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 71

J.A.E.E. VAN NUNEN

CONTRACTING MARKOV DECISION PROCESSES

MATHEMATISCH CENTRUM AMSTERDAM 1976

AMS(MOS) subject classification scheme (1970): 90C45

ISBN 90 6196 129 7

CONTENTS

ACKNOWLEDGEMENTS

This Mathematical Centre Tract contains a slightly adapted version of my thesis. The research leading to this thesis was carried out at the department of mathematics of the Eindhoven University of Technology. My thesis adviser was Prof.dr. J. Wessels. I wish to thank him in particular for the splendid opportunity given to me and for the many stimulating discussions. Much of the research presented in this monograph was done in close cooperation with him.

I feel very grateful to my co-promotor Prof.dr. S.T.M. Ackermans for numerous advises which improved the manuscript considerably. I also owe many thanks to Dr. F.H. Simons for several valuable comments.

It has been my privilege to work within an extremely inspiring environment. Therefore, I would like to extend my sincere thanks to Prof. Wessels and my colleagues Luuk Groenewegen, Kees van Hee and Jacob Wijngaard. Our amity and congenial interest in Markov programming was a great support to me.

I am indebted to the Mathematical Centre for the opportunity to publish my thesis in their series Mathematical Centre Tracts.

CHAPTER 1

INTRODUCTION

In the last two decades much attention has been given to Markov decision processes. Markov decision processes were first introduced by Bellman [2] in 1957, and constitute a special class of dynamic programming problems. In 1960 Howard [35] published his book "Dynamic programming and Markov processes". This publication gave an important impulse to the investigation of Markov decision processes.

We will first give an outline of the decision processes to be investigated.

Consider a system with a countable state space S. The system can be controlled at discrete points in time $t = 0,1,2,\ldots,$ by a decision maker. If at time t the system is observed to be in state $i \in S$, the decision maker can select an action from a nonempty set A. This set is independent of the state $i \in S$ and of the time instant t. If he selects the action $a \in A$ in state $i \in S$, at time t, the system's state at time $t+1$ will be $j \in S$ with probability $p^a(i,j)$ again independent of t. He then earns an immediate (expected) reward $r(i,a)$.

Usually, the problem is to choose the actions "as good as possible" with respect to a given optimality criterion.

We will use the *(expected) total reward criterion.*
So the problem is:

1) to determine a recipe (decision rule) according to which actions should be taken such that the (expected) total reward over an infinite time horizon is maximal;
2) to determine the total reward that may be expected if we act according to that decision rule.

In first instance the following solution techniques for Markov decision processes with a finite state space and a finite action space were available: the policy iteration algorithm developed by Howard [35], linear programming [11], [13], approximation by standard dynamic programming [35].

A disadvantage (especially for large scaled problems) of the former two methods is that each iteration step requires a relatively large amount of computation.
Furthermore, the convergence of the method of standard dynamic programming is very slow.
Hence, the construction of numerical methods for determining optimal solutions has been the subject of much research in this area. Moreover, much attention has been paid to the generalization of the model as described by Bellman and Howard. Both subjects will be studied in this monograph.

MacQueen [46] introduced an improved version of the standard dynamic programming algorithm by constructing, in each iteration step, improved upper and lower bounds for the optimal return vector. His approach yields a rather fast algorithm for solving finite state space finite action space Markov decision processes.
Modifications of the afore mentioned optimization procedures have been given by e.g. Hastings [25], [26] who proposed a Gauss-Seidel-like technique and Reetz [60] who based his optimization procedure on an overrelaxation idea.
Modifications have also been given by Porteus [59], Wessels [74], van Nunen and Wessels [55], and van Nunen [53], [54].

By and by several extensions of the original model were presented. The finite state space and finite action space restriction was dropped. For example Maitra [48] and Derman [14] studied Markov decision processes with a countable state space and finite action space, whereas Blackwell [5] and Denardo [12] already investigated Markov decision processes with a general state and action space.
The restriction of equidistant decision points was dropped as well; see Jewell [37], [38].
As a remaining restriction, however, a bounded reward structure is assumed in the above articles.
This restriction has been released recently, see Lippman [44], [45], Harrison [24], Wessels [75], Hinderer [31], and Hordijk [33].

In this monograph we will investigate Markov decision processes on a countably infinite or finite state space and with a general action space. Furthermore, we allow for an *unbounded reward structure*. We do *not* require the transition probabilities to be *strictly defective*, with respect to the usual supremum norm. We assume the existence of a function $b: S \to \mathbb{R}$ and a positive function $\mu: S \to \mathbb{R}^+ := \{x \in \mathbb{R} \mid x > 0\}$ such that

$$\forall_{i \in S} \forall_{a \in A(i)} \ |r(i,a) - b(i)| < \mu(i) \ ,$$

the function μ will be used to construct a weighting function, or a bounding function. Moreover, we assume

$$\exists_{0<\rho<1} \forall_{i \in S} \forall_{a \in A(i)} \sum_{j \in S} p^a(i,j)\mu(j) \le \rho\mu(i) \ .$$

In order to guarantee the existence of the total expected reward we assume the function b on S to be a charge with respect to the transition probability structure (see section 4.1).
These assumptions on the reward and transition probability structure arise in fact by a combination and a slight extension of the conditions as proposed by Wessels [75] and Harrison [24]. As will be shown in the final chapter the assumptions allow e.g. the investigation of a large class of discounted Markov and Semi-Markov decision Processes. Lippman's assumptions [45] are covered as well, see van Nunen and Wessels [56].

We will develop a set of optimization procedures for solving Markov decision problems, satisfying the described conditions, with respect to the total reward criterion. This will be done by using the concept of *stopping time* (see also Wessels [74]), which results in a unifying approach. This set of methods includes the procedures for finite state, finite action space Markov decision processes as proposed by Howard [35], Reetz [60], Hastings [25] MacQueen [46]. A main role in our approach will be played by the theory of monotone contraction mappings defined on a complete metric space of functions on S. This space will be denoted by V.

The concept of stopping time will be used to define a set of contraction mappings on V. Given a decision rule and given the starting state $i \in S$ we may define the stochastic process $\{s_t \mid t = 0,1,\dots\}$ where s_t denotes the state of the system at time t. Roughly speaking a stopping time is a recipe for terminating the stochastic process $\{s_t \mid t \ge 0\}$. For each stopping time (denoted by δ) we define the mapping U_δ of V by defining $(U_\delta v)(i)$ as the supremum over all decision rules of the expected total reward until the process is stopped according to the stopping time δ, given that the process starts in state $i \in S$, while, in addition, a terminal reward $v(j)$ is earned if the process is stopped in state $j \in S$. U_δ is proved to be a monotone contractive mapping on the complete metric space $\mathcal{V}$.

For stopping times that are *nonzero* (see section 2.2) U_δ will be strictly contracting, its fixed point being equal to the requested optimal expected total reward over an infinite time horizon (denoted by V^*). Hence, the fixed point is independent of the chosen nonzero stopping time. This implies that for each nonzero stopping time δ, V^* may be approximated successively by a sequence $v_n^\delta := U_\delta v_{n-1}^\delta$, starting from any $v_0^\delta \in \mathcal{V}$. So for each nonzero stopping time δ we have

$$v_n^\delta \rightarrow V^* \ .$$

These results may be formulated alternatively as follows: for each nonzero stopping time δ, V^* is the unique solution of the optimality equation

$$v = U_\delta v, \text{ in } \mathcal{V} \ .$$

The class of described methods may be extended.

For a special class of stopping times, which we called transition memoryless stopping times (see section 2.2), the mappings U_δ produce the opportunity of determining in each iteration step a decision rule of a special type for which the supremum by applying U_δ is attained or approximately attained (see chapter 5). Such a decision rule will be called a stationary Markov strategy (denoted by f^∞). We define the mapping L_δ^f of $\mathcal{V}$ in a similar way as we have defined U_δ, with the difference that the expected reward by applying the stationary Markov strategy f^∞ is computed instead of the supremum over all decision rules.

For transition memoryless nonzero stopping times we define for each $\lambda \in \mathbb{N} = \{1,2,\ldots\}$ a mapping $U_\delta^{(\lambda)}$ of $\mathcal{V}$.

If the supremum by computing $U_\delta v$ is attained for a stationary Markov strategy (f^∞) then

$$U_\delta^{(\lambda)} v := (L_\delta^f)^\lambda v := (L_\delta^f)(L_\delta^f)^{\lambda-1} v, \quad \text{with } \lambda \in \mathbb{N} \ ,$$

$U_\delta^{(\infty)}$ is defined by

$$U_\delta^{(\infty)} v := \lim_{n\to\infty} U_\delta^{(n)} v \ .$$

If this supremum is not attained $U_\delta^{(\lambda)}$ is defined by using a Markov strategy for which the supremum is approximated (see chapter 6). $U_\delta^{(\lambda)}$ is neither necessarily contracting nor monotone. However, the monotone contraction property of the mappings U_δ and L_δ^f enables the use of $U_\delta^{(\lambda)}$ as a base for successive approximation methods.
For each transition memoryless nonzero stopping time δ and each $\lambda \in \mathbb{N} \cup \{\infty\}$ a sequence $v_n^{\delta\lambda}$ defined by

$$v_0^{\delta\lambda} \in \mathcal{V}; \qquad v_n^{\delta\lambda} := U_\delta^{(\lambda)} v_{n-1}^{\delta\lambda} := (L_\delta^{f_n})^{\lambda} v_{n-1}^{\delta\lambda}$$

converges to V^*. Here f_n^∞ is chosen such that $L_\delta^{f_n} v_{n-1}^{\delta\lambda}$ approximates $U_\delta v_{n-1}^{\delta\lambda}$ sufficiently well. So each pair (δ,λ) yields a successive approximation of V^*. Moreover, the stationary Markov strategy found in the n-th iteration of such a procedure becomes (ε-)optimal for n sufficiently large.
The vectors $v_{n-1}^{\delta\lambda}$ and $v_n^{\delta\lambda}$ enable us to construct upper and lower bounds for the optimal return V^*. In addition, the availability of upper and lower bounds allows an incorporation of a suboptimality test. The use of upper and lower bounds and a suboptimality test may yield a considerable gain in computation time, see section 7.3.

We conclude this introduction with a short overview of the contents of the subsequent chapters.

In the first three sections of chapter 2 some basic notions required in the sequel are presented. After the introduction of some notations (section 2.1) we discuss in sections 2.2 and 2.3 the concepts of stopping time and weighted supremum norms respectively. The final section of chapter 2 is devoted to some properties of weighted supremum norms.

In chapter 3 we treat Markov reward processes. (stochastic processes without the possibility of making decisions). In section 3.1 the Markov reward model is defined. Reward functions may be unbounded under our assumptions. In section 3.2 the concept of stopping time is used to define the contraction mappings on the complete metric space $\mathcal{V}$ (introduced in section 2.3). A discussion of the assumptions is the topic of the final section 3.3.

The study of Markov decision processes starts in chapter 4. After a description of the model (section 4.1), section 4.2 contains the introduction of decision rules and assumptions. These assumptions will be a natural extension of those in chapter 3. Under our assumptions some results about Markov decision processes will be proved (section 4.3). The final section (4.4) is again devoted to a discussion of the assumptions.

In chapter 5 the concept of stopping time is used to generate a whole set of optimization procedures based on the mappings U_δ. For each decision rule π, not necessarily a stationary Markov strategy, and each stopping time δ a contractive mapping L_δ^π of V will be defined and investigated (section 5.1). Next, (section 5.2) the operator U_δ will be studied. Finally, we will present necessary and sufficient conditions for the stopping times under which we can restrict the attention to stationary Markov strategies only (transition memoryless stopping times).

In chapter 6 we investigate value oriented successive approximations based on the mappings $U_\delta^{(\lambda)}$. The term "value oriented" is used since in each iteration step extra effort is given to obtain better estimates for the total expected reward corresponding to the stationary Markov strategy f_n^∞.

Chapter 7 will be used to construct upper and lower bounds for the optimal reward V^*. In this chapter also a suboptimality test will be introduced. In the third section of this chapter we show how our theory may be used in the special case of a Markov decision process with a finite state space and a finite action space. We indicate the relation with the existing optimization procedures.

In the brief chapter 8 we weaken the assumptions as imposed in chapter 4. This weakened version corresponds to the N-stage contraction assumption introduced by Denardo [12]. It will be proved that N-stage contraction with respect to a given bounding function implies the existence of a new bounding function satisfying the assumptions of chapter 4.

We conclude this monograph with a chapter in which we show that a number of specific Markov decision processes is covered by our theory. We will also show how a number of the existing approximation methods for certain systems of linear equations are included in our treatment of Markov reward processes (chapter 3). The final part of this chapter consists of an example. In this example we treat an inventory problem.

CHAPTER 2

PRELIMINARIES

The goal of this chapter will be the introduction of some of the notions which play an important role throughout this monograph.

First (section 2.1) we will give some notations and we will introduce the measurable spaces relevant for the stochastic processes that will be investigated.

Next (section 2.2) stopping times are introduced. We will allow for randomized stopping times. Several specific stopping times will be described.

In section 2.3 a bounding function μ is introduced. The function μ will be used to define a weighted supremum norm. In the following chapters this bounding function will appear to be one of the tools for handling Markov decision processes with an unbounded reward structure and with a transition probability structure that needs not to be contractive with respect to the usual supremum norm.

Using the bounding function μ a Banach space $\mathcal{W}$ and a complete metric space $\mathcal{V}$ are defined.

Finally (section 2.4), we discuss some properties of bounding functions.

2.1. *Notations*

As mentioned in the introductory chapter we study a system which is observed to be in one of the states from a *state space* S at times $t = 0,1,\ldots$ We assume S to be countably infinite or finite, and represent the states by the integers, starting with zero. So if the state space is finite, S is represented by $\{0,1,\ldots,N\}$, where $N+1$ is the number of states. If S is countably infinite it is represented by $\{0,1,2,\ldots\}$. A *path* is a sequence of states that are subsequently visited.

REMARK 2.1.1. The state 0 is included in the state space in order to be able to deal with processes with defective transition probabilities on $\{1,2,\ldots\}$ or $\{1,2,\ldots,N\}$. This is done in the usual way by defining for $i \geq 1$, $p(i,0) := 1 - \sum_{j\geq 1} p(i,j)$ and $p(0,0) := 1$. It follows that 0 is an absorbing state. Therefore, without loss of generality in the sequel we may and shall assume that in the state space S the state 0 is absorbing and the transition probabilities satisfy $\sum_{j\in S} p(i,j) = 1$ for all $i \in S$.

NOTATIONS 2.1.1.

(i) $S^k := S \times S \times \ldots \times S$, the k-fold Cartesian product of S, so $S^1 = S$. $S^\infty := S \times S \times \ldots$; S^∞ is the set of all paths.

(ii) Let $\alpha \in S^k$, with $k \geq n$, $k,n \in \mathbb{N}$ then $\alpha^{(n)}$ denotes the row vector of the first n components of α.

(iii) k_α is the number of components of α. So $k_\alpha = n$ if and only if $\alpha \in S^n$.

(iv) The i-th component of $\alpha \in S^n$, $n \geq i$ is denoted by $[\alpha]_{i-1}$.

(v) Hence $\alpha \in S^n$ may be written as $\alpha = ([\alpha]_0,[\alpha]_1,\ldots,[\alpha]_{k_\alpha-1})$.

(vi) $\gamma := (\alpha,\beta) := ([\alpha]_0,[\alpha]_1,\ldots,[\alpha]_{k_\alpha-1},[\beta]_0,\ldots,[\beta]_{k_\beta-1})$, $k_\gamma = k_\alpha + k_\beta$, where $\alpha,\beta \in G_\infty$ with $G_\infty := \bigcup_{k=1}^{\infty} S^k$.

(vii) The term (column) *vector* is used hereafter for a real valued function on S.

(viii) The term *matrix* is used hereafter for a real valued function on S^2.

(ix) The (i,j)-th element of a matrix P will be denoted by p(i,j).

(x) P^0 is the identity matrix (with diagonal entries equal to one and other entries equal to zero).

(xi) Matrix multiplication and matrix-vector multiplication are defined as usual (in all cases there will be absolute convergence).

(xii) P^n is the n-fold matrix product $P \times P \times \ldots \times P$, the (i,j)-th entry of P^n is denoted by $p^{(n)}(i,j)$.

(xiii) Let v,w be vectors, then $v \leq w$ if and only if $v(i) \leq w(i)$ for all $i \in S$; $v < w$ if and only if $v \leq w$ and for at least one $i \in S$ $v(i) < w(i)$.

Let $\mathcal{S}_0$ be the σ-field of all subsets of S, then the measurable space $(\Omega_0,\mathcal{F}_0)$ is defined to be the product space, with $\Omega_0 = S^\infty$ and $\mathcal{F}_0$ is the σ-algebra on Ω_0 generated by the finite products of the σ-field $\mathcal{S}_0$.
In order to be able to use the concept of stopping time in an adequate way we extend the measurable space $(\Omega_0,\mathcal{F}_0)$ to the measurable space $(\Omega,\mathcal{F})$. The space $(\Omega,\mathcal{F})$ will play a main role in the sequel. Let the set $E := \{0,1\}$ and let $\mathcal{S}$ be the σ-field of all subsets of $S \times E$ then the measurable space $(\Omega,\mathcal{F})$ is defined to be the product space with $\Omega := (S \times E)^\infty$ and $\mathcal{F}$ is the σ-algebra on Ω generated by the finite products of the σ-field $\mathcal{S}$.
So Ω_0 contains all sequences of the form

$$\omega_0 := (i_0,i_1,i_2,\ldots), \quad i_t \in S ,$$

whereas Ω contains all sequences of the form

$$\omega := ((i_0,d_0),(i_1,d_1),\ldots.), \quad i_t \in S,\ d_t \in E .$$

2.2. *Stopping times*

We now are ready to introduce our notion of stopping time. We will not use the term stopping time in the standard way. However, as follows in the sequel of this monograph, there is a direct relation between our definition and the usual one.

DEFINITION 2.2.1. A (randomized) *stopping time* is a function $\delta\colon G_\infty \to [0,1]$ satisfying

$$(2.2.1)\qquad \delta(0) = 1;\ \forall_{\alpha\in G_\infty}\ [\forall_{k\leq k_\alpha}\ [\delta(\alpha^{(k)}) \neq 0] \Rightarrow [\delta((\alpha,0)) = 1]] .$$

DEFINITION 2.2.2. The set of all (randomized) stopping times is Δ.

REMARK 2.2.1.

(i) Roughly speaking for each $\alpha \in S^k$, we will use $1-\delta(\alpha)$ as the probability that a stochastic process on S is indicated to stopp at time $k-1$ in state $[\alpha]_{k-1}$ given that the states $[\alpha]_0,[\alpha]_1,\ldots,[\alpha]_{k-1}$ have been visited successively.

(ii) From now on we will use the less formal notation $\delta(\alpha,\beta)$, $\delta(i)$ instead of $\delta((\alpha,\beta))$ and $\delta((i))$ respectively.

DEFINITION 2.2.3.

(i) $\delta \in \Delta$ is said to be a *nonrandomized* stopping time if and only if

$$\forall_{\alpha\in G_\infty}\ \delta(\alpha) \in \{0,1\} .$$

(ii) $\delta \in \Delta$ is said to be a *memoryless* stopping time if and only if

$$\forall_{\alpha\in G_\infty}\ [\forall_{k\leq k_\alpha-1}\ [\delta(\alpha^{(k)}) \neq 0] \Rightarrow [\delta(\alpha) = \delta([\alpha]_{k_\alpha-1})]] .$$

(iii) $\delta \in \Delta$ is said to be an *entry time* if and only if δ is nonrandomized and memoryless.

DEFINITION 2.2.4. A nonempty subset $G \subset G_\infty$ is said to be a *goahead set* if and only if

(i) $\forall_{\alpha,\beta\in G_\infty} [(\alpha,\beta) \in G \Rightarrow \alpha \in G]$

(ii) $(0) \in G$

(iii) $\forall_{\alpha\in G} [(\alpha,0) \in G]$.

NOTATIONS 2.2.1.

(i) G_n is the goahead set of those sequences of G_∞ for which the components $[\alpha]_i$ are zero for $i \geq n$, if there are any. So

$$G_n := (\bigcup_{k=1}^{n} S^k) \cup \{(\alpha,\beta) \mid \beta \in \bigcup_{k=1}^{\infty} \{0\}^k , \quad \alpha \in S^n\} .$$

(ii) For a goahead set G we define $G(i)$ by

$$G(i) := \{\alpha \in G \mid [\alpha]_0 = i\}, \quad i \in S .$$

LEMMA 2.2.1. The characteristic function of a goahead set is a nonrandomized stopping time.

PROOF. Straightforward. □

DEFINITION 2.2.5. $\delta \in \Delta$ is said to be a *nonzero* stopping time if and only if

$$\exists_{\varepsilon>0} \forall_{i\in S} \delta(i) > \varepsilon .$$

REMARK 2.2.2.

(i) A nonrandomized stopping time is nonzero if and only if

$$\forall_{i \in S}\ \delta(i) = 1 .$$

(ii) A nonzero stopping time $\delta \in \Delta$, which is an entry time, has the following property

$$\forall_{\alpha \in G_\infty}\ \delta(\alpha) = 1 .$$

DEFINITION 2.2.6. A goahead set is said to be nonzero if and only if $S \subset G$.

DEFINITION 2.2.7. Let $\delta_0, \delta_1 \in \Delta$ then $\delta_0 \leq \delta_1$ if and only if

$$\forall_{\alpha \in G_\infty}\ \delta_0(\alpha) \leq \delta_1(\alpha) .$$

LEMMA 2.2.2. Let Q be an index set and suppose for each $q \in Q$, G_q is a goahead set. Let δ_q be the with G_q corresponding stopping time (δ_q is the characteristic function of G_q).
Let δ^-, δ^+ be defined by

$$\delta^-(\alpha) := \inf_{q \in Q} \delta_q(\alpha),\ \delta^+(\alpha) := \sup_{q \in Q} \delta_q(\alpha)$$

respectively, then δ^-, δ^+ are elements of Δ.
δ^- and δ^+ corresponds to the goahead sets $\bigcap_{q \in Q} G_q$, $\bigcup_{q \in Q} G_q$ respectively.

PROOF. The proof follows by inspection. □

DEFINITION 2.2.8. The nonrandomized *stopping function* $\tau\colon \Omega \to \mathbb{Z}^+ \cup \{\infty\}$ is defined by

(i) $\tau(\omega) = n \Leftrightarrow (d_0 = d_1 = \ldots = d_{n-1} = 0 \wedge d_n = 1)$

(ii) $\tau(\omega) = \infty \Leftrightarrow (d_t = 0 \text{ for all } t \in \mathbb{Z}^+)$.

DEFINITION 2.2.9. A stopping time δ is said to be *transition memoryless* if and only if

$$\forall_{\alpha\in G_\infty}[(k_\alpha > 1) \wedge (\forall_{k<k_\alpha}[\delta(\alpha^{(k)}) \neq 0])] \Rightarrow [\delta(\alpha) = \delta([\alpha]_{k_\alpha-2},[\alpha]_{k_\alpha-1})]$$

DEFINITION 2.2.10. A goahead set is said to be transition memoryless if and only if the corresponding nonrandomized stopping time is transition memoryless.

LEMMA 2.2.3. Memoryless stopping times are transition memoryless.

We will now give some simple examples of nonzero stopping times. The examples 2.2.1-2.2.4 are nonrandomized stopping times, which can be expressed in terms of goahead sets. Example 2.2.5 is a simple illustration of a randomized stopping time. The examples 2.2.2-2.2.5 give transition memoryless stopping times, see also Wessels [74], and van Nunen and Wessels [55].

EXAMPLE 2.2.1. $G := G_n$ or in terms of stopping times $\forall_{\alpha\in G_n}\ \delta(\alpha) = 1$ else $\delta(\alpha) = 0$.

EXAMPLE 2.2.2. The goahead set G_H is defined by

$$G_H(0) := \bigcup_{k=1}^{\infty} \{0\}^k,\ G_H(i) := \{(i,\alpha) \mid \alpha \in \bigcup_{j=0}^{i-1} G_H(j)\} \cup \{i\},\ \text{for}\ i \neq 0.$$

EXAMPLE 2.2.3.

$$G := S \cup (\bigcup_{k=2}^{\infty} B^k) \cup \{(\alpha,\beta) \mid \alpha \in S \cup (\bigcup_{k=2}^{\infty} B^k),\ \beta \in \bigcup_{k=1}^{\infty} \{0\}^k\}$$

with $B \subset S$.

EXAMPLE 2.2.4. The goahead set G_R is defined by

$$G_R(0) := \bigcup_{k=1}^{\infty} \{0\}^k\ ;$$

$$G_R(i) := \{\alpha \mid \alpha \in \bigcup_{k=1}^{\infty} \{i\}^k\} \cup \{(\alpha,\beta) \mid \alpha \in \bigcup_{k=1}^{\infty} \{i\}^k,\ \beta \in G_R(0)\}, \quad i \in S\backslash\{0\}\ .$$

EXAMPLE 2.2.5. δ is given by $\forall_{i\in S\backslash\{0\}}\ \delta(i) = \frac{1}{2}$ else $\delta(\alpha) = 1$.

2.3. *Weighted supremum norms*

DEFINITION 2.3.1. A real valued function μ on S is said to be a *bounding function* if and only if

(i) $\mu(i) > 0$, for all $i \in S\backslash\{0\}$.

(ii) $\mu(0) = 0$.

DEFINITION 2.3.2. Let μ be a bounding function, then W_μ is the set of vectors such that

(2.3.1) $\exists_{M\in\mathbb{R}^+}\ \forall_{i\in S}\ |w(i)| \le M\cdot\mu(i)$.

REMARK 2.3.1. Note that $w(0) = 0$ for each $w \in W_\mu$.

DEFINITION 2.3.3. Let μ be a bounding function. Then, for each $w \in W_\mu$, the μ-norm of w is defined by

$$\|w\|_\mu := \sup_{i\in S\backslash\{0\}} \frac{|w(i)|}{\mu(i)}\ .$$

LEMMA 2.3.1. The space W_μ with this μ-norm (weighted supremum norm) is a Banach space.

PROOF. The proof is straightforward. □

DEFINITION 2.3.4. Let the matrix A be a bounded linear operator in W_μ. The norm of A is defined by

$$\|A\|_\mu := \sup_{\|w\|_\mu=1} \|Aw\|_\mu\ .$$

REMARK 2.3.2.

(i) It is easily verified that if $a(0,1) = a(0,2) = \ldots = 0$

$$\| A \|_\mu = \sup_{i \in S\setminus\{0\}} \mu^{-1}(i) \sum_{j \in S} |a(i,j)| \mu(j) \;,$$

and if this supremum is finite then A is a bounded linear operator.

(ii) The concept of bounding function is studied in more detail in sections 2.4, 3.3, 4.4, and 8.1.

(iii) We refer to Wessels [75], who introduced the concept of weighted supremum norms in this context and to Hinderer [31], who used Wessels' idea of weighted supremum norms for defining bounding functions.

DEFINITION 2.3.5. Let b be a vector with $b(0) = 0$, and let $\rho \in [0,1)$. The set of vectors $V_{\mu,b,\rho}$ is defined by

$$V_{\mu,b,\rho} := \{v \mid (v - (1-\rho)^{-1}b) \in W_\mu\} \;.$$

REMARK 2.3.3. Note that also $v(0) = 0$ for $v \in V$ and $v_1 - v_2 \in W_\mu$ for $v_1, v_2 \in V$.

DEFINITION 2.3.6. The *metric* d_μ on $V_{\mu,b,\rho}$ is defined by

$$d_\mu(v_1,v_2) := \| v_1 - v_2 \|_\mu \quad \text{for any } v_1,v_2 \in V_{\mu,b,\rho} \;.$$

LEMMA 2.3.2. A set $V_{\mu,b,\rho}$ with the metric d_μ is a complete metric space.

Unless explicitely mentioned we fix μ, b, and ρ for the remaining part of this monograph. Referring to these fixed μ, b and ρ we will omit the subscripts μ, b, ρ.

2.4. *Some remarks on bounding functions*

In this section we give some properties of a bounding function μ' with respect to the corresponding spaces $W_{\mu'}$ and $V_{\mu'}$.

LEMMA 2.4.1. Suppose S contains a finite number of elements. Let μ_o be a bounding function and $w_n \in \mathcal{W}_{\mu_o}$ $(n \geq 0)$ then

$$[\| w_n - w \|_{\mu_o} \to 0] \Rightarrow [\| w_n - w \|_{\mu'} \to 0, \text{ for all bounding function } \mu'] .$$

PROOF. For a proof we refer to books on numerical mathematics, see e.g. Collatz [8], Krasnosel'skii [42]. □

LEMMA 2.4.2. Suppose S contains a finite number of elements and μ_o is a bounding function, then

(i) $$[\| B \|_{\mu_o} < 1] \Rightarrow \forall_{\mu'} \exists_{n \in \mathbb{N}} [\| B^n \|_{\mu'} < 1] ,$$

(ii) $$\exists_{n \in \mathbb{N}} [\| B^n \|_{\mu_o} < 1] \Rightarrow \exists_{\mu'} [\| B \|_{\mu'} < 1] ,$$

where B is a matrix.

PROOF. The proof of (i) follows directly from the fact that

$$[\| B \|_{\mu_o} < 1] \Rightarrow \lim_{n \to \infty} B^n = 0 ,$$

where 0 is the matrix with all entries zero. The proof of (ii) can be found in e.g. Krasnosel'skii [42]. □

LEMMA 2.4.3. Suppose μ_o is a bounding function and B is a nonnegative matrix with finite μ_o-norm, then

$$\exists_{n \in \mathbb{N}} [\| B^n \|_{\mu_o} < 1] \Rightarrow \exists_{\mu'} [\| B \|_{\mu'} < 1] .$$

PROOF. For a proof we refer to van Hee and Wessels [29], who proved this theorem for (sub-) Markov matrices, but their proof can easily be extended to this lemma. □

REMARK 2.4.1.

(i) Note that in lemma 2.4.3 S is not required to be finite.

(ii) If S is countably infinite the linear space $\mathcal{W}$ may contain elements that are not bounded. If $\mu(i) \to \infty$ for $i \to \infty$ then it is also permitted for $|w(i)| \to \infty$.

(iii) Note that it is not requested that the μ-norm of b exists. If the μ-norm of b exists then clearly $\mathcal{W} = \mathcal{V}$.

CHAPTER 3

MARKOV REWARD PROCESSES

In this chapter we restrict ourselves to Markov reward processes. So we exhibit our method for the first time in a relatively simple situation.
After defining the model (section 3.1) we introduce the assumptions on the reward structure and the transition probability structure. As mentioned we require neither the reward structure to be bounded nor the transition probabilities to be strictly defective.
Next (section 3.2) we show that each nonzero stopping time $\delta \in \Delta$ defines a contraction mapping (L_δ) on the complete metric space V. The fixed point appears to be independent of the stopping time. It equals the total expected reward over an infinite time horizon. In the final section (3.3) we discuss the assumptions on the reward structure and the transition probabilities in relation to the bounding function μ and the function b.

3.1. *The Markov reward model*

We consider a system that is observed to be in one of the states of S at discrete points in time $t = 0,1,\dots$. If the system's state at time t is $i \in S$, the system's state at time $t+1$ will be $j \in S$ with probability $p(i,j)$, independent of the time instant t.

ASSUMPTION 3.1.1.

(i) $\forall_{i,j\in S}\ 0 \le p(i,j) \le 1$

(ii) $\forall_{i\in S}\ \sum_{j\in S} p(i,j) = 1$

(iii) $p(0,0) = 1$.

For each $i \in S$ the unique probability measure $\mathbb{P}_i$ on (Ω_0,F_0) is defined in the standard way, see e.g. Neveu [52], Bauer [1] by defining the probabilities of cylindrical sets.

$$(3.1.1)\qquad \mathbb{P}_i(\{\omega_0 \mid [\omega_0]_0 = \ell_0,[\omega_0]_1 = \ell_1,\dots,[\omega_0]_n = \ell_n\}) := \delta_{i,\ell_0}\prod_{k=0}^{n-1} p(\ell_k,\ell_{k+1}),$$

where $n \in \mathbb{Z}^+$ and $\delta_{i,j}$ is the Kronecker symbol

$$\delta_{i,j} := \begin{cases} 1 & \text{if } i = j \\ 0 & \text{else .} \end{cases}$$

Given the starting state $i \in S$ and the matrix P (with entries $p(i,j)$) we consider the stochastic process $\{s_{0,n} \mid n \geq 0\}$, where $s_{0,n}(\omega_0) = [\omega_0]_n$. So $s_{0,n}$ is the state of the process at time n. The stochastic process $\{s_{0,n} \mid n \geq 0\}$ is a Markov chain with stationary transition probabilities. See e.g. Ross [62], Feller [17], Karlin [39], Cox and Miller [9], Kemeny and Snell [40].

For each stopping time $\delta \in \Delta$ and each starting state $i \in S$ we define in a similar way the unique probability measure $\mathbb{P}_{i,\delta}$ on (Ω,F) by giving for $n \in \mathbb{Z}^+$

$$(3.1.2) \qquad \mathbb{P}_{i,\delta}(\{\omega \mid [\omega]_0 = (\ell_0,c_0), [\omega]_1 = (\ell_1,c_1),\dots,[\omega]_n = (\ell_n,c_n)\}) :=$$

$$\delta_{i,\ell_0}\left(\prod_{k=0}^{n} [\delta(\ell_0,..,\ell_k)]^{1-c_k}[1-\delta(\ell_0,\ell_1,..\ell_k)]^{c_k}\right)\left(\prod_{k=0}^{n-1} p(\ell_k,\ell_{k+1})\right),$$

with $\ell_k \in S$ and $c_k \in E$.

This defines for each $i \in S$ and $\delta \in \Delta$ a stochastic process $\{(s_n,e_n) \mid n \geq 0\}$, where $s_n(\omega) := i_n$, $e_n(\omega) := d_n$.

So s_n, e_n are the state and the value of e_n at time n.

REMARK 3.1.1. The stochastic process $\{(s_n,e_n) \mid n \geq 0\}$ is not a Markov chain since the value $\delta(\alpha)$ may depend on the complete history $([\alpha]_0,[\alpha]_1,\dots,[\alpha]_{k_\alpha -1})$ for each $\alpha \in G_\infty$.

Formula (3.1.2) shows the connection between $\mathbb{P}_i$ and $\mathbb{P}_{i,\delta}$.

For each $\omega \in \Omega$, $\omega = ((i_0,d_0),(i_1,d_1)..)$ we define ω_0 by $\omega_0 := (i_0,i_1,i_2,..)$.

For $B_0 \in F_0$ we define the set $B \in F$ by

$$B = \{\omega \in \Omega \mid \omega_0 \in B_0\} .$$

It is easily verified that for $B_0 \in F_0$ and each $\delta \in \Delta$ we have for $i \in S$

$$(3.1.3) \qquad \mathbb{P}_i(B_0) = \mathbb{P}_{i,\delta}(B) .$$

Let f_0 be a measurable function on $(\Omega_0,\mathcal{F}_0)$. The function f on $(\Omega,\mathcal{F})$ is then defined such that

$$f(\omega) := f_0(\omega_0) .$$

It follows from formula (3.1.3) that

$$\mathbb{E}_i f_0 = \mathbb{E}_{i,\delta} f ,$$

where $\mathbb{E}_i f_0$, $\mathbb{E}_{i,\delta} f$ denote the expectation of f_0 and f with respect to the probability measures $\mathbb{P}_i$ and $\mathbb{P}_{i,\delta}$ respectively.
In the sequel we omit the subscript 0 in f_0. The process $\{s_{0,n} \mid n \geq 0\}$ will thus be denoted by $\{s_n \mid n \geq 0\}$.

NOTATION 3.1.1. By $\mathbb{E}f$, $\mathbb{E}_\delta f$ we denote the vector with components $\mathbb{E}_i f$, $\mathbb{E}_{i,\delta} f$ respectively.

We now state the assumptions on the reward structure and the transition probabilities of the system considered in this section. Therefore, we first introduce the reward function. At each point in time a reward is earned. We assume this reward to depend on the actual state of the system only. So the reward function r is a vector.

ASSUMPTION 3.1.2.

$$(r - b) \in \mathcal{W} .$$

ASSUMPTION 3.1.3.

$$\sum_{n=0}^{\infty} P^n |b| < \infty \quad .$$

ASSUMPTION 3.1.4.

$$\|P\| < 1 .$$

ASSUMPTION 3.1.5.

$$(Pb - \rho b) \in \mathcal{W} ..$$

REMARK 3.1.2.

(i) Since $b(0) = 0$ and $\mu(0) = 0$, assumption 3.1.2 implies that also $r(0) = 0$.

(ii) If $b \notin W$, then $r \notin W$.

(iii) In terms of potential theory (see e.g. Hordijk [33]) the second assumption states that b is a charge with respect to P, which implies the existence of the total expected reward over an infinite time horizon.

(iv) Assumption 3.1.4 means that the transition probabilities are such that the expectation of $\mu(\underline{s}_1)$, with respect to $\mathbb{P}_i$ is at most $\|\mathbb{P}\|\cdot\mu(i)$. This implies that the process has a tendency to decrease its μ-value.

(v) The final assumption states that, given the starting state $i_o \in S$, the difference between the expected one-stage reward and $\rho r(i_o)$ lies between $-M\mu(i_o)$ and $M\mu(i_o)$ for some $M \in \mathbb{R}^+$ and all $i_o \in S$.

(vi) Note that if $b \in W$ the assumptions 3.1.2-3.1.5 may be replaced by

(a) $r \in W$,

(b) $\|P\| < 1$.

LEMMA 3.1.1.

$$(Pr - \rho b) \in W \ .$$

PROOF. $\|Pr - \rho b\| \le \|Pb - \rho b\| + 2\|r - b\| =: M_1$, which is finite according to the assumptions 3.1.2-3.1.5. □

LEMMA 3.1.2. For M_1 as defined in the proof of lemma 3.1.1, we have

$$\|P^n r - \rho^n b\| \le M_1 \cdot n \cdot \rho_o^{n-1}, \qquad n = 1,2,\dots ,$$

with $\rho_o := \max\{\|P\|, \rho\}$.

PROOF. The proof proceeds by induction. The statement is true for n = 1. Suppose it is true for arbitrary $n \ge 1$. Using the assumptions 3.1.2-3.1.5 we then have

$$\begin{aligned} \|P^{n+1}r - \rho^{n+1}b\| &\le \|P(P^n r - \rho^n b)\| + \|\rho^n Pb - \rho^{n+1}b\| \\ &\le \|P\|(M_1 n\rho_o^{n-1}) + \rho^n\|Pb - \rho b\| \\ &\le M_1 n\rho_o^n + M_1\rho_o^n = M_1(n+1)\rho_o^n \ . \end{aligned}$$ □

LEMMA 3.1.3.

$$\forall_{i \in S} \lim_{n\to\infty} (P^n b)(i) = \lim_{n\to\infty} (P^n r)(i) = 0 .$$

PROOF. The proof is a direct consequence of assumption 3.1.3 and the foregoing lemma. □

For each $n \geq 1$ we define the vector V_n by

$$(3.1.4) \qquad V_n := \mathbb{E} \sum_{k=0}^{n-1} r(s_k) .$$

LEMMA 3.1.4.

$$V_n = \sum_{k=0}^{n-1} P^k r .$$

PROOF. The proof follows by inspection. □

Clearly $V_n(i)$ represents the total expected reward over n time periods when the initial state is $i \in S$.

THEOREM 3.1.1.

$$\lim_{n\to\infty} V_n \in V .$$

PROOF. The convergence of $\sum_{n=0}^{\infty} P^n r$ follows from assumption 3.1.3 and 3.1.2, since,

$$\sum_{n=0}^{\infty} P^n r = \sum_{n=0}^{\infty} P^n b + \sum_{n=0}^{\infty} P^n (r - b) .$$

We now have by lemma 3.1.2,

$$\| \sum_{n=0}^{\infty} P^n r - (1-\rho)^{-1} b \| = \| \sum_{n=0}^{\infty} (P^n r - \rho^n b) \| \leq \sum_{n=0}^{\infty} \| P^n r - \rho^n b \| \leq \sum_{n=0}^{\infty} M_1 n \rho_{\circ}^{n-1} = M_1 (1-\rho_{\circ})^{-2} .$$

□

DEFINITION 3.1.1. The total expected reward vector V is defined by

$$V := \sum_{n=0}^{\infty} P^n r .$$

3.2. *Contraction mappings and stopping times*

LEMMA 3.2.1. Let $v \in \mathcal{V}$ then

(i) $P^n|v|$ exists for all $n \in \mathbb{Z}^+$,

(ii) $\lim\limits_{n\to\infty} (P^n|v|)(i) = 0,\ i \in S.$

PROOF. $v \in \mathcal{V}$ implies that v can be written as $v = (1-\rho)^{-1}b + w$ where $w \in \mathcal{W}$. So $P^n|v| \le (1-\rho)^{-1}P^n|b| + P^n|w|$ which is defined. Moreover, since $\sum_{n=0}^{\infty} P^n|b|$ and $\sum_{n=0}^{\infty} P^n|w|$ exist we find part (ii) of the lemma. □

DEFINITION 3.2.1. The mapping L_1 of $\mathcal{V}$ is defined by

$$L_1 v := r + Pv, \qquad v \in \mathcal{V} .$$

LEMMA 3.2.2.

(i) L_1 maps $\mathcal{V}$ into $\mathcal{V}$.

(ii) L_1 is a monotone mapping.

(iii) The set $\{v \in \mathcal{V} \mid \|v-(1-\rho)^{-1}b\| \le M_1(1-\rho_o)^{-2}\}$ is mapped into itself by L_1.

(iv) L_1 is strictly contracting with contraction radius $\|P\|$.

(v) The unique fixed point of L_1 is V.

PROOF.

(i) $L_1 v = r + Pv = r + P((1-\rho)^{-1}b + w)$, with $w \in \mathcal{W}$. So

$$\begin{aligned}
\|L_1 v - (1-\rho)^{-1}b\| &= \|r + (1-\rho)^{-1}Pb + Pw - (1-\rho)^{-1}b\| \\
&\le \|b + (1-\rho)^{-1}Pb - (1-\rho)^{-1}b\| + \|Pw\| + \|r-b\| \\
&\le \|(1-\rho)^{-1}(Pb-\rho b)\| + \|P\|\cdot\|w\| + \|r-b\| \\
&= (1-\rho)^{-1}\|Pb-\rho b\| + \|P\|\cdot\|w\| + \|r-b\| < \infty .
\end{aligned}$$

The proof of part (ii) is trivial.
(iii) Let $\|v - (1-\rho)^{-1}b\| \le M_1(1-\rho_o)^{-2}$ then

$$\begin{aligned}\|L_1v - (1-\rho)^{-1}b\| &= \|r + Pv - (1-\rho)^{-1}b\| \le \\ &\le \|P(v-(1-\rho)^{-1}b)\| + \|(1-\rho)^{-1}Pb - (1-\rho)^{-1}b + r\| \\ &\le M_1\rho_o(1-\rho_o)^{-2} + (1-\rho_o)^{-1}\|Pb - \rho b\| + \|r-b\| \\ &\le M_1\rho_o(1-\rho_o)^{-2} + (1-\rho_o)^{-1}[\|Pb-\rho b\| + 2\|r-b\|] \\ &\le M_1(1-\rho_o)^{-2} .\end{aligned}$$

(iv) Let $v_1, v_2 \in V$ then v_1, v_2 can be given by $v_1 = (1-\rho)^{-1}b + w_1$ and $v_2 = (1-\rho)^{-1}b + w_2$ respectively where $w_1, w_2 \in W$. So $v_1 - v_2 = w_1 - w_2$, thus

$$\|L_1v_1 - L_1v_2\| = \|P(v_1 - v_2)\| = \|P(w_1 - w_2)\| \le \|P\|\cdot\|w_1 - w_2\| .$$

By choosing v_1 and v_2 such that $w_1 = \mu$ and $w_2 = 0$ equality is obtained.
The last part of the lemma follows directly from

$$L_1V = r + P(\sum_{n=0}^{\infty} P^n r) = r + \sum_{n=1}^{\infty} P^n r = \sum_{n=0}^{\infty} P^n r = V$$

where the interchange of summations is justified since $\sum_{n=0}^{\infty} P^n|r| < \infty$.□

Now we return to the concept of stopping time. Note that the stopping function τ on Ω is a random variable. So we can define the random variable s_τ by

$$(3.2.1) \qquad s_\tau := \begin{cases} s_n & \text{if } \tau = n , \\ 0 & \text{if } \tau = \infty . \end{cases}$$

Given the starting state $i \in S$ and the transition probabilities the distribution of τ is uniquely determined by the choice of $\delta \in \Delta$.

LEMMA 3.2.3. Let $\delta \in \Delta$ be a nonzero stopping time, then

$$\exists_{\gamma>0} \forall_{i\in S} \mathbb{E}_{i,\delta}\tau \geq \gamma .$$

PROOF. The proof follows directly from the definition of τ and the definition of nonzero stopping time. □

DEFINITION 3.2.2. Let $\delta \in \Delta$, the mapping L_δ of V is defined component-wise by

$$(L_\delta v)(i) := \mathbb{E}_{i,\delta}[\sum_{k=0}^{\tau-1} r(s_k) + v(s_\tau)], \quad i \in S .$$

REMARK 3.2.1. Note that as a consequence of the definitions of δ and V, $(L_\delta v)(0) = 0$ for all $v \in V$.

EXAMPLE 3.2.1. Let $\delta \in \Delta$ be the nonrandomized stopping time that corresponds to the goahead set G_1 and let $v \in V$ then

$$(L_\delta v)(i) = r(i) + \sum_{j\in S} p(i,j)v(j) .$$

EXAMPLE 3.2.2. Let $\delta \in \Delta$ be the stopping time that corresponds to the goahead set G_H and $v \in V$ then

$$(L_\delta v)(i) = r(i) + \sum_{j<i} p(i,j)(L_\delta v)(j) + \sum_{j\geq i} p(i,j)v(j) .$$

EXAMPLE 3.2.3. Let $\delta \in \Delta$ be the stopping time that corresponds to the goahead set G_R then

$$(L_\delta v)(i) = (1-p(i,i))^{-1}r(i) + (1-p(i,i))^{-1} \sum_{j\neq i} p(i,j)v(j), \; i \neq 0.$$

DEFINITION 3.2.3. The matrix P_δ is defined to be the matrix with (i,j)-th element $(p_\delta(i,j))$ equal to

$$p_\delta(i,j) := \sum_{n=0}^{\infty} \mathbb{P}_{i,\delta}(s_n = j, \tau = n) .$$

LEMMA 3.2.4. Let $\delta \in \Delta$ be a nonzero stopping time then

$$\rho_\delta := \| P_\delta \| \leq (1 - \inf_{i \in S} \delta(i)) + (\inf_{i \in S} \delta(i)) \| P \| < 1 .$$

PROOF. First note that $\| P_\delta \|$ is finite, since $\| P_\delta \| \leq \sum_{n=0}^{\infty} \| P^n \| < \infty$ (assumption 3.1.4).

For $\delta \in \Delta$ we define the stopping time $\delta_M \in \Delta$ by

$$\delta_M(\alpha) := \begin{cases} 0 & \text{if } \alpha \in \bigcup_{k=M+1}^{\infty} (S \backslash \{0\})^k \\ \delta(\alpha) & \text{else .} \end{cases}$$

Now,

$$\forall_{\varepsilon > 0}\ \exists_{N \in \mathbb{N}}\ \forall_{M > N}\ \left| \| P_{\delta_M} \| - \| P_\delta \| \right| < \varepsilon,$$

since

$$\left| \| P_{\delta_M} \| - \| P_\delta \| \right| \leq \sum_{n=M}^{\infty} \| P^n \| .$$

So it suffices to prove the lemma for stopping times δ_M. This will be done by induction with respect to M.

Let $\Delta_n \subset \Delta$ be the set of stopping times with

$$\Delta_n := \{ \delta \in \Delta \mid \delta(\alpha) = 0 \text{ for all } \alpha \in \bigcup_{k=n+1}^{\infty} (S)^{K} \}.$$

So Δ_0 only contains the stopping times with $\delta(i) = 0$ for all $i \in S\backslash$. For $\delta \in \Delta_0$ we have $\| P_\delta \| = 1$.

Suppose $\delta \in \Delta_1$ is a nonzero stopping time, then

$$\begin{aligned} \sum_{j \in S} P_\delta(i,j)\mu(j) &= \sum_{j \in S} [\mathbb{P}_{i,\delta}(s_0 = j,\ \tau = 0) + \mathbb{P}_{i,\delta}(s_1 = j,\ \tau = 1]\mu(j) \\ &= (1 - \delta(i))\mu(i) + \delta(i) \sum_{j \in S} p(i,j)\mu(j) \\ &\leq [(1 - \delta(i)) + \delta(i)\| P \|]\mu(i) < 1 \end{aligned}$$

Since $\delta \in \Delta_1$ is supposed to be a nonzero stopping time, there exists a number $\varepsilon > 0$ such that $\delta(i) > \varepsilon$, for all $i \in S$ which implies $\| P_\delta \| = \rho_\delta < 1$. Now we state the induction hypothesis: Suppose for arbitrary $n \geq 1$

$$\| P_\delta \| \leq 1 \text{ on } \Delta_n \text{ and } \| P_\delta \| < 1 \text{ if } \delta \in \Delta_n \text{ is nonzero .}$$

Let $\delta \in \Delta_{n+1}$, we define $\delta_i(\alpha) := \delta(i,\alpha)$ for $i \in S$ and $\alpha \in G_\infty$. It is easily verified that $\delta_i \in \Delta_n$.

Now for each $i \in S$ we have

$$\begin{aligned}
\sum_{j\in S} p_\delta(i,j)\mu(j) &= \sum_{j\in S} \sum_{m=0}^{n+1} \mathbb{P}_{i,\delta}(s_m = j,\ \tau = m)\mu(j) \\
&= \sum_{j\in S} [\mathbb{P}_{i,\delta}\ (s_0 = j,\ \tau = 0) + \\
&\quad + \sum_{m=1}^{n+1} \mathbb{P}_{i,\delta}\ (s_m = j,\ \tau = m)]\mu(j) \\
&= (1 - \delta(i))\mu(i) + \delta(i) \sum_{j\in S} \sum_{m=1}^{n+1} \sum_{k\in S} p(i,k)\ \cdot \\
&\quad \cdot \mathbb{P}_{i,\delta}\ (s_m = j,\ \tau = m \mid e_0 = 0,\ s_1 = k)\mu(j) \\
&= (1 - \delta(i))\mu(i) + \\
&\quad + \ \delta(i) \sum_{k\in S} p(i,k) \sum_{j\in S} \sum_{m=0}^{n} \mathbb{P}_{k,\delta_i}(s_m = j,\ \tau = m)\mu(j) \\
&\leq (1 - \delta(i))\mu(i) + \delta(i) \sum_{k\in S} p(i,k)\mu(k) \\
&\leq (1 - \delta(i))\mu(i) + \ \delta(i)\| P \|)\mu(i)\ .
\end{aligned}$$

So if $\delta \in \Delta_{n+1}$ is a nonzero stopping time then

$$\| P_\delta \| = \rho_\delta < 1\ .$$

This completes the proof. □

LEMMA 3.2.5. Let $\delta \in \Delta$ then $L_\delta V = V$.

PROOF. We first mention a property of Markov chains which holds for non-anticipating stopping mechanisms

$$\mathbb{E}_j r(s_k) = \mathbb{E}_{i,\delta}(r(s_{\tau+k})|s_\tau = j) \quad \text{if } \mathbb{P}_{i,\delta}(s_\tau = j) > 0,\ k \in \mathbb{N}.$$

Consider for each $i \in S$

$$\begin{aligned}
(L_\delta V)(i) &= \mathbb{E}_{i,\delta}[\sum_{k=0}^{\tau-1} r(s_k) + V(s_\tau)] \\
&= \mathbb{E}_{i,\delta}[\sum_{k=0}^{\tau-1} r(s_k)] + \sum_{n=0}^{\infty} \sum_{j\in S} \mathbb{P}_{i,\delta}(s_n = j,\ \tau = n) V(j) \\
&= \mathbb{E}_{i,\delta}[\sum_{k=0}^{\tau-1} r(s_k)] + \sum_{n=0}^{\infty} \sum_{j\in S} \mathbb{P}_{i,\delta}(s_n = j,\ \tau = n) \sum_{k=0}^{\infty} \mathbb{E}_j r(s_k) \\
&= \mathbb{E}_{i,\delta}[\sum_{k=0}^{\tau-1} r(s_k)] + \sum_{n=0}^{\infty} \sum_{j\in S} \sum_{k=0}^{\infty} \mathbb{P}_{i,\delta}(s_n = j,\ \tau = n) \mathbb{E}_j r(s_k) \\
&= \mathbb{E}_{i,\delta}[\sum_{k=0}^{\tau-1} r(s_k)] + \\
&\quad + \sum_{n=0}^{\infty} \sum_{j\in S} \sum_{k=0}^{\infty} \mathbb{P}_{i,\delta}(s_\tau = j,\ \tau = n) \mathbb{E}_{i,\delta}[r(s_{\tau+k})|s_\tau = j] \\
&= \mathbb{E}_{i,\delta}[\sum_{k=0}^{\tau-1} r(s_k)] + \sum_{k=0}^{\infty} \mathbb{E}_{i,\delta} r(s_{\tau+k}) = \mathbb{E}_{i,\delta} \sum_{k=0}^{\infty} r(s_k) = V(i)
\end{aligned}$$

where the interchange of summations is justified by the fact that

$$\sum_{n=0}^{\infty} \mathbb{E}_i |r(s_n)| < \infty . \qquad \square$$

LEMMA 3.2.6. Let $\delta \in \Delta$ then

(i) L_δ maps $\mathcal{V}$ into $\mathcal{V}$.

(ii) L_δ is a monotone mapping.

(iii) L_δ is strictly contracting if and only if δ is nonzero.

(iv) The contraction radius of L_δ equals $\rho_\delta = \| P_\delta \|$.

(v) The set $\{v \in \mathcal{V} \mid \| v - (1-\rho)^{-1}b \| \le (1-\rho_o)^{-2}M'\}$ is mapped by L_δ in itself, where $M' := \dfrac{2M_1}{1-\rho_\delta}$ and M_1 is defined as in the proof of lemma 3.1.1 by $M_1 := \| Pb - \rho b \| + 2\| r - b \|$.

PROOF. The proof of (i) follows from lemma 3.2.5 and theorem 3.1.1; since $V \in \mathcal{V}$ each $v \in \mathcal{V}$ may be written as $v = V + w$ with $w \in \mathcal{W}$. Now $L_\delta v = L_\delta(V+w) = L_\delta V + P_\delta w = V + P_\delta w \in \mathcal{V}$. The monotonicity of L_δ is trivial.
To prove (iii) we first note that $v_1, v_2 \in \mathcal{V}$ imply that $(v_1 - v_2)$, $(L_\delta v_1 - L_\delta v_2)$ are elements of $\mathcal{W}$. Moreover v_1, v_2 may be given by $v_1 = V + w_1$, $v_2 = V + w_2$ with $w_1, w_2 \in \mathcal{W}$. So

$$\| L_\delta v_1 - L_\delta v_2 \| = \| P_\delta w_1 - P_\delta w_2 \| \le \| P_\delta \| \cdot \| w_1 - w_2 \| .$$

L_δ is strictly contracting if and only if δ is nonzero follows from lemma 3.2.4. The contraction radius equals $\| P_\delta \|$ as is verified by choosing $v_1 = V + \mu$ and $v_2 = V$. The last assertion follows from $\| V - (1-\rho)^{-1}b \| \le M_1(1-\rho_o)^{-2}$ so each $v \in \{v \in \mathcal{V} \mid \| v - (1-\rho)^{-1}b \| \le (1-\rho_o)^{-2}M'\}$ may be written as $v = V + w$, where the μ-norm of $w \in \mathcal{W}$ is at most

$$\| w \| \le (1-\rho_o)^{-2}M_1 + (1-\rho_o)^{-2}M' .$$

Now

$$\begin{aligned} \| L_\delta v - (1-\rho)^{-1}b \| &= \| L_\delta(V+w) - (1-\rho)^{-1}b \| \le \\ &\le \| V - (1-\rho)^{-1}b \| + \| P_\delta \| \cdot \| w \| \\ &\le M_1 \ (1-\rho_o)^{-2} + \rho_\delta(M_1 + M')(1-\rho_o)^{-2} = \\ &= ((1+\rho_\delta)M_1 + \rho_\delta M')(1-\rho_o)^{-2} \le M'(1-\rho_o)^{-2} . \quad \square \end{aligned}$$

THEOREM 3.2.1. For any nonzero stopping time $\delta \in \Delta$ the mapping L_δ has the unique fixed point V (independent of δ).

PROOF. The proof follows directly from the fact that $\mathcal{V}$ is a complete metric space and the foregoing lemmas. □

LEMMA 3.2.7.

(i) If $\delta_1, \delta_2 \in \Delta$ and $\delta_1 \le \delta_2$ then

$$\rho_{\delta_1} := \| P_{\delta_1} \| \ge \| P_{\delta_2} \| =: \rho_{\delta_2} .$$

(ii) Suppose δ_1, δ_2 are nonrandomized nonzero stopping times corresponding to the goahead sets G^1 and G^2 then

$$G^1 \subset G^2 \Rightarrow \delta_1 \le \delta_2 \text{ and thus } \rho_{\delta_1} \ge \rho_{\delta_2} .$$

(iii) Let Q be a set of indices, let δ_q correspond to the nonzero goahead set G_g then

$$\rho_{\delta^+} \le \sup_{q \in Q} \rho_{\delta_q} \quad \text{and} \quad \rho_{\delta^-} \ge \inf_{q \in Q} \rho_{\delta_q} .$$

LEMMA 3.2.8. Let $\delta \in \Delta$ be nonzero, $v_0^\delta \in \mathcal{V}$ and $v_n^\delta := L_\delta(v_{n-1}^\delta)$ then

(i) $\quad v_n^\delta \to V \qquad$ (in μ-norm)

(ii) $\quad L_\delta v_0^\delta \le v_0^\delta \Rightarrow v_n^\delta \downarrow V \qquad$ (in μ-norm)

(iii) $\quad L_\delta v_0^\delta \ge v_0^\delta \Rightarrow v_n^\delta \uparrow V \qquad$ (in μ-norm)

where the convergence is component-wise.

PROOF. The proof of (i) is a direct consequence of theorem 3.2.1, whereas parts (ii) and (iii) follow from the monotonicity of the mapping L_δ and theorem 3.2.1. □

So the determination of the total expected reward over an infinite time horizon, V, may be done by successive approximation of V by v_n^δ, with arbitrary nonzero stopping time δ and arbitrary element v_0^δ of $\mathcal{V}$.
Particularly if δ_1, δ_H, δ_R are the nonrandomized nonzero stopping times corresponding to the goahead sets G_1, G_H, G_R respectively, the following well-known successive approximations converge to the return vector V

(i) $v_n^{\delta_1} := r + Pv_{n-1}^{\delta_1}$, with $v_0^{\delta_1} \in V$;

(ii) $v_n^{\delta_H}$ is component-wise defined by

$$v_n^{\delta_H}(i) := r(i) + \sum_{j<i} p(i,j)v_n^{\delta_H}(j) + \sum_{j\geq i} p(i,j)v_{n-1}^{\delta_H}(j), \quad i \in S ,$$

with $v_0^{\delta_H} \in V$;

(iii) $v_n^{\delta_R}$ is component-wise defined by

$$v_n^{\delta_R}(i) := (1-p(i,i))^{-1}r(i) + (1-p(i,i)^{-1} \sum_{j\neq i} p(i,j)v_{n-1}^{\delta_R}(j), \; i \in S,$$

$i \neq 0$, with $v_0^{\delta_R} \in V$.

Furthermore if in each of these approximations v_0^{δ} is chosen as required in lemma 3.2.8(ii) or (iii) the convergence will be monotone.

3.3. *A discussion on the assumptions 3.1.2-3.1.5*

The following lemmas will clarify some of the relations between the bounding function μ and the function b (remind remark 3.1.2(vi)).

LEMMA 3.3.1. Suppose $\exists_{M'\in\mathbb{R}^+} \forall_{i\in S} [b(i) \geq -M'\mu(i)]$ then there exist a ρ', $0 < \rho' < 1$ and a bounding function μ' such that

(3.3.1) $\quad \|P\|_{\mu'} \leq \rho' < 1$

(3.3.2) $\quad \|r\|_{\mu'} < \infty$.

PROOF. Choose $M_3 := \max\{2\|Pb - \rho b\|_{\mu}\cdot(1-\rho_o)^{-1}, 2M'\}$, $\rho' := \rho_o + \frac{1}{2}(1-\rho_o)$ and $\mu'(i) := b(i) + M_3\mu(i)$, then clearly μ' is a bounding function. Now

$$\begin{aligned}\|r\|_{\mu'} &= \sup_{i\in S\setminus\{0\}} (|r(i)|)(b(i)+M_3\mu(i))^{-1} \\ &\leq \sup_{i\in S\setminus\{0\}} (|b(i)|)(b(i)+M_3\mu(i))^{-1} \\ &\quad + \sup_{i\in S\setminus\{0\}} (|r(i)-b(i)|)(b(i)+M_3\mu(i))^{-1} < \infty .\end{aligned}$$

Furthermore

$$
\begin{aligned}
P\mu' = P(b+M_3\mu) &\le \rho b + Pb - \rho b + \rho_o M_3\mu \\
&\le \rho b + \| Pb - \rho b\|_\mu \cdot \mu + \rho_o M_3\mu \\
&\le \rho'\mu' \,.
\end{aligned}
$$

□

In a similar way the following lemma can be proved.

LEMMA 3.3.2. Suppose $\exists_{M'\in\mathbb{R}^+} \forall_{i\in S} [b(i) \le M'\mu(i)]$; then there exist a ρ', $0 < \rho' < 1$ and a bounding function μ' such that (3.3.1) and (3.3.2) are satisfied.

The latter two lemmas express that if the reward function is bounded from one side (with respect to the weighting factors $\mu(i)$), a new bounding function μ' can be defined such that the Banach space $W_{\mu'}$ contains b, r, V_n and V. However, for the existence of such a bounding function μ', the condition that b is bounded from one side (with respect to μ) is essential, as is illustrated by the following example.

EXAMPLE 3.3.1. $S := \{0,1,2,\ldots\}$, $p(0,0) = 1$, $\forall_{i\in S\setminus\{0\}}\ p(i,0) = 1-\rho$, $r(0) = 0$. Let $i_0 := \min_{i\in\mathbb{N}} \{(\frac{i}{i+2})^2 > \rho\}$, if i_0 is even then we redefine $i_0 := i_0 + 1$.
For all $0 < i < i_0$ the rewards and transition probabilities are given by $r(i) = 0$, $p(i,i) = \rho$, $p(i,j) = 0$ for $j \ne i$ and $j \ne 0$. For $i > \frac{1}{2}i_0$, we choose

$$
\begin{aligned}
&p(2i,2i+2) = p(2i-1,2i+1) = \rho(1-a_i)\ , \\
&p(2i,2i+1) = p(2i-1,2i+2) = \rho a_i\ ,
\end{aligned}
$$

with

$$
\begin{aligned}
&a_i = \tfrac{1}{2}(1-\rho(\tfrac{i+1}{i})^2);\ r(2i) = (i)^{-2}\rho^{-i}; \\
&r(2i+1) = -(i+1)^{-2}\rho^{-(i+1)};\ p(2i-1,j) = p(2i,j) = 0 \text{ otherwise}; \\
&b = r;\ \mu(i) = 1 \text{ for } i \ne 0.
\end{aligned}
$$

We clearly have $\| P \|_{\mu} = \rho$.

Moreover it is easily verified that

$$\sum_{n=0}^{\infty} \sum_{j \in S} p^{(n)}(2i,j)|r(j)| \le \rho^{-i} \sum_{n=i}^{\infty} n^{-2} ,$$

and

$$\sum_{n=0}^{\infty} \sum_{j \in S} p^{(n)}(2i+1,j)|r(j)| \le \rho^{-(i+1)} \sum_{n=i+1}^{\infty} n^{-2} .$$

It can also be verified that

$$\forall_{i \in S} \ |\sum_{j} p(i,j)b(j) - \rho b(i)| = 0 .$$

So the assumptions 3.1.2-3.1.5 are satisfied.

However, no bounding function μ' exists for which $\rho' < 1$ and $\| r \|_{\mu'} < \infty$. This follows since $r(2i) = (i)^{-2}\rho^{-i}$, $r(2i+1) = -(i+1)^{-2}\rho^{-(i+1)}$ for $i > \frac{1}{2}i_0$ implies that an eventual bounding function μ' should satisfy

$$\mu'(2i) \ge \frac{1}{M} \rho^{-i}(i)^{-2} =: \mu_0(2i)$$

$$\mu'(2i+1) \ge \frac{1}{M} \rho^{-(i+1)}(i+1)^{-2} =: \mu_0(2i+1) ,$$

for some $M \in \mathbb{R}^+$ and $i > \frac{1}{2}i_0$.

Assume the existence of a bounding function μ' and a $\rho' < 1$ such that

$$(3.3.3) \qquad \mu' \ge \frac{1}{\rho'} P\mu' .$$

We define

$$i_1 := \max\{i_0, \min_{i \in \mathbb{N}}\{(\frac{i}{i+1})^2 > \frac{1}{2}(1+\rho')\}\} .$$

Then, substituting μ_0 in the right hand side of (3.3.3) yields, for $i > 2i_1$, the condition

$$\mu'(i) \geq \beta\mu_0(i) =: \mu_1(i) \; ,$$

with $\beta := \frac{1}{\rho'}(\frac{1+\rho'}{2}) > 1$.
Substituting $\mu_1(i)$ in (3.3.3) yields for $i > 2i_1$

$$\mu'(i) \geq \beta\mu_1(i) = \beta^2\mu_0(i) \; .$$

Iterating in this way proves that no bounding function exists.

REMARK 3.3.1.

(i) It is easily verified that the assumptions 3.1.3 and 3.1.4 do not imply assumption 3.1.5. By replacing the rewards in the above example by $|r|$ the assumptions 3.1.3 and 3.1.4 remain satisfied whereas assumption 3.1.5 fails.

(ii) If b is bounded from one side (in μ-norm) it follows from lemma 3.3.1 or 3.3.2 that r is a charge with respect to P, since

$$\| \sum_{n=0}^{\infty} P^n|r| \|_{\mu'} \leq (1-\rho')^{-1}\| r \|_{\mu'} \; ,$$

$$\sum_{n=0}^{\infty} P^n|r|(i) \leq (1-\rho')^{-1}|r(i)|(\mu'(i))^{-1}, \quad i \neq 0 \; .$$

In this case assumption 3.1.3 may be replaced by the assumption that b is bounded from one side (in μ-norm).

CHAPTER 4

MARKOV DECISION PROCESSES

As mentioned in chapter 1 we consider in this and the following chapters Markov decision processes.

In section 4.1 the model is described. Next, in section 4.2, decision rules and the assumptions on the transition probabilities and on the reward structure will be introduced. Again the assumptions will allow for an unbounded reward structure. They are in fact a natural extension of the assumptions 3.1.2-3.1.5 to the case in which decisions are permitted. A number of results about Markov decision processes will be proved under our assumptions (section 4.3). For example, the existence of ε-optimal stationary Markov decision rules will be shown. For Markov decision processes with a bounded reward structure this result has also been obtained by Blackwell [5] and Denardo [12].

Harrison [24] proved the same for discounted Markov decision processes with a bounding function $\mu(i) = 1$ for $i \in S\backslash\{0\}$. Moreover, in section 4.3 we prove the convergence (in μ-norm) of the standard dynamic programming algorithm.

The final section will be devoted again to a discussion on the assumptions.

4.1. *The Markov decision model*

We consider a Markov decision process on the countably infinite or finite state space S at discrete points in time $t = 0,1,\dots$. In each state $i \in S$ the set of *actions* available is A. We allow A to be a general set and suppose $\mathcal{A}$ to be a σ-field on A with $\{a\} \in \mathcal{A}$ if $a \in A$. If the system's actual state is $i \in S$ and action $a \in A$ is selected, then the system's next state will be $j \in S$ with probability $p^a(i,j)$.

ASSUMPTION 4.1.1.

(i) $\qquad \forall_{a\in A}\, \forall_{i,j\in S}\; p^a(i,j) \geq 0$,

(ii) $\qquad \forall_{a\in A}\, \forall_{i\in S} \sum_{j\in S} p^a(i,j) = 1$,

(iii) $\qquad \forall_{a\in A}\; p^a(0,0) = 1$,

(iv) $p^a(i,j)$ as a function of a, is a measurable function on $(A,\mathcal{A})$ for each $i,j \in S$.

If state $i \in S$ is observed at time n and action $a \in A$ has been selected, then an immediate (expected) reward $r(i,a)$ is earned. So from now on the reward function r is a real valued measurable function on $S \times A$.

The objective is to choose the actions at the successive points in time in such a way that the total expected reward over an infinite time horizon is maximal. A precise formulation will be given in the following sections.

It will be shown later on (chapter 9) that our model formulation includes the discounted case (with a discounted factor $\beta < 1$), since β may be supposed to be incorporated in the $p^a(i,j)$. The same holds for semi-Markov decision processes where it is only required that t is interpreted as the number of the decision moment rather than actual time. For semi-Markov decision processes with discounting, the resulting discount factor depends on i,j and $a \in A$ only, and may again be supposed to be incorporated in the transition probabilities $p^a(i,j)$.

4.2. *Decision rules and assumptions*

In the first part of this section we are concerned with the concept of decision rules. Roughly, a decision rule is a recipe for taking actions at each point in time. A decision rule will be denoted by π. The action to be selected at time n, according to π, may be a function of the entire history of the process until that time. We allow for the decision rule π to be such that for each state $i \in S$ actions are selected by a random mechanism. This random mechanism may be a function of the history too.

DEFINITION 4.2.1.

(i) An n-*stage history* h_n of the process is a $(2n+1)$-tuple

$$h_n := (i_0,z_0,i_1,z_1,\dots,z_{n-1},i_n), \quad i_t \in S,\ z_t \in A .$$

(ii) H_n, $n \geq 0$ denotes the set of all possible n-stage histories.

(iii) $\mathcal{S}_n$ is the product σ-field of subsets of H_n generated by $\mathcal{S}_0$ and $\mathcal{A}$.

DEFINITION 4.2.2.

(i) Let q_n be a transition probability of $(H_n, \mathcal{S}_n)$ into $(A, \mathcal{A})$, $n \geq 0$. So

(a) for every $h_n \in H_n$ $q_n(\cdot|h_n)$ is a probability measure on $(A, \mathcal{A})$;

(b) for every $A' \in \mathcal{A}$ $q_n(A'|\cdot)$ is measurable on $(H_n, \mathcal{S}_n)$.

Then a *decision rule* (π) is defined to be a sequence of transition probabilities, $\pi := (q_0, q_1, q_2, \ldots)$.

(ii) The set of all decision rules is denoted by D.

DEFINITION 4.2.3.

(i) A decision rule $\pi = (q_0, q_1, \ldots)$ is called *nonrandomized* if $q_n(\cdot|h_n)$ is a degenerated measure on $(A, \mathcal{A})$ for each $n \geq 0$, i.e. $\exists_{a \in A} [q_n(a|h_n) = 1]$. The set of all nonrandomized decision rules is $\mathcal{N}$.

(ii) A decision rule $\pi = (q_0, q_1, \ldots)$ is said to be *Markov* or *memoryless* if for all $n \geq 0$ $q_n(\cdot|h_n)$ depends on the last component of h_n only.

(iii) The set of all Markov decision rules is denoted by RM.

(iv) A decision rule is said to be a *Markov-strategy* if it is nonrandomized and Markov.

(v) The set of Markov strategies is denoted by $\mathcal{M}$.

(vi) A Markov strategy can thus be identified with a sequence of functions $\{f_n \mid n = 0, 1, \ldots\}$ where f_n is a function from S into A. Such a function is called a (Markov) *policy*. The set of all possible policies is denoted by F.

(vii) A Markov strategy is called *stationary* if all its component policies are identical. We denote by f^∞ the stationary Markov strategy with component f. F^∞ denotes the set of all stationary Markov strategies.

(viii) For $\pi = (f_0, f_1, \ldots) \in \mathcal{M}$ and $g \in F$ we denote by $(g, \pi) := (g, f_0, f_1, \ldots)$ the Markov strategy that applies g first and then applies the policies of π in their given order.

For $n \in \mathbb{N}$ we define the measurable space $((S \times A)^n, \mathcal{S}^n_{0,A})$, where $\mathcal{S}^n_{0,A}$ is the product σ-field generated by $\mathcal{S}_0$ and $\mathcal{A}$. The product space $(\Omega_{0,A}, \mathcal{F}_{0,A})$ is the space with $\Omega_{0,A} = (S \times A)^\infty$ and $\mathcal{F}_{0,A}$ the product σ-field of subsets of $(S \times A)^\infty$ generated by $\mathcal{S}_0$ and $\mathcal{A}$. For each $\pi = (q_0, q_1, \ldots)$ and each $n \in \mathbb{N}$ we define for $((S \times A)^n, \mathcal{S}^n_{0,A})$ and $(S \times A, \mathcal{S}_{0,A})$ the transition probability on $((S \times A)^n, \mathcal{S}^n_{0,A})$ as follows

$$p_{n,n+1}(B \times A' \mid (i_0,z_0),\dots,(i_n,z_n)) :=$$

$$= \sum_{i\in B} p^{z_n}(i_n,i) \int_{A'} q_{n+1}(da \mid (i_0,z_0,\dots,i_n,z_n,i))$$

where $B \in S_0$ and $A' \in A$.

For each decision rule $\pi \in D$ and each starting state $i_0 \in S$ (as a consequence of the Ionescu Tulcea theorem, see e.g., Neveu [52]), the sequence transition probabilities $\{p_{n,n+1}\}_n$ defines a unique probability measure $\mathbb{P}^{\pi}_{i_0}$ on $(\Omega_{0,A},F_{0,A})$. So we may consider the stochastic processes $\{(s_n,a_n) \mid n \geq 0\}$ and $\{s_n \mid n \geq 0\}$ where s_n and a_n are the projections on the n-th state space and n-th action space respectively. This means that s_n is the state and a_n is the action at time n.
It may be verified that for the simplified situation considered in chapter 3, the Ionescu Tulcea theorem would yield the probability measure $\mathbb{P}_{i_0}$.

NOTATIONS 4.2.1.

(i) Let v be a real valued function on $(\Omega_{0,A},F_{0,A})$. We denote by $\mathbb{E}^{\pi}_i v$ the expectation of v with respect to the probability measure $\mathbb{P}^{\pi}_i$.

(ii) $\mathbb{E}^{\pi} v$ denotes the vector with i-th component equal to $\mathbb{E}^{\pi}_i v$.

(iii) If $\pi \in D$ is a stationary Markov strategy $\pi = \{f,f,\dots\}$ we may use $\mathbb{E}^f_i$, $\mathbb{E}^f$ instead of $\mathbb{E}^{\pi}_i$, $\mathbb{E}^{\pi}$ respectively.

(iv) By r^f we denote the vector with components $r(i,f(i))$.

(v) By P^f we denote the matrix with (i,j)-th element equal to $p^{f(i)}(i,j)$.

(vi) For each $\pi \in D$ we define $P^0(\pi) := I$, where I is the identity matrix, and the matrix $P^n(\pi)$, $(n > 0)$ is the matrix with (i,j)-th entry equal to $\mathbb{P}^{\pi}_i(s_n = j)$.
So if $\pi \in M$; $\pi = (f_0,f_1,f_2,\dots)$ then $P^n(\pi) = P^{f_0}P^{f_1}\dots P^{f_{n-1}}$.

The stochastic process $\{s_n,a_n \mid n \geq 0\}$ is not necessarily a Markov process since the decision rules $\pi \in D$ may be such that actions are selected depending on the complete history of the process. If π is Markov then the stochastic process $\{s_n \mid n \geq 0\}$ is a, not necessarily stationary, Markov chain. Moreover, if π is a stationary Markov strategy, then the stochastic process $\{s_n \mid n \geq 0\}$ is a Markov chain with stationary transition probabilities.

As mentioned, we will use the expected total reward over an infinite time horizon as a measure for the effectiveness of a decision rule π. If at time n the history $h_n := \{i_0, z_0, i_1, z_1, \ldots, i_n\}$ has been observed and action $z_n \in A$ is selected at that point in time, then the total reward (return) over n time periods is

$$(4.2.1) \quad V_{n+1}(h_n, z_n) := \sum_{k=0}^{n} r(i_k, z_k) \ .$$

Without assumptions on the reward function r and the transition probabilities $p^a(i,j)$ there is of course no guarantee that under an arbitrary decision rule π, V_n has a finite expectation. In order to guarantee the existence of the expectation of V_n and the total expected reward over an infinite time horizon for an arbitrary decision rule π we generalize the assumptions 3.1.2-3.1.5 to the situation in which decisions (actions) are allowed.

ASSUMPTION 4.2.1.

$$\exists_{M \in \mathbb{R}^+} \forall_{f \in F} \ \| r^f - b \| < M \ .$$

ASSUMPTION 4.2.2.

$$\sup_{\pi \in M} \sum_{n=0}^{\infty} P^n(\pi) |b| < \infty \ .$$

ASSUMPTION 4.2.3.

$$\exists_{\rho_* < 1} \forall_{f \in F} \ \| P^f \| \leq \rho_* \ .$$

ASSUMPTION 4.2.4.

$$\exists_{M \in \mathbb{R}^+} \forall_{f \in F} \ \| P^f b - \rho b \| \leq M \ .$$

REMARK 4.2.1.

(i) Note that if A contains one element only, the assumptions coincide with those in section 3.1.

(ii) As a consequence of the definition of the bounding function μ and the function b assumption 4.2.1 implies that $\forall_{a\in A}\ r(0,a) = 0$. Since for $i = 0$ the reward and the transition probabilities are in fact independent of $a \in A$ the actions may be identified.

(iii) r^f may be written as $r^f = b + y^f$ where $y^f \in \mathcal{W}$.

(iv) We define $\rho_* := \sup_{f\in F} \{\|P^f\|\}$ and $\rho_\circ := \max\{\rho,\rho_*\}$.

LEMMA 4.2.1.

(i) $$\exists_{M\in\mathbb{R}^+}\ \forall_{f,g\in F}\ \|P^f b - \rho r^g\| \le M ,$$

(ii) $$\exists_{M\in\mathbb{R}^+}\ \forall_{f,g\in F}\ \|P^f r^g - \rho b\| \le M ,$$

(iii) $$\exists_{M\in\mathbb{R}^+}\ \forall_{f,g,h\in F}\ \|P^f r^g - \rho r^h\| \le M .$$

PROOF. We will only prove part (iii) since the proof of the other parts proceeds along the same lines

$$\begin{aligned}\|P^f r^g - \rho r^h\| &= \|P^f b + P^f y^g - \rho b - \rho y^h\| \\ &\le \|P^f b - \rho b\| + \|P^f y^g\| + \rho\|y^h\| \\ &\le \|P^f b - \rho b\| + \|P^f\|\cdot\|y^g\| + \rho\|y^h\| < \infty .\end{aligned}$$ □

LEMMA 4.2.2.

$$\forall_{\pi\in\mathcal{M}}\ \sum_{n=0}^{\infty} P^n(\pi)\,|r^{f_n}| < \infty .$$

PROOF. For $\pi = (f_0,f_1,\ldots)$, $P^n(\pi)\,|r^{f_n}| = P^n(\pi)\,|b + y^{f_n}|$

$$\|P^n(\pi)\| \le \prod_{k=0}^{n-1} \|P^{f_k}\| \le \rho_*^n .$$

Since $y^{f_n} \in \mathcal{W}$ for all $f_n \in F$ we have $\|P^n(\pi)y^{f_n}\| \le \rho_*^n M_1$, where M_1 is chosen in accordance with assumption 4.2.1.

So,
$$\sum_{n=0}^{\infty} \| P^n(\pi) y^{f_n} \| \le (1-\rho_*)^{-1} M_1$$

which implies

$$\sum_{n=0}^{\infty} P^n(\pi) |y^{f_n}| \le (1-\rho_*)^{-1} M_1 \cdot \mu \ .$$

This yields the result

$$\sum_{n=0}^{\infty} P^n(\pi) |r^{f_n}| \le \sum_{n=0}^{\infty} P^n(\pi) |b| + \sum_{n=0}^{\infty} P^n(\pi) |y^{f_n}| < \infty \ . \qquad \square$$

REMARK 4.2.2. In terms of potential theory (see e.g. Hordijk [33]). Lemma 4.2.2 says that the reward structure is a charge structure with respect to the transition probabilities.

4.3. *Some properties of Markov decision processes*

In the previous section we have introduced the assumptions on the reward and the probability structure. In this section a number of results in Markov decision theory will be proved under our assumptions. We shall first prove that the total expected reward over an infinite time horizon for every decision rule $\pi \in D$ is an element of the complete metric space V. Let π be an arbitrary decision rule, Given the initial state $i \in S$ the decision rule determines the unique probability measure $\mathbb{P}_i^{\pi}$ on $(\Omega_{0,A}, F_{0,A})$ as described in the previous section.

LEMMA 4.3.1. For each $\pi \in D$, $\| P^n(\pi) \| \le \rho_*^n$.

PROOF. The proof proceeds by induction. The statement is trivial for $n=0,1$. Suppose it is true for some $n \ge 1$, then

$$\sum_{j\in S} \mathbb{P}_i^{\pi}(s_{n+1} = j)\mu(j) = \int_{H_n} \int_{A} \sum_{j\in S} q_n(da|h_n) p^a(s_n,j) \mathbb{P}_i^{\pi}(dh_n)\mu(j)$$

$$= \int_{H_n} \mathbb{P}_i^\pi(dh_n) \int_A q_n(da|h_n) \sum_j p^a(s_n,j)\mu(j)$$

$$\le \int_{H_n} \mathbb{P}_i^\pi(dh_n)\rho_*\mu(s_n)$$

$$= \sum_{\ell\in S} \mathbb{P}_i^\pi(s_n = \ell)\rho_*\mu(\ell) \le \rho_*^{n+1}\mu(i) \ .$$ □

LEMMA 4.3.2. There exists a $M \in \mathbb{R}^+$ such that for all $\pi \in D$, and all $n=0,1,2,\ldots$

$$\|\mathbb{E}^\pi r(s_n,a_n) - \rho^n b\| \le \rho_\circ^{n-1}M(n+1) \ .$$

PROOF. The proof proceeds along the same lines as the proof of lemma 3.1.2. Choose $M := (2M_1 + M_2)$ where M_1 is such that $\forall_{f\in F} \|r^f - b\| \le M_1$ and M_2 is such that $\|P^f r^g - \rho r^h\| \le M_2$ for all $f,g,h \in F$. Then the proposition is true for $n = 0,1$, as follows from lemma 4.2.1(ii). Assume it is true for an arbitrary $n \ge 1$, then we have to prove that it is true for $n+1$. We first note that for all $n \ge 1$ and for all $i \in S$

$$\|\mathbb{E}^\pi[r(s_{n+1},a_{n+1}) - \rho r(s_n,a_n)]\| \le \rho_\circ^n M$$

as follows from

$$\mathbb{E}_i^\pi[r(s_{n+1},a_{n+1}) - \rho r(s_n,a_n)] =$$

$$= \mathbb{E}_i^\pi[\sum_{j\in S} p^{a_n}(s_n,j) \int_A q_{n+1}(da \mid s_0,a_0,\ldots,a_{n-1},s_n,a_n,j)r(j,a) + \\ - \rho r(s_n,a_n)]$$

$$\le \mathbb{E}_i^\pi[\sum_{j\in S} p^{a_n}(s_n,j)(b(j) + M_1\mu(j)) - \rho r(s_n,a_n)]$$

$$\le \mathbb{E}_i^\pi[\rho b(s_n) + M_2\mu(s_n) + \rho_* M_1\mu(s_n) - \rho r(s_n,a_n)]$$

$$\le \mathbb{E}_i^\pi[(M_2 + 2\rho_\circ M_1)\mu(s_n)]$$

$$\le \rho_\circ(M_2 + 2\rho_\circ M_1)\mathbb{E}_i^\pi\mu(s_{n-1}) \le \ldots \le \rho_\circ^n(M_2 + 2\rho_\circ M_1)\mu(i) \le \rho_\circ^n M\mu(i) \ .$$

Using this inequality and the induction hypothesis yields

$$\begin{aligned}\mathbb{E}_i^\pi r(s_{n+1},a_{n+1}) &= \rho\mathbb{E}_i^\pi r(s_n,a_n) + \mathbb{E}_i^\pi[r(s_{n+1},a_{n+1}) - \rho r(s_n,a_n)] \\ &\le \rho\mathbb{E}_i^\pi r(s_n,a_n) + \rho_o^n M\mu(i) \\ &\le \rho^2\mathbb{E}_i^\pi r(s_{n-1},a_{n-1}) + 2\rho_o^n M\mu(i) \\ &\vdots \\ &\le \rho^{n+1}\mathbb{E}_i^\pi r(s_0,a_0) + \rho_o^n(n+1)M\mu(i) \\ &\le \rho^{n+1}b(i) + \rho_o^{n+1}M_1\mu(i) + \rho_o^n(n+1)M\mu(i) \\ &\le \rho^{n+1}b(i) + \rho_o^n(n+2)M\mu(i) \ .\end{aligned}$$

In a similar way it may be proved that

$$\mathbb{E}_i^\pi r(s_{n+1},a_{n+1}) \ge \rho^{n+1}b(i) + \rho_o^n(n+2)M\mu(i) \ . \qquad \square$$

Let the decision rule π be given. The expected n-period reward by using decision rule π given the initial state $i \in S$ is defined by

$$(4.3.1) \qquad V_{\pi,n}(i) := \mathbb{E}_i^\pi \sum_{k=0}^{n-1} r(s_k,a_k) \ .$$

THEOREM 4.3.1. For all $\pi \in D$, define the corresponding total expected reward vector V_π by

$$V_\pi := \lim_{n\to\infty} V_{\pi,n} \ ,$$

then

$$\| V_\pi - (1-\rho)^{-1}b \| \le (1-\rho_o)^{-2}M_o, \quad \text{with } M_o := (\rho_o)^{-1}M$$

where M is defined as in the proof of lemma 4.3.2.

PROOF.

$$\| \sum_{k=0}^{\infty} (\mathbb{E}^{\pi} r(s_k, a_k) - \rho^k b) \| \le \sum_{k=0}^{\infty} \rho_o^k (k+1) M_o \le (1-\rho_o)^{-2} M_o . \qquad \square$$

So for each $\pi \in D$ the corresponding expected total reward over an infinite time horizon exists and is an element of V. We now want to prove the existence of ε-optimal Markov strategies.

DEFINITION 4.3.1.

(i) A decision rule π^* is said to be ε-*optimal* if and only if

$$V_{\pi^*} \ge V_{\pi} - \varepsilon\mu , \quad \text{for all } \pi \in D .$$

(ii) A decision rule π^* is said to be *optimal* if and only if

$$V_{\pi^*} \ge V_{\pi} , \quad \text{for all } \pi \in D .$$

THEOREM 4.3.2. (i) For every $\varepsilon > 0$ and $\pi = (q_0, q_1, \ldots) \in D$ there exists a Markov strategy $\pi^* \in M$ such that for all $i \in S$

$$V_{\pi^*}(i) \ge V_{\pi}(i) - \varepsilon\mu(i) ,$$

(ii) $$\sup_{\pi_o \in M} V_{\pi_o}(i) = \sup_{\pi \in D} V_{\pi}(i) .$$

REMARK 4.3.1.

(i) Part (ii) is proved in a more general setting by van Hee [28]. To prove his theorem van Hee used a result of Derman and Strauch (which was generalized by Hordijk [33]) in which was stated that for fixed $i \in S$ and an arbitrary $\pi \in D$ there is a $\pi^* \in RM$ such that

$$\mathbb{P}_i^{\pi}[s_n = j, a_n \in A'] = \mathbb{P}_i^{\pi^*}[s_n = j, a_n \in A'] \quad \text{for all } j \in S, A' \in \mathcal{A}.$$

We will use a slightly different approach that proceeds along the same lines as a proof given by Blackwell [5] for discounted Markov decision processes with bounded rewards.

(ii) The result obtained by van Hee covers the corresponding results of Blackwell [6] (positive dynamic programming), Strauch [66] (negative dynamic programming), Hordijk [34] (convergent dynamic programming) and our result (contracting dynamic programming).

(iii) For discounted (semi-) Markov decision processes with a finite state space and a finite decision space an elementary proof is given in Wessels and van Nunen [76].

PROOF OF THEOREM 4.3.2. Choose N such that $\sum_{N+1}^{\infty} \rho_\circ^n (n+1)M \le \frac{1}{4}\varepsilon$ where M is as defined in the proof of lemma 4.3.2. Now if π' is another decision rule such that $q_0' = q_0,\ q_1' = q_1, \dots, q_N' = q_N$ then

$$\| V_\pi - V_{\pi'} \| \le 2 \sum_{N+1}^{\infty} \rho_\circ^n (n+1) M \le \tfrac{1}{2}\varepsilon \ .$$

This enables us to assume that π is Markov from some point (say N + 1) on. So π might be represented by $\pi = (q_0, q_1, \dots, q_N, f_{N+1}, f_{N+2}, \dots)$ with $f_k \in F$ for $k > N$. We will show that there exists a decision rule π'' of the form $\pi'' := (q_0, q_1, \dots, q_{N-1}, f_N, f_{N+1}, \dots)$ such that

$$V_{\pi''}(i) \ge V_{\pi'}(i) - \varepsilon'\mu(i), \quad \varepsilon' > 0 \ .$$

Using this fact N times will produce the desired Markov strategy π^*, if ε' is sufficiently small.

For each $j \in S$ we define

$$V_{\pi'}^{N+1}(j) := \sum_{n=N+1}^{\infty} \mathbb{E}_i^{\pi'} [r(s_n, a_n) \mid s_{N+1} = j] \quad \text{for}$$

$$\mathbb{P}_i^{\pi'} (s_{N+1} = j) > 0, \ 0 \text{ otherwise} \ .$$

We determine the action $f_N(i) \in A$ such that

$$\text{(4.3.2)} \quad r(i, f_N(i)) + \sum_{j \in S} p^{f_N(i)}(i,j) V_{\pi'}^{N+1}(j) \ge$$

$$\ge \sup_{a \in A} \{ r(i,a) + \sum_{j \in S} p^a(i,j) V_{\pi'}^{N+1}(j) \} - \varepsilon'\mu(i) \ .$$

However, since

$$V_{\pi'}(i) = V_{\pi,N-1}(i) + \mathbb{E}_i^{\pi}[r(s_N,a_N) + \sum_{j\in S} p^{a_N}(s_N,j)V_{\pi'}^{N+1}(j)]$$

and

$$V_{\pi''}(i) = V_{\pi,N-1}(i) + \mathbb{E}_i^{\pi}[r(s_N,f_N(s_N)) + \sum_{j\in S} p^{f_N(s_N)}(s_N,j)V_{\pi'}^{N+1}(j)] ,$$

it will be clear that π'' with f_N such that (4.3.2) holds has the desired property. □

As a consequence of the foregoing theorem we can state the following corollary.

COROLLARY 4.3.1. For all $i \in S$, all $\pi \in D$ and each $\varepsilon > 0$ we have

$$\sup_{\pi_o \in M} V_{\pi_o}(i) > V_\pi(i) - \varepsilon\mu(i) .$$

So if we look for an ε-optimal decision rule π, and its corresponding total expected return over an infinite time horizon V_π, it is allowed to restrict the investigations to Markov strategies.

LEMMA 4.3.3. For each $v \in V$ and each $\pi \in M$ we have

$$\forall_{i\in S} \lim_{n\to\infty} (P^n(\pi)v)(i) = 0 .$$

PROOF. From lemma 4.3.2 it follows that $\| P^n(\pi)b - \rho^n b\| \to 0$ for $n \to \infty$. Since v may be written as $v = (1-\rho)^{-1}b + w$, with $w \in W$ and since $\| P^n(\pi)\| \le \rho_*^n$ the statement will be clear. □

Now in a similar way as we have introduced the mapping L_1 on V in chapter 3 we introduce the mapping L_1^f of V for each $f \in F$.

DEFINITION 4.3.2. Let $f \in F$, $v \in V$, then the mapping L_1^f of V is defined component-wise by

$$(L_1^f v)(i) := r(i,f(i)) + \sum_{j \in S} p^{f(i)}(i,j)v(j) \ , \quad i \in S \ ,$$

or in vector notation

$$L_1^f v := r^f + P^f v \ .$$

REMARK 4.3.2. For $P := P^f$, $r := r^f$ the results of chapter 3 may be obtained. So we have e.g.

(i) L_1^f is a monotone mapping of V into V.

(ii) L_1^f is a contraction mapping on V with contraction radius $\| P^f \| =: \rho^f$.

(iii) $L_1^f V_{f^\infty} = V_{f^\infty}$.

We now define the well-known optimal return operator U_1 on V (see e.g. Blackwell [4], MacQueen [46], Harrison [24] or van Nunen [53], [54]).

DEFINITION 4.3.3. The mapping U_1 of V is defined component-wise as follows

$$(U_1 v)(i) := \sup_{a \in A} \{r(i,a) + \sum_{j \in S} p^a(i,j)v(j)\}, \quad i \in S \ ,$$

or in vector notation

$$U_1 v := \sup_{f \in F} L_1^f v := \sup_{f \in F} \{r^f + P^f v\} \ .$$

REMARK 4.3.3. It will be clear that $U_1 v(i)$ gives the supremum of the expected return that can be earned if we start in state i at time $t = 0$, use decision $a \in A$ at time $t = 0$ and receive $v(j)$ if, as a result of that decision, state $j \in S$ is reached at time $t = 1$.

THEOREM 4.3.3.

(i) U_1 maps V into V.

(ii) U_1 is a monotone mapping.

(iii) U_1 is a contraction mapping with contraction radius

$$\nu_1 := \sup_{f\in F} \| P^f \| := \sup_{f\in F} \rho^f = \rho_* < 1 .$$

(iv) U_1 maps $\{v \in V \mid \| v - (1-\rho)^{-1} b \| \le (1-\rho_o)^{-2} M_o\}$ into itself, where M_o is defined as in theorem 4.3.1.

(v) U_1 has a unique fixed point V^*; so V^* is the unique solution in V of the optimality equation

$$(4.3.3) \qquad v = U_1 v .$$

(vi) Let $v_0 \in V$ then the sequence $v_n := U_1 v_{n-1}$ converges in μ-norm to V^*.

PROOF. Clearly for any $\varepsilon > 0$ and $v \in V$ there exists an $f \in F$ such that $L_1^f v \ge U_1 v - \varepsilon\mu$. It now follows from remark 4.3.2(i) that U_1 maps V into V. The monotonicity of U_1 is trivial. Let v_1 and $v_2 \in V$. For any $\varepsilon > 0$ choose $f,g \in F$ such that

$$L_1^f v_1 \ge U_1 v_1 - \varepsilon\mu \ ; \quad L_1^g v_2 \ge U_1 v_2 - \varepsilon\mu .$$

Then

$$U_1 v_1 - U_1 v_2 \le L_1^f v_1 + \varepsilon\mu - L_1^f v_2 = P^f(v_1 - v_2) + \varepsilon\mu ,$$

on the other hand

$$U_1 v_1 - U_1 v_2 \ge L_1^g v_1 - L_1^g v_2 - \varepsilon\mu = P^g(v_1 - v_2) - \varepsilon\mu ,$$

which yields

$$\| U_1 v_1 - U_1 v_2 \| \le \max(\| P^f \|, \| P^g \|) \| v_1 - v_2 \| + \varepsilon .$$

Since $\varepsilon > 0$ was chosen arbitrary we have

$$\| U_1 v_1 - U_1 v_2 \| \le \nu_1 \| v_1 - v_2 \| .$$

By using $v_1 := (1-\rho)^{-1}b + \ell\mu$ and $v_2 := (1-\rho)^{-1}b$, with $\ell \in \mathbb{R}^+$, it is easily verified that for any $\varepsilon > 0$

$$\| U_1v_1 - U_1v_2 \| \geq (\nu_1 - \varepsilon)\| v_1 - v_2 \| = (\nu_1 - \varepsilon)\ell$$

for ℓ sufficiently large. So the contraction radius of U_1 equals ν_1. Let $v \in \{v \in V \mid \| v - (1-\rho)^{-1}b \| \leq (1-\rho_o)^{-2}M_o\}$ then for each $\varepsilon > 0$ there exists an $f \in F$ such that $L_1^f v \geq U_1 v - \varepsilon\mu$. However, since for each $f \in F$, L_1^f maps $\{v \in V \mid \| v - (1-\rho)^{-1}b \| \leq (1-\rho_o)^{-2}M_o\}$ into itself, also part (iv) of the theorem is proved. The assertions (v) and (vi) are direct consequences of the fact that U_1 is a contraction mapping from the complete metric space V into V, see e.g. Ljusternik and Sobolew [43]. □

THEOREM 4.3.4.

(i) $V_\pi \leq V^*$ for all $\pi \in D$.

(ii) For each $\varepsilon > 0$ there exists an $f \in F$ such that $V_{f^\infty} \geq V^* - \varepsilon\mu$.

(iii) Let $\pi \in M$ and $g \in F$ then $[V_{(g,\pi)} > V_\pi] \Rightarrow [V_{g^\infty} > V_\pi]$.

(iv) π^* is optimal if and only if V_{π^*} satisfies the optimality equation (4.3.3), so $V_{\pi^*} = V^*$.

PROOF. Suppose $\pi = \{f_0, f_1, \ldots\}$ is an arbitrary element of M. Then, $L_1^{f_0} L_1^{f_1} \ldots L_1^{f_n} V^* \leq U_1^{n+1} V^* = V^*$ for each $n \geq 0$. However, $L_1^{f_0} L_1^{f_1} \ldots L_1^{f_n} V^* = V_{\pi,n} + P^n(\pi)V^*$, so letting $n \to \infty$ we have as a consequence of lemma 4.3.3 that

$$V_\pi = \lim_{n\to\infty} V_{\pi,n} \leq V^* .$$

As a consequence of theorem 4.3.2 the statement (i) is true for all decision rules π since it is true for Markov strategies.

(ii) Choose $f \in F$ in such a way that $L_1^f V^* \geq V^* - \varepsilon(1-\rho_o)\mu$, with $\varepsilon > 0$, then

$$(L_1^f)^n V^* \geq V^* - \varepsilon(1-\rho_o)[1 + \rho_o + \ldots + \rho_o^{n-1}]\mu ,$$

so, letting $n \to \infty$, we have

$$V_{f^\infty} \geq V^* - \varepsilon\mu \ .$$

(iii) $V_{(g,\pi)} = L_1^g V_\pi$. Now since L_1^g is monotone we have

$$(L_1^g)^n V_\pi \geq (L_1^g)^{n-1} V_\pi \geq \ldots \geq L_1^g V_\pi > V_\pi \ ,$$

so again, letting n → ∞, we obtain

$$V_{g^\infty} > V_\pi \ .$$

(iv) If $V_{\pi^*} = V^*$ then trivially π^* is optimal (see part (i) of this theorem). Reversely if π^* is optimal then $L_1^f V_{\pi^*} \leq V_{\pi^*}$ for all $f \in F$ since otherwise f^∞ would be better then π^* and thus π^* would not be optimal. Hence, $U_1 V_{\pi^*} \leq V_{\pi^*}$. However, since U_1 is a monotone operator we have

$$U_1^n V_{\pi^*} \leq V_{\pi^*} \ ,$$

so

$$\lim_{n\to\infty} U_1^n V_{\pi^*} = V^* \leq V_{\pi^*} \ ,$$

this yields the equality

$$V^* = V_{\pi^*} \ . \qquad \square$$

REMARK 4.3.4.

(i) If $V_{f^\infty} \geq V^* - \varepsilon\mu$ then f^∞ is ε-optimal.

(ii) We will often denote V_{f^∞} by V_f.

(iii) From part (iii) of the foregoing theorem we see that Howard's policy improvement routine [35], remains valid in this more general setting.

(iv) A proof of the foregoing theorems for discounted Markov decision processes with a general state and action space but with a bounded reward structure was given by Blackwell [5]. For discounted Markov decision processes with countable state and action space and with a bounded reward structure an elementary proof was given by Ross [62]. For bounded rewards a proof can also be found in Denardo [12].

For two actions $a_0, a_1 \in A$ and $i \in S$ it is possible that $r(i,a_0) = r(i,a_1)$ and $\forall_{j \in S}\ p^{a_0}(i,j) = p^{a_1}(i,j)$. We then say that a_0 and a_1 coincide with respect to $i \in S$, otherwise a_0 and a_1 are different with respect to state $i \in S$.

Let $A(i)$ contain representatives for all classes of coinciding actions with respect to $i \in S$. We say $A(i)$ is finite if $A(i)$ contains only finitely many elements.

LEMMA 4.3.4. If $A(i)$ is finite for each $i \in S$ then there exists an optimal stationary Markov strategy.

PROOF. If, for each $i \in S$, the action set $A(i)$ is finite then, since the supremum which defines the mapping U_1 is defined component-wise there will exist an $f \in F$ such that $L_1^f V^* = V^*$,

$$V_{f^\infty} = \lim_{n\to\infty} (L_1^f)^n V^* = V^* ,$$

which means that f^∞ is optimal. □

REMARK 4.3.5.

(i) A proof of this lemma, for discounted Markov decision processes with a bounded reward structure, is also given by Blackwell [5] and Derman [15].

(ii) If $A(i)$ is not finite for each $i \in S$ then it is easy to construct counter examples showing that an optimal decision rule may not exist, see e.g. Blackwell [5].

Our interest is in fact the computation of the optimal return vector V^* and the determination of an $(\varepsilon-)$optimal stationary Markov strategy. We will prove that as a consequence of the theorems 4.3.3 and 4.3.4 the method of successive approximation may be used to determine the optimal return vector V^* and an $(\varepsilon-)$optimal stationary Markov strategy.

DEFINITION 4.3.4. A function g defined on the power set of F into F is said to be a *choice function* if and only if for each nonempty set $B \subset F$ we have $g(B) \in B$.

From now on we assume g to be an arbitrary but given choice function.

REMARK 4.3.6. For the existence of a choice function, in a general situation as we have defined it here, we need the *axiom of choice*. We introduce the choice function only for notational convenience.
For the processes we described, the full strengtheness is never required since we have to make only countably many choices. In the first place, only at the successive points in time $n = 0,1,\dots$ a policy f has to be selected from a nonempty set $B \subset F$. Secondly, since S may at most be countable, for each $f \in F$ only countably many actions have to be selected.
Usually, in practical situations, the choice function can be given explicitly. This is specifically true in the situation that S and A are finite, since in that situation also F contains finitely many elements.

DEFINITION 4.3.5. For $\varepsilon > 0$ we define the mapping $U_{1,\varepsilon}$ of V by

$$U_{1,\varepsilon}v := L_1^f v \ ,$$

with $f := g(\{f \mid \|U_1 v - L_1^f v\| < \varepsilon\})$, from theorem 4.3.3 it follows that $U_{1,\varepsilon}$ is well defined for $\varepsilon > 0$.
If $\forall_{v \in V}\ \exists_{f \in F}\ U_1 v = L_1^f v$ then ε may be zero in that case we define $U_{1,0}v := L_1^f v$, with $f = g(\{f \mid U_1 v = L_1^f v\})$.

LEMMA 4.3.5. The mapping $U_{1,\varepsilon}$ maps V into V.

PROOF. The proof is evident since for each $f \in F$ L_1^f maps V into V. □

DEFINITION 4.3.6.

(i) A mapping B of V into V is said to be ε-*monotone* ($\varepsilon \geq 0$) if for each $v,w \in V$, with $v \geq w$, we have

$$Bv \geq Bw - \varepsilon\mu \ .$$

(ii) A mapping B of V into V is said to be ε-*contracting* ($\varepsilon \geq 0$) with contraction radius ρ' if for each $v,w \in V$ we have

$$\|Bv - Bw\| \leq \rho'\|v - w\| + \varepsilon \ .$$

LEMMA 4.3.6. The mapping $U_{1,\varepsilon}$ $(\varepsilon > 0)$ has the following properties

(i) $U_{1,\varepsilon}$ is ε-monotone $(\varepsilon > 0)$.

(ii) $U_{1,\varepsilon}$ is ε-contracting with contraction radius ν_1.

If $U_{1,0}$ is defined then $U_{1,0}$ has these properties as well.

PROOF.

(i) Let $v,w \in \mathcal{V}$ be such that $v \geq w$; then since U_1 is a (0-) monotone mapping we have

$$U_{1,\varepsilon}v - U_{1,\varepsilon}w \geq U_1v - \varepsilon\mu - U_1w \geq -\varepsilon\mu \ .$$

(ii) $$U_{1,\varepsilon}v - U_{1,\varepsilon}w \leq U_1v - U_1w + \varepsilon\mu \ ,$$

$$U_{1,\varepsilon}v - U_{1,\varepsilon}w \geq U_1v - \varepsilon\mu - U_1w \ .$$

So

$$\| U_{1,\varepsilon}v - U_{1,\varepsilon}w \| \leq \| U_1v - U_1w \| + \varepsilon \leq \nu_1 \| v - w \| + \varepsilon \ . \qquad \square$$

LEMMA 4.3.7. Let $v_0 \in \mathcal{V}$ and suppose $U_{1,0}$ is defined. Let the sequence v_n be defined by $v_n := U_{1,0}v_{n-1}$ and let $f_n := g(\{f \mid U_1v_{n-1} = L_1^f v_{n-1}\})$, then

(i) $$\| v_n - v^* \| \leq \rho_*^n \| v_0 - v^* \|,$$

(ii) $$\| v_{f_n} - v_n \| \leq (1-\rho_*)^{-1}\rho_* \| v_n - v_{n-1} \| \ ,$$

(iii) If $v_0 \in \mathcal{V}$ is chosen such that $U_1v_0 \geq v_0$ then

$$v_{n-1} \leq v_n \leq v_{f_n} \leq v^* \ .$$

PROOF.

(i) $$\| v_n - v^* \| = \| U_1^n v_0 - v^* \| = \| U_1^n v_0 - U_1^n v^* \| \leq \rho_*^n \| v_0 - v^* \| \ .$$

(ii) Consider

$$(L_1^{f_n})^k v_n - v_n = (L_1^{f_n})^k v_n - (L_1^{f_n})^{k-1} v_n + (L_1^{f_n})^{k-1} v_n - \ldots + L_1^{f_n} v_n - v_n$$

which yields

$$\| (L_1^{f_n})^k v_n - v_n \| \le (1-\rho_*)^{-1} (1-\rho_*^{k-1}) \| L_1^{f_n} v_n - v_n \|$$

$$\le \rho_* (1-\rho_*)^{-1} (1-\rho_*^{k-1}) \| v_n - v_{n-1} \| .$$

By letting $k \to \infty$ we have

$$\| V_{f_n} - v_n \| \le (1-\rho_*)^{-1} \rho_* \| v_n - v_{n-1} \| .$$

The final part follows from the monotonicity of U_1 and $L_1^{f_n}$. □

LEMMA 4.3.8. Let $\varepsilon > 0$ and $v_0 \in V$ such that

$$\forall_{i \in S} \, [v_0(i) \le (U_1 v_0)(i) - \varepsilon\mu(i)] .$$

Let the sequence v_n $(n \ge 0)$ be defined by

$$v_n := L_1^{f_n} v_{n-1} ,$$

where $f_n := g(B_n)$ with $B_n := \{f \in F \mid L_1^f v_{n-1} \ge \max\{v_{n-1}, U_1 v_{n-1} - \varepsilon(1-\rho_\circ)\mu\}\}$, then

(i) $\| v_n - V^* \| < \varepsilon$, for n sufficiently large ,

(ii) $v_{n-1} \le v_n \le V_{f_n} \le V^*$.

PROOF. We define $\delta := (1-\rho_\circ)\varepsilon$. Note that B_n is not empty. Namely, if $v_{n-1} \ge U_1 v_{n-1} - \delta\mu$, then $f_n := f_{n-1}$ already suffices, since

$$v_{n-1} = L_1^{f_{n-1}} v_{n-2} \ge v_{n-2} .$$

Since, $L_1^{f_{n-1}}$ is monotone we have

$$(L_1^{f_{n-1}})^2 v_{n-2} \geq L_1^{f_{n-1}} v_{n-2} = v_{n-1} .$$

Similarly, we have for $k \in \mathbb{N}$

$$(L_1^{f_n})^k v_{n-1} \geq v_n$$

which yields by letting $n \to \infty$

$$v_n \leq V_{f_n^\infty} =: V_{f_n} .$$

However we know already $V^* \geq V_{f_n}$, so

$$v_n \leq V_{f_n} \leq V^* .$$

The inequality $v_n \geq v_{n-1}$ follows directly from the definition of v_n. It now remains to be proven that $v_n \to V^*$ in μ-norm. Since $v_n \geq v_{n-1}$ this convergence will be monotone.

$$\begin{aligned} v_n = L_1^{f_n} v_{n-1} &\geq U_1 v_{n-1} - \delta\mu \geq U_1(U_1 v_{n-2} - \delta\mu) - \delta\mu \\ &\geq U_1^2 v_{n-2} - \rho_\circ \delta\mu - \delta\mu . \end{aligned}$$

Continuing in this way we get,

$$v_n \geq U_1^n v_0 - \delta\mu(1 + \rho_\circ + \ldots + \rho_\circ^{n-1}) \geq U_1^n v_0 - \varepsilon(1 - \rho_\circ^n)\mu .$$

Now, since $U_1^n v_0 \to V^*$ (in μ-norm) we have the final result $\| v_n - V^* \| < \varepsilon$ for n sufficiently large. □

REMARK 4.3.7. It is easy to find a starting vector $v_0 \in V$ satisfying $v_0(i) \leq U_1 v_0(i) - \varepsilon\mu(i)$ namely $v_0 := (1-\rho)^{-1} b - \ell\mu$ for $\ell \in \mathbb{R}^+$ chosen sufficiently large.

LEMMA 4.3.9. Let $v_0 \in V$. Let the sequence v_n be defined by

$$v_n := U_{1,\varepsilon_n} v_{n-1} \ ,$$

with $\varepsilon_n := \nu_1^n$ and let $f_n := g(\{f \in F \mid \| U_1 v_{n-1} - L_1^f v_{n-1} \| \le \varepsilon_n\})$, then

(i) $\lim_{n\to\infty} v_n = V^*$ (in μ-norm) ,

(ii) $\lim_{n\to\infty} V_{f_n} = V^*$ (in μ-norm) ,

PROOF. Let $M := \| v_0 - V^* \|$, then

$$\begin{aligned} \| v_1 - V^* \| &= \| v_1 - U_1 v_0 + U_1 v_0 - V^* \| \\ &\le \| v_1 - U_1 v_0 \| + \| U_1 v_0 - U_1 V^* \| \\ &\le \nu_1 + \nu_1 M = \nu_1 (M+1) \ . \end{aligned}$$

The first part of the proof follows now by induction.
Suppose $\| v_n - V^* \| \le \nu_1^n (M + n)$ for some $n \ge 0$ then

$$\begin{aligned} \| v_{n+1} - V^* \| &\le \| L_1^{f_{n+1}} v_n - U_1 v_n \| + \| U_1 v_n - U_1 V^* \| \\ &\le \nu_1^{n+1} + \nu_1 (\nu_1^n (M + n)) = \nu_1^{n+1} (M + (n + 1)) \ . \end{aligned}$$

So since the induction hypothesis holds for $n = 0,1$ it holds for all n, this implies part (i) of the lemma.

Choose $\varepsilon > 0$.
Note that as a consequence of part (i) there exists a $N \in \mathbb{N}$ such that

(a) $\forall_{n>N} \| v_n - v_{n-1} \| \le \frac{1}{2}\varepsilon (1 - \nu_1)$

(b) $\forall_{n>N} \| U_1 v_{n-1} - v_{n-1} \| \le \frac{1}{2}\varepsilon (1 - \nu_1)$.

Now, as a consequence of (a) we have

$$\begin{aligned}\| (L_1^{f_n})^k v_{n-1} - v_{n-1} \| &= \| (L_1^{f_n})^k v_{n-1} - (L_1^{f_n})^{k-1} v_{n-1} + \\ &\quad + (L_1^{f_n})^{k-1} v_{n-1} - \ldots + L_1^{f_n} v_{n-1} - v_{n-1} \| \\ &\le (1-\nu_1)^{-1}(1-\nu_1^k) \| v_n - v_{n-1} \| < \tfrac{1}{2}\varepsilon \ .\end{aligned}$$

So by letting $k \to \infty$ we have $\| V_{f_n} - v_{n-1} \| < \frac{1}{2}\varepsilon \quad (n > N)$.

Similarly, we deduce from (b)

$$\| V^* - v_{n-1} \| < \tfrac{1}{2}\varepsilon \ , \quad n > N \ .$$

Since ε was chosen arbitrarily this implies part (ii) of the lemma. □

So the computation of V^*, the optimal total expected reward over an infinite time horizon, may be executed by successive approximation of V^* by v_n as described in the lemmas 4.3.8 and 4.3.9. Moreover, for each $\varepsilon > 0$ the Markov strategy f_n^∞ is ε-optimal for n sufficiently large.

4.4. *Remarks on the assumptions 4.2.1-4.2.4.*

The first assumption (4.2.1) stated that the difference between b and r^f is bounded in μ-norm for $f \in F$. This assumption arises by combining the assumptions of Wessels [75] and Harrison [24] on the reward structure and the transition probabilities. As proved in section 3.3 our assumptions are more general than the assumptions of Wessels and Harrison separately. Assumption 4.2.2 is introduced to ensure the existence of the (conditional) expectations of the stochastic variables. If we use an extended notion of expectation (see van Hee and van Nunen [30]) assumption 4.2.2 may be replaced by the weaker assumption

$$\forall_{f,g \in F} \ P^f |r^g| < \infty \ .$$

This assumption was first used by Harrison [23]. Assumption 4.2.3 requires that also for $a \in A$ the expectation of $\mu(s_1)$ is at most $\rho_* \mu(s_0)$. This implies that for each decision rule π the corresponding stochastic process $\{s_n \mid n \ge 0\}$ has a tendency to decrease its μ-value and/or to fade. The final assumption requires that, for each $i \in S$ and each $a \in A$, the expected one-stage reward differs not too much from the immediate reward r(i,a).

For $\mu(i) = 1$, $i \neq 0$ and the usual notion of expectation our assumptions include the assumptions as used by Harrison [24] for discounted Markov decision processes. If $b \in \mathcal{W}$, the Markov decision processes as described by Wessels [75] are obtained. For a specific choice of a bounding function μ and $b \in \mathcal{W}$, the idea of using weighted supremum norms is used by Lippman [44], [45] for semi-Markov decision processes, see also van Nunen and Wessels [56]. Wijngaard [78], used the idea of a specific (exponential) weighted supremum norm for inventory problems with respect to the average reward criterion. Hinderer [31] used the idea of bounding functions for finite horizon Markov decision processes.

For $\mu(i) = 1$, $i \neq 0$ and $b \in \mathcal{W}$ our assumptions reduce to the well-known assumptions for Markov decision processes (see e.g. Denardo [12]) to guarantee the existence of the total expected reward over an infinite time horizon

$$\exists_{M \in \mathbb{R}^+} \forall_{i \in S} \forall_{a \in A} \; |r(i,a)| \leq M$$

$$\exists_{\rho < 1} \forall_{i \neq 0} \forall_{a \in A} \sum_{j \neq 0} p^a(i,j) \leq \rho \; .$$

Finally, we want to extend the lemmas 3.3.1 and 3.3.2 to the case in which decisions are allowed.

LEMMA 4.4.1. If $\exists_{M \in \mathbb{R}^+} \forall_{i \in S} \; b(i) \geq -M\mu(i)$ then there exists a number ρ' with $0 < \rho' < 1$ and a bounding function μ' such that

$$\exists_{M' \in \mathbb{R}^+} \forall_{f \in F} \; \| r^f \|_{\mu'} \leq M' \text{ and } \| P^f \|_{\mu'} \leq \rho' < 1 \; .$$

LEMMA 4.4.2. If $\exists_{M \in \mathbb{R}^+} \forall_{i \in S} \; b(i) \leq M\mu(i)$ then there exists a number ρ' with $0 < \rho' < 1$ and a bounding function μ' such that

$$\exists_{M' \in \mathbb{R}^+} \forall_{f \in F} \; \| r^f \|_{\mu'} \leq M \quad \text{and} \quad \forall_{f \in F} \; \| P^f \|_{\mu'} < \rho' < 1 \; .$$

So if in addition to the assumptions 4.2.1-4.2.4 the rewards are bounded from one side (with respect to the bounding function μ) then it is possible to define a new bounding function μ' such that the Banach space $\mathcal{W}_{\mu'}$ of vectors v with $\| v \|_{\mu'} < \infty$ contains V^* and V_f for $f \in F$. Moreover, the operators L_1^f and U_1 are contraction mappings on $\mathcal{W}_{\mu'}$ with fixed points V_f and V^* respectively. If b is bounded from one side with respect to the bounding function μ, the assumption 4.2.2 may be omitted. In that case it is easily proven that the reward structure is a charge with respect to the transition probabilities (see Hordijk [33] and Groenewegen [21]), i.e. for each $\pi \in M$, $\pi = (f_0, f_1, \ldots)$ we have

$$\sum_{n=0}^{\infty} P^n(\pi) |r^{f_n}| < \infty ,$$

which follows from

$$\| \sum_{n=0}^{\infty} P^n(\pi) |r^{f_n}| \|_{\mu'} \le (1 - \rho')^{-1} \sup_{f \in F} \| r^f \|_{\mu'} < \infty .$$

CHAPTER 5

STOPPING TIMES AND CONTRACTION IN MARKOV DECISION PROCESSES

In this chapter we will use the concept of stopping time as introduced in chapter 2 to generate a whole set of optimization procedures for solving Markov decision processes satisfying the assumptions 4.2.1-4.2.4.
The idea of using stopping times for generating such a set was introduced by Wessels [74]. Wessels used nonrandomized stopping times to generate a set of optimization procedures for finite state space, finite action space discounted Markov decision processes.
In this chapter this set of optimization procedures is extended by allowing randomized stopping times. Furthermore, the class of problems for which the procedures hold is generalized by using the less restrictive assumptions 4.2.1-4.2.4. The set of optimization procedures will include the known solution techniques as introduced by Blackwell [4], Hastings [25], Reetz [60] and van Nunen [54] for discounted Markov decision problems. Recently, again for finite state finite decision space Markov decision processes, Porteus [59] also introduced a set of optimization procedures. He introduced a number of transformations which might be used to investigate transformed Markov decision processes with the same (ε-)optimal policies and the same (possibly transformed) optimal return vector (V^*). A number of the transformations introduced by Porteus are in fact covered by our approach.

After some preliminaries the contraction mappings L_δ^π of V will be introduced and investigated (section 5.1). We will restrict the considerations to nonrandomized decision rules, which is justified by the results of the preceding chapter. Then (section 5.2) the optimal return operator U_δ on V will be defined. For each stopping time $\delta \in \Delta$ the mapping U_δ yields a *policy improvement* procedure on which the determination of an (ε-)optimal Markov strategy f^∞ and the successive approximation of V^* and V_f may be based. We will prove that the sequence v_n^δ, arising by successive application of U_δ on an arbitrary element $v_0^\delta \in V$ converges to V^* if and only if the stopping time δ is nonzero. Moreover, we prove in section 5.2 that by using U_δ attention can be restricted to stationary Markov strategies if and only if the stopping time δ is transition memoryless.
Finally, we will give algorithms which determine an (ε-)optimal Markov strategy and compute the optimal return vector V^* by means of successive approximation.

5.1. *The contraction mapping* L_δ^π

For each $n \in \mathbb{N}$ we define the measurable space $((S \times E \times A)^n, \mathcal{S}_A^n)$, where $\mathcal{S}_A^n$ is the product σ-field generated by $\mathcal{S}$ and $\mathcal{A}$. The product space $(\Omega_A, \mathcal{F}_A)$ is the space with $\Omega_A := (S \times E \times A)^\infty$ and $\mathcal{F}_A$ the product σ-field of subsets of $(S \times E \times A)^\infty$ generated by $\mathcal{S}$ and $\mathcal{A}$.

In a similar way as in chapter 3 we define, for each stopping time $\delta \in \Delta$, each decision rule $\pi := (q_0, q_1, \ldots) \in \mathcal{N}$ and each starting state $i \in S$, a probability measure $\mathbb{P}_{i,\delta}^\pi$ on $(\Omega_A, \mathcal{F}_A)$ in the standard way by giving the probability for cylindrical sets in Ω_A, $\omega_A := ((i_0, d_0, z_0), (i_1, d_1, z_1), \ldots)$.

$$\mathbb{P}_{i,\delta}^\pi(\{\omega_A \mid s_k = i_k, e_k = d_k, a_k = z_k,\ \text{for } k = 0,1,\ldots,n\}) =$$

$$= \delta_{i,i_0} \prod_{k=0}^{n-1} p^{z_k}(i_k, i_{k+1}) \cdot$$

$$\cdot \prod_{k=0}^{n} [(\delta(i_0, i_1, \ldots, i_k)^{1-d_k}(1-\delta(i_0, \ldots, i_k))^{d_k} \cdot q_k(z_k | i_0, z_0, i_1, z_1, \ldots, i_k)]$$

where δ_{i,i_0} is again the Kronecker symbol (see section 3.1).

This probability measure can also be obtained in a similar way as in chapter 4 by using transition probabilities.

For each $\pi = (q_0, q_1, \ldots) \in \mathcal{N}$ each $n \in \mathbb{N}$, and $\delta \in \Delta$ we then define for $(S \times E \times A, \mathcal{S}_A)$ the transition probabilities $p_{n,n+1}^\delta$ on $((S \times E \times A)^n, \mathcal{S}_A^n)$ as follows

$$p_{n,n+1}^\delta(B \times E' \times A' | (i_0, d_0, z_0), \ldots, (i_n, d_n, z_n)) :=$$

$$\sum_{i \in B} \sum_{d \in E'} p^{z_n}(i_n, i)[\delta(i_0, \ldots, i_n, i)]^{1-d}[1 - \delta(i_0, \ldots, i_n, i)]^d \cdot$$

$$\cdot \int_{A'} q_{n+1}(da \,|\, (i_0, z_0, \ldots, i_n, z_n, i)) \;,$$

where $A' \in \mathcal{A}$ and B and E' are elements of the power set of S and E respectively.

For each decision rule $\pi \in N$ and each starting state $i \in S$, as a consequence of the Ionescu Tulcea theorem (see e.g. Neveu [52]) these transition probabilities $p^{\delta}_{n,n+1}$ induce a unique probability measure $\mathbb{P}^{\pi}_{i,\delta}$ on (Ω_A, F_A). So we may consider the stochastic processes $\{s_n \mid n \geq 0\}$, $\{(s_n, e_n) \mid n \geq 0\}$, $\{(s_n, e_n, a_n) \mid n \geq 0\}$, where as in chapter 3 and 4, s_n, e_n, a_n are the projections on the n-th state space, the n-th E space and the n-th action space respectively.

NOTATIONS 5.1.1.

(i) For f a real valued function on (Ω, F) we denote by $\mathbb{E}^{\pi}_{i,\delta} f$ the expectation of f with respect to the probability measure $\mathbb{P}^{\pi}_{i,\delta}$.

(ii) $\mathbb{E}^{\pi}_{\delta} f$ denotes the vector with i-th component equal to $\mathbb{E}^{\pi}_{i,\delta} f$.

(iii) If $\pi \in M$ is a stationary Markov strategy, $\pi = (f,f,\ldots)$ then $\mathbb{E}^{f}_{i,\delta}$, $\mathbb{E}^{f}_{\delta}$ may be used instead of $\mathbb{E}^{\pi}_{i,\delta}$, $\mathbb{E}^{\pi}_{\delta}$ respectively.

Clearly, the probability measure $\mathbb{P}^{\pi}_{i,\delta}$ on (Ω, F) induces a unique probability measure on the coarser measurable space (Ω_0, F_0).

LEMMA 5.1.1. For each $\delta \in \Delta$, $\pi \in N$ and $i \in S$ the probability measure induced by $\mathbb{P}^{\pi}_{i,\delta}$ on the coarser measurable space (Ω_0, F_0) equals $\mathbb{P}^{\pi}_{i}$.

PROOF. The statement follows directly from the definitions of $\mathbb{P}^{\pi}_{i,\delta}$ and $\mathbb{P}^{\pi}_{i}$.□

We now extend the mapping L_{δ} of chapter 3 (defined by using the stopping function τ) for the situation in which decisions are allowed.

DEFINITION 5.1.1. For each $\delta \in \Delta$ and $\pi \in N$ the mapping L^{π}_{δ} of V is defined by

$$(L^{\pi}_{\delta} v)(i) := \mathbb{E}^{\pi}_{i,\delta}\left[\sum_{k=0}^{\tau-1} r(s_k, a_k) + v(s_\tau)\right], \quad i \in S,$$

or in vector notation

$$(L^{\pi}_{\delta}) v := \mathbb{E}^{\pi}_{\delta}\left[\sum_{k=0}^{\tau-1} r(s_k, a_k) + v(s_\tau)\right].$$

REMARK 5.1.1.

(i) We recall that $v(s_\tau) := v(0) = 0$ if $\tau = \infty$.

(ii) Note that for all $\delta \in \Delta$, $\pi \in N$, $v \in V$ we have $(L_\delta^\pi v)(0) = 0$.

LEMMA 5.1.2. If δ is the nonrandomized stopping time corresponding to G_1 and $\pi = \{q_0, q_1, \ldots\} \in N$, then L_δ^π equals $L_1^{f_0}$ as defined in chapter 4 where f_0 is defined by $f_0(i) = z_i$ and $z_i \in A$ is the action for which $q_0(z_i | i) = 1$. Consequently for that δ, L_δ^π is a monotone mapping of $V \to V$, L_δ^π is strictly contracting with contraction radius $\rho_\delta^{f_0}$ and fixed point V_{f_0}.

DEFINITION 5.1.2. For each $\delta \in \Delta$ and $\pi \in N$ the matrix P_δ^π is defined to be the matrix with (i,j)-th element equal to

$$p_\delta^\pi(i,j) := \sum_{n=0}^{\infty} \mathbb{P}_{i,\delta}^\pi(s_n = j, \tau = n) .$$

REMARK 5.1.2. Note that for each $\delta \in \Delta$ and $\pi \in N$ we have $p_\delta^\pi(0,0) = 1$ and $\forall_{j \in S \setminus \{0\}}\ p_\delta^\pi(0,j) = 0$.

LEMMA 5.1.3. Let δ be a nonzero stopping time and let $\pi \in N$, then

$$\rho_\delta^\pi := \| P_\delta^\pi \| \le (1 - \inf_{i \in S} \delta(i)) + \inf_{i \in S} \delta(i) \rho_* < 1 .$$

PROOF. The proof proceeds along the same lines as the proof of lemma 3.2.4.□

LEMMA 5.1.4.

(i) Suppose $\delta_1 \le \delta_2$ then for all $\pi \in N$, $\rho_{\delta_1}^\pi \ge \rho_{\delta_2}^\pi$.

(ii) Suppose δ_1, δ_2 are nonrandomized stopping times G^1 and G^2 are the go-ahead sets corresponding to δ_1 and δ_2. Then

$$G^1 \subset G^2 \Rightarrow \delta_1 \le \delta_2, \text{ and thus } \rho_{\delta_1}^\pi \ge \rho_{\delta_2}^\pi , \quad \text{for all } \pi \in N .$$

(iii) Let Q be an index set and suppose for each $q \in Q$, $\delta_q \in \Delta$ is a nonrandomized stopping time then for all $\pi \in N$

$$\rho^{\pi}_{\delta^+} \le \sup_{q\in Q} \rho^{\pi}_{\delta_q} \quad , \qquad \rho^{\pi}_{\delta^-} \ge \inf_{q\in Q} \rho^{\pi}_{\delta_q} \ .$$

LEMMA 5.1.5. Let $\delta \in \Delta$ and $\pi \in N$, then L^{π}_{δ} maps V into V.

PROOF. From the foregoing chapter we know that there exists a $M' \in \mathbb{R}^+$ such that for all $\pi \in N$ $\|v_\pi - (1-\rho)^{-1}b\| \le M'$. Choose, for each $v \in V$, $M_v \in \mathbb{R}^+$ such that $\|v - (1-\rho)^{-1}b\| = M_v$. So for all $\pi \in N$ we have

$$\|v - v_\pi\| \le \|v - (1-\rho)^{-1}b\| + \|v_\pi - (1-\rho)^{-1}b\| \le M_v + M' \ .$$

Let $\pi := (q_0,q_1,\dots) \in N$ and $i \in S$ let $f_0(i)$ be defined as in lemma 5.1.2. Then π_i is defined by $\pi_i := (q_0',q_1',\dots)$ with

$$q_k'(\cdot\,|\,i_0,z_0,\dots,i_k) := q_{k+1}(\cdot\,|\,i,f_0(i),i_0,z_0,\dots,i_k) \ .$$

Let Δ_N' contain all stopping times with the following property

$$\forall_{\alpha\in \bigcup_{k=N+1}^{\infty} S^k} \quad \delta(\alpha^{(N)}) \neq 0 \Rightarrow \delta(\alpha) = 1 \ .$$

Then it is easily verified that for each $i \in S$, $\pi \in N$ and $\delta \in \Delta_N'$ $\mathbb{P}^{\pi}_{i,\delta}(\tau \le N \vee \tau = \infty) = 1$.
Let $\delta \in \Delta_{N+1}'$, and define for each $i \in S$, $\delta_i \in \Delta$ by $\delta_i(\alpha) := \delta(i,\alpha)$ for each $\alpha \in \bigcup_{k=1}^{\infty} S^k$, then clearly $\delta_i \in \Delta_N'$. Now,

$$L^{\pi}_{\delta}v(i) = \mathbb{E}^{\pi}_{i,\delta}[\sum_{k=0}^{\tau-1} r(s_k,f_k(s_k)) + v(s_\tau)]$$

$$= (1-\delta(i))v(i) + \delta(i)[r(i,f_0(i)) + \mathbb{E}^{\pi}_{i,\delta}[\sum_{k=1}^{\tau-1} r(s_k,a_k) + v(s_\tau)\,|\,e_0 = 0]]$$

$$= (1-\delta(i))v(i) + \delta(i)[r(i,f_0(i)) +$$

$$+ \sum_{j\in S} p^{f_0(i)}(i,j) \cdot \mathbb{E}^{\pi_i}_{j,\delta_i}[\sum_{k=0}^{\tau-1} r(s_k,a_k) + v(s_\tau)]]$$

$$\leq (1-\delta(i))v(i)+\delta(i)[r(i,f_0(i))+\sum_{j\in S} p^{f_0(i)}(i,j)[V_{f_0}(j)+(M'+M_v)\mu(j)]]$$

$$\leq (1-\delta(i))v(i)+\delta(i)[V_{f_0}(i)+\rho_*(M'+M_v)\mu(i)]$$

$$\leq V_{f_0}(i)+(1-\delta(i))(M'+M_v)\mu(i)+\delta(i)\rho_*(M'+M_v)\mu(i)$$

$$\leq V_{f_0}(i)+(M'+M_v)\mu(i) \ .$$

In a similar way it may be proved that

$$L_\delta^\pi v(i) \geq V_{f_0}(i) - (M'+M_v)\mu(i) \ .$$

So $L_\delta^\pi v \in V$.
Since for $\delta \in \Delta_0'$, $L_\delta^\pi v \in V$, as is easily verified, it follows by induction that for all $\delta \in \Delta$, $L_\delta^\pi v \in V$. □

LEMMA 5.1.6. Let $\delta \in \Delta$ and $\pi \in N$, then

(i) L_δ^π is monotone.

(ii) L_δ^π is strictly contracting if and only if δ is a nonzero stopping time; in this case the contraction radius equals ρ_δ^π.

(iii) For each nonzero stopping time, L_δ^π possesses a unique fixed point $v_\delta^\pi \in V$.

(iv) The set $\{v \in V \mid \|v-(1-\rho)^{-1}b\| \leq (1-\rho_0)^{-2}M'\}$ is mapped in itself by L_δ^π where $M' := (1-\rho_\delta^\pi)^{-1}M$ and M is defined as in theorem 4.3.1.

PROOF. Part (i) of the lemma is trivial. To prove part (ii), let $v_1, v_2 \in V$, then v_1 and v_2 can be given by $v_k = (1-\rho)^{-1}b + w_k$ $(k = 1,2)$, where $w_k \in W$. So $\|v_1 - v_2\| = \|w_1 - w_2\|$.
Hence

$$\|L_\delta^\pi v_1 - L_\delta^\pi v_2\| = \|P_\delta^\pi w_1 - P_\delta^\pi w_2\| \leq \|P_\delta^\pi\| \cdot \|w_1 - w_2\| \ .$$

The fact that L_δ^π is strictly contracting if and only if δ is a nonzero stopping time is a consequence of lemma 5.1.3. The contraction radius equals $\|P_\delta^\pi\|$ as is easily verified by choosing $v_1 = (1-\rho)^{-1}b + \mu$, $v_2 = (1-\rho)^{-1}b$.

Part (ii) of the lemma is evident since L_δ^π is a contraction mapping on the complete metric space V (see e.g. Ljusternik and Sobolew [43]). The final statement is clear since

$$V_\delta^\pi \in \{v \in V \mid \| v - (1-\rho)^{-1} b \| \leq (1-\rho_o)^{-2} M'\} .$$

Hence

$$L_\delta^\pi v \leq V_\delta^\pi + \rho_\delta^\pi (1-\rho_o)^{-2} M' \mu \leq$$

$$\leq (1-\rho)^{-1} b + (1-\rho_o)^{-2} M \mu + (1-\rho_o)^{-2} \cdot \rho_\delta^\pi M' \mu$$

$$\leq (1-\rho)^{-1} b + (1-\rho_o)^{-2} M' \mu .$$

Similarly

$$L_\delta^\pi v \geq (1-\rho)^{-1} b - (1-\rho_o)^{-2} M' \mu . \qquad \square$$

LEMMA 5.1.7.

(i) Let $\delta \in \Delta$ be a nonzero stopping time and let $\pi := (f,f,\ldots)$ be a stationary Markov strategy, then L_δ^f has the unique fixed point V_f independent of the stopping time.

(ii) If $\pi \in N$ is not stationary then there exists a situation $\{p^a(i,j), r(i,a)\}$ for which nonzero stopping times δ_1 and $\delta_2 \in \Delta$ exist, such that $L_{\delta_1}^\pi$ and $L_{\delta_2}^\pi$ possess different fixed points.

PROOF.

(i) If π is a stationary Markov strategy then the process $\{s_n \mid n \geq 0\}$ is a Markov chain with rewards only depending on the current state. Theorem 3.2.1 now implies the result.

(ii) Suppose $\pi := (q_0, q_1, \ldots) \in N$ is nonstationary then A contains two or more elements. Let $n_0 \in \mathbb{N}$ be such that

(a) for all $k < n_0$ $\quad q_k(f_0(i) \mid i_0, f_0(i_0), i_1, f_0(i_1), \ldots, i_k) = 1$ (note that $n_0 \geq 1$).

(b) for at least one $j_0 \in S$ and at least one path

$$q_{n_0}(f_0(j_0) \mid i_0, f_0(i_0), i_1, f_0(i_1), \ldots, j_0) = 0 .$$

Let $r(i,z) := b(i)$ for $i \in S\backslash\{j_0\}$ and all $z \in A$, choose $r(j_0,f_0(j_0)) := = b(j_0)$ and $r(j_0,z) = b(j_0) + \ell\mu$ for all $z \in A\backslash f_0(j_0)$. Choose $p^{z_1}(i,j) = p^{z_2}(i,j) > 0$ for all $i,j \in S$ and $z_1,z_2 \in A$ such that the assumptions 4.1.1, 4.2.1-4.2.4 are satisfied. So for each π the stochastic process $\{s_n, n \geq 0\}$ is a Markov chain. Now by choosing δ_1, δ_{n_0} corresponding to G_1, G_{n_0} respectively it is easily verified that the fixed points of the mappings $L^{\pi}_{\delta_1}$ and $L^{\pi}_{\delta_{n_0}}$ are different for ℓ sufficiently large. □

5.2. *The optimal return mapping* U_δ

LEMMA 5.2.1. Let $i \in S$, let $\delta \in \Delta$ and $v \in V$, then for each $\varepsilon > 0$ there exists a decision rule $\pi_i \in N$ such that $L^{\pi_i}_{\delta}v(i) \geq L^{\pi}_{\delta}v(i) - \varepsilon\mu(i)$ for all $\pi \in N$.

PROOF. The existence of a $M \in \mathbb{R}^+$ such that for all $\pi \in N$, $L^{\pi}_{\delta}v \leq (1-\rho)^{-1}b + M\mu$ follows from lemma 5.1.5. So for each $i \in S$ there exists a $\pi_i \in N$ such that $L^{\pi_i}_{\delta}v(i) \geq L^{\pi}_{\delta}v(i) - \varepsilon\mu(i)$. □

DEFINITION 5.2.1. The mapping U_δ of V is defined component-wise by

$$U_\delta v(i) := \sup_{\pi\in N} L^{\pi}_{\delta}v(i) .$$

REMARK 5.2.1. Note that the supremum over $\pi \in N$ is taken component-wise.

THEOREM 5.2.1. Let $\delta \in \Delta$, then

(i) U_δ maps V into V.

(ii) U_δ is monotone.

(iii) U_δ is a contraction mapping if and only if δ is nonzero. The contraction radius ν_δ of U_δ satisfies $\nu_\delta := \sup_{\pi\in N} \rho^{\pi}_{\delta}$.

(iv) The set $\{v \in V \mid \|v - (1-\rho)^{-1}b\| \leq (1-\rho_o)^{-2}M'\}$ is mapped into itself by U_δ, where M' is defined as in lemma 5.1.6 with ρ^{π}_{δ} replaced by ν_δ.

(v) If δ is nonzero then U_δ possesses the unique fixed point V^*, or equivalently V^* is the unique solution in V of the optimality equation

$$v = U_\delta v . \tag{5.2.1}$$

PROOF. Part (i) of the proof follows directly from lemmas 5.1.5 and 5.2.1. Part (ii) is trivial.
(iii) Let $v,w \in V$ and $\varepsilon > 0$. Let $i \in S$, choose $\pi_0, \pi_1 \in N$ such that

$$L_\delta^{\pi_0} v(i) \geq L_\delta^{\pi} v(i) - \varepsilon\mu(i)$$

and

$$L_\delta^{\pi_1} w(i) \geq L_\delta^{\pi} w(i) - \varepsilon\mu(i), \quad \text{for all } \pi \in N .$$

We then have

$$U_\delta v(i) - U_\delta w(i) \leq L_\delta^{\pi_0} v(i) - L_\delta^{\pi_0} w(i) + \varepsilon\mu(i)$$

$$= \sum_{j \in S} P_\delta^{\pi_0} (w(j) - v(j)) + \varepsilon\mu(i)$$

and

$$U_\delta v(i) - U_\delta w(i) \geq L_\delta^{\pi_1} v(i) - L_\delta^{\pi_1} w(i) - \varepsilon\mu(i)$$

$$= \sum_{j} P_\delta^{\pi_1}(i,j)(v(j) - w(j)) - \varepsilon\mu(i) .$$

This implies

$$|U_\delta v(i) - U_\delta w(i)| \leq \max\{\| P_\delta^{\pi_0} \|, \| P_\delta^{\pi_1} \|\} \cdot |v(i) - w(i)| + \varepsilon\mu(i) .$$

Since ε was chosen arbitrarily we have

$$\| U_\delta v - U_\delta w \| \leq \nu_\delta \| v - w \| .$$

It follows from lemma 5.1.3 that U_δ is strictly contracting for nonzero δ. Now we prove that the contraction radius is ν_δ. Let π_0 be such that $\| P_\delta^{\pi_0} \| \geq \nu_\delta - \frac{1}{2}\varepsilon$ and let $i_0 \in S$ be such that $\sum_{j \in S} P_\delta^{\pi_0}(i_0,j)\mu(j) \geq (\nu_\delta - \varepsilon)\mu(i_0)$. Then choose $v := (1-\rho)^{-1}b + \ell\mu$ and $w := (1-\rho)^{-1}b$. Let $M \in \mathbb{R}^+$ be such that $L_\delta^{\pi_0}((1-\rho)^{-1}b) \geq (1-\rho)^{-1}b - M\mu$ (see lemma 5.1.5). Consider $U_\delta v(i_0) - U_\delta w(i_0)$

$$\begin{aligned} U_\delta v(i_0) - U_\delta w(i_0) &\geq L_\delta^{\pi_0} v(i_0) - (1-\rho)^{-1} b(i_0) - M\mu(i_0) \\ &\geq \sum_{j\in S} P_\delta^{\pi_0}(i_0,j)\ell\mu(j) - 2M\mu(i_0) \\ &\geq (\nu_\delta - \varepsilon)\ell\mu(i_0) - 2M\mu(i_0) \geq (\nu_\delta - 2\varepsilon)\ell\mu(i_0) \end{aligned}$$

for ℓ chosen sufficiently large ($\ell > \frac{2M}{\varepsilon}$).
This implies ν_δ is the contraction radius.
Since $\| P_\delta^\pi \| = 1$ if δ is not a nonzero stopping time (see the proof of lemma 5.1.3) we have also proved that U_δ is not strictly contracting in this case.
(iv) It is a direct consequence of lemma 5.1.6 that

$$\{v \in V \mid \| v - (1-\rho)^{-1} b \| \leq (1-\rho_o)^{-2} M'\}$$

is mapped into itself by U_δ.
(v) We first note that for δ nonzero U_δ has a unique fixed point. The final part of the theorem follows by considering $U_\delta V^*$.

$$U_\delta V^*(i) = \sup_{\pi\in N} \mathbb{E}_{i,\delta}^{\pi}[\sum_{k=0}^{\tau-1} r(s_k,a_k) + V^*(s_\tau)] \;.$$

Now choose $\varepsilon > 0$ and $f \in F$ such that $V_f \geq V^* - \varepsilon\mu$. In theorem 4.3.4 we have proved that such an f exists. Let $\pi \in F^\infty$ be defined by $\pi := (f,f,\dots)$.
Then

$$U_\delta V^*(i) \geq \mathbb{E}_{i,\delta}^{f}[\sum_{k=0}^{\tau-1} r(s_k,f(s_k)) + V_f(s_\tau)] - \varepsilon\mu(i) \;.$$

However, (see remark 4.3.2), we already know $L_\delta^f V_f = V_f$, which implies $U_\delta V^*(i) \geq V_f(i) - \varepsilon\mu(i)$ and since ε was chosen arbitrarily we have

$$U_\delta V^*(i) \geq V^*(i) , \qquad i \in S \;.$$

On the other hand $U_\delta V^*(i) \leq V^*(i)$ as can be proved inductively. This proof proceeds in a similar way as the first part of the proof of lemma 5.1.5. □

COROLLARY 5.2.1. Let $\delta \in \Delta$ be nonzero and let $v_0^\delta \in V$. Let the sequence v_n^δ be defined by

$$v_n^\delta := U_\delta v_{n-1}^\delta ,$$

then

$$\lim_{n\to\infty} v_n^\delta = V^* \quad \text{(in } \mu\text{-norm)} .$$

LEMMA 5.2.2.

(i) Let $\delta_1, \delta_2 \in \Delta$ be nonzero and suppose $\delta_1 \le \delta_2$, then $\nu_{\delta_1} \le \nu_{\delta_2}$.

(ii) Let $\delta_1, \delta_2 \in \Delta$ be nonrandomized and nonzero, let G^1 and G^2 be the go-ahead sets corresponding to δ_1 and δ_2 respectively, then

$$G^1 \subset G^2 \Rightarrow \delta_1 \le \delta_2 \text{ and thus } \nu_{\delta_1} \ge \nu_{\delta_2} .$$

In the previous chapter we have proved the existence of (ε-)optimal stationary Markov strategies. Regrettably, the procedure as described in corollary 5.2.1 does not produce such (ε-)optimal Markov strategies. So we should like to characterize nonzero stopping times which allow for the use of stationary Markov strategies only. The following two theorems provide the main step for such a characterization.

THEOREM 5.2.2. If $\delta \in \Delta$ is transition memoryless, $\varepsilon > 0$ and $v \in V$, then there exists an $f \in F$ such that for all $i \in S$

$$L_\delta^f v(i) \ge U_\delta v(i) - \varepsilon\mu(i) .$$

Hence

$$U_\delta v = \sup_{f\in F} L_\delta^f v .$$

PROOF. Let $v \in V$. We will define a new Markov decision process such that $L_\delta^\pi v(i)$ (of the old process) is the total expected reward over an infinite time horizon (for the new process) if the starting state is i and strategy

$\pi \in N$ is used. Hence, for this new Markov decision process attention may be restricted to memoryless Markov strategies, as is proved in chapter 4. This implies that for the determination of $U_\delta v$ in the original problem, the restriction to stationary Markov strategies is permitted too.
We will assume, as is allowed without loss of generality, $\delta(i) = 1$ for all $i \in S$. We define the new Markov decision process in the following way: $\bar{S}$, the new state space, is the union of two copies of S;

$$S^* := \{i^* \mid i \in S\},\ S_* := \{i_* \mid i \in S\} .$$

So the states in S are two times represented in $\bar{S}$. For the states $i_* \in S_* \subset \bar{S}$ and all $a \in A$ we define $p^a(i_*, 0_*) := 1$ and $r(i_*, a) = v(i)$. For the states in S^* we define

$$p^a(i_1^*, i_2^*) := p^a(i_1, i_2)\delta(i_1, i_2) ,$$

$$p^a(i_1^*, i_{2*}) := p^a(i_1, i_2)(1 - \delta(i_1, i_2)) ,$$

$$r(i^*, a) := r(i, a) .$$

It is easily verified that $L_\delta^\pi v(i)$ is just the total expected reward over an infinite time horizon if the process starts in state i and decision rule π is used. □

Transition memoryless stopping times are the only stopping times for which a restriction to memoryless or stationary Markov strategies is always allowed: this fact is expressed in the following theorem.

THEOREM 5.2.3. Suppose the stopping time $\delta \in \Delta$ is not transition memoryless, then there exists a Markov decision process with state space S (i.e. there exists a set A and numbers $\{p^a(i,j)\}$, $\{r(i,a)\}$ such that for this Markov decision process

$$\neg[\forall_{\varepsilon>0}\ \exists_{f \in F}\ [L_\delta^f v \geq U_\delta v - \varepsilon\mu]] .$$

REMARK 5.2.2. In fact $\sup_{f \in F} L_\delta^f v$ may not be defined.

PROOF OF THEOREM 5.2.3. $\delta \in \Delta$ is not transition memoryless, implies the existence of states $i_0, j_0 \in S$ and two paths $\alpha, \gamma \in \bigcup_{k=0}^{\infty} (S\backslash\{0\})^k$ such that $\delta(\alpha, i_0, j_0) < \delta(\gamma, i_0, j_0)$. For $\eta \in (S\backslash\{0\})^0$ we define $\delta(\eta, i_0, j_0) := \delta(i_0, j_0)$ where η is α or γ. Let

$$[1 - \delta(\alpha, i_0, j_0)] =: c[1 - \delta(\gamma, i_0, j_0)]$$

which implies $c > 1$. The case $\delta(\gamma, i_0, j_0) = 1$ requires a slight, self-evident modification.

We will construct a counterexample satisfying the following conditions

(1) $\forall_{i \in S\backslash(i_0)}$ $A(i)$ contains only one element, whereas $A(i_0)$ contains two elements, $A(i_0) := \{1,2\}$.

(2) For all $i \in S$, $i \notin B := \{[\alpha]_0, [\alpha]_1, \ldots, [\alpha]_{k_\alpha - 1}, [\gamma]_0, \ldots [\gamma]_{k_\gamma - 1}, i_0, j_0\}$ we have $p^a(i,i) := \rho$, $p^a(i,j) := 0$ for $j \in S\backslash\{i,0\}$; $p(i,0) := 1 - \rho$; $r(i,a) := 0$; $v(j) := 0$. Moreover, $p(0,0) := 1$.

(3) For all $i \in B$ and all j, $p^a(i,j) := 0$ if $j \notin B \cup \{0\}$, else $p^a(i,j) > 0$.

(4) $\forall_{i,j \in B} \forall_{a \in A(i)} \sum_{j \in B} p^a(i,j) \le \rho$; $\forall_{i \in B} \forall_{a \in A(i)} [p^a(i,0) = 1 - \sum_{j \in B} p^a(i,j)]$.

(5) $\mu(i) = 1$ for $i \neq 0$.

As a consequence of condition (1) the index a in $p^a(i,j)$ and $r(i,a)$ can be omitted if $i \neq i_0$.

For the investigation of $U_\delta v$ the following form has to be maximized with respect to a for $\eta = \alpha, \gamma$ respectively

$$(5.2.2) \qquad r(i_0, a) + \sum_{j \in B} p^a(i_0, j)(U_{\delta_\eta} v)(j)$$

where δ_η is defined by $\delta_\eta(\beta) := \delta(\eta, i_0, \beta)$ for all $\beta \in \bigcup_{k=1}^{\infty} S^k$ and $\eta = \alpha, \gamma$ respectively.

The second term in (5.2.2) may be written as

$$(5.2.3) \qquad p^a(i_0, j_0)(1 - \delta(\eta, i_0, j_0))v(j_0) + \sum_{j \neq j_0} p^a(i_0, j)(1 - \delta(\eta, i_0, j))v(j)$$

$$+ \sum_{j \in B} p^a(i_0, j)\delta(\eta, i_0, j) r(j,a) +$$

$$+ \sum_{j \in B} p^a(i_0, j)\delta(\eta, i_0, j) \sum_{k \in B} p^a(j,k) U_{\delta_{\eta,j}} v(k) .$$

with $\delta_{\eta,j}(\beta) = \delta(\eta,i_0,j,\beta)$ for $\beta \in \bigcup_{k=1}^{\infty} S^k$ and $\eta = \alpha,\gamma$.

Suppose $i_0 \neq j_0$.
Let $p^1(i_0,j_0) := q \neq 0$; $p^2(i_0,j_0) := \frac{1}{2}q$. Let for $j \in B\backslash\{j_0\}$, $p^1(i_0,j) := p^2(i_0,j) := \varepsilon$; for $i \neq i_0$ and all $j \in B$ $p(i,j) := \varepsilon$. We define $v(j_0) := [1 - \delta(\alpha,i_0,j_0)]^{-1}q^{-1}$ and $v(j) := 0$ for $j \neq j_0$; $r(j) := 0$ for $j \in B\backslash\{i_0\}$; we will choose $|r(i_0,a)| < 2$ for $a = 1,2$.

It is now easily verified that $U_{\delta_\eta} v(j)$, $U_{\delta_{\eta,j}} v(k)$ are bounded by $M := v(j_0) + 2(1-\rho)^{-1}$.
So formula 5.2.2 can be given (using 5.2.3) for $\eta = \alpha$ by

$$\begin{cases} r(i_0,1) + 1 + \mathcal{O}(\varepsilon) & \text{for } \varepsilon \to 0 \\ r(i_0,2) + \frac{1}{2} + \mathcal{O}(\varepsilon) & \text{for } \varepsilon \to 0\ . \end{cases}$$

For $\eta = \gamma$ we have

$$\begin{cases} r(i_0,1) + \dfrac{1}{c} + \mathcal{O}(\varepsilon) & \text{for } \varepsilon \to 0 \\ r(i_0,2) + \dfrac{1}{2c} + \mathcal{O}(\varepsilon) & \text{for } \varepsilon \to 0\ . \end{cases}$$

Choose $r(i_0,2) = 0$ and $r(i_0,1) = +\frac{1}{2}(-\frac{1}{2} - \frac{1}{2c})$. Then in state i_0 after path α decision 1 has to be selected whereas in state i_0 after path γ decision 2 is optimal if ε is chosen sufficiently small.

Suppose $j_0 = i_0$.
Let $p^1(i_0,i_0) := q$; $p^2(i_0,i_0) := \frac{1}{2}q$; $V(j) := r(j) := 0$ for $j \in B\backslash\{i_0\}$ $p^1(i_0,j) := p^2(i_0,j) = q^2$, $q << 1$ and choose $v(i_0) := [1 - \delta(a,i_0,i_0)]^{-1}q^{-1}$.
Then formula (5.2.2) can be given for $\eta = \alpha$ by

$$\begin{cases} r(i_0,1) + 1 + \mathcal{O}(q) & \text{for } q \to 0 \\ r(i_0,2) + \frac{1}{2} + \mathcal{O}(q) & \text{for } q \to 0\ . \end{cases}$$

For $\eta = \gamma$ we have

$$\begin{cases} r(i_0,1) + \frac{1}{c} + \mathcal{O}(q) & \text{for } q \to 0 \\ r(i_0,2) + \frac{1}{2c} + \mathcal{O}(q) & \text{for } q \to 0 \, . \end{cases}$$

This implies in a similar way as for $i_0 \neq j_0$ that in state i_0 after path α decision 1 is optimal whereas after path γ decision 2 is optimal in state i_0 (for q chosen sufficiently small). □

COROLLARY 5.2.2. Let $\delta \in \Delta$ be nonzero and transition memoryless and let $v_0^\delta \in V$, then the sequence $\{v_n^\delta\}$, $n \geq 0$ defined by

$$v_n^\delta := \sup_{f \in F} L_\delta^f v_{n-1}^\delta$$

converges in μ-norm to V^*.

PROOF. The statement follows directly from the foregoing theorems 5.2.1 and 5.2.2. □

DEFINITION 5.2.2. For $\varepsilon > 0$ and $\delta \in \Delta$ transition memoryless, we define the mapping $U_{\delta,\varepsilon}$ of V by

$$U_{\delta,\varepsilon} v := L_\delta^f v \, ,$$

with $f := g(\{f \mid \| U_\delta v - L_\delta^f v \| < \varepsilon\})$. From theorem 5.2.2 it follows that $U_{\delta,\varepsilon}$ is well defined for $\varepsilon > 0$. If $\forall_{v \in V} \exists_{f \in F}\ U_\delta v = L_\delta^f v$ then ε may be zero, in that case we define $U_{\delta,0} v := L_\delta^f v$, with $f := g(\{f \mid U_\delta v = L_\delta^f v\})$.

LEMMA 5.2.3. The mapping $U_{\delta,\varepsilon}$ maps V into V.

LEMMA 5.2.4. The mapping $U_{\delta,\varepsilon}$, $\varepsilon > 0$ and δ transition memoryless and nonzero, has the following properties

(i) $U_{\delta,\varepsilon}$ is ε-monotone.

(ii) $U_{\delta,\varepsilon}$ is ε-contracting with contraction radius ν_δ.

If $U_{\delta,0}$ is defined then also $U_{\delta,0}$ has these properties.

LEMMA 5.2.5. Let $\delta \in \Delta$ be transition memoryless and nonzero, let $v_0^\delta \in V$, and suppose $U_{\delta,0}$ is defined. Let the sequence v_n^δ be defined by

$$v_n^\delta := U_{\delta,0} v_{n-1}^\delta$$

and let $f_n := g(\{f \mid U_\delta v_{n-1}^\delta = L_\delta^f v_{n-1}^\delta\})$, then

(i) $$\| v_n^\delta - v^* \| \le \nu_\delta^n \| v_0^\delta - v^* \|$$

(ii) $$\| v_n^\delta - v_{f_n} \| \le (1 - \nu_\delta)^{-1} \nu_\delta \| v_n^\delta - v_{n-1}^\delta \|$$

(iii) if $v_0^\delta \in V$ is chosen such that $U_{\delta,0} v_0^\delta \ge v_0^\delta$ then

$$v_{n-1}^\delta \le v_n^\delta \le v_{f_n} \le v^* .$$

PROOF.

(i) $v_n^\delta = U_\delta^n v_0^\delta$. This implies

$$\| v_n^\delta - v^* \| = \| U_\delta^n v_0^\delta - U_\delta^n v^* \| \le \nu_\delta^n \| v_0^\delta - v^* \| .$$

(ii) $$(L_\delta^{f_n})^k v_n^\delta - v_n^\delta = (L_\delta^{f_n})^k v_n^\delta - (L_\delta^{f_n})^{k-1} v_n^\delta + (L_\delta^{f_n})^{k-1} v_n^\delta - \ldots + L_\delta^{f_n} v_n^\delta - v_n^\delta .$$

So

$$\| (L_\delta^{f_n})^k v_n^\delta - v_n^\delta \| \le (1 - \nu_\delta)^{-1} (1 - (\nu_\delta)^k) \| L_\delta^{f_n} v_n^\delta - v_n^\delta \| .$$

So for k → ∞ we find

$$\| v_{f_n} - v_n^\delta \| \le (1 - \nu_\delta)^{-1} \nu_\delta \| v_n^\delta - v_{n-1}^\delta \| .$$

Part (iii) of the lemma follows from the monotonicity of $U_{\delta,0}$ and $L_\delta^{f_n}$. □

LEMMA 5.2.6. Let $\varepsilon > 0$, let $\delta \in \Delta$ be transition memoryless and nonzero and let $v_0 \in V$ such that $v_0^\delta \le U_\delta v_0^\delta - \varepsilon\mu$. Let the sequence v_n^δ $(n \ge 0)$ be defined by

$$v_n^\delta := L_\delta^{f_n} v_{n-1}^\delta \ ,$$

where $f_n := g(B_n)$, with $B_n := \{f \mid L_\delta^f v_{n-1}^\delta \ge \max\{v_{n-1}^\delta, U_\delta v_{n-1}^\delta - \varepsilon(1-\nu_\delta)\mu\}\}$, then

(i) $\qquad \| v_n^\delta - V^* \| < \varepsilon$ for n sufficiently large

(ii) $\qquad v_{n-1}^\delta \le v_n^\delta \le V_{f_n} \le V^*$.

PROOF. The proof is completely analogous to the proof of lemma 4.3.8. □

REMARK 5.2.3.

(i) Also in this situation it is easy to find a starting vector $v_0^\delta \in V$ such that $U_\delta v_0^\delta \ge v_0^\delta + \varepsilon\mu$ by choosing e.g. $v_0^\delta := (1-\rho)^{-1} b - \ell\mu$ with $\ell \in \mathbb{R}^+$ sufficiently large.

(ii) It follows by combining part (i) and (ii) of the foregoing lemma that $\| V_{f_n} - V^* \| < \varepsilon$ for n sufficiently large.

LEMMA 5.2.7. Let $\delta \in \Delta$ be transition memoryless and nonzero and let $v_0^\delta \in V$. Let the sequence v_n^δ be defined by

$$v_n^\delta := U_{\delta,\varepsilon_n} v_{n-1}^\delta \ ,$$

with $\varepsilon_n := \nu_\delta^n$ and $f_n := g(\{f \mid \| U_\delta v_{n-1}^\delta - L_\delta^f v_{n-1}^\delta \| \le \varepsilon_n\})$, then

(i) $\qquad \lim_{n\to\infty} v_n^\delta = V^* \quad$ (in μ-norm) ,

(ii) $$\lim_{n\to\infty} V_{f_n} = V^* \quad \text{(in } \mu\text{-norm)} .$$

PROOF. The proof is completely analogous to the proof of lemma 4.3.9. □

REMARK 5.2.4.

(i) From the preceding lemmas it follows that for any nonzero transition memoryless stopping time $\delta \in \Delta$ the determination of the optimal return vector V^* can be done by successive approximation of V^* with a sequence v_n^δ as described in lemma 5.2.5-5.2.7. Moreover, for each $\varepsilon > 0$ the policy f_n is ε-optimal for n sufficiently large. So in fact each nonzero transition memoryless stopping time produces a *policy improvement* procedure for solving the Markov decision problem.

(ii) For finite state space Markov decision processes with a bounded reward structure and nonrandomized stopping times, many of the results given in this chapter may be found in Wessels [74] and van Nunen and Wessels [55].

CHAPTER 6

VALUE ORIENTED SUCCESSIVE APPROXIMATION

In the previous chapter we have developed policy improvement procedures for Markov decision processes. Here we will deal mainly with policy improvement value determination procedures. Procedures of this type require extra computational effort, in each iteration step, in order to find better estimates for the actual value vector V_{f_n}.

In section 6.1 we will show that each transition memoryless stopping time $\delta \in \Delta$ (so in fact each policy improvement procedure) generates a whole set of policy improvement value determination procedures. In section 6.2 some aspects of the value oriented methods will be discussed and connections with results of other authors will be given.

From now on we will restrict the considerations to transition memoryless nonzero stopping times. This restriction is made since stationary (Markov) strategies are in fact the relevant decision rules, as is proved in theorem 4.3.4, and transition memoryless stopping times are the only stopping times for which a restriction to stationary strategies is always allowed (see section 5.3). In spite of this restriction, the existing policy improvement procedures, as introduced by Howard [35], Hastings [26], Reetz [60], and the present author [54], are contained in the set of policy improvement procedures generated by transition memoryless stopping times.

6.1. *Policy improvement value determination procedures*

For each $\delta \in \Delta$ that is transition memoryless and nonzero, a set of value oriented procedures will be based on an extension of the contraction mapping U_δ of V.

DEFINITION 6.1.1. The set of all transition memoryless nonzero stopping times is denoted by Δ'.

DEFINITION 6.1.2. For $\varepsilon > 0$, $\delta \in \Delta'$, and $\lambda \in \mathbb{N}$ we define the mapping $U_{\delta,\varepsilon}^{(\lambda)}$ of V by

$$U_{\delta,\varepsilon}^{(\lambda)} v := (L_\delta^f)^\lambda v$$

with $f := g(\{f \mid \|U_\delta v - L_\delta^f v\| < \varepsilon\})$.
From theorem 5.2.2 it follows that $U_{\delta,\varepsilon}^{(\lambda)}$ is well defined for $\varepsilon > 0$.
We define the mapping $U_{\delta,\varepsilon}^{(\infty)}$ of V by

$$U_{\delta,\varepsilon}^{(\infty)} v := \lim_{\lambda\to\infty} U_{\delta,\varepsilon}^{(\lambda)} v = V_f \ .$$

If $\forall_{v\in V}\ \exists_{f\in F}\ U_\delta v = L_\delta^f v$, then ε may be zero. In that case we define $U_{\delta,0}^{(\lambda)} v := (L_\delta^f)^\lambda v$ and $U_{\delta,0}^{(\infty)} v := \lim_{\lambda\to\infty} U_{\delta,0}^{(\lambda)} v = V_f$ with $f := g(\{f \mid U_\delta v = L_\delta^f v\})$.

LEMMA 6.1.1. For $\lambda \in \mathbb{N} \cup \{\infty\}$, and $\delta \in \Delta'$, $U_{\delta,\varepsilon}^{(\lambda)}$ maps V into V.

LEMMA 6.1.2. Let $\delta \in \Delta'$, let $\varepsilon > 0$ and $\lambda \in \mathbb{N} \cup \{\infty\}$, then

(i) $U_{\delta,\varepsilon}^{(\lambda)}$ is not necessarily ε-monotone.

(ii) $U_{\delta,\varepsilon}^{(\lambda)}$ is not necessarily ε-contracting.

PROOF. The statements follow directly from the following example, where the exact value of ε is irrelevant, since only finitely many Markov policies are available. We choose $\varepsilon = 1$. Let $\delta \in \Delta'$ correspond to G_1, $S\backslash\{0\} := \{1,2\}$, $\mu(1) := \mu(2) := 1$, $A(1) := A(2) := \{1,2\}$, $p^1(1,1) := p^1(2,1) := p^2(2,2) := 0$ $p^2(1,2) := 0$, $p^1(1,2) := p^1(2,2) := p^2(2,1) := p^2(1,1) := 0.99$, $r(1,1) := r(2,1) := 10$, $r(1,2) := r(2,2) := 0$, $v_1(1) := v_1(2) := 0$, $v_2(1) := 100$, $v_2(2) := 10$. Now it is easily verified that

$$B_1 := \{f \mid \|U_\delta v_1 - L_\delta^f v_1\| \le 1\} = \{f = \tbinom{1}{1}\}$$

and

$$B_2 := \{f \mid \|U_\delta v_2 - L_\delta^f v_2\| \le 1\} = \{f = \tbinom{2}{2}\} \ .$$

However, for $\lambda \to \infty$ we have

$$U_{\delta,1}^{(\lambda)} v_1 \to \tbinom{1000}{1000}, \quad U_{\delta,1}^{(\lambda)} v_2 \to \tbinom{0}{0} \ . \qquad \square$$

From the foregoing lemma it may seem impossible to use the mapping $U_{\delta,\varepsilon}^{(\lambda)}$ in a similar way as we have used U_δ and $U_{\delta,\varepsilon}$. However, the structure of Markov decision processes enables us to base on $U_{\delta,\varepsilon}^{(\lambda)}$ approximation procedures for solving the Markov decision process. This will be proved in the sequel of this section. We will first assume that $U_{\delta,0}$ is defined. In practical situation this assumption is often satisfied.

LEMMA 6.1.3. Let $\delta \in \Delta'$ and suppose $U_{\delta,0}$ is defined. Let $\lambda \in \mathbb{N} \cup \{\infty\}$ and let $v_0^{\delta\lambda}$ be such that $U_\delta v_0^{\delta\lambda} \geq v_0^{\delta\lambda}$. Let the sequence $v_n^{\delta\lambda}$ be defined by $v_n^{\delta\lambda} := U_{\delta,0}^{(\lambda)} v_{n-1}^{\delta\lambda}$, then

(i) $$v_{n-1}^{\delta\lambda} \leq v_n^{\delta\lambda} \leq v_{f_n} \leq v^*$$

(ii) $$v_n^{\delta\lambda} \nearrow v^* \ , \ \| v_n^{\delta\lambda} - v^* \| \leq \nu_\delta^n \| v_0^{\delta\lambda} - v^* \|$$

where $f_n := g(\{f \mid U_\delta v_{n-1}^{\delta\lambda} = L_\delta^f v_{n-1}^{\delta\lambda}\})$.

PROOF. Since $U_\delta v_0^{\delta\lambda} \geq v_0^{\delta\lambda}$ we have

$$L_\delta^{f_1} v_0^{\delta\lambda} \geq v_0^{\delta\lambda} \ .$$

Since $L_\delta^{f_1}$ is a monotone contraction on V

$$v_0^{\delta\lambda} \leq L_\delta^{f_1} v_0^{\delta\lambda} \leq (L_\delta^{f_1})^2 v_0^{\delta\lambda} \leq \ldots \leq (L_\delta^{f_1})^\lambda v_0^{\delta\lambda} = v_1^{\delta\lambda} \ .$$

Moreover, it follows from the monotonicity of $L_\delta^{f_1}$ that

$$v_{f_1} := \lim_{n\to\infty} (L_\delta^{f_1})^n v_0^{\delta\lambda} \geq U_{\delta,0}^{(\lambda)} v_0^{\delta\lambda} = v_1^{\delta\lambda} \ .$$

Now, since $U_\delta v_1^{\delta\lambda} = L_\delta^{f_2} v_1^{\delta\lambda} \geq L_\delta^{f_1} v_1^{\delta\lambda} \geq v_1^{\delta\lambda}$ it is similarly verified that $v_2^{\delta\lambda} \geq v_1^{\delta\lambda}$ and $v_2^{\delta\lambda} \geq U_\delta^2 v_0^{\delta\lambda}$. The proof proceeds further in an inductive way using the monotonicity and contraction properties of the mappings U_δ and L_δ^f and the fact that $U_\delta^n v_0^{\delta\lambda} \nearrow v^*$ (see lemma 5.2.5). □

REMARK 6.1.1. It is easy to find a $v_0^{\delta\lambda} \in V$ such that $U_\delta v_0^{\delta\lambda} \geq v_0^{\delta\lambda}$ (see remark 5.2.3).

LEMMA 6.1.4. Let $\delta \in \Delta'$ and suppose $U_{\delta,0}$ is defined. Let $\lambda \in \mathbb{N} \cup \{\infty\}$ and $v_0^{\delta\lambda} \in V$. Let the sequence $v_n^{\delta\lambda}$ be defined by

$$v_n^{\delta\lambda} := U_{\delta,0}^{(\lambda)} v_{n-1}^{\delta\lambda}$$

then

(i) $$\lim_{n\to\infty} v_n^{\delta\lambda} = V^* \quad \text{(in } \mu\text{-norm)}$$

(ii) $$\lim_{n\to\infty} V_{f_n} = V^* \quad \text{(in } \mu\text{-norm)} ,$$

where $f_n := g(\{f \mid U_\delta v_{n-1}^{\delta\lambda} = L_\delta^f v_{n-1}^{\delta\lambda}\})$.

PROOF. We will prove the lemma $\lambda \in \mathbb{N}$ since the proof of the lemma for $\lambda = \infty$ is a direct consequence of the foregoing lemma. Namely, $U_{\delta,0}^{(\infty)} v_0^{\delta\infty} = V_{f_1}$. This implies $U_{\delta,0} V_{f_1} \ge L_\delta^{f_1} V_{f_1} = V_{f_1}$. So after the first iteration step the conditions of the foregoing lemma are satisfied.

Proof of part (i) of the lemma: First, we remark that

$$v_n^{\delta\lambda} \le V^* + \nu_\delta^{n\lambda} \| v_0^{\delta\lambda} - V^* \| \cdot \mu ,$$

as follows from

$$v_n^{\delta\lambda} - V^* \le (U_\delta)^{n\lambda} v_0^{\delta\lambda} - U_\delta^{n\lambda} V^* \le \nu_\delta^{n\lambda} \| v_0^{\delta\lambda} - V^* \| \cdot \mu .$$

The proof of $v_n^{\delta\lambda} \ge V^* - \varepsilon\mu$ for n sufficiently large is more complicated.

We will first prove

(6.1.1) $$\forall_{\varepsilon>0}\ \exists_{N\in\mathbb{N}}\ \forall_{n\ge N}\ [U_\delta v_n^{\delta\lambda} - v_n^{\delta\lambda} \ge -\varepsilon\mu] .$$

Let $M := \| U_\delta v_0^{\delta\lambda} - v_0^{\delta\lambda} \|$. Then

$$U_\delta v_0^{\delta\lambda} - v_0^{\delta\lambda} \ge -M\mu$$

and

$$L_\delta^{f_1} (U_\delta v_0^{\delta\lambda}) - L_\delta^{f_1} v_0^{\delta\lambda} \ge -(\rho_\delta^{f_1}) M\mu \ge -\nu_\delta M\mu .$$

Similarly we find

$$(L_\delta^{f_1}) v_1^{\delta\lambda} - (L_\delta^{f_1})^\lambda v_0^{\delta\lambda} \ge -\nu_\delta^\lambda M\mu \ .$$

So

$$U_\delta v_1^{\delta\lambda} - v_1^{\delta\lambda} \ge L_\delta^{f_1} v_1^{\delta\lambda} - v_1^{\delta\lambda} \ge -\nu_\delta^\lambda M\mu \ .$$

By induction it follows that

$$(6.1.2) \qquad U_\delta v_n^{\delta\lambda} - v_n^{\delta\lambda} \ge -\nu_\delta^{n\lambda} M\mu \ .$$

This proves formula (6.1.1).

Let $\varepsilon > 0$, let N be such that $(\lambda - 1)\nu^{N\lambda}(1 - \nu_\delta^\lambda)^{-1}M < \varepsilon$. Then (6.1.2) with $n = N$ implies

$$L_\delta^{f_{N+1}} U_\delta v_N^{\delta\lambda} \ge U_\delta v_N^{\delta\lambda} - \nu_\delta^{N\lambda} M\mu \ge v_N^{\delta\lambda} - 2\nu_\delta^{N\lambda} M\mu \ .$$

Similarly we find

$$v_{N+1}^{\delta\lambda} = (L_\delta^{f_{N+1}})^\lambda v_N^{\delta\lambda} \ge (L_\delta^{f_{N+1}})^{\lambda-1}(v_N^{\delta\lambda} - \nu_\delta^{N\lambda} M\mu) \ge U_\delta v_N^{\delta\lambda} - (\lambda - 1)\nu_\delta^{N\lambda} M\mu.$$

This implies

$$(6.1.3) \qquad U_\delta v_{N+1}^{\delta\lambda} \ge U_\delta^2 v_N^{\delta\lambda} - \nu_\delta(\lambda - 1)\nu_\delta^{N\lambda} M\mu \ .$$

From formula (6.1.2) with $n = N + 1$ we have

$$U_\delta v_{N+1}^{\delta\lambda} \ge v_{N+1}^{\delta\lambda} - \nu_\delta^{(N+1)\lambda} M\mu$$

which yields

$$v_{N+2}^{\delta\lambda} \ge U_\delta v_{N+1}^{\delta\lambda} - (\lambda - 1)\nu_\delta^{(N+1)\lambda} M\mu \ .$$

Substituting the right hand side of formula (6.1.3) in the above formula we get

$$v_{N+2}^{\delta\lambda} \ge U_\delta^2 v_N^{\delta\lambda} - (\lambda - 1)\nu_\delta^{N\lambda+1} M\mu - (\lambda - 1)\nu_\delta^{(N+1)\lambda} M\mu \ .$$

So

$$U_\delta v_{N+2}^{\delta\lambda} \geq U_\delta^3 v_N^{\delta\lambda} - (\lambda - 1)\nu_\delta^{N\lambda+2} M\mu - (\lambda - 1)\nu_\delta^{(N+1)\lambda+1} M\mu \ .$$

However, from formula (6.1.2) we have

$$U_\delta v_{N+2}^{\delta\lambda} \geq v_{N+2}^{\delta\lambda} - \nu_\delta^{(N+2)\lambda} M\mu \ .$$

In a similar way as for $N+1$, $N+2$ this yields

$$v_{N+3}^{\delta\lambda} \geq U_\delta v_{N+2}^{\delta\lambda} - (\lambda - 1)\nu_\delta^{(N+2)\lambda} M\mu$$

$$v_{N+3}^{\delta\lambda} \geq U_\delta^3 v_N^{\delta\lambda} - (\lambda - 1)[\nu_\delta^{(N+2)\lambda} + \nu_\delta^{(N+1)\lambda+1} + \nu_\delta^{N\lambda+2}] M\mu \ .$$

In general

$$v_{N+k}^{\delta\lambda} \geq U_\delta^k v_N^{\delta\lambda} - (\lambda - 1)[\nu_\delta^{(N+k-1)\lambda} + \nu_\delta^{(N+k-2)\lambda+1} + \ldots + \nu_\delta^{N\lambda+k-1}] M\mu$$

$$\geq U_\delta^k v_N^{\delta\lambda} - (\lambda - 1)\nu_\delta^{N\lambda}(1 - (\nu_\delta^\lambda)^{k-1})(1 - \nu_\delta^\lambda)^{-1} M\mu \ .$$

So for $k \to \infty$ we get

$$\liminf_{k\to\infty} v_{N+k}^{\delta\lambda} \geq V^* - (\lambda - 1)\nu^{N\lambda}(1 - \nu_\delta^\lambda)^{-1} M\mu \geq V^* - \varepsilon\mu \ .$$

This completes the proof of part (i) of the lemma.
Note that it follows from this proof that

$$\forall_{\varepsilon>0} \ \exists_{N\in\mathbb{N}} \ \forall_{n>N} \ \| U_\delta v_n^{\delta\lambda} - v_n^{\delta\lambda} \| < \varepsilon \ .$$

(ii) For $\varepsilon > 0$ and n sufficiently large we have

$$\| U_\delta v_n^{\delta\lambda} - v_n^{\delta\lambda} \| < \tfrac{1}{2}\varepsilon(1 - \nu_\delta) \text{ and } \| v_n^{\delta\lambda} - V^* \| < \tfrac{1}{2}\varepsilon \ .$$

So for all $k \in \mathbb{N}$

$$\|(L_\delta^{f_n})^k v_n^{\delta\lambda} - V^*\| \leq \sum_{\ell=0}^{k-1} \nu_\delta^\ell \|L_\delta^{f_n} v_n^{\delta\lambda} - v_n^{\delta\lambda}\| + \|v_n^{\delta\lambda} - V^*\| < \varepsilon .$$

This results in

$$\|V_{f_n} - V^*\| < 2\varepsilon ,$$

for n sufficiently large. □

Until now we have proved the convergence of sequences $v_n^{\delta\lambda}$ to V^* under the restrictive condition that $U_{\delta,0}$ is defined. In the sequel of this section this assumption will be replaced by weaker assumptions.

THEOREM 6.1.1. Let $\delta \in \Delta'$. Let $\lambda \in \mathbb{N} \cup \{\infty\}$ and let $v_0^{\delta\lambda} \in V$. Let the sequence $v_n^{\delta\lambda}$ be defined by

$$v_n^{\delta\lambda} := U_{\delta,\varepsilon_n}^{(\lambda)} v_{n-1}^{\delta\lambda} ,$$

with $\varepsilon_n := \nu_\delta^{n\lambda}$. Then

(i) $\lim_{n\to\infty} v_n^{\delta\lambda} = V^*$ (in μ-norm)

(ii) $\lim_{n\to\infty} V_{f_n} = V^*$ (in μ-norm)

where $f_n := g(\{f \mid \|U_\delta v_{n-1}^{\delta\lambda} - L_\delta^f v_{n-1}^{\delta\lambda}\| < \varepsilon_n\})$.

PROOF. The proof proceeds along the same lines as the proof of lemma 6.1.4 with the exception of the first part where a slight modification is required. Define

$$M := \|L_\delta^{f_1} v_0^{\delta\lambda} - v_0^{\delta\lambda}\|$$

then

$$(L_\delta^{f_1})^\lambda v_0^{\delta\lambda} - (L_\delta^{f_1})^{\lambda-1} v_0^{\delta\lambda} \geq -\nu_\delta^{\lambda-1} M\mu ,$$

as follows from the monotonicity and contraction property of $L_\delta^{f_1}$. Moreover

$$U_\delta v_1^{\delta\lambda} - v_1^{\delta\lambda} \geq L_\delta^{f_2} v_1^{\delta\lambda} - v_1^{\delta\lambda} \geq -\nu_\delta^\lambda \mu + L_\delta^{f_1} v_1^{\delta\lambda} - v_1^{\delta\lambda} \geq -\nu_\delta^\lambda (M+1)\mu \ .$$

In general we find inductively

$$U_\delta v_n^{\delta\lambda} - v_n^{\delta\lambda} \geq L_\delta^{f_{n+1}} v_n^{\delta\lambda} - v_n^{\delta\lambda} \geq -\nu^{n\lambda}(M + n)\mu \ .$$

The final part of the theorem proceeds in a similar way as the proof of lemma 6.1.4. □

LEMMA 6.1.5. Let $\varepsilon > 0$, let $\delta \in \Delta'$. Let $\lambda \in \mathbb{N} \cup \{\infty\}$ and let $v_0^{\delta\lambda} \in V$ such that $v_0^{\delta\lambda} \leq U_\delta v_0^{\delta\lambda} - \varepsilon\mu$. Let the sequence $v_n^{\delta\lambda}$ be defined by

$$v_n^{\delta\lambda} := (L_\delta^{f_n})^\lambda v_{n-1}^{\delta\lambda}$$

where $f_n := g(B_n)$ with $B_n := \{f \mid L_\delta^f v_{n-1}^{\delta\lambda} \geq \max\{v_{n-1}^{\delta\lambda}, U_\delta v_{n-1}^{\delta\lambda} - \varepsilon(1 - \nu_\delta)\mu\}$.
Then

(i) $\quad \| v_n^{\delta\lambda} - v^* \| < \varepsilon$, for n sufficiently large ,

(ii) $\quad v_{n-1}^{\delta\lambda} \leq v_n^{\delta\lambda} \leq v_{f_n} \leq v^*$.

PROOF. The proof may be found straightforwardly by using similar arguments as in lemma 6.1.3 and the foregoing theorem (6.1.1), see also the lemmas 4.3.8 and 5.2.6. □

REMARK 6.1.2. As mentioned earlier it is not difficult to find a starting vector $v_0^{\delta\lambda} \in V$ such that $v_0^{\delta\lambda} \leq U_\delta v_0^{\delta\lambda} - \varepsilon\mu$.

THEOREM 6.1.2. Let $\varepsilon > 0$, let $\delta \in \Delta'$. Let $\lambda \in \mathbb{N} \cup \{\infty\}$, and let $v_0^{\delta\lambda} \in V$. Let the sequence $v_n^{\delta\lambda}$ be defined by

$$v_n^{\delta\lambda} := U_{\delta,\eta}^{(\lambda)} v_{n-1}^{\delta\lambda} \ ,$$

with $\eta := \frac{1}{2}\varepsilon(1-\nu_\delta)^3$. Then

(i) $\| v_n^{\delta\lambda} - V^* \| < 2\varepsilon$ for n sufficiently large ,

(ii) $\| v_{f_n} - V^* \| < 3\varepsilon$ for n sufficiently large .

PROOF. The proof proceeds in a similar way as the proof of lemma 6.1.4. Also in this case it will be clear that

$$v_n^{\delta\lambda} - V^* \le (U_\delta)^{n\lambda} v_0^{\delta\lambda} - V^* \le \nu_\delta^{n\lambda} \| v_0^{\delta\lambda} - V^* \| \cdot \mu \ .$$

So

$$v_n^{\delta\lambda} \le V^* + \nu_\delta^{n\lambda} \| v_0^{\delta\lambda} - V^* \| \cdot \mu \ .$$

Let $M := \| U_\delta v_0^{\delta\lambda} - v_0^{\delta\lambda} \| + \eta$. Then

$$L_\delta^{f_1} v_0^{\delta\lambda} - v_0^{\delta\lambda} \ge - M\mu$$

and

$$(L_\delta^{f_1})^\lambda v_0^{\delta\lambda} - (L_\delta^{f_1})^{\lambda-1} v_0^{\delta\lambda} \ge - \nu_\delta^{\lambda-1} M\mu$$

$$L_\delta^{f_2} v_1^{\delta\lambda} - v_1^{\delta\lambda} \ge U_\delta v_1^{\delta\lambda} - v_1^{\delta\lambda} - \eta\mu$$

$$\ge L_\delta^{f_1} v_1^{\delta\lambda} - v_1^{\delta\lambda} - \eta\mu \ge -\nu_\delta^\lambda M\mu - \eta\mu \ .$$

Inductively we find

$$L_\delta^{f_n} v_{n-1}^{\delta\lambda} - v_{n-1}^{\delta\lambda} \ge -\nu_\delta^{\lambda(n-1)} M\mu - \eta(1 + \nu_\delta^\lambda + \ldots + \nu_\delta^{\lambda(n-2)})\mu$$

$$\ge -\nu_\delta^{\lambda(n-1)} M\mu - \tfrac{1}{2}\varepsilon\mu(1 - \nu_\delta)^2 \ .$$

So for n sufficiently large

(6.1.4) $$U_\delta v_n^{\delta\lambda} - v_n^{\delta\lambda} \ge U_{\delta,\eta} v_n^{\delta\lambda} - v_n^{\delta\lambda} \ge -\varepsilon(1-\nu_\delta)^2\mu \ .$$

So from formula (6.1.4) with n = N we have

$$(L_\delta^{f_{N+1}})^2 v_N^{\delta\lambda} \ge L_\delta^{f_{N+1}} v_N^{\delta\lambda} - \nu_\delta(1-\nu_\delta)^2\varepsilon\mu \ge v_N^{\delta\lambda} - \nu_\delta(1-\nu_\delta)^2\varepsilon\mu - (1-\nu_\delta)^2\varepsilon\mu .$$

Iterating in this way we get

$$v_{N+1}^{\delta\lambda} \ge L_\delta^{f_{N+1}} v_N^{\delta\lambda} - \nu_\delta[\nu_\delta^{\lambda-1}(1-\nu_\delta)^2 + \nu_\delta^{\lambda-2}(1-\nu_\delta)^2 + \dots + (1-\nu_\delta)^2]\varepsilon\mu .$$

Now since

$$L_\delta^{f_{N+1}} v_N^{\delta\lambda} \ge U_\delta v_N^{\delta\lambda} - \tfrac{1}{2}(1-\nu_\delta)^3\varepsilon\mu$$

we get

$$\begin{aligned} v_{N+1}^{\delta\lambda} &\ge U_\delta v_N^{\delta\lambda} - [\nu_\delta(1-\nu_\delta^{\lambda-1}) + \tfrac{1}{2}(1-\nu_\delta)^2](1-\nu_\delta)\varepsilon\mu \\ &\ge U_\delta v_N^{\delta\lambda} - (1-\nu_\delta)\varepsilon\mu . \end{aligned}$$

This yields by applying U_δ

$$(6.1.5) \qquad U_\delta v_{N+1}^{\delta\lambda} \ge U_\delta^2 v_N^{\delta\lambda} - \nu_\delta(1-\nu_\delta)\varepsilon\mu .$$

From formula (6.1.4) with n = N + 1 we have

$$(L_\delta^{f_{N+2}})^2 v_{N+1}^{\delta\lambda} \ge L_\delta^{f_{N+2}} v_{N+1}^{\delta\lambda} - \nu_\delta(1-\nu_\delta)^2\varepsilon\mu \ge v_{N+1}^{\delta\lambda} - (1+\nu_\delta)(1-\nu_\delta)^2\varepsilon\mu$$

and

$$v_{N+2}^{\delta\lambda} \ge L_\delta^{f_{N+2}} v_{N+1}^{\delta\lambda} - \nu_\delta[1 + \nu_\delta + \dots + \nu_\delta^{\lambda-1}](1-\nu_\delta)^2\varepsilon\mu .$$

In a similar way as for n = N this yields

$$v_{N+2}^{\delta\lambda} \ge U_\delta v_{N+1}^{\delta\lambda} - (1-\nu_\delta)\varepsilon\mu .$$

So

$$U_\delta v_{N+2}^{\delta\lambda} \ge U_\delta^2 v_{N+1}^{\delta\lambda} - \nu_\delta(1-\nu_\delta)\varepsilon\mu .$$

By using (6.1.5) we find

$$U_\delta v^{\delta\lambda}_{N+2} \ge U^3_\delta v^{\delta\lambda}_N - \nu^2_\delta (1-\nu_\delta)\varepsilon\mu - \nu_\delta (1-\nu_\delta)\varepsilon\mu$$

and

$$L_\delta^{f_{N+3}} v^{\delta\lambda}_{N+2} \ge U^3_\delta v^{\delta\lambda}_N - [\nu^2_\delta + \nu_\delta + 1](1-\nu_\delta)\varepsilon\mu \ .$$

So for k → ∞ we have

$$\liminf_{k\to\infty} v^{\delta\lambda}_{N+k} \ge V^* - \varepsilon\mu \ .$$

This proves part (i) of the lemma.
The final part follows in a similar way as the proof of part (ii) of lemma 6.1.4. □

REMARK 6.1.3.

(i) Note that the foregoing theorem (for $\lambda = 1$) states that, for any $\delta \in \Delta'$, $\varepsilon > 0$, and $v^\delta_0 \in \mathcal{V}$ we have for $v^\delta_n := U_{\delta,\eta} v^\delta_{n-1}$

(a) $\| v^\delta_n - V^* \| < 2\varepsilon$

(b) $\| v_{f_n} - V^* \| \le 3\varepsilon$,

for n sufficiently large.

(ii) From the foregoing theorems and lemmas it will be clear that for any $\delta \in \Delta'$ and for any $\lambda \in \mathbb{N} \cup \{\infty\}$ the optimal return vector V^* may be approximated by a sequence $v^{\delta\lambda}_n$ as described in the theorems 6.1.1 and 6.1.2 and the lemmas 6.1.3-6.1.5. Moreover, the policy f_n becomes (ε-) optimal for n sufficiently large. So in fact each $\delta \in \Delta'$ and each $\lambda \in \mathbb{N} \cup \{\infty\}$ produce a policy improvement value determination procedure for solving the Markov decision problem.

(iii) Note that for $\lambda = 1$ the sequences $v^{\delta\lambda}_n$ coincide with the corresponding sequences v^δ_n discussed in chapter 5.

6.2. *Some remarks on the value oriented methods*

In this section we shall try to illustrate what really happens when we use a value oriented successive approximation method. Furthermore, connections with existing solution techniques for Markov decision problems will be given, We shall restrict ourselves here to the situation that $U_{\delta,0}$ is defined. In the preceding sections we have seen how the results can be extended if this restriction is released.

In a very simple situation, finite state and decision space discounted Markov decision processes, it has been illustrated by the author [53] that for δ corresponding to G_1 the mapping $U_{\delta,0}^{(\lambda)}$ produced better estimates for V_{f_n} compared to the estimates of the mapping U_δ. This might be expected since in each iteration step extra computational effort is spent to find better estimates for V_{f_n} (the value vector of the actual policy f_n).

For $v \in V$ and $\delta \in \Delta'$, let $f := g(\{f \mid L_\delta^f v = U_\delta v\})$. Then, since $L_\delta^f V_f = V_f$ we have

$$\| U_{\delta,0}^{(\lambda)} v - V_f \| = \| (L_\delta^f)^\lambda v - (L_\delta^f)^\lambda V_f \| \le (\rho_\delta^f)^\lambda \| v - V_f \| \ .$$

This implies that in general $U_\delta^{(\lambda)} v$ is a better estimate for the total expected return V_f under the actual policy f than $U_\delta v$.

Choose $v_0 \in V$ such that $U_\delta v_0 \ge v_0$. It is easily verified that the sequences $v_n^{\delta\lambda} := U_{\delta,0}^{(\lambda)} v_{n-1}^{\delta\lambda}$ and $v_n^\delta := U_{\delta,0} v_{n-1}^\delta$ both started with the chosen v_0 satisfy

$$v_n^\delta \le v_n^{\delta\lambda} \le v^* \ .$$

For some examples we refer to van Nunen [53]. In [53] we have shown for finite state space finite decision space discounted Markov decision processes that, for $\delta \in \Delta$ corresponding to G_1 the value vector $v_n^{\delta\lambda} := U_{\delta,0}^{(\lambda)} v_{n-1}^{\delta\lambda}$ may be interpreted as the total expected reward over $n\lambda$-time periods if the strategy

$$\overbrace{f_n, f_n, \dots, f_n}^{\lambda \times}, \overbrace{f_{n-1}, \dots, f_{n-1}}^{\lambda \times}, \dots, \overbrace{f_1, \dots, f_1}^{\lambda \times}$$

is used and the terminal value vector equals $v_0^{\delta\lambda}$. A similar interpretation may be given if more complicated stopping times are used.

If δ is nonzero and $\lambda = \infty$, then the algorithms as described in theorem 6.1.1 are clearly of the policy iteration type, see e.g. Howard [35], Mine and Osaki [50]. This means that in each iteration step the value vectors V_{f_n} (total expected return over an infinite time horizon) of the actual policy f_n, is computed exactly. The choice of $\delta \in \Delta'$ only determines the way of looking for possible improvement of policies. For any $\delta \in \Delta'$ in the case $\lambda = \infty$, each iteration step brings a strict improvement of the values V_{f_n} until the optimum V^* is reached. This occurs in a finite number of steps if only finitely many Markov policies are available. For $\lambda = \infty$, in the first iteration step, the value vector V_{f_1} is computed exactly. So after one iteration step the condition $U_{\delta,0}v \geq v$, as required in lemma 6.1.3, is satisfied; from then on the convergence will thus be monotone.

Howard's policy iteration algorithm [35], [50], for finite state space finite decision space discounted Markov decision processes equals the example δ corresponding to G_1, $\lambda \equiv \infty$, $\mu(i) = 1$ and $b(i) = 0$ for $i \in S\backslash\{0\}$.

If in this situation δ is replaced by the stopping time corresponding to the goahead set G_H we get Hasting's modified (Gauss-Seidel) policy iteration algorithm [26]. As mentioned, for $\lambda = 1$ and δ corresponding to G_1 the successive approximation methods yield the standard dynamic programming method as described by e.g. Bellman [3], Blackwell [4], [5], MacQueen [46].

In this chapter the procedures have been defined for a fixed number $\lambda \in \mathbb{N} \cup \{\infty\}$. However, it is not essential that λ is fixed for all iteration steps. The value of λ may depend on the number of the iteration step and even on specific aspects of the actual iteration process.

For numerical experiences with a value oriented method we refer to van Nunen [53].

CHAPTER 7

UPPER BOUNDS, LOWER BOUNDS AND SUBOPTIMALITY

In the previous chapters we have proved the convergence of the sequences v_n, v_n^{δ}, $v_n^{\delta\lambda}$, as defined in the preceding chapters to the optimal return vector V^*. We have also proved that the Markov strategy f_n^{∞} found in the n-th iteration step of the actual algorithm is ε-optimal for n sufficiently large. In general the convergence proceeds at a geometric rate, since we have required $\| P^f \| \le \rho_* < 1$ for all $f \in F$ (see lemma 5.2.5). The question now arising is whether one is able to construct better estimates for V_{f_n} and V^* at an earlier stage of the iteration process. We will show that this can be done by extrapolation based on the differences $v_n^{\delta} - v_{n-1}^{\delta}$.
Upper and lower bounds for the value vectors V_{f_n} and V^* enable us to incorporate a test for the suboptimality of policies.
Section 7.1 will be devoted to the construction of upper and lower bounds. Section 7.2 will deal with the concept of suboptimality of policies. In the final section 7.3 we restrict the considerations to finite state space Markov decision processes. The concepts of bounds and suboptimality will be considered with respect to these processes. Several numerical aspects will be discussed; relations with the work of other authors will be given.

7.1. *Upper bounds and lower bounds for* V_{f_n} *and* V^*

We will treat the concept of bounds in a similar way as Porteus [58], [59] did in the finite state case with δ corresponding to G_1, b = 0 and the unweighted supremum norm. We also refer to van der Wal [71] who treated the concept of bounds in a similar way for finite state space Markov games with b = 0 and the unweighted supremum norm.

DEFINITION 7.1.1. Let v_n^{δ} be a sequences of vectors in V, then we define α_n^{δ}, β_n^{δ} for n = 1,2,... by

$$\alpha_n^{\delta} := \inf_{i \in S\setminus\{0\}} [(v_n^{\delta}(i) - v_{n-1}^{\delta}(i))\mu^{-1}(i)] ,$$

$$\beta_n^{\delta} := \sup_{i \in S\setminus\{0\}} [(v_n^{\delta}(i) - v_{n-1}^{\delta}(i))\mu^{-1}(i)] .$$

Moreover, we define z_n^δ, y_n^δ by

$$z_n^\delta := \begin{cases} \inf\limits_{f\in F}\ \inf\limits_{i\neq 0}\ (\mu^{-1}(i) \sum\limits_{j\in S} p_\delta^f(i,j)\mu(j)) & \text{if } \alpha_n^\delta \geq 0 \\ \nu_\delta & \text{if } \alpha_n^\delta < 0 \end{cases}$$

$$y_n^\delta := \begin{cases} \nu_\delta & \text{if } \beta_n^\delta \geq 0 \\ \inf\limits_{f\in F}\ \inf\limits_{i\neq 0}\ (\mu^{-1}(i) \sum\limits_{j\in S} p_\delta^f(i,j)\mu(j)) & \text{if } \beta_n^\delta < 0 \ . \end{cases}$$

LEMMA 7.1.1.

(i) $\qquad \forall_{f\in F}\ z_n^\delta\cdot\alpha_n^\delta\cdot\mu \leq L_\delta^f v_n^\delta - L_\delta^f v_{n-1}^\delta \leq y_n^\delta\cdot\beta_n^\delta\cdot\mu$

(ii) $\qquad (z_n^\delta)^N\cdot\alpha_n^\delta\cdot\mu \leq U_\delta^N v_n^\delta - U_\delta^N v_{n-1}^\delta \leq (y_n^\delta)^N\cdot\beta_n^\delta\cdot\mu\ , \qquad N\in\mathbb{N}$

(iii) $\qquad \forall_{\varepsilon>0}\ z_n^\delta\cdot\alpha_n^\delta\cdot\mu - \varepsilon\mu \leq U_{\delta,\varepsilon}v_n^\delta - U_{\delta,\varepsilon}v_{n-1}^\delta \leq y_n^\delta\cdot\beta_n^\delta\cdot\mu + \varepsilon\mu\ .$

PROOF. Since L_δ^f is a monotone mapping it is straightforwardly verified that

$$L_\delta^f v_n^\delta - L_\delta^f v_{n-1}^\delta = P_\delta^f(v_n^\delta - v_{n-1}^\delta) \leq y_n^\delta\cdot\beta_n^\delta\cdot\mu$$

$$L_\delta^f v_n^\delta - L_\delta^f v_{n-1}^\delta = P_\delta^f(v_n^\delta - v_{n-1}^\delta) \geq z_n^\delta\cdot\alpha_n^\delta\cdot\mu\ .$$

We will prove part (ii) only for N = 1. For general N the proof follows by induction. Choose an arbitrary $\varepsilon > 0$ and choose $f^1, f^2 \in F$ such that

$$L_\delta^{f_1} v_n^\delta \geq U_\delta v_n^\delta - \varepsilon\mu \text{ and } L_\delta^{f_2} v_{n-1}^\delta \geq U_\delta v_{n-1}^\delta - \varepsilon\mu\ .$$

Then

$$\begin{aligned} U_\delta v_n^\delta - U_\delta v_{n-1}^\delta &\leq L_\delta^{f_1} v_n^\delta + \varepsilon\mu - L_\delta^{f_1} v_{n-1}^\delta = P_\delta^{f_1}(v_n^\delta - v_{n-1}^\delta) + \varepsilon\mu \\ &\leq y_n^\delta\cdot\beta_n^\delta\cdot\mu + \varepsilon\mu \end{aligned}$$

and

$$U_\delta v_n^\delta - U_\delta v_{n-1}^\delta \ge L_\delta^{f_2} v_n^\delta - L_\delta^{f_2} v_{n-1}^\delta - \varepsilon\mu = P_\delta^{f_2}(v_n^\delta - v_{n-1}^\delta) - \varepsilon\mu$$

$$\ge z_n^\delta \cdot \alpha_n^\delta \cdot \mu - \varepsilon\mu .$$

Since ε was chosen arbitrarily this implies part (ii).
Part (iii) follows similarly by using the definition of $U_{\delta,\varepsilon}$. □

LEMMA 7.1.2. Let $\delta \in \Delta'$, $v_0^\delta \in V$ and suppose $U_{\delta,0}$ is defined. Let the sequence v_n^δ be defined as in lemma 5.2.5 then

(i) $\beta_n^\delta > 0 \Rightarrow \beta_{n-1}^\delta > 0$ and thus $\nu_\delta = y_n^\delta = y_{n-1}^\delta$

(ii) $\alpha_n^\delta < 0 \Rightarrow \alpha_{n-1}^\delta < 0$ and thus $\nu_\delta = z_n^\delta = z_{n-1}^\delta$.

PROOF.

(i)
$$0 < \beta_n^\delta = \sup_{i \neq 0} [(v_n^\delta(i) - v_{n-1}^\delta(i))\mu^{-1}(i)] =$$
$$= \sup_{i \neq 0} [(U_\delta v_{n-1}^\delta(i) - U_\delta v_{n-2}^\delta(i))\mu^{-1}(i)] \le$$
$$\le \sup_{i \neq 0} [(P^{f_n}(v_{n-1}^\delta - v_{n-2}^\delta)(i))\mu^{-1}(i)] \le \beta_{n-1}^\delta .$$

(ii)
$$0 > \alpha_n^\delta = \inf_{i \neq 0} [(v_n^\delta(i) - v_{n-1}^\delta(i))\mu^{-1}(i)] =$$
$$= \inf_{i \neq 0} [(U_\delta v_{n-1}^\delta(i) - U_\delta v_{n-2}^\delta(i))\mu^{-1}(i)] \ge$$
$$\ge \inf_{i \neq 0} [(P^{f_{n-1}}(v_{n-1}^\delta - v_{n-2}^\delta)(i))\mu^{-1}(i)] \ge \alpha_{n-1}^\delta .$$ □

COROLLARY 7.1.1. Let $\delta \in \Delta'$, $v_0^\delta \in V$ and suppose $U_{\delta,0}$ is defined. Let the sequence v_n^δ be defined as in lemma 5.2.5 then

(i) $$\sum_{k=1}^{\infty} (y_n^\delta)^k \cdot \beta_n^\delta \le (y_{n-1}^\delta) \sum_{k=1}^{\infty} (y_{n-1}^\delta)^k \cdot \beta_{n-1}^\delta$$

(ii) $$\sum_{k=1}^{\infty} (z_n^\delta)^k \cdot \alpha_n^\delta \geq (\alpha_{n-1}^\delta) \sum_{k=1}^{\infty} (z_{n-1}^\delta)^k \alpha_{n-1}^\delta \, .$$

PROOF. For $\beta_n^\delta > 0$ and $\alpha_n^\delta < 0$ the proof of part (i) and (ii) follows from the foregoing two lemmas.
The other cases are trivial. □

LEMMA 7.1.3. Let $\delta \in \Delta'$, let $v_0^\delta \in V$ and suppose $U_{\delta,0}$ is defined. Let the sequence v_n^δ be defined as in lemma 5.2.5.

(i) The sequence u_n^δ $(n \geq 1)$ defined by

$$u_n^\delta := v_n^\delta + y_n^\delta (1 - y_n^\delta)^{-1} \beta_n^\delta \cdot \mu$$

yields monotone nonincreasing upper bounds for V^* and thus for V_{f_n}. Moreover

$$u_n^\delta \downarrow v^* \, .$$

(ii) The sequence ℓ_n^δ $(n \geq 1)$ defined by

$$\ell_n^\delta := v_n^\delta + z_n^\delta (1 - z_n^\delta)^{-1} \alpha_n^\delta \cdot \mu$$

yields monotone nondecreasing lower bounds for V_{f_n} and thus for V^*. Moreover

$$\ell_n^\delta \uparrow v^* \, .$$

(iii) $$\| u_n^\delta - \ell_n^\delta \| \leq 2\nu_\delta^n \; (1 - \nu_\delta)^{-1} \| v_1^\delta - v_0^\delta \| \, .$$

PROOF.

$$U_\delta^N v_n^\delta = v_n^\delta + U_\delta^N v_n^\delta - U_\delta^{N-1} v_n^\delta + U_\delta^{N-1} v_n^\delta - \ldots + U_\delta v_n^\delta - v_n^\delta$$

$$\leq v_n^\delta + \sum_{k=1}^{N} (y_n^\delta)^k \beta_n^\delta \cdot \mu$$

see lemma 7.1.1(ii).

So for $N \to \infty$ we find

$$V^* \le v_n^\delta + y_n^\delta (1 - y_n^\delta)^{-1} \beta_n^\delta \mu \ .$$

The monotonicity of u_n^δ follows from

$$\begin{aligned} u_n^\delta - u_{n-1}^\delta &= v_n^\delta - v_{n-1}^\delta + y_n^\delta (1 - y_n^\delta)^{-1} \beta_n^\delta \mu - y_{n-1}^\delta (1 - y_{n-1}^\delta)^{-1} \beta_{n-1}^\delta \mu \\ &\le (y_{n-1}^\delta) \beta_{n-1}^\delta \mu + (y_{n-1}^\delta)^2 (1 - y_n^\delta)^{-1} \beta_{n-1}^\delta \mu + \\ &\qquad - y_{n-1}^\delta (1 - y_{n-1}^\delta)^{-1} \beta_{n-1}^\delta \mu \le 0 \ . \end{aligned}$$

The statements about the lower bounds follow similarly by considering $(L_\delta^{f_n})^N v_{n-1}^\delta$.

The convergence of u_n^δ and ℓ_n^δ to V^* follows from lemma 5.2.5.

Part (iii) follows by inspection. □

REMARK 7.1.1.

(i) Note that for δ chosen corresponding to G_1 the bounds may be used for the sequence v_n as defined in lemma 4.3.7.

(ii) If in the situation as described in lemma 7.1.3 the starting vector v_0^δ is chosen such that $U_\delta v_0^\delta \ge v_0^\delta$ then it is easily verified that $\alpha_n^\delta \ge 0$ for all $n \ge 1$. This implies $z_n^\delta = z_{n-1}^\delta$ and $y_n^\delta = y_{n-1}^\delta$ for all $n > 1$.

LEMMA 7.1.4. Let the sequence v_n^δ be defined as in lemma 5.2.6, then

(i) The sequence w_n^δ $(n \ge 1)$ defined by

$$w_1^\delta := u_1^\delta + (1 - y_1^\delta)^{-1} \varepsilon (1 - \nu_\delta) \mu$$

$$w_n^\delta(i) := \min\{w_{n-1}^\delta(i), u_n^\delta(i) + (1 - y_n^\delta)^{-1} \varepsilon (1 - \nu_\delta) \mu(i)\}, \ n > 1, \ i \in S \backslash \{0\}$$

where u_n^δ is defined as in lemma 7.1.3, yields monotone nonincreasing upper bounds for V^* and thus for V_{f_n} .

(ii) The sequence x_n^δ, $(n \geq 1)$ defined by

$$x_1^\delta := \ell_1^\delta$$

$$x_n^\delta(i) := \max\{x_{n-1}^\delta(i), \ell_n^\delta(i)\},\ n > 1,\ i \in S\backslash\{0\}$$

where ℓ_n^δ is defined as in lemma 7.1.3, yields monotone nondecreasing lower bounds for V^*.

PROOF. The proof proceeds along the same lines as the proof of the foregoing lemma.
We first remark that $v_n^\delta - v_{n-1}^\delta \geq 0$ for all n since

$$v_n^\delta(i) \geq \max\{v_{n-1}^\delta(i), U_\delta v_{n-1}^\delta(i) - \varepsilon(1-\nu_\delta)\mu(i)\}$$

so for all $n > 1$ $y_n^\delta = y_{n-1}^\delta = \nu_\delta$ and $z_n^\delta = z_{n-1}^\delta$.
Now consider

$$U_\delta^N v_n^\delta = v_n^\delta + \sum_{k=1}^{N} (U_\delta^k v_n^\delta - U_\delta^{k-1} v_n^\delta)$$

$$\leq v_n^\delta + (1-\nu_\delta^N)(1-\nu_\delta)^{-1} \sup_{i\neq 0} [\frac{U_\delta v_n^\delta(i) - v_n^\delta(i)}{\mu(i)}]\mu$$

$$\leq v_n^\delta + (1-\nu_\delta^N)(1-\nu_\delta)^{-1} \sup_{i\neq 0} [\frac{U_\delta v_n^\delta(i) - U_\delta v_{n-1}^\delta(i)}{\mu(i)} + \varepsilon(1-\nu_\delta)]\mu$$

$$\leq v_n^\delta + (1-\nu_\delta^N)(1-\nu_\delta)^{-1}[\beta_n^\delta y_n^\delta + \varepsilon(1-\nu_\delta)]\mu\ .$$

For $N \to \infty$ we obtain the requested result.

For the lower bounds we have

$$(L_\delta^{f_n})^N v_{n-1}^\delta = v_n^\delta + \sum_{k=2}^{N} ((L_\delta^{f_n})^k v_{n-1}^\delta - (L_\delta^{f_n})^{k-1} v_{n-1}^\delta)$$

$$\geq v_n^\delta + z_n^\delta(1 - (z_n^\delta)^{N-1})(1 - z_n^\delta)^{-1}\alpha_n^\delta \cdot \mu\ .$$

So for $N \to \infty$ we have

$$V_{f_n} \geq v_n^\delta + z_n^\delta(1 - z_n^\delta)^{-1}\alpha_n^\delta \cdot \mu = \ell_n^\delta .$$

The monotonicity of w_n^δ, x_n^δ follows from the definition of w_n^δ and x_n^δ whereas $\| w_n^\delta - x_n^\delta \| \leq 2\varepsilon(1-\nu_\delta)^{-1}$ follows from lemma 5.2.6. □

REMARK 7.1.2. From the foregoing proof it will be clear that ℓ_n^δ is a lower bound for V_{f_n}.

LEMMA 7.1.5. Let v_n^δ be defined as in lemma 5.2.7, then

(i) the sequence w_n^δ $(n \geq 1)$ defined by

$$w_1^\delta := u_1^\delta + \nu_\delta(1-y_1^\delta)^{-1}\mu$$

$$w_n^\delta(i) := \min\{w_{n-1}^\delta(i), u_n^\delta(i) + (\nu_\delta)^n(1-y_n^\delta)^{-1}\mu(i)\}, \ (n>1), \ i \in S\backslash\{0\}$$

where u_n^δ is defined as in lemma 7.1.3, yields monotone nonincreasing upper bounds for V^* and thus for V_{f_n}.

(ii) the sequence x_n^δ $(n \geq 1)$ defined by

$$x_1^\delta := \ell_1^\delta$$

$$x_n^\delta(i) := \max\{x_{n-1}^\delta(i), \ell_n^\delta(i)\}, \ n > 1, \ i \in S\backslash\{0\}$$

where ℓ_n^δ is defined as in lemma 7.1.3, yields monotone nondecreasing lower bounds for V^*.

(iii) $\| w_n^\delta - x_n^\delta \| \leq 2(\nu_\delta)^n(1-\nu_\delta)^{-1}(\| v_0^\delta - V^* \| + n)$.

PROOF. The proof of the first two parts proceed in a similar way as the proofs of the foregoing lemmas whereas the final part follows from lemma 5.2.7. □

REMARK 7.1.3.

(i) A lower bound for V_{f_n} may be found in a similar way as in lemma 7.1.4 (see also remark 7.1.2).

(ii) Note that if δ is chosen corresponding to G_1 then the bounds may be used for the sequences v_n and f_n as defined in lemma 4.3.8.

(iii) Upper and lower bounds for the sequences $v_n^{\delta\lambda}$ as defined in chapter 6 can be obtained in a similar way. We will illustrate this by giving the bounds for the sequences $v_n^{\delta\lambda}$ as defined in lemma 6.1.4, i.e.

$$u_n^{\delta\lambda} := v_n^{\delta\lambda} + (1 - y_n^{\delta\lambda})^{-1}\beta_n^{\delta\lambda}\mu$$

$$\ell_n^{\delta\lambda} := v_n^{\delta\lambda} + (1 - z_n^{\delta\lambda})^{-1}\alpha_n^{\delta\lambda}\mu$$

where

$$\beta_n^{\delta\lambda} := \sup_{i\neq 0}[(L_\delta^{f_{n+1}} v_n^{\delta\lambda}(i) - v_n^{\delta\lambda}(i))\mu^{-1}(i)] ,$$

$$\alpha_n^{\delta\lambda} := \inf_{i\neq 0}[(L_\delta^{f_{n+1}} v_n^{\delta\lambda}(i) - v_n^{\delta\lambda}(i))\mu^{-1}(i)]$$

and $y_n^{\delta\lambda}$, $z_n^{\delta\lambda}$ are defined by means of $\beta_n^{\delta\lambda}$ and $z_n^{\delta\lambda}$ in a similar way as y_n^{δ} and z_n^{δ}.

7.2. *The suboptimality of Markov policies and suboptimal actions*

We first want to remark once more that we still restrict ourselves to transition memoryless nonzero stopping times.

DEFINITION 7.2.1.

(i) A Markov policy $f \in F$ is called *suboptimal* if f^∞ is not optimal.

(ii) A Markov policy $f \in F$ is called ε-*suboptimal* if f^∞ is not ε-optimal.

LEMMA 7.2.1. Suppose $u,\ell \in V$ are an upper bound and a lower bound for the value vector V^*, and let $\delta \in \Delta'$, then

(i) $[L_\delta^f u < U_\delta \ell] \Rightarrow f$ is suboptimal

(ii) $[L_\delta^f u < U_\delta \ell - \varepsilon\mu] \Rightarrow f$ is ε-suboptimal .

PROOF.

(i) In chapter 5 we have proved $f \in F$ is optimal if and only if $L_\delta^f V^* = V^*$; moreover we have proved $U_\delta V^* = V^*$. However

$$L_\delta^f V^* \leq L_\delta^f u < U_\delta \ell \leq U_\delta V^* = V^* ,$$

consequently f is suboptimal.

(ii) $$L_\delta^f V^* \leq L_\delta^f u < U_\delta \ell - \varepsilon\mu \leq U_\delta V^* - \varepsilon\mu = V^* - \varepsilon\mu .$$

So

$$(L_\delta^f)^2 V^* \leq L_\delta^f (V^* - \varepsilon\mu) \leq L_\delta^f V^* < V^* - \varepsilon\mu .$$

Iterating in this way yields

$$(L_\delta^f)^n V^* < V^* - \varepsilon\mu$$

which implies that $V_f < V^* - \varepsilon\mu$. □

It would also be nice to make assertions as to the (ε-)suboptimality of actions.

DEFINITION 7.2.2.

(i) An action $a_0 \in A(i)$ is said to be suboptimal with respect to state $i \in S$ if there exists no optimal policy $f \in F$ with $f(i) = a_0$.

(ii) An action $a_0 \in A(i)$ is said to be ε-suboptimal with respect to state $i \in S$ if there exists no ε-optimal policy $f \in F$ such that $f(i) = a_0$.

LEMMA 7.2.2. Let $\delta \in \Delta', u, \ell \in V$ be an upper bound and a lower bound for V^*. Let $i \in S$. For each $f \in F$ we define the set $F_i^f \subset F$ by

$$F_i^f := \{f' \in F \mid f'(j) = f(j) \text{ for all } j \neq i\}$$

then

(i) The action $a \in A(i)$ is suboptimal with respect to $i \in S$ if for each $f \in F$ with $f(i) = a$ there exists an $f' \in F_i^f$ such that

$$L_\delta^f u < L_\delta^{f'} \ell \ .$$

(ii) The action $a \in A(i)$ is ε-suboptimal with respect to $i \in S$ if for each $f \in F$ with $f(i) = a$ there exists an $f' \in F_i^f$ such that

$$L_\delta^f u < L_\delta^{f'} \ell - \varepsilon\mu \ .$$

PROOF. The proof proceeds in a similar way as the proof of the foregoing lemma. □

The assertions of the foregoing lemma may also be expressed in terms of maximizing over the decisions in state $i \in S$ as follows.

COROLLARY 7.2.1. Let $\delta \in \Delta'$, and $u,\ell \in V$ be an upper bound and a lower bound for V^*. Denote by $f_a \in F$ a Markov policy with $f(i) = a$ then

(i) the action $a' \in A(i)$ is suboptimal with respect to $i \in S$ if for all $f_{a'} \in F$

$$L_\delta^{f_{a'}} u(i) < \sup_{a \in A(i)} (L_\delta^{f_a})\ell(i) \ .$$

(ii) The action $a' \in A(i)$ is ε-suboptimal with respect to $i \in S$ if for all $f_{a'} \in F$

$$L_\delta^{f_{a'}} u(i) < \sup_{a \in A(i)} (L_\delta^{f_a})\ell(i) - \varepsilon\mu(i) \ .$$

Especially for computational purposes it would be desirable to maximize component-wise in the determination of $U_\delta v$, $U_{\delta,\varepsilon} v$. Sufficient conditions to allow a component-wise maximization may be found in the following lemma.

LEMMA 7.2.3. Let $\delta \in \Delta'$ and suppose (eventually after renumbering the state space) that

$$(7.2.1) \qquad \forall_{i \in S} \ \forall_{j \geq i} \ \forall_{a \in A(i)} \ [p^a(i,j)\delta(i,j) = 0] \ ,$$

then

$$U_\delta v(i) := \sup_{f\in F} L_\delta^f v(i) = (1 - \delta(i))v(i) +$$

$$\delta(i) \sup_{a\in A(i)} \{r(i,a) + \sum_{j<i} p^a(i,j)[(1-\delta(i,j))v(j)+\delta(i,j)U_\delta v(j)]$$

$$+ \sum_{j\geq i} p^a(i,j)v(j)\} \quad .$$

PROOF. The proof of the lemma follows directly by analysing the transformed problem as described in the proof of theorem 5.2.2. □

REMARK 7.2.1. The transition memoryless nonzero stopping time G_H and G_R as described in the examples 2.2.2 and 2.2.4 and the stopping time δ corresponding to G_1 satisfy the conditions of the foregoing lemma.

In the next section we will consider into more detail situations as described in formula (7.2.1). This will be done for finite state space Markov decision processes.

LEMMA 7.2.4. Let $\delta \in \Delta'$, and let $\ell,u \in V$ be a lower and an upper bound for V^* respectively and suppose (7.2.1) is satisfied, then

(i) the action $a_0 \in A(i)$ is suboptimal if

$$r(i,a_0) + \sum_{j<i} p^{a_0}(i,j)[(1-\delta(i,j))u(j) + \delta(i,j)U_\delta u(j)] + \sum_{j\geq i} p^{a_0}(i,j)u(j)$$

$$< \sup_{a\in A(i)} \{r(i,a) + \sum_{j<i} p^a(i,j)[(1-\delta(i,j))\ell(j)+\delta(i,j)U_\delta \ell(j)] + \sum_{j\geq i} p^a(i,j)\ell(j)\}.$$

(ii) The action $a \in A(i)$ is ε-suboptimal if the first term under (i) is smaller than the second term minus $\varepsilon\mu(i)(\delta(i))^{-1}$.

PROOF. The proof follows by inspection. □

REMARK 7.2.2.

(i) For δ corresponding to G_1 the condition required in lemma 7.2.4(i) reduces to

$$r(i,a_0) + \sum_{j\in S} p^{a_0}(i,j)u(j) < \sup_{a\in A(i)} \{r(i,a) + \sum_{j\in S} p^a(i,j)\ell(j)\} .$$

(ii) The latter suboptimality criterion can be expressed more explicitly if the upper and lower bounds are given in terms of v_n and v_{n-1}, see e.g. MacQueen [47].

(iii) Better criteria can be obtained if we impose additional conditions on the transition probabilities, see e.g. Hübner [36].

7.3. *Some remarks on finite state space Markov decision processes*

For practical purposes (the finite state space) Markov decision processes become more and more important. Many practical problems such as inventory problems, replacement problems and Marketing problems may be described by Markov decision models. We refer to e.g. Scarf [63], Tijms [67], Howard [35], Hastings [25] and Wessels and van Nunen [73].

Sometimes it will be self evident to approximate Markov decision processes with a countable state space by finite state space processes, this is done e.g. by Fox [19].

For the solution of Markov decision problems in the finite state case, linear programming [13], [49], and policy iteration may be used. If the underlying decision space contains only a finite number of elements, both methods yield an optimal solution in a finite number of steps. However, difficulties arise if the state space is large. For example, the policy iteration method requires in each iteration step the solution of a system of linear equations of the size of the number of states. For solving Markov decision problems with a large state space, successive approximation methods that avoid the solution of the large systems of linear equations, become preferable. This is especially true if the concept of extrapolation is used to construct upper and lower bounds, see Schellhaas [64], MacQueen [46], Porteus [59], Finkbeiner and Runggaldier [18], Das Gupta [22] and the present author [54].

As we have seen upper and lower bounds enable us to incorporate a test for the suboptimality of actions. Such a test may also reduce the required computational effort considerably, see e.g. Grinold [20], MacQueen [47], Porteus [58], Hastings and Mello [27], and Hübner [36].
Hinderer [32] derived in a similar way bounds for finite stage dynamic programs in the case $b \in \mathcal{W}$, δ corresponding to G_1 and the usual supremum norm. In [31] Hinderer extended his results by using weighted supremum norms as introduced by Wessels [74].
A number of the described techniques for solving Markov decision problems with respect to the total reward criterion can be used in a modified way for solving these problems with respect to the average reward criterion, see Odoni [57], White [77], Schweitzer [65], Morton [51], Veinott [69] and van der Wal [70]. Also for Markov games similar techniques can be used; see van der Wal [71], [72]. Since periodic Markov decision processes (see e.g. Carton [7], Riis [61]) can be described as ordinary Markov decision processes, by incorporating the period in the state definition the same holds for such processes.
For Markov decision processes with a finite state space the assumptions 4.1.1, 4.2.1-4.2.4 reduce in fact to the well-known assumptions

$$\exists_{M \in \mathbb{R}^+} \forall_{f \in F} \; \| r^f \| < M \quad \text{and} \quad \exists_{n \in \mathbb{N}} \exists_{\rho < 1} \forall_{f \in R} \; \| (P^f)^n \| \le \rho < 1 \quad ,$$

(n-stage contraction), see for instance Denardo [12] or Porteus [59]. For finite state space problems van Hee and Wessels [29] have proved the equivalence between the existence of a bounding function μ' such that $\| P^f \|_{\mu'} \le \rho < 1$ for all $f \in F$ and N-stage contraction. We will discuss the concept of N-stage contraction in chapter 8.
The use of a bounding function may be interpreted as the transformation of the problem in an equivalent one (in a similar way as Porteus did in his recent paper [59] (the similarity transformations)). This can be done by defining transformed rewards $\tilde{r}(i,a) := \mu^{-1}(i)r(i,a)$ and transformed transition probabilities $\tilde{p}^a(i,j) := \mu^{-1}(i)p^a(i,j)\mu(j)$. Then the optimal value vector of the transformed problem $(\tilde{V}^*)$ equals $\tilde{V}^* = \mu^{-1}V^*$, with V^* the optimal return of the original problem as is straightforwardly verified.
Also the use of stopping times for successive approximation procedures may be viewed upon as investigating a transformed problem with

$$\tilde{r}^f := \mathbb{E}_\delta^f \sum_{k=0}^{\tau-1} r(s_k) \quad \text{and} \quad \tilde{P}^f := P_\delta^f \; .$$

As proved in the preceding chapters the optimal return vector $\tilde{V}^*$ then equals V^* of the original problem.

For finite state space Markov decision processes the last mentioned transformation corresponds to the pre-inverse transformation as introduced by Porteus [59].

Let the *norm of the process* be defined by $\sup_{f \in F} \| P^f \|$ and let the *spectral radius of the process* be defined by $\sup_{f \in F} \gamma^f$, with γ^f the spectral radius of P^f (its maximum absolute eigenvalue) then it will be clear that the choice of an adequate transition memoryless nonzero stopping time yields a transformed process with a reduced spectral radius. The choice of the μ-function however may influence the norm of the transformed process but does not alter the spectral radius.

A reduction of the spectral radius improves the performance of the value iteration while a reduction of the norm of the process (using an adequate μ-function) may improve the bounds introduced in section 7.1.

For computational purposes it will be desirable that for each $f \in F$ and each $\delta \in \Delta'$ the matrix with (i,j)-th entry equal to $P^{f(i)}(i,j)\delta(i,j)$ possesses the triangular structure as described in formula (7.2.1). As mentioned earlier, if the stopping times corresponding with G_1, G_R, G_H are used then (7.2.1) is satisfied. For δ corresponding to G_1 the approximation procedure described in lemma 4.3.6 results in

$$v_0 \in V \quad , \forall_{i \in S} \; v_n(i) = \max_{a \in A(i)} \{r(i,a) + \sum_{j \in S} p^a(i,j) v_{n-1}(j)\} \; .$$

For δ corresponding to G_R the procedure described in lemma 5.2.4 may be represented more explicitly by

$$v_0^\delta \in V,$$

$$v_n^\delta(i) = \max_{a \in A(i)} \{\frac{1}{1-p^a(i,i)} r(i,a) + \frac{1}{1-p^a(i,i)} \sum_{j \neq i} p^a(i,j) v_{n-1}^\delta(j)\}$$

which is the well-known Jacobi iteration, see e.g. Porteus [59]. By choosing δ such that δ corresponds to G_R with exception of $\delta(i,i)$, $i \in S$ for which it is allowed that $\delta(i,i) < 1$ for each $i \in S \backslash \{0\}$ we get a successive overrelaxation procedure. An arbitrary choice of δ such that (7.2.1) is satisfied may yield a combination of overrelaxation and Gauss-Seidel-like procedures.

For δ corresponding to G_H the procedure as described in lemma 5.2.4 can be executed component-wise

$$v_0 \in V,\ v_n^\delta(i) = \max_{a\in A(i)} \{r(i,a) + \sum_{j<i} p^a(i,j)v_n^\delta(j) + \sum_{j\geq i} p^a(i,j)v_{n-1}^\delta(j)\} .$$

Of course the way in which the state space is ordered may influence the rate of convergence the process, see e.g. Kushner and Kleinman [41].

As described in the previous sections all those methods allow for the construction of lower and upper bounds and for the use of suboptimality criteria. Which procedure is preferable for solving a given finite state space Markov decision process depends on the problem under consideration. If we want to compare two different procedures it will be necessary to compare the corresponding sequences of upper and lower bounds. However, where the estimates for V^* in the n-th iteration step of a specific algorithm may be better then those of another algorithm (other stopping time) this does not mean unfortunately that it is possible to construct in an easy way, bounds that are better too. We have illustrated this phenomenon by some examples in [54]. Moreover the question whether additional effort spent for the computation by more sophisticated stopping times and the computational gain in the number of requested iterations, is evenly matched, cannot be answered in general. For numerical experiences in these directions we refer to Porteus [58], Schellhaas [64] and Kushner and Kleinman [41].

The use of value oriented procedures, as described in chapter 6, for solving finite state and decision space Markov decision problems may yield a considerable gain in computational effort. This is the case if the policy improvement procedure requires many operations i.e. if the total number of decisions is large compared with the number of states. Furthermore, in practice optimal policies are achieved after a relatively small number of iteration steps, whereas a nonnegligable number of iterations is still required to satisfy the convergence criterion. Especially for the last reason it will be profitable to adapt the value of λ during the iteration process. For numerical experiences with the value oriented method in the case δ corresponding to G_1 we refer to [53].

CHAPTER 8

N-STAGE CONTRACTION

In chapter 4 we have required in fact one stage contraction with respect to the bounding function μ i.e.

$$\exists_{\rho'<1} \forall_{f\in F} \| P^f \|_\mu \leq \rho' .$$

This assumption may be weakened to the case of the so-called N-stage contraction. The extension is comparable with Denardo's [12] concept of N-stage contraction in its strengthened form introduced for the unweighted supremum norm case. We will prove that if the N-stage contraction assumption in its strengthened form is satisfied instead of assumption 4.2.2, this implies the existence of a bounding function μ' such that, the assumptions 4.2.1-4.2.2 are satisfied when μ is replaced by μ'. Moreover we will indicate the convergence (in μ-norm) of the sequences v_n, v_n^δ, $v_n^{\delta\lambda}$ as defined in the preceding chapters under the N-stage contraction assumption in its strengthened form.

8.1. *Convergence under the strengthened N-stage contraction assumption*

First we will introduce the strengthened *N-stage contraction assumption*.

ASSUMPTION 8.1.1.

(i) $\exists_{M\in\mathbb{R}^+} \forall_{f\in F} \| P^f \| \leq M$

(ii) $\exists_{N\in\mathbb{N}} \exists_{\rho^*<1} \forall_{f_1,f_2,\ldots,f_N\in F} \| P^{f_1} P^{f_2} \ldots P^{f_N} \| \leq \rho^*$.

THEOREM 8.1.1. Suppose, the assumptions 4.2.1, 4.2.3, 4.2.4 and 8.1.1 are satisfied then there exists a bounding function μ' such that

(i) $\exists_{M\in\mathbb{R}^+} \forall_{f\in F} \| r^f - b \|_{\mu'} \leq M$

(ii) $\exists_{\rho'<1} \forall_{f\in F} \| P^f \|_{\mu'} \leq \rho'$

(iii) $\exists_{M\in\mathbb{R}^+} \forall_{f\in F} \| P^f b - \rho b \|_{\mu'} \le M$.

PROOF. Choose α such that $\rho^* < \alpha^N < 1$ where $\rho^* < 1$ follows from assumption 8.1.1 and define μ' by

$$\mu' := \sup_{\pi\in\mathcal{M}} \sum_{n=0}^{\infty} \frac{1}{\alpha^n} P^n(\pi)\cdot\mu \ .$$

Then for $\pi = (f_0, f_1, \ldots)$

$$\mu' = \sup_{\pi\in\mathcal{M}} \sum_{m=0}^{N} \frac{1}{\alpha^m} P^m(\pi) \sum_{n=0}^{\infty} \frac{1}{\alpha^{nN}} P^{f_{m+1}} \ldots (P^{f_{m+nN}})\mu \ .$$

However, since $P^{f_0} \ldots P^{f_{N-1}}\mu \le \rho^* \mu$ it holds that

$$P^{f_{m+1}} \ldots P^{f_{m+nN}}\mu \le (\rho^*)^n \mu \ .$$

Hence

$$\mu' \le \sup_{\pi\in\mathcal{M}} \sum_{m=0}^{N} P^m(\pi)(1 - \rho^* \alpha^{-N})^{-1} \alpha^{-m} \mu$$

on the other hand

$$\forall_{f\in F}\ \mu' \ge \mu + \frac{1}{\alpha} P^f \mu' \ ,$$

as follows from the definition of μ' and the fact that Markov strategies dominate all decision rules, for a proof see van Hee [28], We also refer to van Hee and Wessels [29].
This implies $\alpha\mu' \ge P^f \mu'$ for all $f \in F$, so $\forall_{f\in F} \| P^f \|_{\mu'} \le \alpha < 1$. From the definition of μ' we see $\mu' \ge \mu$. Now the proof of part (i) and (iii) is trivial. □

The foregoing theorem proves that the strengthened N-stage contraction assumption implies the existence of a bounding function μ' such that for this new bounding function the assumptions 4.2.1-4.2.4 are satisfied.

REMARK 8.1.1. Also Lippman [45] gives conditions for N-stage contraction. However, it is proved in van Nunen and Wessels [56] that these conditions imply the existence of a bounding function μ' such that for μ' the assumptions 4.2.1-4.2.4, with $b = 0$, are satisfied.

On one hand one could say the strengthened N-stage contraction assumption yields no real extension since it guarantees the existence of a μ' with the described properties. On the other hand the strengthened assumption 8.1.1 also guarantees the convergence of the sequences v_n, v_n^δ, $v_n^{\delta\lambda}$, as defined in the previous chapters, with respect to the original μ-norm. This will be indicated in the sequel of this section by proving that the above statement is true for the sequences as defined in lemmas 4.3.7 and 5.2.5.

THEOREM 8.1.2. Suppose the assumptions 4.2.1, 4.2.3, 4.2.4 and 8.1.1 are satisfied then for each $f \in F$ we have

(i) L_1^f is a mapping $V \to V$.

(ii) $(L_1^f)^N$ is a contraction mapping of V into V with contraction radius ρ', where N and ρ' follow from assumption 8.1.1.

(iii) For each $v_0^f \in V$ the sequence v_n^f defined by

$$v_n^f := L_1^f v_{n-1}^f$$

converges to V_{f^∞}.

(iv) $$\| v_{nN}^f - V_f \| \leq (\rho')^n \| v_0^f - V_f \| .$$

PROOF. The proof of the first two parts of the lemma is straightforward. To prove part (iii), note that if L_1^f has a fixed point it has to be the fixed point of $(L_1^f)^N$.

Let V be the fixed point of $(L_1^f)^N$ then V is a fixed point of L_1^f as follows from

$$\| L_1^f V - V \| = \| (L_1^f)^{N+1} V - (L_1^f)^N V \| \leq \rho' \| L_1^f V - V \| .$$

That $v_n^f \to V$ (in μ-norm) for $n \to \infty$ is proved by choosing $\varepsilon > 0$ and $k \in \mathbb{N}$ such that

$$\| (L_1^f)^{Nk} v_0^f - v \| \le \varepsilon \ .$$

Let $n > Nk$, then n can be written as $n = Nk_1 + \ell$ with $k_1 \ge k$ and $1 \le \ell \le N$,

$$\| (L_1^f)^n v - v \| = \| (L_1^f)^{\ell} ((L_1^f)^{Nk_1} v_0^f) - v \| = \| (L_1^f)^{\ell} ((L_1^f)^{Nk_1} v_0^f) - (L_1^f)^{\ell} v \|$$

$$\le \| P^f \|^{\ell} \| (L_1^f)^{Nk_1} v_0^f - v \| \le \varepsilon C \ .$$

Where $C := \max\{1, (\sup_{f \in F} \| P^f \|)^N\}$.

That V equals V_{f^∞} is easily verified.

Part (iv) follows in the standard way

$$\| v_{nN}^f - V_f \| = \| (L_1^f)^{nN} v_0^f - (L_1^f)^{nN} V_f \| \le (\rho^*)^n \| v_0^f - V_f \| \ . \qquad \square$$

THEOREM 8.1.3. Suppose the assumptions 4.2.1, 4.2.3, 4.2.4 and 8.1.1 are satisfied and suppose $U_{1,0}$ is defined, then

(i) $U_{1,0}$ maps V into V.

(ii) $(U_{1,0})^N$ is a contraction mapping of V into V with contraction radius ρ' where N and ρ' follow from assumption 8.1.1.

(iii) the sequence v_n defined for $v_0 \in V$ by

$$v_n = U_{1,0} v_{n-1}$$

converges to v^* (in μ-norm).

(iv) $$\| v_{nN} - v^* \| \le (\rho^*)^n \| v_0 - v^* \| \ .$$

PROOF. The first part of the lemma is trivial. Part (ii) holds since the N-stage contraction assumption 8.1.1 is satisfied. The final parts follow in a similar way as the final parts of the foregoing theorem. □

We finish this chapter with a theorem concerning the sequence v_n^δ as defined in lemma 5.2.5(iii).

THEOREM 8.1.4. Suppose the assumptions 4.2.1, 4.2.3, 4.2.4 and 8.1.1 are satisfied, and let δ be a transition memoryless nonzero stopping time and suppose $U_{\delta,0}$ is defined, then

(i) $U_{\delta,0}$ maps $\mathcal{V}$ into $\mathcal{V}$.

(ii) Suppose $v_0^\delta \in \mathcal{V}$ is such that $U_{\delta,0}v_0^\delta \geq v_0^\delta$ then the sequence v_n^δ defined by

$$v_n^\delta := U_{\delta,0}v_{n-1}^\delta$$

converges to V^* in μ-norm.

PROOF. The proof follows directly from the fact that

$$\forall_{f \in F}\ (L_\delta^f)^N v_0^\delta \leq U_\delta^N v_0^\delta$$

and the fact that $U_\delta V^* = V^*$. □

CHAPTER 9

SOME EXPLANATORY EXAMPLES

In section 9.1 we will describe how a number of specific Markov decision processes, namely discounted Markov decision processes and (discounted) semi-Markov decision processes are covered by the theory developed in this monograph.

In section 9.2 we will show how the theory developed in chapter 3 can be used to generate successive approximation procedures for solving systems of linear equations of the form

$$Ax = r \ ,$$

where A can be described $A = I - P$, where I is the identity matrix and P satisfies the assumptions of chapter 3. Four special cases will be indicated. In section 9.3 the monograph will be concluded with an example concerning an inventory problem. The problem is the countable state space discounted equivalent of the inventory problem with average reward criterion as treated by Wijngaard [78].

9.1. *Some examples of specific Markov decision processes*

We start with a remark concerning the applicability of our model. In chapter 4 we have introduced the reward $r(i,a)$ as the immediate return if the system is in state $i \in S$ and action $a \in A$ is selected. However in several situations the one stage return will depend on the subsequent transition that occurs aswell. So the one stage return can be composed of two parts, an amount $r^1(i,a)$ depending on the actual state $i \in S$ and the selected action $a \in A$, and an amount $r^2(i,a,j)$ which depends on the next state (j) that is visited aswell. However we can still use our model if we define $r(i,a)$ as an expected one stage return i.e.

$$r(i,a) := r^1(i,a) + \sum_{j \in S} p^a(i,j) r^2(i,a,j) \ .$$

We will now consider *discounted* Markov decision processes, treated for finite state space finite decision space Markov decision processes by e.g. Howard [35] and for general state space, general action space Markov decision processes with a bounded reward structure by e.g. Blackwell [5]. The difference between the models we have discussed and discounted models is that in the latter situation the reward earned at time n is weighted by a factor β^n, where $\beta \geq 0$ is the discountfactor. We denote the transition probabilities for the discounted process by $q^a(i,j)$ for $i,j \in S$ and $a \in A$. For an arbitrary Markov strategy $\pi := (f_0,f_1,\ldots)$ we define $Q^n(\pi)$ with respect to the transition probabilities $q^a(i,j)$ in a similar way as we have defined $P^n(\pi)$. Then the total expected *discounted* reward over an infinite time horizon equals

$$\sum_{n=0}^{\infty} \beta^n Q^n(\pi) r^{f_n} .$$

By incorporating the discount factor β in the transition probabilities i.e. $p^a(i,j) := \beta q^a(i,j)$ the total expected discounted reward equals

$$\sum_{n=0}^{\infty} P^n(\pi) r^{f_n} .$$

So if for these redefined transition probabilities and for the reward structure the assumptions of chapter 4 are satisfied then discounted Markov decision processes are covered by the theory developed in the previous chapters. Usually this discountfactor is supposed to be smaller than one (i.e. $0 \leq \beta < 1$). However, depending on the transition probabilities it may be allowed that $\beta \geq 1$.

For discounted Markov decision processes with $\beta < 1$, $b = 0$ and with $\mu(i) = 1$ if $i \neq 0$ the assumptions of chapter 4 are satisfied if the rewards are bounded.

We now consider semi-Markov decision processes as introduced by Jewell[37], [38].

For semi-Markov decision processes state transitions do not neccessarily occur at equidistant points in time. The time (t) between two state transitions is a random variable with a probability distribution function $F^a_{i,j}(t)$. The probabilities of the state transitions are as in chapter 4. So if immediately after a transition, the state of the system is $i \in S$ and action $a \in A(i)$ is selected, then the system's next state will be j with a probability denoted by $q^a(i,j)$ where the $q^a(i,j)$ satisfy the assumption 4.1.1.

Such a transition will occur before time t_1 with probability $\int_{0^-}^{t_1} dF^a_{i,j}(t)$.
We can consider the embedded Markov decision process, in discrete time by defining the state of the embedded process at time n $(n = 0,1,\dots)$ to be the state immediately after the n-th transition of the original decision process (see e.g. Mine and Osaki [50], Ross [62], De Cani [10]).

For semi-Markov decision processes with respect to the total expected reward criterion (without discounting) the total expected reward over an infinite number of transitions, (using $\pi = (f_0,f_1,\dots)$ and $Q^n(\pi)$ as we did before) equals

$$\sum_{n=0}^{\infty} Q^n(\pi) r^{f_n} .$$

So provided that for the embedded process the assumptions of chapter 4 are satisfied the theory of the preceding chapters can be applied.

In the theory of semi-Markov decision processes with discounting, one assumes that rewards incurred at time t are discounted by a factor β^t. In a similar way as in the ordinary discounted case the discountfactor can be incorporated in the transition probabilities for the embedded process i.e.

$$p^a(i,j) := \int_{0^-}^{\infty} \beta^t dF^a_{i,j}(t)] q^a(i,j) .$$

9.2. *The solution of systems of linear equations*

Suppose the following system of linear equations has to be solved

$$(9.2.1) \qquad Ax = r$$

with $A := I - P$, where I is the identity matrix and P satisfies the conditions as imposed in chapter 3 with $\mu(i) := 1$, $i \neq 0$ and $b(i) := 0$ for $i \in S$.

For each $\delta \in \Delta$ we have proved in chapter 3 that the sequence

$$v_n^\delta := L_\delta v_{n-1}^\delta, \quad v_0^\delta \in V$$

converges to the solution of (9.2.1) if and only if δ is a nonzero stopping time.

So for each nonzero stopping time we have a successive approximation method for solving (9.2.1). For some specific stopping times these methods are already known from numerical mathematics (see e.g. Varga [68]). This will be shown in this section. So the convergence of these numerical methods to the solution of the system of linear equations (9.2.1) follow at one blow from theorem 3.2.1.

EXAMPLE 9.2.1. Let $\delta \in \Delta'$ correspond to the goahead set G_R. We have for each $i \in S$

$$(9.2.2) \quad v_n^\delta(i) := L_\delta v_{n-1}^\delta(i) = \frac{1}{1-p(i,i)}\, r(i) + \frac{1}{1-p(i,i)} \sum_{j \neq i} p(i,j)\, v_{n-1}^\delta(j) \quad \text{with } v_0 \in V .$$

We define D as the diagonal matrix with diagonal entries $(1-p(i,i))$ and define F and E by the strictly upper and strictly lower triangular parts of P. Then clearly $(I-P)$ can be expressed by

$$I - P = D - E - F .$$

Now v_n^δ can be given by

$$v_n^\delta := D^{-1}(E+F)v_{n-1}^\delta + D^{-1}r .$$

This iterative method is known as the point Jacobi or point total step method (see Varga [68]).

EXAMPLE 9.2.2. Let $\delta \in \Delta'$ correspond to the goahead set G_H' with $G_H'(0) = G_H(0)$, $G_H'(i) := \{(\alpha,\beta) \mid \beta \in \bigcup_{j=0}^{i-1} G_H'(j),\ \alpha \in \bigcup_{k=1}^{\infty} \{i\}^k\}$, then v_n^δ can be given componentwise by (see also example (iii) on page 32)

$$(9.2.3)\qquad v_n^{\delta}(i) := \frac{1}{1-p(i,i)}\, r(i) + \frac{1}{1-p(i,i)} \sum_{j<i} p(i,j)\, v_n^{\delta}(j) + \\ + \frac{1}{1-p(i,i)} \sum_{j>i} p(i,j)\, v_{n-1}^{\delta}(j)\ .$$

By using the matrices D, E and F as defined in the previous example, formula (9.2.3) can be given by

$$(D - E)\, v_n^{\delta} = F v_{n-1}^{\delta} + r\ ,$$

or alternatively by

$$v_n^{\delta} = (D - E)^{-1} F v_{n-1}^{\delta} + (D - E)^{-1} r\ .$$

This iteration method is known as the point Gauss-Seidel or point single step iterative method.

EXAMPLE 9.2.3. Let $\delta \in \Delta'$ be the transition memoryless nonrandomized stopping time such that $\delta(\alpha) = \delta([\alpha]_{k_\alpha - 1})$ for $\alpha \in G_H'$ (see example 9.2.2), $\delta(\alpha) = 0$ elsewhere. Moreover choose $\delta(i)$ such that

$$\forall_{\substack{i,j \in S \\ i \neq 0}} \quad \frac{\delta(i,i)(1-p(i,i))}{1-\delta(i,i)p(i,i)} = \frac{\delta(j,j)(1-p(j,j))}{1-\delta(j,j)p(j,j)}\ .$$

Then v_n^{δ} can be expressed component-wise by

$$(9.2.4)\qquad v_n^{\delta}(i) = \frac{\delta(i)}{1-\delta(i)p(i,i)}\, r(i) + \frac{1-\delta(i)}{1-\delta(i)p(i,i)}\, v_{n-1}^{\delta}(i) + \\ + \frac{\delta(i)}{1-\delta(i)p(i,i)} \sum_{j<i} p(i,j)\, v_n^{\delta}(j) + \\ + \frac{\delta(i)}{1-\delta(i)p(i,i)} \sum_{j>i} p(i,j)\, v_{n-1}^{\delta}(j)\ .$$

By defining $\omega := \dfrac{\delta(i)(1-p(i,i))}{1-\delta(i)p(i,i)}$ and using the matrices D, E, F (9.2.4) can be given by

$$(D - \omega E)\, v_n^{\delta} = \omega r + ((1-\omega)D + \omega F)\, v_{n-1}^{\delta}\ .$$

This is known as the point successive overrelaxation method.

EXAMPLE 9.2.4. Suppose the state space is partitioned in blocks $B_1 = \{1,2,\dots,n_1\}$, $B_2 = \{n_1+1,\dots,n_2\}$ and so on. Let P be given by

$$P := \begin{bmatrix} P_{1,1} & P_{1,2} & \cdots & P_{1,n} & \cdots \\ P_{2,1} & P_{2,2} & \cdots & P_{2,n} & \cdots \\ \vdots & & \ddots & \vdots & \\ P_{n,1} & P_{n,2} & \cdots & P_{n,n} & \\ \vdots & & & \vdots & \ddots \end{bmatrix} .$$

where $P_{n,m}$ contains the entries $p(i,j)$ with $i \in B_n$ and $j \in B_m$. We define the matrices D, E and F by

$$D := I - \begin{bmatrix} P_{1,1} & \bigcirc \\ \bigcirc & P_{2,2} \\ & & \ddots \end{bmatrix}$$

$$E = \begin{bmatrix} 0 & 0 & \cdots \\ P_{2,1} & 0 & \cdots \\ P_{3,1} & P_{3,2} & \ddots \\ \vdots & & \ddots \end{bmatrix} \qquad F = \begin{bmatrix} 0 & P_{1,2} & P_{1,3} & \cdots & \\ 0 & 0 & P_{2,3} & P_{2,4} & \cdots \\ \vdots & & 0 & \ddots & \ddots \\ \vdots & & & \ddots & \ddots \end{bmatrix}$$

Now let $\delta \in \Delta'$ be the transition memoryless nonrandomized stopping time with $\delta(i) = 1$ for all $i \in S$, and $\delta(i,j) = 1$ if $j = 0$ or there is some $n \in \mathbb{N}$ such that $i,j \in B_n$, let $\delta(i,j) = 0$ elsewhere. Now it is easily verified that $v_n^\delta := L_\delta v_{n-1}^\delta$ can be given by

$$v_n^\delta = D^{-1}(E + F)v_{n-1}^\delta + D^{-1}r .$$

This iterative method is known as the block Jacobi iterative method (see e.g. Varga [68]).

From the foregoing examples it will be clear that the so called block successive overrelaxation iterative method can be found by combining the ideas used in example 9.2.3 and 9.2.4. Of course several other options (other choices of δ) are available. The previous examples are only given to illustrate how the concept of stopping time can be used to generate the iterative methods for solving systems of linear equations of the form (9.2.1).

9.3. *An inventory problem*

The inventory problem we deal with in this final section will not show the full strength of the described methods.
It only illustrates the power of the concept of bounding function (weighted supremum norm). Therefore, in the remaining part of this section, the vector b is taken equal to the zero vector ($\forall_{i \in S}\ b(i) = 0$). In our example we treat the discrete analogue of the inventory problem studied by Wijngaard [78]. However we will not investigate whether optimal strategies exhibit a specific structure or not.

We will first describe what we will mean here with an inventory problem.
An inventory problem consists of

(i) A state space containing the allowed inventory levels. Here, we assume the allowed inventory levels to be integers. We allow for backlogging which explains why the inventory levels can be negative.

(ii) Decision spaces A(i) containing for each inventory level $i \in \mathbb{Z}$ the possible orders. We assume the orders to be nonnegative integers. So A(i) might be given by e.g. $A(i) = \{0,1,\dots,N\}$ if at level i all orders of size 0 up to N are allowed. We assume the leadtime to be zero.

(iii) A distribution function of the demand (d) per period. We assume this demand to be nonnegative. For $k \in \mathbb{Z}^+$ the probability that the demand equals k is p_k we assume $p_0 \neq 1$. Moreover we assume

$$(9.3.1) \qquad \sum_{k=0}^{\infty} p_k \exp(k) < \infty .$$

(iv) A cost function c_1 on $\mathbb{Z}^+$ into $\mathbb{R}^+$. For each $k \in \mathbb{Z}^+$, $c_1(k)$ gives the ordering costs of k units.
A cost function c_2 on $\mathbb{Z}$ into $\mathbb{R}^+$. For each $k \in \mathbb{Z}$, $c_2(k)$ gives the one period (expected) inventory and stock out cost if at the beginning of the period k units are available (note that $k < 0$ means a shortage) we assume

$$(9.3.2) \quad \begin{cases} \exists_{M \in \mathbb{R}^+} \forall_{k \in \mathbb{Z}^+} [c_1(k)\exp(-k) < M] \\ \exists_{M \in \mathbb{R}^+} \forall_{k \in \mathbb{Z}} [c_2(k)\exp(-|k|) < M] . \end{cases}$$

(v) An optimality criterion. Here we will use the total (expected) discounted reward criterion with discountfactor β $(0 < \beta < 1)$.

We will now adapt the formulation of the inventory problems in such a way that they are covered by the models developed in the preceding chapters. Of course it is possible to label the inventory levels with $\mathbb{N} \cup \{0\}$. However, we will use the state space $S := \mathbb{Z} \cup \{0'\}$ to maintain the correspondence with the inventory problems as treated by Wijngaard. The state 0' represents the fictive absorbing state. We define the bounding function μ on S by

$$\mu(0') := 0; \ \mu(i) = \exp(|i|) \quad \text{for } i \in \mathbb{Z} .$$

We define the transition probabilities $p^a(i,j)$ with $a \in A(i)$, $i,j \in S$ and $i \neq 0'$ by

$$p^a(i,j) := \begin{cases} \beta p_{i+a-j} & \text{for } j \leq i + a , \\ 1 - \beta & \text{for } j = 0' , \\ 0 & \text{else} . \end{cases}$$

For $i = 0'$ we define $p(0',0') = 1$.
We define for $i \in S\backslash\{0'\}$ and $a \in A(i)$ the rewards $r(i,a)$ by

$$r(i,a) = -c_1(a) - c_2(i + a) .$$

THEOREM 9.3.1. Let $m, M, R \in \mathbb{Z}$ such that $m < 0 < M$, $R \leq M - m$ and

$$\sum_{k=0}^{\infty} p_k \exp(k) < \exp(R) .$$

Suppose $A(i)$ is such that

$$i \leq m \Rightarrow \forall_{a \in A(i)} \; a \geq R$$

$$i \leq M \Rightarrow \forall_{a \in A(i)} \; i + a \leq M$$

$$i > M \Rightarrow A(i) = \{0\} .$$

Then

(i) $\qquad \exists_{M' \in \mathbb{R}^+} \forall_{i \in S} \forall_{a \in A(i)} \; |r(i,a)| \leq M'\mu(i)$,

(ii) $\qquad \exists_{M'' \in \mathbb{R}^+} \forall_{f \in F} \; \| P^f \|_{\mu} \leq M''$,

(iii) $\qquad \exists_{\rho' < 1} \exists_{N \in \mathbb{N}} \forall_{f_0, \ldots, f_N \in F} \; \| P^{f_0} P^{f_1} \ldots P^{f_N} \| \leq \rho'$.

PROOF. The parts (i) and (ii) follow straightforwardly. To prove the final part of the theorem we use a result of Wijngaard [78]. From Wijngaard theorem 5.4 and the boundedness of μ on $[m,M]$ it follows that

$$\exists_{M^* \in \mathbb{R}^+} \forall_{n \in \mathbb{N}} \forall_{f_0, \ldots, f_n \in F} \; \beta^{-n} (P^{f_0} P^{f_1} \ldots P^{f_n}) \mu \leq M^* \mu .$$

Multiplying both sides by β^n we have for n sufficiently large that the n-stage assumption 8.1.1 is satisfied. □

REFERENCES

[1] Bauer, H., Wahrscheinlichkeitstheorie und Grundzüge der Maßtheorie.
Berlin, Walter de Gruiter & Co., 1968.

[2] Bellman, R., A Markovian decision process.
J. Math. Mech. 6 (1957), 679-684.

[3] Bellman, R., Dynamic programming.
Princeton (N.J.), Princeton University Press, 1957.

[4] Blackwell, D., Discrete dynamic programming.
Ann. Math. Statist. 33 (1962), 719-726.

[5] Blackwell, D., Discounted dynamic programming.
Ann. Math. Statist. 36 (1965), 226-235.

[6] Blackwell, D., Positive dynamic programming.
Proc. Fifth Berkeley Sympos. Math. Stat. and Prob., Vol. 1 (1967), 415-418.

[7] Carton, D.C., Une application de l'algorithme de Howard pour des phénomènes saisonniers.
Proc. 3rd Intern. Conf. Operations Res., Oslo, (1963), 683-691.

[8] Collatz, L., Funktional Analysis und Numerische Mathematik.
Berlin, Springer-Verlag, 1964.

[9] Cox, D.R. and H.D. Miller, The theory of stochastic processes.
London, Methuen & Co. LTD, 1968.

[10] De Cani, J.S., A dynamic programming algorithm for embedded Markov chains when the planning horizon is at infinity.
Management Sci. 10 (1964), 716-733.

[11] De Ghellinck, G.T. and G.D. Eppen, Linear programming solutions for separable Markovian decision problems.
Management Sci. 13 (1967), 371-394.

[12] Denardo, E.V., Contraction mappings in the theory underlying dynamic programming.
SIAM Rev. 9 (1967), 165-177.

[13] D'Epenoux, F., Sur un problème de production et de stockage dans l'aléatoire.
Rev. Franc. Rech. Opér. 14 (1960), 3-16.

[14] Derman, C., Markovian sequential control processes-denumerable state space.
J. Math. Anal. Appl. 10 (1965), 295-302.

[15] Derman, C., Finite state Markovian decision processes.
Academic Press New York etc., 1970.

[16] Derman, C. and R.E. Strauch, A note on memoryless rules for controlling sequential control processes.
Ann. Math. Statist. 37 (1966), 276-278.

[17] Feller, W., An introduction to probability theory and its applications, Vol. II (2nd ed.).
New York, Wiley, 1971.

[18] Finkbeiner, B. and W. Rungaldier, A value iteration algorithm for Markov renewal programming.
Computing methods in optimization problems 2; Ed. by L. Zadeh. New York etc., Academic Press 1969, pp. 95-104.

[19] Fox, B.L., Finite state approximation to denumerable state dynamic programs.
J. Math. Anal. Appl. 34 (1971), 665-670.

[20] Grinold, R., Elimination of suboptimal actions in Markov decision problems.
Operations Res. 21 (1973), 848-851.

[21] Groenewegen, L.P.J., Convergence results related to the equalizing property in a Markov decision process.
Eindhoven, University of Technology Eindhoven, Dept. of Math., 1975. (Memorandum-COSOR 75-18).

[22] Das Gupta, S., An algorithm to estimate suboptimal present values for unichain Markov processes with alternative reward structures.
Berlin, Springer-Verlag, Berlin etc., 1973, pp. 399-406.
(Lecture Notes in Computer Science, no. 3).

[23] Harrison, J., Countable state discounted Markovian decision processes with unbounded rewards.
Stanford, Dept. of Operations Res., Stanford University, 1970. (Techn. Rep. no. 17.)

[24] Harrison, J., Discrete dynamic programming with unbounded rewards.
Ann. Math. Statist. 43 (1972), 636-644.

[25] Hastings, N.A.J., Some notes on dynamic programming and replacement.
Operations Res. Quart. 19 (1968), 453-464.

[26] Hastings, N.A.J., Bounds on the gain of a Markov decision process.
Operations Res. 19 (1971), 240-248.

[27] Hastings, N.A.J. and J. Mello, Tests for suboptimal actions in discounted Markov programming.
Management Sci. 19 (1973), 1019-1022.

[28] van Hee, K.M., Markov strategies in dynamic programming.
Math. Oper. Res. 3 (1978), 37-41.

[29] van Hee, K.M. and J. Wessels, Markov decision processes and strongly excessive functions.
Stoch. Proc. appl. 8 (1978), 59-76.

[30] van Hee, K.M. and J.A.E.E. van Nunen, A note on the iterated expectation criterion for discrete dynamic programming.
Eindhoven, University of Technology Eindhoven, Dept. of Math., 1976. (Memorandum COSOR 76-03.)

[31] Hinderer, K., Bounds for stationary finite stage dynamic programs with unbounded reward functions.
Hamburg, Institut für Mathematische Stochastik der Universität Hamburg, June 1975 (Report).

[32] Hinderer, K., Estimates for finite stage dynamic programs.
J. Math. Anal. Appl. 55 (1976), 207-238.

[33] Hordijk, A., Dynamic programming and Markov potential theory.
Amsterdam, Mathematisch Centrum, 1974 (Mathematical Centre Tracts, no. 51).

[34] Hordijk, A., Convergent dynamic programming.
Stanford, Dep. Operations Res., Stanford University, 1974. (Techn. Rep. 28).

[35] Howard, R.A., Dynamic programming and Markov processes.
Cambridge (Mass.), M.I.T. Press, 1960.

[36] Hübner, G., Improved procedures for eliminating suboptimal actions in Markov programming by the use of contraction properties.
Transaction of the seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes: Prague, 1974 (incl. 1974 European Meeting of Statisticians). Academia, Prague, (1977)

[37] Jewell, W.S., Markov renewal programming: I. Formulation, finite return models.
Operations Res. 11 (1963), 938-948.

[38] Jewell, W.S., Markov renewal programming: II. Infinite return models, example,
Operations Res., 11 (1963), 949-971.

[39] Karlin, S., A first course in stochastic processes.
New York etc., Academic Press, 1966.

[40] Kemeny, J.G. and J.L. Snell, Finite Markov chains.
Princeton (N.J.), Van Nostrand, 1960.

[41] Kushner, H.J. and A.J. Kleinman, Accelerated procedures for the solution of discrete Markov control problems.
IEEE Transactions on automatic control. A.C.-16 (1971), 147-152.

[42] Krasnosel'skii, M.A., Approximate solutions of operator equations.
Groningen, Wolters-Noordhoff Publ. Comp., 1972.

[43] Ljusternik, L.A. and W.I. Sobolew, Elemente der Funktionalanalysis.
Berlin, Akademie-Verlag, 1968.

[44] Lippman, S.A., Semi-Markov decision processes with unbounded rewards.
Management Sci. 19 (1973), 717-731.

[45] Lippman, S.A., On dynamic programming with unbounded rewards.
Management Sci. 21 (1975), 1225-1233.

[46] MacQueen, J., A modified dynamic programming method for Markovian decision problems.
J. Math. Anal. Appl. 14 (1966), 38-43.

[47] MacQueen, J., A test for suboptimal actions in Markovian decision problems.
Operations Res. 15 (1967), 559-561.

[48] Maitra, A., Dynamic programming for countable state systems.
Sankhya Ser. A 27 (1965), 259-266.

[49] Manne, A.S., Linear programming and sequential decisions.
Management Sci. 6 (1960), 259-267.

[50] Mine, H. and S. Osaki, Markovian decision processes.
New York etc., Elsevier, 1970.

[51] Morton, T.E., Undiscounted Markov renewal programming via modified successive approximations.
Operations Res. 19 (1971), 1081-1089.

[52] Neveu, J., Mathematical foundations of the calculus of probability.
San Francisco, Holden-Day, 1965.

[53] van Nunen, J.A.E.E., A set of successive approximation methods for discounted Markovian decision problems.

[54] van Nunen, J.A.E.E., Improved successive approximation methods for discounted Markov decision processes.
Progress in Operations Research; ed. by A. Prékopa, Amsterdam, North-Holland Publishing Company, 1976, vol. II, pp. 667-682.

[55] van Nunen, J.A.E.E. and J. Wessels, A principle for generating optimization procedures for discounted Markov decision processes.
Progress in Operations Research; ed. by A. Prékopa, Amsterdam, North-Holland Publishing Company, 1976, vol. II, pp. 683-695.

[56] van Nunen, J.A.E.E. and J. Wessels, A note on dynamic programming with unbounded rewards.
Management. Sci. 24 (1978), 576-580.

[57] Odoni, A., On finding the maximal gain for Markov decision processes.
Operations Res. 17 (1969), 857-860.

[58] Porteus, E.L., Some bounds for discounted sequential decision processes.
Management Sci. 18 (1971), 7-11.

[59] Porteus, E.L., Bounds and transformations for discounted finite Markov decision chains.
Operations Res. 23 (1975), 761-784.

[60] Reetz, D., Solution of a Markovian decision problem by successive overrelaxation.
Zeitschrift für Operations Res. 21 (1973), 29-32.

[61] Riis, J.O., Discounted Markov programming in a periodic process.
Operations Res. 13 (1965), 920-929.

[62] Ross, S.M., Applied probability models with optimization applications.
San Francisco, Holden-Day, 1970.

[63] Scarf, H., The optimality of (S,s) policies in the dynamic inventory problem.
Mathematical methods in the social sciences; ed. by K.J. Arrow, S. Karlin and P. Suppes. Stanford, Stanford University Press, 1960; chap. 13.

[64] Schellhaas, H., Zur Extrapolation in Markoffschen entscheidungsmodellen mit Diskontierung.
Zeitschrift für Operations Res. 18 (1974), 91-104.

[65] Schweitzer, P., Iterative solution of the functional equations of undiscounted Markov renewal programming.
J. Math. Anal. Appl. 34 (1971), 495-501.

[66] Strauch, R.E., Negative dynamic programming.
Ann. Math. Statist. 37 (1966), 871-889.

[67] Tijms, H.C., Analysis of (s,S) inventory models.
Amsterdam, Mathematisch Centrum 1972 (Mathematical Centre Tracts, no. 40).

[68] Varga, R.S., Matrix iterative analysis.
Englewood Cliffs, Prentice-Hall, 1962.

[69] Veinott, A.F., On finding optimal policies in discrete dynamic programming with no discounting.
Ann. Math. Statist. 37 (1966), 1284-1294.

[70] van der Wal, J., The solution of an undiscounted completely ergodic Markov decision process by successive approximations.
Computing 17 (1976), 57-62.

[71] van der Wal, J., Discounted Markov Games; successive approximations and stopping times. The int. Journal of Game Theory.
6 (1977), 11-12.

[72] van der Wal, J., The solution of Markov games by successive approximation.
(Master's thesis), Eindhoven, Department of Mathematics, University of Technology, Eindhoven, 1975.

[73] Wessels, J. and J.A.E.E. van Nunen, Dynamic planning of sales promotions by Markov programming.
Proceed. XX Int. Meeting of the Institute of Manag. Sci.
Tel Aviv. Jerusalem Academic Press, 1975, Vol. II, 737-742.

[74] Wessels, J., Stopping times and Markov programming.
Transaction of the seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes: Prague, 1974 (incl. 1974 European Meeting of Statisticians). Academia, Prague, (1977), 575-585.

[75] Wessels, J., Markov programming by successive approximations with respect to weighted supremum norms.
J. Math. Anal. Appl. 58 (1977), 326-335.

[76] Wessels, J. and J.A.E.E. van Nunen, Discounted semi-Markov decision processes: linear programming and policy iteration.
Statistica Neerlandica 29 (1975), 1-7.

[77] White, D.J., Dynamic programming Markov chains, and the method of successive approximations.
J. Math. Anal. Appl. 6 (1963), 373-376.

[78] Wijngaard, J., Stationary Markovian decision problems.
University of Technology Eindhoven, 1975.

SUBJECT INDEX

SELECTED LIST OF SYMBOLS

MATHEMATICAL CENTRE TRACTS

1 T. van der Walt. *Fixed and almost fixed points.* 1963.

2 A.R. Bloemena. *Sampling from a graph.* 1964.

3 G. de Leve. *Generalized Markovian decision processes, part I: model and method.* 1964.

4 G. de Leve. *Generalized Markovian decision processes, part II: probabilistic background.* 1964.

5 G. de Leve, H.C. Tijms, P.J. Weeda. *Generalized Markovian decision processes, applications.* 1970.

6 M.A. Maurice. *Compact ordered spaces.* 1964.

7 W.R. van Zwet. *Convex transformations of random variables.* 1964.

8 J.A. Zonneveld. *Automatic numerical integration.* 1964.

9 P.C. Baayen. *Universal morphisms.* 1964.

10 E.M. de Jager. *Applications of distributions in mathematical physics.* 1964.

11 A.B. Paalman-de Miranda. *Topological semigroups.* 1964.

12 J.A.Th.M. van Berckel, H. Brandt Corstius, R.J. Mokken, A. van Wijngaarden. *Formal properties of newspaper Dutch.* 1965.

13 H.A. Lauwerier. *Asymptotic expansions.* 1966, out of print; replaced by MCT 54.

14 H.A. Lauwerier. *Calculus of variations in mathematical physics.* 1966.

15 R. Doornbos. *Slippage tests.* 1966.

16 J.W. de Bakker. *Formal definition of programming languages with an application to the definition of ALGOL 60.* 1967.

17 R.P. van de Riet. *Formula manipulation in ALGOL 60, part 1.* 1968.

18 R.P. van de Riet. *Formula manipulation in ALGOL 60, part 2.* 1968.

19 J. van der Slot. *Some properties related to compactness.* 1968.

20 P.J. van der Houwen. *Finite difference methods for solving partial differential equations.* 1968.

21 E. Wattel. *The compactness operator in set theory and topology.* 1968.

22 T.J. Dekker. *ALGOL 60 procedures in numerical algebra, part 1.* 1968.

23 T.J. Dekker, W. Hoffmann. *ALGOL 60 procedures in numerical algebra, part 2.* 1968.

24 J.W. de Bakker. *Recursive procedures.* 1971.

25 E.R. Paërl. *Representations of the Lorentz group and projective geometry.* 1969.

26 European Meeting 1968. *Selected statistical papers, part I.* 1968.

27 European Meeting 1968. *Selected statistical papers, part II.* 1968.

28 J. Oosterhoff. *Combination of one-sided statistical tests.* 1969.

29 J. Verhoeff. *Error detecting decimal codes.* 1969.

30 H. Brandt Corstius. *Exercises in computational linguistics.* 1970.

31 W. Molenaar. *Approximations to the Poisson, binomial and hypergeometric distribution functions.* 1970.

32 L. de Haan. *On regular variation and its application to the weak convergence of sample extremes.* 1970.

33 F.W. Steutel. *Preservation of infinite divisibility under mixing and related topics.* 1970.

34 I. Juhász, A. Verbeek, N.S. Kroonenberg. *Cardinal functions in topology.* 1971.

35 M.H. van Emden. *An analysis of complexity.* 1971.

36 J. Grasman. *On the birth of boundary layers.* 1971.

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.

38 W.A. Verloren van Themaat. *Automatic analysis of Dutch compound words.* 1972.

39 H. Bavinck. *Jacobi series and approximation.* 1972.

40 H.C. Tijms. *Analysis of (s,S) inventory models.* 1972.

41 A. Verbeek. *Superextensions of topological spaces.* 1972.

42 W. Vervaat. *Success epochs in Bernoulli trials (with applications in number theory).* 1972.

43 F.H. Ruymgaart. *Asymptotic theory of rank tests for independence.* 1973.

44 H. Bart. *Meromorphic operator valued functions.* 1973.

45 A.A. Balkema. *Monotone transformations and limit laws.* 1973.

46 R.P. van de Riet. *ABC ALGOL, a portable language for formula manipulation systems, part 1: the language.* 1973.

47 R.P. van de Riet. *ABC ALGOL, a portable language for formula manipulation systems, part 2: the compiler.* 1973.

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.

49 H. Kok. *Connected orderable spaces.* 1974.

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.

51 A. Hordijk. *Dynamic programming and Markov potential theory.* 1974.

52 P.C. Baayen (ed.). *Topological structures.* 1974.

53 M.J. Faber. *Metrizability in generalized ordered spaces.* 1974.

54 H.A. Lauwerier. *Asymptotic analysis, part 1.* 1974.

55 M. Hall, Jr., J.H. van Lint (eds.). *Combinatorics, part 1: theory of designs, finite geometry and coding theory.* 1974.

56 M. Hall, Jr., J.H. van Lint (eds.). *Combinatorics, part 2: graph theory, foundations, partitions and combinatorial geometry.* 1974.

57 M. Hall, Jr., J.H. van Lint (eds.). *Combinatorics, part 3: combinatorial group theory.* 1974.

58 W. Albers. *Asymptotic expansions and the deficiency concept in statistics.* 1975.

59 J.L. Mijnheer. *Sample path properties of stable processes.* 1975.

60 F. Göbel. *Queueing models involving buffers.* 1975.

63 J.W. de Bakker (ed.). *Foundations of computer science.* 1975.

64 W.J. de Schipper. *Symmetric closed categories.* 1975.

65 J. de Vries. *Topological transformation groups, 1: a categorical approach.* 1975.

66 H.G.J. Pijls. *Logically convex algebras in spectral theory and eigenfunction expansions.* 1976.

68 P.P.N. de Groen. *Singularly perturbed differential operators of second order.* 1976.

69 J.K. Lenstra. *Sequencing by enumerative methods.* 1977.

70 W.P. de Roever, Jr. *Recursive program schemes: semantics and proof theory.* 1976.

71 J.A.E.E. van Nunen. *Contracting Markov decision processes.* 1976.

72 J.K.M. Jansen. *Simple periodic and non-periodic Lamé functions and their applications in the theory of conical waveguides.* 1977.

73 D.M.R. Leivant. *Absoluteness of intuitionistic logic.* 1979.

74 H.J.J. te Riele. *A theoretical and computational study of generalized aliquot sequences.* 1976.

75 A.E. Brouwer. *Treelike spaces and related connected topological spaces.* 1977.

76 M. Rem. *Associons and the closure statement.* 1976.

77 W.C.M. Kallenberg. *Asymptotic optimality of likelihood ratio tests in exponential families.* 1978.

78 E. de Jonge, A.C.M. van Rooij. *Introduction to Riesz spaces.* 1977.

79 M.C.A. van Zuijlen. *Emperical distributions and rank statistics.* 1977.

80 P.W. Hemker. *A numerical study of stiff two-point boundary problems.* 1977.

81 K.R. Apt, J.W. de Bakker (eds.). *Foundations of computer science II, part 1.* 1976.

82 K.R. Apt, J.W. de Bakker (eds.). *Foundations of computer science II, part 2.* 1976.

83 L.S. van Benthem Jutting. *Checking Landau's "Grundlagen" in the AUTOMATH system.* 1979.

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.

85 J. van Mill. *Supercompactness and Wallman spaces.* 1977.

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.

88 A. Schrijver. *Matroids and linking systems.* 1977.

89 J.W. de Roever. *Complex Fourier transformation and analytic functionals with unbounded carriers.* 1978.

90 L.P.J. Groenewegen. *Characterization of optimal strategies in dynamic games.* 1981.

91 J.M. Geysel. *Transcendence in fields of positive characteristic.* 1979.

92 P.J. Weeda. *Finite generalized Markov programming.* 1979.

93 H.C. Tijms, J. Wessels (eds.). *Markov decision theory.* 1977.

94 A. Bijlsma. *Simultaneous approximations in transcendental number theory.* 1978.

95 K.M. van Hee. *Bayesian control of Markov chains.* 1978.

96 P.M.B. Vitányi. *Lindenmayer systems: structure, languages, and growth functions.* 1980.

97 A. Federgruen. *Markovian control problems; functional equations and algorithms.* 1984.

98 R. Geel. *Singular perturbations of hyperbolic type.* 1978.

99 J.K. Lenstra, A.H.G. Rinnooy Kan, P. van Emde Boas (eds.). *Interfaces between computer science and operations research.* 1978.

100 P.C. Baayen, D. van Dulst, J. Oosterhoff (eds.). *Proceedings bicentennial congress of the Wiskundig Genootschap, part 1.* 1979.

101 P.C. Baayen, D. van Dulst, J. Oosterhoff (eds.). *Proceedings bicentennial congress of the Wiskundig Genootschap, part 2.* 1979.

102 D. van Dulst. *Reflexive and superreflexive Banach spaces.* 1978.

103 K. van Harn. *Classifying infinitely divisible distributions by functional equations.* 1978.

104 J.M. van Wouwe. *Go-spaces and generalizations of metrizability.* 1979.

105 R. Helmers. *Edgeworth expansions for linear combinations of order statistics.* 1982.

106 A. Schrijver (ed.). *Packing and covering in combinatorics.* 1979.

107 C. den Heijer. *The numerical solution of nonlinear operator equations by imbedding methods.* 1979.

108 J.W. de Bakker, J. van Leeuwen (eds.). *Foundations of computer science III, part 1.* 1979.

109 J.W. de Bakker, J. van Leeuwen (eds.). *Foundations of computer science III, part 2.* 1979.

110 J.C. van Vliet. *ALGOL 68 transput, part I: historical review and discussion of the implementation model.* 1979.

111 J.C. van Vliet. *ALGOL 68 transput, part II: an implementation model.* 1979.

112 H.C.P. Berbee. *Random walks with stationary increments and renewal theory.* 1979.

113 T.A.B. Snijders. *Asymptotic optimality theory for testing problems with restricted alternatives.* 1979.

114 A.J.E.M. Janssen. *Application of the Wigner distribution to harmonic analysis of generalized stochastic processes.* 1979.

115 P.C. Baayen, J. van Mill (eds.). *Topological structures II, part 1.* 1979.

116 P.C. Baayen, J. van Mill (eds.). *Topological structures II, part 2.* 1979.

117 P.J.M. Kallenberg. *Branching processes with continuous state space.* 1979.

118 P. Groeneboom. *Large deviations and asymptotic efficiencies.* 1980.

119 F.J. Peters. *Sparse matrices and substructures, with a novel implementation of finite element algorithms.* 1980.

120 W.P.M. de Ruyter. *On the asymptotic analysis of large-scale ocean circulation.* 1980.

121 W.H. Haemers. *Eigenvalue techniques in design and graph theory.* 1980.

122 J.C.P. Bus. *Numerical solution of systems of nonlinear equations.* 1980.

123 I. Yuhász. *Cardinal functions in topology - ten years later.* 1980.

124 R.D. Gill. *Censoring and stochastic integrals.* 1980.

125 R. Eising. *2-D systems, an algebraic approach.* 1980.

126 G. van der Hoek. *Reduction methods in nonlinear programming.* 1980.

127 J.W. Klop. *Combinatory reduction systems.* 1980.

128 A.J.J. Talman. *Variable dimension fixed point algorithms and triangulations.* 1980.

129 G. van der Laan. *Simplicial fixed point algorithms.* 1980.

130 P.J.W. ten Hagen, T. Hagen, P. Klint, H. Noot, H.J. Sint, A.H. Veen. *ILP: intermediate language for pictures.* 1980.

131 R.J.R. Back. *Correctness preserving program refinements: proof theory and applications.* 1980.

132 H.M. Mulder. *The interval function of a graph.* 1980.

133 C.A.J. Klaassen. *Statistical performance of location estimators.* 1981.

134 J.C. van Vliet, H. Wupper (eds.). *Proceedings international conference on ALGOL 68.* 1981.

135 J.A.G. Groenendijk, T.M.V. Janssen, M.J.B. Stokhof (eds.). *Formal methods in the study of language, part I.* 1981.

136 J.A.G. Groenendijk, T.M.V. Janssen, M.J.B. Stokhof (eds.). *Formal methods in the study of language, part II.* 1981.

137 J. Telgen. *Redundancy and linear programs.* 1981.

138 H.A. Lauwerier. *Mathematical models of epidemics.* 1981.

139 J. van der Wal. *Stochastic dynamic programming, successive approximations and nearly optimal strategies for Markov decision processes and Markov games.* 1981.

140 J.H. van Geldrop. *A mathematical theory of pure exchange economies without the no-critical-point hypothesis.* 1981.

141 G.E. Welters. *Abel-Jacobi isogenies for certain types of Fano threefolds.* 1981.

142 H.R. Bennett, D.J. Lutzer (eds.). *Topology and order structures, part 1.* 1981.

143 J.M. Schumacher. *Dynamic feedback in finite- and infinite-dimensional linear systems.* 1981.

144 P. Eijgenraam. *The solution of initial value problems using interval arithmetic; formulation and analysis of an algorithm.* 1981.

145 A.J. Brentjes. *Multi-dimensional continued fraction algorithms.* 1981.

146 C.V.M. van der Mee. *Semigroup and factorization methods in transport theory.* 1981.

147 H.H. Tigelaar. *Identification and informative sample size.* 1982.

148 L.C.M. Kallenberg. *Linear programming and finite Markovian control problems.* 1983.

149 C.B. Huijsmans, M.A. Kaashoek, W.A.J. Luxemburg, W.K. Vietsch (eds.). *From A to Z, proceedings of a symposium in honour of A.C. Zaanen.* 1982.

150 M. Veldhorst. *An analysis of sparse matrix storage schemes.* 1982.

151 R.J.M.M. Does. *Higher order asymptotics for simple linear rank statistics.* 1982.

152 G.F. van der Hoeven. *Projections of lawless sequences.* 1982.

153 J.P.C. Blanc. *Application of the theory of boundary value problems in the analysis of a queueing model with paired services.* 1982.

154 H.W. Lenstra, Jr., R. Tijdeman (eds.). *Computational methods in number theory, part I.* 1982.

155 H.W. Lenstra, Jr., R. Tijdeman (eds.). *Computational methods in number theory, part II.* 1982.

156 P.M.G. Apers. *Query processing and data allocation in distributed database systems.* 1983.

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.

158 J.W. de Bakker, J. van Leeuwen (eds.). *Foundations of computer science IV, distributed systems, part 1.* 1983.

159 J.W. de Bakker, J. van Leeuwen (eds.). *Foundations of computer science IV, distributed systems, part 2.* 1983.

160 A. Rezus. *Abstract AUTOMATH.* 1983.

161 G.F. Helminck. *Eisenstein series on the metaplectic group, an algebraic approach.* 1983.

162 J.J. Dik. *Tests for preference.* 1983.

163 H. Schippers. *Multiple grid methods for equations of the second kind with applications in fluid mechanics.* 1983.

164 F.A. van der Duyn Schouten. *Markov decision processes with continuous time parameter.* 1983.

165 P.C.T. van der Hoeven. *On point processes.* 1983.

166 H.B.M. Jonkers. *Abstraction, specification and implementation techniques, with an application to garbage collection.* 1983.

167 W.H.M. Zijm. *Nonnegative matrices in dynamic programming.* 1983.

168 J.H. Evertse. *Upper bounds for the numbers of solutions of diophantine equations.* 1983.

169 H.R. Bennett, D.J. Lutzer (eds.). *Topology and order structures, part 2.* 1983.