the sets S(i,v) are independent of $v \in V$ for all $i \in \Omega$. Hence we write S(i) for S(i,v), $i \in \Omega$. Let $z = v^o + d\underline{1} \in V$. Regroup the terms to the right of (4.5.3) and use $\sum_j H_{ij}^k = T_i^k$ ($i \in \Omega$, $k \in K^o(i)$) to conclude that (z,u) satisfy the last two equations in (4.5.2). Hence z represents the optimal bias vector z^* and $(\langle g^* \rangle \underline{1}, z, u)$ solves the entire system (4.5.2). □

Under (UNI) we have in view of $g^* = \langle g^* \rangle \underline{1}$, that the optimality equation (1.9.5) reduces to (1.9.8) i.e. $V = \hat{V}$. Hence a vector $v^o \in V$, the maximal gain rate $\langle g^* \rangle$ and the sets $S(i,v^o) = S(i)$, $i \in \Omega$ may be obtained from a single equation scheme, like (4.3.3). Next, it follows from lemma 4.5.1 that the scalar $d[v^o]$ and hence the vector z^*, as well as a solution u to the third equation in (4.5.2) and a bias-optimal policy may be obtained from a simultaneously computed second scheme of the type (4.3.3).

We conclude that whereas in general a system of 6 schemes is required to obtain bias-optimal policies, the number may be reduced to *two* under assumption (UNI). We finally recall that Hordijk and Tijms' method (cf. section 4.3 and 1.8) may be used in case $g^* = \langle g^* \rangle \underline{1}$ to find the optimal bias-vector, though not necessarily a bias-optimal policy.

CHAPTER 5

The optimality equation in average cost denumerable state semi-Markov decision problems, recurrence conditions and algorithms

5.1. INTRODUCTION

In this chapter we consider an undiscounted semi-Markov decision model specified by five objects $(I, A(i), p_{ij}(a), c(i,a), \tau(i,a))$. We are concerned with a dynamic system which at decision epochs beginning with epoch 0 is observed to be in one of the states of the *denumerable* state space I. After observing the state of the system, an action must be chosen. For any state $i \in I$, the set $A(i)$ denotes the set of possible actions for state i. If the system is in state i at any decision epoch and action $a \in A(i)$ is chosen, then regardless of the history of the system, the following happens:

(i) an immediate cost $c(i,a)$ is incurred;

(ii) the time until the next decision epoch is random with mean $\tau(i,a)$;

(iii) at the next decision epoch the system will be in state j with probability $p_{ij}(a)$ where $\sum_{j \in I} p_{ij}(a) = 1$ for all $i \in I$ and $a \in A(i)$.

Unless stated otherwise, we make throughout this chapter the following assumptions.

A1. *For any* $i \in I$, *the set* $A(i)$ *is a compact metric space on which both* $c(i,a)$, $\tau(i,a)$ *and* $p_{ij}(a)$ *for any* $j \in I$ *are continuous.*

A2. *There is a finite number* M *such that* $|c(i,a)| \leq M$ *and* $\tau(i,a) \leq M$ *for all* $i \in I$ *and* $a \in A(i)$.

A3. *There is a positive number* δ *such that* $\tau(i,a) \geq \delta$ *for all* $i \in I$ *and* $a \in A(i)$.

We note that assumption A1 is satisfied when $A(i)$ is finite for all $i \in I$.

A policy π for controlling the system is any (possibly randomized) rule for choosing actions. For any initial state i and policy π, denote by X_n and a_n the state and the action chosen at the nth decision epoch for $n = 0,1,\ldots$. (the 0^{th} decision epoch is at epoch 0). Denote by E_π the

expectation when policy π is used. Let $F = X_{i\in I} A(i)$, i.e. F is the class of all functions f which add to each state $i \in I$ a single action $f(i) \in A(i)$. For any $f \in F$, denote by $f^{(\infty)}$ the stationary policy which prescribes action f(i) whenever the system is in state i. Under each stationary policy $f^{(\infty)}$ the process $\{X_n, n \geq 0\}$ is a Markov-chain with one-step transition probability matrix $P(f) = (p_{ij}(f(i)))$, $i,j \in I$. For $n = 1,2,\ldots$, denote the n-step transition probability matrix of this Markov chain by $P^n(f) = (p_{ij}^n(f))$, $i,j \in I$. For $n = 1$ we write $P(f) = (p_{ij}(f))$.

Finally, the following assumption is made throughout this chapter:

A4. *For any* $f \in F$, *the stochastic matrix* P(f) *has no two disjoint closed sets.*

In this chapter, we are concerned with the optimality equation for the long run average costs. We give a large number of recurrence conditions with respect to the stochastic matrices P(f), $f \in F$, under which the existence of a bounded solution to this optimality equation will be proven. This will be done in section 3. First, however, these recurrence conditions are presented in section 2, and we exhibit several relations between them, thereby mapping out some of the existence conditions that have appeared in the literature, so far.

It is important to note that the existence of a bounded solution to the optimality equation, implies the existence of an optimal stationary policy among the class of all policies and with respect to a strong version of the average cost optimality criterion, which implies essentially weaker versions usually considered in the literature (cf.[46] and th.5.3.2). Further we note that after having established the optimality equation for the average costs, a repeated application of this result yields a sequence of optimality equations that are involved when considering the more sensitive and selective n-discounted optimality criteria, thus showing the existence of stationary n-discounted optimal policies (cf. HORDIJK and SLADKÝ [59] and section 4.1).

Besides the existence of a bounded solution to the optimality equation for the average costs, we will consider the problem of determining such a solution, which in turn yields an optimal stationary policy. In section 4, we shall show that under each of our conditions the value-iteration method can be used to determine a bounded solution to the optimality equation. The policy-iteration method will be considered in section 5. Under condition

C6', to be stated below, we shall prove that the average costs and the relative cost functions of the policies generated by this method, converge to a solution of the optimality equation. This result considerably generalizes a related result in DERMAN [24].

The results in this chapter are based upon FEDERGRUEN and TIJMS [40] and FEDERGRUEN, HORDIJK and TIJMS [42].

5.2. RECURRENCE CONDITIONS AND EQUIVALENCES

In this section we shall formulate a number of recurrence conditions, on the set $P = (P(f), f \in F)$ and prove several relations between these conditions. Before doing this, we note that by $F = X_{i\in I} A(i)$ and assumption A1, the set F is a compact metric space in the product topology where for any $i,j \in I$, the function $p_{ij}(f)$ is continuous on F. We note that in the remainder of this section, we merely use this fact, rather than the product property of F. Also, using the relation

$$(5.2.1) \quad p_{ij}^{m+1}(f) = \sum_{k\in I} p_{ik}(f) p_{kj}^{m}(f) \quad \text{for all } i,j \in I,\ m \geq 1 \text{ and } f \in F$$

and proposition 18 on p.232 in [100], it follows by complete induction that for any $n \geq 1$ and $i,j \in I$ the function $p_{ij}^{n}(f)$ is continuous on F. Further we introduce the following notation. For any $i_0 \in I$, $A \subseteq I$ and $f \in F$, define the taboo probability

$$(5.2.2) \quad t_{i_0 A}^{n}(f) = \sum_{i_1,\ldots,i_n \in I\setminus A} p_{i_0 i_1}(f) \ldots p_{i_{n-1} i_n}(f), \quad n = 1,2,\ldots$$

i.e. $t_{iA}^{n}(f)$ in the probability that the first return to set A, takes more than n transitions, when starting in state i and using policy $f^{(\infty)}$. For any $i \in I$, $A \subseteq I$ and $f \in F$, define the (possibly infinite) mean recurrence time

$$(5.2.3) \quad \mu_{iA}(f) = 1 + \sum_{n=1}^{\infty} t_{iA}^{n}(f)$$

i.e. $\mu_{iA}(f)$ is the excepted number of transitions until the first return to the set A, when starting in state i and using policy $f^{(\infty)}$. Finally, we write $t_{iA}^{n}(f) = t_{ij}^{n}(f)$ and $\mu_{iA}(f) = \mu_{ij}(f)$ for $A = \{j\}$.

Consider now the following simultaneous recurrence conditions on the set $P = (P(f),\ f \in F)$.

C1. *There is a finite set* K *and a finite number* B *such that*

$$\mu_{iK}(f) \leq B \text{ for all } i \in I \text{ and } f \in F.$$

C2. *There is a finite set* K, *an integer* $\nu \geq 1$ *and a number* $\rho > 0$ *such that*

$$\sum_{j\in K} p^{\nu}_{ij}(f) \geq \rho \text{ for all } i \in I \text{ and } f \in F.$$

C3. *There is an integer* $\nu \geq 1$ *and a number* $\rho > 0$ *such that*

$$\inf_{i_1,i_2\in I} \{ \sum_{j\in I} \min[p^{\nu}_{i_1 j}(f), p^{\nu}_{i_2 j}(f)] \} \geq \rho \text{ for all } f \in F \tag{5.2.4}$$

C4. *There is an integer* $\nu \geq 1$ *and a number* $\rho > 0$ *such that for any* $f \in F$ *a probability distribution* $\{\pi_j(f), j\in I\}$ *(say) exists for which*

$$\Big| \sum_{j\in A} p^{n}_{ij}(f) - \sum_{j\in A} \pi_j(f) \Big| \leq (1-\rho)^{\lfloor n/\nu \rfloor} \text{ for all } i \in I,\ A \subset I \text{ and } n \geq 1. \tag{5.2.5}$$

where $\lfloor x \rfloor$ *denotes the largest integer less than or equal to* x.

C5. *For any* $f \in F$ *there is a probability distribution* $\{\pi_j(f), j\in I\}$ *such that*

$$p^{n}_{ij}(f) \to \pi_j(f) \text{ uniformly in } (i,f) \in I \times F \text{ as } n \to \infty \text{ for any } j \in I. \tag{5.2.6}$$

C6. *There is a finite number* B *such that for any* $f \in F$ *a state* s_f *exists for which*

$$\mu_{is_f}(f) \leq B \text{ for all } i \in I.$$

C7. *There is a finite set* K *and a finite number* B *such that for any* $f \in F$ *a state* $s_f \in K$ *exists for which*

$$\mu_{is_f}(f) \leq B \text{ for all } i \in I.$$

C8. *There is an integer* $\nu \geq 1$ *and a number* $\rho > 0$ *such that for any* $f \in F$ *a state* s_f *exists for which*

$$p^{\nu}_{is_f}(f) \geq \rho \text{ for all } i \in I.$$

C9. *There is a finite set* K, *an integer* $\nu \geq 1$ *and a number* $\rho > 0$ *such that for any* $f \in F$ *a state* $s_f \in K$ *exists for which*

$$p^{\nu}_{is_f}(f) \geq \rho \text{ for all } i \in I.$$

We note that in C4 the condition $\sum_{j\in I} |p_{ij}^n(f) - \pi_j(f)| \leq 2(1-\rho)^{\lfloor n/\nu \rfloor}$ for all $i \in I$, $f \in F$ and $n \geq 1$ may be equivalently stated instead of (5.2.5). The condition C1 was considered in [58], cf. also [98]. The condition C2 was introduced in [58] and called the simultaneous Doeblin condition since for each $f \in F$ the stochastic matrix $P(f)$ satisfies the so-called Doeblin condition from Markov chain theory. The conditions C3 and C4 were introduced in [40]. Following Markov chain terminology, the conditions C3 and C4 could be called a simultaneous scrambling condition (cf. [51]) and a simultaneous quasi-compactness (or strong ergodicity) condition (cf. [88]) respectively. Observe that except for C1-C2 each of the above conditions implies assumption A4 in itself.

Further, any $P(f)$ is aperiodic under C3, C4, C5, and C8. Finally, we note that the left side of (5.2.4) denotes the ergodic coefficient of the stochastic matrix $P^{\nu}(f)$ (cf. also (3.3.8)) and that $\{\pi_j(f), j \in I\}$ in C4-C5 denotes the unique stationary probability distribution of $P(f)$.

The following theorem was obtained in a more general setting in HORDIJK [58].

THEOREM 5.2.1.

(i) *The conditions* C1 *and* C2 *are equivalent.*

(ii) *Under condition* C2, *every stochastic matrix* $P(f)$ *has a unique stationary probability distribution* $\{\pi_j(f), j \in I\}$ *(say) such that for any* $j \in I$, *the function* $\pi_j(f)$ *is continuous on* F.

We shall now give the following equivalences:

THEOREM 5.2.2. *If the stochastic matrix* $P(f)$ *is aperiodic for each* $f \in F$, *then condition* C2 *implies condition* C3.

THEOREM 5.2.3. *Condition* C3 *implies condition* C4.

THEOREM 5.2.4.

(i) *The condition* C5 *implies both condition* C2 *and* C9.

(ii) *The conditions* C3, C4, C5, C8 *and* C9 *are equivalent.*

THEOREM 5.2.5.

(i) *The condition* C2 *implies the condition* C7.

(ii) *The condition* C6 *implies the condition* C7.

(iii) *The conditions* C1, C2, C6 *and* C7 *are equivalent.*

(iv) *If the stochastic matrix* $P(f)$ *is aperiodic for each* $f \in F$, *then the conditions* C1-C9 *are equivalent.*

In case the set P consists of a single stochastic matrix, the above equivalences may be found, albeit in a scattered way, in the literature, cf. p. 197 in DOOB [28], p.142 in HUANG, et al. [65], p.226 in ISAACSON and MADSEN [67], and p.185 in NEVEU [88].

We shall now give the proof of the above theorems 5.2.2 - 5.2.5.

Proof of Theorem 5.2.2. Suppose first that C2 with triple (K,ν,ρ) holds and that every $P(f)$ is aperiodic. We shall then verify condition C3. Since for any $f \in F$ the stochastic matrix $P(f)$ satisfies the Doeblin condition, has no two disjoint closed sets and is aperiodic, we have from Markov chain theory (e.g. [28]) that every $P(f)$ has a unique stationary probability distribution $\{\pi_j(f), j \in I\}$ (say) such that

$$(5.2.7) \qquad \lim_{n\to\infty} p_{ij}^n(f) = \pi_j(f) \quad \text{for all } i,j \in I.$$

Since C2 implies $\sum_{j\in K} p_{ij}^n(f) \geq \rho$ for all $i \in I$, $f \in F$ and $n \geq \nu$, we have

$$(5.2.8) \qquad \sum_{j\in K} \pi_j(f) \geq \rho \quad \text{for all } f \in F.$$

Define now

$$(5.2.9) \qquad F_k = \{f \in F \mid \pi_k(f) \geq \frac{\rho}{|K|}\} \quad \text{for } k \in K.$$

where $|K|$ denotes the number of states in K. Then, by (5.2.8),

$$F = \bigcup_{k\in K} F_k.$$

Using part (ii) of th.5.2.1 and the fact that F is a compact metric space, it follows that for any $k \in K$ the set F_k is closed and hence compact. For any $i \in I$ and $k \in K$, define

$$(5.2.10) \qquad n(i,k,f) = \min\{n \geq 1 \mid p_{ik}^n(f) > \frac{\rho}{2|K|}\} \quad \text{for } f \in F_k.$$

By (5.2.7), $n(i,k,f)$ exists and is finite. Using the fact that $P^n(f)$ is continuous on F for each $n \geq 1$, it is immediately verified that for each $i \in I$ and $k \in K$ the set $\{f \in F_k \mid n(i,k,f) \geq \alpha\}$ is closed for any real α, i.e. for each $i \in I$ and $k \in K$ the function $n(i,k,f)$ is upper semi-continuous on the compact set F_k.

Now, by Proposition 10 on p.161 in ROYDEN [100], we have that for each $i \in I$ and $k \in K$ the function $n(i,k,f)$ assumes a finite maximum on F_k. Hence, using the finiteness of K, we can find an integer $\mu \geq 1$ such that

(5.2.11) $n(i,k,f) \leq \mu$ for all $i \in K$, $k \in K$ and $f \in F_k$.

Next define for any $k \in K$

(5.2.12) $m(k,f) = \min\{n \geq 1 \mid p_{kk}^{m}(f) > \frac{\rho}{2|K|} \text{ for all } n \leq m \leq n+\mu\}$ for $f \in F_k$.

We now verify that for each $k \in K$ the set $S_\alpha = \{f \in F_k \mid m(k,f) \geq \alpha\}$ is closed for any real α. Fix $k \in K$ and an integer $\alpha > 1$. Suppose that $f_n \in S_\alpha$ for $n \geq 1$ and that $f_n \to f^*$ as $n \to \infty$. Then we can find a subsequence $\{n_h, h \geq 1\}$ of integers, and integers r and t with $1 \leq r \leq \alpha - 1$ and $r \leq t \leq r + \mu$ such that $p_{kk}^{t}(f_{n_h}) \leq \rho/(2|K|)$ for all $h \geq 1$. Hence, by the fact that $p_{kk}^{t}(f)$ is continuous on F, we find $p_{kk}^{t}(f^*) \leq \rho/(2|K|)$ and so $f^* \in S_\alpha$. We have now proved that for any $k \in K$, the function $m(k,f)$ is upper semi-continuous on the compact set F_k. Hence there exists an integer $N \geq 1$ such that

$$m(k,f) < N \quad \text{for all } k \in K \text{ and } f \in F_k.$$

For any $k \in K$ and $f \in F_k$, we have by (5.2.10)-(5.2.12)

$$p_{ik}^{\mu+m(k,f)}(f) \geq p_{ik}^{n(i,k,f)}(f)\, p_{kk}^{m(k,f)+\mu-n(i,k,f)}(f) > \frac{\rho^2}{4|K|^2} \text{ for all } i \in K.$$

Hence for any $k \in K$ and $f \in F_k$,

$$p_{ik}^{\nu+\mu+m(k,f)}(f) \geq \sum_{j \in K} p_{ij}^{\nu}(f)\, p_{jk}^{\mu+m(k,f)}(f) > \frac{\rho^3}{4|K|^2} \text{ for all } i \in I.$$

Using this result, we now find for any $k \in K$ and $f \in F_k$,

$$\begin{aligned}
&\sum_{j \in I} \min\left[p_{i_1 j}^{\nu+\mu+N}(f), p_{i_2 j}^{\nu+\mu+N}(f)\right] \geq \\
&\geq \sum_{j \in I} \min\left[p_{i_1 k}^{\nu+\mu+m(k,f)}(f)\, p_{kj}^{N-m(k,f)}(f), p_{i_2 k}^{\nu+\mu+m(k,f)}(f)\, p_{kj}^{N-m(k,f)}(f)\right] \geq \\
&\geq \frac{\rho^3}{4|K|^2} \sum_{j \in I} p_{kj}^{N-m(k,f)}(f) = \frac{\rho^3}{4|K|^2} \quad \text{for all } i_1, i_2 \in I,
\end{aligned}$$

which verifies C3.

Proof of Theorem 5.2.3. The proof of this theorem proceeds along the same lines as that of theorem 1 in ANTHONISSE & TIJMS [1], or that of th.2.2.5.

Assume C3 holds with the pair (ν,ρ). Fix $f \in F$ and $A \subseteq I$. For $n = 1,2,\ldots$, define

$$M_n = \sup_{i \in I} \sum_{j \in A} p_{ij}^{n}(f) \quad \text{and} \quad m_n = \inf_{i \in I} \sum_{j \in A} p_{ij}^{n}(f).$$

Using (5.2.1), it follows that

(5.2.13) $M_{n+1} \le M_n$ and $m_{n+1} \ge m_n$ for all $n \ge 1$.

For any number a, let a^+ and a^- be defined as in section 2.1, i.e. $a^+ = \max(a,0)$ and $a^- = \min(a,0)$. Then, $a = a^+ + a^-$ and for any sequence $\{a_j, j \in I\}$ of numbers such that $\sum_{j\in I}|a_j| < \infty$ and $\sum_{j\in I} a_j = 0$ we have $\sum_j a_j^+ = -\sum_j a_j^-$. Further, we note that $(a-b)^+ = a-\min(a,b)$ for any pair of numbers a,b. Fix now $i \in I$ and $n > \nu$. Then,

$$\sum_{j\in A} p_{ij}^n(f) - \sum_{j\in A} p_{rj}^n(f) = \sum_{k\in I} \{p_{ik}^\nu(f) - p_{rk}^\nu(f)\} \sum_{j\in A} p_{kj}^{n-\nu}(f) =$$

$$= \sum_{k\in I} \{p_{ik}^\nu(f) - p_{rk}^\nu(f)\}^+ \sum_{j\in A} p_{kj}^{n-\nu}(f) + \sum_{k\in I} \{p_{ik}^\nu(f) - p_{rk}^\nu(f)\}^- \sum_{j\in A} p_{kj}^{n-\nu}(f) \le$$

$$\le \{M_{n-\nu} - m_{n-\nu}\} \sum_{k\in I} \{p_{ik}^\nu(f) - p_{rk}^\nu(f)\}^+ =$$

$$= \{M_{n-\nu} - m_{n-\nu}\}\{1 - \sum_{k\in I} \min[p_{ik}^\nu(f), p_{rk}^\nu(f)]\} \le$$

$$\le (1-\rho)(M_{n-\nu} - m_{n-\nu}).$$

Since i and r were arbitrarily chosen, it follows that

$$M_n - m_n \le (1-\rho)\{M_{n-\nu} - m_{n-\nu}\} \quad \text{for all } n > \nu.$$

Hence, since $M_n - m_n$ is non-increasing in $n \ge 1$,

(5.2.14) $M_n - m_n \le (1-\rho)^{\lfloor n/\nu \rfloor}$ for all $n \ge 1$.

Together (5.2.11) and (5.2.12) imply that for some finite non-negative number $\pi(A)$

$$\lim_{n\to\infty} M_n = \lim_{n\to\infty} m_n = \pi(A).$$

Further for any $n \ge 1$,

(5.2.15) $m_n \le \pi(A) \le M_n$ and $m_n \le \sum_{j\in A} p_{ij}^n(f) \le M_n$ for all $i \in I$.

It now follows from (5.2.12) and (5.2.13) that

$$\Big|\sum_{j\in A} p_{ij}^n(f) - \pi(A)\Big| \le (1-\rho)^{\lfloor n/\nu \rfloor} \quad \text{for all } n \ge 1.$$

Since this relation holds for any $A \subseteq I$, it follows that $\pi(.)$ is a probability measure on the class of all subsets of I which completes the proof. □

Proof of Theorem 5.2.4. (i) Suppose that condition C5 holds. Since for any $i,j \in I$ and $n \geq 1$ the function $p_{ij}^n(f)$ is continuous on F, it follows from (5.2.6) that for any $j \in I$ the function $\pi_j(f)$ is continuous in $f \in F$. Now, let $\{K_n, n = 1,2,....\}$ be a sequence of finite sets $K_n \subset I$ such that $K_{n+1} \supseteq K_n$ for all $n \geq 1$ and $\lim_{n\to\infty} K_n = I$. Let $a_n(f) = \sum_{j\in K_n} \pi_j(f)$ for $n \geq 1$ and $f \in F$. Then the function $a_n(f)$ is continuous in $f \in F$ for any $n \geq 1$, and moreover, for any $f \in F$ we have $a_{n+1}(f) \geq a_n(f)$ for all $n \geq 1$ and $\lim_{n\to\infty} a_n(f) = 1$. Now, since F is compact, we have by theorem 7.13 in RUDIN [101] that $a_n(f)$ converges to 1 uniformly in $f \in F$ as $n \to \infty$. Hence for each $\varepsilon > 0$ there is a finite integer n such that $a_n(f) \geq 1-\varepsilon$ for all $f \in F$. This shows that we can find a finite set K and a number $\delta > 0$ such that

$$(5.2.16) \qquad \sum_{j\in K} \pi_j(f) \geq \delta \quad \text{for all } f \in F.$$

By (5.2.6) and the finiteness of K, we can find an integer $\nu \geq 1$ such that $p_{ij}^\nu(f) \geq \pi_j(f) - \delta/(2|K|)$ for all $i \in I$, $f \in F$ and $j \in K$ where $|K|$ denotes, once again, the number of states in K. Together this inequality and (5.2.14) imply condition C2. Further we get from (5.2.14) that for any $f \in F$ there is a state s_f such that $\pi_{s_f}(f) \geq \delta/|K|$ and so $p_{is_f}^\nu(f) \geq \delta/(2|K|)$ for all $i \in I$ and $f \in F$. This inequality verifies condition C9 which completes the proof of part (i).

(ii) Since C9 implies C8 and in its turn C8 implies C3 and since C4 implies C5, this part follows by using theorem 5.2.3 and part (i) of theorem 5.2.4. □

Proof of Theorem 5.2.5. To prove the theorem, we shall use a classical perturbation of the stochastic matrices $P(f)$, $f \in F$, which is analogous to the data-transformation (1.8.1) and (1.8.2) with $\sigma = 1$. Fix any number τ with $0 < \tau < 1$ and let $\bar{P} = (\bar{P}(f), f \in F)$ be the set of stochastic matrices $\bar{P}(f) = (\bar{p}_{ij}(f))$, $i,j \in I$ such that for any $f \in F$ and $i,j \in I$:

$$\bar{p}_{ij}(f) = \begin{cases} \tau p_{ij}(f) & \text{for } j \neq i \\ 1 - \tau + \tau p_{ii}(f) & \text{for } j = i \end{cases}$$

Note that, by $p_{ii}(f) \geq 1 - \tau > 0$ for all $i \in I$ and $f \in F$, the stochastic matrix $\bar{P}(f)$ is *aperiodic* for all $f \in F$. Also note that for any $i,j \in I$ the function $\bar{p}_{ij}(f)$ is continuous in $f \in F$ and for any $f \in F$, the stochastic

matrix $\bar{P}(f)$ has no two disjoint closed sets. Define for the stochastic matrices $\bar{P}(f)$ the taboo probabilities $\bar{t}^{n}_{iA}(f)$ and the mean recurrence times $\bar{\mu}_{iA}(f)$ as in (5.2.2) and (5.2.3). By induction on n, it is straightforward to verify that for any $f \in F$

$$\bar{t}^{n}_{ij}(f) = \sum_{k=0}^{n} \binom{n}{k} (1-\tau)^{n-k} \tau^{k} t^{k}_{ij}(f) \quad \text{for all } n = 0,1,\ldots \text{ and } i,j \in I \text{ with } i \neq j, \tag{5.2.17}$$

where $\bar{t}^{0}_{ij}(f) = t^{0}_{ij}(f) = 1$. From the relations (5.2.3) and (5.2.17) we get

$$\bar{\mu}_{ij}(f) = \frac{\mu_{ij}(f)}{\tau} \quad \text{for all } i,j \in I \text{ with } i \neq j \text{ and } f \in F. \tag{5.2.18}$$

We note that this relation is intuitively clear by a direct probabilistic interpretation.

We now prove (i). Suppose that the condition C2 holds with triple (K,ν,ρ). Then, by $\bar{p}_{ij}(f) \geq \tau p_{ij}(f)$ for all $i,j \in I$ and $f \in F$, we have

$$\sum_{j \in K} \bar{p}^{\nu}_{ij}(f) \geq \tau^{\nu} \sum_{j \in K} p^{\nu}_{ij}(f) \geq \tau^{\nu} \rho \quad \text{for all } i \in I \text{ and } f \in F.$$

Hence the condition C2 applies to the set $\bar{P} = (\bar{P}(f), f \in F)$. Moreover we have that any stochastic matrix $\bar{P}(f)$, $f \in F$ is aperiodic. Now, by the combination of th.5.2.2, th.5.2.3 and part (ii) of th.5.2.4, it follows that condition C9 applies to the set $\bar{P}$. Since condition C9 implies C7, we have that condition C7 applies to the set $\bar{P}$. Now, by invoking (5.2.16), it follows that the condition C7 holds for the set $P = (P(f), f \in F)$ as was to be proved.

Next we prove (ii). Suppose that condition C6 holds. Then, by invoking again (5.2.16), we have that condition C6 applies to the set $\bar{P}$. Hence there is a finite number B such that for any $f \in F$ there exists a state s_f such that

$$\bar{\mu}_{is_f}(f) = 1 + \sum_{n=1}^{\infty} \bar{t}^{n}_{is_f}(f) \leq B \text{ for all } i \in I. \tag{5.2.19}$$

Fix now $0 < \gamma < 1$. Since for any $f \in F$ and $i \in I$ the taboo probability $\bar{t}^{n}_{is_f}(f)$ is non-increasing in n, it follows that there is an integer $N \geq 1$ such that

$$\bar{t}^{N}_{is_f}(f) \leq \gamma \quad \text{for all } i \in I \text{ and } f \in F. \tag{5.2.20}$$

(Supposing the contrary to (5.2.20) gives a contradiction with (5.2.19)). Together the inequality (5.2.20) and the fact that $\bar{p}_{kk}(f) \geq 1 - \tau$ for all $k \in I$ and $f \in F$ imply

$$\bar{p}^{N}_{is_f}(f) \geq (1-\tau)^{N-1}(1-\gamma) \quad \text{for all } i \in I \text{ and } f \in F.$$

This shows that condition C8 applies to the set $\bar{P}$. Next, by part (ii) of theorem 5.2.4 condition C9 applies to the set $\bar{P}$. Since C9 implies C7, it follows that condition C7 applies to the set $\bar{P}$. Now by invoking again (5.2.16) we have that condition C7 holds for the stochastic matrices $P(f)$, $f \in F$ as was to be verified.

We obtain part (iii) of the theorem by noting that C7 trivially implies both C1 and C6 and using part (i) of theorem 5.2.1 and parts (i)-(ii) of theorem 5.2.5. Finally, part (iv) of the theorem is an immediate consequence of the theorems 5.2.2-5.2.4 and part (iii) of the theorem.

5.3. THE AVERAGE COSTS OPTIMALITY EQUATION

In this section we shall prove that under each of the conditions C1-C9 the optimality equation for the average costs has a bounded solution. To establish the optimality equation, we shall employ a simple but very useful data-transformation analogous to the one introduced in SCHWEITZER [108], which is the exact analogue of (1.9.9) and (1.9.10). We associate with the semi-Markov model a discrete-time Markov decision model with state space I, the set A(i) as the set of possible actions for state i, one-step costs $\hat{c}(i,a)$, one-step transition times $\hat{\tau}(i,a) \equiv 1$ and one-step transition probabilities $\hat{p}_{ij}(a)$ where, for all $i,j \in I$ and $a \in A(i)$

$$\hat{c}(i,a) = \frac{c(i,a)}{\tau(i,a)} \quad \text{and} \quad \hat{p}_{ij}(a) = \frac{\tau}{\tau(i,a)}\{p_{ij}(a)-\delta_{ij}\} + \delta_{ij}$$

with δ_{ij} representing the Kronecker function, as before, and where τ is a fixed number such that

$$0 < \tau < \delta = \inf_{i,a}\{\tau(i,a)/(1-p_{ii}(a)) \mid p_{ii}(a) < 1\}.$$

Observe that $\delta > 0$ and note that the assumptions A1 - A4 also apply to the transformed model. Further, letting the finite positive number γ be equal to $\sup_{i,a} \tau(i,a)$, it is readily verified that for all $i \in I$ and $a \in A(i)$ we

have that $\{\hat{p}_{ij}(a), j \in I\}$ is a probability distribution with

$$(5.3.1) \quad \hat{p}_{ii}(a) \geq 1 - \frac{\tau}{\delta} > 0 \text{ and } \hat{p}_{ij}(a) \geq \frac{\tau}{\gamma} p_{ij}(a) \quad \text{for } j \neq i.$$

By the first part of (5.3.1), we have that for any $f \in F$ the stochastic matrix $\hat{P}(f)$ is *aperiodic*. This aperiodicity will play a crucial role in the analysis below. Also, letting the finite positive number ϕ be equal to $\min[1-\tau/\delta, \tau/\gamma]$ and using (5.3.1), it is immediately verified that for any set $A \subseteq I$ and all $n \geq 1$,

$$(5.3.2) \quad \sum_{j \in A} \hat{p}^n_{ij}(f) \geq \phi^n \sum_{j \in A} p^n_{ij}(f) \quad \text{for all } i \in I \text{ and } f \in F.$$

By parts (iii)-(iv) of theorem 5.2.5, we have that each of the conditions C1-C9 implies condition C2. Hence, by (5.3.2) we have that under each of the conditions C1-C9, holding for the set $P = (P(f), f \in F)$, condition C2 applies to the set $\hat{P} = (\hat{P}(f), f \in F)$ as well. Together this result, the aperiodicity of the policies in $\hat{P}$ and the theorems 5.2.2-5.2.3, imply that under each of the conditions C1-C9 there is a number $\rho > 0$ and an integer $\nu \geq 1$, such that for any $f \in F$, the stochastic matrix $\hat{P}(f)$ has a unique stationary probability distribution $\{\hat{\pi}_j(f), j \in I\}$ (say) with

$$(5.3.3) \quad \Big| \sum_{j \in A} \hat{p}^n_{ij}(f) - \sum_{j \in A} \hat{\pi}_j(f) \Big| \leq (1-\rho)^{\lfloor n/\nu \rfloor} \quad \text{for all } i \in I,\ A \subseteq I \text{ and } n \geq 1.$$

This result will underly the derivation of the optimality equation for the transformed model (cf. also TIJMS [123]) from which we easily get the optimality equation for the semi-Markov decision model considered. Before showing this, we give the following lemma.

LEMMA 5.3.1. *Let* $\{h_n(.), n \geq 1\}$ *be a sequence of bounded functions on* I *such that, for some bounded function* $h(.)$ *on* I, $\lim_{n\to\infty} h_n(i) = h(i)$ *for all* $i \in I$. *Then, for any* $i \in I$,

$$\lim_{n\to\infty} \min_{a \in A(i)} \{c(i,a) + \sum_{j \in I} p_{ij}(a) h_n(j)\} = \min_{a \in A(i)} \{c(i,a) + \sum_{j \in I} p_{ij}(a) h(j)\}.$$

PROOF. Fix $i \in I$. For any $n \geq 1$, let action a_n minimize $c(i,a) + \sum_j p_{ij}(a) h_n(j)$ for $a \in A(i)$. Observe that, by A1, such a minimizing action exists.

Now, let $\{n_k, k \geq 1\}$ be any infinite sequence of positive integers. Since $A(i)$ is a compact metric space, we can choose an action $a^* \in A(i)$ and a subsequence $\{t_k, k \geq 1\}$ of $\{n_k, k \geq 1\}$ such that $a_{t_k} \to a^*$ as $k \to \infty$. Using the

fact that, by A1, $\sum_{j\in A} p_{ij}(a_{t_k}) \to \sum_{j\in A} p_{ij}(a^*)$ as $k \to \infty$ for any set $A \subseteq I$ and using proposition 18 on p.232 in ROYDEN [100], it follows from

$$c(i,a_{t_k}) + \sum_{j\in I} p_{ij}(a_{t_k})h_{t_k}(j) \leq c(i,a) + \sum_{j\in I} p_{ij}(a)h_{t_k}(j)$$

for all $a \in A(i)$ and $k \geq 1$, that

$$\lim_{k\to\infty} \min_{a\in A(i)} \{c(i,a) + \sum_{j\in I} p_{ij}(a)h_{t_k}(j)\} = \min_{a\in A(i)} \{c(i,a) + \sum_{j\in I} p_{ij}(a)h(j)\}$$

which proves the lemma. □

We now prove the main result of this section.

THEOREM 5.3.2. *Under each of the conditions* C1-C9 *there exists a finite constant* g^* *and a bounded function* $v^*(i)$, $i \in I$ *such that*

$$(5.3.4) \quad v^*(i) = \min_{a\in A(i)} \{c(i,a) - g^*\tau(i,a) + \sum_{j\in I} p_{ij}(a)v^*(j)\} \quad \text{for all } i \in I.$$

The constant g^* *is uniquely determined and the bounded function* $v^*(i)$, $i \in I$ *is uniquely determined up to an additive constant.*

PROOF. Consider first the transformed model. As shown above, there is an integer $\nu \geq 1$ and a number $\rho > 0$ such that for any $f \in F$ the stochastic matrix $\hat{P}(f)$ has a stationary probability distribution satisfying (5.3.3). To verify the optimality equation for the transformed model, consider first the discounted cost criterion. For any $0 < \beta < 1$, define for each policy π (observe that $\hat{c}(i,a)$ is uniformly bounded in i,a),

$$\hat{V}_\beta(i,\pi) = E_\pi[\sum_{n=0}^{\infty} \beta^n \hat{c}(X_n,a_n) \mid X_0 = i] \quad \text{for } i \in I,$$

and let $\hat{V}_\beta(i) = \inf_\pi \hat{V}_\beta(i,\pi)$, $i \in I$. It is known that for any $0 < \beta < 1$ the function $\hat{V}_\beta(i)$, $i \in I$ is the unique bounded solution to: (cf. e.g. MAITRA [82])

$$(5.3.5) \quad \hat{V}_\beta(i) = \min_{a\in A(i)} \{\hat{c}(i,a) + \beta \sum_{j\in I} \hat{p}_{ij}(a)\, \hat{V}_\beta(j)\}, \quad i \in I,$$

and, moreover,

$$(5.3.6) \quad \hat{V}_\beta(i,f_\beta^{(\infty)}) = \hat{V}_\beta(i) \quad \text{for all } i \in I,$$

for any $f_\beta \in F$ such that $f_\beta(i)$ minimizes the right-side of (5.3.5) for all $i \in I$. For any $0 < \beta < 1$ and $f \in F$, we have

$$(5.3.7)\qquad \hat{V}_\beta(i,f^{(\infty)}) = \sum_{n=0}^{\infty} \beta^n \sum_{j\in I} \hat{p}_{ij}^{\,n}(f)\hat{c}(j,f(j)) \quad \text{for all } i \in I$$

where $\hat{p}_{ij}^{\,0}(f) = \delta_{ij}$. From (5.3.3), it follows that for any $f \in F$, $i, k \in I$ and $n \geq 1$ the total variation of the signed measure $\mu(A) = \sum_{j\in A} \hat{p}_{ij}^{\,n}(f) - \sum_{j\in A} \hat{p}_{kj}^{\,n}(f)$ is bounded by $4(1-\rho)^{\lfloor n/\nu \rfloor}$. Using this result and letting B be any finite number such that $|\hat{c}(i,a)| \leq B$ for all i,a, it follows from (5.3.7) that, for any $0 < \beta < 1$ and all $f \in F$,

$$|\hat{V}_\beta(i,f^{(\infty)}) - \hat{V}_\beta(k,f^{(\infty)})| \leq 4B \sum_{n=0}^{\infty} (1-\rho)^{\lfloor n/\nu \rfloor} \leq \frac{4B\nu}{\rho} \quad \text{for all } i,k \in I.$$

Hence, by (5.3.6),

$$|\hat{V}_\beta(i) - \hat{V}_\beta(k)| \leq \frac{4B\nu}{\rho} \quad \text{for all } i,k \in I \text{ and all } 0 < \beta < 1.$$

Now, by using lemma 5.3.1 and by making an obvious modification on the proof of theorem 6.18 in ROSS [98], there exists a finite constant g and a bounded function v(i), $i \in I$ such that

$$(5.3.8)\qquad v(i) = \min_{a\in A(i)} \{\hat{c}(i,a) - g + \sum_{j\in I} \hat{p}_{ij}(a)v(j)\} \quad \text{for all } i \in I.$$

We shall now verify that $g^* = g$ and $v^*(i) = \tau v(i)$, $i \in I$ satisfy (5.3.4). To do this, observe that (5.3.8) can be equivalently written as

$$v(i) \leq \frac{c(i,a)}{\tau(i,a)} - g + \frac{\tau}{\tau(i,a)} \sum_{j\in A} p_{ij}(a)v(j) + (1 - \frac{\tau}{\tau(i,a)})v(i)$$

for all $i \in I$ and $a \in A(i)$

where for any $i \in I$ the equality holds for at least one $a \in A(i)$. Multiplying both sides of this inequality by $\tau(i,a) > 0$, we find

$$0 \leq c(i,a) - g\tau(i,a) + \tau \sum_{j\in I} p_{ij}(a)v(j) - \tau v(i), \quad i \in I \text{ and } a \in A(i),$$

where for any $i \in I$ the equality sign holds for at least one $a \in A(i)$. This proves that $g^* = g$ and $v^*(i) = \tau v(i)$, $i \in I$ satisfy the optimality equation (5.3.4). By theorem 6.17 in ROSS [98], we have that the constant g^* in (5.3.4) is uniquely determined and, by lemma 3 in HORDIJK, SCHWEITZER and TIJMS [61], we have that the function $v^*(i)$, $i \in I$ in (5.3.4) is uniquely determined up to an additive constant.

For any policy π, define for all $i \in I$ and $n \geq 1$,

$$V_n(i,\pi) = E_\pi[\sum_{k=0}^{n} c(X_k,a_k)\,|\,X_0=i] \text{ and } T_n(i,\pi) = E_\pi[\sum_{k=0}^{n} \tau(X_k,a_k)\,|\,X_0=i].$$

Define a policy π^* to be *average cost optimal in the strong sense if*

$$\limsup_{n\to\infty} \frac{V_n(i,\pi^*)}{T_n(i,\pi^*)} \le \liminf_{n\to\infty} \frac{V_n(i,\pi)}{T_n(i,\pi)} \tag{5.3.9}$$

for all $i \in I$ and any policy π.

An examination of the proof of theorem 7.6 in ROSS [98] gives the following theorem.

THEOREM 5.3.2. *Let* $\{g^*, v^*(i), i \in I\}$ *be any bounded solution to the optimality equation* (5.3.4) *and let* $f_0 \in F$ *be such that* $f_0(i)$ *minimizes the right side of* (5.3.4) *for all* $i \in I$. *Then*

$$\liminf_{n\to\infty} \frac{V_n(i,\pi)}{T_n(i,\pi)} \ge g^* \text{ for all } i \in I \text{ and any policy } \pi$$

and

$$\lim_{n\to\infty} \frac{V_n(i,f_0^{(\infty)})}{T_n(i,f_0^{(\infty)})} = g^* \text{ for all } i \in I,$$

so the stationary policy $f_0^{(\infty)}$ *is average cost optimal in the strong sense.*

, Although strong optimality as in (5.3.9) is immediate when the optimality equation has a bounded solution, this criterion may be difficult to verify directly. In the literature the lim sup and lim inf average cost criteria are usually considered when the optimality equation cannot be established. However, the relations (2)-(4) in FLYNN [46] show that these criteria are essentially weaker than the criterion (5.3.9).

Finally, letting Z(t) be the total costs incurred up to time t and using theorem 7.5 in ROSS [98], we have under each of the conditions C1-C9 that for any $f \in F$,

$$\lim_{t\to\infty} t^{-1} E_{f(\infty)}[Z(t)\,|\,X_0=i] = \lim_{n\to\infty} \frac{V_n(i,f^{(\infty)})}{T_n(i,f^{(\infty)})} =$$

$$= \textstyle\sum_{j\in I} c(j,f(j))\ \pi_j(f) \ /\ \sum_{j\in I} \tau(j,f(j))\ \pi_j(f) \quad \text{for all } i \in I \tag{5.3.10}$$

where $\{\pi_j(f), j \in I\}$ is the unique stationary probability distribution of $P(f)$.

5.4. THE VALUE-ITERATION METHOD

In this section it is assumed that one of the conditions C1-C9 holds. We shall show that a bounded solution to the optimality equation (5.3.4) may be obtained by using value-iteration. We note that value-iteration may not work when some of the one-step transition probability matrices associated with the stationary policies are periodic. However, in the proof of theorem 5.3.2 we have seen that by the data-transformation given in section 3 any semi-Markov decision problem can be transformed into a discrete-time Markov decision problem with aperiodic transition probability matrices so that any bounded solution $\{g;\ v(i),\ i \in I\}$ to the optimality equation of the transformed model gives a bounded solution $\{g^*=g;\ v^*(i) = \tau v(i),\ i \in I\}$ to (5.3.4). Therefore we assume in the remainder of this section that $\tau(i,a) = 1$ for all i,a and that $P(f)$ is *aperiodic* for all $f \in F$.

Let $\{g^*;\ v^*(i),\ i \in I\}$ be any bounded set of numbers satisfying

$$(5.4.1) \qquad v^*(i) = \min_{a \in A(i)} \{c(i,a) - g^* + \sum_{j \in I} p_{ij}(a) v^*(j)\} \quad \text{for all } i \in I.$$

For any given bounded function $v_0(i)$, $i \in I$, define for $n = 1,2,\ldots$ the bounded function $v_n(i)$, $i \in I$ by the value-iteration equation

$$(5.4.2) \qquad v_n(i) = \min_{a \in A(i)} \{c(i,a) + \sum_{j \in I} p_{ij}(a) v_{n-1}(j)\} \quad \text{for } i \in I.$$

Observe that, by A1, the minimum on the right side of (5.4.2) is attained for all i. The asymptotic behaviour of the sequence $\{v_n(i) - ng^*,\ n \geq 1\}$ for $i \in I$ has been studied in [61] where the action sets $A(i)$ were taken to be finite. This finiteness is in fact only used to verify relation (18) in [61]; however, by invoking lemma 32 on p.178 in [100], it follows that the results in [61] also apply when for any $i \in I$ the set $A(i)$ is a compact metric space such that both $c(i,a)$ and $p_{ij}(a)$ for $j \in I$ are continuous on $A(i)$. Since the assumptions 1 - 5 in [61] are satisfied, we have for some constant c that

$$(5.4.3) \qquad \lim_{n \to \infty} \{v_n(i) - ng^*\} = v^*(i) + c \quad \text{for all } i \in I.$$

Hence, by choosing some state i_0 and defining $y_n = v_n(i_0) - v_{n-1}(i_0)$ and $w_n(i) = v_n(i) - v_n(i_0)$ for $i \in I$ and $n \geq 1$, it follows that the bounded numbers $\{y_n;\ w_n(i),\ i \in I\}$ converge as $n \to \infty$ to a bounded solution of (5.4.1).

it was pointed out in remark 2 in chapter 2, that even in the finite state space case, C3' with $\nu \geq 2$ is essentially stronger than C3 with $\nu \geq 2$.

5.5. THE POLICY ITERATION METHOD

Throughout this section, it is assumed that the following strengthening of condition C6 holds.

C6': There is a finite number B and a state s such that $\mu_{is}(f) \leq B$ for all $i \in I$ and all $f \in F$.

Condition C6' was first introduced in ROSS [98]. Since C6' implies C6, it follows from the combination of th.5.2.5 (iii) and th.5.2.1 (ii) that for each $f \in F$, the stochastic matrix P(f) has a unique stationary probability distribution $\{\pi_j(f), j \in I\}$ (say). We further suppose that the assumptions A1 and A2 together with the assumption A3' hold where

A3': There are finite numbers $\varepsilon > 0$ and M such that $\tau(i,a) \leq M$ for all i,a and $\sum_{j\in I} \tau(i,f(j))\pi_j(f) \geq \varepsilon$ for all $f \in F$.

This assumption slightly weakens A3 and allows instantaneous transitions. Using ideas from a convergence proof given in [19] for a policy iteration approach to controlled Markov processes with a general state space, it will be shown that both the average costs and the relative cost functions of the stationary policies generated by the policy iteration method converge so that the limits constitute a bounded solution to the optimality equation (5.3.4). Partial convergence results of this type were obtained in DERMAN [24] under the restrictive additional assumption of no transient states under any P(f), $f \in F$.

We first give some preliminaries. Let the state s be as in condition C6'. We have for any $f \in F$ that

$$\pi_i(f) = \sum_{j\in I} p_{ji}(f)\pi_j(f) \quad \text{for all } i \in I. \tag{5.5.1}$$

Moreover, we have from Markov chain theory

$$\pi_i(f) = \sum_{n=0}^{\infty} {}_s p_{si}^n(f)/\mu_{ss}(f) \quad \text{for all } i \in I \tag{5.5.2}$$

where ${}_s p_{ij}^0 = \delta_{ij}$ for all i,j and

$${}_s p_{ij}^n(f) = P_{f(\infty)}\{X_n=j,\ X_k \neq s \text{ for } 1 \leq k \leq n | X_0=i\} \text{ for } i,j \in I \text{ and } n \geq 1. \tag{5.5.3}$$

Next, we note that, by choosing $f_n \in F$ such that $f_n(i)$ minimizes the right side of (5.4.2) for all i, and when defining the average costs vector g(f) by (5.3.10), it follows by making minor modifications on standard arguments used in HASTINGS [53] and ODONI [89] (cf. also (1.8.13) and (1.8.14)) that, for all $n \geq 1$:

$$\inf_{i\in I}\{v_n(i)-v_{n-1}(i)\} \leq g^* \leq g(f_n) \leq \sup_{i\in I}\{v_n(i)-v_{n-1}(i)\},$$

where $\inf_i\{v_n(i)-v_{n-1}(i)\}$ and $\sup_i\{v_n(i)-v_{n-1}(i)\}$ are non-decreasing and non-increasing respectively in $n \geq 1$.

Finally, consider the special case where condition C3 with $\nu=1$ holds. Let $\mathcal{B}$ be the class of all bounded functions on I and define the mapping $T: \mathcal{B} \to \mathcal{B}$ by

$$Tu(i) = \min_{a\in A(i)} \{c(i,a) + \sum_{j\in I} p_{ij}(a)u(j)\}$$

and define $sp[u] = \sup_i u(i) - \inf_i u(i)$ for $u \in \mathcal{B}$. Then a repetition of the proof of theorem 2.2.4 and 2.2.5 shows that, for some number $\rho > 0$ $sp[Tu - Tw] \leq (1-\rho)\, sp[u - w]$ for all $u,w \in \mathcal{B}$ i.e. T is contracting with respect to the sp[.] "norm". Next, using this result and the existence of a bounded solution to (5.4.1), it is readily verified (cf. the proof of th. 2.2.1) that $|v_n(i)-ng^*-v^*(i)| \leq (1-\rho)^n\, sp[v_0-v^*]$ for all $i \in I$ and $n \geq 1$, i.e. in this case the convergence in (5.4.3) is geometrically fast and uniform in i.

Establishing the *rate* of convergence in the *general* case where in the transformed model C3 holds with $\nu \geq 1$ (i.e. where either one of the conditions C1-C9 applies to the original model) remains an outstanding problem. The analysis given in section 1.6 exploits the finite dimensionality of the state space heavily, and cannot be used in the *infinite* state space case. Note from the proof of th.2.2.5 that the T-operator can be shown to be (ν-step) contracting with respect to the quasi-norm sp[.] on B, in case condition C3' applies (cf. condition (S) in (2.2.4)):

C3': There is an integer $\nu \geq 1$ and a number $\rho > 0$ such that

$$\inf_{i_1\neq i_2}\{\sum_{j\in I}\min[P(f_\nu)\ldots P(f_1)_{i_1 j}; P(h_\nu)\ldots P(h_1)_{i_2 j}]\} \geq \rho$$

for all $(f_1,\ldots,f_\nu)$ and $(h_1,\ldots,h_\nu) \in F$.

Hence, using the proof of th.2.2.1 it follows that under C3', value-iteration is guaranteed to exhibit a geometric rate of convergence. Finally,

Observe that

$$(5.5.4)\qquad \mu_{is}(f) = 1 + \sum_{n=1}^{\infty} \sum_{j\in I} {}_s p_{ij}^{n}(f) \quad \text{for all } i \in I.$$

Further, for any $f \in F$, define

$$(5.5.5)\qquad g(f) = \sum_{i\in I} c(i,f(i))\pi_i(f) \Big/ \sum_{i\in I} \tau(i,f(i))\pi_i(f),$$

and

$$(5.5.6)\qquad w_i(f) = \sum_{n=0}^{\infty} \sum_{j\in I} \{c(j,f(j)) - g(f)\,\tau(j,f(j))\}\, {}_s p_{ij}^{n}(f), \quad i \in I.$$

We note that $g(f)$ gives the long-run average expected costs per unit time under policy $f^{(\infty)}$ for each initial state. Also $w_i(f)$ for $i \in I$ can be interpreted as a relative cost function.

It is immediately verified from (5.5.2) and (5.5.4)-(5.5.6) that, for any $f \in F$, the function $w_i(f)$, $i \in I$ is bounded and has the property that

$$(5.5.7)\qquad w_s(f) = 0.$$

Consider now for fixed $f \in F$ the following system of linear equations in $\{g;\ v_i,\ i \in I\}$,

$$(5.5.8)\qquad v_i = c(i,f(i)) - g\tau(i,f(i)) + \sum_{j\in I} p_{ij}(f(i))v_j \quad \text{for } i \in I.$$

We recall the following well-known theorem (see [18] and DERMAN & VEINOTT [27]).

THEOREM 5.5.1. *For any* $f \in F$,

(a) *The set of numbers* $\{g = g(f);\ v_i = w_i(f),\ i \in I\}$ *is a bounded solution to* (5.5.8).

(b) *In any bounded solution* $\{g;\ v_i,\ i \in I\}$ *to* (5.5.8), $g = g(f)$.

(c) *For any two bounded solutions* $\{g;v\}$ *and* $\{g;u\}$ *to* (5.5.8) *there is a constant* c *such that* $v_i - u_i = c$ *for all* $i \in I$.

(d) *For any* $j \in I$, *there is a unique bounded solution* $\{g;\ v\}$ *to* (5.5.8) *such that* $v_j = 0$.

In general it will be difficult to solve the system of equations (5.5.8). However, in a number of applications the particular structure of the problem may be exploited to solve these equations, cf. [18].

By the assumptions A2 and A3', we have

LEMMA 5.5.2. *The set of numbers* $\{g(f), f \in F\}$ *is bounded.*

For any $f \in F$ and any bounded solution $\{g(f);\ v_i(f),\ i \in I\}$ to (5.5.8), define

$$(5.5.9)\quad T(i,a,v(f)) = c(i,a) - g(f)\tau(i,a) + \sum_{j\in I} p_{ij}(a)v_j(f) \quad \text{for } i \in I \text{ and } a \in A(i).$$

Observe that

$$(5.5.10)\quad T(i, f(i), v(f)) = v_i(f) \quad \text{for all } i \in I \text{ and } f \in F.$$

The following lemma shows how the stationary policy $f^{(\infty)}$ can be improved to a stationary policy $h^{(\infty)}$ whose average costs are less than or equal to that of $f^{(\infty)}$.

LEMMA 5.5.3. *Let* $f \in F$ *and let* $\{g(f);v(f)\}$ *be any bounded solution to* (5.5.8). *Suppose* $h \in F$ *is such that*

$$(5.5.11)\quad T(i, h(i), v(f)) \le v_i(f) \quad \text{for all } i \in I.$$

Then $g(h) \le g(f)$.

PROOF. The proof is standard. Multiply both sides of the inequality (5.5.11) with $\pi_i(f)$ and sum over $i \in I$. Next the desired result follows after an interchange of the order of summation which is justified by the boundedness of $v(f)$ and using the steady-state equation (5.5.1) for policy $h^{(\infty)}$.

We now formulate the policy-iteration method.

Policy Iteration Method

Step 0. Initialize with any $f_1 \in F$.

Step 1. Let $f^{(\infty)}$ be the current policy. Determine the unique bounded solution $\{g(f);\ w(f)\}$ to the system of linear equations (5.5.8) in which $v_s = 0$.

Step 2. Determine $f' \in F$ such that $T(i, f'(i), w(f)) = \min_{a\in A(i)} T(i,a,w(f))$ for all $i \in I$.

Let $\{f_n^{(\infty)},\ n \ge 1\}$ be the sequence of stationary policies generated by the policy iteration method. Observe that, by part (c) of theorem 5.5.1, f_{n+1} is independent of the particular choice of the bounded solution to (5.5.8) with $f = f_n$. By lemma 5.5.3,

$$(5.5.12)\quad g(f_{n+1}) \le g(f_n) \quad \text{for all } n \ge 1.$$

We shall prove that the bounded numbers $\{g(f_n);\ w_i(f_n),\ i \in I\}$ converge as

$n \to \infty$ to a bounded solution of the optimality equation (5.3.4). To do this, we shall use a modified semi-Markov decision model specified by the five objects $(\bar{I}, \bar{A}(i), \bar{p}_{ij}(a), \bar{c}(i,a), \bar{\tau}(i,a))$ where, for some artificial state ∞ and action a_∞ (say),

$$\bar{I} = I \cup \{\infty\}, \quad \bar{A}(i) = A(i) \quad \text{for } i \in I, \quad \bar{A}(\infty) = \{a_\infty\},$$

$$\bar{c}(i,a) = c(i,a), \quad \bar{\tau}(i,a) = \tau(i,a) \quad \text{for } i \in I \text{ and } a \in A(i),$$

$$\bar{c}(\infty,a_\infty) = \bar{\tau}(\infty,a_\infty) = 0, \quad \bar{p}_{\infty s}(a_\infty) = 1, \quad \bar{p}_{\infty j}(a_\infty) = 0 \text{ for } j \neq s,$$

$$\bar{p}_{ij}(a) = \begin{cases} p_{ij}(a) & \text{for } i,j \in I,\ a \in A(i),\ j \neq s \\ p_{is}(a) & \text{for } i \in I,\ a \in A(i),\ j = \infty. \end{cases}$$

In fact this modified model is identical to the original semi-Markov decision model, except that before any transition to state s there first occurs a transition to state ∞ after which an instantaneous transition occurs to state s involving no costs. For the modified model, denote by $\bar{F}$ the class of all functions h which add to each state $i \in \bar{I}$ a single action $h(i) \in \bar{A}(i)$ and associate with any $h \in \bar{F}$ the stochastic matrix $\bar{P}(h) = (\bar{p}_{ij}(h(i)))$, $i,j \in \bar{I}$. Since $h(\infty) = a_\infty$ for all $h \in \bar{F}$, there is a one-to-one correspondence between F and $\bar{F}$. For any $f \in F$, denote by $\bar{f}$ the unique element in $\bar{F}$ such that $\bar{f}(i) = f(i)$ for all $i \in I$. It is immediate that there is a finite number B (say) such that under any stochastic matrix $\bar{P}(\bar{f})$, $f \in F$ the expected number of transitions required before the first return to state ∞ is bounded by B for any starting state $i \in \bar{I}$. Hence condition C6' with state s replaced by state ∞ also applies to the modified model. This result together with the fact that $\bar{A}(\infty)$ consists of a *single* action will play a crucial role in the convergence proof below. Further, for any $f \in F$, the stochastic matrix $\bar{P}(\bar{f})$ has a unique stationary probability distribution $\{\bar{\pi}_j(f),\ j \in I\}$. Using the steady-state equation, we have for any $f \in F$ that $\bar{\pi}_s(f) = \bar{\pi}_\infty(f)$ and $\bar{\pi}_i(f) = \pi_i(f)/\{1+\pi_s(f)\}$ for all $i \in I$. Hence the assumptions A1, A2 and A3' also apply to the modified model. Further, letting

$$\bar{g}(f) = \sum_{i \in \bar{I}} \bar{c}(i,\bar{f}(i))\bar{\pi}_i(f) \Big/ \sum_{i \in \bar{I}} \bar{\tau}(i,\bar{f}(i))\bar{\pi}_i(f) \quad \text{for } f \in F$$

it follows that

(5.5.13) $\bar{g}(f) = g(f)$ for all $f \in F$.

Further, for any $f \in F$ define

$$\bar{w}_i(f) = \sum_{n=0}^{\infty} \sum_{j \in \bar{I}} \{\bar{c}(j,\bar{f}(j)) - \bar{g}(f)\bar{\tau}(j,\bar{f}(j))\}_{\infty}\bar{p}^n_{ij}(\bar{f}), \quad i \in \bar{I}$$

where the definition of ${}_{\infty}\bar{p}^n_{ij}(\bar{f})$ is analogous to that of ${}_{s}p^n_{ij}(f)$ in (5.5.3). Then, as above, the bounded function $\bar{w}_i(f)$, $i \in \bar{I}$ has for any $f \in F$ the property

(5.5.14) $\bar{w}_{\infty}(f) = 0.$

Since theorem 5.5.1 also applies to the modified model, we have for any $f \in F$ that $\{g = \bar{g}(f);\ v_i = \bar{w}_i(f),\ i \in \bar{I}\}$ is the unique bounded solution to

$$(5.5.15) \quad v_i = \bar{c}(i,\bar{f}(i)) - g\bar{\tau}(i,\bar{f}(i)) + \sum_{j \in \bar{I}} \bar{p}_{ij}(\bar{f}(i))v_j, \quad i \in \bar{I},$$

with the property that $v_{\infty} = 0$. Further, using (5.5.13) - (5.5.15), it is immediately verified that for any $f \in F$, $\bar{w}_{\infty}(f) = \bar{w}_s(f)$ and that $\{g = g(f);\ v_i = \bar{w}_i(f),\ i \in I\}$ is a bounded solution to (5.5.8) having the property that $v_s = 0$. By the parts (a) and (d) of theorem 5.5.1, it now follows that

(5.5.16) $\bar{w}_i(f) = w_i(f)$ for all $i \in I$ and $f \in F$.

Using the relations (5.5.14) and (5.5.16) it is now straightforward to verify that the following correspondence exists between any pair of sequences $\{f_n^{(\infty)}, n \geq 1\}$, with $f_n \in F$, and $\{h_n^{(\infty)}, n \geq 1\}$, with $h_n \in \bar{F}$, that are generated by the policy iteration method in the original and modified model respectively

(5.5.17) $h_n = \bar{f}_n$ for all $n \geq 1$ when $h_1 = \bar{f}_1$.

The above relationships will be used to prove the convergence results for the policy-iteration method. Before doing this, we give the following lemma.

LEMMA 5.5.4. *Let* $\{u_n,\ n \geq 1\}$ *be a bounded sequence of numbers such that for any* $\varepsilon > 0$ *there is an integer* $N(\varepsilon)$ *for which* $u_{n+m} \leq u_n + \varepsilon$ *for all* $n,m \geq N(\varepsilon)$. *Then the sequence* $\{u_n\}$ *is convergent.*

PROOF. Let $u = \liminf_{n\to\infty} u_n$ and let $U = \limsup_{n\to\infty} u_n$. Choose $\varepsilon > 0$. Then, $U \leq u_n + \varepsilon$ for all $n \geq N(\varepsilon)$, so, $U \leq u + \varepsilon$ which proves the lemma since ε was arbitrarily chosen. □

We now prove the convergence results for the policy-iteration method.

THEOREM 5.5.5. *Let* $\{f_n^{(\infty)}, n \geq 1\}$ *with* $f_n \in F$ *be any sequence of stationary policies generated by the policy-iteration method applied on the semi-Markov decision model considered. Let* $g^* = \inf_{f \in F} g(f)$. *Then*

$$\lim_{n \to \infty} g(f_n) = g^* \tag{5.5.18}$$

and, for some bounded function w_i^*, $i \in I$,

$$\lim_{n \to \infty} w_i(f_n) = w_i^* \quad \text{for all } i \in I. \tag{5.5.19}$$

Moreover, the bounded numbers $\{g^*; w_i^*, i \in I\}$ *satisfy the optimality equation*

$$w_i^* = \min_{a \in A(i)} \{c(i,a) - g^* \tau(i,a) + \sum_{j \in I} p_{ij}(a) w_j^*\} \quad \text{for all } i \in I. \tag{5.5.20}$$

PROOF. Suppose that we have already verified (5.5.18) and (5.5.19). Using the construction of f_n and the relations (5.5.8) and (5.5.9), we have for all $n \geq 2$

$$w_i(f_n) = c(i, f_n(i)) - g(f_n)\tau(i, f_n(i)) + \sum_{j \in I} p_{ij}(f_n(i)) w_j(f_n), \quad i \in I \tag{5.5.21}$$

and

$$c(i, f_n(i)) - g(f_{n-1})\tau(i, f_n(i)) + \sum p_{ij}(f_n(i)) w_j(f_{n-1}) =$$

$$= \min_{a \in A(i)} \{c(i,a) - g(f_{n-1})\tau(i,a) + \sum_{j \in I} p_{ij}(a) w_j(f_{n-1})\}, \quad i \in I. \tag{5.5.22}$$

Since I is denumerable and A(i) is a compact metric space for any $i \in I$, we can choose a $f^* \in F$ and an infinite sequence $\{n_k, k \geq 1\}$ such that

$$\lim_{k \to \infty} f_{n_k}(i) = f^*(i) \quad \text{for all } i \in I.$$

Now, taking $n = n_k$ in (5.5.21) and (5.5.22), letting $k \to \infty$ and using A1 together with the same arguments as in the proof of lemma 5.3.1 we easily get the result (5.5.20) where $f^*(i)$ minimizes the right-side of (5.5.20) for all $i \in I$. It remains to prove (5.5.18) and (5.5.19). We shall first prove these relations under the assumption

(5.5.23) the action set A(s) consists of a single action.

Next, using the modified model, we shall verify that (5.5.18) and (5.5.19) also hold without the assumption (5.5.23). Now suppose that (5.5.23) holds. Fix $n \geq 1$. By (5.5.23), we have $f_{n+1}(s) = f_n(s)$ and so, by (5.5.9) and part

(a) of theorem 5.5.1,

$$T(s,f_{n+1}(s),w(f_n)) = c(s,f_n(s)) - g(f_n)\tau(s,f_n(s)) + \sum_{j\in I} p_{sj}(f_n(s))w_j(f_n) = w_s(f_n).$$

Hence, by (5.5.7),

(5.5.24) $T(s,f_{n+1}(s),w(f_n)) = 0.$

Using the abbreviated notation

$$a_n(i) = c(i,f_{n+1}(i)) - g(f_n)\tau(i,f_{n+1}(i)) \quad \text{for } i \in I$$

we have, by (5.5.9)

(5.5.25) $T(i,f_{n+1}(i),w(f_n)) = a_n(i) + \sum_{j\in I} p_{ij}(f_{n+1})w_j(f_n)$ for $i \in I$.

By the construction of f_{n+1} and (5.5.10),

(5.5.26) $w_j(f_n) \geq T(j,f_{n+1}(j),w(f_n))$ for all $j \in I$.

Using (5.5.24) - (5.5.26) and (5.5.3), we have for any $i \in I$

$$T(i,f_{n+1}(i),w(f_n)) \geq a_n(i) + \sum_{j\neq s} p_{ij}(f_{n+1})T(j,f_{n+1}(j),w(f_n)) =$$

$$= a_n(i) + \sum_{j\in I} {}_s p^1_{ij}(f_{n+1})T(j,f_{n+1}(j),w(f_n)) =$$

$$= a_n(i) + \sum_{j\in I} {}_s p^1_{ij}(f_{n+1})a_n(j) + \sum_{j\in I} {}_s p^1_{ij}(f_{n+1}) \sum_{h\in I} p_{jh}(f_{n+1})w_h(f_n).$$

Continuing in this way, we find by induction on m that for any $m \geq 1$

$$T(i,f_{n+1}(i),w(f_n)) \geq \sum_{k=0}^{m} \sum_{j\in I} a_n(j)\, {}_s p^k_{ij}(f_{n+1}) +$$

$$+ \sum_{j\in I} {}_s p^m_{ij}(f_{n+1}) \sum_{h\in I} p_{jh}(f_{n+1})w_h(f_n), \quad i \in I.$$

We now observe that, by condition C6' and relation (5.5.4),

$$\lim_{m\to\infty} \sum_{j\in I} {}_s p^m_{ij}(f) = 0 \quad \text{for all } i \in I \text{ and } f \in F.$$

Using this result, (5.5.4) and the boundedness of the functions $a_n(i)$ and $w_i(f_n)$, $i \in I$, it now follows that

(5.5.27) $T(i,f_{n+1}(i),w(f_n)) \geq \sum_{k=0}^{\infty} \sum_{j\in I} a_n(j)\, {}_s p^k_{ij}(f_{n+1})$ for all $i \in I$.

Putting $\Delta_n = g(f_n) - g(f_{n+1})$, it follows from (5.5.6) and (5.5.27) that, for any $i \in I$,

$$w_i(f_{n+1}) - T(i, f_{n+1}(i), w(f_n)) \le \Delta_n \sum_{k=0}^{\infty} \sum_{j \in I} \tau(j, f_{n+1}(j))\, s p_{ij}^k(f_{n+1}),$$

where the various interchanges of the summations involved, are justified by the absolute convergence of these series. Next, using the boundedness of $\tau(i,a)$, relation (5.5.4) and condition C6', there is some finite number B such that

(5.5.28) $\quad w_i(f_{n+1}) - T(i, f_{n+1}(i), w(f_n)) \le \Delta_n B \quad$ for all $i \in I$ and $n \ge 1$.

Hence, by (5.5.26) and (5.5.28), $w_i(f_{n+1}) - w_i(f_n) \le \Delta_n B$ for all $i \in I$ and $n \ge 1$ which implies

(5.5.29) $\quad w_i(f_{n+m}) - w_i(f_n) \le \{g(f_n) - g(f_{n+m})\}B \quad$ for all $i \in I$ and $n, m \ge 1$.

Since the sequence $\{g(f_n),\ n \ge 1\}$ is bounded from below and non-increasing (see lemma 5.5.2 and (5.5.12)), it follows that $\lim_{n\to\infty} g(f_n)$ exists and is finite. Further, it is immediate from (5.5.6) and (5.5.4) that $w_i(f)$ is bounded in $f \in F$ for each $i \in I$. Now, using (5.5.29) and lemma 5.5.4, we obtain (5.5.19) for some bounded function w_i^*, $i \in I$. To prove (5.5.18), observe that, by (5.5.26),

$$0 \le w_i(f_n) - w_i(f_{n+1}) + w_i(f_{n+1}) - T(i, f_{n+1}(i), w(f_n)) \quad \text{for all } i \in I \text{ and } n \ge 1,$$

and so, by (5.5.19) and (5.5.28),

(5.5.30) $\quad \lim_{n\to\infty} \{w_i(f_n) - T(i, f_{n+1}(i), w(f_n))\} = 0 \quad$ for all $i \in I$.

Choose now $f \in F$. By the definition of f_{n+1} and (5.5.9), we have for all $i \in I$ and $n \ge 1$,

$$c(i, f(i)) - g(f_n)\tau(i, f(i)) + \sum_{j \in I} p_{ij}(f) w_j(f_n) \ge T(i, f_{n+1}(i), w(f_n)) - w_i(f_n) + w_i(f_n).$$

Multiply both sides of this inequality by $\pi_i(f)$ and sum over $i \in I$. After an interchange of the order of summation, justified by the boundedness of the functions involved, and using (5.5.1), we get

$$\sum_{i \in I} \{c(i, f(i)) - g(f_n)\tau(i, f(i))\}\pi_i(f) \ge \sum_{j \in I} \{T(i, f_{n+1}(i), w(f_n)) - w_i(f_n)\}\pi_i(f).$$

Next, letting $n \to \infty$ and using the bounded convergence theorem and the relations (5.5.30) and (5.5.5), we find $g(f) \geq \liminf_{n\to\infty} g(f_n)$ which implies (5.5.18) since $f \in F$ was arbitrarily chosen. We now have verified (5.5.18) and (5.5.19) under the assumption (5.5.23). Finally, using the modified model for which condition C6' with state ∞ instead of state s applies, and where $\bar{A}(\infty)$ consists of a single action, the above proof shows, using the relations (5.5.13), (5.5.14), (5.5.16) and (5.5.17), that (5.5.18) and (5.5.19) equally hold without the assumption (5.5.23). This completes the proof.

Part II. Stochastic games

CHAPTER 6

On N-person stochastic games with denumerable state space

6.1. INTRODUCTION

In the previous chapters, we considered Markov Decision and Renewal Problems in which a *single* decision maker controls the development of some Markovian system. However in many stochastic control problems arising in various applications such as the modelling of economic markets, the description of biological systems etc. (cf. SOBEL [117]), the system is simultaneously controlled by more than one decision maker. As a consequence, these problems have to be modelled using stochastic games, and the following three chapters will be devoted to the latter.

This chapter considers non-cooperative N-person stochastic games with a countable state space and compact metric action spaces. We concentrate upon the average return per unit time criterion for which both the existence of an equilibrium policy and solutions to the optimality equation are established, under a number of recurrence conditions with respect to the transition probability matrices associated with the stationary policies.

These results are obtained by showing that the average return criterion arises as a (first) sensitive discount optimality criterion. More specifically, we show that under each one of the aforementioned recurrency conditions, average return equilibrium policies appear as limit policies of sequences of total discounted return equilibrium policies where the discount factor tends to one. The first results on this topic are due to STERN [119].

Accordingly, after giving some preliminaries and notation in section 2, we first establish in section 3 the existence of a total discounted return equilibrium policy for each discount factor $\alpha \in [0,1)$ (an existing proof in [118] appears to be incorrect). Related work on the discounted model with infinite state space may be found for example in HIMMELBERG et al. [56], IDZIK [66] and WHITT [132].

In section 4, the existence of an average return equilibrium policy and a solution to the optimality equation are established, whereas in section 5, we review and extend the results that are known for the case where both the state space and the action spaces are finite. Finally, in section 6, we deal with the case of perfect information. The results in this chapter have been taken from FEDERGRUEN [30].

6.2. PRELIMINARIES AND NOTATION

This chapter treats a N-person non-cooperative stochastic game specified by the objects S, $A^i(s)$, q and r.

The set S is countable, and for each i = 1,...,N and $s \in S$, $A^i(s)$ is a compact metric space where the set S denotes the state space of some system and $A^i(s)$ denotes the set of actions, available to player i, in state s.

We define A as the union of all $A^i(s)$ ($s \in S$; i = 1,...,N) and C as

$$C = X_{i=1}^{N} A. \tag{6.2.1}$$

q associates with each pair $(s,\underline{a}) \in S \times C$ a probability distribution $q_{s.}(\underline{a})$ on the elements of S which is measurable in $\underline{a}$; and r^i is a bounded real-valued measurable function on $S \times C$, for all i = 1,...,N.

A stochastic game may be considered as a sequence $\gamma_1,\gamma_2,\ldots$ of non-cooperative games played by the N players, where $s \in S$ indexes the set $\{\Gamma_s \mid s \in S\}$ from which γ_t (t = 1,2,...) is drawn. Note that all the players' actions in $\gamma_t = s$ (t = 1,2,...; $s \in S$) constitute a vector $\underline{a} = [a^1,\ldots,a^N] \in$ $\in C(s)$ where

$$C(s) = X_{i=1}^{N} A^i(s), \quad s \in S. \tag{6.2.2}$$

When $\gamma_t = s$, i.e., when the system is in state s and the vector $\underline{a} \in C(s)$ denotes all the players' actions in γ_t, then the one-step expected reward to player i, is given by $r^i(s;\underline{a})$ and the system moves to state t with probability $q_{st}(\underline{a})$.

For each $s \in S$, and i = 1,...,N let $\Xi(A^i(s))$ denote the set of all signed measures on $\mathcal{B}_{A^i(s)}$, the Borel subsets of $A^i(s)$, endowed with the weak topology (cf. VARADARAJAN [126], p.16-17). This corresponds to weak convergence within the sets $\Xi(A^i(s))$. The sets belonging to the base by which this topology is defined satisfy the Hausdorff postulates for neighbourhoods, and are in addition locally convex (cf. p.205 in ROYDEN [100]).

As a consequence we obtain that $\Xi(A^i(s))$ is a linear Hausdorff locally convex topological space.

Let $M(A^i(s))$ be the subspace of all *probability* measures on $\mathcal{B}_{A^i(s)}$, with the induced topology. It then follows from th.3.4 in [126] that $M(A^i(s))$ can be metrized as a compact convex metric subspace of $\Xi(A^i(s))$, since $A^i(s)$ is a compact metric space.

Next we define for each $s \in S$, $\Xi(C(s)) = X_{i=1}^N \Xi(A^i(s))$, and $M(C(s)) = X_{i=1}^N M(A^i(s))$, $i = 1,\ldots,N$.

Note that $\Xi(C(s))$ is again a linear Hausdorff locally convex topological space, and that $M(C(s))$ is again a compact convex metrizable subspace of $\Xi(C(s))$, $s \in S$. Finally, we observe that $M(C(s))$ can be identified as the space of all product probability measures on $\mathcal{B}_{C(s)}$, the product σ-field in $C(s)$. Moreover, for any sequence $\{\underline{\mu}_n\}_{n=1}^{\infty}$ with $\underline{\mu}_n \in M(C(s))$, $n = 1,2,\ldots$, it follows from th.3.2 in BILLINGSLEY [9] that

$$(6.2.3)\qquad \int_{C(s)} v(\underline{a})\, d\,\underline{\mu}_n(\underline{a}) \to \int_{C(s)} v(\underline{a})\, d\,\underline{\mu}(\underline{a}), \quad \text{as } n \to \infty$$

for all real-valued and continuous functions $v(.)$ on $C(s)$

if and only if $\underline{\mu}_n \to \underline{\mu}$ (in the product topology).

We note that any continuous function on the compact metric space $C(s)$ is bounded. We use the (abbreviated) notation $[\underline{\mu}^{-i},\nu]$ for the N-person randomized action $[\mu^1,\ldots,\mu^{i-1},\nu,\mu^{i+1},\ldots,\mu^N]$ that results from $\underline{\mu} = [\mu^1,\ldots,\mu^N]$ when the i-th player changes from μ^i to ν, the other players continuing to use their respective actions in $\underline{\mu}$. Defining $r^i(s;\underline{\mu}) = E_{\underline{\mu}} r^i(s;\underline{a})$ and $q_{st}(\underline{\mu}) = E_{\underline{\mu}} q_{st}(\underline{a})$ for all $\underline{\mu} = [\mu^1,\ldots,\mu^N] \in M(C(s))$, $s \in S$, $i = 1,\ldots,N$, we obtain

$$(6.2.4)\qquad r^i(s;\underline{\mu}) = \int_{C(s)} r^i(s;\underline{a})\, d\underline{\mu}(\underline{a}) =$$

$$= \int_{A^1(s)} \cdots \int_{A^N(s)} r^i(s;a^1,\ldots,a^N)\, d\mu^1(a^1)\ldots d\mu^N(a^N)$$

$$(6.2.5)\qquad q_{st}(\underline{\mu}) = \int_{C(s)} q_{st}(\underline{a})\, d\underline{\mu}(\underline{a}) =$$

$$= \int_{A^1(s)} \cdots \int_{A^N(s)} q_{st}(a^1,\ldots,a^N)\, d\mu^1(a^1)\ldots d\mu^N(a^N)$$

where the second equality in (6.2.4) and (6.2.5) follows from Fubini's theorem. Observe that $r^i(s;\underline{\mu})$ and $q_{st}(\underline{\mu})$ are both multilinear in $\underline{\mu}$, i.e.,

for all $\lambda \in [0,1]$:

(6.2.6) $r^i(s;\mu^1,\ldots,\lambda\mu^j+(1-\lambda)\nu^j,\ldots,\mu^N) = \lambda r^i(s;\mu^1,\ldots,\mu^j,\ldots,\mu^N) +$

$+ (1-\lambda)\, r^i(s;\mu^1,\ldots,\nu^j,\ldots,\mu^N)$

(6.2.7) $q_{st}(\mu^1,\ldots,\lambda\mu^j+(1-\lambda)\nu^j,\ldots,\mu^N) = \lambda q_{st}(\mu^1,\ldots,\mu^j,\ldots,\mu^N) +$

$+ (1-\lambda)\, q_{st}(\mu^1,\ldots,\nu^j,\ldots,\mu^N)$.

Hereafter we assume that for each $s \in S$,

(6.2.8) $r^i(s;\underline{a})$ and $q_{st}(\underline{a})$ are continuous on $C(s)$, for all $i = 1,\ldots,N$ and $t \in S$.

Observe from (6.2.3) that (6.2.8) implies that, for each $s \in S$, the one-step expected rewards and transition probabilities are continuous on the space of all *randomized* N players' actions $M(C(s))$ as well:

(6.2.9) $r^i(s;\underline{\mu})$ and $q_{st}(\underline{\mu})$ are continuous on $M(C(s))$ for all $i = 1,\ldots,N$ and $t \in S$.

Let $F^i = X_{s\in S} M(A^i(s))$ be the set of all *decision rules for player* i, $(i = 1,\ldots,N)$, i.e., of all functions f^i mapping each state s into an action $f^i(s) \in M(A^i(s))$. A policy for a player i is a sequence $\pi^i = (f^{i(1)}, f^{i(2)}, \ldots)$ of decision rules. Using policy π^i means that $f^{i(n)}$ is employed at time n; thus if the system is observed in state s at time n, then player i chooses action $f^{i(n)}(s)$, the s-th component of $f^{i(n)}$. We write $f^{i(\infty)}$ for the *stationary policy* $(f^i, f^i, \ldots)$ for player i. As a consequence we let F^i represent the class of all stationary policies for player i as well.

A stationary policy $f^{i(\infty)} \in F^i$ is said to be *pure* if in each state $s \in S$ it prescribes a specific action in $A^i(s)$ with probability one. Finally, the set of all policies for player i is denoted by Π^i, and $\underline{\Pi} = X_{i=1}^N \Pi^i$ represents the class of all N players' policies, with $\underline{F} = X_{i=1}^N F^i$ the subset of the *stationary* N players' policies. We associate with each stationary policy $\underline{f}^{(\infty)} \in \underline{F}$, the transition probability matrix $P(\underline{f})$ (tpm), i.e.,

$$P(\underline{f})_{st} = q_{st}(\underline{f}(s))$$

with the n-th power $P^n(\underline{f})$ indicating the matrix of n-step transition probabilities, i.e. $P^n(\underline{f}) = P(\underline{f})\, P^{n-1}(\underline{f})$, $n \geq 2$.

For any policy $\underline{\pi} = [\pi^1,\ldots,\pi^N] \in \underline{\Pi}$ we define $V_\alpha^i(\underline{\pi};s)$ and $g^i(\underline{\pi};s)$ as the total expected α-discounted return, and the long-run average return per unit time to player i, when the initial state is s:

$$(6.2.10)\quad V_\alpha^i(\underline{\pi};s) = E_{\underline{\pi}}\{\sum_{k=0}^{\infty}\alpha^k r^i(s_k;\underline{a}_k) \mid s_0=s\};\ i = 1,\ldots,N;\ s\in S;\ 0 \le \alpha < 1$$

$$(6.2.11)\quad g^i(\underline{\pi};s) = \limsup_{t\to\infty} \frac{1}{t+1} E_{\underline{\pi}}\{\sum_{k=0}^{t} r^i(s_k;\underline{a}_k) \mid s_0=s\};\ i = 1,\ldots,N;\ s \in S$$

where $E_{\underline{\pi}}$ indicates the expectation given the players' common policy $\underline{\pi} \in \underline{\Pi}$ is used and where $\{s_k;k=0,1,2,\ldots\}$ and $\{\underline{a}_k;k=0,1,\ldots\}$ denote the stochastic processes of the states and actions that result from policy $\underline{\pi}$.

A N-tuple of policies $\underline{\pi}^* = [\pi^{*1},\ldots,\pi^{*N}] \in \underline{\Pi}$ is said to be an *α-discounted equilibrium point of policies* (α-DEP) if, *simultaneously for every initial state of the system* s,

$$(6.2.12)\quad V_\alpha^i(\underline{\pi}^*;s) \ge V_\alpha^i(\underline{\pi};s) \text{ for all } i = 1,\ldots,N \text{ and } \underline{\pi} \in \underline{\Pi}^{-i}(\underline{\pi}^*):$$

where

$$(6.2.13)\quad \underline{\Pi}^{-i}(\underline{\pi}^*) = \{\underline{\pi} = [\pi^1,\ldots,\pi^N] \in \underline{\Pi} \mid \pi^j = \pi^{*j},\ j \ne i\}.$$

Similarly we define $\underline{\pi}^*$ as an *average return equilibrium point of policies (AEP)*, if *simultaneously for every initial state* s,

$$(6.2.14)\quad g^i(\underline{\pi}^*;s) \ge g^i(\underline{\pi};s) \text{ for all } i = 1,\ldots,N \text{ and } \underline{\pi} \in \underline{\Pi}^{-i}(\underline{\pi}^*).$$

Hence, whenever the players choose an α-DEP (AEP) $\underline{\pi}^*$, none of them, whatever the initial state of the system, can increase his own total expected α-discounted return (expected average return per unit time) by changing to some other policy $\pi^i \ne \pi^{*i} \in \Pi^i$, the other players continuing to use their respective policies in $\underline{\pi}^*$.

Note that we do not consider *history-dependent* policies, i.e. policies which prescribe for each time t, a randomized action in dependence on the entire history $H_t = (s_0,\underline{a}_0;s_1,\ldots,s_{t-1},\underline{a}_{t-1},s_t)$ of the system up to time t, rather than in dependence on the current state s_t alone. The justification for our confining ourselves to the class $\underline{\Pi}$ is provided by [62], who showed as an adaptation of the corresponding result in DERMAN & STRAUCH [26] that whenever a policy $\underline{\pi}^*$ is an α-DEP or AEP within $\underline{\Pi}$, it is an equilibrium policy within the broader class of history-dependent policies as well.

We conclude this section by observing that if the sets $A^i(s)$ (i=1,..., N; s$\in$S) are convex compact subsets of some linear metric space themselves,

such that for all $i = 1,\ldots,N$ $r^i(s;\underline{a})$ is linear or even concave in the i-th component of $\underline{a}$ (cf. (6.2.6) and (6.2.7)) then the existence of a *pure* instead of a randomized stationary α-DEP or AEP is guaranteed under the same conditions, as follows from an examination of the analysis below.

6.3. EXISTENCE OF STATIONARY α-DEP'S

In this section we prove the existence of a stationary α-DEP for each $\alpha \in [0,1)$. For each policy $\underline{f}^{(\infty)} \in \underline{F}$, the total expected α-discounted return to player i, when starting in state $s \in S$, is denoted by

$$(6.3.1) \qquad V_\alpha^i(\underline{f}^{(\infty)};s) = \sum_{n=0}^{\infty} \alpha^n \sum_{t\in S} P^n(\underline{f})_{st}\, r^i(t;\underline{f}(t)).$$

The following lemma proves that $V_\alpha^i(\underline{f}^{(\infty)};s)$ is a continuous function on $\underline{F}$ for all $i = 1,\ldots,N$; $s \in S$ and $\alpha \in [0,1)$:

LEMMA 6.3.1. *Fix* $s \in S$, $1 \le i \le N$ *and* $\alpha \in [0,1)$. *Then* $V_\alpha^i(\underline{f}^{(\infty)};s)$ *is continuous on* $\underline{F}$.

PROOF. We first observe that since $\underline{F}$ is metrizable, it suffices to show that $\lim_{n\to\infty} V_\alpha^i(\underline{f}_n^{(\infty)};s) = V_\alpha^i(\underline{f}^{(\infty)};s)$ whenever $\{\underline{f}_n\}_{n=1}^{\infty} \to \underline{f}$, with $\underline{f}_n, \underline{f} \in \underline{F}$. Let M be such that

$$(6.3.2) \qquad |r^i(s;\underline{a})| \le M \quad \text{for all } s \in S, \text{ and } \underline{a} \in C(s).$$

It is then easily verified that

$$(6.3.3) \qquad |V_\alpha^i(\underline{h}^{(\infty)};s)| \le M/(1-\alpha) \quad \text{for all } \underline{h}^{(\infty)} \in \underline{F} \text{ and } s \in S.$$

Next, observe by complete induction that as a consequence of (6.2.8) and (6.2.3) $P^k(\underline{f})_{st}$ is continuous on $\underline{F}$ for all $s,t \in S$ and $k = 1,2,\ldots$. This, in turn, implies using proposition 18 on p.232 in ROYDEN [100] that for each $\ell = 0,1,\ldots$

$$(6.3.4) \qquad \lim_{n\to\infty} \textstyle\sum_t P^\ell(\underline{f}_n)_{st}\, r^i(t;\underline{f}_n(t)) = \sum_t P^\ell(\underline{f})_{st}\, r^i(t;\underline{f}(t)).$$

Finally, pick $\varepsilon > 0$ and choose K such that $\alpha^K \le \varepsilon(1-\alpha)/4M$. Let $H_{\underline{h}}^k(s) = \sum_{\ell=0}^{k-1} \alpha^\ell \sum_{t\in S} P^\ell(\underline{h})_{st}\, r^i(t;\underline{h}(t))$ for all $k = 1,2,\ldots$ and $\underline{h} \in \underline{F}$. Observe that for each $\underline{h} \in \underline{F}$:

$$(6.3.5) \qquad V_\alpha^i(\underline{h}^{(\infty)};s) = H_{\underline{h}}^K(s) + \alpha^K \textstyle\sum_{t\in S} P^K(\underline{h})_{st}\, V_\alpha^i(\underline{h}^{(\infty)};t).$$

In view of (6.3.4) there exists an integer N_0 such that $|H^K_{\underline{f}_n}(s) - H^K_{\underline{f}}(s)| \leq \varepsilon/2$, for all $n \geq N_0$. We thus obtain that for all $n \geq N_0$:

$$|V^i_\alpha(\underline{f}_n^{(\infty)};s) - V^i_\alpha(\underline{f}^{(\infty)};s)| \leq |H^K_{\underline{f}_n}(s) - H^K_{\underline{f}}(s)| +$$

$$+ \alpha^K \Big| \sum_{t\in S} P^K(\underline{f}_n)_{st} V^i_\alpha(\underline{f}_n^{(\infty)};t) - \sum_{t\in S} P^K(\underline{f})_{st} \cdot V^i_\alpha(\underline{f}^{(\infty)};t) \Big| \leq$$

$$\varepsilon/2 + \frac{\varepsilon(1-\alpha)}{4M} \frac{2M}{(1-\alpha)} = \varepsilon. \quad \square$$

We now turn to the existence of an α-DEP.

For a compact, metric state space and under somewhat stronger continuity assumptions with respect to the one-step expected rewards, and transition probability functions, the issue of the existence of an α-DEP was first dealt with by SOBEL [118]. Unfortunately there seem to be a number of serious errors which invalidate the approach. Although with a considerable amount of additional work, the proof in [118] can be rectified for the case of a denumerable state space, we prefer to give a different proof.

The extension of theorem 6.3.4 to a more general state space remains an outstanding problem. Difficulties arise e.g. in view of $\underline{F}$ becoming non-metrizable when the state space is over-countable, and as a consequence more general and less tractable convergence concepts are required. For the most recent development on this topic, we refer to WHITT [132].

Our approach uses an extension of the Kakutani fixed point theorem which was obtained independently by GLICKSBERG [48] and FAN [29]. First, for each compact set U, let 2^U denote the class of all (non-empty) closed subsets of U. A point to set mapping Φ: $U \to 2^U$ (with U satisfying the first countability axiom) is said to be upper semi-continuous, if for each sequence $\{x_n\}_{n=1}^\infty$, $x_n \in U$:

$$(6.3.6) \quad \{x_n\}_{n=1}^\infty \to x,\ y_n \in \Phi(x_n),\ \{y_n\}_{n=1}^\infty \to y \Rightarrow y \in \Phi(x).$$

LEMMA 6.3.2. *Given an upper semi-continuous point to convex set mapping* Φ: $U \to 2^U$, *defined on a convex compact subset* U *of a linear Hausdorff locally convex topological space, there exists a point* $x \in \Phi(x)$. $\square$

Observe from the analysis in section 1, that $X_{s\in S} \Xi(C(s))$, the space of all functions f mapping each state s into a N-tuple of (finite, signed) measures $f(s) \in \Xi(C(s))$, endowed with the product topology, is again a linear Hausdorff locally convex topological space, with $\underline{F}$, the countable topological

product of the spaces $M(C(s))$ $(s \in S)$, a metrizable subspace which is in addition convex and compact, as a consequence of Tychonoff's theorem. The fixed point theorem in lemma 6.3.2 will be applied by constructing a point to set mapping on $\underline{F}$, as a subspace of $X_{s\in S} \Xi(C(s))$.

We finally need the following lemma, the proof of which follows from th. 6-f in BLACKWELL [11]:

LEMMA 6.3.3. *Fix* $0 \leq \alpha < 1$. *A stationary policy* $\underline{f}^{(\infty)} = [f^{1(\infty)},\ldots,f^{N(\infty)}]$ *is an* α-DEP, *iff* $V_\alpha^i(\underline{f}^{(\infty)};s)$ *satisfies the optimality equation:*

$$(6.3.7)\quad V_\alpha^i(\underline{f}^{(\infty)};s) = \max_{\mu\in M(A^i(s))} \{r^i(s;[\underline{f}^{-i}(s),\mu]) + \alpha \sum_{t\in S} q_{st}([\underline{f}^{-i}(s),\mu])\, V_\alpha^i(\underline{f}^{(\infty)};t)\}$$

for all $s \in S$, $i = 1,\ldots,N$.

THEOREM 6.3.4. *There exists a stationary* α-DEP *for each* $\alpha \in [0,1)$.

PROOF. We first observe that for each $\underline{f} \in \underline{F}$ and $i = 1,\ldots,N$ there exists, as a result of (6.2.8) a $h \in F^i$ such that for all $s \in S$:

$$(6.3.8)\quad r^i(s;[\underline{f}^{-i}(s),h(s)]) + \alpha \sum_{t\in S} q_{st}([\underline{f}^{-i}(s),h(s)])\, V_\alpha^i(\underline{f}^{(\infty)};t) = \max_{\mu\in M(A^i(s))} \{r^i(s;[\underline{f}^{-i}(s),\mu]) + \alpha \sum_{t\in S} q_{st}([\underline{f}^{-i}(s),\mu]) V_\alpha^i(\underline{f}^{(\infty)};t)\}.$$

For any $i = 1,\ldots,N$ and $\underline{f} \in \underline{F}$, let $\Phi^i(\underline{f})$ denote the set of all $h \in F^i$ that satisfy (6.3.8) for all $s \in S$, and define the point-to-convex set mapping

$$\Phi\colon \underline{F} \to 2^{\underline{F}}\colon \underline{f} \to \Phi(\underline{f}) = X_{i=1}^N \Phi^i(\underline{f}).$$

We next show the upper semi-continuity of this point-to-set mapping. Fix $\{\underline{f}_n\}_{n=1}^\infty$, $\{\underline{h}_n\}_{n=1}^\infty$ with (1) $\underline{f}_n$, $\underline{h}_n \in \underline{F}$, (2) $\lim_{n\to\infty} \underline{f}_n = \underline{f}$; $\lim_{n\to\infty} \underline{h}_n = \underline{h}$ and (3) $\underline{f}_n \in \Phi(\underline{h}_n)$.

Substitute $\underline{f}_n$ for $\underline{f}$ and h_n^i for h in (6.3.8) and let n tend to infinity. It then follows that h^i satisfies (6.3.8) for $\underline{f}$, and this for all $i = 1,\ldots N$ and $s \in S$, as a consequence of (6.2.8), lemma 6.3.2, the boundedness of $V_\alpha^i(\underline{f}_n^{(\infty)};s)$ and proposition 18 on p.232 in ROYDEN [100].

As a consequence of the upper semi-continuity of Φ, and the fact that Φ is a point-to-convex set mapping of a convex compact subset $\underline{F}$ of the linear Hausdorff locally convex topological space $X_{s\in S} \Xi(C(s))$ into itself, it

follows from lemma 6.3.2 that there exists a $\underline{f}^* \in \underline{F}$ such that $\underline{f}^* \in \Phi(\underline{f}^*)$ which implies (6.3.7) and hence proves the theorem (cf. lemma 6.3.3). □

6.4. THE EXISTENCE OF AVERAGE RETURN EQULIBRIUM POLICIES (AEP's)

In this section, we will show that the existence of stationary average return equilibrium policies (AEP's) is guaranteed when either one of the recurrence conditions C1 - C9, in combination with assumption A4, as introduced in chapter 5, is imposed on $\mathcal{P} = \{P(\underline{f}) | \underline{f} \in \underline{F}\}$, the set of all transition probability matrices, associated with the stationary policies. For ease of reference, we first restate these conditions, in the context of this stochastic games model. For any $s \in S$, $A \subseteq S$ and $\underline{f} \in \underline{F}$, define the possibly infinite recurrence time $\mu_{sA}(\underline{f})$ as in chapter 5 (cf. (5.2.3)) i.e. $\mu_{sA}(\underline{f})$ represents the expected number of transitions until the first visit to the set A, when starting in state s and when the N players use the policy $\underline{f}^{(\infty)} \in \underline{F}$.

We convene, once again, to write $\mu_{sA}(\underline{f}) = \mu_{st}(\underline{f})$ for $A = \{t\}$:

C1. There is a finite set $K \subseteq S$ and a finite number B such that

$$\mu_{sK}(\underline{f}) \leq B \quad \text{for all } s \in S \text{ and } \underline{f} \in \underline{F}$$

C2. There is a finite set K, an integer $\nu \geq 1$ and a number $\rho > 0$ such that

$$\sum_{t \in K} P^{\nu}(\underline{f})_{st} \geq \rho \quad \text{for all } s \in S \text{ and } \underline{f} \in \underline{F}$$

C3. There is an integer $\nu \geq 1$ and a number $\rho > 0$ such that

$$\inf_{s_1, s_2 \in S} \{\sum_{t \in S} \min[P^{\nu}(\underline{f})_{s_1 t}; P^{\nu}(\underline{f})_{s_2 t}]\} \geq \rho \quad \text{for all } \underline{f} \in \underline{F} \tag{6.4.1}$$

C4. There is an integer $\nu \geq 1$ and a number $\rho > 0$ such that for any $\underline{f} \in \underline{F}$ a probability distribution $\{\pi_t(\underline{f}), t \in S\}$ (say) exists for which

$$\left| \sum_{t \in A} P^{n}(\underline{f})_{st} - \sum_{t \in A} \pi_t(\underline{f}) \right| \leq (1-\rho)^{\lfloor n/\nu \rfloor} \quad \text{for all } s \in S,\ A \subseteq S \text{ and } n \geq 1, \tag{6.4.2}$$

where $\lfloor x \rfloor$ denotes the largest integer less than or equal to x.

C5. For any $\underline{f} \in \underline{F}$ there is a probability distribution $\{\pi_t(\underline{f}), t \in S\}$ such that

$$P^{n}(\underline{f})_{st} \to \pi_t(\underline{f}) \text{ uniformly in } (s, \underline{f}) \in S \times \underline{F} \text{ as } n \to \infty, \text{ for any } t \in S \tag{6.4.3}$$

C6. There is a finite number B such that for any $\underline{f} \in \underline{F}$ a state $s^*_{\underline{f}}$ exists for which

$$\mu_{ss^*_{\underline{f}}}(\underline{f}) \leq B \text{ for all } s \in S$$

C7. There is a finite set K and a finite number B, such that for any $\underline{f} \in \underline{F}$ a state $s^*_{\underline{f}} \in K$ exists for which

$$\mu_{ss^*_{\underline{f}}}(\underline{f}) \leq B \text{ for all } s \in S$$

C8. There is an integer $\nu \geq 1$ and a number $\rho > 0$ such that for any $\underline{f} \in \underline{F}$ a state $s^*_{\underline{f}}$ exists for which

$$P^{\nu}(\underline{f})_{ss^*_{\underline{f}}} \geq \rho \text{ for all } s \in S$$

C9. There is a finite set K, an integer $\nu \geq 1$ and a number $\rho > 0$ such that for any $\underline{f} \in \underline{F}$ a state $s^*_{\underline{f}} \in K$ exists for which

$$P^{\nu}(\underline{f})_{ss^*_{\underline{f}}} \geq \rho \text{ for all } s \in S$$

Moreover, the following assumption is made throughout this section (cf. A4 in chapter 5).

A. For any $\underline{f} \in \underline{F}$, the stochastic matrix $P(\underline{f})$ *has no two disjoint* closed sets.

which is automatically implied by conditions C3 - C9. We refer to chapter 5, section 2, for a detailed investigation of the various ways in which some of these simultaneous recurrence conditions generalize well-known conditions from Markov chain theory, as well as for an analysis of the various relationships that exist among these conditions. We merely note that the special case of C3 with $\nu = 1$, can be formulated in a simpler way:

LEMMA 6.4.1. C3 *with* $\nu = 1 \iff$ *there is a number* $\rho > 0$ *such that for each four elements* $(s_1, s_2, \underline{a}_1, \underline{a}_2)$ *with* $s_1 \neq s_2$ *and* $\underline{a}_1 \in C(s_1)$, $\underline{a}_2 \in C(s_2)$:

$$\sum_{t \in S} \min\{q_{s_1 t}(\underline{a}_1), q_{s_2 t}(\underline{a}_2)\} \geq \rho. \tag{6.4.4}$$

PROOF. Fix $s_1, s_2 \in S$ and $\underline{\mu}_1 \in M(C(s_1))$, $\underline{\mu}_2 \in M(C(s_2))$ and observe that, as a consequence of (6.4.4):

$$\rho \leq E_{\underline{\mu}_1, \underline{\mu}_2}\Big[\sum_{t \in S} \min\{q_{s_1 t}(\underline{a}_1), q_{s_2 t}(\underline{a}_2)\}\Big] =$$

$$= \sum_{t \in S} E_{\underline{\mu}_1, \underline{\mu}_2} \min\{q_{s_1 t}(\underline{a}_1), q_{s_2 t}(\underline{a}_2)\} \leq \sum_{t \in S} \min\{q_{s_1 t}(\underline{\mu}_1), q_{s_2 t}(\underline{\mu}_2)\}, \tag{6.4.5}$$

where the interchange of expectation and summation is justified by the non-negativity of $\min\{q_{s_1t}(\underline{a}_1), q_{s_2t}(\underline{a}_2)\}$, and where the inequality part follows from Jensen's inequality and the concaveness of $\min(.,.)$ on R^2. Note finally that (6.4.5) coincides with the special case of (6.4.2) where $\nu = 1$. □

Moreover, we point out that in STERN [119], the existence of a stationary AEP was recently proven under the following special case of C6 (cf. C6' in chapter 5):

C6'. There exists a state s^* and a number B, such that the mean recurrence times $\mu_{ss^*}(\underline{f}) \leq B$ for all $s \in S$, $\underline{f} \in \underline{F}$.

In addition to the conditions C1 - C9, we introduce the recurrence condition D (cf. HORDIJK [58], p.106) which is of an entirely different type, and under which the existence of a stationary AEP will be shown to hold equally well:

D: There exists a number R such that for each player $i = 1,\ldots,N$ and for any combination of stationary policies $\{f^1,\ldots,f^{i-1},f^{i+1},\ldots,f^N\}$ of the other players, there is a policy $f^i \in F^i$ for player i, for which the mean recurrence time $\mu_{st}(\underline{f})$ from any state s to any state t, under policy $\underline{f} = [f^1,\ldots,f^N]$ is bounded by R, i.e. for each $\{f^1,\ldots,f^{i-1}, f^{i+1},\ldots,f^N\}$ with $f^j \in F^j$ for all $j \neq i$, there exists a $f^i \in F^i$ such that

(6.4.6) $\mu_{st}(\underline{f}) \leq R$ for all $s,t \in S$, with $\underline{f} = [f^1,\ldots,f^N]$.

Theorem 6.4.2 below gives the main result of this chapter.

THEOREM 6.4.2. *A stationary* AEP *exists under each one of the conditions* C1 - C9, *or condition* D.

PROOF. Once again, the possible periodicity of some of the transition probability matrices in $P = \{P(\underline{f}) \mid \underline{f} \in \underline{F}\}$ may cause unnecessary problems in establishing the existence proof. We therefore employ, once again, a generalization of the data-transformation (1.8.1) and (1.8.2) with $\sigma = 1$, so as to transform our model into an "equivalent" one in which all of the tpm's are aperiodic. Consider the transformed N-person stochastic game which has the same state and action spaces for all players and the same reward structures. Only the transition probability functions alter as follows:

(6.4.7) $\tilde{q}_{st}(\underline{a}) = \tau[q_{st}(\underline{a}) - \delta_{st}] + \delta_{st}$; $s,t \in S$; $\underline{a} \in C(s)$

where τ is a fixed number such that

$$0 < \tau < \tau_0 = \inf_{s,\underline{a}} \{1/(1-q_{ss}(\underline{a})) \mid q_{ss}(\underline{a}) < 1\}$$

and where δ_{st} represents once again, the Kronecker function. Note that (6.2.8) and assumption A hold in the transformed model as well, and that for all $s \in S$ and $\underline{a} \in C(s)$ we have that $\{\tilde{q}_{st}(\underline{a}), t \in S\}$ is a probability distribution, with

(6.4.8) $\tilde{q}_{ss}(\underline{a}) \geq 1 - \tau/\tau_0 > 0$ and $\tilde{q}_{st}(\underline{a}) \geq \tau\, q_{st}(\underline{a})$ for $s \neq t$.

Moreover both (6.4.7) and (6.4.8) remain true when replacing the pure action vectors $\underline{a} \in C(s)$ by the randomized action vectors $\underline{\mu} \in M(s)$, $s \in S$. Due to the first part of (6.4.8) we have that for any $\underline{f} \in \underline{F}$ the t.p.m. $\tilde{P}(\underline{f})$ is *aperiodic*. In addition, one immediately verifies

(6.4.9) $\tilde{P}^n(\underline{f})_{st} \geq (\tau^*)^n\, P(\underline{f})_{st}$, for all $s,t \in S$; $n \geq 1$ and $\underline{f} \in \underline{F}$

where $\tau^* = \min[1 - \tau/\tau_0; \tau]$.

Next, let $\tilde{V}_\alpha^i(\underline{f}^{(\infty)};s)$ denote the total expected α-discounted reward, for player i, in the transformed model, when starting in state s, and when the N players use policy $\underline{f}^{(\infty)} \in \underline{F}$. For each α $(0 \leq \alpha < 1)$ we choose a specific α-DEP $\underline{f}_\alpha \in \underline{F}$ (with respect to the transformed model). Finally, we fix a state $s^* \in S$ and define:

(6.4.10) $\tilde{v}_\alpha^i(s) = \tilde{V}_\alpha^i(\underline{f}_\alpha^{(\infty)};s) - \tilde{V}_\alpha^i(\underline{f}_\alpha^{(\infty)};s^*)$; $i = 1,\ldots,N$; $s \in S$.

We next need the following result, the proof of which will be deferred to the end of the proof of the theorem:

(6.4.11) *Under either one of the conditions* C1-C9 *or* D *(applying to the original model), the family of functions* $\{\tilde{v}_\alpha^i(.) \mid 0 \leq \alpha < 1\}$ *is uniformly bounded for all* $i = 1,\ldots,N$.

Next, observe that $|(1-\alpha)\, \tilde{V}_\alpha^i(\underline{f}_\alpha^{(\infty)};s^*)| \leq M$ for all $\alpha \in [0,1)$ and $i = 1,\ldots,N$. This, together with (6.4.11) and the fact that for all $s \in S$, any sequence of points in the compact metric space $M(C(s))$ has a convergent subsequence, imply, using the diagonalization procedure, the existence of N constants g^i, N bounded functions $\tilde{v}^i(.)$, a policy $\underline{f}^{(\infty)} \in \underline{F}$ and a sequence $\{\alpha_k\}_{k=1}^\infty$, with $\alpha_k \in [0,1)$ and $\lim_{k\to\infty} \alpha_k = 1$, such that:

(a) $\lim_{k\to\infty} \underline{f}_{\alpha_k} = \underline{f}$

(b) $\lim_{k\to\infty} (1-\alpha_k)\, \tilde{V}^i_{\alpha_k}(\underline{f}^{(\infty)}_{\alpha_k};s^*) = g^i;\ i = 1,\ldots,N$

(c) $\lim_{k\to\infty} \tilde{v}^i_{\alpha_k}(s) = \tilde{v}^i(s)$, for all $s \in S$, $i = 1,\ldots,N$.

Now, fix $i \in \{1,\ldots,N\}$ and $s = s_o \in S$ and subtract $\tilde{V}^i_{\alpha_k}(\underline{f}^{(\infty)}_{\alpha_k};s^*)$ from both sides of (6.3.7) with $\alpha = \alpha_k$, and $s = s_o$, in order to obtain (cf. (6.4.10)):

$$\tilde{v}^i_{\alpha_k}(s_o) = \max_{\mu\in M(A^i(s_o))}\{r^i(s_o;[\underline{f}^{-i}_{\alpha_k}(s_o),\mu]) - (1-\alpha_k)\tilde{V}^i_{\alpha_k}(\underline{f}^{(\infty)}_{\alpha_k};s^*) + \textstyle\sum_{t\in S} \tilde{q}_{s_o t}([\underline{f}^{-i}_{\alpha_k}(s_o),\mu])\tilde{v}^i_{\alpha_k}(t)\} \tag{6.4.12}$$

where $f^i_{\alpha_k}(s_o)$ attains the maximum on the right side of (6.4.12). Letting k tend to infinity in (6.4.12) we obtain for all $s \in S$:

$$g^i+\tilde{v}^i(s) = \max_{\mu\in M(A^i(s))}\{r^i(s;[\underline{f}^{-i}(s),\mu]) + \textstyle\sum_{t\in S} \tilde{q}_{st}([\underline{f}^{-i}(s),\mu])\tilde{v}^i(t)\} \tag{6.4.13}$$

with $f^i(s)$ attaining the maximum on the right hand side of (6.4.13); all of this as a consequence of (a),(b) and (c), (6.3.1) and proposition 18 on p.232 in ROYDEN [100]. Next, using (6.4.7), one verifies that the functions $v^i(.) = \tau\tilde{v}^i(.)$ satisfy (6.4.13) with $\tilde{q}_{st}(.)$ replaced by $q_{st}(.)$, i.e. for all $s \in S$ and $i = 1,\ldots,N$:

$$g^i+v^i(s) = \max_{\mu\in M(A^i(s))}\{r^i(s;[\underline{f}^{-i}(s),\mu]) + \textstyle\sum_{t\in S} q_{st}([\underline{f}^{-i}(s),\mu])v^i(t)\}. \tag{6.4.14}$$

Next, it follows from th.6.17 in ROSS [98] that policy $\underline{f}^{(\infty)}$ is an AEP and that $g^i(\underline{f}^{(\infty)};s) = g^i$ for all $s \in S$ and $i = 1,\ldots,N$.

This leaves us with the proof of (6.4.11). Under condition D, (6.4.11) is immediate from theorem 12.8 in HORDIJK [58]. Next, using the combination of th.5.2.2, th.5.2.4 part (ii) and th.5.2.5 part (iii) it follows that if either one of the conditions C1 - C9 applies to the original model, then C2 applies to this model as well. In view of (6.4.9), C2 then applies to the transformed model, in which C2 $\Rightarrow$ C4, in view of the introduced aperiodicity and th.5.2.5 part (iv). It then follows from (6.4.2), that for *any* $\underline{f} \in \underline{F}$, $s \in S$ and $n \geq 1$, and *some* integer $\nu \geq 1$ and positive number $\rho < 1$, the

total variation of the signed measure $\sigma(A) = \sum_{t\in A} \tilde{P}^n(\underline{f})_{st} - \sum_{t\in A} \tilde{P}^n(\underline{f})_{s^*t}$ is bounded by $4\ (1-\rho)^{\lfloor n/\nu \rfloor}$. Finally it follows from (6.3.1) that, for any $0 \leq \alpha < 1$ and $i = 1,\ldots,N$ and all $\underline{f} \in \underline{F}$:

$$|\tilde{V}^i_\alpha(\underline{f}^{(\infty)};s) - \tilde{V}^i_\alpha(\underline{f}^{(\infty)};s^*)| \leq 4M \sum_{n=0}^{\infty} (1-\rho)^{\lfloor n/\nu \rfloor} = 4M\nu\rho^{-1}, \text{ for all } s\in S$$

thus proving (6.4.11) under each of the C-conditions. □

The proof of theorem 6.4.2 also shows the following corollary.

COROLLARY 6.4.3. *Under any one of the conditions* C1 - C9 *and condition* D, *each limit policy obtained from a sequence of stationary* α-DEP's *with discount-factor tending to one, is an* AEP.

6.5. STOCHASTIC GAMES WITH A FINITE STATE AND ACTION SPACE

In this section, we finally consider the N-person stochastic games with finite state and action space, as studied in ROGERS [97] and SOBEL [117]. We first need the following supplementary notation:

Let $A^i(s) = \{1,\ldots,K^i(s)\}$ and let f^i_{sk}, for any policy $\underline{f} \in \underline{F}$, denote the probability with which the kth alternative ($1 \leq k \leq K^i(s)$) is chosen by player i when entering state $s \in S$.

For any policy $\underline{f} \in \underline{F}$, we define $P^*(\underline{f})$ as the Cesaro limit of the sequence $\{P^n(\underline{f})\}_{n=1}^{\infty}$ and the fundamental matrix $Z(\underline{f}) = [I-P(\underline{f})+P^*(\underline{f})]^{-1}$. For each $i = 1,\ldots,N$ let the bias-vector $w^i(\underline{f})$ be defined by (cf. BLACKWELL [10]):

$$w^i(\underline{f})_s = \sum_{t\in S} Z(\underline{f})_{st}[r^i(t;\underline{f}(t)) - g^i(\underline{f}^{(\infty)};t)];\ s \in S.$$

Finally, let $R(\underline{f}) = \{t \in S | P^*(\underline{f})_{tt} > 0\}$ denote the set of recurrent states for $P(\underline{f})$.

Observe that for each $\underline{f} \in \underline{F}$, $g^i(\underline{f}^{(\infty)};s) = \Sigma_t\ P^*(\underline{f})_{st}\ r^i(t;\underline{f}(t))$ for all $i = 1,\ldots,N$, $s \in S$, and that: (cf. [85])

$$(6.5.1)\quad V^i_\alpha(\underline{f}^{(\infty)};s) = \frac{g^i(\underline{f}^{(\infty)};s)}{1-\alpha} + w^i(\underline{f})_s + o^i(\alpha;\underline{f})_s, \text{ for all } i = 1,\ldots,N,\ s \in S,\ \alpha \in [0,1),$$

where $o^i(\alpha;\underline{f}) = \sum_{\ell=1}^{\infty} (1-\alpha)^\ell\ V^\ell(\underline{f})$, such that

$$(6.5.2)\quad \|o^i(\alpha;\underline{f})\| \leq \omega^i(\alpha;\underline{f}) \overset{\text{def}}{=} \sum_{\ell=1}^{\infty} (1-\alpha)^\ell \|V^\ell(\underline{f})\|.$$

Note that $\omega^i(\alpha;\underline{f})$, as a Taylor series in $(1-\alpha)$, is continuous for $\alpha = 1$ and hence decreases monotonically to 0 as $\alpha \uparrow 1$.

Denote by $n(\underline{f})$ the number of subchains (closed, irreducible sets of states) for $P(\underline{f})$ and let $C^m(\underline{f})$ indicate the mth subchain $(1 \leq m \leq n(\underline{f}))$. Finally, let $\underline{F}_p \subseteq \underline{F}$ denote the *finite* set of pure and stationary policies and define (cf. SCHWEITZER & FEDERGRUEN [109]):

$$R^* = \{s \mid s \in R(\underline{f}) \text{ for some policy } \underline{f} \in \underline{F}_p\},$$

the set of states that are recurrent under some pure policy.

Although the existence of an α-DEP is always guaranteed, it is known from a well-known counterexample by GILLETTE [47] that even in the two person-zero sum case an AEP does not need to exist when for some of the policies $\underline{f}^{(\infty)} \in \underline{F}$, $P(\underline{f})$ is multichained (i.e. $n(\underline{f}) \geq 2$). This seeming contrast with the Markov Decision Processes (MDPs) with finite state and action space is explained by the fact that in stochastic games, as distinct from the former, an essential use is made of the set of all randomized actions, whereas in addition the above result perfectly corresponds with what is known to be the case in MDPs with a finite state space, but arbitrary compact action spaces (cf. BATHER [2]). Under the assumption that for each $\underline{f}^{(\infty)} \in \underline{F}_p$, $P(\underline{f})$ is unichained, the existence of an AEP was first proved in ROGERS [97] and SOBEL [117]. Moreover, in SOBEL [117], as a still stronger property, the existence of a (g,w)- or bias-equilibrium policy $\underline{f}^* \in \underline{F}$ was treated, which we believe should be defined as an AEP $\underline{f}^*$, for which:

$$(6.5.3) \quad w^i(\underline{f}^*)_s \geq w^i(\underline{h})_s \text{ for all } i = 1,\ldots,N,\ s \in S \text{ and } \underline{h} \in \Pi^{-i}(\underline{f}^*) \cap \Pi^i_{AEP}(\underline{f}^*),$$

where

$$\Pi^i_{AEP}(\underline{f}^*) = \{\underline{h} \in \underline{F} \mid g^i(\underline{h})_s = g^i(\underline{f}^*)_s \text{ for all } s \in S\};\ i = 1,\ldots,N$$

(the definition 3 in [117] does not extend the (g,w)-optimality notion in Markov Decision Theory; moreover, with the definition in [117], a (g,w)-optimal policy does not even need to exist in the case $N = 1$, i.e., in the case of an MDP).

In SOBEL [117], the question of the existence of a (g,w)-equilibrium policy was treated using the Brouwer fixed-point theorem with respect to the point-to-point mapping Φ: $\underline{F} \to \underline{F}$, with for all $i = 1,\ldots,N$; $s \in S$ and $k \in A$:

$$\Phi(\underline{f})^i_{sk} = (f^i_{sk} + \phi^i_{sk}(\underline{f}))/(1 + \sum_{\ell \in A} \phi^i_{s\ell}(\underline{f})),$$

where

$$\phi^i_{sk}(\underline{f}) = a^i_{sk} + b^i_{sk} + c^i_{sk},$$

$$(1)\quad a^i_{sk} = \max\{0, \sum_{t \in S} q_{st}([\underline{f}^{-i}(s),k])\ g^i(\underline{f}^{(\infty)};t) - g^i(\underline{f}^{(\infty)};s)\},$$

$$(2)\quad b^i_{sk} = \begin{cases} 0 & \text{if } \sum_s \sum_k a^i_{sk} > 0, \\ \max\{0, r^i(s;[\underline{f}^{-i}(s),k]) + \sum_t q_{st}([\underline{f}^{-i}(s),k])\ w^i(\underline{f})_t - & \\ \quad - g^i(\underline{f}^{(\infty)};s) - w^i(\underline{f})_s\}, & \text{otherwise.} \end{cases}$$

$$(3)\quad c^i_{sk} = \begin{cases} 0 & \text{if } \sum_s \sum_k b^i_{sk} > 0, \\ \max\{0, \sum_t q_{st}([\underline{f}^{-i}(s),k])\ z^i(\underline{f})_t - w^i(\underline{f})_s - z^i(\underline{f})_s\}, & \text{otherwise} \end{cases}$$

where $z^i(\underline{f}) = -Z(\underline{f})\ w^i(\underline{f})$.

Unfortunately, the mapping Φ may be discontinuous in $\underline{f}$, since the $\phi^i_{sk}(\underline{f})$ can be discontinuous in those $\underline{f}$ that satisfy, for all $i = 1,\ldots,N$, $s \in S$ the functional equation:

$$(6.5.4)\quad g^i(\underline{f}^{(\infty)};s) = \max_{k \in A^i(s)} \sum_t q_{st}([\underline{f}^{-i}(s),k])\ g^i(\underline{f}^{(\infty)};t),$$

or the functional equation

$$(6.5.5)\quad w^i(\underline{f})_s + g^i(\underline{f}^{(\infty)};s) = \max_{k \in A^i(s)} \{r^i(s;[\underline{f}^{-i}(s),k]) + \sum_t q_{st}([\underline{f}^{-i}(s),k]) w^i(\underline{f})_t\},$$

but for which, in any sphere in $\underline{F}$ containing $\underline{f}$, policies $\underline{h}$ can be found that do *not* satisfy (6.5.4) (or (6.5.5) respectively). An example of this kind is easy to construct.

While under the assumption in SOBEL [117] that $P(\underline{f})$ is unichained for every policy $\underline{f} \in \underline{F}_p$, the proof in [117] can be rectified in order to show the existence of an AEP (merely by redefining $\phi^i_{sk}(\underline{f}) = b^i_{sk}$ since in this case only criterion (2) is needed), we observe that this result follows immediately from theorem 6.4.2 and the observation that with S a finite state space, condition C2 is automatically satisfied.

We note that in both the counterexamples (to the existence of an AEP) by BATHER [2], example 2.3 and GILLETTE [47], the matrix $P^*(\underline{f})$ is discontinuous in $\underline{f} \in \underline{F}$.

In this section we show in fact that the existence of an AEP is guaranteed, if either $P^*(\underline{f})$ is a continuous (matrix)-function on $\underline{F}$, or if the Markov Decision Process that results for any player $i \in \{1,\ldots,N\}$ when the other players have chosen some stationary policy, is a *communicating* system (cf. BATHER [4] and condition B2 below). Moreover we show that the former property is met under condition B1 below which is an assumption upon the chain structure of the pure (stationary) policies.

In addition, the approach used in this section has again the advantage of showing that AEPs appear as limit policies from a sequence of α-DEPs with discount factor α tending to one.

Let $\underline{f}_1,\ldots,\underline{f}_L$ be an enumeration of $\underline{F}_p$, and consider the following equivalence relation on (cf. the conditions A_1' and A_2' in section 2 of chapter 2):

$$C = \{C^m(\underline{f}_r) \mid 1 \leq r \leq L;\ 1 \leq m \leq n(\underline{f}_r)\}.$$

Let $C \simeq C'$ if there exists $\{C^{(1)} = C, C^{(2)},\ldots,C^{(n)} = C'\}$ with $C^{(i)} \in C$, and $C^{(i)} \cap C^{(i+1)} \neq \emptyset$, for $i = 1,\ldots,n-1$.

Let $C^{(1)},\ldots,C^{(n^*)}$ be the corresponding equivalence classes on C, and let $R^{*(1)},\ldots,R^{*(n^*)}$ be the corresponding partition of R^* (cf. (6.5.2)):

$$R^{*(\ell)} = \bigcup_{\{(m,r) \mid C^m(\underline{f}_r) \in C^{(\ell)}\}} C^m(\underline{f}_r).$$

The following lemma shows that under assumption B1, all policies in $\underline{F}$ have the same number of subchains, i.e. $n(\underline{f})$ is constant on $\underline{F}$:

B1: Every (pure) policy $\underline{f} \in \underline{F}_p$ has exactly one subchain within each $R^{*(\ell)}$, $\ell = 1,\ldots,n^*$.

LEMMA 6.5.1. *If* B1 *holds, then all the policies in* $\underline{F}$ *have the same number of subchains.*

PROOF. Fix $\underline{f}_o \in \underline{F}$. We prove that $P(\underline{f}_o)$ has exactly one subchain within each $R^{*(\ell)}$ $(\ell = 1,\ldots,n^*)$ by showing subsequently:

(1) $R(\underline{f}_o) \subseteq R^*$;

(2) any subchain of $P(\underline{f}_o)$ is contained within one of the sets $R^{*(\ell)}$;

(3) in every one of the sets $R^{*(\ell)}$ there is exactly one subchain of $P(\underline{f}_o)$.

(1) and (2) follow immediately from parts (b) and (c) of Th. 3.2 in [109], so that (3) remains to be shown.

Fix ℓ ($1 \leq \ell \leq n$) and assume first that $R(\underline{f}_o) \cap R^{*(\ell)} = \emptyset$. It then follows from lemma 2.2 in [109] that there exists a pure policy $\underline{h} \in \underline{F}_p$, with $R(\underline{h}) \subseteq R(\underline{f}_o)$, such that $R(\underline{h}) \cap R^{*(\ell)} = \emptyset$, contradicting B1. Finally, observe that for any pair $\underline{f}_1, \underline{f}_2 \in \underline{F}_p$, the subchains of $\underline{f}_1$ and $\underline{f}_2$ that are contained within $R^{*(\ell)}$ must intersect, since it would otherwise be possible to construct a $\underline{f}_3 \in \underline{F}_p$ with two subchains within $R^{*(\ell)}$, contradicting B1, and verify that this property implies that $P(\underline{f})$ cannot have two or more subchains within $R^{*(\ell)}$. □

REMARK. Assume that every policy in $\underline{F}_p$ is unichained (cf. SOBEL [117], ROGERS [97]) and observe that this assumption implies for any pair $\underline{f}_1, \underline{f}_2 \in \underline{F}_p$ that their subchains must intersect, so that all the subchains in C belong to the same equivalence class, i.e. $n^* = 1$.

It hence follows that the assumption in SOBEL [117] and ROGERS [97] is identical with the special case of B1 where $n^* = 1$.

We next introduce condition B2:

B2: For every $i \in \{1,\ldots,N\}$, for every pair of states $s,t \in S$, and for every combination $\{f^j \in F^j | j \neq i\}$ of policies of the other players, there is a policy $f^i \in F^i$ for player i and an integer ℓ such that $P(\underline{f})^{\ell}_{st} > 0$; where $\underline{f} = [f^1,\ldots,f^N]$.

Note that B2 can be seen as an extension of the assumption that the system is communicating (cf. BATHER [2], HORDIJK [58]). It is easily verified (cf. BATHER [3], p.526) that under assumption B2 the seemingly stronger condition (6.5.6) is satisfied.

(6.5.6) For every $i \in \{1,\ldots,N\}$ and for every combination $\{f^j \in F^j | j \neq i\}$ of policies of the other players there is a policy $f^i \in F^i$ for player i, such that $P(\underline{f})$ is an irreducible Markov Chain, where $\underline{f} = [f^1,\ldots,f^N]$.

Using the fact that in an irreducible Markov Chain the mean recurrence time from any state s to any state t is finite one concludes that B2 is in fact the relaxation of condition D to the finite state space model, such that the existence of stationary AEP is guaranteed under this condition. Theorem 6.5.2 below shows that the same applies to condition B1.

THEOREM 6.5.2. *There exists a stationary* AEP *if either* B1 *or* B2 *holds.*

PROOF. The theorem merely has to be proved under B1. Fix $i = 1,\ldots,N$; $s \in S$. It follows from lemma 6.5.1 that $n(\underline{f})$ is constant on $\underline{F}$, and hence from th.5 in SCHWEITZER [105] that $P^*(\underline{f})$ is continuous in $\underline{f} \in \underline{F}$, which in its turn invokes, by their very definition, the continuity of $g^i(\underline{f}^{(\infty)};s)$ and $w^i(\underline{f})_s$ in $\underline{f} \in \underline{F}$, for all $s \in S$.

We first fix an α-DEP $\underline{f}_\alpha \in \underline{F}$, for each $\alpha \in [0,1)$. Inserting (6.5.1) into both sides of (6.2.12) and multiplying both sides of the resulting inequality by $(1-\alpha)$ we obtain for all $h \in F^i$

$$(6.5.7)\quad g^i(\underline{f}_\alpha^{(\infty)};s) + (1-\alpha)\, w^i(\underline{f}_\alpha)_s + (1-\alpha)\, o^i(\alpha;\underline{f}_\alpha)_s \geq$$

$$\geq g^i([\underline{f}_\alpha^{-i},h]^{(\infty)};s) + (1-\alpha)\, w^i([\underline{f}_\alpha^{-i},h])_s + (1-\alpha)\, o^i(\alpha;[\underline{f}_\alpha^{-i},h])_s.$$

It next follows from the fact that $\underline{F}$ is a compact metric space that one can find a policy $\underline{f}^{*(\infty)} \in \underline{F}$, and a sequence $\{\alpha_k\}_{k=1}^{\infty}$, with $\alpha_k \in [0,1)$ and $\lim_{k\to\infty} \alpha_k = 1$, such that $\lim_{k\to\infty} \underline{f}_{\alpha_k} = \underline{f}^*$. We further show:

$$(6.5.8)\quad \lim_{k\to\infty} (1-\alpha_k)\, o^i(\alpha_k;\underline{f}_{\alpha_k})_s = 0 = \lim_{k\to\infty} (1-\alpha_k)\, o^i(\alpha_k;[\underline{f}_{\alpha_k}^{-i},h])_s.$$

Merely proving the first equality in (6.5.8) (the proof of the second one being analogous), we observe that for each $\alpha \in [0,1)$, $o^i(\alpha;\underline{f})_s$ is continuous in $\underline{f} \in \underline{F}$, as a result of lemma 6.3.1, relation (6.5.1) and the continuity of $g^i(\underline{f}^{(\infty)};s)$ and $w^i(\underline{f})_s$ in $\underline{f} \in \underline{F}$.

(6.5.8) then follows from the fact that for any $\underline{h} \in \underline{F}$, $|(1-\alpha)\, o^i(\alpha;\underline{h})_s|$ is bounded by $|(1-\alpha)\, \omega^i(\alpha;\underline{h})_s|$ which decreases monotonically to zero (cf. (6.5.2)), as $\alpha \uparrow 1$, using e.g. Dini's theorem (cf. ROYDEN [100], p.162).

Finally, let k tend to infinity on both sides of (6.5.7) with $\alpha = \alpha_k$, and use (6.5.8) as well as the continuity of $g^i(\underline{f}^{(\infty)};s)$ and $w^i(\underline{f})_s$ in $\underline{f} \in \underline{F}$, in order to obtain:

$$(6.5.9)\quad g^i(\underline{f}^{*(\infty)};s) \geq g^i([\underline{f}^{*-i},h]^{(\infty)};s), \quad \text{for all } i = 1,\ldots,N;$$
$$s \in S \text{ and } h \in F^i.$$

Consider next the "decision problem" that arises when all players but player i tie themselves down to their respective policies in $\underline{f}^*$, and observe from (6.5.9) that in this decision problem, f^{*i} is a maximal gain policy to player i within F^i. It then follows from theorem 2 in BLACKWELL [10], that f^{*i} is also optimal within Π^i.

6.6. N-PERSON GAMES WITH PERFECT INFORMATION

We finally turn to the question under which condition(s) a pure instead of a randomized AEP exists, for every choice of the one-step expected rewards $r^i(s;\underline{a})$.

So far the only stochastic games known to have this property are the so-called two person-zero sum games with perfect information, in which in each state of the system one of the two players has not more than one alternative.

The existence of a pure AEP for this class of stochastic games was first treated by GILLETTE [47]. Unfortunately an incorrect extension of the Hardy-Littlewood theorem was used, as has been pointed out by LIGGETT & LIPPMAN [78].

The existence of a pure AEP, and, as an even stronger result, the existence of a pure bias-equilibrium policy may, however be derived from the fact that a pure stationary α-DEP exists for each $\alpha \in [0,1)$, where the latter has already been proved by SHAPLEY [115].

Since $\underline{F}_p$ is a finite set, we can therefore find a policy $\underline{f}^* = (f^{*1}, f^{*2}) \in \underline{F}_p$ and a sequence $\{\alpha_n\}_{n=1}^{\infty}$, with $\alpha_n \uparrow 1$, such that $\underline{f}^*$ is an α_n-DEP for $n = 1,2,\ldots$ Let $r(s;\underline{a}) = r^1(s;\underline{a}) = -r^2(s;\underline{a})$ and $V_\alpha(\underline{h};s) = V_\alpha^1(\underline{h}^{(\infty)};s) = -V_\alpha^2(\underline{h}^{(\infty)};s)$, and observe that $V_\alpha(\underline{h},s) = \sum_t [I-\alpha P(\underline{h})]^{-1}_{st} r(t;\underline{h}(t))$ is a rational function in α for all $\underline{h} \in \underline{F}_p$ and $s \in S$.

Since $V_\alpha([h^1,f^{*2}];s) - V_\alpha(\underline{f}^*;s)$ and $V_\alpha([f^{*1},h^2];s) - V_\alpha(\underline{f}^*;s)$ are also rational functions in α, for all h^1,h^2 and $s \in S$, and hence are either identically equal to zero or have a finite number of zeros, there exists an $\tilde{\alpha}(h^1,h^2,s)$ such that, for all $\alpha > \tilde{\alpha}(h^1,h^2,s)$:

$$(6.6.1) \qquad V_\alpha([h^1,f^{*2}];s) \leq V_\alpha(\underline{f}^*;s) \leq V_\alpha([f^{*1},h^2];s).$$

Since S and $\underline{F}_p$ are finite, we thus obtain an α^* such that $\underline{f}^*$ is an α-DEP for all $\alpha > \alpha^*$. It then follows by comparing the Laurent series expansion for $V_\alpha(\underline{f}^*)$ and $V_\alpha([h^1,f^{*2}])$ as well as the one of $V_\alpha(\underline{f}^*)$ and $V_\alpha([f^{*1},h^2])$ that $\underline{f}^*$ is a bias-equilibrium policy, and more generally an equilibrium policy under all of the sensitive discount optimality criteria (cf. MILLER & VEINOTT [85]).

REMARK. The proof in LIGGETT & LIPPMAN [78] for the existence of a pure AEP is more complicated than the one above; moreover, it requires an addi-

tional argument. More specifically, instead of th.5 in BLACKWELL [10] we need the stronger result that in each Markov Decision Model there exists a discount factor α^* such that any policy that is α-optimal for some $\alpha > \alpha^*$ is α-optimal for all $\alpha > \alpha^*$, which is immediate from the proof of th.5. Relation (5) in [78] should be adapted in this sense.

One might wonder whether the existence of a pure AEP is also guaranteed in the case of two-person, *nonzero*-sum, or even more generally in the case of N person games with perfect information. The following two-person game is, however a counterexample, which is due to VRIEZE & WANROOIJ [130]. Let $S = \{1,2\}$ and $A^1(1) = A^2(2) = \{1,2\}$ with $A^2(1) = A^1(2) = \{1\}$. Let $r^2(1;(1,1)) = r^1(2;(1,1)) = 1$ and $r^2(1;(2,1)) = r^1(2;(1,2)) = -1$, the other rewards being zero, and let

$$q_{11}(1,1) = q_{21}(1,1) = 2/3 \text{ and } q_{11}(2,1) = q_{21}(1,2) = 1/3.$$

CHAPTER 7

On the functional equations in undiscounted and sensitive discounted stochastic games

7.1. INTRODUCTION AND SUMMARY

After having dealt in the previous chapter with the general N person stochastic games model, we now turn in the last two chapters of this book to the special case where the state and action spaces are finite, with *two* players and where the games are zero-sum. In fact we will consider a slight generalization of the latter, which we will denote as a two-person zero-sum Stochastic Renewal Game *(SRG)*. $\Omega = \{1,\ldots,N\}$ indicates, once again, the finite state space, and in each state $i \in \Omega$, $K(i)$ and $M(i)$ represent the finite sets of actions available to player 1 and 2 resp. When the actions $k \in K(i)$ and $\ell \in M(i)$ are chosen in state i, then

(1) the probability that state j is the next state to be observed, is given by $P_{ij}^{k,\ell} \geq 0$ $(\sum_{j=1}^{N} P_{ij}^{k,\ell} = 1)$

(2) the period of time until the next observation of state, is a random variable t, with conditional probability distribution function $F_{ij}^{k,\ell}(.)$ given that j is the next state of the system

(3) for each $x \geq 0$, $R_{ij}^{k,\ell}(x)$ denotes the expected income earned by player 1 from player 2, during the first x units of time, given that state j is the next state of the system and $t \geq x$.

The discrete time case, where each transition takes exactly *one* unit of time, is known as the stochastic games-model (cf. e.g. [90], [115]) and will be denoted as the *SDG-case*. When one of the two players has only one action in each state of the system, the SRG and SDG model reduce to a Markov Renewal Program *(MRP)*, and a pure Markov Decision Problem *(MDP)* resp. If the payoffs are discounted at the interest rate $r > 0$, the SRG-game is called the *r-discount game*. Let $V(r)$ denote the vector the i-th component of which indicates the value of the r-discount game with initial state

$i \in \Omega$. The existence of a value for the r-discount game goes back to SHAPLEY [115].

In a recent paper, BEWLEY and KOHLBERG [6] gave a description of the asymptotic behaviour of $V(r)$, as the interest rate r decreases to zero, by deriving a series expansion of $V(r)$, for all r sufficiently small. When there is no reason to discount future rewards, or whenever the infinite stage game model serves as an approximation to the model where the planning horizon is finite though large, the *average return per unit time criterion*, in one of its possible specifications (cf. BEWLEY and KOHLBERG [7]) is the first criterion to be considered.

In section 6.5 we recalled from GILLETTE [47] that one or both players may fail to have equilibrium policies with respect to the average return per unit time criterion, and we pointed out a number of recurrency conditions under which the existence of an *AEP* (cf. (6.2.14)) is guaranteed for each possible combination of rewards.

In this chapter we show that in undiscounted SRG's, a pair of functional vector equations arises which is the natural analogue of the corresponding ones in Markov Decision Theory (cf. (1.9.4) and (1.9.5)). We show that, in complete analogy to the structure of MRP's, the existence of a solution to this pair of functional equations is a necessary condition for the existence of a stationary AEP.

We give a constructive proof, showing that a specific class of successive approximation schemes converges to a solution of this pair of functional equations *(f.e.)*.

For the case where the optimal average return per unit time is independent of the initial state of the system, these successive approximation schemes provide an algorithm to locate AEPs whenever existing, as will be pointed out in chapter 8. Conversely and in contrast with what is known to be the case in ordinary MRP's, it is shown that the existence of a solution to the pair of functional equations only needs to be *sufficient* for the existence of an AEP when the asymptotic average value (cf. section 3) is independent of the initial state of the system.

This is explained by showing that a pair of policies which satisfies the two optimality equations for some solution pair, is merely guaranteed to meet some *partial* optimality result (cf. prop.7.3.4).

The above results are obtained in section 3, after giving the notation in section 2.

In section 4, we give some properties of the optimality equations, both for the general multichain and for the unichain-case. Since only the tails of the streams of rewards matter when considering the average return per unit time criterion, more sensitive optimality criteria are needed to make further selections within the class of AEPs. As a consequence, we next consider the extension to the SRG-model of the sensitive discount and cumulative average optimality criteria that have been formulated and studied in the literature on MDPs (cf. MILLER and VEINOTT [85], VEINOTT [127], SLADKŶ [116], DENARDO [22] as well as chapter 4). In section 6, we show that in addition to the above mentioned pair of f.e. an entire sequence of coupled f.e. arises when considering these sensitive optimality criteria. We prove that this sequence has a solution when all of the tpm's associated with the pure stationary policies are unichained. Moreover, we extend the results obtained for the average return per unit time criterion to the entire set of sensitive optimality criteria.

7.2. NOTATION AND PRELIMINARIES

For each finite set S, let $\|S\|$ denote the number of elements, it contains. If $A = [A_{ij}]$ is a matrix, let val A indicate the value of the corresponding matrix game. In this chapter we will use a matrix norm which is different from the one employed so far (cf. (2.1.4)). Accordingly, let $|A| = \max_{i,j}|A_{ij}|$, and observe that for any pair of matrices A,B of equal dimensions:

(7.2.1) $\quad |\text{val } A - \text{val } B| \leq |A - B|$.

(Let (x^A,y^A) and (x^B,y^B) be equilibrium pairs of actions in the matrix games A and B; then $\min_{i,j}(A_{ij}-B_{ij}) \leq x^B (A-B)y^A = x^B Ay^A - x^B By^A \leq$ val A - val B $\leq x^A Ay^B - x^A By^B = x^A (A-B)y^B \leq \max_{i,j}(A_{ij}-B_{ij})$).

For each state $i \in \Omega$, let $K(i) = \{x \in E^{\|K(i)\|} \mid x \geq 0, \sum_{k=1}^{\|K(i)\|} x_k = 1\}$ denote the set of all randomized actions available to player 1 in state i. Similarly $M(i) = \{y \in E^{\|M(i)\|} \mid y \geq 0, \sum_{\ell=1}^{\|M(i)\|} y_\ell = 1\}$ indicates the set of all randomized actions available to player 2 in state $i \in \Omega$. For every $i \in \Omega$, any tableau of numbers $[c_i^{k,\ell}]$, $k = 1,\ldots,\|K(i)\|$; $\ell = 1,\ldots,\|M(i)\|$ and for each pair of closed convex subsets $\tilde{K}(i) \subseteq K(i)$ and $\tilde{M}(i) \subseteq M(i)$, we denote by

(7.2.2) $\quad [\tilde{K}(i),\tilde{M}(i)]^{[c_i^{k,\ell}]}$

the two-person zero-sum game which has $\tilde{K}(i)$ and $\tilde{M}(i)$ as the action sets for player 1 and 2 resp. and where the payoff to player 1, is given by

$$\sum_{k=1}^{\|K(i)\|} \sum_{\ell=1}^{\|M(i)\|} x_k \, c_i^{k,\ell} \, y_\ell ,$$

when the players choose action $x \in \tilde{K}(i)$ and $y \in \tilde{M}(i)$ resp. The minimax value of this game is indicated by $\mathrm{val}_{[\tilde{K}(i)\tilde{M}(i)]}[c_i^{k,\ell}]$.

When $\tilde{K}(i) = K(i)$ and $\tilde{M}(i) = M(i)$ we use the abbreviated notation $[c_i^{k,\ell}]$ to indicate the game in (7.2.2). The following lemma is immediate from KARLIN ([70], pp.63):

LEMMA 7.2.1. *Fix* $i \in \Omega$. *Let* $\tilde{K}(i)$ *and* $\tilde{M}(i)$ *be closed convex polyhedral subsets of* $K(i)$ *and* $M(i)$. *Then the sets of optimal actions in any one of the two-person zero-sum games in* (7.2.2) *are again closed convex polyhedral subsets of* $\tilde{K}(i)$ *and* $\tilde{M}(i)$ *and thus of* $K(i)$ *and* $M(i)$.

Note that a stationary strategy f (h) for player 1 (2) is characterized by a tableau $[f_{ik}]$ ($[h_{i\ell}]$) satisfying $f_{ik} \geq 0$ and $\sum_{k \in K(i)} f_{ik} = 1$ ($h_{ik} \geq 0$ and $\sum_{\ell \in M(i)} h_{i\ell} = 1$), where $f_{ik}(h_{i\ell})$ is the probability that the k-th (ℓ-th) alternative is chosen when entering state $i \in \Omega$. We let $\Phi(\Psi)$ denote the set of all stationary policies for player 1 (2).

When a positive interest rate r is introduced into the model, income earned at time t is discounted by the factor e^{-rt} and the associated SRG will be referred to as the r-discount game. When in state i the players choose action $k \in K(i)$ and $\ell \in M(i)$, the one-step expected r-discounted reward for player 1 is given by:

$$(7.2.3) \qquad \rho_i^{k,\ell}(r) = \sum_j p_{ij}^{k,\ell} \int_0^\infty \int_0^x e^{-rt} \, d\, R_{ij}^{k,\ell}(t) \, d\, F_{ij}^{k,\ell}(x)$$

and let $q_i^{k,\ell} = \rho_i^{k,\ell}(0)$ denote the one-step expected *(undiscounted)* reward. Let

$$\tau_{ij}^{k,\ell} = \int_0^\infty x \, d\, F_{ij}^{k,\ell}(x) < \infty$$

denote the expected *conditional* holding time in state i, when the players choose actions $k \in K(i)$, $\ell \in M(i)$ and given that the next state observed is state j. Likewise, let

$$T_i^{k,\ell} = \sum_j H_{ij}^{k,\ell} \quad \text{where} \quad H_{ij}^{k,\ell} = P_{ij}^{k,\ell}\, \tau_{ij}^{k,\ell}$$

denote the expected *unconditional* holding time in state i, when $k \in K(i)$, and $\ell \in M(i)$ are the actions chosen, and assume $T_i^{k,\ell} > 0$ ($i \in \Omega$, $k \in K(i)$, $\ell \in M(i)$).

Like in chapter 1 we associate with each pair $(f,h) \in \Phi \times \Psi$ the N-component reward vector $q(f,h)$, the holding time vector $T(f,h)$, and the matrices $P(f,h)$ and $H(f,h)$:

$$q(f,h)_i = \sum_{k\in K(i)} \sum_{\ell\in M(i)} f_{ik} \cdot q_i^{k,\ell} \cdot h_{i\ell} \quad ; i \in \Omega$$

$$T(f,h)_i = \sum_{k\in K(i)} \sum_{\ell\in M(i)} f_{ik} \cdot T_i^{k,\ell} \cdot h_{i\ell} \quad ; i \in \Omega$$

$$P(f,h)_{ij} = \sum_k \sum_\ell f_{ik} \cdot P_{ij}^{k,\ell} \cdot h_{i\ell} \quad ; i,j \in \Omega$$

$$H(f,h)_{ij} = \sum_k \sum_\ell f_{ik} \cdot H_{ij}^{k,\ell} \cdot h_{i\ell} \quad ; i,j \in \Omega.$$

For any pair $(f,h) \in \Phi \times \Psi$ define the stochastic matrix $\Pi(f,h)$ as the Cesaro limit of the sequence $\{P^n(f,h)\}_{n=1}^{\infty}$. Denote by $n(f,h)$ the number of subchains (closed, irreducible sets of states) for $P(f,h)$ and let $R(f,h) = \{i \in \Omega \mid \Pi(f,h)_{ii} > 0\}$ i.e. $R(f,h)$ is the set of recurrent states for $P(f,h)$. $C^m(f,h)$ ($m = 1,\ldots,n(f,h)$) denotes the m-th subchain of $P(f,h)$; $\Phi_i^m(f,h)$ is the probability of absorption in $C^m(f,h)$ when starting in state i, and $\pi^m(f,h)$ represents the (unique) stationary probability distribution of $P(f,h)$ on $C^m(f,h)$. Finally, $g(f,h)$ represents the gain rate vector when the two players use policy $f \in \Phi$, and $h \in \Psi$.

Next, we recall that $V(r)$, the value vector of the r-discount game, satisfies the equation (cf. SHAPLEY [115]):

$$(7.2.4) \qquad V(r)_i = \text{val}[\rho_i^{k,\ell}(r) + \sum_j m_{ij}^{k,\ell}(r)\, V(r)_j], \quad i \in \Omega,\ r > 0$$

where

$$m_{ij}^{k,\ell}(r) = P_{ij}^{k,\ell} \int_0^\infty e^{-rt}\, d\, F_{ij}^{k,\ell}(t) \geq 0.$$

BEWLEY and KOHLBERG [6] recently showed for the discrete time case (SDG's) that $V(r)$ may be expressed as a real fractional power or *Puiseux* series in r, for all interest rates r that are sufficiently close to 0. More specifically, there exists an integer $L \geq 1$ and constants $a_i^{(k)}$ ($i \in \Omega$; $k = -\infty,\ldots,L$) such that:

(7.2.5) $\quad V(r) = \sum_{k=-\infty}^{L} a^{(k)} r^{-k/L}$

This result carries easily over to the general SRG-case. We henceforth assume that $\rho_i^{k,\ell}(r)$ and $m_{ij}^{k,\ell}(r)$ have a Taylor series expansion ($i \in \Omega$; $k \in K(i)$; $\ell \in M(i)$):

LEMMA 7.2.2. *V(r) has a Puiseux series expansion as in* (7.2.5).

PROOF. The proof goes along the lines of section 11 in [6]. Note that $\sum_j m_{ij}^{k,\ell}(r) < 1$ for all $r > 0$ and all $i \in \Omega$; $k \in K(i)$ and $\ell \in M(i)$. Observe next, from a standard contraction mapping argument that

(7.2.6) the equation $x_i = val[\rho_i^{k,\ell}(r) + \sum_j m_{ij}^{k,\ell}(r)\, x_j]$, $i \in \Omega$ has a (unique) solution for all values of the parameters $\rho_i^{k,\ell}(r)$ and $m_{ij}^{k,\ell}(r)$ such that $m_{ij}^{k,\ell}(r) \geq 0$ and $\sum_j m_{ij}^{k,\ell}(r) < 1$.

Since (7.2.6) is a sentence in elementary algebra (cf. cor.9.2 in [6]) it follows from Tarski's principle (cf. section 11 in [6]) that (7.2.6) is true over any real closed field, if it is true over the reals. Finally, the set of all real Puiseux series was shown to be a closed ordered field (cf. section 10 in [6]) which completes the proof that V(r) has an expansion of the type $\sum_{k=-\infty}^{K} a^{(k)} r^{-k/L}$ for some pair of integers $K,L \geq 1$. The fact that $a^{(k)} = 0$, for all $k > L$ finally follows from the proof of th. 7.2 in [6] and the observation that for all r sufficiently small,

(7.2.7) $\quad \sum_j m_{ij}^{k,\ell}(r) \leq 1 - \tfrac{1}{2} r\, T_{min}, \quad i \in \Omega,\ k \in K(i),\ \ell \in M(i)$

where $T_{min} = \min_{i,k,\ell} T_i^{k,\ell} > 0$. To verify (7.2.7) note that $e^{-rt} \leq 1 - \tfrac{1}{2}rt$ for all r sufficiently close to zero, and use the definition of $m_{ij}^{k,\ell}(r)$. □

We recall from the example in section 14 of [6] that in general V(r) cannot be expressed as a rational function or *Laurent* series in r (i.e. the special case of (7.2.5) where L = 1) as is known to be the case in ordinary MRP's (cf. [22], [85]). The vector $a^{(L)}$ in (7.2.5) is called the *asymptotic average value* vector. Finally it was shown in [6] that in the *discrete-time case* (of SDG's), $a^{(L)} = \lim_{n\to\infty} v(n)/n$ where v(n) is the vector, whose i-th component denotes the value of the n-step game with initial state i.

7.3. THE AVERAGE RETURN CRITERION; A PAIR OF FUNCTIONAL EQUATIONS

In this section we are concerned with the average return per unit time criterion, i.e. we evaluate any pair of (possibly non-stationary) policies for players 1 and 2, by considering for each initial state $i \in \Omega$:

$$(7.3.1) \qquad g(\varphi,\psi)_i = \liminf_{n\to\infty} (E_{\varphi,\psi} \sum_{r=1}^{n} \rho_r)/(E_{\varphi,\psi} \sum_{r=1}^{n} \tau_r); \quad i \in \Omega,$$

where $\rho_r(\tau_r)$ denotes the payoff to player 1 (the length of the period) in between of the r-1-st and the r-th observation of state. $E_{\varphi,\psi}$ indicates the expectation, given the players' policies φ and ψ. A number of equivalent criteria have been formulated in [8].

It is known from GILLETTE [47] that one or both players may fail to have gain- optimal policies. For the discrete-time case, we pointed out in section 5 of the previous chapter, that the existence of a stationary AEP is guaranteed for every possible combination of one-step expected rewards $q_i^{k,\ell}$ if the matrix function $\Pi(f,h)$ is continuous on $\Phi \times \Psi$. In addition we remarked that the latter, in its turn, is guaranteed to hold when the number of subchains $n(f,h)$ is continuous, i.e. constant on $\Phi \times \Psi$, and a (finitely verifiable) sufficient condition with respect to the chain structure of the set of *pure* policies, was provided by lemma 6.5.1. Lemma 7.3.1 shows that these results carry over to the general SRG-case:

LEMMA 7.3.1. *Let* $n(f,h)$ *be constant on* $\Phi \times \Psi$. *Then there exists a stationary AEP.*

PROOF. Note first of all that an equilibrium pair of stationary policies exists in the r-discount game, for all $r > 0$. For any pair of policies $f,h \in \Phi \times \Psi$ let $V(f,h)(r)$ denote the total expected return vector in the r-discount game, and consider its Laurent series expansion in powers of the interest rate r (cf. DENARDO [22]):

$$V(f,h)(r) = r^{-1}g(f,h) + w(f,h) + o(r;f,h)$$

We next observe from lemma 1 and corollary 2 in DENARDO [22] that both $g(f,h)$ and $w(f,h)$ may be written as rational functions of $q(f,h)$; $P(f,h)$; $\Phi^m(f,h)$ and $\pi^m(f,h)$; $Z(f,h) = [I-P(f,h) + \Pi(f,h)]^{-1}$; $T(f,h)$ as well as of bilinear functions in the first two (three) terms of the power series expansions of $\rho_i^{k,\ell}(r)$ (and $m_{ij}^{k,\ell}(r)$). Using theorem 5 in SCHWEITZER [105] we note

that all of these functions are continuous on $\Phi \times \Psi$, in view of $n(f,h)$ being constant. Conclude that both $g(f,h)$ and $w(f,h)$ are continuous on $\Phi \times \Psi$, as well, whence the existence of a stationary AEP can be proved in complete analogy to the proof of theorem 6.5.1. □

As a special case of the condition in lemma 7.3.1, we have that stationary AEPs exist, if

(U): every pair of pure stationary policies is unichained.

In this section, we show that the following pair of optimality equations arises when analyzing the average return criterion:

$$(7.3.3) \quad g_i = \mathrm{val}[\textstyle\sum_{j=1}^{N} p_{ij}^{k,\ell} g_j], \quad i \in \Omega$$

$$(7.3.4) \quad v_i = \mathrm{val}_{[K(i,g),M(i,g)]}[q_i^{k,\ell} - \textstyle\sum_j H_{ij}^{k,\ell} g_j + \sum_j p_{ij}^{k,\ell} v_j], \quad i \in \Omega,$$

where for each $i \in \Omega$, and each solution g^* to (7.3.3), $K(i,g^*)$ and $M(i,g^*)$ are the sets of optimal actions in the matrix game in (7.3.3) with $g = g^*$. Note from lemma 7.2.1 that the sets $K(i,g)$ and $M(i,g)$ are in fact the convex hulls of a finite number of extreme points such that the games to the right of (7.3.4) may be interpreted as simple matrix games. Observe finally that (7.3.3) and (7.3.4) are the natural extension of the optimality equations (1.9.7) and (1.9.8) in undiscounted MRPs.

We say that a pair of policies (f^*,h^*) satisfies the optimality equations (7.3.3) and (7.3.4), if for some solution (g^*,v^*), $(f^*(i), h^*(i))$ is an equilibrium pair of actions in the matrix games to the right of (7.3.3) and (7.3.4). First we show that a solution pair to the equations (7.3.3) and (7.3.4) exists whenever a stationary AEP exists. Our proof is a constructive one; in fact we show that a certain class of successive approximation schemes converge to a solution pair of (7.3.3) and (7.3.4). These schemes are the natural analogue of a value iteration scheme in undiscounted MDPs which is due to HORDIJK and TIJMS [60] and which was presented in section 1.8 (cf. also section 4.3). First of all, observe that $a^{(L)}$, the asymptotic average value vector (cf. (7.2.5)) is a solution to (7.3.3):

$$(7.3.5) \quad a_i^{(L)} = \mathrm{val}[\textstyle\sum_j p_{ij}^{k,\ell} a_j^{(L)}], \quad i \in \Omega$$

as is easily verified by inserting (7.2.5) into both sides of (7.2.4), multiplying the resulting equality by $r > 0$, letting r tend to zero, and by interchanging the limit and value-operation (cf. (7.2.1)).

We next consider a related SRG, with Ω as state space. For each $i \in \Omega$, $k \in K(i)$, $\ell \in M(i)$ let,

$$\tilde{R}_{ij}^{k,\ell}(x) = R_{ij}^{k,\ell}(x) - a_j^{(L)} x; \quad i,j \in \Omega;\ k \in K(i);\ \ell \in M(i)$$

denote the income functions, and verify using (7.2.3) that

$$(7.3.6) \qquad \tilde{q}_i^{k,\ell} = q_i^{k,\ell} - \sum_j H_{ij}^{k,\ell}\, a_j^{(L)}; \quad i \in \Omega;\ k \in K(i),\ \ell \in M(i).$$

Both the transition probabilities and the transition time distributions remain unaltered. Moreover we restrict in each state $i \in \Omega$ the set of (randomized) actions available for player 1 to $K(i,a^{(L)})$ and the set of actions for player 2 to $M(i,a^{(L)})$. $\tilde{V}(r)$, $\tilde{a}^{(k)}$, $k = -\infty,\ldots,\tilde{L}$ and for each $f \in X_{i=1}^{N} K(i,a^{(L)})$, $h \in X_i M(i,a^{(L)})$ the quantities $\tilde{q}(f,h)$ and $\tilde{g}(f,h)$ are defined in complete analogy to $V(r)$, $a^{(k)}$, $k = -\infty,\ldots,L$; $q(f,h)$ and $g(f,h)$. Before introducing the successive approximation schemes we first need the following theorem:

THEOREM 7.3.2. *Assume there exists a stationary* AEP $(f^*,h^*) \in \Phi \times \Psi$ *in the original stochastic renewal game* (SRG). *Then*

(a) $a^{(L)} = g(f^*,h^*)$

(b) $a^{(k)} = 0$, $k = 1,\ldots,L-1$

(c) *every policy* $f \in \Phi$ $(h \in \Psi)$ *which is gain-optimal for player* 1 *(or* 2*) in the original* SRG, *is gain-optimal in the transformed* SRG

(d) *there exists a constant* $B > 0$, *and an integer* $\tilde{L} \geq 1$, *such that for all* $r, s > 0$ *sufficiently small:*

$$(7.3.7) \qquad \|\tilde{V}(r) - \tilde{V}(s)\| \leq B\left|r^{1/\tilde{L}} - s^{1/\tilde{L}}\right|.$$

PROOF.

(a),(b): go along the lines of lemma 7.1.1 in [8]:

Let $\sum_{k=-\infty}^{1} A^{(k)} r^{-k}$ $[\sum_{k=-\infty}^{1} B^{(k)} r^{-k}]$ be the Laurent series expansion of $W^1(r)$ $[W^2(r)]$, the total discounted return to player 1 [2] in the MRP that results when player 2 [1] ties himself down to policy $h^*[f^*]$. Since $f^*[h^*]$ is gain-optimal in this MRP conclude that $A^{(1)} = B^{(1)} = g(f^*,h^*)$. Finally, parts (a) and (b) follow from the inequalities $W^2(r) \leq V(r) \leq W^1(r)$.

(c) Fix a stationary AEP (f^*,h^*) and $i \in \Omega$; recall (e.g. from th.1 in [23]) that for any $f \in \Phi$ and $h \in \Psi$:
$P(f,h^*)\, g(f^*,h^*) \leq P(f^*,h^*)\, g(f^*,h^*) = g(f^*,h^*) \leq P(f^*,h)\, g(f^*,h^*)$
such that: $[P(f,h^*)\, a^{(L)}]_i \leq [P(f^*,h^*)\, a^{(L)}]_i = a_i^{(L)} \leq [P(f^*,h)\, a^{(L)}]_i$, thus proving that $f^*(i) \in K(i,a^{(L)})$ and $h^*(i) \in M(i,a^{(L)})$ for all $i \in \Omega$, or in other words the feasibility of f^* and h^* in the transformed game. We next show, that (f^*,h^*) is an AEP in the transformed SRG, with $\tilde{g}(f^*,h^*) = 0$, by proving:

(7.3.8)
$$\text{(i)}\ \tilde{g}(f,h^*) = g(f,h^*) - a^{(L)},\ \text{for all } f \in X_i\, K(i,a^{(L)})$$
$$\text{(ii)}\ \tilde{g}(f^*,h) = g(f^*,h) - a^{(L)},\ \text{for all } h \in X_i\, M(i,a^{(L)}).$$

Confining ourselves to (7.3.8) (i) (the proof of (ii) being analogous) fix $f \in X_i\, K(i,a^{(L)})$, and observe by iterating the equality $a^{(L)} = P(f,h^*)\, a^{(L)}$, that:

$$a^{(L)} = \begin{cases} c^{(m)} & \text{for all } i \in C^m(f,h^*);\ m = 1,\ldots,n(f,h^*) \\ \sum_{m=1}^{n(f,h^*)} \Phi_i^m(f,h^*)\, c^{(m)}, & \text{for all } i \in \Omega\backslash R(f,h^*). \end{cases}$$

Then,

$$\tilde{g}^{(m)}(f,h^*) = \frac{\langle \pi^m(f,h^*), q(f,h^*) - H(f,h^*)a^{(L)} \rangle}{\langle \pi^m(f,h^*), T(f,h^*) \rangle}$$

$$= g^{(m)}(f,h^*) - c^{(m)}\left(\sum_i \pi^m(f,h^*)_i \sum_j H(f,h^*)_{ij}\right)/\langle \pi^m(f,h^*), T(f,h^*) \rangle$$

$$= g^{(m)}(f,h^*) - c^{(m)};\ m = 1,\ldots,n(f,h^*)$$

and conclude that $\tilde{g}(f,h^*)_i = \sum_{m=1}^{n(f,h^*)} \Phi_i^m(f,h^*)\, \tilde{g}^{(m)}(f,h^*) = g(f,h^*)_i - a_i^{(L)}$, for all $i \in \Omega$.

(d) The proof of part (b) shows that $\tilde{a}^{(\tilde{L})} = 0$ as well as the existence of a stationary AEP in the transformed model, and the latter implies by applying part (a) to the transformed game, that $\tilde{a}^{(k)} = 0$ for $k = 1,\ldots,\tilde{L}-1$ as well i.e. for all r sufficiently small:

$$\tilde{V}(r) = \sum_{k=0}^{\infty} a^{(-k)} r^{k/\tilde{L}}. \tag{7.3.9}$$

Now by applying the mean value theorem and using the fact that $\tilde{V}(r)$ as a power series in $r^{1/\tilde{L}}$ has a continuous derivative at $r = 0$, we obtain the desired result.

We next introduce the following successive approximation scheme:

$$(7.3.10)\quad y(n)_i = \mathrm{val}_{[K(i,a^{(L)}),M(i,a^{(L)})]}[\tilde{\rho}_i^{k,\ell}(r_n) + \sum_j m_{ij}^{k,\ell}(r_n)y(n-1)_j]$$

where $\{r_n\}_{n=1}^{\infty}$ is a sequence of interest rates, with $\lim_{n\to\infty} r_n = 0$. Under the assumption that a stationary AEP exists, the following theorem exhibits the existence of a solution pair to the optimality equations (7.3.3) and (7.3.4) by showing in analogy to th.1 in HORDIJK and TIJMS [60] that the sequence $\{y(n)\}_{n=1}^{\infty}$ converges under specific conditions on $\{r_n\}_{n=1}^{\infty}$:

THEOREM 7.3.3. *Assume the original* SRG *has a stationary* AEP. *Then:*

(a) $(a^{(L)},\tilde{a}^{(0)})$ *is a solution pair to the* f.e. (7.3.4) *and* (7.3.5)

(b) *Let* $\{r_n\}_{n=1}^{\infty}$ *satisfy the conditions:*

(1) $(1-r_1') \cdots (1-r_n') \to 0$, *as* $n \to \infty$

(2) $\sum_{j=2}^{n} (1-r_n') \cdots (1-r_j') \, |r_j'^{1/\tilde{L}} - r_{j-1}'^{1/\tilde{L}}| \to 0$, as $n \to \infty$

where $r_j' = \frac{1}{2} r_j \min_{i,k,\ell} T_i^{k,\ell}$. Then $\lim_{n\to\infty} y(n) = \tilde{a}^{(0)}$.

PROOF.

(a) part (a) follows immediately from part (b) by letting n tend to infinity on both sides of (7.3.10) and by observing that the value of a matrix game depends continuously upon its entries (cf. (7.2.1)).

(b) Note that

$$\tilde{V}(r_n)_i = \mathrm{val}_{[K(i,a^{(L)}),M(i,a^{(L)})]}[\tilde{\rho}_i^{k,\ell}(r_n) + \sum_j m_{ij}^{k,\ell}(r_n)\tilde{V}(r_n)_j], \quad i \in \Omega$$

and conclude from (7.2.1) that $|y(n+1)_i - \tilde{V}(r_n)_i| \leq (\max_{i,k,\ell} \sum_j m_{ij}^{k,\ell}(r_n))\cdot \|y(n) - \tilde{V}(r_n)\| \leq (1-\frac{1}{2}r_n \cdot T_{\min})\|y(n) - \tilde{V}(r_n)\| = (1-r_n')\|y(n) - \tilde{V}(r_n)\|$, for all $i \in \Omega$, where in (7.2.7) the second inequality was shown to hold for all r_n sufficiently close to zero, i.e. for all $n \geq n_0$ (say). As a consequence of theorem 7.3.2 part (d), we may fix an integer $n_1 \geq n_0$ such that for all $m \geq n_1$: $\|\tilde{V}(r_{m+1}) - \tilde{V}(r_m)\| \leq B|r_{m+1}^{1/\tilde{L}} - r_m^{1/\tilde{L}}| = B'|r_{m+1}'^{1/\tilde{L}} - r_m'^{1/\tilde{L}}|$, where $B' = (2/T_{\min})^{1/\tilde{L}}B$. We conclude that for all $m = 1,2,\ldots$:

$$\|y(n_1+m) - \tilde{V}(r_{n_1+m})\| \leq (1-r'_{n_1+m-1})\|y(n_1+m-1) - \tilde{V}(r'_{n_1+m-1})\|$$

$$+ B'(1-r'_{n_1+m-1}) \, |r'^{1/\tilde{L}}_{n_1+m} - r'^{1/\tilde{L}}_{n_1+m-1}|$$

and by iterating this inequality, we finally obtain:

$$\|y(n_1+m) - \tilde{V}(r_{n_1+m})\| \leq (1-r'_{n_1+m-1}) \cdots (1-r'_{n_1+1})\|y(n_1) - \tilde{V}(r_{n_1})\|$$
$$+ B' \sum_{j=n_1}^{n_1+m-1} (1-r'_{n_1+m-1}) \cdots (1-r'_j) \left| r'^{1/\tilde{L}}_{j+1} - r'^{1/\tilde{L}}_{j} \right|.$$

It follows from (7.3.9) that $\lim_{n\to\infty} \tilde{V}(r_n) = \tilde{a}^{(0)}$ which in view of the above inequality and the properties imposed upon $\{r_n\}_{n=1}^{\infty}$, enables us to conclude that $\lim_{n\to\infty} y(n)-\tilde{a}^{(0)} = \lim_{n\to\infty} y(n) - \tilde{V}(r_n) = 0$. □

We note that with $\tilde{q}_i^{k,\ell}$ redefined by $\tilde{q}_i^{k,\ell} = q_i^{k,\ell} - a_i^{(L)} \tau_i^{k,\ell}$ (instead of (7.3.6)) the analysis of th.7.3.2 and th.7.3.3 leads just as well to the existence of a solution to the following pair of f.e., whenever a stationary AEP exists:

$$(7.3.11) \quad g_i = \text{val}[\textstyle\sum_j p_{ij}^{k,\ell} g_j], \quad i \in \Omega$$

$$(7.3.12) \quad v_i = \text{val}_{[K(i,g),M(i,g)]}[q_i^{k,\ell} - g_i \tau_i^{k,\ell} + \textstyle\sum_j p_{ij}^{k,\ell} v_j], \quad i \in \Omega.$$

We next observe that conditions (1) and (2) of theorem 7.3.3, part (b) are satisfied for any choice:

$$r_n = n^{-b}, \quad \text{with } 0 < b \leq 1$$

as will be verified in lemma 8.2.1.

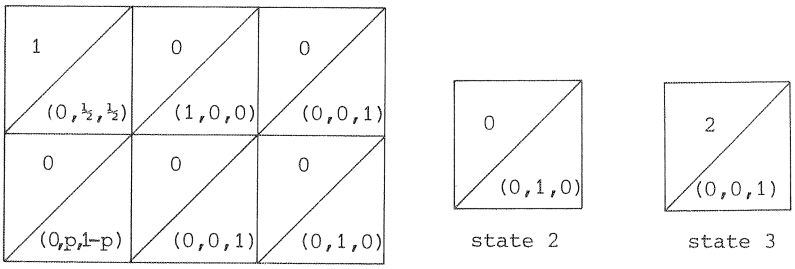
In addition, we note that when the asymptotic average value is independent of the initial state of the system, i.e. when $a_i^{(L)} = \langle a^{(L)} \rangle$ for all $i \in \Omega$, the f.e. (7.3.3) and (7.3.4), as well as (7.3.11) and (7.3.12) reduce to the single (vector)-equation:

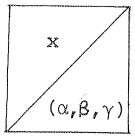
$$(7.3.13) \quad v_i^* = \text{val}[q_i^{k,\ell} - \langle g^* \rangle \tau_i^{k,\ell} + \textstyle\sum_j p_{ij}^{k,\ell} v_j^*], \quad i \in \Omega$$

the discrete-time version of which has been considered in HOFFMAN and KARP [57]. In this case, the convergence result of part (b) of the previous theorem leads to a method for approximating the asymptotic average value by lower and upper bounds as well as for finding for both players and any $\varepsilon > 0$ stationary policies which are ε-optimal with respect to the average return criterion (cf. chapter 8).

EXAMPLE 1 *shows that whereas the existence of a solution pair to the* f.e. (7.3.3) *and* (7.3.4) *is a necessary condition for the existence of a stationary* AEP, *it may fail to be sufficient:*

EXAMPLE 1. (all $\tau_{ij}^{k,\ell} = 1$; $i,j \in \Omega$; $k \in K(i)$, $\ell \in M(i)$).



The notation 

means that if the players choose the row and column corresponding to this box, then player 2 pays player 1 the amount x and the next state is 1, with probability α, 2 with probability β and 3 with probability $\gamma = 1-\alpha-\beta$. Take $p = \frac{3}{4}$. We first verify that $a^{(L)} = [1,0,2]$. Note, by $\lim_{n\to\infty} \frac{v(n)}{n} = a^{(L)}$ that $a_2^{(L)} = 0$ and $a_3^{(L)} = 2$. Next, we show that $x = 1$ is the *unique* solution to the equation:

$$(7.3.14) \quad x = \mathrm{val}\begin{bmatrix} 1 & x & 2 \\ 0.5 & 2 & 0 \end{bmatrix}$$

by distinguishing between the cases $x > 1$ and $x < 1$. Finally recall from (7.3.5) that $a_1^{(L)}$ is a solution to (7.3.14). We next verify that $a^{(L)}$ in combination with any vector $v^* \in E^3$ satisfying $v_1^* = \frac{1}{2}v_2^* + \frac{1}{2}v_3^* - \frac{1}{2}$, is a solution pair to (7.3.3) and (7.3.4). Note that $K(1,a^{(L)}) = \{[1,0]\}$ and $M(1,a^{(L)}) = \{[y_1,y_2,y_3] \mid y_3 = 0, y_1+y_2 = 1, y_1 \geq \frac{2}{3}\}$, such that in this example (7.3.4) becomes:

$$v_1^* = \min\{\tfrac{1}{2}v_2^* + \tfrac{1}{2}v_3^*;\ -\tfrac{1}{3} + \tfrac{1}{3}v_1^* + \tfrac{1}{3}v_2^* + \tfrac{1}{3}v_3^*\};\ v_2^* = v_2^*;\ v_3^* = v_3^*.$$

We next verify that there is no stationary AEP in this game. Note that a stationary policy for player 1 is completely specified by the probability x with which action 1 in state 1 is chosen. Likewise, a stationary policy for player 2, is specified by the probability vector (y_1,y_2,y_3) with which

the available actions in state 1 are randomized.

$$\text{If } x = 1 \text{ and } y_2 = 1\text{: } g(f,h)_1 = 0$$

$$\text{if } x = 1 \text{ and } y_2 < 1\text{: } g(f,h)_1 = (y_1+2y_3)/(y_1+y_3)$$

$$\text{if } x < 1 \text{: } g(f,h)_1 = \frac{x(\frac{1}{2}y_1-2y_2+2y_3)+\frac{1}{2}y_1+2y_2}{1-xy_2}$$

which shows that no pair of stationary policies is an AEP.

(Note first that only pairs of policies with $x < 1$, can be candidates for an AEP, since with $x = 1$, player 2 reacts optimally by putting $y_2 = 1$, whereas with the choice $y_2 = 1$, player 1 is strictly better off when choosing $x < 1$. Next observe that with $x < 1$, player 1 can guarantee himself $\min\{\frac{1}{2}x+\frac{1}{2};2;2x\}$ which allows him to come arbitrarily close to one, without actually reaching the value one itself).

We conclude that in general, and in contrast to what is known to be the case for ordinary MRPs, a policy pair (f^*,h^*) *which satisfies the optimality equations* (7.3.3) *and* (7.3.4) *for some solution pair* (g^*,v^*), *does not need to be an AEP.*

Example 1, with $p = \frac{1}{2}$ shows that this may even be the case when stationary AEPs do exist:

Note that with $p = \frac{1}{2}$, $g^* = [1,0,2]$satisfies (7.3.3), by verifying

$$1 = \text{val}\begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 0 \end{bmatrix},$$

and conclude that $K(1,g^*) = \{[x,1-x] \mid \frac{1}{2} \le x \le 1\}$ and $M(1,g^*) = \{[1,0,0]\}$. The optimality equation (7.3.4) thus becomes:

$$v_1^* = \max\{\tfrac{1}{2}v_1^* + \tfrac{1}{2}v_2^*;\ \tfrac{1}{2}v_1^* + \tfrac{1}{2}v_2^* - \tfrac{1}{2}\};\ v_2^* = v_2^*;\ v_3^* = v_3^*$$

such that g^* in combination with any vector $v^* \in E^3$ satisfying $v_1^* = \frac{1}{2}v_2^*+\frac{1}{2}v_3^*$, is a solution pair to (7.3.3) and (7.3.4). Verify that any pair of policies (f^*,h^*) with $\frac{1}{2} \le f_{11}^* < 1$ and $h_{11}^* = 1$ is a stationary AEP, whereas the only pair of policies which satisfies the optimality equations (7.3.3) and (7.3.4) for (g^*,v^*) has $f_{11}^* = 1$ and $h_{11}^* = 1$ and is *not* an AEP. Observe finally (by considering the gain rate of one of the AEPs) that $g^* = a^{(L)}$.

We conclude that even when stationary AEPs do exist, such policy pairs do not necessarily need to be found within the class of (pairs of) policies that satisfy the optimality equations for some solution pair (the existence

of which follows from th.7.3.3).

Whereas the above examples illustrate that no *full* optimality results may be obtained for policy pairs that satisfy the optimality equations (7.3.3) and (7.3.4) for some solution pair (g^*,v^*), in proposition 7.3.4 below a *restricted optimality result* is derived.

PROPOSITION 7.3.4. *Let* (f^*,h^*) *be a policy pair which satisfies the optimality equations* (7.3.3) *and* (7.3.4) *for some solution pair* (g^*,v^*). *Then* $g(f^*,h^*) \leq g(f^*,h)$ *for all policies* h, *having the property:*

(7.3.15) $g_i^* = [P(f^*,h)g^*]_i \Rightarrow$ h(i) is an *optimal* action in the matrix game in (7.3.3) with $g = g^*$; $i \in \Omega$

with the same restricted optimality result holding for player 1, *when player* 2 *ties himself down to policy* h^*.

PROOF. Fix a policy h which satisfies (7.3.15). Recall from the proof of part (a) of th.7.3.2 that g^* is constant on each of the subchains of $P(f^*,h)$ and conclude that:

(1) $g_i^* \leq [P(f^*,h)g^*]_i$, with

(2) $v_i^* \leq q(f^*,h)_i - g_i^* T(f^*,h)_i + [P(f^*,h)v^*]_i$, for all states $i \in R(f^*,h)$ for which (1) holds with strict equality. Apply the proof of lemma 4, part (a) in [23] to verify that $g^* \leq g(f^*,h)$, with strict equality holding for $h = h^*$. □

Let in example 1, $f^k(h^k)$, $k = 1,2$ be the pure policy for player 1 (2) which has $1 = f_{1k}^k = (h_{1k}^k)$. Note that both for $p = \frac{1}{2}$ and $p = \frac{3}{4}$, (f^1,h^1) satisfies the restricted optimality result of prop.7.3.4, but fails to be an AEP, since $0 = g(f^1,h^2)_1 < g(f^1,h^1)_1 = 1$. Observe that h^2 satisfies $g^* = P(f^1,h^2)g^*$ but $h^2(1)$ is *not* an *optimal* action in the matrix game in (7.3.3).

Finally note that, whereas a stationary AEP does not need to satisfy *both* optimality equations (7.3.3) and (7.3.4) for any solution pair (g^*,v^*) (cf. example 1 with $p = \frac{1}{2}$), it will certainly have to satisfy the first f.e. for $g^* = a^{(L)}$.

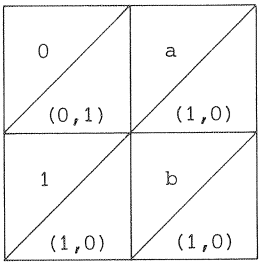
REMARK 1. In ordinary MRP's, a policy f, in order to be maximal gain, needs to satisfy the *second* optimality equation (7.3.4) *only* in its *recurrent* states (cf. lemma 1.4.2). In the general SDG or SRG model however, we could not weaken the prerequisite in proposition 7.3.4, to the assumption:

(a) $(f^*(i), h^*(i))$ is an equilibrium pair of actions in the matrix game in (7.3.3) for *every* $i \in \Omega$

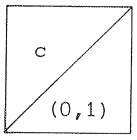
(b) $(f^*(i), h^*(i))$ is an equilibrium pair of actions in the matrix game in (7.3.4) for *every* $i \in R(f^*,h^*)$

even when confining ourselves to the *restricted* optimality result in prop. 7.3.4, as is illustrated by example 2:

EXAMPLE 2. Consider the *SDG*-model:



state 1



state 2

Let $f^k(h^k)$; k = 1,2 be defined as above. Take a = 0, b = 2, c = 1. Note that (f^2,h^1) is an AEP such that $a^{(L)} = [1,1]$ and verify that in this example $(a^{(L)};[0,1])$ is a solution pair to (7.3.3) and (7.3.4). Note that (f^1,h^1) satisfies (7.3.3) in every $i \in \Omega$, and (7.3.4) in every $i \in R(f^1,h^1) = \{2\}$; however $0 = g(f^1,h^2)_1 < g(f^1,h^1)_1 = 1$ in spite of h^2 satisfying condition (7.3.15) in proposition 7.3.4.

Prop.7.3.4 makes clear that a policy pair which satisfies (7.3.3) and (7.3.4) may fail to be an AEP, only when one of the sets $K(i,g^*)$ or $M(i,g^*)$ $(i \in \Omega)$ is a *strict* subset of $K(i)$ or $M(i)$. As a consequence no problems arise when the asymptotic average value is independent of the initial state of the system:

COROLLARY 7.3.5. *Assume* $a_i^{(L)} = \langle a^{(L)} \rangle$ *for all* $i \in \Omega$. *Then the following statements are equivalent:*

(I) $a^{(k)} = 0$, *for* $k = 1,\ldots,L-1$

(II) *there exists a stationary* AEP

(III) *the functional equations* (7.3.3) *and* (7.3.4) *have a solution pair* $(a^{(L)},v^*)$.

In addition, under either one of (I), (II) *or* (III), *any policy pair which satisfies the funct. eq.* (7.3.4) *and* (7.3.5) *for some solution pair* $(a^{(L)},v^*)$ *is an* AEP.

PROOF. We shall prove (II) ⇒ (I) ⇒ (III) ⇒ (II). By theorem 7.3.2 part (b), we have (II) ⇒ (I). Use (7.2.3), (7.3.6) and $a^{(L)} = \langle a^{(L)} \rangle \underline{1}$, to verify

(7.3.16) $\tilde{\rho}_i^{k,\ell}(r) = \rho_i^{k,\ell}(r) + \frac{\langle a^{(L)} \rangle}{r} \{\sum_j m_{ij}^{k,\ell}(r) - 1\}$; $i \in \Omega$, $k \in K(i)$, $\ell \in M(i)$

and

(7.3.17) $\tilde{q}_i^{k,\ell} = q_i^{k,\ell} - \langle a^{(L)} \rangle T_i^{k,\ell}$; $i \in \Omega$; $k \in K(i)$, $\ell \in M(i)$.

Subtract $\langle a^{(L)} \rangle / r$ from both sides of (7.2.4), to obtain

$$V(r)_i - \frac{\langle a^{(L)} \rangle}{r} = \text{val}[\tilde{\rho}_i^{k,\ell}(r) + \sum_j m_{ij}^{k,\ell}(r)\{V(r)_j - \frac{\langle a^{(L)} \rangle}{r}\}], \quad i \in \Omega$$

and conclude that $\tilde{V}(r) = V(r) - \frac{a^{(L)}}{r}$, satisfying

(7.3.18) $\tilde{V}(r)_i = \text{val}[\tilde{\rho}_i^{k,\ell}(r) + \sum_j m_{ij}^{k,\ell}(r)\, \tilde{V}(r)_j]$, $i \in \Omega$.

It follows from (7.3.9) that $\tilde{V}(r)$ is continuous in $r = 0$. Using this, $\lim_{r \downarrow 0} \tilde{\rho}_i^{k,\ell}(r) = \tilde{q}_i^{k,\ell}$ and $\lim_{r \downarrow 0} m_{ij}^{k,\ell}(r) = P_{ij}^{k,\ell}$ as well as (7.3.17), we obtain that $[a^{(L)}, \tilde{a}^{(0)}]$ satisfy (7.3.3) and (7.3.4) by letting r tend to zero in (7.3.18).

(III) ⇒ (II): follows from prop.7.3.4, by taking any pair of policies which satisfies the funct. e.q. (7.3.3) and (7.3.4) for $(a^{(L)},v^*)$, thus proving the last assertion of the corollary, at one blow. □

REMARK 2. The implication (III) ⇒ (II) even holds for a denumerable state space (cf. e.g. th.6.4.2). Observe that when the asymptotic average value does depend upon the initial state of the system, (I) and (II) do not need to be equivalent, i.e. (I) may fail to imply (II); as an example take the Big Match (cf. [47]) which has even a Laurent series expansion for V(r), i.e. which has L = 1 (cf. [7], section 8).

7.4. SOME PROPERTIES OF THE SOLUTION SPACE OF THE OPTIMALITY EQUATIONS

In this section, we discuss a number of properties of the functional equations (7.3.3) and (7.3.4) which we will need in the following section. We first observe that in general (7.3.3) and (7.3.4) may fail to have a

solution pair, just like there may fail to be (stationary) AEPs. As an example, take ex.2 with a = 1, b = c = 0, which appeared first in STERN [119] and was used in BEWLEY and KOHLBERG ([7], sect. 11). Note from [7], that this example has $\underline{0}$ as its asymptotic average value vector, but has no stationary AEP and apply cor.7.3.5 (or alternatively note that in this example L = 2, and $a^{(1)} = [1,0]$; and apply cor.7.3.5).

Next, whenever a solution pair (g^*,v^*) exists to the optimality equations (7.3.3) and (7.3.4), the v^*-part of the solution is obviously not uniquely determined (note e.g. that if (g^*,v^*) is a solution pair, then so is $(g^*,v^*+c\underline{1})$ for any scalar c; cf. also [109], where a complete characterization of the solution space was given, for the case of ordinary MRPs). In addition, since a pair of policies (f^*,h^*) which satisfies the optimality equations, does need to be an AEP, it is still unclear to us whether the g^*-part is always uniquely determined. All of the above difficulties arise, in view of the chain structure being discontinuous on $\Phi \times \Psi$ in the general multichain case. Theorem 7.4.1 below gives a number of characterizations with respect to the optimality equations, under condition (U). Since in the following section, optimality equations of a slightly more general structure will appear, we formulate and derive our results with respect to the f.e.:

$$(7.4.1) \qquad x_i = \mathrm{val}_{[\hat{K}(i),\hat{M}(i)]}[a_i^{k,\ell} + \textstyle\sum_j p_{ij}^{k,\ell}\, x_j], \quad i \in \Omega$$

$$(7.4.2) \qquad y_i = \mathrm{val}_{[\tilde{K}(i,x^*),\tilde{M}(i,x^*)]}[c_i^{k,\ell} - \textstyle\sum_j H_{ij}^{k,\ell}\, x_j + \sum_j p_{ij}^{k,\ell}\, y_j], \quad i \in \Omega$$

where for each $i \in \Omega$, $\hat{K}(i)$ and $\hat{M}(i)$ are closed convex polyhedral subsets of $K(i)$ and $M(i)$, and where for each solution x^* to (7.4.1), $\tilde{K}(i,x^*)$ and $\tilde{M}(i,x^*)$ are the sets of optimal actions in the matrix games in (7.4.1) with $x = x^*$; $a_i^{k,\ell}$ and $c_i^{k,\ell}$ are given quantities ($i \in \Omega$, $k \in K(i)$, $\ell \in M(i)$).

THEOREM 7.4.1.

(a) (7.4.1) *has a solution* x^*, *if and only if*

(7.4.3) *the* SDG *with* $\tilde{q}_i^{k,\ell} = a_i^{k,\ell}$ *has a stationary* AEP, *and* $\underline{0}$ *as its asymptotic average value vector*

(b) *Assume condition* (U) *to be satisfied. Then if* (7.4.3) *holds:*

(1) *The solution* x^* *to* (7.4.1) *is unique up to a multiple of* $\underline{1}$, *such that the sets* $\tilde{K}(i) = K(i,x^*)$ *and* $\tilde{M}(i) = M(i,x^*)$, $i \in \Omega$, *are uniquely determined.*

(2) *A solution* (x^*,y^*) *to* (7.4.1) *and* (7.4.2) *exists, where* x^* *is uniquely determined by:*

$$(7.4.4.)\quad x_i^* = x_i^o + \max_{f\in X_i\tilde{K}(i)} \min_{h\in X_i\tilde{M}(i)} \frac{<\pi^1(f,h),c(f,h)-H(f,h)x^o>}{<\pi^1(f,h),T(f,h)>}$$

$$= x_i^o + \min_{h\in X_i\tilde{M}(i)} \max_{f\in X_i\tilde{K}(i)} \frac{<\pi^1(f,h),c(f,h)-H(f,h)x^o>}{<\pi^1(f,h),T(f,h)>}$$

where x^o *denotes some solution to* (7.4.1). *Moreover,* y^* *is unique up to a multiple of* $\underline{1}$.

PROOF.

(a) immediate from corollary 7.3.5

(b) (1): Let x^o,x^1 be two solutions to (7.4.1) and let (f^o,h^o) and (f^1,h^1) be two pairs of policies which satisfy (7.4.1) for x^o and x^1 resp. Note that: $x^o \le a(f^o,h^1) + P(f^o,h^1)\,x^o$ and $x^1 \ge a(f^o,h^1) + P(f^o,h^1)\,x^1$ and subtract the second inequality from the first one, in order to obtain: $x^o - x^1 \le P(f^o,h^1)\,[x^o-x^1]$, and by iterating the latter:

$$(7.4.5)\quad [x^o,x^1]_i \le c_1 = <\pi^1(f^o,h^1),\, x^o - x^1>,\ i \in \Omega.$$

Similarly, we obtain

$$<\pi^1(f^1,h^o),\, x^o - x^1> = c_2 \le [x^o,x^1]_i,\ i \in \Omega.$$

We finally show $c_1 = c_2$, which proves part (a). Multiply both sides of (7.4.5) by $\pi^1(f^o,h^1)$ in order to conclude that $x_i^o = x_i^1 = c_1$, for all $i \in R(f^o,h^1)$. Similarly we obtain $x_i^o - x_i^1 = c_2$ for all $i \in R(f^1,h^o)$ which implies $c_1=c_2=c$, as a consequence of $R(f^1,h^o) \cap R(f^o,h^1) \neq \emptyset$, in view of assumption (U).

(2): Fix a solution x^o to (7.4.1), and consider the SRG, which has Ω as its state space, $\tilde{K}(i)$ and $\tilde{M}(i)$ as the sets of (randomized) alternatives available to player 1 and 2 and with one-step expected rewards $\tilde{q}_i^{k,\ell} = c_i^{k,\ell} - \sum_j H_{ij}^{k,\ell}\, x_j^o$ and unaltered transition probabilities and transition time distributions. Note from lemma 7.2.1 that each of the sets $\tilde{K}(i)$ and $\tilde{M}(i)$ may be considered as the set of randomizations of a finite number of (pure) alternatives. This, in combination with condition (U), implies as a result of lemma 7.3.1, and cor.7.3.5 the existence of a solution to the f.e.:

$$y_i^* = \mathrm{val}_{[\tilde{K}(i),\tilde{M}(i)]}[c_i^{k,\ell} - \textstyle\sum_j H_{ij}^{k,\ell} x_j^o - g_i^o T_i^{k,\ell} + \sum_j P_{ij}^{k,\ell} y_j^*],\ i \in \Omega$$

where g^o is the asymptotic average value vector in this stochastic game with $g^o_i = \langle g^o \rangle$, $i \in \Omega$ in view of our unichainedness-assumption. This implies that $x^* = x^o + g^o$ is a solution to (7.4.1), thus showing the existence of a solution pair to (7.4.1) and (7.4.2). We next show that the x^*-part is uniquely determined, and derive its explicit expression. The fact that the y^*-part is unique up to a multiple of $\underline{1}$ follows as in part (b) (1). Let (x^*,y^*) be a solution to (7.4.1) and (7.4.2), and let (f^*,h^*) be a policy pair which satisfies (7.4.1) and (7.4.2) for (x^*,y^*). Let $g^o = x^* - x^o$. Then for each $h \in X_i \tilde{M}(i)$:

$$y^* \geq c(f^*,h) - H(f^*,h)\, x^o - \langle g^o \rangle\, T(f^*,h) + P(f^*,h)y^*$$

Multiply this inequality by $\Pi(f^*,h)$ and conclude that:

$$(7.4.6) \qquad g^o \geq \frac{\langle \pi^1(f^*,h), c(f^*,h) - H(f^*,h)x^o \rangle}{\langle \pi^1(f^*,h), T(f^*,h) \rangle}$$

with strict equality holding for $h = h^*$. Likewise, one can show that:

$$g^o \leq \langle \pi^1(f,h^*),\ c(f,h^*) - H(f,h^*)x^o \rangle / \langle \pi^1(f,h^*), T(f,h^*) \rangle$$

for all $f \in X_i \tilde{K}(i)$, with strict equality for $f = f^*$, thus completing the proof of part (b). □

7.5. SENSITIVE DISCOUNT AND CUMULATIVE AVERAGE OPTIMALITY

In this section, we consider a sequence of increasingly selective optimality criteria, which appears as the natural extension to the SRG-model of the sensitive discount (or cumulative average) optimality criteria, as formulated and studied in Markov Decision Theory (cf. e.g. [22],[85],[127]).

We call a policy pair (φ^*,ψ^*) a n-discount equilibrium pair of policies *(n-EP)* $(n = -1,0,\ldots)$, if:

$$(7.5.1) \qquad \limsup_{r \downarrow 0} r^{-n}[V(\varphi^*,\psi^*)(r) - V(\varphi^*,\psi)(r)] \leq 0 \leq$$

$$\liminf_{r \downarrow 0} r^{-n}[V(\varphi^*,\psi^*)(r) - V(\varphi,\psi^*)(r)]$$

where $V(\varphi,\psi)(r)$ denotes the total discounted return to player 1, when the players use policies φ,ψ and when the rewards are discounted at rate r.

We restrict our analysis to the sensitive discount criteria for the

discrete-time case of SDGs, in order to avoid too burdensome a notation. The extension of our results to the general SRG-case, is immediate and the analysis of the cumulative average optimality criteria is analogous to the one given below, with the same sequence of f.e. associated (note that for n = -1, equivalence of the two criteria was proven in BEWLEY and KOHLBERG [7]). E.g. whereas in the general SRG-model the expressions in the various f.e., to be considered below, become more complicated functions of the terms in the expansions of $\rho_i^{k,\ell}(r)$ and $m_{ij}^{k,\ell}(r)$ (cf. DENARDO [22]), the structure of each consecutive pair of f.e. is exactly identical to the one of (7.4.1) and (7.4.2). Consider the following sequence of optimality equations:

$$(7.5.2)\quad g_i = \mathrm{val}\Big[\sum_j p_{ij}^{k,\ell}\, g_j\Big],\quad i \in \Omega$$

$$(7.5.3)\quad x(0)_i = \mathrm{val}_{[K(i,g),M(i,g)]}\Big[q_i^{k,\ell} - g_i + \sum_j p_{ij}^{k,\ell}\, x(0)_j\Big],\quad i \in \Omega$$

$$(7.5.4)\quad x(m)_i = \mathrm{val}_{[K^{(m)}(i,X(m-1)),M^{(m)}(i,X(m-1))]}\Big[-x(m-1)_i + \sum_j p_{ij}^{k,\ell}\, x(m)_j\Big],$$

$$m = 1,2,\ldots,\ i \in \Omega$$

where X(m) denotes the m+2-tuple of vectors $(g,x(0),\ldots,x(m))$, $m = 0,1,\ldots$. In addition for all $m = 1,2,\ldots$ and $i \in \Omega$ and any solution X(m-1) to the first m+1 f.e. in (7.5.2)-(7.5.4), $K^{(m)}(i,X(m-1))$ and $M^{(m)}(i,X(m-1))$ denote the sets of optimal actions in the m+1-st f.e.

For each stationary pair of policies (f,h) let:

$$(7.5.5)\quad V(f,h)(r) = r^{-1}g(f,h) + \sum_{k=0}^{\infty} x^{(k)}(f,h)r^k$$

represent the Laurent series expansion of the total discounted return associated with (f,h). Finally, if x is a vector, we say x is *lexicographically non-negative*, written $x \succeq 0$, if the first nonvanishing element of x is positive. Similarly, x is called *lexicographically positive*, written $x \succ 0$ if $x \succeq 0$ and $x \neq 0$. We write $x \succeq (\succ) y$ or $y \preceq (\prec) x$ if $x-y \succeq (\succ) 0$.

THEOREM 7.5.1.

(a) *Let* (f^*,h^*) *be a stationary* n-EP ($n = -1,0,\ldots$). *Then*

(1) *There exists a* n+3*-tuple* $(g^*,x(0),\ldots,x(n+1))$ *which satisfies* (7.5.2), (7.5.3) *and the first* n+1 f.e. *of* (7.5.4).

(2) *In the Puiseux series expansion of* V(r), *we have:*

$$(7.5.6)\qquad a^{(-\ell L)} = \begin{cases} g(f^*,h^*), & \text{for } \ell = -1 \\ x^{(\ell)}(f^*,h^*), & \text{for } \ell = 0,\ldots,n \end{cases}$$

$$a^{(-\ell L-p)} = 0, \text{ for } \ell = -1,\ldots,n;\quad p = 1,\ldots,L-1$$

(b) *Let* (f^*,h^*) *be a stationary* N-EP. *Then*

(1) (f^*,h^*) *is a* n-EP *for all* $n \geq N$

(2) V(r) *has a Laurent series expansion.*

PROOF.

(a)(1) For $n = -1$ the assertion holds as a consequence of th.7.3.3; hence we assume $n \geq 0$. Note that $h^*(f^*)$ is a n-optimal policy in the MDP which results for player 2 (1) when player 1 (2) ties himself down to policy $f^*(h^*)$. Use th.4 of [127] to conclude that for all $i \in \Omega$, $f \in \Phi$, $h \in \Psi$:

$$(7.5.7)\qquad [\textstyle\sum_j P(f,h^*)_{ij} g(f^*,h^*)_j ; q(f,h^*)_i - g(f^*,h^*)_i + \sum_j P(f,h^*)_{ij} x^{(0)}(f^*,h^*)_j$$

$$;\ldots; -x^{(n-1)}(f^*,h^*)_i + \textstyle\sum_j P(f,h^*)_{ij} x^{(n)}(f^*,h^*)_j] \leq$$

$$[g(f^*,h^*)_i ; x^{(0)}(f^*,h^*)_i ;\ldots; x^{(n)}(f^*,h^*)_i] \leq$$

$$[\textstyle\sum_j P(f^*,h)_{ij} g(f^*,h^*)_j ; q(f^*,h)_i - g(f^*,h^*)_i + \sum_j P(f^*,h)_{ij} x^{(0)}(f^*,h^*)_j$$

$$;\ldots; -x^{(n-1)}(f^*,h^*)_i + \textstyle\sum_j P(f^*,h)_{ij} x^{(n)}(f^*,h^*)_j]$$

with strict equality holding for $h = h^*$, and $f = f^*$. One easily concludes that $X(n) = [g(f^*,h^*);\ldots;x^{(n)}(f^*,h^*)]$ satisfy (7.5.2), (7.5.3) and the first n f.e. in (7.5.4). To prove that there exists a solution to the n+1-st f.e. in (7.5.4) as well, note that (f^*,h^*) is an AEP in the stochastic game, which has Ω as its state space, $K^{(n+1)}(i,X(n))$ and $M^{(n+1)}(i,X(n))$ as the action spaces in state $i \in \Omega$, and with one-step expected reward vectors $\tilde{q}(f,h) = -x^{(n)}(f^*,h^*)$, and transition probabilities $\tilde{P}^{k,\ell}_{ij} = P^{k,\ell}_{ij}$. (cf. e.g. DENARDO [22], p.491). Finally note that this stochastic game has $\underline{0}$ as its asymptotic average vector, and apply th.7.3.3.

(2) We prove part (a)(2) by complete induction with respect to ℓ. Note that for $\ell = -1$, the equalities $a^{(-\ell L)} = g(f^*,h^*)$ and $a^{(-\ell L-p)} = 0$ for

$p = 1,\ldots,L-1$ follow from the fact that (f^*,h^*) is a stationary AEP, using theorem 7.3.2. Assume that (7.5.6) holds for all $\ell = -1,\ldots,\ell^*<n$. Let

$$\sum_{k=-\infty}^{1} A^{(k)} r^{-k} [\sum_{k=-\infty}^{1} B^{(k)} r^{-k}]$$

be the Laurent series expansion of $W^1(r)[W^2(r)]$, the total discounted return to player 1 [2] in the MDP that results when player 2 [1] ties himself down to policy $h^*[f^*]$ and note that

$$(7.5.8) \quad A^{(1)} = B^{(1)} = g(f^*,h^*) \text{ and } A^{(-k)} = B^{(-k)} = x^{(k)}(f^*,h^*) \quad \text{for all } k = 0,\ldots,\ell^*+1$$

In view of $f^*[h^*]$ being ℓ^*+1- optimal in this MDP. Observe that, $W^2(r) \leq V(r) \leq W^1(r)$, and conclude from (7.5.8) and the induction assumption that the coefficients of the terms with power strictly less than ℓ^*+1, in $W^1(r)$, $V(r)$ and $W^2(r)$ coincide. Since $A^{(-\ell^*-1)} = B^{(-\ell^*-1)}$ we conclude that (7.5.6) holds for $\ell = \ell^*+1$ as well.

(b) (1): It follows from VEINOTT ([127], p.1646) that $f^*[h^*]$, since being N-optimal, is n-optimal for all $n \geq N$ in the MDP that results when player 2 [1] ties himself down to policy $h^*[f^*]$.

(b) (2): Immediate from (a) (2) and (b) (1). □

We observe that part (b) of th.7.5.1 may not be extended to the general SRG-model, since it does not even hold in the general MRP-case (cf. [22], p.489). However, part (b) generalizes proposition 6.4 in [7], where it was shown that $V(r)$ has a Laurent series expansion, if there exists a uniformly *discount optimal* pair of policies, i.e. a pair (f^*,h^*) which is optimal in the r-discount game for all r sufficiently small. We next observe, that whereas the existence of a solution to the first n+1 f.e. in (7.5.2)-(7.5.4) is a *necessary* condition for the existence of a stationary n-EP, it certainly may fail to be sufficient, as was pointed out for the case $n = -1$, in section 3.

In analogy to prop.7.3.4, the following *partial* optimality result may be obtained for any policy pair which satisfies (7.5.2), (7.5.3) and the first n+1 f.e. in (7.5.4) for some solution $(g^*,x^*(0),\ldots,x^*(n+1))$:

PROPOSITION 7.5.2. *Fix* $n = -1,0,\ldots$. *Let* $(g^*,x^*(0),\ldots,x^*(n+1))$ *be a solution to* (7.5.2), (7.5.3) *and the first* n+1 f.e. *of* (7.5.4), *and let*

$(f^*,h^*) \in \Phi \times \Psi$ *be a policy pair which satisfies these optimality equations for this solution. Then*

$$[g(f^*,h^*)_i;x^{(0)}(f^*,h^*)_i;\ldots;x^{(n)}(f^*,h^*)_i] \preceq [g(f^*,h)_i;\ldots;x^{(n)}(f^*,h)_i]$$

holds in every $i \in \Omega$, *for those policies* h *for which:*

$$(7.5.9)\quad \sum_j P(f^*,h)_{ij}\, g^*_j = g^*_i \Rightarrow h(i) \in M(i,g^*)$$

$$\{h(i) \in M(i,g^*) \text{ and } x^*(0)_i = q(f^*,h)_i - g^*_i + \sum_j P(f^*,h)_{ij}\, x^*(0)_j\} \Rightarrow h(i) \in M^{(1)}(i,X^*(0)),$$

$$\{h(i) \in M^{(m)}(i,X^*(m-1)) \text{ and } x^*(m)_i = -x^*(m-1)_i + \sum_j P(f^*,h)_{ij}\, x^*(m)_j\} \Rightarrow h(i) \in M^{(m+1)}(i,X^*(m)),$$

$$1 \le m \le n+1$$

with the same restricted optimality result holding for policy f^*. □

We finally turn to the case where condition (U) is satisfied:

THEOREM 7.5.3. *Assume condition* (U) *holds. Then*

(a) *there exists a solution to the entire sequence of* f.e. (7.5.2),(7.5.3) *and* (7.5.4).

(b) *Fix* $n = 0,1,\ldots$. *In the solution* $(g^*,x^*(0),\ldots,x^*(n))$ *to* (7.5.2), (7.5.3) *and the first* n f.e. *of* (7.5.4), *we have* $(g^*,x^*(0),\ldots,x^*(n-1))$ *uniquely determined (explicit expressions of which may be obtained by a repeated application of* th.7.4.1), *whereas* $x^*(n)$ *is unique up to a multiple of* $\underline{1}$.

PROOF. Part (a) follows from part (b), and part (b) is proven by complete induction with respect to n. Note that for $n = 0$, the assertion follows as a special case of th.7.4.1. Assume, it holds for some $n = 0,1,\ldots$. We then have in particular that $x^*(n-1)$ (or g^* when $n = 0$) is uniquely determined and that the f.e.:

$$(7.5.10)\quad x(n)_i = \mathrm{val}_{[K^{(n)}(i,X(n-1));M^{(n)}(i,X(n-1))]}[-x(n-1)_i + \sum_j p^{k,\ell}_{ij}\, x(n)_j],$$

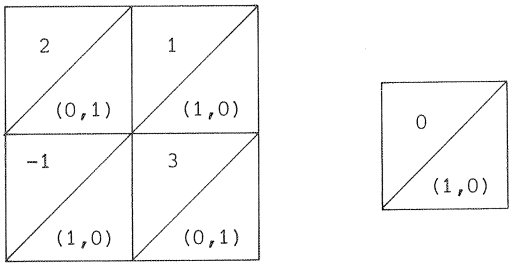
$$i \in \Omega$$

(or (7.5.3) in case $n = 0$) has a solution. Apply th.7.4.1 to the combination

of (7.5.10) (or (7.5.3) in case n = 0) and the n+1-st f.e. in (7.5.4), to verify that the assertion holds for n+1 as well. □

In SOBEL ([117], th.2), it was asserted that a stationary 1-EP always exists under condition (U). In chapter 6 we pointed out that the proof of this theorem is incorrect, and the next example shows that the asserted result itself *may fail to hold:*

EXAMPLE 3.



Note that $g(f,h) \stackrel{def}{=} r(x,y) = (5xy-2x-4y+3)/(2+2xy-x-y)$ where $x = f_{11}$ and $y = h_{11}$. Let

$$(7.5.11)\quad \phi(x) \stackrel{def}{=} \min_y r(x,y) = \min\left\{\frac{3x-1}{1+x}\ ;\ \frac{2x-3}{x-2}\right\} = \frac{3x-1}{1+x}$$

$$(7.5.12)\quad \psi(y) = \max_x r(x,y) = \max\left\{1;\ \frac{4y-3}{y-2}\right\}$$

where the second equality in (7.5.11) and (7.5.12) follows from the observation that with one of the two players tieing himself down to a specific (randomized) strategy, a *pure* strategy can be used by his rival when optimizing the resulting MDP.

Conclude that

$$(7.5.13)\quad \{1\} = \{x^* \in [0,1] \mid \phi(x^*) = \max_x \phi(x)\}, \text{ and}$$

$$[1/3,1] = \{y^* \in [0,1] \mid \psi(y^*) = \min_y \psi(y)\}$$

such that $\{f^* \mid f^*_{11} = 1\}$ and $\{h^* \mid h^*_{11} \geq 1/3\}$ are the sets of optimal (stationary) policies for players 1 and 2, with respect to the average return per unit time criterion. Note however that none of these policy pairs is 1-EP, since $x^{(0)}(f^*,h) = 0$ and $g(f^*,h) = 1$ when $h_{11} = 0$, whereas $x^{(0)}(f^*,h^*)_1 > 0$ for all policies $h^* \in \{h^* \mid h^*_{11} \geq 1/3\}$, which are gain optimal for player 2.

CHAPTER 8

Successive approximation methods in two-person zero-sum stochastic games

8.1. INTRODUCTION AND SUMMARY

In this final chapter we turn to the computational aspects of solving *SDG's* (two-person zero-sum stochastic games). Moreover, we concentrate upon the undiscounted version of the game. The discounted version is relatively easy to deal with in view of contraction mapping theory forcing itself to the front. In fact, the idea of solving the discounted version via successive approximations, goes back to SHAPLEY [115] thereby preceding the work on successive approximation methods within the restricted area of Markov Decision Theory.

The *undiscounted* version of the problem is more difficult to solve. In chapter 6, we pointed out that one or both players may fail to have optimal policies, as follows from an example in GILLETTE [47]. Moreover, within the same chapter we obtained a series of recurrency conditions with respect to the tpm's associated with the stationary policies, under which the existence of a stationary AEP is guaranteed. So far, very little attention has been paid to the actual computation of the asymptotic average value g^* and the determination of optimal policies for both players.

In view of the fact that the value of the discounted (and a fortiori, the undiscounted version of the) game does not necessarily lie within the same ordered field as the parameters of the problem (cf. BEWLEY & KOHLBERG [6]) we cannot expect to find algorithms that are finite in the sense that they involve a finite number of field operations.

So far, the literature has provided two algorithms (HOFFMAN & KARP [57] and POLLATSCHEK & AVI-ITZHAK [93]). It was shown that the first algorithm converges to a stationary AEP, if the tpm of each pure stationary policy pair is unichained and has no transient states. Although the second algorithm seems to compare favorably with the first one, as far as net running

time and the required number of iterations is concerned, it is still unknown under which conditions its convergence is guaranteed.

Observe that successive approximation methods tend to tackle much larger problems than methods that are based on Policy Iteration or mathematical programming. In this chapter we provide two successive approximation methods for locating optimal policies for both players. (The second method has also been treated quite recently by VAN DER WAL [124]). In both algorithms we obtain in addition at each step of the iteration procedure, upper and lower bounds for the asymptotic average value which converge to the latter as the number of iteration steps tends to infinity.

The first algorithm is an adaptation of the modified value-iteration method of HORDIJK and TIJMS [60], as presented in section 1.8 and as used in sections 4.3 and 7.3. Its convergence is guaranteed whenever condition (H) below is satisfied:

(H) (a) a stationary AEP exists

(b) the asymptotic average value g^* is independent of the initial state of the system.

The second algorithm is based upon the more elementary value-iteration method, and may succesfully be applied whenever condition (U) below holds:

(U) the tpm of each of the pure stationary policy pairs is unichained.

We recall that (U) $\Rightarrow$ (H) (cf. e.g. section 6.4). Corollary 7.3.2 showed us that under (H) the solution of the game reduces completely to the problem of finding a solution to a single (vector)-functional equation (cf. (7.3.13)). Under (U) this optimality equation has a solution v^*, which is unique up to a multiple of $\underline{1}$ (cf. theorem 7.4.1). Thus putting $v_N^* = 0$, we obtain in this case, lower and upper bounds for the components of v^* as well. These bounds generalize the ones obtained under (U) in the one player case (cf. corollary 2.3.2).

At each step of the procedure, both methods merely require the solution of N relatively small Linear Programs (the size of which is determined by the number of actions in K(i) and M(i), $i \in \Omega$). Especially for large scale systems, i.e. when $N >> 1$, this compares favorably with the techniques used in [57] and [93] which require at each step of the procedure, the solution of a system of at least N equations.

One might wish to extend these methods to the general SRG-case, as discussed in chapter 7. In the one-player case, this extension was possible

due to the data-transformations (1.9.9) and (1.9.10) which turn every undiscounted MRP into an equivalent MDP. In the two-player case, the analogue of this data-transformation, will generally fail to work. The only exception is provided by the case where the expected holding times $T_i^{k,\ell}$ ($i \in \Omega$, $k \in K(i)$, $\ell \in M(i)$) satisfy the *separability* assumption:

(SEP) (8.1.1) $T_i^{k,\ell} = \alpha_i^k \beta_i^\ell$, with α_i^k, $\beta_i^\ell > 0$; $i \in \Omega$, $k \in K(i)$, $\ell \in M(i)$.

This will be shown in the appendix, section 8.4. (8.1.1) holds e.g. in case the transition time between two successive observations of the state of the system, merely depends upon the current state, possibly in combination with the action chosen by one of the two players. Establishing an algorithm for the *general* SRG-case remains however, an outstanding problem.

In section 2 and 3 we present our two methods. Throughout this chapter we use the notation and preliminary results as given in section 7.2. Finally, the results in this chapter have been distilled from FEDERGRUEN [32].

8.2. A MODIFIED VALUE-ITERATION TECHNIQUE

Throughout this section, we assume (H) to hold, which implies in view of corollary 7.3.5 the existence of a solution pair (g^*, v^*) to the optimality equation (7.3.13):

$$(8.2.1) \qquad v_i = \mathrm{val}[q_i^{k,\ell} - g^* + \textstyle\sum_j p_{ij}^{k,\ell} v_j], \quad i \in \Omega$$

where g^* denotes the asymptotic average value of the game. Define V as the set of solutions to (8.2.1). Recall from (7.3.9) that V(r), as defined by (7.2.4), has for some integer $L \geq 1$, a Puiseux series expansion of the special type:

$$(8.2.2) \qquad V(r) = g^*/r + \textstyle\sum_{k=-\infty}^{0} a^{(k)} r^{-k/L}, \quad \text{for all } r \text{ sufficiently close to } 0.$$

Applying the proof of th.7.3.3 to the SDG-case, we conclude that any scheme

$$(8.2.3) \qquad y(n+1)_i = \mathrm{val}[q_i^{k,\ell} - g^* + (1+r_n)^{-1} \textstyle\sum_j p_{ij}^{k,\ell} y(n)_j], \quad i \in \Omega,$$

with $y(0)$ a given N-vector, has $\lim_{n\to\infty} y(n) = a^{(0)}$, provided that the sequence $\{r_n\}_{n=1}^{\infty}$ satisfies the conditions:

(8.2.4) $\quad (1-r_n) \ldots (1-r_2) \to 0, \quad$ as $n \to \infty$

(8.2.5) $\quad \sum_{j=2}^{n} (1-r_n) \ldots (1-r_j)\left|r_j^{1/L} - r_{j-1}^{1/L}\right| \to 0, \quad$ as $n \to \infty$

where in addition $a^{(0)}$ is a solution to the optimality equation (8.2.1).

LEMMA 8.2.1. *Conditions* (8.2.4) *and* (8.2.5) *are satisfied for any choice:*

$$r_n = n^{-b} \qquad \textit{with} \qquad 0 < b \leq 1.$$

PROOF. Note using the mean value theorem that $n^b - (n-1)^b \leq 1$ for all $n = 1,2,\ldots$ and use this inequality in order to verify that:

$$(1-r_n) \ldots (1-r_2) = \frac{(n^b-1)}{n^b} \frac{((n-1)^b-1)}{(n-1)^b} \ldots \frac{(2^b-1)}{2^b} \leq \frac{2^b-1}{n^b}$$

which proves (8.2.4). Next we apply the mean value theorem to verify that

$$\sum_{j=3}^{n} (1-r_n) \ldots (1-r_j)\left|r_j^{1/L} - r_{j-1}^{1/L}\right| \leq$$

$$bL^{-1} \sum_{j=3}^{n} \frac{(n^b-1)}{n^b} \ldots \frac{(j^b-1)}{j^b} (j-1)^{-b/L-1} \leq bL^{-1}n^{-b} \sum_{j=2}^{n} j^{b(1-1/L)-1} \leq$$

$$bL^{-1}n^{-b} \int_1^n x^{b(1-1/L)-1}\, dx \leq$$

$$\begin{cases} bn^{-b}\, \ell n(n), & \text{if } L = 1 \\ (L-1)^{-1}n^{-b/L}, & \text{otherwise} \end{cases}$$

which proves (8.2.5). □

REMARK 1. For the MDP- i.e. one player - case, lemma 8.2.1 indicates a larger range of permitted values for b, than the one that was obtained in [60] (p. 206, remark) using a different analysis.

Observe that the sequence $\{y(n)\}_{n=1}^{\infty}$ cannot be computed in view of g^* being unknown. We circumvent this numerical difficulty as in WHITE [131], i.e. we define the sequences $\{\hat{y}(n)\}_{n=1}^{\infty}$ and $\{G(n)\}_{n=1}^{\infty}$ by:

(8.2.6) $\quad \hat{y}(n+1)_i = y(n+1)_i - y(n+1)_N = \text{val}[q_i^{k,\ell} + (1+r_n)^{-1} \sum_j p_{ij}^{k,\ell} \hat{y}(n)_j]$

$$- G(n+1); \quad i \in \Omega; \quad n = 0,1,2,\ldots.$$

(8.2.7) $G(n+1) = \text{val}[q_N^{k,\ell} + (1+r_n)^{-1} \sum_j p_{Nj}^{k,\ell} \hat{y}(n)_j]$; $i \in \Omega$; $n = 0,1,2,\ldots$

where $\hat{y}(0)_i = y(0)_i - y(0)_N$; $i \in \Omega$.

THEOREM 8.2.2. *For all* $n = 1,2,\ldots$ *let*

(8.2.8) $L(n+1) = \min_i\{\hat{y}(n+1)_i + G(n+1) - (1+r_n)^{-1}\hat{y}(n)_i\}$

$U(n+1) = \max_i\{\hat{y}(n+1)_i + G(n+1) - (1+r_n)^{-1}\hat{y}(n)_i\}$

(a) *Let* (f^*,h^*) *be a stationary AEP and for any* $n = 1,2,\ldots$ *let* $(f_n,h_n) \in \Phi \times \Psi$ *be any pair of policies which attain the* N *equilibria to the right of* (8.2.6) *simultaneously. Then*

(1) $L(n) \le G(n) \le U(n)$ $n = 1,2,\ldots$

(2) $L(n+1) \le g(f_n,h^*)_i \le g^* \le g(f^*,h_n)_i \le U(n+1)$; $i \in \Omega$

(b) *If* $\{r_n\}_{n=1}^{\infty}$ *satisfies* (8.2.4) *and* (8.2.5), *then*

$\lim_{n\to\infty} L(n) = \lim_{n\to\infty} G(n) = \lim_{n\to\infty} U(n) = g^*$

$\lim_{n\to\infty} \hat{y}(n) = a^{(0)} - \langle a_N^{(0)}\rangle \underline{1}$, *satisfying* (8.2.1).

PROOF.

(a) (1) Note from (8.2.6) that $\hat{y}(n)_N = 0$ for all $n = 0,1,2,\ldots$, hence

$$L(n) \le \text{val}[q_N^{k,\ell} + (1+r_n)^{-1} \sum_j p_{Nj}^{k,\ell} \hat{y}(n)_j] = G(n) \le U(n)$$

(2) The inner inequalities are immediate from the fact that (f^*,h^*) is a stationary AEP. We next prove the left most inequality $L(n+1) \le g(f_n,h^*)$, the proof of $g(f^*,h_n) \le U(n+1)$ being analogous. Note that for all $j \in \Omega$:

$$L(n+1) + (1+r_n)^{-1}\hat{y}(n)_j \le \text{val}[q_j^{k,\ell} + (1+r_n)^{-1} \sum_r p_{jr}^{k,\ell}\hat{y}(n)_r] \le$$

$$\le q(f_n,h^*)_j + (1+r_n)^{-1}[P(f_n,h^*)\hat{y}(n)]_j,$$

and multiply both sides of this inequality by $\Pi(f_n,h^*)_{ij} \ge 0$ and sum on $j \in \Omega$.

(b) Recall that $\lim_{n\to\infty} y(n) = a^{(0)} \in V$. Next we observe that if $v \in V$ then so is $v + c\underline{1}$ for all scalars c. Hence,

$$\lim_{n\to\infty}\hat{y}(n) = \lim_{n\to\infty} y(n) - \langle y(n)_N\rangle \underline{1} = a^{(0)} - \langle a^{(0)}_N\rangle \underline{1} \in V.$$

This in combination with the fact that the "val"-operator is (Lipschitz) continuous (cf. (7.2.1)) imply:

$$\lim_{n\to\infty} L(n) = \min_i \{val[q_i^{k,\ell} + \sum_j p_{ij}^{k,\ell}\, a_j^{(0)}] - a_i^{(0)}\} = \min_i g^* =$$

$$g^* = \max_i \{val[q_i^{k,\ell} + \sum_j p_{ij}^{k,\ell}\, a_j^{(0)}] - a_i^{(0)}\} = \lim_{n\to\infty} U(n)$$

which together with part (a)(1) completes the proof of (b). □

REMARK 2. When taking $r_n = n^{-b}$ for some b, with $0 < b \leq 1$, the approach to the limits in part (b) of the above theorem, exhibits a convergence rate which is of the order

$$\begin{cases} O(n^{-b}\ell n\, n), & \text{if } L = 1 \\ O(n^{-b/L}), & \text{otherwise} \end{cases}$$

as follows from the proof of lemma 8.2.1 and theorem 7.3.3. We note that the bounds for g^* in part (a)(2) generalize the bounds ODONI [89] and HASTINGS [53] obtained for the MDP-case.

We summarize this section by specifying an algorithm which approximates g^*, as well as a solution $v \in V$, and which finds for any $\varepsilon > 0$, ε-optimal policies for both players:

ALGORITHM 1.

step 0: Fix a sequence $\{r_n\}_{n=1}^{\infty}$ satisfying (8.2.4) and (8.2.5), e.g. take $r_n = n^{-b}$, with $0 < b \leq 1$. Set $n = 0$, fix $y(0) \in E^N$ and $\varepsilon > 0$.

step 1: Calculate $\hat{y}(n+1)$, $L(n+1)$, $G(n+1)$, and $U(n+1)$ from $\hat{y}(n)$ using (8.2.6), (8.2.7) and (8.2.8).

step 2: If $U(n+1) - L(n+1) < \varepsilon$, determine a stationary policy pair (f_n, h_n) which attains the N equilibria to the right of (8.2.6) simultaneously, use $f_n(h_n)$ as an ε-optimal policy for player 1(2); $G(n+1)$ as an ε-approximation for g^* and $\hat{y}(n+1)$ as an approximation for a solution $v \in V$. Otherwise, increment n by one and return to step 1.

8.3. VALUE-ITERATION; A SUFFICIENT CONDITION FOR CONVERGENCE

In this section we discuss the asymptotic behaviour of the sequence:

(8.3.1) $\quad v(n+1)_i = Qv(n)_i, \quad i \in \Omega$

where the operator $Q: E^N \to E^N$ is defined by $Qx_i = \text{val}[q_i^{k,\ell} + \sum_j p_{ij}^{k,\ell} x_j]$, $i \in \Omega$, and where $v(0) \in E^N$ is a given N-vector.

Note that the Q operator is monotonic, and satisfies the basic properties:

(8.3.2) $\quad Q(x+c\underline{1}) = Qx + c\underline{1}$ for all scalars c; $x \in E^N$

(8.3.3) $\quad (x-y)_{\min} \leq (Qx-Qy)_{\min} \leq (Qx-Qy)_{\max} \leq (x-y)_{\max}; \quad x,y \in E^N$

where (8.3.3) is easily verified by applying the Q-operator to both sides of the inequalities $y + (x-y)_{\min}\underline{1} \leq x$ and $x \leq y + (x-y)_{\max}\underline{1}$, using its monotonicity as well as (8.3.2).

Note that $v(n)_i$ may be interpreted as the value of the n-stage game when starting in state i and given some final amount $v(0)_j$ is earned by player 1 from player 2, when ending up in state j. Whereas we still have $\lim_{n\to\infty} \frac{v(n)}{n} = g^*$ (cf. BEWLEY & KOHLBERG [6], th.3.2) the difference $\{v(n) - ng^*\}_{n=1}^{\infty}$ does not need to be bounded, in sharp contrast to the behaviour in the one player - model (cf. BROWN [13], th.4.3).

In fact, BEWLEY & KOHLBERG [8] proved the existence of a number $B > 0$ and a Puiseux series in n,

$$W(n) = ng^* + \sum_{k=-\infty}^{\hat{L}-1} b^{(k)} n^{k/\hat{L}}$$

such that $\|v(n) - W(n)\| < B \log (n+1)$, $n = 1,2,\ldots$ and they gave an example in which the distance between $\{v(n)\}_{n=1}^{\infty}$ and the field of Puiseux series is indeed of $O(\log n)$.

LEMMA 8.3.1. *$\{v(n) - ng^*\}_{n=1}^{\infty}$ is bounded under condition (H).*

PROOF. Note from corollary 7.3.5 that (H) implies the existence of a solution $v \in V$. Next use (7.2.1) in order to conclude that:

$$|v(n)_i - ng^* - v_i| = |\text{val}[q_i^{k,\ell} + \textstyle\sum_j p_{ij}^{k,\ell} v(n-1)_j]$$

$$- \text{val}[q_i^{k,\ell} + \textstyle\sum_j p_{ij}^{k,\ell} (v+(n-1)g^*)]| \leq \|v(n-1)-(n-1)g^*-v\|, \quad i \in \Omega. \ \square$$

It was pointed out in section 1.5 which deals with the *one* -player case that even in case $\{v(n)-ng^*\}_{n=1}^{\infty}$ is bounded, the sequence may fail to converge

provided that some of the tpm's of the pure stationary policy pairs happen to be periodic (in section 1.5 we recalled the necessary and sufficient condition for $\{v(n)-ng^*\}_{n=1}^{\infty}$ to converge in the MDP-case for all $v(0) \in E^N$).

In this section we first apply a data-transformation which turns our SDG into an equivalent one, and next analyze the behaviour of (8.3.1) under condition (U) which is a stronger version of (H) (cf. section 1). The data-transformation is the natural analogue of (1.8.1) and (1.8.2) with $\sigma = 1$ and has been employed in some of the proofs of chapter 6.

$$\tilde{q}_i^{k,\ell} = q_i^{k,\ell}; \quad i \in \Omega,\ k \in K(i),\ \ell \in M(i), \tag{8.3.4}$$

$$\tilde{P}_{ij}^{k,\ell} = \tau(P_{ij}^{k,\ell} - \delta_{ij}) + \delta_{ij}; \quad i,j \in \Omega;\ k \in K(i),\ \ell \in M(i) \tag{8.3.5}$$

where $0 < \tau < 1$.

Verify that $\tilde{P}_{ij}^{k,\ell} \geq 0$ and $\sum_j \tilde{P}_{ij}^{k,\ell} = 1$ and use the arguments of section 1.8 to conclude that the gain rate vector of each pair of policies in $\Phi \times \Psi$ remains unaltered by the data-transformation. In addition, each of the tpm's of the stationary policy pairs in the transformed model, has a *positive diagonal*, which obviously implies aperiodicity. Let $\tilde{Q}$ be the value iteration operator in the transformed model, and define $\tilde{V}$ as the solution set of its optimality equation (8.2.1). Finally let $\{\tilde{v}(n)\}_{n=1}^{\infty} = \{\tilde{Q}^n v(0)\}_{n=1}^{\infty}$.

LEMMA 8.3.2.

(a) $\tilde{V} = \{v \in E^N \mid \tau\, v \in V\}$

(b) *If* $(f^*,h^*) \in \Phi \times \Psi$ *satisfy the optimality equation* (8.2.1) *in the original [transformed] model, for some* $v \in V$, $[\tilde{v} \in \tilde{V}]$ *then they will satisfy the optimality equation in the transformed [original] model for* $\tau^{-1}\, v\ [\tau\tilde{v}]$ *as well.*

PROOF. Consider an arbitrary two-person zero-sum game $[c_i^{k,\ell}]$ for some $i \in \Omega$. Then for any constant a and positive number b:

$$\text{(i)} \quad \text{val}[c_i^{k,\ell}+a] = \text{val}[c_i^{k,\ell}] + a \tag{8.3.6}$$

$$\text{(iii)} \quad \text{val}[bc_i^{k,\ell}] = b\ \text{val}[c_i^{k,\ell}]$$

with the set of equilibrium pairs of action remaining unaltered, both by the translation, and by the (positive) scalar multiplication. Use (8.3.6) while rewriting (8.2.1) as

$$0 = \mathrm{val}[q_i^{k,\ell} - g^* + \sum_j (P_{ij}^{k,\ell} - \delta_{ij})v_j], \quad i \in \Omega \text{ or}$$

$$0 = \mathrm{val}[q_i^{k,\ell} - g^* + \sum_j \tau(P_{ij}^{k,\ell} - \delta_{ij})(\tau^{-1}v_j)], \; i \in \Omega. \qquad \square$$

Next, as in section 2.3, we restrict the analysis of the $\tilde{Q}$-operator on the N-1-dimensional subspace $\bar{E}^N = \{x \in E^N | x_N = 0\}$, considering the following reduction $\bar{Q}$ of the $\tilde{Q}$-operator:

$$\bar{Q}: \bar{E}^N \to \bar{E}^N: x \to \tilde{Q}x - \langle \tilde{Q}x_N \rangle \underline{1}.$$

Accordingly define $\bar{v}(n)_i = \tilde{v}(n)_i - \tilde{v}(n)_N$, $i \in \Omega$ (Note the similarity with the reduction in WHITE [131] and of $\{y(n)\}_{n=1}^{\infty}$ to $\{\hat{y}(n)\}_{n=1}^{\infty}$ in (8.2.6)). Now, under (U), $v \in V$, and hence in view of lemma 8.3.2 $v \in \tilde{V}$ are unique up to a multiple of $\underline{1}$; i.e. on $\bar{E}^N$ there exists a *unique* solution $v \in V$, which we will denote by v^*. Likewise, let $\tilde{v}^* = \tau^{-1}v^*$ be the unique member of $\tilde{V}$ in E^N. In section 1.8 we introduced the concept of a Lyapunov function, and pointed out how th.10.4. in ZANGWILL [133] can be used in order to study the asymptotic behaviour of iterative schemes of the type $\bar{v}(n+1) = \bar{Q}v(n)$; $n \geq 0$.

We next observe that both $\Lambda_1(x)$ and $\Lambda_2(x)$ are Lyapunov-functions on $\bar{E}^N$ with $\tilde{v}^*$ as origin, where (cf. (1.8.17) and (1.8.18)):

(8.3.7) $$\Lambda_1(x) = \mathrm{sp}[x - \tilde{v}^*]$$

$$\Lambda_2(x) = \mathrm{sp}[\bar{Q}x - x] = \mathrm{sp}[\tilde{Q}x - x]$$

by verifying that both for k = 1,2 (cf. (1.8.16), (a) and (b)):

(8.3.8) (i) $\Lambda_k(x)$ is continuous on $\bar{E}^N$

(ii) $\Lambda_k(x) \geq 0$; $\Lambda_k(x) = 0 \iff x = \tilde{v}^*$.

The verification for k = 1 is immediate. The non-negativity of $\Lambda_2(x)$ is immediate as well, its continuity follows from the continuity of the value operator (cf. (7.2.1)) and finally $\mathrm{sp}[\tilde{Q}x - x] = 0 \iff$ there exists a scalar g, such that $\tilde{Q}x - x = \langle g \rangle \underline{1} \iff x \in \tilde{V} \cap \bar{E}^N \iff x = \tilde{v}^*$. Note finally that $\Lambda_2(x)$ has the advantage of being computable at each point x in $\bar{E}^N$.

We next recall that in the transformed model, and as a consequence of assumption (U) the tpm's of all stationary policy pairs are unichained, and in addition have all diagonal entries strictly positive. In th.2.2.4, part (4) we proved that this implies the following "scrambling-type" condition

for all pairs of N-tuples of pure policy pairs $\{(f_1,h_1);\ldots;(f_N,h_N)\}$ and $\{(f_1',h_1');\ldots;(f_N',h_N')\}$:

$$(8.3.9)\qquad \sum_{j=1}^{N} \min[\tilde{P}(f_N,h_N)\ldots\tilde{P}(f_1,h_1)_{i_1 j};\tilde{P}(f_N',h_N')\ldots\tilde{P}(f_1',h_1')_{i_2 j}] > 0$$
$$\text{for all } i_1 \neq i_2.$$

Observe that for all $i_1,i_2 \in \Omega$ the expression to the left of the above inequality is a continuous function on $[X_{\ell=1}^{N} \Phi \times \Psi]^2$ which can be embedded as a compact subset of a Euclidean space. Hence there exists a uniform scrambling coefficient $\alpha > 0$, such that

$$(8.3.10)\qquad \sum_{j=1}^{N} \min[\tilde{P}(f_N,h_N)\ldots\tilde{P}(f_1,h_1)_{i_1 j};\tilde{P}(f_N',h_N')\ldots\tilde{P}(f_1',h_1')_{i_2 j}] > \alpha$$
$$\text{for all } i_1 \neq i_2;\ (f_\ell,h_\ell) \text{ and } (f_\ell',h_\ell') \in \Phi \times \Psi \ (\ell = 1,\ldots,N).$$

This enables us to prove the convergence of $\{\bar{v}(n)\}_{n=1}^{\infty}$ under (U). Let

$$\ell(n+1) = [\tilde{Q}\,\bar{v}(n) - \bar{v}(n)]_{\min} = [\tilde{Q}\,\tilde{v}(n) - \tilde{v}(n)]_{\min}, \text{ and}$$

$$u(n+1) = [\tilde{Q}\,\bar{v}(n) - \bar{v}(n)]_{\max} = [\tilde{Q}\,\tilde{v}(n) - \tilde{v}(n)]_{\max} \quad \text{for all } n = 0,1,\ldots$$

Finally define $g(n+1) = [\tilde{Q}\,\bar{v}(n)]_N$.

To prove convergence of $\{\bar{v}(n)\}_{n=1}^{\infty}$ we merely have to show that either one of the functions $\Lambda_1(x)$ or $\Lambda_2(x)$ satisfy (cf. (1.8.16)):

(8.3.11) (i) $\Lambda_k(\bar{Q}x) \leq \Lambda_k(x)$ for all $x \in \bar{E}^N$; $k = 1,2$,

(ii) for some integer $J \geq 1$, $\Lambda_k(\bar{Q}^J x) < \Lambda_k(x)$ for all x with $\Lambda_k(x) > 0$; $k = 1,2$.

THEOREM 8.3.3. *Suppose that* (U) *holds.*

(a) *both* $\Lambda_1(x)$ *and* $\Lambda_2(x)$ *satisfy* (8.3.11) *with* $J = N$; *hence* $\lim_{n\to\infty} \bar{v}(n) = \tilde{v}^*$ *for all* $v(0) \in E^N$.

(b) $\ell(n) \leq \ell(n+1) \leq g(n+1) \leq u(n+1) \leq u(n)$ *for all* $n = 1,2,\ldots$.
$\lim_{n\to\infty} \ell(n) = \lim_{n\to\infty} g(n) = \lim_{n\to\infty} u(n) = g^*$.

(c) *Let* $(f^*,h^*) \in \Phi \times \Psi$ *be an* AEP, *and for all* $n = 1,2,\ldots$, *let* $(f_n,h_n) \in \Phi \times \Psi$ *be any pair of policies which attain the* N *equilibria to the right of*

$$\tilde{v}(n+1) = \tilde{Q}\,\tilde{v}(n); \quad n \geq 0$$

simultaneously. Then $\ell(n+1) \leq g(f_n,h^*) \leq g^* \leq g(f^*,h_n) \leq u(n+1)$.

PROOF.

(a) We merely show that $\Lambda_1(x)$ satisfies (8.3.11), the proof for $\Lambda_2(x)$ being analogous. Use (8.3.3) to verify that $\Lambda_1(\bar{Q}x) = sp[\bar{Q}x-\tilde{v}^*] = sp[\tilde{Q}x-\tilde{Q}\tilde{v}^*] \leq sp[x-\tilde{v}^*] = \Lambda_1(x)$. Next we obtain part (ii) of condition (8.3.11) by showing that for all $x \in E^N$:

$$(8.3.12)\quad \Lambda_1(\bar{Q}^N x) = \Lambda_1(\tilde{Q}^N x) = sp[\tilde{Q}^N x-\tilde{Q}^N v^*] \leq (1-\alpha)\Lambda_1(x),$$

where the proof of (8.3.12) goes along lines with the proof of theorem 2.2.5, using (8.3.10).

(b) The proof of $\ell(n+1) \leq g(n+1) \leq u(n+1)$ is analogous to the proof of part (a)(1) in theorem 8.2.2. Next note that $\ell(n+1) = [\tilde{Q}\bar{v}(n) - \bar{v}(n)]_{min} = [\tilde{Q}(\tilde{Q}\bar{v}(n-1)) - \tilde{Q}\bar{v}(n-1)]_{min} \geq [\tilde{Q}\bar{v}(n-1) - \bar{v}(n-1)]_{min} = \ell(n)$, where the inequality part follows from (8.3.3). The monotonicity of $\{u(n)\}_{n=1}^{\infty}$ is shown in complete analogy.

(c) cf. the proof of theorem 8.2.2 part (a)(2). □

Observe that (8.3.12) is stronger than condition (8.3.11) part (ii), since the latter does not require the existence of some integer $J \geq 1$, for which

$$\sup_{x\in E^N} \Lambda_k(\bar{Q}^J x)/\Lambda_k(x) < 1; \quad k = 1,2.$$

In fact (8.3.12) shows that the approach to all of the limits in parts (a) and (b) of the above theorem exhibit a *geometric* rate of convergence, which is considerably better than the rates we obtained in section 2, for algorithm 1 (cf. Remark 2). In this particular case, it is even possible to show (along lines with the proof of theorem 2.2.5) that $\bar{Q}$ is a N-step contraction mapping on E^N, i.e. $sp[\tilde{Q}^N x - \tilde{Q}^N y] \leq (1-\alpha)\, sp[x-y]$ for all $x,y \in E^N$ and the latter leads to the following bounds on v^*:

$$(8.3.13)\quad \bar{v}(nN+r)_i - \alpha^{-1}(1-\alpha)^N sp[\tilde{v}(n)-\tilde{v}(0)] \leq \tau^{-1}v_i^* \leq$$

$$\bar{v}(nN+r)_i + \alpha^{-1}(1-\alpha)^N sp[\tilde{v}(n)-\tilde{v}(0)]; \; i \in \Omega; \; n = 1,2,\ldots; \; r = 0,\ldots,N-1.$$

(for a proof cf. theorem 2.3.1).

Finally we conclude this section by specifying as in section 2 an algorithm which approximates g^* as well as some $v \in V$, and for any $\varepsilon > 0$, ε-optimal policies for both players:

ALGORITHM 2.

step 0: Fix $0 < \tau < 1$ and transform the SDG with $(q_i^{k,\ell}; P_{ij}^{k,\ell})$ into an equivalent SDG with $(\tilde{q}_i^{k,\ell}; \tilde{P}_{ij}^{k,\ell})$ using the transformation formulae (8.3.4) and (8.3.5). Set $n = 0$; fix $v(0) \in E^N$ and $\varepsilon > 0$

step 1 and step 2: as in algorithm 1, merely replacing $\hat{y}(n)$, $L(n)$, $G(n)$, $U(n)$ by $\bar{v}(n)$, $\ell(n)$, $g(n)$ and $u(n)$; $n = 1,2,\ldots$.

Note that in this case (8.3.13) may be used as a stopping criterion for getting ε-approximations for v^*.

8.4. APPENDIX: ON REDUCING UNDISCOUNTED SRGs TO EQUIVALENT UNDISCOUNTED SDGs

In section 1.9 we pointed out that successive approximation methods for Markov Renewal Programs could be obtained by transforming the MRP-model into an equivalent undiscounted MDP-model. When trying to obtain a similar reduction for the SRG-case, thereby establishing an algorithm to solve the undiscounted version of this game, it is tempting to consider the natural extension of the data-transformation (1.9.9) and (1.9.10):

$$\tilde{q}_i^{k,\ell} = q_i^{k,\ell}/T_i^{k,\ell}; \quad i \in \Omega;\ k \in K(i),\ \ell \in M(i) \tag{8.4.1}$$

$$\tilde{P}_{ij}^{k,\ell} = \delta_{ij} + (\tau/T_i^{k,\ell})[P_{ij}^{k,\ell} - \delta_{ij}]; \quad i,j \in \Omega;\ k \in K(i),\ \ell \in M(i)$$

$$\tilde{T}_i^{k,\ell} = 1; \quad i \in \Omega;\ k \in K(i);\ \ell \in M(i).$$

with

$$0 < \tau \leq \min\{1/(1-P_{ii}^{k,\ell}) \mid i \in \Omega,\ k \in K(i),\ \ell \in M(i) \text{ with } P_{ii}^{k,\ell} < 1\}. \tag{8.4.2}$$

Note that as a consequence of (8.4.2) $\tilde{P}_{ij}^{k,\ell} \geq 0$ and $\sum_j \tilde{P}_{ij}^{k,\ell} = 1$ for all $i,j \in \Omega$, $k \in K(i)$, $\ell \in M(i)$ such that it is possible to define a (related) discrete-time SDG which has $\{\tilde{q}_i^{k,\ell}\}$ as its one-step expected rewards and $\{\tilde{P}_{ij}^{k,\ell}\}$ as its set of one-step transition probabilities.

We recall from (7.3.13) that under (H), the solution of our SRG-model reduces to the problem of finding a vector $v \in E^N$ that satisfies the functional equation (cf. also (8.2.1) and corollary 7.3.5):

$$v_i = \text{val}[q_i^{k,\ell} - g^* T_i^{k,\ell} + \sum_j P_{ij}^{k,\ell} v_j]; \quad i \in \Omega. \tag{8.4.3}$$

Lemma 8.4.1 below shows that the above proposed reduction method works, in

case the holding times $T_i^{k,\ell}$ satisfy the separability assumption (SEP) in (8.1.1). Let V denote the set of solutions to (8.4.3) and let $\tilde{V}$ be the set of solutions to the optimality equation in the transformed SDG. Likewise, all other quantities of interest in the transformed model will be marked off by a $(\sim)$-symbol.

LEMMA 8.4.1. *Suppose* (H) *and* (SEP) *in* (8.1.1) *hold. For each pair of policies* $f \in \Phi$ *and* $h \in \Psi$ *define* $\tilde{f}$, $\hat{f} \in \Phi$ *and* $\tilde{h}$, $\hat{h} \in \Psi$ *by:*

$$(8.4.4)\quad \tilde{f}_{ik} = f_{ik}\alpha_i^k / \textstyle\sum_{r\in K(i)} f_{ir}\alpha_i^r;\ i \in \Omega,\ k \in K(i)$$

$$\hat{f}_{ik} = f_{ik}(\alpha_i^k)^{-1} / \textstyle\sum_{r\in K(i)} f_{ir}(\alpha_i^r)^{-1};\ i \in \Omega,\ k \in K(i)$$

$$\tilde{h}_{i\ell} = h_{i\ell}\beta_i^\ell / \textstyle\sum_{r\in M(i)} h_{ir}\beta_i^r;\ \ i \in \Omega,\ \ell \in M(i)$$

$$\hat{h}_{i\ell} = h_{i\ell}(\beta_i^\ell)^{-1} / \textstyle\sum_{r\in M(i)} h_{ir}(\beta_i^r)^{-1};\ i \in \Omega,\ \ell \in M(i).$$

Then,

(a) $\quad g(f,h) = \tilde{g}(\tilde{f},\tilde{h})$ and $\tilde{g}(f,h) = g(\hat{f},\hat{h})$

i.e. if (f,h) *is an* AEP *in the original* [transformed] *model, then* $(\tilde{f},\tilde{h})$ [$(\hat{f},\hat{h})$] *is an* AEP *in the transformed* [*original*] *model. In other words, there exists a computationally tractible one- to one correspondence between the sets of stationary* AEPs *in the two models.*

(b) $\quad \tilde{V} = \{v \in E^N \mid \tau v \in V\}.$

PROOF. We first consider an arbitrary matrix game $[c_i^{k,\ell}]$ for some $i \in \Omega$, in relationship with its "transformed" version $[c_i^{k,\ell}/T_i^{k,\ell}]$. Assuming that $\mathrm{val}[c_i^{k,\ell}] = 0$, we prove the following two properties

$$(8.4.5)\quad \mathrm{val}[c_i^{k,\ell}/T_i^{k,\ell}] = 0.$$

(8.4.6) If $x^* \in K(i)$ is an optimal action in the original {transformed} matrix game, then $\tilde{x} = [x_k^* \alpha_i^k / \sum_r x_r^* \alpha_i^r]_{k\in K(i)}$ {and $\hat{x} = [x_k^*(\alpha_i^k)^{-1} / \sum_r x_r^*(\alpha_i^r)^{-1}]_{k\in K(i)}$} is an optimal action in the transformed {original} model.

A similar one to one correspondence exists between the sets of optimal actions for player 2.

Part (b) then follows by rewriting (8.4.3) in a homogeneous way, i.e. $0 = \text{val}[q_i^{k,\ell} - g^* T_i^{k,\ell} + \sum_j \tau(P_{ij}^{k,\ell} - \delta_{ij})\,(\tau^{-1}v)_j]$, $i \in \Omega$, invoking (8.4.5). The proof of part (a) follows from (8.4.6) and the observation that in the system

$$0 = [P(f,h)-I]g \tag{8.4.7}$$

$$0 = \{q(f,h)_i - g_i T(f,h)_i + [P(f,h)-I]v_i\}, \quad i \in \Omega$$

the g-part is uniquely determined as g(f,h).

This leaves us with the proof of (8.4.5) and (8.4.6). Let (x^*,y^*) be a pair of equilibrium actions in the original matrix game. Then, for all $y \in M(i)$:

$$\sum_k \sum_\ell \tilde{x}_k \frac{c_i^{k,\ell}}{\alpha_i^k \beta_i^\ell}\, y_\ell = \frac{(\sum_r y_r \beta_i^r)}{(\sum_r x_r^* \alpha_i^r)} \sum_k \sum_\ell x_k^*\, c_i^{k,\ell} \left[\frac{y_\ell \beta_i^\ell}{\sum_r y_r \beta_i^r}\right] \geq 0 \tag{8.4.8}$$

where the inequality follows from x^* being optimal in the original game. Likewise, with $\tilde{y} = [y_\ell^* \beta_i^\ell / \sum_r y_r^* \beta_i^r]_{\ell \in M(i)}$ we obtain

$$\sum_k \sum_\ell x_k \frac{c_i^{k,\ell}}{\alpha_i^k \beta_i^\ell}\, \tilde{y}_\ell \leq 0 \quad \text{for all } x \in K(i) \tag{8.4.9}$$

such that (8.4.5) and (8.4.6) follow from the combination of (8.4.8) and (8.4.9). □

We conclude that g^*, $v \in V$ *and* ε*-optimal policies for both players can be computed by applying algorithm* 1 *under* (H), *or algorithm* (2) *under* (U) *to the transformed* SDG, *and by exploiting the one to one correspondences exhibited by lemma* 8.4.1. *Note in addition, that by choosing* τ *strictly less than the upperbound in* (8.4.2) *the transformation in step* 0 *of algorithm* 2 *becomes superfluous.*

The above described reduction fails, if the expected holding times fail to satisfy (SEP). This is due to (8.4.5) and (8.4.6) breaking down in general, examples of which can easily be constructed. As a consequence, establishing an algorithm for the general SRG-case remains an outstanding problem.

REFERENCES

[1] ANTHONISSE, Jac M. & H.C. TIJMS, (1977) *Exponential convergence of products of stochastic matrices*, J.M.A.A. 59, 360-364.

[2] BATHER, J., (1973) *Optimal decision procedures for finite Markov Chains*, part I, Adv. Appl. Prob. 5, 328-339.

[3] ———— , (1973) *Optimal decision procedures for finite Markov Chains*, part II, Adv. Appl. Prob. 5, 521-540.

[4] ———— , (1973) *Optimal decision procedures for finite Markov Chains*, part III, Adv. Appl. Prob. 5, 541-553.

[5] BELLMAN, R., (1957) *A Markovian decision process*, J. Math. Mech. 6, 679-684.

[6] BEWLEY, T. & E. KOHLBERG, (1976) *The asymptotic theory of stochastic games*, Math. of O.R. 1, 197-208.

[7] ———— & ———— , (1976) *The asymptotic solution of a recursive equation arising in stochastic games*, Math. of O.R. 1, 321-336.

[8] ———— & ———— , (1978) *On stochastic games with stationary optimal strategies*, Math. of O.R. 3, 104-125.

[9] BILLINGSLEY, P., (1968) *Convergence of probability measures*, Wiley, New York.

[10] BLACKWELL, D., (1962) *Discrete dynamic programming*, Ann. Math. Stat. 33, 719-726.

[11] ———— , (1968) *Discounted dynamic programming*, Ann. Math. Stat. 39, 216-235.

[12] BOWERMAN, B., (1974) *Nonstationary Markov decision processes and related topics in nonstationary Markov Chains*, Ph.D. thesis, Iowa State University.

[13] BROWN, B., (1965) *On the iterative method of dynamic programming on a finite state space discrete time Markov Process*, Ann. Math. Stat. 36, 1279-1285.

[14] CHATTERJEE, S. & E. SENETA, (1977) *Towards consensus: some convergence theorems on repeated averaging*, J. Appl. Prob. 14, 89-97.

[15] COLLATZ, L., (1964) *Funktional Analysis und Numerische Mathematik,* Berlin, Springer Verlag.

[16] DALKEY, N., (1969) *The Delphi method: an experimental study of group opinion,* The Rand Corporation, RM-5888.

[17] DE GROOT, M., (1974) *Reaching a Consensus,* J. Am. Stat. Ass. 69, 118-121.

[18] DE LEVE, G., A. FEDERGRUEN & H.C. TIJMS, (1977) *A general Markov decision method* I : *Model and techniques,* Adv. Appl. Prob. 9, 296-315.

[19] ———— , ———— & ———— , (1977) *A general Markov decision method* II: *Applications,* Adv. Appl. Prob. 9, 316-335.

[20] DENARDO, E., (1967) *Contraction Mappings in the theory underlying dynamic programming,* SIAM Review 9, 165-177.

[21] ———— , (1970) *Computing a bias-optimal policy in discrete-time Markov decision problems,* Op. Res. 18, 279-289.

[22] ———— , (1971) *Markov Renewal Programs with small interest rates,* Ann. Math. Stat. 42, 477-496.

[23] ———— & B. FOX, (1968) *Multichain Markov Renewal Programs,* SIAM J. Appl. Math. 16, 468-487.

[24] DERMAN, C., (1966) *Denumerable state Markovian decision processes - average cost criterion,* Ann. Math. Stat. 37, 1545-1553.

[25] ———— , (1970) *Finite state Markovian Decision Processes,* Academic Press, New York.

[26] ———— , R. STRAUCH, (1966) *A note on memoryless rules for controlling sequential processes,* Ann. Math. Stat. 37, 276-278.

[27] ———— , A. VEINOTT, Jr., (1967) *A solution to a countable system of equations arising in Markovian decision processes,* Ann. Math. Stat. 38, 582-584.

[28] DOOB, J., (1953) *Stochastic Processes,* John Wiley, New York.

[29] FAN, K., (1952) *Fixed point and minimax theorems in locally convex topological linear spaces,* Proc. Nat. Acad. Sci. 38, 121-126.

[30] FEDERGRUEN, A., (1978) *On N-person stochastic games with denumerable state space*, Adv. Appl. Prob. 10, 452-471.

[31] ———————— , (1980) *On the functional equations in undiscounted and sensitive discounted stochastic games*, Zeitschr. f. Op. Res. A24, 243-262.

[32] ———————— , (1980) *Successive approximation methods in undiscounted stochastic games*, Op. Res. 28, 794-809.

[33] ———————— , (1981) *On non-stationary Markov chains with converging transition matrices*, Stoch. Proc. Appl. 11, 187-192.

[34] ———————— , & P.J. SCHWEITZER, (1978) *Discounted and undiscounted value iteration in Markov decision processes: a survey*, in Puterman, M.L. (ed.), *Dynamic Programming and Its Applications*, Academic Press, New York, 23-52.

[35] ———————— & ———————— , (1978) *Turnpike properties in undiscounted Markov Decision Problems* (forthcoming).

[36] ———————— & ———————— , (1980) *Non-stationary Markov decision problems with converging parameters*, J. Opt. Th. Appl. 34, 207-241.

[37] ———————— & ———————— , (1984) *Successive approximation methods for solving nested functional equations in Markov decision theory*, Math. of O.R. (to appear).

[38] ———————— & ———————— , (1978) *Data-transformations for Markov Renewal Programming* (forthcoming).

[39] ———————— & ———————— , (1978) *Lyapunov functions in Markov Renewal Programming* (forthcoming).

[40] ———————— & H.C. TIJMS, (1978) *The optimality equation in average cost denumerable state semi-Markov decision problems, recurrency conditions and algorithms*, J. Appl. Prob. 15, 356-373.

[41] ———— , A. HORDIJK & H.C. TIJMS, (1978) *Recurrence conditions in denumerable state Markov decision processes,* in Puterman, M.L. (ed.), *Dynamic Programming and Its Applications*, Academic Press, New York, 1-22.

[42] ———— , ———— & ———— , (1978) *A note on simultaneous recurrence conditions on a set of denumerable stochastic matrices,* J. Appl. Prob. 15, 842-847.

[43] ———— , P.J. SCHWEITZER & H.C. TIJMS, (1978) *Contraction mappings underlying undiscounted Markov decision problems,* J.M.A.A. 65, 711-730.

[44] ———— , ———— & ———— , (1977) *Value iteration in undiscounted Markov decision problems,* part I: *Asymptotic behaviour,* part II: *Geometric Convergence,* part III: *Algorithms,* in Tijms, H.C. & J. Wessels (eds.), *Markov decision theory,* Proceedings of the advanced seminar on Markov decision theory held in Amsterdam, September 13-17, 1976, Math. Centre Tract no.93, 119-140, 141-151, 153-159.

[45] FINKBEINER, B. & W. RUNGALDIER, (1969) *A value iteration algorithm for Markov Renewal Programming, in Computing methods in optimization problems* 2, ed. by L. Zadeh, New York, Academic Press, 95-104.

[46] FLYNN, J., (1976) *Conditions for the equivalence of optimality criteria in dynamic programming,* Ann. Stat. 4, 936-953.

[47] GILLETTE, D., (1957) *Stochastic Games with zero stop probabilities in M. Dresher et. al.* (eds.) *Contributions to the theory of Games,* Vol. III (Princeton Univ. Press), Princeton, New Jersey, 179-188.

[48] GLICKSBERG, I., (1952) *A further generalization of the Kakutani fixed point theorem with application to Nash equilibrium points,* Proc. Amer. Math. Soc. 3 (1952), 170-174.

[49] GOFFIN, J., (1977) *On convergence rates of subgradient optimization methods,* Math. Progr. 13, 329-347.

[50] GRINOLD, R., (1973) *Elimination of suboptimal actions in Markov decision problems,* Op. Res. 21, 848-851.

[51] HAJNAL, J., (1958) *Weak ergodicity in non homogeneous Markov Chains,* Proc. Cambridge Philos. Soc. 54, 233-246.

[52] HASTINGS, N., (1969) *Optimization of discounted Markov decision problems,* Op. Res. Quart. 20, 499-500.

[53] ———— , (1971) *Bounds on the gain of a Markov decision process,* Op. Res. 19, 240-244.

[54] ———— , (1976) *A test for nonoptimal actions in undiscounted finite Markov decision chains,* Man. Sci. 23, 87-92.

[55] ———— & J. MELLO, (1973) *Tests for suboptimal actions in discounted Markov programming,* Man. Sci. 19, 1019-1022.

[56] HIMMELBERG, C. , T. PARTHASARATHY, T. RAGHAVAN & F. van VLECK, (1976) *Existence of* p*-equilibrium and optimal stationary strategies in stochastic games,* Proc. Am. Math. Soc. 245-251.

[57] HOFFMAN, A. & R. KARP, (1966) *On non-terminating stochastic games,* Man. Sci. 12, 359-370.

[58] HORDIJK, A., (1974) *Dynamic Programming and Potential Theory,* Mathematical Centre Tract 51, Mathematisch Centrum, Amsterdam.

[59] ———— & K. SLADKŶ, (1977) *Sensitive optimality criteria in countable state dynamic programming,* Math. of O.R. 2, 1-14.

[60] ———— & H.C. TIJMS, (1975) *A modified form of the iterative method of dynamic programming,* Ann. Stat. 3, 203-208.

[61] ———— , P.J. SCHWEITZER & H.C. TIJMS, (1975) *The asymptotic behaviour of the minimal total expected cost for the denumerable state Markov decision model,* J. Appl. Prob. 12, 298-305.

[62] ———— , O. VRIEZE & G. WANROOIJ, (1983) *Semi-Markov strategies in stochastic games,* Int. J. Game Th. 12, 81-90.

[63] HOWARD, R., (1960) *Dynamic Programming and Markov processes,* John Wiley, New York.

[64] ———— & J. MATHESON, (1972) *Risk sensitive Markov decision processes,* Man. Sci. 18, 356-369.

[65] HUANG, C., D. ISAACSON & B. VINOGRADE, (1976) *The rate of convergence of certain nonhomogeneous Markov Chains,* Z. Wahrscheinlichkeitsth. 35, 141-146.

[66] IDZIK, A., (1976) *Two-stage noncooperative discounted stochastic games*, Computation Centre of the Polish Academy of Science, Warsaw.

[67] ISAACSON, D. & R. MADSEN, (1974) *Positive columns for stochastic matrices*, J. Appl. Prob. 11, 829-835.

[68] ———— & ———— , (1976) *Markov Chains, theory and applications*, John Wiley, New York.

[69] JEWELL, W., (1963) *Markov Renewal Programming*, Op. Res. 11, 938-971.

[70] KARLIN, S., (1959) *Mathematical methods and the theory of games*, Vol. I, Addison-Wesley, London.

[71] KEMENY, J. & J. SNELL, (1961) *Finite Markov chains*, Van Nostrand, Princeton.

[72] KOEHLER, G., A WHINSTON & G. WRIGHT, (1975) *Optimization over Leontief substitution systems*, North Holland, Amsterdam.

[73] KUSHNER, H. & A. KLEINMAN, (1971) *Accelerated procedures for the solution of discrete Markov control problems*, IEEE Trans. Automatic Control AC-16, 147-152.

[74] LANERY, E., (1967) *Etude asymptotique des systèmes Markoviens à commande*, Rev. Inf. Rech. Op. 1, 3-56.

[75] ———— , (1968) *Compléments à l'étude asymptotique des systèmes Markoviens à commande*, I.R.I.A. Rocquencourt, France.

[76] LEMBERSKY, M., (1974) *On maximal rewards and ε-optimal policies in continuous time Markov decision chains*, Ann. Stat. 2, 159-169.

[77] ———— , (1974) *Preferred rules in continuous time Markov decision processes*, Man. Sci. 21, 348-357.

[78] LIGGETT, T. & S. LIPPMAN, (1969) *Stochastic games with perfect information and time average payoff*, SIAM Review 11, 604-607.

[79] LIPPMAN, S., (1975) *On dynamic programming with unbounded rewards*, Man. Sci. 21, 1225-1233.

[80] LUENBERGER, D., (1973) *Introduction to linear and nonlinear programming*, Addison-Wesley, Reading, Massachusetts.

[81] MacQUEEN, J., (1967) *A test for suboptimal actions in Markovian decision problems*, Op. Res. 15, 559-561.

[82] MAITRA, A., (1968) *Discounted dynamic programming on compact metric spaces*, Sankhya Ser. A. 30, 211-216.

[83] MANDL, P., (1967) *An iterative method for maximizing the characteristic root of positive matrices*, Rev. Roum. Math. Pures et Appl., Tome XII, no.9, 1317-1322.

[84] MANNE, A., (1960) *Linear programming and sequential decisions*, Man. Sci. 6, 259-267.

[85] MILLER, B. & A. VEINOTT, Jr., (1969) *Discrete dynamic programming with a small interest rate*, Ann. Math. Stat. 40, 366-370.

[86] MORTON, T. & W. WECKER, (1977) *Discounting, ergodicity and convergence for Markov decision processes*, Man. Sci. 23, 890-900.

[87] MURRAY, W., (1972) *Numerical methods for unconstrained optimization*, Academic Press, New York.

[88] NEVEU, J., (1965) *Mathematical foundations of the calculus of probability*, Holden-Day, San Francisco.

[89] ODONI, A., (1969) *On finding the maximal gain for Markov decision processes*, Op. Res. 17, 857-860.

[90] PARTHASARATHY, T. & M. STERN, (1976) *Markov Games a survey*, University of Illinois, Chicago.

[91] PAZ, A., (1971) *Introduction to probabilistic automata*, Academic Press, New York.

[92] PLISKA, S., (1976) *Optimization of multitype branching processes*, Man. Sci. 23, 117-125.

[93] POLLATSCHEK, M. & B. AVI-ITZHAK, (1969) *Algorithms for stochastic games with geometrical interpretation*, Man. Sci. 15, 399-415.

[94] PORTEUS, E., (1971) *Some bounds for discounted sequential decision processes*, Man. Sci. 18, 7-11.

[95] ———————, (1975) *Bounds and transformations for discounted finite Markov decision chains*, Op. Res. 23, 761-784.

[96] REETZ, D., (1973) *Solution of a Markovian decision problem by successive overrelaxation*, Zeitschr.f. Op. Res. 17, 29-32.

[97] ROGERS, P., (1969) *Nonzero-sum stochastic games*, Report ORC 69-8, Operations Res. Center, Univ. of California, Berkeley, Calif.

[98] ROSS, S., (1970) *Applied probability models with optimization applications*, Holden-Day, San Francisco, Calif.

[99] ROTHBLUM, U., (1975) *Multiplicative Markov decision chains*, Yale University report, School of Management and Organization, New Haven, Conn.

[100] ROYDEN, H., (1968) *Real analysis*, 2nd ed., MacMillan, New York.

[101] RUDIN, W., (1964) *Principles of mathematical analysis*, (2nd ed.), McGraw-Hill, New York.

[102] RUSSEL, C., (1972) *An optimal policy for operating a multiple purpose reservoir*, Op. Res. 20, 1181-1189.

[103] SCHELLHAAS, H., (1974) *Zur extrapolation in Markoffschen entscheidungsmodellen mit Diskontierung*, Zeitschr. f. Op. Res. 18, 91-104.

[104] SCHWEITZER, P.J., (1965) *Perturbation theory and Markovian decision processes*, Ph.D. dissertation, M.I.T. Operations Research Center Report 15.

[105] ———————— , (1968) *Perturbation theory and finite Markov chains*, J. Appl. Prob. 5, 401-413.

[106] ———————— , (1968) *A turnpike theorem for undiscounted Markovian decision processes*, presented at ORSA/TIMS, national meeting, May 1968, San Francisco, Calif.

[107] ———————— , (1971) *Multiple policy improvements in undiscounted Markov renewal programming*, Op. Res. 19, 784-793.

[108] ———————— , (1971) *Iterative solution of the functional equations for undiscounted Markov renewal programming*, J.M.A.A. 34, 495-501.

[109] ———————— , A. FEDERGRUEN, (1978) *The functional equations of undiscounted Markov renewal programming*, Math. of O.R. 3, 308-321.

[110] ———————— & ———————— , (1977) *The asymptotic behavior of undiscounted value iteration in Markov decision problems*, Math. of O.R. 2, 360-381.

[111] ———— & ———— , (1979) *Geometric convergence of value-iteration in multichain Markov renewal programming*, Adv. Appl. Prob. 11, 188-217; *Correction*, Adv. Appl. Prob. 11, 456.

[112] ———— & ———— , (1982) *Variational characterizations in Markov renewal programs*, Working paper 482A, Graduate School of Business, Columbia University, New York.

[113] SENETA, E., (1973) *Nonnegative matrices*, Allen & Unwin, London.

[114] SHAPIRO, J., (1968) *Turnpike planning horizons for a Markovian decision model*, Man. Sci. 14, 292-300.

[115] SHAPLEY, L., (1953) *Stochastic games*, Proc. Nat. Acad. Sci. U.S.A. 39, 1095-1100.

[116] SLADKÝ, K., (1974) *On the set of optimal controls for Markov chains with rewards*, Kybernetica 10, 350-367.

[117] SOBEL, M., (1971) *Noncooperative stochastic games*, Ann. Math. Stat. 42, 1930-1935.

[118] ———— , (1973) *Continuous stochastic games*, J. Appl. Prob. 10, 597-604.

[119] STERN, M., (1975) *On stochastic games with limiting average payoff*, Ph.D. dissertation, Dept. of Math., Univ. of Illinois, Chicago, Circle Campus.

[120] SU, Y. & R. DEININGER, (1972) *Generalization of White's method of successive approximations to periodic Markovian decision processes*, Op. Res. 20, 318-326.

[121] TAYLOR, H., (1965) *Markovian sequential replacement processes*, Ann. Math. Stat. 36, 1677-1694.

[122] TIJMS, H.C. (1974) *An iterative method of approximating average cost optimal* (s,S) *inventory policies*, Zeitschr. f. Op. Res. 18, 215-223.

[123] ———— , (1975) *On dynamic programming with arbitrary state space and the average return as criterion*, Math. Centre report BW 55/75.

[124] VAN DER WAL, J., (1980) *Successive approximations for average reward Markov games*, Int. J. Game Th. 9, 13-24.

[125] VAN NUNEN, J., (1976) *A set of successive approximation methods for discounted Markovian decision problems*, Zeitschr. f. Op. Res. 20, 203-209.

[126] VARADARAJAN, V., (1958) *Weak convergence of measures on separable metric spaces*, Sankhya, Series A, 15-22.

[127] VEINOTT, A. Jr., (1969) *Discrete dynamic programming with sensitive discount optimality criteria*, Ann. Math. Stat. 40, 1635-1660.

[128] VERKHOVSKY, B., (1977) *Smoothing system design and parametric Markovian programming*, in H.C. Tijms & J. Wessels (eds.) *"Markov decision Theory"*, Proceedings of the advanced seminar on Markov decision theory held in Amsterdam, september 13-17, 1976; Mathematical Centre Tract no.93, 105-117.

[129] ———— & V. SPIVAK, (1972) *Water systems optimal design and controlled stochastic processes*, Ekonomika 1, Mathematicheskie Metody, Vol. VIII, N6, 966-972.

[130] VRIEZE, O. & G. WANROOIJ, (1975) private communication.

[131] WHITE, D., (1963) *Dynamic programming, Markov chains and the method of successive approximations*, J.M.A.A.6, 373-376.

[132] WHITT, W., (1980) *Representation and approximation of noncooperative sequential games*, SIAM J. Control Optim. 18, 33-48.

[133] ZANGWILL, W., (1969) *Nonlinear programming, a unified approach*, Prentice Hall, inc.; Englewood Cliffs, N.J.

TITLES IN THE SERIES MATHEMATICAL CENTRE TRACTS

(An asterisk before the MCT number indicates that the tract is under preparation).

A leaflet containing an order form and abstracts of all publications mentioned below is available at the Mathematisch Centrum, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands. Orders should be sent to the same address.

MCT 1 T. VAN DER WALT, *Fixed and almost fixed points*, 1963. ISBN 90 6196 002 9.

MCT 2 A.R. BLOEMENA, *Sampling from a graph*, 1964. ISBN 90 6196 003 7.

MCT 3 G. DE LEVE, *Generalized Markovian decision processes, part I: Model and method*, 1964. ISBN 90 6196 004 5.

MCT 4 G. DE LEVE, *Generalized Markovian decision processes, part II: Probabilistic background*, 1964. ISBN 90 6196 005 3.

MCT 5 G. DE LEVE, H.C. TIJMS & P.J. WEEDA, *Generalized Markovian decision processes, Applications*, 1970. ISBN 90 6196 051 7.

MCT 6 M.A. MAURICE, *Compact ordered spaces*, 1964. ISBN 90 6196 006 1.

MCT 7 W.R. VAN ZWET, *Convex transformations of random variables*, 1964. ISBN 90 6196 007 X.

MCT 8 J.A. ZONNEVELD, *Automatic numerical integration*, 1964. ISBN 90 6196 008 8.

MCT 9 P.C. BAAYEN, *Universal morphisms*, 1964. ISBN 90 6196 009 6.

MCT 10 E.M. DE JAGER, *Applications of distrubutions in mathematical physics*, 1964. ISBN 90 6196 010 X.

MCT 11 A.B. PAALMAN-DE MIRANDA, *Topological semigroups*, 1964. ISBN 90 6196 011 8.

MCT 12 J.A.Th.M. VAN BERCKEL, H. BRANDT CORSTIUS, R.J. MOKKEN & A. VAN WIJNGAARDEN, *Formal properties of newspaper Dutch*, 1965. ISBN 90 6196 013 4.

MCT 13 H.A. LAUWERIER, *Asymptotic expansions*, 1966, out of print; replaced by MCT 54.

MCT 14 H.A. LAUWERIER, *Calculus of variations in mathematical physics*, 1966. ISBN 90 6196 020 7.

MCT 15 R. DOORNBOS, *Slippage tests*, 1966. ISBN 90 6196 021 5.

MCT 16 J.W. DE BAKKER, *Formal definition of programming languages with an application to the definition of ALGOL 60*, 1967. ISBN 90 6196 022 3.

MCT 17 R.P. VAN DE RIET, *Formula manipulation in ALGOL 60, part 1*, 1968. ISBN 90 6196 025 8.

MCT 18 R.P. VAN DE RIET, *Formula manipulation in ALGOL 60, part 2*, 1968. ISBN 90 6196 038 X.

MCT 19 J. VAN DER SLOT, *Some properties related to compactness*, 1968. ISBN 90 6196 026 6.

MCT 20 P.J. VAN DER HOUWEN, *Finite difference methods for solving partial differential equations*, 1968. ISBN 90 6196 027 4.

MCT 21 E. WATTEL, *The compactness operator in set theory and topology*, 1968. ISBN 90 6196 028 2.

MCT 22 T.J. DEKKER, *ALGOL 60 procedures in numerical algebra, part 1*, 1968. ISBN 90 6196 029 0.

MCT 23 T.J. DEKKER & W. HOFFMANN, *ALGOL 60 procedures in numerical algebra, part 2*, 1968. ISBN 90 6196 030 4.

MCT 24 J.W. DE BAKKER, *Recursive procedures*, 1971. ISBN 90 6196 060 6.

MCT 25 E.R. PAËRL, *Representations of the Lorentz group and projective geometry*, 1969. ISBN 90 6196 039 8.

MCT 26 EUROPEAN MEETING 1968, *Selected statistical papers, part I*, 1968. ISBN 90 6196 031 2.

MCT 27 EUROPEAN MEETING 1968, *Selected statistical papers, part II*, 1969. ISBN 90 6196 040 1.

MCT 28 J. OOSTERHOFF, *Combination of one-sided statistical tests*, 1969. ISBN 90 6196 041 X.

MCT 29 J. VERHOEFF, *Error detecting decimal codes*, 1969. ISBN 90 6196 042 8.

MCT 30 H. BRANDT CORSTIUS, *Exercises in computational linguistics*, 1970. ISBN 90 6196 052 5.

MCT 31 W. MOLENAAR, *Approximations to the Poisson, binomial and hypergeometric distribution functions*, 1970. ISBN 90 6196 053 3.

MCT 32 L. DE HAAN, *On regular variation and its application to the weak convergence of sample extremes*, 1970. ISBN 90 6196 054 1.

MCT 33 F.W. STEUTEL, *Preservation of infinite divisibility under mixing and related topics*, 1970. ISBN 90 6196 061 4.

MCT 34 I. JUHÁSZ, A. VERBEEK & N.S. KROONENBERG, *Cardinal functions in topology*, 1971. ISBN 90 6196 062 2.

MCT 35 M.H. VAN EMDEN, *An analysis of complexity*, 1971. ISBN 90 6196 063 0.

MCT 36 J. GRASMAN, *On the birth of boundary layers*, 1971. ISBN 90 6196 064 9.

MCT 37 J.W. DE BAKKER, G.A. BLAAUW, A.J.W. DUIJVESTIJN, E.W. DIJKSTRA, P.J. VAN DER HOUWEN, G.A.M. KAMSTEEG-KEMPER, F.E.J. KRUSEMAN ARETZ, W.L. VAN DER POEL, J.P. SCHAAP-KRUSEMAN, M.V. WILKES & G. ZOUTENDIJK, *MC-25 Informatica Symposium* 1971. ISBN 90 6196 065 7.

MCT 38 W.A. VERLOREN VAN THEMAAT, *Automatic analysis of Dutch compound words*, 1971. ISBN 90 6196 073 8.

MCT 39 H. BAVINCK, *Jacobi series and approximation*, 1972. ISBN 90 6196 074 6.

MCT 40 H.C. TIJMS, *Analysis of (s,S) inventory models*, 1972. ISBN 90 6196 075 4.

MCT 41 A. VERBEEK, *Superextensions of topological spaces*, 1972. ISBN 90 6196 076 2.

MCT 42 W. VERVAAT, *Success epochs in Bernoulli trials (with applications in number theory)*, 1972. ISBN 90 6196 077 0.

MCT 43 F.H. RUYMGAART, *Asymptotic theory of rank tests for independence*, 1973. ISBN 90 6196 081 9.

MCT 44 H. BART, *Meromorphic operator valued functions*, 1973. ISBN 90 6196 082 7.

MCT 45 A.A. BALKEMA, *Monotone transformations and limit laws* 1973. ISBN 90 6196 083 5.

MCT 46 R.P. VAN DE RIET, *ABC ALGOL, A portable language for formula manipulation systems, part 1: The language*, 1973. ISBN 90 6196 084 3.

MCT 47 R.P. VAN DE RIET, *ABC ALGOL, A portable language for formula manipulation systems, part 2: The compiler*, 1973. ISBN 90 6196 085 1.

MCT 48 F.E.J. KRUSEMAN ARETZ, P.J.W. TEN HAGEN & H.L. OUDSHOORN, *An ALGOL 60 compiler in ALGOL 60, Text of the MC-compiler for the EL-X8*, 1973. ISBN 90 6196 086 X.

MCT 49 H. KOK, *Connected orderable spaces*, 1974. ISBN 90 6196 088 6.

MCT 50 A. VAN WIJNGAARDEN, B.J. MAILLOUX, J.E.L. PECK, C.H.A. KOSTER, M. SINTZOFF, C.H. LINDSEY, L.G.L.T. MEERTENS & R.G. FISKER (eds), *Revised report on the algorithmic language ALGOL 68*, 1976. ISBN 90 6196 089 4.

MCT 51 A. HORDIJK, *Dynamic programming and Markov potential theory*, 1974. ISBN 90 6196 095 9.

MCT 52 P.C. BAAYEN (ed.), *Topological structures*, 1974. ISBN 90 6196 096 7.

MCT 53 M.J. FABER, *Metrizability in generalized ordered spaces*, 1974. ISBN 90 6196 097 5.

MCT 54 H.A. LAUWERIER, *Asymptotic analysis, part 1*, 1974. ISBN 90 6196 098 3.

MCT 55 M. HALL JR. & J.H. VAN LINT (eds), *Combinatorics, part 1: Theory of designs, finite geometry and coding theory*, 1974. ISBN 90 6196 099 1.

MCT 56 M. HALL JR. & J.H. VAN LINT (eds), *Combinatorics, part 2: Graph theory, foundations, partitions and combinatorial geometry*, 1974. ISBN 90 6196 100 9.

MCT 57 M. HALL JR. & J.H. VAN LINT (eds), *Combinatorics, part 3: Combinatorial group theory*, 1974. ISBN 90 6196 101 7.

MCT 58 W. ALBERS, *Asymptotic expansions and the deficiency concept in statistics*, 1975. ISBN 90 6196 102 5.

MCT 59 J.L. MIJNHEER, *Sample path properties of stable processes*, 1975. ISBN 90 6196 107 6.

MCT 60 F. GÖBEL, *Queueing models involving buffers*, 1975. ISBN 90 6196 108 4.

*MCT 61 P. VAN EMDE BOAS, *Abstract resource-bound classes, part 1*, ISBN 90 6196 109 2.

*MCT 62 P. VAN EMDE BOAS, *Abstract resource-bound classes, part 2*, ISBN 90 6196 110 6.

MCT 63 J.W. DE BAKKER (ed.), *Foundations of computer science*, 1975. ISBN 90 6196 111 4.

MCT 64 W.J. DE SCHIPPER, *Symmetric closed categories*, 1975. ISBN 90 6196 112 2.

MCT 65 J. DE VRIES, *Topological transformation groups 1 A categorical approach*, 1975. ISBN 90 6196 113 0.

MCT 66 H.G.J. PIJLS, *Locally convex algebras in spectral theory and eigenfunction expansions*, 1976. ISBN 90 6196 114 9.

*MCT 67 H.A. LAUWERIER, *Asymptotic analysis, part 2*, ISBN 90 6196 119 X.

MCT 68 P.P.N. DE GROEN, *Singularly perturbed differential operators of second order*, 1976. ISBN 90 6196 120 3.

MCT 69 J.K. LENSTRA, *Sequencing by enumerative methods*, 1977. ISBN 90 6196 125 4.

MCT 70 W.P. DE ROEVER JR., *Recursive program schemes: Semantics and proof theory*, 1976. ISBN 90 6196 127 0.

MCT 71 J.A.E.E. VAN NUNEN, *Contracting Markov decision processes*, 1976. ISBN 90 6196 129 7.

MCT 72 J.K.M. JANSEN, *Simple periodic and nonperiodic Lamé functions and their applications in the theory of conical waveguides*, 1977. ISBN 90 6196 130 0.

MCT 73 D.M.R. LEIVANT, *Absoluteness of intuitionistic logic*, 1979. ISBN 90 6196 122 X.

MCT 74 H.J.J. TE RIELE, *A theoretical and computational study of generalized aliquot sequences*, 1976. ISBN 90 6196 131 9.

MCT 75 A.E. BROUWER, *Treelike spaces and related connected topological spaces*, 1977. ISBN 90 6196 132 7.

MCT 76 M. REM, *Associations and the closure statement*, 1976. ISBN 90 6196 135 1.

MCT 77 W.C.M. KALLENBERG, *Asymptotic optimality of likelihood ratio tests in exponential families*, 1977. ISBN 90 6196 134 3.

MCT 78 E. DE JONGE & A.C.M. VAN ROOIJ, *Introduction to Riesz spaces*, 1977. ISBN 90 6196 133 5.

MCT 79 M.C.A. VAN ZUIJLEN, *Empirical distributions and rank statistics*, 1977. ISBN 90 6196 145 9.

MCT 80 P.W. HEMKER, *A numerical study of stiff two-point boundary problems*, 1977. ISBN 90 6196 146 7.

MCT 81 K.R. APT & J.W. DE BAKKER (eds), *Foundations of computer science II*, part 1, 1976. ISBN 90 6196 140 8.

MCT 82 K.R. APT & J.W. DE BAKKER (eds), *Foundations of computer science II*, part 2, 1976. ISBN 90 6196 141 6.

MCT 83 L.S. BENTHEM JUTTING, *Checking Landau's "Grundlagen" in the AUTOMATH system*, 1979. ISBN 90 6196 147 5.

MCT 84 H.L.L. BUSARD, *The translation of the elements of Euclid from the Arabic into Latin by Hermann of Carinthia (?) books vii-xii*, 1977. ISBN 90 6196 148 3.

MCT 85 J. VAN MILL, *Supercompactness and Wallman spaces*, 1977. ISBN 90 6196 151 3.

MCT 86 S.G. VAN DER MEULEN & M. VELDHORST, *Torrix I, A programming system for operations on vectors and matrices over arbitrary fields and of variable size*. 1978. ISBN 90 6196 152 1.

*MCT 87 S.G. VAN DER MEULEN & M. VELDHORST, *Torrix II*, ISBN 90 6196 153 X.

MCT 88 A. SCHRIJVER, *Matroids and linking systems*, 1977. ISBN 90 6196 154 8.

MCT 89 J.W. DE ROEVER, *Complex Fourier transformation and analytic functionals with unbounded carriers*, 1978. ISBN 90 6196 155 6.

MCT 90 L.P.J. GROENEWEGEN, *Characterization of optimal strategies in dynamic games*, 1981 . ISBN 90 6196 156 4.

MCT 91 J.M. GEYSEL, *Transcendence in fields of positive characteristic*, 1979. ISBN 90 6196 157 2.

MCT 92 P.J. WEEDA, *Finite generalized Markov programming*, 1979. ISBN 90 6196 158 0.

MCT 93 H.C. TIJMS & J. WESSELS (eds), *Markov decision theory*, 1977. ISBN 90 6196 160 2.

MCT 94 A. BIJLSMA, *Simultaneous approximations in transcendental number theory*, 1978. ISBN 90 6196 162 9.

MCT 95 K.M. VAN HEE, *Bayesian control of Markov chains*, 1978. ISBN 90 6196 163 7.

MCT 96 P.M.B. VITÁNYI, *Lindenmayer systems: Structure, languages, and growth functions*, 1980. ISBN 90 6196 164 5.

MCT 97 A. FEDERGRUEN, *Markovian control problems; functional equations and algorithms*, 1983 . ISBN 90 6196 165 3.

MCT 98 R. GEEL, *Singular perturbations of hyperbolic type*, 1978. ISBN 90 6196 166 1.

MCT 99 J.K. LENSTRA, A.H.G. RINNOOY KAN & P. VAN EMDE BOAS, *Interfaces between computer science and operations research*, 1978. ISBN 90 6196 170 X.

MCT 100 P.C. BAAYEN, D. VAN DULST & J. OOSTERHOFF (eds), *Proceedings bicentennial congress of the Wiskundig Genootschap, part 1*, 1979. ISBN 90 6196 168 8.

MCT 101 P.C. BAAYEN, D. VAN DULST & J. OOSTERHOFF (eds), *Proceedings bicentennial congress of the Wiskundig Genootschap, part 2*, 1979. ISBN 90 6196 169 6.

MCT 102 D. VAN DULST, *Reflexive and superreflexive Banach spaces*, 1978. ISBN 90 6196 171 8.

MCT 103 K. VAN HARN, *Classifying infinitely divisible distributions by functional equations*, 1978. ISBN 90 6196 172 6.

MCT 104 J.M. VAN WOUWE, *Go-spaces and generalizations of metrizability*, 1979. ISBN 90 6196 173 4.

MCT 105 R. HELMERS, *Edgeworth expansions for linear combinations of order statistics*, 1982. ISBN 90 6196 174 2.

MCT 106 A. SCHRIJVER (ed.), *Packing and covering in combinatorics*, 1979. ISBN 90 6196 180 7.

MCT 107 C. DEN HEIJER, *The numerical solution of nonlinear operator equations by imbedding methods*, 1979. ISBN 90 6196 175 0.

MCT 108 J.W. DE BAKKER & J. VAN LEEUWEN (eds), *Foundations of computer science III*, part *1*, 1979. ISBN 90 6196 176 9.

MCT 109 J.W. DE BAKKER & J. VAN LEEUWEN (eds), *Foundations of computer science III*, part *2*, 1979. ISBN 90 6196 177 7.

MCT 110 J.C. VAN VLIET, *ALGOL 68 transput*, part *I*: *Historical review and discussion of the implementation model*, 1979. ISBN 90 6196 178 5.

MCT 111 J.C. VAN VLIET, *ALGOL 68 transput*, part *II*: *An implementation model*, 1979. ISBN 90 6196 179 3.

MCT 112 H.C.P. BERBEE, *Random walks with stationary increments and renewal theory*, 1979. ISBN 90 6196 182 3.

MCT 113 T.A.B. SNIJDERS, *Asymptotic optimality theory for testing problems with restricted alternatives*, 1979. ISBN 90 6196 183 1.

MCT 114 A.J.E.M. JANSSEN, *Application of the Wigner distribution to harmonic analysis of generalized stochastic processes*, 1979. ISBN 90 6196 184 X.

MCT 115 P.C. BAAYEN & J. VAN MILL (eds), *Topological Structures II*, part *1*, 1979. ISBN 90 6196 185 5.

MCT 116 P.C. BAAYEN & J. VAN MILL (eds), *Topological Structures II*, part *2*, 1979. ISBN 90 6196 186 6.

MCT 117 P.J.M. KALLENBERG, *Branching processes with continuous state space*, 1979. ISBN 90 6196 188 2.

MCT 118 P. GROENEBOOM, *Large deviations and asymptotic efficiencies*, 1980. ISBN 90 6196 190 4.

MCT 119 F.J. PETERS, *Sparse matrices and substructures, with a novel implementation of finite element algorithms*, 1980. ISBN 90 6196 192 0.

MCT 120 W.P.M. DE RUYTER, *On the asymptotic analysis of large-scale ocean circulation*, 1980. ISBN 90 6196 192 9.

MCT 121 W.H. HAEMERS, *Eigenvalue techniques in design and graph theory*, 1980. ISBN 90 6196 194 7.

MCT 122 J.C.P. BUS, *Numerical solution of systems of nonlinear equations*, 1980. ISBN 90 6196 195 5.

MCT 123 I. YUHÁSZ, *Cardinal functions in topology - ten years later*, 1980. ISBN 90 6196 196 3.

MCT 124 R.D. GILL, *Censoring and stochastic integrals*, 1980. ISBN 90 6196 197 1.

MCT 125 R. EISING, *2-D systems, an algebraic approach*, 1980. ISBN 90 6196 198 X.

MCT 126 G. VAN DER HOEK, *Reduction methods in nonlinear programming*, 1980. ISBN 90 6196 199 8.

MCT 127 J.W. KLOP, *Combinatory reduction systems*, 1980. ISBN 90 6196 200 5.

MCT 128 A.J.J. TALMAN, *Variable dimension fixed point algorithms and triangulations*, 1980. ISBN 90 6196 201 3.

MCT 129 G. VAN DER LAAN, *Simplicial fixed point algorithms*, 1980. ISBN 90 6196 202 1.

MCT 130 P.J.W. TEN HAGEN et al., *ILP Intermediate language for pictures*, 1980. ISBN 90 6196 204 8.

MCT 131 R.J.R. BACK, *Correctness preserving program refinements: Proof theory and applications*, 1980. ISBN 90 6196 207 2.

MCT 132 H.M. MULDER, *The interval function of a graph*, 1980. ISBN 90 6196 208 0.

MCT 133 C.A.J. KLAASSEN, *Statistical performance of location estimators*, 1981. ISBN 90 6196 209 9.

MCT 134 J.C. VAN VLIET & H. WUPPER (eds), *Proceedings international conference on ALGOL 68*, 1981. ISBN 90 6196 210 2.

MCT 135 J.A.G. GROENENDIJK, T.M.V. JANSSEN & M.J.B. STOKHOF (eds), *Formal methods in the study of language*, part *I*, 1981. ISBN 90 6196 211 0.

MCT 136 J.A.G. GROENENDIJK, T.M.V. JANSSEN & M.J.B. STOKHOF (eds), *Formal methods in the study of language*, part *II*, 1981. ISBN 90 6196 213 7.

MCT 137 J. TELGEN, *Redundancy and linear programs*, 1981. ISBN 90 6196 215 3.

MCT 138 H.A. LAUWERIER, *Mathematical models of epidemics*, 1981. ISBN 90 6196 216 1.

MCT 139 J. VAN DER WAL, *Stochastic dynamic programming, successive approximations and nearly optimal strategies for Markov decision processes and Markov games*, 1980. ISBN 90 6196 218 8.

MCT 140 J.H. VAN GELDROP, *A mathematical theory of pure exchange economies without the no-critical-point hypothesis*, 1981.
ISBN 90 6196 219 6.

MCT 141 G.E. WELTERS, *Abel-Jacobi isogenies for certain types of Fano three-folds*, 1981.
ISBN 90 6196 227 7.

MCT 142 H.R. BENNETT & D.J. LUTZER (eds), *Topology and order structures*, part 1, 1981.
ISBN 90 6196 228 5.

MCT 143 H. J.M. SCHUMACHER, *Dynamic feedback in finite- and infinite dimensional linear systems*, 1981.
ISBN 90 6196 229 3.

MCT 144 P. EIJGENRAAM, *The solution of initial value problems using interval arithmetic. Formulation and analysis of an algorithm*, 1981.
ISBN 90 6196 230 7.

MCT 145 A.J. BRENTJES, *Multi-dimensional continued fraction algorithms*, 1981. ISBN 90 6196 231 5.

MCT 146 C. VAN DER MEE, *Semigroup and factorization methods in transport theory*, 1982. ISBN 90 6196 233 1.

MCT 147 H.H. TIGELAAR, *Identification and informative sample size*, 1982.
ISBN 90 6196 235 8.

MCT 148 L.C.M. KALLENBERG, *Linear programming and finite Markovian control problems*, 1983. ISBN 90 6196 236 6.

MCT 149 C.B. HUIJSMANS, M.A. KAASHOEK, W.A.J. LUXEMBURG & W.K. VIETSCH, (eds), *From A to Z, proceeding of a symposium in honour of A.C. Zaanen*, 1982. ISBN 90 6196 241 2.

MCT 150 M. VELDHORST, *An analysis of sparse matrix storage schemes*, 1982.
ISBN 90 6196 242 0.

MCT 151 R.J.M.M. DOES, *Higher order asymptotics for simple linear Rank statistics*, 1982. ISBN 90 6196 243 9.

MCT 152 G.F. VAN DER HOEVEN, *Projections of Lawless sequences*, 1982.
ISBN 90 6196 244 7.

MCT 153 J.P.C. BLANC, *Application of the theory of boundary value problems in the analysis of a queueing model with paired services*, 1982.
ISBN 90 6196 247 1.

MCT 154 H.W. LENSTRA, JR. & R. TIJDEMAN (eds), *Computational methods in number theory, part I*, 1982.
ISBN 90 6196 248 X.

MCT 155 H.W. LENSTRA, JR. & R. TIJDEMAN (eds), *Computational methods in number theory, part II*, 1982.
ISBN 90 6196 249 8.

MCT 156 P.M.G. APERS, *Query processing and data allocation in distributed database systems*, 1983.
ISBN 90 6196 251 X.

MCT 157 H.A.W.M. KNEPPERS, *The covariant classification of two-dimensional smooth commutative formal groups over an algebraically closed field of positive characteristic*, 1983.
ISBN 90 6196 252 8.

MCT 158 J.W. DE BAKKER & J. VAN LEEUWEN (eds), *Foundations of computer science IV, Distributed systems*, part 1, 1983.
ISBN 90 6196 254 4.

MCT 159 J.W. DE BAKKER & J. VAN LEEUWEN (eds), *Foundations of computer science IV, Distributed systems*, part 2, 1983.
ISBN 90 6196 255 0.

MCT 160 A. REZUS, *Abstract automath*. 1983.
ISBN 90 6196 256 0.

MCT 161 G.F. HELMINCK, *Eisenstein series on the metaplectic group, An algebraic approach*, 1983.
ISBN 90 6196 257 9.

MCT 162 J.J. DIK, *Tests for preference*, 1983.
ISBN 90 6196 259 5

MCT 163 H. SCHIPPERS, *Multiple grid methods for equations of the second kind with applications in fluid mechanics*, 1983.
ISBN 90 6196 260 9.

MCT 164 F.A. VAN DER DUYN SCHOUTEN, *Markov decision processes with continuous time parameter*, 1983.
ISBN 90 6196 261 7.

MCT 165 P.C.T. VAN DER HOEVEN, *On point processes*, 1983.
ISBN 90 6196 262 5.

MCT 166 H.B.M. JONKERS, *Abstraction, specification and implementation techniques, with an application to garbage collection*, 1983.
ISBN 90 6196 263 3.

MCT 167 W.H.M. ZIJM, *Nonnegative matrices in dynamic programming*, 1983.
ISBN 90 6196 264 1

MCT 168 J.H. EVERTSE, *Upper bounds for the numbers of solutions of diophantine equations*, 1983.
ISBN 90 6196 265 X.

MCT 169 H.R. BENNETT & D.J. LUTZER (eds), *Topology and order structures*, part 2, 1983.
ISBN 90 6196 266 8.

An asterisk before the number means "to appear"