Printed at the Mathematical Centre, Kruislaan 413, Amsterdam, The Netherlands.

The Mathematical Centre, founded 11 February 1946, is a non-profit institution for the promotion of pure and applied mathematics and computer science. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

MATHEMATICAL CENTRE TRACTS 139

STOCHASTIC DYNAMIC PROGRAMMING

SUCCESSIVE APPROXIMATIONS AND NEARLY OPTIMAL STRATEGIES FOR MARKOV DECISION PROCESSES AND MARKOV GAMES

J. VAN DER WAL

SECOND EDITION

MATHEMATISCH CENTRUM AMSTERDAM 1984

1980 Mathematics subject classification: 90C47, 90D15

ISBN 90 6196 218 8

First printing 1981
Second edition 1984

ACKNOWLEDGEMENT

This study contains my thesis which has been written at the Department of Mathematics of the Eindhoven University of Technology.
I am most grateful to my thesis advisor Prof. Jaap Wessels for his guidance and his interest during the years that this research has been carried out.
I thank Mrs. Elsina Baselmans for her really excellent typing. And finally I thank the Mathematical Centre for letting me publish this monograph in their series of Mathematical Centre Tracts.

CONTENTS

CHAPTER 1

GENERAL INTRODUCTION

In this introductory chapter first (section 1) an informal description is given of the Markov decision processes and Markov games that will be studied. Next (section 2) we consider the optimality equations, also called the functional equations of dynamic programming. The optimality equations are the central point in practically each analysis of these decision problems. In section 3 a brief overview is given of the existing algorithms for the determination or approximation of the optimal value of the decision process. Section 4 indicates aims and results of this monograph while summarizing the contents of the following chapters. Then (section 5) we formally introduce the Markov decision process to be studied (the formal model description of the Markov game will be given later). We define the various strategies that will be distinguished, and introduce the criterion of total expected rewards and the criterion of average rewards per unit time. Finally, in section 6 some notations are introduced.

1.1. INFORMAL DESCRIPTION OF THE MODELS

This monograph deals with Markov decision processes and two-person zero-sum Markov (also called stochastic) games. Markov decision processes (MDP's) and Markov games (MG's) are mathematical models for the description of situations where one or more decision makers are controlling a dynamical system, e.g. in production planning, machine replacement or economics. In these models it is assumed that the Markov property holds. I.e., given the present state of the system, all information concerning the past of the system is irrelevant for its future behaviour.
Informally, an MDP can be described as follows:

Informal description of the MDP model

There is a dynamical system and a set of possible states it can occupy, called the *state space*, denoted by S. Here we only consider the case that S is finite or countably infinite.
Further, there is a set of actions, called the *action space*, denoted by A. At discrete points in time, t = 0,1,..., say, the system is observed by a controller or *decision maker*. At each decision epoch, the decision maker - having observed the present state of the system - has to choose an action from the set A. As a joint result of the state $i \in S$ and the action $a \in A$ taken in state i, the decision maker earns a (possibly negative) reward r(i,a), and the system moves to state j with probability p(i,a,j), $j \in S$, with $\sum_{j \in S} p(i,a,j) = 1$.

The situation in the two-person zero-sum game is very similar. Only, now there are two decision makers instead of one - usually called players - and two action sets, A for player I and B for player II. In the cases we consider, A and B are assumed to be finite. At each decision epoch, the players each choose - independently of the other - an action. As a result of the actions a of player I and b of player II in state i, player I receives a (possibly negative) payoff r(i,a,b) from player II (which makes the game zero-sum), and the system moves to state j with probability p(i,a,b,j), $j \in S$, with $\sum_{j \in S} p(i,a,b,j) = 1$.

The aim of the decision maker(s) is to control the system in such a way as to optimize some criterion function. Here two criteria will be considered, viz. the criterion of total expected rewards (including total expected discounted rewards), and the criterion of average rewards per unit time.

1.2. THE FUNCTIONAL EQUATIONS

Starting point in practically each analysis of MDP's and Markov games are the functional equations of dynamic programming.
Let us denote the optimal-value function for the total-reward MDP by v^*, i.e. $v^*(i)$ is the optimal value of the total-reward MDP for initial state i, $i \in S$. Then v^* is a solution of the optimality equation

$$(1.1) \qquad v(i) = (Uv)(i) := \max_{a \in A} \{r(i,a) + \sum_{j \in S} p(i,a,j)v(j)\} , \quad i \in S .$$

Or in functional notation

$$(1.2) \qquad v = Uv .$$

A similar functional equation arises in the total-reward Markov game. In that case $(Uv)(i)$ is the game-theoretical value of the matrix game with entries

$$r(i,a,b) + \sum_{j \in S} p(i,a,b,j)v(j) , \quad a \in A , \quad b \in B .$$

In many publications on MDP's and MG's the operator U is a contraction. For example, in SHAPLEY [1953], where the first formulation of a Markov game is given, there is an absorbing state, * say, where no more returns are obtained, with $p(i,a,b,*) > 0$ for all i, a and b. Since S, A and B are in Shapley's case finite, this implies that the game will end up in *, and that U is a contraction and hence has a unique fixed point. Shapley used this to prove that this fixed point is the value of the game and that there exist optimal stationary strategies for both players.
In many of the later publications the line of reasoning is similar to that in Shapley's paper.

In the average reward MDP the optimal-value function, g^* say, usually called the gain of the MDP, is part of a solution of a pair of functional equations in g and v:

$$(1.3) \qquad g(i) = \max_{a \in A} \sum_{j \in S} p(i,a,j)g(j) ,$$

$$(1.4) \qquad v(i) + g(i) = \max_{a \in \bar{A}(i)} \{r(i,a) + \sum_{j \in S} p(i,a,j)v(j)\} ,$$

where $\bar{A}(i)$ denotes the set of maximizers in (1.3).

In the first paper on MDP's, BELLMAN [1957] considered the average-reward MDP with finite state and action spaces. Under an additional condition, guaranteeing that g^* is a constant function (i.e. the gain of the MDP is independent of the initial state), Bellman studied the functional equations (1.3) and (1.4) and the dynamic programming recursion

(1.5) $\quad v_{n+1} = Uv_n \ , \quad n = 0,1,\ldots \ ,$

where U is defined as in (1.1).
He proved that $v_n - ng^*$ is bounded, i.e., the optimal n-stage reward minus n times the optimal average reward is bounded. Later BROWN [1965] proved that $v_n - ng^*$ is bounded for every MDP, and only around 1978 a relatively complete treatment of the behaviour of $v_n - ng^*$ has been given by SCHWEITZER and FEDERGRUEN [1978], [1979].

The situation in the average-reward Markov game is more complicated. In 1957, GILLETTE [1957] made a first study of the finite state and action average-reward MG. Under a rather restrictive condition, which implies the existence of a solution to a pair of functional equations similar to (1.3) and (1.4) with g a constant function, he proved that the game has a value and that stationary optimal strategies for both players exist. He also described a game for which the pair of functional equations has no solution. BLACKWELL and FERGUSON [1968] showed that this game does have a value; only recently it has been shown by MONASH [1979] and, independently, by MERTENS and NEYMAN [1980] that every average-reward MG with finite state and action spaces has a value.

1.3. REVIEW OF THE EXISTING ALGORITHMS

An important issue in the theory of MDP's and MG's is the determination, usually approximation, of v^* (in the average-reward case g^*) and the determination of (nearly-) optimal, preferably stationary, strategies. This also is the main topic in this study.

Since in the total-reward case, for the MDP as well as for the MG, v^* is a solution of an optimality equation of the form $v = Uv$, one can try to approximate v^* by the standard successive approximation scheme

$$v_{n+1} = Uv_n \ , \quad n = 0,1,\ldots \ .$$

If U is a contraction, as in Shapley's case, then v_n will converge to v^*. Further, the contractive properties of U enable us to obtain bounds on v^* and nearly optimal stationary strategies; see for the MDP a.o. MAC QUEEN [1966], PORTEUS [1971], [1975] and Van NUNEN [1976a], and for the MG a.o. CHARNES and SCHROEDER [1967], KUSHNER and CHAMBERLAIN [1969] and Van der WAL [1977a].

For this contracting case various other successive approximation schemes have been proposed. Viz., for the MDP the Gauss-Seidel method by HASTINGS [1968] and an overrelaxation algorithm by REETZ [1973], and for the MG the Gauss-Seidel method by KUSHNER and CHAMBERLAIN [1969]. As has been shown by WESSELS [1977a], Van NUNEN and WESSELS [1976], Van NUNEN [1976a], Van NUNEN and STIDHAM [1978] and Van der WAL [1977a], these algorithms can be described and studied very well in terms of the go-ahead functions by which they may be generated.

The so-called value-oriented methods, first mentioned by PORTEUS [1971], and extensively studied by Van NUNEN [1976a], [1976c], are another type of algorithms. In the value-oriented approach each optimization step is followed by a kind of extrapolation step. Howard's classic policy iteration algorithm [HOWARD,1960] can be seen as an extreme element of this set of methods, since in this algorithm each optimization step is followed by an extrapolation in which the value of the maximizing policy is determined. The finite contracting MDP can also be solved by a linear programming approach, see d'EPENOUX [1960]. Actually, the policy iteration method is equivalent to a linear program where it is allowed to change more than one basic variable at a time, cf. WESSELS and Van NUNEN [1975].

If U is not a contraction, then the situation becomes more complicated. For example, v_n need no longer converge to v^*. And even if v_n converges to v^*, it is in general not possible to decide whether v_n is already close to v^* and to detect nearly-optimal (stationary) strategies from the successive approximations scheme.

For the average reward MDP there exists by now, as mentioned before, a relatively complete treatment of the method of standard successive approximations, see SCHWEITZER and FEDERGRUEN [1978], [1979].

Alternatively, one can use Howard's policy iteration method [HOWARD,1960], which, in a slightly modified form, always converges, see BLACKWELL [1962]. Furthermore, several authors have studied the relation between the average-reward MDP and the discounted MDP with discountfactor tending to one, see e.g. HOWARD [1960], BLACKWELL [1962], VEINOTT [1966], MILLER and VEINOTT [1969] and SLADKY [1974]. This has resulted for example in Veinott's extended version of the policy iteration method which yields strategies that are stronger than merely average optimal.
Another algorithm that is based on the relation between the discounted and the average-reward MDP, is the unstationary successive approximations method of BATHER [1973] and HORDIJK and TIJMS [1975]. In this algorithm the average-reward MDP is approximated by a sequence of discounted MDP's with discountfactor tending to one.
Also, there is the method of value-oriented successive approximations, which has been proposed for the average-reward case, albeit without convergence proof, by MORTON [1971].
And finally, one may use the method of linear programming, cf. De GHELLINCK [1960], MANNE [1960], DENARDO and FOX [1968], DENARDO [1970], DERMAN [1970], HORDIJK and KALLENBERG [1979] and KALLENBERG [1980].

The situation is essentially different for the average-reward MG. In general, no nearly-optimal Markov strategies exist, which implies that nearly-optimal strategies cannot be obtained with the usual dynamic programming methods. Only in special cases the methods described above will be of use, see e.g. GILLETTE [1957], HOFFMAN and KARP [1966], FEDERGRUEN [1977], and Van der WAL [1980].

1.4. SUMMARY OF THE FOLLOWING CHAPTERS

Roughly speaking one may say that this monograph deals mainly with various dynamic programming methods for the approximation of the value and the determination of nearly-optimal stationary strategies in MDP's and MG's. We study the more general use of several dynamic programming methods, which were previously used only in more specific models (e.g. the contracting MDP). This way we fill a number of gaps in the theory of dynamic programming for MDP's and MG's.

Our intentions and results are described in some more detail in the following summary of the various chapters.
The contents of this book can be divided into three parts. Part 1, chapters 2-5, considers the total-reward MDP, part 2, chapters 6-9, deals with the average-reward MDP, and in part 3, chapters 10-13, some two-person zero-sum MG's are treated.

In chapter 2 we study the total-reward MDP with countable state space and general action space. First it is shown that it is possible to restrict the considerations to randomized Markov strategies. Next some properties are given of the various dynamic programming operators. Then the finite-stage MDP and the optimality equation are considered. These results are used to prove that one can restrict oneself even to pure Markov strategies (in this general setting this result is due to Van HEE [1978a]).
This chapter will be concluded with a number of results on the existence or nonexistence of nearly-optimal strategies with certain special properties, e.g. stationarity. Some of the counterexamples may be new, and it seems that also theorem 2.22 is new.
In chapter 3 the various successive approximation methods are introduced for the MDP model of chapter 2. First a review is given of several results for the method of standard successive approximations. Then, in this general setting, the set of successive approximation algorithms is formulated in terms of go-ahead functions, introduced and studied for the contracting MDP by WESSELS [1977a], Van NUNEN and WESSELS [1976], Van NUNEN [1976a], and Van NUNEN and STIDHAM [1978]. Finally, the method of value-oriented successive approximations is introduced. This method was first mentioned for the contracting MDP by PORTEUS [1971], and studied by Van NUNEN [1976c]. In general, these methods do not converge.
Chapter 4 deals with the so-called strongly convergent MDP (cf. Van HEE and Van der WAL [1977] and Van HEE, HORDIJK and Van der WAL [1977]). In this model it is assumed that the sum of all absolute rewards is finite, and moreover that the sum of the absolute values of the rewards from time n onwards tends to zero if n tends to infinity, uniformly in all strategies. It is shown that this condition guarantees the convergence of the successive approximation methods generated by nonzero go-ahead functions, i.e., the convergence of v_n to v^*. Further, we study under this condition the value-oriented method and it is shown that the monotonic variant, and therefore also the policy iteration method, always converges.

In chapter 5 the contracting MDP is considered. We establish the (essential) equivalence of four different models for the contracting MDP, and we review some results on bounds for v^* and on nearly-optimal strategies.
Further, for the discounted MDP with finite state and action spaces, some Laurent series expansions are given (for example for the total expected discounted reward of a stationary strategy) and the more sensitive optimality criteria are formulated (cf. MILLER and VEINOTT [1969]). The results of this chapter are needed in chapters 6-8 and 11.
In chapter 6 the average-reward MDP with finite state and action spaces is introduced. This chapter serves as an introduction to chapters 7-9, and for the sake of self-containedness we review several results on the existence of optimal stationary strategies, the policy iteration method and the method of standard successive approximations.
Chapter 7 deals with the more sensitive optimality criteria in the discounted and the average-reward MDP and re-establishes the equivalence of k-discount optimality and (k+1)-order average optimality. This equivalence was first shown by LIPPMAN [1968] (for a special case) and by SLADKY [1974]. We reprove this result using an unstationary successive approximation algorithm. As a bonus of this analysis a more general convergence proof is obtained for the algorithm given by BATHER [1973] and some of the algorithms given by HORDIJK and TIJMS [1975].
In chapter 8 it is shown that in the policy iteration algorithm the improvement step can be replaced by a maximization step formulated in terms of go-ahead functions (cf. WESSELS [1977a] and Van NUNEN and WESSELS [1976]). In the convergence proof we use the equivalence of average and discounted optimality criteria that has been established in chapter 7. A special case of the policy iteration methods obtained in this way is Hastings' Gauss-Seidel variant, cf. HASTINGS [1968].
Chapter 9 considers the method of value-oriented successive approximations, which for the average-reward MDP has been first formulated, without convergence proof, by MORTON [1971]. Under two conditions: a strong aperiodicity assumption (which is no real restriction) and a condition guaranteeing that the gain is independent of the initial state, it is shown that the method yields arbitrary close bounds on g^*, and nearly-optimal stationary strategies.
Chapter 10 gives an introduction to the two-person zero-sum Markov game. It will be shown that the finite-stage problem can be 'solved' by a

dynamic programming approach, so that we can restrict ourselves again to (randomized) Markov strategies. We also show that the restriction to Markov strategies in the nonzero-sum game may be rather unrealistic.
In chapter 11 the contracting MG is studied. For the successive approximation methods generated by nonzero go-ahead functions we obtain bounds on v^* and nearly-optimal stationary strategies. These results are very similar to the ones in the contracting MDP (chapter 5). Further, for this model the method of value-oriented successive approximations is studied, which contains the method of HOFFMAN and KARP [1966] as a special case.
Chapter 12 deals with the so-called positive MG. In this game it is assumed that $r(i,a,b) \geq c > 0$ for all i, a and b and some constant c, thus the second player looses at least an amount c in each step. However, he can restrict his losses by terminating the game at certain costs (modeled as a transition to an extra absorbing state in which no more payoffs are obtained). We show that in this model the method of standard successive approximations provides bounds on v^* and nearly-optimal stationary strategies for both players.
Finally, in chapter 13, the method of standard successive approximations is studied for the average-reward Markov game with finite state (and action) space(s). Under two restrictive conditions, which imply that the value of the game is independent of the initial state, it is shown that the method yields good bounds on the value of the game, and nearly-optimal stationary strategies for both players.

1.5. FORMAL DESCRIPTION OF THE MDP MODEL

In this section a formal characterization is given of the MDP. The formal model of the Markov game will be given in chapter 10.

Formally, an MDP is characterized by the following objects.

S: a nonempty finite or countably infinite set S, called the state space, together with the σ-field $\mathcal{S}$ of all its subsets.

A: an arbitrary nonempty set A, called the action space, with a σ-field $\mathcal{A}$ containing all one-point sets.

p: a transition probability function $p: S \times A \times S \to [0,1]$, called the *transition law*. I.e., $p(i,a,\cdot)$ induces for all $(i,a) \in S \times A$ a proba-

bility measure on $(S,\mathcal{S})$ and $p(i,\cdot,j)$ is $\mathcal{A}$-measurable for all $i,j \in S$.

r: a real-valued function r on $S \times A$ called the *reward function*, where we require that $r(i,\cdot)$ is $\mathcal{A}$-measurable for all $i \in S$.

At discrete points in time, $t = 0,1,\ldots$ say, a decision maker, having observed the state of the MDP, chooses an action, as a result of which he earns some immediate reward according to the function r and the MDP reaches a new state according to the transition law p.

In the sequel also state-dependent action sets, notation $A(i)$, $i \in S$, will be encountered. This can be modeled in a similar way. We shall not pursue this here.
Also it is assumed that $p(i,a,\cdot)$ is, for all i and a, a probability measure, whereas MDP's are often formulated in terms of defective probabilities. Clearly, these models can be fitted in our framework by the addition of an extra absorbing state.

In order to control the system the decision maker may choose a decision rule from a set of control functions satisfying certain measurability conditions. To describe this set, define

$$H_0 := S\,, \quad H_n := (S \times A)^n \times S\,, \quad n = 1,2,\ldots\,.$$

So, H_n is the set of possible histories of the system starting at time 0 upto time n, i.e., the sequence of preceding states of the system, the actions taken previously and the present state of the system. We assume that this information is available to the decision maker at time n.
On H_n we introduce the product σ-field $\mathcal{H}_n$ generated by $\mathcal{S}$ and $\mathcal{A}$.
Then a decision rule π the decision maker is allowed to use, further called *strategy*, is any sequence $\pi_0,\pi_1,\ldots$ such that the function π_n, which prescribes the action to be taken at time n, is a transition probability from H_n into A. So, let $\pi_n(C|h_n)$ denote for all sets $C \in \mathcal{A}$ and for all histories $h_n \in H_n$ the probability that at time n given the history h_n an action from the set C will be chosen, then $\pi_n(C|\cdot)$ is $\mathcal{H}_n$-measurable for all $C \in \mathcal{A}$ and $\pi_n(\cdot|h_n)$ is a probability measure on $(A,\mathcal{A})$ for all $h_n \in H_n$.
Notation: $\pi = (\pi_0,\pi_1,\ldots)$. Thus we allow for randomized and history-dependent strategies. The set of all strategies will be denoted by Π.
A subset of Π is the set RM of the so-called *randomized Markov strategies*.
A strategy $\pi \in \Pi$ belongs to RM if for all $n = 1,2,\ldots$, for all

$h_n = (i_0,a_0,\ldots,i_n) \in H_n$ and for all $C \in \mathcal{A}$, the probability $\pi_n(C|h_n)$ depends on h_n only through the present state i_n.

The set M of all pure Markov strategies, or shortly *Markov strategies*, is the set of all $\pi \in RM$ for which there exists a sequence $f_0,f_1,\ldots$ of mappings from S into A such that for all $n = 0,1,\ldots$ and for all $(i_0,a_0,\ldots,i_n) \in H_n$ we have

$$\pi_n(\{f_n(i_n)\} \mid (i_0,a_0,\ldots,i_n)) = 1 \ .$$

Usually a Markov strategy will be denoted by the functions $f_0,f_1,\ldots$ characterizing it: $\pi = (f_0,f_1,\ldots)$.

A mapping f from S into A is called a *policy*. The set of all policies is denoted by F.

A *stationary strategy* is any strategy $\pi = (f,f_1,f_2,\ldots) \in M$ with $f_n = f$ for all $n = 1,2,\ldots$; notation $\pi = f^{(\infty)}$. When it is clear from the context that a stationary strategy is meant, we usually write f instead of $f^{(\infty)}$.

Note that since it has been assumed that $\mathcal{A}$ contains all one-point sets, any sequence $f_0,f_1,\ldots$ of policies actually gives a strategy $\pi \in M$.

Each strategy $\pi = (\pi_0,\pi_1,\ldots) \in \Pi$ generates a sequence of transition probabilities p_n from H_n into $A \times S$ as follows: For all $C \in \mathcal{A}$ and $D \in \mathcal{S}$

$$p_0(C \times D \mid i_0) = \int_C \pi_0(da \mid i_0) \sum_{j \in D} p(i_0,a,j) \ , \quad i \in S$$

and for $n = 1,2,\ldots$

$$p_n(C \times D \mid (i_0,a_0,\ldots,i_n)) = \int_C \pi_n(da \mid (i_0,a_0,\ldots,i_n) \sum_{j \in D} p(i_n,a,j) \ ,$$

$$(i_0,a_0,\ldots,i_n) \in H_n \ .$$

Endow $\Omega := (S \times A)^\infty$, the set of possible realizations of the process, with the product σ-field generated by $\mathcal{S}$ and $\mathcal{A}$. Then for each $\pi \in \Pi$, the sequence of transition probabilities $\{p_n\}$ defines for each initial state $i \in S$ a probability measure $\mathbb{P}_{i,\pi}$ on Ω and a stochastic process $\{(X_n,A_n), n = 0,1,\ldots\}$, where X_n denotes the state of the system at time n and A_n the action chosen at time n.

The expectation operator with respect to the probability measure $\mathbb{P}_{i,\pi}$ will be denoted by $\mathbb{E}_{i,\pi}$.

Now we can define the *total expected reward*, when the process starts in state $i \in S$ and strategy $\pi \in \Pi$ is used:

$$(1.6) \qquad v(i,\pi) := \mathbb{E}_{i,\pi} \sum_{n=0}^{\infty} r(X_n,A_n) \ ,$$

whenever the expectation at the right hand side is well-defined.
In order to guarantee this, we assume the following condition to be fulfilled throughout chapters 2-5, where the total-reward MDP is considered.

CONDITION 1.1. *For all* $i \in S$ *and* $\pi \in \Pi$

$$(1.7) \qquad u(i,\pi) := \mathbb{E}_{i,\pi} \sum_{n=0}^{\infty} r^+(X_n,A_n) < \infty \ ,$$

where

$$r^+(i,a) := \max \{0,r(i,a)\} \ , \quad i \in S \ , \quad a \in A \ .$$

Condition 1.1 allows us to interchange expectation and summation in (1.6), and implies

$$(1.8) \qquad \lim_{n\to\infty} v_n(i,\pi) = v(i,\pi) \ ,$$

where

$$(1.9) \qquad v_n(i,\pi) := \mathbb{E}_{i,\pi} \sum_{k=0}^{n-1} r(X_k,A_k) \ .$$

The *value* of the total-reward MDP is defined by

$$(1.10) \qquad v^*(i) := \sup_{\pi\in\Pi} v(i,\pi) \ , \quad i \in S \ .$$

REMARK 1.2. An alternative criterion is that of *total expected discounted rewards*, where it is assumed that a unit reward earned at time n is worth only β^n at time 0, with β, $0 \le \beta < 1$, the discountfactor.
The total expected β-discounted reward when the process starts in state $i \in S$ and strategy $\pi \in \Pi$ is used, is defined by

$$(1.11) \qquad v_\beta(i,\pi) := \mathbb{E}_{i,\pi} \sum_{n=0}^{\infty} \beta^n r(X_n,A_n) \ ,$$

whenever the expectation is well-defined.

The discounted MDP can be fitted into the framework of total expected rewards by incorporating the discountfactor into the transition probabilities and adding an extra absorbing state, * say, as follows:
Let the discounted MDP be characterized by the objects S, A, p, r and β, then define a transformed MDP characterized by $\hat{S},\hat{A},\hat{p},\hat{r}$ with $\hat{S} = S \cup \{*\}$, $* \notin S$, $\hat{A} = A$, $\hat{r}(i,a) = r(i,a)$, $\hat{r}(*,a) = 0$, $\hat{p}(i,a,j) = \beta p(i,a,j)$, $\hat{p}(i,a,*) = 1 - \beta$ and $\hat{p}(*,a,*) = 1$ for all $i,j \in S$ and $a \in A$.
Then, clearly, for all $i \in S$ and $\pi \in \Pi$ the total expected reward in the transformed MDP is equal to the total expected β-discounted reward in the original problem.
Therefore we shall not consider the discounted MDP explicitly, except for those cases where we want to study the relation between the average reward MDP and the β-discounted MDP with β tending to one.

The second criterion that is considered is the criterion of *average reward per unit time*.
The average reward per unit time for initial state $i \in S$ and strategy $\pi \in \Pi$ is defined by (cf. (1.9))

$$(1.12) \qquad g(i,\pi) := \liminf_{n\to\infty} n^{-1} v_n(i,\pi) .$$

Since this criterion is considered only for MDP's with finite state and action spaces, $g(i,\pi)$ is always well-defined.
The value of the average-reward MDP is defined by

$$(1.13) \qquad g^*(i) := \sup_{\pi\in\Pi} g(i,\pi) , \quad i \in S .$$

1.6. NOTATIONS

This introductory chapter will be concluded with a number of notations and conventions.

$\mathbb{R}$:= the set of real numbers,

$\bar{\mathbb{R}} := \mathbb{R} \cup \{-\infty\}$.

For any $x \in \bar{\mathbb{R}}$ we define

$$x^+ := \max\{x,0\} ,$$

$$x^- := \min\{x,0\} .$$

So,

$$x = x^+ + x^- \quad \text{and} \quad |x| = x^+ - x^- .$$

The set of all real-valued functions on S is denoted by V:

(1.14) $\quad V := \{v: S \to \mathbb{R}\}$

and $\bar{V}$ denotes the set

(1.15) $\quad \bar{V} := \{v: S \to \bar{\mathbb{R}}\}$.

For any v and $w \in \bar{V}$ we write

$$v < 0 \quad \text{if} \quad v(i) < 0 \quad \text{for all } i \in S ,$$

and

$$v < w \quad \text{if} \quad v(i) < w(i) \quad \text{for all } i \in S .$$

Similarly, if < is replaced by $\leq$, =, $\geq$ or >.
For a function v from S into $\bar{\mathbb{R}} \cup \{+\infty\}$ we write

$$v < \infty \quad \text{if} \quad v(i) < \infty \quad \text{for all } i \in S, \text{ so if } v \in \bar{V} .$$

For any $v \in \bar{V}$ define the elements v^+ and v^- in $\bar{V}$ by

(1.16) $\quad v^+(i) := (v(i))^+ , \quad i \in S$

and

(1.17) $\quad v^-(i) := (v(i))^- , \quad i \in S .$

For any $v \in V$ the function $|v| \in V$ is defined by

(1.18) $\quad |v|(i) := |v(i)| , \quad i \in S .$

The unit function on S is denoted by e:

(1.19) $\quad e(i) = 1 \quad \text{for all } i \in S ,$

If, in an expression defined for all $i \in S$, the subscript or argument corresponding to the state i is omitted, then the corresponding function on S is meant. For example, $v(\pi)$, $u(\pi)$ and $g(\pi)$ are the elements in $\bar{V}$ with i-th component $v(i,\pi)$, $u(i,\pi)$ and $g(i,\pi)$, respectively. Similarly, if in $\mathbb{P}_{i,\pi}(\cdot)$ or $\mathbb{E}_{i,\pi}(\cdot)$ the subscript i is omitted, then we mean the corresponding function on S.

Let $\mu \in V$ satisfy $\mu \geq 0$, then the mapping $\| \ \|_\mu$ from V into $\mathbb{R} \cup \{+\infty\}$ is defined by

(1.20) $\quad \|v\|_\mu := \inf \{c \in \mathbb{R} \mid |v| \leq c\mu\} , \quad v \in V ,$

where, by convention, the infimum of the empty set is equal to $+\infty$.
The subspaces V_μ of V and V_μ^+ of $\bar{V}$ are defined by

(1.21) $\quad V_\mu := \{v \in V \mid \|v\|_\mu < \infty\}$

and

(1.22) $\quad V_\mu^+ := \{v \in \bar{V} \mid v^+ \in V_\mu\} .$

The space V_μ with norm $\|v\|_\mu$ is a Banach space.

In the analysis of the MDP a very important role will be played by the Markov strategies and therefore by the policies. For that reason the following notations are very useful. For any $f \in F$ let the real-valued function $r(f)$ on S and the mapping $P(f)$ from $S \times S$ into $[0,1]$ be defined by

(1.23) $\quad (r(f))(i) := r(i,f(i)) , \quad i \in S$

and

(1.24) $\quad (P(f))(i,j) := p(i,f(i),j) , \quad i,j \in S .$

Further we define, for all $v \in \bar{V}$ for which the expression at the right hand side is well defined,

(1.25) $\quad (P(f)v)(i) := \sum_{j \in S} p(i,f(i),j)v(j) , \quad i \in S , \ f \in F ,$

(1.26) $\quad \tilde{U}v := \sup_{f \in F} P(f)v ,$

(1.27) $\quad L(f)v := r(f) + P(f)v , \quad f \in F ,$

(1.28) $\quad Uv := \sup_{f \in F} L(f)v ,$

(1.29) $\quad L^+(f)v := (r(f))^+ + P(f)v , \quad f \in F ,$

(1.30) $\quad U^+v := \sup_{f \in F} L^+(f)v ,$

(1.31) $\quad L^{abs}(f)v := |r(f)| + P(f)v , \quad f \in F ,$

(1.32) $\quad U^{abs}v := \sup_{f \in F} L^{abs}(f)v ,$

where the suprema are defined componentwise.

Finally, we define the following functions on S:

$$u^*(i) := \sup_{\pi \in \Pi} u(i,\pi) \ , \quad i \in S \ , \tag{1.33}$$

$$z(i,\pi) := \mathbb{E}_{i,\pi} \sum_{n=0}^{\infty} |r(X_n,A_n)| \ , \quad i \in S \ , \ \pi \in \Pi \ , \tag{1.34}$$

$$z^*(i) := \sup_{\pi \in \Pi} z(i,\pi) \ , \quad i \in S \ , \tag{1.35}$$

$$w(i,\pi) := \mathbb{E}_{i,\pi} \sum_{n=0}^{\infty} r^-(X_n,A_n) \ , \quad i \in S \ , \ \pi \in \Pi \ , \tag{1.36}$$

$$w^*(i) := \sup_{\pi \in \Pi} w(i,\pi) \ , \quad i \in S \ . \tag{1.37}$$

CHAPTER 2

THE GENERAL TOTAL-REWARD MDP

2.1. INTRODUCTION

In this chapter we will perform a first analysis on the general total-reward MDP model formulated in section 1.5.
Throughout this chapter we assume condition 1.1:

(2.1) $\quad u(\pi) < \infty$ for all $\pi \in \Pi$.

A major issue in this chapter is the proof of the following result due (in this general setting) to Van HEE [1978a]:

(2.2) $\quad \sup_{\pi \in M} v(i,\pi) = v^*(i)$, $i \in S$.

I.e., when optimizing $v(i,\pi)$ one needs to consider only Markov strategies. The proof given here is essentially Van Hee's, but the steps are somewhat more elementary.
While establishing (2.2) we will obtain a number of results of independent interest.

First (in section 2) an extension of a theorem of DERMAN and STRAUCH [1966] given by HORDIJK [1974] is used to prove that for a fixed initial state i any strategy $\pi \in \Pi$ can be replaced by a strategy $\pi' \in RM$ which yields the same marginal distributions for the process $\{(X_n,A_n), n = 0,1,\ldots\}$. This implies that in the optimization of $v(i,\pi)$, $u(i,\pi)$, etc., one needs to consider only randomized Markov strategies. Hordijk's result is even stronger and also implies that $u^* < \infty$.
Further it is shown in this preliminary section that the mappings $P(f)$, $\tilde{U}$, $L(f)$, U, $L^+(f)$ and U^+ defined in (1.25)-(1.30), are in fact operators on $V^+_{u^*}$, i.e. they map $V^+_{u^*}$ into itself. These operators will play an important role in our further analysis, particularly in the study of successive

approximation methods.
A first use of these operators is made in section 3, where it is shown that the finite-horizon MDP can be treated by a dynamic programming approach. This implies that in the finite-horizon MDP one needs to consider only Markov strategies.
The results for the finite-horizon case imply that also $u(i,\pi)$ is optimized within the set of Markov strategies:

$$\sup_{\pi \in M} u(i,\pi) = u^*(i) \ , \quad i \in S \ . \tag{2.3}$$

Next, in section 4, we consider the optimality equation

$$v = Uv \ , \tag{2.4}$$

and we show that v^* is a (in general not unique) solution of this equation. In section 5 it is shown that, if $v^* \leq 0$, the fact that v^* satisfies (2.4) implies the existence of a nearly-optimal Markov strategy uniformly in the initial state. I.e., there exists a Markov strategy π such that

$$v(\pi) \geq v^* - \varepsilon e \ . \tag{2.5}$$

In section 6 we prove (2.2) using the fact that in finite-stage MDP's one may restrict oneself to Markov strategies and using the existence of a uniformly nearly-optimal Markov strategy in ∞-stage MDP's with a nonpositive value.
Finally, in section 7, we present a number of results on nearly-optimal strategies. One of our main results is: if A is finite, then for each initial state $i \in S$ there exists a nearly-optimal stationary strategy.

2.2. *SOME PRELIMINARY RESULTS*

In this section we first want to prove that we can restrict ourselves to randomized Markov strategies and that condition 1.1 implies that $u^* < \infty$. To this end we use the following generalization of a result of DERMAN and STRAUCH [1966], given by HORDIJK [1974, theorem 13.2].

LEMMA 2.1. *Let $\pi^{(1)},\pi^{(2)},\ldots$ be an arbitrary sequence of strategies and let $c_1,c_2,\ldots$ be a sequence of nonnegative real numbers with $\Sigma_{k=1}^{\infty} c_k = 1$.*

Then there exists for each $i \in S$ *a strategy* $\pi \in RM$ *such that*

$$(2.6) \qquad \mathbb{P}_{i,\pi}(X_n = j,\ A_n \in C) = \sum_{k=1}^{\infty} c_k \mathbb{P}_{i,\pi(k)}(X_n = j,\ A_n \in C) ,$$

for all $j \in S$, *all* $C \in \mathcal{A}$ *and all* $n = 0,1,\ldots$.

PROOF. Let $(\pi_0,\pi_1,\ldots) \in RM$ be defined by

$$\pi_n(C|j) = \frac{\sum_{k=1}^{\infty} c_k \mathbb{P}_{i,\pi(k)}(X_n = j,\ A_n \in C)}{\sum_{k=1}^{\infty} c_k \mathbb{P}_{i,\pi(k)}(X_n = j)}$$

for all $j \in S$, for all $n = 0,1,\ldots$ and all $C \in \mathcal{A}$, whenever the denominator is nonzero. Otherwise, let $\pi_n(\cdot|j)$ be an arbitrary probability measure on $(A,\mathcal{A})$.
Then one can prove by induction that $\pi = (\pi_0,\pi_1,\ldots)$ satisfies (2.6) for all $j \in S$, all $C \in \mathcal{A}$ and all $n = 0,1,\ldots$. For details, see HORDIJK [1974]. □

The special case of this lemma with $c_1 = 1$, $c_n = 0$, $n = 0,1,\ldots$, shows that any strategy $\pi^{(1)} \in \Pi$ can be replaced by a strategy $\pi \in RM$ having the same marginal distributions for the process $\{(X_n,A_n),\ n = 0,1,\ldots\}$. This leads to the following result:

COROLLARY 2.2. *For each initial state* $i \in S$ *and each* $\pi \in \Pi$ *there exists a strategy* $\hat{\pi} \in RM$ *such that*

$$v(i,\pi) = v(i,\hat{\pi}) .$$

Therefore

$$\sup_{\pi \in \Pi} v(i,\pi) = \sup_{\pi \in RM} v(i,\pi) .$$

Similarly, if v *is replaced by* v_n, u *or* z.

Since for corollary 2.2 to hold with v replaced by u, condition 1.1 is not needed, it follows from this corollary that condition 1.1 is equivalent to:

$$u(\pi) < \infty \quad \text{for all } \pi \in RM .$$

Another way in which one can use lemma 2.1 is the following.
Suppose that in order to control the process we want to use one strategy out of a countable set $\{\pi^{(1)},\pi^{(2)},\ldots\}$. In order to decide which strategy to play, we start with a random experiment which selects strategy $\pi^{(k)}$ with probability c_k. Then formally this compound decision rule is not a strategy in the sense of section 1.5 (as the prescribed actions do not depend on the history of the process only, but also on the outcome of the random experiment). Lemma 2.1 now states that, although this decision rule is not a strategy, there exists a strategy $\pi \in RM$ which produces the same marginal distributions for the process as the compound strategy described above. Using lemma 2.1 in this way, we can prove the following theorem.

THEOREM 2.3. *For all* $i \in S$,

$$u^*(i) < \infty .$$

PROOF. Suppose that for some $i \in S$ we have $u^*(i) = \infty$. Then there exists a sequence $\pi^{(1)},\pi^{(2)},\ldots$ of strategies with $u(i,\pi^{(k)}) \geq 2^k$. Now, applying lemma 2.1 with $c_k = 2^{-k}$, $k = 1,2,\ldots$, we find a strategy $\pi \in RM$ satisfying (2.6). For this strategy π we then have

$$u(i,\pi) = \sum_{k=1}^{\infty} c_k u(i,\pi^{(k)}) \geq \sum_{k=1}^{\infty} 2^{-k} 2^k = \infty .$$

But this would contradict condition 1.1. Hence $u^*(i) < \infty$ for all $i \in S$. □

Since, clearly, $v(\pi) \leq u(\pi)$ for all $\pi \in \Pi$, theorem 2.3 immediately yields

COROLLARY 2.4. *For all* $i \in S$,

$$v^*(i) < \infty .$$

In the second part of this section we study the mappings $P(f)$, $L(f)$, etc. It will be shown that these mappings are in fact operators on the space $V^+_{u^*}$. First we prove

LEMMA 2.5. *For all* $f \in F$,

$$L^+(f)u^* \leq u^* .$$

PROOF. Choose $f \in F$ and $\varepsilon > 0$ arbitrarily. As we shall show at the end of the proof, there exists a strategy $\pi \in \Pi$ satisfying $u(\pi) \geq u^* - \varepsilon e$. Further, the decision rule: "use policy f at time 0 and continue with strategy π at time 1 (pretending the process to restart at time 1)" is also an element of Π. Thus, denoting this strategy by $f \circ \pi$, we have

$$L^+(f)u^* \leq (r(f))^+ + P(f)(u(\pi) + \varepsilon e)$$
$$= (r(f))^+ + P(f)u(\pi) + \varepsilon e = u(f \circ \pi) + \varepsilon e \leq u^* + \varepsilon e .$$

Since $\varepsilon > 0$ and $f \in F$ are chosen arbitrarily, the assertion follows.
It remains to be shown that for all $\varepsilon > 0$ there exists a strategy $\pi \in \Pi$ satisfying $u(\pi) \geq u^* - \varepsilon e$.
Certainly, there exists for all $i \in S$ a strategy $\pi^i \in \Pi$ which satisfies $u(i,\pi^i) \geq u^*(i) - \varepsilon$. But then the strategy $\pi \in \Pi$, with

$$\pi(C \mid (i_0,\ldots,i_n)) = \pi^{i_0}(C \mid (i_0,\ldots,i_n))$$

for all $C \in \mathcal{A}$, all $n = 0,1,\ldots$ and all $(i_0,\ldots,i_n) \in H_n$, satisfies

$$u(\pi) \geq u^* - \varepsilon e .$$ □

From this lemma we obtain

THEOREM 2.6. *$P(f)$, $\tilde{U}$, $L(f)$, U, $L^+(f)$ and U^+ are (for all $f \in F$) operators on $V^+_{u^*}$, i.e., they are properly defined on $V^+_{u^*}$ and they map $V^+_{u^*}$ into itself.*

PROOF. Since for all $v \in V^+_{u^*}$ (and all $f \in F$)

$$U^+v = \max \{P(f)v, \tilde{U}v, L(f)v, Uv, L^+(f)v, U^+v\} ,$$

it is sufficient to prove the theorem for U^+.
That U^+ is properly defined on $V^+_{u^*}$ and maps $V^+_{u^*}$ into itself follows from lemma 2.5, since for all $v \in V^+_{u^*}$,

$$U^+v \leq U^+v^+ \leq U^+\|v^+\|_{u^*}u^* \leq U^+ \max\{1, \|v^+\|_{u^*}\}u^*$$
$$\leq \max\{1, \|v^+\|_{u^*}\}u^* .$$ □

Similarly, one may prove

THEOREM 2.7. *If* $z^* < \infty$, *then* $P(f)$, $\tilde{U}$, $L(f)$, U, $L^+(f)$, $U^+(f)$, $L^{abs}(f)$ *and* U^{abs} *are operators on* V_{z^*}.

2.3. THE FINITE-STAGE MDP

In this section we study the finite-stage MDP. It is shown that the value of this MDP as well as a nearly-optimal Markov strategy can be determined by a dynamic programming approach.

We consider an MDP in which the system is controlled at the times $t = 0,1,\ldots,n-1$ only, and if - as a result of the actions taken - the system reaches state j at time n, then there is a terminal payoff $v(j)$, $j \in S$. This MDP will be called the n-*stage MDP with terminal payoff* v ($v \in V$).

By $v_n(i,\pi,v)$ we denote the total expected reward in the n-stage MDP with initial state i and terminal payoff v when strategy $\pi \in \Pi$ is used,

$$(2.7) \qquad v_n(i,\pi,v) := \mathbb{E}_{i,\pi}\left[\sum_{k=0}^{n-1} r(X_k,A_k) + v(X_n)\right],$$

provided the expression is properly defined. To ensure that this is the case some condition on v is needed. We make the following assumption which will hold throughout this section.

CONDITION 2.8.

$$\sup_{\pi\in\Pi} \mathbb{E}_\pi v^+(X_n) < \infty, \quad n = 1,2,\ldots .$$

Note that it follows from lemma 2.1 that condition 2.8 is equivalent to

$$\mathbb{E}_\pi v^+(X_n) < \infty, \quad n = 1,2,\ldots \text{ for all } \pi \in RM .$$

Now let us consider the following dynamic programming scheme

$$(2.8) \qquad \begin{cases} v_0 := v \\ v_{n+1} := Uv_n, \quad n = 0,1,\ldots . \end{cases}$$

We will show that v_n is just the value of the n-stage MDP with terminal payoff v and that this scheme also yields a uniformly ε-optimal Markov

strategy. In order to do this we first prove by induction formulae (2.9)-(2.11).

(2.9) $L^+(f)v_{n-1}^+ < \infty$ for all $f \in F$ and $n = 1,2,\ldots$.

(2.10) $v_n < \infty$, $n = 1,2,\ldots$.

(2.11) For all $\varepsilon > 0$ there exist policies $f_0,f_1,\ldots$ such that

$$L(f_{n-1}) \ldots L(f_0)v \geq v_n - \varepsilon(1-2^{-n})e \ .$$

That (2.9)-(2.11) hold for n = 1 can be shown along exactly the same lines as the proof of the induction step and is therefore omitted.
Let us continue with the induction proof. Assuming that (2.9) - (2.11) hold for n = t, we prove them to hold for n = t + 1.
Let $f \in F$ be arbitrary and $f_{t-1},f_{t-2},\ldots,f_0$ be a sequence of policies satisfying (2.11) for n = t. Denote by π the (t+1)-stage strategy $\pi = (f,f_{t-1},f_{t-2},\ldots,f_0)$ (we specify π only for the first t + 1 stages). Then

$$\begin{aligned} L^+(f)v_t^+ &\leq L^+(f)[L(f_{t-1}) \ldots L(f_0)v + \varepsilon(1-2^{-t})e]^+ \\ &\leq L^+(f)[L^+(f_{t-1}) \ldots L^+(f_0)v^+ + \varepsilon(1-2^{-t})e] \\ &= L^+(f)L^+(f_{t-1}) \ldots L^+(f_0)v^+ + \varepsilon(1-2^{-t})e \\ &= \mathbb{E}_\pi \left[\sum_{k=0}^{t} r^+(X_k,A_k) + v^+(X_{t+1}) \right] + \varepsilon(1-2^{-t})e \ . \end{aligned}$$

So, by condition 1.1 and condition 2.8 formula (2.9) holds for n = t + 1. And also

$$\begin{aligned} v_{t+1} = Uv_t &\leq U^+v_t^+ \\ &\leq \sup_{\pi\in\Pi} \mathbb{E}_\pi \sum_{k=0}^{t} r^+(X_k,A_k) + \sup_{\pi\in\Pi} \mathbb{E}_\pi v^+(X_{t+1}) + \varepsilon(1-2^{-t})e < \infty \ , \end{aligned}$$

by theorem 2.3 and condition 2.8. Thus (2.10) also holds for n = t + 1. But $v_{t+1} < \infty$ implies the existence of a policy f_t such that

$$L(f_t)v_t \geq v_{t+1} - \varepsilon 2^{-t-1}e \ .$$

So,

$$L(f_t)L(f_{t-1}) \ldots L(f_0)v \geq L(f_t)(v_t - \varepsilon(1-2^{-t})e)$$

$$\geq L(f_t)v_t - \varepsilon(1-2^{-t})e \geq v_{t+1} - \varepsilon(1-2^{-t-1})e \ .$$

Which proves (2.11) for $n = t+1$.

This completes the proof of the induction step, thus (2.9)-(2.11) hold for all n.

In particular we see that for all $n = 1,2,\ldots$ a Markov strategy $\pi^{(n)} = (f_{n-1}, f_{n-2}, \ldots, f_0)$ exists such that

$$(2.12) \qquad v_n(\pi^{(n)},v) \geq v_n - \varepsilon e \ .$$

Hence, as $\varepsilon > 0$ is arbitrary,

$$(2.13) \qquad \sup_{\pi \in M} v_n(\pi,v) \geq v_n \ .$$

So, what remains to be shown is that

$$(2.14) \qquad v_n(\pi,v) \leq v_n \quad \text{for all } \pi \in \Pi \ .$$

Using lemma 2.1 one easily shows that it is sufficient to prove (2.14) for all $\pi \in RM$ (take $c_1 = 1$, $c_n = 0$, $n = 1,2,\ldots$).

Let $\pi = (\pi_0, \pi_1, \ldots) \in RM$ be arbitrary, and let $\pi^{\leftarrow k}$ denote the strategy $(\pi_k, \pi_{k+1}, \ldots)$.

Then we have for all $k = 0,1,\ldots,n-1$ and all $i \in S$

$$v_{n-k}(i,\pi^{\leftarrow k},v) = \int_A \pi_k(da|i)[r(i,a) + \sum_j p(i,a,j)v_{n-k-1}(j,\pi^{\leftarrow k+1},v)]$$

$$\leq \sup_{a \in A} \{r(i,a) + \sum_j p(i,a,j)v_{n-k-1}(j,\pi^{\leftarrow k+1},v)\} \ .$$

Hence,

$$v_{n-k}(\pi^{\leftarrow k},v) \leq Uv_{n-k-1}(\pi^{\leftarrow k+1},v) \ ,$$

and by the monotonicity of U and $v_0(\pi^{\leftarrow n},v) = v_0$ we have

$$v_n(\pi,v) \leq Uv_{n-1}(\pi^{\leftarrow 1},v) \leq \ \ldots \ \leq U^n v_0(\pi^{\leftarrow n},v) = U^n v_0 = v_n \ .$$

As $\pi \in RM$ was arbitrary, this proves (2.14) for all $\pi \in RM$ and thus, as we argued before, (2.14) holds for all $\pi \in \Pi$.

Summarizing the results of this section we see that we have proved

THEOREM 2.9. *If* $v \in V$ *satisfies condition 2.8, then for all* $n = 1,2,\ldots$

(i) $\sup_{\pi\in\Pi} v_n(\pi,v) = \sup_{\pi\in M} v_n(\pi,v) = U^n v$;

(ii) *for all* $\varepsilon > 0$ *there exists a strategy* $\hat{\pi} \in M$ *satisfying*

$$v_n(\hat{\pi},v) \geq U^n v - \varepsilon e \ .$$

Note that the n-stage MDP with terminal payoff v is properly defined and can be treated by the dynamic programming scheme (2.8) under conditions less strong than conditions 1.1 and 2.8.
It is sufficient that

$$\mathbb{E}_\pi \sum_{k=0}^{n-1} r^+(X_k,A_k) < \infty \quad \text{for all } \pi \in \Pi \ (\pi \in RM)$$

and that

$$\mathbb{E}_\pi v^+(X_k) < \infty \ , \quad k = 1,2,\ldots,n \ , \quad \text{for all } \pi \in \Pi \ (\pi \in RM) \ .$$

So, for example, theorem 2.9 also applies when r and v are bounded but $u^* = \infty$.

From these results for the finite-stage MDP we immediately obtain the following result for the ∞-stage MDP.

THEOREM 2.10. *For all* $i \in S$

$$u^*(i) = \sup_{\pi\in M} u(i,\pi) \ .$$

PROOF. For all $\varepsilon > 0$ there exists a strategy $\hat{\pi} \in \Pi$ such that $u(i,\hat{\pi}) \geq u^*(i) - \varepsilon/2$. Then there also exists a number n such that

$$u_n(i,\hat{\pi}) \geq u^*(i) - \varepsilon \ ,$$

where $u_n(i,\pi)$ is defined by

$$(2.15) \qquad u_n(i,\pi) := \mathbb{E}_{i,\pi} \sum_{k=0}^{n-1} r^+(X_k,A_k) \ .$$

Now we can apply theorem 2.9 (i) to the n-stage MDP with terminal payoff

$v = 0$ and rewards r^+ instead of r to obtain

$$\sup_{\pi \in M} u_n(i,\pi) \geq u_n(i,\hat{\pi}) \geq u^*(i) - \varepsilon \ .$$

Thus, with $u(i,\pi) \geq u_n(i,\pi)$ for all $\pi \in \Pi$, also

$$\sup_{\pi \in M} u(i,\pi) \geq u^*(i) - \varepsilon \ .$$

As this holds for all $\varepsilon > 0$, the assertion follows. □

2.4. THE OPTIMALITY EQUATION

As we already remarked in chapter 1, the functional equation

(2.16) $\quad v = Uv$

plays an important role in the analysis of the MDP. Equation (2.16) is also called the *optimality equation*. Note that in general Uv is not properly defined for every $v \in V$.

THEOREM 2.11. *v^* is a solution of (2.16).*

PROOF. First observe that by theorem 2.6 Uv^* is properly defined as $v^* \leq u^*$. In order to prove the theorem we follow the line of reasoning in ROSS [1970, theorem 6.1]. The proof consists of two parts: first we prove $v^* \geq Uv^*$ and then $v^* \leq Uv^*$.

Let $\varepsilon > 0$ and let $\pi \in \Pi$ be a uniformly ε-optimal strategy, i.e., $v(\pi) \geq v^* - \varepsilon e$. That such a strategy exists can be shown along the same lines as in the proof of theorem 2.5. Let f be an arbitrary policy. Then the decision rule: "use policy f at time 0 and continue with strategy π at time 1, pretending the process started at time 1" is again a strategy. We denote it by $f \circ \pi$.

So we have

$$v^* \geq v(f \circ \pi) = L(f)v(\pi) \geq L(f)v^* - \varepsilon e \ .$$

As $f \in F$ and $\varepsilon > 0$ are arbitrary, also

$$v^* \geq Uv^* \ .$$

In order to prove $v^* \leq Uv^*$, let $\pi = (\pi_0, \pi_1, \ldots)$ be an arbitrary randomized

Markov strategy and let $\pi^{+1} \in RM$ be the strategy $(\pi_1,\pi_2,\ldots)$. Then we have

$$v(i,\pi) = \int_A \pi_0(da|i)[r(i,a + \sum_j p(i,a,j)v(j,\pi^{+1})]$$

$$\leq \int_A \pi_0(da|i)[r(i,a) + \sum_j p(i,a,j)v^*(j)]$$

$$\leq \int_A \pi_0(da|i)(Uv^*)(i) = (Uv^*)(i) .$$

Taking the supremum with respect to $\pi \in RM$ we obtain, with corollary 2.2, $v^* \leq Uv^*$, which completes the proof. □

In general, the solution of (2.16) is not unique. For example, if $r(i,a) = 0$ for all $i \in S$, $a \in A$, then $v^* = 0$, and any constant vector solves (2.16).
In chapters 4 and 5 we will see that, under certain conditions, v^* is the unique solution of (2.16) within a Banach space. This fact has important consequences for the method of successive approximations.

From theorem 2.11 we immediately have

THEOREM 2.12 (cf. BLACKWELL [1967, theorem 2]). *If* $v \geq 0$, v *satisfies condition 2.8 and* $v \geq Uv$, *then* $v \geq v^*$.

PROOF. By theorem 2.9 (ii) and the monotonicity of U we have for all $\pi \in \Pi$ and all $n = 1,2,\ldots$

$$v_n(\pi) \leq U^n 0 \leq U^n v = U^{n-1}Uv \leq U^{n-1}v \leq \ldots \leq v .$$

So,

$$v(\pi) = \lim_{n\to\infty} v_n(\pi) \leq v .$$

Hence

$$v^* = \sup_{\pi\in\Pi} v(\pi) \leq v .$$ □

Note that in the conditions of the theorem we can replace "v satisfies condition 2.8" by "$\tilde{U}v^+ < \infty$", because $v \geq Uv$ and $\tilde{U}v^+ < \infty$ already imply that the scheme (2.8) is properly defined.

For the case $v^* \geq 0$ we obtain from theorem 2.12 the following characterization of v^*.

COROLLARY 2.13. *If* $v^* \geq 0$, *then* v^* *is the smallest nonnegative solution of the optimality equation.*

2.5. THE NEGATIVE CASE

In this section we will see that the fact that v^* solves the optimality equation implies the existence of uniformly nearly-optimal Markov strategies, if $v^* \leq 0$ or if v^* satisfies the weaker asymptotic condition (2.18) below.

From $v^* = Uv^*$ we have the existence of a sequence of policies $f_0, f_1, \ldots$ satisfying

$$L(f_n)v^* \geq v^* - \varepsilon 2^{-n-1} e \ , \quad n = 0,1,\ldots \ . \tag{2.17}$$

Then we have

THEOREM 2.14. *Let* $\pi_\varepsilon = (f_0, f_1, \ldots)$ *be a Markov strategy with* f_n *satisfying (2.17) for all* $n = 0,1,\ldots$. *If*

$$\limsup_{n\to\infty} \mathbb{E}_{\pi_\varepsilon} v^*(X_n) \leq 0 \ , \tag{2.18}$$

then π_ε *is uniformly* ε*-optimal, i.e.* $v(\pi) \geq v_\varepsilon^* - \varepsilon e$.

PROOF. $v^* \leq u^*$. So, by theorem 2.6, $\tilde{U}v^* \in V_{u^*}^+$, and, by induction, $\tilde{U}^n v^* \in V_{u^*}^+$, hence $\tilde{U}^n v^* < \infty$ for all $n = 1,2,\ldots$. So, $v_n(\pi_\varepsilon, v^*)$ is properly defined for all n, and we have

$$\begin{aligned} v(\pi_\varepsilon) &= \lim_{n\to\infty} v_n(\pi_\varepsilon) \geq \limsup_{n\to\infty} v_n(\pi_\varepsilon, v^*) \\ &= \limsup_{n\to\infty} L(f_0) \ldots L(f_{n-1}) v^* \\ &\geq \limsup_{n\to\infty} \{L(f_0) \ldots L(f_{n-2}) v^* - \varepsilon 2^{-n} e\} \geq \ldots \geq \\ &\geq \limsup_{n\to\infty} \{v^* - \varepsilon(2^{-1} + 2^{-2} + \ldots + 2^{-n}) e\} = v^* - \varepsilon e \ . \end{aligned}$$

□

An important consequence of this theorem is the following corollary which is used in the next section to prove that in the optimization of $v(i,\pi)$ we can restrict ourselves to Markov strategies.

COROLLARY 2.15. *If* $v^* \leq 0$, *then there exists a uniformly* ε*-optimal Markov strategy. In particular, there exists for all* $\varepsilon > 0$ *a strategy* $\pi \in M$ *satisfying*

$$w(\pi) \geq w^* - \varepsilon e \ .$$

(For the definition of $w(\pi)$ and w^*, see (1.36) and (1.37)).

As a special case of theorem 2.14 we have

THEOREM 2.16 (cf. HORDIJK [1974, theorem 6.3.c)]. *If* f *is a policy satisfying*

$$L(f)v^* = v^*$$

and

$$\limsup_{n\to\infty} \mathbb{E}_f v^*(X_n) \leq 0 \ ,$$

then f *is uniformly optimal:* $v(f) = v^*$.

As a corollary to this theorem we have

COROLLARY 2.17 (cf. STRAUCH [1966, theorem 9.1]). *If* A *is finite and for all* $f \in F$

$$\limsup_{n\to\infty} \mathbb{E}_f v^*(X_n) \leq 0 \ ,$$

then there exists a uniformly optimal stationary strategy.

PROOF. By the finiteness of A and theorem 2.11 there exists a policy f satisfying $L(f)v^* = v^*$; then the assertion follows with theorem 2.16. □

We conclude this section with the following analogue of theorem 2.12 and corollary 2.13:

THEOREM 2.18.

(i) *If* $v \leq 0$ *and* $v \leq Uv$ *then* $v \leq v^*$.

(ii) *If* $v^* \leq 0$ *then* v^* *is the largest nonpositive solution of (2.16).*

PROOF.

(i) As $v \leq 0$, v clearly satisfies condition 2.8. And as $v \leq Uv$ we can find policies f_n, $n = 0,1,\ldots$, satisfying

$$L(f_n)v \geq v - \varepsilon 2^{-n-1} e ,$$

where $\varepsilon > 0$ can be chosen arbitrarily small.

Then, analogous to the proof of theorem 2.14 we have for $\pi = (f_0,f_1,\ldots)$

$$v(\pi) = \lim_{n\to\infty} v_n(\pi) \geq \limsup_{n\to\infty} v_n(\pi,v) \geq v - \varepsilon e .$$

So also $v^* \geq v - \varepsilon e$ and, as ε is arbitrary, $v^* \geq v$.

(ii) Immediately from (i). □

2.6. THE RESTRICTION TO MARKOV STRATEGIES

In this section we use the results of the previous sections, particularly corollary 2.2, theorem 2.9 and corollary 2.15, to prove that we can restrict ourselves to Markov strategies in the optimization of $v(i,\pi)$.

THEOREM 2.19 (Van HEE [1978a]). *For all* $i \in S$

$$\sup_{\pi\in M} v(i,\pi) = v^*(i) .$$

PROOF. The proof proceeds as follows. First observe that there exists a randomized Markov strategy $\hat{\pi}$ which is nearly optimal for initial state i (corollary 2.2). Then there is a number n such that practically all positive rewards (for initial state i and strategy $\hat{\pi}$) are obtained before time n. From time n onwards we consider the negative rewards only. For this "negative problem" there exists (by corollary 2.15) a uniformly nearly optimal Markov strategy $\bar{\pi}$. Finally, consider the n-stage MDP with terminal payoff $w(\bar{\pi})$. For this problem there exists (by theorem 2.9) a nearly optimal Markov strategy $\pi^{(n)}$.

Then the Markov strategy: "use $\pi^{(n)}$ until time n and $\bar{\pi}$ afterwards, pretending the process restarts at time n" is nearly optimal in the ∞-stage MDP.

So, fix state $i \in S$ and choose $\varepsilon > 0$. Let $\hat{\pi} \in RM$ be ε-optimal for initial state i: $v(i,\hat{\pi}) \geq v^*(i) - \varepsilon$. Now split up $v(i,\hat{\pi})$ into three terms, as follows:

$$(2.19)\qquad v(i,\hat{\pi}) = v_n(i,\hat{\pi}) + \mathbb{E}_{i,\hat{\pi}} \sum_{k=n}^{\infty} r^+(X_k,A_k) - \mathbb{E}_{i,\hat{\pi}} \sum_{k=n}^{\infty} r^-(X_k,A_k)$$

with n so large that

$$(2.20)\qquad \mathbb{E}_{i,\hat{\pi}} \sum_{k=n}^{\infty} r^+(X_k,A_k) \leq \varepsilon \ .$$

Next, let $\bar{\pi} = (\bar{f}_0,\bar{f}_1,\ldots) \in M$ satisfy (cf. corollary 2.15)

$$(2.21)\qquad w(\bar{\pi}) \geq w^* - \varepsilon e \ .$$

If we now replace $\hat{\pi}$ by $\bar{\pi}$ from time n onwards, i.e., replace $\hat{\pi}_t$ by $\bar{f}_{t-n}$, $t = n,n+1,\ldots$, and ignore the positive rewards from time n onwards, then we obtain an n-stage MDP with terminal payoff $w(\bar{\pi})$ in which we use strategy $\hat{\pi}$. For this n-stage problem, by theorem 2.9 there exists a Markov strategy $\tilde{\pi} = (f_0,f_1,\ldots)$ which is ε-optimal for initial state i. Hence

$$v_n(i,\tilde{\pi},w(\bar{\pi})) \geq v_n(i,\hat{\pi},w(\bar{\pi})) - \varepsilon \ .$$

Finally, consider the Markov strategy $\pi^* = (f_0,f_1,\ldots,f_{n-1},\bar{f}_0,\bar{f}_1,\ldots)$, the strategy which plays $\tilde{\pi}$ upto time $n-1$ and then switches to $\bar{\pi}$. For this strategy we have

$$(2.22)\qquad v(i,\pi^*) = v_n(i,\tilde{\pi},w(\bar{\pi})) + \mathbb{E}_{i,\pi^*} \sum_{k=n}^{\infty} r^+(X_k,A_k)$$

$$\geq v_n(i,\tilde{\pi},w(\bar{\pi})) \geq v_n(i,\hat{\pi},w(\bar{\pi})) - \varepsilon \ .$$

Since $\hat{\pi}^{\leftarrow n} := (\hat{\pi}_n,\hat{\pi}_{n+1},\ldots)$ is again a strategy,

$$w(\bar{\pi}) \geq w^* - \varepsilon e \geq w(\hat{\pi}^{\leftarrow n}) - \varepsilon e \ .$$

So

$$(2.23)\qquad v_n(i,\hat{\pi},w(\bar{\pi})) \geq v_n(i,\hat{\pi},w(\hat{\pi}^{\leftarrow n})) - \varepsilon \ .$$

With (2.20) it follows that

$$(2.24)\qquad v_n(i,\hat{\pi},w(\hat{\pi}^{\leftarrow n})) = v_n(i,\hat{\pi}) + \mathbb{E}_{i,\hat{\pi}} \sum_{k=n}^{\infty} r^-(X_k,A_k) \geq v(i,\hat{\pi}) - \varepsilon \ .$$

Hence, from (2.22)-(2.24) and $v(i,\hat{\pi}) \geq v^*(i) - \varepsilon$,

$$v(i,\pi^*) \geq v^*(i) - 4\varepsilon \ .$$

As $\varepsilon > 0$ is arbitrary, the proof is complete. □

So, for each initial state there exists a nearly-optimal Markov strategy. If $v^* \leq 0$, then even a uniformly ε-optimal Markov strategy exists. (Note that this uniformity was essential in order to obtain (2.23).) In the next section (example 2.26) we will see that in general a uniformly nearly-optimal Markov strategy does not exist.

2.7. NEARLY OPTIMAL STRATEGIES

In this section we derive (and review) a number of results on nearly-optimal strategies. In the previous sections we already obatined some results on the existence of nearly optimal strategies (theorems 2.14, 2.16 and 2.19, and corollaries 2.15 and 2.17).
One of the most interesting (and as far as we know new) results in given in theorem 2.22: if A is finite, then for each state i there exists an ε-optimal stationary strategy. If S is also finite, then there even exists a uniformly optimal stationary strategy.
Further some examples are given showing that in general uniformly nearly-optimal Markov, or randomized Markov, strategies do not exist.

The first question we address concerns the existence of nearly-optimal stationary strategies.
In general, ε-optimal stationary strategies do not exist, as is shown by the following example.

EXAMPLE 2.20. $S = \{1\}$, $A = (0,1]$, $r(i,a) = -a$, $p(1,a,1) = 1$, $a \in A$.
Clearly, $v^* = 0$, but for all $f \in F$ we have $v(f) = -\infty$.

In this example the nonfiniteness of A is essential.
If A is finite, then we have the following two theorems which we believe to be new in the setting considered here.

THEOREM 2.21. *If* S *and* A *are finite, then there exists a uniformly optimal stationary strategy.*

The proof of this theorem is postponed until chapter 5, section 5.

Using theorem 2.21 we prove

THEOREM 2.22. *If* A *is finite, then for each* $\varepsilon > 0$ *and for each initial state* $i \in S$ *there exists an* ε*-optimal stationary strategy.*

PROOF. The proof is rather involved. Roughly, it goes like this. First let π be an ε-optimal Markov strategy for initial state i. Then we construct a finite set B such that, if the process starts in state i and strategy π is used, nearly all positive rewards are obtained before the system leaves B. From that moment on we consider only the negative rewards. For this negative problem by corollary 2.17 there exists a uniformly optimal stationary strategy h_1: $w(h_1) = w^*$.
Next we consider the finite state MDP with state space B, where as soon as the system leaves B and reaches a state $j \notin B$ we obtain a terminal reward $w^*(j)$. For this MDP by theorem 2.21 there exists an optimal stationary strategy h_2.
Finally we prove that the stationary strategy f with $f(i) = h_2(i)$ for $i \in B$ and $f(i) = h_1(i)$ if $i \notin B$ is nearly optimal for initial state i.
So, fix $i \in S$ and choose $\varepsilon > 0$. Then there exists (by theorem 2.19) an ε-optimal Markov strategy π for initial state i: $v(i,\pi) \geq v^*(i) - \varepsilon$.
Next we construct a finite set $B \subset S$, with $i \in B$, such that practically all positive rewards (for initial state i and strategy π) are obtained before the system first leaves B.
Let n_0 be such that (cf. (2.15)) $u(i,\pi) - u_{n_0}(i,\pi) \leq \varepsilon$ and define for $n = 0,1,\ldots,n_0-1$

$$u^{(n)}(\pi) := \mathbb{E}_\pi \sum_{k=n}^{n_0-1} r^+(X_k,A_k) \ .$$

With $\pi = (f_0,f_1,\ldots)$ this implies

$$u^{(n)}(\pi) = r^+(f_n) + P(f_n)u^{(n+1)}(\pi) \ .$$

Clearly, for all $j \in S$ and $n = 0,1,\ldots,n_0-2$, there exists a finite set $B_{n+1}(j)$ such that only a fraction ε_0 (to be specified below) is lost:

$$r^+(j,f_n(j)) + \sum_{k \in B_{n+1}(j)} p(j,f_n(j),k)u^{(n+1)}(k,\pi) \geq (1-\varepsilon_0)u^{(n)}(j,\pi) \ .$$

Define

$$B_0 := \{i\} \quad \text{and} \quad B_{n+1} := B_n \cup \bigcup_{j \in B_n} B_{n+1}(j) \ , \quad n = 0,1,\ldots,n_0-2 \ ,$$

and

$$B := B_{n_0-1} \ .$$

Now we will show that indeed nearly all positive rewards are obtained before B is left (if ε_0 is chosen sufficiently small).
Let τ be the first-exit time from the set B:

$$\tau(i_0,i_1,\ldots) = \inf \{n \mid i_n \notin B\} \ ,$$

for all $i_0 \in B$ and $i_1,i_2,\ldots \in S$.
Then for the sum of all positive rewards until the first exit from B we have, with $u^{(n_0)}(\pi) = 0$,

$$\mathbb{E}_{i,\pi} \sum_{n=0}^{\tau-1} r^+(X_n,A_n) \geq \mathbb{E}_{i,\pi} \sum_{n=0}^{\min(\tau-1,n_0-1)} r^+(X_n,A_n)$$

$$= \sum_{n=0}^{n_0-1} \sum_{j \in B} \mathbb{P}_{i,\pi}(X_n = j \ , \tau > n) r^+(j,f_n(j))$$

$$\geq \sum_{n=0}^{n_0-1} \sum_{j \in B} \mathbb{P}_{i,\pi}(X_n = j \ , \tau > n)[(1-\varepsilon_0)u^{(n)}(j,\pi) +$$

$$- \sum_{k \in B} p(j,f_n(j),k)u^{(n+1)}(k,\pi)]$$

$$= (1-\varepsilon_0) \sum_{n=0}^{n_0-1} \sum_{j \in B} \mathbb{P}_{i,\pi}(X_n = j \ , \tau > n)u^{(n)}(j,\pi) +$$

$$- \sum_{n=1}^{n_0} \sum_{k \in B} \mathbb{P}_{i,\pi}(X_n = k \ , \tau > n)u^{(n)}(k,\pi)$$

$$= u_{n_0}(i,\pi) - \varepsilon_0 \sum_{n=0}^{n_0-1} \sum_{j \in B} \mathbb{P}_{i,\pi}(X_n = j \ , \tau > n)u^{(n)}(j,\pi)$$

$$\geq u_{n_0}(i,\pi) - \varepsilon_0 \sum_{n=0}^{n_0-1} \sum_{j \in B} \mathbb{P}_{i,\pi}(X_n = j)u^{(n)}(j,\pi)$$

$$\geq u_{n_0}(i,\pi) - \varepsilon_0 \sum_{n=0}^{n_0-1} \mathbb{E}_{i,\pi} u^{(n)}(X_n,\pi) \geq$$

$$\geq u_{n_0}(i,\pi) - \varepsilon_0 n_0 u_{n_0}(i,\pi) ,$$

as clearly

$$u_{n_0}(i,\pi) \geq \mathbb{E}_{i,\pi} u^{(n)}(X_n,\pi) , \quad n = 1,2,\ldots,n_0-1 .$$

So, choosing $\varepsilon_0 = \varepsilon / n_0 u_{n_0}(i,\pi)$, we have

$$\mathbb{E}_{i,\pi} \sum_{n=0}^{\tau-1} r^+(X_n,A_n) \geq u_{n_0}(i,\pi) - \varepsilon \geq u(i,\pi) - 2\varepsilon . \tag{2.25}$$

Next consider the MDP where, once the process has left B, we continue with a stationary strategy h_1, satisfying $w(h_1) = w^*$ (which exists by corollary 2.17), counting the negative rewards only. This MDP is essentially equivalent to an MDP with finite state space $\hat{S} := B \cup \{*\}$ ($* \notin V$), action space $\hat{A} = A$, rewards $\hat{r}(i,a)$ and transition probabilities $\hat{p}(i,a,j)$, defined by

$$\begin{cases} \hat{r}(i,a) := r(i,a) + \sum_{j \notin B} p(i,a,j) w^*(j) , & i \in B , \\ \hat{r}(*,a) := 0 , \\ \hat{p}(i,a,j) = p(i,a,j) , & i,j \in B , \\ \hat{p}(i,a,*) = \sum_{j \notin B} p(i,a,j) , & i \in B , \\ \hat{p}(*,a,*) = 1 , & \text{for all } a \in A . \end{cases} \tag{2.26}$$

So, as soon as the system leaves B it is absorbed in state * and we therefore adapt the immediate rewards.

For this finite-state MDP there exists by theorem 2.21 an optimal stationary strategy h_2.

Now we want to show that the stationary strategy f defined by

$$\begin{cases} f(j) = h_2(j) , & j \in B , \\ f(j) = h_1(j) , & j \notin B , \end{cases}$$

is nearly optimal for initial state i.

Before we do this, we have to derive some important inequalities for the MDP defined by (2.26).

Denote the total expected reward for initial state $j \in B$ and strategy $\tilde{\pi} \in \Pi$ by $\hat{v}(j,\tilde{\pi})$ (formally we should say $\hat{\tilde{\pi}} \in \hat{\Pi}$, but there is a clear correspondence

between strategies in Π and in $\hat{\Pi}$). Then for all $j \in B$ and $\tilde{\pi} \in \Pi$

$$\hat{v}(j,h_2) \geq \hat{v}(j,\tilde{\pi}) \ .$$

Particularly,

$$\hat{v}(j,h_2) \geq \hat{v}(j,h_1) \geq w^*(j) \ , \tag{2.27}$$

where the second inequality follows from

$$\hat{v}(j,h_1) = \mathbb{E}_{j,h_1}\left[\sum_{n=0}^{\tau-1} r(X_n,A_n) + w^*(X_\tau)\right]$$

$$\geq \mathbb{E}_{j,h_1}\left[\sum_{n=0}^{\tau-1} r^-(X_n,A_n) + w(X_\tau,h_1)\right] = w(j,h_1) = w^*(j) \ .$$

And

$$\begin{aligned}\hat{v}(i,h_2) \geq \hat{v}(i,\pi) &= \mathbb{E}_{i,\pi}\left[\sum_{n=0}^{\tau-1} r(X_n,A_n) + w^*(X_\tau)\right] \\ &\geq \mathbb{E}_{i,\pi}\left[\sum_{n=0}^{\tau-1} r(X_n,A_n) + \sum_{n=\tau}^{\infty} r^-(X_n,A_n)\right] \\ &= \mathbb{E}_{i,\pi}\sum_{n=0}^{\infty} r(X_n,A_n) - \mathbb{E}_{i,\pi}\sum_{n=\tau}^{\infty} r^+(X_n,A_n) \\ &\geq v(i,\pi) - 2\varepsilon \geq v^*(i) - 3\varepsilon \ .\end{aligned} \tag{2.28}$$

Finally, we can prove that f is 3ε-optimal for initial state i in the original MDP.

We will prove that $v(i,f) \geq \hat{v}(i,h_2)$, which by (2.28) is sufficient.

To this end we define the stopping times $\tau_1,\tau_2,\ldots$, where τ_n is the time of the n-th switch from B to $S \setminus B$ or vice versa. I.e., for any $\zeta = (i_0,i_1,\ldots)$ and $n \geq 1$

$$\tau_n(\zeta) = \inf\{k > \tau_{n-1}(\zeta) \mid \text{if } i_{\tau_{n-1}(\zeta)} \in B \text{ then } i_k \notin B \text{ else } i_k \in B\},$$

with $\tau_0(\zeta) = 0$.

Clearly $\tau_n \geq n$. So

$$v(i,f) = \lim_{n\to\infty} \mathbb{E}_{i,f} \sum_{k=0}^{\tau_n-1} r(X_k,A_k) \ .$$

Thus, since $w^* \le 0$,

$$(2.29) \qquad v(i,f) \ge \limsup_{n\to\infty} \mathbb{E}_{i,f}\left[\sum_{k=0}^{\tau_n-1} r(X_k,A_k) + w^*(X_{\tau_n})\right] .$$

Further, we have for all $j \in B$, using (2.27) and $f = h_2$ on B,

$$(2.30) \qquad \mathbb{E}_{j,f}\left[\sum_{k=0}^{\tau_1-1} r(X_k,A_k) + w^*(X_{\tau_1})\right]$$

$$= \mathbb{E}_{j,h_2}\left[\sum_{k=0}^{\tau_1-1} r(X_k,A_k) + w^*(X_{\tau_1})\right] = \hat{v}(j,h_2) \ge w^*(j) .$$

And for all $j \notin B$,

$$(2.31) \qquad \mathbb{E}_{j,f}\left[\sum_{k=0}^{\tau_1-1} r(X_k,A_k) + w^*(X_{\tau_1})\right]$$

$$\ge \mathbb{E}_{j,h_1}\left[\sum_{k=0}^{\tau_1-1} r^-(X_k,A_k) + w(X_{\tau_1},h_1)\right] = w(j,h_1) = w^*(j) .$$

Also, for $\zeta = (i_0,i_1,\ldots)$,

$$\tau_n(\zeta) = \tau_{n-1}(\zeta) + \tau_1(i_{\tau_{n-1}(\zeta)},i_{\tau_{n-1}(\zeta)+1},\ldots) .$$

Thus, for $n = 1,2,\ldots$, we get from (2.30) and (2.31),

$$(2.32) \qquad \mathbb{E}_{j,f}\left[\sum_{k=0}^{\tau_n-1} r(X_k,A_k) + w^*(X_{\tau_n})\right]$$

$$= \mathbb{E}_{j,f}\left[\sum_{k=0}^{\tau_{n-1}-1} r(X_k,A_k) + \mathbb{E}_{X_{\tau_{n-1}},f}\left[\sum_{\ell=0}^{\tau_1-1} r(\tilde{X}_\ell,\tilde{A}_\ell) + w^*(\tilde{X}_{\tau_1})\right]\right]$$

$$\ge \mathbb{E}_{j,f}\left[\sum_{k=0}^{\tau_{n-1}-1} r(X_k,A_k) + w^*(X_{\tau_{n-1}})\right] ,$$

where we write (X_k,A_k) and $(\tilde{X}_\ell,\tilde{A}_\ell)$ in order to distinguish between the process starting at time 0 and the process starting at time τ_{n-1}. Repeatedly applying (2.32) we obtain

$$v(i,f) \geq \limsup_{n\to\infty} \mathbb{E}_{i,f}\left[\sum_{k=0}^{\tau_n-1} r(X_k,A_k) + w^*(X_{\tau_n})\right]$$

$$\geq \mathbb{E}_{i,f}\left[\sum_{k=0}^{\tau_1-1} r(X_k,A_k) + w^*(X_{\tau_1})\right] = \hat{v}(i,h_2) \ ,$$

as $\tau_1 = \tau$ for initial state i.
So, with (2.28),

$$v(i,f) \geq v^*(i) - 3\varepsilon \ ,$$

which, as $\varepsilon > 0$ is arbitrary, completes the proof. □

As we remarked before, there does not necessarily exist a uniformly nearly optimal stationary (or even a Markov or randomized Markov) strategy. This will be shown in the examples 2.24-2.26.
However, if all rewards are nonnegative, the so-called positive dynamic programming case, we have the following theorem due to ORNSTEIN [1969].

THEOREM 2.23. *If* $r(i,a) \geq 0$ *for all* $i \in S$, $a \in A$, *then for every* $\varepsilon > 0$ *a stationary strategy* f *exists, satisfying*

(2.33) $\quad v(f) \geq (1-\varepsilon)v^* \ .$

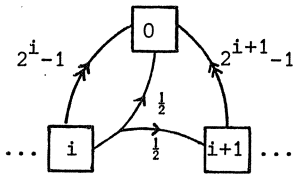
PROOF. For the very ingenious proof see ORNSTEIN [1969].

Note, that in theorem 2.23 the action space A need not be finite.
So, in the positive dynamic programming case there does exist a stationary strategy that is uniformly ε-optimal in the multiplicative sense of (2.33).
Clearly, if v^* is bounded, then theorem 2.23 also implies (for the positive case) the existence of a stationary strategy f which is uniformly ε-optimal in the additive sense:

(2.34) $\quad v(f) \geq v^* - \varepsilon e \ .$

In general, however, even if A is finite, a stationary strategy satisfying (2.34) need not exist.
This is shown by the following example given by BLACKWELL [1967].

EXAMPLE 2.24. $S := \{0,1,2,\ldots\}$, $A = \{1,2\}$. State 0 is absorbing: $r(0,a) = 0$, $p(0,a,0) = 1$. In state i, $i = 1,2,\ldots$, we have $r(i,1) = 0$, $p(i,1,i+1) = p(i,1,0) = \frac{1}{2}$ and $r(i,2) = 2^i - 1$, $p(i,2,0) = 1$. So, in state i you either receive $2^i - 1$ and the system moves to state 0, or you receive nothing and the system moves to states 0 and $i+1$, each with probability $\frac{1}{2}$.



Clearly, $v^*(0) = 0$ and $v^*(i) = 2^i$, $i = 1,2,\ldots$. Now let f be a stationary strategy. Then either $f(i) = 1$ for all $i = 1,2,\ldots$, thus $v(f) = 0$, or $f(i) = 2$ for at least one i, i_0 say. But then $v(i_0,f) = v^*(i_0) - 1$. Hence no stationary strategy can be ε-optimal in the sense of (2.34) for $0 \le \varepsilon < 1$.

However, if we consider also 'randomized' stationary strategies, then a stationary strategy that is uniformly ε-optimal in the sense of (2.34) does exist, at least in this example.

We call a strategy $\pi \in RM$, $\pi = (\pi_0,\pi_1,\ldots)$ randomized stationary if $\pi_n = \pi_0$, $n = 1,2,\ldots$. In this example π is completely characterized by the probability p_i by which action 1 is chosen in state i, $i \in S$. If $p_i = p$ for all $i \in S$, then we have

$$v(i,\pi) = (1-p)(2^i - 1) + p\cdot\tfrac{1}{2}(1-p)(2^{i+1} - 1) + p\cdot\tfrac{1}{2}\cdot p\cdot\tfrac{1}{2}(1-p)(2^{i+2} - 1) + \ldots$$
$$= 2^i - (1-p)(1-\tfrac{1}{2}p)^{-1} .$$

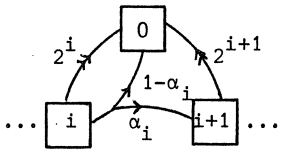
Thus, if $0 < \varepsilon < 1$ and $p > 1 - \frac{1}{2}\varepsilon$ (then $(1-p)(1-\frac{1}{2}p)^{-1} < \varepsilon$), then we have

$$v(\pi) \ge v^* - \varepsilon e .$$ □

This example demonstrates that it may happen that there exists no 'pure' stationary strategy that is ε-optimal in the sense of (2.34), whereas a randomized stationary strategy having this property does exist.

The following example which is only a slight modification of example 2.24 shows that in general also a randomized stationary strategy satisfying (2.34) need not exist.

EXAMPLE 2.25. All data are identical to those in example 2.24, except for the following: $r(i,2) = 2^i$ and $p(i,1,i+1) = \alpha_i = b_i/2b_{i+1}$, with $b_i = 1 + i^{-1}$, $p(i,1,0) = 1 - \alpha_i$, $i = 1,2,\ldots$. Thus, for $i = 1,2,\ldots$,



$$v^*(i) = \sup \{2^i , \alpha_i 2^{i+1} , \alpha_i\alpha_{i+1} 2^{i+2} , \ldots\}$$

$$= 2^i \sup \{1 , b_i/b_{i+1} , b_i/b_{i+2} , \ldots\} = 2^i b_i = 2^i(1 + i^{-1}) ,$$

as $b_i \downarrow 1$ for $i \to \infty$.

Let π be any randomized stationary strategy, then π is again completely characterized by the probabilities p_i by which action 1 is chosen in state i, $i = 1,2,\ldots$.

In order that $v(\pi) \geq v^* - e$, it is certainly necessary that for all $i = 1,2,\ldots$

$$(1 - p_i)2^i + p_i\alpha_i v^*(i+1) \geq v^*(i) - 1 ,$$

or, after some algebra,

$$1 - p_i \leq i2^{-i} .$$

Since otherwise,

$$v(i,\pi) = (1 - p_i)2^i + p_i\alpha_i v(i+1,\pi) \leq (1 - p_i)2^i + p_i\alpha_i v^*(i+1) < v^*(i) - 1 .$$

But then, using $\alpha_i \leq 2/3$, $p_i \leq 1$ and $1 - p_i \leq i2^{-i}$, $i = 1,2,\ldots$, we get

$$v(i,\pi) = (1 - p_i)2^i + p_i\alpha_i(1 - p_{i+1})2^{i+1} + p_i\alpha_i p_{i+1}\alpha_{i+1}(1 - p_{i+2})2^{i+2} + \ldots$$

$$\leq i + (\tfrac{2}{3})(i+1) + (\tfrac{2}{3})^2(i+2) + \ldots = 3i + 6 .$$

So, for $i \geq 4$,

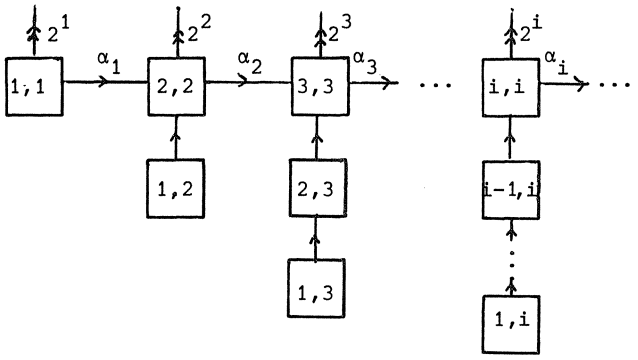
$$v(i,\pi) < v^*(i) - 1 .$$

Hence, no randomized stationary strategy π can satisfy

$$v(\pi) \geq v^* - \varepsilon e .$$

Extending this example, one may show that there need not even exist a randomized Markov strategy π satisfying $v(\pi) \geq v^* - \varepsilon e$. Recall that the possibility of restriction to randomized Markov strategies (corollary 2.2) holds only componentwise.

EXAMPLE 2.26. $S := \{0\} \cup \{(k,i) \mid k,i = 1,2,\ldots, k \leq i\}$, $A = \{1,2\}$.



All rewards and transition probabilities are zero except for the following: $r((i,i),2) = 2^i$, $p((k,i),a,(k+1,i)) = 1$, $p((i,i),1,(i+1,i+1)) = \alpha_i$, $p((i,i),1,0) = 1 - \alpha_i$, $p((i,i),2,0) = 1$, $p(0,a,0) = 1$, $i = 1,2,\ldots, k < i$ and $a \in A$. Here α_i is defined as in example 2.25. The states (i,i) play the same role as in the states i in example 2.25.
Clearly,

$$v^*((k,i)) = 2^i(1 + i^{-1}) \quad , \quad k \leq i \, , \quad i = 1,2,\ldots \, .$$

Let us fix i and look for a randomized Markov strategy π that is 1-optimal for all states $(i,i),(i,i+1),(i,i+2),\ldots,$ simultaneously.
The relevant actions for state $(i,i+\ell)$ are the actions at time t in state $(i+t,i+t)$ for $t \geq \ell$.
Denote by $p_{i+\ell}$ the probability that in state $(i+\ell,i+\ell)$ action 1 is taken at time ℓ. Then one easily verifies that again we need to have

$$1 - p_t \leq t2^{-t} \quad , \quad t = i,i+1,\ldots \, .$$

But this again implies $v((i,i),\pi) \leq 3i + 6$, contradicting for $i \geq 4$ the 1-optimality of π for initial state (i,i).
Hence, there does not exist a uniformly ε-optimal strategy $\pi \in RM$ in the additive sense of (2.34).
In Van HEE, HORDIJK and Van der WAL [1977] an example is given with both positive and negative rewards and finite action space in which neither in the additive nor in the multiplicative sense a uniformly ε-optimal strategy exists.

CHAPTER 3

SUCCESSIVE APPROXIMATIONS FOR THE TOTAL-REWARD MDP

3.1. INTRODUCTION

Our main interest when studying the total reward MDP (or any other decision process) is to determine the value of the MDP and to determine a (nearly) optimal strategy.
In this chapter some methods are studied by which the first part of this problem, the approximation of v^*, might be solved.
We know that v^* is a solution of the optimality equation

(3.1) $\quad Uv = v$

(theorem 2.11). So we can try to approximate a solution of this equation (hoping it to be the right one), for example by the so-called method of standard successive approximations

(3.2) $\quad v_{n+1} = Uv_n \ , \quad n = 0,1,\ldots \ .$

This method will be studied in section 2. Further some (partly known) conditions are given which guarantee that v_n converges to v^*. In general, as is shown in example 3.2, the method of standard successive approximation need not converge.
There are several other iterative methods by which a solution of an equation like (3.1) might be approximated. See, for example, VARGA [1962, chapter 3] for a description of the Jacobi, the Gauss-Seidel and the overrelaxation methods for the solution of a simple matrix equation. The latter three methods, Jacobi iteration, Gauss-Seidel iteration (see HASTINGS [1968]) and overrelaxation (see REETZ [1973]) have also been studied for contracting MDP's.
It is possible to describe all these successive approximation methods in terms of go-ahead functions. This has the advantage that one can study the

convergence of these methods simultaneously. The go-ahead function approach has been introduced for the contracting case by WESSELS [1977a], Van NUNEN and WESSELS [1976], Van NUNEN [1976a] and Van NUNEN and STIDHAM [1978]. In section 3 we introduce the go-ahead function technique for the general total reward MDP model of section 1.5. The corresponding dynamic programming operators are studied in section 4, where it is shown that v^* is indeed a fixed point of each of these operators.
In section 5 a subset is considered of the set of all go-ahead functions which still contains all those go-ahead functions, by which the algorithms we are interested in can be described. It is shown that for this set of go-ahead functions one needs to consider in the optimization step only Markov strategies.
A second set of algorithms for the approximation of v^* is the set of value-oriented successive approximation methods. These methods have been first mentioned for the contracting MDP by PORTEUS [1971] and have been extensively studied and shown to converge by Van NUNEN [1976a]. In the value-oriented approach each optimization step is followed by some kind of extrapolation. We will consider these methods in section 6.

3.2. STANDARD SUCCESSIVE APPROXIMATIONS

In this section we first study the method of standard successive approximations for the solution of the optimality equation (3.1) hoping to obtain an approximation of v^*.

Standard successive approximations.

Choose $v_0 \in V^+_{u^*}$ (v_0 is often called the scrapvalue)
Determine for $n = 0,1,\ldots$

$$v_{n+1} = Uv_n .$$

From theorem 2.6 it follows that this scheme is properly defined.
Another way of looking at the method of standard successive approximations is to consider it as an approximation of the ∞-horizon MDP by finite-stage MDP's. Then the question is: can we approximate the ∞-horizon MDP by a sequence of finite-stage MDP's (with terminal payoff v_0), i.e. does v_n converge to v^*.

We will say that the method of standard successive approximations for scrap-value v_0 converges if $\lim_{n\to\infty} U^n v_0$ exists and is equal to v^*.
Define v_∞ by

$$v_\infty := \operatorname{liminf}_{n\to\infty} U^n 0 , \tag{3.3}$$

where the liminf is taken componentwise.
Then we have the following well-known result.

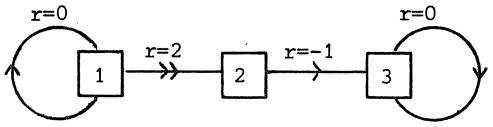
LEMMA 3.1 (SCHÄL [1975, formula (2.5)]).

$$v_\infty \geq v^* .$$

PROOF. For all $\pi \in \Pi$ we have $v_n(\pi) \leq U^n 0$, so also $v(\pi) \leq v_\infty$ (from $v_n(\pi) \to v(\pi)$ if $n \to \infty$). Hence, $v^* = \sup_{\pi\in\Pi} v(\pi) \leq v_\infty$. □

In general, v_∞ may be larger than v^* as the following simple example shows.

EXAMPLE 3.2 (Van HEE, HORDIJK and Van der WAL [1977]). $S := \{1,2,3\}$, $A = \{1,2\}$, $r(1,1) = r(3,a) = 0$, $r(1,2) = 2$, $r(2,a) = -1$, $p(1,1,1) = p(1,2,2) = p(2,a,3) = p(3,a,3) = 1$, $a \in A$.



Clearly, $v^*(1) = 1$, but $(U^n 0)(1) = 2$ for all $n \geq 1$. So $v^*(1) < v_\infty(1)$.

The problem in this example is that we can postpone negative rewards, which in the ∞-horizon MDP are unavoidable. This leads us to the following theorem.

THEOREM 3.3 (Van HEE, HORDIJK and Van der WAL [1977, theorem 3.5]).
If

$$\operatorname{liminf}_{n\to\infty} \inf_{\pi\in M} \mathbb{E}_\pi v^*(X_n) \geq 0$$

then

$$v_\infty = v^* .$$

PROOF. By lemma 3.1 it is sufficient to prove that $\operatorname{limsup}_{n\to\infty} U^n 0 \leq v^*$.
We have

$$U^n 0 \leq U^n v^* - \inf_{\pi \in M} \mathbb{E}_\pi v^*(X_n) = v^* - \inf_{\pi \in M} \mathbb{E}_\pi v^*(X_n) .$$

So,

$$\limsup_{n \to \infty} U^n 0 \leq v^* - \liminf_{n \to \infty} \inf_{\pi \in M} \mathbb{E}_\pi v^*(X_n) \leq v^* . \qquad \square$$

COROLLARY 3.4 (cf. BLACKWELL [1967, theorem 3]). *If for all* $i \in S$ *and* $a \in A$ *we have* $r(i,a) \geq 0$ *(the positive dynamic programming case), then the method of standard successive approximations with scrapvalue* 0 *converges.*

As we have seen in example 3.2, finiteness of A is in general not sufficient for the convergence of $U^n 0$ to v^*. In case all $r(i,a) \leq 0$, however, one has the following result.

THEOREM 3.5 (STRAUCH [1966, theorem 9.1]). *If* A *is finite and* $r(i,a) \leq 0$ *for all* $i \in S$ *and* $a \in A$, *then the method of standard successive approximations with scrapvalue* 0 *converges.*

PROOF. By lemma 3.1 and theorem 2.18 it is sufficient to prove $Uv_\infty \geq v_\infty$ (clearly $v_\infty \leq 0$). Therefore choose some arbitrary state $i \in S$. Then, by the finiteness of A, there exist an action a and a subsequence $\{U^{n_k} 0\}$ of $\{U^n 0\}$, such that

$$r(i,a) + \sum_j p(i,a,j)(U^{n_k} 0)(j) = (U^{n_k+1} 0)(i) , \quad k = 0,1,\ldots .$$

As the sequence v_n is monotonically nonincreasing ($U0 \leq 0$, so by the monotonicity of U also $U^{n+1} 0 \leq U^n 0$), we have

$$\sum_j p(i,a,j)(U^{n_k} 0)(j) \to \sum_j p(i,a,j) v_\infty(j) \quad (k \to \infty) .$$

Hence,

$$(Uv_\infty)(i) \geq r(i,a) + \sum_j p(i,a,j) v_\infty(j)$$

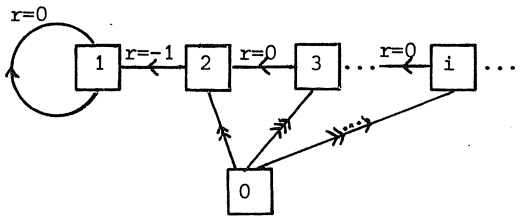
$$= \lim_{k \to \infty} \{ r(i,a) + \sum_j p(i,a,j)(U^{n_k} 0)(j) \} = v_\infty(i) .$$

As i has been chosen arbitrarily, the proof is complete. $\square$

Clearly, theorem 3.5 also holds if the action set in state i varies with i but is still finite for all $i \in S$.

If A is not finite, however, then the method of standard successive approximations (with $v_0 = 0$ and $r(i,a) \leq 0$) need not converge, as is shown by the following example.

EXAMPLE 3.6 (STRAUCH [1966, example 6.1]). $S := \{0,1,2,\ldots\}$, $A = \{2,3,4,\ldots\}$. For all $a \in A$: $r(i,a) = 0$, $i \neq 2$, $r(2,a) = -1$, $p(1,a,1) = 1 = p(i+1,a,i)$, $i \geq 1$, and $p(0,a,a) = 1$. Clearly $(U^n 0)(0) = 0$ for all n, but $v^*(0) = -1$. So $v_\infty(0) > v^*(0)$.



A nice result, from which lemma 3.1 and corollary 3.4 follow immediately, is the following. (Recall the definition of w^* in (1.37).)

THEOREM 3.7. *The method of standard successive approximations converges for all scrapvalues v_0, with $w^* \leq v_0 \leq v^*$. I.e., for all these scrapvalues we have*

$$\lim_{n\to\infty} U^n v_0 \quad \textit{exists and is equal to } v^* .$$

PROOF. By the monotonicity of U we have for all $w^* \leq v_0 \leq v^*$

$$U^n w^* \leq U^n v_0 \leq U^n v^* = v^* , \quad n = 1,2,\ldots ,$$

hence

$$\operatorname*{limsup}_{n\to\infty} U^n w^* \leq \operatorname*{limsup}_{n\to\infty} U^n v_0 \leq v^* .$$

So it is sufficient to prove

$$\operatorname*{liminf}_{n\to\infty} U^n w^* \geq v^* .$$

Let $\pi = (f_0, f_1, \ldots) \in M$ be an arbitrary strategy. Then for all n

$$v(\pi) = \mathbb{E}_\pi \left[\sum_{k=0}^{n-1} r(X_k, A_k) + \sum_{k=n}^{\infty} r^+(X_k, A_k) + \sum_{k=n}^{\infty} r^-(X_k, A_k) \right] .$$

Further we have, with $\pi^{\leftarrow n} = (f_n, f_{n+1}, \ldots)$,

$$\mathbb{E}_\pi \sum_{k=n}^{\infty} r^-(X_k,A_k) = \mathbb{E}_\pi w(X_n,\pi^{\leftarrow n}) \le \mathbb{E}_\pi w^*(X_n) \ .$$

So for all $n = 1,2,\ldots$

$$U^n w^* \ge \mathbb{E}_\pi \left[\sum_{k=0}^{n-1} r(X_k,A_k) + w^*(X_n) \right]$$

$$\ge \mathbb{E}_\pi \left[\sum_{k=0}^{n-1} r(X_k,A_k) + \sum_{k=n}^{\infty} r^-(X_k,A_k) \right] = v(\pi) - \mathbb{E}_\pi \sum_{k=n}^{\infty} r^+(X_k,A_k) \ .$$

By condition 1.1 we have

$$\mathbb{E}_\pi \sum_{k=n}^{\infty} r^+(A_k,X_k) \to 0 \quad (n \to \infty) \ ,$$

hence

$$\liminf_{n\to\infty} U^n w^* \ge v(\pi) \quad \text{(componentwise)}.$$

Taking the (pointwise) supremum with respect to $\pi \in M$, we obtain

$$\liminf_{n\to\infty} U^n w^* \ge v^* \ .$$

So, for all $w^* \le v_0 \le v^*$,

$$\lim_{n\to\infty} U^n w^* = \lim_{n\to\infty} U^n v_0 = v^* \ . \qquad \square$$

In the next chapter we consider a fairly general condition, which implies that

$$\limsup_{n\to\infty} \sup_{\pi\in\Pi} \mathbb{E}_\pi \left| v^*(X_n) \right| = 0 \ ,$$

and thus, by theorem 3.3, implies the convergence of the method of standard successive approximations for scrapvalue 0. This condition excludes the possibility of postponing negative rewards which is essential in the counterexamples 3.2 and 3.6.

3.3. SUCCESSIVE APPROXIMATION METHODS AND GO-AHEAD FUNCTIONS

Besides the method of standard successive approximations considered in the previous section, there are several other successive approximation techniques one could try to use to determine (approximate) the value of an MDP. Three well-known variants of the method of standard successive approximations are the following.

Jacobi or total step iteration.

Choose v_0.
Determine for $n = 0,1,\ldots$

$$v_{n+1}(i) = \sup_{a\in A} \{r(i,a) + \sum_{j\neq i} p(i,a,j)v_n(j) + p(i,a,i)v_{n+1}(i)\} .$$

Gauss-Seidel iteration.

Choose v_0.
Determine for $n = 0,1,\ldots$

$$v_{n+1}(i) = \sup_{a\in A} \{r(i,a) + \sum_{j>i} p(i,a,j)v_n(j) + \sum_{j\leq i} p(i,a,j)v_{n+1}(j)\} .$$

Successive overrelaxation method.

Choose v_0.
Determine for $n = 0,1,\ldots$

$$v_{n+1}(i) = (1-\alpha)v_n(i) + \alpha \sup_{a\in A} \{r(i,a) + \sum_{j>i} p(i,a,j)v_n(j) +$$

$$+ \sum_{j<i} p(i,a,j)v_{n+1}(j) + cv_{n+1}(i) + [p(i,a,i) - c]v_n(i)\} ,$$

with $0 \leq \alpha \leq 1$ and $0 \leq c \leq \inf_{i,a} p(i,a,i)$.

These methods are known from numerical analysis. For example, they can be used for the iterative solution of systems of linear equations, see VARGA [1962, chapter 3].
In the context of MDP's the Gauss-Seidel method has been introduced by HASTINGS [1968] and the method of successive overrelaxation by REETZ [1973] (the special case $\alpha = 1$).
For each of these algorithms, one wants to investigate whether v_n converges to v^*. In order to avoid the necessity of treating these algorithms one

after another, we would like to have a unifying notation which enables us to study these algorithms simultaneously.
Such a unifying notation is the description of successive approximation methods by go-ahead functions as introduced by WESSELS [1977a] and further elaborated by Van NUNEN and WESSELS [1976], Van NUNEN [1976a] and Van NUNEN and STIDHAM [1978]. In order to see that the go-ahead function approach is very natural, consider for example the improvement step in the Gauss-Seidel iteration. In words, we could describe this step as follows.
"In order to obtain $v_{n+1}(i)$ take action a in state i; if the next state is a state $j > i$, then you stop and receive a terminal reward $v_n(j)$, if the next state is a state $j \leq i$, then you go ahead to obtain $v_{n+1}(j)$."

We see that Jacobi iteration, Gauss-Seidel iteration and standard successive approximations are algorithms which can be described by a (go-ahead) function δ from S^2 into $\{0,1\}$; if for a pair of states i,j you have $\delta(i,j) = 1$, then you go ahead after a transition from i to j, and if $\delta(i,j) = 0$, then you stop.
In the successive overrelaxation algorithm, however, the situation is different. First, it has to be decided whether the iteration process will start, which happens with probability α, and then: if in state i action a has been taken and the system makes a transition from state i to i, then we go ahead with probability $c / p(i,a,i)$ and we stop with probability $(p(i,a,i) - c) / p(i,a,i)$.
So in this case the choice between going ahead and stopping has to be made by a random experiment, which at time 1 (and thereafter) also depends on the action a.
Thus the overrelaxation algorithm can be described by a (go-ahead) function δ from $S \cup S \times A \times S$ into $[0,1]$, with $\delta(i) = \alpha$, $\delta(i,a,j) = 1$ if $j < i$, $\delta(i,a,j) = 0$ if $j > i$ and $\delta(i,a,i) = c / p(i,a,i)$, $i \in S$, $a \in A$.

DEFINITION 3.8. *A go-ahead function δ is a map from*

$$S \cup \bigcup_{n=1}^{\infty} (S \times A)^n \cup \bigcup_{n=1}^{\infty} (S \times A)^n \times S$$

into $[0,1]$ *which is measurable with respect to the σ-field generated by S and A.*

The interpretation is as follows:
Let $(i_0,a_0,i_1,\ldots)$ be a realization of the process, then the observation of the process (and its earnings) is stopped at time n before action a_n is chosen with probability $1-\delta(i_0,a_0,\ldots,i_n)$ (provided the observations have not been stopped before) and it is stopped after action a_n is chosen - but before it is executed - with probability $1-\delta(i_0,a_0,\ldots,i_n,a_n)$ (if the observations did not terminate before).
We define the go-ahead function also on $(S\times A)^n$ since this can be used to restore the equal row-sum property in the case of (essentially) sub-stochastic transition matrices (arising e.g. from semi-Markov decision problems with discounting), see Van NUNEN and STIDHAM [1978].

In order to be able to cope with the fact that a go-ahead function not only takes the values 0 and 1, we have to incorporate this random aspect of the go-ahead device in the probability space. Therefore we extend the space $\Omega = (S\times A)^\infty$ to a space $\Omega_0 := (S\times E\times A\times E)^\infty$, where $E := \{0,1\}$. On E we consider the σ-field $\mathcal{E}$ of all subsets, and on $(S\times E\times A\times E)^n$ the σ-field generated by $\mathcal{S}$, $\mathcal{A}$ and $\mathcal{E}$.
As in section 1.5 we can now generate for all $\pi = (\pi_0,\pi_1,\ldots) \in \Pi$, transition probabilities p_0^δ from S into $E\times A\times E\times S$ and p_n^δ from $S\times(E\times A\times E\times S)^n$ into $E\times A\times E\times S$, $n = 1,2,\ldots$, by e.g.

$$p_0^\delta(\{1\}\times C\times\{0\}\times D \mid i_0) = \delta(i_0)\int_C \pi_0(da \mid i_0)(1-\delta(i,a))\sum_{j\in D} p(i_0,a,j) ,$$

and for $n = 1,2,\ldots$

$$p_{n+1}^\delta(\{1\}\times C\times\{0\}\times D \mid i_0,a_0,z_0,i_1,\ldots,z_n,i_{n+1}) = \delta(i_0,a_0,\ldots,i_{n+1})\cdot$$

$$\cdot\int_C \pi_n(da \mid i_0,a_0,\ldots,i_{n+1})(1-\delta(i_0,y_0,\ldots,i_{n+1}))\sum_{j\in D} p(i_{n+1},a,j) ,$$

for all $C \in \mathcal{A}$ and $D \in \mathcal{S}$. Here $y_n = 0$ if the observation of the process stops immediately after i_n has been observed (if it did not stop before) and $y_n = 1$ if we go ahead. And $z_n = 0$ if the observations terminate after action a_n is selected but before it is executed, $z_n = 1$ if the observations continue.
Further we endow Ω_0 with the product σ-field generated by $\mathcal{S}$, $\mathcal{A}$ and $\mathcal{E}$. Then for each $\pi \in \Pi$ the sequence of transition probabilities $\{p_n^\delta,\ n = 0,1,\ldots\}$ defines for each initial state $i \in S$ a probability measure $\mathbb{P}_{i,\pi}^\delta$ on Ω_0 and

a stochastic process $\{(X_n,Y_n,A_n,Z_n),\ n = 0,1,\ldots\}$, where Y_n and Z_n are the outcomes of the random experiments immediately after state X_n has been observed and A_n has been chosen.
We denote by $\mathbb{E}^{\delta}_{i,\pi}$ the expectation operator with respect to $\mathbb{P}^{\delta}_{i,\pi}$.

Next, define the function τ on Ω_0 by

$$\tau(i_0,y_0,a_0,z_0,i_1,y_1,\ldots) = \inf\{n \mid y_n z_n = 0\} .$$

So τ is a stopping time which denotes the time upon which the observation of the process is stopped.
For any $\pi \in \Pi$ and go-ahead function δ, we define the operator $L_\delta(\pi)$ for all $v \in V$ for which the expectation is properly defined by

$$(3.4)\qquad (L_\delta(\pi)v)(i) := \mathbb{E}^{\delta}_{i,\pi}\left[\sum_{n=0}^{\tau-1} r(X_n,A_n) + v(X_\tau)\right], \quad i \in S ,$$

where $v(X_\tau)$ is defined 0 if $\tau = \infty$.
We will see later that $L_\delta(\pi)v$ is properly defined for all $v \in V^+_{u^*}$ (or $v \in V^+_{z^*}$ if $z^* < \infty$).
Further we define $U_\delta v$ by

$$(3.5)\qquad U_\delta v := \sup_{\pi\in\Pi} L_\delta(\pi)v .$$

Note that the improvement step of the algorithms described at the beginning of this section, can now be formulated as

$$v_{n+1} = U_\delta v_n ,$$

with δ the corresponding go-ahead function.

Further, define the operators $L^+_\delta(\pi)$, $L^{abs}_\delta(\pi)$, U^+_δ and U^{abs}_δ by

$$L^+_\delta(\pi)v := \mathbb{E}^{\delta}_{\pi}\left[\sum_{n=0}^{\tau-1} r^+(X_n,A_n) + v(X_\tau)\right],$$

$$U^+_\delta v := \sup_{\pi\in\Pi} L^+_\delta(\pi)v ,$$

$$L^{abs}_\delta(\pi)v := \mathbb{E}^{\delta}_{\pi}\left[\sum_{n=0}^{\tau-1} |r(X_n,A_n)| + v(X_\tau)\right],$$

$$U_\delta^{abs} v := \sup_{\pi \in \Pi} L_\delta^{abs}(\pi) v ,$$

for all $v \in V$ for which the expectations are properly defined.

3.4. THE OPERATORS $L_\delta(\pi)$ AND U_δ

In this section it will be proved that $L_\delta(\pi)$ and U_δ are for all go-ahead functions δ operators on $V_{u^*}^+$ and if $z^* < \infty$ also operators on $V_{z^*}^+$.
The main result of this section is that for all go-ahead functions δ

$$U_\delta v^* = v^* .$$

In order to prove this we need the following basic inequality.
For all $\pi^{(1)}$ and $\pi^{(2)} \in \Pi$ and for all go-ahead functions δ

(3.6) $$L_\delta(\pi^{(1)}) v(\pi^{(2)}) \leq v^* .$$

This result is intuitively clear. Playing strategy $\pi^{(1)}$ until time τ and then switching over to $\pi^{(2)}$ can never yield more than v^*. However, this decision rule is in general (if δ does not take on only the values 0 and 1) not a strategy in the sense of section 1.5. This is caused by the measurability problems which arise from fitting $\pi^{(1)}$ and $\pi^{(2)}$ together at a time, that is determined by the outcomes of a series of random experiments upon which a strategy may not depend. So (3.6) still needs a proof.

The line of reasoning we follow is simple. It only has to be shown that the decisionmaker cannot benefit from knowledge about the outcomes of these random experiments, or any other data that are independent of the future behaviour of the process.
Therefore, let (S,A,p,r) characterize our original MDP and let $(\hat{S},\hat{A},\hat{p},\hat{r})$ be another MDP with $\hat{S} = S$, $\hat{A} = A \times B$, where B is some arbitrary space, $\hat{r}(i,(a,b)) = r(i,a)$ and $\hat{p}(i,(a,b),j) = p(i,a,j)$ for all $i,j \in S$ and $(a,b) \in A \times B$. Let further $\mathcal{B}$ be the σ-field containing all subsets of B and $\hat{\mathcal{A}}$, the σ-field on $\hat{A}$, be the product σ-field generated by $\mathcal{A}$ and $\mathcal{B}$.
So the transition probabilities and the immediate rewards depend on $(a,b) \in A \times B$ only through the first coordinate. (In order to prove (3.6) we will let B contain the outcomes of the random experiments.) To see that the

two MDP's are essentially equivalent, observe the following. Any (randomized) Markov strategy in $(\hat{S},\hat{A},\hat{p},\hat{r})$ induces a (randomized) Markov strategy in (S,A,p,r) and conversely, each (randomized) Markov strategy in (S,A,p,r) yields a whole set of (randomized) Markov strategies in $(\hat{S},\hat{A},\hat{p},\hat{r})$, where these corresponding strategies have the same value.
Marking all objects corresponding to the MDP $(\hat{S},\hat{A},\hat{p},\hat{r})$ by a $\hat{}$ we obtain the following important lemma.

LEMMA 3.9.

$$\hat{u}^* = u^* \ , \quad \hat{v}^* = v^* \quad \textit{and} \quad \hat{z}^* = z^* \ .$$

PROOF. By corollary 2.2 we can restrict ourselves to the consideration of randomized Markov strategies. So the result is immediate from the observed relation between randomized Markov strategies in the two problems. □

THEOREM 3.10. *For all* $\pi^{(1)}$ *and* $\pi^{(2)} \in \Pi$ *and for all go-ahead functions* δ

(i) $\quad L_\delta(\pi^{(1)})v(\pi^{(2)}) \le v^* \ ;$

(ii) $\quad L^+_\delta(\pi^{(1)})u(\pi^{(2)}) \le u^* \ ;$

(iii) $\quad L^{abs}_\delta(\pi^{(1)})z(\pi^{(2)}) \le z^* \ .$

PROOF. We will apply lemma 3.9 with $B = \{0,1\}$. The triple $\pi^{(1)},\delta,\pi^{(2)}$ yields a strategy in $(\hat{S},\hat{A},\hat{p},\hat{r})$, namely the strategy $\hat{\pi}$ defined as follows.
If $b_0 = b_1 = \ldots = b_{n-1} = 1$, then

$$\hat{\pi}_n(C \times \{1\} \mid i_0,a_0,b_0,i_1,\ldots,i_{n-1},a_{n-1},b_{n-1},i_n) =$$

$$= \delta(i_0,a_0,\ldots,i_n) \int_{a\in C} \delta(i_0,a_0,\ldots,i_n,a)\pi^{(1)}_n(da \mid i_0,a_0,\ldots,i_n)$$

and

$$\hat{\pi}(\cup \times \{0\} \mid i_0,\ldots,i_n) = [1-\delta(i_0,\ldots,i_n)]\pi^{(2)}_0(C|i_n) +$$

$$+ \delta(i_0,\ldots,i_n) \int_{a\in A} [1-\delta(i_0,\ldots,i_n,a)]\pi^{(1)}_n(da \mid i_0,\ldots,i_n)\pi^{(2)}_0(C|i_n) \ .$$

And if $b_0 = b_1 = \ldots = b_{t-1} = 1$, $b_t = 0$, $t \le n-1$, then

$$\hat{\pi}(C \times \{0\} \mid i_0,\ldots,i_n) = \pi^{(2)}_{n-t}(C \mid i_t,a_t,\ldots,i_n) ,$$

and

$$\hat{\pi}(C \times \{1\} \mid i_0,\ldots,i_n) = 0 .$$

So $\inf \{n \mid b_n = 0\}$ corresponds with the stopping time τ in the original MDP upon which we switch from strategy $\pi^{(1)}$ to $\pi^{(2)}$. Hence, clearly

$$L_\delta(\pi^{(1)})v(\pi^{(2)}) = \hat{v}(\hat{\pi}) \leq \hat{v}^* = v^* \quad \text{(by lemma 3.9)}.$$

Similarly, one obtains (ii) and (iii). □

COROLLARY 3.11. *For all $\pi \in \Pi$ and all go-ahead functions δ:*

(i) $L_\delta(\pi)$, $L^+_\delta(\pi)$, U_δ *and* U^+_δ *are operators on* $V^+_{u^*}$.

(ii) *If* $z^* < \infty$, *then* $L_\delta(\pi)$, $L^+_\delta(\pi)$, $L^{abs}_\delta(\pi)$, U_δ, U^+_δ *and* U^{abs}_δ *are operators on* V_{z^*}.

PROOF.

(i) One may easily verify that it follows from the monotonicity of $L^+_\delta(\pi)$ and U^+_δ and from $L_\delta(\pi)v \leq L^+_\delta(\pi)v$, that it is sufficient to prove

(3.7) $$L^+_\delta(\pi)u^* \leq u^* , \quad \text{for all } \pi \in \Pi .$$

Let $\pi^{(2)}$ be a strategy with $u(\pi^{(2)}) \geq u^* - \varepsilon e$, then we have for all $\pi \in \Pi$

$$L^+_\delta(\pi)u^* \leq L^+_\delta(\pi)u(\pi^{(2)}) + \varepsilon e) \leq L^+_\delta(\pi)u(\pi^{(2)}) + \varepsilon e \leq u^* + \varepsilon e$$

(by theorem 3.10(ii)).

As $\varepsilon > 0$ can be chosen arbitrarily, we also have (3.7).

(ii) It is sufficient to prove $L^{abs}_\delta(\pi)z^* \leq z^*$ for all $\pi \in \Pi$, the proof of which is identical to the proof of (3.7). □

Similarly we can prove

COROLLARY 3.12. *For all $\pi \in \Pi$ and all go-ahead functions δ we have*

$$L_\delta(\pi)v^* \leq v^* ,$$

hence

$$U_\delta v^* \le v^* .$$

In order to prove $U_\delta v^* \ge v^*$, which together with corollary 3.12 would yield $U_\delta v^* = v^*$, we need the following lemma.

LEMMA 3.13. *For all* $\pi \in \Pi$ *and for all go-ahead functions* δ *we have*

$$\mathbb{E}_\pi^\delta v^*(X_\tau) \ge \mathbb{E}_\pi^\delta \sum_{n=\tau}^{\infty} r(X_n, A_n) .$$

PROOF. Let for any $n \ge 0$ and $(i_0, a_0, i_1, \ldots, a_n)$ the strategy $\pi^{(i_0,a_0,i_1,\ldots,a_n)}$ be defined by

$$\pi_0^{(i_0,a_0,i_1,\ldots,a_n)}(\{a_n\} \mid i_n) = 1 ,$$

and for $k = 1,2,\ldots$

$$\pi_k^{(i_0,a_0,i_1,\ldots,a_n)}(C \mid i_n, a_n, i_{n+1}, \ldots, i_{n+k}) =$$

$$= \pi_{n+k}(C \mid i_0, a_0, i_1, \ldots, i_n, a_n, \ldots, i_{n+k}) , \quad C \in A .$$

Then

$$\mathbb{E}_\pi^\delta \sum_{n=\tau}^{\infty} r(X_n, A_n) = \mathbb{E}_\pi^\delta v(X_\tau, \pi^{(X_0,A_0,\ldots,X_\tau,A_\tau)}) \le \mathbb{E}_\pi^\delta v^*(X_\tau) ,$$

since, clearly, for all $n \ge 0$ and all $i_0, a_0, \ldots, i_n, a_n$

$$v(i_n, \pi^{(i_0,a_0,\ldots,a_n)}) \le v^*(i_n) . \qquad \square$$

From this lemma one immediately has

COROLLARY 3.14. *For all* $\pi \in \Pi$ *and all go-ahead functions* δ *we have*

$$L_\delta(\pi) v^* \ge v(\pi) ,$$

whence also

$$U_\delta v^* \ge \sup_{\pi \in \Pi} v(\pi) = v^* .$$

PROOF. For all $\pi \in \Pi$ and δ,

$$L_\delta(\pi)v^* = \mathbb{E}_\pi^\delta\left[\sum_{n=0}^{\tau-1} r(X_n,A_n) + v^*(X_\tau)\right]$$

$$\geq \mathbb{E}_\pi^\delta\left[\sum_{n=0}^{\tau-1} r(X_n,A_n) + \sum_{n=\tau}^{\infty} r(X_n,A_n)\right] = v(\pi) \ . \qquad \square$$

And finally we obtain from corollaries 3.12 and 3.14,

THEOREM 3.15. *For all go-ahead functions* δ

$$U_\delta v^* = v^* \ , \quad U_\delta^+ u^* = u^* \quad \text{and} \quad U_\delta^{abs} z^* = z^* \ .$$

So it makes sense to study the following successive approximation procedures

$$\begin{cases} \text{Choose } v_0 \text{ (in } V_{u^*}^+ \text{ or, if } z^* < \infty, \text{ in } V_{z^*}). \\ \text{Determine for } n = 0,1,\ldots \\ \\ v_{n+1} = U_\delta v_n \ . \end{cases}$$

Clearly, in order to have v_n converge to v^* one needs conditions on v_0 and the MDP (the reward structure for example). But we do also need a condition on δ. For example, if in the successive overrelaxation algorithm of section 3.3 we have $\alpha = 0$, then $U_\delta v_0 = v_0$ for any $v_0 \in V$, so the method will never converge to v^* if $v_0 \neq v^*$.
Therefore it seems natural to consider go-ahead functions satisfying the following definition.

DEFINITION 3.16. *A go-ahead function* δ *is called nonzero if*

$$\alpha_\delta := \inf_{i\in S} \inf_{a\in A} \delta(i)\delta(i,a) > 0 \ .$$

Note that for the go-ahead function δ which corresponds to the overrelaxation algorithm with $\alpha = 0$ we have $\alpha_\delta = 0$.

3.5. THE RESTRICTION TO MARKOV STRATEGIES IN $U_\delta v$

In general it will not be possible to consider only Markov strategies in the optimization of $L_\delta(\pi)v$, since δ may be history dependent.
An interesting question is now for which go-ahead functions δ can we restrict ourselves to the consideration of Markov strategies, i.e. for which δ do we have

$$(3.8) \qquad U_\delta v = \sup_{\pi \in M} L_\delta(\pi)v \ ,$$

where the supremum is taken componentwise.
In this section we show that for a certain class of go-ahead functions (3.8) does hold.
WESSELS [1977a] and Van NUNEN [1976a] have shown for action-independent go-ahead functions that in the contracting case one can restrict the attention to stationary strategies in the maximization of $U_\delta v$ if $\delta(i_0,\ldots,i_{n+1}) = \delta(i_n,i_{n+1})$ for all $n = 1,2,\ldots$. Go-ahead functions having this property they called "transition memoryless". Van NUNEN and STIDHAM [1978] remarked that this result can be extended to action-dependent go-ahead functions for which $\delta(i_0,\ldots,a_n) = \delta(i_n,a_n)$ and $\delta(i_0,\ldots,a_n,i_{n+1}) = \delta(i_n,a_n,i_{n+1})$ for all $n = 1,2,\ldots$.

DEFINITION 3.17. *A go-ahead function δ is called Markov, if for all $n = 0,1,\ldots$ and all $i_0,a_0,i_1,\ldots$ the probabilities $\delta(i_0,a_0,\ldots,a_n)$ and $\delta(i_0,a_0,\ldots,a_n,i_{n+1})$ only depend on the last two or three coordinates, respectively, and on n. I.e., there exist functions $\delta_0,\delta_1,\ldots$ from $S\times A \cup S\times A\times S$ into $[0,1]$ such that $\delta(i_0,\ldots,i_n,a_n) = \delta_n(i_n,a_n)$ and $\delta(i_0,\ldots,i_n,a_n,i_{n+1}) = \delta_n(i_n,a_n,i_{n+1})$ for all $n = 0,1,\ldots$.*

There is some similarity between the effects of the go-ahead function and the transition law. And as a stochastic process on S^∞ is an (inhomogenous) Markov process if the probabilities $\mathbb{P}(X_{n+1} = j \mid X_0 = i_0,\ldots,X_n = i_n)$ depend on i_n, j and n only, it seems natural to use the term Markov for the go-ahead functions of definition 3.17.
In the terminology of Wessels and Van Nunen one might use the term "time dependent transition memoryless".

Using the similarity between go-ahead functions and transition laws, we prove the following result.

THEOREM 3.18. *If δ is a Markov go-ahead function and $v \in V_{u^*}^+$ (or, if $z^* < \infty$, $v \in V_{z^*}$), then for all $i \in S$*

$$\sup_{\pi \in M} (L_\delta(\pi)v)(i) = (U_\delta v)(i) .$$

PROOF. The line of proof is essentially the same as in the proofs by WESSELS [1977a] and Van NUNEN [1976a] for the result that, for transition memoryless go-ahead functions, one can restrict the attention to stationary strategies in the contracting case.
Incorporating the effects of the go-ahead function in the rewards and transition probabilities, we construct an MDP characterized by $(\hat{S},\hat{A},\hat{p},\hat{r})$ which corresponds in a natural way to the problem of optimizing $L_\delta(\pi)$.
Define $\hat{S} := \{(i,t) \mid i \in S,\ t = 0,1,\ldots\} \cup \{*\}$ and $\hat{A} := A$. Assuming (without loss of generality) that $\delta(i) = 1$ for all $i \in S$, we define for $t = 0,1,\ldots$, $i \in S$ and $a \in A$

$$\begin{cases} \hat{r}((i,t),a) = [1 - \delta_t(i,a)]v(i) + \\ \qquad\qquad + \delta_t(i,a)[r(i,a) + \sum_j p(i,a,j)[1 - \delta_{t+1}(i,a,j)]v(j)] , \\ \hat{p}((i,t),a,(j,t+1)) = \delta_t(i,a)p(i,a,j)\delta_{t+1}(i,a,j) , \quad j \in S , \\ \hat{p}((i,t),a,*) = 1 - \sum_j \delta_t(i,a)p(i,a,j)\delta_{t+1}(i,a,j) , \\ \hat{r}(*,a) = 0 \quad \text{and} \quad \hat{p}(*,a,*) = 1 . \end{cases}$$

As for $v \in V_{u^*}^+$ we have $L_\delta^+(\pi)v^+ < \infty$ for all $\pi \in \Pi$, one easily observes that also the MDP $(\hat{S},\hat{A},\hat{p},\hat{r})$ satisfies condition 1.1. Further, any strategy π for the problem of optimizing $L_\delta(\cdot)v$ yields a strategy $\hat{\pi}$ for the initial states $(i,0)$ in $(\hat{S},\hat{A},\hat{p},\hat{r})$, with $(L_\delta(\pi)v)(i) = \hat{v}((i,0),\hat{\pi})$, and any strategy $\hat{\pi}$ for the MDP $(\hat{S},\hat{A},\hat{p},\hat{r})$ induces a strategy π with $\hat{v}((i,0),\hat{\pi}) = (L_\delta(\pi)v)(i)$. And as Markov strategies in $(\hat{S},\hat{A},\hat{p},\hat{r})$ are also Markov strategies in the original MDP, we have for all $i \in S$

$$(U_\delta v)(i) = \sup_{\hat{\pi} \in \hat{\Pi}} \hat{v}((i,0),\hat{\pi}) = \sup_{\hat{\pi} \in \hat{M}} \hat{v}((i,0),\hat{\pi}) = \sup_{\pi \in M} (L_\delta(\pi)v)(i) . \quad \square$$

Observe that the three algorithms presented in section 3.3 all correspond to Markov go-ahead functions. In fact, the corresponding go-ahead functions belong to an even more restricted class: the set of stationary go-ahead functions.

DEFINITION 3.19. *A go-ahead function* δ *is called stationary if for all* $n = 1,2,\ldots$ *we have* $\delta(i_0,a_0,\ldots,i_n,a_n) = \delta(i_n,a_n)$ *and* $\delta(i_0,a_0,\ldots,i_n,a_n,i_{n+1}) = \delta(i_n,a_n,i_{n+1})$.

Wessels and Van Nunen use the term transition memoryless for these go-ahead functions. For the same reason as mentioned before (the similarity between go-ahead functions and transition laws) we prefer the term stationary.
If δ is stationary, then one can construct an MDP characterized by $(\hat{S},\hat{A},\hat{p},\hat{r})$ which corresponds to the problem of finding $U_\delta v$, which is considerably simpler than the MDP in the proof of theorem 3.18. Namely (assuming again without loss of generality $\delta(i) = 1$ for all $i \in S$), the MDP $(\hat{S},\hat{A},\hat{p},\hat{r})$ with for all $i \in S$ and $a \in A$,

$$
(3.9)\qquad \begin{cases} \hat{S} := S \cup \{*\}\ , \quad \hat{A} := A\ , \\ \hat{r}(i,a) := [1-\delta(i,a)]v(i) + \\ \qquad\qquad + \delta(i,a)[r(i,a) + \sum_j p(i,a,j)[1-\delta(i,a,j)]v(j)]\ , \\ \hat{p}(i,a,j) := \delta(i,a)p(i,a,j)\delta(i,a,j)\ , \quad j \in S\ , \\ \hat{p}(i,a,*) := 1 - \sum_j \delta(i,a)p(i,a,j)\delta(i,a,j)\ , \\ \hat{r}(*,a) := 0 \quad \text{and} \quad \hat{p}(*,a,*) = 1\ . \end{cases}
$$

We see that there is a one-to-one correspondence between strategies in the original MDP and the part of the strategies for $(\hat{S},\hat{A},\hat{p},\hat{r})$ on S. So, if in the MDP $(\hat{S},\hat{A},\hat{p},\hat{r})$ we can restrict ourselves to stationary strategies, then also

$$\sup_{f\in F} L_\delta(f)v = U_\delta v\ , \quad \text{componentwise.}$$

For example, we have

THEOREM 3.20. *Let* δ *be a stationary go-ahead function, then either of the following two conditions guarantees that*

$$
(3.10)\qquad \sup_{f\in F} (L_\delta(f)v)(i) = (U_\delta v)(i) \quad \textit{for all } i \in S\ .
$$

(i) $r(i,a) \geq 0$ *for all* $i \in S$ *and* $a \in A$ *and* $v \geq 0$.
(ii) A *is finite and* $v \in V^+_{u^*}$.

PROOF. Immediate from theorems 2.22 and 2.23. □

If the action set in state i, A(i), depends on i, then we may replace in theorem 3.20(i) the condition A is finite by A(i) is finite for all $i \in S$.

In chapter 4 we consider another condition which guarantees (3.10) to hold.

3.6. VALUE-ORIENTED SUCCESSIVE APPROXIMATIONS

Another variant of the method of standard successive approximations is the method of value-oriented (standard) successive approximations.
In all successive approximation methods considered in the previous sections, the algorithm consisted of a sequence of optimization steps. In the value-oriented methods, each optimization step is followed by some kind of extrapolation.

Value-oriented standard successive approximations

Choose $v_0 \in V^+_{u^*}$, $\lambda \in \{1,2,\ldots\}$ and a sequence $\{d_n, n = 0,1,\ldots\}$ of strictly positive real-valued functions on S ($d_n(i) > 0$ for all $i \in S$) with $d_n \to 0$ $(n \to \infty)$.
Determine for $n = 0,1,\ldots$ a policy f_{n+1} such that

$$L(f_{n+1})v_n \geq Uv_n - d_n \tag{3.11}$$

and define

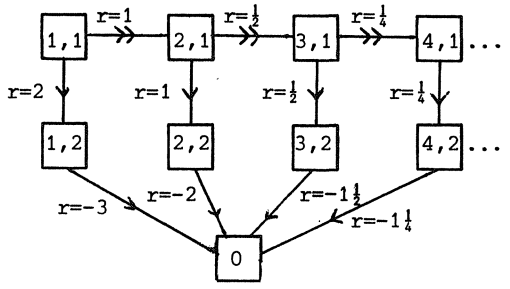
$$v_{n+1} = L^{\lambda}(f_{n+1})v_n \,. \tag{3.12}$$

The reason why we consider arbitrary functions d_n and not just functions $\varepsilon_n e$ will become clear in section 4.8.
So, after each optimization step we determine a policy f_{n+1}, then v_{n+1} is obtained by using f_{n+1} during λ periods of time in the λ-stage MDP with terminal pay-off v_n. This can be seen as a kind of extrapolation.
Note that if A is finite, we do not need the functions d_n and (3.11) can be replaced by $L(f_{n+1})v_n = Uv_n$.
For the contracting MDP this method has been first mentioned (without convergence proof) by PORTEUS [1971]. Van NUNEN [1976a,1976c] has proved that in the contracting case the value-oriented method converges, i.e., v_n

converges to v^* for suitably chosen scrap-values v_0.
The value-oriented variant can also be formulated for the successive approximations methods generated by a go-ahead function δ. In that case (3.11) and (3.12) are replaced by $L_\delta(\pi^{(n+1)})v_n \geq U_\delta v_n - d_n$ and $v_{n+1} = L_\delta^\lambda(\pi^{(n+1)})v_n$, respectively.
For the contracting MDP the combination of go-ahead functions and value-oriented methods has been studied by Van NUNEN [1976a] and Van NUNEN and WESSELS [1977b].
Here we only consider the value-oriented variant of the method of standard successive approximations.

It is clear that the value-oriented method will not converge in general, since for $v_0 = 0$ the method of standard successive approximations not even needs to converge. But even if the method of standard successive approximations does converge, then the value-oriented method may not converge.

EXAMPLE 3.21. $S = \{0\} \cup \{(i,k) \mid i \in \{1,2,\ldots\}, k \in \{1,2\}\}$, $A = \{1,2\}$,



$r((i,1),1) = 2^{-i+2}$, $r((i,1),2) = 2^{-i+1}$, $p((i,1),1,(i,2)) = p((i,1),2,(i+1,1)) = 1$, $r((i,2),a) = -1 - 2^{-i+2}$, $p((i,2),a,0) = p(0,a,0) = 1$, $r(0,a) = 0$, $a \in A$.
Consider the case $v_0 = 0$, $\lambda = 2$, $d_n = 0$, $n = 0,1,\ldots$. Clearly the method of standard successive approximations converges as $U0 = v^*$. However, as one may verify, the sequence v_n obtained for the value-oriented variant converges to a vector $\hat{v}$ with $\hat{v}((i,1)) = v^*((i,1)) - 1$.

Conversely, the method of value-oriented successive approximations may converge in cases where the method of standard successive approximations does not converge. For example, consider the MDP of example 3.2. With $v_0 = 0$ the method of standard successive approximations does not converge whereas the value-oriented method converges for all $\lambda > 1$.

The question of convergence is somewhat more transparent in the following monotonic version of the value-oriented method.

Monotone value-oriented standard successive approximations

Choose some $v_0 \in V^+_{u^*}$ for which there exists a policy f such that $L(f)v_0 \geq v_0$; choose $\lambda \in \{1,2,\ldots\}$ and a sequence of real-valued functions $\{d_n, n = 0,1,\ldots\}$ on S with $d_n > 0$ for all $n = 0,1,\ldots$ and $\lim_{n\to\infty} d_n = 0$.
Determine for $n = 0,1,\ldots$ a policy f_{n+1} such that

$$L(f_{n+1})v_n \geq \max\{v_n, Uv_n - d_n\} \tag{3.13}$$

and define

$$v_{n+1} = L^\lambda(f_{n+1})v_n .$$

Since $L(f_n)v_n \geq v_n$ for all $n = 0,1,\ldots$ (as one may easily show by induction) there exists for all n a policy f_{n+1} satisfying (3.13). It is also clear (from $v_0 \in V^+_{u^*}$) that Uv_n is properly defined for all n.
As $v_{n+1} \geq v_n$ for all n, the sequence $\{v_n\}$ converges to a limit, $\hat{v}$ say. And also $\hat{v} \in V^+_{u^*}$.
The question remains, when do we have $\hat{v} = v^*$. In chapter 4, section 8, we consider a rather general condition, which guarantees that the monotone variant converges for all $v_0 \in V_{z^*}$ for which there exists a policy f such that $L(f)v_0 \geq v_0$.
Here we only prove the following result (cf. theorem 3.7).

THEOREM 3.22. *For all* $v_0 \in V$ *with* $w^* \leq v_0 \leq v^*$ *for which there exists a policy* f *with* $L(f)v_0 \geq v_0$, *the monotone value-oriented standard successive approximations method converges.*

PROOF. Let $\{v_n, n = 0,1,\ldots\}$ be the sequence generated by the method and $\hat{v} = \lim_{n\to\infty} v_n$. As $v_0 \leq v^*$, we have by induction $v_n \leq v^*$ for all $n = 1,2,\ldots$.
Namely, suppose $v_n \leq v^*$, then $v_{n+1} = L^\lambda(f_{n+1})v_n \leq U^\lambda v_n \leq U^\lambda v^* = v^*$. Thus also $\hat{v} \leq v^*$.
But from (3.13) and the monotonicity of $L(f_{n+1})$ we also have

$$v_{n+1} = L^\lambda(f_{n+1})v_n \geq L(f_{n+1})v_n \geq Uv_n - d_n .$$

Letting n tend to infinity, we get with the monotonicity of v_n and $d_n \to 0$ $(n \to \infty)$

$$\hat{v} \geq U\hat{v} \ .$$

Hence also

$$\hat{v} \geq \lim_{n\to\infty} U^n\hat{v} \geq \lim_{n\to\infty} U^n w^* = v^* \quad \text{(by theorem 3.7)}.$$

So $\hat{v} = v^*$, which completes the proof. □

COROLLARY 3.23. *If for all* $i \in S$ *and* $a \in A$ *we have* $r(i,a) \geq 0$, *then the monotone value-oriented method converges for scrapvalue* $v_0 = 0$.

CHAPTER 4

THE STRONGLY CONVERGENT MDP

4.1. *INTRODUCTION*

In chapters 2 and 3 we analysed the general total-reward MDP. We have seen that in general the method of (standard) successive approximations does not converge, and that in general nearly-optimal stationary strategies do not exist.

In this chapter we study the total-reward MDP under some additional assumptions concerning the absolute values of the income streams. Mostly we have assumptions at least as strong as

CONDITION 4.1.

$$(4.1) \qquad \text{(i)} \quad z^* = \sup_{\pi \in \Pi} \mathbb{E}_\pi \sum_{n=0}^{\infty} |r(X_n, A_n)| < \infty \ ,$$

and

$$(4.2) \qquad \text{(ii)} \quad \lim_{n \to \infty} \sup_{\pi \in M} \sum_{k=n}^{\infty} \mathbb{E}_\pi |r(X_k, A_k)| = 0 \ .$$

Condition 4.1(ii) is also called the *uniform tail condition*.

It will be clear that (4.2) implies that the ∞-horizon MDP can be approximated by finite-stage MDP's, so that the method of standard successive approximations with scrapvalue 0 will converge.

A main point in this introductory section is to show that condition 4.1 is equivalent to a so-called strong convergence condition, and implies

$$(4.3) \qquad \lim_{n \to \infty} \sup_{\pi \in M} \mathbb{E}_\pi z^*(X_n) = 0 \ .$$

From this in sections 2, 3 and 4 we obtain the equivalence of conservingness and optimality, the convergence of the method of standard successive

approximations for any scrapvalue $v_0 \in V_{z^*}$ and the convergence of the policy iteration method. In section 5 it is shown how the strong convergence condition relates to the concept of Liapunov functions as introduced in the context of MDP's by HORDIJK [1974]. In section 6 the convergence is studied of the successive approximations methods, generated in section 3.3 by means of go-ahead functions. In section 7 it is shown that for stationary go-ahead functions one needs to consider only stationary strategies in the optimization of $L_\delta(\cdot)v$. Finally, in section 8, we consider the method of value-oriented successive approximations.

The results of sections 1, 3, 4, 5 and part of section 2 can also be found in Van HEE, HORDIJK and Van der WAL [1977] and/or Van HEE and Van der WAL [1977]. Related results can be found in SCHÄL [1975] and STIDHAM [1978]. Stidham also compares the conditions in Van HEE, HORDIJK and Van der WAL [1977], SCHÄL [1975] and STIDHAM [1978].

In the remainder of this introductory section we establish the equivalence of condition 4.1 to the strong convergence condition (condition 4.2).
To formulate this condition define Φ as the set of all sequences $\varphi = (\varphi_0,\varphi_1,\ldots)$ with $\varphi_n \in V$ for all $n = 0,1,\ldots$, with $\varphi_0 \geq e$, $\varphi_{n+1} \geq \varphi_n$ for all $n = 0,1,\ldots$, and with $\lim_{n\to\infty} \varphi_n = \infty$ (pointwise). (So, $\varphi_0(i) \geq 1$ and $\varphi_n(i) \uparrow \infty$ $(n \to \infty)$ for all $i \in S$.)
For all $\varphi \in \Phi$ we define

$$(4.4) \qquad z_\varphi(i,\pi) := \mathbb{E}_{i,\pi} \sum_{n=0}^{\infty} \varphi_n(i)\,|r(X_n,A_n)| \,, \quad i \in S \,, \ \pi \in \Pi \,,$$

$$(4.5) \qquad z_\varphi^*(i) := \sup_{\pi\in M} z_\varphi(i,\pi) \,.$$

CONDITION 4.2 (Strong-convergence condition). *There exists a $\varphi \in \Phi$ for which*

$$z_\varphi^* < \infty \,.$$

An MDP which satisfies the strong convergence condition is called *strongly convergent*.

In order to show that condition 4.2 is equivalent to condition 4.1, we first derive the following lemma from which we see that condition 4.2 implies condition 4.1:

LEMMA 4.3. *For all* $\varphi = (\varphi_0,\varphi_1,\ldots) \in \Phi$ *and all* $n = 0,1,\ldots$ *we have*

$$\text{(i)} \qquad \mathbb{E}_{i,\pi} \sum_{k=n}^{\infty} |r(X_k,A_k)| \le (\varphi_n(i))^{-1} z_\varphi(i,\pi) , \quad i \in S , \ \pi \in M ,$$

and thus

$$\text{(ii)} \qquad \sup_{\pi\in M} \mathbb{E}_{i,\pi} \sum_{k=n}^{\infty} |r(X_k,A_k)| \le (\varphi_n(i))^{-1} z^*_\varphi(i) , \quad i \in S .$$

PROOF. It is sufficient to prove (i) since (ii) follows from (i) immediately. For all $i \in S$ and all $\pi \in M$,

$$\mathbb{E}_{i,\pi} \sum_{k=n}^{\infty} |r(X_k,A_k)| \le \mathbb{E}_{i,\pi} \sum_{k=n}^{\infty} (\varphi_n(i))^{-1} \varphi_k(i) |r(X_k,A_k)|$$

$$\le (\varphi_n(i))^{-1} \mathbb{E}_{i,\pi} \sum_{k=0}^{\infty} \varphi_k(i) |r(X_k,A_k)| = (\varphi_n(i))^{-1} z_\varphi(i,\pi) . \qquad \square$$

So we see that, if $z^*_\varphi < \infty$ for some $\varphi \in \Phi$, then condition 4.1 is satisfied.

THEOREM 4.4. *An MDP is strongly convergent if and only if condition 4.1 holds.*

PROOF. As we already remarked, the "only if" part follows immediately from lemma 4.3.
In order to prove the "if" part we construct a sequence φ for which $z^*_\varphi < \infty$. First define

$$b_n(i) := \sup_{\pi\in M} \sum_{k=n}^{\infty} \mathbb{E}_{i,\pi} |r(X_k,A_k)| , \quad i \in S , \ n = 0,1,\ldots .$$

Clearly, $b_n \ge b_{n+1}$. Next, for $i \in S$ define

$$N_0(i) := 0$$

and

$$N_{\ell+1}(i) := \max \{\min \{n \mid b_n(i) \le 2^{-\ell-1}\}, N_\ell(i) + 1\} , \quad \ell = 0,1,\ldots .$$

And finally define φ by

$$\varphi_n(i) := \ell + 1 \quad \text{if} \quad N_\ell(i) \le n < N_{\ell+1}(i) , \quad i \in S , \ n = 0,1,\ldots .$$

Then

$$\sup_{\pi\in M} \sum_{k=N_\ell(i)}^{N_{\ell+1}(i)-1} \varphi_k(i)\, \mathbb{E}_{i,\pi} |r(X_k,A_k)| \le (\ell+1)2^{-\ell} , \quad \ell = 1,2,\ldots .$$

Consequently, for all $i \in S$,

$$z_\varphi^*(i) \le \sup_{\pi\in M} \sum_{k=0}^{N_1(i)-1} \mathbb{E}_{i,\pi} |r(X_k,A_k)| + \sum_{\ell=1}^{\infty} (\ell+1)2^{-\ell} \le z^*(i) + 3 < \infty ,$$

which completes the proof. □

Next we show that condition 4.1 (or condition 4.2) implies (4.3). To prove this we need the following lemma.

LEMMA 4.5. *For all* $n = 0,1,\ldots$

$$\sup_{\pi\in M} \mathbb{E}_\pi z^*(X_n) = \sup_{\pi\in M} \mathbb{E}_\pi \sum_{k=n}^{\infty} |r(X_k,A_k)| .$$

PROOF. Let $\pi = (f_0,f_1,\ldots) \in M$ be arbitrary, then conditioning on X_n yields

$$\mathbb{E}_\pi \sum_{k=n}^{\infty} |r(X_k,A_k)| = \mathbb{E}_\pi z(X_n,\pi^{\leftarrow n}) \le \mathbb{E}_\pi z^*(X_n) ,$$

where $\pi^{\leftarrow n} = (f_n,f_{n+1},\ldots)$.
So, it is sufficient to show that

$$\sup_{\pi\in M} \mathbb{E}_\pi \sum_{k=n}^{\infty} |r(X_k,A_k)| \ge \sup_{\pi\in M} \mathbb{E}_\pi z^*(X_n) .$$

Let $\varepsilon > 0$ be arbitrary, then by theorem 2.23 a stationary strategy f exists with

$$z(f) \ge z^*(1-\varepsilon) .$$

Then, with $\tilde{\pi} = (f_0,f_1,\ldots,f_{n-1},f,f,\ldots)$,

$$\mathbb{E}_{\tilde{\pi}} \sum_{k=n}^{\infty} |r(X_k,A_k)| = \mathbb{E}_{\tilde{\pi}} z(X_n,f) = \mathbb{E}_\pi z(X_n,f) \ge (1-\varepsilon)\mathbb{E}_\pi z^*(X_n) .$$

Hence also

$$\sup_{\pi\in M} \mathbb{E}_\pi \sum_{k=n}^{\infty} |r(X_k,A_k)| \geq (1-\varepsilon) \sup_{\pi\in M} \mathbb{E}_\pi z^*(X_n) \ .$$

As $\varepsilon > 0$ is arbitrary, the proof is complete. □

THEOREM 4.6. *If condition 4.1 or condition 4.2 holds, then*

(i) $\lim_{n\to\infty} \sup_{\pi\in M} \mathbb{E}_\pi z^*(X_n) = 0$;

(ii) $\lim_{n\to\infty} \tilde{U}^n v = 0$ *for all* $v \in V_{z^*}$ *(for the definition of* $\tilde{U}$ *see* (1.27)*)*.

PROOF. (i) follows immediately from lemmas 4.5 and 4.3(ii), and (ii) follows from (i), with $\tilde{U}^n v = \sup_{\pi\in M} \mathbb{E}_\pi v(X_n)$ and $|v| \leq cz^*$ for some $c \in \mathbb{R}$. □

An important consequence of theorem 4.6 is

THEOREM 4.7. *If condition 4.1 or 4.2 holds, then* v^* *is the unique solution of the optimality equation* $Uv = v$ *within* V_{z^*}.

PROOF. Clearly $v^* \in V_{z^*}$, and by theorem 2.11 v^* solves the optimality equation. So it remains to prove the uniqueness.
From theorem 2.7 we know that U and $\tilde{U}$ map V_{z^*} into itself. Let $\hat{v}$ be a solution of $Uv = v$, with $\hat{v} \in V_{z^*}$, and therefore $|\hat{v} - v^*| \in V_{z^*}$ as well. Then

$$|\hat{v} - v^*| = |U\hat{v} - Uv^*| \leq \tilde{U}|\hat{v} - v^*| \ . \tag{4.6}$$

The inequality in (4.6) holds for any two functions v and w in V_{z^*}, as follows from

$$Uv - Uw = \sup_{f\in F} [L(f)v - Uw] \leq \sup_{f\in F} [L(f)v - L(f)w]$$

$$= \tilde{U}(v-w) \leq \tilde{U}|v-w| \ .$$

Iterating (4.6) yields

$$|\hat{v} - v^*| \leq \tilde{U}^n|\hat{v} - v^*| \ .$$

So, letting n tend to infinity, and using theorem 4.6(ii), we obtain

$$|\hat{v} - v^*| = 0 \ ,$$

which proves the uniqueness within V_{z^*}. □

4.2. CONSERVINGNESS AND OPTIMALITY

In this section it is investigated whether the concepts (nearly-) conserving and (nearly-) optimal coincide if the MDP is strongly convergent (i.e., if condition 4.1 or 4.2 is satisfied).
A policy f is called *conserving* if

$$L(f)v^* = v^* , \tag{4.7}$$

and called ε-*conserving* if

$$L(f)v^* \geq v^* - \varepsilon e , \quad \varepsilon > 0 . \tag{4.8}$$

So, conserving policies preserve the possibility of ultimately obtaining v^*. However, as we see from example 3.2, a conserving policy need not yield an optimal stationary strategy. Namely, let f be a policy with $f(1) = 1$, then $L(f)v^* = v^*$ but $v(1,f) = 0 < v^*(1) = 1$.
For a conserving policy f to yield an optimal stationary strategy, it is necessary that f is also *equalizing*, i.e. that

$$\mathbb{E}_f v^*(X_n) \to 0 \quad (n \to \infty) ,$$

and this condition is not satisfied in example 3.2.
From theorem 4.6(i) and the fact that $|v^*| \leq z^*$ it follows that for a strongly convergent MDP all strategies are equalizing, so in this case conservingness and optimality coincide.
The case of ε-conservingness is somewhat more complicated.
The notions conserving and equalizing were introduced by DUBINS and SAVAGE [1965], and have been used in the context of MDP's by HORDIJK [1974] and GROENEWEGEN [1978] to characterize optimal strategies.

THEOREM 4.8. *If the MDP is strongly convergent, then we have:*

(i) *if* f *is conserving, then* $v(f) = v^*$,
(ii) *for all* $\varepsilon > 0$ *there exists a Markov strategy* π *satisfying*

$$v(\pi) \geq v^* - \varepsilon e .$$

PROOF.
(i) Iterating (4.7) yields

$$v_n(f) = L^n(f)0 = L^n(f)v^* - \mathbb{E}_f v^*(X_n) = v^* - \mathbb{E}_f v^*(X_n) .$$

Letting n tend to infinity the result now follows with $|v^*| \le z^*$ from theorem 4.6 (see also theorem 2.16).

(ii) Construct $\pi = (f_0, f_1, \ldots) \in M$ along the lines of (2.17). Then use theorem 2.14, since (2.18) follows from theorem 4.6(i) with $|v^*| \le z^*$. □

The strong-convergence condition implies that one needs to consider only stationary strategies in the optimization of $v(i,\pi)$ for a fixed initial state.

THEOREM 4.9. *If the MDP is strongly convergent, then for any* $\varepsilon > 0$ *and any initial state* $i \in S$ *a stationary strategy* f *exists satisfying*

$$v(i,f) \ge v^*(i) - \varepsilon .$$

PROOF. Let $\varphi \in \Phi$ be such that

$$z_\varphi^* < \infty .$$

Let n be so large that

$$(\varphi_n(i))^{-1} z_\varphi^*(i) < \frac{\varepsilon}{3} ,$$

and let f be $\frac{\varepsilon}{3n}$-conserving, i.e.

$$L(f)v^* \ge v^* - \frac{\varepsilon}{3n} e .$$

Then

$$v(i,f) \ge v_n(i,f) - \mathbb{E}_{i,f} \sum_{k=n}^{\infty} |r(X_k, A_k)|$$

$$\ge (L^n(f)0)(i) - \frac{\varepsilon}{3} = (L^n(f)v^*)(i) - \mathbb{E}_{i,f} v^*(X_n) - \frac{\varepsilon}{3}$$

$$\ge v^*(i) - n\frac{\varepsilon}{3n} - \frac{\varepsilon}{3} - \frac{\varepsilon}{3} = v^*(i) - \varepsilon .$$ □

In theorems 4.10 and 4.11 some results on the existence of uniformly nearly-optimal strategies for the strongly convergent MDP are given.

THEOREM 4.10. *If for some* $\varphi \in \Phi$ *we have* $z_\varphi^* < \infty$ *and* $(\varphi_n(i))^{-1} z_\varphi^*(i)$ *converges to zero uniformly on* S, *then for every* $\varepsilon > 0$ *a stationary strategy* f *exists satisfying* $v(f) \ge v^* - \varepsilon e$.

PROOF. The proof is almost identical to the proof of theorem 4.9. □

THEOREM 4.11. *Let* $\varphi \in \Phi$. *If* $\varphi_n \to \infty$ *uniformly on* S, *and if* $z_\varphi^* < \infty$, *then for every* $\varepsilon > 0$ *a stationary strategy* f *exists satisfying*

$$v(f) \geq v^* - \varepsilon z_\varphi^* .$$

PROOF. Fix $\varepsilon > 0$. Choose n such that

$$\varphi_n \geq \frac{3}{\varepsilon} e ,$$

and a policy f satisfying

$$L(f)v^* \geq v^* - \frac{\varepsilon}{3n} z^* .$$

Then

$$v(f) = L^n(f)0 + \mathbb{E}_f v(X_n,f) = L^n(f)v^* + \mathbb{E}_f[v(X_n,f) - v^*(X_n)]$$

$$\geq v^* - \frac{\varepsilon}{3n} \sum_{k=0}^{n-1} P^k(f)z^* - 2\mathbb{E}_f z^*(X_n)$$

$$\geq v^* - \frac{\varepsilon}{3} z^* - 2\frac{\varepsilon}{3} z_\varphi^* \geq v^* - \varepsilon z_\varphi^* ,$$

where we used

$$P^k(f)z^* \leq z^* \leq z_\varphi^* ,$$

and lemmas 4.5 and 4.3(ii). □

Note the following. If state j can be reached from state i, i.e., for some $n \geq 0$ and some $\pi \in M$ we have

$$\mathbb{P}_{i,\pi}(X_n = j) > 0 ,$$

then clearly

$$\sum_{k=0}^{\infty} \varphi_{k+n}(i)\mathbb{E}_{j,\pi^\circ} |r(X_k,A_k)| < \infty \quad \text{for all } \pi^\circ \in M.$$

So, certainly (as $\varphi_{k+n} \geq \varphi_k$)

$$\sum_{k=0}^{\infty} \varphi_k(i) \mathbb{E}_{j,\pi^\circ} |r(X_k, A_k)| < \infty \quad \text{for all } \pi^\circ \in M .$$

Hence, if the MDP is strongly convergent, and if there exists a state $i \in S$ from which all states can be reached, then there also exists a $\varphi \in \Phi$ with $\varphi_n \to \infty$ uniformly on S, for which $z_\varphi^* < \infty$.

In chapter 5, where the contracting MDP is studied, theorem 4.11 will be applied. Theorem 4.10 can be used in the discounted MDP with bounded rewards.

4.3. *STANDARD SUCCESSIVE APPROXIMATIONS*

Consider the standard successive approximations scheme

$$(4.9) \quad \begin{cases} \text{Choose } v_0 \in V_{z^*}. \\ \text{Determine for } n = 0,1,\ldots \\ \qquad v_{n+1} := Uv_n . \end{cases}$$

By theorem 2.7 this scheme is properly defined for all $v_0 \in V_{z^*}$. If the MDP is strongly convergent, then we see from theorem 4.6(i) that the condition in theorem 3.3 holds, hence the method of standard successive approximations with scrapvalue 0 converges.
The following theorem states: if the MDP is strongly convergent, then the scheme (4.9) converges for all $v_0 \in V_{z^*}$.

THEOREM 4.12. *If the MDP is strongly convergent then for all* $v_0 \in V_{z^*}$

$$\lim_{n\to\infty} v_n = v^* .$$

PROOF. The proof is similar to the proof of theorem 4.7.
Let $v_0 \in V_{z^*}$ then

$$|v_{n+1} - v^*| = |Uv_n - Uv^*| \le \tilde{U}|v_n - v^*| \le \ldots \le \tilde{U}^{n+1}|v_0 - v^*| .$$

Since $v_0 - v^* \in V_{z^*}$ we have by theorem 4.6(ii)

$$\tilde{U}^n|v_0 - v^*| \to 0 \quad (n \to \infty) .$$

Hence

$$\lim_{n\to\infty} v_n = v^* \ . \qquad \square$$

4.4. THE POLICY ITERATION METHOD

Another technique, which can be used for approximating v^* is the policy iteration method introduced by HOWARD [1960] for the discounted MDP with finite state and action spaces.

Policy iteration method

Choose some initial policy f_0 and a sequence of constants $\{\varepsilon_n, n = 0,1,\ldots\}$ with $\varepsilon_n > 0$ and $\lim_{n\to\infty} \varepsilon_n = 0$.
Define $v_0 := v(f_0)$.
Determine for $n = 0,1,\ldots$ a policy f_{n+1} satisfying

$$L(f_{n+1})v_n \geq \max\{v_n, Uv_n - \varepsilon_n e\} \ , \tag{4.10}$$

and define

$$v_{n+1} = v(f_{n+1}) \ .$$

In this section it will be shown that, if the MDP is strongly convergent, the policy iteration method converges, i.e.,

$$\lim_{n\to\infty} v_n = v^* \quad \text{(pointwise)}. \tag{4.11}$$

In the remainder of this section we assume that MDP under consideration to be strongly convergent.
In order to prove (4.11) we first show that the sequence v_n converges monotonically to a limit, $\hat{v}$ say, with $\hat{v} \in V_{z^*}$ and $\hat{v} \leq v^*$.

LEMMA 4.13. *If for some* $v \in V_{z^*}$ *and some* $f \in F$ *we have* $L(f)v \geq v$, *then* $v(f) \geq L(f)v$.

PROOF. Iterating $L(f)v \geq v$ we get by the monotonicity of $L(f)$

$$v(f) = \lim_{n\to\infty} [L^n(f)v - P^n(f)v] \geq L(f)v - \limsup_{n\to\infty} P^n(f)v \ .$$

From theorem 4.6(ii) it follows that $P^n(f)v \to 0$ $(n \to \infty)$. Hence

$$v(f) \geq L(f)v \ . \qquad \square$$

Using (4.10), we immediately obtain from lemma 4.13:

COROLLARY 4.14. *Let f_0 be an arbitrary policy, and let $\{v_n\}$ be a sequence obtained by the policy iteration method with initial policy f_0, then* $\lim_{n\to\infty} v_n$ *exists. Further it is clear that*

$$\lim_{n\to\infty} v_n \leq v^* \quad \textit{and} \quad \left|\lim_{n\to\infty} v_n\right| \leq z^* \ .$$

LEMMA 4.15. *Let $\{v_n\}$ be a sequence as in corollary 4.14 and let $\hat{v}$ be defined by* $\hat{v} := \lim_{n\to\infty} v_n$, *then*

$$U\hat{v} \leq \hat{v} \ .$$

PROOF. By (4.10) and lemma 4.13 we have

$$v_{n+1} \geq L(f_{n+1})v_n \geq Uv_n - \varepsilon_n e \ .$$

So, with the monotonicity of v_n, we obtain letting $n \to \infty$,

$$\hat{v} \geq U\hat{v} \ . \qquad \square$$

Now we can prove

THEOREM 4.16. *Let $\{v_n\}$ be as in lemma 4.15. Then*

$$\lim_{n\to\infty} v_n = v^* \ .$$

PROOF. By corollary 4.14 the limit $\hat{v}$ of the sequence v_n satisfies $\hat{v} \leq v^*$. It remains to be shown that $\hat{v} \geq v^*$. Since $\hat{v} \in V_{z^*}$ (corollary 4.14) it follows from lemma 4.15, the monotonicity of U and theorem 4.12, that

$$v^* = \lim_{n\to\infty} U^n\hat{v} \leq \hat{v} \ .$$

Hence $\hat{v} = v^*$. $\square$

4.5. STRONG CONVERGENCE AND LIAPUNOV FUNCTIONS

Consider a sequence of functions $\ell_1,\ell_2,\ldots$ from S into $[0,\infty]$ satisfying for all $f \in F$

$$(4.12)\qquad \begin{cases} \ell_1 \geq |r(f)| + P(f)\ell_1 , \\ \ell_{n+1} \geq \ell_n + P(f)\ell_{n+1} , \quad n = 0,1,\ldots . \end{cases}$$

A set of *finite* functions $\ell_1,\ell_2,\ldots,\ell_m$ satisfying (4.12) upto n = m, is called a *system of Liapunov functions of order* m for the MDP.
In the context of MDP's Liapunov functions are first studied by HORDIJK [1974, chapters 4 and 5] and [1976]. The relation between the existence of a system of Liapunov functions for the MDP and the method of standard successive approximations has been studied by Van HEE, HORDIJK and Van der WAL [1977]. It is shown for example, that the existence of a system of Liapunov functions of order 2 implies the convergence of the standard successive approximations to v^*, for all scrapvalues in V_{z^*}.
In this section we consider the relation between the existence of Liapunov functions of order m and special sequences $\varphi \in \Phi$ for which z^*_φ is finite.

First define the sequence $\{y_n\}$

$$(4.13)\qquad \begin{cases} y_1 := z^* , \\ y_{n+1} := \sup_{\pi\in M} \mathbb{E}_\pi \sum_{k=0}^{\infty} y_n(X_k) , \quad n = 0,1,\ldots . \end{cases}$$

So, y_n may be equal to ∞.
Then we have the following result.

THEOREM 4.17. *Let $\ell_1,\ell_2,\ldots,\ell_m$ be a system of Liapunov functions of order* m *for the MDP. Then for the functions $y_1,\ldots,y_m$ defined in* (4.13) *we have*

$$y_n \leq \ell_n , \quad n = 1,2,\ldots,m .$$

PROOF. The proof proceeds for fixed m by induction on n. First we examine the case n = 1. From

$$\ell_1 \geq |r(f)| + P(f)\ell_1 = L^{abs}(f)\ell_1 \quad \text{for all } f \in F ,$$

we have for any $\pi = (f_0,f_1,\ldots) \in M$ and all $k = 0,1,\ldots$

$$(4.14) \qquad z_k(\pi) := L^{abs}(f_0)\cdots L^{abs}(f_{k-1})0 \le L^{abs}(f_0)\cdots L^{abs}(f_{k-1})\ell_1$$

$$\le L^{abs}(f_0)\cdots L^{abs}(f_{k-2})\ell_1 \le \ldots \le \ell_1 .$$

Hence also

$$z^* = \sup_{\pi\in M} \lim_{k\to\infty} z_k(\pi) \le \ell_1 ,$$

which completes the proof for $n = 1$. Note that this is a special case of theorem 2.12.

Now assume $y_n \le \ell_n$ for some $n < m$. Then one easily shows along the lines of (4.14) that

$$\ell_{n+1} \ge \sup_{\pi\in M} \mathbb{E}_\pi \sum_{k=0}^{\infty} \ell_n(X_k) .$$

Hence

$$\ell_{n+1} \ge \sup_{\pi\in M} \mathbb{E}_\pi \sum_{k=0}^{\infty} \ell_n(X_k) \ge \sup_{\pi\in M} \mathbb{E}_\pi \sum_{k=0}^{\infty} y_n(X_k) = y_{n+1} .$$

So $\ell_n \ge y_n$ for all $n = 1,2,\ldots,m$. □

We see that the existence of a system of Liapunov functions of order m implies that y_m is finite. The following theorem relates the finiteness of y_m to a special sequence $\varphi \in \Phi$ for which z^*_φ is still finite.

THEOREM 4.18. *For all* $m = 1,2,\ldots$

$$(4.15) \qquad y_m \ge \sup_{\pi\in M} \mathbb{E}_\pi \sum_{k=0}^{\infty} \binom{m+k-1}{k} |r(X_k,A_k)| .$$

PROOF. The proof proceeds by induction on m. For $m = 1$ formula (4.15) holds by definition (with equality). Assume (4.15) holds for $m = n$. Then for all $\pi = (f_0,f_1,\ldots) \in M$

$$y_{n+1} \ge \sum_{k=0}^{\infty} P(f_0)\cdots P(f_{k-1})y_n ,$$

where $P(f_0)P(f_{-1})$ is defined to be the identity operator. So

$$y_{n+1} \geq \sum_{k=0}^{\infty} P(f_0) \cdots P(f_{k-1}) \sum_{t=0}^{\infty} \binom{n+t-1}{t} P(f_k) \cdots P(f_{k+t-1}) \, |r(f_{k+t})|$$

$$= \sum_{k=0}^{\infty} \sum_{t=0}^{\infty} \binom{n+t-1}{t} P(f_0) \cdots P(f_{k+t-1}) \, |r(f_{k+t})|$$

$$\underset{k+t=s}{=} \sum_{s=0}^{\infty} \sum_{t=0}^{s} \binom{n+t-1}{t} P(f_0) \cdots P(f_{s-1}) \, |r(f_s)|$$

$$= \sum_{s=0}^{\infty} \binom{n+s}{s} P(f_0) \cdots P(f_{s-1}) \, |r(f_s)|$$

$$= \mathbb{E}_{\pi} \sum_{k=0}^{\infty} \binom{n+1+k-1}{k} |r(X_k, A_k)| \, .$$

Here we used

$$\sum_{t=0}^{s} \binom{n+t-1}{t} = \binom{n+s}{s} \, .$$

Taking the supremum with respect to $\pi \in M$ we obtain (4.15) for $m = n+1$. Hence (4.15) holds for all $m = 1,2,\ldots$. □

So the existence of a system of Liapunov functions of order m implies that z_{φ}^{*} is finite for the sequence $\varphi = (\varphi_0, \varphi_1, \ldots)$ with

$$\varphi_k = \binom{m+k-1}{k} e \, , \quad k = 0,1,\ldots \, .$$

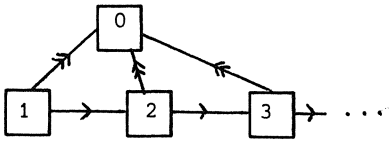
Specifically, if there exists a system of Liapunov functions of order 2, then $y_2 < \infty$ implies by (4.15) that the MDP is strongly convergent.

One might ask whether y_m is finite if and only if

$$\text{(4.16)} \qquad \sup_{\pi \in M} \mathbb{E}_{\pi} \sum_{k=0}^{\infty} \binom{m+k-1}{k} |r(X_k, A_k)| < \infty \, .$$

The following example shows that this is not the case.

EXAMPLE 4.19. $S = \{0,1,2,\ldots\}$, $A = \{1,2\}$. State 0 is absorbing with $r(0,a) = 0$, $a = 1,2$. In state $i \geq 1$, we have $r(i,1) = 0$, $p(i,1,i+1) = 1$, $r(i,2) = i^{-1}$, $p(i,2,0) = 1$. So with $\varphi_n = n+1$ (the case $m = 2$) we have for



all $\pi \in M$

$$\mathbb{E}_\pi \sum_{n=0}^{\infty} (n+1)\left|r(X_n,A_n)\right| \le e \ .$$

However,

$$y_1(i) = i^{-1} \ , \quad i = 1,2,\ldots \ ,$$

so for strategy f with f(i) = 1 for all $i \in S$, we have

$$\mathbb{E}_{i,f} \sum_{n=0}^{\infty} y_1(X_n) = i^{-1} + (i+1)^{-1} + \ldots = \infty \ .$$

Hence $z_\varphi^* < \infty$ for the sequence $\varphi = (\varphi_0,\varphi_1,\ldots)$ with $\varphi_n = (n+1)e$ does not imply $y_2 < \infty$.

A slightly stronger condition than (4.16), however, implies that y_m is finite.

THEOREM 4.20. *Let $\varphi \in \Phi$ be a sequence with $\varphi_n = b_n^{m-1} e$, $n = 0,1,\ldots$ (so $b_0 \ge 1$ and $b_{n+1} \ge b_n$), satisfying*

$$b := \sum_{n=0}^{\infty} b_n^{-1} < \infty \quad \textit{and} \quad z_\varphi^* < \infty \ ,$$

then the functions $y_1,\ldots,y_m$ are finite

PROOF. By induction on k it will be shown that

$$(4.17) \qquad \sup_{\pi\in M} \mathbb{E}_\pi y_k(X_n) \le z_\varphi^* b^{k-1} b_n^{k-m} \ , \quad k = 1,2,\ldots,m-1 \ , \ n = 0,1,\ldots \ .$$

Once we have (4.17) for k = m - 1 and for all n = 0,1,..., we immediately obtain

$$(4.18) \qquad y_m = \sup_{\pi\in M} \mathbb{E}_\pi \sum_{n=0}^{\infty} y_{m-1}(X_n) \le \sum_{n=0}^{\infty} z_\varphi^* b^{m-2} b_n^{-1} = z_\varphi^* b^{m-1} < \infty \ .$$

So indeed, it is sufficient to prove (4.17).
First consider the case k = 1. As $y_1 = z^*$, we have by lemmas 4.5 and 4.3(ii)

$$\sup_{\pi\in M} \mathbb{E}_\pi y_1(X_n) = \sup_{\pi\in M} \mathbb{E}_\pi \sum_{t=n}^{\infty} \left|r(X_t,A_t)\right| \le z_\varphi^* b_n^{1-m} \ , \quad n = 0,1,\ldots \ .$$

So, (4.17) holds for k = 1 and all n = 0,1,... . Now assume that (4.17) holds for $k = k_0 < m-1$ and all n = 0,1,... . Then

$$\sup_{\pi\in M} \mathbb{E}_\pi y_{k_0+1}(X_n) = \sup_{\pi\in M} \mathbb{E}_\pi \sum_{t=n}^{\infty} y_{k_0}(X_t)$$

$$\leq \sum_{t=n}^{\infty} \sup_{\pi\in M} \mathbb{E}_\pi y_{k_0}(X_t) \leq \sum_{t=n}^{\infty} z^*_\varphi b_t^{k_0-1} b^{k_0-m}$$

$$\leq z^*_\varphi b^{k_0-1} \sum_{t=n}^{\infty} b_t^{-1} b_n^{k_0-m+1} \leq z^*_\varphi b^{k_0} b_n^{k_0+1-m} .$$

Hence, (4.17) also holds for $k = k_0 + 1$ and thus by induction for k = m - 1 and all n = 0,1,... . Then (4.18) completes the proof. □

4.6. THE CONVERGENCE OF $U_\delta^n v$ TO v^*

In this section we consider the set of algorithms introduced in chapter 3, sections 3-5, by means of go-ahead functions.
The main result of this section is, that if the MDP is strongly convergent, we have

$$\lim_{n\to\infty} U_\delta^n v = v^* \quad \text{for all } v \in V_{z^*} ,$$

for any nonzero go-ahead function δ.

Define the operators $\tilde{L}_\delta(\pi)$ and $\tilde{U}_\delta$ on V_{z^*} analogously to $\tilde{L}(f)$ and $\tilde{U}$ by

$$\tilde{L}_\delta(\pi) v = \mathbb{E}_\pi^\delta v(X_\tau)$$

$$\tilde{U}_\delta v = \sup_{\pi\in\Pi} \tilde{L}_\delta(\pi) v , \quad v \in V_{z^*} .$$

Then for all $v \in V_{z^*}$,

$$|U_\delta v - v^*| = |U_\delta v - U_\delta v^*| \leq \tilde{U}_\delta |v - v^*| , \tag{4.19}$$

since for all $v,w \in V_{z^*}$

$$U_\delta v - U_\delta w = \sup_{\pi\in\Pi} [L_\delta(\pi) v - U_\delta w] \leq$$

$$\leq \sup_{\pi\in\Pi} [L_\delta(\pi)v - L_\delta(\pi)w] = \sup_{\pi\in\Pi} \tilde{L}_\delta(\pi)(v-w) \leq \tilde{U}_\delta|v-w| \ .$$

Iterating (4.19) yields for all n = 1,2,...

$$|U_\delta^n v - v^*| \leq \tilde{U}_\delta^n |v - v^*| \ .$$

So, we see that if $v \in V_{z^*}$ (thus also $|v-v^*| \in V_{z^*}$), then a sufficient condition for the convergence of $U_\delta^n v$ to v^* is the convergence of $\tilde{U}_\delta^n z^*$ to zero.
In order to prove that if the MDP is strongly convergent, then $\tilde{U}_\delta^n z^*$ converges to zero for any nonzero go-ahead function δ, we first have to derive two lemmas.

LEMMA 4.21. *If* $v \geq 0$ *and* $\tilde{U}v \leq v$, *then we have*

(4.20) $\quad \tilde{U}_\delta v \leq v$

for all go-ahead functions δ.

PROOF. In order to prove this lemma we construct an optimal stopping problem which has value v from which we will conclude that $\tilde{U}_\delta v \leq v$ for all δ.
Define the MDP characterized by $(\hat{S},\hat{A},\hat{p},\hat{r})$, with

$$\hat{S} := S \cup \{*\} \quad (* \notin S) \ , \quad \hat{A} := A \cup \{+\} \quad (+ \notin A) \ ,$$

$$\hat{r}(i,a) = 0 \ , \quad \hat{p}(i,a,j) = p(i,a,j) \ , \quad a \in A \ , \quad i,j \in S \ ,$$

$$\hat{r}(i,+) = v(i) \ , \quad \hat{r}(*,\hat{a}) = 0 \ , \quad \hat{p}(i,+,*) = \hat{p}(*,\hat{a},*) = 1 \ , \quad \hat{a} \in \hat{A} \ .$$

In this newly defined MDP the action denoted by the character + corresponds to stopping and transfers the system to the absorbing state *. Denote all objects in this MDP by a $\hat{}$, and define $\hat{v}$ by $\hat{v}(i) = v(i)$, $i \in S$, and $\hat{v}(*) = 0$. Then clearly

$$\hat{L}(f)\hat{v} \leq \hat{v} \quad \text{for all } f \in \hat{F} \quad \text{and} \quad \hat{L}(f^+)\hat{v} = \hat{v} \ ,$$

where f^+ is a policy with $f^+(i) = +$ for all $i \in S$.
Hence, with theorem 2.12

$$\hat{v}^* = \hat{v} \ .$$

Since any $\pi \in \Pi$ (extended with the behaviour in *) is a strategy in the stopping problem (without ever stopping) we have (by theorem 3.15)

$$\hat{L}_\delta(\pi)v \leq v \ .$$

Further, we have on S

$$\hat{L}_\delta(\pi)v = \tilde{L}_\delta(\pi)v \ .$$

Hence

$$\tilde{U}_\delta v = \sup_{\pi\in\Pi} \tilde{L}_\delta(\pi)v \leq v \ . \qquad \square$$

Note that lemma 4.21 not only holds for nonzero go-ahead functions. For nonzero go-ahead functions a somewhat stronger result then (4.20) holds.

LEMMA 4.22. *If* $v \geq 0$ *and* $\tilde{U}v \leq v$, *then we have*

$$\tilde{U}_\delta v \leq (1-\alpha_\delta)v + \alpha_\delta \tilde{U}v \tag{4.21}$$

for any go-ahead function δ. *(For the definition of* α_δ *see definition 3.16.)*

PROOF. Intuitively (4.21) follows straightforwardly from the optimal stopping problem constructed in (4.20). Namely from $\tilde{U}v \leq v$ it follows that the sooner you stop the better. Since α_δ is a lower bound for the probability that you stop after time 0, $(1-\alpha_\delta)v + \alpha_\delta\tilde{U}v$ is an upper bound on $\tilde{U}_\delta v$. Formally (4.21) can be proved as follows by conditioning on X_0, A_0 and X_1. Define for all $i \in S$ and $a \in A$ by $\delta^{(i,a)}$ the go-ahead function with for all n, $i_0,a_0,i_1,\ldots$

$$\delta^{(i,a)}(i_0,a_0,\ldots,i_n) = \delta(i,a,i_0,a_0,\ldots,i_n)$$

and

$$\delta^{(i,a)}(i_0,\ldots,a_n) = \delta(i,a,i_0,\ldots,a_n) \ .$$

And define by $\pi^{(i,a)}$ the strategy with for all n, all $i_0,a_0,i_1,\ldots$ and all $C \in A$

$$\pi_n^{(i,a)}(C \mid i_0,\ldots,i_t) = \pi_{n+1}^{(i,a)}(i,a,i_0,\ldots,i_t) \ .$$

Then we have for all $i \in S$ and $\pi \in \Pi$

$$(\tilde{L}_\delta(\pi)v)(i) = \int_A \pi_0(da \mid i)\{[1-\delta(i)\delta(i,a)]v(i) + \delta(i)\delta(i,a)\sum_j p(i,a,j)\cdot \cdot(\tilde{L}_{\delta(i,a)}(\pi^{(i,a)})v)(j)\} \ . \tag{4.22}$$

By lemma 4.21

$$(4.23)\qquad \tilde{L}_{\delta(i,a)}(\pi^{(i,a)})v \le v \quad \text{for all } i \in S,\ a \in A .$$

Further,

$$(4.24)\qquad \sum_j p(i,a,j)v(j) \le (\tilde{U}v)(i) \quad \text{for all } i \in S,\ a \in A ,$$

and

$$\begin{aligned}(4.25)\qquad & [1-\delta(i)\delta(i,a)]v(i) + \delta(i)\delta(i,a)(\tilde{U}v)(i)\\ &= (1-\alpha_\delta)v(i) + \alpha_\delta(\tilde{U}v)(i) + [\delta(i)\delta(i,a)-\alpha_\delta](\tilde{U}v-v)(i)\\ &\le (1-\alpha_\delta)v(i) + \alpha_\delta(\tilde{U}v)(i) .\end{aligned}$$

Substituting subsequently (4.23), (4.24) and (4.25) into (4.22) yields

$$\begin{aligned}(\tilde{L}_\delta(\pi)v)(i) &\le \int_A \pi_0(da \mid i)[(1-\alpha_\delta)v(i) + \alpha_\delta(\tilde{U}v)(i)]\\ &= (1-\alpha_\delta)v(i) + \alpha_\delta(\tilde{U}v)(i) .\end{aligned}$$

Taking the supremum with respect to $\pi \in \Pi$ finally yields (4.21). □

Now we can prove our main result.

THEOREM 4.23. *If the MDP is strongly convergent, then*

$$\lim_{n\to\infty} U_\delta^n v = v^* \quad \textit{for all } v \in V_{z^*}$$

for any nonzero go-ahead function δ.

PROOF. As has been remarked before, it is sufficient to prove

$$\tilde{U}_\delta^n z^* \to 0 \quad (n \to \infty) .$$

Clearly for all $m = 0,1,\ldots$

$$\tilde{U}^m z^* \ge 0 .$$

Further it follows from $\tilde{U}z^* \le z^*$ with the monotonicity of $\tilde{U}$ that

$$\tilde{U}(\tilde{U}^m z^*) \le \tilde{U}^m z^* .$$

So we can apply lemma 4.22 to obtain

$$\tilde{U}_\delta(\tilde{U}^m z^*) \le (1-\alpha_\delta)\tilde{U}^m z^* + \alpha_\delta \tilde{U}^{m+1} z^* \ , \quad m = 0,1,\dots \ .$$

Hence,

$$\tilde{U}_\delta^n z^* \le \tilde{U}_\delta^{n-1}[(1-\alpha_\delta)z^* + \alpha_\delta \tilde{U} z^*] \le (1-\alpha_\delta)\tilde{U}_\delta^{n-1} z^* + \alpha_\delta \tilde{U}_\delta^{n-1}\tilde{U} z^*$$

$$\le \dots \le \sum_{k=0}^{n} \binom{n}{k} (1-\alpha_\delta)^{n-k} \alpha_\delta^k \tilde{U}^k z^* \ .$$

Since $\alpha_\delta > 0$ for a nonzero go-ahead function δ and since $\tilde{U}^n z^* \to 0$ $(n \to \infty)$, we also have $\tilde{U}_\delta^n z^* \to 0$ $(n \to \infty)$. □

For the case of nonrandomized go-ahead functions theorem 4.23 (with $v = 0$) has already been given by Van HEE [1978b]. Recall that this is the case in which there are no measurability problems when fitting two strategies together at time τ.

In general the method of successive approximations need not converge for a nonzero go-ahead function δ.
The following lemma states that, if in the optimization of $v(\pi)$ one needs to consider only stationary strategies, then

$$\liminf_{n\to\infty} U_\delta^n 0 \ge v^* \ .$$

This will enable us to show that in two special cases:
(i) the positive dynamic programming case;
(ii) the case that A is finite and $v^* \ge 0$,
we have

$$\lim_{n\to\infty} U_\delta^n 0 = v^* \ ,$$

for any nonzero go-ahead function δ.

LEMMA 4.24. *If for initial state* $i \in S$ *we have*

$$\sup_{f\in F} v(i,f) = v^*(i) \ ,$$

then for any nonzero go-ahead function δ

$$\liminf_{n\to\infty} (U_\delta^n 0)(i) \ge v^*(i) \ .$$

PROOF. It is sufficient to prove

$$\lim_{n\to\infty} (L_\delta^n(f)0)(i) = v(i,f) \quad \text{for all } f \in F \text{ with } v(i,f) > -\infty ,$$

since this implies

$$\operatorname{liminf}_{n\to\infty} (U_\delta^n 0)(i) \geq \sup_{f\in F} \lim_{n\to\infty} (L_\delta^n(f)0)(i) = \sup_{f\in F} v(i,f) = v^*(i) .$$

Assume $v^*(i) > -\infty$, otherwise the result is trivial. And let f be any policy with $v(i,f) > -\infty$. Now consider the MDP in which in each state $j \in S$ all actions except f(j) are eliminated, i.e. the MDP $(S,\hat{A},\hat{p},\hat{r})$ with for all $i,j \in S$

$$\hat{A} = \{1\} , \quad \hat{r}(i,1) = r(i,f(i)) \quad \text{and} \quad \hat{p}(i,1,j) = p(i,f(i),j) .$$

If $v(f) > -\infty$, then clearly this MDP is strongly convergent. So for this problem the method of successive approximations with scrapvalue 0 converges for any nonzero go-ahead function δ. Thus

$$L_\delta^n(f)0 \to v(f) \quad (n \to \infty) .$$

If $v(j,f) = -\infty$ for some j, then we can restrict S to the set of states k for which $v(k,f) > -\infty$, since $v(j,f) = -\infty$ implies

$$\mathbb{P}_{i,f}(X_n = j) = 0 \quad \text{for all } n ,$$

and follow the same reasoning.
Hence

$$(L_\delta^n(f)0)(i) \to v(i,f) \quad (n \to \infty)$$

for all f with $v(i,f) > -\infty$, and the proof is complete. □

THEOREM 4.25. *For each of the following two conditions we have*

$$\lim_{n\to\infty} U_\delta^n v = v^* \quad \textit{for all } v \textit{ with } 0 \leq v \leq v^* ,$$

for any nonzero go-ahead function δ:
(i) $r(i,a) \geq 0$ *for all* $i \in S$, $a \in A$;
(ii) A *is finite and* $v^* \geq 0$.

PROOF. By the monotonicity of U_δ we have for each of the two conditions

$$U_\delta^n 0 \le U_\delta^n v \le U_\delta^n v^* = v^* ,$$

for all v with $0 \le v \le v^*$ and all $n = 0,1,\ldots$.
So it is sufficient to show that

$$\liminf_{n\to\infty} U_\delta^n 0 \ge v^* .$$

Hence (i) follows immediately from theorem 2.23 and lemma 4.24 and (ii) follows from theorem 2.22 and lemma 4.24. □

We conjecture that for all nonzero δ we have

$$\lim_{n\to\infty} U_\delta^n v = v^* \quad \text{for all } v \text{ with } w^* \le v \le v^* ,$$

compare theorem 3.7.

4.7. STATIONARY GO-AHEAD FUNCTIONS AND STRONG CONVERGENCE

In section 3.5 the question has been raised whether

$$\sup_{f\in F} (L_\delta(f)v)(i) = (U_\delta v)(i) \quad \text{for all } i \in S , \tag{4.26}$$

if δ is a stationary go-ahead function.
And it has been shown that if, for the transformed MDP $(\hat{S},\hat{A},\hat{p},\hat{r})$ defined in (3.9), we have

$$\sup_{f\in F} \hat{v}(i,f) = \hat{v}^*(i) \quad \text{for all } i \in S , \tag{4.27}$$

then (4.26) holds (see also theorem 3.20).
In this section the following result will be shown.

THEOREM 4.26. *If the MDP is strongly convergent and δ is a stationary go-ahead function, then*

$$\sup_{f\in F} (L_\delta(f)v)(i) = (U_\delta v)(i) \quad \textit{for all } i \in S \textit{ and all } v \in V_{z^*} .$$

PROOF. We will show that the MDP defined in (3.9) is strongly convergent. By theorem 4.9 this implies that (4.27) holds, which - as has been argued in section 3.5 - proves the theorem. From (3.9) it follows that it suffices to consider the case $v = z^*$.

We use the indicator function notation on Ω_0 (= $(S \times E \times A \times E)^\infty$). I.e., for any subset B of Ω_0 we consider the function I_B on Ω_0, defined by $I_B(\omega) = 1$ if $\omega \in B$ and 0 elsewhere, $\omega \in \Omega_0$. All objects concerning the transformed MDP will be marked by a hat. Let $\hat{\pi}$ be an arbitrary strategy in $\hat{M}$ and π the corresponding strategy in M, then we have for the case $v = z^*$,

$$\begin{aligned}(4.28)\qquad \hat{\mathbb{E}}_{\hat{\pi}} |\hat{r}(X_k,A_k)| &= \mathbb{E}_\pi^\delta I_{\{\tau \geq k\}}((X_0,Y_0,A_0,\ldots))\,|(1-\delta(X_k,A_k))z^*(X_k) + \\ &\quad + \delta(X_k,A_k)[r(X_k,A_k) + \sum_j p(X_k,A_k,j)(1-\delta(X_k,A_k,j)z^*(j)]| \\ &\leq \mathbb{E}_\pi^\delta I_{\{\tau \geq k\}}((X_0,Y_0,A_0,\ldots))[(1-\delta(X_k,A_k))z^*(X_k) + \delta(X_k,A_k)\cdot \\ &\quad \cdot [|r(X_k,A_k)| + \sum_j p(X_k,A_k,j)z^*(j)] + \\ &\quad - \delta(X_k,A_k)\sum_j p(X_k,A_k,j)\delta(X_k,A_k,j)z^*(j)] .\end{aligned}$$

Further, since $U^{abs}z^* \leq z^*$,

$$(4.29)\qquad |r(i,a)| + \sum_j p(i,a,j)z^*(j) \leq z^*(i) \quad \text{for all } i \in S \text{ and } a \in A ,$$

and

$$\begin{aligned}(4.30)\qquad &\mathbb{E}_\pi^\delta I_{\{\tau \geq k\}}((X_0,Y_0,A_0,\ldots))\delta(X_k,A_k)\sum_j p(X_k,A_k,j)\delta(X_k,A_k,j)z^*(j) \\ &= \mathbb{E}_\pi^\delta I_{\{\tau \geq k+1\}}((X_0,Y_0,A_0,\ldots))z^*(X_{k+1}) .\end{aligned}$$

Substituting (4.29) and (4.30) in (4.28) yields

$$\begin{aligned}\hat{\mathbb{E}}_{\hat{\pi}} |\hat{r}(X_k,A_k)| &\leq \mathbb{E}_\pi^\delta I_{\{\tau \geq k\}}((X_0,Y_0,A_0,\ldots))z^*(X_k) + \\ &\quad - \mathbb{E}_\pi^\delta I_{\{\tau \geq k+1\}}((X_0,Y_0,A_0,\ldots))z^*(X_{k+1}) .\end{aligned}$$

Hence

$$\begin{aligned}(4.31)\qquad \hat{\mathbb{E}}_{\hat{\pi}} \sum_{k=n}^{\infty} |\hat{r}(X_k,A_k)| &\leq \mathbb{E}_\pi^\delta I_{\{\tau \geq n\}}((X_0,Y_0,A_0,\ldots))z^*(X_n) + \\ &\quad - \liminf_{k\to\infty} \mathbb{E}_\pi^\delta I_{\{\tau \geq k\}}((X_0,Y_0,A_0,\ldots))z^*(X_k) .\end{aligned}$$

Also

$$\begin{aligned}(4.32)\qquad &\mathbb{E}_\pi^\delta I_{\{\tau \geq k\}}((X_0,Y_0,A_0,\ldots))z^*(X_k) \leq \mathbb{E}_\pi^\delta z^*(X_k) \\ &= \mathbb{E}_\pi z^*(X_k) \leq \tilde{U}^k z^* \to 0 \quad (k \to \infty) \quad \text{(theorem 4.6)}.\end{aligned}$$

Finally, substitution of (4.32) in (4.31) yields

$$\sup_{\hat{\pi}\in\hat{M}} \hat{E}_{\hat{\pi}} \sum_{k=n}^{\infty} |\hat{r}(X_k,A_k)| \leq \tilde{U}^n z^* \ .$$

So condition 4.1 holds and the transformed MDP is strongly convergent by theorem 4.4, which completes the proof. □

4.8. VALUE-ORIENTED SUCCESSIVE APPROXIMATIONS

In this section the convergence is studied of the method of value-oriented successive approximations, which was introduced in section 3.6, for the strongly convergent MDP. It will be shown that the monotone value-oriented method converges for all $v_0 \in V_{z^*}$ for which there exists a policy f satisfying $L(f)v_0 \geq v_0$.
Further, two conditions will be given, each of which guarantees that the nonmonotonic version converges.

THEOREM 4.27. *If the MDP is strongly convergent, then the monotone value-oriented method defined in section 3.6 converges for all* $v_0 \in V_{z^*}$ *for which there exists an* $f \in F$ *satisfying* $L(f)v_0 \geq v_0$.

PROOF. Note that if the MDP is strongly convergent, then the policy iteration method of section 4.4 is just the monotone value-oriented method with "$\lambda = \infty$". Namely, for the strongly convergent MDP we have

$$\lim_{\lambda\to\infty} L^{\lambda}(f)v = v(f) \quad \text{for all } f \in F \text{ and all } v \in V_{z^*} \ .$$

To prove the theorem we follow the line of reasoning of section 4.4.
As remarked in section 3.6

$$\lim_{n\to\infty} v_n \quad \text{exists.}$$

Further,

$$v_n = L^{\lambda}(f_n)\cdots L^{\lambda}(f_1)v_0 = L^{\lambda}(f_n)\cdots L^{\lambda}(f_1)v^* + P^{\lambda}(f_n)\cdots P^{\lambda}(f_1)(v_0 - v^*)$$

$$\leq v^* + \tilde{U}^{n\lambda}(v_0 - v^*) \ .$$

Hence, with theorem 4.6,

$$\hat{v} := \lim_{n\to\infty} v_n \leq v^* \ .$$

Also $\hat{v} \in V_{z^*}$ and, by lemma 4.15, $U\hat{v} \leq \hat{v}$, from which one proves (as in theorem 4.16) that $\hat{v} = v^*$. □

Now let us consider the nonmonotonic value-oriented method. Let $\{f_n\}$ and $\{v_n\}$ be sequences of policies and value functions obtained from the method of value-oriented standard successive approximations. So

$$L(f_{n+1})v_n \geq Uv_n - d_n \ , \quad n = 0,1,... \tag{4.33}$$

where $\{d_n\}$ is the sequence of strictly positive real-valued functions on S with $d_n \to 0$ $(n \to \infty)$.
And

$$v_{n+1} = L^{\lambda}(f_{n+1})v_n \ , \quad n = 0,1,... \ . \tag{4.34}$$

In order to investigate whether v_n converges to v^*, we follow the line of reasoning in the proofs by Van NUNEN [1976a] and ROTHBLUM [1979].
Clearly

$$\limsup_{n\to\infty} v_n \leq \lim_{n\to\infty} U^{n\lambda}v_0 = v^* \ .$$

Further,

$$\begin{aligned} U^k v_n = U^{k-1}Uv_n &\leq U^{k-1}v_{n+1} + \tilde{U}^{k-1}(Uv_n - v_{n+1}) \\ &\leq U^{k-2}v_{n+2} + \tilde{U}^{k-2}(Uv_{n+1} - v_{n+2}) + \tilde{U}^{k-1}(Uv_n - v_{n+1}) \leq \ldots \\ &\leq v_{n+k} + \sum_{m=0}^{k-1} \tilde{U}^{k-m-1}(Uv_{n+m} - v_{n+m+1}) \ , \end{aligned}$$

as follows from

$$Uv \leq Uw + \tilde{U}(v-w) \quad \text{for all } v,w \in V_{z^*} \ .$$

So, since

$$\lim_{n\to\infty} U^k v_n = v^* \ ,$$

it is sufficient for the convergence of v_n to v^* that

$$(4.35)\qquad \limsup_{n\to\infty}\ \limsup_{k\to\infty}\ \sum_{m=0}^{k-1} \tilde{U}^{k-m-1}(Uv_{n+m} - v_{n+m+1}) \le 0 \ .$$

THEOREM 4.28. *Each of the following two conditions guarantees that the method of value-oriented successive approximations converges:*

(i) $\sum_{n=0}^{\infty} \tilde{U}^n z^* < \infty$ *and* $d_n \le \varepsilon_n z^*$, *with* $\varepsilon_{n+1} \le \varepsilon_n$, $n = 0,1,\ldots$ *and* $\sum_{n=0}^{\infty} \varepsilon_n < \infty$.

(ii) $\sum_{n=0}^{\infty} (n+1)\tilde{U}^n z^* < \infty$ *and* $d_n \le \varepsilon_n z^*$, *with* $\varepsilon_{n+1} \le \varepsilon_n$, $n = 0,1,\ldots$ *and* $\varepsilon_n \to 0$ $(n \to \infty)$.

PROOF. We will show that each of the conditions (i) and (ii) implies (4.35), which, as has been argued before, is sufficient. Therefore we first derive some inequalities for $Uv_{n+m} - v_{n+m+1}$.
From (4.33) and (4.34) follows for all $n = 0,1,\ldots$

$$(4.36)\qquad v_{n+1} = L^{\lambda}(f_{n+1})v_n = \sum_{k=1}^{\lambda-1} (L^{k+1}(f_{n+1})v_n - L^k(f_{n+1})v_n) + L(f_{n+1})v_n$$

$$= \sum_{k=0}^{\lambda-2} P^k(f_{n+1})(L(f_{n+1})v_n - v_n) + L(f_{n+1})v_n$$

$$\ge \sum_{k=0}^{\lambda-2} P^k(f_{n+1})(Uv_n - v_n - d_n) + Uv_n - d_n \ .$$

And

$$(4.37)\qquad Uv_{n+1} - v_{n+1} \ge L(f_{n+1})v_{n+1} - v_{n+1} = P^{\lambda}(f_{n+1})(L(f)v_n - v_n)$$

$$\ge P^{\lambda}(f_{n+1})(Uv_n - v_n - d_n) \ .$$

Repeated application of (4.37) yields

$$(4.38)\qquad Uv_{n+m} - v_{n+m} \ge -[P^{\lambda}(f_{n+m})d_{n+m-1} + P^{\lambda}(f_{n+m})P^{\lambda}(f_{n+m-1})d_{n+m-2} +$$

$$+ \ldots + P^{\lambda}(f_{n+m}) \cdots P^{\lambda}(f_{m+1})d_m] + P^{\lambda}(f_{n+m}) \cdots P^{\lambda}(f_{m+1})(Uv_m - v_m) \ .$$

And from (4.36) and (4.38) we obtain

$$(4.39)\qquad Uv_{n+m} - v_{n+m+1} \le d_{n+m} - \sum_{k=0}^{\lambda-2} P^k(f_{n+m+1})(Uv_{n+m} - v_{n+m} - d_{n+m})$$

$$\le d_{n+m} + \sum_{k=0}^{\lambda-2} P^k(f_{n+m+1})[d_{n+m} + P^\lambda(f_{n+m})d_{n+m-1} + \ldots +$$

$$+ P^\lambda(f_{n+m}) \cdots P^\lambda(f_{m+1})d_m + P^\lambda(f_{n+m}) \cdots P^\lambda(f_{m+1})(v_m - Uv_m)] .$$

Let $K \ge 2$ be such that $|v_0| \le \frac{1}{2}Kz^*$, then $|v_k - Uv_k| \le Kz^*$ for all $k = 0,1,\ldots$ (since $v_0 \in V_{z^*}$ such a K exists). Then substitution in (4.39) of

$$v_m - Uv_m \le Kz^* ,$$

$$d_k \le \varepsilon_k z^* , \quad k = 0,1,\ldots ,$$

$$P(f)v \le \tilde{U}v \quad \text{for all } v \in V_{z^*} \text{ and } f \in F$$

$$\varepsilon_{m+k} \le \varepsilon_m \quad \text{for all } m,k = 0,1,\ldots$$

and

$$\tilde{U}(v+w) \le \tilde{U}v + \tilde{U}w \quad \text{for all } v,w \in V_{z^*}$$

gives

$$Uv_{n+m} - v_{n+m+1} \le \varepsilon_{n+m}z^* + \sum_{k=0}^{\lambda-2} \tilde{U}^k[\varepsilon_{n+m}z^* + \tilde{U}^\lambda\varepsilon_{n+m-1}z^* + \ldots + \tilde{U}^{n\lambda}\varepsilon_m z^* + \tilde{U}^{n\lambda}Kz^*]$$

$$\le \varepsilon_{n+m}z^* + \sum_{k=0}^{\lambda-2} \tilde{U}^k\varepsilon_{n+m}z^* + \sum_{k=0}^{\lambda-2} \tilde{U}^k\tilde{U}^\lambda\varepsilon_{n+m-1}z^* + \ldots + \sum_{k=0}^{\lambda-2} \tilde{U}^k\tilde{U}^{n\lambda}(\varepsilon_m + K)z^*$$

$$\le \varepsilon_n z^* + \sum_{k=0}^{\lambda-2} \tilde{U}^k\varepsilon_n z^* + \sum_{k=0}^{\lambda-2} \tilde{U}^{k+\lambda}\varepsilon_{n-1}z^* + \ldots + \sum_{k=0}^{\lambda-2} \tilde{U}^{k+n\lambda}(\varepsilon_0 + K)z^* .$$

With $\tilde{U}^k z^* \le z^*$ for all k this simplifies to

$$Uv_{n+m} - v_{n+m-1} \le \lambda[\varepsilon_n z^* + \tilde{U}^\lambda\varepsilon_{n-1}z^* + \ldots + \tilde{U}^{(n-1)\lambda}\varepsilon_1 z^* + \tilde{U}^{n\lambda}(\varepsilon_0 + K)z^*] .$$

So,

$$\sum_{m=0}^{k-1} \tilde{U}^{k-m-1}(Uv_{n+m} - v_{n+m-1}) \le \lambda \sum_{m=0}^{k-1} \tilde{U}^{k-m-1}[\varepsilon_n z^* + \tilde{U}^\lambda\varepsilon_{n-1}z^* + \ldots + \tilde{U}^{n\lambda}(\varepsilon_0 + K)z^*]$$

$$\le \lambda\left[\sum_{m=0}^{\infty} \tilde{U}^m\varepsilon_n z^* + \sum_{m=0}^{\infty} \tilde{U}^m\tilde{U}^\lambda\varepsilon_{n-1}z^* + \ldots + \sum_{m=0}^{\infty} \tilde{U}^m\tilde{U}^{n\lambda}(\varepsilon_0 + K)z^*\right] .$$

Or, with

$$c_k := \sum_{m=k}^{\infty} \tilde{U}^m z^* ,$$

$$\sum_{m=0}^{k-1} \tilde{U}^{k-m-1}(Uv_{n+m} - v_{n+m+1}) \le \lambda[c_0\varepsilon_n + c_1\varepsilon_{n-1} + \ldots + c_n\varepsilon_0 + c_nK] . \tag{4.40}$$

Finally, we have to show that each of the conditions (i) and (ii) guarantees that the right-hand side in (4.40) tends to zero if n tends to infinity.

(i) Fix some state $i \in S$ and some $\varepsilon > 0$. From

$$\sum_{n=0}^{\infty} \tilde{U}^n z^* < \infty$$

it follows that $c_n(i) \to 0$ $(n \to \infty)$. So, we can choose integers k_0 and $n_0 \ge k_0$ such that

$$c_{k_0+1}(i) \sum_{n=0}^{\infty} \varepsilon_n \le \frac{\varepsilon}{3} ,$$

and

$$Kc_{n_0}(i) \le \frac{\varepsilon}{3} \quad \text{and} \quad \varepsilon_{n_0-k_0} \sum_{k=0}^{k_0} c_k(i) \le \frac{\varepsilon}{3} .$$

Then for $n > n_0$ with $\varepsilon_{k+1} \le \varepsilon_k$ and $c_{k+1} \le c_k$ for all $k = 0,1,\ldots$

$$\sum_{k=0}^{n} c_k(i)\varepsilon_{n-k} + c_n(i)K = \sum_{k=0}^{k_0} c_k(i)\varepsilon_{n-k} + \sum_{k=k_0+1}^{n} c_k(i)\varepsilon_{n-k} + c_n(i)K$$

$$\le \varepsilon_{n-k_0} \sum_{k=0}^{k_0} c_k(i) + c_{k_0+1}(i) \sum_{k=0}^{\infty} \varepsilon_k + c_n(i)K \le \varepsilon .$$

So, since i and $\varepsilon > 0$ can be chosen arbitrarily, the right hand side in (4.40) tends to zero if n tends to infinity and thus (4.35) holds. Hence $v_n \to v^*$ $(n \to \infty)$.

(ii) $$\sum_{k=0}^{\infty} c_k \le \sum_{k=0}^{\infty} \sum_{n=k}^{\infty} \tilde{U}^n z^* = \sum_{n=0}^{\infty} \sum_{k=0}^{n} \tilde{U}^n z^* = \sum_{n=0}^{\infty} (n+1)\tilde{U}^n z^* < \infty .$$

So the proof can be given in exactly the same way as in (i) with the roles of ε_k and $c_k(i)$ reversed. □

CHAPTER 5

THE CONTRACTING MDP

5.1. *INTRODUCTION*

The most intensively studied MDP's, at least with respect to computational procedures, are the contracting models. Of these the most common model is the discounted MDP with bounded reward structure (cf. SHAPLEY [1953], HOWARD [1960] and BLACKWELL [1962,1965]. In the case of a countable state space the reward structure is typically unbounded. For example, in inventory control models and queueing models part of the costs will tend to infinity if the stock or the number of customers in the queue increases.
In order to be able to deal with unbounded rewards, we assume the existence of a nonnegative function μ on S, called a *bounding function*. It is assumed that all rewards are bounded with respect to this function, i.e., $r(f) \in V_\mu$ for all $f \in F$ (or bounded from above: $r(f) \in V_\mu^+$). Further it is assumed that the transition matrices are contractions with respect to the μ-norm: $P(f)\mu \leq \rho\mu$ for some $\rho < 1$ and all $f \in F$.
The use of bounding functions in this way has been introduced by WESSELS [1977b]. Bounding functions used as strongly excessive functions also appear in VEINOTT [1969], in a lemma due to Hoffman, and in WIJNGAARD [1975].
Another idea for coping with the unbounded reward structure has been introduced by HARRISON [1972]. He considers the discounted MDP and assumes (in essence) the existence of a function b such that $r(f) - b$ is bounded for all $f \in F$ and that for some $\rho < 1$ also $P(f)b - \rho b$ is bounded.
These two ideas are combined in the contracting MDP model of Van NUNEN [1976a] and the slightly extended model of Van NUNEN and WESSELS [1977a].
In this chapter we first consider four different models for a contracting MDP. It will be shown that these models are equivalent with respect to the important features in the ∞-horizon problem (section 2). Next we relate the contraction model to the strongly convergent MDP of chapter 4 (section 3).

In section 4 some results are reviewed with respect to bounds on v^* and nearly-optimal stationary strategies for the successive approximation algorithms generated by nonzero go-ahead functions. In chapter 11 very similar results will be obtained for the contracting Markov game.
The discounted MDP with finite state and action spaces is studied in some detail in section 5. We derive Laurent series expansions for $v_\beta(f)$ and $r(h) + \beta P(h)v_\beta(f)$ in $1-\beta$ when β tends to 1 (cf. MILLER and VEINOTT [1969]). In section 6 various more sensitive optimality criteria are formulated for the case that β tends to 1 (cf. BLACKWELL [1962] and VEINOTT [1966]). The results of the latter two sections will be used extensively in chapters 6 - 8.

5.2. THE VARIOUS CONTRACTIVE MDP MODELS

In this section subsequently four different models for the contracting MDP will be studied. It will be shown that these models are equivalent with respect to the ∞-horizon behaviour.
The first, and most general, model is the following.

Model I

Define $\bar{r}(i) := \sup_{a \in A} r(i,a)$, $i \in S$. (Then $\bar{r} < \infty$ by condition 1.1.) There exists a nonnegative real-valued function μ on S ($\mu \in V$) such that:

(i) For some constants ρ_1 and M_1, with $0 \leq \rho_1 < 1$ and $M_1 \geq 0$,

(5.1) $\quad P(f)\bar{r} - \rho_1 \bar{r} \leq M_1 \mu$ for all $f \in F$.

(ii) For some constant ρ_2, with $0 \leq \rho_2 < 1$,

(5.2) $\quad P(f)\mu \leq \rho_2 \mu$ for all $f \in F$.

(iii) There exists a policy $h \in F$ and constants $M_2, M_3 \geq 0$, such that

(5.3) $\quad \rho_1 \bar{r} - P(h)\bar{r} \leq M_2 \mu$ and $\bar{r} - r(h) \leq M_3 \mu$.

This model is somewhat more general than the model studied in Van NUNEN and WESSELS [1977a], where it is assumed that $\bar{r} - r(h) \leq M_3 \mu$ for all $h \in F$, and than the model of Van NUNEN [1976a], who assumes that (5.3) holds for all $h \in F$.

First the model I assumptions will be analyzed, from which we see that we can a priori eliminate in each state a number of suboptimal actions. After this elimination the MDP (with now state-dependent action sets) fits into the contracting model of Van NUNEN [1976a]. Further two data transformations will be considered. The first one transforms Van Nunen's model into the model of WESSELS [1977b] and the second one transforms Wessels' model into the 'standard' discounted model.

Now let us consider model I.
From (5.1)-(5.3) one can already obtain some bounds on v^*. By the definition of $\bar{r}$ we have for all $\pi = (f_0,f_1,\ldots) \in M$,

$$(5.4) \qquad v_n(\pi) = r(f_0) + P(f_0)r(f_1) + \ldots + P(f_0)\cdots P(f_{n-2})r(f_{n-1})$$

$$\leq \bar{r} + P(f_0)\bar{r} + \ldots + P(f_0)\cdots P(f_{n-2})\bar{r} .$$

Define $\rho_* := \max\{\rho_1,\rho_2\}$, then we have the following lemma.

LEMMA 5.1 (cf. Van NUNEN [1976a, lemma 3.1.2]). *For all* $f_0,\ldots,f_k \in F$ *we have (within Model I)*

$$(5.5) \qquad P(f_0)\cdots P(f_k)\bar{r} \leq \rho_1^{k+1}\bar{r} + (k+1)\rho_*^k M_1\mu \qquad k = 1,2,\ldots .$$

PROOF. From (5.1), (5.2) and the definition of ρ_*,

$$P(f_0)\cdots P(f_k)\bar{r} \leq P(f_0)\cdots P(f_{k-1})(\rho_1\bar{r} + M_1\mu)$$

$$\leq \rho_1 P(f_0)\cdots P(f_{k-1})\bar{r} + \rho_2^k M_1\mu$$

$$\leq \ldots \leq \rho_1^{k+1}\bar{r} + (\rho_2^k + \rho_1\rho_2^{k-1} + \ldots + \rho_1^k)M_1\mu$$

$$\leq \rho_1^{k+1}\bar{r} + (k+1)\rho_*^k M_1\mu .$$ □

Substitution of (5.5) into (5.4) yields for all $\pi \in M$,

$$(5.6) \qquad v_n(\pi) \leq \sum_{k=0}^{n-1} \rho_1^k\bar{r} + \sum_{k=0}^{n-2}(k+1)\rho_*^k M_1\mu .$$

On the other hand, we have for any policy $h \in F$ satisfying (5.3),

$$(5.7)\qquad v_n(h) = r(h) - \bar{r} + P(h)(r(h) - \bar{r}) + \ldots + P^{n-1}(h)(r(h) - \bar{r}) +$$

$$+ \bar{r} + P(h)\bar{r} + \ldots + P^{n-1}(h)\bar{r}$$

$$\geq (I + P(h) + \ldots + P^{n-1}(h))(-M_3\mu) + (I + P(h) + \ldots + P^{n-1}(h))\bar{r}$$

$$\geq -\sum_{k=0}^{n-1} \rho_2^k M_3\mu + \sum_{k=0}^{n-1} P^k(h)\bar{r} .$$

Further, the following analogon of lemma 5.1 holds.

LEMMA 5.2 (cf. Van NUNEN [1976a, lemma 3.1.2]). *For any policy* $h \in F$ *satisfying* (5.3) *we have*

$$(5.8)\qquad P^k(h)\bar{r} \geq \rho_1^k \bar{r} - k\rho_*^{k-1} M_2\mu , \quad k = 1,2,\ldots .$$

PROOF. Similar to the proof of lemma 5.1. □

So, from (5.7), (5.8) and (5.2), also

$$(5.9)\qquad v_n(h) \geq \sum_{k=0}^{n-1} \rho_1^k \bar{r} - \sum_{k=0}^{n-1} \rho_2^k M_3\mu - \sum_{k=0}^{n-2} (k+1)\rho_*^k M_2\mu .$$

Letting n tend to infinity in formulae (5.6) and (5.9) yields, since by theorem 2.19 one has to consider only Markov strategies,

THEOREM 5.3. *For an MDP satisfying the assumptions of model I we have*

$$-(1-\rho_2)^{-1} M_3\mu - (1-\rho_*)^{-2} M_2\mu \leq v^* - (1-\rho_1)^{-1}\bar{r} \leq (1-\rho_*)^{-2} M_1\mu .$$

The second inequality in theorem 5.3 implies, with lemma 5.1 and (5.2), also that

$$(5.10)\qquad \limsup_{n\to\infty} \mathbb{E}_\pi v^*(X_n) \leq 0 \quad \text{for all } \pi \in M .$$

Hence, by theorem 2.14, a uniformly ε-optimal Markov strategy (in the additive sense) exists.
(5.10) can also be used to prove the following result for stationary strategies.

THEOREM 5.4. *For an MDP satisfying the assumptions of model I, there exists for all* $\varepsilon > 0$ *a policy* $f \in F$ *satisfying* $v(f) \geq v^* - \varepsilon\mu$.

PROOF. Since v^* satisfies the optimality equation, there exists for all $\varepsilon > 0$ a policy f for which

$$L(f)v^* \geq v^* - \varepsilon(1-\rho_2)\mu \ .$$

Then, for all n,

$$(5.11) \quad L^n(f)v^* \geq L^{n-1}(f)[v^* - \varepsilon(1-\rho_2)\mu] = L^{n-1}(f)v^* - \varepsilon(1-\rho_2)P^{n-1}(f)\mu$$

$$\geq L^{n-1}(f)v^* - \varepsilon(1-\rho_2)\rho_2^{n-1}\mu \geq \ldots \geq$$

$$\geq v^* - \varepsilon(1-\rho_2)(1+\rho_2+\ldots+\rho_2^{n-1})\mu \geq v^* - \varepsilon\mu \ .$$

Further,

$$v(f) = \lim_{n\to\infty} L^n(f)0 = \lim_{n\to\infty}[L^n(f)v^* - P^n(f)v^*] \ .$$

So, with (5.10) and (5.11),

$$v(f) \geq \liminf_{n\to\infty} L^n(f)v^* - \limsup_{n\to\infty} P^n(f)v^* \geq v^* - \varepsilon\mu \ . \qquad \square$$

This result enables us to eliminate some of the suboptimal actions from the MDP, so that after the action elimination procedure all policies will satisfy (5.3) (with different constants M_2 and M_3).
From theorem 5.4 it follows that, if in each state $i \in S$ all actions $a \in A$ satisfying

$$(5.12) \quad r(i,a) + \sum_{j\in S} p(i,a,j)v^*(j) < v^*(i) - K\mu(i)$$

are eliminated ($K > 0$ is some arbitrary constant), then the value of the MDP will remain unchanged. However, the set of actions in state i may now be different for each i. Moreover, there will still exist for all $\varepsilon > 0$ a stationary "$\varepsilon\mu$-optimal" strategy.
Using this idea we introduce the following elimination procedure.

A priori action elimination procedure

Eliminate in each state $i \in S$ those actions for which

$$(5.13) \quad r(i,a) + \sum_{j\in S} p(i,a,j)[(1-\rho_1)^{-1}\bar{r}(j) + (1-\rho_*)^{-2}M_1\mu(j)] <$$

$$< (1-\rho_1)^{-1}\bar{r}(i) - (1-\rho_2)^{-1}M_3\mu(i) - (1-\rho_*)^{-2}M_2\mu(i) - \mu(i) \ .$$

One may easily verify that it follows from theorem 5.3 that all actions satisfying (5.13) also satisfy (5.12) with K = 1. So, using the action elimination procedure we obtain an MDP which, with respect to the value and (nearly-) optimal stationary strategies, is equivalent to the original model I MDP.

For the remaining actions in state i we have

$$r(i,a) - \bar{r}(i) + (1-\rho_1)^{-1}[\sum_{j\in S} p(i,a,j)\bar{r}(j) - \rho_1\bar{r}(i)] \geq -M_4\mu(i) ,$$

with

$$M_4 = (1-\rho_*)^{-2}\rho_2 M_1 + (1-\rho_*)^{-2} M_2 + (1-\rho_2)^{-1} M_3 + 1 .$$

So, clearly

(5.14) $$r(i,a) - \bar{r}(i) \geq -[(1-\rho_1)^{-1} M_1 + M_4]\mu(i) ,$$

and

(5.15) $$\sum_{j\in S} p(i,a,j)\bar{r}(j) - \rho_1\bar{r}(i) \geq -(1-\rho_1)M_4\mu(i) .$$

Thus after the a priori action elimination procedure we obtain an MDP (with state dependent action sets A(i)), which satisfies the following conditions.

Model II

Define $\bar{r}(i) := \sup_{a\in A(i)} r(i,a)$.
There exists a nonnegative real-valued function μ on S such that:

(i) For some constant $M_1 \geq 0$,

(5.16) $$\bar{r} - r(f) \leq M_1\mu \quad \text{for all } f \in F .$$

(ii) For some constants ρ_1 and M_2, with $0 \leq \rho_1 < 1$ and $M_2 \geq 0$,

(5.17) $$|P(f)\bar{r} - \rho_1\bar{r}| \leq M_2\mu \quad \text{for all } f \in F .$$

(iii) For some constant ρ_2, with $0 \leq \rho_2 < 1$,

(5.18) $$P(f)\mu \leq \rho_2\mu \quad \text{for all } f \in F .$$

This is precisely the model studied by Van NUNEN [1976a].

For the model II MDP we have the following result.

THEOREM 5.5. *For an MDP satisfying the assumptions of model II, we have for all* $\pi \in \Pi$ *and all* $n = 1,2,\ldots$

$$\text{(i)} \quad -\sum_{k=0}^{n-1} \rho_2^k M_1\mu - \sum_{k=0}^{n-2} (k+1)\rho_*^k M_2\mu \le v_n(\pi) - \sum_{k=0}^{n-1} \rho_1^k \bar{r} \le \sum_{k=0}^{n-2} (k+1)\rho_*^k M_2\mu ;$$

$$\text{(ii)} \quad -(1-\rho_2)^{-1} M_1\mu - (1-\rho_*)^{-2} M_2\mu \le v(\pi) - (1-\rho_1)^{-1}\bar{r} \le (1-\rho_*)^{-2} M_2\mu ,$$

with again $\rho_* = \max\{\rho_1,\rho_2\}$.

PROOF.

(i) For all $\pi \in M$ (i) follows analogously to (5.4) and (5.7). Since in the maximization of $v_n(\pi)$ and (here) also in the minimization one may restrict oneself to Markov strategies, (i) holds for all $\pi \in \Pi$.

(ii) is obtained from (i) by letting n tend to infinity. □

So, all strategies in model II have an ∞-horizon reward of $(1-\rho_1)^{-1}\bar{r}$ plus some term which is bounded in μ-norm.

Using the following transformation of the immediate rewards, which is due to PORTEUS [1975], we obtain from the model II MDP a new MDP that fits into the framework of the model studied by WESSELS [1977b]. See also Van NUNEN and WESSELS [1977a].

$$\text{(5.19)} \quad \tilde{r}(i,a) = r(i,a) - (1-\rho_1)^{-1}[\bar{r}(i) - \sum_{j\in S} p(i,a,j)\bar{r}(j)] .$$

Then it follows immediately from (5.16) and (5.17) that

$$\tilde{r}(f) \in V_\mu \quad \text{for all } f \in F .$$

Combined with (5.18), this implies that the newly obtained MDP satisfies the conditions of WESSELS [1977b].

Further, let $\tilde{v}_n(\pi)$ and $\tilde{v}(\pi)$ denote the n-period and ∞-horizon total expected rewards in the MDP with $\tilde{r}$ instead of r. Then for all $\pi = (f_0,f_1,\ldots) \in M$,

$$\text{(5.20)} \quad \tilde{v}_n(\pi) = \tilde{r}(f_0) + P(f_0)\tilde{r}(f_1) + \ldots + P(f_0)\cdots P(f_{n-2})\tilde{r}(f_{n-1})$$

$$= r(f_0) - (1-\rho_1)^{-1}(\bar{r} - P(f_0)\bar{r}) +$$

$$+ P(f_0)[r(f_1) - (1-\rho_1)^{-1}(\bar{r} - P(f_1)\bar{r})] +$$

$$+ \ldots + P(f_0) \cdots P(f_{n-2})[r(f_{n-1}) - (1-\rho_1)^{-1}(\bar{r} - P(f_{n-1})\bar{r})] =$$

$$= r(f_0) + P(f_0)r(f_1) + \ldots + P(f_0) \cdots P(f_{n-2})r(f_{n-1}) +$$

$$- (1-\rho_1)^{-1}[\bar{r} - P(f_0)\bar{r} + P(f_0)\bar{r} + \ldots - P(f_0) \cdots P(f_{n-1})\bar{r}]$$

$$= v_n(\pi) - (1-\rho_1)^{-1}(\bar{r} - P(f_0) \cdots P(f_{n-1})\bar{r}) \ .$$

Since $P(f_0) \cdots P(f_{n-1})\bar{r}$ tends to zero if n tends to infinity (cf. lemmas 5.1 and 5.2), we obtain from (5.20)

THEOREM 5.6. *For all* $\pi \in \Pi$,

$$\tilde{v}(\pi) = v(\pi) - (1-\rho_1)^{-1}\bar{r} \ .$$

PROOF. For $\pi \in M$ the result follows from (5.20) with $n \to \infty$. For arbitrary $\pi \in \Pi$ the result can be obtained in an analogous way, see Van NUNEN and WESSELS [1977a]. □

So, the model II MDP and the transformed problem obtained from it via (5.19) are equivalent with respect to the ∞-horizon behaviour, since the total expected rewards differ only by a strategy-independent amount $(1-\rho_1)^{-1}\bar{r}$. The successive approximations, however, may differ as we see from (5.20), since the term $P(f_0) \cdots P(f_{n-1})\bar{r}$ is not independent of the strategy. Therefore, if we are not interested in the finite-horizon behaviour, we can just as well perform transformation (5.19) which leads to a third and somewhat simpler model.

Model III

A nonnegative real-valued function μ on S exists such that:

(i) For some constant $M \geq 0$

$$|r(f)| \leq M\mu \quad \text{for all } f \in F \ .$$

(ii) For some constant ρ, with $0 \leq \rho < 1$,

$$P(f)\mu \leq \rho\mu \quad \text{for all } f \in F \ .$$

As remarked before, this is the contracting model considered in WESSELS [1977b].

With a second data transformation an MDP of the model III type can be transformed into the 'standard' discounted model with bounded rewards. This so-called similarity transformation is due to VEINOTT [1969], and applied to this model it can be found in Van NUNEN and WESSELS [1977a]. Let Λ be the diagonal matrix defined by

$$\Lambda(i,i) := \mu(i) ,$$

and let Λ^- be the diagonal matrix with

$$\Lambda^-(i,i) := \begin{cases} \mu(i)^{-1} & \text{if } \mu(i) > 0 , \\ 0 & \text{elsewhere.} \end{cases}$$

Now consider the following transformation

$$(5.21) \quad \begin{cases} \hat{r}(i,a) := \Lambda^-(i,i) r(i,a) , & i \in S , \ a \in A , \\ \hat{p}(i,a,j) := \Lambda^-(i,i) p(i,a,j) \Lambda(j,j) , & i,j \in S , \ a \in A . \end{cases}$$

So for policies,

$$\hat{r}(f) = \Lambda^- r(f)$$

and

$$\hat{P}(f) = \Lambda^- P(f) \Lambda .$$

Then for all $f \in F$,

$$(5.22) \quad |\hat{r}(f)| = \Lambda^- |r(f)| \leq M \Lambda^- \mu \leq Me ,$$

and

$$(5.23) \quad \hat{P}(f)e = \Lambda^- P(f) \Lambda e = \Lambda^- P(f) \mu \leq \rho \Lambda^- \mu \leq \rho e .$$

Further, let $\hat{v}_n(\pi)$ and $\hat{v}(\pi)$ denote the n-period and ∞-horizon total expected rewards, respectively, in the MDP obtained from a model III MDP via transformation (5.21). Then we have

$$(5.24) \quad \begin{aligned} \hat{v}_n(\pi) &= \hat{r}(f_0) + \hat{P}(f_0)\hat{r}(f_1) + \ldots + \hat{P}(f_0) \cdots \hat{P}(f_{n-2}) r(f_{n-1}) \\ &= \Lambda^- r(f_0) + \Lambda^- P(f_0) \Lambda \Lambda^- r(f_1) + \ldots + \\ &\qquad + \Lambda^- P(f_0) \Lambda \cdots \Lambda^- P(f_{n-2}) \Lambda \Lambda^- r(f_{n-1}) . \end{aligned}$$

If $\mu(i) > 0$, then $\Lambda\Lambda^-(i,i) = 1$ and if $\mu(i) = 0$, then $\Lambda\Lambda^-(i,i) = 0$, so for

all $v \in V_\mu$

$$\Lambda\bar{\Lambda} v = v$$

(if $v \in V_\mu$ and $\mu(i) = 0$ then also $v(i) = 0$). Thus

$$\Lambda\bar{\Lambda} r(f) = r(f) \quad \text{and} \quad \Lambda\bar{\Lambda} P(f)v = P(f)v \quad \text{for all } f \in F \text{ and } v \in V_\mu .$$

Substitution of this into (5.24) yields

$$\hat{v}_n(\pi) = \bar{\Lambda} v_n(\pi) .$$

And, with $n \to \infty$, also

$$\hat{v}(\pi) = \bar{\Lambda} v(\pi) .$$

So this second data transformation leads to a completely equivalent model: essentially the same n-period rewards and thus the same successive approximations, and the same ∞-horizon rewards.
Only, note that (5.23) states that the matrices $\hat{P}(f)$ are no longer stochastic. This can be simply repaired by the addition of an extra absorbing state, but we will not do this explicitly here.
Thus from a model III MDP we obtain by transformation (5.21) the equivalent

Model IV

(i) There exists a constant $M \geq 0$ such that

$$|r(f)| \leq Me \quad \text{for all } f \in F .$$

(ii) There exists a constant ρ, with $0 \leq \rho < 1$, such that

$$P(f)e \leq \rho e \quad \text{for all } f \in F .$$

An example of a model IV MDP is obtained if in the finite state discounted MDP the discountfactor is incorporated in the transition probabilities, but the additional state is not introduced.
Model IV is slightly simpler to deal with than model III. However, for two reasons we prefer not to transform a model III MDP into a model IV MDP:
(i) the reward structure in a countable state MDP is typically unbounded;
(ii) the 'transition probabilities' in a model IV MDP obtained via (5.21) no longer have this physical interpretation with respect to the original MDP.

Whenever in the following two sections we speak of a contracting MDP we will mean an MDP of the model III type.

5.3. CONTRACTION AND STRONG CONVERGENCE

This section deals with the relation between the contraction assumption (model III) and the strong-convergence condition.
For a model III contracting MDP we have for all $\pi = (f_0, f_1, \ldots) \in M$,

$$\mathbb{E}_\pi |r(X_n, A_n)| = P(f_0) \cdots P(f_{n-1}) |r(f_n)|$$

$$\leq P(f_0) \cdots P(f_{n-1}) M\mu \leq \rho^n M\mu , \quad n = 1,2,\ldots ,$$

which yields the following result.

THEOREM 5.7 (cf. Van HEE, HORDIJK and Van der WAL [1977]). *An MDP that is contracting in the sense of model III satisfies the strong convergence condition for a sequence* $\varphi = (\varphi_0, \varphi_1, \ldots) \in \Phi$ *with* $\varphi_n = \lambda^n e$, $n = 0,1,\ldots$, *where* λ *is any constant satisfying* $1 < \lambda < \rho^{-1}$.

PROOF.

$$z_\varphi^* = \sup_{\pi \in M} \sum_{n=0}^{\infty} \lambda^n |r(X_n, A_n)| \leq M(1 - \lambda\rho)^{-1} \mu < \infty . \qquad \square$$

Further we have

THEOREM 5.8. *If S is finite, then the following two conditions are equivalent:*

(i) *The MDP is contracting (in the sense of model III).*
(ii) *The MDP is strongly convergent.*

PROOF. By theorem 5.7 it only remains to be shown that (ii) implies (i). By the strong convergence and the finiteness of S a constant α, $0 \leq \alpha < 1$, and an integer n_0 exist such that

$$\sup_{\pi \in M} \mathbb{E}_\pi z^*(X_{n_0}) \leq \alpha^{n_0} z^* .$$

Now, following WALTER [1976], we define

$$\nu := z^* + \alpha^{-1} \tilde{U} z^* + \alpha^{-2} \tilde{U}^2 z^* + \ldots + \alpha^{-n_0+1} \tilde{U}^{n_0-1} z^* .$$

Then for all $f \in F$

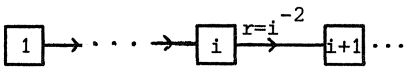
$$P(f)\nu \le \tilde{U}\nu \le \tilde{U}z^* + \alpha^{-1}\tilde{U}^2 z^* + \ldots + \alpha^{-n_0+1}\tilde{U}^{n_0} z^*$$

$$\le \tilde{U}z^* + \alpha^{-1}\tilde{U}^2 z^* + \ldots + \alpha^{-n_0+2}\tilde{U}^{n_0-1} z^* + \alpha z^* = \alpha\nu \ .$$

Further, $|r(f)| \le z^* \le \nu$ for all $f \in F$. Hence the MDP satisfies the assumptions of model III with $\mu = \nu$, $M = 1$ and $\rho = \alpha$. □

If S is countable, then strong convergence need not imply contraction, as we see in the following example.

EXAMPLE 5.9. $S = \{1,2,\ldots\}$, $A = \{1\}$, $p(i,1,i+1) = 1$ and $r(i,1) = i^{-2}$. This MDP is clearly strongly convergent; take for example $\varphi_n(i) = \sqrt{i+n}$. In order that this MDP is contracting with bounding function μ, the function μ has to satisfy for all $i \in S$



$$r(i,1) \le M\mu(i) \quad \text{or} \quad \mu(i) \ge M^{-1}i^{-2} \ ,$$

but also

$$(P^{i-1}\mu)(1) = \mu(i) \le \rho^{i-1}\mu(1) \quad \text{for some } \rho < 1 \ .$$

This is impossible, so this MDP is not contractive in the sense of model III. One may verify that also in the sense of model II this example is not contracting.

For some further discussion on the relation between contraction, bounding functions and the spectral radius of the MDP, see Van HEE and WESSELS [1978] and ZIJM [1978].

5.4. CONTRACTION AND SUCCESSIVE APPROXIMATIONS

In this section we will consider in some detail the various successive approximation methods for the model III MDP. First the set of algorithms generated by means of (nonzero stationary) go-ahead functions is considered. The use of these go-ahead functions in the case of contraction has been extensively studied by WESSELS [1977a], Van NUNEN and WESSELS [1976,1977b], Van NUNEN [1977a] and Van NUNEN and STIDHAM [1978]. All results presented here can be found in one of these papers. We review these results here for the sake of completeness and for later reference in chapter 11.

From theorems 5.7 and 4.12 it follows that the method of standard successive approximations converges. Further we have from lemma 4.22

THEOREM 5.10. *For any go-ahead function* δ, *for all* $\pi \in \Pi$ *and for all* $v,w \in V_\mu$

(i) $\quad \|L_\delta(\pi)v - L_\delta(\pi)w\|_\mu \leq (1 - \alpha_\delta + \alpha_\delta\rho)\|v - w\|_\mu$;

(ii) $\quad \|U_\delta v - U_\delta w\| \leq (1 - \alpha_\delta + \alpha_\delta\rho)\|v - w\|_\mu$.

PROOF. We only prove (i) since the proof of (ii) is very similar.

$$L_\delta(\pi)v - L_\delta(\pi)w = \tilde{L}_\delta(\pi)(v - w) \leq \|v - w\|_\mu \tilde{L}_\delta(\pi)\mu \leq \|v - w\|_\mu \tilde{U}_\delta\mu .$$

With $\tilde{U}\mu \leq \rho\mu$, it follows from lemma 4.22 that

$$\tilde{U}_\delta\mu \leq (1 - \alpha_\delta + \alpha_\delta\rho)\mu .$$

So

$$L_\delta(\pi)v - L_\delta(\pi)w \leq (1 - \alpha_\delta + \alpha_\delta\rho)\|v - w\|_\mu \mu .$$

Reversing the roles of v and w yields

$$|L_\delta(\pi)v - L_\delta(\pi)w| \leq (1 - \alpha_\delta + \alpha_\delta\rho)\|v - w\|_\mu \mu ,$$

from which the proof is immediate. □

Since the space V_μ is a Banach space we have

COROLLARY 5.11. *If* $\alpha_\delta > 0$, *then* $L_\delta(\pi)$ *and* U_δ *are contractions on* V_μ *with radius less than or equal to* $(1 - \alpha_\delta + \alpha_\delta\rho)$, *and thus have unique fixed points (within* V_μ*).*
The fixed point of $L_\delta(f)$ *is* $v(f)$, *the fixed point of* U_δ *is* v^*.

Note that in general the fixed point of $L_\delta(\pi)$ will be unequal to $v(\pi)$.

A very important consequence of the contraction assumption is that it allows for extrapolations that yield bounds on v^* and enable us to recognize nearly-optimal strategies.
In order to formulate the results, we use the following notations. Let δ be a nonzero stationary go-ahead function (if one is interested in convergence of successive approximations and stationary strategies it is reasonable to consider only these go-ahead functions). Define

$$\|w\|_\mu^{\max} := \inf \{c \in \mathbb{R} \mid w \le c\mu\} \quad \text{for all } w \in V_\mu ,$$

$$\|w\|_\mu^{\min} := \sup \{c \in \mathbb{R} \mid w \ge c\mu\} \quad \text{for all } w \in V_\mu ,$$

$$\rho_\delta^{\max}(f) := \|\tilde{L}_\delta(f)\mu\|_\mu^{\max} \quad \text{for all } f \in F ,$$

$$\rho_\delta^{\min}(f) := \|\tilde{L}_\delta(f)\mu\|_\mu^{\min} \quad \text{for all } f \in F ,$$

$$\rho_\delta^{\max} := \sup_{f\in F} \rho_\delta^{\max}(f) ,$$

$$\rho_\delta^{\min} := \inf_{f\in F} \rho_\delta^{\min}(f) .$$

Further, define for all $v \in V_\mu$,

$$\rho_{\delta,v}(f) := \begin{cases} \rho_\delta^{\max}(f) & \text{if } \|L_\delta(f)v - v\|_\mu^{\min} < 0 \\ \rho_\delta^{\min}(f) & \text{if } \|L_\delta(f)v - v\|_\mu^{\min} \ge 0 , \end{cases}$$

and

$$\rho_{\delta,v}^* := \begin{cases} \rho_\delta^{\max} & \text{if } \|U_\delta v - v\|_\mu^{\max} > 0 \\ \rho_\delta^{\min} & \text{if } \|U_\delta v - v\|_\mu^{\max} \le 0 . \end{cases}$$

Then we have the following theorem.

THEOREM 5.12. *Let δ be a nonzero stationary go-ahead function and let $v \in V_\mu$ and $f \in F$ be arbitrary. Then*

(i) $$v(f) \ge L_\delta(f)v + \rho_{\delta,v}(f)(1 - \rho_{\delta,v}(f))^{-1}\|L_\delta(f)v - v\|_\mu^{\min}\mu ,$$

(ii) $$v^* \le U_\delta v + \rho_{\delta,v}^*(1 - \rho_{\delta,v}^*)^{-1}\|U_\delta v - v\|_\mu^{\max}\mu .$$

PROOF.

(i) By corollary 5.11,

$$v(f) = \lim_{n\to\infty} L_\delta^n(f)v .$$

Further, for all $n \ge 1$,

$$L_\delta^n(f)v \geq L_\delta^{n-1}(f)(v + \|L_\delta(f)v - v\|_\mu^{\min}\,\mu)$$

$$\geq L_\delta^{n-1}(f)v + \rho_{\delta,v}^{n-1}(f)\|L_\delta(f)v - v\|_\mu^{\min}\,\mu$$

$$\geq \ldots \geq L_\delta(f)v + [\rho_{\delta,v}(f) + \ldots + \rho_{\delta,v}^{n-1}(f)]\|L_\delta(f)v - v\|_\mu^{\min}\,\mu\,.$$

Hence (i) follows by letting n tend to infinity.

(ii) Having noted that $v(\pi)$ is optimized by stationary strategies, (ii) follows in an analogous way. □

Now consider for a nonzero go-ahead function δ the following successive approximation scheme:

(5.25)
$$\begin{cases} \text{Choose } v_0 \in V_\mu. \\ \text{Determine for } n = 0,1,\ldots \\ \qquad v_{n+1} = U_\delta v_n\,. \end{cases}$$

Then it follows from theorem 5.10 that

$$\|v_{n+1} - v_n\|_\mu \leq (1 - \alpha_\delta + \alpha_\delta\rho)^n\|U_\delta v_0 - v_0\|_\mu \to 0 \quad (n \to \infty)\,.$$

Further, for all $v \in V_\mu$ and all $\varepsilon > 0$ a policy f exists satisfying

$$L_\delta(f)v \geq U_\delta v - \varepsilon\mu\,.$$

That such a uniformly nearly-optimal policy indeed exists can be shown for example with theorems 5.7 and 4.11 and the proof of theorem 4.26, since for the sequence φ mentioned in theorem 5.7 also $z_\varphi^* \in V_\mu$.
Thus there also exists a policy f_n such that

$$\|L_\delta(f_n)v_n - v_n\|_\mu$$

is small if n is large.
From this and theorem 5.12 it follows that we can obtain bounds on v^* and nearly-optimal stationary strategies from the successive approximation scheme (5.25).

A second type of algorithms is formed by the set of value-oriented methods. These methods converge for any sequence $d_n \leq \varepsilon_n\mu$ with $\varepsilon_n \downarrow 0$, as follows from theorem 5.7 and a slightly changed variant of theorem 4.28(ii) (using

$z^* \leq M(1-\rho)^{-1}\mu)$. Though, as has been shown by Van NUNEN [1976a], the mapping that generates v_{n+1} from v_n is neither necessarily monotone, nor necessarily contracting, one may easily show that v_n converges to v^* exponentially fast (compare the proof of theorem 4.28(ii)).
Further, one can use theorem 5.12, for the δ corresponding to the method of standard successive approximations, to obtain bounds on v^* and nearly-optimal stationary strategies.

5.5. *THE DISCOUNTED MDP WITH FINITE STATE AND ACTION SPACES*

In this section we study a special contracting MDP, namely the discounted MDP with finite state space $S = \{1,2,\ldots,N\}$ and finite action space. Because of the relation with the average-reward MDP (which will be studied in chapters 6.9) we consider in particular the case that the discount factor approaches 1.
Moreover, we prove theorem 2.21, the proof of which we postponed.

The total expected β-discounted reward, when strategy π is used, is defined by (see (1.11))

$$(5.26) \qquad v_\beta(\pi) := \mathbb{E}_\pi \sum_{n=0}^{\infty} \beta^n r(X_n,A_n) \ , \quad \pi \in \Pi \ , \quad 0 \leq \beta < 1 \ .$$

Clearly, the expectation is properly defined for all $0 \leq \beta < 1$, since

$$\mathbb{E}_\pi \sum_{n=0}^{\infty} \beta^n r^+(X_n,A_n) \leq (1-\beta)^{-1} \max_{i,a} r^+(i,a)e < \infty \ .$$

Further, define

$$(5.27) \qquad v_\beta^* := \sup_{\pi\in\Pi} v_\beta(\pi) \ .$$

As already remarked in section 1.5 this discounted MDP can be fitted into the general model by the addition of an absorbing state, $*$ say. Defining the function μ on $S \cup \{*\}$ by $\mu(i) = 1$, $i \in S$ and $\mu(*) = 0$, one easily verifies that this extended MDP is contracting in the sense of model III.
We will not incorporate the discountfactor into the transition probabilities, since we want to study the case of a varying discountfactor, but we do use the fact that the discounted MDP is contracting.

In order to study the β-discounted MDP, it is convenient to define the operators $L_\beta(f)$ and U_β on V by

$$L_\beta(f)v := r(f) + \beta P(f)v \tag{5.28}$$

and

$$U_\beta v := \max_{f \in F} L_\beta(f)v \ . \tag{5.29}$$

As A is finite, there exists for each $\beta \in [0,1)$ a policy f_β satisfying the optimality equation

$$L_\beta(f_\beta)v_\beta^* = v_\beta^* \ . \tag{5.30}$$

Further, by corollary 5.11,

$$v_\beta(f) = \lim_{n\to\infty} L_\beta^n(f)v \quad \text{for all } v \in V \ . \tag{5.31}$$

Hence

$$v_\beta(f_\beta) = \lim_{n\to\infty} L_\beta^n(f_\beta)v_\beta^* = v_\beta^* \ , \tag{5.32}$$

which is merely a special case of theorem 4.8(i). This leads to the following result which has already been proved by SHAPLEY [1953].

THEOREM 5.13. *If* S *and* A *are finite, then there exists for all* $\beta \in [0,1)$ *an optimal stationary strategy for the* β*-discounted MDP, i.e., a policy* f_β *satisfying*

$$v_\beta(f_\beta) = v_\beta^* \ .$$

An important consequence of this result is, that it enables us to prove theorem 2.21, the proof of which has been postponed.

THEOREM 2.21. *If* S *and* A *are finite and condition 1.1 holds, then there exists a stationary strategy* f*, satisfying* $v(f) = v^*$.

PROOF. Since S and A are finite, there are only finitely many policies. So, let $\{\beta_n, n = 0,1,\ldots\}$ be a sequence of discount factors with β_n tending to 1, then there exists a subsequence $\{\beta_{n_k}, k = 0,1,\ldots\}$ and a policy f^* such that

$$v_{\beta_{n_k}}(f^*) = v_{\beta_{n_k}}^* \quad \text{for } k = 0,1,\ldots \ .$$

Further, let π be an arbitrary strategy, then we have

$$v(\pi) = \lim_{\beta \uparrow 1} v_\beta(\pi) \ .$$

Thus for all $\pi \in \Pi$

$$v(\pi) = \lim_{\beta \uparrow 1} v_\beta(\pi) = \lim_{k \to \infty} v_{\beta_{n_k}}(\pi) \leq \lim_{k \to \infty} v_{\beta_{n_k}}(f^*) = v(f^*) \ .$$

Hence,

$$v(f^*) = v^* \ . \qquad \square$$

In the remainder of this section the case that the discountfactor tends to 1 will be studied. Our interest in this case is caused by the relationship that exists between the average-reward MDP and the discounted MDP with discountfactor close to 1, see e.g. BLACKWELL [1962], VEINOTT [1966] and MILLER and VEINOTT [1969].
First let us derive a Laurent series expansion in powers of $1 - \beta$ for $v_\beta(f)$, for β tending to 1. Miller and Veinott already derived the Laurent series expansion in $\beta^{-1}(1 - \beta)$, but for our purpose (particularly chapter 8) it is more convenient to deal with the expansion in $1 - \beta$.

For any stationary strategy f we have

$$v_\beta(f) = \sum_{n=0}^{\infty} \beta^n P^n(f) r(f) \ . \tag{5.33}$$

Define

$$P^*(f) := \lim_{n \to \infty} n^{-1} \sum_{k=0}^{n-1} P^k(f) r(f) \ . \tag{5.34}$$

Then

$$P^*(f) r(f) = \lim_{n \to \infty} n^{-1} \sum_{k=0}^{n-1} P^k(f) r(f) = \lim_{n \to \infty} n^{-1} v_n(f) = g(f) \ , \tag{5.35}$$

where $g(f)$ is the average reward per unit time for strategy f (cf. (1.12)).
Hence, (5.33) can be rewritten as follows,

$$v_\beta(f) = \sum_{n=0}^{\infty} \beta^n [P^n(f) - P^*(f) + P^*(f)] r(f) =$$

$$= (1 - \beta)^{-1} g(f) + \sum_{n=0}^{\infty} \beta^n [P^n(f) - P^*(f)] r(f) \ .$$

Since

$$P(f)P^*(f) = P^*(f)P(f) = P^*(f) \quad (5.36)$$

also

$$P^n(f) - P^*(f) = [P(f) - P^*(f)]^n , \quad n = 1,2,\ldots .$$

This gives

$$v_\beta(f) = (1-\beta)^{-1} g(f) - g(f) + \sum_{n=0}^{\infty} \beta^n [P(f) - P^*(f)]^n r(f) \quad (5.37)$$

$$= (1-\beta)^{-1} g(f) - g(f) + [I - \beta P(f) + \beta P^*(f)]^{-1} r(f) .$$

That $I - \beta P(f) + \beta P^*(f)$ is nonsingular for all β, with $0 \leq \beta \leq 1$ (1 included), can be seen as follows. Let x satisfy

$$[I - \beta P(f) + \beta P^*(f)]x = 0 , \quad (5.38)$$

then premultiplication with $P^*(f)$ yields, with (5.36),

$$P^*(f)x = 0 .$$

Substituting this in (5.38) yields $x = \beta P(f)x$, so for $\beta < 1$ we have $x = 0$. If $\beta = 1$, then iterating and averaging $x = P(f)x$ yields

$$x = n^{-1} \sum_{k=0}^{n-1} P^k(f)x ,$$

so, with $n \to \infty$, $x = P^*(f)x = 0$.
From the nonsingularity of $I - P(f) + P^*(f)$ we obtain a Laurent series expansion for $I - \beta P(f) + \beta P^*(f)$ in $1 - \beta$, for β sufficiently close to 1, in the following way.
Writing Q for $P(f) - P^*(f)$, we have

$$(1 - \beta Q)^{-1} = (I - Q + (1-\beta)Q)^{-1} = (I + (1-\beta)(I-Q)^{-1}Q)^{-1}(I-Q)^{-1} \quad (5.39)$$

$$= \sum_{k=0}^{\infty} (-1)^k (1-\beta)^k [(I-Q)^{-1}Q]^k (I-Q)^{-1} .$$

So, for β sufficiently close to 1, we have the following expansion for $v_\beta(f)$,

$$v_\beta(f) = \sum_{k=-1}^{\infty} (1-\beta)^k c_k(f) , \quad (5.40)$$

with

$$c_{-1}(f) = g(f) ,$$

$$c_0(f) = (I - Q)^{-1} r(f) - g(f) ,$$

$$c_k(f) = (-1)^k[(I - Q)^{-1}Q]^k (I - Q)^{-1} r(f) , \quad k = 1,2,\ldots .$$

It is notationally convenient when comparing stationary strategies for discountfactors close enough to 1 to use the following two partial orderings on F:

$$(5.41) \qquad f \succcurlyeq h \iff \begin{cases} \text{For all } i \in S \\ c_{-1}(i,f) \geq c_{-1}(i,h) \\ \text{and for all } k = 0,1,\ldots \\ c_\ell(i,f) = c_\ell(i,h) , \ \ell = -1,\ldots,k-1 \Rightarrow c_k(i,f) \geq c_k(i,h) , \end{cases}$$

$$(5.42) \qquad f \succ h \iff f \succcurlyeq h \text{ and not } h \succcurlyeq f .$$

Then we have

LEMMA 5.14. *For any two policies* f *and* h *there exists a constant* $\beta(f,h)$, *with* $0 \leq \beta(f,h) < 1$, *such that for all* β *with* $\beta(f,h) \leq \beta < 1$

$$f \succcurlyeq h \iff v_\beta(f) \geq v_\beta(h) .$$

PROOF. Immediately from

$$v_\beta(f) - v_\beta(h) = \sum_{k=-1}^{\infty} (1 - \beta)^k [c_k(f) - c_k(h)]$$

and the definition of $f \succcurlyeq h$. □

An immediate consequence of this lemma is

THEOREM 5.15 (cf. BLACKWELL [1962]). *There exists a constant* β_0, *with* $0 \leq \beta_0 < 1$, *such that for all* f *and* h *and all* $\beta_0 \leq \beta < 1$

$$f \succcurlyeq h \iff v_\beta(f) \geq v_\beta(h) .$$

Moreover there exists a stationary strategy f^* *such that for all* $f \in F$ *we have* $f^* \succcurlyeq f$.

PROOF. The first assertion follows from lemma 5.14 and the fact that (since S and A are finite) there are only finitely many policies. The second assertion follows along the same lines as the proof of theorem 2.21 in this section, namely, take f^* such that $v_{\beta_n}(f^*) = v^*_{\beta_n}$ for some sequence $\beta_n \uparrow 1$. □

We conclude this section with some results concerning Laurent series expansions for $L_\beta(f)v_\beta(h)$, in particular for $h = f^*$, where f^* is a stationary strategy that is optimal for all β sufficiently close to 1.
For all f and $h \in F$ we have

$$L_\beta(f)v_\beta(h) = r(f) + \beta P(f)v_\beta(h)$$

$$= r(f) + [P(f) - (1-\beta)P(f)] \sum_{k=-1}^{\infty} (1-\beta)^k c_k(h) .$$

This yields the following expansion for $L_\beta(f)v_\beta(h)$:

$$L_\beta(f)v_\beta(h) = \sum_{k=-1}^{\infty} (1-\beta)^k d_k(f,h) , \tag{5.43}$$

with

$$d_{-1}(f,h) = P(f)c_{-1}(h) , \tag{5.44}$$

$$d_0(f,h) = r(f) + P(f)c_0(h) - P(f)c_{-1}(h) , \tag{5.45}$$

$$d_k(f,h) = P(f)c_k(h) - P(f)c_{k-1}(h) , \quad k = 1,2,\ldots . \tag{5.46}$$

With $S = \{1,2,\ldots,N\}$ we can interpret $d(\cdot)$ and $c_k(\cdot)$ as column vectors in $\mathbb{R}^N$ and $P(f)$ as an $(N \times N)$-matrix.
For all f and $h \in F$ let us denote by $C(f)$ and $D(f,h)$ the $(N \times \infty)$-matrices (N is the number of states in S) with columns $c_k(f)$ and $d_k(f,h)$, $k = -1,0,1,\ldots$, respectively.
For equally sized matrices define the following two partial orderings:

$P \succcurlyeq Q \iff$ In each row of the matrix $P - Q$ the first nonzero element (if any) is positive. (5.47)

$P \succ Q \iff P \succcurlyeq Q$ and not $Q \succcurlyeq P$. (5.48)

So $f \succcurlyeq h$ [$f \succ h$] is equivalent to $C(f) \succcurlyeq C(h)$ [$C(f) \succ C(h)$].

Then we have the following theorem.

THEOREM 5.16

(i) *For all* $f \in F$

$$D(f,f) = C(f) .$$

(ii) *For all* f *and* $h \in F$

$$D(f,h) \succcurlyeq C(h) \Rightarrow f \succcurlyeq h .$$

(iii) *If* f^* *satisfies* $f^* \succcurlyeq f$ *for all* $f \in F$, *then*

$$D(f,f^*) \preccurlyeq C(f^*) \quad \textit{for all } f \in F .$$

PROOF.

(i) Follows immediately from

$$L_\beta(f)v_\beta(f) = v_\beta(f) .$$

(ii) If $D(f,h) \succcurlyeq C(h)$, then for all β sufficiently close to 1

$$L_\beta(f)v_\beta(h) \geq v_\beta(h) .$$

So, with the monotonicity of $L_\beta(f)$ and (5.32), for all β sufficiently close to 1

$$v_\beta(f) = \lim_{n\to\infty} L_\beta^n(f)v_\beta(h) \geq L_\beta(f)v_\beta(h) \geq v_\beta(h) .$$

Hence, by theorem 5.15, $f \succcurlyeq h$.

(iii) $f^* \succcurlyeq f$ for all $f \in F$ implies $v_\beta(f^*) = v_\beta^*$ for all β sufficiently close to 1. So

$$U_\beta v_\beta(f^*) = v_\beta(f^*) ,$$

and for all $f \in F$

$$L_\beta(f)v_\beta(f^*) \leq v_\beta(f^*) ,$$

for β close enough to 1. Hence

$$D(f,f^*) \preccurlyeq C(f^*) \quad \text{for all } f \in F . \qquad \square$$

5.6. SENSITIVE OPTIMALITY

In the literature various criteria of optimality have been introduced for the case that the discountfactor tends to 1.
BLACKWELL [1962] studied this problem, and he introduced the following two concepts of optimality.
He called a strategy π *nearly optimal* if

$$(5.49) \qquad v_\beta^* - v_\beta(\pi) \to 0 \quad (\beta \uparrow 1) \ ,$$

and a strategy π *optimal* if

$$(5.50) \qquad v_\beta^* = v_\beta(\pi) \quad \text{for all } \beta \text{ close enough to } 1.$$

(We shall use these concepts only in this section.)
VEINOTT [1969] introduced the following more sensitive optimality criteria.
A strategy $\hat{\pi}$ is called k-*discount optimal*, $k \in \{-1,0,1,\ldots\}$, if

$$(5.51) \qquad \liminf_{\beta \uparrow 1} (1-\beta)^{-k} [v_\beta(\hat{\pi}) - v_\beta(\pi)] \geq 0 \quad \text{for all } \pi \in \Pi \ .$$

Finally, a strategy is called ∞-*discount optimal* if it is k-discount optimal for all $k = -1,0,1,\ldots$.
Clearly, a nearly optimal strategy in the sense of (5.49) is 0-discount optimal. Substituting for π in (5.51) a strategy f^* satisfying $v_\beta(f^*) = v_\beta^*$ for β sufficiently close to 1, we see that a 0-discount optimal strategy is nearly optimal in the sense of (5.49). So these two concepts are equivalent.
Further we see that optimality in the sense of (5.50) is equivalent to ∞-discount optimality.
In chapter 7 it will be shown that there is a close relationship between k-discount optimality and more sensitive optimality criteria in the average-reward case (cf. SLADKY [1974]).
The relation between the discounted MDP when the discountfactor tends to 1 and the average-reward MDP, and in particular the policy iteration method for the average-reward case, has been studied in various publications.
BLACKWELL [1962] showed that Howard's policy iteration method for the average reward MDP [HOWARD, 1960] yields, under certain conditions, a nearly optimal policy. VEINOTT [1966] extended Howard's method in such a way that it always produces a nearly optimal stationary strategy. A further extension of the policy iteration method by MILLER and VEINOTT [1969] yields k-discount optimal policies for all $k = -1,0,\ldots,\infty$.

In chapter 8 we use the concept of go-ahead functions to derive variants of the policy iteration method that also yield k-discount optimal stationary strategies.

CHAPTER 6

INTRODUCTION TO THE AVERAGE-REWARD MDP

In the chapters 6 - 9 we consider the average-reward MDP. Throughout these four chapters both the state space and the action space are assumed to be finite, and the states will be labeled 1,2,...,N, so $S = \{1,2,...,N\}$. Further, condition 1.1 no longer holds.
This chapter serves as an introduction to the average-reward MDP and reviews some results on these processes. In particular, results on the existence of optimal stationary strategies (section 1), on the policy iteration method (section 2), and on the method of standard successive approximations (section 3).

6.1. OPTIMAL STATIONARY STRATEGIES

In this section it will be shown that an optimal stationary strategy exists for the average reward per unit time criterion. Namely, a (the) strategy f^* that satisfies $f^* \geq f$ for all $f \in F$ (for the existence of such a policy f^* see theorem 5.15). (For the average optimality of a policy h the condition $h \geq f$ for all f is, however, not necessary.)
Recall that the average reward per unit time g for a strategy $\pi \in \Pi$ has been defined by (see (1.12))

$$(6.1) \qquad g(\pi) = \liminf_{n\to\infty} n^{-1} v_n(\pi) = \liminf_{n\to\infty} n^{-1} \mathbb{E}_\pi \sum_{k=0}^{n-1} r(X_k, A_k) \ .$$

For a stationary strategy $f \in F$ we have (cf. (5.35))

$$(6.2) \qquad g(f) = P^*(f) r(f) \ ,$$

where (cf. (5.34))

$$(6.3) \qquad P^*(f) = \lim_{n\to\infty} n^{-1} \sum_{k=0}^{n-1} P^k(f) \ .$$

We want to show that

$$(6.4) \qquad g^* := \sup_{\pi\in\Pi} g(\pi) = \max_{f\in F} g(f) \ .$$

In order to show that any policy f^* satisfying $f^* \succcurlyeq f$ for all $f \in F$ (cf. (5.41) and theorem 5.15) is average optimal we need the following lemma.

LEMMA 6.1 (cf. BROWN [1965]). *Let f^* be a policy satisfying $f^* \succcurlyeq f$ for all $f \in F$, then for all sufficiently large $K \in \mathbb{R}$*

$$(6.5) \qquad L(f)[Kg(f^*) + c_0(f^*)] \le L(f^*)[Kg(f^*) + c_0(f^*)]$$

$$= U[Kg(f^*) + c_0(f^*)] = (K+1)g(f^*) + c_0(f^*) \ .$$

PROOF. Let f be an arbitrary policy, then we have from theorem 5.16 and (5.44) and (5.45):
For all $i \in S$

$$(P(f)g(f^*))(i) \le (P(f^*)g(f^*))(i) = g(i,f^*)$$

and if

$$(P(f)g(f^*))(i) = g(i,f^*) \ ,$$

then

$$r(i,f(i)) + (P(f)c_0(f^*))(i) - g(i,f^*)$$

$$\le r(i,f^*(i)) + (P(f^*)c_0(f^*))(i) - g(i,f^*) = c_0(i,f^*) \ .$$

So, for all K sufficiently large,

$$L(f)[Kg(f^*) + c_0(f^*)] = KP(f)g(f^*) + r(f) + P(f)c_0(f^*)$$

$$\le KP(f^*)g(f^*) + r(f^*) + P(f^*)c_0(f^*)$$

$$= L(f^*)[Kg(f^*) + c_0(f^*)] = (K+1)g(f^*) + c_0(f^*) \ . \qquad \square$$

With this lemma we can prove the following well-known result.

THEOREM 6.2. *Let f^* be a policy satisfying $f^* \succcurlyeq f$ for all $f \in F$ (such a policy exists by theorem 5.15). Then*

$$g(f^*) = g^* \ (= \sup_{\pi\in\Pi} g(\pi)) \ .$$

PROOF. Let π be an arbitrary strategy, and let K_0 be a constant such that (6.5) holds for all $K \geq K_0$. Then

$$
\begin{aligned}
(6.6) \qquad g(\pi) &= \liminf_{n\to\infty} n^{-1} v_n(\pi) \leq \liminf_{n\to\infty} n^{-1} U^n 0 \\
&= \liminf_{n\to\infty} n^{-1} U^n [K_0 g(f^*) + c_0(f^*)] \\
&= \liminf_{n\to\infty} n^{-1} [(K_0 + n) g(f^*) + c_0(f^*)] = g(f^*) .
\end{aligned}
$$

Hence

$$
g^* = \sup_{\pi\in\Pi} g(\pi) \leq g(f^*) .
$$

Clearly, $g^* \geq g(f^*)$, so the proof is complete. □

Note that (6.6) also holds if liminf is replaced by limsup (apart from the first equation). So f^* remains optimal if we use the maximality of

$$
\limsup_{n\to\infty} n^{-1} v_n(\pi)
$$

as a criterion.

So we see from theorem 6.2 that, when we are looking for an optimal or nearly-optimal strategy, we can restrict ourselves to stationary strategies. This is done in the policy iteration algorithm.

6.2. *THE POLICY ITERATION METHOD*

Before formulating the policy iteration method we give the following characterization of $g(f)$ and $c_0(f)$.

LEMMA 4.3 (BLACKWELL [1962]). *The system of linear equations in* g *and* v, $g, v \in V$,

$$
\begin{aligned}
&(i) \quad P(f)g = g \\
(6.7) \qquad &(ii) \quad L(f)v = v + g \\
&(iii) \quad P^*(f)v = 0
\end{aligned}
$$

has the unique solution $g = g(f)$, $v = c_0(f)$.

PROOF. First we show that $(g(f),c_0(f))$ solves (6.7). That $g(f)$ and $c_0(f)$ satisfy (i) and (ii) follows from (5.44) and (5.45) with $f = h$ and by theorem 5.16(i).

To prove $P^*(f)c_0(f) = 0$, premultiply $v_\beta(f)$ with $P^*(f)$, which yields

$$P^*(f)v_\beta(f) = P^*(f)\sum_{k=0}^{\infty}\beta^k P^k(f)r(f) = \sum_{k=0}^{\infty}\beta^k P^*(f)P^k(f)r(f)$$

$$= \sum_{k=0}^{\infty}\beta^k P^*(f)r(f) = (1-\beta)^{-1}\, g(f) \ ,$$

where we used (5.36). Also

$$P^*(f)v_\beta(f) = P^*(f)[(1-\beta)^{-1}\, g(f) + c_0(f) + \mathcal{O}(1-\beta)]$$

$$= (1-\beta)^{-1}\, g(f) + P^*(f)c_0(f) + \mathcal{O}(1-\beta) \quad (\beta \uparrow 1) \ .$$

So

$$P^*(f)c_0(f) = 0 \ .$$

To prove the uniqueness of the solution $(g(f),c_0(f))$, let us assume that (g^0,v^0) and (g^1,v^1) both solve (6.7). Iterating and averaging (i) we get

$$P^*(f)g^0 = g^0 \quad \text{and} \quad P^*(f)g^1 = g^1 \ .$$

And premultiplying (ii) by $P^*(f)$ we obtain

$$P^*(f)r(f) = P^*(f)g^0 = P^*(f)g^1 \ .$$

So (with (6.2))

$$g^0 = g^1 = g(f) \ .$$

To prove $v^0 = v^1$, subtract $L(f)v^1 = v^1 + g(f)$ from $L(f)v^0 = v^0 + g(f)$ to obtain

$$P(f)(v^0 - v^1) = (v^0 - v^1) \ .$$

Iterating and averaging this equality yields

$$v^0 - v^1 = P^*(f)(v^0 - v^1) \ .$$

But from (iii) we have

$$P^*(f)v^0 = P^*(f)v^1 = 0 \ .$$

Hence $v^0 = v^1$, which proves that the solution of (6.7) is unique. □

In the sequel we often write v(f) instead of $c_0(f)$.

Now let us formulate Howard's policy iteration algorithm for the average reward case [HOWARD, 1960] with the modification due to BLACKWELL [1962] that guarantees convergence.

Policy iteration algorithm

Choose $f \in F$.

Value determination step

Determine the unique solution (g(f),v(f)) of (6.7).

Policy improvement step

Determine for each $i \in S$ the set

$$A(i,f) := \{a \in A \mid \sum_{j \in S} p(i,a,j)g(j,f) = \max_{a_0 \in A} \sum_{j \in S} p(i,a_0,j)g(j,f)\}$$

and subsequently

$$B(i,f) := \{a \in A(i,f) \mid r(i,a) + \sum_{j \in S} p(i,a,j)v(j,f) =$$

$$= \max_{a_0 \in A(i,f)} \{r(i,a_0) + \sum_{j \in S} p(i,a_0,j)v(j,f)\}\} .$$

Replace policy f by a policy h with $h(i) \in B(i,f)$ and $h(i) = f(i)$ if $f(i) \in B(i,f)$ for all $i \in S$, and return to the value determination step. Repeat until the policy f cannot be improved anymore, i.e., until $f(i) \in B(i,f)$ for all $i \in S$.

For the policy iteration method we have the following convergence result.

THEOREM 6.4 (see BLACKWELL [1962, theorem 4]). *Let policy* h *be an improvement of the policy* f *obtained by the policy improvement step of the policy iteration algorithm, then*

(i) *if* $h = f$, *then* $g(f) = g^*$,
(ii) *if* $h \neq f$, *then* $h \gneqq f$.

From (i) *and* (ii) *it follows, as* F *is finite by the finiteness of* S *and* A, *that the policy iteration method converges, i.e., it yields an average optimal policy after finitely many iterations.*

PROOF. For a proof see BLACKWELL [1962]. We don't give the proof here, since theorem 6.4 is merely a special case of theorem 8.7 which we prove in chapter 8.

VEINOTT [1966] and MILLER and VEINOTT [1969] have shown that the policy iteration method can be extended in such a way that the algorithm terminates with a policy which not only maximizes g(f) but also some (or all) subsequent terms of the Laurent series expansion for $v_\beta(f)$.

HASTINGS [1968] introduced a modified version of the policy iteration for the case that all P(f) are irreducible. (P(f) is irreducible if for each pair $i,j \in S$ there exists a number n such that $(P^n(f))(i,j) > 0$.) In that case $P^*(f)$ will have equal rows and g(f) will be independent of the initial state, so $A(i,f) = A(i)$ for all $i \in S$.
Hastings showed that the standard successive approximation step in the definition of B(i,f) can be replaced by a Gauss-Seidel step.

In chapter 8 the concept of go-ahead functions is used to study this and other variants of the (standard) policy iteration method, as well as several variants of the extended versions of this method as formulated by VEINOTT [1966] and MILLER and VEINOTT [1969]. It will also be shown that these algorithms converge (not only if P(f) is irreducible), and that the extended versions again yield more sensitive optimal strategies.

Closely related to the policy iteration method are the linear programming formulations. After d'EPENOUX [1960] introduced linear programming for the discounted MDP, De GHELLINCK [1960] and MANNE [1960], independently, gave the linear programming formulation for the average-reward criterion in the unichain case. (The case that for each policy f the underlying Markov chain has one recurrent subchain and possibly some transient states.) The multichain case has been attacked a.o. by DENARDO and FOX [1968], DENARDO [1970] and DERMAN [1970]. Recently their results have been improved considerably by HORDIJK and KALLENBERG [1979].

6.3. SUCCESSIVE APPROXIMATIONS

Another method to determine optimal or nearly-optimal policies, is the method of standard successive approximations:

$$\begin{cases} \text{Choose } v_0. \\ \text{Determine for } n = 0,1,\ldots \\ \qquad v_{n+1} = Uv_n \ . \end{cases}$$

From lemma 6.1 we immediately have the following result to to BROWN [1965].

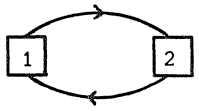
THEOREM 6.5. *$v_n - ng^*$ is bounded in n for all $v_0 \in V$.*

PROOF. Let K_0 be so large that (6.5) holds for all $K \geq K_0$. Then using the finiteness of S, we have for all n = 1,2,...

$$\begin{aligned} \|v_n - ng^*\|_e &\leq \|U^n[K_0 g^* + c_0(f^*)] - ng^*\|_e + \|\tilde{U}^n[v_0 - K_0 g^* - c_0(f^*)]\|_e \\ &\leq \|K_0 g^* - c_0(f^*)\|_e + \|v_0 - K_0 g^* - c_0(f^*)\|_e < \infty \ . \end{aligned}$$ □

In general, however, $v_n - ng^*$ need not converge.

EXAMPLE 6.6. S := {1,2}, A = {1}, r(1,1) = 2, r(2,1) = 0, p(1,1,2) = p(2,1,1) = 1.



For this MDP clearly $g^* = (1,1)^T$, but $U^n 0 - ng^*$ oscillates between $(1,-1)^T$ and 0.

Further, if $v_n - ng^*$ does not converge, then it need not be that for sufficiently large n a policy f_n satisfying $L(f_n)v_n = v_{n+1}$ is average optimal. This is shown by an example of LANERY [1967].
In case of convergence, however, we have the following result.

THEOREM 6.7. *Let $v_n - ng^*$ converge. Then, if n is sufficiently large, a policy f_n satisfying $L(f_n)v_n = v_n$ is average optimal.*

PROOF. Define

$$\hat{v} := \lim_{n\to\infty} [v_n - ng^*] \ .$$

Then

$$L(f_n)v_n = L(f_n)[ng^* + \hat{v} + o(1)] = v_{n+1} = (n+1)g^* + \hat{v} + o(1) \quad (n \to \infty) .$$

Hence, since F is finite, we have for n sufficiently large

(6.8) $$P(f_n)g^* = g^*$$

and

(6.9) $$L(f_n)\hat{v} = \hat{v} + g^* .$$

Iterating and averaging (6.8), we get $P^*(f_n)g^* = g^*$.
So, premultiplication of (6.9) by $P^*(f_n)$ yields

$$P^*(f_n)r(f_n) = g(f_n) = P^*(f_n)g^* = g^* . \qquad \square$$

WHITE [1963] has shown that $v_n - ng^*$ converges if there exists a specific state $i_0 \in S$ and an integer r such that

(6.10) $$P(f_1) \cdots P(f_r)(i,i_0) > 0 ,$$

for all policies $f_1,\ldots,f_r$ and all $i \in S$.
DENARDO [1973] proved convergence of $v_n - ng^*$ under the weaker hypothesis that all P(f) are unichained (one recurrent class and possibly some transient states) and aperiodic. Note that the matrix in example 6.6 is periodic.
The general multichained case with periodicities has been studied by BROWN [1965] and LANERY [1967]. Finally, a relatively complete treatment has been given by SCHWEITZER and FEDERGRUEN [1978, 1979]. The latter two authors established e.g. that $v_n - ng^*$ converges if all P(f) are aperiodic (even under weaker conditions) and that there always exists an integer J, the "essential period of the MDP", such that

$$U^{nJ+m} v_0 - nJg^* \text{ converges for all } m = 0,1,\ldots,J-1 .$$

The latter result (with incorrect proofs) was also given by Brown and by Lanery.

Periodicity, however, need not be a problem. SCHWEITZER [1971] has given a data transformation which transforms any MDP into an equivalent MDP that is aperiodic.

Aperiodicity transformation

Let the MDP be characterized by S, A, p and r. Construct a new MDP with S, A, $\hat{p}$ and $\hat{r}$ as follows:

Choose $\alpha \in (0,1)$ and define

$$(6.11)\quad \begin{cases} \hat{r}(i,a) = (1-\alpha)r(i,a) , & i \in S,\ a \in A , \\ \hat{p}(i,a,i) = \alpha + (1-\alpha)p(i,a,i) , & i \in S,\ a \in A , \\ \hat{p}(i,a,j) = (1-\alpha)p(i,a,j) , & i,j \in S,\ j \neq i \text{ and } a \in A . \end{cases}$$

We will show that the two MDP's are indeed equivalent. Denote all objects in the transformed MDP by a $\hat{}$. Then for all $f \in F$

$$\hat{P}(f) = \alpha I + (1-\alpha)P(f) ,$$

so, clearly, $\hat{P}(f)$ is aperiodic for all $f \in F$, and

$$\hat{r}(f) = (1-\alpha)r(f) .$$

One easily verifies that

$$(6.12)\quad \hat{P}(f)g(f) = g(f)$$

and

$$(6.13)\quad \hat{r}(f) + \hat{P}(f)v(f) = v(f) + (1-\alpha)g(f) .$$

Further we have

$$\hat{P}(f)P^*(f) = \alpha P^*(f) + (1-\alpha)P(f)P^*(f) = P^*(f) ,$$

so also

$$\hat{P}^*(f)P^*(f) = P^*(f) .$$

And

$$\hat{P}^*(f)P(f) = \hat{P}^*(f)[1-\alpha)^{-1}\hat{P}(f) - \alpha I] = \hat{P}^*(f) ,$$

hence

$$\hat{P}^*(f)P^*(f) = \hat{P}^*(f) .$$

This implies

$$\hat{P}^*(f) = P^*(f)$$

thus

$$(6.14)\quad \hat{P}^*(f)v(f) = 0 .$$

So it follows from (6.12)-(6.14) and lemma 4.3 that the transformed MDP is equivalent to the original ∞-horizon MDP, with

$$(\hat{g}(f),\hat{v}(f)) = ((1-\alpha)g(f),v(f)) \ .$$

The finite horizon MDP's, however, are different.

So from now on it may be assumed that the MDP under consideration is aperiodic in the strong sense of (6.11), i.e., all $p(i,a,i)$ are strictly positive. And thus that $v_n - ng^*$ converges (which clearly implies $v_{n+1} - v_n \to g^*$). So theorem 6.7 applies. However, in order to obtain an appropriate algorithm, one has to be able to verify whether n is already so large that $v_{n+1} - v_n$ is close to g^* and that f_n is nearly optimal.
If $g^*(i)$ is independent of the initial state, which, for example, is the case if all $P(f)$ are unichained, then the following lemma makes it possible to recognize near-optimality.

LEMMA 6.8 (cf. HASTINGS [1968] and HORDIJK and TIJMS [1975]). *Let $v \in V$ be arbitrary and let f be a policy satisfying*

$$L(f)v = Uv \ .$$

Then

$$\min_{i\in S} (Uv-v)(i)e \le g(f) \le g^* \le \max_{i\in S} (Uv-v)(i)e \ .$$

PROOF. For all $h \in F$ we have

$$P^*(h)(L(h)v - v) = P^*(h)r(h) = g(h) \ . \tag{6.15}$$

So, with $h = f$,

$$g(f) = P^*(f)(L(f)v - v) = P^*(f)(Uv - v)$$

$$\ge P^*(f) \min_{i\in S} (Uv-v)(i)e = \min_{i\in S} (Uv-v)(i)e \ .$$

Clearly, $g(f) \le g^*$, and applying (6.15) with $h = f^*$ $(g(f^*) = g^*)$ we obtain

$$g^* = g(f^*) = P^*(f^*)(L(f^*)v - v) \le P^*(f^*)(Uv - v)$$

$$\le P^*(f) \max_{i\in S} (Uv-v)(i)e = \max_{i\in S} (Uv-v)(i)e \ . \qquad \square$$

If g^* is constant and $v_n - ng^*$ converges, then $Uv_n - v_n$ converges to g^*, so

$$\max_{i\in S} (Uv_n - v_n)(i) - \min_{i\in S} (Uv_n - v_n)(i) \to 0 \quad (n \to \infty) .$$

So in this case lemma 6.8 shows us that the method of standard successive approximations yields (arbitrarily close) bounds on g^* and nearly-optimal stationary strategies.
It is also clear that lemma 6.8 is not of much help if g^* is not constant.

One may also try to use the method of value-oriented successive approximations. For the average-reward case this method has been proposed by MORTON [1971], who, however, does not give a convergence proof.
In chapter 9 we study the value-oriented method under the so-called strong aperiodicity assumption that $P(f) \geq \alpha I$ for some $\alpha > 0$ and all f (cf. (6.11)), and under various conditions concerning the chain structure of the MDP, all guaranteeing that g^* is constant.

Another variant of the method of standard successive approximations has been introduced by BATHER [1973] and by HORDIJK and TIJMS [1975]. This method approximates the average-reward MDP by a sequence of discounted MDP's with discountfactor tending to 1.

(6.16)
Choose $v_0 \in V$.
Determine for $n = 0,1,\ldots$

$$v_{n+1} = U_{\beta_n} v_n ,$$

where $\{\beta_n\}$ is a sequence of discount factors tending to 1.

HORDIJK and TIJMS proved, that if g^* is constant, $v_{n+1} - v_n$ converges to g^* if the sequence $\{\beta_n\}$ satisfies the following two conditions:

$$\beta_1\beta_2 \cdots \beta_n \to 0 \quad (n \to \infty) ,$$

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

A possible choice for $\{\beta_n\}$ is $\beta_n = 1 - n^{-b}$, $0 < b \leq 1$, $n = 1,2,\ldots$. The convergence, however, is rather slow, namely of order $n^{-b} \ln n$. BATHER [1973] has considered the special case $\beta_n = 1 - n^{-1}$.

In chapter 7 we introduce a nonstationary variant of the method of standard successive approximations to study the relation between the more sensitive optimality criteria in the average-reward case and the discounted case. This nonstationary method turns out to be equivalent to the method of Hordijk and Tijms for sequences $\beta_n = 1 - \frac{k}{n}$. From our analysis it will follow that for these sequences the method (6.16) also converges if $g^*(i)$ depends on the initial state.

CHAPTER 7

SENSITIVE OPTIMALITY

7.1. INTRODUCTION

In this chapter we consider some more sensitive optimality criteria for the average-reward MDP with finite state space $S = \{1,2,\ldots,N\}$ and finite action space.
The criterion of average reward per unit time is often rather unsatisfactory, since the criterion value depends only on the tail of the income stream, and not on the rewards during the first, say 1000, periods.
In order to overcome this problem, one may consider more sensitive optimality criteria.
One of these is the criterion of average overtaking optimality introduced by VEINOTT [1966].

DEFINITION 7.1. *A strategy $\hat{\pi} \in \Pi$ is called average overtaking optimal, if for all $\pi \in \Pi$*

$$(7.1) \qquad \liminf_{n\to\infty} n^{-1} \sum_{m=1}^{n} [v_m(\hat{\pi}) - v_m(\pi)] \geq 0 \ .$$

Veinott proved that an average overtaking optimal oplicy is nearly optimal in the sense of Blackwell, formula (5.49), and therefore also 0-discount optimal in the sense of (5.51). Veinott conjectured the reverse to be true as well. This conjecture was proved to be correct by DENARDO and MILLER [1968]. LIPPMAN [1968] proved that average overtaking optimality and 0-discount optimality are equivalent (not only for stationary strategies).
A stronger criterion than (7.1) is the following, introduced by DENARDO and ROTHBLUM [1979].

DEFINITION 7.2. *A strategy* $\hat{\pi} \in \Pi$ *is called overtaking optimal if for all* $\pi \in \Pi$

$$\liminf_{n\to\infty} [v_n(\hat{\pi}) - v_n(\pi)] \geq 0 \ .$$

In general, there need not exist an overtaking optimal policy, since for two average overtaking optimal strategies $\pi^{(1)}$ and $\pi^{(2)}$, the difference $v_n(\pi^{(1)}) - v_n(\pi^{(2)})$ may oscillate around 0. BROWN [1965] gives an example where this oscillation is not caused by the periodicity of the transition matrices. Denardo and Rothblum proved that under certain conditions an overtaking optimal strategy does exist.
An extension of the concept of average overtaking optimality has been given by SLADKY [1974].
Define for n = 0,1,...

$$(7.2) \quad \begin{cases} v_n^{(0)}(\pi) := v_n(\pi) \ , \\ v_n^{(k)}(\pi) := \sum_{\ell=0}^{n-1} v_\ell^{(k-1)}(\pi) \ , \quad k = 1,2,\ldots \ . \end{cases}$$

Then

$$(7.3) \quad v_n^{(k)}(\pi) = \sum_{\ell_k=0}^{n-1} \sum_{\ell_{k-1}=0}^{\ell_k-1} \cdots \sum_{\ell_1=0}^{\ell_2-1} v_{\ell_1}^{(0)}(\pi) = \sum_{\ell=0}^{n-k} \binom{n-\ell-1}{k-1} v_\ell^{(0)}(\pi) \ .$$

DEFINITION 7.3 (SLADKY [1974]). *A strategy* $\hat{\pi} \in \Pi$ *is called* k*-order average optimal, if for all* $\pi \in \Pi$

$$\liminf_{n\to\infty} n^{-1}[v_n^{(k)}(\hat{\pi}) - v_n^{(k)}(\pi)] \geq 0 \ .$$

So, a 0-order average-optimal strategy is average optimal and a 1-order average-optimal strategy is average overtaking optimal. Sladky has shown that a strategy π is k-order average-optimal if and only if it is (k - 1)-discount optimal.
Here (in section 2) we will prove this result for stationary strategies following a somewhat different line of reasoning. The case of arbitrary strategies is notationally more complicated. As a byproduct of our approach we obtain a successive approximations algorithm yielding k-order average optimal policies; the problem to recognize these policies, however, remains.

In section 3 we obtain a relation between this algorithm and the algorithms by BATHER [1973] and HORDIJK and TIJMS [1975] as formulated in (6.16).

7.2. THE EQUIVALENCE OF k-ORDER AVERAGE OPTIMALITY AND (k - 1)-DISCOUNT OPTIMALITY

In this section we show that a policy is k-order average optimal if and only if it is (k - 1)-discount optimal. Part of the results in this section can be found in Van der WAL and ZIJM [1979].
In order to prove this we study the following dynamic programming scheme

$$(7.4)\qquad \begin{cases} v_0^{(k)} := 0 , \\ v_{n+1}^{(k)} := \max\limits_{f\in F} \{ \binom{n}{k} r(f) + P(f) v_n^{(k)} \} , \quad n = 0,1,\ldots , \end{cases}$$

where $\binom{n}{k} := 0$ if $k > n$.
The reason why we study this scheme will become clear from the following analysis.
Let $\pi = (f_0,f_1,\ldots)$ be an arbitrary Markov strategy, and let $v_n^{(k)}(\pi)$ be defined as in (7.2). Then, by definition

$$v_0^{(0)}(\pi) = 0 ,$$

from which we obtain with (7.2) inductively,

$$v_n^{(k)}(\pi) = 0 \quad \text{for all } n \le k , \ n,k = 0,1,\ldots .$$

Further we have for all $n \ge k$ the following recursion

$$(7.5)\qquad v_{n+1}^{(k)}(\pi) = \sum_{\ell=0}^{n} v_\ell^{(k-1)}(\pi) = \sum_{\ell=0}^{n-1} v_\ell^{(k-1)}(\pi) + v_n^{(k-1)}(\pi)$$

$$= v_n^{(k)}(\pi) + v_n^{(k-1)}(\pi) .$$

From (7.5) we can obtain the following lemma, which gives a recursion similar to (7.4) for an arbitrary strategy.

LEMMA 7.4. *Let* $\pi = (f_0,f_1,\ldots)$ *be an arbitrary Markov strategy, then for all* $n,k = 0,1,\ldots$

$$(7.6)\qquad v_{n+1}^{(k)}(\pi) = \binom{n}{k} r(f_0) + P(f_0) v_n^{(k)}(\pi^{\leftarrow 1}) \ ,$$

with $\pi^{\leftarrow 1} = (f_1, f_2, \ldots)$.

PROOF. With $\binom{\ell}{m} = 0$ for all $\ell < m$, we see that (7.6) holds for all points (n,k) with $n < k$. Clearly, (7.6) also holds for $k = 0$, since in that case (7.6) reduces to

$$v_{n+1}(\pi) = r(f_0) + P(f_0) v_n(\pi^{\leftarrow 1}) \ .$$

We will prove that (7.6) holds for all $n,k \geq 0$ by induction on n and k simultaneously.
Assume that (7.6) holds for the pairs (n_0-1,k_0) and (n_0-1,k_0-1), then we have with (7.5)

$$v_{n_0+1}^{(k_0)}(\pi) = v_{n_0}^{(k_0)}(\pi) + v_{n_0}^{(k_0-1)}(\pi)$$

$$= \binom{n_0-1}{k_0} r(f_0) + P(f_0) v_{n_0-1}^{(k_0)}(\pi^{\leftarrow 1}) + \binom{n_0-1}{k_0-1} r(f_0) + P(f\ ) v_{n_0-1}^{(k_0-1)}(\pi^{\leftarrow 1}) \ .$$

Applying (7.5) with π replaced by $\pi^{\leftarrow 1}$ we obtain using

$$(7.7)\qquad \binom{\ell}{m} = \binom{\ell-1}{m} + \binom{\ell-1}{m-1} \quad \text{for all } \ell,m = 1,2,\ldots$$

that

$$v_{n_0+1}^{(k_0)}(\pi) = \binom{n_0}{k_0} r(f_0) + P(f_0) v_{n_0}^{(k_0)}(\pi^{\leftarrow 1}) \ .$$

So, (7.6) holds for (n_0,k_0). As (7.6) holds for all $n < k$ and also for $k = 0$, it follows by induction that (7.6) holds for all $n,k = 0,1,\ldots$. □

For a stationary strategy this yields

$$(7.8)\qquad v_{n+1}^{(k)}(f) = \binom{n}{k} r(f) + P(f) v_n^{(k)}(f) \ , \quad f \in F \ .$$

The similarity with the scheme (7.4) is clear. Before we study this scheme, we first analyze the recursion (7.8) in somewhat more detail.
To this end define for all $f \in F$ (cf. (5.40))

$$(7.9)\qquad D_n^{(k)}(f) := \binom{n}{k+1} g(f) + \binom{n-1}{k} c_0(f) + \ldots + \binom{n-k-1}{0} c_k(f) \quad \text{if } n > k$$

and

$$D_n^{(k)}(f) := 0 \quad \text{if } n \le k .$$

We will show that $v_n^{(k)}(f) - D_n^{(k)}(f)$ is bounded if n tends to infinity (for fixed k and f). To prove this we need the following lemma.

LEMMA 7.5. *For all* $n \ge k$ *and all* $f \in F$

$$\binom{n}{k} r(f) + P(f) D_n^{(k)}(f) = D_{n+1}^{(k)}(f) .$$

PROOF. For all $n \ge k$ and all $f \in F$ we have, with (7.7),

$$\begin{aligned}\binom{n}{k} r(f) + P(f) D_n^{(k)}(f) = \binom{n}{k+1} P(f) g(f) + \binom{n}{k}[r(f) + P(f) c_0(f)] + \\ + \binom{n-1}{k-1}[P(f) c_1(f) - P(f) c_0(f)] + \\ + \ldots + \binom{n-k+1}{1}[P(f) c_{k-1}(f) - P(f) c_{k-2}(f)] + \\ + \binom{n-k}{0}[P(f) c_k(f) - P(f) c_{k-1}(f)] + \\ + \left[\binom{n-k-1}{0} - \binom{n-k}{0}\right] P(f) c_k(f) .\end{aligned}$$

Hence, with (5.44)-(5.46) for h = f, theorem 5.16(i), and $\binom{n-k-1}{0} - \binom{n-k}{0} = 0$,

$$\begin{aligned}\binom{n}{k} r(f) + P(f) D_n^{(k)}(f) = \binom{n}{k+1} g(f) + \binom{n}{k}[g(f) + c_0(f)] + \\ + \binom{n-1}{k-1} c_1(f) + \ldots + \binom{n-k}{0} c_k(f) = D_{n+1}^{(k)}(f) ,\end{aligned}$$

where we used (7.7) once more with $(\ell,m) = (n+1,k+1)$. □

Now we can prove

THEOREM 7.6. *For all* $k = 0,1,\ldots$ *and* $f \in F$

$$v_n^{(k)}(f) = D_n^{(k)}(f) + O(1) \quad (n \to \infty) . \tag{7.10}$$

PROOF. For all $n > k$ and all $f \in F$,

$$\begin{aligned}v_{n+1}^{(k)}(f) - D_{n+1}^{(k)}(f) &= \binom{n}{k} r(f) + P(f) v_n^{(k)}(f) - \binom{n}{k} r(f) - P(f) D_n^{(k)}(f) \\ &= P(f)[v_n^{(k)}(f) - D_n^{(k)}(f)] .\end{aligned}$$

Hence

$$v_n^{(k)}(f) - D_n^{(k)}(f) = P^{n-k-1}(f)[v_{k+1}^{(k)}(f) - D_{k+1}^{(k)}(f)] = O(1) \quad (n \to \infty) .$$

Note, that if P(f) is aperiodic, then

$$v_n^{(k)}(f) - D_n^{(k)}(f) \quad \text{converges for } n \to \infty .$$

Theorem 7.6 enables us to compare stationary strategies for k-order average optimality. In order to consider also nonstationary strategies (as we have to according to definition 7.3), we consider the dynamic programming scheme (7.4).
For this scheme one can easily prove inductively that

$$v_n^{(k)} \geq v_n^{(k)}(\pi) \quad \text{for all } \pi \in M ,$$

and along similar lines as in section 2.3 one can then show

$$(7.11) \qquad v_n^{(k)} \geq v_n^{(k)}(\pi) \quad \text{for all } \pi \in \Pi .$$

To prove a similar asymptotic result as (7.10) for $v_n^{(k)}$ we need the following lemma.

LEMMA 7.7. *For each* $k = 0,1,\ldots$ *there exists an integer* $n_0 > k$ *such that for all* $n \geq n_0$

$$(7.12) \qquad \max_{f \in F} \{\binom{n}{k} r(f) + P(f) D_n^{(k)}(f^*)\} = D_{n+1}^{(k)}(f^*) ,$$

where f^* *is a policy satisfying* $f^* \succcurlyeq f$ *for all* $f \in F$ *(cf. theorem 5.15).*

PROOF. With (7.7) we get for $f \in F$ and $n > k$,

$$(7.13) \qquad \binom{n}{k} r(f) + P(f) D_n^{(k)}(f^*) = \binom{n}{k+1} P(f) g(f^*) + \binom{n}{k}[r(f) + P(f) c_0(f^*)] +$$

$$+ \binom{n-1}{k-1}[P(f) c_1(f^*) - P(f) c_0(f^*)] +$$

$$+ \ldots + \binom{n-k}{0}[P(f) c_k(f^*) - P(f) c_{k-1}(f^*)] .$$

Since $\binom{n}{k+1} = \frac{n-k}{k+1} \binom{n}{k}$ and for all ℓ and m, $\binom{\ell}{m} = \frac{\ell}{m} \binom{\ell-1}{m-1}$, we see that the subsequent terms on the right hand side in (7.13) decrease by an order n.

So, if n is sufficiently large, say $n \geq n_0$, then in order to maximize the left-hand side of (7.13) we can maximize separately the subsequent terms on the right-hand side. I.e., first maximize $P(f)g(f^*)$, next $\binom{n}{k}[r(f) + P(f)c_0(f^*)]$, etc. Then it follows with (5.44)-(5.46) for $h = f^*$ and theorem 5.16(iii) and (i) that (7.13) is maximal for $f = f^*$. Finally, (7.12) follows from lemma 7.5 with $f = f^*$. □

Now we can obtain the asymptotic behaviour of $v_n^{(k)}$.

THEOREM 7.8. *For all* $k = 0,1,\ldots$

$$v_n^{(k)} = D_n^{(k)}(f^*) + O(1) \quad (n \to \infty) ,$$

where f^* *is again a policy as mentioned in theorem 5.15.*

PROOF. From (7.11) and theorem 7.6 we have

$$v_n^{(k)} \geq D_n^{(k)}(f^*) + O(1) \quad (n \to \infty) .$$

So it suffices to prove

$$v_n^{(k)} \leq D_n^{(k)}(f^*) + O(1) \quad (n \to \infty) .$$

To prove this, define

$$\Delta_n^{(k)} := v_n^{(k)} - D_n^{(k)}(f^*) , \quad n > k .$$

Then we have for all $n \geq n_0$ (the constant mentioned in lemma 7.7),

$$\begin{aligned} v_{n+1}^{(k)} &= \max_{f \in F} \{\binom{n}{k} r(f) + P(f) D_n^{(k)}(f^*) + P(f)\Delta_n^{(k)}\} \\ &\leq \max_{f \in F} \{\binom{n}{k} r(f) + P(f) D_n^{(k)}(f^*)\} + \max_{f \in F} P(f)\Delta_n^{(k)} \\ &= D_{n+1}^{(k)}(f^*) + \tilde{U}\Delta_n^{(k)} . \end{aligned}$$

So,

$$\Delta_{n+1}^{(k)} \leq \tilde{U}\Delta_n^{(k)} .$$

Hence,

$$v_n^{(k)} - D_n^{(k)}(f^*) \leq \tilde{U}^{n-n_0}\Delta_{n_0}^{(k)} = O(1) \quad (n \to \infty) ,$$

which completes the proof. □

Finally, we can prove

THEOREM 7.9. *A policy* f *is* $(k-1)$*-discount optimal if and only if* f *is* k*-order average optimal.*

PROOF.

(i) First we prove the 'if' part. Let f be k-order average optimal, then certainly

$$\liminf_{n\to\infty} n^{-1}[v_n^{(k)}(f) - v_n^{(k)}(f^*)] \geq 0 .$$

So, with theorem 7.6,

$$\liminf_{n\to\infty} n^{-1}[D_n^{(k)}(f) - D_n^{(k)}(f^*)] \geq 0 .$$

Thus, in order that f is k-order optimal we certainly need $c_\ell(f) = c_\ell(f^*)$, $\ell = -1,\ldots,k-1$ (cf. (7.6)). Hence a k-order average optimal policy is also $(k-1)$-discount optimal.

(ii) To prove the 'only if' part, let f be a $(k-1)$-discount optimal policy. Then $c_\ell(f) = c_\ell(f^*)$, $\ell = -1,\ldots,k-1$. So

$$D_n^{(k)}(f) = D_n^{(k)}(f^*) + O(1) \quad (n \to \infty) .$$

Hence, for all $\pi \in \Pi$,

$$v_n^{(k)}(f) - v_n^{(k)}(\pi) \geq v_n^{(k)}(f) - v_n^{(k)} = D_n^{(k)}(f) - D_n^{(k)}(f^*) + O(1) =$$

$$= O(1) \quad (n \to \infty) .$$

Dividing by n we see that f is indeed k-order average optimal. □

As mentioned before, SLADKY [1974] has proved that theorem 7.9 holds for arbitrary strategies.

More or less as a byproduct of our analysis we have obtained the dynamic programming scheme (7.4). We end this section with some remarks about this scheme. We obtained

$$(7.14) \qquad v_n^{(k)} = \binom{n}{k+1} g(f^*) + \binom{n-1}{k} c_0(f^*) + \ldots + \binom{n-k}{1} c_{k-1}(f^*) + O(1) \quad (n \to \infty)$$

and one may even show that if e.g. all P(f) are aperiodic (cf. SCHWEITZER and FEDERGRUEN [1979]) the term $O(1)$ can be replaced by $w_k + O(\rho^n)$ for some w and some $\rho < 1$, i.e.,

$$(7.15)\qquad v_n^{(k)} = \binom{n}{k+1} g(f^*) + \ldots + \binom{n-k}{1} c_{k-1}(f^*) + w_k + O(\rho^n) \qquad (n \to \infty) .$$

From (7.14) we have

$$\frac{v_n^{(k)}}{\binom{n}{k+1}} = g(f^*) + O\left(\frac{1}{n}\right) \qquad (n \to \infty) .$$

From (7.15), however, we can obtain

$$\sum_{\ell=0}^{k+1} (-1)^\ell \binom{k+1}{\ell} v_{n-\ell}^{(k)} = g(f^*) + O(\rho^n) \qquad (n \to \infty)$$

and for example, if $k > 0$,

$$c_0(f^*) = \sum_{\ell=0}^{k} (-1)^\ell \binom{k}{\ell} v_{n-\ell}^{(k)} - (n-k) \sum_{\ell=0}^{k+1} (-1)^\ell \binom{k+1}{\ell} v_{n-\ell}^{(k)} + O(n\rho^n) \qquad (n \to \infty) .$$

Further, a policy f_n maximizing

$$\binom{n}{k} r(f) + P(f) v_n^{(k)}$$

will be $(k-2)$-discount optimal for n sufficiently large ($n \geq n_0$). This follows from the fact that (for $n \geq n_0$) policy f_n satisfies

$$\begin{cases} P(f_n) g(f^*) = g(f^*) \\ r(f_n) + P(f_n) c_0(f^*) = c_0(f^*) + g(f^*) \\ P(f_n) c_\ell(f^*) - P(f_n) c_{\ell-1}(f^*) = c_\ell(f^*) , \quad \ell = 1,\ldots,k-1 , \end{cases}$$

and the fact that the solution $(g, c_0, c_1, \ldots, c_{k-2})$ of the system

$$\begin{cases} P(f_n) g = g \\ r(f_n) + P(f_n) c_0 = c_0 + g \\ P(f_n) c_\ell - P(f_n) c_{\ell-1} = c_\ell , \quad \ell = 1,\ldots,k-2 \\ P^*(f_n) c_{k-2} = 0 \end{cases}$$

is unique and equal to $(g(f_n), c_0(f_n), \ldots, c_{k-2}(f_n))$. So $g(f_n) = g(f^*)$ and $c_\ell(f_n) = c_\ell(f^*)$ for $\ell = 0,1,\ldots,k-2$.
If, moreover, $v_n^{(k)} - D_n^{(k)}(f^*)$ converges for $n \to \infty$, then the stationary strategy f_n will be even $(k-1)$-discount optimal if n is sufficiently large.

The argument is similar as in theorem 6.7. However, we cannot simply verify whether for a specific (large) n the policy f_n is already ℓ-order average optimal (we encountered the same problem in the case of standard successive approximations if g^* is not constant).

7.3. EQUIVALENT SUCCESSIVE APPROXIMATIONS METHODS

In this section we want to show that the dynamic programming scheme (7.4) is equivalent with the successive approximations method of HORDIJK and TIJMS [1975] for a special choice of their sequence $\{\beta_n\}$.
In order to do so, define for fixed k and all n = k+1,k+2,...

$$\hat{v}_n^{(k)}(\pi) := v_n^{(k)}(\pi) \,/\, \binom{n-1}{k} \quad \text{for all } \pi \in \Pi ,$$

and

$$\hat{v}_n^{(k)} := v_n^{(k)} \,/\, \binom{n-1}{k} .$$

Then for all $\pi = (f_0,f_1,\ldots) \in M$, with $\pi^{+1} = (f_1,f_2,\ldots)$,

$$\begin{aligned} \hat{v}_{n+1}^{(k)}(\pi) &= r(f_0) + P(f_0) \frac{v_n^{(k)}(\pi^{+1})}{\binom{n}{k}} \\ &= r(f_0) + \frac{\binom{n-1}{k}}{\binom{n}{k}} P(f_0) \frac{v_n^{(k)}(\pi^{+1})}{\binom{n-1}{k}} \\ &= r(f_0) + \beta_n^{(k)} P(f_0) \hat{v}_n^{(k)}(\pi^{+1}) , \end{aligned}$$

where

$$\beta_n^{(k)} := \frac{\binom{n-1}{k}}{\binom{n}{k}} = 1 - \frac{k}{n} . \tag{7.16}$$

Similarly, we obtain

$$\hat{v}_{n+1}^{(k)} = \max_{f \in F} \{ r(f) + \beta_n^{(k)} \hat{v}_n^{(k)} \} .$$

So the dynamic programming scheme (7.4) is equivalent to the following algorithm:

(7.17)
$$\begin{cases} \text{Define } w_k^{(k)} := 0. \\ \text{Determine for } n = k, k+1, \ldots \\ w_{n+1}^{(k)} := \max_{f \in F} \{ r(f) + \beta_n^{(k)} w_n^{(k)} \} , \\ \text{with } \beta_n^{(k)} \text{ defined by (7.16).} \end{cases}$$

This is merely a special case of the set of algorithms introduced by HORDIJK and TIJMS [1975]. For k = 1, this is the method introduced by BATHER [1973]. Further we see from (7.14) that

$$w_n^{(k)} = \hat{v}_n^{(k)} = \frac{\binom{n}{k+1} g(f^*) + \binom{n-1}{k} c_0(f^*) + \ldots + \binom{n-k}{1} c_{k-1}(f^*) + O(1)}{\binom{n-1}{k}}$$

$$= \frac{n}{k+1} g^* + c_0(f^*) + O(\tfrac{1}{n}) \qquad (n \to \infty) .$$

So,

$$w_{n+1}^{(k)} - w_n^{(k)} = \frac{n+1}{k+1} g^* + c_0(f^*) - \frac{n}{k+1} g^* - c_0(f^*) + O(\tfrac{1}{n})$$

$$= \frac{1}{k+1} g^* + O(\tfrac{1}{n}) \qquad (n \to \infty) .$$

From the equivalence of the schemes (7.4) and (7.17) and the observations made in the preceding section, we see that a policy that maximizes

$$r(f) + \beta_n^{(k)} w_n^{(k)} ,$$

will not only be average optimal, but even (k - 1)-order average optimal, provided n is sufficiently large. For this to be true, g^* need not be independent of the initial state as has been assumed by Bather and by Hordijk and Tijms.

CHAPTER 8

POLICY ITERATION, GO-AHEAD FUNCTIONS AND SENSITIVE OPTIMALITY

8.1. INTRODUCTION

This chapter deals with policy iteration algorithms for the determination of sensitive optimal policies for the average-reward MDP with finite state space $S := \{1,2,\ldots,N\}$ and finite action space A.
There are two variants of Howard's policy iteration algorithm [HOWARD, 1960] (with the modification by BLACKWELL [1962]) which have been introduced by HASTINGS [1968] and MILLER and VEINOTT [1969]. Hastings replaced the policy improvement step by a Gauss-Seidel step, cf. section 3.3, and proved convergence for the case that all P(f) are unichained. Miller and Veinott extended the improvement step in such a way that the algorithm can produce not only average-optimal but also n-order average optimal policies.
In this chapter we will formulate, in terms of certain stationary go-ahead functions, a set of policy iteration algorithms, which includes the methods of Hastings and Miller and Veinott. It will be shown that these methods all converge in finitely many steps, and yield n-order average-optimal policies, where n will depend on the algorithm under consideration.
The line of reasoning is very similar to that in MILLER and VEINOTT [1969] and VEINOTT [1966].

We restrict ourselves to the consideration of stationary go-ahead functions, for which, as we have seen in sections 3.5 and 4.7, the restriction to policies is permitted under certain conditions. Further we require that $\delta(i) = \delta(i,a) = 1$ for all $a \in A$ and $i \in S$. Thus the go-ahead function is completely characterized by the function $\delta(i,a,j)$, and stopping only occurs immediately after a transition has been made.
From a practical point of view this is not a serious restriction, since in the policy iteration algorithm one is mostly not interested in extrapola-

tions, and thus not very interested in preserving the equal row sums. Recall that it was precisely for this reason that one wanted to have the possibility to choose $\delta(i)$ and $\delta(i,a)$ less than 1. The reason for the restriction to these stationary go-ahead functions is that it considerably simplifies the notations, which already will become rather complex. The extension to the case $0 < \alpha_\delta < 1$ (for stationary go-ahead functions) is, however, straightforward.
Further we have to require

$$(8.1) \qquad \mathbb{E}_f^\delta \tau < \infty \quad \text{for all } f \in F$$

In the algorithms in section 3.3 the only situation in which $\mathbb{E}_f^\delta \tau$ may not be finite, is the case that for some $i \in S$ and $a \in A$ we have $p(i,a,i) = \delta(i,a,i) = 1$. If, in this case, one takes $\delta(i,a,i) = 1 - \varepsilon$, then (8.1) holds again.

In section 2 the policy iteration algorithm, with improvement step formulated in terms of a go-ahead function, is studied for a fixed discount factor $\beta < 1$. This analysis gives us an idea of how to construct the improvement step for the average-reward MDP, treating it again as the limiting case for β tending to 1. Next (in section 3) a Laurent series expansion is obtained for

$$L_{\beta,\delta}(h)v_\beta(f) = \mathbb{E}_h^\delta \left[\sum_{k=0}^{\tau-1} \beta^k r(X_k,A_k) + \beta^\tau v_\beta(X_\tau,f) \right]$$

in $(1-\beta)$ for β tending to 1. In section 4 we derive from this expansion the policy improvement step for the policy iteration algorithm. The convergence proof will be given in section 5.

8.2. SOME NOTATIONS AND PRELIMINARIES

Define for all $f \in F$ the operator $L_{\beta,\delta}(f)$ on V $(= \mathbb{R}^N)$ by

$$(8.2) \qquad L_{\beta,\delta}(f)v := \mathbb{E}_f^\delta \left[\sum_{k=0}^{\tau-1} \beta^k r(X_k,A_k) + \beta^\tau v(X_\tau) \right] .$$

We only consider policies, since for stationary go-ahead functions we can restrict ourselves to stationary strategies in the maximization of $L_{\beta,\delta}(\cdot)v$. Clearly, $L_{\beta,\delta}(f)$ is monotone. Further $L_{\beta,\delta}(f)$ is a contraction with a con-

traction radius of at most β for go-ahead functions with $\delta(i) = \delta(i,a) = 1$ or $\tau \geq 1$. Namely, for all $v,w \in V$

$$\|L_{\beta,\delta}(f)v - L_{\beta,\delta}(f)w\|_e = \|\mathbb{E}_f^{\delta}\beta^{\tau}(v-w)(X_\tau)\|_e \leq \|\mathbb{E}_f^{\delta}\beta^{\tau}\|_e\|v-w\|_e \leq \beta\|v-w\|_e \quad (8.3)$$

Further it is obvious that

$$L_{\beta,\delta}(f)v_\beta(f) = v_\beta(f) .$$

So

$$\lim_{k\to\infty} L^k_{\beta,\delta}(f)v = v_\beta(f) \quad \text{for all } v \in V . \quad (8.4)$$

And we also have:

$$\text{if } L_{\beta,\delta}(f)v \geq v , \text{ then } v_\beta(f) \geq L_{\beta,\delta}(f)v . \quad (8.5)$$

Namely,

$$v_\beta(f) = \lim_{k\to\infty} L^k_{\beta,\delta}(f)v \geq \ldots \geq L_{\beta,\delta}(f)v .$$

Now let us consider the policy iteration algorithm for a fixed $\beta < 1$ with the improvement step formulated in terms of a go-ahead function.

Policy iteration algorithm

Let f be the actual policy.

Determine

$$\max_{h\in F} L_{\beta,\delta}(h)v_\beta(f) \quad (8.6)$$

and replace f by a maximizer of (8.6), until

$$\max_{h\in F} L_{\beta,\delta}(h)v_\beta(f) = v_\beta(f) . \quad (8.7)$$

In order to formulate some results for this algorithm it is convenient to define the partial ordering $\succ$ on V by

$$v \succ w \text{ if and only if } v \geq w \text{ and } v \neq w . \quad (8.8)$$

From (8.5) we see that if

$$L_{\beta,\delta}(h)v_\beta(f) \succ v_\beta(f)$$

then

$$v_\beta(h) \succ v_\beta(f) \ .$$

So, since there are only finitely many policies, after finitely many replacements (8.7) will hold. Then we have

$$L_{\beta,\delta}(h)v_\beta(f) \leq v_\beta(f) \quad \text{for all } h \in F \ ,$$

hence, analogously to (8.5),

$$v_\beta(h) \leq v_\beta(f) \quad \text{for all } h \in F \ ,$$

or f is optimal for discount factor β:

$$v_\beta(f) = v_\beta^* \ .$$

However, we are not interested in the case of a fixed discount factor β, but in the case that β approaches 1 (cf. chapter 7).
The following two lemmas for the case $\beta \uparrow 1$ give us already an idea of the kind of policy improvement step, for the determination of an n-order average-optimal policy, we should look for.

LEMMA 8.1. *If for all* β *sufficiently close to* 1

$$L_{\beta,\delta}(h)v_\beta(f) \succ v_\beta(f) \ ,$$

then for all β *close enough to* 1

$$v_\beta(h) \succ v_\beta(f) \ ,$$

hence also $h \succ f$.

LEMMA 8.2. *If*

$$L_{\beta,\delta}(h)v_\beta(f) \leq v_\beta(f) + O((1-\beta)^{n+1}e) \quad (\beta \uparrow 1) \ , \tag{8.9}$$

then

$$v_\beta(h) \leq v_\beta(f) + O((1-\beta)^{n}e) \quad (\beta \uparrow 1) \ . \tag{8.10}$$

Hence, if (8.9) *holds for all* $h \in F$, *then* f *is* (n-1)*-discount optimal and* n*-order average optimal.*

PROOF. In order to prove that (8.9) implies (8.10), first observe that, similar to (8.3), we have for all $v,w \in V$,

$$L_{\beta,\delta}(h)v - L_{\beta,\delta}(h)w \leq \beta \|(v-w)^+\|_e e \ ,$$

and thus

$$\| (L_{\beta,\delta}(h)v - L_{\beta,\delta}(h)w)^+ \|_e \leq \beta \| (v-w)^+ \|_e \ .$$

Hence,

$$\begin{aligned} v_\beta(h) &= \lim_{k\to\infty} L^k_{\beta,\delta}(h) v_\beta(f) \\ &= \lim_{k\to\infty} \sum_{\ell=0}^{k-1} [L^{\ell+1}_{\beta,\delta}(h) v_\beta(f) - L^{\ell}_{\beta,\delta}(h) v_\beta(f)] + v_\beta(f) \\ &\leq \lim_{k\to\infty} \sum_{\ell=0}^{k-1} \beta^\ell \| (L_{\beta,\delta}(h) v_\beta(f) - v_\beta(f))^+ \|_e \, e + v_\beta(f) \\ &= (1-\beta)^{-1} \, O((1-\beta)^{n+1} e) + v_\beta(f) \\ &= v_\beta(f) + O((1-\beta)^n e) \quad (\beta \uparrow 1) \ . \end{aligned}$$

□

Lemmas 8.1 and 8.2 suggest that we should try to find a policy iteration algorithm of the following form:

Algorithm

Let f be the actual policy

Policy improvement step

Find a policy h, such that

(8.11) $L_{\beta,\delta}(h)v_\beta(f) \succ v_\beta(f)$ for all β close enough to 1 ;

replace f by h.

Repeat the policy improvement step until a policy f is found which satisfies

(8.12) $L_{\beta,\delta}(h)v_\beta(f) \leq v_\beta(f) + O((1-\beta)^{n+1} e)$ for all $h \in F$.

In section 4 we will see that it is possible to check, in a relatively simple way, whether (8.12) holds.

Observe, that since there are only finitely many policies, one will have

$$L_{\beta,\delta}(h)v_\beta(f) \leq v_\beta(f) \quad \text{for all } h \in F \text{ and all } \beta \text{ close to 1,}$$

after only finitely many executions of the policy improvement step. So after finitely many improvement steps one obtains a policy f satisfying (8.12).

8.3. THE LAURENT SERIES EXPANSION OF $L_{\beta,\delta}(h)v_\beta(f)$

In order to find a policy h which satisfies (8.11) or to check whether a policy f satisfies (8.12), we will derive the Laurent series expansion of $L_{\beta,\delta}(h)v_\beta(f)$ in $1-\beta$ for $\beta \uparrow 1$. It will turn out that this Laurent series expansion has a relatively simple form because of the simple structure of the go-ahead functions: stationary, with $\delta(\cdot) = \delta(\cdot,\cdot) = 1$. To exploit this structure we split up the transition probabilities in two parts:

$$\bar{p}_\delta(i,a,j) := p(i,a,j)\delta(i,a,j)$$

and

$$\tilde{p}_\delta(i,a,j) := p(i,a,j)(1-\delta(i,a,j)) \ , \quad a \in A \ , \ i,j \in S \ .$$

So, with probability $\bar{p}_\delta$ a transition is made after which one goes ahead, and with probability $\tilde{p}_\delta$ the transition is followed by stopping.
Further define the matrices $\bar{P}_\delta(h)$ and $\tilde{P}_\delta(h)$ by

$$\bar{P}_\delta(h)(i,j) := \bar{p}_\delta(i,h(i),j) \ ,$$

$$\tilde{P}_\delta(h)(i,j) := \tilde{p}_\delta(i,h(i),j) \ , \quad i,j \in S \ , \ h \in F \ .$$

So,

$$P(f) = \bar{P}_\delta(f) + \tilde{P}_\delta(f) \ .$$

From the analogy with the MDP defined in (3.9) it follows that for all $h \in F$ and $v \in V$

$$L_{\beta,\delta}(h)v = r(h) + \beta\tilde{P}_\delta(h)v + \beta\bar{P}_\delta(h)L_{\beta,\delta}(h)v \ .$$

Hence,

$$L_{\beta,\delta}(h)v = [I - \beta\bar{P}_\delta(h)]^{-1}[r(h) + \beta\tilde{P}_\delta(h)v] \ .$$

Further it has been assumed that

$$\mathbb{E}_h^\delta \tau = e + \bar{P}_\delta(h)e + \bar{P}_\delta^2(h)e + \ldots < \infty \ .$$

So $\bar{P}_\delta^n(h) \to 0$ $(n \to \infty)$, hence $I - \bar{P}_\delta(f)$ is nonsingular. Thus, with (5.39) we obtain the following expansion of $[I - \beta\bar{P}_\delta(h)]^{-1}$ in powers on $1-\beta$ for $\beta \uparrow 1$

$$[I - \beta\bar{P}_\delta(h)]^{-1} = \sum_{k=0}^{\infty} (1-\beta)^k(-1)^k\{[I - \bar{P}_\delta(h)]^{-1}\,\bar{P}_\delta(h)\}^k[I - \bar{P}_\delta(h)]^{-1} \ .$$

And with the expansion (5.40) for $v_\beta(f)$,

$$(8.13)\qquad L_{\beta,\delta}(h)v_\beta(f) = \sum_{k=0}^{\infty}(1-\beta)^k(-1)^k\{[I-\bar{P}_\delta(h)]^{-1}\,\bar{P}_\delta(h)\}^k[I-\bar{P}_\delta(h)]^{-1}\cdot$$

$$\cdot\left[r(h) + [1-(1-\beta)]\tilde{P}_\delta(h)\sum_{\ell=-1}^{\infty}(1-\beta)^\ell c_\ell(f)\right] \quad (\beta \uparrow 1)\,.$$

To simplify this expression we define the following notations:

$$r_\delta(h) := [I-\bar{P}_\delta(h)]^{-1}\,r(h)$$

(the expected reward until time τ);

$$P_\delta(h) := [I-\bar{P}_\delta(h)]^{-1}\,\tilde{P}_\delta(h)$$

($P_\delta(h)(i,j)$ is the probability of stopping in state j);

$$Q_\delta(h) := [I-\bar{P}_\delta(h)]^{-1}$$

($Q_\delta(h)(i,j)$ is the expected number of visits to state j before time τ);

$$R_\delta(h) := [I-\bar{P}_\delta(h)]^{-1}\,\bar{P}_\delta(h)$$

($R_\delta(h)(i,j)$ is the expected number of visits to state j after time 0 and before time τ).

Substituting this into (8.13) yields

$$L_{\beta,\delta}(h)v_\beta(f) = \sum_{k=0}^{\infty}(1-\beta)^k(-1)^k\,R_\delta^k(h)\cdot$$

$$\cdot\left[r_\delta(h) + [1-(1-\beta)]P_\delta(h)\sum_{\ell=-1}^{\infty}(1-\beta)^\ell c_\ell(f)\right].$$

Or

$$L_{\beta,\delta}(h)v_\beta(f) = \sum_{k=-1}^{\infty}(1-\beta)^k\,d_k(h,f)\,,$$

with

$$(8.14)\qquad d_{-1}(h,f) := P_\delta(h)c_{-1}(f)$$

and, for k = 0,1,...,

$$(8.15)\qquad d_k(h,f) := (-1)^k R_\delta^k(h) r_\delta(h) + \sum_{\ell=0}^{k+1} (-1)^\ell R_\delta^\ell(h) P_\delta(h) c_{k-\ell}(f) + \sum_{\ell=0}^{k} (-1)^{\ell+1} R_\delta^\ell(h) P_\delta(h) c_{k-\ell-1}(f) ,$$

where the dependence of $d_k(h,f)$ on δ has been suppressed.
The expression (8.15) can be simplified further to

$$(8.16)\qquad d_0(h,f) = r_\delta(h) + P_\delta(h) c_0(f) - Q_\delta(h) P_\delta(h) c_{-1}(f)$$

and

$$(8.17)\qquad d_k(h,f) = - R_\delta(h) d_{k-1}(h,f) + P_\delta(h)(c_k(f) - c_{k-1}(f)) , \quad k = 1,2,\ldots .$$

So, if one wants to maximize $L_{\beta,\delta}(\cdot)v_\beta(f)$ for all β sufficiently close to 1, then one has to maximize first $d_{-1}(\cdot,f)$, next one has to maximize $d_0(\cdot,f)$ over those policies which maximize $d_{-1}(\cdot,f)$, etc. In section 4 it will be shown that this can be done in a relatively simple way.

For later use we want to rewrite the equations (8.14), (8.16) and (8.17). Therefore, define for all f and $h \in F$

$$(8.18)\qquad \psi_k(h,f) := d_k(h,f) - c_k(f) , \quad k = -1,0,\ldots$$

($\psi_k(h,f)$ depends of course also on δ).
Clearly, since $L_{\beta,\delta}(f)v_\beta(f) = v_\beta(f)$, we have

$$\psi_k(f,f) = 0 , \quad k = -1,0,\ldots .$$

Substituting (8.18) into (8.14), (8.16) and (8,17) yields

$$(8.19)\qquad P_\delta(h) c_{-1}(f) = c_{-1}(f) + \psi_{-1}(h,f) ,$$

$$(8.20)\qquad r_\delta(h) + P_\delta(h) c_0(f) - Q_\delta(h)(c_{-1}(f) + \psi_{-1}(h,f)) = c_0(f) + \psi_0(h,f) ,$$

$$(8.21)\qquad - R_\delta(h)(c_{k-1}(f) + \psi_{k-1}(h,f)) + P_\delta(h)(c_k(f) - c_{k-1}(f)) = c_k(f) + \psi_k(h,f) .$$

Or, upon premultiplication with $I - \bar{P}_\delta(h)$,

$$(8.22)\qquad \tilde{P}_\delta(h) c_{-1}(f) + \bar{P}_\delta(h)(c_{-1}(f) + \psi_{-1}(h,f)) = c_{-1}(f) + \psi_{-1}(h,f) ,$$

$$(8.23)\qquad r(h) + \tilde{P}_\delta(h)c_0(f) - (c_{-1}(f) + \psi_{-1}(h,f)) + \bar{P}_\delta(h)(c_0(f) + \psi_0(h,f)) = c_0(f) + \psi_0(h,f) ,$$

$$(8.24)\qquad - \bar{P}_\delta(h)(c_{k-1}(f) + \psi_{k-1}(h,f)) + \tilde{P}_\delta(h)(c_k(f) - c_{k-1}(f)) + \bar{P}_\delta(h)(c_k(f) - \psi_k(h,f)) = c_k(f) + \psi_k(h,f) .$$

In the next section we will see how to obtain from equations (8.22)-(8.24) a characterization of the set of policies which subsequently maximize $d_{-1}(\cdot,f)$, $d_0(\cdot,f)$, ... until $d_n(\cdot,f)$.

This section is concluded with a restatement of lemmas 8.1 and 8.2 in terms of the functions $\psi_k(h,f)$.
Define the matrix $\Psi_n(h,f)$ as the $(N \times (n+2))$-matrix with columns $\psi_{-1}(h,f)$, $\psi_0(h,f),\ldots,\psi_n(h,f)$, and define the matrix $\Psi_\infty(h,f)$ as the $(N \times \infty)$-matrix with columns $\psi_{-1}(h,f),\psi_0(h,f),\ldots$. Then lemmas 8.1 and 8.2 state that

if $\Psi_\infty(h,f) \succ 0$, then $h \succ f$,

and

if $\Psi_n(h,f) \preccurlyeq 0$ for all $h \in F$, then t is n-order average optimal.

8.4. THE POLICY IMPROVEMENT STEP

In this section a policy inprovement step will be constructed which, given a policy $f \in F$, either produces a policy $h \in F$ satisfying $\Psi_\infty(h,f) \succ 0$, or signals that $\Psi_n(h,f) \preccurlyeq 0$ for all $h \in F$.

In section 3 it has been shown, that in order to maximize $L_{\beta,\delta}(\cdot)v_\beta(f)$ for all β sufficiently close to 1, one has to maximize subsequently the terms $\psi_{-1}(\cdot,f),\psi_0(\cdot,f),\ldots$ (or $d_{-1}(\cdot,f),d_0(\cdot,f),\ldots$). Since in the maximization of $\psi_0(\cdot,f)$ only those policies need to be considered that maximize $\psi_{-1}(\cdot,f)$, we first derive a characterization of the set of policies maximizing $\psi_{-1}(\cdot,f)$.
Let h maximize $\psi_{-1}(\cdot,f)$,

$$(8.25)\qquad \psi_{-1}(h,f) = \max_{g\in F} \psi_{-1}(g,f) =: \psi_{-1}(f) .$$

The existence of such a uniformly maximizing policy follows from the fact that the MDP which is equivalent to the problem of maximizing $P_\delta(\cdot)c_{-1}(f)$ (constructed along the lines of (3.9)) is contracting. This follows from $\bar{P}_\delta^n(\cdot) \to 0$ $(n \to \infty)$ and the finiteness of S and A.
For h we have

$$(8.26) \qquad P_\delta(h)c_{-1}(f) = c_{-1}(f) + \psi_{-1}(f)$$

and

$$(8.27) \qquad \tilde{P}_\delta(h)c_{-1}(f) + \bar{P}_\delta(h)(c_{-1}(f) + \psi_{-1}(f)) = c_{-1}(f) + \psi_{-1}(f) \ .$$

Now, define $\gamma_{-1}(i,a,f)$ for all $i \in S$ and $a \in A$ by

$$(8.28) \qquad \sum_{j\in S} \tilde{p}_\delta(i,a,j)c_{-1}(j,f) + \sum_{j\in S} \bar{p}_\delta(i,a,j)(c_{-1}(j,f) + \psi_{-1}(j,f))$$

$$= c_{-1}(i,f) + \psi_{-1}(i,f) + \gamma_{-1}(i,a,f) \ .$$

Then it follows from (8.27) that $\gamma_{-1}(i,h(i),f) = 0$ for all $i \in S$. Further, $\gamma_{-1}(i,a,f) \leq 0$ for all $i \in S$, $a \in A$. Namely, suppose $\gamma_{-1}(\cdot,\cdot,f) > 0$ for some pair (i,a_i), then we have for the policy $\hat{h}$ defined by $\hat{h}(i) = a_i$, and $\hat{h}(j) = h(j)$, $j \neq i$, with $\gamma_{-1}(h,f)(i) = \gamma_{-1}(i,h(i),f)$,

$$\tilde{P}_\delta(\hat{h})c_{-1}(f) + \bar{P}_\delta(\hat{h})(c_{-1}(f) + \psi_{-1}(f)) = c_{-1}(f) + \psi_{-1}(f) + \gamma_{-1}(\hat{h},f) \ .$$

So, after premultiplication with $[I - \bar{P}_\delta(\hat{h})]^{-1}$, and some reordering

$$P_\delta(\hat{h})c_{-1}(f) = c_{-1}(f) + \psi_{-1}(f) + [I - \bar{P}_\delta(\hat{h})]^{-1} \gamma_{-1}(\hat{h},f) \ .$$

Since $\gamma_{-1}(h,f) \succ 0$, also

$$[I - \bar{P}_\delta(\hat{h})]^{-1} \gamma_{-1}(\hat{h},f) = \sum_{k=0}^{\infty} \bar{P}_\delta^k(\hat{h})\gamma_{-1}(\hat{h},f) \geq \gamma_{-1}(\hat{h},f) \succ 0 \ .$$

But this would contradict the definition of $\psi_{-1}(f)$. So, $\gamma_{-1}(i,a,f) \leq 0$ for all $i \in S$ and $a \in A$.
Similarly, let g be a policy with $\gamma_{-1}(g,f) \prec 0$, then

$$\psi_{-1}(g,f) = \tilde{P}_\delta(g)c_{-1}(f) + \bar{P}_\delta(g)[c_{-1}(f) + \psi_{-1}(g,f)] - c_{-1}(f)$$

$$\leq \tilde{P}_\delta(g)c_{-1}(f) + \bar{P}_\delta(g)[c_{-1}(f) + \psi_{-1}(f)] - c_{-1}(f)$$

$$= c_{-1}(f) + \psi_{-1}(f) + \gamma_{-1}(g,f) - c_{-1}(f) \prec \psi_{-1}(f) \ .$$

Hence the set of policies which maximize $\psi_{-1}(\cdot,f)$ can be characterized by the actions having $\gamma_{-1}(i,a,f) = 0$. I.e., a policy g has $\psi_{-1}(g,f) = \psi_{-1}(f)$ if and only if $\gamma_{-1}(i,g(i),f) = 0$ for all $i \in S$.

This result can also be obtained from the observation that

$$\max_{g\in F} \{\tilde{P}_\delta(g)c_{-1}(f) + \bar{P}_\delta(g)(c_{-1}(f) + \psi_{-1}(f))\} = c_{-1}(f) + \psi_{-1}(f)$$

is just the optimality equation for the MDP which is equivalent to the problem of maximizing $P_\delta(\cdot)c_{-1}(f)$.
Now define for all $i \in S$

$$A_{-1}(i,f) := \{a \in A \mid \gamma_{-1}(i,a,f) = 0\}$$

and define the restricted policy set $F_{-1}(f)$ by

$$F_{-1}(f) := \{g \in F \mid g(i) \in A_{-1}(i,f) \text{ for all } i \in S\} .$$

So $F_{-1}(f)$ is just the set of maximizing policies of $\psi_{-1}(\cdot,f)$.
Next one has to maximize $\psi_0(\cdot,f)$ over all policies which maximize $\psi_{-1}(\cdot,f)$, i.e., over the policies in $F_{-1}(f)$.
From now on we are, strictly speaking, considering a restricted MDP with state-dependent action sets $A_{-1}(i,f)$, $i \in S$. Define

$$\psi_0(f) := \max_{g\in F_{-1}(f)} \psi_0(g,f) ,$$

and define $\gamma_0(i,a,f)$ for all $i \in S$ and $a \in A_{-1}(i,f)$ by

$$\begin{aligned} (8.29) \qquad & r(i,a) + \sum_{j\in S} \tilde{p}_\delta(i,a,j)c_0(j,f) - c_{-1}(i,f) - \psi_{-1}(i,f) + \\ & + \sum_{j\in S} \bar{p}_\delta(i,a,j)(c_0(j,f) + \psi_0(j,f)) \\ & = c_0(i,f) + \psi_0(i,f) + \gamma_0(i,a,f) \end{aligned}$$

(compare (8.23)).
Further define

$$A_0(i,f) := \{a \in A_{-1}(i,f) \mid \gamma_0(i,a,f) = 0\} , \quad i \in S$$

and

$$F_0(f) := \{g \in F \mid g(i) \in A_0(i,f) \text{ for all } i \in S\} .$$

Then similar reasoning as before shows that $F_0(f)$ is just the set of maximizing policies of $\psi_0(\cdot,f)$ within $F_{-1}(f)$.

Continuing in this way, define for $k = 1,2,\ldots$

$$\psi_k(f) := \max_{g \in F_{k-1}(f)} \psi_k(g,f) ,$$

and define $\gamma_k(i,a,f)$ by (cf. (8.24)

$$\begin{aligned} (8.30) \quad & - \sum_{j \in S} \bar{p}_\delta(i,a,j)\,(c_{k-1}(j,f) + \psi_{k-1}(j,f)) \\ & + \sum_{j \in S} \tilde{p}_\delta(i,a,j)\,(c_k(j,f) - c_{k-1}(j,f)) + \\ & + \sum_{j \in S} \bar{p}_\delta(i,a,j)\,(c_k(j,f) + \psi_k(j,f)) = \\ & = c_k(i,f) + \psi_k(i,f) + \gamma_k(i,a,f) . \end{aligned}$$

Further define

$$A_k(i,f) := \{a \in A_{k-1}(i,f) \mid \gamma_k(i,a,f) = 0\} , \quad i \in S ,$$

and

$$F_k(f) := \{g \in F \mid g(i) \in A_k(i,f) \text{ for all } i \in S\} .$$

Then $F_k(f)$ is again the set of maximizers of $\psi_k(\cdot,f)$ within $F_{k-1}(f)$, i.e., if $g \in F_{k-1}(f)$, then $\psi_k(g,f) = \psi_k(f)$ if and only if $g \in F_k(f)$.

Now the policy iteration algorithm for the determination of an n-order average-optimal policy can be formulated as follows.

Policy iteration algorithm for an n-order average optimal policy

Let f be the actual policy.

Policy improvement step

Determine for all $i \in S$ the set $A_n(i,f)$, and replace f by a policy which satisfies for all $i \in S$

$$(8.31) \qquad h(i) \in A_n(i,f) \quad \text{and} \quad h(i) = f(i) \quad \text{if} \quad f(i) \in A_n(i,f) .$$

Repeat the policy improvement step until a policy f is found which satisfies

$$f(i) \in A_n(i,f) \quad \text{for all } i \in S .$$

In the next section it will be shown that the policy h, obtained from f by the policy improvement step, satisfies:

(i) if $h \neq f$, i.e., $f(i) \notin A_n(i,f)$ for some $i \in S$, then $\Psi_\infty(h,f) \succ 0$, so $h \succ f$;

(ii) if $h = f$, i.e., $f(i) \in A_n(i,f)$ for all $i \in S$, then f is n-order average optimal.

8.5. THE CONVERGENCE PROOF

In order to show that for a policy h obtained from f by the improvement step (8.31) one has

$$\Psi_\infty(h,f) \succ 0$$

(and hence $h \succ f$), unless $h = f$, we first need the following lemma.

LEMMA 8.3. *Let f be an arbitrary policy, then*

(i) $\psi_{-1}(f) \geq 0$.

Moreover, if for some $i \in S$ *and* $k \in \{-1,0,\ldots\}$ *we have* $\psi_{-1}(i,f) = \ldots = \psi_k(i,f) = 0$, *then*

(ii) $\psi_{-1}(j,f) = \ldots = \psi_k(j,f) = 0$

for all $j \in S(i,f) := \{\ell \in S \mid \bar{p}_\delta(i,f(i),\ell) > 0\}$,

(iii) $f(i) \in A_k(i,f)$,

(iv) $\psi_{k+1}(i,f) \geq 0$.

PROOF.

(i) From

$$L_{\beta,\delta}(f)v_\beta(f) = v_\beta(f)$$

it follows that

$$d_{-1}(f,f) = P_\delta(f)c_{-1}(f) = c_{-1}(f) .$$

Hence

$$\psi_{-1}(f) = \max_{g\in F} P_\delta(g)c_{-1}(f) - c_{-1}(f) \geq P_\delta(f)c_{-1}(f) - c_{-1}(f) = 0 .$$

Next we prove (ii)-(iv) by induction.
First the case k = -1. Assume $\psi_{-1}(i,f) = 0$. Then subtraction of equation (8.22), with h = f, for state i from equation (8.28) yields, with $\psi_{-1}(f,f) = 0$,

$$\sum_{j\in S} \bar{p}_\delta(i,f(i),j)\psi_{-1}(j,f) = \psi_{-1}(i,f) + \gamma_{-1}(i,f(i),f) .$$

Hence, with $\psi_{-1}(i,f) = 0$, $\psi_{-1}(f) \geq 0$ and $\gamma_{-1}(i,f(i),f) \leq 0$,

$$\gamma_{-1}(i,f(i),f) = 0 , \quad \text{so } f(i) \in A_{-1}(i,f)$$

and

$$\sum_{j\in S} \bar{p}_\delta(i,f(i),j)\psi_{-1}(j,f) = 0 , \quad \text{so } \psi_{-1}(j,f) = 0 \text{ for all } j \in S(i,f) . \tag{8.32}$$

Now, let

$$S_{-1}(f) := \{j \in S \mid \psi_{-1}(j,f) = 0\} ,$$

then it follows from (8.32) that $S_{-1}(f)$ is closed under $\bar{P}_\delta(f)$. Further $f(j) \in A_{-1}(j,f)$ for all $j \in S_{-1}(f)$. Next, let $\bar{f}$ be any policy with $\bar{f}(f) = f(j)$ for all $j \in S_{-1}(f)$ and $\bar{f}(j) \in A_{-1}(j,f)$ elsewhere, then $\bar{f} \in S_{-1}(f)$. If the process starts in $S_{-1}(f)$ and policy $\bar{f}$ is used, then the system will not leave $S_{-1}(f)$ before time τ, therefore only actions from f will be used. Hence

$$\psi_0(i,f) = \max_{g\in F_{-1}(f)} \psi_0(i,g,f) \geq \psi_0(i,\bar{f},f) = \psi_0(i,f,f) = 0 .$$

This completes the proof for k = -1.
Now let us assume that (ii)-(iv) hold for k = m - 1, and define the sets $S_\ell(f)$, $\ell = 0,1,\ldots$, by

$$S_\ell(f) := \{j \in S \mid \psi_{-1}(j,f) = \ldots = \psi_\ell(j,f) = 0\} . \tag{8.33}$$

Then it follows from the induction assumption that $f(j) \in A_{m-1}(j,f)$ for all $j \in S_{m-1}(f)$, that $\psi_{m-1}(j,f) = 0$ for all j belonging to a set S(i,f) for some $i \in S_{m-1}(f)$, and that $\psi_m(j,f) \geq 0$ for all $j \in S_{m-1}(f)$.
We will prove (ii)-(iv) for k = m, so assume $\psi_{-1}(i,f) = \ldots = \psi_m(i,f) = 0$.
The proof is almost identical to the proof of the case k = -1. Subtracting

(8.24), with h = f, for state i from equation (8.30) yields, with $\psi_{m-1}(j,f) = 0$ for all $j \in S(i,f)$ and $\psi_{m-1}(f,f) = 0$,

$$(8.34) \qquad \gamma_m(i,f(i),f) = \sum_{j \in S} \bar{p}_\delta(i,f(i),j)\psi_m(j,f) \ .$$

(For m = 0 one has to subtract (8.23), with h = f, for state i from (8.29) which also yields (8.34).)
Since $f(i) \in A_{m-1}(i,f)$, we have $\gamma_m(i,f(i),f) \leq 0$, and since $\psi_{m-1}(j,f) = 0$ for all $j \in S(i,f)$ by the induction assumption, also $\psi_m(j,f) \geq 0$ for all $j \in S(i,f)$.
This implies that both sides in (8.34) must be equal to zero.
So $f(i) \in A_m(i,f)$ and $\psi_m(j,f) = 0$ for all $j \in S(i,f)$. Hence $S_m(f)$ is closed under $\bar{P}_\delta(f)$ and the same reasoning as for k = -1 yields $\psi_{m+1}(i,f) \geq 0$.
This completes the proof of (ii)-(iv). □

This lemma yields the following corollary.

COROLLARY 8.4.

(i) *Denote by* $\Psi_\infty(f)$ *the* $(N \times \infty)$*-matrix with columns* $\psi_{-1}(f), \psi_0(f), \psi_1(f), \ldots$, *then*

$$\Psi_\infty(f) \geqslant 0 \ .$$

(ii) *Let* h *be the (a) policy obtained from* f *by the policy improvement step* (8.31), *then*

$$\psi_{-1}(i,f) = \ldots = \psi_n(i,f) = 0 \quad \textit{only if} \quad h(i) = f(i) \ .$$

Finally, we can prove that each policy improvement step yields a better policy.

THEOREM 8.5. *Let* h *be a policy obtained from* f *by the policy improvement step* (8.31), *then*

$$\Psi_\infty(h,f) \geqslant 0 \ ,$$

and

$$\Psi_\infty(h,f) = 0 \quad \textit{only if} \quad h = f \ .$$

PROOF. Let $S_n(f)$ be defined as in (8.33). Then

$$(8.35) \qquad L_{\beta,\delta}(h)v_\beta(f)(j) > v_\beta(j,f)$$

for all $j \notin S_n(f)$ and for all β close enough to 1.
Further, $h(i) = f(i)$ for all $i \in S_n(f)$. So, for all $i \in S_n(f)$,

$$(L_{\beta,\delta}(h)v_\beta(f) - v_\beta(f))(i) = r(i,f(i)) + \beta \sum_{j\in S} \tilde{p}_\delta(i,f(i),j)v_\beta(j,f) +$$

$$+ \beta \sum_{j\in S} \bar{p}_\delta(i,f(i),j)L_{\beta,\delta}(h)v_\beta(f)(j) - v_\beta(i,f) \ .$$

Also,

$$r(i,f(i)) + \beta \sum_{j\in S} \tilde{p}_\delta(i,f(i),j)v_\beta(j,f) + \beta \sum_{j\in S} \bar{p}_\delta(i,f(i),j)v_\beta(j,f) = v_\beta(i,f) \ .$$

Together this yields

$$(8.36) \qquad (L_{\beta,\delta}(h)v_\beta(f) - v_\beta(f))(i) =$$

$$= \beta \sum_{j\in S} \bar{p}_\delta(i,f(i),j)(L_{\beta,\delta}(h)v_\beta(f) - v_\beta(f))(j)$$

$$\geq \beta \sum_{j\in S_n(f)} \bar{p}_\delta(i,f(i),j)(L_{\beta,\delta}(h)v_\beta(f) - v_\beta(f))(j) \ ,$$

for all β close enough to 1.
Iterating (8.36) (on $S_n(f)$) and letting the number of iterations tend to infinity yields

$$(8.37) \qquad (L_{\beta,\delta}(h)v_\beta(f) - v_\beta(f))(i) \geq 0$$

for all $i \in S_n(f)$ and all β sufficiently close to 1.
From (8.35) and (8.37) we obtain

$$\Psi_\infty(h,f) \succcurlyeq 0 \ ,$$

and from (8.35)

$$\Psi_\infty(h,f) = 0 \quad \text{only if} \quad S_n(f) = S \ .$$

But $S_n(f) = S$ implies $\Psi_n(f) = 0$, so, with lemma 8.3(iii), $f(i) \in A_n(i,f)$ for all $i \in S$, hence $h = f$. □

Since there are only finitely many policies and since $\Psi_\infty(h,f) \succ 0$ implies $h \succ f$, it follows from the transitivity of the relation $\succ$ for policies, that the algorithm must terminate after finitely many policy improvements with a policy f satisfying $f(i) \in A_n(i,f)$ for all $i \in S$.

It remains to be shown that this policy f is then n-order average optimal. By lemma 8.2 it suffices to prove that for all $g \in F$

$$L_{\beta,\delta}(g)v_\beta(f) \leq v_\beta(f) + O((1-\beta)^{n+1}e) \quad (\beta \uparrow 1)$$

or

$$\Psi_n(g,f) \preccurlyeq 0 \quad \text{for all } g \in F .$$

To prove this, consider the following analogon of lemma 8.3.

LEMMA 8.6. *If* $f \in F_n(f)$, *so* $\Psi_n(f) = 0$, *then for all* $g \in F$

(i) $\quad \psi_{-1}(g,f) \leq 0 ,$

and further, if for some $k < n$, *some* $i \in S$ *and some* $g \in F$ *we have* $\psi_{-1}(i,g,f) = \ldots = \psi_k(i,g,f) = 0$, *then*

(ii) $\quad \psi_{-1}(j,g,f) = \ldots = \psi_k(j,g,f) = 0$ *for all* $j \in S(i,g) ,$

(iii) $\quad g(i) \in A_k(i,f) ,$

(iv) $\quad \psi_{k+1}(i,g,f) \leq 0 .$

PROOF. The proof is similar to the proof of lemma 8.3.

(i) $\quad \psi_{-1}(g,f) \leq \max_{h \in F} \psi_{-1}(h,f) = \psi_{-1}(f) = 0 .$

The proof of (ii)-(iv) proceeds again by induction on k.
First the case k = -1. So assume $\psi_{-1}(i,g,f) = 0$. Subtracting (8.22), with h = g, for state i from (8.28) we obtain with $\psi_{-1}(i,g,f) = 0$ and $\psi_{-1}(f) = 0$,

$$(8.38) \qquad \sum_{j \in S(i,g)} \bar{p}_\delta(i,g(i),j)\psi_{-1}(j,g,f) = \gamma_{-1}(i,g(i),f) .$$

The left-hand side in (8.38) is nonnegative and the right-hand side is non-positive, so both sides must be equal to zero. Hence, $\psi_{-1}(j,g,f) = 0$ for all $j \in S(i,g)$ and $g(i) \in A_{-1}(i,f)$.
Let $\bar{g}$ be an arbitrary policy in $F_{-1}(f)$ with $\bar{g}(j) = g(j)$ for all j with $g(j) \in A_{-1}(i,f)$, then

$$\psi_0(i,g,f) = \psi_0(i,\bar{g},f) \leq \max_{h \in F_{-1}(f)} \psi_0(i,h,f) = \psi_0(i,f) = 0 .$$

This completes the proof for k = -1.

The case $k \geq 0$ is completely analogous to the case $k \geq 0$ in lemma 8.3, and is therefore omitted. □

From lemma 8.6 we immediately have

THEOREM 8.7. *If* $f \in F_n(f)$, *then* $\Psi_n(g,f) \preceq 0$ *for all* $g \in F$, *hence* f *is n-order average optimal.*

Theorems 8.5 and 8.7 together imply that the policy iteration algorithm for the determination of an n-order average-optimal policy, formulated in (8.31), terminates after finitely many policy improvements with an n-order average-optimal policy.
For the case $n = 0$, this generalizes the result of HASTINGS [1968] to the case of state-dependent gains. Further we see that the algorithm of MILLER and VEINOTT [1969] corresponds to the special case $\delta(i,a,j) = 0$ for all $i,j \in S$ and $a \in A$.

CHAPTER 9

VALUE-ORIENTED SUCCESSIVE APPROXIMATIONS FOR THE AVERAGE-REWARD MDP

9.1. INTRODUCTION

This chapter deals with the method of value-oriented standard successive approximations for the average-reward MDP with finite state space $S = \{1,2,\ldots,N\}$ and finite action space A. As has been shown in chapters 3 and 4, value-oriented methods can be used for the approximation of the value of a total-reward MDP (theorems 3.22, 4.27 and 4.28). For the average-reward MDP the value-oriented method has been first mentioned by MORTON [1971], however, without convergence proof. Here it will be shown that the value-oriented method converges under a strong aperiodicity assumption and various conditions on the chain structure of the MDP, which have in common that they all guarantee that the gain g^* of the MDP is independent of the initial state.
The contents of this chapter (except for section 8) can be found in Van der WAL [1980a].
Let us first formulate the method.

Value-oriented standard successive approximations

Choose $v_0 \in V\ (= \mathbb{R}^N)$ and $\lambda \in \{1,2,\ldots\}$.
Determine for $n = 0,1,\ldots$ a policy f_{n+1} such that

$$L(f_{n+1})v_n = Uv_n \ , \tag{9.1}$$

and define

$$v_{n+1} = L^{\lambda}(f_{n+1})v_n \ . \tag{9.2}$$

For $\lambda = 1$ this is just the method of standard successive approximations. As we have seen in the total-reward case the method of value-oriented standard successive approximations lays somewhere in between the method of standard successive approximations and the policy iteration method. At the end of this first section we will see that also in the average-reward case the value-oriented method becomes very similar to the policy iteration method if λ is large.

In general, the sequences $\{f_n\}$ and $\{v_n\}$ are (given v_0 and λ) not unique. Let $\{f_n, v_n\}$ be an arbitrary sequence pair which can be obtained when using the value-oriented standard successive approximations method. Throughout this chapter this sequence will be held fixed. The results that will be obtained hold for all sequences which might result from applying the value-oriented method.

Except for section 8 we work under the following assumption.

Strong aperiodicity assumption

There exists a constant $\alpha > 0$ such that

(9.3) $$P(f) \geq \alpha I \quad \text{for all } f \in F ,$$

where I denotes the identity matrix.

Recall that in section 6.3 it has been shown, that any average-reward MDP can be transformed into an equivalent MDP, which satisfies this strong aperiodicity assumption, by means of Schweitzer's aperiodicity transformation (see SCHWEITZER [1971]).

Moreover, we always use a condition which guarantees that g^* is independent of the initial state: the irreducibility condition in section 3, the unichain condition in sections 4 and 5, and in sections 6 and 7 the conditions of communicatingness and simply connectedness, respectively.

Let us already consider the unichain case in somewhat more detail.
An MDP is called *unichained* if for all $f \in F$ the matrix $P(f)$ is unichained, i.e., the Markov chain corresponding to $P(f)$ has only one recurrent subchain and possibly some transient states.
For a unichained MDP the gain $g(f)$ corresponding to a policy f is independent of the initial state, since $g(f) = P^*(f)r(f)$ and, in this case, $P^*(f)$

is a matrix with equal rows. Hence (cf. theorem 6.2), also g^* is independent of the initial state.
It will be convenient to denote for all $f \in F$ by g_f the scalar with $g(f) = g_f e$, and to denote by g_* the scalar with $g^* = g_* e$.
As we already remarked, the value-oriented method becomes for large λ very similar to the policy iteration method. For the unichain case this can be easily seen as follows. For all n

$$v_{n+1} = L^{\lambda}(f_{n+1})v_n = L^{\lambda}(f_{n+1})v(f_{n+1}) + P^{\lambda}(f_{n+1})(v_n - v(f_{n+1}))$$

$$= v(f_{n+1}) + \lambda g_{f_{n+1}} e + P^{\lambda}(f_{n+1})(v_n - v(f_{n+1})) \ .$$

If λ tends to infinity, then, by the strong aperiodicity assumption, $P^{\lambda}(f_{n+1})$ converges to the matrix with equal rows $P^*(f_{n+1})$. Thus, $P^{\lambda}(f_{n+1})(v_n - v(f_{n+1}))$ converges to a constant vector. So, if λ is large, then also the difference between v_{n+1} and $v(f_{n+1})$ is nearly a constant vector. Hence, if λ is sufficiently large, there will exist a policy $\bar{f}$ which maximizes both $L(f)v_{n+1}$ and $L(f)v(f_{n+1})$. Further, we see that the policy improvement step of the policy iteration algorithm, formulated in section 6.2, reduces to the maximization of $L(f)v(f_{n+1})$ if $g(f_{n+1})$ is a constant vector, which happens to be the case if the MDP is unichained. So indeed in the unichain case the value-oriented method and the policy iteration method become very similar for large λ. Approximative algorithms based on this idea can be found in MORTON [1971] and Van der WAL [1976].

In this chapter it will be shown that under various conditions the value-oriented method converges. So we have to show that the value-oriented method enables us to find for all $\varepsilon > 0$ a so-called ε-optimal policy, i.e. a policy f which satisfies $g(f) \geq g^* - \varepsilon e$.

First (in section 2) some preliminary inequalities are given. Next, the irreducible case (the unichain case without transient states) is dealt with (section 3). The unichain case is treated in section 4 and in section 5 it is shown that in the unichain case the value-oriented method converges ultimately exponentially fast. Sections 6 and 7 relax the unichain condition to communicatingness (cf BATHER [1973]) and simply connectedness (cf. PLATZMAN [1977]). Finally, in section 8, an example of a unichained MDP is presented which shows that, if instead of strong aperiodicity we only

assume that all chains are aperiodic, then the value-oriented method may cycle between suboptimal policies.

9.2. SOME PRELIMINARIES

Let $\{f_n\}$ and $\{v_n\}$ be the fixed sequences (section 9.1) obtained from the value-oriented method.
Define ℓ_n and u_n, $n = 0,1,\ldots$, by

$$\ell_n := \min_{i \in S} (Uv_n - v_n)(i) \tag{9.4}$$

and

$$u_n := \max_{i \in S} (Uv_n - v_n)(i) \ . \tag{9.5}$$

Then lemma 6.8 states that

$$\ell_n e \leq g(f_{n+1}) \leq g^* \leq u_n e \ . \tag{9.6}$$

So, what we would like to show is that $u_n - \ell_n$ tends to zero if n tends to infinity, so that both ℓ_n and u_n converge to g_* (for which of course g^* has to be independent of the initial state).
A first result in this direction is given by the following lemma.

LEMMA 9.1. *The sequence* $\{\ell_n, n = 0,1,\ldots\}$ *is monotonically nondecreasing.*

PROOF. For all n

$$\begin{aligned} Uv_n - v_n &= L(f_{n+1})v_n - v_n \geq L(f_n)v_n - v_n \\ &= L^\lambda(f_n)L(f_n)v_{n-1} - L^\lambda(f_n)v_{n-1}) = P^\lambda(f_n)(L(f_n)v_{n-1} - v_{n-1}) \\ &\geq P^\lambda(f_n)\ell_{n-1}e = \ell_{n-1}e \ . \end{aligned}$$

Hence also $\ell_n \geq \ell_{n-1}$. □

In the special case of $\lambda = 1$, also the sequence $\{u_n\}$ is monotone (actually nonincreasing, see ODONI [1969]). This, however, need not be the case if $\lambda > 1$.

EXAMPLE 9.2. $S = \{1,2\}$, $A(1) = \{1\}$, $A(2) = \{1,2\}$. Furthermore, $p(1,1,1) = 1$ and $r(1,1) = 100$. In state 2 action 1 has $r(2,1) = 0$ and $p(2,1,1) = 0{,}9$ and action 2 has $r(2,2) = 10$ and $p(2,2,1) = 0{,}1$. So action 1 has the higher probability of reaching state 1 but action 2 has the higher immediate reward.
Now take $v_0 = 0$ and $\lambda = 2$. Then $Uv_0 - v_0 = (100,10)^T$, so $u_0 = 100$ and we get $v_1 = (200,29)^T$. Next we compute $Uv_1 - v_1 = (100,153.9)^T$, thus $u_1 = 153.9 > u_0$.

Our approach in the following section will be as follows. First we examine the sequence $\{\ell_n\}$ for which it will be shown that $\ell_n \uparrow g_*$. Next it will be shown that u_n converges to g_* as well. Hence $u_n - \ell_n$ tends to zero, which, by (9.6), implies that f_n becomes nearly optimal in the long run.

9.3. THE IRREDUCIBLE CASE

This section deals with the irreducible MDP, i.e., the case that for all $f \in F$ the matrix $P(f)$ is irreducible. The analysis of this case is considerably simpler than in the unichain case with transient states, which is to be considered in sections 4 and 5.
So, throughout this section, it is assumed that the matrices $P(f)$ all have one recurrent subchain and no transient states.
Define for $n = 0,1,\ldots$ the vector $g_n \in \mathbb{R}^N$ by

$$g_n = Uv_n - v_n \ . \tag{9.7}$$

Then (compare the proof of lemma 9.1)

$$g_{n+1} \geq P^{\lambda}(f_{n+1})g_n \ . \tag{9.8}$$

And consequently, for all $k = 1,2,\ldots,$

$$g_{n+k} \geq P^{\lambda}(f_{n+k}) \cdots P^{\lambda}(f_{n+1})g_n \ . \tag{9.9}$$

Define

$$\gamma := \min_{i,j \in S} \min_{h_1,\ldots,h_{N-1}} P(h_1)P(h_2) \cdots P(h_{N-1})(i,j) \ . \tag{9.10}$$

In lemma 9.4 it will be shown that the aperiodicity assumption and the

irreducibility assumption together imply that $\gamma > 0$. Then the following lemma implies that ℓ_n converges to g_* exponentially fast.

LEMMA 9.3. *If* $k\lambda \geq N-1$, *then for all* n

$$g_* - \ell_{n+k} \leq (1-\gamma)(g_* - \ell_n) .$$

PROOF. Let j_0 satisfy $g_n(j_0) = u_n$. Then for all $i \in S$ and all $h_1,\ldots,h_{N-1}$

$$P(h_1)\cdots P(h_{N-1})g_n(i) = \sum_{j\in S} P(h_1)\cdots P(h_{N-1})(i,j)g_n(j)$$

$$= \sum_{j\neq j_0} P(h_1)\cdots P(h_{N-1})(i,j)g_n(j) + P(h_1)\cdots P(h_{N-1})(i,j_0)u_n$$

$$\geq \sum_{j\neq j_0} P(h_1)\cdots P(h_{N-1})(i,j)\ell_n + P(h_1)\cdots P(h_{N-1})(i,j_0)u_n$$

$$\geq (1-\gamma)\ell_n + \gamma u_n \geq (1-\gamma)\ell_n + \gamma g_* .$$

So,

$$P(h_1)\cdots P(h_{N-1})g_n \geq [(1-\gamma)\ell_n + \gamma g_*]e .$$

Then also for all $m > N-1$ and all $h_1,\ldots,h_m$

$$P(h_1)\cdots P(h_m)g_n \geq [(1-\gamma)\ell_n + \gamma g_*]e .$$

Hence, with (9.9), for all k such that $k\lambda \geq N-1$,

$$g_{n+k} \geq P^{\lambda}(f_{n+k})\cdots P^{\lambda}(f_{n+1})g_n \geq [(1-\gamma)\ell_n + \gamma g_*]e .$$

Thus

$$\ell_{n+k} \geq (1-\gamma)\ell_n + \gamma g_*$$

or

$$g_* - \ell_{n+k} \leq (1-\gamma)(g_* - \ell_n) . \qquad \square$$

LEMMA 9.4. $\gamma > 0$.

PROOF. Since S and F are finite, it is sufficient to prove that

$$P(h_1)\cdots P(h_{N-1})(i,j) > 0$$

for all $i,j \in S$ and all $h_1,h_2,\ldots,h_{N-1} \in F$.

Let $h_1, h_2, \ldots, h_{N-1}$ be an arbitrary sequence of policies. For this sequence define for all $n = 0,1,\ldots,N-1$ and all $i \in S$ the subsets $S(i,n)$ of S by

$$S(i,0) := \{i\}$$

$$S(i,n) := \{j \in S \mid P(h_1) \cdots P(h_n)(i,j) > 0\} , \quad n = 1,2,\ldots,N-1 .$$

Then it has to be shown that $S(i,N-1) = S$ for all $i \in S$.
Clearly, $S(i,n) \subset S(i,n+1)$, since (by definition) $j \in S(i,n)$ implies $P(h_1) \cdots P(h_n)(i,j) > 0$ and (9.3) implies $P(h_{n+1})(j,j) > 0$, hence

$$P(h_1) \cdots P(h_{n+1})(i,j) \geq P(h_1) \cdots P(h_n)(i,j) P(h_{n+1})(j,j) > 0 .$$

It remains to be shown that the sets $S(i,n)$ are strictly increasing as long as $S(i,n) \neq S$.
Suppose $S(i,n+1) = S(i,n)$. Then we have for all $j \in S(i,n)$ and all $k \notin S(i,n)$ that $P(h_{n+1})(j,k) = 0$, otherwise $k \in S(i,n+1)$. So $S(i,n)$ is closed under $P(h_{n+1})$. Since $P(h_{n+1})$ is irreducible, this implies $S(i,n) = S$. Hence $S(i,n)$ is strictly increasing until $S(i,n) = S$, so ultimately for $n = N - 1$ one will have $S(i,n) = S$. □

So, by lemmas 9.3 and 9.4, we now know that ℓ_n converges to g_* exponentially fast. Thus f_n will be ε-optimal for n sufficiently large. The problem, however, is to recognize this. Therefore, we want that also u_n converges to g_*. From (5.35) and (5.36)

$$P^*(f_{n+1}) g_n = g_{f_{n+1}} e .$$

Define κ by

$$\kappa := \min_{i,j \in S} \min_{f \in F} P^*(f)(i,j) . \tag{9.11}$$

Clearly, from the irreducibility assumption, one has $\kappa > 0$. Then we have the following lemma.

LEMMA 9.5. *For all* $n = 0,1,\ldots$

$$u_n - \ell_n \leq \kappa^{-1}(g_* - \ell_n) .$$

PROOF. The assertion is immediate from

$$g_* e \geq g_{f_{n+1}} e = P^*(f_{n+1}) g_n \geq (1-\kappa)\ell_n e + \kappa u_n e . \qquad \square$$

Summarizing the results for the irreducible MDP one has

THEOREM 9.6.

(i) ℓ_n *converges monotonically and exponentially fast to* g_*.

(ii) u_n *converges exponentially fast, though not necessarily monotonically, to* g_*.

So, in the irreducible case and under the strong aperiodicity assumption the value-oriented method converges exponentially fast.

9.4. THE GENERAL UNICHAIN CASE

In this section the irreducibility assumption of the previous section is replaced by the weaker unichain condition. For the unichained MDP lemma 9.4 no longer holds and the constant κ, defined in (9.11), may be zero, so lemma 9.5 can no longer be used. Thus the approach will have to be different from the one in the preceding section.
First we will derive a similar lemma as lemma 9.4, which enables us to show that the span of v_n is bounded (theorem 9.10), where the span of a vector v, notation sp(v), is defined by

$$\mathrm{sp}(v) := \max_{i\in S} v(i) - \min_{i\in S} v(i) .$$

Next it is shown that the boundedness of $\mathrm{sp}(v_n)$ implies that ℓ_n converges to g_* and finally we show that there must exist a subsequence of $\{u_n\}$ which converges to g_*.
So, throughout this section the MDP under consideration is assumed to be unichained.
Define

$$\eta := \min_{i,j\in S} \min_{h_1,\ldots,h_{N-1}} \sum_{k\in S} \min \{P(h_1)\cdots P(h_{N-1})(i,k), P(h_1)\cdots P(h_{N-1})(j,k)\} . \tag{9.12}$$

Then the unichain condition and the strong aperiodicity assumption yield the following result.

LEMMA 9.7. $\eta > 0$.

This lemma states that for all $h_1,\ldots,h_{N-1}$ any two states i and j have a common successor at time N - 1. Conditions of the type $\eta > 0$ are called scrambling conditions (cf. e.g. HAJNAL [1958], MORTON and WECKER [1977] and ANTHONISSE and TIJMS [1977]), and give contraction in the span-norm.

PROOF OF LEMMA 9.7. The line of reasoning is similar to the one in the proof of lemma 9.4. Again, let $h_1,\ldots,h_{N-1}$ be an arbitrary sequence of policies and define S(i,n), n = 0,1,...,N-1, as in the proof of lemma 9.4. Then, clearly, $S(i,n) \subset S(i,n+1)$, and if $S(i,n) = S(i,n+1)$, then S(i,n) is closed under $P(h_{n+1})$. Now it has to be shown that $S(i,N-1) \cap S(j,N-1)$ is nonempty for all pairs $i,j \in S$.
Suppose $S(i,N-1) \cap S(j,N-1)$ is empty. Then S(i,N-1) and S(j,N-1) are both proper subsets of S, so there must exist numbers m and n, $m,n < N-1$, such that $S(i,m) = S(i,m+1)$ and $S(j,n) = S(j,n+1)$. But this implies that S(i,m) is closed under $P(h_{m+1})$ and that S(j,n) is closed under $P(h_{n+1})$, and since $S(i,N-1) \cap S(j,N-1)$ is empty, $S(i,m) \cap S(j,n)$ is also empty. So, let f be a policy with $f(s) = h_{m+1}(s)$ for $s \in S(i,m)$ and $f(s) = h_{n+1}(s)$ for $s \in S(j,n)$, then P(f) has at least two disjoint, nonempty closed subchains: S(i,m) and S(j,n), which contradicts the unichain condition. Hence $S(i,N-1) \cap S(j,N-1)$ is nonempty, or

$$\sum_{k\in S} \min \{P(h_1)\cdots P(h_{N-1})(i,k)\ ,\ P(h_1)\cdots P(h_{N-1})(j,k)\} > 0\ .$$

Since S and F are finite, also $\eta > 0$. □

Further we have

LEMMA 9.8. *For all* $v \in \mathbb{R}^N$ *and all* $h_1,\ldots,h_{N-1} \in F$ *we have*

$$(9.13)\qquad sp(P(h_1)\cdots P(h_{N-1})v) \leq (1-\eta)\,sp(v)\ .$$

PROOF. Let i and j be a maximal and minimal component of $P(h_1)\cdots P(h_{N-1})v$, respectively. Then, writing Q instead of $P(h_1)\cdots P(h_{N-1})$, we have

$$\begin{aligned} sp(Qv) &= (Qv)(i) - (Qv)(j) = \sum_{k\in S} [Q(i,k) - Q(j,k)]v(k) \\ &= \sum_{k\in S} [Q(i,k) - \min\{Q(i,k),Q(j,k)\}]v(k) + \\ &\qquad - \sum_{k\in S} [Q(j,k) - \min\{Q(i,k),Q(j,k)\}]v(k) \leq \end{aligned}$$

$$\leq (1-\eta)\max_{k\in S} v(k) - (1-\eta)\min_{k\in S} v(k) = (1-\eta)sp(v) \ . \qquad \square$$

Define K_0 by

$$K_0 := \max_{i,a} r(i,a) - \min_{i,a} r(i,a) \ ,$$

then one has the following lemma.

LEMMA 9.9. *For all* $v \in \mathbb{R}^N$ *and all* $h_1,\ldots,h_{N-1} \in F$ *we have*

$$sp(L(h_1)\cdots L(h_{N-1})v) \leq (N-1)K_0 + (1-\eta)sp(v) \ .$$

PROOF. By definition of K_0 one has $sp(r(f)) \leq K_0$ for all $f \in F$. Further, $sp(P(f)v) \leq sp(v)$ for all $v \in \mathbb{R}^N$ and $f \in F$. So

$$\begin{aligned} sp(L(h_1)\cdots L(h_{N-1})v) &= sp(r(h_1) + P(h_1)r(h_2) + \\ &\quad + \ldots + P(h_1)\cdots P(h_{N-2})r(h_{N-1}) + P(h_1)\cdots P(h_{N-1})v) \\ &\leq sp(r(h_1)) + \ldots + sp(P(h_1)\cdots P(h_{N-2})r(h_{N-1})) + \\ &\quad + sp(P(h_1)\cdots P(h_{N-1})v) \\ &\leq (N-1)K_0 + (1-\eta)sp(v) \ . \end{aligned} \qquad \square$$

In order to prove that $sp(v_n)$ is bounded, we introduce the following notation

$$w_{k\lambda+p} = L^p(f_{k+1})v_k \ , \quad k = 0,1,\ldots \text{ and } p = 0,1,\ldots,\lambda-1 \ . \tag{9.14}$$

Then it follows from lemma 9.9 that for all $\ell = 0,1,\ldots$ and all $q = 0,1,\ldots,N-2$,

$$\begin{aligned} sp(w_{\ell(N-1)+q}) &\leq (N-1)K_0 + (1-\eta)sp(w_{(\ell-1)(N-1)+q}) \leq \\ &\leq \ldots \leq (N-1)K_0 + (1-\eta)(N-1)K_0 + \ldots + (1-\eta)^{\ell-1}(N-1)K_0 + \\ &\quad + (1-\eta)^{\ell} sp(w_q) \ . \end{aligned}$$

Further, it follows from the proof of lemma 9.9 that

$$sp(w_q) = L^q(f_1)v_0 \leq qK_0 + sp(v_0) \leq (N-1)K_0 + sp(v_0) \ , \quad q = 0,1,\ldots,N-2 \ .$$

So, for all $\ell = 0,1,\ldots$ and all $q = 0,1,\ldots,N-2$

$$sp(w_{\ell(N-1)+q}) \leq \eta^{-1}(N-1)K_0 + sp(v_0) .$$

Hence

$$(9.15) \qquad sp(w_m) \leq \eta^{-1}(N-1)K_0 + sp(v_0) , \quad m = 0,1,\ldots .$$

This yields

THEOREM 9.10. $sp(v_n)$ *is bounded.*

PROOF. Immediately from $v_n = w_{n\lambda}$ and (9.15). □

Before this can be used to prove that $\ell_* := \lim_{n\to\infty} \ell_n$ is equal to g_*, we have to derive a number of inequalities. For g_n, defined by (9.7), one has (9.8) and (9.9) and also

$$v_{n+1} - v_n = L^{\lambda}(f_{n+1})v_n - v_n = L^{\lambda}(f_{n+1})v_n - L^{\lambda-1}(f_{n+1})v_n + \ldots - v_n$$

$$= [P^{\lambda-1}(f_{n+1}) + \ldots + P(f_{n+1}) + I]g_n , \quad n = 0,1,\ldots .$$

So,

$$(9.16) \qquad v_{n+k} - v_n = \sum_{t=n}^{n+k-1} [P^{\lambda-1}(f_{t+1}) + \ldots + P(f_{t+1}) + I]g_t ,$$

$n = 0,1,\ldots$, $k = 1,2,\ldots$.

Let us consider $v_{m+q} - v_m$, where m and q are arbitrary for the time being. From (9.9) we obtain, with the strong aperiodicity assumption, that for all n and k (with α as in (9.3))

$$g_{n+k}(i) \geq \alpha^{k\lambda} g_n(i) + (1 - \alpha^{k\lambda})\ell_n ,$$

and

$$(9.17) \qquad g_{n+k}(i) \geq \alpha^{k\lambda-p}(P^p(f_{n+1})g_n)(i) + (1 - \alpha^{k\lambda-p}) \min_{j\in S} (P^p(f_{n+1})g_n)(j)$$

$$\geq \alpha^{k\lambda-p}(P^p(f_{n+1})g_n)(i) + (1 - \alpha^{k\lambda-p})\ell_n$$

$$\geq \alpha^{k\lambda}(P^p(f_{n+1})g_n)(i) + (1 - \alpha^{k\lambda})\ell_n , \quad p = 0,1,\ldots,\lambda-1 .$$

Now suppose that $i_0 \in S$ satisfies

$$g_{m+q}(i_0) = \ell_{m+q} \quad (\leq \ell_*) \ .$$

Then (9.17), with $n+k = m+q$ and $n = t$ $(m \leq t < m+q)$, yields

$$(9.18) \qquad (P^p(f_{t+1})g_t)(i_0) \leq \alpha^{-\lambda(m+q-t)}[g_{m+q}(i_0) - (1-\alpha^{\lambda(m+q-t)})\ell_t]$$

$$\leq \alpha^{-\lambda(m+q-t)}[g_{m+q}(i_0) - (1-\alpha^{\lambda(m+q-t)})\ell_m]$$

$$= \ell_m + \alpha^{-\lambda(m+q-t)}[g_{m+q}(i_0) - \ell_m]$$

$$\leq \ell_m + \alpha^{-\lambda q}[\ell_* - \ell_m] \ , \quad p = 0,1,\ldots,\lambda-1 \ .$$

Hence, with (9.16) and (9.18),

$$(9.19) \qquad (v_{m+q} - v_m)(i_0) \leq q\lambda\ell_m + q\lambda\alpha^{-q\lambda}(\ell_* - \ell_m) \ .$$

On the other hand we have $u_n \geq g_*$ for all n. Hence there must exist a state $j_0 \in S$ which has $g_{m+k}(j_0) \geq g_*$ for at least $N^{-1}q$ of the indices $m+k \in \{m,m+1,\ldots,m+q-1\}$.
So, for this state j_0,

$$(9.20) \qquad (v_{m+q} - v_m) \geq N^{-1} q\, g_* + (q\lambda - N^{-1}q)\ell_m \geq q\lambda\ell_m + N^{-1}q(g_* - \ell_*) \ .$$

Then it follows from (9.19) and (9.20) that

$$(9.21) \qquad sp(v_{m+q} - v_m) \geq N^{-1}q(g_* - \ell_*) - q\lambda\alpha^{-q\lambda}(\ell_* - \ell_m) \ .$$

Now we are ready to prove

THEOREM 9.11. $\ell_* = g_*$.

PROOF. Clearly $\ell_* \leq g_*$. Assume $\ell_* < g_*$. By theorem 9.10 there exists a constant K_1 such that

$$(9.22) \qquad sp(v_n) \leq K_1 \quad \text{for all } n = 0,1,\ldots \ .$$

Now choose q such that $N^{-1}q(g_* - \ell_*) \geq 2K_1 + K_2$, where K_2 is some positive constant. Next, choose m such that $q\lambda\alpha^{-q\lambda}(\ell_* - \ell_m) < K_2$. Then, it follows from (9.21) that

$$sp(v_{m+q} - v_m) > 2K_1 + K_2 - K_2 = 2K_1 \ .$$

Hence, using (9.22) with $n = m$,

$$sp(v_{m+q}) \geq sp(v_{m+q} - v_m) - sp(v_m) > 2K_1 - K_1 = K_1 ,$$

which contradicts (9.22) for $n = m+q$.
Therefore we must have $\ell_* = g_*$. □

So, we now know that ℓ_n converges to g_*, and, by (9.6), that f_n becomes nearly optimal if n becomes large. In order to be able to recognize that f_n is nearly optimal one needs (at least) the following result.

THEOREM 9.12. *g_* is a limitpoint of the sequence $\{u_n\}$.*

PROOF. We know that $u_n \geq g_*$. Further it follows from the boundedness of $sp(v_n)$ (theorem 9.10) that also $\{sp(g_n)\}$ or $\{u_n - \ell_n\}$ is bounded. Hence also $\{u_n\}$ is bounded. Now, suppose the smallest limitpoint of $\{u_n\}$ to be strictly larger than g_*. Then one may construct, using a similar reasoning as in (9.20) and in the proof of theorem 9.11, a violation of the boundedness of $\{sp(v_n)\}$. Hence g_* is a limitpoint of $\{u_n\}$. □

So, if all P(f) are unichained and the strong aperiodicity assumption holds, then we see, from theorems 9.11 and 9.12 and from (9.6), that the method of value-oriented standard successive approximations converges. I.e., the method yields an approximation of the gain g^* of the MDP and nearly-optimal stationary strategies.

In the next section it will be shown that g_* is not only a limitpoint of $\{u_n\}$, but that u_n converges to g_*, exponentially fast.

9.5. GEOMETRIC CONVERGENCE FOR THE UNICHAIN CASE

For the irreducible case we have obtained that $sp(g_n)$ converges to zero geometrically. For the general unichain case it has only been shown that there exists a subsequence of $\{g_n\}$ for which $sp(g_n)$ converges to zero.
In this section it will be shown that also in the unichain case $sp(g_n)$ converges to zero exponentially fast.
So, the MDP under consideration is again assumed to be unichained.

Since $sp(v_n)$ is bounded, also $v_n - v_n(N)e$ is bounded. Further, g_*e is a limitpoint of $\{g_n\}$. And since there are only finitely many policies, there

exists a subsequence of $\{v_n\}$ (and $\{g_n\}$) with $g_{n_k} \to g_*e$, $f_{n_k+1} = f$ and $v_{n_k} - v_{n_k}(N)e \to v$ $(k \to \infty)$ for some $f \in F$ and $v \in \mathbb{R}^N$.
Then for all k

$$\max_{h\in F} L(h)v_{n_k} - v_{n_k} = L(f_{n_k+1})v_{n_k} - v_{n_k} = L(f)v_{n_k} - v_{n_k} = g_{n_k} .$$

Letting k tend to infinity yields

$$\max_{h\in F} L(h)v - v = L(f)v - v = g_*e , \tag{9.23}$$

where it has been used that for all $h \in F$,

$$L(h)v_{n_k} - v_{n_k} = L(h)(v_{n_k} - v_{n_k}(N)e) - (v_{n_k} - v_{n_k}(N)e) .$$

Then we have the following lemma.

LEMMA 9.13. *Let* $\varepsilon > 0$ *be such, that* $L(h)v - v \geq g_*e - \varepsilon e$ *implies* $L(h)v - v = g_*e$ *(clearly such an* ε *exists by the finiteness of* F*). Then*

$$sp(v_n - v) \leq \varepsilon \quad \textit{and} \quad L(f_{n+1})v = v + g_*e$$

imply

$$sp(v_{n+1} - v) \leq \varepsilon \quad \textit{and} \quad L(f_{n+2})v = v + g_*e .$$

Before this lemma is proved, note the following. Since $v_{n_k} - v_{n_k}(N)e - v$ tends to zero, there exists a number m such that $sp(v_m - v) \leq \varepsilon$ and $L(f_m)v = v + g_*e$. Then, as a consequence of lemma 9.13,

$$sp(v_n - v) \leq \varepsilon \quad \text{and} \quad L(f_n)v = v + g_*e \quad \text{for all } n \geq m .$$

But that implies for all q = 1,2,...

$$\begin{aligned} v_{m+q} &= L^\lambda(f_{m+q}) \cdots L^\lambda(f_{m+1})v_m \\ &= L^\lambda(f_{m+q}) \cdots L^\lambda(f_{m+1})v + P^\lambda(f_{m+q}) \cdots P^\lambda(f_{m+1})(v_m - v) \\ &= v + q\lambda g_*e + P^\lambda(f_{m+q}) \cdots P^\lambda(f_{m+1})(v_m - v) . \end{aligned} \tag{9.24}$$

So, by lemma 9.8, $sp(v_{m+q} - v)$ decreases in q exponentially fast to zero. And also g_{m+q} converges to g_*e exponentially fast, since

$$g_{m+q} = L(f_{m+q+1})v_{m+q} - v_{m+q}$$

$$= L(f_{m+q+1})v_{m+q} - L(f_{m+q+1})v + v + g_*e - v_{m+q}$$

$$= [P(f_{m+q+1}) - I](v_{m+q} - v) + g_*e \ .$$

PROOF OF LEMMA 9.13.

$$v_{n+1} = L^{\lambda}(f_{n+1})v_n = L^{\lambda}(f_{n+1})v + P^{\lambda}(f_{n+1})(v_n - v)$$

$$= v + \lambda g_*e + P^{\lambda}(f_{n+1})(v_n - v) \ .$$

So,

$$sp(v_{n+1} - v) = sp(P^{\lambda}(f_{n+1})(v_n - v)) \leq sp(v_n - v) \leq \varepsilon \ .$$

And

$$L(f_{n+2})v - v = L(f_{n+2})v_{n+1} + P(f_{n+2})(v - v_{n+1}) - v$$

$$\geq L(f)v_{n+1} + P(f_{n+2})(v - v_{n+1}) - v$$

$$= L(f)v + P(f)(v_{n+1} - v) + P(f_{n+2})(v - v_{n+1}) - v$$

$$= g_*e + [P(f) - P(f_{n+2})](v_{n+1} - v) \geq g_*e - \varepsilon e \ ,$$

since for any two stochastic matrices P_1 and P_2 and for any $w \in \mathbb{R}^N$ one has $(P_1 - P_2)w \geq -sp(w)$. Hence also

$$L(f_{n+2})v - v = g_*e \ .$$

□

9.6. THE COMMUNICATING CASE

In section 4 the convergence proof for the unichain case has been given in two stages. First the unichain assumption and the strong aperiodicity assumption were used to prove that $sp(v_n)$ is bounded (lemmas 9.7-9.9 and theorem 9.10). And in the second stage we used the boundedness of $\{sp(v_n)\}$ and $u_n \geq g_*$ to prove that $\ell_n \to g_*$ and that $u_{n_k} \to g_*$ $(k \to \infty)$ for some subsequence $\{u_{n_k},\ k = 0,1,\ldots\}$. From this it will be clear that the method of value-oriented successive approximations will converge whenever $\{sp(v_n)\}$ is bounded and the gain of the MDP is independent of the initial state (if the strong aperiodicity assumption holds).

In this section the communicating MDP will be considered.
An MDP is called *communicating* if there exists for any two states i and j a policy f and a number r such that $P^r(f)(i,j) > 0$ (cf. BATHER [1973]).
Many practical problems are communicating, but need not be unichained. On the other hand, an MDP may be unichained but not communicating since some states may be transient under all policies.
Throughout this section the MDP considered is assumed to be communicating.
Clearly, if the MDP is communicating, the gain is independent of the initial state. We will show that also in the communicating case $\{sp(v_n)\}$ is bounded. Therefore, define

$$\hat{K} := \max_{i,a} |r(i,a)| ,$$

$$L_n := \min_{i\in S} v_n(i) , \quad U_n := \max_{i\in S} v_n(i) , \quad n = 0,1,\ldots ,$$

$$\theta := \min_{i,j\in S} \min_{a\in A} \{p(i,a,j) \mid p(i,a,j) > 0\} .$$

In order to prove that $\{sp(v_n)\}$ is bounded we need the following lemmas.

LEMMA 9.15. *For all* n = 0,1,...

(i) $L_{n+1} \geq L_n - \lambda\hat{K}$,

(ii) $U_{n+1} \leq U_n + \lambda\hat{K}$.

PROOF.
(i) For all n = 0,1,...

$$v_{n+1} = L^{\lambda}(f_{n+1})v_n = r(f_{n+1}) + P(f_{n+1})r(f_{n+1}) +$$
$$+ \ldots + P^{\lambda-1}(f_{n+1})r(f_{n+1}) + P^{\lambda}(f_{n+1})v_n$$
$$\geq - \lambda\hat{K}e + P^{\lambda}(f_{n+1})v_n \geq - \lambda\hat{K}e + L_n e .$$

Hence also

$$L_{n+1} \geq - \lambda\hat{K} + L_n .$$

Similarly one obtains (ii). □

LEMMA 9.16. *If* $sp(v_{n+N-1}) \geq sp(v_n)$, *then for all* m *with* $n \leq m < n+N-1$,

$$L_{m+1} - L_m \leq \lambda\hat{K}(2N - 3) .$$

PROOF. From lemma 9.15 we obtain

$$\begin{aligned} sp(v_{n+N-1}) &= U_{n+N-1} - L_{n+N-1} \\ &= \sum_{k=n}^{n+N-2} [(U_{k+1} - U_k) - (L_{k+1} - L_k)] + U_n - L_n \\ &= \sum_{k=n}^{n+N-2} (U_{k+1} - U_k) - \sum_{\substack{k=n \\ k\neq m}}^{n+N-2} (L_{k+1} - L_k) - (L_{m+1} - L_m) + sp(v_n) \\ &\leq \lambda\hat{K}(N-1) + \lambda\hat{K}(N-2) - (L_{m+1} - L_m) + sp(v_n) \\ &= \lambda\hat{K}(2N-3) - (L_{m+1} - L_m) + sp(v_n) \ . \end{aligned}$$

Hence, if $sp(v_{n+N-1}) \geq sp(v_n)$, then $L_{m+1} - L_m \leq \lambda\hat{K}(2N-3)$. □

LEMMA 9.17. *If* $sp(v_{n+N-1}) \geq sp(v_n)$ *and* $v_{m+1}(i) \leq C + L_{m+1}$ *for some* $i \in S$, *for some constant* C *and some* m *with* $n \leq m < n+N-1$, *then*

$$v_m(j) \leq L_m + \alpha^{1-\lambda}\theta^{-1}[C + 2\lambda\hat{K}(N-1)] \ ,$$

for all $j \in S$ *for which an action* $a \in A$ *with* $p(i,a,j) > 0$ *exists* (α *is the constant in* (9.3)).

PROOF. For all $m = 0,1,\ldots$

$$v_{m+1} = L^{\lambda}(f_{m+1})v_m = L^{\lambda-1}(f_{m+1})Uv_m$$

and

$$Uv_m \geq -\hat{K}e + \max_{f\in F} P(f)v_m \ .$$

So,

$$\begin{aligned} v_{m+1}(i) &\geq (L^{\lambda-1}(f_{m+1}))(-\hat{K}e + \max_{f\in F} P(f)v_m)(i) \\ &= -\hat{K} + (L^{\lambda-1}(f_{m+1}) \max_{f\in F} P(f)v_m)(i) \\ &\geq -\lambda\hat{K} + (P^{\lambda-1}(f_{m+1}) \max_{f\in F} P(f)v_m)(i) \\ &\geq -\lambda\hat{K} + \alpha^{\lambda-1} \max_{a\in A} \sum_{k\in S} p(i,a,k)v_m(k) + (1-\alpha^{\lambda-1})L_m \ . \end{aligned}$$

Thus also

$$C + L_{m+1} \geq v_{m+1}(i) \geq -\lambda\hat{K} + \alpha^{\lambda-1} \max_{a\in A} \sum_{k\in S} p(i,a,k) v_m(k) + (1-\alpha^{\lambda-1}) L_m \ .$$

Then lemma 9.16 yields

$$C + \lambda\hat{K}(2N-3) + L_m \geq -\lambda\hat{K} + \alpha^{\lambda-1} \max_{a\in A} \sum_{k\in S} p(i,a,k) v_m(k) + (1-\alpha^{\lambda-1}) L_m \ .$$

Or,

$$\max_{a\in A} \sum_{k\in S} p(i,a,k)\,(v_m(k) - L_m) \leq \alpha^{1-\lambda}[C + 2\lambda\hat{K}(N-1)] \ .$$

Hence, if for some $j \in S$ and $a \in A$ we have $p(i,a,j) > 0$, so $p(i,a,j) \geq \theta$, then certainly

$$\theta(v_m(j) - L_m) \leq \alpha^{1-\lambda}[C + 2\lambda\hat{K}(N-1)] \ ,$$

which proves this lemma. □

Next we show that, if $sp(v_{n+N-1}) \geq sp(v_n)$, then $sp(v_n)$ cannot be arbitrarily large.

Define

$$C_0 := 0 \ ,$$

$$C_n := \alpha^{1-\lambda}\,\theta^{-1}[C_{n-1} + 2\lambda\hat{K}(N-1)] \ , \quad n = 1,2,\ldots,N-1 \ .$$

Then the following lemma holds.

LEMMA 9.18. *If* $sp(v_{n+N-1}) \geq sp(v_n)$, *then* $sp(v_n) \leq C_{N-1}$.

PROOF. Let $i \in S$ be such that $v_{n+N-1}(i) = L_{n+N-1}$, and define the sets $S(i,t)$, $t = 0,1,\ldots,N-1$, by

$$S(i,0) := \{i\} \ ,$$

$$S(i,t+1) := \{j \in S \mid \text{there exists a state } j \in S(i,t) \text{ and an action } a \in A \text{ such that } p(j,a,k) > 0\} \ ,$$

$(t = 0,1,\ldots,N-2)$.

From $p(j,a,j) \geq \alpha > 0$ for all $j \in S$ and $a \in A$ we have $S(i,t) \subset S(i,t+1)$. Further it follows from the communicatingness that ultimately $S(i,N-1) = S$ (cf. the proof of lemma 9.7). Then lemma 9.17 yields (with $C = 0$)

$$v_{n+N-2}(j) - L_{n+N-2} \leq C_1 \quad \text{for all } j \in S(i,1) .$$

Next we obtain with lemma 9.17

$$v_{n+N-3}(j) - L_{n+N-3} \leq C_2 \quad \text{for all } j \in S(i,2) .$$

Continuing in this way we get

$$v_n(j) - L_n \leq C_{N-1} \quad \text{for all } j \in S(i,N-1) = S .$$

Hence

$$sp(v_n) \leq C_{N-1} . \qquad \square$$

Finally, it can be shown that also in the communicating case $\{sp(v_n)\}$ is bounded.

THEOREM 9.19.

$$sp(v_t) \leq \max \{sp(v_0) + 2\lambda\hat{K}(N-2) , C_{N-1} + 2\lambda\hat{K}(N-1)\} \quad \textit{for all } t = 0,1,\ldots .$$

PROOF. For all n we either have $sp(v_{n+N-1}) < sp(v_n)$ or $sp(v_{n+N-1}) \geq sp(v_n)$. But, if $sp(v_{n+N-1}) \geq sp(v_n)$, then (by lemma 9.18) $sp(v_n) \leq C_{N-1}$, and thus with repeated application of lemma 9.15

$$sp(v_{n+N-1}) = U_{n+N-1} - L_{n+N-1} \leq U_{n+N-2} - L_{n+N-2} + 2\lambda\hat{K}$$

$$\leq \ldots \leq U_n - L_n + 2\lambda\hat{K}(N-1) = sp(v_n) + 2\lambda\hat{K}(N-1) .$$

Hence for all n

$$sp(v_{n+N-1}) \leq \max \{sp(v_n) , C_{N-1} + 2\lambda\hat{K}(N-1)\} ,$$

which immediately yields for all $t = p + q(N-1)$, $q = 0,1,\ldots$, and $p = 0,1,\ldots,N-2$

$$sp(v_t) \leq \max \{sp(v_p) , C_{N-1} + 2\lambda\hat{K}(N-1)\} .$$

Finally,

$$sp(v_p) \leq sp(v_0) + 2\lambda\hat{K}p \leq sp(v_0) + 2\lambda\hat{K}(N-2) , \quad p = 0,1,\ldots,N-2$$

gives for all $t = 0,1,\ldots$

$$sp(v_t) \leq \max \{sp(v_0) + 2\lambda\hat{K}(N-2) , C_{N-1} + 2\lambda\hat{K}(N-1)\} . \qquad \square$$

So, also in the communicating case, $sp(v_n)$ is bounded. And, as has been argued at the beginning of this section, that implies that the value-oriented method converges, i.e., yields bounds on g_* and nearly-optimal stationary strategies.

As in section 5, it can be shown that g_* is not only a limitpoint of $\{u_n\}$ but that u_n converges to g_*. (One may verify that lemma 9.13 also holds in the communicating case.) However, since in the communicating case (9.13) not necessarily holds, we cannot conclude that the convergence is again geometric. It is evident that, if there is a unique policy satisfying $L(f)v = v + g_*e$, then g_n converges to g_*e exponentially fast. One might conjecture that the rate of convergence of g_n to g_*e is always geometric.

9.7. *SIMPLY CONNECTEDNESS*

A weaker condition, which still assumes that the gain of the MDP is independent of the initial state and, as will be shown, that $\{sp(v_n)\}$ is bounded, is the condition of simply connectedness, introduced by PLATZMAN [1977]. Platzman used this condition to prove the convergence of the method of standard successive approximations.
An MDP is called *simply connected* if the state space S is the union of two sets S° and $\bar{S}$, where S° is a communicating class (i.e., all states in S° can be reached from one another) and $\bar{S}$ is transient under any policy.
Observe that, if the MDP is simply connected, the gain is again independent of the initial state. In order to prove that simply connectedness also implies that $\{sp(v_n)\}$ is bounded again, define

$$L_n^\circ := \min_{i\in S^\circ} v_n(i) , \quad L_n := \min_{i\in S} v_n(i) ,$$

$$U_n^\circ := \max_{i\in S^\circ} v_n(i) , \quad U_n := \max_{i\in S} v_n(i) .$$

Clearly, S° is closed under any $P(f)$. Further, let k be the minimal value of ℓ for which $\ell\lambda + 1$ is at least equal to the number of states in $\bar{S}$. Then for some constant $\zeta > 0$ and for all $i \in \bar{S}$ and all $h_1,\ldots,h_k \in F$

$$\sum_{j\in S^\circ} P^\lambda(h_1)\cdots P^\lambda(h_k)(i,j) \geq \zeta .$$

Hence for all $i \in \bar{S}$

$$v_{n+k}(i) \leq k\lambda\hat{K} + \zeta U_n^{\circ} + (1-\zeta)U_n \ .$$

And, since we also have for all $i \in S^{\circ}$

$$v_{n+k}(i) \leq k\lambda\hat{K} + U_n^{\circ} \leq k\lambda\hat{K} + \zeta U_n^{\circ} + (1-\zeta)U_n \ ,$$

one may conclude

$$U_{n+k} \leq k\lambda\hat{K} + \zeta U_n^{\circ} + (1-\zeta)U_n \ .$$

Similarly, one shows

$$L_{n+k} \geq -k\lambda\hat{K} + \zeta L_n^{\circ} + (1-\zeta)L_n \ .$$

Hence

$$\text{sp}(v_{n+k}) \leq 2k\lambda\hat{K} + \zeta(U_n^{\circ} - L_n^{\circ}) + (1-\zeta)\text{sp}(v_n) \ . \tag{9.25}$$

From the preceding section we know that $U_n^{\circ} - L_n^{\circ}$ is bounded, so (9.25) can be rewritten as

$$\text{sp}(v_{n+k}) \leq K^0 + (1-\zeta)\text{sp}(v_n) \ , \quad n = 0,1,\ldots \ .$$

From this one easily shows that

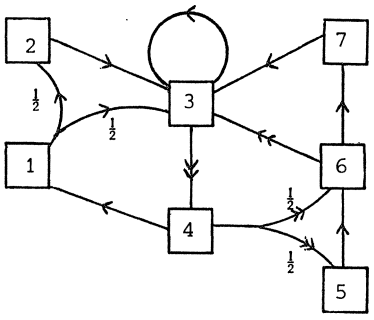
$$\text{sp}(v_n) \leq \zeta^{-1}K^0 + 2\lambda\hat{K}(k-1) + \text{sp}(v_0) \ , \ n = 0,1,\ldots \ .$$

So, indeed, if the MDP is simply connected, then $\{\text{sp}(v_n)\}$ is bounded. Since simply connectedness implies constant gain, one may argue in identically the same way as in section 6 that the value-oriented method converges. Finally, we can make the same remark as at the end of section 6 (with the condition of communicatingness replaced by simply connectedness).

9.8. *SOME REMARKS*

(i) The proofs in the preceding sections depend heavily on the strong aperiodicity assumption. One might wonder whether mere aperiodicity, as in the standard successive approximations case, would not suffice. The following example demonstrates one of the problems one can get under the weaker assumption: all P(f) are aperiodic (and unichained).

EXAMPLE 9.21. $S = \{1,2,3,4,5,6,7\}$, $A(3) = A(4) = A(6) = \{1,2\}$, $A(1) = A(2) = A(5) = A(7) = \{1\}$. So there are eight different policies which can be characterized by the triples (a_3,a_4,a_6), where a_i is the action in state i, i = 3,4,6. Clearly, P(f) is unichained and aperiodic for all f. Now let us consider, for the case $\lambda = 2$, the sequence of policies $(1,2,1),(2,1,2),(1,2,1)$, etc. Then the matrix $P^2((1,2,1))P^2((2,1,2)) =: Q$ is no longer unichained, since $Q(1,1) = Q(4,4) = 1$. This could never happen under the strong aperiodicity assumption.

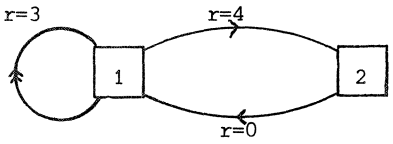


Now we will show that this feature gives difficulties for the convergence.
Choose $r(1,1) = 2$, $r(2,1) = r(3,1) = 4$, $r(3,2) = 6$, $r(4,1) = 4$, $r(4,2) = r(5,1) = 6$, $r(6,1) = 2$, $r(6,2) = r(7,1) = 0$.
Then the policies (1,2,1) and (2,1,2) both have gain 4 and the optimal gain is $4\frac{2}{7}$ for policy (2,2,2).
Choose $v_0 = (1,4,2,0,0,0,0)^T$, then, as will be shown, cycling may occur between the nonoptimal policies (1,2,1) and (2,1,2).
Computing Uv_0 yields $L((a_3,a_4,a_6))v_0 = Uv_0$ for all policies (a_3,a_4,a_6) with $a_4 = 2$. Choose among the maximizers $f_1 = (1,2,1)$, then $v_1 = L^2(f_1)v_0 = (8,10,10,10,8,4,6)^T$. Now any policy (a_3,a_4,a_6) with $a_3 = a_6 = 2$ satisfies $L((a_3,a_4,a_6))v_1 = Uv_1$. Choosing $f_2 = (2,1,2)$ we get $v_2 = (17,20,18,16,16,16,16)^T = v_0 + 16e$.
So, indeed, cycling may occur between the suboptimal policies (1,2,1) and (2,1,2) in which case ℓ_n will not converge to g_* (but $\ell_n = 2$ for all n).

In this example, however, there is some ambiguity in the choice of the maximizing policies. The question remains whether cycling may occur if we use for breaking ties the rule: "do not change an action unless there is a strictly better one".

(ii) For the method of standard successive approximations we have that $\{v_n - ng^*\}$ is bounded, even if some or all policies are periodic. The following example shows that in the value-oriented method $\{v_n - n\lambda g^*\}$ may be unbounded.

EXAMPLE 9.22. $S = \{1,2\}$, $A(1) = \{1,2\}$, $A(2) = \{1\}$, $r(1,1) = 4$, $r(1,2) = 3$, $r(2,1) = 0$, $p(1,1,2) = p(1,2,1) = p(2,1,1) = 1$. For the case $\lambda = 2$, $v_0 = 0$, one has $L(f_1)v_0 = Uv_0$ for the policy with $f(1) = 1$. Thus $v_1 = 4e$, and $v_n = 4ne$. Since $g_* = 3$, we have $v_n - \lambda n g_* e = -ne$ which is clearly unbounded.



(iii) We conjecture that the value-oriented method always converges if g^* is independent of the initial state (provided the strong aperiodicity assumption holds).

(iv) Instead of choosing λ as a constant in advance one may use a different λ in each iteration. Probably it is sensible to start with small values of λ and to let λ increase if $sp(g_n)$ decreases.

CHAPTER 10

INTRODUCTION TO THE TWO-PERSON ZERO-SUM MARKOV GAME

In the MDP model there is one decision maker earning rewards from a system he (partly) controls. In many real-life situations, however, there are several decision makers having conflicting interests, e.g. in economics and disarmament. Such decision problems can be modeled as so-called Markov games. A special case of these Markov games (MG's) is the two-person zero-sum MG introduced by SHAPLEY [1953]. In this game there are two decision makers who have completely opposite interests. Shapley called these games stochastic games. The term Markov game stems from ZACHRISSON [1964]. An elementary treatment of two-person zero-sum MG's can be found in Van der WAL and WESSELS [1976].

Chapters 10-13 deal with two-person zero-sum MG's. In this introductory chapter first (section 1) the model of the two-person zero-sum MG is formulated. Next (in section 2) the finite-stage MG is treated, and it is shown that one may again restrict the attention to history-independent stategies. In section 3 a two-person nonzero-sum MG is considered. It is shown that in nonzero-sum games the restriction to history-independent strategies is sometimes rather unrealistic. Section 4 contains an introduction to the infinite-horizon MG and summarizes the contents of chapters 11-13.

10.1. THE MODEL OF THE TWO-PERSON ZERO-SUM MARKOV GAME

Informally, the model of the two-person zero-sum MG has already been formulated in section 1.1. Formally, the MG can be introduced along similar lines as the MDP in section 1.5.

The two-person zero-sum MG is characterized by the following objects: A nonempty finite or countably infinite set S, finite nonempty sets A and B, a function $p: S \times A \times B \times S \to [0,1]$ with $\sum_{j \in S} p(i,a,b,j) = 1$ for all

$(i,a,b) \in S \times A \times B$, and a function $r: S \times A \times B \to \mathbb{R}$. We think of S as the *state space* of some dynamical system which is controlled at discrete points in time, $t = 0,1,\ldots$, say, and of A and B as the *action sets* for player I and player II, respectively. At each time t the two players, having observed the present state of the system (as well as all preceding states and previously taken actions) simultaneously choose an action from the sets A and B, respectively. As a result of the chosen actions, a by player I and b by player II, the system moves to state j with probability $p(i,a,b,j)$ and player I receives from player II a (possibly negative) amount $r(i,a,b)$. The function p is called the *transition law* and the function r the *reward function*.

Similar as in section 1.5 the sets of strategies for the two players can be defined.
Define the sets of histories of the system:

$$H_0 := S \ , \quad H_n := (S \times A \times B)^n \times S \ , \quad n = 1,2,\ldots \ .$$

Then a *strategy* π for player I is any sequence $\pi_0,\pi_1,\ldots$ such that π_n is a transition probability from H_n into A. So for each history $h_n \in H_n$ the function π_n determines the probabilities $\pi_n(\{a\} \mid h_n)$ that action a will be chosen at time n if h_n is the history of the system upto time n. The set of history-dependent strategies for player I is denoted by Π.
Similarly we can define a strategy γ for player II. The set of strategies for player II is denoted by Γ.
In the case of the MDP a very important role has been played by the pure Markov strategies. In the game-situation it is clear that in general one can not restrict the attention to pure Markov strategies, since already in the matrix game one has to consider randomized actions. In the MG the role of the pure Markov strategies in the MDP is played by the randomized Markov strategies.
Since no concepts are needed for pure strategies we will use the following definitions and notations.
A *policy* f for player I is any function from $S \times A$ into $[0,1]$ satisfying $\sum_{a \in A} f(i,a) = 1$ for all $i \in S$. The set of all policies is denoted by F_I.
Similarly, a policy h for player II is any map from $S \times B$ into $[0,1]$ with $\sum_{b \in B} h(i,b) = 1$, $i \in S$. The set of policies for player II is denoted by F_{II}.

A strategy π for player I is called a *randomized Markov strategy* or shortly *Markov strategy* if the probabilities $\pi_n(\{a\} \mid h_n)$ depend on h_n only through the present state. So a Markov strategy for player I is completely characterized by the policies f_n satisfying $\pi_n(\{a\} \mid (i_1,\ldots,i_n)) = f_n(i_n,a)$, $n = 0,1,\ldots$, $a \in A$ and $(i_0,\ldots,i_n) \in H_n$. Mostly we write $\pi = (f_0,f_1,\ldots)$. The set of all Markov strategies for player I is denoted by M_I. Similarly, one defines the set M_{II} of Markov strategies for player II.
Finally, a *stationary strategy* for player I is any strategy $\pi = (f,f_1,f_2,\ldots)$ with $f_n = f$ for all $n = 1,2,\ldots$; notation $f^{(\infty)}$, or - if no confusion will arise - f. Similar for player II.

As in section 1.5, any initial state $i \in S$ and any pair of strategies $\pi \in \Pi$, $\gamma \in \Gamma$, define a probability measure on $(S \times A \times B)^\infty$, denoted by $\mathbb{P}_{i,\pi,\gamma}$, and a stochastic process $\{(X_n,A_n,B_n),\ n = 0,1,\ldots\}$, where X_n is the state of the system and A_n and B_n are the actions chosen at time n by players I and II, respectively. The expectation with respect to $\mathbb{P}_{i,\pi,\gamma}$ is denoted by $\mathbb{E}_{i,\pi,\gamma}$.

10.2. THE FINITE-STAGE MARKOV GAME

This section deals with the finite-horizon MG. It will be shown that - as in the case of the finite-stage MDP - this game can be treated by a dynamic programming approach.
The n-period MG is played as follows: the two players are controlling the system at times 0,1 upto n - 1 only, and if - as a result of the actions at time n - 1 - the system reaches state j at time n, then player I receives a final payoff v(j), $j \in S$, from player II and the game terminates.
This game will be called the n-*stage Markov game with terminal payoff* v.
The total expected n-stage reward for player I in this game, when the initial state is i and strategies π and γ are played is defined by

$$(10.1) \qquad v_n(i,\pi,\gamma,v) := \mathbb{E}_{i,\pi,\gamma}\left[\sum_{k=0}^{n-1} r(X_k,A_k,B_k) + v(X_n)\right],$$

provided the expectation at the right-hand side is properly defined.
The reward for player II is equal to $-v_n(i,\pi,\gamma,v)$.
The ensure that the expectation in (10.1) is properly defined, we make the following assumption.

CONDITION 10.1. *For all* $\pi \in \Pi$ *and* $\gamma \in \Gamma$,

(i) $$\mathbb{E}_{i,\pi,\gamma} \sum_{k=0}^{n-1} r^+(X_k,A_k,B_k) < \infty \ , \quad i \in S \ ,$$

(ii) $$\mathbb{E}_{i,\pi,\gamma} v^+(X_k) < \infty \ , \quad k = 1,2,\ldots,n \ , \quad i \in S \ .$$

Strictly speaking we need condition 10.1(ii) for k = n only. However, if one wants to use a dynamic programming approach, then (ii) is needed also for k = 1,...,n-1.

Our aim is to show that the n-stage MG with terminal payoff v has a value, i.e., that for each $i \in S$ a real number $v_n(i,v)$ exists such that

(10.2) $$\sup_{\pi\in\Pi} \inf_{\gamma\in\Gamma} v_n(i,\pi,\gamma,v) = \inf_{\gamma\in\Gamma} \sup_{\pi\in\Pi} v_n(i,\pi,\gamma,v) =: v_n(i,v) \ .$$

This number $v_n(i,v)$ is called the value of the game.
Further we show that player I has an optimal Markov strategy, i.e., a strategy $\pi^{(n)}$ satisfying

(10.3) $$v_n(\pi^{(n)},\gamma,v) \geq v_n(v) \quad \text{for all } \gamma \in \Gamma \ ,$$

and that for all $\varepsilon > 0$ player II has an ε-optimal Markov strategy, i.e., a strategy $\gamma^{(n)}$ satisfying

(10.4) $$v_n(\pi,\gamma^{(n)},v) \leq v_n(v) + \varepsilon e \quad \text{for all } \pi \in \Pi \ .$$

We will see later what causes the asymmetry in (10.3) and (10.4).
The value as well as the (nearly-) optimal Markov strategies will be determined by a dynamic programming scheme. The approach is very similar to the one in section 2.3 for the finite-stage MDP.

First let us introduce a few more notations.
For any pair of policies $f \in F_I$, $h \in F_{II}$, define the immediate reward function r(f,h) by

(10.5) $$r(f,h)(i) := \sum_{a\in A} \sum_{b\in B} f(i,a)h(i,b)r(i,a,b) \ , \quad i \in S \ .$$

Further, define the operators P(f,h), L(f,h) and U on suitable subsets of $\bar{V}$ (cf. (1.15)) by

$$(10.6)\qquad (P(f,h)w)(i) := \sum_{a\in A}\sum_{b\in B} f(i,a)h(i,b) \sum_{j\in S} p(i,a,b,j)w(j) ,$$

$$i \in S , \quad f \in F_I , \quad h \in F_{II} ,$$

$$(10.7)\qquad L(f,h)w := r(f,h) + P(f,h)w , \quad f \in F_I, \ h \in F_{II} ,$$

and

$$(10.8)\qquad Uw := \max_{f\in F_I} \inf_{h\in F_{II}} L(f,h)w .$$

The operator U defined in (10.8) plays the same role in the analysis of the MG as the operator U defined in (1.28) does in the MDP. For that reason the capital U is used again. Throughout chapters 10-13 the operator U will be the one defined in (10.8), so no confusion will arise.
Observe that in (10.8) we write $\inf_{h\in F_{II}}$ instead of $\min_{h\in F_{II}}$. The reason for this is the same as the one which causes the asymmetry in (10.3) and (10.4).
Note that $(L(f,h)w)(i)$ is precisely the expected amount player I will obtain in the 1-stage game with terminal payoff v when i is the initial state and policies f by player I and h by player II are used. In fact, $(L(f,h)w)(i)$ depends of f and h only through $f(i,\cdot)$ and $h(i,\cdot)$.
Also observe that for a given initial state, i say, the 1-stage game is merely a matrix game. To solve this game one has to determine the value and optimal randomized actions for the matrix game with entries

$$(10.9)\qquad r(i,a,b) + \sum_{j\in S} p(i,a,b,j)w(j) .$$

So we see that $(Uw)(i)$ is just the value of the 1-stage game with terminal reward w and initial state i.
There is one small problem: one or more of the entries (10.9) may be equal to $-\infty$ (in the situations considered here there are always conditions on w that guarantee that the entries in (10.9) are properly defined and that they are less that $+\infty$).
Suppose that player II uses all actions in B with at least some arbitrary small probability. Then player I is forced to use only those actions a (if any) for which (10.9) is finite for all $b \in B$. Otherwise, player I would loose an infinite amount. One easily verifies that this implies that the value of the original matrix game is equal to the value of the truncated matrix game in which player I can use only those actions a for which (10.9) is finite for all $b \in B$.

EXAMPLE 10.2. $A = B = \{1,2\}$. The notation is as follows: If both players take action 1, then player I receives 1; if player I takes action 2 and player II action 1, then player I looses an infinite amount; etc. Clearly, the value of the game is 0 and player I has an optimal strategy, namely action 1, whereas player II has only an ε-optimal strategy, namely use action 1 with probability $\varepsilon > 0$ and 2 with probability $1 - \varepsilon$.

$$\begin{pmatrix} 1 & 0 \\ -\infty & 2 \end{pmatrix}$$

So, if the matrix contains entries equal to $-\infty$, then player II may have no optimal randomized action. This is the reason why we have to write $\inf_{h \in F_{II}}$ in (10.8) and the cause of the asymmetry in (10.3) and (10.4).

It is well-known that the value and optimal randomized actions for a matrix game in which all elements are finite can be found by linear programming.

Now let us consider the following dynamic programming scheme

$$(10.10) \quad \begin{cases} v_0 := v \ , \\ v_{k+1} := Uv_k \ , \quad k = 0,1,\ldots,n-1 \ . \end{cases}$$

Following the approach of section 2.2 one may prove by induction the following results:

(i) $P(f,h)v_k^+ < \infty$ for all $f \in F_I$, $h \in F_{II}$ and $k = 0,1,\ldots,n-1$.

(ii) $v_k < \infty$ for all $k = 1,2,\ldots,n$.

(iii) There exist policies $f_0,\ldots,f_{n-1}$ for player I satisfying for all $k = 0,1,\ldots,n-1$

$$L(f_k,h)v_k \geq v_{k+1} \quad \text{for all } h \in F_{II} \ .$$

Then for the Markov strategy $\pi^{(n)} = (f_{n-1},\ldots,f_0)$ we have

$$(10.11) \quad v_n(\pi^{(n)},\gamma,v) \geq v_n \quad \text{for all } \gamma \in M_{II} \ ,$$

since, let $\gamma = (\tilde{h}_{n-1},\ldots,\tilde{h}_0) \in M_{II}$ be arbitrary, then

$$\begin{aligned} v_n(\pi^{(n)},\gamma,v) &= L(f_{n-1},\tilde{h}_{n-1}) \cdots L(f_0,\tilde{h}_0)v_0 \\ &\geq L(f_{n-1},\tilde{h}_{n-1}) \cdots L(f_1,\tilde{h}_1)v_1 \geq \ldots \geq v_n \ . \end{aligned}$$

(iv) There exist for all $\varepsilon > 0$ policies $h_{n-1},\ldots,h_0$ for player II satisfying for $k = 0,1,\ldots,n-1$,

$$L(f,h_k)v_k \leq v_{k+1} + \varepsilon 2^{-k-1}e \quad \text{for all } f \in F_I .$$

Then for the Markov strategy $\gamma^{(n)} = (h_{n-1},\ldots,h_0)$ we have

$$v_n(\pi,\gamma^{(n)},v) < v_n + \varepsilon e \quad \text{for all } \pi \in M_I , \tag{10.12}$$

since, let $\pi = (\tilde{f}_{n-1},\ldots,\tilde{f}_0) \in M_I$ be arbitrary, then

$$\begin{aligned} v_n(\pi,\gamma^{(n)},v) &= L(\tilde{f}_{n-1},h_{n-1}) \cdots L(\tilde{f}_0,h_0)v_0 \\ &\leq L(\tilde{f}_{n-1},h_{n-1}) \cdots L(\tilde{f}_1,h_1)(v_1 + \varepsilon 2^{-1}e) \\ &= L(\tilde{f}_{n-1},h_{n-1}) \cdots L(\tilde{f}_1,h_1)v_1 + \varepsilon 2^{-1}e \\ &\leq \ldots \leq v_n + \varepsilon(1-2^{-n})e < v_n + \varepsilon e . \end{aligned}$$

The line of proof is almost identical to the one in section 2.2 and is therefore omitted.
As a fairly straightforward generalization of the result of DERMAN and STRAUCH [1966] (cf. also lemma 2.1) one has that, if one of the players uses a Markov strategy, any strategy of the other player can be replaced by a (randomized) Markov strategy giving the same marginal distributions for the process, see e.g. GROENEWEGEN and WESSELS [1976]. Thus (10.11) and (10.12) generalize to all $\gamma \in \Gamma$ and $\pi \in \Pi$, respectively.
This yields the following result.

THEOREM 10.3. *If for* v *condition 10.1 holds, then the* n*-stage MG with terminal payoff* v *can be solved by the dynamic programming scheme* (10.10). *I.e., the game has the value* $v_n = U^n v$, *there exists an optimal Markov strategy for player I and for all* $\varepsilon > 0$ *there exists an* ε*-optimal Markov strategy for player II which can be determined from the scheme* (10.10).

PROOF. From the foregoing it is clear that it suffices to prove that v_n is the value of the game.
From (10.11) and (10.12) we have

$$\sup_{\pi\in\Pi} \inf_{\gamma\in\Gamma} v_n(\pi,\gamma,v) \geq \inf_{\gamma\in\Gamma} v_n(\pi^{(n)},\gamma,v) \geq v_n ,$$

and for all $\varepsilon > 0$

$$\inf_{\gamma\in\Gamma}\sup_{\pi\in\Pi} v_n(\pi,\gamma,v) \le \sup_{\pi\in\Pi} v_n(\pi,\gamma^{(n)},v) \le v_n + \varepsilon e \ .$$

Since clearly

$$\sup_{\pi\in\Pi}\inf_{\gamma\in\Gamma} v_n(\pi,\gamma,v) \le \inf_{\gamma\in\Gamma}\sup_{\pi\in\Pi} v_n(\pi,\gamma,v) \ ,$$

this yields

$$\sup_{\pi\in\Pi}\inf_{\gamma\in\Gamma} v_n(\pi,\gamma,v) = \inf_{\gamma\in\Gamma}\sup_{\pi\in\Pi} v_n(\pi,\gamma,v) = v_n \ ,$$

which completes the proof. □

Note that, if for all $k = 0,1,\ldots,n-1$ and all $a \in A$ and $b \in B$

$$\sum_{j\in S} p(i,a,b,j)v_k(j) > -\infty \ ,$$

then player II has an optimal randomized action in each matrix game and hence there exists also for player II an optimal Markov strategy for the n-stage MG with terminal payoff v.

10.3. TWO-PERSON NONZERO-SUM MARKOV GAMES AND THE RESTRICTION TO MARKOV STRATEGIES

In the preceding section it has been shown that the finite-stage two-person zero-sum MG can be solved by a dynamic programming approach. One might wonder whether such a dynamic programming approach can also be used in the nonzero-sum case. For this it is necessary that one can restrict the attention to Markov strategies. We will present an example that shows that this restriction to Markov strategies may be rather unrealistic.

In the two-person nonzero-sum MG there is one difference compared to the zero-sum case, namely, there are two reward functions. If in state i actions a by player I and b by player II are used, then player I receives a reward $r^{I}(i,a,b)$ and player II receives a reward $r^{II}(i,a,b)$. (In the zero-sum case we have $r^{I}(i,a,b) + r^{II}(i,a,b) = 0$.) Further, if there is a terminal payoff (as in the finite-stage game), then we have to specify a terminal payoff for each of the two players, v^{I} and v^{II}, say.

This gives us two n-stage payoff functions, namely

$$v_n^I(\pi,\gamma,v^I) := \mathbb{E}_{\pi,\gamma}\left[\sum_{k=0}^{n-1} r^I(X_k,A_k,B_k) + v^I(X_n)\right]$$

and

$$v_n^{II}(\pi,\gamma,v^{II}) := \mathbb{E}_{\pi,\gamma}\left[\sum_{k=0}^{n-1} r^{II}(X_k,A_k,B_k) + v^{II}(X_n)\right] ,$$

for players I and II, respectively.
A pair of strategies (π^*,γ^*) is a Nash-equilibrium pair (cf. NASH [1951]) if

$$v_n^I(\pi^*,\gamma^*,v^I) \geq v_n^I(\pi,\gamma^*,v^I) \quad \text{for all } \pi \in \Pi$$

and

$$v_n^{II}(\pi^*,\gamma^*,v^{II}) \geq v_n^{II}(\pi^*,\gamma,v^{II}) \quad \text{for all } \gamma \in \Gamma .$$

So, if the players use π^* and γ^*, then neither of them can improve his expected payoff by switching to another strategy.

The basic element in this game is the so-called bimatrix game. Each bimatrix game has at least one Nash-equilibrium pair of randomized actions. With a "double" dynamic programming scheme it is possible to obtain also a Nash-equilibrium pair of Markov strategies for the n-stage nonzero-sum game, see e.g. Van der WAL and WESSELS [1977].
However, there may be several Nash equilibrium pairs and in general different pairs of equilibrium strategies will have different values (this in contrast to the zero-sum case where the equilibrium value is unique).
So, one is not just interested in finding some Nash-equilibrium pair, but one wants to have an equilibrium pair for which the equilibrium values are (in some sense) acceptable for both players. An extra difficulty is the fact that there may also exist equilibrium pairs in Markov strategies that cannot be found by a dynamic programming approach and even equilibrium pairs in history-dependent strategies.

This section will be concluded with an example that shows that the values corresponding to a Nash-equilibrium pair of history-dependent strategies may be superior to the values of all Nash-equilibrium pairs of Markov strategies. A similar example for the infinite-horizon case can be found in Van der WAL and WESSELS [1977].

EXAMPLE 10.4. $S := \{1\}$, $A = B = \{1,2,3\}$. So the game is merely a repeated bimatrix game. The rewards are given in the bimatrix, where the notation is as follows. If both players take action 1, then they both receive 10; if player I takes action 1 and player II takes action 2, then player I receives 0 and player II receives 15; etc.

$$\begin{pmatrix} 10,10 & 0,15 & 0,6 \\ 15,0 & 6,6 & 0,6 \\ 6,0 & 6,0 & 0,0 \end{pmatrix}$$

Let us first consider the case that this bimatrix game is played only once. Then it is clear that it is reasonably attractive for both players if they both take action 1. However, this is not a Nash-equilibrium pair. Since, if your opponent takes action 1, the best you can do is take action 2 which yields you 15 instead of 10. The only pairs of Nash-equilibrium strategies are the pairs of randomized actions which only use actions 2 and 3. Among these, the most attractive pair is the one in which both players take action 2, yielding 6 to each of them.

Now consider the case that this bimatrix game is played twice. It seems clear that at the second stage both players should choose action 2, however, once we assume this, the two-stage nonzero-sum game reduces to a bimatrix game which is almost identical to the one-stage game; the only difference is that all entries in the bimatrix are enlarged by 6. In this game both players will choose again action 2 yielding for the 2-stage game a total reward of 12 to each of them.

But suppose both players use the following strategy: at the first stage take action 1; at the second stage take action 2 if the opponent also took action 1 at stage 1, otherwise take action 3. Then they both take action 1 at stage 1 and action 2 at stage 2, so they both receive 16. As one may easily verify, this pair of strategies is indeed a Nash-equilibrium pair. So, there exists a Nash-equilibrium pair of history-dependent strategies which is superior to all equilibrium pairs in Markov strategies.

Note that these equilibrium pairs of history-dependent strategies cannot be found by an ordinary dynamic programming scheme like (10.10). For this reason we will not consider the nonzero-sum MG any further.

10.4. INTRODUCTION TO THE ∞-STAGE MARKOV GAME

The following three chapters will deal with infinite-horizon two-person zero-sum Markov games. As in the case of the MDP two criteria are considered: the total expected reward and the average reward per unit time.

For any two strategies $\pi \in \Pi$ and $\gamma \in \Gamma$ and any initial state $i \in S$ the total expected reward for player I is defined by

$$(10.13) \qquad v(i,\pi,\gamma) := \mathbb{E}_{i,\pi,\gamma} \sum_{n=0}^{\infty} r(X_n,A_n,B_n) \ ,$$

and the average reward per unit time for player I is defined by

$$(10.14) \qquad g(i,\pi,\gamma) := \liminf_{n\to\infty} v_n(i,\pi,\gamma,0) \ ,$$

provided that the expectations are properly defined.
For player II the total expected reward is of course equal to $-v(i,\pi,\gamma)$. The average reward for player II is defined equal to $-g(i,\pi,\gamma)$ which makes the criterion asymmetric, but for the game we will consider this is irrelevant. If one would like to have a symmetric criterion, then one can take $\frac{1}{2}$ liminf + $\frac{1}{2}$ limsup. Also, one could take limsup instead of liminf in (10.14).

Clearly,

$$\sup_{\pi\in\Pi} \inf_{\gamma\in\Gamma} v(\pi,\gamma) \le \inf_{\gamma\in\Gamma} \sup_{\pi\in\Pi} v(\pi,\gamma)$$

and

$$\sup_{\pi\in\Pi} \inf_{\gamma\in\Gamma} g(\pi,\gamma) \le \inf_{\gamma\in\Gamma} \sup_{\pi\in\Pi} g(\pi,\gamma) \ ,$$

where supinf and infsup are taken componentwise.
We say that the infinite-horizon game with the criterion of total expected rewards has the value v^* if

$$\sup_{\pi\in\Pi} \inf_{\gamma\in\Gamma} v(\pi,\gamma) = \inf_{\gamma\in\Gamma} \sup_{\pi\in\Pi} v(\pi,\gamma) = v^* \ .$$

Similar, the average-reward MG is said to have the value g^* if

$$\sup_{\pi\in\Pi} \inf_{\gamma\in\Gamma} g(\pi,\gamma) = \inf_{\gamma\in\Gamma} \sup_{\pi\in\Pi} g(\pi,\gamma) = g^* \ .$$

The concept of the (infinite-horizon) MG has been introduced by SHAPLEY [1953]. Shapley considered the criterion of total expected rewards. He assumed the state space to be finite and further he assumed the existence of an unspecified state, * say, with $r(*,a,b) = 0$, $p(*,a,b,*) = 1$ and $p(i,a,b,*) > 0$ for all $i \in S$, $a \in A$ and $b \in B$. This assumption guarantees that the system eventually reaches the state *, and that the income from time n onwards decreases exponentially fast if $n \to \infty$. So the game can be approximated by finite-horizon games, i.e., by the method of standard successive approximations.
Shapley used that fact that in this case U is a contraction mapping to prove that the ∞-horizon MG has a value v^* which is precisely the unique fixed point of U. Moreover, he proved that policies f^* and h^* satisfying

$$L(f,h^*)v^* \leq v^* \leq L(f^*,h)v^* \quad \text{for all } f \in F_I, \ h \in F_{II},$$

yield optimal stationary strategies for the ∞-stage game:

$$v(\pi,h^*) \leq v^* \leq v(f^*,\gamma) \quad \text{for all } \pi \in \Pi, \ \gamma \in \Gamma.$$

The fact that U is a contraction also implies that the method of standard successive approximations yields bounds on v^* and nearly-optimal stationary strategies for the two players, see e.g. CHARNES and SCHROEDER [1967] and Van der WAL [1977a].
In chapter 11 we consider a generalization of Shapley's model, namely, the contracting MG with countable state space, cf. chapter 5. It will turn out that many of the results obtained for the contracting MDP can be generalized to the contracting game. Several algorithms will be considered, e.g. the Gauss-Seidel method, which for the MG has been introduced by KUSHNER and CHAMBERLAIN [1969] and the value-oriented approach.

Another infinite-horizon MG model that has been considered in the literature is the so-called positive MG: the game where $r(i,a,b) \geq 0$ for all i, a and b, see e.g. KUSHNER and CHAMBERLAIN [1969] and MAITRA and PARTHASARATHY [1971]. Kushner and Chamberlain consider the case that r(i,a,b) is bounded away from zero and that player II can terminate the play. Under this condition they established the existence of a value and nearly-optimal strategies that can be found by successive approximations. Maitra and Parthasarathy assume that $v(\pi,\gamma) < \infty$ for all π and γ and prove (among other things), for the case that S is finite, the existence of a value, of a nearly-optimal

stationary strategy for player I and of an optimal stationary strategy for player II.
In chapter 12 we consider the case that r(i,a,b) is bounded away from zero and that at certain bounded costs player II can terminate the play in any state. It will be shown that this game has features which are very similar to the contracting MG.

The infinite-horizon MG at the criterion of average reward per unit time has been first considered by GILLETTE [1957] for the case of a finite state space. He showed that if for some integer r

$$P(f_1,h_1)P(f_2,h_2)\cdots P(f_r,h_r)(i,j) > 0$$

for all $i,j \in S$ and all $f_k \in F_I$, $h_k \in F_{II}$, $k = 1,\ldots,r$, then the average-reward MG has a value and optimal stationary strategies for both players exist. Also he gives an example in which

$$\max_{f\in F_I} \min_{h\in F_{II}} g(f,h) < \min_{h\in F_{II}} \max_{f\in F_I} g(f,h) .$$

So, in general, there need not exist stationary optimal strategies.
This example, called the big match, has been further investigated by BLACKWELL and FERGUSON [1968]. They showed that also within the set of Markov strategies this game has no value, but if one also considers the history-dependent strategies, then the game does have a value. Special cases of infinite-stage average-reward MG's have also been considered by RIOS and YANEZ [1966], ROGERS [1969], SOBEL [1971], KOHLBERG [1974] and FEDERGRUEN [1977]. Only recently it has been shown by MONASH [1979] and independently by MERTENS and NEYMAN [1980] that every average-reward two-person zero-sum MG with finite state space has a value.
In chapter 13 we will study two special cases of the average reward MG for which the value of the game is independent of the initial state. And we show that in these cases the method of standard successive approximations converges, i.e., yields bounds on the value g^* and nearly-optimal stationary strategies for both players.

CHAPTER 11

THE CONTRACTING MARKOV GAME

11.1. INTRODUCTION

In section 10.2 the finite-stage two-person zero-sum MG has been studied. From this analysis it follows that there must also be a lot of similarity between the infinite-horizon two-person zero-sum MG at the criterion of total expected rewards and the total-reward MDP. In this chapter it will be shown how several ideas developed for the contracting MDP can be extended to the contracting game.

In the MG to be considered in this chapter the state space is assumed to be countable and the action spaces are finite. Further, the following condition is assumed to hold throughout this chapter.

Contraction assumption

There exists a nonnegative vector $\mu \in V$ such that

(i) For some constant $M \geq 0$

$$|r(f,h)| \leq M\mu \quad \text{for all } f \in F_I,\ h \in F_{II}. \tag{11.1}$$

(ii) For some constant ρ, with $0 \leq \rho < 1$,

$$P(f,h)\mu \leq \rho\mu \quad \text{for all } f \in F_I,\ h \in F_{II}. \tag{11.2}$$

We call this infinite-horizon game the *contracting* MG.
Taking the function μ such that $\mu(*) = 0$ and $\mu(i) = 1$, $i \neq *$, it is clear that the contracting MG generalizes Shapley's game (cf. section 10.4). The contracting model of this chapter is the same as the one studied in Van der WAL and WESSELS [1977].
Note also that the contraction assumption is a straightforward generalization of the model III assumptions in section 5.2.

In the remainder of this introductory section it is shown that the contraction assumption implies that $v(\pi,\gamma)$ is properly defined and finite for all π and γ, that the operators $L(f,h)$ and U are contractions on the Banach space V_μ with respect to the μ-norm, and that the unique fixed point of U within V_μ is the value of the infinite-stage MG, thus generalizing results in SHAPLEY [1953]. Next (section 2) the method of standard successive approximations is considered. Sections 3 and 4 deal with variants of this method which can be generated by go-ahead functions, section 5 considers generalizations of the policy iteration method and the value-oriented method. Finally, section 6 gives some possible extensions, e.g. the extension of results for the strongly-convergent MDP to strongly-convergent games.

First it will be shown that the contraction assumption implies that for any two strategies $\pi \in \Pi$, $\gamma \in \Gamma$, the total expected reward $v(\pi,\gamma)$ is properly defined and that $v(\pi,\gamma) \in V_\mu$.
Define for all $f \in F_I$ and $h \in F_{II}$ the operator $L^{abs}(f,h)$ on V_μ by

$$L^{abs}(f,h)v = |r(f,h)| + P(f,h)v \ . \tag{11.3}$$

It is immediately clear that $L^{abs}(f,h)$, and of course also $L(f,h)$ and U, map V_μ into itself.
For example, for any $v \in V_\mu$,

$$\|L^{abs}(f,h)v\|_\mu \le \| |r(f,h)| + P(f,h)\|v\|_\mu \mu\|_\mu \le M + \rho\|v\|_\mu < \infty \ .$$

From the analysis of the finite-stage MDP $(S,A\times B,p,r)$ in section 2.3 (we let the decision maker choose both a and b) it follows that

$$\sup_{\pi\in\Pi}\sup_{\gamma\in\Gamma} \mathbb{E}_{\pi,\gamma} \sum_{k=0}^{n-1} |r(X_k,A_k,B_k)|$$

$$= \sup_{\pi\in M_I}\sup_{\gamma\in M_{II}} \mathbb{E}_{\pi,\gamma} \sum_{k=0}^{n-1} |r(X_k,A_k,B_k)|$$

$$= \sup_{f_0,\ldots,f_{n-1}\in F_I} \sup_{h_0,\ldots,h_{n-1}\in F_{II}} L^{abs}(f_0,h_0)\cdots L^{abs}(f_{n-1},h_{n-1})0 \ .$$

Further,

$$L^{abs}(f_0,h_0)\cdots L^{abs}(f_{n-1},h_{n-1})0 \le L^{abs}(f_0,h_0)\cdots L^{abs}(f_{n-2},h_{n-2})M\mu$$

$$\le \rho^{n-1}M\mu + L^{abs}(f_0,h_0)\cdots L^{abs}(f_{n-2},h_{n-2})0$$

$$\le \ldots \le (1+\rho+\ldots+\rho^{n-1})M\mu \ ,$$

for all $n = 1,2,\ldots$ and all $f_0,\ldots,f_{n-1} \in F_I$, $h_0,\ldots,h_{n-1} \in F_{II}$. Hence, letting n tend to infinity,

$$v(\pi,\gamma) \le (1-\rho)^{-1} M\mu \quad \text{for all } \pi \in \Pi \ , \ \gamma \in \Gamma \ . \tag{11.4}$$

So, $v(\pi,\gamma)$ is properly defined and belongs to V_μ.

Next observe that $L(f,h)$ and U are contractions on the Banach space V_μ with respect to the μ-norm. Namely, for all $f \in F_I$ and $h \in F_{II}$ and for all v and $w \in V_\mu$

$$\|L(f,h)v - L(f,h)w\|_\mu = \|P(f,h)(v-w)\|_\mu$$

$$\le \|P(f,h)\|v-w\|_\mu\mu\|_\mu \le \rho\|v-w\|_\mu \ . \tag{11.5}$$

And let for arbitrary $v,w \in V_\mu$ the policies f_v, h_v, f_w and h_w satisfy

$$L(f,h_v)v \le L(f_v,h)v \quad \text{for all } f \in F_I \text{ and } h \in F_{II}$$

and

$$L(f,h_w)w \le L(f_w,h)w \quad \text{for all } f \in F_I \text{ and } h \in F_{II} \ .$$

Then

$$Uv - Uw = L(f_v,h_v)v - L(f_w,h_w)w \le L(f_v,h_w)v - L(f_v,h_w)w =$$

$$= P(f_v,h_w)(v-w) \le \rho\|v-w\|_\mu\,\mu \ .$$

Similarly,

$$Uw - Uv \le \rho\|v-w\|_\mu\,\mu \ .$$

Hence,

$$\|Uv - Uw\|_\mu \le \rho\|v-w\|_\mu \quad \text{for all } v,w \in V_\mu \ . \tag{11.6}$$

So, U has a unique fixed point in V_μ which is denoted, somewhat prematurely, by v^*.

THEOREM 11.1.

(i) *The unique fixed point of the operator* U *is the value of the* ∞*-horizon Markov game.*

(ii) *Let the policies* f^* *and* h^* *satisfy*

$$(11.7) \qquad L(f,h^*)v^* \le v^* \le L(f^*,h)v^* \quad \textit{for all } f \in F_I,\ h \in F_{II},$$

then the stationary strategies f^* *and* g^* *are optimal in the* ∞*-horizon game for players I and II, respectively.*

PROOF. It suffices to prove

$$(11.8) \qquad v(\pi,g^*) \le v^* \le v(f^*,\gamma) \quad \text{for all } \pi \in \Pi,\ \gamma \in \Gamma.$$

Let us prove $v(\pi,g^*) \le v^*$ first.

From (11.7) we have for all $\pi = (f_0,f_1,\ldots,) \in M_I$ and all $n = 1,2,\ldots$

$$\begin{aligned} v_n(\pi,g^*) &:= v_n(\pi,g^*,0) = L(f_0,g^*)\cdots L(f_{n-1},g^*)0 \\ &\le L(f_0,g^*)\cdots L(f_{n-1},g^*)(v^* + \|v^*\|_\mu \mu) \\ &\le L(f_0,g^*)\cdots L(f_{n-1},g^*)v^* + \rho^n\|v^*\|_\mu \mu \\ &\le L(f_0,g^*)\cdots L(f_{n-2},g^*)v^* + \rho^n\|v^*\|_\mu \mu \\ &\le \ldots \le v^* + \rho^n\|v^*\|_\mu \mu . \end{aligned}$$

So, with $n \to \infty$,

$$v(\pi,g^*) \le v^* \quad \text{for all } \pi \in M_I .$$

Further, it follows from the extension of Derman and Strauch's result by GROENEWEGEN and WESSELS [1976] (cf. section 10.2) that

$$\sup_{\pi\in\Pi} v(\pi,g^*) = \sup_{\pi\in M_I} v(\pi,g^*) .$$

Hence,

$$v(\pi,g^*) \le v^* \quad \text{for all } \pi \in \Pi .$$

Similarly, one proves the second inequality in (11.8). □

Theorem 11.1 is merely a straightforward generalization of Shapley's results for the finite state space case ([SHAPLEY, 1953]).

11.2. THE METHOD OF STANDARD SUCCESSIVE APPROXIMATIONS

In this section it will be shown that, as in the case of the contracting MDP, the method of standard successive approximations with scrapvalue $v_0 \in V_\mu$ yields bounds on v^* and nearly-optimal stationary strategies for the two players. The results of this section can be found in Van der WAL and WESSELS [1977] and improve or extend results in CHARNES and SCHROEDER [1967] and Van der WAL [1977a]. Compare also section 5.4 with $\delta \equiv 1$.

Standard successive approximations

(11.9)
Choose $v_0 \in V_\mu$.
Determine for $n = 0,1,\ldots$

$$v_{n+1} = Uv_n$$

and policies $f_n \in F_I$ and $h_n \in F_{II}$ satisfying

$$L(f,h_n)v_n \leq v_{n+1} \leq L(f_n,h) \quad \text{for all } f \in F_I,\ h \in F_{II}\ .$$

Since U is a contraction we immediately have that v_n converges to v^* in μ-norm, namely

$$\|v_n - v^*\|_\mu = \|U^n v_0 - U^n v^*\|_\mu \leq \rho^n \|v_0 - v^*\|_\mu \to 0 \quad (n \to \infty)\ .$$

In order to show that the standard successive approximation scheme yields bounds on v^* and nearly-optimal stationary strategies, we need the following notations (cf. section 5.4).
Define

$$\|w\|_\mu^{\max} := \inf \{c \in \mathbb{R} \mid w \leq c\mu\} \quad \text{for all } w \in V_\mu\ ,$$

$$\|w\|_\mu^{\min} := \sup \{c \in \mathbb{R} \mid w \geq c\mu\} \quad \text{for all } w \in V_\mu\ ,$$

$$\rho_I^{\max}(f) := \| \max_{h \in F_{II}} P(f,h)\mu \|_\mu^{\max} \quad \text{for all } f \in F_I\ ,$$

$$\rho_I^{\min}(f) := \| \min_{h \in F_{II}} P(f,h)\mu \|_\mu^{\min} \quad \text{for all } f \in F_I\ ,$$

$$\rho_{II}^{\max}(h) := \| \max_{f \in F_I} P(f,h)\mu \|_\mu^{\max} \quad \text{for all } h \in F_{II}\ ,$$

$$\rho_{II}^{\min}(h) := \| \min_{f \in F_I} P(f,h)\mu \|_{\mu}^{\min} \quad \text{for all } h \in F_{II} ,$$

and for the policies f_n and g_n satisfying (11.9)

$$\rho_{I,n}(f_n) := \begin{cases} \rho_I^{\max}(f_n) & \text{if } \|v_{n+1} - v_n\|_{\mu}^{\min} < 0 , \\ \rho_I^{\min}(f_n) & \text{if } \|v_{n+1} - v_n\|_{\mu}^{\min} \geq 0 , \end{cases}$$

$$\rho_{II,n}(h_n) := \begin{cases} \rho_{II}^{\max}(h_n) & \text{if } \|v_{n+1} - v_n\|_{\mu}^{\max} \geq 0 , \\ \rho_{II}^{\min}(h_n) & \text{if } \|v_{n+1} - v_n\|_{\mu}^{\max} < 0 . \end{cases}$$

Using these notations one has the following results (cf. theorem 5.12).

THEOREM 11.2

(i) *For all* $\gamma \in \Gamma$

$$v(f_n,\gamma) \geq v_{n+1} + \rho_{I,n}(f_n)(1 - \rho_{I,n}(f_n))^{-1} \|v_{n+1} - v_n\|_{\mu}^{\min} \mu .$$

(ii) *For all* $\pi \in \Pi$

$$v(\pi,h_n) \leq v_{n+1} + \rho_{II,n}(h_n)(1 - \rho_{II,n}(h_n))^{-1} \|v_{n+1} - v_n\|_{\mu}^{\max} \mu .$$

(iii)
$$\rho_{I,n}(f_n)(1 - \rho_{I,n}(f_n))^{-1} \|v_{n+1} - v_n\|_{\mu}^{\min} \mu \leq v^* - v_{n+1}$$
$$\leq \rho_{II,n}(h_n)(1 - \rho_{II,n}(h_n))^{-1} \|v_{n+1} - v_n\|_{\mu}^{\max} \mu .$$

PROOF.

(i) From the result of GROENEWEGEN and WESSELS [1976] it follows that it suffices to prove (i) for all $\gamma \in M_{II}$.

Let $\gamma = (\tilde{h}_0, \tilde{h}_1, \ldots) \in M_{II}$ be arbitrary, then

$$(11.10) \quad v(f_n,\gamma) = \lim_{k \to \infty} L(f_n,\tilde{h}_0) \cdots L(f_n,\tilde{h}_k) 0 = \lim_{k \to \infty} L(f_n,\tilde{h}_0) \cdots L(f_n,\tilde{h}_k) v_n .$$

For all $k = 1,2,\ldots,$

$$L(f_n,\tilde{h}_0) \cdots L(f_n,\tilde{h}_k) v_n \geq L(f_n,\tilde{h}_0) \cdots L(f_n,\tilde{h}_{k-1}) U v_n$$
$$\geq L(f_n,\tilde{h}_0) \cdots L(f_n,\tilde{h}_{k-1}) (v_n + \|v_{n+1} - v_n\|_{\mu}^{\min} \mu) .$$

Further,

$$P(f_n,h)\|v_{n+1} - v_n\|_\mu^{\min} \mu \geq \rho_{I,n}(f_n)\|v_{n+1} - v_n\|_\mu^{\min} \mu .$$

Hence,

$$\begin{aligned}(11.11)\quad & L(f_n,\tilde{h}_0)\cdots L(f_n,\tilde{h}_k)v_n \\ & \geq L(f_n,\tilde{h}_0)\cdots L(f_n,\tilde{h}_{k-1})v_n + \rho_{I,n}^k(f_n)\|v_{n+1} - v_n\|_\mu^{\min} \mu \\ & \geq \ldots \geq v_{n+1} + (\rho_{I,n}(f_n) + \ldots + \rho_{I,n}^k(f_n))\|v_{n+1} - v_n\|_\mu^{\min} \mu .\end{aligned}$$

So from (11.10) and (11.11) one obtains for all $\gamma \in M_{II}$

$$v(f_n,\gamma) \geq v_{n+1} + \rho_{I,n}(f_n)(1 - \rho_{I,n}(f_n))^{-1}\|v_{n+1} - v_n\|_\mu^{\min} \mu .$$

Similarly one proves (ii). Then (iii) follows immediately from (i) and (ii). □

Since $v_{n+1} - v_n$ tends to zero, if n tends to infinity, it follows from theorem 11.2 that the method of standard successive approximations yields good approximations of v^* and nearly-optimal stationary strategies for both players.

11.3. GO-AHEAD FUNCTIONS

In this section, following the approach of chapter 3, we generate by means of nonzero go-ahead functions a set of variants of the method of standard successive approximations.

For the two-person game a *go-ahead function* is any function δ from

$$S \cup \bigcup_{n=1}^{\infty} (S \times A \times B)^n \cup \bigcup_{n=1}^{\infty} (S \times A \times B)^n \times S \quad \text{into } [0,1] .$$

The interpretation is the same as for the MDP. E.g., $\delta(i_0,a_0,b_0,\ldots,i_n)$ denotes the probability that the observations of the process will continue after the history $i_0,a_0,b_0,\ldots,i_n$, given that the observations have not been stopped before.

A go-ahead function δ is called *nonzero* if

$$(11.12)\quad \alpha_\delta := \inf_{i\in S} \min_{a\in A} \min_{b\in B} \delta(i)\delta(i,a,b) > 0 .$$

In order to describe e.g. the overrelaxation method and to define the L_δ and U_δ operators we have to incorporate the random experiments - the outcomes of which determine whether the observation of the process continues - into the stochastic process. Therefore we extend the space $(S \times A \times B)^\infty$ to the space $(S \times E \times A \times B \times E)^\infty$ with $E := \{0,1\}$ again.

As in section 3.3 one may define for each initial state $i \in S$, any go-ahead function δ and any pair of strategies $\pi \in \Pi$ and $\gamma \in \Gamma$ the probability measure $\mathbb{P}^\delta_{i,\pi,\gamma}$ and the stochastic process $\{(X_n, Y_n, A_n, B_n, Z_n),\ n = 0,1,\ldots\}$, where X_n, A_n and B_n are the state and actions at time n, where $Y_n = 1$ if the observations continue after X_n has been observed and $Y_n = 0$ otherwise, and where $Z_n = 1$ if the observations continue after the selection of A_n and B_n and $Z_n = 0$ otherwise.

The expectation with respect to $\mathbb{P}^\delta_{i,\pi,\gamma}$ is denoted by $\mathbb{E}^\delta_{i,\pi,\gamma}$.

Next define the stopping time τ on $(S \times E \times A \times B \times E)^\infty$ by

$$\tau(i_0,y_0,a_0,b_0,z_0,i_1,y_1,\ldots) = \inf \{n \mid y_n z_n = 0\} .$$

So τ denotes again the time upon which the observations of the process are stopped.

Further, we define for any $\pi \in \Pi$, any $\gamma \in \Gamma$ and any go-ahead function δ the operators $L_\delta(\pi,\gamma)$ and U_δ on V_μ:

$$(11.13) \qquad L_\delta(\pi,\gamma)v := \mathbb{E}^\delta_{\pi,\gamma}\left[\sum_{n=0}^{\tau-1} r(X_n,A_n,B_n) + v(X_\tau)\right] ,$$

with $v(X_\tau) = 0$, by definition, if $\tau = \infty$;

$$(11.14) \qquad U_\delta v := \sup_{\pi\in\Pi} \inf_{\gamma\in\Gamma} L_\delta(\pi,\gamma)v ,$$

where the supinf is taken componentwise.

That $L_\delta(\pi,\gamma)$ and U_δ are properly defined on V_μ, map V_μ into itself and are, if $\alpha_\delta > 0$, even contractions on V_μ can be seen as follows.

Define the operator $\tilde{L}_\delta(\pi,\gamma)$ on V_μ by

$$(11.15) \qquad \tilde{L}_\delta(\pi,\gamma)v := \mathbb{E}^\delta_{\pi,\gamma} v(X_\tau) , \quad \pi \in \Pi , \ \gamma \in \Gamma .$$

Then we can apply lemma 4.22 with $v = \mu$ on the MDP with state space S, action space A × B, rewards r and transition law p, to obtain

$$(11.16) \qquad \tilde{L}_\delta(\pi,\gamma)\mu \le (1-\alpha_\delta)\mu + \alpha_\delta \max_{f\in F_I} \max_{h\in F_{II}} P(f,h)\mu \le [1-\alpha_\delta(1-\rho)]\mu .$$

Hence

$$L_\delta(\pi,\gamma)v = \left| \mathbb{E}^\delta_{\pi,\gamma} \left[\sum_{n=0}^{\tau-1} r(X_n,A_n,B_n) + v(X_\tau) \right] \right|$$

$$\leq \mathbb{E}^\delta_{\pi,\gamma} \sum_{n=0}^{\infty} |r(X_n,A_n,B_n)| + \|v\|_\mu \, \tilde{L}_\delta(\pi,\gamma)\mu$$

$$\leq M(1-\rho)^{-1}\mu + \|v\|_\mu [1 - \alpha_\delta(1-\rho)]\mu \ .$$

Thus $L_\delta(\pi,\gamma)v$ is properly defined on V_μ.
And

$$L_\delta(\pi,\gamma)v \in V_\mu \quad \text{for all } \pi \in \Pi \text{ and } \gamma \in \Gamma \ ,$$

whence also

$$U_\delta v \in V_\mu \ .$$

Further, for all $\pi \in \Pi$, $\gamma \in \Gamma$ and $v,w \in V_\mu$,

$$\|L_\delta(\pi,\gamma)v - L_\delta(\pi,\gamma)w\|_\mu = \|\tilde{L}_\delta(\pi,\gamma)(v-w)\|_\mu$$

$$\leq \| \|v-w\|_\mu \, \tilde{L}_\delta(\pi,\gamma)\mu \|_\mu \leq [1 - \alpha_\delta(1-\rho)] \|v-w\|_\mu \ .$$

So, if $\alpha_\delta > 0$, then $L_\delta(\pi,\gamma)$ is a contraction on V_μ with respect to the μ-norm.
Similar as in section 1 (the proof of (11.6)) one may show that, if $\alpha_\delta > 0$, also U_δ is a contraction on V_μ. Hence, for any nonzero δ, the operator U_δ has a unique fixed point in V_μ, v_δ say.

Our next step is to prove that $v_\delta = v^*$, so that it makes sense to use successive approximation methods generated by nonzero go-ahead functions.
We can follow the line of reasoning of section 3.4. First one may show, in a similar way, that for all π^1 and $\pi^2 \in \Pi$

$$L_\delta(\pi^1,h^*)v(\pi^2,h^*) \leq \sup_{\pi\in\Pi} v(\pi,h^*) = v^* \ ,$$

where h^* denotes an optimal policy for player II.

Thus for all $\pi \in \Pi$

$$L_\delta(\pi,h^*)v^* \leq v^*$$

and

$$U_\delta v^* \leq v^* .$$

Similarly, we obtain for all $\gamma \in \Gamma$

$$L_\delta(f^*,\gamma)v^* \geq v^* ,$$

so

$$U_\delta v^* \geq v^* .$$

Hence

$$U_\delta v^* = v^* . \tag{11.17}$$

So v^* is a fixed point of U_δ in V_μ which implies $v_\delta = v^*$ (δ nonzero).
As a consequence we have for all nonzero go-ahead functions δ

$$\lim_{n\to\infty} U_\delta^n v = v^* , \quad v \in V_\mu . \tag{11.18}$$

11.4. STATIONARY GO-AHEAD FUNCTIONS

In this section it will be shown that as in the case of the contracting MDP (cf. section 5.4) any nonzero stationary go-ahead function generates a successive approximation algorithm that yields bounds on v^* and nearly-optimal stationary strategies for the two players.
Similarly as in definition 3.19 a go-ahead funcyion is called *stationary* if for all $n = 1,2,\ldots$ and all $i_0,a_0,b_0,i_1,\ldots$

$$\delta(i_0,a_0,b_0,\ldots,i_n,a_n,b_n) = \delta(i_n,a_n,b_n)$$

and

$$\delta(i_0,a_0,b_0,\ldots,i_n,a_n,b_n,i_{n+1}) = \delta(i_n,a_n,b_n,i_{n+1}) .$$

First we show that for stationary go-ahead functions one can restrict the attention to stationary strategies in the determination of $U_\delta v$, i.e., we show that for every $v \in V_\mu$ there exist policies $f \in F_I$ and $h \in F_{II}$ satisfying

$$L_\delta(\pi,h)v \leq U_\delta v \leq L_\delta(f,\gamma)v \quad \text{for all } \pi \in \Pi , \ \gamma \in \Gamma . \tag{11.19}$$

To prove this construct the Markov game $(\hat{S},\hat{A},\hat{B},\hat{p},\hat{r})$ which is essentially equivalent to the problem of the determination of $U_\delta v$ as follows (the line of reasoning is essentially the same as for the MDP): Assume, without loss of generality, $\delta(i) = 1$ for all $i \in S$ and define

$$\hat{S} := S \cup \{*\} , \quad * \notin S ,$$

$$\hat{A} := A , \quad \hat{B} := B ,$$

$$\hat{p}(i,a,b,j) := \delta(i,a,b)p(i,a,b,j)\delta(i,a,b,j) ,$$

$$\hat{p}(i,a,b,*) := 1 - \sum_{j\in S} \hat{p}(i,a,b,j) , \quad \hat{p}(*,a,b,*) = 1 ,$$

$$\hat{r}(i,a,b) := [1 - \delta(i,a,b)]v(i) + \delta(i,a,b)[r(i,a,b) + \sum_{j\in S} p(i,a,b,j)[1 - \delta(i,a,b,j)]v(j)] ,$$

$$\hat{r}(*,a,b) := 0 .$$

One easily verifies that, with the bounding function $\hat{\mu}$ on $\hat{S}$ defined by $\hat{\mu}(i) = \mu(i)$, $i \in S$, and $\hat{\mu}(*) = 0$, this Markov game is also contracting. So it follows from theorem 11.1 that this game has a value and that both players have stationary optimal strategies. Then the restrictions of these stationary optimal strategies to the states in S, f and h say, satisfy (11.19).

Now consider for a stationary go-ahead function δ the following successive approximation procedure

(11.20)
$$\begin{cases} \text{Choose } v_0 \in V_\mu . \\ \text{Determine for } n = 0,1,\ldots \\ \qquad v_{n+1} = U_\delta v_n , \\ \text{and policies } f_n \text{ and } h_n \text{ satisfying} \\ \qquad L_\delta(\pi,h_n)v_n \le v_{n+1} \le L_\delta(f_n,\gamma)v_n \quad \text{for all } \pi \in \Pi , \ \gamma \in \Gamma . \end{cases}$$

From (11.16) and (11.18) one easily shows that v_n converges to v^* exponentially fast. In order to obtain again, as in theorems 5.12 and 11.2, the MacQueen bounds for v^* and the strategies f_n and h_n we need the following notations.

Define

$$\rho^{\max}_{I,\delta}(f) := \| \max_{h\in F_{II}} \tilde{L}_\delta(f,h)\mu\|^{\max}_\mu , \quad f \in F_I ,$$

$$\rho^{\min}_{I,\delta}(f) := \| \min_{h\in F_{II}} \tilde{L}_\delta(f,h)\mu\|^{\min}_\mu , \quad f \in F_I ,$$

$$\rho^{\max}_{II,\delta}(h) := \| \max_{f\in F_I} \tilde{L}_\delta(f,h)\mu\|^{\max}_\mu , \quad h \in F_{II} ,$$

$$\rho^{\min}_{II,\delta}(h) := \| \min_{f\in F_I} \tilde{L}_\delta(f,h)\mu\|^{\min}_\mu , \quad h \in F_{II} ,$$

and for the policies f_n and g_n satisfying (11.20)

$$\rho_{I,\delta,n}(f_n) := \begin{cases} \rho^{\max}_{I,\delta}(f_n) & \text{if } \|v_{n+1} - v_n\|^{\min}_\mu < 0 , \\ \rho^{\min}_{I,\delta}(f_n) & \text{if } \|v_{n+1} - v_n\|^{\min}_\mu \geq 0 , \end{cases}$$

$$\rho_{II,\delta,n}(h_n) := \begin{cases} \rho^{\max}_{II,\delta}(h_n) & \text{if } \|v_{n+1} - v_n\|^{\max}_\mu \geq 0 , \\ \rho^{\min}_{II,\delta}(h_n) & \text{if } \|v_{n+1} - v_n\|^{\max}_\mu < 0 . \end{cases}$$

Then one has the following result (cf. theorem 11.2).

THEOREM 11.3. *Let* $\{v_n\}$, $\{f_n\}$ *and* $\{h_n\}$ *be the sequences obtained in* (11.20). *Then we have* (δ *nonzero*)

(i) *for all* $\gamma \in \Gamma$,

$$v(f_n,\gamma) \geq v_{n+1} + \rho_{I,\delta,n}(f_n)(1 - \rho_{I,\delta,n}(f_n))^{-1}\|v_{n+1} - v_n\|^{\min}_\mu \mu ,$$

(ii) *for all* $\pi \in \Pi$,

$$v(\pi,h_n) \leq v_{n+1} + \rho_{II,\delta,n}(h_n)(1 - \rho_{II,\delta,n}(h_n))^{-1}\|v_{n+1} - v_n\|^{\max}_\mu \mu ,$$

(iii)
$$\rho_{I,\delta,n}(f_n)(1 - \rho_{I,\delta,n}(f_n))^{-1}\|v_{n+1} - v_n\|^{\min}_\mu \mu \leq v^* - v_{n+1}$$
$$\leq \rho_{II,\delta,n}(h_n)(1 - \rho_{II,\delta,n}(h_n))^{-1}\|v_{n+1} - v_n\|^{\max}_\mu \mu .$$

PROOF.

(i) Using the result of GROENEWEGEN and WESSELS [1976] one may easily show that

$$v(f_n,\gamma) \geq \min_{h\in F_{II}} v(f_n,h) \quad \text{for all } \gamma \in \Gamma .$$

Since further for all nonzero δ, all $h \in F_{II}$ and all $v \in V_\mu$

$$v(f_n,h) = \lim_{k\to\infty} L_\delta^k(f_n,h)v ,$$

the proof follows along the same lines as the proof of theorem 11.2. Similarly one obtains (ii), and (iii) follows immediately from (i) and (ii). □

In general, the amount of work that has to be done in order to obtain $U_\delta v$ is of the same order as the amount needed to solve the original ∞-stage game. However, for special stationary go-ahead functions, e.g. those corresponding to the "game variants" of the algorithms formulated in section 3.3, $U_\delta v$ can be computed componentwise by solving simple matrix games. In that case the amount of work becomes the same as for the computation of Uv.

11.5. POLICY ITERATION AND VALUE-ORIENTED METHODS

In this section it will be shown how the policy iteration method and the method of value-oriented standard successive approximations can be generalized for the contracting Markov game.

For the contracting MG (with finite state space) POLLATSCHEK and AVI-ITZHAK [1969] have suggested the following straightforward generalization of Howard's policy iteration method.

(11.21)
Choose $v_0 \in V_\mu$.
Determine for $n = 0,1,\ldots$ policies f_n and h_n satisfying

$$L(f,h_n)v_n \leq Uv_n \leq L(f_n,h)v_n \quad \text{for all } f \in f_I ,\ h \in F_{II} ,$$

and define

$$v_{n+1} = v(f_n,h_n) .$$

Pollatschek and Avi-Itzhak proved that under a rather conservative condition v_n converges to v^*. (One may easily show that $\rho < \frac{1}{3}$ guarantees that v_n

converges to v^*.) RAO, CHANDRASEKARAN and NAIR [1973] claimed that the algorithm would always converge, however, their proof is incorrect. And, as the following example, given in Van der WAL [1977b], demonstrates, their proof cannot be repaired.

EXAMPLE 11.4. $S = \{1,2\}$, $A(1) = B(1) = \{1,2\}$, $A(2) = B(2) = \{1\}$.

r(1,a,b)		b = 1	b = 2
a	1	3	6
	2	2	1

p(1,a,b,1)		b = 1	b = 2
a	1	$\frac{3}{4}$	$\frac{1}{4}$
	2	$\frac{3}{4}$	$\frac{3}{4}$

Further, $r(2,1,1) = 0$ and $p(2,1,1,2) = 1$. So, state 2 is absorbing. Taking μ such that $\mu(1) = 1$, $\mu(2) = 0$, one immediately sees that the game is contracting. Now, choose $v_0 = 0$. Then, in order to determine policies f_0 and h_0 satisfying (11.21) for $n = 0$, one has to solve for state 1 the matrix game

$$\begin{pmatrix} 3 & 6 \\ 2 & 1 \end{pmatrix}.$$

Clearly, this game has value 3 and the policies f_0 and h_0 with $f_0(1,1) = h_0(1,1) = 1$ are optimal. So v_1 has $v_1(1) = 12$, $v_1(2) = 0$.
Next, in order to obtain f_1 and h_1 we have to solve in state 1 the matrix game

$$\begin{pmatrix} 3 + \frac{3}{4}\cdot 12 & 6 + \frac{1}{4}\cdot 12 \\ 2 + \frac{3}{4}\cdot 12 & 1 + \frac{3}{4}\cdot 12 \end{pmatrix} = \begin{pmatrix} 12 & 9 \\ 11 & 10 \end{pmatrix}.$$

The value of this game is 10 and the optimal policies f_1 and h_1 have $f_1(1,2) = h_1(1,2) = 1$. So $v_2(1) = 4$, $v_2(2) = 0$.
In the third iteration step one has to solve the matrix game

$$\begin{pmatrix} 3 + \frac{3}{4}\cdot 4 & 6 + \frac{1}{4}\cdot 4 \\ 2 + \frac{3}{4}\cdot 4 & 1 + \frac{3}{4}\cdot 4 \end{pmatrix} = \begin{pmatrix} 6 & 7 \\ 5 & 4 \end{pmatrix},$$

which has value 6 for the policies f_2 and h_2 with $f_2(1,1) = h_2(1,1) = 1$. Thus $v_3(1) = 12$ and $v_3(2) = 0$, and $v_3 = v_1$.
Continuing in this way we get $v_{2n} = v_2$, $v_{2n+1} = v_1$, $n = 1,2,\ldots$. So we see that in this example Pollatschek and Avi-Itzhak's generalization of Howard's policy iteration cycles.

Another generalization of Howard's method has been proposed by HOFFMAN and KARP [1966] for the average-reward MG. The same idea can also be used for the total-reward case, see POLLATSCHEK and AVI-ITZHAK [1969] and RAO, CHANDRASEKARAN and NAIR [1973].
To describe this algorithm we define for all $h \in F_{II}$ the operator U_h on V_μ by

$$U_h v = \max_{f \in F_I} L(f,h)v \ .$$

Then Hoffman and Karp's variant can be formulated as follows.

Choose $v_0 \in V_\mu$.
Determine for $n = 0,1,\ldots$ a policy h_n satisfying

$$L(f,h_n)v_n \leq Uv_n \quad \text{for all } f \in F$$

and determine

$$v_{n+1} := \lim_{k\to\infty} U_{h_n}^k v_n \ .$$

Observe that to obtain v_{n+1} one has to solve a whole MDP exactly, e.g. by Howard's policy iteration method.
However, it is not necessary that one actually determines the value of the MDP. As in the MDP we can use a value-oriented variant in which v_{n+1} is taken equal to $U_{h_n}^\lambda v_n$ for some λ.
The algorithm then becomes (see Van der WAL [1977b]):

(11.22)

Choose $v_0 \in V_\mu$ and $\lambda \in \{1,2,\ldots\infty\}$.
Determine for $n = 0,1,\ldots$ a policy h_n satisfying

$$L(f,h_n)v_n \leq Uv_n \quad \text{for all } f \in F$$

and determine

$$v_{n+1} := U_{h_n}^\lambda v_n \ .$$

For the monotone version of this algorithm, where one starts with a scrap-value v_0 for which $Uv_0 \leq v_0$, the convergence proof can be found in Van der WAL [1977b]. The line of reasoning is exactly the same as for the MDP (cf. theorem 3.22)

As has been pointed out by ROTHBLUM [1979] one may follow in the nonmonotonic case the line of proof given by Van NUNEN [1976c] for the MDP.

THEOREM 11.5 (ROTHBLUM [1979]).
The sequence $\{v_n\}$ *obtained in* (11.22) *converges to* v^*.

PROOF. See ROTHBLUM [1979]. □

11.6. THE STRONGLY CONVERGENT MARKOV GAME

Having seen the large similarity between the contracting MDP and the contracting MG it is natural to ask whether more results for the total-reward MDP can be translated to the case of the total-reward MG.
That one cannot just translate all results is immediately clear if we consider e.g. lemma 3.1. It is obvious that for the general total-reward MDP lemma 3.1 need not hold since the MG is in a sense a combination of a maximization and a minimization problem.
However, for the strong-convergence case most results for the MDP also hold for the MG.
An MG is called strongly convergent if there exists a sequence $\varphi = (\varphi_0,\varphi_1,\ldots) \in \Phi$ (cf. section 4.1) for which

$$(11.23)\qquad \sup_{\pi\in\Pi}\sup_{\gamma\in\Gamma} \mathbb{E}_{i,\pi,\gamma} \sum_{n=0}^{\infty} \varphi_n(i)\,|r(X_n,A_n,B_n)| < \infty\,, \quad i \in S\,.$$

Just as for the MDP, the existence of a function $\varphi \in \Phi$ for which (11.23) is finite is equivalent to the following pair of conditions:

$$(i)\qquad z^* := \sup_{\pi\in\Pi}\sup_{\gamma\in\Gamma} \mathbb{E}_{\pi,\gamma} \sum_{n=0}^{\infty} |r(X_n,A_n,B_n)| < \infty\,,$$

$$(ii)\qquad \sup_{\pi\in\Pi}\sup_{\gamma\in\Gamma} \mathbb{E}_{\pi,\gamma} \sum_{k=n}^{\infty} |r(X_k,A_k,B_k)| \to 0 \quad (n\to\infty)\,.$$

One may show that the method of standard successive approximations converges using the game-equivalent of theorem 4.6 with instead of the operator $\tilde{U}$ an operator $\underset{\sim}{\tilde{U}}$ defined as

$$\underset{\sim}{\tilde{U}}v := \sup_{\pi\in\Pi}\sup_{\gamma\in\Gamma} \tilde{L}(\pi,\gamma)v\,.$$

Similarly one may translate the results of section 4.2. E.g., if

$$L(f,h)v^* \geq v^* \quad \text{for all } h \in F_{II}$$

then f is an optimal stationary strategy for player I.
Further, one may show that for nonzero go-ahead functions δ the corresponding method of successive approximations converges, using lemmas 4.21 and 4.22 and theorem 4.23 with the operator $\underset{\sim}{\tilde{U}}_\delta$ defined by

$$\underset{\sim}{\tilde{U}}_\delta v := \sup_{\pi\in\Pi} \sup_{\gamma\in\Gamma} \tilde{L}_\delta(\pi,\gamma)v \ .$$

CHAPTER 12

THE POSITIVE MARKOV GAME WHICH CAN BE TERMINATED BY THE MINIMIZING PLAYER

12.1. INTRODUCTION

In the general positive MG it is assumed that $r(i,a,b)$ is nonnegative for all i, a and b. So, for all $n = 0,1,\ldots,\infty$, the expected n-stage reward is for any pair of strategies properly defined and so is any successive approximation scheme with scrapvalue $v \geq 0$. (By reversing the roles of the two players the game can be made to satisfy condition 10.1 for all $n = 1,2,\ldots,\infty$. For historical reasons, however, we prefer to treat the game as a positive game.)

In this chapter we shall analyze a special positive MG having the following properties:

(12.1)

(i) there is a specific state, $*$ say, for which $r(*,a,b) = 0$ and $p(*,a,b,*) = 1$ for all $a \in A$, $b \in B$;

(ii) there is a constant $c > 0$ such that $r(i,a,b) \geq c$ for all $i \neq *$ and all $a \in A$ and $b \in B$;

(iii) there exists a constant $C \geq c$ and in each state $i \in S$ an action, $b(i)$ say, for player II such that $p(i,a,b(i),*) = 1$ and $r(i,a,b(i)) \leq C$ for all $a \in A$.

So, the state $*$ is absorbing and the actual play can be considered to have stopped once state $*$ has been reached. Further, as long as the system has not yet reached $*$, player II looses at least an amount of c in each step. However, by taking at time 0 the appropriate action in the initial state he can force the system into state $*$, thus restricting his total loss to at most C.

This is a special case of the positive MG considered by KUSHNER and CHAMBERLAIN [1969].

In the sequel it will be shown that the method of standard successive approximations yields bounds for the value of the game and nearly-optimal stationary strategies for the two players. As we will see, this specific positive game has a very similar structure as the contracting MG. The results of this chapter can be found in Van der WAL [1979]. Part of the results were already given by KUSHNER and CHAMBERLAIN [1969].

First let us consider in some detail the general positive MG with $r(i,a,b) \geq 0$ for all $a \in A$, $b \in B$ and A and B finite.
For this game consider the following standard successive approximation scheme:

$$(12.2)\quad \begin{cases} \text{Define } v_0 := 0. \\ \text{Determine for } n = 0,1,\ldots \\ \qquad v_{n+1} = Uv_n \ . \end{cases}$$

From the nonnegativity of the reward structure it follows that $U0 \geq 0$. So, by the monotonicity of U, the sequence v_n converges monotonically to a - not necessarily finite - limit, v_∞ say:

$$v_\infty = \lim_{n\to\infty} v_n \ .$$

Since A and B are finite, it follows that for all $i \in S$ the value of the matrix game with entries

$$r(i,a,b) + \sum_{j\in S} p(i,a,b,j)v_n(j) \ , \quad a \in A \ , \ b \in B \ ,$$

converges to the value of the matrix game with entries

$$r(i,a,b) + \sum_{j\in S} p(i,a,b,j)v_\infty(j) \ , \quad a \in A \ , \ b \in B \ ,$$

even if some of the entries are equal to $+\infty$ (cf. section 10.2).
So for all $i \in S$

$$v_{n+1}(i) \to (Uv_\infty)(i) \quad (n \to \infty) \ ,$$

which implies

$$(12.3)\qquad Uv_\infty = v_\infty \ .$$

From (12.3) it is almost immediate that v_∞ is the value of the infinite-stage positive MG and that player II has an optimal stationary strategy.

Namely, by playing an optimal strategy for the n-stage game first and playing an arbitrary strategy thereafter, player I guarantees himself an expected income of at least v_n. Thus

$$(12.4) \qquad \sup_{\pi\in\Pi} \inf_{\gamma\in\Gamma} v(\pi,\gamma) \geq \lim_{n\to\infty} v_n = v_\infty \ .$$

On the other hand, there exists a policy h^* for player II satisfying

$$L(f,h^*)v_\infty \leq v_\infty \quad \text{for all } f \in F_I \ .$$

Hence

$$(12.5) \qquad v(\pi,h^*) \leq v_\infty \quad \text{for all } \pi \in \Pi \ .$$

To prove (12.5) observe that

$$v(\pi,h^*) = \lim_{n\to\infty} v_n(\pi,h^*,0)$$

and that (cf. section 10.2)

$$v_n(\pi,h^*,0) \leq v_n(\pi,h^*,v_\infty) \leq U^n v_\infty = v_\infty \ .$$

Now it follows from (12.4) and (12.5) that v_∞ is the value of the infinite-stage game and that h^* is an optimal stationary strategy for player II. The value of the game is further denoted by v^*.
For the positive MDP, ORNSTEIN [1969] has proved the existence of a uniformly ε-optimal strategy in the multiplicative sense, see theorem 2.23. This result has been partly generalized to the case of positive games by KUMAR and SHIAU [1979]. To be precise, they proved that, if v^* is bounded, a stationary strategy f exists such that

$$v(f,\gamma) \geq v^* - \varepsilon e \quad \text{for all } \gamma \in \Gamma \ .$$

Although v_n converges to v^* the scheme (12.2) is in general not of much use, since it does not provide an upper bound on v^* and there is no possibility to recognize whether e.g. the policies obtained at the n-stage of the successive approximation scheme are nearly optimal. Therefore, we further consider the specific positive MG satisfying properties (12.1(i)-(iii)).

In the sequel we will not explicitly incorporate the specific state * into the state space and further S will denote the set of states unequal to *. So, in the sequel we have defective transition probabilities. The reason

for doing this is that it somewhat simplifies our notations. Further, the game is said to have terminated if the system has reached * (left S).

In section 3 it will be shown that the method of standard successive approximations yields arbitrary close bounds on v^* and nearly-optimal stationary strategies for the two players. Before we can do this, first some results have to be derived concerning the duration of the game (time till absorption in *). This will be done in section 2.

12.2. SOME PRELIMINARY RESULTS

This section studies for a number of situations the asymptotic behaviour of the probability that the actual play has not yet been terminated.

Let $\hat{h}$ be the policy for player II that takes in each state i the specific action b(i) that terminates the play immediately. Then, clearly,

$$(12.6) \qquad Uv \leq \sup_{f \in F_I} L(f,\hat{h})v \leq Ce \quad \text{for all } v \geq 0 .$$

and

$$(12.7) \qquad v^* \leq \sup_{\pi \in \Pi} v(\pi,\hat{h}) \leq Ce .$$

Also, by (12.1(ii)),

$$(12.8) \qquad v^* \geq U0 \geq ce .$$

Denote for all $i \in S$, $\pi \in \Pi$ and $\gamma \in \Gamma$ by $p_n(i,\pi,\gamma)$ the probability that, given the initial state i and the strategies π and γ, the system has not yet reached the absorbing state * at time n, n = 1,2,... .
Further, let $\gamma_n(v)$ be an arbitrary optimal strategy for player II in the n-stage game with terminal payoff v, and let $\gamma_n(\pi,v)$ be an optimal strategy for player II if it is already known that player I will use strategy π.
Then we have the following lemma.

LEMMA 12.1. *If* $v \geq 0$, *then for all* $\pi \in \Pi$

$$(i) \qquad p_n(i,\pi,\gamma_n(v)) \leq \min \{1 , C / (nc + \min_{i \in S} v(i)\} ,$$

$$(ii) \qquad p_n(i,\pi,\gamma_n(\pi,v)) \leq \min \{1 , C / (nc + \min_{i \in S} v(i)\} .$$

PROOF.

(i) Clearly,

$$v_n(i,\pi,\gamma_n(v),v) \geq p_n(i,\pi,\gamma_n(v))(nc + \min_{i\in S} v(i)) .$$

By (12.6) also $U^n v \leq Ce$, so

$$v_n(i,\pi,\gamma_n(v),v) \leq U^n v \leq Ce .$$

Hence,

$$Ce \geq p_n(i,\pi,\gamma_n(v))(nc + \min_{i\in S} v(i)) ,$$

from which (i) follows immdiately.

(ii) The proof of (ii) is similar. □

So,

$$p_n(i,\pi,\gamma_n(v)) = \mathcal{O}(\tfrac{1}{n}) \quad (n \to \infty) \quad \text{for all } v \geq 0 .$$

In corollary 12.4 we will see that for certain specific strategies $\gamma_n(v)$ the probability $p_n(i,\pi,\gamma_n(v))$ decreases even exponentially fast.

Lemma 12.1 enables us to prove the following results.

THEOREM 12.2.

(i) *Any strategy* f^* *with* $L(f^*,h)v^* \geq v^*$ *for all* $h \in F_{II}$ *is optimal for the* ∞*-stage game, i.e.,*

$$v(f^*,\gamma) \geq v^* \quad \textit{for all } \gamma \in \Gamma .$$

(ii) *For all* $v \geq 0$

$$\lim_{n\to\infty} U^n v = v^* .$$

PROOF.

(i) By (12.7),

$$v^*(i) \leq v_n(i,f^*,\gamma_n(f^*,0),v^*)$$

$$\leq v_n(i,f^*,\gamma_n(f^*,0),0) + p_n(i,f^*,\gamma_n(f^*,0))C .$$

So, by lemma 12.1(ii),

$$v_n(f^*,\gamma_n(f^*,0),0) \geq v^* + \mathcal{O}(\tfrac{1}{n}) \quad (n \to \infty) .$$

Hence for all $\gamma \in \Gamma$

$$v(f^*,\gamma) = \lim_{n\to\infty} v_n(f^*,\gamma,0) \geq \lim_{n\to\infty} v_n(f^*,\gamma_n(f^*,0),0) \geq v^* .$$

(ii) Clearly, for all $v \geq 0$,

$$\liminf_{n\to\infty} U^n v \geq \lim_{n\to\infty} U^n 0 = v^* ,$$

so it suffices to prove

$$\limsup_{n\to\infty} U^n v \leq v^* .$$

By (12.6) it is even sufficient to prove that

$$\limsup_{n\to\infty} U^n Ce \leq v^* .$$

Let $\pi^{(n)}$ be an optimal strategy for player I in the n-stage game with terminal payoff Ce. Then

$$\begin{aligned}(U^n Ce)(i) &\leq v_n(i,\pi^{(n)},h^*,Ce) \\ &\leq v_n(i,\pi^{(n)},h^*,v^*) + p_n(i,\pi^{(n)},h^*,v^*)C \\ &\leq v^*(i) + O(\tfrac{1}{n}) \quad (n\to\infty) .\end{aligned}$$

The latter inequality follows from the optimality of h^* for the n-stage game with terminal payoff v^* with $U^n v^* = v^*$ and lemma 12.1(i).
Hence

$$\limsup_{n\to\infty} U^n v \leq \limsup_{n\to\infty} U^n Ce \leq v^* ,$$

which completes the proof. □

If one wants to find bounds on v^* and nearly-optimal stationary strategies for the two players, then the inequalities in lemma 12.1 are too weak. We will show that for certain optimal n-stage strategies $\gamma_n(v)$ the probability $p_n(i,\pi,\gamma_n(v))$ tends to zero exponentially fast.

Let $v \geq 0$ be arbitrary and let $\{h_n^v\}$ be a sequence of policies satisfying

$$L(f,h_n^v)U^n v \leq U^{n+1}v \quad \text{for all } f \in F_I , \quad n = 0,1,\ldots .$$

Define

$$\gamma_n^*(v) := (h_{n-1}^v, h_{n-2}^v, \ldots, h_0^v) .$$

Then $\gamma_n^*(v)$ is not only optimal in the n-stage MG with terminal payoff v but $\gamma_n^*(v)$ is also optimal in the k-stage MG with terminal payoff $U^{n-k}v$ for all $k < n$.

Define further for all $v \geq 0$ and all $n = 0,1,\ldots$

$$p_n(v) := \min \{1, C / (nc + \min_{i \in S} v(i))\} .$$

Then for all $n,m \geq 0$, all $\pi \in \Pi$ and all $v \geq 0$

$$p_n(i,\pi,\gamma_{n+m}^*(v)) \leq p_n(U^m v) . \tag{12.9}$$

Now we can prove that $p_n(i,\pi,\gamma_n^*(v))$ decreases exponentially fast.

LEMMA 12.3. *For all* $n,m \geq 0$, *for all* $\pi \in \Pi$ *and all* $v \geq 0$

$$p_{n+m}(i,\pi,\gamma_{n+m}^*(v)) \leq p_n(i,\pi,\gamma_{n+m}^*(v)) p_m(v) , \quad i \in S .$$

PROOF. From lemma 11.1(i) and (12.9) we have for all $i \in S$

$$p_{n+m}(i,\pi,\gamma_{n+m}^*(v)) \leq$$

$$\leq \sum_{j \in S} \mathbb{P}_{i,\pi,\gamma_{n+m}^*(v)}(X_n = j) \sum_{k \in S} \sup_{\pi' \in \Pi} \mathbb{P}_{j,\pi',\gamma_m^*(v)}(X_m = k)$$

$$\leq p_n(i,\pi,\gamma_{n+m}^*(v)) \sup_{j \in S} \sup_{\pi' \in \Pi} p_m(j,\pi',\gamma_m^*(v))$$

$$\leq p_n(i,\pi,\gamma_{n+m}^*(v)) p_m(v) . \qquad \square$$

Since $v \geq w$ implies $p_n(v) \geq p_n(w)$, lemma 12.3 yields the following corollary.

COROLLARY 12.4. *If* $v \geq 0$ *and* $n = km + \ell$, $k,\ell,m = 0,1,\ldots$, *then for all* π

$$p_n(i,\pi,\gamma_n^*(v)) \leq (p_m(0))^k p_\ell(0) .$$

If moreover $Uv \geq v$, *then*

$$p_n(i,\pi,\gamma_n^*(v)) \leq (p_m(v))^k p_\ell(v) .$$

PROOF. Straightforward. $\square$

12.3. BOUNDS ON v^* AND NEARLY-OPTIMAL STATIONARY STRATEGIES

Corollary 12.4 enables us to obtain a better upperbound on v^*.

THEOREM 12.5. *Let* $v \in V$ *satisfy* $0 \leq v \leq Uv$ *and let* m *be such that* $p_m(v) < 1$. *Then*

$$Uv \leq v^* \leq Uv + (1 - p_m(v))^{-1} \sum_{k=1}^{m} p_k(v) \|Uv - v\|_e e .$$

PROOF. By theorem 12.2(ii) we have

$$v^* = \lim_{n\to\infty} U^n v = Uv + \sum_{n=1}^{\infty} (U^{n+1}v - U^n v) .$$

So, by the monotonicity of U, we have $v^* \geq Uv$.
Further, let the policies f_t^v satisfy for all $h \in F_{II}$

$$L(f_t^v, h) U^t v \geq U^{t+1} v , \quad t = 0,1,\ldots ,$$

and define

$$\pi_n^*(v) = (f_{n-1}^v, \ldots, f_0^v) , \quad n = 1,2,\ldots .$$

Then for all $n = 1,2,\ldots$

$$\begin{aligned} U^{n+1}v - U^n v &= L(f_n^v, h_n^v) \cdots L(f_1^v, h_1^v) Uv - L(f_{n-1}^v, h_{n-1}^v) \cdots L(f_0^v, h_0^v) v \\ &\leq L(f_n^v, h_{n-1}^v) \cdots L(f_1^v, h_0^v) Uv - L(f_n^v, h_{n-1}^v) \cdots L(f_1^v, h_0^v) v \\ &\leq \sup_{i\in S} p_n(i, \pi_{n+1}^*(v), \gamma_n^*(v)) \|Uv - v\|_e e . \end{aligned}$$

Hence, by corollary 11.4

$$\begin{aligned} v^* &\leq Uv + \sum_{n=1}^{\infty} \sup_{i\in S} p_n(i, \pi_{n+1}^*(v), \gamma_n^*(v)) \|Uv - v\|_e e \\ &\leq Uv + \sum_{n=1}^{\infty} p_n(v) \|Uv - v\|_e e \\ &\leq Uv + \sum_{\ell=0}^{\infty} \sum_{k=1}^{m} (p_m(v))^{\ell} p_k(v) \|Uv - v\|_e e \\ &= Uv + (1 - p_m(v))^{-1} \sum_{k=1}^{m} p_k(v) \|Uv - v\|_e e . \end{aligned}$$

□

Since for all $v_0 \geq 0$ we have $U^n v_0 \to v^*$ (theorem 12.2(ii)) also $U^{n+1}v_0 - U^n v_0 \to 0$ $(n \to \infty)$. From the proof of theorem 12.2 one easily sees that the convergence of $U^n v_0$ to v^* is uniform in the initial state, hence $\|U^{n+1}v_0 - U^n v_0\|_e$ tends to zero if n tends to infinity. So theorem 12.5 can be used to obtain good bounds on v^* (take $v = U^n v_0$ for n sufficiently large).

The fact that if $v \geq 0$ all sensible strategies for player II terminate the n-stage game with terminal payoff v, also leads to the following result.

THEOREM 12.6. *If for some* $v \geq 0$ *we have*

$$L(f,h)v \geq v \quad \textit{for all } h \in F_{II} ,$$

then

$$v(f,\gamma) \geq \min_{h \in F_{II}} L(f,h)v \geq v \quad \textit{for all } \gamma \in \Gamma .$$

PROOF. Let $\gamma_n(f,0) = (h_{n-1},\ldots,h_0)$ be an optimal reply to f for player II in the n-stage game with terminal payoff 0, then for all $\gamma \in \Gamma$

$$\begin{aligned} v_n(f,\gamma,0) &\geq v_n(f,\gamma_n(f,0),0) = L(f,h_{n-1}) \cdots L(f,h_0)0 \\ &\geq L(f,h_{n-1}) \cdots L(f,h_0)v - \sup_{i \in S} p(i,f,\gamma_n(f,0))\|v\|_e \\ &\geq L(f,h_{n-1})v - \sup_{i \in S} p(i,f,\gamma_n(f,0))\|v\|_e . \end{aligned}$$

The result now follows with lemma 11.1(i) by letting n tend to infinity. □

COROLLARY 12.7. *Let* f^* *satisfy*

$$L(f^*,h)v^* \geq v^* \quad \textit{for all } h \in F_{II} ,$$

then the strategy f^* *is optimal for player I for the infinite-stage game:*

$$v(f^*,\gamma) \geq v^* \quad \textit{for all } \gamma \in \Gamma .$$

Further we see that theorem 12.6 in combination with theorem 12.5 enables us to obtain from a monotone standard successive approximation scheme a nearly-optimal stationary strategy for player I.

Next it will be shown how a nearly-optimal stationary strategy for player II can be found.

THEOREM 12.8. *Let* $v \in V$ *and* $h \in F_{II}$ *satisfy*

(i) $\alpha e \leq v \leq Ce$ *for some* $\alpha > 0$,

(ii) $L(f,h)v \leq v + \varepsilon e$ *for some* $0 \leq \varepsilon < c$ *and all* $f \in F_I$.

Then a constant ρ, *with* $0 \leq \rho \leq 1 - \frac{c-\varepsilon}{C}$, *exists satisfying*

$$P(f,h)v \leq \rho v \quad \textit{for all } f \in F_I ,$$

which implies

$$v(\pi,h) \leq \max_{f \in F_I} L(f,h)v + \rho(1-\rho)^{-1} \frac{\varepsilon}{\alpha} v \quad \textit{for all } \pi \in \Pi . \tag{12.10}$$

PROOF. For all $f \in F_I$ we have

$$P(f,h)v = L(f,h)v - r(f,h) \leq v + \varepsilon e - ce \leq \rho v ,$$

for some $\rho \leq 1 - \frac{c-\varepsilon}{C}$.

Now let $\pi = (f_0, f_1, \ldots)$ be an arbitrary Markov strategy for player I (it suffices to consider only Markov strategies), then

$$\begin{aligned}
v_n(\pi,h,0) &= L(f_0,h) \cdots L(f_{n-1},h)0 \\
&\leq L(f_0,h) \cdots L(f_{n-1},h)v \leq L(f_0,h) \cdots L(f_{n-2},h)(v + \varepsilon e) \\
&= L(f_0,h) \cdots L(f_{n-2},h)v + P(f_0,h) \cdots P(f_{n-2},h)\varepsilon e \\
&\leq L(f_0,h) \cdots L(f_{n-2},h)v + P(f_0,h) \cdots P(f_{n-2},h) \frac{\varepsilon}{\alpha} v \\
&\leq L(f_0,h) \cdots L(f_{n-2},h)v + \rho^{n-1} \frac{\varepsilon}{\alpha} v \\
&\leq \ldots \leq L(f_0,h)v + (\rho + \rho^2 + \ldots + \rho^{n-1}) \frac{\varepsilon}{\alpha} v .
\end{aligned}$$

Letting n tend to infinity and taking the maximum with respect to f_0 yields (12.10). □

Clearly the right hand side in (12.10) is also an upperbound on v^* and this remains true if $\max_{f \in F_I} L(f,h)v$ is replaced by Uv.

Now consider the successive approximation scheme:

Choose $v_0 \geq 0$ such that $Uv_0 \geq v_0$.

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

$$v_{n+1} = Uv_n .$$

Then, since $U^n v_0$ converges to v^* uniformly in the initial state and since $U^{n+1}v_0 \geq U^n v_0$ by the monotonicity of U, we can apply theorems 12.5, 12.6 and 12.8 with $v = U^n v_0$ for n sufficiently large to obtain good bounds on v^* and nearly-optimal stationary strategies for the two players.
Note that the function v in theorem 12.8 is strongly excessive with respect to the set of transition "matrices" P(f,h), where h - the policy mentioned in the theorem - is held fixed. So the resulting MDP with h fixed is contracting in the sense of chapter 5 and (12.10) is rather similar to theorem 5.12(ii).

To obtain a lowerbound on v^* we have used $Uv \geq v$ (theorem 12.5). Also for the near-optimality of f in theorem 12.6 the monotonicity has been used. The following theorem demonstrates how in the nonmonotonic case as well a lowerbound on v^* and a nearly-optimal stationary strategy for player I can be found.

THEOREM 12.9. *Let the policy* $f \in F_I$ *satisfy for some* v, *with* $0 \leq v \leq Ce$,

$$L(f,h)v \geq v - \varepsilon e \quad \textit{for all } h \in F_{II} ,$$

where $\varepsilon \geq 0$ *is some constant, then*

$$(12.11) \qquad v(f,\gamma) \geq \min_{h \in F_{II}} L(f,h)v - \varepsilon \frac{(C-c)C}{c^2} e \quad \textit{for all } \gamma \in \Gamma .$$

PROOF. Let the stationary strategy $\tilde{h}$ be an optimal reply to f in the ∞-stage game. That such an optimal stationary strategy exists follows from the fact that if player I uses a fixed stationary strategy, then the remaining minimization problem for player II is an MDP. (Formally, this needs a proof since player II may choose his actions dependent of previous actions of player I, but this will not be worked out here.) Considered as a maximization problem this is a negative MDP for which by corollary 2.17 an optimal stationary strategy exists.
Define

$$\tilde{v} := v(f,\tilde{h}) .$$

Then $L(f,\tilde{h})\tilde{v} = \tilde{v}$, which yields

$$P(f,\tilde{h})\tilde{v} \leq \tilde{v} - ce .$$

So, with $ce \leq \tilde{v} \leq Ce$, thus $\tilde{v}/C \leq e$, also

$$P(f,\tilde{h})\tilde{v} \le \tilde{v} - \frac{c}{C}\tilde{v} = (1 - \frac{c}{C})\tilde{v} .$$

One easily argues that

$$p_n(i,f,\tilde{h}) \to 0 \quad (n \to \infty) ,$$

so

$$v(f,\tilde{h}) = \lim_{n\to\infty} L^n(f,\tilde{h})v .$$

Further,

$$L^n(f,\tilde{h})v \ge L^{n-1}(f,\tilde{h})(v - \varepsilon e) = L^{n-1}(f,\tilde{h})v - P^{n-1}(f,\tilde{h})\varepsilon e$$

$$\ge \ldots \ge L(f,\tilde{h})v - \varepsilon[P(f,\tilde{h}) + \ldots + P^{n-1}(f,\tilde{h})]e .$$

With $e \le c^{-1}\tilde{v}$ (from $\tilde{v} \ge ce$) it follows that

$$P^k(f,\tilde{h})e \le c^{-1}P^k(f,\tilde{h})\tilde{v} \le c^{-1}(1 - \frac{c}{C})^k\tilde{v} \le \frac{C}{c}(1 - \frac{c}{C})^k e .$$

So for all n

$$L^n(f,\tilde{h})v \ge L(f,\tilde{h})v - \varepsilon \sum_{k=1}^{n-1} \frac{C}{c}(1 - \frac{c}{C})^k e .$$

Thus, letting n tend to infinity, it follows from the optimality of $\tilde{h}$ that

$$v(f,\gamma) \ge \min_{h\in F_{II}} L(f,h)v - \varepsilon \frac{(C-c)C}{c^2} e \quad \text{for all } \gamma \in \Gamma . \qquad \square$$

The right hand side in (12.11) is clearly also a lowerbound on v^*, so theorems 12.8 and 12.9 can be combined to obtain good bounds on v^* and nearly-optimal stationary strategies for both players.

CHAPTER 13

SUCCESSIVE APPROXIMATIONS FOR THE AVERAGE-REWARD MARKOV GAME

13.1. INTRODUCTION AND SOME PRELIMINARIES

This chapter deals with the average-reward Markov game with finite state space $S := \{1,2,\ldots,N\}$ and finite action spaces A and B for players I and II, respectively. In general, these games neither have a value within the class of stationary strategies nor within the class of Markov strategies. This has been shown by GILLETTE [1957] and by BLACKWELL and FERGUSON [1968], respectively. Gillette, and afterwards HOFFMAN and KARP [1966] have proved that the game does have a value within the class of stationary strategies, if for each pair of stationary strategies the underlying Markov chain is irreducible. This condition has been weakened by ROGERS [1969] and by SOBEL [1971], who still demand the underlying Markov chains to be unichained but allow for some transient states. FEDERGRUEN [1980] has shown that the unichain restriction may be replaced by the condition that the underlying Markov chains corresponding to a pair of (pure) stationary strategies all have the same number of irreducible subchains. Only recently MONASH [1979], and independently MERTENS and NEYMAN [1980], have shown that every average-reward MG with finite state and action spaces has a value within the class of history-dependent strategies.

In this chapter we consider for two situations the method of standard successive approximations:

$$(13.1)\quad \begin{cases} \text{Choose } v_0 \in \mathbb{R}^N. \\ \text{Determine for } n = 0,1,\ldots \\ \qquad v_{n+1} = Uv_n\ . \end{cases}$$

In the first case it is assumed that for each pair of pure stationary strategies the underlying Markov chain is unichained. In the second case it is assumed that the functional equation

(13.2) $\quad Uv = v + ge$

has a solution $v \in \mathbb{R}^N$, $g \in \mathbb{R}$, say.
In both cases we further assume the strong aperiodicity assumption to hold, i.e., for some $\alpha > 0$

(13.3) $\quad P(f,h) \geq \alpha I$ for all $f \in F_I$, $h \in F_{II}$.

At the end of this section it will be shown that the latter assumption is - as in the case of the MDP - no real restriction.
In section 2 we will see that the unichain assumption implies that the function equation (13.2) has a solution, so the first case is merely an example of the second. The fact that (13.2) has a solution $(\tilde{v},g_*)$ implies (corollary 13.2) that the game has a value independent of the initial state, namely g_*e, and that both players have optimal stationary strategies. So, in the two cases considered here, the value of the game will be independent of the initial state. This value is further denoted by g_*e.

In sections 2 (the unichain case) and 3 (the case that (13.2) has a solution) it is shown that the method of standard successive approximations formulated in (13.1) yields good bounds on the value of the game and nearly-optimal stationary strategies for the two players.
The results of this chapter can be found in Van der WAL [1980b].

Before we are going to study the unichain case some preliminaries are considered.
First observe that, since S is finite,

$$r(i,a,b) + \sum_{j \in S} p(i,a,b,j)v(j)$$

is finite for all $v \in \mathbb{R}^N$, all $a \in A$, $b \in B$ and all $i \in S$. Hence one may write for all $v \in V$

$$Uv = \max_{f \in F_I} \min_{h \in F_{II}} L(f,h)v .$$

So, both players have optimal policies in the 1-stage game with terminal payoff v.

Since in the two situations treated in this chapter the value of the game is independent of the initial state the following basic lemma (cf. lemma 6.8) is very useful.

LEMMA 13.1. *Let* $v \in \mathbb{R}^N$ *be arbitrary, then*

(i) $$\inf_{\gamma\in\Gamma} g(f,\gamma) \geq \min_{h\in F_{II}} \min_{i\in S} (L(f,h)v - v)(i)e ,$$

(ii) $$\sup_{\pi\in\Pi} g(\pi,h) \leq \max_{f\in F_I} \max_{i\in S} (L(f,h)v - v)(i)e .$$

PROOF. We only prove (i), the proof of (ii) being similar.
Let player I play the stationary strategy f, then the extension of the Derman and Strauch theorem by GROENEWEGEN and WESSELS [1976] says that player II may restrict himself to Markov strategies. So, let $\gamma = (h_0,h_1,\ldots)$ be an arbitrary Markov strategy for player II. Then

$$g(f,\gamma) = \liminf_{n\to\infty} n^{-1} v_n(f,\gamma,0) = \liminf_{n\to\infty} n^{-1} v_n(f,\gamma,v) . \tag{13.4}$$

Further,

$$\begin{aligned} v_n(f,\gamma,v) &= L(f,h_0)\cdots L(f,h_{n-1})v \\ &= L(f,h_0)\cdots L(f,h_{n-2})(v + L(f,h_{n-1})v - v) \\ &= L(f,h_0)\cdots L(f,h_{n-2})v + P(f,h_0)\cdots P(f,h_{n-2})\cdot \\ &\qquad \cdot (L(f,h_{n-1}v - v) \\ &\geq L(f,h_0)\cdots L(f,h_{n-2})v + P(f,h_0)\cdots P(f,h_{n-2})\cdot \\ &\qquad \cdot \min_{h\in F_{II}} \min_{i\in S} (L(f,h)v - v)(i)e \\ &= L(f,h_0)\cdots L(f,h_{n-2})v + \min_{h\in F_{II}} \min_{i\in S} (L(f,h)v - v)(i)e \\ &\geq \ldots \geq v + n \min_{h\in F_{II}} \min_{i\in S} (L(f,h)v - v)(i)e . \end{aligned} \tag{13.5}$$

Now (i) follows immediately from (13.4) and (13.5). □

Lemma 13.1 yields the following corollary.

COROLLARY 13.2.

(i) *If for some* $g \in \mathbb{R}$ *and* $v \in \mathbb{R}^N$ *we have*

$$Uv = v + g_* e$$

then $g_* e$ *is the value of the game, and policies* f_v *and* h_v *satisfying*

$$L(f,h_v)v \leq Uv \leq L(f_v,h)v \quad \textit{for all } f \in F_I\,,\ h \in F_{II}$$

yield optimal stationary strategies for players I and II, respectively.

(ii) *Let* $v \in \mathbb{R}^N$ *be arbitrary and the policies* f_v *and* h_v *satisfy*

$$L(f,h_v)v \leq Uv \leq L(f_v,h)v \quad \textit{for all } f \in F_I\,,\ h \in F_{II}\,,$$

then f_v *and* h_v *are both* $\mathrm{sp}(Uv - v)$*-optimal. I.e., let* g^* *denote the value of the game, then*

$$g(f_v,\gamma) \geq g^* - \mathrm{sp}(Uv - v)e \quad \textit{for all } \gamma \in \Gamma$$

and

$$g(\pi,h_v) \leq g^* + \mathrm{sp}(Uv - v)e \quad \textit{for all } \pi \in \Pi\,.$$

(iii) *For all* $v \in V$

$$\min_{i\in S} (Uv - v)(i)e \leq g^* \leq \max_{i\in S} (Uv - v)(i)e\,.$$

PROOF. The proof follows immediately from

$$g^* \geq \inf_{\gamma\in\Gamma} g(f_v,\gamma) \geq \min_{i\in S} (Uv - v)(i)e$$

and

$$g^* \leq \sup_{\pi\in\Pi} g(\pi,h_v) \leq \max_{i\in S} (Uv - v)(i)e$$

$$= \min_{i\in S} (Uv - v)(i)e + \mathrm{sp}(Uv - v)e\,. \qquad \square$$

So corollary 13.2(ii) and (iii) show that it makes sense to study the successive approximation scheme (13.1) if the value of the game is independent of the initial state. And further that the method yields good bounds for the value g^* and nearly-optimal stationary strategies for the two players if

(13.6) $\quad sp(v_{n+1} - v_n) \to 0 \quad (n \to \infty)$.

In the next two sections we will use the strong aperiodicity assumption to prove that (13.6) holds for the two cases we are interested in.
Before this will be done, we first show that the strong aperiodicity assumption - as in the MDP case - is not a serious restriction.
Let our MG be characterized by (S,A,B,p,r), then one may use the data-transformation of SCHWEITZER [1971] again (cf. section 6.3) to obtain an equivalent MG characterized by $(S,A,B,\hat{p},\hat{r})$ with

$$\hat{p}(i,a,b,i) := \alpha + (1-\alpha)p(i,a,b,i) ,$$

$$\hat{p}(i,a,b,j) := (1-\alpha)p(i,a,b,j) , \quad j \neq i ,$$

$$\hat{r}(i,a,b) := (1-\alpha)r(i,a,b) ,$$

for all $i \in S$, $a \in A$, $b \in B$, where α is some constant with $0 < \alpha < 1$.
Writing $\hat{L}$, $\hat{U}$ and $\hat{g}$ for the operators L and U and the function g in the transformed MG, we obtain

$$\hat{L}(f,h)v - v = (1-\alpha)r(f,h) + [\alpha I + (1-\alpha)P(f,h)]v - v$$

$$= (1-\alpha)[r(f,h) + P(f,h)v - v] = (1-\alpha)(L(f,h)v - v) .$$

Whence, with $1-\alpha > 0$, also

$$\hat{U}v - v = (1-\alpha)(Uv - v) .$$

So, if the functional equation (13.2) of the original MG has a solution, $(g_*,\tilde{v})$ say, then the functional equation (13.2) of the transformed game, $\hat{U}v = v + ge$, has a solution $((1-\alpha)g_*,\tilde{v})$. So $(1-\alpha)g_*e$ is the value of the transformed game.
Conversely, if $\hat{U}\hat{v} = \hat{v} + \hat{g}_*e$, then $U\hat{v} = \hat{v} + (1-\alpha)^{-1}\hat{g}_*e$. Further, let for example the policy $\hat{f}$ satisfy

$$\hat{L}(\hat{f},h)v - v \geq \hat{g}_*e - \varepsilon e \quad \text{for all } h \in F_{II} ,$$

which implies by corollary 13.2(ii) that $\hat{f}$ is ε-optimal in the transformed game. Then $\hat{f}$ is $(1-\alpha)^{-1}\varepsilon$-optimal in the original game as follows from corollary 13.2(ii), with

$$L(\hat{f},h)v - v = (1-\alpha)^{-1}(\hat{L}(\hat{f},h)v - v)$$

$$\geq (1-\alpha)^{-1}\hat{g}_*e - (1-\alpha)^{-1}\varepsilon e = g_*e - (1-\alpha)^{-1}\varepsilon e .$$

So we see that the two problems are equivalent with respect to those features that interest us: the value and (nearly-) optimal stationary strategies.

13.2. THE UNICHAINED MARKOV GAME

In this section we consider the unichained MG, i.e., the case that for each pair of pure stationary strategies the underlying Markov chain consists of one recurrent subchain and possibly some transient states. Further it is assumed that the strong aperiodicity assumption, (13.3), holds.
It is shown that in this case the method of standard successive approximations (13.1) converges, i.e., that

$$sp(v_{n+1} - v_n) \to 0 \quad (n \to \infty) \ .$$

The line of reasoning is similar as in section 9.4. First we derive a scrambling condition like lemma 9.7 from which, along the lines of lemma 9.8, it follows that $sp(v_{n+1} - v_n)$ converges to zero even exponentially fast.

LEMMA 13.3. *There exists a constant* η, *with* $0 < \eta \leq 1$, *such that for all* $\pi, \tilde{\pi} \in M_I$ *and* $\gamma, \tilde{\gamma} \in M_{II}$ *and for all* $i,j \in S$

$$(13.7) \qquad \sum_{k \in S} \min \{ \mathbb{P}_{i,\pi,\gamma}(X_{N-1} = k) \ , \ \mathbb{P}_{i,\tilde{\pi},\tilde{\gamma}}(X_{N-1} = k) \} \geq \eta \ .$$

(Recall that N *is the number of states in* S.*)*

PROOF. The proof is very similar to the proof of lemma 9.7.
First it is shown that the left hand side in (13.7) is positive for any four pure Markov strategies $\pi = (f_1, f_2, \ldots)$, $\tilde{\pi} = (\tilde{f}_1, \tilde{f}_2, \ldots)$, $\gamma = (h_1, h_2, \ldots)$ and $\tilde{\gamma} = (\tilde{h}_1, \tilde{h}_2, \ldots)$.
Fix these four strategies and define for all $i \in S$ and all $n = 0,1,\ldots,N-1$ the sets $S(i,n)$ and $\tilde{S}(i,n)$ by

$$S(i,0) := \tilde{S}(i,0) := \{i\} \ ,$$

$$S(i,n) := \{ j \in S \mid P(f_1,h_1) \cdots P(f_n,h_n) > 0 \} \ , \quad n = 1,\ldots,N-1$$

$$\tilde{S}(i,n) := \{ j \in S \mid P(\tilde{f}_1,\tilde{h}_1) \cdots P(\tilde{f}_n,\tilde{h}_n) > 0 \} \ , \quad n = 1,\ldots,N-1$$

Clearly the sets $S(i,n)$ and $\tilde{S}(i,n)$ are monotonically nondecreasing in n.
For example, if $j \in S(i,n)$, then

$$P(f_1,h_1) \cdots P(f_n,h_n)(i,j) > 0 ,$$

and, by the strong aperiodicity assumption,

$$P(f_{n+1},h_{n+1})(j,j) > 0 ,$$

so also

$$P(f_1,h_1) \cdots P(f_n,h_n)P(f_{n+1},h_{n+1})(i,j) > 0 ,$$

hence $j \in S(i,n+1)$
Further, if $S(i,n) = S(i,n+1)$ $[\tilde{S}(i,m) = \tilde{S}(i,m+1)]$, then the set $S(i,n)$ $[\tilde{S}(i,m+1)]$ is closed under $P(f_{n+1},h_{n+1})$ $[P(\tilde{f}_{m+1},\tilde{h}_{m+1})]$.
In order to prove that the left hand side in (13.7) is positive, we have to prove that the intersection

$$S(i,N-1) \cap \tilde{S}(j,N-1)$$

is nonempty for all $i,j \in S$.
Suppose to the contrary that for some pair (i_0,j_0) this intersection is empty. Then for some $n,m < N-1$

$$S(i_0,n) \text{ is closed under } P(f_{n+1},h_{n+1})$$

and

$$\tilde{S}(j_0,m) \text{ is closed under } P(\tilde{f}_{m+1},\tilde{h}_{m+1}) ,$$

and further $S(i_0,n)$ and $\tilde{S}(j_0,m)$ are disjoint.
But this implies that we can construct from f_{n+1} and $\tilde{f}_{m+1}$ and from h_{n+1} and $\tilde{h}_{m+1}$ policies f and h for which $P(f,h)$ has at least two nonempty disjoint subchains, which contradicts the unichain assumption.
Hence

$$S(i,N-1) \cap \tilde{S}(j,N-1)$$

is nonempty for all $i,j \in S$.
Since there are only finitely many pure $(N-1)$-stage Markov strategies there must exist a constant $\eta > 0$ for which (13.7) holds for all pure Markov strategies π, $\tilde{\pi}$, γ and $\tilde{\gamma}$. Moreover, it can be shown that the minimum of the left hand side of (13.7) within the set of Markov strategies is equal to the minimum within the set of pure Markov strategies. So the proof is complete. □

Next, this lemma will be used to prove that $sp(v_{n+1} - v_n)$ tends to zero exponentially fast.

Let $\{f_k\}$ and $\{h_k\}$ be sequences of policies satisfying for all $k = 0,1,\ldots$

$$L(f,h_k)v_k \leq v_{k+1} \leq L(f_k,h)v_k \quad \text{for all } f \in F_I\,,\ h \in F_{II}\,.$$

Then for all n

$$\begin{aligned} v_{n+2} - v_{n+1} &= L(f_{n+1},h_{n+1})v_{n+1} - L(f_n,h_n)v_n \\ &\leq L(f_{n+1},h_n)v_{n+1} - L(f_{n+1},h_n)v_n \\ &= P(f_{n+1},h_n)(v_{n+1} - v_n)\,. \end{aligned} \tag{13.8}$$

So,

$$v_{n+N} - v_{n+N-1} \leq P(f_{n+N-1},h_{n+N-2}) \cdots P(f_{n+1},h_n)(v_{n+1} - v_n)\,. \tag{13.9}$$

Similarly,

$$v_{n+N} - v_{n+N-1} \geq P(f_{n+N-2},h_{n+N-1}) \cdots P(f_n,h_{n+1})(v_{n+1} - v_n)\,. \tag{13.10}$$

Now let π, $\tilde{\pi}$, γ and $\tilde{\gamma}$ denote the (N-1)-stage Markov strategies $(f_{n+N-1},f_{n+N-2},\ldots,f_{n+1})$, $(f_{n+N-2},\ldots,f_n)$, $(h_{n+N-2},\ldots,h_n)$ and $(h_{n+N-1},\ldots,h_{n+1})$, respectively. Then we have from (13.9) and (13.10) for all $i,j \in S$

$$\begin{aligned} &(v_{n+N} - v_{n+N-1})(i) - (v_{n+N} - v_{n+N-1})(j) \\ &\leq \sum_{k\in S} [\mathbb{P}_{i,\pi,\gamma}(X_{N-1} = k) - \mathbb{P}_{j,\tilde{\pi},\tilde{\gamma}}(X_{N-1} = k)](v_{n+1} - v_n)(k) \\ &= \sum_{k\in S} [\mathbb{P}_{i,\pi,\gamma}(X_{N-1} = k) - \min\{\mathbb{P}_{i,\pi,\gamma}(X_{N-1} = k), \mathbb{P}_{j,\tilde{\pi},\tilde{\gamma}}(X_{N-1} = k)\}](v_{n+1} - v_n)(k) + \\ &- \sum_{k\in S} [\mathbb{P}_{j,\tilde{\pi},\tilde{\gamma}}(X_{N-1} = k) - \min\{\mathbb{P}_{i,\pi,\gamma}(X_{N-1} = k), \mathbb{P}_{j,\tilde{\pi},\tilde{\gamma}}(X_{N-1} = k)\}](v_{n+1} - v_n)(k) \\ &\leq \sum_{k\in S} [\mathbb{P}_{i,\pi,\gamma}(X_{N-1}=k) - \min\{\mathbb{P}_{i,\pi,\gamma}(X_{N-1}=k), \mathbb{P}_{j,\tilde{\pi},\tilde{\gamma}}(X_{N-1}=k)\}] \max_{\ell\in S}(v_{n+1} - v_n)(\ell) + \\ &- \sum_{k\in S} [\mathbb{P}_{j,\tilde{\pi},\tilde{\gamma}}(X_{N-1}=k) - \min\{\mathbb{P}_{i,\pi,\gamma}(X_{N-1}=k), \mathbb{P}_{j,\tilde{\pi},\tilde{\gamma}}(X_{N-1}=k)\}] \min_{\ell\in S}(v_{n+1} - v_n)(\ell) \\ &= [1 - \sum_{k\in S} \min\{\mathbb{P}_{i,\pi,\gamma}(X_{N-1} = k), \mathbb{P}_{j,\tilde{\pi},\tilde{\gamma}}(X_{N-1} = k)\}]\, sp(v_{n+1} - v_n)\,. \end{aligned}$$

Hence for all n = 0,1,...

(13.11) $\quad sp(v_{n+N} - v_{n+N-1}) \leq (1-\eta)\, sp(v_{n+1} - v_n)$,

where η is the constant in lemma 13.3.
This immediately leads to

THEOREM 13.4. *If the Markov game is unichained and the strong aperiodicity assumption holds, then we have for the standard successive approximation scheme* (13.1)

(i) $\quad v_{n+1} - v_n = g_* e + O((1-\eta)^{\frac{n}{N-1}}) \quad (n \to \infty)$,

with η *the constant in lemma 13.3.*
Further, for some $v^* \in V$,

(ii) $\quad v_n = n g_* e + v^* + O((1-\eta)^{\frac{n}{N-1}}) \quad (n \to \infty)$

and

(iii) $\quad Uv^* = v^* + g_* e$.

PROOF. (i) follows immediately from (13.11). Then (ii) follows from (i) and (iii) follows from (ii). That the constant involved is equal to g_* is immediate from corollary 13.2. □

So, if the MG is unichained and the strong aperiodicity assumption holds (for example as a result of Schweitzer's datatransformation), then the method of standard successive approximations yields an ε-band on the value of the game and ε-optimal stationary strategies for both players for all $\varepsilon > 0$ and this even exponentially fast.

13.3. THE FUNCTIONAL EQUATION $Uv = v + ge$ HAS A SOLUTION

In this section it will be shown that, if the functional equation $Uv = v + ge$ has a solution, (g_*, v^*) say, and if the strong aperiodicity assumption holds, then $v_{n+1} - v_n$ converges to $g_* e$. By corollary 13.2 this implies that the method of standard successive approximations (13.1) yields a good approximation of the value of the game and nearly-optimal stationary strategies for the two players.

The line of reasoning will be similar to the one in chapter 9. A different proof can be found in Van der WAL [1980b].

Define for all n = 0,1,...

$$g_n := v_{n+1} - v_n \ , \tag{13.12}$$

$$\ell_n := \min_{i \in S} g_n(i) \ , \tag{13.13}$$

$$u_n := \max_{i \in S} g_n(i) \ . \tag{13.14}$$

It will be shown that $\{\ell_n\}$ is monotonically nondecreasing with limit g_* and that $\{u_n\}$ is monotonically nonincreasing also with limit g_*.
Therefore we first need the following lemma.

LEMMA 13.5. *For all* $v,w \in \mathbb{R}^N$

$$\min_{i \in S} (v-w)(i)e \leq Uv - Uw \leq \max_{i \in S} (v-w)(i)e \ ,$$

so

$$sp(Uv - Uw) \leq sp(v-w) \ .$$

PROOF. Let f_v, h_v, f_w and h_w satisfy for all $f \in F_I$ and $h \in F_{II}$

$$L(f,h_v)v \leq Uv \leq L(f_v,h)v$$

and

$$L(f,h_w)w \leq Uw \leq L(f_w,h)w \ .$$

Then

$$Uv - Uw \geq L(f_w,h_v)v - L(f_w,h_v)w = P(f_w,h_v)(v-w) \geq \min_{i \in S} (v-w)(i)e \ .$$

Similarly one establishes the second inequality. □

From this lemma one immediately obtains the following corollary.

COROLLARY 13.6.

(i) *For all* n = 0,1,...

$$\ell_n \leq \ell_{n+1} \leq g_* \leq u_{n+1} \leq u_n \ ;$$

(ii) $sp(v_n)$ *is bounded in* n.

PROOF. (i) follows from lemma 13.5 with $v = v_{n+1}$ and $w = v_n$ and corollary 13.2.

(ii) $$sp(v_n) = sp(U^n v_0) \le sp(U^n v_0 - U^n v^*) + sp(U^n v^*)$$
$$\le sp(v_0 - v^*) + sp(v^*) \ . \qquad \square$$

Now we will follow the line of reasoning in section 9.4 to prove that $\ell_* := \lim_{n\to\infty} \ell_n$ is equal to g_*.

From an inequality like (13.10) one immediately obtains for all n and k

$$g_{n+k} \ge \alpha^k g_n + (1-\alpha^k)\ell_n e \ .$$

Especially, if

$$g_{m+q}(i_0) = \ell_{m+q} \ ,$$

then (cf. (9.17) and (9.18)) for all $0 \le p \le q$

$$\ell_* \ge g_{m+q}(i_0) \ge \alpha^p g_{m+q-p}(i_0) + (1-\alpha^p)\ell_{m+q-p}$$
$$\ge \alpha^p g_{m+q-p}(i_0) + (1-\alpha^p)\ell_m \ ,$$

where the last inequality follows from corollary 13.6(i).
So for all $0 \le p \le q$

$$g_{m+q-p}(i_0) \le \alpha^{-p}(\ell_* - \ell_m) + \ell_m \le \alpha^{-q}(\ell_* - \ell_m) + \ell_m \ .$$

Hence, cf. (9.19),

$$(13.15) \qquad v_{m+q}(i_0) - v_m(i_0) = \sum_{t=0}^{q-1} g_{m+t}(i_0) \le q\alpha^{-q}(\ell_* - \ell_m) + q\ell_m \ .$$

On the other hand we have $u_n \ge g_*$ for all n. Hence, there exists a state $j_0 \in S$ which has $g_{m+k}(j_0) \ge g_*$ for at least $N^{-1}q$ of the indices $m+k \in \{m, m+1, \ldots, m+q-1\}$. So for this state j_0

$$(13.16) \qquad v_{m+q}(j_0) - v_m(j_0) \ge N^{-1} q\, g_* + (q - N^{-1}q)\ell_m = q\ell_m + N^{-1}q(g_* - \ell_m) \ .$$

Together, (13.15) and (13.16) yield

$$(13.17) \qquad sp(v_{m+q} - v_m) \ge N^{-1}q(g_* - \ell_m) - q\alpha^{-q}(\ell_* - \ell_m) \ .$$

Now we can prove

LEMMA 13.7. $\ell_* = g_*$.

PROOF. The proof is practically identical to the proof of theorem 9.11. We therefore delete it. □

THEOREM 13.8.

$$\lim_{n\to\infty} (v_{n+1} - v_n) = g_* e \ .$$

PROOF. From lemma 13.7 we have already

$$\lim_{n\to\infty} \min_{i\in S} (v_{n+1} - v_n)(i) = g_* \ .$$

Similarly to lemma 13.7 one may prove that also

$$\lim_{n\to\infty} u_n = g_* \ ,$$

which completes the proof. □

So we see that, if the functional equation $Uv = v + ge$ has a solution and if the strong aperiodicity assumption holds, then the method of standard successive approximations yields an ε-band on the value of the game and nearly-optimal stationary strategies for both players.

REFERENCES

ANTHONISSE, J. and H. TIJMS (1977), Exponential convergence of products of stochastic matrices, J. Math. Anal. Appl. 59, 360-364.

BATHER, J. (1973), Optimal decision procedures for finite Markov chains, Part II, Adv. Appl. Prob. 5, 521-540.

BELLMAN, R. (1957), A Markov decision process, J. Math. Mech. 6. 679-684.

BLACKWELL, D. (1962), Discrete dynamic programming, Ann. Math. Statist. 33 , 719-726.

BLACKWELL, D. (1965), Discounted dynamic programming, Ann. Math. Statist. 36, 226-235.

BLACKWELL, D. (1967), Positive dynamic programming, in Proceedings of the 5th Berkeley symposium on Mathematical Statistics and Probability, Vol. I, 415-418.

BLACKWELL, D. and T. FERGUSON (1968), The big match, Ann. Math. Statist. 39, 159-163.

BROWN, B. (1965), On the iterative method of dynamic programming on a finite state space discrete time Markov process, Ann. Math. Statist. 36, 1279-1285.

CHARNES, A. and R. SCHROEDER (1967), On some stochastic tactical antisubmarine games, NRLQ 14, 291-311.

DENARDO, E. (1967), Contraction mappings in the theory underlying dynamic programming, Siam Rev. 9, 165-177.

DENARDO, E. (1970), On linear programming in a Markov decision problem, Man. Science 16, 281-288.

DENARDO, E. (1973), A Markov decision problem, in Mathematical Programming, eds. T. Hu and S. Robinson, Acad. Press, New York, 33-68.

DENARDO, E. and B. FOX (1968), Multichain Markov renewal programs, Siam J. on Appl. Math. 16, 468-487.

DENARDO, E. and B. MILLER (1968), An optimality condition for discrete dynamic programming with no discounting, Ann. Math. Statist. 39, 1220-1227.

DENARDO, E. and U. ROTHBLUM (1979), Overtaking optimality for Markov decision chains, Math. Oper. Res. 4, 144-152.

DERMAN, C. (1970), Finite stage Markovian decision processes, Acad. Press, New York.

DERMAN, C. and R. STRAUCH (1966), A note on memoryless rules for controlling sequential decision processes, Ann. Math. Statist. 37, 276-278.

DUBINS, L. and L. SAVAGE (1965), How to gamble if you must, McGraw-Hill, New York.

d'EPENOUX, F. (1960), Sur un problème de production et de stockage dans l'aléatoire, Rev. Française de Rech. Oper. 14, 3-16.

FEDERGRUEN, A. (19), Successive approximation methods in undiscounted stochastic games, Oper. Res. 28, 794-809.

GHELLINCK, G. de (1960), Les problèmes de décisions sequentielle, Cah. du Centre d'Etudes de Rech. Opér. 2, 161-179.

GILLETTE, D. (1957), Stochastic games with zero stop probabilities, in Contributions to the theory of games, Vol. III, eds. M. Dresher, A. Tucker and P. Wolfe, Princeton Univ. Press, Princeton, New Jersey, 179-187.

GROENEWEGEN, L. (1978), Characterization of optimal strategies in dynamic games, Doctoral dissertation, Eindhoven Univ. of Technology.

GROENEWEGEN, L. and J. WESSELS (1976), On the relation between optimality and saddle-conservation in Markov games, Eindhoven Univ. of Technology, Dept. of Maths., Memorandum COSOR 76-14.

HAJNAL, J. (1958), Weak ergodicity in nonhomogeneous Markov chains, Proc. Cambridge Phil. Soc. 54, 233-246.

HARRISON, J. (1972), Discrete dynamic programming with unbounded rewards, Ann. Math. Statist. 43, 636-644.

HASTINGS, N. (1968), Some notes on dynamic programming and replacement, Oper. Res. Q. 19, 453-464.

HASTINGS, N. (1971), Bounds on the gain of a Markov decision process, Oper. Res. 19, 240-243.

HEE, K. van (1978a), Markov strategies in dynamic programming, Math. Oper. Res. 3, 37-41.

HEE, K. van (1978b), Bayesian control of Markov chains, Doctoral dissertation, stelling 5 (in Dutch), Eindhoven Univ. of Technology.

HEE, K. van, A. HORDIJK and J. van der WAL (1977), Successive approximations for convergent dynamic programming, in Markov decision theory, eds. H. Tijms and J. Wessels, Math. Centre Tract 93, Mathematisch Centrum, Amsterdam, 183-211.

HEE, K. van and J. van der WAL (1977), Strongly convergent dynamic programming: some results, in Dynamische Optimierung, ed. M. Schäl, Bonner Math. Schriften nr. 98, Bonn, 165-172.

HEE, K. van and J. WESSELS (1978), Markov decision processes and strongly excessive functions, Stoch. Proc. Appl. 8, 59-76.

HOFFMAN, A. and R. KARP (1966), On nonterminating stochastic games, Man. Science 12, 359-370.

HORDIJK, A. (1974), Dynamic programming and Markov potential theory, Math. Centre Tract 51, Mathematisch Centrum, Amsterdam.

HORDIJK, A. (1976), Regenerative Markov decision models, in Stochastic systems: modeling, identification and optimization, Math. Progr. Studies 6, North-Holland, Amsterdam, 49-72.

HORDIJK, A. and L. KALLENBERG (1979), Linear programming and Markov decision chains, Man. Science 25, 352-362.

HORDIJK, A. and H. TIJMS (1975), A modified form of the iterative method of dynamic programming, Ann. of Statist. 3, 203-208.

HOWARD, R. (1960), Dynamic programming and Markov processes, Wiley, New York.

KALLENBERG, L. (1980), Linear programming and finite Markovian control problems, Doctoral dissertation, Univ. of Leiden.

KOHLBERG, E. (1974), Repeated games with absorbing states, Ann. of Statist. 2, 724-738.

KUMAR, P. and T. SHIAU (1979), Randomized strategies in zero-sum discrete-time stochastic games, Univ. of Maryland, Dept. of Maths., Baltimore County.

KUSHNER, H. and S. CHAMBERLAIN (1969), Finite state stochastic games: existence theorems and computational procedures, IEEE, Trans. Aut. Control 14, 248-254.

LANÉRY, E. (1967), Etude asymptotique des systèmes Markoviens à commande, Rev. Inf. Rech. Opér. 1, 3-56.

LIPPMAN, S. (1969), Criterion equivalence in discrete dynamic programming, Oper. Res. 17. 920-923.

MACQUEEN, J. (1966), A modified dynamic programming method for Markovian decision problems, J. Math. Anal. Appl. 14, 38-43.

MAITRA, A. and T. PARTHASARATHY (1971), On stochastic games II, J. Opt. Theory Appl. 8, 154-160.

MANNE, A. (1960), Linear programming and sequential decisions, Man. Science 6, 259-267.

MERTENS, J. and A. NEYMAN (1980), Stochastic games, Univ. Catholique de Louvain, Dept. of Maths.

MILLER, B. and A. VEINOTT (1969), Discrete dynamic programming with a small interest rate, Ann. Math. Statist. 40, 366-370.

MONASH, C. (1979), Stochastic games: the minimax theorem, Harvard Univ., Cambridge, Massachusetts.

MORTON, T. (1971), Undiscounted Markov renewal programming via modified successive approximations, Oper. Res. 19, 1081-1089.

MORTON, T. and W. WECKER (1977), Ergodicity and convergence for Markov decision processes, Man. Science 23, 890-900.

NASH, J. (1951), Non-cooperative games, Ann. of Maths. 54, 286-295.

NUNEN, J. van (1976a), Contracting Markov decision processes, Math. Centre Tract 71, Mathematisch Centrum Amsterdam.

NUNEN, J. van (1976b), Improved successive approximation methods for discounted Markov decision processes, in Progress in Operations Research, ed. A. Prékopa, North-Holland, Amsterdam, 667-682.

NUNEN, J. van (1976c), A set of successive approximation methods for discounted Markovian decision problems, Zeitschr. Oper. Res. 20, 203-208.

NUNEN, J. van and S. STIDHAM (1978), Action-dependent stopping times and Markov decision processes with unbounded rewards, to appear in OR Spectrum.

NUNEN, J. van and J. WESSELS (1976), A principle for generating optimization procedures for discounted Markov decision processes, Colloquia Mathematica Societatis Bolyai Janos, Vol. 12, 683-695, North Holland, Amsterdam.

NUNEN, J. van and J. WESSELS (1977a), Markov decision processes with unbounded rewards, in Markov decision theory, eds. H. Tijms and J. Wessels, Math. Centre Tract 93, Mathematisch Centrum Amsterdam, 1-24.

NUNEN, J. van and J. WESSELS (1977b), The generation of successive approximations for Markov decision processes using stopping times, in Markov decision theory, eds. H. Tijms and J. Wessels, Math. Centre Tract 93, Mathematisch Centrum Amsterdam, 25-37.

ODONI, A. (1969), On finding the maximal gain for Markov decision processes, Oper. Res. 17, 857-860.

ORNSTEIN, D. (1969), On the existence of stationary optimal strategies, Proc. Amer. Math. Soc. 20, 563-569.

PLATZMAN, L. (1977), Improved conditions for convergence in undiscounted Markov renewal programming, Oper. Res. 25, 529-533.

POLLATSCHEK, M. and B. AVI-ITZHAK (1969), Algorithms for stochastic games with geometrical interpretation, Man. Science 15, 399-415.

PORTEUS, E. (1971), Some bounds for discounted sequential decision processes, Man. Science 18, 7-11.

PORTEUS, E. (1975), Bounds and transformations for discounted finite Markov decision chains, Oper. Res. 23, 761-784.

RAO, S., R. CHANDRASEKARAN and K. NAIR (1973), Algorithms for discounted stochastic games, J. Opt. Theory Appl. 11, 627-637.

REETZ, D. (1973), Solution of a Markovian decision problem by successive overrelaxation, Zeitschr. Oper. Res. 17, 29-32.

RIOS, S. and I. YANEZ (1966), Programmation sequentielle en concurrence, in Research papers in statistics, ed. F. David, Wiley, New York, 289-299.

ROGERS, P. (1969), Nonzero-sum stochastic games, Berkeley, Univ. of California, Oper. Res. Centre, Report ORC 69-8.

ROSS, S. (1970), Applied probability models with optimization applications, Holden Day, San Francisco.

ROTHBLUM, U. (1979), Iterated successive approximation for sequential decision processes, Yale University, School of Org. and Man., New Haven, Connecticut.

SCHÄL, M. (1975), Conditions for optimality in dynamic programming and for the limit of n-stage policies to be optimal, Zeitschr. Wahrsch. Th. verw. Gebieten 32, 179-196.

SCHWEITZER, P. (1971), Iterative solution of the functional equations of undiscounted Markov renewal programming, J. Math. Anal. Appl. 34, 495-501.

SCHWEITZER, P. and A. FEDERGRUEN (1978), The asymptotic behaviour of undiscounted value iteration in Markov decision problems, Math. Oper. Res. 2, 360-382.

SCHWEITZER, P. and A. FEDERGRUEN (1979), Geometric convergence of value-iteration in multichain Markov decision problems, Adv. Appl. Prob. 11, 188-217.

SHAPLEY, L. (1953), Stochastic games, Proc. Nat. Acad. Sci. 39, 1095-1100.

SLADKY, K. (1974), On the set of optimal controls for Markov chains with rewards, Kybernetica 10, 350-367.

SOBEL, M. (1971), Noncooperative stochastic games, Ann. Math. Statist. 42, 1930-1935.

STIDHAM, S. (1978), On the convergence of successive approximations in dynamic programming with non-zero terminal reward, Raleigh, North Carolina State University, Technical Report no. 78-9.

STRAUCH, R. (1966), Negative dynamic programming, Ann. Math. Statist. 37, 871-889.

VARGA, R. (1962), Matrix iterative analysis, Prentice-Hall, Englewood Cliffs.

VEINOTT, A. (1966), On finding optimal policies in discrete dynamic programming with no discounting, Ann. Math. Statist. 37, 1284-1294.

VEINOTT, A. (1969), Discrete dynamic programming with sensitive discount optimality criteria, Ann. Math. Statist. 40, 1635-1660.

WAL, J. van der (1976), A successive approximation algorithm for an undiscounted Markov decision process, Computing 17, 157-162.

WAL, J. van der (1977a), Discounted Markov games: successive approximations and stopping times, Int. J. Game Theory 6, 11-22.

WAL, J. van der (1977b), Discounted Markov games: generalized policy iteration method, J. Opt. Theory Appl. 25, 125-138.

WAL, J. van der (1979), Positive Markov games with stopping actions, in Game theory and related topics, eds. O. Moeschlin and D. Pallaschke, North-Holland, Amsterdam, 117-126.

WAL, J. van der (1980a), The method of value oriented successive approximations for the average reward Markov decision process, OR Spectrum 1, 233-242.

WAL, J. van der (1980b), Successive approximations for average reward Markov games, Int. J. Game Theory 9, 13-24.

WAL, J. van der and J. WESSELS (1976), On Markov games, Stat. Neerlandica 30, 51-71.

WAL, J. van der and J. WESSELS (1977), Successive approximation methods for Markov games, in Markov decision theory, eds. H. Tijms and J. J. Wessels, Math. Centre Tract 93, Mathematisch Centrum Amsterdam, 39-55.

WAL, J. van der and H. ZIJM (1979), Note on a dynamic programming recursion, Eindhoven Univ. of Technology, Dept. of Maths., Memorandum COSOR 79-12.

WALTER, W. (1976), A note on contraction, Siam Rev. 18, 107-111.

WESSELS, J. (1977a), Stopping times and Markov programming, in Transactions of the 7-th Prague conference of Information theory, Statistical decision functions and Random processes, Academia, Prague, 575-585.

WESSELS, J. (1977b), Markov programming by successive approximations with respect to weighted supremum norms, J. Math. Anal. Appl. 58, 326-335.

WESSELS, J. and J. van NUNEN (1975), Discounted semi-Markov decision processes: linear programming and policy iteration, Stat. Neerlandica 29, 1-7.

WHITE, D. (1963), Dynamic programming, Markov chains and the method of successive approximations, J. Math. Anal. Appl. 6, 373-376.

WIJNGAARD, J. (1975), Stationary Markovian decision problems, Doctoral dissertation, Eindhoven, Univ. of Technology.

ZACHRISSON, L. (1964), Markov games, in Advances in game theory, eds. M. Dresher, L. Shapley and A. Tucker, Princeton, New Jersey, 211-253.

ZIJM, H. (1978), Bounding functions for Markov decision processes in relation to the spectral radius, Oper. Res. Verfahren 33, 461-472.

SYMBOL INDEX

SUBJECT INDEX

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*, . 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 II, 1983.
ISBN 90 6196 266 8.

An asterisk before the number means "to appear"